DOCTORAL THESIS

Unification of Decision Support Techniques: Mitigating Statistical Paradoxes for Enabling Trustworthy Decision Making

Rahul Sharma

TALLINNA TEHNIKAÜLIKOOL
TALLINN UNIVERSITY OF TECHNOLOGY
TALLINN 2023

TALLINN UNIVERSITY OF TECHNOLOGY
DOCTORAL THESIS
49/2023

Unification of Decision Support Techniques: Mitigating Statistical Paradoxes for Enabling Trustworthy Decision Making

RAHUL SHARMA

TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies
Department of Software Science

The dissertation was accepted for the defence of the degree of Doctor of Philosophy (Computer Science) on 2 October 2023

Supervisor: Prof. Dr. Dirk Draheim,
Information Systems Group
Department of Software Science
School of Information Technologies
Tallinn University of Technology
Tallinn, Estonia

Opponents: Dr. Divesh Srivastava,
Database Research Department
AT&T Labs
New Jersey, USA

Professor Ladjel Bellatreche, Ph.D., Ph.D.,
Laboratoire d’Informatique et d’Automatique pour les Systèmes
École Nationale Supérieure de Mécanique et d’Aérotechnique
Poitiers, France

Defence of the thesis: 30 October 2023, Tallinn

Declaration:
Hereby, I declare that this doctoral thesis, my original investigation and achievement, submitted for the doctoral degree at Tallinn University of Technology, has not been submitted for any academic degree elsewhere

Rahul Sharma ______________________
signature

ISSN 2585-6898 (publication)
ISBN 978-9916-80-055-3 (publication)
ISSN 2585-6901 (PDF)
ISBN 978-9916-80-056-0 (PDF)
Printed by Koopia Niini & Rauam

TALLINNA TEHNIKAÜLIKOOL
DOKTORITÖÖ
49/2023

Otsuste toetamise tehnikate ühtlustamine: statistiliste paradokside mõju maandamine usaldusväärsete otsuste tegemise võimaldamiseks

RAHUL SHARMA

TAL TECH
KIRJASTUS

Contents

List of Publications

The list of the author's publications. These publications serve as the foundation for the research presented in this thesis.

I R. Sharma, M. Kaushik, S. A. Peious, S. B. Yahia, and D. Draheim. Expected vs. unexpected: Selecting right measures of interestingness. In M. Song, I.-Y. Song, G. Kotsis, A. M. Tjoa, and I. Khalil, editors, *Proceedings of DaWaK 2020 – the 22nd International Conference on Big Data Analytics and Knowledge Discovery*, pages 38–47, Cham, 2020. Springer International Publishing

II R. Sharma, M. Kaushik, S. A. Peious, A. Bazin, S. A. Shah, I. Fister, S. B. Yahia, and D. Draheim. A novel framework for unification of association rule mining, online analytical processing and statistical reasoning. *IEEE Access*, 10:12792–12813, 2022

III R. Sharma, M. Kaushik, S. A. Peious, M. Shahin, A. S. Yadav, and D. Draheim. Towards unification of statistical reasoning, OLAP and association rule mining: Semantics and pragmatics. In A. Bhattacharya, J. Lee Mong Li, D. Agrawal, P. K. Reddy, M. Mohania, A. Mondal, V. Goyal, and R. Uday Kiran, editors, *Proceedings of DASFAA 2022 – the 27th International Conference on Database Systems for Advanced Applications*, pages 596–603, Cham, 2022. Springer International Publishing

IV R. Sharma, H. Garayev, M. Kaushik, S. A. Peious, P. Tiwari, and D. Draheim. Detecting Simpson's paradox: A machine learning perspective. In C. Strauss, A. Cuzzocrea, G. Kotsis, A. M. Tjoa, and I. Khalil, editors, *Proceedings of DEXA 2022 – the 33rd International Conference on Database and Expert Systems Applications*, pages 323–335, Cham, 2022. Springer International Publishing

V R. Sharma, M. Kaushik, S. A. Peious, M. Bertl, A. Vidyarthi, A. Kumar, and D. Draheim. Detecting Simpson's paradox: A step towards fairness in machine learning. In S. Chiusano, T. Cerquitelli, R. Wrembel, K. Nørvåg, B. Catania, G. Vargas-Solar, and E. Zumpano, editors, *Proceedings of ADBIS 2022 – the 26th International Conference on New Trends in Database and Information Systems*, pages 67–76, Cham, 2022. Springer International Publishing

VI R. Sharma, M. Kaushik, S. A. Peious, M. Shahin, A. Vidyarthi, P. Tiwari, and D. Draheim. Why not to trust big data: Discussing statistical paradoxes. In U. K. Rage, V. Goyal, and P. K. Reddy, editors, *Proceedings of DASFAA 2022 International Workshops – the 27th International Conference on Database Systems for Advanced Applications*, pages 50–63, Cham, 2022. Springer International Publishing

VII R. Sharma, M. Kaushik, S. A. Peious, M. Shahin, A. Vidyarthi, and D. Draheim. Existence of the Yule-Simpson effect: An experiment with continuous data. In *Proceedings of Confluence 2022 – the 12th International Conference on Cloud Computing, Data Science & Engineering*, pages 351–355, 2022

VIII R. Sharma. On statistical paradoxes and overcoming the impact of bias in expert systems: towards fair and trustworthy decision making. *SSRN*, pages 1–37, July 2023. doi:10.2139/ssrn.4506432

Author's Contributions to the Publications

I **First Author:** The author conducted a preliminary study on association rule mining and measures of interestingness, prepared the figures, and wrote the manuscript.

II **First Author:** The author defined the research problem, designed and carried out experimentation, analyzed the results, prepared the figures, and wrote the manuscript.

III **First Author:** The author identified the research problem, designed mathematical formation and semantic correspondences between decision support techniques, prepared figures and wrote the manuscript.

IV **First Author:** The author identified the research problem and extensively discussed the identification of statistical paradoxes, particularly Simpson's Paradox, in categorical datasets. Subsequently, the author designed and executed experiments, generated informative figures and meticulously composed the manuscript.

V **First Author:** The author identified the research problem and thoroughly explored the identification of Simpson's Paradox in continuous datasets. Subsequently, the author designed and executed experiments, created illustrative figures and authored the manuscript.

VI **First Author:** The author emphasized the research problems, conducted a preliminary study, and performed an in-depth analysis of Simpson's Paradox in a case study. Additionally, the author designed and executed experiments, prepared figures, and composed the manuscript.

VII **First Author:** The author highlighted the research problem and demonstrated the role of statistical paradoxes using several real-world and synthetic datasets. The author prepared figures and wrote the manuscript.

VIII **First Author:** The author adeptly formulated the research problem, designed and executed the experimentation, developed an application, analyzed the results, prepared the figures and wrote the manuscript.

Abbreviations

ACIF	All combination Influencing Factor
AI	Artificial Intelligence
ARM	Association Rule Mining
CEVs	Conditional Expected Values
CFQs	Constrained Frequent Set
CAP	Consistency, Availability and Partition Tolerance
CBR	Case-Based Reasoning
DIRECT	Discovering and Reconciling Conflicts
DMDSS	Data Mining Decision Support System
DST	Decision Support Technique
DSS	Decision Support System
FIA	Freedom of Information Act
GBP	Grid Based Pruning
GUI	Graphical User Interface
HPC	High Performance Computing
IADSS	Intelligent Agent Assisted DSS
IDSS	Integrated Decision Support System
KDD	Knowledge Discovery in Databases
ML	Machine Learning
MSMiner	Multi-strategy Data Mining Platform
NJ	New Jersey
OLAP	Online Analytical Processing
PVM	Parallel Virtual Machine
SAS	Statistical Software Suite
SAP	System Applications and Products in Data Processing
SAP-BW	SAP Business Warehouse
SPSS	Statistical Package for the Social Sciences
SR	Statistical Reasoning
uARMSolver	universal Association Rule Mining Solver
UDS	User Defined Dataset

Terms

Decision support technique

A *decision support technique* (DST) is an approach that assists individuals and organizations in making informed decisions. As such, a decision support technique is typically based on a specific original rationale and consists of a specific data model, specific measures and specific algorithms. Typical examples of DSTs are Online Analytical Processing (OLAP) (rationale: interactive analysis of dependencies in data; relational data model), association rule mining (ARM) (rationale: automatic reporting of dependencies in data; data model: proprietary frequent itemset apparatus) and statistical reasoning (SR) (rational: statistical reasoning *per se*; data model: random variables).

Decision support tool

A *decision support tool* is a software application or system that aids individuals or organizations in the decision-making process.

Statistical paradox

A *statistical paradox* is a phenomenon where a trend or relationship observed in data appears to contradict common sense, i.e., uninformed human expectation or prediction based on the same data, for example, when a trend or relationship observed within subgroups of data vanishes or reverses when the subgroups are combined or when additional variables are considered.

Summary

This thesis presents a comprehensive compilation of eight published research articles, summarizing their key findings, methodologies, and contributions. Copies of these articles are included in the appendix for easy access and further exploration.

The thesis is structured as follows. Section 1 provides an introduction and overview of the research problem's relevance and the underlying motivation for this research. Subsequently, in Section 2, the research aim, focus and specific research questions (RQs) are thoroughly discussed, offering readers a clear understanding of the research direction and objectives. Section 3 offers an overview of existing literature and related research, setting the context for the current research. The research methodologies used throughout the thesis are summarized in Section 4, providing readers with an understanding of the research approach. Furthermore, the major contributions of this dissertation are highlighted and discussed in Section 5. In Section 6, the framework for the unification of decision support techniques (DSTs), the framework for mitigating the impact of bias resulting from statistical paradoxes, and a web-based application developed during the research are presented. The research results are then presented in Section 7, demonstrating their contribution to evaluating the artifacts created during the study. Moreover, Section 8 outlines the future direction of this research, suggesting potential areas for further exploration and improvement. Finally, the thesis concludes in Section 9 with a concise summary, emphasizing the key findings and concluding the research.

1 Introduction

1.1 Problem Relevance

Since the 17th century, statistical reasoning (SR) has played a crucial role in deriving valuable insights from data [85]. It has been an integral part of various decision support tools like SPSS (Statistical Package for the Social Sciences) and SAS (Statistical Analysis System) [85, 54, 67]. However, the advent of the information technology revolution in the 1990s brought forth a diverse range of powerful decision support techniques (DSTs) [14]. Each of these techniques possesses its own rationale, objectives, and perspectives and has gained significant importance in both research and practice. Presently, many popular DSTs such as association rule mining (ARM) [1, 2, 33], online analytical processing (OLAP) [15], and decision trees [62] are extensively utilized in research and practice for data mining, busi-ness intelligence, and machine learning (ML).

From a technical standpoint, these DSTs have some common objectives. However, they are designed and developed with distinct business perspectives, incorporating a variety of independently developed methods and algorithms, each with its own mathematical formalizations and algorithms.

Therefore, the variations in methodologies, terminology, and data representation systematically construct barriers that restrict the smooth transfer of results and insights across different domains. In particular, unlocking the potential of statistical results in their application scenarios is often challenging. Henceforth, we refer to these fundamental inconsistencies between DSTs as artificial gaps.

Addressing the artificial gaps between DSTs requires bridging the conceptual and methodological differences between DSTs. This can involve developing interoperability frameworks, creating common data representations, establishing semantic correspondences, and devising techniques for translating results and insights across different DSTs. By overcoming these gaps, we can unlock the synergistic benefits of integrating various DSTs, leading to a unified decision support system (DSSs) [30].

On the other hand, the emergence of various DSTs has overlooked fundamental statistical challenges discussed and identified centuries ago in statistics. These challenges include the identification and mitigation of confounding effects and statistical paradoxes [60, 91, 57, 7]. The primary objective of DSTs is to assist decision-makers by offering unbiased and reliable data-driven recommendations. However, the existence of confounders and statistical paradoxes in benchmark and real-life datasets can easily drive a decision-support tool to generate biased outcomes.

In statistics, the discussion on bias, confounders and statistical paradoxes is not new; they have been discussed for centuries, such as Simpson's paradox [91, 83], Berkson's paradox [7], and Lord's paradox [87]. These paradoxes are extreme cases of confounding that challenge common assumptions and can lead to surprising conclusions. Moreover, their effects are not only limited to the outcome of DSTs; they can have significant effects in various domains that involve data analysis. These effects can be attributed to various factors such as measurement errors, confounding variables, disparities in data distribution, and non-linear relationships. Therefore, in DSTs, it is essential to consider these factors when analyzing data to avoid the adverse impacts of paradoxes.

To address these two significant and distinct challenges in DSTs, this dissertation aims to strengthen DSTs to develop unified, fair, and trustworthy decision-support applications. In particular, to address the first challenge in DSTs, this research provides an elaboration of semantic correspondences between the three popular DSTs, i.e., ARM, OLAP, and SR. Subsequently, a novel framework for the unification of SR, OLAP, and ARM is suggested.

Further, to address the second research challenge, the dissertation emphasizes addressing the challenges posed by statistical paradoxes in DSTs. To achieve this, the thesis aims to discuss several measures to handle confounding effects and deal with the severe impacts of statistical paradoxes in DSTs. Next, the author suggests a framework for mitigating bias in training datasets. To provide evidence for the relevance of such a framework, the author conducts a series of experiments with three different measures on multiple real-world and benchmark datasets. To showcase the practical effectiveness of the proposed framework, the author has created a user-friendly web-based application. This application not only incorporates the example measures the author discussed but also integrates them into the outlined framework for bias mitigation. The author's assertion is that this application can be an invaluable tool for data scientists and researchers, as it has the capability to detect and address confounding effects automatically. The author contends that the suggested framework and application hold significant potential for future extensions beyond their current scope of application.

In light of the problem's relevance, the next two sections, 1.2 and 1.3, provide a detailed explanation of the two primary research areas.

1.2 On Establishing Semantic Correspondences Between DSTs

Integration of different DSTs to develop unified decision support tools has been discussed by several researchers. In 1997, Kamber et al. [38] discussed the integration of two important data mining techniques, i.e., OLAP and ARM, and referred to it as metarule-guided mining, which involved the use of pre-defined rule templates that are customizable by the user. This approach aimed to streamline the data mining process and increase efficiency by providing a framework for users to follow. Later, Han et al. [32] proposed DB-Miner for interactive mining, which provides a wide range of data mining operations such as association, generalization, characterization, classification, and prediction. In 2002, Imielinski et al. [36] presented cubegrades, which represents a generalization of association rules. Cubegrades showcase the impact of specializing (rolldown), generalizing (roll-up), and mutating (changing the dimensions) on a set of measures or aggregates within a cube. In Zhu's work [92], an innovative approach called online analytical mining of association rules was proposed. The research presented an algorithm for conducting inter-dimensional ARM, intra-dimensional ARM, and hybrid ARM. These techniques enable the discovery of association rules across different dimensions of data, both within a single dimension and between multiple dimensions. Building on the foundation of OLAP technologies, Zhu also designed a method for performing multi-level ARM. This approach allows for the exploration and extraction of association rules at multiple levels of granularity within a hierarchical structure, enabling a more comprehensive analysis of the data. Moreover, the potential application of association rules has recently been discussed with "cryptocurrency blockchain data" by [49].

Despite the existing research in this area, one notable gap in the state-of-the-art is the lack of elaboration on the concept of semantic correspondences between DSTs. While there are extensive studies on integrating DSTs, the specific notion of semantic correspondences, which aims to establish meaningful connections and relationships between different DSTs, remains underdeveloped. To address this gap, this thesis aims to delve into the concept of semantic correspondences between the three DSTs, i.e., ARM, OLAP and SR. Furthermore, it can enable decision-makers to work with cross-platform decision support tools and check their results from different viewpoints.

In pursuit of establishing semantic mappings between DSTs, the author places significant emphasis on probability theory. Specifically, the author focuses on conditional expected values (CEVs), which play a central role in our consideration as they correspond to *sliced average aggregates* in the context of OLAP. Additionally, they have the potential to correspond to *ratio-scale confidences* in a generalized ARM setting [20]. With a solid understanding of the semantic correspondences between the three DSTs, the author firmly believes in the potential to design highly beneficial next-generation features for advanced decision-support tools. By establishing meaningful connections and correspondences between DSTs, it will be easy to unlock new possibilities and capabilities that enhance the effectiveness and utility of DSSs.

1.3 Addressing the Severe Impact of Statistical Paradoxes in DSTs

In the realm of DSTs, it is crucial to recognize that biased results can arise due to the presence of confounding variables within the data. The implications of statistical paradoxes and their influence on analyses have been extensively examined by renowned mathematicians and statisticians, as evidenced by notable works such as Yule, Pearl, and Berkson [91, 57, 7]. Statistical paradoxes, including Simpson's paradox and Berkson's paradox [83, 7], draw attention to the presence of confounding effects within data analysis. These effects arise when the association between two variables is distorted or modified by the influence of a third variable, known as a confounder. The presence of a confounder can create an illusion of association or lead to the misinterpretation of causal relationships. This highlights the importance of recognizing and accounting for confounding effects in DSTs to obtain trustworthy and meaningful conclusions.

For instance, an artificial intelligence-based recruitment tool utilized by Amazon [18] appears to have failed to assess candidates for software development roles in a manner that is free from gender bias. This is because the tool was trained using resumes from a time when the technology industry was predominantly male, leading to a bias towards male candidates. There are many similar examples [55, 37, 8] which clearly highlight the incompetence of DSTs and the need for addressing fundamental statistical challenges.

Various frameworks such as [90, 6, 9, 5, 66] and best practices exist to reduce bias in

ML, but they are often designed for specific types of bias, such as gender or racial bias and struggle with confounding effects and statistical paradoxes in classical DSTs. It is difficult for any framework to fully recognize and account for the effects of various statistical paradoxes.

To improve the trustworthiness of mainstream DSTs, it is important to effectively mitigate the severe impacts of confounding, causality, and statistical paradoxes. Addressing paradoxical outcomes and handling statistical paradoxes pose significant challenges for bias mitigation frameworks, particularly when working with different DSTs. Therefore, to overcome these challenges, this thesis highlights several methods to identify and adjust the impacts of statistical paradoxes and propose a comprehensive framework that specifically targets the mitigation of statistical paradoxes in DSTs.

2 Aim and Focus

Whilst there is a lot of discussion and research about developing advanced DSTs [93, 21, 89, 11], the notion of identifying semantic correspondences between DSTs and developing a common framework to utilize their varied features is not discussed much in state of the art. Next, the author has also identified a lack of attention and focus on improving the trustworthiness of DSTs. This means that limited efforts have been made to address and overcome the fundamental statistical challenges that have been discussed for centuries.

Based on these two major research gaps, this thesis aims to answer two primary RQs and six supplementary RQs as detailed in Section 2.1.

2.1 Research Questions

- RQ-1: How to bridge the artificial gaps between different DSTs?
  - RQ-1.1 What are the semantic correspondences between the three major decision support techniques, i.e., statistical reasoning (SR), online analytical processing (OLAP) and association rule mining (ARM)?
  - RQ-1.2 How to provide a systematic interpretation of results between different decision support techniques? In how far can we consider SR, OLAP, and ARM as synonymous?
  - RQ-1.3 How to develop a common framework for integrating SR, OLAP and ARM?
- RQ-2: How to systematically assess the impact of statistical paradoxes in multivariate data? How to utilize these assessments for better decision-making?
  - RQ-2.1 How to identify the existence of the Yule-Simpson effect in multivariate data?
  - RQ-2.2 How to adjust the impact of the Yule-Simpson effect in multivariate data?
  - RQ-2.3 How to develop a platform to handle statistical paradoxes in multivariate data and recommend appropriate adjustments for improved decision-making?

In this thesis, a comprehensive analysis of eight articles has been conducted to address two primary and six supplementary research questions. The articles have made substantial contributions to the existing knowledge in the field, thereby significantly contributing to the existing knowledge in the field.

The first publication [I], “Expected vs. Unexpected: Selecting Right Measures of Interestingness", provides a preliminary study to identify the most typical and useful roles of the measures of interestingness in ARM. The second publication [II], “A Novel Framework for Unification of Association Rule Mining, Online Analytical Processing and Statistical Reasoning", undertook an extensive analysis of various strategies aimed at bridging the gap between the three popular DSTs: SR, OLAP, and ARM. Our contribution lies in elaborating the semantic correspondences between the foundations of SR, OLAP and ARM, i.e., probability theory, relational algebra and the itemset apparatus, respectively. Next, the author has introduced a novel framework that aims to unify DSTs and developed a corresponding tool to validate this concept. This tool facilitates the unified utilization of DSTs within a conventional decision support process, offering clarity on how operations from SR, ARM, and OLAP can complement each other in enhancing data comprehension, data visualization, and decision-making processes.

The third publication [III], “Towards unification of statistical reasoning, OLAP and association rule mining: Semantics and pragmatics", strives to overcome the artificial gaps that exist between three DSTs: SR, OLAP and ARM. By establishing these semantic correspondences, the author proposes that the unification of DSTs can form the basis for designing advanced multi-paradigm data mining tools in the future.

The fourth publication [IV], “Detecting Simpson’s paradox: A machine learning perspective", centers on addressing a specific instance of a statistical paradox known as Simpson’s paradox in categorical data. Through real-world case studies, the author highlights the profound impact of this paradox. Furthermore, the author presents an algorithm designed to detect Simpson’s paradox and identify the confounding variables within categorical datasets.

The fifth Publication [V], “Detecting Simpson’s paradox: A step towards fairness in machine learning", discusses ways to identify the impact of Simpson’s paradox on linear trends, particularly in relation to continuous values. Subsequently, the author provides a practical demonstration of its effects using three benchmark training datasets commonly employed in ML.

The sixth publication [VI], “Why not to trust big data: Discussing statistical paradoxes", and the seventh publication [VII], “Existence of the Yule-Simpson effect: An experiment with continuous data" provide a preliminary study and further insights about different statistical paradoxes and emphasis on the statistical evaluation and human experts in the loop towards developing trustworthy DSTs.

The eighth publication [VIII], “On statistical paradoxes and overcoming the impact of bias in expert systems: towards fair and trustworthy decision making”, highlights the significance of addressing statistical paradoxes within DSTs and aims to contribute to the development of fair and reliable DSTs. To this end, a framework is proposed for mitigating the impact of statistical paradoxes in DSTs. Additionally, various measures for adjusting the influence of confounders are discussed. To validate the effectiveness and utility of the proposed framework, a web-based application has been developed. The current version of the application allows for the investigation of potential confounders by detecting instances of Simpson’s paradox and offers a feature for adjusted observations. In order to provide empirical evidence supporting the relevance of the framework and the application, a series of experiments have been conducted on both real-world and benchmark datasets.

Table 1 provides a concise summary of the author’s publications, mapping their respective research questions and contributions. Publication [I], [II], [III], contributes towards answering RQ1, RQ1.1 , RQ 1.2 and RQ 1.3. Rest five publications [IV], [V], [VI], [VII], [VIII]

Table 1: Mapping among RQs, publications and their contributions

Research Questions	*Publications*	*Contributions*
RQ1, RQ1.1	[I], [II], [III]	C1
RQ1.2	[II], [III]	C1
RQ1.3	[II]	C1, C2
RQ2, RQ2.1	[IV], [V], [VI], [VII], [VIII]	C3
RQ2.2	[VIII]	C4, C5, C6
RQ2.3	[VIII]	C4, C5, C6

are focused on answering the RQ2, RQ2.1, RQ2.2, and RQ2.3.

3 Related Research

The work presented in this section offers an overview of the related information gathered from the existing literature. It establishes the foundation by synthesizing and summarizing the current state of knowledge in the field. Moreover, it highlights the gaps and limitations in the existing literature that pertain to the research questions introduced in Section 2.1. By identifying these gaps, the thesis aims to contribute to the knowledge base in a meaningful way. It intends to address these research gaps by conducting original research and providing novel insights and solutions.

3.1 On the Unification of Decisions Support Techniques

The evolution of DSTs has been driven by advancements in technology, changes in business environments, and the growing complexity of decision-making processes. Over the years, DSTs have evolved from traditional manual approaches to more sophisticated and automated systems. However, some fundamental challenges are yet to be answered. This section provides an overview of the studies exploring the integration of DSTs. The author conducts a thorough examination of relevant research articles pertaining to the integration of various DSTs. Some of these research works are further discussed as follows:

Wang et al. [89] presented a new architecture that integrates knowledge discovery in databases (KDD) techniques into existing decision support systems (DSSs). The paper discusses integrating different techniques in group DSSs using three types of decision support agents. Rupnik et al. [65] developed a data mining decision support system (DMDSS) that combines classification, clustering, and association rules. Zhuang et al. [94] proposed a methodology that integrates data mining and case-based reasoning to develop a pathology test ordering system. Data mining is used to extract knowledge from past data and used in decision support. Liu et al. [50] conducted a survey in 2010 to assess the development of an integrated decision support system (IDSS). IDSS combines four DSTs: knowledge-based systems, data mining, intelligent agents, and web technology, to help users interpret decision alternatives and discover patterns in large data sets.

These works are mainly focused on developing and integrating DSSs with DSTs. However, none of them discuss semantic correspondences between DSTs. The author also examines the current state of the art for integrating OLAP and ARM, with some works concentrating on intra-dimensional association rules and others on inter-dimensional association rules.

Kamber et al. [38] made a notable contribution by addressing the relationship between ARM and OLAP. Their work introduced a meta-rule-guided mining approach specifically designed for extracting association rules from multi-dimensional data cubes. The author introduced four algorithms that efficiently analyze OLAP data cubes for discovering valuable patterns and associations within multi-dimensional datasets. Later, DBMiner was proposed by Han et al. [32] for interactive mining. It offers a broad range of data mining functions, including association, generalization, characterization, classification, and prediction. In 2002, Imielinski et al. [36] introduced cubegrades, a generalization of association rules. Cubegrades highlights the impact of roll-up, roll-down, and changing dimensions on measures or aggregates within a cube. Zhu [92] proposed online analytical mining of association rules and presented algorithms for inter-dimensional, intra-dimensional, and hybrid ARM. These techniques enable the discovery of association rules across different dimensions of data, both within a single dimension and between multiple dimensions. Building on the foundation of OLAP technologies, Zhu also designed a method for multilevel ARM, enabling a more comprehensive analysis of data.

This thesis provides an elaboration on the semantic correspondences between ARM, OLAP and SR with their foundations, i.e., itemset apparatus, relational algebra, and probability theory, respectively. The work presented in publications [II] and [III] also characterizes the degree and type of synonymity among SR, OLAP and ARM. To achieve semantic correspondences between SR and OLAP, the author compared the expected value of an item X, i.e., $\mathsf{E}(X)$, with the output of the AVG query in OLAP. The author demonstrates that the average of a random variable Y with condition X (Conditional Expected values) and the conditional average of an OLAP query provide the same outcome. The author also illustrated the high-level abstraction of the framework, which was also implemented as a tool that first recognizes different kinds of data (discretized, numerical, categorical) and then develops generalized association rules for the various combinations of influencing factors and target columns.

Additionally, the author has made significant contributions to related research, particularly in the areas of generalizing ARM to continuous values and Big Data. The contributions are supported by several research papers coauthored by the author, including systematic literature reviews on the potential and applications of NARM [43, 42, 40], analysis of human perceptions in discretizing numerical attributes [45, 44], impact-driven discretization of numerical factors, and utilization of swarm-intelligence algorithms for mining numerical association rules [41, 39]. Furthermore, the author also contributed as a coauthor to the research that explores the application of Big Data analytics in association rule mining and investigates factors in diverse datasets using cluster-based association rule mining techniques [73, 72]. These contributions collectively enhance the field of ARM, making it applicable to a wider range of data types and larger datasets [71].

3.2 Statistical Paradoxes

In DSTs, statistical paradoxes refer to situations that produce unexpected or counter-intuitive results that may not align with human expectations or common sense. These paradoxes can have a significant impact on any individual and organization. This section gives an overview of studies that investigate the existence and impact of statistical paradoxes in data. DSTs, such as ARM and OLAP, can produce biased results due to confounding variables in data [48, 86]. The role of statistical paradoxes and their impact has been discussed deeply in classical data analysis by expert mathematicians and statisticians [91, 57, 7]. Therefore, understanding causal relationships hand in hand with evaluating the existence of statistical paradoxes is an essential step forward toward developing trustworthy and fair DSTs.

Statistical paradoxes are fundamentally related to a range of various statistical concepts such as partial correlations [27], p-technique [13], suppressor variables [16], conditional independence [19], propensity score matching [64], causal inference [57, 59], and mediator variables [51] as well as statistical challenges such as ecological fallacy [63, 47] and Lord's paradox [87].

Statistical paradoxes, such as Simpson's paradox and Berkson's paradox, highlight the existence of confounding effects. These effects emerge when the relationship between two variables is distorted or altered due to the presence of a third variable, known as the confounder. Confounding effects can lead to misleading or counterintuitive conclusions if not properly addressed or accounted for in the statistical analysis. In mathematical statistics, causality and confounding are two related concepts that researchers widely discuss. This is evident in the works of established researchers such as Otte and Pearl [56, 58].

The work of Pearl [58, 57] has made a substantial impact on the advancement of probabilistic reasoning and causal modelling in the field of AI. Pearl's contributions include the development of an extensive framework for causal inference, which focuses on reasoning about causal relationships between variables. This framework provides valuable tools and methodologies for understanding cause-and-effect relationships, enabling researchers and practitioners to uncover the underlying mechanisms driving observed phenomena. Otte [56] discussed how probabilistic causality relates to Simpson's paradox. Otte's discussion revolves around the concept of probabilistic causality, which suggests that a cause doesn't always generate a unique effect but instead alters probabilities. Schield [69] proposed using Cornfield's conditions to assess the presence of confounding variables that affect both the dependent variable (target variable) and independent variables (impact factors). Spellman [84] gave an example of how different information can lead to different conclusions. Schaller [68] talked about how individuals form opinions and conclusions about others with only a small amount of information. In the work [19], the idea of conditional independence and how it affects statistical inference is presented and discussed. In Cartwright's study, the connection between scientific law and causal necessity in philosophy was investigated [12]. Fiedler discussed sampling issues, pseudo-contingencies, and inductive reasoning in social psychology, including cognitive consistency, social cognition, and implicit social cognition [23, 26, 24, 25]. Alipourfard et al. [3] found Simpson's paradox in social and behavioral data [4]. Blyth [10] discussed Simpson's paradox and the *sure-thing principle*, two essential concepts that can help decision-makers avoid incorrect decisions based on incomplete or misleading data. Hernán [34] provided examples illustrating Simpson's paradox across various contexts. They emphasized the significance of understanding confounding variables, selection bias, and effect modification to accurately interpret statistical results and draw conclusions. Kievit et al. [46] found Simpson's paradox in psychological science. They created an R package to detect confounding effects in continuous data. Freitas et al. [28] proposed an algorithm for detecting instances of Simpson's paradox. However, Curley et al. [17] explained the role of Simpson's paradox and its implications for decision-making. In 2010, Greenland [31] investigated the relationship between Simpson's paradox and Bayesian non-collapsibility. He used an example of adding constants in contingency tables. However, according to Tu et al. [87], various statistical paradoxes, such as Simpson's paradox, suppression effects and Lord's paradox, are all manifestations of the same phenomenon referred to as the reversal paradox.

Tu et al. [87] claimed that different statistical paradoxes, including Simpson's paradox, Lord's paradox, and suppression effects, are manifestations of the same phenomenon known as the reversal paradox.

During the investigation of frameworks aimed at mitigating bias, it became evident that most of these frameworks were primarily designed to address specific types of bias and lacked the capability to effectively handle confounding effects and statistical paradoxes within classical DSTs, such as ARM. Frameworks such as such as [90, 6, 9, 5, 66] have been developed to tackle specific biases, such as gender or racial bias. While these frameworks have made notable contributions in their respective domains, they often encounter challenges when it comes to recognizing and accounting for the effects of various statistical paradoxes.

The existing literature highlights that considerable discussions have contributed to advancing our understanding of these complex phenomena and their implications in data analysis and decision-making. However, this work highlights the importance of detecting statistical paradoxes in data. The author provided a detailed discussion of the impact of Simpson's paradox with the help of examples which are presented in [IV], [V], [VI], [VII]. This work presented in [VIII] utilizes two measures (one for continuous data and one for categorical data) for investigating confounders via detecting instances of Simpson's paradox in regard to the stratification of Pearson correlation and presents one measure for adjusting the impact of confounders, which generalizes standard back-door adjustment to continuous data.

4 Research Methodology

The primary methodological foundation of the research conducted in this thesis stems from the principles of design science research [35, 53, 52]. Design science research methodology is widely used in the fields of information systems and computer science to create innovative artifacts or design solutions that effectively tackle specific challenges or problems. This thesis makes six significant contributions in the form of distinct artifacts and methods. These contributions are specifically designed to answer the identified RQs and fill the existing technological and knowledge gaps. The development of these artifacts aims to strengthen existing DSTs to foster fair and trustworthy decision-making processes.

The central framework of this thesis is built upon a collection of eight research articles. These research articles contribute valuable insights and evidence to support the findings and conclusions presented in this thesis. Among these scholarly articles, one has been published in a high-impact Q1 journal, highlighting its significance and rigorous evaluation [II]. Additionally, another article is a technical report available on the esteemed social science research network repository (SSRN) [VIII]. The remaining six articles [I], [III], [IV], [V], [VII], [VI], are published in reputable conference proceedings, further emphasizing their scholarly contributions. The collection of these scholarly research articles forms a solid foundation of evidence and insights, lending strong support to the research findings and conclusions explained in this thesis.

In this thesis, the author analyzes the evaluation results, identifies strengths and weaknesses, and improves the outcomes. This leads to refining and improving the artifact or design solution based on the feedback received during the evaluation phase. By adopting the design science research methodology throughout the process, the author has applied relevant theories and methodologies to guide decision-making. The goal is to ensure the rigor and validity of the research while contributing to the advancement of knowledge and practice in the field.

The research design process used in this thesis is illustrated in Figure 1.

4.1 Building

In the research design, constructing and assessing iterations are essential aspects. Building iterations involve the iterative process of developing and refining artifacts [35]. The process begins by identifying a problem and comprehending existing literature, theories,

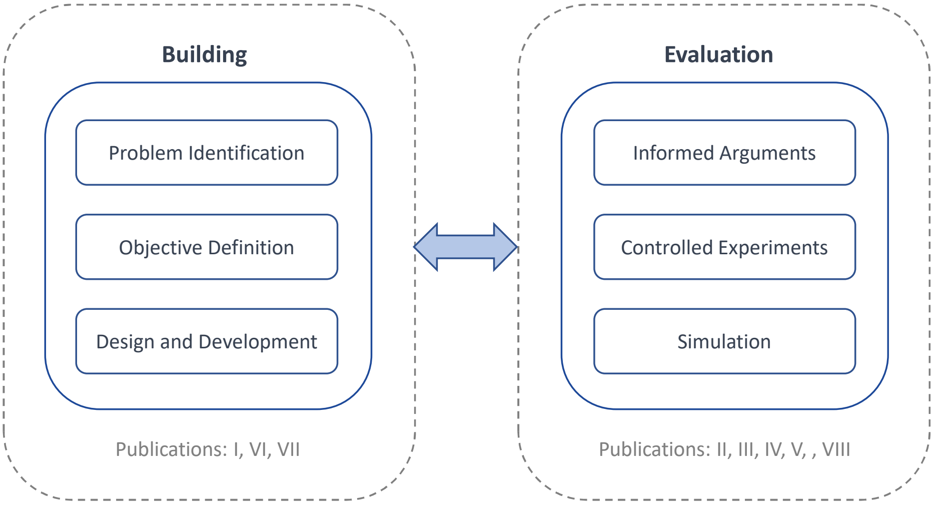


Figure 1: Design science process: The primary methodological foundation of the research from the principles of design science research [35, 53, 52]

and context. Subsequently, potential solutions are conceptualized, leading to the development and implementation of an artifact or prototype. Lastly, the solution's effectiveness is evaluated and validated through experiments, simulations, and informed arguments.

Towards building the artifacts, as per the design science research methodologies [35, 53, 52], the author has used the following three research methodologies for building the artifacts. The publications [I], [VI], and [VII] have been utilized in identifying the problem and defining the clear research objectives.

4.1.1 Problem Identification In design science research, the problem identification phase marks the initial step in the research process. By following this step, the author of the thesis meticulously identifies and defines a specific problem or challenge requiring attention. This critical phase involves a comprehensive examination of existing issues and gaps within the field of study. Through an in-depth literature review, the author gains a profound understanding of the current state of knowledge and pinpoints the precise research problem that the study aims to address and resolve. This phase lays the foundational groundwork for subsequent research activities, guiding the research direction and providing the context for further investigation and analysis.

4.1.2 Objective Definition After the problem identification phase, the author proceeds to define clear objectives for the study. These objectives outline the desired outcomes or results that the research intends to accomplish. The objectives provide guidance and serve as a basis for designing the artifacts or solutions in later phases of the research process. By aligning the objectives with the identified problem, the author ensures that the study remains focused and purposeful, ultimately contributing to the advancement of knowledge and addressing the research challenge effectively.

4.1.3 Design and Development In the design and development phase, the author takes on the critical task of designing and developing artifacts or solutions to effectively address the identified problem. This phase entails the creation of prototypes, models, algorithms, or frameworks that embody the author's proposed solution. Here, the design and development phase is an iterative process of refining and improving the artifacts based on evaluation and feedback. This iterative approach ensures that the artifacts continuously evolve to meet the desired outcomes and successfully address the identified problem.

4.2 Evaluation

This thesis extensively discusses the evaluation research methodologies based on design science research applied to the published research articles [35, 52, 53]. Detailed explanations of these methodologies are provided in Sections 4.2.1, 4.2.2, and 4.2.3. Through these sections, readers of the thesis can gain a comprehensive understanding of the evaluation methodologies employed in this research work.

4.2.1 Informed Arguments In the field of design science research, informed arguments play a crucial role in supporting the development and validation of innovative artifacts or design solutions [35]. These arguments are constructed upon a blend of theoretical foundations, empirical evidence, and logical reasoning, working together to enhance the credibility and persuasiveness of the research findings. By leveraging this comprehensive approach, we provide robust support for the development and validation of our design contributions.

In the publication [III], the author worked to answer RQ1 and provide an in-depth analysis of three different DSTs, i.e., SR, OLAP and ARM and investigate the semantic correspondence between them. Further, In the next article [II], the author introduces a novel framework aiming to unify SR, OLAP, and ARM. The primary objective of this framework is to provide an integrated approach to decision support by bringing together these technologies.

The publications [VI] and [VII] provide a preliminary investigation into statistical paradoxes, particularly Simpson's paradox, within DSTs, the author explores the role of Simpson's paradox in DSTs and establishes the foundation for the need to detect and address this paradox. The framework provided in [VIII] likely builds upon the preliminary study mentioned earlier, and it aims to provide a means to identify the impacts of statistical paradoxes toward developing fair and trustworthy DSTs.

4.2.2 Controlled Experiments Controlled experiments serve as a widely adopted research methodology in various fields, including design science research, to investigate cause-and-effect relationships and evaluate the effects of specific interventions or treatments [35]. In such experiments, researchers deliberately manipulate and control independent variables while observing and measuring the resulting impact on dependent variables. In this research, the primary objective of using controlled experiments as a research methodology is to control independent variables while measuring the effects on dependent variables. Ultimately, the goal is to establish causal relationships and draw valid conclusions about the effectiveness of the three discussed measures.

The publications [VIII] [V], [IV], delve into the exploration of various measures aimed at identifying and mitigating the impact of confounding variables in several benchmark and real-life datasets that contain both categorical and continuous variables. Through rigorous experimentation with different datasets, this research aimed to assess the effectiveness of these measures.

4.2.3 Simulation In design science research, simulation plays a pivotal role in thoroughly evaluating the functionality and effectiveness of specific artifacts or systems [35]. By employing synthetic data in a controlled and virtual environment, the author aims to gain a comprehensive understanding of the performance and capabilities of the artifacts under examination. This process allows for a detailed assessment of the artifact's or system's functionality and aids in informing further improvements and refinements.

The article [VIII] implemented a web-based application to detect the existence of confounding effects and the instances of Simpson's paradox. Further, three measures are discussed to identify and adjust the impact confounding variable in several benchmark datasets. According to the design science research methodology, the author accomplished the model development, implementation, experimentation and analyzing the outcome of the artifacts.

5 Contributions

The dissertation presents both conceptual and practical solutions, addressing two primary and six supplementary research questions, concluding with a total of six significant contributions. The mapping of contributions and their respective evaluation methodologies is given in Table 2.

The main contributions of this research can be summarized as follows:

- *C1:* Contribution (C1) highlighted a significant gap among the three most popular DSTs, namely SR, OLAP, and ARM. To address this gap, the author analyzed a range of approaches aimed at bridging the gap between the three DSTs.

  The author contributed by elaborating on the semantic correspondences between the foundations of SR, OLAP and ARM, i.e., probability theory, relational algebra and the itemset apparatus, respectively. The support of an itemset corresponds to the probability of a corresponding event and the confidence of an association rule corresponds to the conditional probability of two corresponding events. Furthermore, the OLAP average aggregate function corresponds to conditional expected values, which closes the loop between ARM, OLAP and probability theory with respect to the most important constructs in ARM and OLAP. By providing semantic mappings between three DSTs discussed in [III], [II] and [I], the author answers RQ1.1 fully and partially answers RQ1.2.

- *C2:* Contribution (C2) suggested a novel framework for the unification of DSTs and implemented a tool to validate the concept of unification. The tool provides unified usage of DSTs in a classical decision support process. It clarifies how far the operations of SR, ARM, and OLAP can complement each other in understanding data, data visualization and decision-making. The tool was developed on the basis of an open-source framework and tested with two real datasets and one synthetic dataset. The results and performance of the tool show valuable contributions toward developing the next-generation DSSs. The programming code and other instructions on how to use the proposed tool are available in the GitHub repository [1]. This particular contribution serves as a response to RQ1.3, as discussed in [II].

- *C3:* Contribution (C3) is the identification of confounding variables and instances of statistical paradoxes, i.e., Simpson's paradox, in multivariate datasets used in DSTs. With this, the author addresses the issue of statistical paradoxes in big data and

[1] https://github.com/rahulgla/unification

their implications on benchmark datasets commonly employed in ML. The author's investigation into Simpson's paradox includes an analysis of its presence and implications in various datasets through a case study and workshop paper published in conferences [VI] and [VII]. These publications provide partial answers to RQ2 and RQ2.1. Additionally, the author addresses the identification of confounding variables and Simpson's paradox instances in continuous and categorical datasets in two other articles [IV] and [V].

- *C4:* Contribution (C4) utilized stratification of Pearson correlation to identify potential confounders in categorical and continuous data. Furthermore, the author generalizes back-door adjustment techniques and uses propensity weighting to adjust the impact of confounders effectively [VIII]. This contribution enhances the understanding and application of confounding adjustments in continuous datasets.
- *C5:* Contribution (C5) introduced a novel framework designed to effectively address statistical paradoxes and confounding effects. The framework consists of three main components and two essential sub-components, visually presented in Figure 4. The framework has been carefully crafted and thoroughly evaluated to enhance accuracy and reliability, minimizing biased outcomes in DSTs. Additionally, the framework offers a promising approach for promoting fair and trustworthy decision-making processes in various domains. The publication [VIII] comprehensively presents the details of the framework and provides in-depth insights into its functionality. Additionally, it addresses and answers RQ2.2 and partially RQ2.3.
- *C6:* Contribution (C6) demonstrated the practical utility of the proposed framework. To showcase its effectiveness, the author has developed a user-friendly web-based application that seamlessly integrates the discussed example measures for bias mitigation. This application, presented in publication [VIII], serves as a valuable tool for data scientists and researchers, offering automated detection and mitigation of confounding effects. By providing a streamlined approach to address and overcome challenges in data analysis, the application makes a significant contribution to the field. The programming code and usage guidelines for the proposed tool can be found in the GitHub repository [2]. Additionally, with this, the author also answers RQ2.3.

6 The Contributed Frameworks and Application

6.1 Framework for the Unification of Decision Support Techniques [II]

In this section, the author introduces the framework for unifying three DSTs. Inspired by the data science process by O'Neil et al. [70], the proposed framework is designed with a modular approach, allowing each module to be interchangeable. Figure 2 provides a high-level abstraction of the framework, while a more detailed overview, based on the process of KDD [22], is illustrated in Figure 3.

The framework comprises seven major components. The graphical user interface (GUI) facilitates communication between decision-makers and the framework for processing raw data. Data pre-processing involves various operations and checks, such as discretization, data cleaning to identify corrupt data, reviewing data types, and transforming data into useful formats. The all-combination influencing factor (ACIF) generator is a tool within the framework that empowers decision-makers to select target columns and influencing

[2] https://github.com/rahul-sharmaa/SimpsonP

Table 2: Mapping among the contributions of the dissertation and corresponding evaluation methodologies

Contribution	Summary	Evaluation Methodology
C1	Establish semantic mappings between DSTs	Informed Arguments
C2	Provide framework for unification of SR, OLAP and ARM	Informed Arguments
C3	Present approaches for identifying the existence of confounding effects and well-known statistical paradoxes.	Informed Arguments
C4	Present measures for adjusting the impact of confounding effects	Controlled Experiment
C5	Provide a framework for mitigating the impact of bias resulting from statistical paradoxes	Informed Arguments
C6	Provide a web-based application to detect and adjust the impact of confounders	Simulations

factors, generating different combinations of data items. The decision support engine consists of multiple DSTs, allowing decision-makers to choose one or more techniques for data processing and gaining insights. Pattern evaluation utilizes different methods from SR, OLAP, and ARM to discover meaningful information. Lastly, the semantic mapper is a manual process that maps the results of DSTs and reports various semantic correspondences between them. A detailed description of the framework is provided in [II].

To illustrate the usefulness of the proposed framework, the author created a working instance [61] using ASP.NET, an open-source framework for web application development. This implementation serves as a prime example of a next-generation decision support tool, showcasing the successful adoption of the proposed framework. The programming code and detailed instructions on how to use this tool can be accessed in the GitHub repository [82].

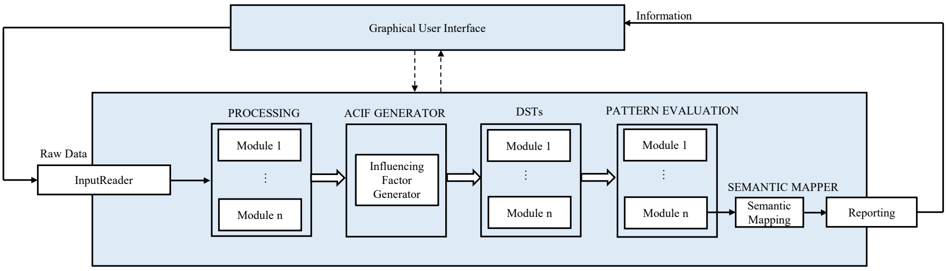


Figure 2: A top-level abstraction of the framework aiming to unify DSTs [II]

6.2 The Framework for Mitigating Bias in Training Datasets [VIII]

Addressing paradoxical outcomes and handling statistical paradoxes is indeed challenging for bias mitigation frameworks, especially when dealing with large and complex datasets. Successfully navigating these challenges necessitates a combination of technical expertise, domain knowledge, and thoughtful consideration of the trade-offs between fairness and accuracy. Improving existing frameworks requires a continuous focus on expanding

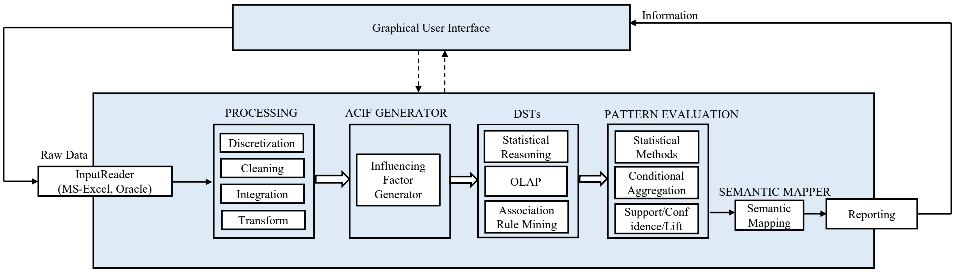


Figure 3: A detailed illustration of the framework aiming to unify DSTs [II]

coverage, enhancing flexibility, fostering collaboration with domain experts, offering guidance on balancing trade-offs, and increasing transparency. To address these challenges, the author proposes a comprehensive framework.

The framework consists of three main components and two sub-components, providing a robust approach to handle bias and statistical paradoxes effectively. This approach aims to promote fair and trustworthy decision-making in data analysis and mitigate the impact of biases in DSTs. Figure 4 provides a graphical representation of this proposed framework.

1. Data pre-processing: The first step involves identifying and removing any errors or inconsistencies in the data. This involves techniques such as data cleaning, normalization, dealing with missing values and outlier detection.

2. Bias mitigation techniques: The second step involves using various techniques to mitigate bias in the dataset. This involves techniques such as data augmentation, where new data is generated to balance the representation of different classes or categories in the data. Another technique uses weighting schemes to give more weight to underrepresented classes or categories.

3. Evaluation: The evaluation aims to ensure that the bias mitigation techniques effectively improve fairness and provide an accurate outcome. This could involve comparing the performance of a DST on biased and unbiased datasets by utilizing different metrics.

   (a) Incorporating domain knowledge: The step involves incorporating domain knowledge into the data analysis process. This involves using expert knowledge to guide the selection of relevant variables and features and using various adjustment techniques. The goal of incorporating domain knowledge is to understand the causes of several statistical paradoxes and take appropriate steps to adjust their impact and to further improve the quality and relevance of the dataset.

   (b) Adjustments in datasets: Uneven distribution of data between two or more groups is one of the reasons for bias. Therefore, by balancing the input variables across different groups in the data, an ML model is less likely to make biased decisions. Balancing the dataset ensures that an ML model is equally exposed to all groups.

The suggested framework is designed to show the following advantages. A detailed description of the framework is provided in [VIII].

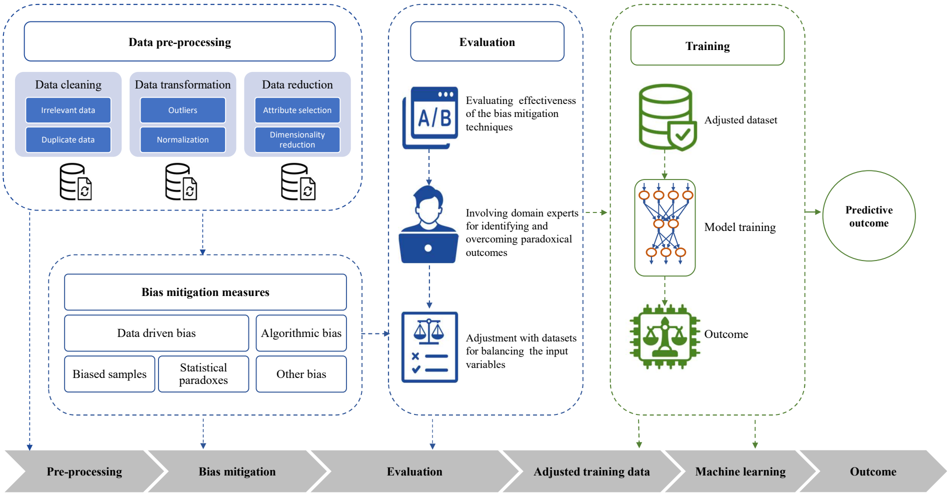


Figure 4: Framework for mitigating the impact of bias resulting from statistical paradoxes [VIII]

- This framework is designed to handle and address bias caused by unexpected relationships between data groups, improving fairness and reliability in decision-making.

- The proposed framework takes a more comprehensive approach compared to existing frameworks that typically focus on specific techniques for reducing bias. It offers support for multiple bias mitigation techniques that cover various strategies for addressing different types of biases.

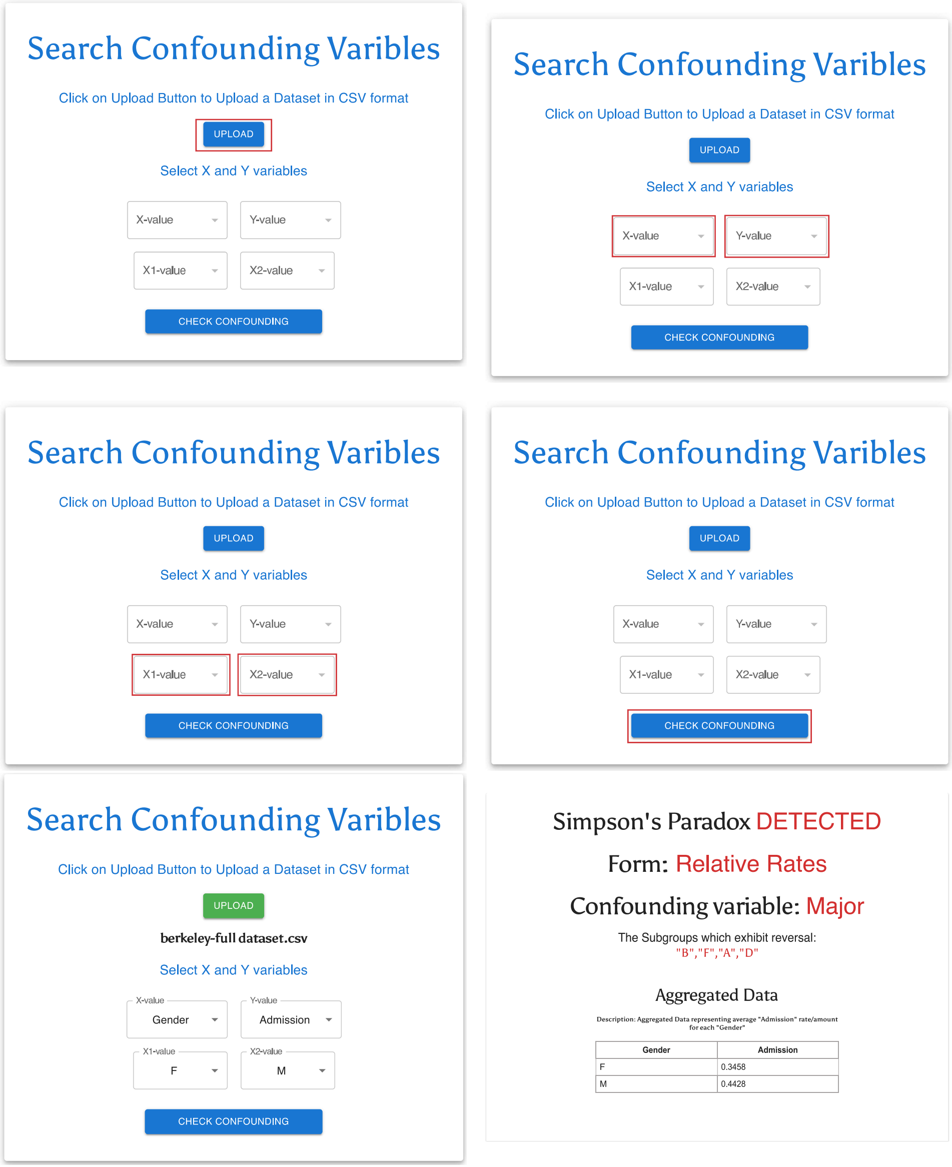
- The proposed framework allows the balancing and adjustment of training data to avoid statistical paradoxes.

- The proposed framework focuses on involving domain experts in AI development and deployment for understanding social and ethical values.

6.3 The Web-based Applications to Identify the Impacts of Confounding Variables [VIII]

By utilizing the suggested framework, the author has developed a web-based application aimed at identifying the impacts of confounding variables and addressing statistical paradoxes [VIII]. At present, the application performs a systematic identification of confounding variables in categorical and continuous datasets. Additionally, the application can detect the presence of Simpson's paradox within multivariate datasets. The application is developed using Python 3.10 programming language and the FastAPI framework, harnessing its advantages of rapid development and high-performance capabilities. In Figure 5, screenshots of the application's user interface are provided, highlighting its user-friendly design and simplicity.

The application offers a seamless user experience, requiring just a few straightforward steps to accomplish its purpose. First, users can easily import a dataset into the application. Next, select the necessary parameters related to their analysis. Once the dataset and parameters are set, users can simply click on the 'Check Confounding' button to detect any confounding variables within the dataset and identify instances of Simpson's paradox. This user-friendly approach ensures that users, regardless of their technical expertise, can easily harness the power of the application to gain valuable insights from their data. The programming code and comprehensive usage guidelines for the proposed application can be accessed from the GitHub repository [3]. The repository contains all the necessary resources to facilitate a smooth understanding and effective implementation of the application.


Search Confounding Varibles
Click on Upload Button to Upload a Dataset in CSV format
UPLOAD
Select X and Y variables
X-value
Y-value
X1-value
X2-value
CHECK CONFOUNDING

berkeley-full dataset.csv

X-value: Gender
Y-value: Admission
X1-value: F
X2-value: M

Simpson's Paradox DETECTED

Form: Relative Rates

Confounding variable: Major

The Subgroups which exhibit reversal:
"B","F","A","D"

Aggregated Data

Description: Aggregated Data representing average "Admission" rate/amount for each "Gender"

Gender	Admission
F	0.3458
M	0.4428

Figure 5: Application's graphical user interface [VIII]

[3] https://github.com/rahul-sharmaa/SimpsonP

7 Results and Artifact's Evaluation

Evaluation serves as a critical component of the research process, ensuring that the artifact's performance and effectiveness are thoroughly assessed. Demonstrating the utility, quality, and efficacy of a design artifact requires a rigorous evaluation process employing well-executed methods.

This section aims to demonstrate the evaluation of the research by presenting the results of the research questions (RQs). It highlights how these results contribute to filling gaps in the existing body of knowledge, advancing our understanding of the subject. Moreover, the section emphasizes how the outcomes of the RQs align with the study's objectives, validating the research's relevance and significance. By addressing these RQs, the research provides valuable insights and practical contributions to the field, making it a meaningful and impactful endeavor.

Table 3 presents a comprehensive summary of the results obtained from RQ1, along with the corresponding publications that address these questions. Three research articles are relevant to RQ1 and its corresponding sub-research questions. Table 3 provides a clear overview of the outcomes achieved in relation to the specified research inquiries, showcasing the specific contributions made in each publication.

Table 3: Summary of the RQ1 results with publication and current knowledge gaps

Results with Publications	Current Knowledge Gaps
In publication [II], a novel framework for the unification of DSTs is presented to develop next-generation decision support tools.	In Section 3.1, the author highlighted that existing studies lack a unified framework to integrate and utilize the outcomes of one DST in another DST.
The publication [II] implemented a sample tool for unifying DSTs based on the proposed framework.	Existing studies discussed in Section 3.1 lack any tool or application that delivers the unification of DSTs.
The publication [III] presented the semantic correspondences between the foundations of SR, OLAP, and ARM. This can be considered as a step towards the unification of different DSTs	The previous research efforts discussed in Section 3.1, focused on integrating OLAP and ARM but did not specifically address the correspondences and unification between DSTs.
The publications [III] and [I] discovered that SR, OLAP, and ARM operations complement each other in data understanding, visualization, and personalized decision-making.	It remains unclear from existing studies whether SR, OLAP, and ARM can be considered synonymous. As a result, there is a significant research gap concerning the exploration of correspondences and the development of a comprehensive framework for integrating DSTs.

Table 4 aligns the results of RQ2 with the identified knowledge gaps, providing a clear overview of how the research addresses and bridges these gaps. Five research articles are relevant to RQ2, each offering valuable insights that contribute to filling the identified knowledge gaps. The table highlights the specific outcomes of RQ2 and the corresponding publications, providing a comprehensive understanding of the research's impact on the existing body of knowledge.

Table 4: Summary of the RQ2 results with publication and current knowledge gaps

Results with Publications	Current Knowledge Gaps
The publication [VIII] made a significant contribution to the development of fair and trustworthy DSTs by proposing a framework that effectively mitigates the impact of biases resulting from statistical paradoxes.	In Section 3.2, an overview of the related literature revealed the existence of various frameworks, but none of them fully recognized and accounted for the effects of different statistical paradoxes.
Publication [VIII] provides a comprehensive investigation on measures for adjusting the impact of confounding variables. Additionally, it introduces a web-based application to identify and adjust the impact of confounding variables.	Upon reviewing the information in Section 3.2, it becomes evident that in DSTs, no existing studies combined both theoretical insights and practical solutions concerning the issue of statistical paradoxes.
Publications [VI] and [VII] present a preliminary study on statistical paradoxes. These contributions lay the groundwork for further exploration and understanding of statistical paradoxes and their implications in various datasets.	The lack of analysis of real-life examples in the reviewed literature is evident. This underscores the need for further research and investigation into the practical implications and real-world applications of statistical paradoxes.
The Publication [IV] contributed by identifying the confounding variable and detecting Simpson's paradox within categorical datasets by using the stratification Pearson correlation.	It is worth noting that the literature review conducted in the article and summarised in 3.2 also highlighted the absence of a key solution for detecting statistical paradoxes in multivariate data.
The publication [V] discussed the ways to identify the impact of Simpson's paradox in continuous data and experimented with three datasets.	The existing literature lacks a comprehensive solution for effectively detecting Simpson's paradox in continuous datasets. This gap in the literature highlighted the need for further research and the development of robust methodologies specifically tailored to address statistical paradoxes.

Through this evaluation, the author concludes that this research successfully addresses all the research questions, resulting in the creation of three IT artifacts: two frameworks and one web-based application. These artifacts serve as valuable contributions to the field, offering innovative solutions and practical tools to tackle the identified challenges.

8 Future Work

This dissertation addressed two major challenges in the development of unified and trustworthy decision-support techniques. Moving forward, there are potential future steps to further improve and enhance the outcomes of this research. By undertaking these future steps, the research can continue to advance and make valuable contributions to the field of decision support and bias mitigation.

8.1 Unification of DSTs

This work establishes a foundation for uncovering semantic correspondences between DSTs and developing a unified framework for their usage. Building upon the established foundation, future research should aim to expand the scope of identifying semantic correspondences between DSTs. This expansion should involve investigating and uncovering semantic correspondences among a broader range of DSTs beyond the three currently examined. By exploring additional DSTs, researchers can develop a more comprehensive and inclusive unified framework that encompasses a wider spectrum of decision-support tools. Following future directions will contribute to the advancement of decision support tools, ensuring their relevance, effectiveness, and applicability in various domains.

- *Expansion of Semantic Correspondences:* Future research should embark on exploring and uncovering semantic correspondences between a wider array of DSTs. This endeavor will significantly contribute to the advancement of decision support tools, facilitating the development of cutting-edge frameworks that are more comprehensive and versatile. By investigating additional DSTs, researchers can unlock new possibilities and applications in the field, enhancing the effectiveness and adaptability of DSSs.

- *Performance Optimization:* Efforts must be made to address performance issues that arise when dealing with large datasets. Leveraging high-performance computing (HPC) infrastructure can significantly enhance the scalability and efficiency of the tool, empowering it to handle larger datasets more effectively. This optimization is crucial to ensure that the tool maintains its reliability and efficiency, even in the face of growing data volumes.

- *Advanced Platform Development:* The proposed tool can be further enriched by incorporating additional features, such as Pearson correlation and regression. These enhancements will empower decision-makers with enhanced analytical capabilities, facilitating the ability to make well-informed and unbiased decisions with a more comprehensive understanding of the data.

8.2 Statistical Paradoxes on DSTs

In this research, the author investigates well-known statistical paradoxes as evidence of expert system bias. Expert system bias poses a significant challenge to successful decision-making as it directly leads to biased decisions. The demonstrated importance of addressing statistical paradoxes in DSTs sets the direction for future research and development toward the development of fair and trustworthy DSTs. By pursuing the following future directions, researchers can advance the field of bias mitigation, improve the reliability of DSSs, and contribute to the development of trustworthy AI systems. These efforts will help ensure accurate, fair, and accountable decision-making processes with wider societal implications.

- *Mitigation of Multiple Statistical Paradoxes:* To advance the effectiveness of the proposed framework in mitigating statistical paradoxes, further research is required. This research should focus on expanding its application beyond Simpson's paradox and encompassing a wider range of statistical paradoxes. Developing robust strategies to handle these paradoxes in diverse datasets necessitates a deeper understanding of statistical concepts. By gaining a more comprehensive understanding, researchers can devise effective and tailored approaches for addressing different paradoxes encountered in various datasets.

- *Handling Complex and High-Dimensional Datasets:* Mitigating statistical paradoxes in complex and high-dimensional datasets, which exhibit non-linear and interactive relationships between variables, presents a significant challenge. To tackle these complexities, future research should concentrate on developing advanced techniques that can effectively address these intricacies and offer reliable bias mitigation. The focus should be on devising approaches capable of navigating the intricate relationships and complexities inherent in such datasets, ensuring accurate and robust bias mitigation strategies.

- *Expansion of the Web-Based Tool:* The current web-based tool, built upon the proposed framework, has demonstrated its efficacy in identifying confounding variables and detecting Simpson's paradox in both categorical and continuous datasets. To expand its capabilities, further enhancements and refinements are necessary. These improvements will enable the tool to handle a broader range of datasets with diverse characteristics and effectively address multiple statistical paradoxes. By extending its functionality, the tool will become more versatile and valuable for researchers and practitioners in mitigating bias and enhancing decision-making across various data scenarios.

9 Conclusion

This Ph.D. thesis makes a noteworthy contribution to the field of information system studies by presenting novel insights into two distinct and equally important research areas in decision support techniques. Based on that, the thesis aimed to answer two main and six supplementary research questions to foster fair and reliable decision-making procedures across diverse domains.

- RQ-1: How to bridge the artificial gaps between different DSTs?
  - RQ-1.1 What are the semantic correspondences between the three major decision support techniques, i.e., statistical reasoning (SR), online analytical processing (OLAP) and association rule mining (ARM)?
  - RQ-1.2 How to provide a systematic interpretation of results between different decision support techniques? In how far can we consider SR, OLAP, and ARM as synonymous?
  - RQ-1.3 How to develop a common framework for integrating SR, OLAP and ARM?

- RQ-2: How to systematically assess the impact of statistical paradoxes in multivariate data? How to utilize these assessments for better decision-making?
  - RQ-2.1 How to identify the existence of the Yule-Simpson effect in multivariate data?
  - RQ-2.2 How to adjust the impact of the Yule-Simpson Effect in multivariate data?

- RQ-2.3 How to develop a platform to handle statistical paradoxes in multivariate data and recommend appropriate adjustments for improved decision-making?

To address these questions, this thesis draws on research and contributions from eight articles published between 2019 and 2023. These research articles offer rigorous analysis, empirical investigations, and experiments using various benchmark and real-life datasets. By leveraging these studies, the thesis enriches the existing body of knowledge in the field, providing valuable insights and advancements in decision support techniques and bias mitigation.

To address the first primary research question and its three supplementary sub-research questions, articles [II], [III], and [I] explored various approaches to bridge the gap between three popular DSTs: SR, OLAP, and ARM. Specifically, they elucidated the semantic correspondences between their foundations, namely probability theory, relational algebra, and itemset apparatus, respectively. Additionally, the paper [II] introduced a new framework to unify DSTs and implemented a tool to validate the concept of unification of DSTs. These research efforts collectively contribute to the advancement of unified decision support techniques and provide valuable insights into the integration of different DSTs.

To address the second research question and its three supplementary research questions, article [VIII], [IV], [V], [VI], and [VII] explore various methods for identifying and adjusting the impact of confounders and statistical paradoxes. These efforts aim to strengthen existing DSTs, promoting fair and trustworthy decision-making processes. Moreover, this dissertation proposes an additional framework to mitigate the impacts of statistical paradoxes in expert systems. To validate the effectiveness and usefulness of this framework, a web-based application has been developed. Currently, the application enables the investigation of possible confounders by detecting instances of Simpson's paradox and provides a feature for adjusted observations. These contributions collectively advance the field of decision support and provide valuable tools and methodologies for addressing biases in data analysis and decision-making.

Overall, this research delivers six significant contributions to the fields of decision support and bias mitigation.

1. Elaborating semantic correspondences between the foundations of SR, OLAP, and ARM, revealing their interconnections.

2. Proposing a novel framework that unifies DSTs and provides a tool to validate this unification concept, facilitating unification among different decision support techniques.

3. Identifying confounding variables and instances of statistical paradoxes, such as Simpson's paradox, in multivariate datasets.

4. Generalizing back-door adjustment techniques and utilizing propensity weighting to effectively mitigate the impact of confounders, promoting fairness and reliability in decision-making.

5. Proposing a novel framework to address statistical paradoxes and confounding effects, thereby enhancing the accuracy and trustworthiness of DSTs.

6. Developing a web-based tool that automates the detection and mitigation of confounding effects, providing valuable assistance to data scientists and researchers in bias mitigation and data analysis.

This research has successfully addressed all research questions and achieved the primary objective of the study. Through the six significant contributions to the field of decision support and bias mitigation, this work has provided valuable insights and solutions to various challenges in the domain. The proposed unification framework and the developed web-based tool exemplify the practical applications of the research findings. By effectively handling confounding variables, statistical paradoxes and promoting fairness in decision-making, this study has demonstrated its comprehensive approach in enhancing the accuracy and reliability of decision-support techniques. Thus, the objectives of this research were fully met, resulting in a substantial contribution to the existing body of knowledge in this area.

List of Figures

List of Tables

References

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining Association Rules Between Sets of Items in Large Databases. *ACM SIGMOD Record*, 22(2):207–216, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of VLDB'1994 – the 20th International Conference on Very Large Data Bases*, page 487–499. Morgan Kaufmann, 1994.

[3] N. Alipourfard, P. G. Fennell, and K. Lerman. Can you trust the trend? discovering Simpson's paradoxes in social data. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 19–27, New York, NY, USA, 2018. Association for Computing Machinery.

[4] N. Alipourfard, P. G. Fennell, and K. Lerman. Using Simpson's paradox to discover interesting patterns in behavioral data. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, pages 2–11. AAAI Publications, 2018.

[5] N. Bantilan. Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation, 2017.

[6] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15, 2019.

[7] J. Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

[8] M. Bertl, P. Ross, and D. Draheim. A survey on AI and decision support systems in psychiatry – uncovering a dilemma. *Expert Systems with Applications*, 202(117464):1–14, 2022.

[9] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. Technical Report MSR-TR-2020-32, Microsoft, May 2020.

[10] C. R. Blyth. On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366, June 1972.

[11] N. Bolloju, M. Khalifa, and E. Turban. Integrating knowledge management into enterprise environments for the next generation decision support. *Decision Support Systems*, 33(2):163–176, 2002.

[12] N. Cartwright. Causal Laws and Effective Strategies. *Noûs*, 13(4):419, Nov. 1979.

[13] R. B. Cattell. P-technique factorization and the determination of individual dynamic structure. *Journal of Clinical Psychology*, 8(1):5–10, 1952.

[14] S. Chaudhuri and U. Dayal. Data warehousing and OLAP for decision support. In *Proceedings of SIGMOD'97 – the 1997 ACM SIGMOD International Conference on Management of Data*, page 507–508. Association for Computing Machinery, 1997.

[15] E. Codd, S. Codd, and C. Salley. *Providing OLAP to User-Analysts: An IT Mandate*. E. F. Codd and Associates, 1993.

[16] A. J. Conger. A revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and psychological measurement*, 34(1):35–46, 1974.

[17] S. P. Curley and G. J. Browne. Normative and Descriptive Analyses of Simpson's Paradox in Decision Making. *Organizational Behavior and Human Decision Processes*, 84(2):308–333, Mar. 2001.

[18] J. Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G, Oct. 2018.

[19] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979.

[20] D. Draheim. Future perspectives of association rule mining based on partial conditionalization (DEXA'2019 keynote). In *Proceedings of DEXA'2019 – the 30th International Conference on Database and Expert Systems Applications*, volume 11706 of *Lecture Notes in Computer Science*. Springer, 2019.

[21] W. Fan, H. Lu, S. E. Madnick, and D. Cheung. Direct: a system for mining data value conversion rules from disparate data sources. *Decision Support Systems*, 34(1):19–39, 2002.

[22] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–37, 1996.

[23] K. Fiedler. Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, 107(4):659–676, 2000.

[24] K. Fiedler. The ultimate sampling dilemma in experience-based decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1):186–203, 2008.

[25] K. Fiedler, P. Freytag, and T. Meiser. Pseudocontingencies: An integrative account of an intriguing cognitive illusion. *Psychological Review*, 116(1):187–206, 2009.

[26] K. Fiedler, E. Walther, P. Freytag, and S. Nickel. Inductive Reasoning and Judgment Interference: Experiments on Simpson's Paradox. *Personality and Social Psychology Bulletin*, 29(1):14–27, Jan. 2003.

[27] R. A. Fisher. III. The influence of rainfall on the yield of wheat at Rothamsted. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213(402-410):89–142, 1925.

[28] A. A. Freitas. On objective measures of rule surprisingness. In *European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 1–9. Springer, 1998.

[29] D. Ghosh, M. Pandey, C. Gautam, A. Vidyarthi, R. Sharma, and D. Draheim. Utilizing continuous time markov chain for analyzing video-on-demand streaming in multimedia systems. *Expert Systems with Applications*, 223:119857, 2023.

[30] G. A. Gorry and S. M. S. Morton. A framework for management information systems. Technical Report 510-71, Alfred P. Sloan School of Management, Massachusetts Institute of Technology, Cambridge, February 1971.

[31] S. Greenland. Simpson's paradox from adding constants in contingency tables as an example of bayesian non collapsibility. *The American Statistician*, 64(4):340–344, 2010.

[32] J. Han, Y. Fu, W. Wang, J. Chiang, O. R. Zaïane, and K. Koperski. DBMiner: Interactive mining of multiple-level knowledge in relational databases. In *Proceedings of SIGMOD'96 – the 1996 ACM SIGMOD International Conference on Management of Data*, page 550. Association for Computing Machinery, 1996.

[33] J. Han, J. Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1):53–87, 2004.

[34] M. A. Hernán, D. Clayton, and N. Keiding. The Simpson's paradox unraveled. *International Journal of Epidemiology*, 40(3):780–785, June 2011.

[35] A. R. Hevner, S. T. March, J. Park, and S. Ram. Design science in information systems research. *MIS quarterly*, 28(1):75–105, 2004.

[36] T. Imielinski, L. Khachiyan, and A. Abdulghani. Cubegrades: Generalizing association rules. *Data Mining and Knowledge Discovery*, 6(3):219–257, 2002.

[37] A. Julia, L. Jeff, K. Lauren, and M. Surya. Machine Bias – propublica.org. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016. [Accessed 06-Feb-2023].

[38] M. Kamber, J. Han, and J. Y. Chiang. Metarule-guided mining of multi-dimensional association rules using data cubes. In *Proceedings of KDD'97 – the 3rd International Conference on Knowledge Discovery and Data Mining*, KDD'97, page 207–210. AAAI Press, 1997.

[39] M. Kaushik. Swarm-intelligence algorithms for mining numerical association rules: An exhaustive multi-aspect analysis of performance assessment data. *SSRN Electronic Journal*, 2023.

[40] M. Kaushik, R. Sharma, I. F. Jr.2, and D. Draheim. Numerical association rule mining: A systematic literature review, 2023.

[41] M. Kaushik, R. Sharma, S. A. Peious, and D. Draheim. Impact-driven discretization of numerical factors: Case of two- and three-partitioning. In *Big Data Analytics*, pages 244–260, Cham, 2021. Springer International Publishing.

[42] M. Kaushik, R. Sharma, S. A. Peious, M. Shahin, S. Ben Yahia, and D. Draheim. On the potential of numerical association rule mining. In *Proceedings of FDSE'2020 – the 7th International Conference on Future Data and Security Engineering*, volume 12466 of *Lecture Notes in Computer Science*, pages 3–20. Springer Singapore, 2020.

[43] M. Kaushik, R. Sharma, S. A. Peious, M. Shahin, S. B. Yahia, and D. Draheim. A systematic assessment of numerical association rule mining methods. *SN Computer Science*, 2(5):1–13, 2021.

[44] M. Kaushik, R. Sharma, M. Shahin, S. A. Peious, and D. Draheim. An analysis of human perception of partitions of numerical factor domains. In *Information Integration and Web Intelligence*, pages 137–144, Cham, 2022. Springer.

[45] M. Kaushik, R. Sharma, A. Vidyarthi, and D. Draheim. Discretizing numerical attributes: An analysis of human perceptions. In *New Trends in Database and Information Systems*, pages 188–197, Cham, 2022. Springer.

[46] R. Kievit, W. Frankenhuis, L. Waldorp, and D. Borsboom. Simpson's paradox in psychological science: a practical guide. *Frontiers in Psychology*, 4:513, 2013.

[47] G. King and M. Roberts. Ei: a (n r) program for ecological inference. *Harvard University*, 2012.

[48] H.-J. Lenz and B. Thalheim. OLAP databases and aggregation functions. In *Proceedings Thirteenth International Conference on Scientific and Statistical Database Management. SSDBM 2001*, pages 91–100, 2001.

[49] I. Liiv. *Association Rules*, pages 27–43. Springer Singapore, Singapore, 2021.

[50] S. Liu, A. H. Duffy, R. I. Whitfield, and I. M. Boyle. Integration of decision support systems to improve decision support performance. *Knowledge and Information Systems*, 22(3):261–286, 2010.

[51] D. P. MacKinnon, A. J. Fairchild, and M. S. Fritz. Mediation analysis. *Annual Review of Psychology*, 58(1):593–614, 2007.

[52] S. T. March and G. F. Smith. Design and natural science research on information technology. *Decision Support Systems*, 15(4):251–266, 1995.

[53] M. L. Markus, A. Majchrzak, and L. Gasser. A design theory for systems that support emergent knowledge processes. *MIS Quarterly*, 26(3):179–212, 2002.

[54] N. H. Nie, D. H. Bent, and C. H. Hull. *SPSS: Statistical Package for the Social Sciences*. McGraw-Hill, 1970.

[55] C. O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, New York, first edition edition, 2016.

[56] R. Otte. Probabilistic Causality and Simpson's Paradox. *Philosophy of Science*, 52(1):110–125, Mar. 1985.

[57] J. Pearl. Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association*, 95(450):428–431, 2000.

[58] J. Pearl. Simpson's paradox: An anatomy. *UCLA Cognitive Systems Laboratory, Technical Report.*, 2011.

[59] J. Pearl. Understanding simpson's paradox. *SSRN Electronic Journal*, 68, 01 2013.

[60] L. A. Pearson Karl and B.-M. Leslie. Genetic (reproductive) selection: Inheritance of fertility in man, and of fecundity in thoroughbred racehorses. *Philosophical Transactions of the Royal Society of London: Series A*, 192:257–330, Dec. 1899.

[61] S. A. Peious, R. Sharma, M. Kaushik, S. A. Shah, and S. B. Yahia. Grand reports: a tool for generalizing association rule mining to numeric target values. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 28–37. Springer, 2020.

[62] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, Mar 1986.

[63] W. S. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–357, 1950.

[64] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[65] R. Rupnik, M. Kukar, and M. Krisper. Integrating data mining and decision support through data mining based decision support system. *Journal of Computer Information Systems*, 47(3):89–104, 2007.

[66] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani. Aequitas: A bias and fairness audit toolkit, 2019.

[67] SAS Institute. *SAS User's guide: statistics; version 5 ed*. SAS InstituteInc., Cary, NC, 3. [print.] edition, 1987.

[68] M. Schaller. Sample size, aggregation, and statistical reasoning in social inference. *Journal of Experimental Social Psychology*, 28(1):65–85, Jan. 1992.

[69] M. Schield. Simpson's paradox and cornfield's conditions. *ASA Proceedings of the Section on Statistical Education*, 1999:106–111, 08 1999.

[70] R. Schutt and C. O'Neil. *Doing Data Science: Straight Talk from the Frontline*. O'Reilly Media, 2013.

[71] M. Shahin, S. Arakkal Peious, R. Sharma, M. Kaushik, S. Ben Yahia, S. A. Shah, and D. Draheim. Big data analytics in association rule mining: A systematic literature review. In *International Conference on Big Data Engineering and Technology (BDET)*, page 40–49. Association for Computing Machinery, 2021.

[72] M. Shahin, M. R. Heidari Iman, M. Kaushik, R. Sharma, T. Ghasempouri, and D. Draheim. Exploring factors in a crossroad dataset using cluster-based association rule mining. *Procedia Computer Science*, 201:231–238, 2022. The 13th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 5th International Conference on Emerging Data and Industry 4.0 (EDI40).

[73] M. Shahin, S. Saeidi, S. A. Shah, M. Kaushik, R. Sharma, S. A. Peious, and D. Draheim. Cluster-based association rule mining for an intersection accident dataset. In *2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, pages 1–6, 2021.

[74] R. Sharma. On statistical paradoxes and overcoming the impact of bias in expert systems: towards fair and trustworthy decision making. *SSRN*, pages 1–37, July 2023. doi:10.2139/ssrn.4506432.

[75] R. Sharma, H. Garayev, M. Kaushik, S. A. Peious, P. Tiwari, and D. Draheim. Detecting Simpson's paradox: A machine learning perspective. In C. Strauss, A. Cuzzocrea, G. Kotsis, A. M. Tjoa, and I. Khalil, editors, *Proceedings of DEXA 2022 – the 33rd International Conference on Database and Expert Systems Applications*, pages 323–335, Cham, 2022. Springer International Publishing.

[76] R. Sharma, M. Kaushik, S. A. Peious, A. Bazin, S. A. Shah, I. Fister, S. B. Yahia, and D. Draheim. A novel framework for unification of association rule mining, online analytical processing and statistical reasoning. *IEEE Access*, 10:12792–12813, 2022.

[77] R. Sharma, M. Kaushik, S. A. Peious, M. Bertl, A. Vidyarthi, A. Kumar, and D. Draheim. Detecting Simpson's paradox: A step towards fairness in machine learning. In S. Chiusano, T. Cerquitelli, R. Wrembel, K. Nørvåg, B. Catania, G. Vargas-Solar, and E. Zumpano, editors, *Proceedings of ADBIS 2022 – the 26th International Conference on New Trends in Database and Information Systems*, pages 67–76, Cham, 2022. Springer International Publishing.

[78] R. Sharma, M. Kaushik, S. A. Peious, M. Shahin, A. Vidyarthi, and D. Draheim. Existence of the Yule-Simpson effect: An experiment with continuous data. In *Proceedings of Confluence 2022 – the 12th International Conference on Cloud Computing, Data Science & Engineering*, pages 351–355, 2022.

[79] R. Sharma, M. Kaushik, S. A. Peious, M. Shahin, A. Vidyarthi, P. Tiwari, and D. Draheim. Why not to trust big data: Discussing statistical paradoxes. In U. K. Rage, V. Goyal, and P. K. Reddy, editors, *Proceedings of DASFAA 2022 International Workshops – the 27th International Conference on Database Systems for Advanced Applications*, pages 50–63, Cham, 2022. Springer International Publishing.

[80] R. Sharma, M. Kaushik, S. A. Peious, M. Shahin, A. S. Yadav, and D. Draheim. Towards unification of statistical reasoning, OLAP and association rule mining: Semantics and pragmatics. In A. Bhattacharya, J. Lee Mong Li, D. Agrawal, P. K. Reddy, M. Mohania, A. Mondal, V. Goyal, and R. Uday Kiran, editors, *Proceedings of DASFAA 2022 – the 27th International Conference on Database Systems for Advanced Applications*, pages 596–603, Cham, 2022. Springer International Publishing.

[81] R. Sharma, M. Kaushik, S. A. Peious, S. B. Yahia, and D. Draheim. Expected vs. unexpected: Selecting right measures of interestingness. In M. Song, I.-Y. Song, G. Kotsis, A. M. Tjoa, and I. Khalil, editors, *Proceedings of DaWaK 2020 – the 22nd International Conference on Big Data Analytics and Knowledge Discovery*, pages 38–47, Cham, 2020. Springer International Publishing.

[82] R. Sharma and S. A. Peious. Towards unification of decision support technologies: Statistical reasoning, OLAP and association rule mining. `https://github.com/rahulgla/unification`, December 21, 2021.

[83] E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

[84] B. A. Spellman, C. M. Price, and J. M. Logan. How two causes are different from one: The use of (un)conditional information in Simpson's paradox. *Memory & Cognition*, 29(2):193–208, Mar. 2001.

[85] S. M. Stigler. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press, 1986.

[86] B. Thalheim and H.-J. Lenz. A formal framework of aggregation for the OLAP-OLTP model. *The International Journal of Universal Computer Science (J.UCS)*, 2009.

[87] Y.-K. Tu, D. Gunnell, and M. S. Gilthorpe. Simpson's paradox, Lord's paradox, and Suppression effects are the same phenomenon–the reversal paradox. *Emerging themes in epidemiology*, 5(1):1–9, 2008.

[88] A. Vidyarthi, R. Agarwal, D. Gupta, R. Sharma, D. Draheim, and P. Tiwari. Machine learning assisted methodology for multiclass classification of malignant brain tumors. *IEEE Access*, 10:50624–50640, 2022.

[89] H. Wang. Intelligent agent-assisted decision support systems: Integration of knowledge discovery, knowledge analysis, and group decision support. *Expert Systems with Applications*, 12(3):323–335, 1997.

[90] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, 2020.

[91] G. U. Yule. Notes on the theory of association of attributes in statistics. *Biometrika*, 2(2):121–134, 02 1903.

[92] H. Zhu. *On-Line Analytical Mining of Association Rules*, *Master's thesis*. School of Computer Science, Simon Fraser University, British Columbia, Canada, 1998.

[93] Y. Zhu, C. Bornhövd, D. Sautner, and A. P. Buchmann. Materializing web data for OLAP and DSS. In *Web-Age Information Management*, pages 201–215. Springer Berlin, Heidelberg, 2000.

[94] Z. Y. Zhuang, L. Churilov, F. Burstein, and K. Sikaris. Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners. *European Journal of Operational Research*, 195(3):662–675, 2009.

Acknowledgements

I am filled with deep gratitude towards my supervisor, Dirk Draheim, for his exceptional advice, unwavering support, and patient guidance throughout my PhD study. His vast knowledge and extensive research experience have been a constant source of inspiration during my academic research. I extend my profound gratitude to Tallinn University of Technology and all individuals associated with the Department of Software Science, who have provided invaluable support and resources, directly and indirectly, enabling the successful completion of my research.

I want to sincerely thank Dr. R K Agarwal, Director General, Ajay Kumar Garg Engineering College, Ghaziabad, India, for his invaluable support and guidance throughout my research journey.

I want to take a moment to express my sincere gratitude to the three most influential and supportive women in my life. These remarkable women are none other than my beloved mother, loving wife, and cherished daughter. Their unwavering encouragement and support have been absolutely invaluable to me throughout my academic journey, and I am deeply grateful for their unwavering commitment to my success. Their constant presence and unwavering love have been a source of inspiration and motivation for me, and I can never thank them enough for the ways in which they have positively impacted my life.

Abstract

Unification of Decision Support Techniques: Mitigating Statistical Paradoxes for Enabling Trustworthy Decision Making

This dissertation strengthens the decision support techniques by providing a framework for unifying their conceptual foundations and addressing the impact of confounding variables and statistical paradoxes. By doing so, this research aims to enhance the effectiveness and reliability of decision-support techniques in various domains.

The research provides six significant contributions to address existing technological and knowledge gaps to foster fair and trustworthy decision-making processes. These contributions are the outcomes of two primary research questions and six supplementary research questions answered within the thesis. The thesis utilizes design science research methodology to create innovative artifacts and methods, providing new insights to widen understanding of the domain under the research.

The first contribution provides ways to establish semantic correspondences between the three major decision support techniques, i.e., statistical reasoning, online analytical processing and association rule mining. It examines various approaches to bridge the gap between them. The second contribution is a novel framework for unifying decision-support techniques for developing a unified platform to interpret results from one DST to another. The third contribution discusses two measures for identifying confounding effects in categorical and continuous datasets. The fourth contribution discusses the measure for adjusting the confounding effects.

Further, the fifth contribution provides a framework for mitigating the impact of bias resulting from statistical paradoxes. The sixth contribution is a web-based application that automatically detects and addresses confounding effects. This application is an invaluable tool for data scientists and researchers, offering automated detection and mitigation of confounding effects and providing a streamlined approach to effectively addressing and overcoming such data analysis challenges. The author argues that the suggested framework and application possess substantial potential for further extensions beyond their current scope of application.

Kokkuvõte

Otsuste toetamise tehnikate ühtlustamine: statistiliste paradoksside mõju maandamine usaldusväärsete otsuste tegemise võimaldamiseks

See väitekiri tugevdab otsuste tegemist toetavaid tehnikaid, pakkudes raamistikku nende tehnikate kontseptuaalsete aluste ühendamiseks ning käsitledes ühismõjurite *(confounding variable)* ja statistiliste paradokside keerukust. Sellest tulenevalt, on käesoleva uuringu eesmärgiks tõhustada otsuste tegemist toetavaid tehnikaid ja nende usaldusväärsust erinevates valdkondades.

Uurimistöö raames valmis kuus olulist tulemit olemasolevate lünkade kõrvaldamiseks tehnoloogias ja teadmistes, et täiustada õiglasi ja usaldusväärseid otsustamist toetavaid protsesse. Loodud artefaktid on käesoleva lõputöö raames vastatud kahe peamise uurimisküsimuse ja kuue täiendava uurimisküsimuse tulemused. Lõputöös kasutatakse uuenduslike artefaktide ja meetodide loomiseks disainiteaduse uurimismetoodikat *(design science)*, pakkudes uusi teadmisi selle uurimuse keskseks oleva valdkonna mõistmiseks.

Esimene tulem võimaldab luua semantilisi vastavusi kolme peamise otsuste tegemist toetava tehnika vahel. Nendeks on statistiline põhjendamine, veebipõhine analüütiline töötlemine ja assotsiatsioonireeglite kaevandamine. Nende tehnikate vaheliste erinevuste ületamiseks uuritakse erinevaid lähenemisviise. Teiseks tulemiks on uudne raamistik otsustamist toetavate tehnikate ühendamiseks ühtse platvormi väljatöötamiseks, et võimaldada tulemuste tõlgendamist ühest DST-st teise. Kolmas tulem käsitleb kahte meedet andmetöötluses kaasnevate mõjude tuvastamiseks kategoorilistes ja pidevates andmekogumites. Neljandas tulemis käsitletakse meetmeid ühismõjurite reguleerimiseks.

Lisaks annab lõputöö viies tulem raamistiku statistilistest paradoksidest tulenevate mõjude leevendamiseks. Kuuendaks tulemiks on veebipõhine rakendus, mis tuvastab ja käsitleb automaatselt ühismõjureid. See rakendus on väärtuslik tööriist eelkõige andmeteadlastele, kuid ka teistele teadlastele, pakkudes ühismõjurite automaatset tuvastamist ja leevendamist ning sujuvamat lähenemisviisi selliste andmeanalüüsi väljakutsete tõhusaks ületamiseks ja lahendamiseks. Töö autor usub, et pakutud raamistikul ja rakendusel on märkimisväärne potentsiaal laieneda edaspidiselt väljaspoole nende praegust rakendusala.

Appendix 1

[I]

R. Sharma, M. Kaushik, S. A. Peious, S. B. Yahia, and D. Draheim. Expected vs. unexpected: Selecting right measures of interestingness. In M. Song, I.-Y. Song, G. Kotsis, A. M. Tjoa, and I. Khalil, editors, *Proceedings of DaWaK 2020 – the 22nd International Conference on Big Data Analytics and Knowledge Discovery*, pages 38–47, Cham, 2020. Springer International Publishing

Expected vs. Unexpected: Selecting Right Measures of Interestingness

Rahul Sharma[1( )], Minakshi Kaushik[1], Sijo Arakkal Peious[1], Sadok Ben Yahia[2], and Dirk Draheim[1]

[1] Information Systems Group, Tallinn University of Technology, Akadeemia tee 15a, 12618 Tallinn, Estonia
{rahul.sharma,minakshi.kaushik,sijo.arakkal,dirk.draheim}@taltech.ee

[2] Software Science Department, Tallinn University of Technology, Akadeemia tee 15a, 12618 Tallinn, Estonia
sadok.ben@taltech.ee

Abstract. Measuring interestingness in between data items is one of the key steps in association rule mining. To assess interestingness, after the introduction of the classical measures (support, confidence and lift), over 40 different measures have been published in the literature. Out of the large variety of proposed measures, it is very difficult to select the appropriate measures in a concrete decision support scenario. In this paper, based on the diversity of measures proposed to date, we conduct a preliminary study to identify the most typical and useful roles of the measures of interestingness. The research on selecting useful measures of interestingness according to their roles will not only help to decide on optimal measures of interestingness, but can also be a key factor in proposing new measures of interestingness in association rule mining.

Keywords: Knowledge discovery in databases · Association rule mining · Measures of interestingness

1 Introduction

In knowledge discovery in data (KDD), association rule mining (ARM) is one of the most established data mining techniques. It is commonly used to find out interesting patterns between data items in large transactional data sets. In ARM, association rules are accompanied by measures of interestingness (support, confidence, lift etc.)[1]; all of these measures of interestingness use different methods (frequency, probability and counts) to calculate frequent itemsets in data sets. The frequency of items represents basic interestingness in association rules. A main origin of measures of interestingness is from common mathematical and information theories such as found in statistics, e.g., Yule's Q method, Yule's Y method, correlation coefficient and odds ratio. Out of the 40 different measures of interestingness available in the literature, no single measure of interestingness is perfect to calculate interestingness in every ARM task. In this paper, based on

M. Song et al. (Eds.): DaWaK 2020, LNCS 12393, pp. 38–47, 2020.
https://doi.org/10.1007/978-3-030-59065-9_4

the diversity of measures proposed to date, we are identifying their roles, classifying their usefulness from several perspectives to start an extended discussion on different properties of measures of interestingness.

Issues in Selecting Measures of Interestingness in ARM

(i) A large number of measures of interestingness are available to choose and many of these measures are not useful in each ARM task.
(ii) The classical measures of interestingness generate a lot of rules, most of these rules are irrelevant and redundant in many scenarios.
(iii) Based on the meaning of measure of interestingness, it's hard to decide on the appropriate measure in a concrete decision support scenario.
(iv) Various interestingness evaluation methods seem not to be rationalized. Some literature seems to simply combine several kinds of interestingness evaluations to new kinds of measures.

This paper is structured as follows. In Sect. 2, we describe expectedness and unexpectedness with respect to the roles of different measures in ARM. Section 3 focuses on the different properties for selecting the right measures of interestingness. Section 4 presents the conclusions and future work.

2 Expectedness and Unexpectedness in ARM

A simple ARM task using classical measures for a data set containing d items potentially generates $3^d - 2^d + 1$ possible association rules and most of these association rules are expected, obvious and duplicate. Take association rules for the data items {Milk, Bread, Butter} as an example. In the association rule in Eq. (1), it can be easily understood that the association of these three items is rather obvious. In ARM, obvious or common association rules can be referred to as *expected* association rules.

$$\{Milk, Bread\} \Rightarrow \{Butter\} \tag{1}$$

The main objective of ARM is to find the interesting association rules, hidden patterns and – most importantly – unexpected association rules in the data set. The association rules generated using the following combination of {Milk, Diaper, Beer} is not as obvious andy more and creates a rather novel pattern of items; in ARM, these types of association rules can be identified as unexpected association rules:

$$\{Milk, Diaper\} \Rightarrow \{Beer\} \tag{2}$$

Based on the variety of definitions of interestingness, the interestingness of an association rule can be categorized via the following nine properties [8]: (1) conciseness, (2) coverage, (3) reliability, (4) peculiarity, (5) diversity, (6) novelty, (7) surprisingness, (8) utility and (9) actionability. Descriptions of all of these properties are summarized in Table 1. Based on these nine definitions of interestingness, the measures of interestingness can be classified into three major

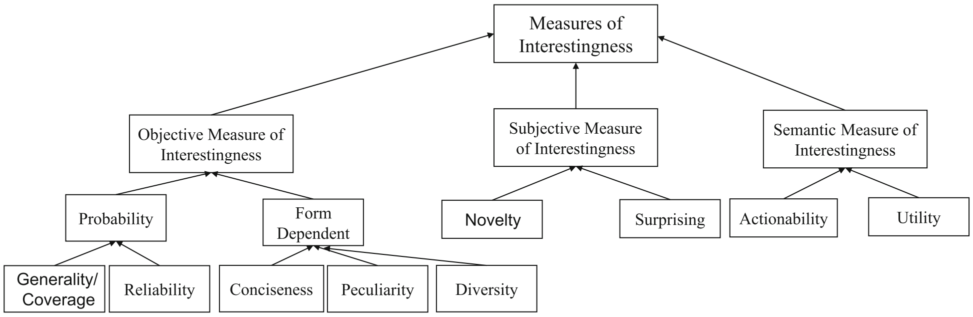

Measures of Interestingness
Objective Measure of Interestingness
Subjective Measure of Interestingness
Semantic Measure of Interestingness
Probability
Form Dependent
Novelty
Surprising
Actionability
Utility
Generality/ Coverage
Reliability
Conciseness
Peculiarity
Diversity

Fig. 1. Types of measures of interestingness.

categories: (1) objective measures of interestingness, (2) subjective measures of interestingness and (3) semantic measures of interestingness [14,18]. Figure 1 is showing all the different types of measures of interestingness.

2.1 Objective Measures of Interestingness for Expected Association Rules

Every transactional data set has some hidden patterns that can be easily identified by using predictive performance or statistical significance. In ARM, such kind of patterns may be referred to as expected patterns and can be computed using objective measures of interestingness. Objective measures mainly focus on the statistics and use statistical strength (probability, count etc.) to assess the degree of interest. As per the definition of interestingness, reliability, generality, conciseness, diversity and peculiarity are based only on data and patterns; therefore, these properties are the foundation of objective measures of interestingness [8]. Support, confidence, lift, conviction and improvement are some examples of objective measures of interestingness.

2.2 Subjective Measures of Interestingness for Unexpected Association Rules

Association rule mining based on common statistical approaches sometimes produces rather obvious or trivial rules. Therefore, the research of Padmanabhan and Tuzhilin [18] first explored the problem of interestingness through the notion of unexpectedness [18,19]. Subjective measures of interestingness usually determine the unexpected association rules in knowledge discovery. Unexpected patterns are opposite to the person's existing knowledge and contradict their expectations and existing knowledge [18].

Finding unexpected patterns in association rule mining is not an easy task, it needs a substantial amount of background information from domain experts [7]. For example, the association rule in Eq. (3) will rather not be considered interesting, even in cases where the rule has a particularly high support and high

Table 1. Interestingness properties in ARM, summarized and apobted from [2–6,9,10,12,15,16,19,21,23,24,26–28].

Property	Description
Conciseness [4,19]	A small number of attribute-value pairs in a pattern represents the conciseness of the pattern and a set of small number of patterns refers to a concise pattern set
Generality/Coverage [2,27]	The generality/coverage property in ARM covers most of the general patterns in ARM
Reliability [16,24]	Association rules or patterns based on common and popular relationships can be identified as reliable association rules or patterns
Peculiarity [3,28]	Peculiarity refers to unexpected behaviour of patterns. A pattern is said to be peculiar if it is significantly different from all other discovered patterns
Diversity [9]	For a pattern, diversity refers to the degree of differences between its elements; for a pattern set, diversity refers to the degree of differences in between the patterns
Novelty [21]	Combinations of unexpected items which create a pattern unknown to a person are known as novel patterns in ARM. These types of patterns can be discovered but can not be identified easily
Surprisingness [5,10,23]	Patterns which are opposite to a person's existing knowledge or expectations or create contradictions are known as surprising patterns in ARM
Utility [6,15]	Patterns which contribute to reaching a goal are called patterns with utility. Patterns with utility allow the user to define utility functions to get particular information from data
Actionability/Applicability [12,26]	Patterns with actionability allow a person to do a specific task for their benefits. These types of patterns usually reflect the person's action to solve a domain problem [12]

confidence, because the relationship expressed by the rule might be rather obvious to the analyst. As opposed to this, the association rule between *Milk* and *Shaving Blades* in Eq. (4) might be much more interesting, because the relationship is rather unexpected and might offer a unique opportunity for selling to the retail store.

$$\{Bread\} \Rightarrow \{Milk\} \tag{3}$$

$$\{Milk\} \Rightarrow \{Shaving\ Blades\} \tag{4}$$

Unexpectedness in Association Rule Mining. Many different definitions of unexpectedness have been proposed in the literature. In [18], unexpectedness has been defined with respect to association rules and beliefs. An association rule $P \Rightarrow Q$ is unexpected in regards to the belief $X \Rightarrow Y$ on a data set D if it follows the following rules:

- (i) $Q \wedge Y \models FALSE$ (This property states that Q and Y logically contradict each other.)
- (ii) This property states that set $P \wedge X$ has a large subset of tuples in the data set D.
- (iii) Rule $P, X \Rightarrow Q$ holds. As per the property (i), Q and Y logically contradict each other, therefore it logically follows that $P, X \Rightarrow \neg$ Y.

2.3 Semantic Measures of Interestingness

In ARM, semantic measures are a special kind of subjective measures of interestingness which include utility, application-specific semantics of patterns and domain knowledge of the person.

Utility: A utility function reflects the clear goal of the user. For example, to check the occurrence of a rare disease, a doctor might select association rules that correspond to low support rules over those with higher. A user with additional domain knowledge can use a utility-based approach. The domain knowledge of the user does not relate to his personal knowledge and expectations from data.

Actionability: In ARM, there is no widespread way to measure the actionability, i.e., it is up to the ability of an organization to do something useful with a discovered pattern; therefore, a pattern can be referred to as interesting if it is both actionable and unexpected. Generally, actionability is associated with a pattern selection strategy, whereas existing measures of interestingness are dependent on applications.

3 Properties for Selecting Objective Measures of Interestingness

It is important to care for applying consistent sets of measures of interestingness, as sometimes a wrong selection of measures may produce conflicting results. To select appropriate objective measures of interestingness, 15 key properties have been introduced in the literature [8,11,20,24]. Some of these properties are well known and some of the properties are not as popular. These properties are very useful to select appropriate measures for an ARM task.

Piatetsky-Shapiro [20] proposed three basic properties that need to be followed by every objective measure R

Property P1: "$R = 0$ if X, Y are two statistically independent data items, i.e., $\mathsf{P}(XY) = \mathsf{P}(X)\mathsf{P}(Y)$". This property states that accidentally occurred patterns or association rules are not interesting.

Property P2: “R monotonically increases with $\mathsf{P}(XY)$ when $\mathsf{P}(X)$ and $\mathsf{P}(Y)$ are same”. *P2* states that if a rule $X \Rightarrow Y$ have more positive correlation then the rule is more interesting.

Property P3: “R monotonically decreases when other parameters $\mathsf{P}(X)$, $\mathsf{P}(Y)$, $\mathsf{P}(X, Y)$ remain unchanged.”

Tan et al. [24] based on 2 × 2 contingency tables, Tan et al. [24] proposed five more properties for probability-based objective measures.

Property O1: “A measure of interestingness R is *symmetric under variable permutation* if it is preserved under the transformation $\Rightarrow_p$ of variable permutation, where $\Rightarrow_p$ is defined as matrix transpose as usual.”

$$\begin{array}{|c|c|c|}\hline & B & \neg B \\ \hline A & x & y \\ \hline \neg A & r & s \\ \hline \end{array} \Rightarrow_p \begin{array}{|c|c|c|}\hline & B & \neg B \\ \hline A & x & r \\ \hline \neg A & y & s \\ \hline \end{array}$$

Property O2: “R is same in row and column scaling. This property is known as the row-and-column scaling invariance.”

$$\begin{array}{|c|c|c|}\hline & B & \neg B \\ \hline \text{A} & \text{x} & \text{y} \\ \hline \neg A & r & s \\ \hline \end{array} \Rightarrow \begin{array}{|c|c|c|}\hline & B & \neg B \\ \hline A & k_3 k_1 x & k_4 k_1 y \\ \hline \neg A & k_3 k_2 r & k_4 k_2 s \\ \hline \end{array}$$

Property O3: “R is anti-symmetric under row and column permutation.”

$$\begin{array}{|c|c|c|}\hline & B & \neg B \\ \hline A & x & y \\ \hline \neg A & r & s \\ \hline \end{array} \Rightarrow \begin{array}{|c|c|c|}\hline & B & \neg B \\ \hline A & r & s \\ \hline \neg A & x & y \\ \hline \end{array}$$

Property O4: “R should remain same under both row and column permutation. This is inversion invariance which shows a special case of the row/column permutation where both rows and columns are swapped simultaneously.”

$$\begin{array}{|c|c|c|}\hline & B & \neg B \\ \hline A & x & y \\ \hline \neg A & r & s \\ \hline \end{array} \Rightarrow \begin{array}{|c|c|c|}\hline & B & \neg B \\ \hline A & s & r \\ \hline \neg A & y & x \\ \hline \end{array}$$

Property O5: “This property represents the null invariance.”

$$\begin{array}{|c|c|c|}\hline & B & \neg B \\ \hline A & x & y \\ \hline \neg A & r & s \\ \hline \end{array} \Rightarrow \begin{array}{|c|c|c|}\hline & B & \neg B \\ \hline A & x & y \\ \hline \neg A & r & s+k \\ \hline \end{array}$$

Lenca et al. [11] proposed five more properties to evaluate measures of interestingness. In these properties, Q1, Q4 and Q5 properties are preferred over the Q2, Q3 properties

Property Q1: “An interesting measure R is constant if there is no counterexample to the rule”. As per this property all the association rules with confidence 1 should have same interestingness value.

Property Q2: “R decreases with $\mathsf{P}(X\neg Y)$ in a linear, concave, or convex fashion around 0+.” This property describes that the value of interestingness decreases with respect to the counterexamples.

Property Q3: “R increases as the total number of records increases.”

Property Q4: “The threshold is easy to fix.” This property focuses on selecting the easy threshold to separate the interesting association rules from uninteresting association rules.

Property Q5: “The semantics of the measures are easy to express.” As per this property, semantics of the interestingness measures should be understandable.

Hamilton et al. [8] have also proposed two more properties to select the right measures of interestingness.

Property S1: “An interesting measure R should be an increasing function of support if the margins in the contingency table are fixed.”

Property S2: “An Interesting measure R should be an increasing function of confidence if the margins in the contingency table are fixed.”

3.1 Towards Selecting Optimal Measures of Interestingness

All three categories of measures (objective, subjective and semantic) consist of many different measures; therefore, it is very difficult to select appropriate measures for an ARM task. Table 2 might be a useful step in the selection of optimal measures of interestingness.

With respect to objective measures of interestingness, Tan et al. and Lenca et al. [11,24] proposed a ranking method to select measures. The ranking method is based on a specific data set that allows specific patterns having greatest standard deviations in all of the rankings. Lenca et al. [11] proposed also another approach to select measures; in this approach, a value and a weight is assigned to each important property in purpose of selecting measures. In the approach proposed by Vaillant et al. [25], objective measures of interestingness are grouped according to their properties and outcomes.

Table 2. Suggested approaches for selecting optimal measures of interestingness.

Objective Measures of Interestingness	Subjective Measures of Interestingness	Semantic Measures of Interestingness
Ranking method based on data sets [24]	Approaches based on formal specification of user knowledge [10, 13, 23]	Utility-based [22]
Ranking method based on properties of measures of interestingness [11]	Eliminating uninteresting patterns [21]	Actionable patterns [13]
Clustering method based on data sets [25]	Constraining the search space [17]	–
Clustering method based on properties of measures of interestingness [25]	–	–

In subjective measures of interestingness, user knowledge and data are the two crucial factors in deciding on optimal measures. Based on existing and vague knowledge of the user, Liu et al. [13] proposed different subjective measures. The approach proposed by Sahar et al. [21] is about eliminating uninteresting patterns; in this approach, there is no specific measure of interestingness. The method proposed by Padmanabhan et al. [17] is about constraining the search space , here, user belief is used as a constraint in mining association rules. In this method, a user's belief is mined as an association rules and if existing knowledge contradicts to the mined belief, it is referred to as a surprising pattern.

With respect to selecting optimal semantic measures of interestingness, [22] have proposed an approach that is about patterns with utility, here, "Interestingness (of a pattern) = probability + utility" [22]. In the actionability approach proposed by [13], a user provides some patterns in the form of fuzzy rules to represent both the possible actions and the situations in which they are likely to be taken.

4 Conclusion

In ARM, it is clear that no single measure of interestingness is suitable for all ARM tasks – a combination of subjective measures and objective measures seem to be the future in ARM. Selecting optimal measures of interestingness is still an open research problem. In this paper, we have conducted a preliminary study of properties that have been proposed to select optimal measures of interestingness.

We have summarized the role of expected and unexpected association rules in data mining and discussed the importance of the degree of user-involvement within the ARM process. Based on this preliminary work, we aim to design a user interface that supports the decision maker in selecting optimal measures of interestingness. The findings should also be helpful in efforts of designing new measures of interestingness in the future.

Acknowledgements. This work has been conducted in the project "ICT programme" which was supported by the European Union through the European Social Fund.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. ACM SIGMOD Record **22**(2), 207–216 (1993). https://doi.org/10.1145/170036.170072
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of VLDB'1994 - the 20th International Conference on Very Large Data Bases, p. 487–499. Morgan Kaufmann (1994)
3. Barnett, V., Lewis, T.: Outliers in Statistical Data. Wiley, 3rd edn (1994)
4. Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., Lakhal, L.: Mining minimal non-redundant association rules using frequent closed itemsets. In: Lloyd, J., et al. (eds.) CL 2000. LNCS (LNAI), vol. 1861, pp. 972–986. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-44957-4_65
5. Bay, S.D., Pazzani, M.J.: Detecting change in categorical data: mining contrast sets. In: Proceedings of KDD 1999 - the 5th ACM International Conference on Knowledge Discovery and Data Mining, pp. 302–306. ACM (1999)
6. Chan, R., Yang, Q., Shen, Y.D.: Mining high utility itemsets. In: Proceedings of ICDM'2003-the 3rd IEEE International Conference on Data Mining, pp. 19–26. IEEE, USA (2003)
7. Hartmann, S., Küng, J., Chakravarthy, S., Anderst-Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) DEXA 2019. LNCS, vol. 11706. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-27615-7
8. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: a survey. ACM Comput. Surv. **38**(3) 56–63 (2006). https://doi.org/10.1145/1132960.1132963
9. Hilderman, R.J., Hamilton, H.J.: Knowledge Discovery and Measures of Interest. Kluwer (2001)
10. Lee, Y.C., Hong, T.P., Lin, W.Y.: Mining association rules with multiple minimum supports using maximum constraints. Int. J. Approximate Reason. **40**(1–2), 44–54 (2005). https://doi.org/10.1016/j.ijar.2004.11.006
11. Lenca, P., Meyer, P., Vaillant, B., Lallich, S.: A multicriteria decision aid for interestingness measure selection. Technical report LUSSI-TR-2004-01-EN, École Nationale Supérieure des Télécommunications de Bretagne (2004)
12. Ling, C.X., Chen, T., Yang, Q., Cheng, J.: Mining optimal actions for profitable CRM. In: Proceedings of ICDM'2002 - the 2nd IEEE International Conference on Data Mining, pp. 767–770. IEEE (2002)
13. Liu, B., Hsu, W., Chen, S.: Using general impressions to analyze discovered classification rules. In: Proceedings of KDD 1997 - The 3rd International Conference on Knowledge Discovery and Data Mining, pp. 31–36. AAAI (1997)

14. Liu, B., Hsu, W., Chen, S., Ma, Y.: Analyzing the subjective interestingness of association rules. IEEE Intell. Syst. Appl. **15**(5), 47–55 (2000). https://doi.org/10.1109/5254.889106
15. Lu, S., Hu, H., Li, F.: Mining weighted association rules. Intell. Data Anal. **5**(3), 211–225 (2001)
16. Ohsaki, M., Kitaguchi, S., Okamoto, K., Yokoi, H., Yamaguchi, T.: Evaluation of rule interestingness measures with a clinical dataset on hepatitis. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS (LNAI), vol. 3202, pp. 362–373. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30116-5_34
17. Padmanabhan, B., Tuzhilin, A.: A belief-driven method for discovering unexpected patterns. In: Proceedings of KDD 1998 - The 4th International Conference on Knowledge Discovery and Data Mining, pp. 94–100. AAAI (1998)
18. Padmanabhan, B., Tuzhilin, A.: Unexpectedness as a measure of interestingness in knowledge discovery. Decision Support Syst. **27**(3), 303–318 (1999). https://doi.org/10.1016/S0167-9236(99)00053-6
19. Padmanabhan, B., Tuzhilin, A.: Small is beautiful: discovering the minimal set of unexpected patterns. In: Proceedings of KDD'2000 - The 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 54–63. AAAI (2000). http://doi.acm.org/10.1145/347090.347103
20. Piatetsky-Shapiro, G.: Discovery, analysis, and presentation of strong rules. In: Piatetsky-Shapiro, G., Frawley, W.J. (eds.) Knowledge Discovery in Databases, pp. 229–248. AAAI/MIT Press (1991)
21. Sahar, S.: Interestingness via what is not interesting. In: Proceedings of ACM SIGKDD'1999–The 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 332–336 (1999)
22. Shen, Y.D., Zhang, Z., Yang, Q.: Objective-oriented utility-based association mining. In: Proceedings of ICDM'2002–The 2nd IEEE International Conference on Data Mining, pp. 426–433. IEEE Computer Society (2002)
23. Silberschatz, A., Tuzhilin, A.: What makes patterns interesting in knowledge discovery systems. IEEE Trans. Knowl. Data Eng. **8**(6), 970–974 (1996). https://doi.org/10.1109/69.553165
24. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: Proceedings of ACM SIGKDD' 2002–The 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 2, pp. 32–41 (2002). https://doi.org/10.1145/775052.775053
25. Vaillant, B., Lenca, P., Lallich, S.: A Clustering of Interestingness Measures. In: Suzuki, E., Arikawa, S. (eds.) DS 2004. LNCS (LNAI), vol. 3245, pp. 290–297. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30214-8_23
26. Jensen, C.S., et al. (eds.) EDBT 2002. LNCS, vol. 2287. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45876-X
27. Webb, G.I., Brain, D.: Generality is predictive of prediction accuracy. In: Proceedings of PKAW'2002- Pacific Rim Knowledge Acquisition Workshop, pp. 117–130 (2002)
28. Zhong, N., Yao, Y.Y., Ohishima, M.: Peculiarity oriented multidatabase mining. IEEE Trans. Knowl. Data Eng. **15**(4), 952–960 (2003)

Appendix 2

R. Sharma, M. Kaushik, S. A. Peious, A. Bazin, S. A. Shah, I. Fister, S. B. Yahia, and D. Draheim. A novel framework for unification of association rule mining, online analytical processing and statistical reasoning. *IEEE Access*, 10:12792–12813, 2022

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 23, 2021, accepted January 7, 2022, date of publication January 12, 2022, date of current version February 4, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3142537

A Novel Framework for Unification of Association Rule Mining, Online Analytical Processing and Statistical Reasoning

RAHUL SHARMA[1], (Graduate Student Member, IEEE), MINAKSHI KAUSHIK[1], SIJO ARAKKAL PEIOUS[1], ALEXANDRE BAZIN[2], SYED ATTIQUE SHAH[1], IZTOK FISTER, JR.[3], (Member, IEEE), SADOK BEN YAHIA[4], AND DIRK DRAHEIM[1]

[1]Information Systems Group, Tallinn University of Technology, 12616 Tallinn, Estonia
[2]Lorraine Research Laboratory in Computer Science and Its Applications (LORIA), CNRS, Inria, Université de Lorraine, 54000 Nancy, France
[3]Faculty of Electrical Engineering and Computer Science, University of Maribor, 2000 Maribor, Slovenia
[4]Software Science Department, Tallinn University of Technology, 12616 Tallinn, Estonia

Corresponding author: Rahul Sharma (rahul.sharma@taltech.ee)

This work has been partially conducted in the project "ICT programme" (information and communications technology programme) which was supported by the European Union through the European Social Fund.

ABSTRACT Statistical reasoning was one of the earliest methods to draw insights from data. However, over the last three decades, association rule mining and online analytical processing have gained massive ground in practice and theory. Logically, both association rule mining and online analytical processing have some common objectives, but they have been introduced with their own set of mathematical formalizations and have developed their specific terminologies. Therefore, it is difficult to reuse results from one domain in another. Furthermore, it is not easy to unlock the potential of statistical results in their application scenarios. The target of this paper is to bridge the artificial gaps between association rule mining, online analytical processing and statistical reasoning. We first provide an elaboration of the semantic correspondences between their foundations, i.e., itemset apparatus, relational algebra and probability theory. Subsequently, we propose a novel framework for the unification of association rule mining, online analytical processing and statistical reasoning. Additionally, an instance of the proposed framework is developed by implementing a sample decision support tool. The tool is compared with a state-of-the-art decision support tool and evaluated by a series of experiments using two real data sets and one synthetic data set. The results of the tool validate the framework for the unified usage of association rule mining, online analytical processing, and statistical reasoning. The tool clarifies in how far the operations of association rule mining and online analytical processing can complement each other in understanding data, data visualization and decision making.

INDEX TERMS Association rule mining, data mining, online analytical processing, statistical reasoning.

I. INTRODUCTION

Decision support techniques play an essential role in today's business environment. Since the 17th century, statistical reasoning (SR) has been used extensively to shape business decisions [1] and it was the earliest method to draw insights from data. With the emergence of decision support systems (DSSs) in the 1970s [2], SR is frequently used in DSSs and decision support tools, just take SPSS (Statistical Package for the Social Sciences) [3] or SAS (Statistical Analysis System) [4] as examples. With the rise of information technology in the 1990s, online analytical processing (OLAP) [5] and association rule mining (ARM) [6] have emerged as powerful decision support techniques (DSTs) [7], both with their specific rationales, objectives, and attitudes. Over the years, both OLAP and ARM have gained massive ground in practice (Cognos, SAP-BW resp. RapidMiner, Orange – to name a few) and, similarly, massive attention in the research community. Unfortunately, both OLAP and ARM have been introduced together with their own genuine mathematical formalizations and developed their specific terminologies. This makes it hard to reuse results from one domain in another; in particular, it is not always easy to unlock the potential of statistical results in OLAP and ARM application scenarios. OLAP represents relational data [8] in multi-dimensional views using roll-ups, drill-downs, slices, dices, etc.

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wang.

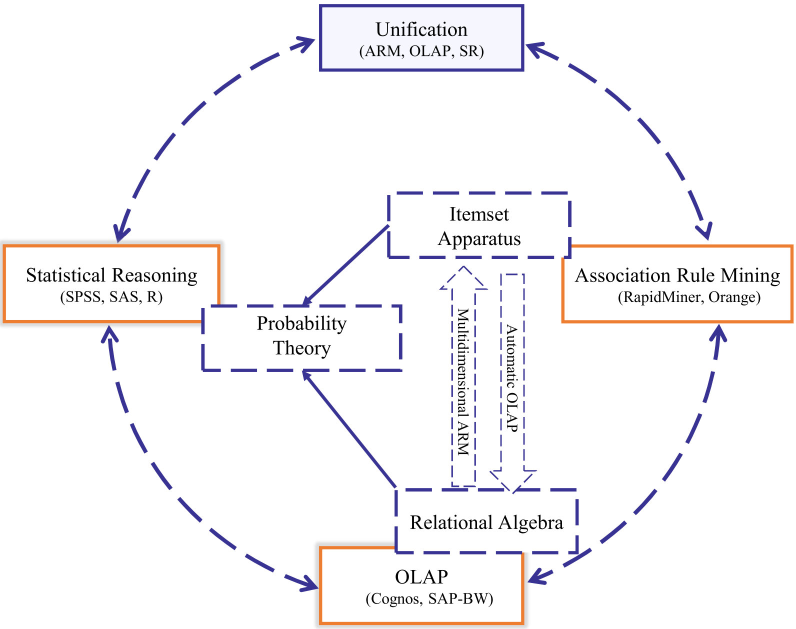


FIGURE 1. Semantic correspondences between association rule mining, online analytical processing and statistical reasoning.

In contrast, ARM relies on the notion of itemsets and frequent itemsets [9] in transaction databases. The correspondences between OLAP and ARM might seem rather simple, but it is neither fully elaborated in the state of the art nor implemented in practice. Because of the strong involvement of SR, OLAP, and ARM in decision-making, this paper aims to bridge the artificial gaps between them. We contribute by elaborating the semantic correspondences between the foundations of SR, OLAP, and ARM, i.e., probability theory, relational algebra, and the itemset apparatus.

In Fig. 1, a graphical representation of the process of determining the semantic correspondence between the SR, OLAP, and ARM is shown. The solid rectangles are used to indicate the selected DSTs, and the blue dashed lines rectangles are used to indicate the foundations of DSTs. The adoption of concepts in between OLAP and ARM (and vice versa) is referred to as automatic OLAP [10] and multi-dimensional ARM [11], respectively. In Table 1 and Table 2, we provide a list of abbreviations and frequently used symbols that are being used throughout the paper.

In the process of establishing semantic correspondences between the three DSTs, probability theory and, in particular, conditional expected values (CEVs) are at the center of our considerations. CEVs correspond to *sliced average aggregates* in OLAP and would correspond to potential *ratio-scale confidences* in a generalized ARM [12]. Based on the semantic correspondences between the DSTs, we are convinced that it is possible to design advantageous next-generation features of advanced decision support tools. A series of popular decision support tools is given in Fig. 2. We use software polls by KDnuggets [13] in the years 2017, 2018, and 2019 to measure the popularity of these tools. The popularity percentages of the tools demonstrate that a diverse range of tools is popular in practice and that they have also gained massive attention in the research community.

Kamber *et al.* [11] addressed the integration of OLAP and ARM as soon as 1997. They have provided the notion of metarule-guided mining, which entails utilizing user-defined rule templates to direct the mining process. Later, Han *et al.* [14] have proposed DBMiner for interactive mining, which provides a wide range of data mining operations such as association, generalization, characterization, classification, and prediction. We also identify several approaches for integrating different DSTs, and there is significant research specifically on the integration of OLAP and ARM in state-of-the-art. We appraise all of these decision support frameworks and different ways of integrating DSTs; however, the concept of semantic correspondences between DSTs is yet to be elaborated in state-of-the-art. A detailed discussion on a variety of decision support frameworks and various approaches for the integration of DSTs is given in Sect. II. Elaborating the semantic correspondences between DSTs will be helpful to fill the artificial gaps between DSTs. Furthermore, it can enable decision-makers to work with cross-platform decision support tools and check their results from different viewpoints.

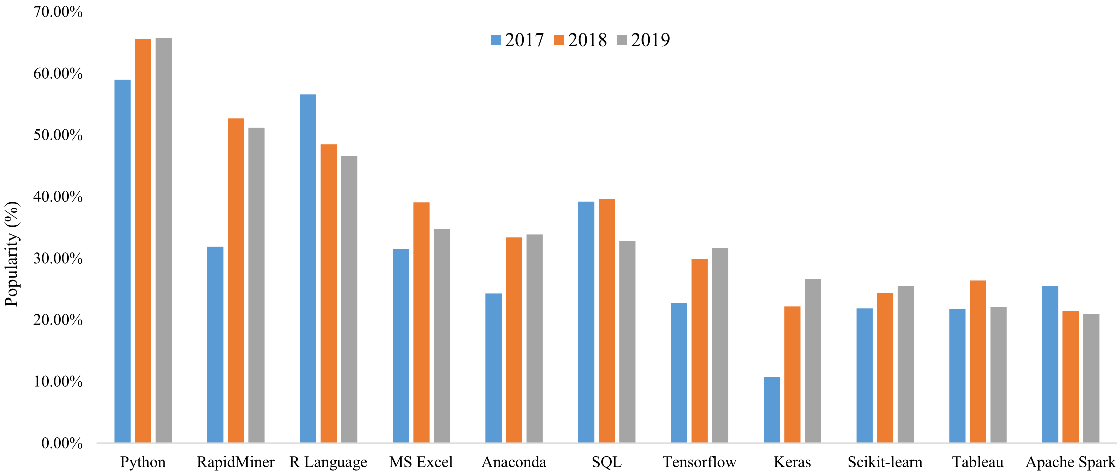


FIGURE 2. A series of popular decision support tools, together with their polarities according to opinion polls by KDnuggets [13] in 2017, 2018, and 2019.

TABLE 1. Abbreviations and acronyms.

Abbreviations	Summary
ACIF	All combination Influencing Factor
ARM	Association Rule Mining
CEVs	Conditional Expected Values
CFQs	Constrained Frequent Set
CAP	Consistency, Availability and Partition Tolerance
CBR	Case Based Reasoning
DIRECT	Discovering and Reconciling Conflicts
DMDSS	Data Mining Decision Support System
DST	Decision Support Technique
DSS	Decision Support System
FIA	Freedom of Information Act
GBP	Grid Based Pruning
GUI	Graphical User Interface
IADSS	Intelligent Agent Assisted DSS
IDSS	Integrated Decision Support System
KDD	Knowledge Discovery in Databases
MSMiner	Multi-strategy Data Mining Platform
NJ	New Jersey
OLAP	Online Analytical Processing
PVM	Parallel Virtual Machine
SAS	Statistical Software Suite
SAP	System Applications and Products in Data Processing
SAP-BW	SAP Business Warehouse
SPSS	Statistical Package for the Social Sciences
SR	Statistical Reasoning
uARMSolver	universal Association Rule Mining Solver
UDS1	User Defined Dataset

TABLE 2. List of frequent symbols.

Notation	Summary
$\mathbb{N}$	The set of natural numbers
$\mathbb{R}$	The set of real numbers
$\mathbb{B}$	Boolean values
$\mathbb{D}$	Discrete values
P	Probability
Σ	Events
Ω	Set of outcomes
T	Set of Transaction
D	OLAP Cube Dimension
σ	Relational Algebra
Δ	OLAP cube domain

The research on elaborating semantic correspondences between the three DSTs is significant due to the following reasons:

1) DSTs are developed independently for intended user groups and intended use cases.
2) Specific terminologies and functions of DSTs create artificial gaps between them and their tools.
3) Interpretation of results from one DST domain to another is not easily possible.
4) Artificial gaps between DSTs force decision-makers to use a variety of DSTs and decision support tools.
5) Various approaches for integrating DSTs are discussed in state of the art; however, correspondences between DSTs are obfuscated.

We observed that elaborating semantic correspondence between DSTs is necessary to bridge various artificial gaps between them. Therefore, in this paper, we elaborate semantic correspondences between the foundations of SR, OLAP, and ARM, i.e., between probability theory, relational algebra, and the itemset apparatus. In particular, we formally establish the correspondence between (i) the support of an itemset and the probability of a corresponding event and (ii) the confidence of an association rule and the conditional probability of two corresponding events. And (iii), the OLAP average aggregate function turns out to correspond to conditional expected values, which closes the loop between ARM, OLAP, and probability theory with respect to the most important constructs in ARM and OLAP.

Based on the semantic correspondences between the DSTs, we propose a novel framework for the unification of DSTs. The framework provides a way to develop various next-generation decision support tools. To validate the proposed framework, we implement a sample tool by combining the operations of three DSTs. The tool's outcomes establish semantic correspondence between SR, OLAP, and ARM and provide various useful data visualization methods. The tool is implemented on ASP.NET. In the tool, we use *'all combinations of influencing factors'* (ACIF) function to select the target column and influencing factors to generate all possible combinations of data items. The programming code and other instructions on how to use the proposed tool are available in the GitHub repository [15]. We have named the tool *grand report* [12], [16]; a *grand report* provides a complete printout of generalized association rules, which can also be seen as the entire unfolding of a pivot table [17]. An instance of the tool is hosted and available on the web.[1] The tool is straightforward to use, and it provides unified usages of DSTs.

The key contributions of the paper are as follows:

1) Elaboration of semantic correspondences between the three DSTs, i.e., SR, OLAP, ARM, and their foundations, i.e., probability theory, relational algebra, and the itemset apparatus, respectively.
2) We characterize to what extent and how far SR, OLAP, and ARM can be considered synonymous.
3) A novel framework for the unification of DSTs is presented to develop next-generation decision support tools.
4) A sample tool is presented to implement the unification of DSTs. The tool provides unified usages of DSTs.
5) The tool is tested on various datasets and compared to a state-of-the-art decision support tool. The comparison and the tool's outcome demonstrate the tool's superior performance.

The paper is organized as follows: In Sect. II, we review current work related to the unification of SR, OLAP, and ARM. Then, in Sect. III, we discuss the main concepts of mainstream SR, OLAP, and ARM. In Sect. IV, we elaborate semantic correspondences between the foundations of SR, OLAP, and ARM, i.e., probability theory, relational algebra, and the itemset apparatus. Subsequently, in Sect. V, we provide the framework for the unification of SR, OLAP, and ARM. A description of its implementation and experiments to showcase the relevance of the proposed framework are given. Finally, a discussion on future work and a conclusion are provided in Sect. VI and Sect. VII, respectively.

II. EXISTING WORK

In this section, previous work related to semantic correspondences between DSTs and various approaches for the integration of DSTs is explored.

The classical DSSs [2] were developed to assist managerial decisions by presenting several combinations of information. With the emergence of OLAP [5], knowledge discovery in databases (KDD) [36] and ARM [6], [37], many authors have proposed a variety of advanced DSSs. In the 1990s, web-based DSSs have been very popular [38]. Later, organizations have started taking advantage of different DSTs in DSSs [19]. We examine eighteen different research articles that discuss the integration of DSSs with different DSTs. A summary of these articles is given in Table 3. Wang [18] presented a novel architecture to integrate KDD techniques into existing DSSs. The authors have discussed the integration of different KDD techniques in group DSSs via three different types of decision support agents. In 2002, Fan *et al.* [19] provided a simple classification scheme for data value conflicts and presented an approach for discovering data conversion rules from data automatically. Bolloju *et al.* [20] provided a method for combining decision support and knowledge management to present an integrative framework for developing enterprise decision support environments. They used *model mart* and *model warehouse* as repositories.

In 2007, Rupnik *et al.* [24] discussed a method for combining DSS and data mining methods. The authors developed a data mining decision support system (DMDSS) that incorporates classification, clustering, and association rules. To investigate the use of data mining technology in DSS, Charest *et al.* [28] presented a theoretical, conceptual, and technological framework for the development of an intelligent data mining assistant by employing case-based reasoning and formal DL-ontology paradigms. Zhuang *et al.* [29] proposed a novel methodology to integrate data mining and case-based reasoning to develop a pathology test ordering system. In this paper, data mining concepts were used to extract the knowledge from past data, and then it was used in decision support.

In 2010, Liu *et al.* [30] conducted a survey to determine the efforts being made to develop an integrated decision support system (IDSS). IDSS combines four DSTs: knowledge-based systems, data mining, intelligent agents, and web technology. IDSS assists users in interpreting decision alternatives, and it also discovers hidden interesting patterns in large amounts of data using data mining tools. Gandhi *et al.* [39] demonstrated a DSS architecture (DSSA) that combines various data mining techniques. In this architecture, data mining tools were used to identify a set of features and patterns that domain experts can use to make decisions.

The majority of these works are inclined towards developing new DSSs and integrating DSSs with DSTs. However, the concept of semantic correspondences between DSTs is not discussed in any of these works. Therefore, we also explore the state of the art for the integration of OLAP and ARM. Some of these works focus on intra-dimensional association rules, while others are concerned with inter-dimensional association rules. Almost all intra-dimensional approaches use repeated predicates from a single data dimension.

[1] http://grandreport.me

TABLE 3. Existing approaches for the integration of different decision support techniques in DSSs.

Study	Year	Approach	Summary
Wang et al. [18]	1997	Intelligent agent-assisted DSS (IADSS)	A novel architecture is presented to integrate different KDD techniques in classical DSSs.
Fan et al. [19]	2002	A method for mining data value conversion rules from various data sources	In the process of integrating business data from multiple sources, a system called Discovering and Reconciling Conflicts (DIRECT) was presented .
Bolloju et al. [20]	2002	For the next generation DSSs, authors presented Integrating knowledge management into enterprise environments.	The authors proposed a method for combining decision support and knowledge management for developing enterprise decision support environments.
Heinrichs et al. [21]	2003	Integrating web-based data mining tools with DSS for knowledge management	Authors have highlighted how the knowledge workers in organizations can integrate data mining tools into their information and knowledge management requirements.
Cho et al. [22]	2003	Data mining for selection of insurance sales agents	For insurance managers, an intelligent DSS, Intelligent Agent Selection Assistant for Insurance, was presented.
Jukić et al. [23]	2006	Exploration of a large data warehouse using qualified association-rule mining	Mine fact tables and captures the correlations in data within data warehouses.
Rupnik et al. [24]	2007	DMDSS: a data mining-based DSS that combines data mining and decision support.	To support decision support procedures, a data mining-based DSS was provided.
March et al. [25]	2007	Integrated DSS: A data warehousing perspective	The authors have examined how data warehouses may be used for integration, implementation, intelligence, and innovation.
Shi et al. [26]	2007	MSMiner	The authors have presented an OLAP platform based on a data warehouse and integrated it with different data mining algorithms.
Domenica et al. [27]	2007	Stochastic programming and scenario generation within a simulation framework: An information systems perspective	This study discussed how sophisticated models and software realizations might be integrated with information systems and DSS tools.
Charest et al. [28]	2008	Intelligent data mining assistant using case-based reasoning	To study effectively deploying DM technology, the authors showed a framework for an intelligent data mining assistant by using case-based reasoning and formal DL ontology paradigms.
Zhuang et al. [29]	2009	Data mining and case-based reasoning for intelligent decision support	The authors used data mining and CBR approaches to provide information from previous data.
Liu et al. [30]	2010	IDSS	By utilizing data mining methods, IDSS assists users in interpreting decision alternatives and discovering hidden patterns in massive amounts of data.
Peng et al. [31],	2011	Incident information management framework	The authors presented a three level framework with three modules: data mining module, multi-criteria-decision-making module and data integration module.
Ltifi et al. [32].	2013	KDD-based human-centered design strategy for designing DSS	Based on KDD and other human-centered design principles, a human-centered design strategy for designing dynamic DSS was proposed
Dong et al. [33]	2014	A framework of Web-based decision support systems for portfolio selection with OLAP and PVM	The authors concentrated on developing a framework for a Web-based DSS.
Fister et al. [34]	2020	uARMSolver a framework	The paper introduces uARMSolver, a new software framework for ARM.
Aidan et al. [35]	2021	Knowledge Graph	Authors demonstrate how knowledge can be represented and extracted using a combination of deductive and inductive procedures.

A summary of different OLAP and ARM integration approaches is given in Table 4.

In 1998, Ng *et al.* [41], and Zhu [10] have proposed different ways to integrate ARM and OLAP together; however, their research was centered towards multi-dimensional ARM, automatic OLAP, and other specific sets of problems. The mainstream ARM was developed to find frequent items, while OLAP represented a multi-dimensional view of data using different OLAP operations. Therefore, the popularity of ARM for transactional datasets and the progress of OLAP [44] in a multi-dimensional environment attracted many authors to propose possible ways to integrate the ARM and OLAP. In 1997, Kamber *et al.* [11] first addressed the relationship between ARM and OLAP and proposed a meta-rule-guided mining approach for mining association rules from a multi-dimensional data cube. In this paper, Kamber *et al.* [11] have presented four algorithms that explore an OLAP data cube for meta-rule-guided mining of multi-dimensional association rules. Imielinski *et al.* [40] have presented cubegrades, a generalization of association rules which display how a set of measures (aggregates) is affected by specializing (rolldown), generalizing (roll-up) and mutating (which is a change in the cube's dimensions). In this paper, cubegrades are shown as more expressive than association rules in capturing associations and trends.

To support the adhoc mining in association rules, Lakshmanan *et al.* [42] proposed an idea of constrained frequent set queries (CFQs) and extended the architecture proposed by Ng *et al.* [41]. In addition, they introduced a new notion of quasi-succinctness and developed a heuristic technique for non-quasi-succinct constraints. Ng *et al.* [41] proposed architecture for exploratory mining of association

TABLE 4. Integration of OLAP and ARM in data mining.

Integration of ARM with OLAP	Approach	Summary
Kamber et al. [11]	Metarule-guided mining approach	Authors started by looking at the relationship between ARM and OLAP and then proposed a meta-rule-guided mining approach for extracting association rules from a multi-dimensional data cube.
Han et al. [14]	DBMiner: a software for various data mining techniques	DBMiner provides a way to combine data mining techniques with database technologies to uncover knowledge at different levels.
Imielinski et al. [40]	Cubegrades	To mine multi-dimensional association rules, a new concept of cubegrades with a novel grid-based pruning (GBP) technique was proposed as a generalized ARM technique. cubegrades was also shown as more expressive than associations rules in capturing associations and trends.
Raymond et al. [41]	Multidimensional Architecture, CAP algorithm, antimonocity, and succinctness	An architecture was proposed for multi-dimensional data mining, and CAP algorithm was used for the maximum degree of pruning. Two new rule pruning properties, antimonocity, and succinctness, were proposed to push the constraints deep inside the mining process.
Lakshmanan et al. [42]	Constrained Frequent Set (CFQs), the notion of quasi-succinctness, a heuristic technique for non-quasi-succinct constraints.	An idea of constrained frequent set queries (CFQs) was proposed In addition, they introduced a new notion of quasi-succinctness and developed a heuristic technique for non-quasi-succinct constraints.
Zhu et al. [10]	Multi-dimensional and Multi-level ARM methods	Proposed online analytical mining of association rules using the concept of multi-dimensional and multi-level ARM. Here, associations are divided into intra-dimensional, inter-dimensional, and constrained-based associations.
Nguyen et al. [43]	Exclusive Confidence and Support was proposed with a new algorithm	The authors proposed an architecture that allows constraint-based and human-centered exploratory mining of association rules. Two new measures, exclusive confidence and natural confidence, were discussed.

rules that is constraint-based and human-centered. To push the constraints deep inside the mining process, this paper presents a new algorithm (CAP) and two new rule pruning properties; antimonocity and succinctness. To generalize ARM within arbitrary n-ary relations and boolean tensors, Nguyen *et al.* [43] proposed exclusive confidence and natural confidence measures. They have also designed a complete, scalable algorithm that computes the exclusive measures. Kamber *et al.* [11] extended the constrained gradient analysis "cubegrades" presented by Imielinski *et al.* [40]. In this paper, the authors have addressed various issues and methods on efficient mining of multi-dimensional, constrained gradients in multi-dimensional data cubes. They have also defined the constraints as significant constraints, probe constraints, and gradient constraints.

Zhu [10] proposed online analytical mining of association rules and presented a step-by-step method and algorithm for inter-dimensional ARM, intra-dimensional ARM, and hybrid ARM. Based on OLAP technologies, they also designed a method to perform multi-level ARM. Chen *et al.* [45] developed an OLAP and data warehousing-based platform for weblog records (WLRs), which supports multi-level and multi-dimensional ARM. Finally, Cerf *et al.* [46] have presented an n-array algorithm for n-array relations, which was used to extract constrained-based closed n-sets.

In the state of the art, integration of DSTs and DSSs frameworks are broadly discussed. However, the correspondences between the foundation of DSTs are obfuscated. Therefore, we aim to elaborate semantic correspondences between the foundations of the three popular DSTs and bridge the artificial gaps between them.

III. PRELIMINARIES

This section provides background information about the three popular DSTs, i.e., SR, OLAP, ARM and their foundation, i.e., probability theory, relational algebra, and itemset mining. In Sect. III-A, we discuss the concepts of SR. Then, in Sect. III-B, the concepts of classical ARM are discussed, and in Sect. III-C, we discuss the basic concepts of OLAP.

A. STATISTICAL REASONING (SR)

With the development of probability theory [1] by thinkers like Gerolamo Cardano, Blaise Pascal, and Pierre de Fermat, statistics has evolved as an essential framework for developing DSS [47] and DSTs; therefore, most of the DSTs have been developed with the core concepts of SR. Since 1970, extensive use of computer systems has made it possible to do large statistical computations that have not been possible manually. In the 19^{th} and 20^{th} centuries, statistics had its victory by evolving into the primary scientific tool – think about classical thermodynamics and its elaboration through statistical mechanics and quantum physics. In the natural sciences, statistics have become the necessary foundation in economics, and many Nobel prizes correspond with the probabilistic variants of game theory. So, it could be said that statistics is the language of science. However, even more, statistics was a crucial driver in the industrial revolution, by helping to optimize production, think about Student's t-distribution.

Moreover, statistics is at the core of optimizing production; think of Six Sigma alone. All this is true, but since 1970, we have seen the next wave of SR. Statistics has left the scientific laboratories and entered the everyday decision-making

TABLE 5. Types of input data used in various decision support techniques.

Decision Support techniques	Types of Input data
Statistical Reasoning: Classification	$X_1 : T_1 \times \ldots \times X_m : T_m$
Statistical Reasoning: Multivariate Data Analysis (Regression)	$X_1 : \mathbb{R} \times \ldots \times X_m : \mathbb{R}$
Online Analytical Processing (OLAP)	$\underbrace{X_1 : D_1 \times \ldots \times X_m : D_m}_{\text{discrete}} \times \underbrace{Y_1 : \mathbb{R} \times \ldots \times Y_{m'} : \mathbb{R}}_{\text{numerical}}$
Association Rule Mining (Bitmap)	$X_1 : \mathbb{B} \times \ldots \times X_m : \mathbb{B}$
Association Rule Mining in Tools	$\underbrace{X_{1_1} : \mathbb{B} \times \ldots \times X_{n_1} : \mathbb{B}}_{X_1 : D_1} \times \cdots \times \underbrace{X_{m_1} : \mathbb{B} \times \ldots \times X_{n_m} : \mathbb{B}}_{X_m : D_m}$

processes in our organizations. Here, SR is the tool of highly specialized experts in highly specialized tasks but becomes available to a broader range of decision-makers. This movement is precisely about what has been expressed by ''The Future of Data Analysis'' by Tukey [48]. It means that systematic decision-making becomes more and more pervasive. In our opinion, this also explains the emergence of ARM and OLAP, which are two immensely successful approaches that complement, extend (but also overlap) the established SR toolkit. Moreover, the journey has just begun, as the current interest in *data science* proves – in 2015, Donoho [49] showed the evolution of data science from statistics. In Table 5, we provide different combinations of data used in SR, OLAP, and ARM. $\mathbb{R}$ is used to represent numerical type data, $\mathbb{D}$ is used to represent discrete type data and $\mathbb{B}$ is used to represent bitmap data.

B. ASSOCIATION RULE MINING (ARM)

To understand the relationship between different data items in transactional datasets and to find out interesting patterns and correlations, Agrawal *et al.* [6] presented the central concept of ARM using binary representations of data items as shown in Table 5. However, ARM is also presented for numerical data items as quantitative ARM [50], numerical ARM [51], [52].

ARM is highly effective in discovering relations and interesting associations among data items using different measures of interestingness [6], [53] and it is a prevalent technique that plays a crucial role in market basket data analysis, bioinformatics, ocean, land, and medical diagnosis.

In the original settings, association rules are extracted from transactional datasets composed of a set $I = \{i_1, \ldots, i_n\}$ of n binary attributes called *items* and a set $D = \{t_1, \ldots, t_n\}$, $t_k \subseteq I$, of *transactions* called database. An *association rule* is a pair of itemsets (X, Y), often denoted by an implication of the form $X \Rightarrow Y$, where X is called the antecedent (or premise) and Y is called the consequent (or conclusion), $X \cap Y = \emptyset$. To select interesting association rules, the following are the most popular measures of interestingness in ARM.

Definition 1: The *Support* of an itemset X with respect to a set of transactions T, denoted by *Supp*(X), is the ratio of transactions that contain all items of X (number of transactions that satisfy X) [54]:

$$\text{Supp}(X) = \frac{|\{t \in T \mid X \subseteq t\}|}{|T|}$$

Definition 2: The *confidence* of an association rule $X \Rightarrow Y$ concerning a set of transaction T, denoted by $\mathit{Conf}(X \Rightarrow Y)$ is the percentage of transactions that contains X which also includes Y. Technically, the confidence of an AR is an estimation of the conditional probability of Y over X:

$$\mathit{Conf}(X \Rightarrow Y) = \frac{\mathit{Supp}(X \cup Y)}{\mathit{Supp}(X)}.$$

Definition 3: The *lift* of an association rule $X \Rightarrow Y$, denoted by $\mathit{Lift}(X \Rightarrow Y)$, is used to measure misleading rules that satisfy minimum support and minimum confidence threshold. The Lift measure is also used to calculate the deviation between an antecedent X and a consequent Y, which is the ratio of the joint probability of X and Y divided by the product of their marginal probabilities.

$$\mathit{Lift}(X \Rightarrow Y) = \frac{\mathit{Supp}(X \cup Y)}{\mathit{Supp}(X) \times \mathit{Supp}(Y)}$$

In ARM, when the number of association rules is too large to be presented to a data mining expert or even treated by a computer, measures of interestingness can filter the interesting association rules. After support, confidence, and lift, more than fifty different measures of interestingness are in the literature [53], [55], [56]. These measures of interestingness are discussed in detail in the literature [57], [58]. Initially, ARM was limited to large transactional datasets. Still, later, Han *et al.*, Lu *et al.*, Imielinski *et al.*, and Nguyen *et al.* [40], [43], [59], [60] presented different views on multi-level and multi-dimensional ARM. Over the years, different ARM frameworks [34] and the use of ARM in varied application scenarios [61], [62] have also been discussed in the state of the art [63].

1) MULTIDIMENSIONAL VIEW OF ARM

More recently, ARM has been adapted to the multidimensional case [43] and multitask-based ARM [64]. In multidimensional setting of ARM, datasets are composed of a set $\mathcal{D} = \{S_1, \ldots, S_n\}$ of *dimensions*, and an n-ary relation between them, i.e. they are formally tuples $(S_1, \ldots, S_n, R)$ with $R \subseteq S_1 \times \cdots \times S_n$. In "Multitask-based" ARM, highly frequent association rules for different ARM tasks are referred as "single-task" rules which are Later combined together to generate the global results, i.e, "multitask rules".

Multidimensional association rules are rules between two so-called *associations* that generalize the notion of itemset. They are defined as the Cartesian products of subsets of dimensions. The set of dimensions used in an association X is called its *domain* and is noted $dom(X)$. For example, $X = \{Milk, Bread\} \times \{Winter\}$ is an association on the domain $dom(X) = \{products, seasons\}$. We use $\pi_{S_i}(X)$ to denote the projection of the association X on the dimension S_i, e.g. $\pi_{products}(X) = \{\text{Milk, Bread}\}$ and $\pi_{seasons}(X) = \{\text{Winter}\}$.

In the multi-dimensional case, the generalization of the notion of support is the *relative support*. The *support of an association X relative to a set $D \supseteq dom(X)$ of dimensions* is defined as

$$Supp_D(X) = \left|\{t \in \prod_{S_d \in \mathcal{D}\setminus\{D\}} S_d \quad | \quad \exists u \in \prod_{S_i \in D\setminus dom(X)} S_i \text{ such that } \forall x \in X, x.u.t \in \mathcal{R}\}\right| \qquad (1)$$

Using the relative support, two variants of confidence, the *exclusive confidence* and *natural confidence* are defined for multidimensional association rules:

$$Conf_{natural}(X \Rightarrow Y) = \frac{Supp_{dom(X\cup Y)}(X \cup Y)}{Supp_{dom(X\cup Y)}(X)}$$
$$Conf_{exclusive}(X \Rightarrow Y) = \frac{Supp_{dom(X\cup Y)}(X \cup Y) \times P}{Supp_{dom(X)}(X)}$$

with $P = |\prod_{S_i \in dom(X\cup Y)\setminus dom(X)} \pi_{S_i}(Y)|$.

In Table 6, the multidimensional association rule $\{Milk\} \Rightarrow \{Bread\} \times \{Spring\}$ has a natural support of $\frac{1}{4}$ because

$$Supp_{\{products,seasons\}}(\{Milk, Bread\} \times \{Spring\}) = |\{c_2\}| = 1$$
$$Supp_{\{products,seasons\}}(\{Milk\}) = |\{c_1, c_2, c_3, c_4\}| = 4. \qquad (2)$$

This rule can also be expressed in first-order logic, i.e.

$$\{Milk\} \Rightarrow \{Bread\} \times \{Spring\} \equiv \forall X, Y, \neg purchase (X, Milk, Y) \vee (purchase(X, Bread, Spring) \wedge purchase(X, Milk, Spring)). \qquad (3)$$

C. ONLINE ANALYTICAL PROCESSING (OLAP)

Historically, OLAP is not a new idea; it has persisted over the decades. Initially, in 1962, Kenneth Iverson proposed the foundation of OLAP in his book "A Programming Language" [65]. In 1975, Information Resources Inc. launched the first OLAP product named "Express", which was acquired by Oracle Inc. in 1995. In 1993, Edgar F. Codd used the term OLAP and set up 12 policies for an OLAP product in his paper "Providing OLAP (Online Analytical Processing) to user-analysts: An IT mandate" [5]. In OLAP, it is essential to have a multi-dimensional cube. Therefore, we show a sample OLAP cube with three dimensions (D_1, D_2, D_3) in Fig. 3. Practically, an OLAP cube consists four types of functions; First, OLAP operations, i.e., RollUp, Drill Down, Slice, Dice, and Pivot. Second is aggregation operations, i.e., SUM, AVG, COUNT, MIN, MAX, calculate trends, ranking, percentiles, attribute-based grouping, compare aggregates, etc. The third is the OLAP operator, i.e., "Force" and "Extract," which convert a dimension into a measure and a measure into a dimension. Fourth is the capability to handle uncertain data within the OLAP model.

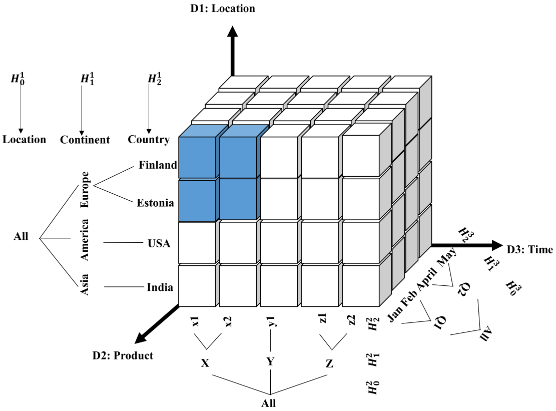


FIGURE 3. A sample OLAP data cube with three dimensions (D1: location, D2: product and D3: time).

IV. SEMANTIC CORRESPONDENCE BETWEEN SR, OLAP AND ARM

In this section, we establish semantic correspondence between SR, OLAP, and ARM. We use probability theory with conditional expected values (CEVs) as the center of our mappings. First, we provide semantic correspondence between SR, i.e., probability theory and ARM, and then we provide semantic correspondence between SR and OLAP.

Definition 4 (σ-Algebra): Given a set Ω, a *σ-Algebra Σ over* Ω is a set of subsets of Ω, i.e., $\Sigma \subseteq \mathsf{P}(\Omega)$, such that the following conditions hold true:

1) $\Omega \in \Sigma$
2) If $A \in \Sigma$ then $\Omega\backslash A \in \Sigma$
3) For all countable subsets of Σ, i.e., $A_0, A_1, A_2 \ldots \in \Sigma$ it holds true that $\bigcup_{i\in\mathbb{N}_0} A_i \in \Sigma$

Definition 5 (Probability Space): A *probability space* $(\Omega, \Sigma, \mathsf{P})$ consists of a set of outcomes Ω, σ algebra of (random) events Σ over the set of outcomes Ω and a probability function P: $\Sigma \to \mathbb{R}$, also called probability measure, such that the following axioms hold true:

TABLE 6. A multidimensional binary dataset in which *customers* (c_1 to c_4) buy *products* (Milk, Bread, Diapers, Beer) during *seasons* (Winter, Spring, Summer).

	Milk	Bread	Diapers	Beer	Milk	Bread	Diapers	Beer	Milk	Bread	Diapers	Beer
c_1	1	1	0	0	1	0	1	0	0	0	1	0
c_2	0	0	1	1	1	1	0	0	1	0	1	0
c_3	1	1	1	1	0	1	1	1	1	1	0	0
c_4	1	1	0	1	1	0	0	1	0	0	1	1
	Winter				Spring				Summer			

1) $\forall A \in \Sigma . 0 \leq \mathsf{P}(A) \leq 1$ (i.e., $\mathsf{P}\colon \Sigma \rightarrow [0, 1]$)

2) $\mathsf{P}(\Omega) = 1$

3) (Countable Additivity): For all countable sets of pairwise disjoint events, i.e., $A_0, A_1, A_2 \ldots \in \Sigma$ with $A_i \cap A_j = \emptyset$ for all $i \neq j$, it holds true that

$$\mathsf{P}\Big(\bigcup_{i=0}^{\infty} A_i\Big) = \sum_{i=0}^{\infty} \mathsf{P}(A_i)$$

Definition 6 (Conditional Probability): Given two events $\{X, Y\} \in \Sigma$ of *probability space* $(\Omega, \Sigma, \mathsf{P})$. If $\mathsf{P}(X) \neq 0$ then we define conditional probability of Y given X as:

$$\mathsf{P}(Y|X) = \frac{\mathsf{P}(X \cap Y)}{\mathsf{P}(X)}$$

Definition 7 (Expected Value): Given a real-valued discrete random variable $X\colon \Omega \rightarrow I$ with indicator set $I = \{i_0, i_1, i_2, \ldots, i_n\} \subseteq \mathbb{R}$ based on $(\Omega, \Sigma, \mathsf{P})$, the *expected value* $\mathsf{E}(X)$, or *expectation* of X (where E can also be denoted as E_P in so-called explicit notation) is defined as follows:

$$\mathsf{E}(X) = \sum_{n=0}^{\infty} i_n \cdot \mathsf{P}(X = i_n)$$

Definition 8 (Conditional Expected Value): Given a real-valued discrete random variable $Y\colon \Omega \rightarrow I$ with indicator set $I = \{i_0, i_1, i_2, \ldots\} \subseteq \mathbb{R}$ based on a probability space $(\Omega, \Sigma, \mathsf{P})$ and an event $X \in \Sigma$, the *expected value* $\mathsf{E}(Y)$ of Y conditional on X (where E can also be denoted as E_P in so-called explicit notation) is defined as follows:

$$\mathsf{E}(Y|X) = \sum_{n=0}^{\infty} i_n \cdot \mathsf{P}(Y = i_n | X) \tag{4}$$

A. ANCHORING ASSOCIATION RULE MINING IN PROBABILITY THEORY

We follow the concepts and notation and their formalization as originally introduced by Agrawal *et al.* in their 1993 paper [6] as closely as possible. First, there is a *whole itemset* $\mathfrak{I} = \{I_1, I_2, \ldots, I_n\}$ consisting of a *total number* n of items $I_1, I_2, \ldots, I_n$. A subset $X \subseteq \mathfrak{I}$ of the whole itemset is called an *itemset*. Next, we introduce the notion of a *set of transactions* T (*that fits the itemset* $\mathfrak{I}$) as a relation as follows:

$$T \subseteq \mathit{TID} \times \underbrace{\{0, 1\} \times \cdots \times \{0, 1\}}_{n-\text{times}} \tag{5}$$

Here, *TID* is a finite set of transaction identifiers. For the sake of convenience, we assume that it has the form $\mathit{TID} = \{1, \ldots, N\}$. Actually, we need to impose a uniqueness constraint on *TID*, i.e., we require that T is right-unique, i.e., a function given as,

$$T \in \mathit{TID} \longrightarrow \underbrace{\{0, 1\} \times \cdots \times \{0, 1\}}_{n-\text{times}} \tag{6}$$

Given (6), we have that N in $\mathit{TID} = \{1, \ldots, N\}$ equals the size of T, i.e., $N = |T|$. Henceforth, we refer to T interchangeably both as a relation and as a function, according to (5) resp. (6). For example, we use $t = \langle i, i_1, \ldots i_n \rangle$ to denote an arbitrary transaction $t \in T$; similarly, we use $T(i)$ to denote the i-th transaction of T more explicitly etc. Given this formalization of the transaction set T, it is correct to say that T is a binary relation between TID and the whole itemset. In that, $I_1, I_2, \ldots, I_n$ need to be thought of as column labels, i.e., there is exactly one bitmap column for each of the n items in $\mathfrak{I}$, compare with (5) and (6). Similarly, Agrawal *et al.* have called the single transaction a bit vector and introduced the notation $t[k]$ for selecting the value of the transaction t in the k-th column of the bitmap table (in counting the columns of the bitmap table, the *TID* column is omitted, as it merely serves the purpose of providing transaction identities), i.e., given a transaction $\langle tid, i_1, \ldots i_n \rangle \in T$, we define $\langle tid, i_1, \ldots i_n \rangle[k] = i_k$. Less explicit, with the help of the usual tuple projection notation π_j, we can define $t[k] = \pi_{k+1}(t)$. Let us call a pair $\langle \mathfrak{I}, T \rangle$ of a whole itemset $\mathfrak{I}$ and a set of transaction T that fits $\mathfrak{I}$ as described above an *ARM frame*. Henceforth, we assume an ARM frame $\langle \mathfrak{I}, T \rangle$ as given.

We have said that a transaction is a bit vector. For the sake of convenience, let us introduce some notation that allows us to treat a transaction as an itemset. Given a transaction $t \in T$ we denote the *set of all items that occur in* t as $\{t\}$ and we define it as follows:

$$\{t\} = \{I_k \in \mathfrak{I} \mid t[k] = 1\} \tag{7}$$

The {t} notation provided by (7) will prove helpful later, as it allows us to express properties about transactions without the need to use bit-vector notation, i.e., without the need to maintain item numbers k of items I_k.

Given an $I_j \in \mathfrak{I}$ and a transaction $t \in T$, Agrawal *et al.* says [6] that I_j *is bought by* t if and only if $t[j] = 1$. Similarly, we can say that t *contains* I_j in such case. Next, given an itemset $X \subseteq \mathfrak{I}$ and a transaction $t \in T$, Agrawal *et al.* says that t *satisfies* X if and only if $t[j] = 1$ for all $I_j \in X$. Similarly, we can say that t *contains all of the items of* X in such case. Next, we can see that t satisfies X if and only if $X \subseteq \{t\}$. Henceforth, we use $X \subseteq \{t\}$ to denote that t satisfies X.

Given an itemset $X \subseteq \Im$, the relative number of all transactions that satisfy X is called the *support of* X and is denoted as $Supp(X)$, i.e., we define:

$$Supp(X) = \frac{|\{t \in T \mid X \subseteq \{t\}\}|}{|T|} \tag{8}$$

Again, it makes perfect sense to talk about the support of an itemset X as the relative number of all transactions that each contain all of the items of X.

An ordered pair of itemsets $X \subseteq \Im$ and $Y \subseteq \Im$ is called an *association rule*, and is denoted by $X \Rightarrow Y$. Now, the relative number of all transactions that satisfy Y among all of those transactions that satisfy X is called the *confidence of* $X \Rightarrow Y$, and is denoted as $Conf(X \Rightarrow Y)$, i.e., we define:

$$Conf(X \Rightarrow Y) = \frac{|\{\, t \in T \mid Y \subseteq \{t\} \wedge X \subseteq \{t\}\,\}|}{|\{t \in T \mid X \subseteq \{t\}\}|} \tag{9}$$

Usually, the confidence of an association rule is introduced via support of itemsets as follows:

$$Conf(X \Rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X)} \tag{10}$$

It can easily be checked that (9) and (10) are equivalent.

B. SEMANTIC CORRESPONDENCE BETWEEN ARM AND SR)

Next, we map the concepts defined in ARM to probability theory. Given an ARM frame $F = \langle \Im, T \rangle$ next we map the concepts defined in ARM to probability space $(\Omega_F, \Sigma_F, \mathsf{P}_F)$. First, we define the set of outcomes Ω_F to be the set of transactions T. Next, we define Σ_F to be the power set of Ω_F. Finally, given an event $X \in \Sigma_F$, we define the probability of X as the relative size of X, as follows:

$$\Omega_F = T \tag{11}$$

$$\Sigma_F = \mathbb{P}(T) \tag{12}$$

$$\mathsf{P}_F(X) = \frac{|X|}{|T|} \tag{13}$$

In the sequel, we drop the indices from Ω_F, Σ_F, and P_F, i.e., we simply use Ω, Σ, and P to denote them, but always keep in mind that we actually provide correspondence from ARM frames F to corresponding probability spaces $(\Omega_F, \Sigma_F, \mathsf{P}_F)$. The idea is simple. Each transaction is modeled as an outcome and, as usual, also a basic event. Furthermore, each set of transactions is an event.

We step forward with item and itemsets. For each item $I \in \Im$ we introduce the *event that item* I *is contained in a transaction*, and we denote that event as $[\![I]\!]$. Next, for each itemset $X \subseteq \Im$, we introduce the *event that all of the items in* X *are contained in a transaction* and we denote that event as $[\![X]\!]$. We define:

$$[\![I]\!] = \{\, t \mid I \in \{t\}\,\} \tag{14}$$

$$[\![X]\!] = \bigcap_{I \in X} [\![I]\!] \tag{15}$$

As usual, we identify an event $[\![I]\!]$ with the characteristic random variable $[\![I]\!] : \Omega \longrightarrow \{0, 1\}$ and use $\mathsf{P}([\![I]\!])$ and $\mathsf{P}([\![I]\!] = 1)$ as interchangeable.

1) FORMAL CORRESPONDENCE OF ARM SUPPORT AND CONFIDENCE TO PROBABILITY THEORY

Based on the correspondence provided by (11) through (15), we can see how ARM *Support* and *Confidence* translate into probability theory.

Lemma 1 (Mapping ARM Support *to Probability Theory):* Given an itemset $X \subseteq \Im$, we have that:

$$Supp(X) = \mathsf{P}([\![X]\!]) \tag{16}$$

Proof: According to (15), we have that $\mathsf{P}([\![X]\!])$ equals

$$\mathsf{P}(\bigcap_{I \in X} [\![I]\!]) \tag{17}$$

Due to (14), we have that (17) equals

$$\mathsf{P}\Big(\bigcap_{I \in X} \{\, t \in T \mid I \in \{t\}\,\}\Big) \tag{18}$$

We have that (18) equals

$$\mathsf{P}(\{\, t \in T \mid \bigwedge_{I \in X} I \in \{t\}\,\}) \tag{19}$$

We have that (19) equals

$$\mathsf{P}(\{t \in T \mid X \subseteq \{t\}\}) \tag{20}$$

According to (13), we have that (20) equals

$$\frac{|\{t \in T \mid X \subseteq \{t\}\}|}{|T|} \tag{21}$$

According to (8), we have that (21) equals $Supp(X)$

Lemma 2 (Mapping ARM Confidence *to Probability Theory):* Given an itemset $X \subseteq \Im$, we have that:

$$Conf(X \Rightarrow Y) = \mathsf{P}\big(\, [\![Y]\!] \;\big|\; [\![X]\!]\,\big)$$

Proof: Omitted.

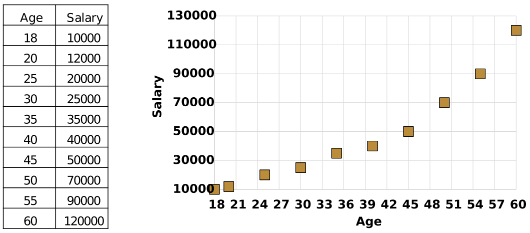
In Table 7, we provide one to one mapping in between the operations of ARM and SR, i.e., probability theory. A set of items in ARM $\Im = \{I_1, I_2, \ldots, I_m\}$ are equivalent to the set of events $\Im = \{I_1 \subseteq \Omega, \ldots, I_m \subseteq \Omega\}$ in probability theory. Transactions T in ARM are equivalent to the set of outcomes Ω in probability space $(\Omega, \Sigma, \mathsf{P})$. Support of an itemset X in ARM is equivalent to the relative probability of the itemset X. Confidence of an association rule $X \Rightarrow Y$ is equivalent to the conditional probability of Y in the presence of X.

C. ANCHORING OLAP IN PROBABILITY THEORY

Decision-makers are using OLAP to explore data in a multidimensional view. It helps to compute different aggregate summaries using various OLAP operations (COUNT, SUM, Drill-Down, Roll-up, Slice, Dice, etc.). For example, Fig. 4 demonstrates age and salary records in a two-dimensional space. In OLAP, data exploration starts from a high granularity level to a lower granularity level or vice versa. The sample data cube is given in Fig. 3 consists of *time*, *location* and *product* dimensions. An OLAP dimension comprises

TABLE 7. Semantic correspondences between association rule mining and statistical reasoning (probability theory).

Association Rule Mining Terminology	Statistical Reasoning (Probability Theory)
Set of Items $\Im = \{I_1, I_2, \ldots, I_m\}$	Set of Events $\Im = \{I_1 \subseteq \Omega, \ldots, I_m \subseteq \Omega\}$
Transactions $T \subseteq I_0 \times \underbrace{\{0,1\} \times \ldots \times \{0,1\}}_{m-times}$	Ω
$t \in T$ *satisfies* itemset $X \subseteq \Im$: $t[k] = 1 \Rightarrow I_k \in X$	$t \in \cap X$
Support of itemset $X \subseteq \Im$: $\mathrm{Supp}(X) = \frac{\lvert\{t \in T \mid X \subseteq t\}\rvert}{\lvert T\rvert}$	$\mathsf{P}(\bigcap_{I \in X} [\![I]\!])$
Confidence of association rule $X \Rightarrow Y$: $\mathrm{Conf}(X) = \frac{\mathrm{Supp}(X \cup Y)}{\mathrm{Supp}(X)}$ $X \subseteq T, Y \subseteq T$, usually: $X \cap Y = \emptyset, \lvert Y\rvert = 1$	Conditional Probability $\mathsf{P}(\,[\![Y]\!] \mid [\![X]\!]\,)$
Lift of $X \Rightarrow Y$ wrt the outermost margin: $\frac{\mathrm{Supp}(X \cup Y)}{\mathrm{Supp}(X) \times \mathrm{Supp}(Y)}$	$\frac{\mathsf{P}(\cap Y \mid \cap X)}{\mathsf{P}(\cap Y)}$



Age	Salary
18	10000
20	12000
25	20000
30	25000
35	35000
40	40000
45	50000
50	70000
55	90000
60	120000

FIGURE 4. A sample representation of age and salary records in two dimensional space.

organized attributes in a hierarchical structure to show the different data granularity levels. For example, the time dimension in Fig. 3 may have the following hierarchy: *Month* → *Quarter* → *Year*. Here, the dimension attribute *Year* shows a high level of granularity, and *Month* shows a lower level of granularity. Based on the sample OLAP cube given in Fig. 3, first, we provide standard notions and definitions and then provide semantic correspondence between OLAP and SR, i.e., probability theory.

1) OLAP CUBE: BASIC NOTATIONS AND DEFINITIONS

Let an OLAP cube C be a multi-dimensional data cube with four-tuple $C = \{\Delta, D, H, M\}$ where Δ represents the OLAP cube domain, D is a non-empty set of n dimensions, H is a set of dimension hierarchy and M is a non-empty set of quantitative measures, i.e., numerical or additive values of a cell. We have considered the following properties concerning the OLAP cube.

- In an OLAP cube C, dimension set $D = \{D_1, D_2 \ldots D_i \ldots D_n\}$, dimension D_i consists of a set of different hierarchy levels H_i, where $i \leq n$.
- A hierarchy level $H^i_j \in H_i$ is a non-empty set of members A_{ij}. $H^i_j (j \geq 0)$ is the j^{th} hierarchical level in D_i. E.g., in Fig. 3, the set of hierarchical level of dimension D_1 is $H_1 = \{H^1_0, H^1_1, H^1_2\} = \{Location, Continent, Country\}$, and in the dimension D_1, the set of members at level H^1_2 is $A_{12} = \{India, USA, Estonia, Finland\}$

Definition 9: Sub Cube: A sub cube C' is part of the main OLAP cube with a non-empty set D' of m dimensions. $D' = \{D_1, D_2 \ldots D_i \ldots D_m\}$ and $m \leq n$. According to D', a tuple $\{\Theta_1 \ldots \Theta_m\}$ is a sub cube C' if $D' \subseteq D$ and $\Theta_1 \subseteq A_{ij}$ for all $i \in \{1 \ldots m\}$ and $\Theta_i \neq null$.

E.g., If in Fig. 3, a dimension set $D' = \{D1, D2\} \in D$ is a sub cube then $(\Theta_1, \Theta_2) = \{Europe, x_1, x_2\}$ will be a sub cube.

Definition 10: Aggregate Measure: A Measure M in a data cube C is the SUM of measure M of all facts in the cube.

E.g., “Total Sales” in Fig. 3 can be evaluated by its sum-based aggregate measure. The aggregate expression $TotalSales(India, \{x_1, x_2, y_1\})$ represents the SUM of total sales turnover for the products (x_1, x_2, y_1) in India.

Definition 11: Intra Dimension Predicate: A dimension predicate A_i in a dimension D_i is its member as a value represented as $a_i \in A_{i_j}$.

E.g., In Fig. 3, a dimension predicate a_1 in dimension D_1 is $a_1 \in \{Asia, America, Europe\}$.

Definition 12: Inter Dimension Predicate: Let data cube C have a sub cube C’ with a non empty set of dimensions $D' = \{D_1, D_2 \ldots D_i \ldots D_m\}$ and $D' \subseteq D$. When the value of dimension predicates $\{A_1 \ldots A_m\}$ belongs to two or more dimensions where $(2 \leq m \leq n)$, then it is referred to as inter dimension predicates.

E.g., In Fig. 3, dimension predicate $\{a_1, a_2\} \in \{D_1, D_2\}$ then $a_1 \in \{Asia, America, Europe\}$ and $a_2 \in \{X, Y, Z\}$.

2) SEMANTIC CORRESPONDENCE BETWEEN OLAP AND SR

As discussed in Sect. III-C, an OLAP cube consists of various operations (Roll-Up, Drill-Down, Slice, Dice, Pivot, SUM, AVG, MIN, etc.). We have that the OLAP conditional operations (Slice, Dice, Drill- Down, Roll-up) on bitmap (Binary) columns correspond to conditional probabilities. Those conditional operations on numerical columns correspond to conditional expected values in probability theory. For example, we model a sample OLAP Table 8 in probability theory. We consider that Table 8 is equivalent to the set of outcomes.

TABLE 8. A sample OLAP table.

City	Profession	Education	Age Group	Freelancer	Salary
New York	Lawyer	Master	25–30	0	3, 800
Seattle	IT	Bachelor	18–25	1	4, 200
Boston	Lawyer	PhD	40–50	1	12, 700
L.A.	Chef	High School	30–40	0	3, 700
...	...	...	...	...	...

Ω in probability space (Ω, Σ, P), a row r is an element of Ω, i.e. $r \in \Omega$ and each column c is equivalent to a random variable $\mathbb{R}$. We consider numerical columns as *finite real-valued* random variables (For Example: Salary $\in \Omega \subseteq \mathbb{R}$) and bitmap columns are considered as events (For Example: Freelancer $\subseteq \Omega$). The following is a probabilistic interpretation of the OLAP Table 8.

- City: $\Omega \longrightarrow$ {*Boston, L.A., New York.......*}
- Profession: $\Omega \longrightarrow$ {*Chef, Construction.....*}
- Education Level: $\Omega \longrightarrow$ {*High School.....*}
- Age Group: $\Omega \longrightarrow$ {*18–20, 25–30... >65*}
- Freelancer: $\Omega \longrightarrow \{0, 1\}$
- Salary: $\Omega \longrightarrow I_{\text{Salary}} \subseteq \mathbb{R}(|I_{\text{Salary}}| \in \mathbb{N})$

3) SEMANTIC CORRESPONDENCE BETWEEN OLAP AVERAGES AND SR

In many cases and as per Codd *et al.* [5], decision-makers use SQL queries to interact with OLAP. Therefore, we start with simple OLAP queries mapped with probability theory. We have a simple OLAP average query; (SELECT AVG(Salary) FROM Table 8). If the number of rows of Table 8 is represented by $|\Omega|$ and the number of rows that contain a value i in column C are equivalent to $\#_C(i)$ then AVG(Salary) FROM Table 8 will compute the average of all the salaries, i.e., a fraction of the sum of the column (Salary) and the total number of rows in the table.

In probability theory, the *average* of a random variable X is the *Expected Value* of $X = \mathsf{E}[X]$. We compare the expected value of X, i.e., $\mathsf{E}(X)$ with the output of the AVG query in OLAP. We have *OLAP Query:*

$$(SELECT\ AVG(Salary)\ FROM\ Table\ 8) \tag{22}$$

$$\text{Expected Value: } \mathsf{E}(\text{Salary}) = \sum_{i \in I_{\text{Salary}}} i \cdot \mathsf{P}(\text{Salary} = i) \tag{23}$$

$$= \sum_{i \in I_{\text{Salary}}} i \cdot \frac{\#_{\text{Salary}}(i)}{|\Omega|} = \frac{\sum_{r \in \Omega} \text{Salary}(r)}{|\Omega|} \tag{24}$$

As per (23) and (24), the average of a random variable X in probability theory and simple averages of an OLAP query provide the same outcome. Hence, we say that an average query in OLAP corresponds to expected values in probability theory.

4) SEMANTIC CORRESPONDENCE BETWEEN OLAP CONDITIONAL AVERAGES AND SR

The conditional average queries in OLAP calculate averages of a column with a WHERE clause. For example, we have an average SQL query with some conditions where the target column is numerical and conditional variables have arbitrary values. We have *OLAP Query:*

$$SELECT\ AVG\ (Salary)\ FROM\ Table\ 8$$
$$WHERE\ City = \ Seattle\ AND\ Profession\ = \ IT; \tag{25}$$

In probability theory, we compute the conditional average of a random number using its conditional expectation. For example, as per Def. 8, the conditional expectation of a random number Y with condition X is given as:

$$\mathsf{E}(Y|X) = \sum_{n=0}^{\infty} i_n \cdot \mathsf{P}(Y \ = i_n |\ X)$$
$$f(i) = E(Y = i_n|X) \tag{26}$$

Here, the value $E(Y = i_n|X)$ is dependent on the value of i. Therefore, we say that $E(Y = i_n|X)$ is a function of i, which is given in (26). We compare the conditional expected value of $E(Y = i_n|X)$ with the output of the conditional AVG query in OLAP. We have *OLAP Query:*

$$SELECT\ AVG\ (Salary)\ FROM\ Table\ 8$$
$$WHERE\ City = \ Seattle\ AND\ Profession\ = \ IT;$$
$$\text{Conditional Expected Value: } \mathsf{E}(Salary|City$$
$$= Seattle \cap Profession = IT) \tag{27}$$
$$\mathsf{E}(Y|X) = \sum_{i \in I_C} i \cdot \mathsf{P}(\mathrm{Y} = i \mid \mathrm{X}) \tag{28}$$

As per (27) and (28), the average of a random variable Y with condition X (Conditional Expected values) and the conditional average of an OLAP query provide the same outcome. Hence, we can say that a conditional average query in OLAP corresponds to the conditional expected values in probability theory. In Fig. 5, we demonstrate the semantic correspondence between the features of SR, OLAP, and ARM. At the top level, we consider OLAP and its features. In the middle, we have probability theory and its features, which work as the middle layer between OLAP, ARM and at the bottom layer, we provide ARM and its measures. In OLAP, we have conditional averages over binary columns,

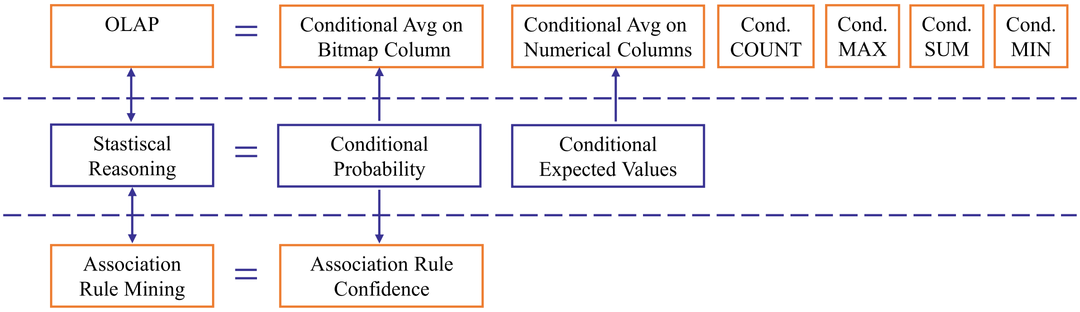

OLAP
=
Conditional Avg on Bitmap Column
Conditional Avg on Numerical Columns
Cond. COUNT
Cond. MAX
Cond. SUM
Cond. MIN
Stastiscal Reasoning
=
Conditional Probability
Conditional Expected Values
Association Rule Mining
=
Association Rule Confidence

FIGURE 5. Demonstration of semantic correspondence between statistical reasoning, OLAP and association rule mining.

TABLE 9. Semantic correspondence between statistical reasoning, OLAP and association rule mining.

Concepts	**Statistical Reasoning**	**OLAP**	**Association Rule Mining**
Background	Probability Space	Database Table	Transaction Dataset
Data Notion	$(\Omega, \Sigma, \mathsf{P})$	$\{C_1 \times C_2 \ldots \times C_n\}$	$I = \{i_1, i_2 \ldots i_n\}$
Data Implementation	Table	Table	Bitmap Table *(in classical ARM)* and Table with discrete value columns *(in practical ARM tools)*
Average of a bitmap column $Y : \{0,1\}$ under dicing w.r.t setting bitmap columns $X_1 \ldots X_m$ to *truth values* (i.e., 1)	Conditional Probability $\mathsf{P}(Y\|X_1, \ldots, X_m) = \frac{\mathsf{P}(Y \cap X_1 \cap \ldots \cap X_m)}{\mathsf{P}(X_1 \cap \ldots \cap X_m)}$	SELECT $Avg(Y)$ FROM T WHERE $X_1 = 1$ AND ... AND $X_m = 1$	Confidence $\mathit{Conf}\big(X_1, \ldots, X_m \Rightarrow Y\big)$ *(in classical ARM)*
Average of a bitmap column $Y : \{0,1\}$ under dicing w.r.t setting discrete columns $X_1 \ldots X_m$ to *arbitrary values*	Conditional Probability $\mathsf{P}(Y\|X_1{=}x_1, \ldots, X_m = x_m) = \frac{\mathsf{P}(Y \cap X_1=x_1 \cap \ldots \cap X_m=x_m)}{\mathsf{P}(X_1 \cap X_1=x_1 \cap \ldots \cap X_m=x_m)}$	SELECT $Avg(Y)$ FROM T WHERE $X_1 = x_1$ AND ... AND $X_m = x_m$	Confidence $\mathit{Conf}\big(X_1{=}x_1, \ldots, X_m = x_m \Rightarrow Y\big)$ *(in practical ARM tools)*
Average of a numerical column $Y : \{i_0, \ldots, i_k\} \subseteq \mathbb{R}$ under dicing w.r.t setting discrete columns $X_1 \ldots X_m$ to *arbitrary values*	Conditional Expected Value $\mathsf{E}(Y\|X_1, \ldots, X_m) = \sum_{n=0}^{k} i_n \cdot \mathsf{P}(Y{=}i_n\|X_1{=}x_1, \ldots, X_m{=}x_m)$	SELECT $Avg(Y)$ FROM T WHERE $X_1 = x_1$ AND ... AND $X_m = x_m$	-
Tools	SPSS, SAS, R	IBM Cognos, Palo, Mondrian, OLAP server, Pivot Table	Rapidminer, Orange, Weka

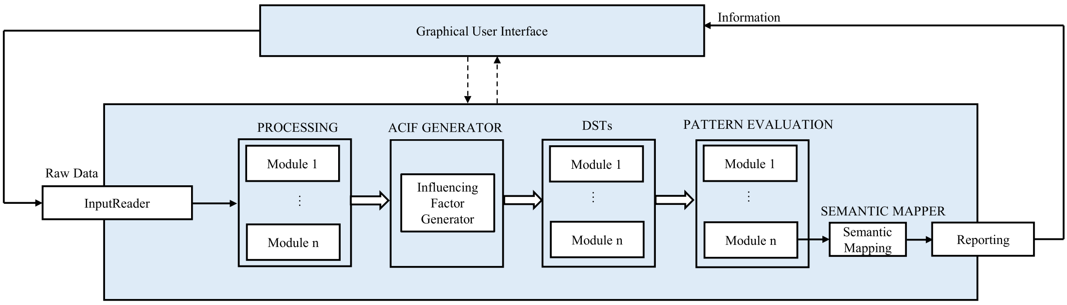


FIGURE 6. A high level abstraction of the framework for the unification of decision support techniques.

conditional averages over numerical columns, and different other conditional aggregates like Max, Min, Sum, etc. In OLAP, conditional averages on binary columns correspond to conditional probability, and they also correspond to confidence in ARM. However, conditional averages on numerical columns in OLAP correspond to conditional expected values

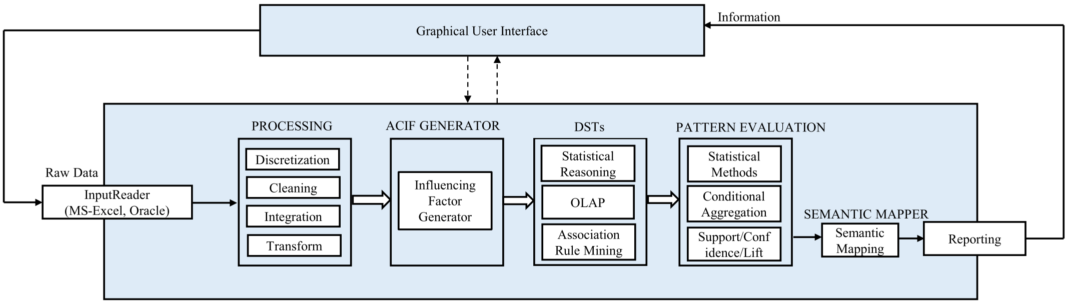


FIGURE 7. A detailed overview of the framework for the unification of decision support techniques.

in probability theory. Based on these semantic correspondences between SR, OLAP, and ARM, we are convinced that DSTs have common features with different names. However, they are being used differently. Therefore, the unification of SR, OLAP, and ARM will provide an advanced novel framework for next-generation decision support tools. In Table 9, we provide a list of semantic correspondence between the features of SR, OLAP, and ARM.

V. THE FRAMEWORK, EVALUATION AND EXPERIMENTS

In this section, the framework for the unification of three DSTs is presented. As a data science process provided by Schutt and O'Neil [66], the proposed framework is modular in design and every module in the framework can be displaced. In Fig. 6, we illustrate the high-level abstraction of the framework and based on the process of knowledge discovery in databases (KDD) [36], a detailed overview of the proposed framework is given in Fig. 7.

The framework consists mainly of seven major components. The Graphical User Interface (GUI) allows decision-makers to communicate with the framework to process the raw data. The data pre-processing includes various operations and checks, including discretization, cleaning, e.g., checking for corrupt data, reviewing the types of data, transforming and integrating data in useful formats, etc. The ACIF generator in the framework is developed for decision-makers to select the target columns and influencing factors to generate different combinations of data items. The decision support engine is a set of multiple DSTs, allowing decision-makers to select one or more techniques to process the data and get insights. The Pattern evaluation is used to find interesting information using different methods from SR, OLAP, and ARM. The semantic mapper is a manual process to map the results of DSTs and reports different semantic correspondences between them. A brief description of all the significant components of the proposed framework is given in Table 10.

A. IMPLEMENTATION OF THE PROPOSED FRAMEWORK

To demonstrate the usability of the proposed framework, an instance of the framework is developed using ASP.NET, an open-source framework for developing web applications. The resulting tool is an example of a next-generation decision support tool implemented by adopting the proposed framework. A summary of technologies and framework used for the implementation of the tool is given in Table 11. The programming code and other instructions on how to use the proposed tool are available in the GitHub repository [15]. The AJAX request methods are used throughout the tool's implementation to establish a connection between the client and server. JSON serialization and deserialization functions convert .NET objects (strings) to JSON format and JSON format to .NET objects. We use Oracle database and Microsoft Excel as databases and for OLAP, we have used relational algebra in the tool.

The tool first recognizes different kinds of data (discretized, numerical, categorical) and then develops generalized association rules for the various combinations of influencing factors and target columns. In the tool, if the selected target column is numerical, then the aggregate function is used, and the average value of the target column is calculated against the chosen influencing factors by the following SQL query; *Select AVG (target column) from table group by influencing factors*. If the specified target column is numerical, the aggregate function is employed in the tool, and the average value of the target column is determined against the chosen influencing factors using the SQL statement; *Select AVG (target column) from table group by influencing factors*. If the selected column is categorical, the tool uses the following SQL query to determine the conditional probability of the target column; *Select conditional probability of target column under influencing factor from table group by target column and influencing factors*. Both support and lift are calculated for numerical and categorical target columns. For the numerical target column, the order of columns is support, lift, an average value of the target column, and then influencing factors. For the categorical target column, the columns are listed in the following order: support, lift, conditional probability, target column, and influencing variables.

1) ACIF GENERATOR

In the tool, we have developed a function for ACIF generator and implemented it in the proposed framework.

TABLE 10. Summary of the components used to develop the framework for the unification of DSTs.

Component	Objective
Graphical User Interface	GUI is Used to communicate between decision-makers and the framework
Data pre-processing	The data pre-processing step includes various operations and checks, including discretization and cleaning (e.g., checking for corrupt data, reviewing the types of data, transforming and integrating data in useful formats, etc.)
ACIF	The ACIF generator is used to compute all the combinations of influencing factors
Decision Support Engine (DSE)	DSE is a set of DSTs used in the framework
Pattern Evaluation	In pattern evaluation, a set of measures from different DSTs are used to evaluate the usefulness of the patterns between the data items
Semantic Mapper	Semantic mapper is a process to compare the outcome of DSTs
Reporting	The reporting process is used to generate various reports for the outcome of DSTs

TABLE 11. Summary of the technologies and framework used for the implementation of the tool.

Description	Technologies
Programming Language	C#
Development Framework	ASP.NET
Requesting data through the web server	AJAX
Data access to the Oracle database	Oracle Data Provider (ODP), ODP.NET
Data access to the Microsoft Excel file	OLE DB

The ACIF generator is developed to select the target column and influencing factors to generate all possible combinations of the selected target column and influencing factors. First, the generator identifies the column combinations from the dataset and generates reports for the target column and influencing factors. The pseudo-code for the ACIF generator and ACIF report generator is given in Listing 1. In the pseudo-code, the *CREATE_COMBINATIONS* function is defined to pass the information of influencing columns and the number of columns. This function calculates the possible combinations of the selected influencing factors. In line 15, the *GENERATE_REPORT* function is defined to generate the reports for various combinations of influencing factors against target columns. This function passes the information about the table name, target columns and influencing columns. In this function, the SQL statement is used to retrieve the support, lift, conditional averages and influencing factors from the data source.

2) MATHEMATICAL DESCRIPTION OF THE ACIF GENERATOR

Let T be a database Table with multiple columns $C = \{X_1 : T_1, \ldots, X_n\colon T_n\}$, where $X_1 \ldots X_n$ represent column names and $T_1 \ldots T_n$ represent column types. To calculate various operations of SR, OLAP and ARM for T,

$$\begin{aligned}
&\forall 1 \leq \psi \leq n\\
&\forall D = \{X'_1\colon D_1, \ldots, X'_{\psi-1}\colon D_{\psi-1}\} \subseteq\\
&C(D_i = d_1, \ldots, d_{ni})\\
&\forall d'_1 \in D_1, \ldots, d'_{\psi-1} \in D_{\psi-1}
\end{aligned}$$

Here, D is the subset of C and the influencing factor.

$$\forall Y\colon \mathbb{R} \in C \; or \; Y = X_{ij}\colon B, X_i\colon d_i \in C$$

Y is the target column, $\mathbb{R}$ is the real-valued numbers then.

Support:

$$\mathsf{P}(Y, X'_1 = d_1, \ldots, X'_{\psi-1} = d_{\psi-1}) \tag{29}$$

Average:

$$\mathsf{E}(Y \mid X'_1 = d_1, \ldots, X'_{\psi-1} = d_{\psi-1}) \tag{30}$$

Lift:

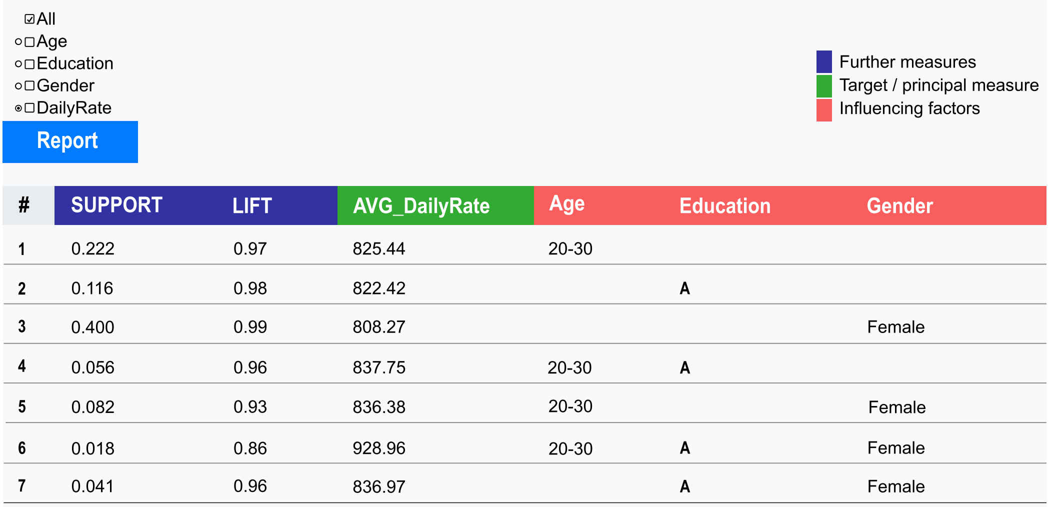
$$\frac{\mathsf{E}(Y \mid X'_1 = d_1, \ldots, X'_{\psi-1} = d_{\psi-1})}{\mathsf{E}(Y)} \tag{31}$$

B. EXPERIMENTS ON THE PROPOSED FRAMEWORK

The experiment section demonstrates the potential of the introduced framework. The tool is evaluated on two real datasets and one synthetic dataset. The tool is tested on a computer with an Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz, 1800 Mhz, 4 Core(s), 8 Logical Processor(s), 16 GB RAM and Windows 10 x 64 operating system. The programming code, datasets, and other necessary instructions about the tool are available in the GitHub repository [15].

The datasets are summarized in Table 12, in which we highlight the number of records, number of attributes, and number of numeric attributes. The first Dataset, New Jersey (NJ) School Teacher Salaries (2016) [67] contains 138, 715 records and 15 attributes, while another real dataset, DC public government employees [68] contains 33, 424 records, which are huge in numbers to check the performance of the tool. In the table, we have described the dataset attributes with their types. Dataset NJ Teacher Salaries (2016) consists of salary, job, and experience data for the teachers and employees in New Jersey schools. The data are sourced from the (NJ) Department of Education. The second real dataset is a list of DC public government employees and their salaries in 2011. The second data set is sourced from the washington times via freedom of information act (FOIA) requests. We have also tested the proposed tool on the sample dataset UDS1 [69]. This dataset contains 1, 470 records with different combinations of numerical, categorical, and discretized attributes. A feature list obtained by parsing the UDS1 dataset is summarized in Table 13.

In the datasets, the target column is the one for which we are computing ARM operations, i.e., support, confidence, lift and OLAP averages with respect to an influencing factor. An influencing factor is an attribute that impacts



#	SUPPORT	LIFT	AVG_DailyRate	Age	Education	Gender
1	0.222	0.97	825.44	20-30		
2	0.116	0.98	822.42		A	
3	0.400	0.99	808.27			Female
4	0.056	0.96	837.75	20-30	A	
5	0.082	0.93	836.38	20-30		Female
6	0.018	0.86	928.96	20-30	A	Female
7	0.041	0.96	836.97		A	Female

FIGURE 8. **A sample report comparing the results of OLAP and ARM measures is as follows: the ARM operations (support, confidence, lift) and OLAP operations (averages) are displayed. A sample dataset is used to generate the report, which includes all possible combinations of influencing factors and numerical target columns.**

Listing 1 Pseudo-Code to Find the ACIF and Generate ACIF Reports

```
function CREATE_COMBINATIONS(influencing_Columns[], numberofColumns)
    if numberofColumns == \text{0}~then
        return []
    return_Values = []
    for i = \text{1}~to LENGTH(influencing_Columns) do
        colName = influencingColumns[i]
        partialLst = REMOVE_COLUMN(i,influencingColumns)
        for each: j in CREATE_COMBINATIONS(partialLst, numberofColumns - 1) do
            APPEND_TO(return_Values,ADD_FIRST(colName,j))
        end for
    end for
    return return_Values
end function

function GENERATE_REPORT(table_Name, target_Column, influencing_Columns[])
    for i = \text{1}~to LENGTH(influencing_Columns) do
        column_Combination = call:CREATE_COMBINATIONS(influencingColumns,i)
        for each: Combinations in column_Combination
            "SELECT COUNT(*)/ (SELECT COUNT(*) FROM "+table_Name+") AS SUPPORT,
AVG("+target_Column+") AS LIFT,
AVG("+target_Column+") AS AVG_targetColumn, "+ Combinations +"
FROM "+table_Name +"
GROUP BY "+ Combinations +"
ORDER BY "+ Combinations;"
        end for
    end for
end function
```

the target columns. Therefore, we also denote the WHERE clause as an influencing factor for the target column in OLAP computations. The UDS1 dataset consists four columns with different data types; age is discretized, gender is categorical, education is categorical and DailyRate is numerical. The column Age has the age groups as $20 - 30$, $30 - 40$, etc.,

TABLE 12. Summary of datasets used to evaluate performance of the proposed tool.

Dataset Type	Title	Records	Total Attributes	Numerical Attribute
Real	NJ Teacher Salaries (2016) [67]	138715	15	5
Real	Public government employees	33424	6	2
Synthetic	UDS1	1470	4	1

TABLE 13. A summary of different attributes obtained by parsing the datasets.

S.No.	Feature	Type
1	Age	DISCRETIZED
2	DailyRate	NUMERICAL
3	Education	CATEGORICAL
4	Gender	CATEGORICAL

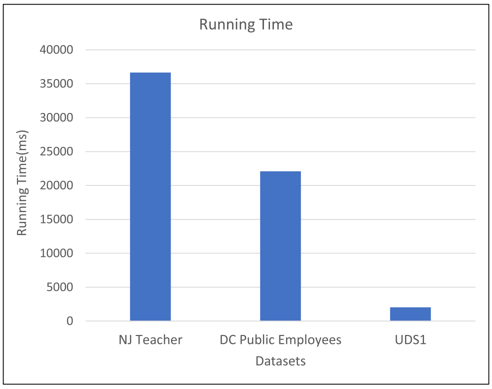


FIGURE 9. Running time and performance variation of the proposed tool induced by the number of records in the datasets.

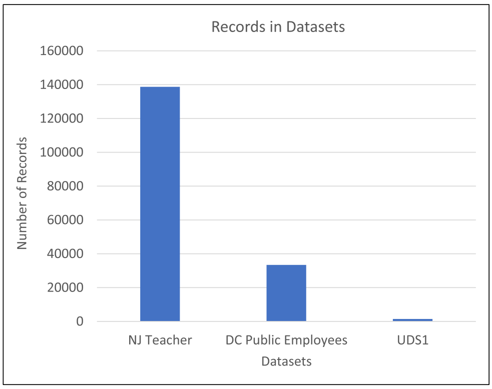


FIGURE 10. Number of records in datasets.

and gender has two categorical values; Male and female. Education has five categorical levels A, B, C, D, and E. For example, if we select education as the target column and its values are A, B, C, D, and E. Here, education is a factor and its values are instances of the factor. The tool calculates the conditional probability for each instance in the generated report. For example, suppose we select DailyRate as the target column and age, gender and education as influencing factors. In this case, all possible combinations of the target column are generated against all selected influencing factors.

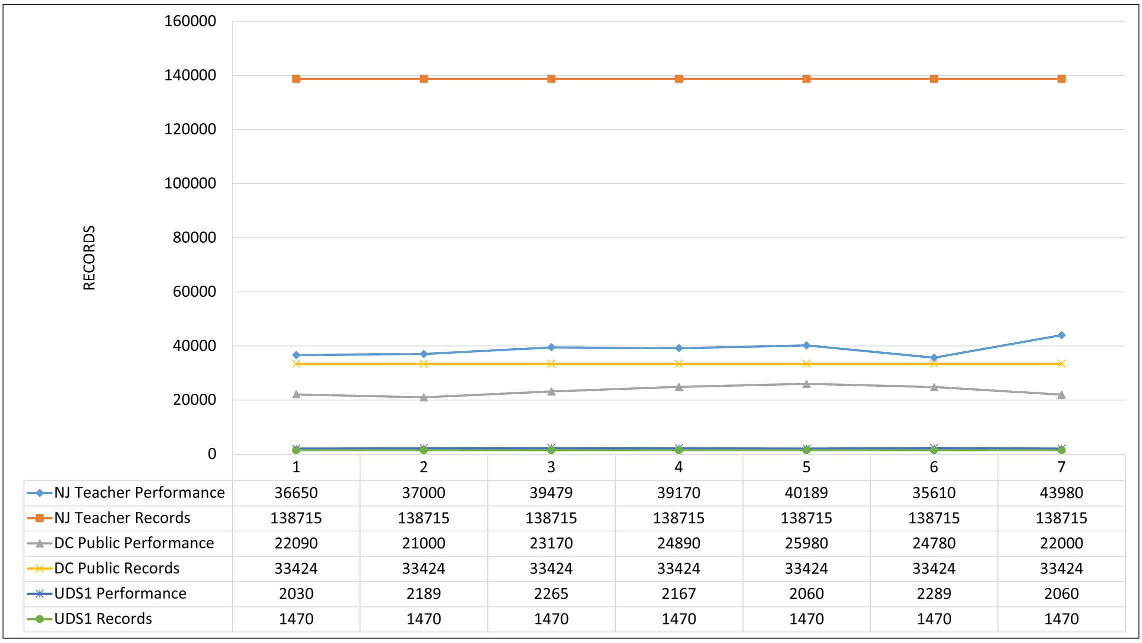
At the first step, the tool checks for the types of input data. Then it generates generalized association rules concerning the possible combinations of influencing factors and target columns. In the second step, the tool provides aggregate values, the conditional probability of the target column for each combination of influencing factors and target column. For SR, the tool calculates conditional probability and the mean value for the numeric target column concerning the influencing factors. For ARM operations, the tool calculates the support, confidence, and lift. For OLAP operations, the tool computes conditional averages. An overview of the computation of different SR, OLAP, and ARM operations is given in Fig. 8. In the report, the blue color code shows ARM operations. The green color code displays the target column, and the red color code indicates the influencing factors.

We have analyzed the performance of the proposed tool with three datasets. The performance of the tool varies with the number of records. If the number of records in a dataset is high, the tool has higher running time and slow performance. In Fig. 9, the performance variation induced by the number of records in a dataset is shown. The Dataset NJ Teacher has a huge number of records; therefore, its running time is $36,650$ milliseconds. Dataset DC Public Employees has $33,424$ records. Therefore, its running time is $22,090$ milliseconds, and the sample dataset UDS1 has a small number of records, i.e., $1,470$; therefore, its running time is $2,030$ milliseconds. Running time and performance variation of the proposed tool induced by the number of records in the datasets is shown in Fig. 10. A summary of records in datasets and performance variation of the tool with the datasets is given in Fig. 11.

C. ADVANTAGES OF THE PROPOSED TOOL OVER EXISTING DECISION SUPPORT TOOLS

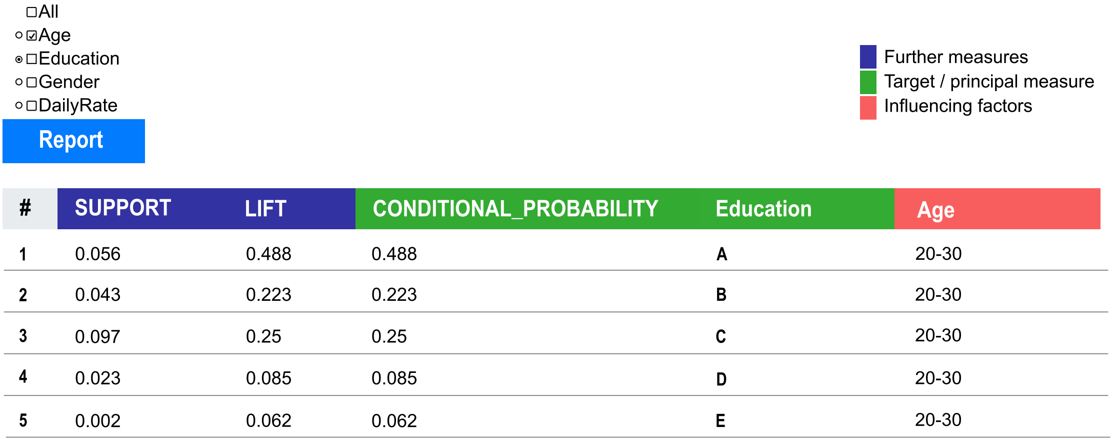
In this section, we compare the capabilities of the proposed tool with one of the state of the art decision support platforms, i.e., RapidMiner [70].

Unlike any other decision support tool, the proposed tool altogether computes SR operations, i.e., conditional



	1	2	3	4	5	6	7
NJ Teacher Performance	36650	37000	39479	39170	40189	35610	43980
NJ Teacher Records	138715	138715	138715	138715	138715	138715	138715
DC Public Performance	22090	21000	23170	24890	25980	24780	22000
DC Public Records	33424	33424	33424	33424	33424	33424	33424
UDS1 Performance	2030	2189	2265	2167	2060	2289	2060
UDS1 Records	1470	1470	1470	1470	1470	1470	1470

FIGURE 11. Performance summary of the tool under two real datasets and one synthetic dataset.



☐All
○☑Age
◉☐Education
○☐Gender
○☐DailyRate

Report

🟦 Further measures
🟩 Target / principal measure
🟥 Influencing factors

#	SUPPORT	LIFT	CONDITIONAL_PROBABILITY	Education	Age
1	0.056	0.488	0.488	A	20-30
2	0.043	0.223	0.223	B	20-30
3	0.097	0.25	0.25	C	20-30
4	0.023	0.085	0.085	D	20-30
5	0.002	0.062	0.062	E	20-30

FIGURE 12. In the proposed tool: a sample project for generating all possible combinations of influencing factors against target columns.

probability, OLAP operation, i.e., conditional averages, and ARM operations, i.e., support, confidence, and lift, see Fig. 12. In addition, the tool computes the average value of a numerical target column against all possible combinations of influencing factors. In Fig. 8, a sample report is given for generating all possible combinations of influencing factors against the target column. However, in RapidMiner, to calculate the average value of a numerical target column against all possible combinations of influencing factors, a decision-maker needs to create multiple connections for all the possible combinations of influencing factors. Moreover, a decision-maker must create a new project for each dataset and repeatedly modify its columns and combinations. Therefore, in Fig. 13, we provide a sample use case to generate all possible combinations of influencing factors against the target column in RapidMiner.

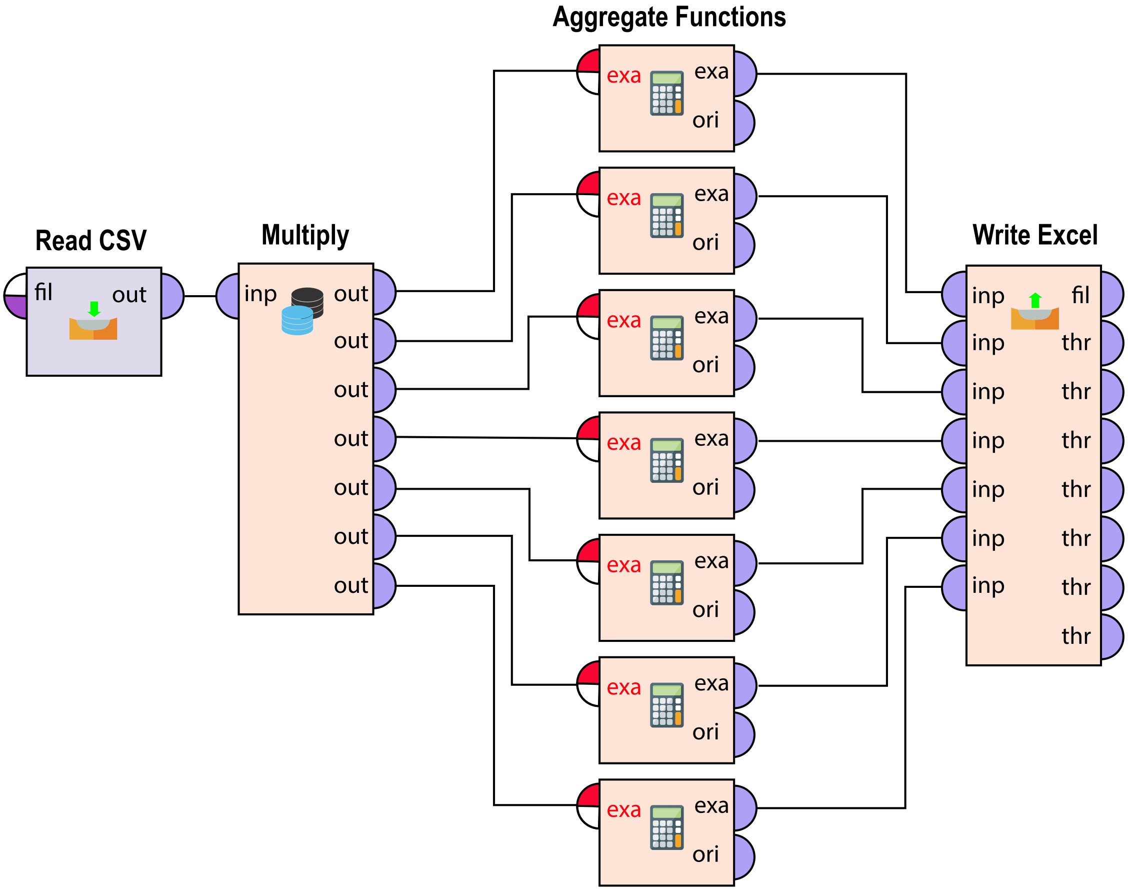


FIGURE 13. In RapidMiner: a sample project for generating all possible combinations of influencing factors against target columns.

TABLE 14. A sample list of premises and conclusions generated by RapidMiner for the influencing factors and target column.

Premises	Conclusion
Education = D	Gender, Age = 30-40
Gender	Age = 30-40, Education = D

Additionally, in RapidMiner, the influencing factors and their values are stored in a single column called the 'conclusion' column as "influencing factors=value". The target column and its values are stored in the 'premises' column as "Target Column=value". A sample list of premises and conclusions generated by RapidMiner for the influencing factors and target column is displayed in Table 14. The representation of the target factors and influencing factors is difficult to understand. It is hard for decision-makers to identify each factor and its instance from the multiple tables. However, the proposed tool creates a separate column for each factor to identify the target column and influence factors quickly. In the tool, a decision-maker can select the target column and all influencing factors at once to generate all combinations of target factors and all influencing factors.

VI. FUTURE WORK

This paper provides a foundation for uncovering the semantic correspondences between DSTs and utilizing them to develop a framework for the unified usages of DSTs. However, the research is yet limited in scope to find the semantic correspondences between the three DSTs only; therefore, in the near future, more DSTs can be investigated to identify the semantic correspondences between them to develop cutting-edge frameworks for next-generation decision support tools. The unified usage of DSTs will not only be helpful in building robust frameworks for a variety of decision support tools but also open a new domain of research for hybrid DSTs.

Furthermore, we intend to build an advanced platform by implementing additional features in the proposed tool, e.g., Pearson correlation, regression, etc. We are also working on a new measure to identify any instance of Simpson's paradox in Big Data. Implementing these measures in the proposed platform will enable decision-makers to determine the genuine and unbiased impact factors.

The proposed tool has some performance issues with large datasets momentarily; therefore, we plan to scale up the performance of the tool by utilizing high-performance computing (HPC) infrastructure and making it available to the decision-makers. We intend to build it as a trustworthy platform and grow as a service provider in the near future.

VII. CONCLUSION

In this paper, we analyzed a series of approaches to overcome the divide between the three most popular DSTs, i.e., SR, OLAP and ARM. We contributed by elaborating the semantic correspondences between the foundations of SR, OLAP and ARM, i.e., probability theory, relational algebra and the itemset apparatus, respectively. The support of an itemset corresponds to the probability of a corresponding event and the confidence of an association rule corresponds to the conditional probability of two corresponding events. Furthermore, the OLAP average aggregate function corresponds to conditional expected values, which closes the loop between ARM, OLAP and probability theory with respect to the most important constructs in ARM and OLAP. We have proposed a novel framework for the unification of DSTs and implemented a tool to validate the concept of unification. The tool provides unified usage of DSTs in a classical decision support process and clarifies in how far the operations of SR, ARM, and OLAP can complement each other in understanding data, data visualization and decision making. The tool was developed on the basis of an open-source framework and tested with two real datasets and one synthetic dataset. The results and performance of the tool show valuable contributions towards developing the next-generation DSSs.

REFERENCES

[1] S. M. Stigler, *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, MA, USA: Harvard Univ. Press, 1986.

[2] G. A. Gorry and S. M. S. Morton, "A framework for management information systems," Alfred P. Sloan School Manage., Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. 510-71, Feb. 1971.

[3] N. H. Nie, D. H. Bent, and C. H. Hull, *SPSS: Statistical Package for the Social Sciences*. New York, NY, USA: McGraw-Hill, 1970.

[4] *SAS User's Guide: Statistics; Version*, 5th ed., SAS Institute, Cary, NC, USA, 1987.

[5] E. Codd, S. Codd, and C. Salley, *Providing OLAP to User-Analysts: An IT Mandate*. East Falmouth, MA, USA: E. F. Codd and Associates, 1993.

[6] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, 1993.

[7] S. Chaudhuri and U. Dayal, "Data warehousing and OLAP for decision support," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 1997, pp. 507–508, doi: 10.1145/253260.253373.

[8] Q. Wang, J. You, B. Zou, Y. Chen, X. Huang, and L. Jia, "Reduced quotient cube: Maximize query answering capacity in OLAP," *IEEE Access*, vol. 9, pp. 141524–141535, 2021.

[9] W. Thurachon and W. Kreesuradej, "Incremental association rule mining with a fast incremental updating frequent pattern growth algorithm," *IEEE Access*, vol. 9, pp. 55726–55741, 2021.

[10] H. Zhu, "On-line analytical mining of association rules," M.S. thesis, Brit. Columbia, Canada, School Comput. Sci., Simon Fraser Univ., British, CO, Canada, 1998.

[11] M. Kamber, J. Han, and J. Y. Chiang, "Metarule-guided mining of multidimensional association rules using data cubes," in *Proc. KDD 3rd Int. Conf. Knowl. Discovery Data Mining (KDD)*, 1997, pp. 207–210.

[12] D. Draheim, "Future perspectives of association rule mining based on partial conditionalization (DEXA'2019 keynote)," in *Proc. DEXA 30th Int. Conf. Database Expert Syst. Appl.* in Lecture Notes in Computer Science, vol. 11706. Cham, Switzerland: Springer, 2019, pp. 40–49.

[13] G. Piatetsky. *Top Analytics, Data Science and Machine Learning Software*. Accessed: Dec. 21, 2021. [Online]. Available: https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html

[14] J. Han, Y. Fu, W. Wang, J. Chiang, O. R. Zaïane, and K. Koperski, "DBMiner: Interactive mining of multiple-level knowledge in relational databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 1996, p. 550, doi: 10.1145/233269.280356.

[15] R. Sharma and S. A. Peious. *Towards Unification of Decision Support Technologies: Statistical Reasoning, OLAP and Association Rule Mining*. Accessed: Dec. 21, 2021. [Online]. Available: https://github.com/rahulgla/unification

[16] S. A. Peious, R. Sharma, M. Kaushik, S. A. Shah, and S. B. Yahia, "Grand reports: A tool for generalizing association rule mining to numeric target values," in *Proc. DaWaK 22nd Int. Conf. Data Warehousing Knowl. Discovery* in Lecture Notes in Computer Science, vol. 12393. Cham, Switzerland: Springer, 2020, pp. 28–37.

[17] R. P. Salas, G. D. Edelson, P. S. Kleppner, and R. S. Shaver, "Data processing apparatus and method for a reformattable multidimensional spreadsheet," U.S. Patent 5 317 686, May 31, 1994.

[18] H. Wang, "Intelligent agent-assisted decision support systems: Integration of knowledge discovery, knowledge analysis, and group decision support," *Expert Syst. Appl.*, vol. 12, no. 3, pp. 323–335, Apr. 1997.

[19] W. Fan, H. Lu, S. E. Madnick, and D. Cheung, "DIRECT: A system for mining data value conversion rules from disparate data sources," *Decis. Support Syst.*, vol. 34, no. 1, pp. 19–39, Dec. 2002.

[20] N. Bolloju, M. Khalifa, and E. Turban, "Integrating knowledge management into enterprise environments for the next generation decision support," *Decis. Support Syst.*, vol. 33, no. 2, pp. 163–176, Jun. 2002.

[21] J. H. Heinrichs and J.-S. Lim, "Integrating web-based data mining tools with business models for knowledge management," *Decis. Support Syst.*, vol. 35, no. 1, pp. 103–112, Apr. 2003.

[22] V. Cho and E. W. T. Ngai, "Data mining for selection of insurance sales agents," *Expert Syst.*, vol. 20, no. 3, pp. 123–132, Jul. 2003. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0394.00235

[23] N. Jukić and S. Nestorov, "Comprehensive data warehouse exploration with qualified association-rule mining," *Decis. Support Syst.*, vol. 42, no. 2, pp. 859–878, Nov. 2006.

[24] R. Rupnik, M. Kukar, and M. Krisper, "Integrating data mining and decision support through data mining based decision support system," *J. Comput. Inf. Syst.*, vol. 47, no. 3, pp. 89–104, 2007.

[25] S. T. March and A. R. Hevner, "Integrated decision support systems: A data warehousing perspective," *Decis. Support Syst.*, vol. 43, no. 3, pp. 1031–1043, Apr. 2007.

[26] Z. Shi, Y. Huang, Q. He, L. Xu, S. Liu, L. Qin, Z. Jia, J. Li, H. Huang, and L. Zhao, "MSMiner—A developing platform for OLAP," *Decis. Support Syst.*, vol. 42, no. 4, pp. 2016–2028, Jan. 2007.

[27] N. Di Domenica, G. Mitra, P. Valente, and G. Birbilis, "Stochastic programming and scenario generation within a simulation framework: An information systems perspective," *Decis. Support Syst.*, vol. 42, no. 4, pp. 2197–2218, Jan. 2007.

[28] M. Charest, S. Delisle, O. Cervantes, and Y. Shen, "Bridging the gap between data mining and decision support: A case-based reasoning and ontology approach," *Intell. Data Anal.*, vol. 12, no. 2, pp. 211–236, Apr. 2008.

[29] Z. Y. Zhuang, L. Churilov, F. Burstein, and K. Sikaris, "Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners," *Eur. J. Oper. Res.*, vol. 195, no. 3, pp. 662–675, Jun. 2009.

[30] S. Liu, A. H. B. Duffy, R. I. Whitfield, and I. M. Boyle, "Integration of decision support systems to improve decision support performance," *Knowl. Inf. Syst.*, vol. 22, no. 3, pp. 261–286, Mar. 2010.

[31] Y. Peng, Y. Zhang, Y. Tang, and S. Li, "An incident information management framework based on data integration, data mining, and multi-criteria decision making," *Decis. Support Syst.*, vol. 51, no. 2, pp. 316–327, 2011.

[32] H. Ltifi, C. Kolski, M. B. Ayed, and A. M. Alimi, "A human-centred design approach for developing dynamic decision support system based on knowledge discovery in databases," *J. Decis. Syst.*, vol. 22, no. 2, pp. 69–96, Apr. 2013, doi: 10.1080/12460125.2012.759485.

[33] J. Dong, H. S. Du, S. Wang, K. Chen, and X. Deng, "A framework of web-based decision support systems for portfolio selection with OLAP and PVM," *Decis. Support Syst.*, vol. 37, no. 3, pp. 367–376, Jun. 2004, doi: 10.1016/S0167-9236(03)00034-4.

[34] I. Fister and I. Fister, Jr., "UARMSolver: A framework for association rule mining," 2020, *arXiv:2010.10884*.

[35] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C.-N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann, "Knowledge graphs," *ACM Comput. Surv.*, vol. 54, no. 4, pp. 1–37, May 2022, doi: 10.1145/3447772.

[36] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, p. 37, 1999.

[37] H. X. Li and L. D. Xu, "Feature space theory – a mathematical foundation for data mining," *Knowl.-Based Syst.*, vol. 14, nos. 5–6, pp. 253–257, Aug. 2001, doi: 10.1016/S0950-7051(01)00103-4.

[38] Y. Zhu, C. Bornhövd, D. Sautner, and A. P. Buchmann, "Materializing web data for OLAP and DSS," in *Proc. WAIM 1st Int. Conf. Web-Age Inf. Manage.* Berlin, Germany: Springer, 2000, pp. 201–214.

[39] K. Gandhi, B. Schmidt, and A. H. C. Ng, "Towards data mining based decision support in manufacturing maintenance," *Proc. CIRP*, vol. 72, pp. 261–265, Jan. 2018.

[40] T. Imielinski, L. Khachiyan, and A. Abdulghani, "Cubegrades: Generalizing association rules," *Data Mining Knowl. Discovery*, vol. 6, no. 3, pp. 219–257, 2002.

[41] R. T. Ng, L. V. S. Lakshmanan, J. Han, and A. Pang, "Exploratory mining and pruning optimizations of constrained associations rules," *ACM SIGMOD Rec.*, vol. 27, no. 2, pp. 13–24, 1998.

[42] L. V. S. Lakshmanan, R. Ng, J. Han, and A. Pang, "Optimization of constrained frequent set queries with 2-variable constraints," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 1999, pp. 157–168, doi: 10.1145/304182.304196.

[43] K.-N. T. Nguyen, L. Cerf, M. Plantevit, and J.-F. Boulicaut, "Multi-dimensional association rules in Boolean tensors," in *Proc. SDM 11th SIAM Int. Conf. Data Mining*. Philadelphia, PA, USA: SIAM, 2011, pp. 570–581.

[44] S. Chaudhuri and U. Dayal, "An overview of data warehousing and OLAP technology," *SIGMOD Rec.*, vol. 26, no. 1, pp. 65–74, Mar. 1997, doi: 10.1145/248603.248616.

[45] Q. Chen, U. Dayal, and M. Hsu, "An OLAP-based scalable web access analysis engine," in *Proc. DaWak 2nd Int. Conf. Data Warehousing Knowl. Discovery*. Berlin, Germany: Springer-Verlag, 2000, pp. 210–223.

[46] L. Cerf, J. Besson, C. Robardet, and J.-F. Boulicaut, "Closed patterns meet n-ary relations," *ACM Trans. Knowl. Discovery From Data*, vol. 3, no. 1, pp. 1–36, Mar. 2009, doi: 10.1145/1497577.1497580.

[47] J. P. Shim, M. Warkentin, J. F. Courtney, D. J. Power, R. Sharda, and C. Carlsson, "Past, present, and future of decision support technology," *Decis. Support Syst.*, vol. 33, no. 2, pp. 111–126, 2002.

[48] J. W. Tukey, "Exploratory data analysis," in *Addison-Wesley Series in Behavioral Science*. Reading, MA, USA: Addison-Wesley, 1977.

[49] D. Donoho, "50 years of data science," *J. Comput. Graph. Statist.*, vol. 26, no. 4, pp. 745–766, 2017, doi: 10.1080/10618600.2017.1384734.

[50] R. Srikant and R. Agrawal, "Mining quantitative association rules in large relational tables," *ACM SIGMOD Rec.*, vol. 25, no. 2, pp. 1–12, Jun. 1996, doi: 10.1145/235968.233311.

[51] M. Kaushik, R. Sharma, S. A. Peious, M. Shahin, S. Ben Yahia, and D. Draheim, "On the potential of numerical association rule mining," in *Proc. FDSE 7th Int. Conf. Future Data Secur. Eng.* in Lecture Notes in Computer Science, vol. 12466. Singapore: Springer, 2020, pp. 3–20.

[52] M. Kaushik, R. Sharma, S. A. Peious, M. Shahin, S. B. Yahia, and D. Draheim, "A systematic assessment of numerical association rule mining methods," *Social Netw. Comput. Sci.*, vol. 2, no. 5, pp. 1–13, Sep. 2021.

[53] L. Geng and H. J. Hamilton, "Interestingness measures for data mining: A survey," *ACM Comput. Surv.*, vol. 38, no. 3, pp. 1–32, Sep. 2006, doi: 10.1145/1132960.1132963.

[54] C. D. Larose and D. T. Larose, *Discovering Knowledge in Data*. Hoboken, NJ, USA: Wiley, 2014, ch. Association Rules, pp. 247–265. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118874059.ch12

[55] R. Sharma, M. Kaushik, S. A. Peious, S. B. Yahia, and D. Draheim, "Expected vs. unexpected: Selecting right measures of interestingness," in *Proc. DaWaK 22nd Int. Conf. Data Warehousing Knowl. Discovery* in Lecture Notes in Computer Science, vol. 12393. Springer, 2020, pp. 38–47.

[56] B. Liu, W. Hsu, and S. Chen, "Using general impressions to analyze discovered classification rules," in *Proc. KDD 3rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. Palo Alto, CA, USA: AAAI Press, 1997, pp. 31–36.

[57] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal, "Mining minimal non-redundant association rules using frequent closed itemsets," in *Proc. CL 1st Int. Conf. Comput. Log.* Berlin, Germany: Springer, 2000, pp. 972–986.

[58] R. J. Hilderman and H. J. Hamilton, *Knowledge Discovery and Measures of Interest. The Springer International Series in Engineering and Computer Science*. New York, NY, USA: Springer, 2001.

[59] J. Han and Y. Fu, "Discovery of multiple-level association rules from large databases," in *Proc. VLDB 21th Int. Conf. Very Large Data Bases*, 1995, pp. 420–431.

[60] H. Lu, L. Feng, and J. Han, "Beyond intratransaction association analysis: Mining multidimensional intertransaction association rules," *ACM Trans. Inf. Syst.*, vol. 18, no. 4, pp. 423–454, 2000, doi: 10.1145/358108.358114.

[61] I. Fister and I. Fister, "Association rules over time," 2020, *arXiv:2010.03834*.

[62] P. Fournier-Viger, J. Li, J. C.-W. Lin, T. T. Chi, and R. U. Kiran, "Mining cost-effective patterns in event logs," *Knowl.-Based Syst.*, vol. 191, Mar. 2020, Art. no. 105241. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705119305581

[63] M. Shahin, S. A. Peious, R. Sharma, M. Kaushik, S. B. Yahia, S. A. Shah, and D. Draheim, "Big data analytics in association rule mining: A systematic literature review," in *Proc. 3rd Int. Conf. Big Data Eng. Technol. (BDET)*, Jan. 2021, pp. 40–49, doi: 10.1145/3474944.3474951.

[64] P. Y. Taser, K. U. Birant, and D. Birant, "Multitask-based association rule mining," *Turkish J. Elect. Eng. Comput. Sci.*, vol. 28, no. 2, pp. 933–955, 2020.

[65] K. E. Iverson, *A Programming Language*. Hoboken, NJ, USA: Wiley, 1962.

[66] R. Schutt and C. O'Neil, *Doing Data Science: Straight Talk From the Frontline*. Sebastopol, CA, USA: O'Reilly Media, 2013.

[67] S. Naik. (2016). *NJ Teacher Salaries*. [Online]. Available: https://data.world/sheilnaik/nj-teacher-salaries-2016

[68] M. Kalish. (2011). *DC Public Employee Salaries*. [Online]. Available: https://data.world/codefordc/dc-public-employee-salaries-2011

[69] R. Sharma and S. A. Peious. (2020). *UDS1*. [Online]. Available: https://github.com/rahulgla/unification/blob/master/UDS1.xlsx

[70] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, "YALE: Rapid prototyping for complex data mining tasks," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2006, pp. 935–940.

RAHUL SHARMA (Graduate Student Member, IEEE) received the B.Tech. and M.Tech. degrees in computer science engineering from Dr. A. P. J. Abdul Kalam Technical University, India. He is currently pursuing the Ph.D. degree in computer science engineering with the Information Systems Group, Tallinn University of Technology, Estonia. He was an Assistant Professor with the Department of Information Technology, Ajay Kumar Garg Engineering College, Ghaziabad, India. His research interests include association rule mining, data science, big data, machine learning, and deep learning.

MINAKSHI KAUSHIK received the B.Tech. and M.Tech. degrees in computer science engineering from Dr. A. P. J. Abdul Kalam Technical University, India. She is currently pursuing the Ph.D. degree in computer science engineering with the Information Systems Group, Tallinn University of Technology, Estonia. Her research interests include association rule mining, big data, machine learning, and deep learning.

SIJO ARAKKAL PEIOUS received the master's degree in computer application from Annamalai University, India, in 2011, and the master's degree in e-governance technologies and services from the Tallinn University of Technology, Estonia, in 2019, where he is currently pursuing the Ph.D. degree with the Department of Software Science. His research interests include association rule mining, big data, machine learning, and deep learning.

ALEXANDRE BAZIN received the Ph.D. degree from Université Pierre et Marie Curie, in 2014. He is currently working at the Lorraine Research Laboratory in Computer Science and its Applications (LORIA), Nancy, France, as a Postdoctoral Researcher. His research interests include lattice theory, symbolic approaches to pattern mining, and explainable artificial intelligence.

SYED ATTIQUE SHAH received the Ph.D. degree from the Institute of Informatics, Istanbul Technical University, Istanbul, Turkey. During his Ph.D. degree, he studied as a Visiting Scholar with the National Chiao Tung University, Taiwan, The University of Tokyo, Japan, and the Tallinn University of Technology, Estonia, where he completed the major content of his thesis. He worked as an Assistant Professor and the Chairperson at the Department of Computer Science, BUITEMS, Quetta, Pakistan. He was also engaged as a Lecturer at the Data Systems Group, Institute of Computer Science, University of Tartu, Estonia. His research interests include big data analytics, cloud computing, information management, and the Internet of Things.

IZTOK FISTER, JR. (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Maribor, Slovenia. He is currently an Assistant Professor at the University of Maribor. He has published more than 120 research articles in referred journals, conferences, and book chapters. His research interests include data mining, pervasive computing, optimization, and sport science. He has acted as a program committee member of more than 30 international conferences. Furthermore, he is a member of the editorial boards of five different international journals.

SADOK BEN YAHIA received the Habilitation degree to lead researches in computer sciences from the University of Montpellier, in 2009. His experience in teaching computer science and information systems is around 20 years. He was a Teaching Assistant with the Faculty of Sciences, Tunis, for two years, an Assistant Professor for seven years, and an Associate Professor for four years. Since 2013, he has been a Full Professor with the Faculty of Sciences. He has been a Professor with the Tallinn University of Technology (TalTech), since 2019. His research interests mainly include combinatorial aspects in big data and their applications to different fields, such as data mining, combinatorial analytics (maximum clique problem and minimal transversals), and smart cities (information aggregation and dissemination and traffic prediction).

DIRK DRAHEIM received the Ph.D. degree from Freie Universität Berlin and the Habilitation degree from Universität Mannheim, Germany. He is currently a Full Professor of information systems and the Head of the Information Systems Group, Tallinn University of Technology, Estonia. The Information Systems Group conducts research in large and ultra-large-scale IT systems. He is also an initiator and a leader of numerous digital transformation initiatives.

Appendix 3

[III]

R. Sharma, M. Kaushik, S. A. Peious, M. Shahin, A. S. Yadav, and D. Draheim. Towards unification of statistical reasoning, OLAP and association rule mining: Semantics and pragmatics. In A. Bhattacharya, J. Lee Mong Li, D. Agrawal, P. K. Reddy, M. Mohania, A. Mondal, V. Goyal, and R. Uday Kiran, editors, *Proceedings of DASFAA 2022 – the 27th International Conference on Database Systems for Advanced Applications*, pages 596–603, Cham, 2022. Springer International Publishing

Towards Unification of Statistical Reasoning, OLAP and Association Rule Mining: Semantics and Pragmatics

Rahul Sharma[1]( ), Minakshi Kaushik[1], Sijo Arakkal Peious[1], Mahtab Shahin[1], Amrendra Singh Yadav[2], and Dirk Draheim[1]

[1] Information Systems Group, Tallinn University of Technology, Tallinn, Estonia
{rahul.sharma,minakshi.kaushik,sijo.arakkal,mahtab.shahin, dirk.draheim}@taltech.ee
[2] Vellore Institute of Technology - VIT Bhopal, Bhopal, India

Abstract. Over the last decades, various decision support technologies have gained massive ground in practice and theory. Out of these technologies, statistical reasoning was used widely to elucidate insights from data. Later, we have seen the emergence of online analytical processing (OLAP) and association rule mining, which both come with specific rationales and objectives. Unfortunately, both OLAP and association rule mining have been introduced with their own specific formalizations and terminologies. This made and makes it always hard to reuse results from one domain in another. In particular, it is not always easy to see the potential of statistical results in OLAP and association rule mining application scenarios. This paper aims to bridge the artificial gaps between the three decision support techniques, i.e., statistical reasoning, OLAP, and association rule mining and contribute by elaborating the semantic correspondences between their foundations, i.e., probability theory, relational algebra, and the itemset apparatus. Based on the semantic correspondences, we provide that the unification of these techniques can serve as a foundation for designing next-generation multi-paradigm data mining tools.

Keywords: Data mining · Association rule mining · Online analytical processing · Statistical reasoning

1 Introduction

Nowadays, decision-makers and organizations are using a variety of modern and old decision support techniques (DSTs) with their specific features and limited scope of work. However, in the era of big data and data science, the huge volume and variety of data generated by billions of internet devices demand advanced

This work has been partially conducted in the project "ICT programme" which was supported by the European Union through the European Social Fund.

A. Bhattacharya et al. (Eds.): DASFAA 2022, LNCS 13245, pp. 596–603, 2022.
https://doi.org/10.1007/978-3-031-00123-9_48

DSTs that can handle a variety of decision support tasks. Currently, no single DST can fulfill this demand. Therefore, to provide advanced decision support capabilities, this paper contributes by elaborating the semantic correspondences between the three popular DSTs, i.e., statistical reasoning (SR) [13], online analytical processing (OLAP) [3] and association rule mining (ARM) [1,11]. These correspondences between SR, ARM and OLAP, and vice versa, appear to be easy, but none of these have been implemented in practice, nor they have been discussed in the state of the art. However, substantial research has been done over the years to enhance OLAP, data warehousing, and data mining approaches [7]. In particular, in data mining, Kamber et al. [8], Surjeet et al. [2] have presented different ways to integrate OLAP and ARM together. Later, Han et al. [5] have proposed DBMiner for interactive mining. In the state of the art, the adoption of concepts in between OLAP and ARM (and vice versa) are referred to as automatic OLAP [14] and multi-dimensional ARM [8]. We appraise all approaches for the integration of the OLAP and ARM. However, the concept of semantic correspondences between DSTs is yet to be elaborated in the state-of-the-art. To establish semantic correspondences between the three DSTs, we use probability theory and conditional expected values (CEVs) at the center of our considerations. CEVs correspond to *sliced average aggregates* in OLAP and would correspond to potential *ratio-scale confidences* in a generalized ARM [4]. Elaborating these concepts between DSTs will enable decision-makers to work with cross-platform decision support tools [6,10] and check their results from different viewpoints.

The paper is structured as follows: In Sect. 2, we elaborate semantic mapping between the SR and ARM, i.e., between probability theory and itemset apparatus. In Sect. 3, we discuss the semantic mapping between the SR and OLAP, i.e., between probability theory and relational algebra. Conclusion is given in Sect. 4.

2 Semantic Mapping Between SR and ARM

We stick to the original ARM concepts and notation provided by Agrawal et al. [1]. However, ARM is also presented for numerical data items as quantitative ARM [12], numerical ARM [9].

In classical ARM, first, there is a *whole itemset* $\mathfrak{I} = \{I_1, I_2, \ldots, I_n\}$ consisting of a *total number* n of items $I_1, I_2, \ldots, I_n$. A subset $X \subseteq \mathfrak{I}$ of the whole itemset is called an *itemset*. We then introduce the concept of a *set of transactions* T (*that fits the itemset* $\mathfrak{I}$) as a relation as follows:

$$T \subseteq TID \times \underbrace{\{0,1\} \times \cdots \times \{0,1\}}_{n-\text{times}} \quad (1)$$

Here, TID is a finite set of transaction identifiers. For the sake of simplicity, we assume that it has the form $TID = \{1, \ldots, N\}$. In fact, we must impose a uniqueness constraint on TID, i.e., we require that T is right-unique, i.e., a function given as,

$$T \in TID \longrightarrow \underbrace{\{0,1\} \times \cdots \times \{0,1\}}_{n-\text{times}} \tag{2}$$

Given (2), we have that N in $TID = \{1, \ldots, N\}$ equals the size of T, i.e., $N = |T|$. Henceforth, we refer to T interchangeably both as a relation and as a function, according to (1) resp. (2). For example, we use $t = \langle i, i_1, \ldots i_n \rangle$ to denote an arbitrary transaction $t \in T$; similarly, we use $T(i)$ to denote the i-th transaction of T more explicitly etc. Given this formalization of the transaction set T, it is correct to say that T is a binary relation between TID and the whole itemset. In that, $I_1, I_2, \ldots, I_n$ need to be thought of as column labels, i.e., there is exactly one bitmap column for each of the n items in $\mathfrak{I}$, compare with (1) and (2). Similarly, Agrawal et al. have called the single transaction a bit vector and introduced the notation $t[k]$ for selecting the value of the transaction t in the k-th column of the bitmap table (in counting the columns of the bitmap table, the TID column is omitted, as it merely serves the purpose of providing transaction identities), i.e., given a transaction $\langle tid, i_1, \ldots i_n \rangle \in T$, we define $\langle tid, i_1, \ldots i_n \rangle [k] = i_k$. Less explicit, with the help of the usual tuple projection notation π_j, we can define $t[k] = \pi_{k+1}(t)$. Let us call a pair $\langle \mathfrak{I}, T \rangle$ of a whole itemset $\mathfrak{I}$ and a set of transaction T that fits $\mathfrak{I}$ as described above an *ARM frame*. Henceforth, we assume an ARM frame $\langle \mathfrak{I}, T \rangle$ as given.

A transaction, as previously stated, is a bit vector. For the sake of simplicity, Let's start with some notation that makes it possible to treat a transaction as an itemset. Given a transaction $t \in T$ we denote the *set of all items that occur in* t as $\{t\}$ and we define it as follows:

$$\{t\} = \{I_k \in \mathfrak{I} \mid t[k] = 1\} \tag{3}$$

The {t} notation provided by (3) will prove helpful later because it allows us to express transaction properties without having to use bit-vector notation, i.e., without having to keep track of item numbers k of items I_k.

Given an $I_j \in \mathfrak{I}$ and a transaction $t \in T$, Agrawal says [1] that I_j *is bought by* t if and only if $t[j] = 1$. Similarly, we can say that t *contains* I_j in such case. Next, given an itemset $X \subseteq \mathfrak{I}$ and a transaction $t \in T$, Agrawal says that t *satisfies* X if and only if $t[j] = 1$ for all $I_j \in X$. Similarly, we can say that t *contains all of the items of* X in such case. Next, we can see that t satisfies X if and only if $X \subseteq \{t\}$. Henceforth, we use $X \subseteq \{t\}$ to denote that t satisfies X.

Given an itemset $X \subseteq \mathfrak{I}$, the relative number of all transactions that satisfy X is called the *support of* X and is denoted as $Supp(X)$, i.e., we define:

$$Supp(X) = \frac{|\{t \in T \mid X \subseteq \{t\}\}|}{|T|} \tag{4}$$

It's perfectly reasonable to discuss an itemset's support once more. X as the relative number of all transactions that each contain all of the items of X.

An ordered pair of itemsets $X \subseteq \mathfrak{I}$ and $Y \subseteq \mathfrak{I}$ is called an *association rule*, and is denoted by $X \Rightarrow Y$. Now, the relative number of all transactions that satisfy Y among all of those transactions that satisfy X is called the *confidence of* $X \Rightarrow Y$, and is denoted as $Conf(X \Rightarrow Y)$, i.e., we define:

$$Conf(X \Rightarrow Y) = \frac{|\{\, t \in T \mid Y \subseteq \{t\} \,\wedge\, X \subseteq \{t\}\,\}|}{|\{t \in T \mid X \subseteq \{t\}\}|} \tag{5}$$

Usually, the confidence of an association rule is introduced via supports of itemsets as follows:

$$Conf(X \Rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X)} \tag{6}$$

It can easily be checked that (5) and (6) are equivalent.

2.1 Semantic Mapping Between Association Rule Mining and SR (Probability Theory)

Here, we compare probability theory to the concepts defined in ARM. Given an ARM frame $F = \langle \mathfrak{I}, T \rangle$. next we map the concepts defined in ARM to probability space $(\varOmega_F, \varSigma_F, \mathsf{P}_F)$. First, we define the set of outcomes $\varOmega_F$ to be the set of transactions T. Next, we define $\varSigma_F$ to be the power set of $\varOmega_F$. Finally, given an event $X \in \varSigma_F$, we define the probability of X as the relative size of X, as follows:

$$\varOmega_F = T \tag{7}$$

$$\varSigma_F = \mathbb{P}(T) \tag{8}$$

$$\mathsf{P}_F(X) = \frac{|X|}{|T|} \tag{9}$$

In the sequel, we drop the indices from $\varOmega_F$, $\varSigma_F$, and P_F, i.e., we simply use $\varOmega$, $\varSigma$, and P to denote them, but always keep in mind that we actually provide a mapping from ARM frames F to corresponding probability spaces $(\varOmega_F, \varSigma_F, \mathsf{P}_F)$. The idea is simple. Each transaction is modeled as an outcome and, as usual, also a basic event. Furthermore, each set of transactions is an event.

We step forward with item and itemsets. For each item $I \in \mathfrak{I}$ we introduce the *event that item* I *is contained in a transaction*, and we denote that event as $[\![I]\!]$. Next, for each itemset $X \subseteq \mathfrak{I}$, we introduce the *event that all of the items in* X *are contained in a transaction* and we denote that event as $[\![X]\!]$. We define:

$$[\![I]\!] = \{\, t \mid I \in \{t\} \,\} \tag{10}$$

$$[\![X]\!] = \bigcap_{I \in X} [\![I]\!] \tag{11}$$

As usual, we identify an event $[\![I]\!]$ with the characteristic random variable $[\![I]\!] : \varOmega \longrightarrow \{0,1\}$ and use $\mathsf{P}([\![I]\!])$ and $\mathsf{P}([\![I]\!]{=}1)$ as interchangeable.

2.2 Formal Mapping of ARM Support and Confidence to Probability Theory

Based on the mapping provided by (7) through (11), we can see how ARM *Support* and *Confidence* translate into probability theory.

Lemma 1 (Mapping ARM *Support* to Probability Theory) *Given an itemset $X \subseteq \mathfrak{I}$, we have that:*

$$\mathit{Supp}(X) = \mathsf{P}(\llbracket X \rrbracket) \tag{12}$$

Proof. According to (11), we have that $\mathsf{P}(\llbracket X \rrbracket)$ equals

$$\mathsf{P}(\bigcap_{I \in X} \llbracket I \rrbracket) \tag{13}$$

Due to (10), we have that (13) equals

$$\mathsf{P}\Big(\bigcap_{I \in X} \{\, t \in T \mid I \in \{t\}\,\}\Big) \tag{14}$$

We have that (14) equals

$$\mathsf{P}(\{\, t \in T \mid \bigwedge_{I \in X} I \in \{t\}\,\}) \tag{15}$$

We have that (15) equals

$$\mathsf{P}(\{t \in T \mid X \subseteq \{t\}\}) \tag{16}$$

According to (9), we have that (16) equals

$$\frac{|\{t \in T \mid X \subseteq \{t\}\}|}{|T|} \tag{17}$$

According to (4), we have that (17) equals $\mathit{Supp}(X)$ □

Lemma 2 (Mapping ARM *Confidence* to Probability Theory) *Given an itemset $X \subseteq \mathfrak{I}$, we have that:*

$$\mathit{Conf}(X \Rightarrow Y) = \mathsf{P}\big(\llbracket Y \rrbracket \,\big|\, \llbracket X \rrbracket\big) \tag{18}$$

Proof. Omitted.

With these mappings, we provide that a set of items in ARM $\mathfrak{I} = \{I_1, I_2, \ldots, I_m\}$ are equivalent to the set of events $\mathfrak{I} = \{I_1 \subseteq \Omega, \ldots, I_m \subseteq \Omega\}$ in probability theory. Transactions T in ARM are equivalent to the set of outcomes Ω in probability space $(\Omega, \Sigma, \mathsf{P})$. Support of an itemset X in ARM is equivalent to the relative probability of the itemset X. Confidence of an association rule $X \Rightarrow Y$ is equivalent to the conditional probability of Y in the presence of X.

3 Semantic Mapping Between SR and OLAP

As per our findings, conditional operations on bitmap (Binary) columns correspond to conditional probabilities, whereas conditional operations on numerical columns correspond to conditional expected values, e.g., we model a sample OLAP Table 1 in probability theory. We consider that Table 1 is equivalent to the set of outcomes Ω in probability space (Ω, Σ, P), a row r is an element of Ω, i.e. $r \in \Omega$ and each column c is equivalent to a random variable $\mathbb{R}$. We consider numerical columns as *finite real-valued* random variables (For Example: Salary $\in \Omega \subseteq \mathbb{R}$) and bitmap columns are considered as events (For Example: Freelancer $\subseteq \Omega$). The following is a probabilistic interpretation of the OLAP Table 1.

Table 1. A sample OLAP table.

City	Profession	Education	Age group	Freelancer	Salary
New York	Lawyer	Master	25–30	0	3.800
Seattle	IT	Bachelor	18–25	1	4.200
Boston	Lawyer	PhD	40–50	1	12.700
L.A	Chef	High School	30–40	0	3.700
...	...	...	...	...	...

3.1 Semantic Mapping Between OLAP Averages and SR

Generally, decision-makers use SQL queries to interact with OLAP [3]. Therefore, we use OLAP queries to be mapped with SR, i.e., probability theory. We have a simple OLAP average query; (SELECT AVG(Salary) FROM Table 1). If the number of rows of Table 1 is represented by $|\Omega|$ and the number of rows that contain a value i in column C are equivalent to $\#_C(i)$ then AVG(Salary) FROM Table 1 will compute the average of all the salaries, i.e., a fraction of the sum of the column (Salary) and the total number of rows in the table. In probability theory, the *average* of a random variable X is the *Expected Value* of $X = \mathsf{E}[X]$. We compare the expected value of X, i.e., $\mathsf{E}(X)$ with the output of the AVG query in OLAP. We have:

$$OLAP - Query\ (SELECT\ AVG\ (Salary)\ FROM\ Table\ 1) \tag{19}$$

$$\text{Expected Value: } \mathsf{E}(\text{Salary}) = \sum_{i \in I_{\text{Salary}}} i \cdot \mathsf{P}(\text{Salary} = i) \tag{20}$$

$$= \sum_{i \in I_{\text{Salary}}} i \cdot \frac{\#_{\text{Salary}}(i)}{|\Omega|} = \frac{\sum\limits_{r \in \Omega} \text{Salary}(r)}{|\Omega|} \tag{21}$$

As per Eq. 20 and Eq. 21, the average of a random variable X in probability theory and simple averages of an OLAP query provides the same outcome. Hence, we say that an average query in OLAP corresponds to expected values in probability theory. The conditional average queries in OLAP calculate averages of a column with a WHERE clause. For example, we have an average SQL query with some conditions where the target column is numerical and conditional variables have arbitrary values. We have: SELECT AVG(Salary) FROM Table 1 WHERE City = *Seattle* AND Profession=*IT*;. In probability theory, we compute the conditional average of a random number using its conditional expectation. Therefore, the conditional expectation of a random number Y with condition X is given as:

$$\mathsf{E}(Y|X) = \sum_{n=0}^{\infty} i_n \cdot \mathsf{P}(Y \;= i_n|\, X) \tag{22}$$

$$f(i) = E(Y = i_n|X) \tag{23}$$

Here, the value $E(Y = i_n|X)$ is dependent on the value of i. Therefore, we say that $E(Y = i_n|X)$ is a function of i, which is given in Eq. 23. We compare the conditional expected value of $E(Y = i_n|X)$ with the output of the conditional AVG query in OLAP. We have:

$$OLAP\ Query : SELECT\ AVG(Salary)\ FROM\ Table\ 1$$
$$WHERE\ City = Seattle\ AND\ Profession = IT; \tag{24}$$

$$\text{Conditional Expected Value: } \mathsf{E}(\text{Salary} \,|\, \text{City}{=}\textit{Seattle} \cap \text{Profession}{=}\textit{IT}) \tag{25}$$

$$\mathsf{E}(Y|X) = \sum_{i \in I_C} i \cdot \mathsf{P}(Y{=}i \,|\, X) \tag{26}$$

As per Eq. 25 and Eq. 26, the average of a random variable Y with condition X (Conditional Expected values) and the conditional average of an OLAP query provides the same outcome. Hence, we can say that a conditional average query in OLAP corresponds to the conditional expected values in probability theory. Based on these mappings in OLAP, conditional averages on binary columns correspond to conditional probability and they also correspond to confidence in ARM.

4 Conclusion

In this paper, we elaborated semantic correspondences between the three DSTs, i.e., SR, OLAP and ARM. We identify that SR, OLAP, and ARM operations complement each other in data understanding, visualization, and making individualized decisions. In the proposed mappings, it is identified that OLAP and ARM have common statistical reasoning, exploratory data analysis methods and offer similar solutions for decision support problems. Based on these findings, we can review current obstacles in each of SR, OLAP and ARM. Furthermore, the semantic correspondences between the three DSTs will be helpful in designing certain next-generation hybrid decision support tools.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. ACM SIGMOD Rec. **22**(2), 207–216 (1993). https://doi.org/10.1145/170036.170072
2. Chaudhuri, S., Dayal, U.: Data warehousing and olap for decision support. In: Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, SIGMOD 1997, pp. 507–508. Association for Computing Machinery, New York (1997). https://doi.org/10.1145/253260.253373
3. Codd, E.F.: Providing olap (on-line analytical processing) to user-analysts: An it mandate. Available from Arbor Software's web site-http://www.arborsoft.com/papers/coddTOC.html (1993)
4. Hartmann, S., Küng, J., Chakravarthy, S., Anderst-Kotsis, G., Tjoa, A.M., Khalil, I. (eds.): DEXA 2019. LNCS, vol. 11706. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-27615-7
5. Han, J., Fu, Y., Wang, W., Chiang, J., Zaïane, O.R., Koperski, K.: DBMiner: interactive mining of multiple-level knowledge in relational databases. In: Proceedings of SIGMOD'96 - the 1996 ACM SIGMOD International Conference on Management of Data, p. 550. Association for Computing Machinery (1996). https://doi.org/10.1145/233269.280356
6. Heinrichs, J.H., Lim, J.S.: Integrating web-based data mining tools with business models for knowledge management. Decis. Support Syst. **35**(1), 103–112 (2003). https://doi.org/10.1016/S0167-9236(02)00098-2
7. Imieliński, T., Khachiyan, L., Abdulghani, A.: Cubegrades: generalizing association rules. Data Min. Knowl. Disc. **6**(3), 219–257 (2002)
8. Kamber, M., Han, J., Chiang, J.: Metarule-guided mining of multi-dimensional association rules using data cubes. In: Proceedings of VLDB'1994 - the 20th International Conference on Very Large Data Bases, KDD 1997, pp. 207–210. AAAI Press (1997)
9. Kaushik, M., Sharma, R., Peious, S.A., Shahin, M., Yahia, S.B., Draheim, D.: A systematic assessment of numerical association rule mining methods. SN Comput. Sci. **2**(5), 1–13 (2021)
10. Arakkal Peious, S., Sharma, R., Kaushik, M., Shah, S.A., Yahia, S.B.: Grand reports: a tool for generalizing association rule mining to numeric target values. In: Song, M., Song, I.-Y., Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) DaWaK 2020. LNCS, vol. 12393, pp. 28–37. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59065-9_3
11. Sharma, R., Kaushik, M., Peious, S.A., Yahia, S.B., Draheim, D.: Expected vs. unexpected: selecting right measures of interestingness. In: Song, M., Song, I.-Y., Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) DaWaK 2020. LNCS, vol. 12393, pp. 38–47. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59065-9_4
12. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. SIGMOD Rec. **25**(2), 1–12 (1996)
13. Stigler, S.M.: The History of Statistics: The Measurement of Uncertainty Before 1900. Harvard University Press (1986)
14. Zhu, H.: On-line analytical mining of association rules. In: Master's thesis. Simon Fraser University, Burnaby, Brithish Columbia, Canada (1998)

Appendix 4

[IV]

R. Sharma, H. Garayev, M. Kaushik, S. A. Peious, P. Tiwari, and D. Draheim. Detecting Simpson's paradox: A machine learning perspective. In C. Strauss, A. Cuzzocrea, G. Kotsis, A. M. Tjoa, and I. Khalil, editors, *Proceedings of DEXA 2022 – the 33rd International Conference on Database and Expert Systems Applications*, pages 323–335, Cham, 2022. Springer International Publishing

Detecting Simpson's Paradox: A Machine Learning Perspective

Rahul Sharma[1]( ), Huseyn Garayev[2], Minakshi Kaushik[1], Sijo Arakkal Peious[1], Prayag Tiwari[3], and Dirk Draheim[1]

[1] Information Systems Group, Tallinn University of Technology,
Akadeemia tee 15a, 12618 Tallinn, Estonia
{rahul.sharma,minakshi.kaushik,dirk.draheim}@taltech.ee

[2] University of Tartu, Tartu, Estonia
hugara@taltech.ee

[3] Department of Computer Science, Aalto University, Espoo, Finland
prayag.tiwari@aalto.fi

Abstract. The size of data collected around the world is growing exponentially, and it has become popular as big data. The volume and velocity of big data are facilitating the transition of machine learning (ML), deep learning (DL) and artificial intelligence (AI) from research laboratories to real life. There are numerous other claims made about Big Data. Can we, however, rely on data blindly? What happens when a dataset used to train ML models has a hidden statistical paradox? Data, like fossil fuels, is valuable, but it must be refined carefully for accurate outcomes. Statistical paradoxes are hard to observe in classical data cleaning and analysis techniques. Still, they are required to be investigated separately in training datasets. In this paper, we discuss the impact of Simpson's paradox on categorical data and demonstrate its effects on AI and ML application scenarios. Next, we provide an algorithm to automatically identify the confounding variable and detect Simpson's paradox within categorical datasets. The algorithm experiments on datasets from two real-world case studies. The outcome of the algorithm uncovers the existence of the paradox and indicates that Simpson's paradox is severely harmful in automatic data analysis, especially in AI, ML and DL.

Keywords: Big data · Artificial intelligence · Deep learning · Machine learning · Data science · Simpson's paradox · Explainable AI

1 Introduction

Human decision-making has always relied on data, but with the advancement of big data technologies, artificial intelligence (AI), data science, machine learning (ML), and deep learning (DL) have gained significant traction in artificial decision-making. These techniques are now widely used in medical sciences, social sciences, and politics, and they substantially impact human life and decisions, either directly or indirectly. In most AI use cases, ML-based trained artificial

C. Strauss et al. (Eds.): DEXA 2022, LNCS 13426, pp. 323–335, 2022.
https://doi.org/10.1007/978-3-031-12423-5_25

systems are used to provide quick and precise results. Still, in some cases, the existence of statistical paradox, causal inference and uneven data distribution can mislead an AI application. Statistical paradoxes are not new to being discussed in statistics and mathematics. These terms are widely used in statistics and have been around for over a century. Expert mathematicians and statisticians adequately discussed various statistical paradoxes (e.g., Simpson's Paradox, Berkson's Paradox, Latent Variables, Law of Unintended Consequences, Tea Leaf Paradox, etc.) and addressed their severe impacts on classical data analysis. However, in modern decision support techniques, specifically AI, ML and DL, causal relationships, data fallacies and statistical paradoxes are not yet appropriately addressed.

A statistical paradox can exist in a wide variety of data. Kügelgen et al. [33] recently emphasized the importance of statistical analysis of real data and demonstrated evidence of Simpson's paradox in COVID-19 data analysis. They claim Italy's overall case fatality rate (CFR) was higher than China's. However, in every age group, China had a higher fatality rate than Italy. These observations raise numerous concerns about data accuracy and analysis. Heather et al. [20] have addressed the existence of Simpson's paradox. In psychological science, Kievit et al. [17] examined the instances of Simpson's paradox. In [14], Kaushik et al. have discussed some measures to find the impact of one numerical variable on another numerical variable. Alipourfard et al. [2] have discovered the existence of Simpson's paradox in social data and behavioural data [3]. The instances Simpson's paradox have also been discussed in various data mining techniques [10,11,13], e.g., association rule mining [1] and numerical association rule mining [15,16,31]. Therefore, understanding data, especially big data, is more critical than processing.

Most of the statistical paradoxes are fundamentally linked to various statistical challenges and mathematical logic, including causal inference [22,23], the ecological fallacy [19,26], Lord's paradox [32], propensity score matching [27], suppressor variables [8], conditional independence [9], partial correlations [12], p-technique [6], mediator variables [21], etc.

In this paper, we concentrate on a specific case of a statistical paradox called Simpson's paradox in categorical data and demonstrate its impact with some real-world case studies. Next, we provide an algorithm to detect Simpson's paradox and identify the confounding variables in categorical values. In statistics, a confounder is described as a statistic variable that influences both the dependent and independent variables, resulting in a spurious relationship. The algorithm is experimented on two datasets to detect confounder and the paradox. The paper is organized as follows.

In Sect. 2, we discuss Simpson's Paradox. In Sect. 3, we propose an algorithm for automatically detecting the Simpson's Paradox in categorical values. In Sect. 4, two real-life datasets are used to demonstrate the impact of the paradox experimentally. Finally, a discussion and conclusion is provided in Sect. 5 and Sect. 6, respectively.

2 Simpson's Paradox

In the year 1899, Karl Pearson et al. [24] demonstrated a statistical paradox in marginal and partial associations between continuous variables. Later in 1903, Udny Yule [35] explained "the theory of association of attributes in statistics" and revealed the existence of an association paradox with categorical variables. In a technical paper published in 1951 [29], Edward H. Simpson described the phenomenon of reversing results. However, in 1972, Colin R. Blyth coined the term "Simpsons Paradox" [5]. Therefore, this paradox is known by different names and is famous as the Yule-Simpson effect, amalgamation paradox, or reversal paradox [25]. Simpson's paradox can exist in any dataset irrespective of its size and type [18]. The paradox demonstrates the importance of having human experts in the loop during an automatic data analysis.

Table 1. Original Simpson's example with 2 × 2 contingency table [29]: the type of association for the entire population ($N = 52$) reverses at the level of sub-populations of men and women.

	Population $N = 52$			Men (M)= 20			Women (F) $= 32$		
	Success (S)	Failure ($\neg S$)	Success rate %	Success	Failure	Success Rate %	Success	Failure	Success rate %
T	20	20	50%	8	5	≈61%	12	15	≈44%
$\neg T$	6	6	50%	4	3	≈57%	2	3	≈40%

We start the discussion on the paradox by using the original example and numbers from Simpson's article [29]. In this example, analysis for medical treatment is demonstrated. Table 1 summarises the effect of the medical treatment for the entire population ($N = 52$) as well as for men and women separately in subgroups. The treatment appears effective for both men and women subgroups (Men: 61% vs 57% and Women: 44% vs 40%); however, the treatment seems ineffective at the whole population level.

We can demonstrate the above example via probability theory and conditional probabilities. Let $T = treatment$, $S = success$, $M = Men$ and $F = Women$ then,

$$\mathsf{P}(S \mid T) = \mathsf{P}(S \mid \neg T) \tag{1}$$

However, the probability for men and women is:

$$\mathsf{P}(S \mid T, M) > \mathsf{P}(S \mid \neg T, M) \tag{2}$$

$$\mathsf{P}(S \mid T, F) > \mathsf{P}(S \mid \neg T, F) \tag{3}$$

Based on Eq. 1, 2 and 3, one should use the treatment or not? As per the success rate for the men and women populations, the treatment is a success, but overall, the treatment is a failure. This reversal of results between groups population and the total population has been referred to as Simpson's Paradox. In statistics, this concept has been discussed widely and named differently by several authors [24,35].

2.1 Impacts of Simpson's Paradox

Simpson's paradox exists in different types of data in different forms. However, classically it is expressed via 2 × 2 contingency tables. Let a 2 × 2 contingency table for treatment (T) and success (S) in the i^{th} sub-population is represented by a four-dimensional vector of real numbers $D = (a_i, b_i, c_i, d_i)$. Then

Table 2. 2 × 2 Contingency table with sub population groups D1 and D2.

	Population $D = D_1 + D_2$		Sub-population D_1		Sub-population D_2	
	Success (S)	Failure $(\neg S)$	Success (S)	Failure $(\neg S)$	Success (S)	Failure $(\neg S)$
Treatment (T)	$a_1 + a_2$	$b_1 + b_2$	a_1	b_1	a_2	b_2
No-Treat. $(\neg T)$	$c_1 + c_2$	$d_1 + d_2$	c_1	d_1	c_2	d_2

$$D = \sum\nolimits_{i=1}^{N} D_i = \left(\sum a_i, \sum b_i, \sum c_i, \sum d_i\right) \tag{4}$$

is the aggregate dataset over N sub populations [30]. This can be read as given in Table 2.

Definition 1. *Consider n groups of data such that group i has A_i trials and $0 \leq X_{A_i} \leq A_i$ "successes". Similarly, consider another similar n groups of data such that group i has B_i trials and $0 \leq Y_{B_i} \leq B_i$ "successes". Then, Simpson's paradox appear if:*

$$\frac{X_{A_i}}{A_i} \leq \frac{Y_{B_i}}{B_i} \text{ for all } i = 1, 2, \ldots, n \text{ but } \frac{\sum_{i=1}^{n} X_{A_i}}{\sum_{i=1}^{n} A_i} \geq \frac{\sum_{i=1}^{n} Y_{B_i}}{\sum_{i=1}^{n} B_i} \tag{5}$$

We could also flip the inequalities and still have the paradox since A and B are chosen arbitrarily.

$$\frac{X_{A_i}}{A_i} \geq \frac{Y_{B_i}}{B_i} \text{ for all } i = 1, 2, \ldots, n \text{ but } \frac{\sum_{i=1}^{n} X_{A_i}}{\sum_{i=1}^{n} A_i} \leq \frac{\sum_{i=1}^{n} Y_{B_i}}{\sum_{i=1}^{n} B_i} \tag{6}$$

We use the following example to show the working of the Eqs. 5 and 6.

$$\frac{10}{20} = \frac{X_{A_1}}{A_1} > \frac{Y_{B_1}}{B_1} = \frac{30}{70} \; and \; \frac{10}{50} = \frac{X_{A_2}}{A_2} > \frac{Y_{B_2}}{B_2} = \frac{10}{60} \text{ yet}$$

$$\frac{10+10}{20+50} = \frac{20}{70} = \frac{X_{A_1} + X_{A_2}}{A_1 + A_2} < \frac{Y_{B_1} + Y_{B_2}}{B_1 + B_2} = \frac{30+10}{70+60} = \frac{40}{130}$$

3 Detecting Simpson's Paradox

Based on the type of trends reversed in various types of data, Simpson's paradox cases are explored into two categories: classification, which involves the relative rates of binary outcomes in two groups, and regression, which involves the sign of a correlation between two variables [34]. Here, we provide an algorithm to detect the paradox in the first case, i.e. for categorical values. In the algorithm, the Pearson correlation index is used to find the relationships between two variables which allows for measuring the strength of the linear association between two variables. The output value of the Pearson correlation lies between -1 and 1. Values greater than 0 imply a positive correlation. The value 1 indicates the exact positive association, while 0 means no correlation. Values less than 0 suggest a negative association, and -1 indicates a clear negative association. The Pearson correlation coefficient is represented by r In Eq. 7. Here, x and y are input vectors, $\bar{x}$ and $\bar{y}$ are means of the variables, respectively.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}} \tag{7}$$

3.1 Algorithm for Detecting the Simpson's Paradox in Categorical Data (Relative Rates)

We formally describe the algorithm for detecting the Simpson's Paradox in linear trends in Algorithm 1. In the algorithm, the primary step is to convert the values of the categorical input variables to binary values. The first variable category is substituted by 0, and the second category is replaced by 1. This conversion allows the Pearson correlation index function to identify the relationship between categorical variables or between categorical and numerical (continuous) variables. We input X - categorical variable by which we condition, $X1$ - the first category of variable X, $X2$ - the second category of variable X, Y - continuous or categorical variable (with two categories) which is aggregated. Table 3 illustrates the form of an example dataset before and after the pre-processing step.

Further, the algorithm calculates the correlation index between X and Y variables with the values of the corresponding columns in the dataset. This way, we obtain information on the sign of the relationship between the variables. Next, we traverse the list of remaining categorical variables, calculate the Pearson index conditioning on each subgroup (category), count the ratio of subgroups where the correlation index reversed relative to the index in aggregated data and store the value key pairs in an array. Subsequently, we get the array element where the value (ratio) is the highest. The maximal value 1 implies the Simpson's paradox occurrence with the corresponding key of the array element being the confounding variable. Cases where the maximal ratio is less than 1 imply the absence of Simpson's paradox. However, they are also regarded as a partial occurrence of the bias and are considered in the further steps. The performance of the algorithm strongly correlates with the size of the datasets.

Algorithm 1: Identification of Simpson's Paradox in Relative Rates

```
Input: A dataset $D$ with categorical variable $x$ and $y$
Output: a pair of confounding variable and ratio of reversed association
d[x] = Preprocess(d[x]) /*conversion of categorical column to binary */
d[y] = Preprocess(d[y])
aggreg_index = Pearson(d[x] , d[y]) /*calculate correlation index between
  columns */
indexes = [] /*initialize index array to store key value pairs: the
  key is column and value is the number of reversed subgroups */
cols = columns(D) /*initialize array of all columns of D */
foreach column ∈ cols do
    if Column Is Not Categorical(column) then
        Continue
    end
    else
        subgroups = Categories(column) // get the categories of a column
        coefficients = [] // initialize empty array to store the correlation indexes
        foreach subgroup ∈ subgroups do
            disaggreg_index = Pearson( D[x]: where D[column] = subgroup,
              D[y]: where D[column] = subgroup) calculate corr. index between
              columns for current subgroup
            Add index of disaggregated to correlation indexes array
        end
    end
    reversed_subgroups = RatioReversedSubgroups(aggreg_index, coefficients)
      /*calculate ratio of the correlation indexes reversed with
      respect to the correlation index for the aggregated data */
    Add column, reversed_subgroups values into indexes
end
Store the max values of indexes pairs into result
Return result
```

Table 3. Illustration of the form of an example dataset before and after the preprocessing step.

Gender	Result
Male	Success
Female	Success
Male	Failure

Gender	Result
0	1
1	1
0	0

4 Experiments and Datasets

The Algorithm is implemented in Python on a personal computer with an Intel(R) Core(TM) i5-8265U CPU @ 1.60 GHz, 1800 Mhz, 4 Core(s), 8 Logical Processor(s), 16 GB RAM and Windows 10 × 64 operating system. We evaluate the algorithm with two real-world case studies with categorical data. The programming code, datasets, and other necessary instructions about the algorithms are available in the GitHub repository [28].

4.1 UC Berkeley Admissions Dataset Fall 1973

UC Berkeley admissions dataset [4] is a classic example of Simpson's Paradox. This dataset contained 12,763 graduate applicants (males and females) to UC-Berkeley in Fall 1973. The dataset was provided by UC-Berkeley researchers to investigate any possible cases of gender bias in the admissions. In the dataset, the admission rate for females is less than for males when data is aggregated; however, when we consider each major separately, female admission rates exceed the rates for males in most subgroups.

The aggregate data given in Table 4 demonstrate significant bias in favour of male applicants; however, data from each department given in Table 5 reveals an opposite story and bias in favour of Female applicants. Figure 1 demonstrate some hidden patterns in the dataset. As per the graph, it is clear that the overall number of women applicants is significantly less than the total men applicants. However, their rejection rate is high as compared to the male applicants. To analyze these hidden patterns and find the possible existence of Simpson's paradox in data, we use the original UC-Berkeley admission dataset having 12763 records with four attributes: *Student_id*, *Gender*, *Major* and *Admission*.

Table 4. Existence of Simpson's Paradox: a case study from UC-Berkeley admission dataset (fall 1973) [4].

	Applications	Admitted	Rejected	Admission %
Men	8442	3738	4704	44%
Women	4321	1494	2827	35%

Table 5. UC-Berkeley admission dataset (fall 1973): Percentage of acceptance rate of men and women in different departments.

Gender	Departments					
	A	B	C	D	E	F
Men	62.06%	63.04%	36.92%	33.09%	27.75%	5.90%
Women	82.41%	68%	34.06%	34.93%	23.92%	7.04%

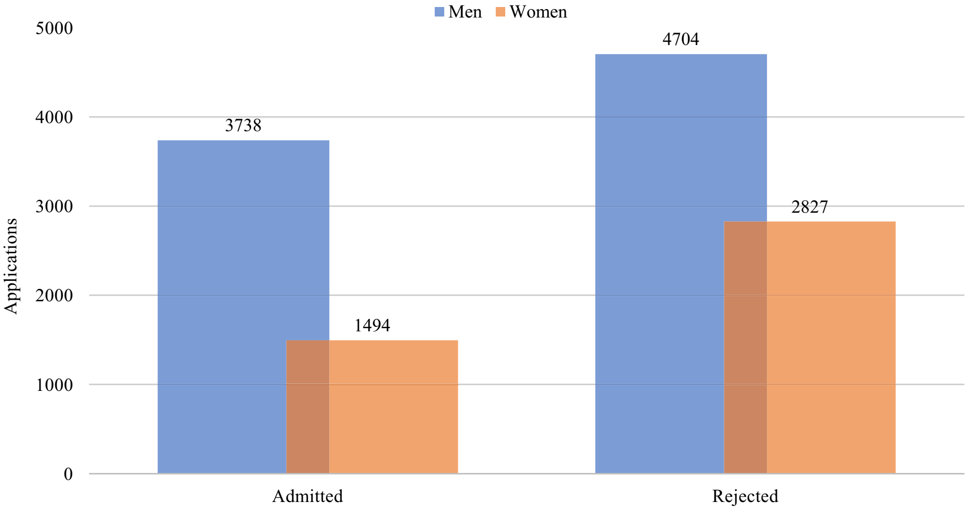


Fig. 1. Graphical representation of information in the UC-Berkeley admission dataset demonstrates hidden patterns and unbalanced data distribution.

In the algorithm, *Gender* attribute is set as X variable and *Admission* attribute is set as Y variable. To detect the paradox, the algorithm first calculates the Pearson correlation between *Gender* and *Admission* variables. In the prepossessing step, the values of *gender* variable, i.e., *Female* and *Male* are categorised by the binary values 1 and 0, similarly, the values of *admission* variable, i.e., *Failure* and *Success* are categorised by the binary values 0 and 1, respectively. Next, the algorithm traverses the complete list of variables to identify the possible confounding variable and compute the ratio of the subgroup reversals. The algorithm returns a confounder and the existence of Simpson's paradox in the dataset. As per the computation, the correlation index between the *Gender* and *Admission* variable is negative for "B, F, A, D" majors, whereas it is positive for the whole population.

4.2 Kidney Stone Treatment Dataset

We use another dataset from a real-world medical case study published by Charig et al. [7] in "The British Medical Journal" in 1986. In this study, the success rate of two different types of treatments to remove the large and small size of kidney stones are compared. In Table 6, Treatment A entails a classical open surgical procedure and treatment B entails an advanced closed surgical procedure. For both small kidney stones and large kidney stones, treatment A, i.e., open surgical procedures (*Success Rate* Small Stone Size 93%, Large Stone Size 73%) performs better than the treatment B (*Success Rate:* Small Stone Size 87%, Large Stone Size 69%), However, when the data for both the treatments is combined, the treatment B (*Success Rate:* 83%) outperforms the treatment A (*Success Rate:* 73%). Table 6 demonstrates the success rates of the treatments in detail.

Table 6. Kidney Stone Dataset: Information about the success rate of the treatments with different sizes of stones. Treatment A outperforms treatment B for large and small kidney stones, but for both kidney stones together, treatment B exceeds treatment A.

Stone size	Treatment (A) = 350			Treatment (B) = 350		
	Success (S)	Failure (F)	Success rate %	Success (S)	Failure (F)	Success rate %
Small	81	6	≈93%	234	36	≈87%
Large	192	71	≈73%	55	25	≈69%
Both	273	78	≈78%	289	61	≈83%

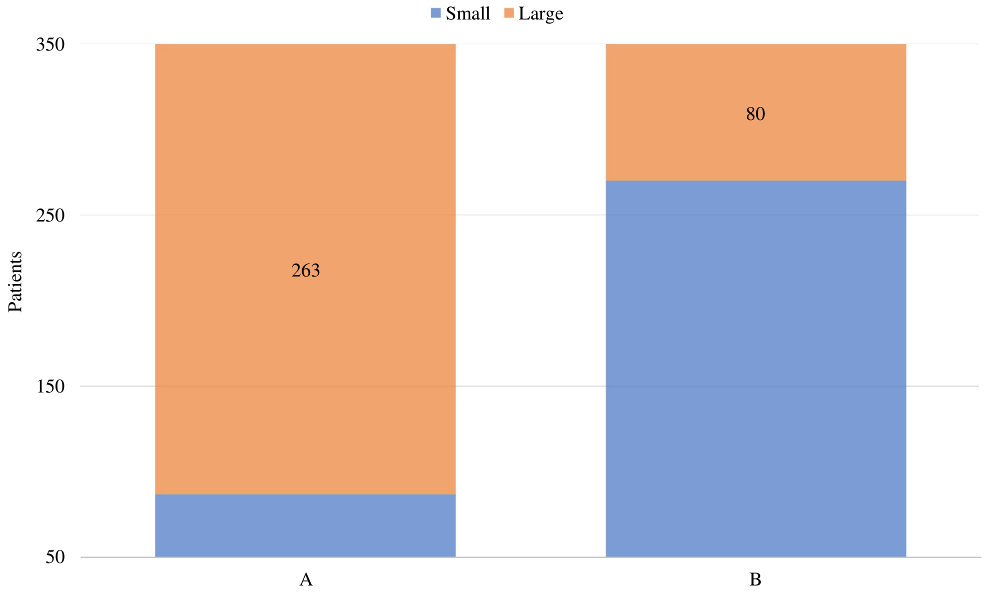


Fig. 2. Graphical representation of information in the kidney stone dataset demonstrates the hidden patterns and unbalanced data distribution for treatments A and B.

Figure 2 demonstrate the graphical representation of the hidden information in the dataset. As per the graphs, it is a perfect case of uneven distribution of sample data for both the treatments. Analyzing this dataset with the algorithm returns a confounder and the existence of Simpson's paradox. As per the computation, the correlation index between the *Treatment A* and *Treatment B* in groups is opposite to the correlation for both the treatments.

5 Discussion and Future Work

The existence of Simpson's paradoxes in real-world studies provides a direction for understanding the impact of causality in artificial decision-making. We noticed that data mining algorithms used in AI, ML and DL focus mainly on identifying the correlations in aggregate data rather than identifying the genuine causal relationships between all the data items. Therefore, understanding statistical paradoxes and evaluating causality in each combination of data items is an essential step toward fair ML models. In future, we plan to simplify the impacts of Simpson's paradox in different types of data (Continuous values) and address various other statistical paradoxes (e.g., Berkson's paradox) in datasets. Further, we intend to develop a simple framework to identify the existence of statistical paradoxes in various types of data.

6 Conclusion

In AI, ML and DL, dealing with causality and statistical paradoxes is still a challenging phenomenon. In most AI use cases, ML-based trained artificial systems are used to provide quick and precise results. Still, in some cases, the existence of statistical paradox, causal inference and uneven data distribution can easily mislead the outcome of artificial systems. In this paper, we focused on addressing a specific case of a statistical paradox called Simpson's paradox in categorical data and demonstrated its impact with some real-world case studies. We provided an algorithm to detect Simpson's paradox and identify the confounding variables in categorical datasets. This algorithm can be utilized to develop a platform that unifies most aspects related to detecting a confounding variable, Simpson's paradox. The algorithm is evaluated on two real-world case study datasets. The algorithm performed well in each experiment, and its running time is proportional to the size of a dataset.

Acknowledgements. This work has been partially conducted in the project "ICT programme" which was supported by the European Union through the European Social Fund.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of VLDB'1994 - the 20th International Conference on Very Large Data Bases, pp. 487–499. Morgan Kaufmann (1994)
2. Alipourfard, N., Fennell, P.G., Lerman, K.: Can you trust the trend? Discovering Simpson's paradoxes in social data. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, pp. 19–27. Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3159652.3159684
3. Alipourfard, N., Fennell, P.G., Lerman, K.: Using Simpson's paradox to discover interesting patterns in behavioral data. In: Proceedings of the Twelfth International AAAI Conference on Web and Social Media. AAAI Publications (2018)
4. Bickel, P.J., Hammel, E.A., O'Connell, J.W.: Sex bias in graduate admissions: data from Berkeley. Science **187**(4175), 398–404 (1975). https://doi.org/10.1126/science.187.4175.398
5. Blyth, C.R.: On Simpson's paradox and the sure-thing principle. J. Am. Stat. Assoc. **67**(338), 364–366 (1972)
6. Cattell, R.B.: P-technique factorization and the determination of individual dynamic structure. J. Clin. Psychol. **8**, 5–10 (1952)
7. Charig, C.R., Webb, D.R., Payne, S.R., Wickham, J.E.: Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. BMJ **292**(6524), 879–882 (1986). https://doi.org/10.1136/bmj.292.6524.879
8. Conger, A.J.: A revised definition for suppressor variables: a guide to their identification and interpretation. Educ. Psychol. Meas. **34**(1), 35–46 (1974)
9. Dawid, A.P.: Conditional independence in statistical theory. J. Roy. Stat. Soc. Ser. B (Methodol.) **41**(1), 1–15 (1979). https://doi.org/10.1111/j.2517-6161.1979.tb01052.x
10. Draheim, D.: DEXA'2019 keynote presentation: future perspectives of association rule mining based on partial conditionalization, Linz, Austria, August 2019. https://doi.org/10.13140/RG.2.2.17763.48163
11. Draheim, D.: Future perspectives of association rule mining based on partial conditionalization. In: Hartmann, S., Küng, J., Chakravarthy, S., Anderst-Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) Proceedings of DEXA'2019 - the 30th International Conference on Database and Expert Systems Applications. LNCS, vol. 11706, p. xvi. Springer, Heidelberg (2019)
12. Fisher, R.A.: III. The influence of rainfall on the yield of wheat at Rothamsted. Philos. Trans. R. Soc. London Ser. B **213**(402–410), 89–142 (1925). Containing Papers of a Biological Character
13. Freitas, A.A., McGarry, K.J., Correa, E.S.: Integrating Bayesian networks and Simpson's paradox in data mining. In: Texts in Philosophy. College Publications (2007)
14. Kaushik, M., Sharma, R., Peious, S.A., Draheim, D.: Impact-driven discretization of numerical factors: case of two- and three-partitioning. In: Srirama, S.N., Lin, J.C.-W., Bhatnagar, R., Agarwal, S., Reddy, P.K. (eds.) BDA 2021. LNCS, vol. 13147, pp. 244–260. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-93620-4_18

15. Kaushik, M., Sharma, R., Peious, S.A., Shahin, M., Ben Yahia, S., Draheim, D.: On the potential of numerical association rule mining. In: Dang, T.K., Küng, J., Takizawa, M., Chung, T.M. (eds.) FDSE 2020. CCIS, vol. 1306, pp. 3–20. Springer, Singapore (2020). https://doi.org/10.1007/978-981-33-4370-2_1
16. Kaushik, M., Sharma, R., Peious, S.A., Shahin, M., Yahia, S.B., Draheim, D.: A systematic assessment of numerical association rule mining methods. SN Comput. Sci. **2**(5), 1–13 (2021). https://doi.org/10.1007/s42979-021-00725-2
17. Kievit, R., Frankenhuis, W., Waldorp, L., Borsboom, D.: Simpson's paradox in psychological science: a practical guide. Front. Psychol. **4**, 513 (2013). https://doi.org/10.3389/fpsyg.2013.00513
18. Kim, Y.: The 9 pitfalls of data science. Am. Stat. **74**(3), 307 (2020). https://doi.org/10.1080/00031305.2020.1790216
19. King, G., Roberts, M.: EI: A(n R) program for ecological inference. Harvard University (2012)
20. Ma, H.Y., Lin, D.K.J.: Effect of Simpson's paradox on market basket analysis. J. Chin. Stat. Assoc. **42**(2), 209–221 (2004). https://doi.org/10.29973/JCSA.200406.0007
21. MacKinnon, D.P., Fairchild, A.J., Fritz, M.S.: Mediation analysis. Ann. Rev. Psychol. **58**(1), 593–614 (2007). https://doi.org/10.1146/annurev.psych.58.110405.085542. pMID: 16968208
22. Pearl, J.: Causal inference without counterfactuals: comment. J. Am. Stat. Assoc. **95**(450), 428–431 (2000)
23. Pearl, J.: Understanding Simpson's paradox. SSRN Electron. J. **68** (2013). https://doi.org/10.2139/ssrn.2343788
24. Pearson Karl, L.A., Leslie, B.M.: Genetic (reproductive) selection: inheritance of fertility in man, and of fecundity in thoroughbred racehorses. Philos. Trans. R. Soc. Lond. Ser. A **192**, 257–330 (1899)
25. Quinlan, J.: Combining instance-based and model-based learning. In: Machine Learning Proceedings 1993, pp. 236–243. Elsevier (1993). https://doi.org/10.1016/B978-1-55860-307-3.50037-X
26. Robinson, W.S.: Ecological correlations and the behavior of individuals. Am. Sociol. Rev. **15**(3), 351–357 (1950)
27. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)
28. Sharma, R., Peious, S.A.: Towards unification of decision support technologies: statistical reasoning. OLAP and Association Rule Mining. https://github.com/rahulgla/unification
29. Simpson, E.H.: The interpretation of interaction in contingency tables. J. Roy. Stat. Soc.: Ser. B (Methodol.) **13**(2), 238–241 (1951)
30. Sprenger, J., Weinberger, N.: Simpson's paradox. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy, Summer 2021 edn. Metaphysics Research Lab, Stanford University (2021)
31. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. In: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, pp. 1–12 (1996)
32. Tu, Y.K., Gunnell, D., Gilthorpe, M.S.: Simpson's Paradox, Lord's Paradox, and Suppression Effects are the same phenomenon-the reversal paradox. Emerg. Themes Epidemiol. **5**(1), 1–9 (2008)
33. Von Kugelgen, J., Gresele, L., Scholkopf, B.: Simpson's paradox in COVID-19 case fatality rates: a mediation analysis of age-related causal effects. IEEE Trans. Artif. Intell. **2**(1), 18–27 (2021). https://doi.org/10.1109/tai.2021.3073088

34. Xu, C., Brown, S.M., Grant, C.: Detecting Simpson's paradox. In: The Thirty-First International Flairs Conference (2018)
35. Yule, G.U.: Notes on the theory of association of attributes in statistics. Biometrika **2**(2), 121–134 (1903)

Appendix 5

[V]

R. Sharma, M. Kaushik, S. A. Peious, M. Bertl, A. Vidyarthi, A. Kumar, and D. Draheim. Detecting Simpson's paradox: A step towards fairness in machine learning. In S. Chiusano, T. Cerquitelli, R. Wrembel, K. Nørvåg, B. Catania, G. Vargas-Solar, and E. Zumpano, editors, *Proceedings of ADBIS 2022 – the 26th International Conference on New Trends in Database and Information Systems*, pages 67–76, Cham, 2022. Springer International Publishing

Detecting Simpson's Paradox: A Step Towards Fairness in Machine Learning

Rahul Sharma[1]( ), Minakshi Kaushik[1], Sijo Arakkal Peious[1],
Markus Bertl[2], Ankit Vidyarthi[3], Ashwani Kumar[4],
and Dirk Draheim[1]

[1] Information Systems Group, Tallinn University of Technology,
Akadeemia tee 15a, 12618 Tallinn, Estonia
{rahul.sharma,minakshi.kaushik,sijo.arakkal,dirk.draheim}@taltech.ee
[2] Department of Health Technologies, Tallinn University of Technology,
Akadeemia tee 15a, 12618 Tallinn, Estonia
mbertl@taltech.ee
[3] Jaypee Institute of Information Technology, Noida, India
[4] Sreyas Institute of Engineering and Technology, Hyderabad, India

Abstract. In the last two decades, artificial intelligence (AI) and machine learning (ML) have grown tremendously. However, understanding and assessing the impacts of causality and statistical paradoxes are still some of the critical challenges in their domains. Currently, these terms are widely discussed within the context of explainable AI (XAI) and algorithmic fairness. However, they are still not in the mainstream AI and ML application development scenarios. In this paper, first, we discuss the impact of Simpson's paradox on linear trends, i.e., on continuous values, and then we demonstrate its effects via three benchmark training datasets used in ML. Next, we provide an algorithm for detecting Simpson's paradox. The algorithm has experimented with the three datasets and appears beneficial in detecting the cases of Simpson's paradox in continuous values. In future, the algorithm can be utilized in designing a certain next-generation platform for fairness in ML.

Keywords: Big data · Artificial intelligence · Machine learning · Data science · Simpson's paradox · Explainable AI

1 Introduction

The outcomes of artificial intelligence (AI) and machine learning (ML) applications are explicitly dependent on the correctness of algorithms and training datasets. However, like statistics and mathematics, handling statistical paradoxes, cause and effect together in datasets is still not in the mainstream AI and ML application development. There are many cases where the outcome of

This work has been partially conducted in the project "ICT programme" which was supported by the European Union through the European Social Fund.

S. Chiusano et al. (Eds.): ADBIS 2022, CCIS 1652, pp. 67–76, 2022.
https://doi.org/10.1007/978-3-031-15743-1_7

AI applications is observed to be biased [13,18]. Like fossil fuels, data is considered a new fuel in the 21st century, but it needs to be properly cleaned for fair results. Nowadays, increased usages of AI and ML in healthcare, social media, digital advertising, search engines, etc., directly or indirectly impact human life and their decisions. Therefore, understanding causal relationships and evaluating the existence of statistical paradoxes should be an essential part of AI application development scenarios for better, fair and unbiased AI applications.

Statistical paradoxes, causality, selection bias, confounding and information bias have been debated in statistics and mathematics for a long time; expert statisticians and mathematicians have effectively handled their severe consequences. Several statistical paradoxes include Simpson's Paradox, Tea Leaf Paradox, Berkson's Paradox, Latent Variables, Law of Unintended Consequences, etc. The term "paradox" denotes a fundamental link between several statistical issues and mathematical reasoning, e.g., causal inference [19,20], Lord's paradox [28], propensity score matching [24], suppressor variables [4], the ecological fallacy [16,23], conditional independence [5], p-technique [3] and partial correlations [9], mediator variables [17], etc.

Handling statistical paradoxes and causality will not only build trust in artificial applications but also serve as the foundation for fairness in AI. In this paper, we explicitly focus on Simpson's paradox, which has also been discussed in various data mining techniques [10], e.g., association rule mining [1,6,7] and numerical association rule mining [14,15,27]. The main aim of this article is to develop an algorithm for detecting Simpson's paradox in continuous values. The algorithm is tested with the three benchmark datasets and appears beneficial in detecting the cases of Simpson's paradox in linear trends. In the future, the algorithm may be used to create a specific next-generation platform for trustworthy AI and fairness in ML.

The paper is organized as follows. In Sect. 2, we discuss the background of Simpson's Paradox. In Sect. 3, we discuss ways to detect Simpson's paradox and propose an algorithm for detecting the paradox in linear trends. In Sect. 4, three benchmark datasets are used to experiment with the algorithm. Finally, a discussion on future work and conclusion is given in Sect. 5 and Sect. 6, respectively.

2 Yule-Simpson's Paradox

In the year 1899, Karl Pearson et al. [21] demonstrated a statistical paradox in marginal and partial associations between continuous variables. Further, in 1903, Udny Yule [29] presented "the theory of association of attributes in statistics" and revealed the existence of an association paradox with categorical variables. Later in 1951, Edward H. Simpson [26] presented the concept of reversing results and in 1972, Colin R. Blyth coined the term "Simpsons Paradox" [2]. Therefore, this paradox is known by different names and is well-known as the Yule-Simpson effect, amalgamation paradox, or reversal paradox [22].

We have used a real-world dataset from Simpson's article [26] to discuss the paradox. In this example, analysis for medical treatment is described. Table 1 provides the number that shows the effect of the medical treatment for the entire population ($N = 52$) as well as for men and women separately in subgroups. The treatment is suitable for both male and female subgroups; however, the treatment appears unsuitable for the entire population.

Table 1. A real life case of Simpson's Paradox: The numbers in the table are taken from Simpson's original article [26].

	Full population N = 52			Women (F) = 20			Men (M), N = 32		
	Success (S)	Failure ($\neg S$)	Succ.%	Success (S)	Failure ($\neg S$)	Succ.%	Success (S)	Failure ($\neg S$)	Succ.%
T	20	20	50%	8	5	61%	12	15	44%
$\neg T$	6	6	50%	4	3	57%	2	3	40%

The Simpson's paradox scenario can also be described via probability theory and conditional probabilities. Let $T = treatment$, $S = successful$, $M = Male$, and $F = Female$ then the $\mathsf{P}(S \mid T)$ can be described as:

$$\mathsf{P}(S \mid T) = \mathsf{P}(S \mid \neg T) \tag{1}$$

$$\mathsf{P}(S \mid T, M) > \mathsf{P}(S \mid \neg T, M) \tag{2}$$

$$\mathsf{P}(S \mid T, F) > \mathsf{P}(S \mid \neg T, F) \tag{3}$$

This reversal of results between the male, female population and the entire population has been referred to as Simpson's Paradox. In statistics, these concepts have been discussed widely and named differently by several authors [21,29].

3 Detecting Simpson's Paradox

The Simpson's paradox instances are investigated for both categorical and continuous values. However, we investigate the paradox in linear trends. The Pearson correlation index is used in the algorithm to determine the correlations between two variables and further define the function of the confounding variable. A confounder can be defined as a factor that affects both the dependent and independent variables, resulting in an incorrect association. The Pearson correlation index allows us to measure the strength of the linear association between two variables. In Eq. 4, Pearson correlation coefficient is represented by r and x, y are input vectors, $\overline{x}$ and $\overline{y}$ are the means of the variables, respectively. The output value always lies between -1 and 1. Values greater than 0 imply a positive correlation, while the values 1 and 0 indicate the exact positive association and no correlation, respectively. Values less than 0 suggest a negative association, and −1 indicates a clear negative association.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}} \tag{4}$$

3.1 Algorithm

To identify an instance of the Simpson's paradox in a continuous dataset with n continuous variable and m discrete variables, we compute a correlation matrix $(n \times n)$ for all the data. Then for m discrete variable with k_m levels, an additional $(n \times n)$ matrix needs to be calculated for each level of variables. Therefore, we need to calculate the $1 + \sum_i^m = k_i$ correlation matrices of size $(n \times n)$ and compare it with the lower half of $\sum_i^m = k_i$ for subgroup levels.

The algorithm's initial step is to determine the correlation between x and y variables with the values of the corresponding columns in the dataset. In this way, we learn the direction of the relationship between the variables. We next walk through the list of remaining variables, compute the Pearson index conditional on each subgroup (category), count the percentage of subgroups where the correlation index is reversed with respect to the correlation index in the aggregate data, and store the value key pairs in an array. We further get the array element where the value (ratio) is the highest. A value greater than 0 implies the existence of Simpson's paradox and the maximal value of 1 indicates a full reversal effect.

4 Experiments

Python programming language is used to implement the algorithm on a personal computer with the Windows 10 × 64 operating system and an Intel(R) Core(TM) i5-8265U CPU running at 1.60 GHz, 1800 MHz, 4 Cores, and 8 Logical Processors. We evaluate the algorithm with three benchmark datasets that are widely used to train various ML models. The performance of the algorithm strongly correlates with the size of the datasets. The programming code, datasets, and other necessary instructions for the algorithm are available in the GitHub repository [25].

4.1 Datasets

Iris dataset, Miles per gallon (MPG) dataset and Penguin dataset are used to demonstrate the presence of Simpson's paradox in continuous data.

Algorithm 1: Identification of the confounding variable in continuous values to identify Simpson's paradox

```
Input: dataset $D$, variable $x$, variable $y$
Output: a pair consisting of confounding variable and ratio of reversed
        association signs
aggreg_index = Pearson(d[x], d[y]) // calculate corr. index between columns
indexes = [] // initialize index array to store key value pairs where the key is
  column and value is the number of reversed subgroups
cols = columns(D) // initialize array of all columns of D
foreach column $\in$ cols do
    if Column Is Not Categorical(column) then
        Continue
    end
    else
        subgroups = Categories(column) // get the categories of a column
          coefficients = [] // initialize empty array to store the correlation
          indexes foreach subgroup $\in$ subgroups do
            disaggreg_index = Pearson( D[x]: where D[column] = subgroup,
              D[y]: where D[column] = subgroup) calculate corr. index between
              columns for current subgroup
            Add index of disaggregated to correlation indexes array
        end
    end
    reversed_subgroups = RatioReversedSubgroups(aggreg_index, coefficients)
      // calculate ratio of the correlation indexes reversed with respect to the
      correlation index for the aggregated data
    Add {column, reversed_subgroups} values into indexes
end
Store the max values of indexes pairs into result
Return result
```

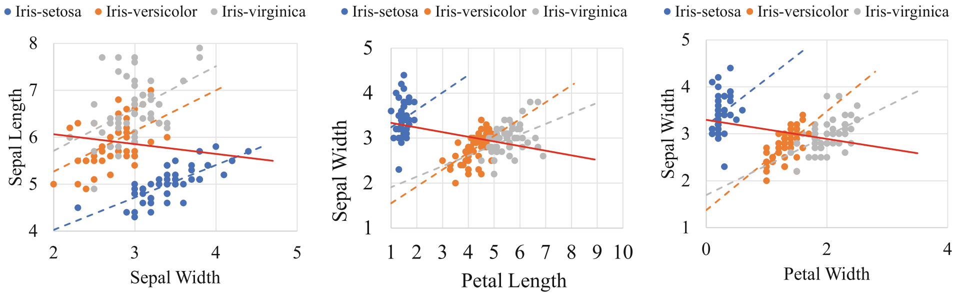


Fig. 1. Scatter plot with trend lines for Iris dataset.

Iris Dataset: In 1936 Ronald Fisher introduced the iris dataset in one of his research papers [8]. In this dataset, there are 50 data samples for the three different iris species, i.e., 'Setosa', 'Versicolor', and 'Virginicare'. In the dataset, species names are categorical, while length and breadth are continuous values.

We visualize the possible associations between the length and breadth of each pair of candidate attributes to identify the instances of Simpson's paradox. Table 2 demonstrates the Pearson correlation index returned by the algorithm 1 between two continuous variables ('sepal length' and 'sepal width').

We identify the existence of Simpson's paradox for three pairs of measurements. 1. sepal length and width, 2. sepal width and petal length, and 3. sepal width and petal width. In Fig. 1, the correlation between sepal width and sepal length is positive (dashed line) for each species. However, the correlation between sepal width and sepal length for the entire population is negative (solid red trend line). Similarly, the pair of petal length and width and the pair of petal width and sepal width have positive trends for each species. However, the overall trend for the length and width for the entire population is negative in both cases.

Table 2. The Pearson correlation index for Iris dataset by the Algorithm 1.

Agg. correlation	Variable 1	Variable 2	Sub group	Group correlation
–0.1093	Sepal length	Sepal width	Iris-setosa	0.7467
–0.1093	Sepal length	Sepal width	Iris-versicolor	0.5259
–0.1093	Sepal length	Sepal width	Iris-virginica	0.4572

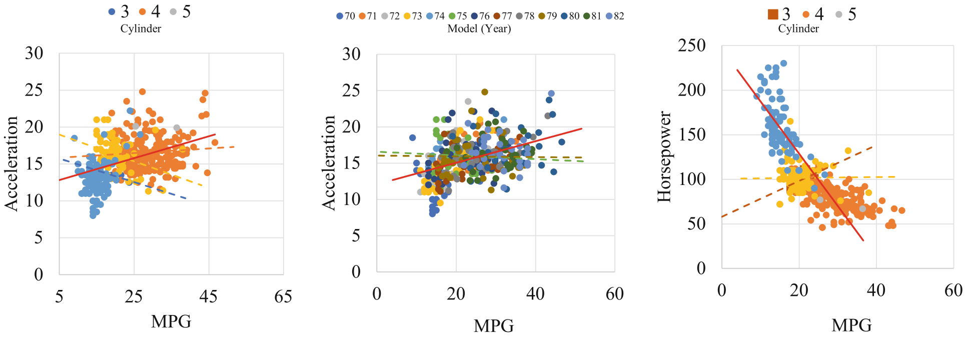


Fig. 2. Scatter plot with trend lines for MPG dataset.

The MPG Dataset: Ross Quinlan used the Auto MPG dataset in 1993 [22]. The dataset contains 398 automobile records from 1970 to 1982, including the vehicle's name, MPG, number of cylinders, horsepower, and weight. The dataset includes three multi-valued discrete attributes and five continuous attributes. We visualize the relationship between MPG, acceleration and horsepower for two categorical attributes (number of cylinders and model year). The goal of analyzing the dataset is to know the factors that influence each car's overall fuel consumption. The dataset consists of fuel consumption in mpg, horsepower, number of cylinders, displacement, weight, and acceleration.

Table 3. The Pearson correlation index for MPG dataset by the proposed algorithm.

Agg. correlation	Variable 1	Variable 2	Sub group	Group correlation
0.4230	mpg	Acceleration	Cylinders	–0.8190
0.4230	mpg	Acceleration	Cylinders	–0.3410
0.4230	mpg	Acceleration	Model year	–0.0510

In the MPG dataset, the existence of Simpson's paradox is discovered in three pairs of measurements. 1. MPG with acceleration according to the engine cylinders, 2. MPG with acceleration with respect to their model year, and 3. MPG with horsepower according to the engine cylinders. In Fig. 2, it is visualized that there is a negative correlation between MPG and acceleration for three-cylinder engines and six-cylinder engines; however, the overall trend between MPG and acceleration is positive (solid red line). Similarly, the overall trend is the opposite for MPG with acceleration with respect to the model year and MPG with horsepower according to the engine cylinders. Table 3 demonstrates the Pearson correlation index returned by the Algorithm 1 between two continuous variables ('mpg' and 'acceleration').

Penguin Dataset. Palmer penguins dataset [11,12] is also a well-known dataset used as an alternative to the Iris dataset. The dataset contains the descriptions of three species of penguins (Adelie, Chinstrap, Gentoo) in the islands of Palmer, Antarctica. The dataset contains 344 data rows with columns: 'species', 'island' , 'culmen_length_mm', 'culmen_depth_mm', 'flipper_length_mm' , 'body_mass_g' and 'sex'. To investigate the instances of Simpson's paradox in the dataset, we set x as 'culmen_length_mm' and y as 'culmen_depth_mm'. As per the results from the algorithm, there is an instance of Simpson's paradox in data as the association between the culmen_length and culmen_depth reverses when data is disaggregated by the species. Figure 3 demonstrates a positive correlation between the culmen_length_mm and culmen_depth_mm of each species. However, it is negative for the aggregate data.

5 Discussion and Future Work

The presence of Simpson's paradoxes in benchmark datasets provides a direction to understand the causality in decision making. We noticed that most ML and deep learning algorithms focus only on identifying correlations rather than identifying the real or causal relationships between data items. Therefore, understanding and evaluating causality is an important term to be discussed in big data, data science, AI and ML. In future, we plan to develop a framework to simplify the impacts of Simpson's paradox and address various other statistical paradoxes (e.g., Berkson's paradox) that have severe implications for big data, data science, AI and ML.

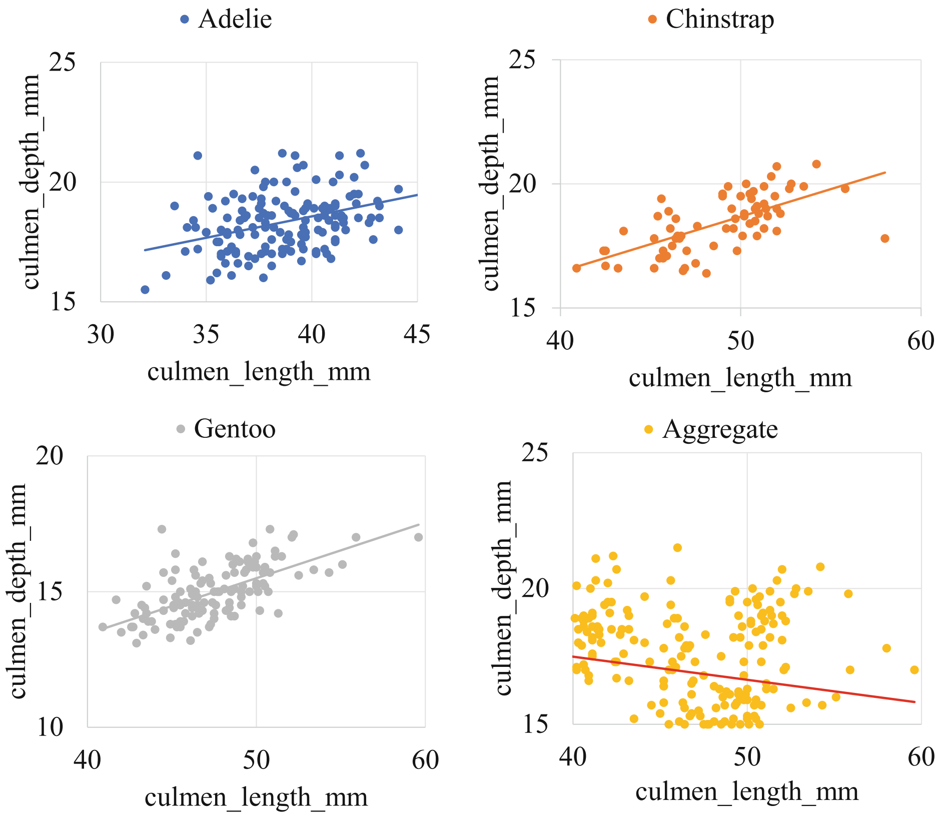


Fig. 3. Scatter plot with trend lines for Penguin dataset.

6 Conclusion

Handling statistical paradoxes is a complex challenge in automatic data mining, specifically in AI and ML techniques. In this paper, we discussed a strong need for statistical evaluations of datasets and demonstrated the impacts of Simpson's paradox on AI and ML via some benchmark training datasets. We argue that if confounding effects are not properly addressed in automatic data mining, the outcomes of data analysis can be completely opposite. However, with the right tools and data analysis, a good analyst or data scientist can handle it in a better way. Further, we provided an algorithm to detect Simpson's paradox in linear trends (continuous values). The algorithm is evaluated on three benchmark datasets and performed well in each experiment. This algorithm can be a part of developing a platform to detect Simpson's paradox in different data (continuous, categorical) and enable data scientists to explore the impacts of confounding variables.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of VLDB 1994 - the 20th International Conference on Very Large Data Bases, pp. 487–499. Morgan Kaufmann (1994)
2. Blyth, C.R.: On Simpson's paradox and the sure-thing principle. J. Am. Stat. Assoc. **67**(338), 364–366 (1972)
3. Cattell, R.B.: P-technique factorization and the determination of individual dynamic structure. J. Clin. Psychol. **8**, 5–10 (1952)
4. Conger, A.J.: A revised definition for suppressor variables: a guide to their identification and interpretation. Educ. Psychol. Meas. **34**(1), 35–46 (1974)
5. Dawid, A.P.: Conditional independence in statistical theory. J. Roy. Stat. Soc. Ser. B (Methodol.) **41**(1), 1–15 (1979). https://doi.org/10.1111/j.2517-6161.1979.tb01052.x
6. Draheim, D.: DEXA 2019 keynote presentation: future perspectives of association rule mining based on partial conditionalization, Linz, Austria, 28th August 2019. https://doi.org/10.13140/RG.2.2.17763.48163
7. Draheim, D.: Future perspectives of association rule mining based on partial conditionalization. In: Hartmann, S., Küng, J., Chakravarthy, S., Anderst-Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) Proceedings of DEXA'2019 - the 30th International Conference on Database and Expert Systems Applications. LNCS, vol. 11706, p. xvi. Springer, Heidelberg (2019)
8. Fisher, R.A.: The use of multiple measurement in taxonomic problems. Ann. Eugenics **7**(2), 179–188 (1936). https://doi.org/10.1111/j.1469-1809.1936.tb02137.x
9. Fisher, R.A.: Iii. the influence of rainfall on the yield of wheat at rothamsted. Phil. Trans. Roy. Soc. Lond. Ser. B Containing Papers Biol. Charact. **213**(402–410), 89–142 (1925)
10. Freitas, A.A., McGarry, K.J., Correa, E.S.: Integrating bayesian networks and simpson's paradox in data mining. In: Texts in Philosophy. College Publications (2007)
11. Gorman, K.B., Williams, T.D., Fraser, W.R.: Ecological sexual dimorphism and environmental variability within a community of antarctic penguins (genus pygoscelis). PLOS ONE **9**(3), 1–14 (2014). https://doi.org/10.1371/journal.pone.0090081
12. Horst, A.M., Hill, A.P., Gorman, K.B.: palmerpenguins: Palmer Archipelago (Antarctica) penguin data (2020). https://doi.org/10.5281/zenodo.3960218, https://allisonhorst.github.io/palmerpenguins/, r package version 0.1.0
13. Julia, A., Jeff, L., Surya, M., Lauren, K.: Machine Bias, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=TiqCeZIj4uLbXl91e3wM2PnmnWbCVOvS
14. Kaushik, M., Sharma, R., Peious, S.A., Shahin, M., Ben Yahia, S., Draheim, D.: On the potential of numerical association rule mining. In: Dang, T.K., Küng, J., Takizawa, M., Chung, T.M. (eds.) FDSE 2020. CCIS, vol. 1306, pp. 3–20. Springer, Singapore (2020). https://doi.org/10.1007/978-981-33-4370-2_1
15. Kaushik, M., Sharma, R., Peious, S.A., Shahin, M., Yahia, S.B., Draheim, D.: A systematic assessment of numerical association rule mining methods. SN Comput. Sci. **2**(5), 1–13 (2021). https://doi.org/10.1007/s42979-021-00725-2
16. King, G., Roberts, M.: Ei: a (n r) program for ecological inference. Harvard University (2012)

17. MacKinnon, D.P., Fairchild, A.J., Fritz, M.S.: Mediation analysis. Ann. Rev. Psychol. **58**(1), 593–614 (2007). https://doi.org/10.1146/annurev.psych.58.110405.085542
18. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, New York (2016)
19. Pearl, J.: Causal inference without counterfactuals: comment. J. Am. Stat. Assoc. **95**(450), 428–431 (2000)
20. Pearl, J.: Understanding Simpson's paradox. SSRN Electron. J. **68** (2013). https://doi.org/10.2139/ssrn.2343788
21. Pearson Karl, L.A., Leslie, B.M.: Genetic (reproductive) selection: inheritance of fertility in man, and of fecundity in thoroughbred racehorses. Phil. Trans. Roy. Soc. Lond. Ser. A **192**, 257–330 (1899)
22. Quinlan, J.: Combining instance-based and model-based learning. In: Machine Learning Proceedings 1993, pp. 236–243. Elsevier (1993). https://doi.org/10.1016/B978-1-55860-307-3.50037-X
23. Robinson, W.S.: Ecological correlations and the behavior of individuals. Am. Sociol. Rev. **15**(3), 351–357 (1950)
24. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)
25. Sharma, R., Peious, S.A.: Towards unification of decision support technologies: Statistical reasoning, OLAP and association rule mining. https://github.com/rahulgla/unification
26. Simpson, E.H.: The interpretation of interaction in contingency tables. J. Roy. Stat. Soc. Ser. B (Methodol.) **13**(2), 238–241 (1951)
27. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. In: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, pp. 1–12 (1996)
28. Tu, Y.K., Gunnell, D., Gilthorpe, M.S.: Simpson's paradox, lord's paradox, and suppression effects are the same phenomenon-the reversal paradox. Emerg. Themes Epidemiol. **5**(1), 1–9 (2008)
29. Yule, G.U.: Notes on the theory of association of attributes in statistics. Biometrika **2**(2), 121–134 (1903)

Appendix 6

[VI]

R. Sharma, M. Kaushik, S. A. Peious, M. Shahin, A. Vidyarthi, P. Tiwari, and D. Draheim. Why not to trust big data: Discussing statistical paradoxes. In U. K. Rage, V. Goyal, and P. K. Reddy, editors, *Proceedings of DASFAA 2022 International Workshops – the 27th International Conference on Database Systems for Advanced Applications*, pages 50–63, Cham, 2022. Springer International Publishing

Why Not to Trust Big Data: Discussing Statistical Paradoxes

Rahul Sharma[1]( ), Minakshi Kaushik[1], Sijo Arakkal Peious[1], Mahtab Shahin[1], Ankit Vidyarthi[2], Prayag Tiwari[3], and Dirk Draheim[1]

[1] Information Systems Group, Tallinn University of Technology,
Akadeemia tee 15a, 12618 Tallinn, Estonia
{rahul.sharma,minakshi.kaushik,sijo.arakkal,mahtab.shahin,
dirk.draheim}@taltech.ee

[2] Jaypee Institute of Information Technology, Noida, India

[3] Department of Computer Science, Aalto University, Espoo, Finland
prayag.tiwari@aalto.fi

Abstract. Big data is driving the growth of businesses, data is the money, big data is the fuel of the twenty-first century, and there are many other claims over Big Data. Can we, however, rely on big data blindly? What happens if the training data set of a machine learning module is incorrect and contains a statistical paradox? Data, like fossil fuels, is valuable, but it must be refined carefully for the best results. Statistical paradoxes are difficult to observe in datasets, but they are significant to analyse in every small or big dataset. In this paper, we discuss the role of statistical paradoxes on Big data. Mainly we discuss the impact of Berkson's paradox and Simpson's paradox on different types of data and demonstrate how they affect big data. We provide that statistical paradoxes are more common in a variety of data and they lead to wrong conclusions potentially with harmful consequences. Experiments on two real-world datasets and a case study indicate that statistical paradoxes are severely harmful to big data and automatic data analysis techniques.

Keywords: Big data · Artificial intelligence · Machine learning · Data science · Simpson's paradox · Explainable AI

1 Introduction

Data has always been critical in making decisions. Earlier, statistics and mathematics have been used to draw insights from data. However, in the last two decades, with the emergence of social media and big data technologies, data science, artificial intelligence (AI) and machine learning (ML) techniques have gained massive ground in practice and theory. These decision support techniques

This work has been partially conducted in the project "ICT programme" which was supported by the European Union through the European Social Fund.

U. K. Rage et al. (Eds.): DASFAA 2022 Workshops, LNCS 13248, pp. 50–63, 2022.
https://doi.org/10.1007/978-3-031-11217-1_4

are being used widely to develop intelligent applications and acquire deeper insights from structured and unstructured data. In most AI use cases, ML based trained artificial systems provide fast and accurate outcomes; however, it does not guarantee accurate results for every use case in real life. Moreover, like statistics, understanding causal relationships and evaluating the existence of statistical paradoxes in the training dataset is not in the mainstream data science, AI and ML application scenarios. AI, machine learning, and big data are now widely used in medical sciences, social sciences, and politics, and they have a direct or indirect impact on human life and decisions. Therefore, understanding causal relationships and evaluating the existence of statistical paradoxes is essential for fair decision making [13,14].

Statistical reasoning and probability theory are the foundation of many AI, big data and data science techniques, e.g., random forest [7], support vector machines [11], etc. Therefore, it is usual to have causal relationships and statistical paradoxes in these decision support techniques. A paradox can be a statement that leads to an apparent self-contradictory conclusion. Even the most well-known and documented paradoxes frequently confound domain specialists because they fundamentally violate common sense.

There are many statistical paradoxes (e.g., Simpson's Paradox, Berkson's Paradox, Latent Variables, Law of Unintended Consequences, Tea Leaf Paradox, etc.). Statistical paradoxes are not new to be discussed in statistics and mathematics; expert mathematicians and statisticians adequately addressed the severe impact of paradoxes. However, in modern decision support techniques, specifically AI and data science, causal relationships, data fallacies and statistical paradoxes are not appropriately addressed. In this article, we discuss the impact of Berkson's Paradox, Yule-Simpson paradox and causal inference on big data. We highlight several hidden problems in data that are not yet discussed in big data mining. We use two benchmark datasets for machine learning and a case study to demonstrate the existence of Simpson's paradox in different types of data.

The paper is organised as follows. In Sect. 2, we discuss why not trust data science, AI, ML and big data. In Sec. 3, we discuss two statistical paradoxes and discuss their impacts on big data mining. In Sect. 4, we use two benchmark datasets for machine learning to demonstrate the effects of Simpson's paradox. In Sect. 5 we provide a case study to analyse the impact of Simpson's paradox in real life. Finally, a discussion and conclusion is provided in Sect. 6 and Sect. 7.

2 Why Not to Trust on Data Science, AI, ML and Big Data

In AI, ML and Data Science, observing trends, mean and correlation between two variables for making decisions is not always correct. E.g., suppose in a city, the Covid-19 infection rate of smokers is less than the infection rate of the non-smokers. Can we claim that smoking prevent Covid-19? It is a perfect case of poor data science where all the variables and features in the dataset are not appropriately observed. In today's world, data literacy may not seem exciting when compared to machine learning algorithms or big data mining, but it should be the foundation for all data mining processes.

Datasets, irrespective of their size and type, are not self-explanatory. It's all numbers and statistics responsible for creating stories out of datasets. Therefore, it's essential to validate a dataset statistically and evaluate the existence of any statistical paradoxes. AI, ML, big data and data science based techniques generate knowledge from data. Therefore, decision support techniques are easily prone to statistical paradoxes and can not be trusted.

3 Statistical Paradoxes

Statistical paradoxes aren't something that hasn't been discussed before. These terms are widely used in statistics and have been around for over a century. Statistical paradoxes are fundamentally related with various statistical challenges and mathematical logic including causal inference [27,28], the ecological fallacy [24,31], Lord's paradox [36], propensity score matching [32], suppressor variables [10], conditional independence [12], partial correlations [16], p-technique [8] and mediator variables [26]. The instances of statistical paradoxes specifically Simpson's paradox have been discussed in various data mining techniques [17], e.g., association rule mining [2] and numerical association rule mining [20,21,34]

More recently, Kügelgen et al. [37] pointed out the importance of statistical analysis of real data and demonstrated instances of Simpson's paradox in Covid-19 data analysis. They provide that the overall case fatality rate (CFR) was higher in Italy than in China. However, in every age group, the fatality rate was higher in China than in Italy. These observations raise many questions on the accuracy of data and its analysis. Heather et al. [25] have addressed the existence of Simpson's paradox. In psychological science, Kievit et al. [22] examined the instances of Simpson's paradox. Alipourfard et al. [3] have discovered the existence of Simpson's paradox in social data and behavioural data [4]. Therefore, understanding data, especially big data, is more critical than its processing. In the following two sections, we discuss Berkson's paradox and Yule-Simpson's Paradox to demonstrate their vast impact on big datasets.

3.1 Berkson Paradox

Berkson's paradox can make it appear as if there is a relationship between two independent variables when there is no relationship between the variables. In 1946, despite diabetes being a risk factor for cholecystitis, Berkson [5] observed a negative correlation between cholecystitis and diabetes in hospital patients. Berkson state that If at least one of two independent events occurs, they become conditionally dependent. In other words, two independent events become conditionally dependent, given that at least one of them occurs. Statistically, Berkson's paradox and Simpson's paradox are very close to each other. Berkson's paradox is a type of selection bias caused by systematically observing some events more than others.

$$if\ 0 < P(A) < 1,\ 0 < P(B) < 1\ and \tag{1}$$

$$P(A|B) = P(A)\ then \tag{2}$$

$$P(A|B, A \cup B) = P(A)\ Hence \tag{3}$$

$$P(A|B, A \cup B) > P(A) \tag{4}$$

As given in Eq. 1 to Eq. 4, $P(A|B)$, a conditional probability, is the probability of observing event A given that B is true. The probability of A given both B and $(A\ or\ B)$ is smaller than the probability of A given $(A\ or\ B)$.

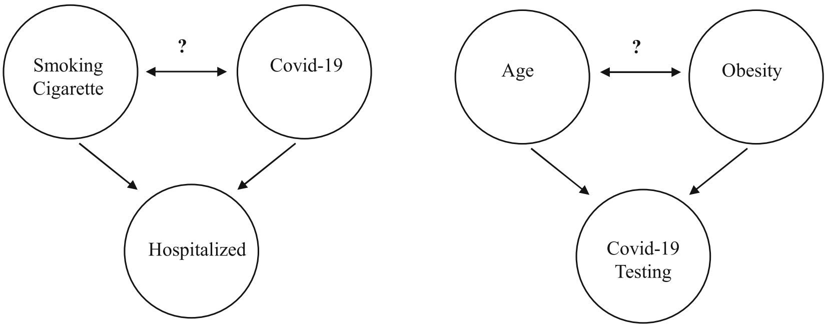


Fig. 1. Berkson's paradox: two noticeable example of Covid-19 which introduce a collider.

As we all know, smoking cigarettes is a well-known risk factor for respiratory diseases. However, recently Wenzel T. [9] observed a negative co-relation between Covid-19 severity and smoking cigarettes. In another observation, Griffith et al. [18] describe it as a Collider Bias or Berkson's paradox. In Fig. 1, we demonstrate an example of collider. Here Smoking cigarettes, Covid-19 are two independent variables, but they collide with another random variable, hospitalised. Here, the variable hospitalised is collider for both smoking cigarettes and Covid-19.

3.2 Yule-Simpson's Paradox

In the year 1899, Karl Pearson et al. [29] demonstrated a statistical paradox in marginal and partial associations between continuous variables. Later in 1903, Udny Yule [38] explained "the theory of association of attributes in statistics" and revealed the existence of an association paradox with categorical variables.

In a technical paper published in 1951 [33], Edward H. Simpson described the phenomenon of reversing results. However, in 1972, Colin R. Blyth coined the term “Simpsons Paradox” [6]. Therefore, this paradox is known with different names and it is popular as the Yule-Simpson effect, amalgamation paradox, or reversal paradox [30].

We start the discussion on the paradox by using the real-world dataset from Simpson’s article [33]. In this example, analysis for medical treatment is demonstrated. Table 1 summarises the effect of the medical treatment for the entire population ($N = 52$) as well as for men and women separately in subgroups. The treatment appears effective for both male and female subgroups; however, the treatment seems ineffective at the whole population level.

Table 1. 2 × 2 contingency table with sub population groups D1 and D2.

	Population $D = D_1 + D_2$		Sub-population D_1		Sub-population D_2	
	Success (S)	Failure ($\neg S$)	Success (S)	Failure ($\neg S$)	Success (S)	Failure ($\neg S$)
Treatment (T)	$a_1 + a_2$	$b_1 + b_2$	a_1	b_1	a_2	b_2
No Treatment ($\neg T$)	$c_1 + c_2$	$d_1 + d_2$	c_1	d_1	c_2	d_2

Definition 1. *Consider D groups of data such that group D_1 has A_i trials and $0 \leq a_i \leq A_i$ “successes”. Similarly, consider an analogous D groups of data such that group D_2 has B_i trials and $0 \leq b_i \leq B_i$ “successes” Then, Simpson’s paradox occurs if*

$$\frac{a_1}{A_1} \geq \frac{b_1}{B_1} \mathit{and} \frac{a_2}{A_i} \geq \frac{b_2}{B_2} \; \mathit{for\ all}\ i = 1, 2, \ldots, n \; \mathit{but} \; \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} A_i} \leq \frac{\sum_{i=1}^{n} b_i}{\sum_{i=1}^{n} B_i} \tag{5}$$

we use the following example to show how this equation works.

$$\frac{10}{20} > \frac{30}{70} \mathit{and} \frac{10}{50} > \frac{10}{60} \text{ but } \frac{10+10}{20+50} < \frac{30+10}{70+60}, \tag{6}$$

We could also flip the inequalities and still have the paradox since A and B are chosen arbitrarily.

Classically the paradox is expressed via contingency tables. Let a 2 × 2 contingency table for treatment (T) and success (S) in the $i^t h$ sub-population is represented by a four-dimensional vector of real numbers $D = (a_1, b_1, a_2, b_2)$. Then

$$D = \sum\nolimits_{i=1}^{N} D_i = \left(\sum a_i, \sum b_i, \sum c_i, \sum d_i\right) \tag{7}$$

is the aggregate dataset over N sub populations. This can be read as given in Table 1.

We can also demonstrate the Simpson's paradox scenario via probability theory and conditional probabilities. Let $T = treatment$, $S = successful$, $M = Male$, and $F = Female$ then,

$$\mathsf{P}(S \mid T) = \mathsf{P}(S \mid \neg T) \tag{8}$$

$$\mathsf{P}(S \mid T, M) > \mathsf{P}(S \mid \neg T, M) \tag{9}$$

$$\mathsf{P}(S \mid T, \neg M) > \mathsf{P}(S \mid \neg T, \neg M) \tag{10}$$

Based on Eq. 8, 9 and 10, one should use the treatment or not? As per the success rate for the male and female population, the treatment is a success, but overall, the treatment is a failure. This reversal of results between groups population and the total population has been referred to as Simpson's Paradox. In statistics, this concept has been discussed widely and named differently by several authors [29,38].

4 Existence of Simpson's Paradox in Big Data

Simpson's paradox can exist in any dataset irrespective of its size and type [23]. The paradox demonstrates the importance of having human experts in the loop to examine and query Big Data results. In this section, we present datasets to analyse the presence and implications of Simpson's paradox on big data.

To identify an instance of the Simpson paradox in a continuous dataset with n continuous variable and m discrete variables, we can compute a correlation matrix $(n \times n)$ for all the data. Then for m discrete variable with k_m levels, an additional $(n \times n)$ matrix needs to be calculated for each level of variables as follows. Therefore, we need to calculate the $1 + \sum_i^m = k_i$ correlation matrices of size $(n \times n)$ and compare it with the lower half of $\sum_i^m = k_i$ for subgroup levels. We have also discussed the measures to find the impact of one numerical variable to another numerical variable [19].

4.1 Datasets

We use the iris dataset and miles per gallon (mpg) dataset, the two benchmark datasets for machine learning to demonstrate the presence of Simpson's paradox in data.

Iris Dataset: Ronald Fisher introduced the iris dataset in a research paper [15]. It consists three types of iris species (Setosa, Versicolor, Virginicare), each with 50 data samples. The species names are categorical attributes, length and width are continuous attributes.

In order to identify the existence of Simpson's paradox in the iris datasets, we first visualise the relationship between the length and width of each pair of candidate attributes. As shown in Fig. 2, in the iris dataset, we identify the

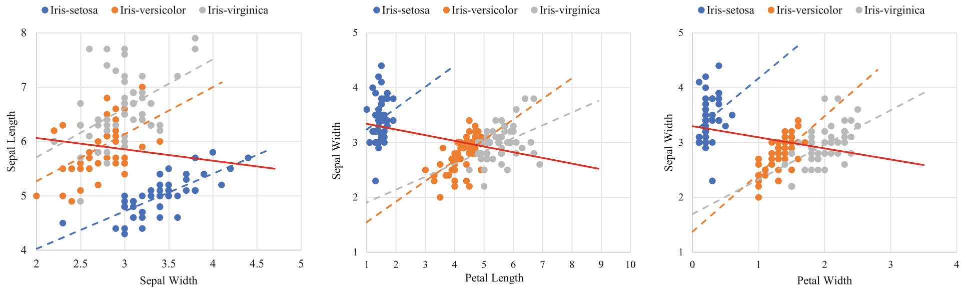


Fig. 2. Simpson's paradox in Iris dataset: there is a positive correlation between the three pairs of sepal length and petal width for the Iris-setosa, Iris-versicolor and Iris-virginicare (dashed lines). However, the overall trend for the length and width for the entire population is negative (solid red line) in all three combinations. (Color figure online)

existence of Simpson's paradox for three pairs of measurements. 1. sepal length and width, 2. sepal width and petal length, and 3. sepal width and petal width.

In Fig. 2, the correlation between sepal width and sepal length is positive (dashed line) for each species. However, the correlation between sepal width and sepal length for the entire population is negative (solid red trend line). Similarly, the pair of petal length, width, and the pair of petal width and sepal width have positive trends for each species; however, the overall trend for the length and width for the entire population is negative in both cases. Therefore, this is a clear case of Simpson's paradox in the iris dataset.

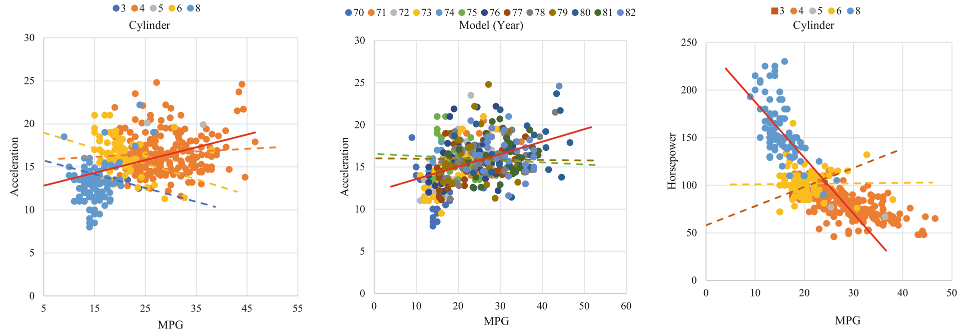


Fig. 3. Simpson's paradox in auto MPG dataset: there is a negative correlation between MPG and acceleration for three cylinders engines and six cylinders engines; however, the overall trend between MPG and acceleration is positive (solid red line). Similarly, the overall trend is positive for MPG and acceleration with respect to the model year. However, the overall trend between MPG and horsepower according to the engine cylinders is negative. (Color figure online)

The MPG Dataset: Ross Quinlan used the Auto MPG dataset in 1993 [30]. The dataset contains 398 automobile records from 1970 to 1982, including the vehicle's name, MPG, number of cylinders, horsepower, and weight. The dataset includes three multi-valued discrete attributes and five continuous attributes.

In order to identify the existence of Simpson's paradox in the MPG datasets, we visualise the relationship between MPG, acceleration and horsepower for two categorical attributes (number of cylinders and model year). The goal of analysing the dataset is to know the factors that influence each car's overall fuel consumption. The dataset consists of fuel consumption in mpg, horsepower, number of cylinders, displacement, weight, and acceleration.

In the MPG dataset, we identify the existence of Simpson's paradox in three pairs of measurements. 1. MPG with acceleration according to the engine cylinders, 2. MPG with acceleration with respect to their model year, and 3. MPG with horsepower according to the engine cylinders. In the Fig. 3, it is visualised that there is a negative correlation between MPG and acceleration for three cylinders engines and six cylinders engines; however, the overall trend between MPG and acceleration is positive (solid red line). Similarly, the overall trend is opposite for MPG with acceleration with respect to the model year and MPG with horsepower according to the engine cylinders.

5 Analysis Simpson's Paradox in Real Life: A Case Study

The case study is from the California Department of developmental services (CDDS), United States of America [35]. As per the annual reports published by the department, the average annual expenditures on Hispanic residents were approximately one-third (1/3) of the average expenditures on White non-Hispanic residents. According to the marginal analysis, it was a solid gender discrimination case. However, a conditional analysis of ethnicity and age found no evidence of ethnic discrimination. Furthermore, except for one age group, the trends were completely opposite. The average annual expenditures on White non-Hispanic residents were less than the expenditures on Hispanic residents. Therefore, it is a perfect case of Simpson's paradox in real life.

Table 2. Number of residents by ethnicity and percentage of expenditures.

Ethnicity	Sum of Expend. ($)	% of Expend.	# of Residents	% of Residents
American Indian	145753	0.81	4	0.4
Asian	2372616	13.13	129	12.9
Black	1232191	6.82	59	5.9
Hispanic	4160654	23.03	376	37.6
Multi Race	115875	0.64	26	2.6
Native Hawaiian	128347	0.71	3	0.3
Other	6633	0.04	2	0.2
White not Hispanic	9903717	54.82	401	40.1
Total	18065786	100%	1000	100%

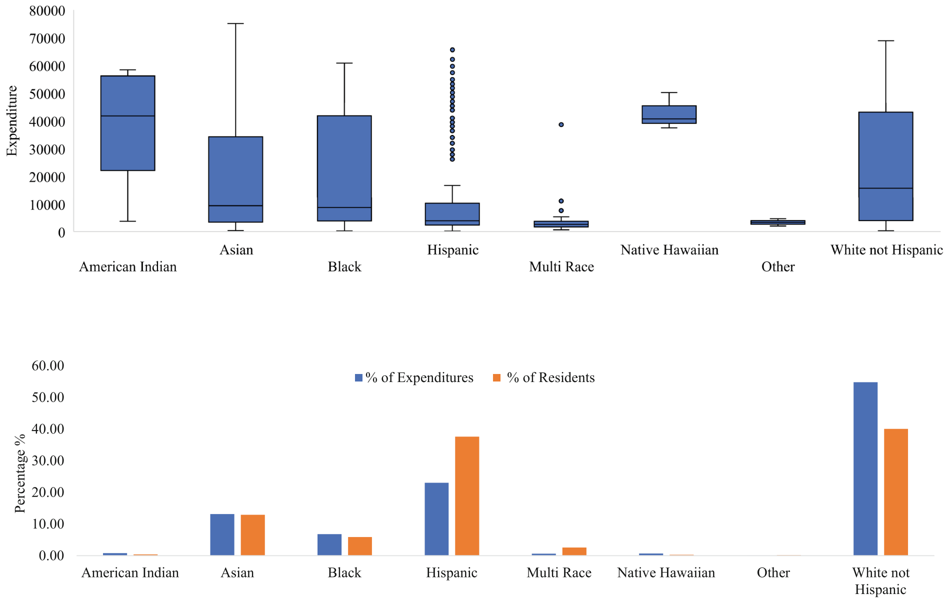


Fig. 4. Distribution of expenditure as per the ethnic groups.

5.1 The Dataset

We use the same dataset to analyse the original claims. The dataset is publicly available at [1]. The dataset mainly consists various information of one thousand disabled residents (DRs) under six important variables (ID, age group, gender, expenditures, ethnicity). Each DR has a unique identification, i.e., "ID". The state department uses AGE to decide the financial needs and other essential needs of DRs. The age groups of the residents are divided into six age groups. (0–5 years old, 6–12 years old, 13–17 years old, 18–21 years old, 22–50 years old, and 51 years old). These groups are based on the amount of financial assistance required at each stage of age. E.g., The 0–5 age group (preschool age) has the fewest needs and thus requires the least funding.

The"Expenditures" variable represents the annual expenditures made by the state to support each resident and their family. Information about the expenditures, the number of residents and their percentage as per ethnicity is given in Table 2. The expenditures include all the expenses, including psychological services, medical fees, transportation and housing costs such as rent (especially for adult residents). As far as the case is concerned, "ethnicity" is the most important demographic variable in the dataset. The dataset includes eight ethnic groups.

As demonstrated in Fig. 4, the population difference between the Hispanic and the White non-Hispanic people is significantly less. However, there is a big difference between the distribution of assistance to the Hispanic and the White non-Hispanic group. Therefore, these two populations are selected for the case study for further investigation.

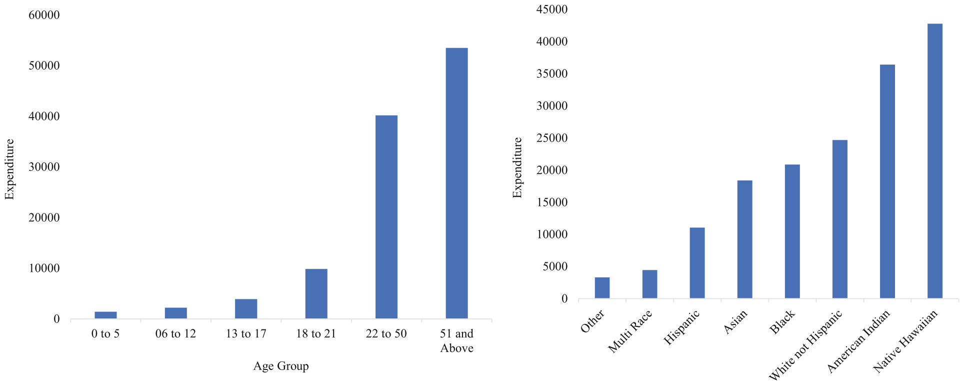


Fig. 5. 1. Average expenditure by age group, 2. Average expenditure by ethnicity.

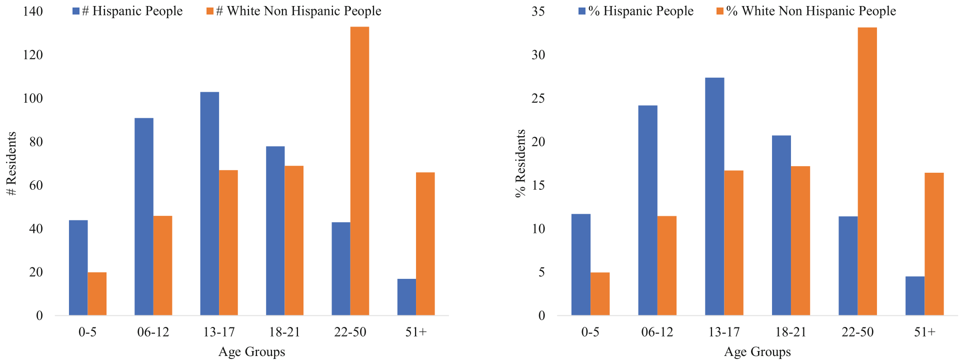


Fig. 6. 1. Hispanic and white non Hispanic residents with their age groups, 2. Percentage of Hispanic and white non Hispanic residents according the age groups.

5.2 Data Analysis

We begin the data analysis by comparing the total amount of expenditure in relation to different ethnic groups. As per the bar chart given in Fig. 5, It is clear that the average expenditure on Hispanic residents is significantly lower than the White non-Hispanic residents. Moreover, the analysis of average expenditure by the age groups shows that the average expenditure was very high for the older age groups. As per Fig. 5, it is also a clear case of age discrimination. However, age is not considered a factor for the discrimination because older people are eligible to get higher expenditures (Fig. 7).

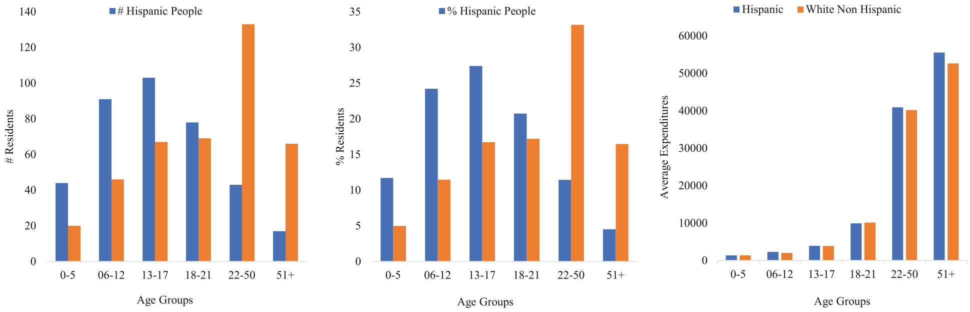


Fig. 7. Average expenditures by ethnicity and age groups.

The overall Hispanic population receiving assistance is younger than the white non-Hispanic population receiving assistance. As the age is showing discriminatory behaviour, therefore, we compare the average amount of funds received by the two observed ethnic groups as per their age groups in Fig. 6. It is clear that the number of beneficiaries from the Hispanic group is higher in the lower age groups, while the number of beneficiaries from the white non-Hispanic group is higher in the older groups. As white non-Hispanic are older people, therefore, they are receiving more support.

Now we see an opposite picture of the case, in Fig. 6. The aggregated data shows that white non-Hispanic people have more support from the department; however, for most of the age groups except one age group, the average expenditure for the Hispanics was higher. So, we are witnessing Simpson's paradox!. The age group variable proved to be lurking in this case, without which we can not show any results in marginal data.

6 Discussion

The existence of statistical paradoxes in benchmark datasets and in real-life case studies provides a direction to understand the causality in decision making. We noticed that most machine learning and deep learning algorithms focus only on identifying correlations rather than identifying the real or causal relationships between data items. Therefore, understanding and evaluating causality is an important term to be discussed in big data, Data Science, AI and ML.

7 Conclusion

Handling statistical paradoxes is a complex challenge in AI, ML and Big Data. Different paradoxes state the possibilities of errors in the outcomes of automatic data analysis conducted for AI, Ml and big data based applications. In this paper, we discussed the existence of Berkson's paradox and demonstrate the existence of Simpson's paradox and in two real datasets. Statistical paradoxes in data reflect the importance of probabilities and causal inference and seek a manual inspection of datasets. We argue that if confounding effects are not properly addressed in datasets, outcomes of an data analysis can be completely opposite. However, with the right tools and data analysis, a good analyst or data scientist can handle it in a better way. The statistical paradoxes confirm essential statistical evaluation for datasets and demonstrate the importance of human experts in the loop to examine and query Big datasets.

References

1. California Department of Developmental Services CDDS expenditures. https://kaggle.com/wduckett/californiaddsexpenditures
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of VLDB 1994 - The 20th International Conference on Very Large Data Bases, pp. 487–499. Morgan Kaufmann (1994)
3. Alipourfard, N., Fennell, P.G., Lerman, K.: Can you trust the trend? Discovering Simpson's paradoxes in social data. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, pp. 19–27. Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3159652.3159684
4. Alipourfard, N., Fennell, P.G., Lerman, K.: Using Simpson's paradox to discover interesting patterns in behavioral data. In: Proceedings of the Twelfth International AAAI Conference on Web and Social Media. AAAI Publications (2018)
5. Berkson, J.: Limitations of the application of fourfold table analysis to hospital data. Biometrics Bull. **2**(3), 47–53 (1946). http://www.jstor.org/stable/3002000
6. Blyth, C.R.: On Simpson's paradox and the sure-thing principle. J. Am. Stat. Assoc. **67**(338), 364–366 (1972)
7. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
8. Cattell, R.B.: P-technique factorization and the determination of individual dynamic structure. J. Clin. Psychol. (1952)
9. Commission, E., Centre, J.R., Wenzl, T.: Smoking and COVID-19: a review of studies suggesting a protective effect of smoking against COVID-19. Publications Office (2020). https://doi.org/10.2760/564217
10. Conger, A.J.: A revised definition for suppressor variables: a guide to their identification and interpretation. Educ. Psychol. Measur. **34**(1), 35–46 (1974)
11. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
12. Dawid, A.P.: Conditional independence in statistical theory. J. Roy. Stat. Soc.: Ser. B (Methodol.) **41**(1), 1–15 (1979). https://doi.org/10.1111/j.2517-6161.1979.tb01052.x
13. Draheim, D.: DEXA'2019 keynote presentation: future perspectives of association rule mining based on partial conditionalization, Linz, Austria, 28th August 2019. https://doi.org/10.13140/RG.2.2.17763.48163
14. Draheim, D.: Future perspectives of association rule mining based on partial conditionalization. In: Hartmann, S., Küng, J., Chakravarthy, S., Anderst-Kotsis, G., A Min Tjoa, Khalil, I. (eds.) Database and Expert Systems Applications. LNCS, vol. 11706, p. xvi. Springer, Heidelberg (2019) (2019)
15. Fisher, R.A.: The use of multiple measurement in taxonomic problems. Ann. Eugen. **7**(2), 179–188 (1936). https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

16. Fisher, R.A.: III. The influence of rainfall on the yield of wheat at rothamsted. Philos. Trans. R. Soc. London Ser. B Containing Papers Biological Character **213**(402–410), 89–142 (1925)
17. Freitas, A.A., McGarry, K.J., Correa, E.S.: Integrating Bayesian networks and Simpson's paradox in data mining. In: Texts in Philosophy. College Publications (2007)
18. Griffith, G.J., et al.: Collider bias undermines our understanding of COVID-19 disease risk and severity. Nat. Commun. **11**(1), 5749 (2020). https://doi.org/10.1038/s41467-020-19478-2
19. Kaushik, M., Sharma, R., Peious, S.A., Draheim, D.: Impact-driven discretization of numerical factors: case of two- and three-partitioning. In: Srirama, S.N., Lin, J.C.-W., Bhatnagar, R., Agarwal, S., Reddy, P.K. (eds.) BDA 2021. LNCS, vol. 13147, pp. 244–260. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-93620-4_18
20. Kaushik, M., Sharma, R., Peious, S.A., Shahin, M., Ben Yahia, S., Draheim, D.: On the potential of numerical association rule mining. In: Dang, T.K., Küng, J., Takizawa, M., Chung, T.M. (eds.) FDSE 2020. CCIS, vol. 1306, pp. 3–20. Springer, Singapore (2020). https://doi.org/10.1007/978-981-33-4370-2_1
21. Kaushik, M., Sharma, R., Peious, S.A., Shahin, M., Yahia, S.B., Draheim, D.: A systematic assessment of numerical association rule mining methods. SN Comput. Sci. **2**(5), 1–13 (2021). https://doi.org/10.1007/s42979-021-00725-2
22. Kievit, R., Frankenhuis, W., Waldorp, L., Borsboom, D.: Simpson's paradox in psychological science: a practical guide. Front. Psychol. **4**, 513 (2013). https://doi.org/10.3389/fpsyg.2013.00513
23. Kim, Y.: The 9 pitfalls of data science. Am. Stat. **74**(3), 307–307 (2020). https://doi.org/10.1080/00031305.2020.1790216
24. King, G., Roberts, M.: EI: A (n R) program for ecological inference. Harvard University (2012)
25. Ma, H.Y., Lin, D.K.J.: Effect of Simpson's paradox on market basket analysis. J. Chin. Stat. Assoc. **42**(2), 209–221 (2004). https://doi.org/10.29973/JCSA.200406.0007
26. MacKinnon, D.P., Fairchild, A.J., Fritz, M.S.: Mediation analysis. Annu. Rev. Psychol. **58**(1), 593–614 (2007). https://doi.org/10.1146/annurev.psych.58.110405.085542. pMID: 16968208
27. Pearl, J.: Causal inference without counterfactuals: comment. J. Am. Stat. Assoc. **95**(450), 428–431 (2000)
28. Pearl, J.: Understanding Simpson's paradox. SSRN Electron. J. **68** (2013). https://doi.org/10.2139/ssrn.2343788
29. Pearson Karl, L.A., Leslie, B.M.: Genetic (reproductive) selection: inheritance of fertility in man, and of fecundity in thoroughbred racehorses. Philos. Trans. R. Soc. Lond. Ser. A **192**, 257–330 (1899)
30. Quinlan, J.: Combining instance-based and model-based learning. In: Machine Learning Proceedings 1993, pp. 236–243. Elsevier (1993). https://doi.org/10.1016/B978-1-55860-307-3.50037-X
31. Robinson, W.S.: Ecological correlations and the behavior of individuals. Am. Sociol. Rev. **15**(3), 351–357 (1950)
32. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)
33. Simpson, E.H.: The interpretation of interaction in contingency tables. J. Roy. Stat. Soc.: Ser. B (Methodol.) **13**(2), 238–241 (1951)

34. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. In: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, pp. 1–12 (1996)
35. Taylor, S.A., Mickel, A.E.: Simpson's paradox: a data set and discrimination case study exercise. J. Stat. Educ. **22**(1), 8 (2014). https://doi.org/10.1080/10691898.2014.11889697
36. Tu, Y.K., Gunnell, D., Gilthorpe, M.S.: Simpson's paradox, lord's paradox, and suppression effects are the same phenomenon-the reversal paradox. Emerg. Themes Epidemiol. **5**(1), 1–9 (2008)
37. Von Kugelgen, J., Gresele, L., Scholkopf, B.: Simpson's paradox in COVID-19 case fatality rates: a mediation analysis of age-related causal effects. IEEE Trans. Artif. Intell. **2**(1), 18–27 (2021). https://doi.org/10.1109/tai.2021.3073088
38. Yule, G.U.: Notes on the theory of association of attributes in statistics. Biometrika **2**(2), 121–134 (1903)

Appendix 7

[VII]

R. Sharma, M. Kaushik, S. A. Peious, M. Shahin, A. Vidyarthi, and D. Draheim. Existence of the Yule-Simpson effect: An experiment with continuous data. In *Proceedings of Confluence 2022 – the 12th International Conference on Cloud Computing, Data Science & Engineering*, pages 351–355, 2022

2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence) | 978-1-6654-3701-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/Confluence52989.2022.9734211

Existence of the Yule-Simpson Effect: An Experiment with Continuous Data

1st Rahul Sharma
Information Systems Group
Tallinn University of Technology
Tallinn, Estonia
rahul.sharma@taltech.ee

2nd Minakshi Kaushik
Information Systems Group
Tallinn University of Technology
Tallinn, Estonia
minakshi.kaushik@taltech.ee

3rd Sijo Arakkal Peious
Information Systems Group
Tallinn University of Technology
Tallinn, Estonia
sijo.arakkal@taltech.ee

4th Mahtab Shahin
Information Systems Group
Tallinn University of Technology
Tallinn, Estonia
mahtab.shahin@taltech.ee

5thAnkit.Vidyarthi
Jaypee Institute of Information Technology
Noida, India
dr.ankit.vidyarthi@gmail.com

6thDirk.Draheim
Information Systems Group
Tallinn University of Technology
Tallinn, Estonia
dirk.draheim@taltech.ee

***Abstract*—In today's world, artificial intelligence based smart applications and smart medical devices are developed with big data-based trained datasets. However, what if a training dataset used to train a machine learning module is incorrect and has a statistical paradox. Statistical paradoxes are complicated to observe in data but are very important to analyze in every training datasets. This article discusses Simpson's paradox and its effects on various datasets. We provide that Simpson's paradox is more common in a variety of data and it leads to wrong conclusions potentially with harmful consequences. We provide a mathematical analysis of Simpson's paradox and analyse its effects on continuous data. Experiments on real-world and synthetic datasets clearly show that the paradox severely impacts big data.**

***Index Terms*—Simpson Paradox, Big Data , Artificial Intelligence, Data Science, Explainable AI**

I. Introduction

In artificial intelligence (AI) use cases, the machine learning (ML) based trained artificial systems provide fast and accurate outcomes for the trained data model. However, it does not guarantee accurate outcomes for every use case in real life. Therefore, understanding causal relationship in data is essential for drawing the proper conclusions.

Data has always been critical in making any decision. Moreover, processing petabytes of data manually in the age of big data is impossible. Manual data analysis techniques are becoming obsolete as the volume of data grows. As a result, it is difficult to conduct statistical analysis on each dataset. Recently, Kügelgen et al. [1] pointed out the importance of statistical analysis of real data and demonstrated instances of Simpson's paradox in Covid-19 data analysis. They provide that the overall case fatality rate (CFR) was higher in Italy than China. However, in every age group, the fatality rate was higher in China than in Italy. These observations raise many questions on the accuracy of data and its analysis. Therefore, understanding data, especially big data, is more critical than its processing.

In statistics, understanding statistical paradoxes are essential for drawing the proper conclusions from data. There are many statistical paradoxes (e.g., Braess's Paradox, Moravec's Paradox, Law of Unintended Consequences and Tea Leaf Paradox). However, Simpson's paradox is one of the known statistical paradoxes in statistics. The paradox is not a new concept to be discussed in the statistics. It is available in different forms and with several names [2] (Reversal paradox, Yule-Simpson effect, Simpson's paradox, amalgamation paradox). This statistical phenomenon was first pointed out by Karl G. Pearson in the year 1899 [3] and later in the year 1903 by George U. Yule [4]. A similar phenomenon was discussed in a short paper by Edward H. Simpson in 1903 [5]. In the year 1972, Colin R. Blyth [6] called it "Simpson's paradox". In this article, we bring the impact of Yule Simpson's effect in big data and discuss its impact on a continuous dataset.

The paper is organized as follows. In Sec. II, an overview of the background work is discussed. In Sect. III, we discuss how data can lie. In Sec. IV, discussion on Simpson's paradox, vector representation and mathematical interpretation of Simpson's paradox is given. In Sect. V, we use a dataset to show the impact of Simpson's paradox on big data. Discussion and future work is presented in Sec. VI. Finally, a conclusion is provided in Sect. VII.

II. Background

Statistical paradoxes are not new to be discussed. In statistics, they are discussed widely and exist since more than a century. In a technical paper published in 1951 [5], Edward H. Simpson described the phenomenon of reversing results, but statisticians Karl Pearson et al. in 1899 [3] and Udny Yule in 1903 [4] had mentioned similar effects previously. Udny Yule reported the existence of association paradox with categorical

This work has been partially conducted in the project "ICT programme" which was supported by the European Union through the European Social Fund.

TABLE I
SIMPSON'S PARADOX: A SAMPLE DATASET FROM SIMPSON'S EXAMPLE [5]

	Full Population			**Women (M), N=20**			**Men (M), N=32**		
	Success (S)	Failure ($\neg S$)	Succ.%	Success (S)	Failure ($\neg S$)	Succ.%	Success (S)	Failure ($\neg S$)	Succ.%
Treatment (T)	20	20	50%	8	5	61%	12	15	44%
Control ($\neg T$)	6	6	50%	4	3	57%	2	3	40%

variables and Karl demonstrated the paradox in marginal and partial associations between continuous variables. In 1972, Colin R. Blyth coined the term "Simpsons Paradox" [6]. The paradox is also known as the Yule–Simpson effect, amalgamation paradox, or reversal paradox [7].

Simpson's paradox is conceptually related to many statistical challenges and techniques [8], including causal inference [9], [10], the ecological fallacy [11], [12], Lord's paradox [13], propensity score matching [14], suppressor variables [15], conditional independence [16], partial correlations [17], p-technique [18] and mediator variables [19].

The implications of Simpson's paradox are severe in various ML techniques, e.g., in association rule mining (ARM) [20], [21], the instances of Simpson's paradox are discussed by Dirk [22], [23]. He has discussed "Future perspectives of association rule mining based on partial conditionalization".

III. HOW BIG DATA CAN LIE

Data does not speak itself. It consists numbers and statistics, which tell the story about the data. All datasets, irrespective of their type and size, are somehow based on numbers and statistics. Therefore, it is very much possible to have an instance of statistical paradoxes in data. Simpson's paradox is one of the known statistical paradoxes; however, there are other statistical paradoxes (e.g., Braess's Paradox, Moravec's Paradox, Law of Unintended Consequences, Tea Leaf Paradox) that can be in a dataset. Machine Learning models are susceptible to cognitive paradoxes between training and testing because they create knowledge from data. Compared to algorithms or big data processing [24], data literacy may not seem exciting, but it should form the basis for any every data processing [25].

IV. SIMPSON'S PARADOX

We start the discussion on the paradox by using the real-world dataset from Simpson's article [5]. In this example, analysis for medical treatment is demonstrated. Table I summarizes effect of the medical treatment for the entire population $(N = 52)$ as well as for men and women separately in subgroups. The treatment appears effective for both male and female subgroups; however, the treatment appears ineffective at the whole population level.

Definition 1: Consider D groups of data such that group D_1 has A_i trials and $0 \leq a_i \leq A_i$ "successes". Similarly, consider an analogous D groups of data such that group D_2 has B_i trials and $0 \leq b_i \leq B_i$ "successes". Then Simpson's paradox occurs if

$$\frac{a_1}{A_1} \geq \frac{b_1}{B_1} and \frac{a_2}{A_i} \geq \frac{b_2}{B_2} \text{ for all } i = 1, 2, \ldots, n \text{ but } \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} A_i} \leq \frac{\sum_{i=1}^{n} b_i}{\sum_{i=1}^{n} B_i}, \quad (1)$$

We use the following example to show how this equation works.

$$\frac{10}{20} > \frac{30}{70} and \frac{10}{50} > \frac{10}{60} \text{ but } \frac{10+10}{20+50} < \frac{30+10}{70+60}, \quad (2)$$

We could also flip the inequalities and still have the paradox since A and B are chosen arbitrarily.

Classically the paradox is generally expressed via contingency tables. Let a 2×2 contingency table for treatment (T) and sucess (S) in the $i^t h$ sub-population is represented by a four-dimensional vector of real numbers $D = (a_1, b_1, a_2, b_2)$ then;

$$D = \sum\nolimits_{i=1}^{N} D_i = \left(\sum a_i, \sum b_i, \sum c_i, \sum d_i\right) \quad (3)$$

is the aggregate dataset over N sub populations. This can be read simply in Table II.

A. Vector Interpretation of Simpson's Paradox

A simple case of the Simpson's paradox can be illustrated by a two-dimensional vector space [26]. We use a simple example given in the Eq. 2 to draw the vector.

We have a vector $\overrightarrow{V}_1 = (A_1, a_1)$ with a slope of $\frac{a_1}{A_1}$ and another vector$\overrightarrow{V}_2 = (A_2, a_2)$ with a slope of $\frac{a_2}{A_2}$. If the success rate of two vectors $\overrightarrow{V}_1$ and $\overrightarrow{V}_2$ are combined then according to the rule of parallelograms, the results will be sum of the vectors, i.e., $(A_1 + A_2, a_1 + a_2)$ with slope $\frac{a_1+a_2}{A_1+A_2}$. The longer vectors represent a higher success rate.

Simpson's paradox provides that although the vector $\overrightarrow{V}_1$ has a smaller slope than the vector $\overrightarrow{Z}_1(B_1, b_1)$, and $\overrightarrow{V}_2$ has a smaller slope than the vector $\overrightarrow{Z}_2$ but the sum of two vectors $\overrightarrow{V}_1 + \overrightarrow{V}_2$ can potentially have a larger slope than the sum $\overrightarrow{Z}_1 + \overrightarrow{Z}_2$. In Fig. 1, the vectors with smaller slopes are represented in blue, and the vectors with a larger slope are represented in red. The dashed lines represent the sum of the vectors.

TABLE II
A 2 × 2 CONTINGENCY TABLE WITH SUB POPULATION GROUPS D1 AND D2

	Population $D = D_1 + D_2$		**Sub-population** D_1		**Sub-population** D_2	
	Success (S)	Failure ($\neg S$)	Success (S)	Failure ($\neg S$)	Success (S)	Failure ($\neg S$)
Treatment (T)	$a_1 + a_2$	$b_1 + b_2$	a_1	b_1	a_2	b_2
No Treatment ($\neg T$)	$c_1 + c_2$	$d_1 + d_2$	c_1	d_1	c_2	d_2

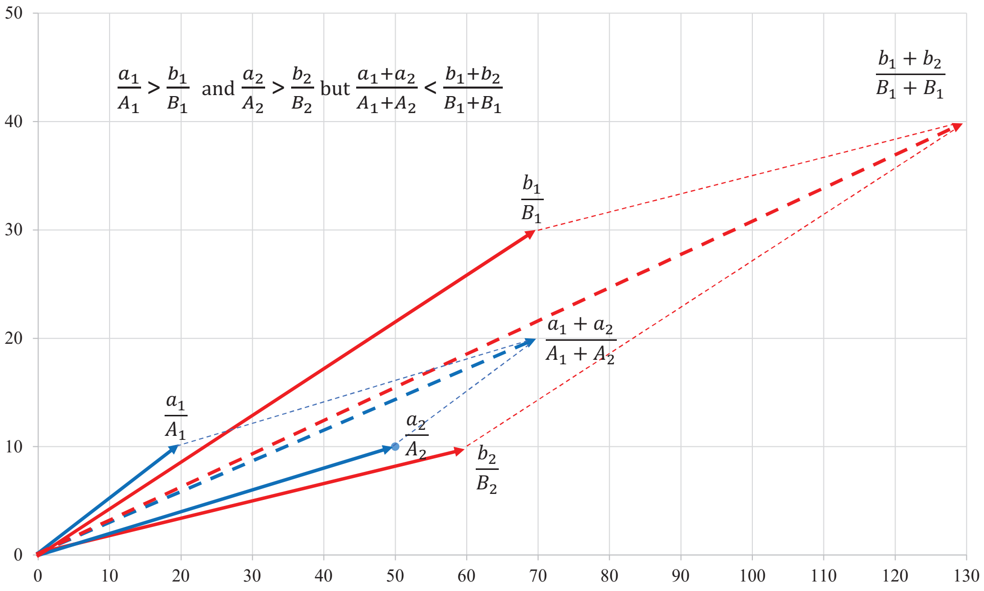


Fig. 1. Vector representation of Simpson's paradox. The dashed lines are showing the sum of the respective vectors

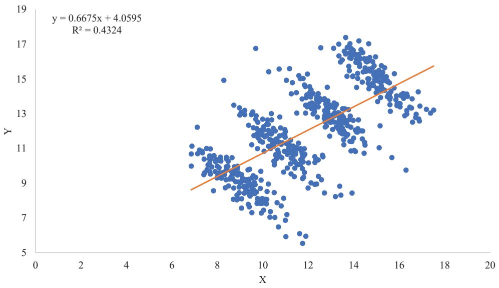


Fig. 2. Trend in a population group: The solid red line is used to show the overall trend

We also can demonstrate this Simpson's paradox scenario via probability theory and conditional probabilities. If we assume; $T = treatment$, $S = successful$, $M = Male$, and $F = Female$ then;

$$\mathsf{P}(S \mid T) = \mathsf{P}(S \mid \neg T) \tag{4}$$

$$\mathsf{P}(S \mid T, M) > \mathsf{P}(S \mid \neg T, M) \tag{5}$$

$$\mathsf{P}(S \mid T, \neg M) > \mathsf{P}(S \mid \neg T, \neg M) \tag{6}$$

Based on Eq. 4 and 5, 6, one can decide to use the treatment or not use the treatment. As per the success rate for male and female population, the treatment is a success, but overall, the treatment is a failure. This reversal of results between groups population and the total population has been referred to as Simpson's Paradox. In statistics, this concept has been

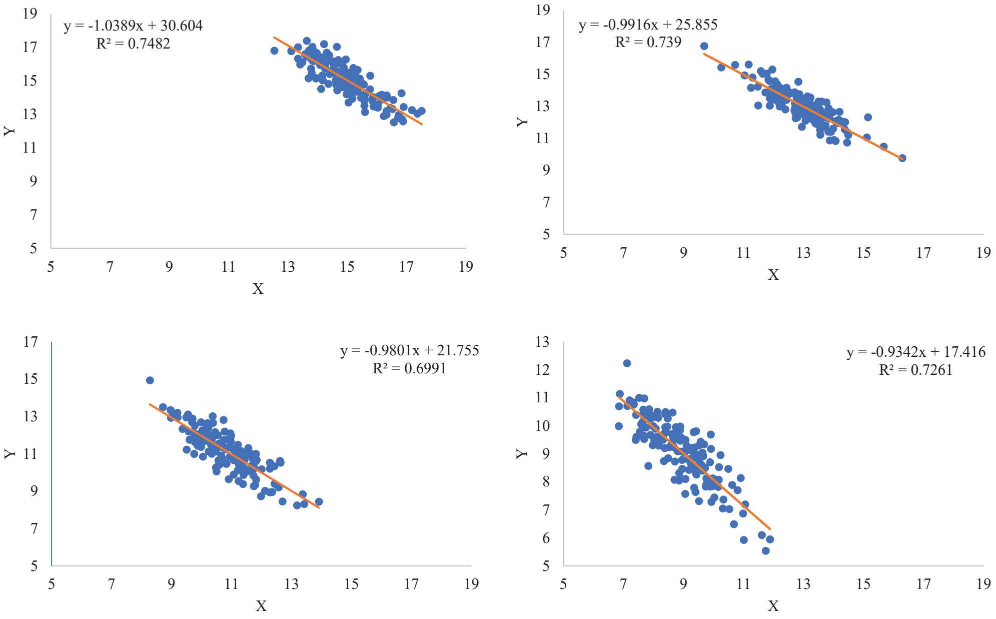


Fig. 3. Reversal of trends while the total population is divided into sub groups

discussed widely and named differently by several authors [3], [4].

V. Existence of Simpson's Paradox in Big Data

Simpson's paradox can exist in any dataset irrespective of its size and type. The paradox demonstrates the importance of having human experts in the loop to examine and query Big Data results. In this section, we present a dataset to analyse the presence and implications of Simpson's paradox on big data.

A. *Simpson's Paradox in Continuous Data*

The instances of Simpson's paradox can be easily discovered by a data scientists in categorical and continuous data [27], [28]. Here, we consider the co-variance between two variables $\sigma(x, y)$ as an example to show the existence of Simpson's paradox in continuous data. We have also discussed the measures to find the impact of one numerical variable to another numerical variable [29]. We use a synthetic dataset to calculate the co-variance between (xy). The data is publicly available at [30]. The X and Y variables have different signs for sub populations and the entire population.

As provided in Fig. 2, the correlation between X and Y is positive in the complete dataset, but as per Fig. 3, the correlation between X and Y is negative within each subgroup. Therefore, this is a case of Simpson's paradox in continuous data. To identify the Simpson's paradox in a continuous dataset with n continuous variable and m discrete variables, we can compute a correlation matrix $(n \times n)$ for all the data. Then for m discrete variable with k_m levels, an additional $(n \times n)$ matrix needs to be calculated for each level of variables as follows. Therefore, we need to calculate the $1 + \sum_i^m = k_i$ correlation matrices of size $(n \times n)$ and compare it with the lower half of $\sum_i^m = k_i$ for subgroup levels.

VI. Discussion and Future Work

Simpson's paradox emerged as a well-known problem in Big Data, AI and ML. It demonstrates a picture not to believe in trends and rejects the trust surrounding the AI and ML applications. Recently, some useful research on explainable artificial intelligence (XAI) discussed various ways to handle the confounding effects and Simpson's paradox, which is the utmost need for next-generation AI and ML applications.

VII. Conclusion

In this paper, a dataset for continuous data is used to demonstrate the existence of Simpson's paradox in continuous data. We provide that if the confounding effects are not addressed appropriately in a datasets then conclusions obtained from that datasets may be totally wrong. Without enough statistical knowledge, it's challenging to know which view of the relationship between two variables makes more sense – the one with or without the third variable. Simpson's paradox is a complex problem for Big data, AI and ML, but with the right tools and data analysis, a good analyst or data scientist can handle it in a better way.

References

[1] J. Von Kugelgen, L. Gresele, and B. Scholkopf, "Simpson's paradox in COVID-19 case fatality rates: A mediation analysis of age-related causal effects," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 1, p. 18–27, Feb 2021.

[2] I. J. Good and Y. Mittal, "The amalgamation and geometry of two-by-two contingency tables," *The Annals of Statistics*, vol. 15, no. 2, Jun. 1987.

[3] L. A. Pearson Karl and B.-M. Leslie, "Genetic (reproductive) selection: Inheritance of fertility in man, and of fecundity in thoroughbred racehorses," *Philosophical Transactions of the Royal Society of London: Series A*, vol. 192, pp. 257–330, Dec. 1899.

[4] G. U. Yule, "Notes on the theory of association of attributes in statistics," *Biometrika*, vol. 2, no. 2, pp. 121–134, 02 1903.

[5] E. H. Simpson, "The interpretation of interaction in contingency tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 13, no. 2, pp. 238–241, 1951.

[6] C. R. Blyth, "On Simpson's paradox and the sure-thing principle," *Journal of the American Statistical Association*, vol. 67, no. 338, pp. 364–366, Jun. 1972.

[7] J. Quinlan, "Combining instance-based and model-based learning," in *Machine Learning Proceedings 1993*. Elsevier, 1993, pp. 236–243.

[8] D. Draheim, *Generalized Jeffrey Conditionalization – A Frequentist Semantics of Partial Conditionalization*. Heidelberg New York Berlin: Springer, 2017.

[9] J. Pearl, "Understanding simpson's paradox," *SSRN Electronic Journal*, vol. 68, 01 2013.

[10] ——, "Causal inference without counterfactuals: Comment," *Journal of the American Statistical Association*, vol. 95, no. 450, pp. 428–431, 2000. [Online]. Available: http://www.jstor.org/stable/2669380

[11] W. S. Robinson, "Ecological correlations and the behavior of individuals," *American Sociological Review*, vol. 15, no. 3, pp. 351–357, 1950.

[12] G. King and M. Roberts, "Ei: a (n r) program for ecological inference," *Harvard University*, 2012.

[13] Y.-K. Tu, D. Gunnell, and M. S. Gilthorpe, "Simpson's paradox, lord's paradox, and suppression effects are the same phenomenon–the reversal paradox," *Emerging themes in epidemiology*, vol. 5, no. 1, pp. 1–9, 2008.

[14] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[15] A. J. Conger, "A revised definition for suppressor variables: A guide to their identification and interpretation," *Educational and psychological measurement*, vol. 34, no. 1, pp. 35–46, 1974.

[16] A. P. Dawid, "Conditional independence in statistical theory," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 41, no. 1, pp. 1–15, 1979. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1979.tb01052.x

[17] R. A. Fisher, "Iii. the influence of rainfall on the yield of wheat at rothamsted," *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, vol. 213, no. 402-410, pp. 89–142, 1925.

[18] R. B. Cattell, "P-technique factorization and the determination of individual dynamic structure." *Journal of Clinical Psychology*, 1952.

[19] D. P. MacKinnon, A. J. Fairchild, and M. S. Fritz, "Mediation analysis," *Annual Review of Psychology*, vol. 58, no. 1, pp. 593–614, 2007, pMID: 16968208. [Online]. Available: https://doi.org/10.1146/annurev.psych.58.110405.085542

[20] R. Agrawal, T. Imieliński, and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases," *ACM SIGMOD Record*, vol. 22, no. 2, pp. 207–216, 1993.

[21] R. Sharma, M. Kaushik, S. A. Peious, S. B. Yahia, and D. Draheim, "Expected vs. unexpected: Selecting right measures of interestingness," in *Proceedings of DaWaK'2020 – the 22nd International Conference on Data Warehousing and Knowledge Discovery*, ser. Lecture Notes in Computer Science, vol. 12393. Springer, 2020, pp. 38–47.

[22] D. Draheim, "Future perspectives of association rule mining based on partial conditionalization," in *Proceedings of DEXA'2019 - the 30th International Conference on Database and Expert Systems Applications*, ser. LNCS, S. Hartmann, J. Küng, S. Chakravarthy, G. Anderst-Kotsis, A Min Tjoa, and I. Khalil, Eds., vol. 11706. Heidelberg New York Berlin: Springer, 2019, p. xvi.

[23] ——, "DEXA'2019 keynote presentation: Future perspectives of association rule mining based on partial conditionalization." Linz, Austria, 28th August 2019. https://doi.org/10.13140/RG.2.2.17763.48163.

[24] M. Shahin, S. Arakkal Peious, R. Sharma, M. Kaushik, S. Ben Yahia, S. A. Shah, and D. Draheim, "Big data analytics in association rule mining: A systematic literature review," in *Proceedings of BDET'2021 - the 3rd International Conference on Big Data Engineering and Technology*. Association for Computing Machinery, 2021, pp. 40–49. [Online]. Available: https://doi.org/10.1145/3474944.3474951

[25] S. A. Peious, R. Sharma, M. Kaushik, S. A. Shah, and S. B. Yahia, "Grand reports: A tool for generalizing association rule mining to numeric target values," in *Proceedings of DaWaK'2020 – the 22nd International Conference on Data Warehousing and Knowledge Discovery*, ser. Lecture Notes in Computer Science, vol. 12393. Springer, 2020, pp. 28–37.

[26] J. Kocik, "Proof without words: Simpson's paradox," *Mathematics Magazine*, vol. 74, no. 5, p. 399, Dec. 2001.

[27] M. Kaushik, R. Sharma, S. A. Peious, M. Shahin, S. B. Yahia, and D. Draheim, "A systematic assessment of numerical association rule mining methods," *SN Computer Science*, vol. 2, no. 5, pp. 1–13, 2021.

[28] M. Kaushik, R. Sharma, S. A. Peious, M. Shahin, S. Ben Yahia, and D. Draheim, "On the potential of numerical association rule mining," in *Proceedings of FDSE'2020 – the 7th International Conference on Future Data and Security Engineering*, ser. Lecture Notes in Computer Science, vol. 12466. Springer, 2020, pp. 3–20.

[29] M. Kaushik, R. Sharma, S. A. Peious, and D. Draheim, "Impact-driven discretization of numerical factors: Case of two- and three-partitioning," in *Big Data Analytics*. Cham: Springer International Publishing, 2021, pp. 244–260.

[30] R. Sharma, "How Big Data Can Lie: An Experiment with Data Scientists," https://github.com/rahulgla/Simpson-s-paradox.

Appendix 8

[VIII]

R. Sharma. On statistical paradoxes and overcoming the impact of bias in expert systems: towards fair and trustworthy decision making. *SSRN*, pages 1–37, July 2023. doi:10.2139/ssrn.4506432

On statistical paradoxes and overcoming the impact of bias in expert systems: towards fair and trustworthy decision making

Rahul Sharma[a]

[a]*Information Systems Group, Tallinn University of Technology, Tallinn, Estonia*

Abstract

In the past two decades, there has been tremendous progress in promoting various decision support techniques (DSTs), e.g., data mining (association rule mining), artificial intelligence (AI), machine learning (ML), and deep learning (DL) across various fields, including healthcare, autonomous driving, personal assistant technology, businesses, education, and justice. However, despite many success stories and advantages, these techniques are often considered biased, unfair, and untrustworthy. In this paper, we examine some of the well-known statistical paradoxes as witnesses of expert system bias. Expert system bias challenges successful decision-making, as it is a direct source of biased decisions. Unfortunately, statistical paradoxes are extreme forms of bias, but their roles have not been discussed adequately in DSTs. In this paper, we aim to discuss how to handle confounding effects and deal with the severe impacts of statistical paradoxes in DSTs. Further, we outline a framework for mitigating bias in training datasets. To provide evidence for the relevance of such a framework, we conduct a series of experiments with three different measures on multiple real-world and benchmark datasets. First, we utilise the stratification of Pearson correlation for identifying potential confounders. Second, we utilize inverse propensity weighting and generalise back-door adjustment techniques for continuous data for adjusting the impact of confounders. To demonstrate the practical utility of the proposed framework, we have developed a user-friendly web-based application. The application incorporates the example measures discussed and integrates them into the outlined framework for bias mitigation. We claim that this application can serve as a valuable tool for data scientists and researchers by automatically detecting and addressing confounding effects. We argue that the suggested framework and application hold immense potential for further extensions beyond their current use.

Keywords: Decision support techniques, trustworthy artificial intelligence, machine learning, statistical paradoxes, bias mitigation framework

1. Introduction

Data-driven decisions have always required the correct assessment and analysis of data. In the past, mathematical and statistical methods have been used broadly to derive insights from data. However, in the last two decades, with the emergence of big data, decision support techniques (DSTs) from data mining, artificial intelligence (AI), data science, and machine learning (ML) have gained massive ground in practice and theory. Nowadays, these techniques play an important role in politics, social science, and medical sciences [1, 2] and significantly influence people's lives and decisions, either directly or indirectly [3, 4, 5].

Presently, in a majority of AI use cases, ML-based trained decision support systems deliver quick responses; despite many advantages and success stories, they are still not fully reliable, fair and trustworthy. In several instances, the results of AI applications have been found partially or fully biased [6, 7, 8]. There are many such examples which have raised serious questions about classical data mining techniques [9].

In 2021, Michael Gentzel [10] highlighted the use of biased facial recognition technology by law enforcement in liberal democracies. In 2018, Joy Buolamwini [11], a member of the MIT Media Lab's Civic Media group, found that a commercially available facial recognition system was significantly more likely to misidentify darker-skinned individuals than lighter-skinned individuals. This was partly due to the fact that the dataset used to train the system was overwhelmingly composed of lighter-skinned individuals, which led to a bias in the system's predictions. In 2016, Julia et al. [7] from ProPublica found that a widely used algorithm for predicting recidivism among criminal defendants was biased against African American defendants. The algorithm was based on data from past criminal defendants, which included a large number of African American defendants who had been arrested for low-level offences. This led to the algorithm predicting that African American defendants were more likely to re-offend, even when they were less likely to do so in reality. Many such examples raise severe questions about the accuracy of AI applications and their machine-learning models.

On the other hand, some of the DSTs, e.g., association rule mining, statistical reasoning and OLAP, have a similar set of objectives. Still, they have been introduced and are used with their own set of mathematical formalizations and have developed their specific terminologies [12]. There are numerous reasons which affect the performance and trustworthiness of DSTs, specifically AI applications [13].

In Fig. 1, some common types of bias that generally impact AI applications are given [14]. Here, we highlight statistical paradoxes as extreme forms of data-driven bias in AI applications. We argue that the existence of statistical paradoxes has not yet been addressed appropriately in the mainstream AI application development scenario.

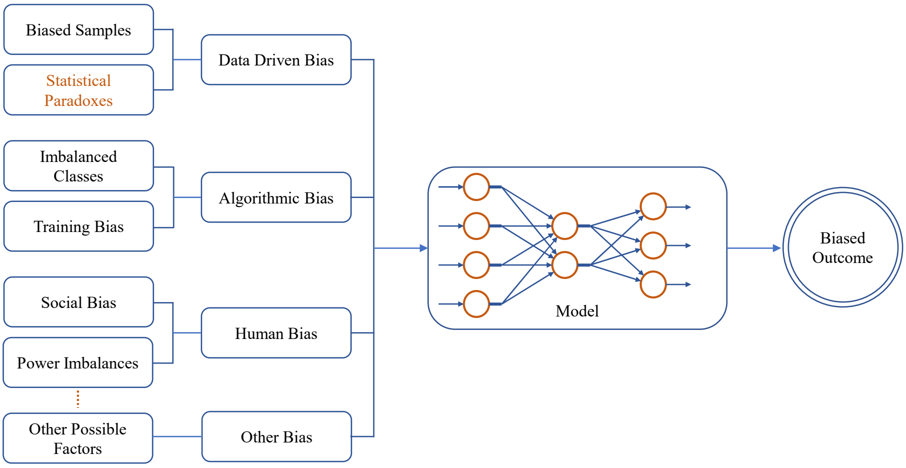


Figure 1: Existence of bias in DSTs is a critical concern that requires extensive exploration. This figure shows different types of biases [14] that are common in AI systems and emphasises statistical paradoxes as extreme forms of data-driven bias

The research on overcoming the impact of bias in the development of various DSTs is highly relevant due to the following reasons:

1. Existence of bias in DSTs can lead to bias and inequalities against certain groups of people.
2. Addressing bias can lead to improved model performance and greater accuracy in predictions. This can significantly improve the accuracy and reliability of DST-based applications, making them more useful and effective in the real world.

By shedding light on statistical paradoxes in several DSTs, we aim to ease the development of strategies to mitigate bias, ultimately leading to more effective and trustworthy decision support systems.

The role of statistical paradoxes and their impact has been discussed deeply in classical data analysis by expert mathematicians and statisticians [15, 16, 17].

Table 1 presents a list of well-known statistical paradoxes that are already addressed in statistics and currently are potential threats to DSTs. In more general, statistical reasoning and probability theory is the foundation of AI and data science, e.g., Naive Bayes classifiers, random forest [18], support vector machines [19], etc. Consequently, causal relationships are usually accompanied by statistical paradoxes in AI and ML-based applications. Therefore, understanding causal relationships hand in hand with evaluating the existence of statistical paradoxes is an essential step forward towards developing trustworthy and fair DSTs.

This paper aims to strengthen DSTs to develop fair and trustworthy decision support applications, which can ultimately lead to an increase in trust and fairness in the usages of DSTs in various areas such as healthcare, finance, and justice. The paper contributes as follows:

1. We suggest a framework for mitigating bias in multivariate training datasets, which elaborates stages of pre-processing, bias mitigation, evaluation, and adjustment of training data.
2. To provide evidence for the relevance of the proposed framework, we conduct a series of experiments with three measures as follows:
   (a) two measures (one for continuous data and one for categorical data) for investigating confounders via detecting instances of Simpson's paradox in regard to stratification of Pearson correlation.
   (b) a (novel) measure for adjusting the impact of confounders, which generalizes standard back-door adjustment to continuous data.

   The experiments are based on multiple real-world and benchmark datasets to evaluate the efficacy and practicality of the measures and the proposed framework.
3. To showcase the usefulness of the proposed framework, a web-based application has been developed that automatically detects and provides possible adjustments to address the impact of potential confounders. We argue that this web application can serve as a valuable tool for data scientists and researchers by automatically detecting and addressing confounding effects to enhance the fairness and trustworthiness of AI applications.

In future work, we plan to extend this research to address various other statistical challenges, thereby increasing its potential impact.

The paper proceeds as follows. In Sect. 2, we discuss existing literature and its usefulness to the proposed work. In Sect. 3, we discuss several statistical paradoxes and explain Simpson's paradox in more depth. Sect. 4 provides detailed information about the Stratification of Pearson correlation for identifying confounders in continuous values. In Sect. 5, two measures for adjusting the confounders are discussed. In Sect. 6, we elaborate the proposed framework. In Sect. 7, an instance of the proposed framework is developed as a web-based application. Sect. 8 provides detailed information about the conducted experiments. In Sect. 9, we provide an exhaustive discussion in terms of relevance, implications, previous work, limitations and future work. We finish the paper with a conclusion in Sect. 10.

2. Related work

Statistical paradoxes such as Simpson's paradox and Berkson's paradox imply confounding effects, which occur when the relationship between two variables is influenced by the presence of a third variable. This third variable is called the confounder. Whenever such third variables are hidden factors, i.e., not present in the data, they are called latent variables. In mathematical statistics, causality and confounding are two related concepts which are widely discussed by established researchers [27, 28]. Causality refers to a cause-and-effect relationship between two events, where the first event (cause) is responsible for the occurrence of the second event (effect). For example, smoking (cause) can lead to lung cancer (effect). Therefore, in AI systems, it is important to identify and control confounding variables in order to ensure that the true causal relationship between the variables of interest is accurately determined.

In literature, Pearl [28, 16] had a significant impact on the development of probabilistic reasoning and causal modelling for AI. He has provided an exhaustive framework for causal inference, i.e., dealing with reasoning about causal relationships. Otte [27] discussed the relationship between probabilistic causality and Simpson's paradox. His discussion relies on the concept of probabilistic causality, which refers to the idea that a cause does not always produce a unique effect but rather alters the probabilities of effects. Schield [29] discussed how Cornfield's conditions could be used to assess the presence of confounding variables that affect both the dependent variable (target variable) and independent variables (impact factors).

Spellman et al. [30] presented a hypothetical scenario involving two possible causes of a particular outcome and showed how the usage of several kinds of information (conditional vs unconditional) could lead to different conclusions. Schaller [31] discusses the role of "evidence, sample size, aggregation, and statistical reasoning in social inference". He discussed how people make judgments and inferences about others based on limited information. Dawid [32] discussed the concept of conditional independence and its implications for statistical inference. Cartwright [33] explored the philosophical concept of causal necessity and its relationship to scientific laws.

Table 1: List of statistical paradoxes which are particularly relevant to AI applications.

Statistical Paradox	Explanation
Simpson's paradox [20]	Simpson's Paradox (often also called: Yule-Simpon's paradox) is named after British mathematician Edward Simpson, who explained a phenomenon in which an obvious trend occurring in a multitude of data sets of a partition reverses as soon as the data sets of the partition are combined.
Berkson's paradox [21]	Berkson's paradox (also called: collider bias, selection bias, sampling bias or ascertainment bias) occurs when the presence of a third variable confounds the relationship between two other variables.
Lord's paradox [22]	Lord's paradox occurs when a machine learning model is trained with too many features or variables, some of which might be correlated with each other rather than with the outcome variable.
Base rate fallacy [23]	The base rate fallacy can manifest in machine learning models when the prior probability of an event or outcome is not taken into account in the model's predictions.
Will Rogers paradox [24]	The Will Rogers phenomenon describes a situation where, as the rarity of an event increases, the accuracy of the model in predicting that event decreases, even though the overall accuracy of the model increases.
Accuracy paradox [25]	The accuracy paradox in AI refers to the phenomenon that a machine learning model may achieve high accuracy on a dataset but still fails to perform well in the real world due to a lack of precision and recall.
Braess's paradox [26]	Braess's paradox is a domain-specific paradox that occurs in transportation networks where adding an additional route can actually increase congestion and travel time for everyone.

Fiedler [34, 35, 36, 37] discussed sampling issues, pseudo contingencies, and inductive reasoning in social psychology, including cognitive consistency, social cognition, and implicit social cognition.

Kievit et al. [38] examined the instances of Simpson's paradox in psychological science and proposed an R package for continuous data to check the confounding effects. They argue that Simpson's paradox may occur in a wide variety of research designs, methods, and questions, in particular, within the social and medical sciences. Alipourfard et al. [39] have discovered the existence of Simpson's paradox in social data and behavioural data [40]. Freitas et al. [41] proposed an algorithm for detecting instances of Simpson's paradox. In [42], Blyth discussed Simpson's paradox and the *sure-thing principle* as two essential concepts in decision theory. Blyth argued that Simpson's paradox and the sure-thing principle are related and that understanding their principles can help decision-makers to avoid making incorrect decisions based on incomplete or misleading data. Curley et al. [43] explained the role of Simpson's paradox and its implications for decision-making. Greenland [44] explored the relationship between Simpson's paradox and Bayesian non-collapsibility, using an example of adding constants in contingency tables. Hernán et al. [45] provided several examples illustrating how Simpson's paradox can arise in different contexts. The author emphasises the importance of understanding confounding variables, selection bias, and effect modification to interpret statistical results properly and draw accurate conclusions. Tu et al. claimed that statistical paradoxes such as "Simpson's paradox, Lord's paradox, and suppression effects are the same phenomenon – the reversal paradox" [46].

The existing literature indicates that in statistics and mathematics, there have been significant discussions addressing confounding, causality and several types of statistical paradoxes. However, these concepts still need to be integrated with mainstream DSTs.

3. Impact of statistical paradoxes on DSTs

3.1. Statistical paradoxes

In DSTs, the impact of statistical paradoxes refers to situations where a decision support system or tool produces unexpected or counter-intuitive results that may not align with human expectations or common sense. The existence of these paradoxes in DSTs can have a significant impact on any individual or organization. For example, Amazon's AI-based recruiting tool [47] did not rate candidates for software development positions in a gender-neutral manner. This was due to the fact that the tool was trained on resumes from more than a decade ago, which favoured male candidates since the tech industry had earlier been male-dominated.

Statistical paradoxes are fundamentally related to a range of various statistical concepts such as partial correlations [48], p-technique [49], suppressor variables [50], conditional independence [32], propensity score matching [51], causal inference [16, 52], and mediator variables [53] as well as statistical challenges such as ecological fallacy [54, 55] and Lord's paradox [46]. Therefore, they are not only limited to AI systems – they can occur in any field that deals with data analysis in terms of a variety of factors, including confounding variables, measurement errors, non-linear relationships and uneven distribution of data. A list of known statistical paradoxes which are harmful to AI applications is given in Table 1. In the sequel, we concentrate on explaining Simpson's paradox as an example. Simpson's paradox presents an extreme case of confounding, has been widely studied in statistics and has severe consequences in AI applications.

3.2. Simpson's paradox

Simpson's paradox is a statistical phenomenon that occurs when the relationship between two variables appears to disappear or even reverse when analysed at different levels of aggregation. This paradox can lead to incorrect conclusions about the relationship between the variables and have significant implications in the design and interpretation of ML and statistical models. The paradox was first discussed in 1899 by Karl Pearson [56] between continuous variables. Later in 1903, Udny Yule further explored the theory of associations in statistics [15] and discovered the paradox in categorical variables. Further, Edward H. Simpson in 1951 [20] described the theory behind the reversal of results. The term "Simpson's paradox" was later coined by Colin R. Blyth in 1972 [42]. This paradox is also known by other names such as the Yule–Simpson effect, amalgamation paradox, or reversal paradox [57].

3.2.1. Most basic case of Simpson's paradox: the $2 \times 2 \times 2$ *contingency table*

We start discussing Simpson's paradox by using the original dataset from Simpson's article [20]. This data set presents the most basic case of a $2 \times 2 \times 2$ contingency table, which is the case of three events representing a target variable, an impact factor and a confounding variable. This basic case can then be generalized in various ways to two arbitrary random variables and we will discuss two of such generalizations in Sects. 3.2.2, 3.2.3. The original data set of [20] is presented in Table 2. Simpon's paradox shows in the data as follows. In terms of both males and females (i.e., in both strata *Male* and *Female*), the treatment shows a positive effect as follows:

$$61.5\% \approx \mathsf{P}_{\mathit{Male}}(\mathit{Alive} \mid \mathit{Treated}) \;>\; \mathsf{P}_{\mathit{Male}}(\mathit{Alive} \mid \mathit{Untreated}) \approx 57.1\% \qquad (1)$$

$$44.\bar{4}\% = \mathsf{P}_{\mathit{Female}}(\mathit{Alive} \mid \mathit{Treated}) \;>\; \mathsf{P}_{\mathit{Female}}(\mathit{Alive} \mid \mathit{Untreated}) = 40\% \qquad (2)$$

Now, based on (1) and (2), we might tend to conclude that the treatment has a positive effect all over – as it has a positive effect for both males and females.

Table 2: Original data set from Simpon's paper (N=52) [20]. The table layout has been re-arranged as compared to [20] to ease the discussion: both independent variables (impact factor 'treatment' and confounder 'gender') are reflected as outermost rows resp. columns and the dependent variable (target variable 'alive/dead') as innermost columns – the data is the same as in [20].

	Male		*Female*	
	Alive	Dead	Alive	Dead
Treated	8	5	12	15
Untreated	4	3	2	3

However, when analyzing the whole population, we have that the treatment does not have an effect as follows:

$$50\% = \mathsf{P}(\mathit{Alive} \mid \mathit{Treated}) \;=\; \mathsf{P}(\mathit{Alive} \mid \mathit{Untreated}) = 50\% \tag{3}$$

Why? The reason for (1), (2) and (3) is in the fact that more than 50% of males survive, in general, whereas less than 50% of females survive, in general. This fact outbalances and effectively (exactly) erases the individual effects of the treatment in the groups of males and females. We have that:

$$\begin{aligned} \mathsf{P}_{\mathit{Male}}(\mathit{Alive}) &= 60\% \qquad (4) \\ \mathsf{P}_{\mathit{Female}}(\mathit{Alive}) &\approx 43.8\% \qquad (5) \end{aligned}$$

Each triple of the form (1), (2) and (3) is said to form an instance of Simpson's paradox.

We can easily construct more extreme instances of Simpson's paradox, i.e., in which the treatment effect is not only erased but even reversed, for example, Table 3, which yields the following facts:

$$\begin{aligned} 48.1\% \approx \mathsf{P}(\mathit{Alive} \mid \mathit{Treated}) &< \mathsf{P}(\mathit{Alive} \mid \mathit{Untreated}) \approx 51.9\% \qquad (6) \\ 76.9\% \approx \mathsf{P}_{\mathit{Male}}(\mathit{Alive} \mid \mathit{Treated}) &>> \mathsf{P}_{\mathit{Male}}(\mathit{Alive} \mid \mathit{Untreated}) \approx 14.3\% \qquad (7) \\ 48.1\% \approx \mathsf{P}_{\mathit{Female}}(\mathit{Alive} \mid \mathit{Treated}) &>> \mathsf{P}_{\mathit{Female}}(\mathit{Alive} \mid \mathit{Untreated}) = 20\% \qquad (8) \end{aligned}$$

With the data in Table 3, the treatment effect has been reversed in the whole population (negative) as opposed to the gender strata (positive), although the significance of the positive treatment effect has increased significantly in both strata.

Table 3: More extreme data set as compared to the original data of Simpon's paper (N=52). The perceived treatment effect is not only erased in the whole population but even reversed as compared to the strata of males and females; also, the treatment effect in both strata is way more significant.

	Male		*Female*	
	Alive	Dead	Alive	Dead
Treated	10	3	13	14
Untreated	1	6	1	4

Interestingly, the high significance of the positive treatment effect for males (measured, e.g., as lift [58] as used in association rule mining [58, 59], i.e., as the quotient of success rates of treated males and untreated males: $5.4 \approx 76.9\%/14.3\%$), can still be over-outbalanced by the relatively less significant treatment effect for women (lift: $2.4 \approx 48.1\%/20\%$), which can be explained by the fact that in our concrete population, we have relatively more women than men:

$$61.5\% \approx \mathsf{P}(\mathit{Female}) \quad > \quad \mathsf{P}(\mathit{Male}) \approx 38.5\% \tag{9}$$

Again, each triple of the form (6), (7) and (8) is said to form instances of Simpson's paradox.

We summarize the discussion of Sect. 3.2 in Def. 1.

Definition 1 (Simpson's paradox (basic case of $2 \times 2 \times 2$ contingency table))**.** Any triple of events Y, X, C (called *target variable*, *impact factor*, *confounder*) is called an *instance of Simpson's paradox*, given that the following holds:

$$\begin{align} \mathsf{P}(Y|X) &\;\leq (\geq)\;\; \mathsf{P}(Y|\overline{X}) \tag{10}\\ \mathsf{P}_C(Y|X) &\;> (<)\;\; \mathsf{P}_C(Y|\overline{X}) \tag{11}\\ \mathsf{P}_{\overline{C}}(Y|X) &\;> (<)\;\; \mathsf{P}_{\overline{C}}(Y|\overline{X}) \tag{12}\end{align}$$

3.2.2. Generalizing Simpson's paradox: the case of $2 \times 2 \times n$ *contingency tables*

In [60], which is considered a standard example of Simpson's paradox in the literature, the paradox shows in a data set that generalizes the confounding variable to arbitrary categorical data, resulting into a $2 \times 2 \times n$ contingency table. The data is about student admission to the University of Berkeley. The target value is the event of admission (vs. non-admission) of a student to a study program, the impact factor is gender (male vs. female), and the confounding variable is the University department (out of n possible departments).

Given an event Y as target variable, an event X as impact factor, and a categorical random variable $C : \Omega \longrightarrow \mathfrak{C} = \{c_1, \ldots, c_n\}$ as confounding variable, we say that a Simpson's paradox is perceived in cases, where

$$\mathsf{P}(Y|X) \leq (\geq)\mathsf{P}(Y|\overline{X}), \tag{13}$$

however

$$\mathsf{P}_{C=c_i}(Y|X) > (<)\mathsf{P}_{C=c_i}(Y|\overline{X}) \tag{14}$$

for a *sufficiently large* number of categories $c_i \in \mathfrak{C}$. Here, the notion of Simpson's paradox becomes more fuzzy. It is not clear what a sufficiently large number of categories would be. In the extreme case, we would require that the trend changes in each of the categories. Otherwise, we could define a threshold of categories (that would be typically larger than at least 50%). It becomes clear that the notion of Simpson's paradox is more a perception and – as such – it might not be exactly definable anymore. The whole situation rather longs for the definition of continuous, gradual measures of degrees of impact.

3.2.3. Generalizing Simpson's paradox: the case of continuous target variables

The notion of Simpson's paradox can be generalized to examples with continuous target variables. In [61], a data set of salaries (continuous target variable), and two categorical factors, i.e., salaries and field of jobs, has been constructed – see Table (4). Salaries and jobs can be considered interchangeably as either an impact factor or a confounder. The data set can be considered as revealing Simpson's paradox as follows. The average salary in Seattle is the highest among all cities of the data set[1]. However, with respect to each of the five fields of jobs (IT, law, commerce, medicine, education), individually, people earn the least (on average) in Seattle. How comes? Again, this is not a contradiction. It just means that in Seattle, relatively more people are working in one of the high-paid job fields.

We consider an instance of Simpson paradox, such as expressed in the data set of Table 4 as an extreme form of confounding. Ultimately we are interested in adjusting impact factors for the influence of confounders as well as the development of impact measures on the basis of such adjustments. In [61], two of such measures (one for adjustment, one for measuring the effect of adjusted impact) have been introduced.

[1] Again, note that the data set is completely artificial

Table 4: Artificial example data set with a continuous target variable (salary), and two factors (cities, field jobs) that can interchangeably serve as either impact factor or confounder [61].

	IT	*Law*	*Commerce*	*Medicine*	*Education*	**Avg. Salary**
Seattle	7.000	3.700	3.600	3.500	3.400	5.100
Boston	7.400	3.900	3.800	3.700	3.600	3.900
Tucson	7.300	3.800	3.700	3.600	3.500	3.900
Washington	7.300	3.800	3.700	3.600	3.500	3.900
Philadelphia	7.300	3.800	3.700	3.600	3.500	3.900
Avg. Salary	7.149	3.797	3.707	3.616	3.504	3.900

4. Startification of Pearson correlation

Paradoxes can also manifest in correlations, leading to misleading or counterintuitive results. In 1899, Pearson [56] illustrated that marginal and partial associations between continuous variables could diverge, leading to the emergence of spurious correlations. These spurious associations can easily be identified by analyzing the trend line, conducting correlation analysis, and inspecting scatter plots for each group and the overall relationship observed in the aggregate data.

To identify confounding variables in categorical data, the relationship between two variables in each group is compared to the aggregate association across all groups. This can be done using multiple ways, e.g., logistic regression, chi-squared test for independence and two-way tables. To better comprehend the relationship between variables in each group, bar graphs or mosaic plots can be used to understand the relationship between variables. In continuous values, confounding variables can be identified by examining the trend line, correlation analysis and visual inspection of scatter plots between each group and comparing it to the overall relationship in the aggregate data. It can easily be accomplished by fitting a regression model to each group's data and then comparing the slopes and intercepts of these models.

In Fig. 2, we utilize a synthetic dataset to compute the covariance between the two variables (X, Y) to demonstrate the marginal and partial associations. The scatter plots between X and Y display different signs within three sub-populations $A, BandC$ and across the entire population. The Pearson correlation between two sets of data can be measured via Eq. 15. Here x and y represent the input vectors, while $\bar{x}$ and $\bar{y}$ are the means of the corresponding variables. The value of r lies between -1 to 1, values greater than 0 indicate a positive correlation, the value 1 represents a perfect positive correlation and value 0 indicates no correlation. Negative values less than 0 suggest a negative correlation, and the value of -1 implies a clear negative association.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}} \tag{15}$$

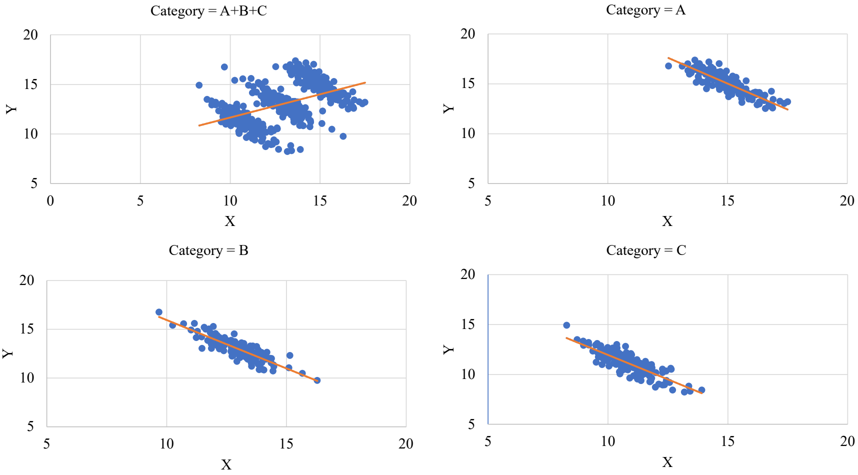


Figure 2: Pearson correlations for two variables (X, Y) demonstrating the opposite trends for the marginal and partial associations

5. On the adjustment of the impact of confounders

The adjustment of the impact of confounders is claimed to be handled in various ways. Donald Rubin et al. [51, 62] proposed propensity score weighting in estimating the causal effect. Judea Pearl developed a framework for reasoning about causal relationships, called causal inference [63, 64]. The regression model proposed by Ronald Fisher et al. [65] is also used for controlling the causal effects. Similarly, there are several other methods, e.g., stratification and meta-analysis, which are used to adjust the impact of confounders [66].

In line with the proposed methods, In Sect. 5.1, we discuss some familiar adjustments in detail and provide a discussion on coerced categorical adjustment with numerical target variables.

5.1. Familiar Adjustments

In [67], Judea Pearl stated: *"If we do have measurements of the third variable, then it is very easy to deconfound the true and spurious effects. For instance, if the confounding variable Z is age, we compare the treatment and control groups in every age group separately. We can then take an average of the effects, weighting each age group according to its percentage in the target population. This method of compensation is familiar to all statisticians; it is called »adjusting for Z« or »controlling for Z.«"* [67]

When we assume that the effect in this statement of Judea Pearl's is binary, we have that the described scenario is the case of $2 \times 2 \times n$ contingency tables as discussed in Sect. 3.2.2.

Law of total probabilities. Given a random variable $z : \Omega \longrightarrow \mathbb{R}$ and a partition $p_1, \ldots, p_n$ of Ω, we have that

$$\mathsf{P}(z) = \sum_i \mathsf{P}(p_i)\mathsf{P}(z|p_i) \tag{16}$$

(16) is called *law of total probabilities.* Given a random variable

$$y : \Omega \longrightarrow \{v_1, \ldots, v_n\},$$

we have that $(y = v_1), \ldots, (y = v_n)$ forms a partition of Ω. In terms of y, the law of total probabilities, therefore, shows as:

$$\mathsf{P}(z) = \sum_i \mathsf{P}(y = v_i)\mathsf{P}(z|y = v_i) \tag{17}$$

Together with a further event x, we have that

$$\mathsf{P}_x(z) = \sum_i \mathsf{P}_x(y = v_i)\mathsf{P}_x(z|y = v_i) \tag{18}$$

Due to the fact that $\mathsf{P}_x(a|b) = \mathsf{P}(a|x, b)$ for any events a and b, we can rewrite (18) as

$$\mathsf{P}(z|x) = \sum_i \mathsf{P}(y = v_i|x)\mathsf{P}(z|x, y = v_i) \tag{19}$$

Given an event $z \subseteq \Omega$, called *target variable (or dependent variable)*, an event x, called *impacting variable (or impact factor)*, and a categorical random variable $y : \Omega \longrightarrow \{v_1, \ldots, v_n\}$, called the *confounder (or confounding variable, or confounding factor)*, we define the *adjustment of the conditional probability* $\mathsf{P}(z|x)$ *to the (impact of) the confounder* y, denoted by $\widehat{\mathsf{P}^y}(z|x)$, as follows:

$$\widehat{\mathsf{P}^y}(z|x) = \sum_{i \leq n} \mathsf{P}(y{=}v_i)\mathsf{P}_x(z|x, y{=}v_i) \tag{20}$$

Again, note that the adjustment of $\mathsf{P}(z|x)$ is achieved by the *coercion* of $\mathsf{P}(y = v_i|x)$ in (19) to the value $\mathsf{P}(y = v_i)$ in (20). This coercion is the only (but crucial) difference between the probability $\mathsf{P}(z|x)$ and the adjusted probability $\widehat{\mathsf{P}^y}(z|x)$.

5.2. Coerced Categorical Adjustment with Numerical Target Variable

Law of total expectations. Given a random variable $z : \Omega \longrightarrow \mathbb{R}$ and a partition $p_1, \ldots, p_n$ of Ω, we have that

$$\mathsf{E}(z) = \sum_i \mathsf{P}(p_i)\mathsf{E}(z|p_i) \tag{21}$$

(21) is called *law of total expectations.* Given a random variable $y : \Omega \longrightarrow \{v_1, \ldots, v_n\}$, we have that $(y = v_1), \ldots, (y = v_n)$ forms a partition of Ω. In terms of y, the law of total expectation, therefore, shows as:

$$\mathsf{E}(z) = \sum_i \mathsf{P}(y = v_i)\mathsf{E}(z|y = v_i) \tag{22}$$

Together with a further event x, we have that

$$\mathsf{E}_x(z) = \sum_i \mathsf{P}_x(y = v_i)\mathsf{E}_x(z|y = v_i) \tag{23}$$

Due to the fact that $\mathsf{P}_x(a|b) = \mathsf{P}(a|x, b)$ for any events a and b, we can rewrite (23) as

$$\mathsf{E}(z|x) = \sum_i \mathsf{P}(y = v_i|x)\mathsf{E}_x(z|x, y = v_i) \tag{24}$$

Given a numerical random variable $z : \Omega \longrightarrow \mathbb{R}$, called *target variable (or dependent variable)*, an event x, called *impacting variable (or impact factor)*, and a categorical random variable $y : \Omega \longrightarrow \{v_1, \ldots, v_n\}$, called the *confounder (or confounding variable, or confounding factor)*, we define the *adjustment of the conditional expectation* $\mathsf{E}(z|x)$ *to the (impact of) the confounder* y, denoted by $\widehat{\mathsf{E}^y}(z|x)$, as follows:

$$\widehat{\mathsf{E}^y}(z|x) = \sum_{i \leq n} \mathsf{P}(y{=}v_i)\mathsf{E}_x(z|x, y{=}v_i) \tag{25}$$

Again, note that the adjustment of $\mathsf{E}(z|x)$ is achieved by the *coercion* of $\mathsf{P}(y = v_i|x)$ in (24) to the value $\mathsf{P}(y = v_i)$ in (25). This coercion is the only (but crucial) difference between the expectation $\mathsf{E}(z|x)$ and the adjusted expectation $\widehat{\mathsf{E}^y}(z|x)$.

Often, a conditional expectation $\mathsf{E}(z|x)$ is denoted as μ_x, when z can be assumed as granted from the context. In accordance with that, we also denote an adjusted expectation $\widehat{\mathsf{E}^y}(z|x)$ as $\widehat{\mu}_x^y$ or even shorter as $\widehat{\mu}_x$, when z (and y) are known from the context.

Accordingly, given a numerical random variable $z : \Omega \longrightarrow \mathbb{R}$, an event x, and a series $\vec{y}$ of confounding categorical random variables $y_1 : \Omega \longrightarrow I_1$ to $y_m : \Omega \longrightarrow I_m$, we first define the multivariate random variable $y : \Omega \longrightarrow I_1 \times \ldots \times I_m$ as usual, i.e., $\mathsf{P}(y = \langle v_1, \ldots, v_m \rangle) = \mathsf{P}(y_1 = v_1, \ldots, y_m = v_m)$; then, we define the adjustment $\widehat{\mathsf{E}^{\vec{y}}}(z|x)$ as follows:

$$\widehat{\mathsf{E}^{\vec{y}}}(z|x) = \widehat{\mathsf{E}^y}(z|x) \tag{26}$$

Note, that (25) is a generalization of (20) from a target event z to a numerical target random variable. In [61, 68], partial conditionalization [69] has been generalized from partial conditional probabilities to partial conditional expectations. Then, the adjusted expectation has been explained as a partial conditional expectation. Then, the quotient of the adjusted expectation $\widehat{\mathsf{E}^y}(z|x)$ and the marginal expectation $\mathsf{E}(z)$ has been called the *genuine impact of x (onto z).* In terms of association rule mining, the genuine impact could also be called *adjusted lift*, as the quotient $\mathsf{P}(z|x)/\mathsf{P}(x)$ is known as *lift* in association rule mining (for the specialized cases that z and x are events).

5.3. Inverse Propensity Score Weighting

The propensity score, denoted as $P(A|X)$, represents the conditional probability of receiving treatment A (exposure) given a set of observed covariates X. It can be estimated using a logistic regression model or other suitable models.

$$P(A = 1|X) = Pr(A = 1|X) \tag{27}$$

Inverse Propensity Score(IPS) weighting involves assigning weights to each observation based on the inverse of its estimated propensity score. The weight assigned to an individual i in the treatment group ($A = 1$) is given by $1/P(A = 1|X_i)$, and the weight assigned to an individual j in the control group ($A = 0$) is given by $1/(1 - P(A = 1|X_j))$.

$$\mathrm{W}_i = \frac{1}{P(A = 1|X_i)} \tag{28}$$

$$\mathrm{W}_j = \frac{1}{1 - P(A = 1|X_j)} \tag{29}$$

Once the inverse propensity scores are calculated, the weighted estimation of the treatment effect can be obtained by weighting the outcome variable Y by the inverse propensity scores. For example, the weighted average treatment effect (W) can be calculated as follows:

$$\mathrm{W} = \frac{\sum[Y_i \cdot w_i]}{\sum w_i} \tag{30}$$

Where Y_i is the outcome variable for individual i, w_i is the corresponding inverse propensity score weight, and the summation is performed over all individuals in the sample. By incorporating these probability equations into the IPS weighting framework, confounding effects can be adjusted to derive unbiased treatment effect estimates.

6. The proposed framework

Mitigating bias in DSTs requires a multi-faceted approach involving careful data collection, analysis, modelling, ongoing monitoring, and refinement. There are several frameworks [70, 71, 72, 73, 74] and best practices that can be used to mitigate bias in AI systems. Table 5 provides an overview of some popular bias mitigation frameworks together with their capabilities to utilise various bias mitigation techniques.

6.1. Issues with existing bias mitigation frameworks

As shown in Table 5, all of these frameworks provide valuable tools for mitigating bias in machine learning. However, most of them are developed to address specific types of bias and have limited capabilities to handle confounding effects and statistical paradoxes in classical DSTs, e.g. association rule mining. Therefore, to fully address these challenges, it is crucial to incorporate domain expertise and supplement these frameworks with additional capabilities to handle confounding effects and statistical paradoxes in various DSTs. The following are the main challenges with the current bias mitigation frameworks:

1. No framework perfectly identifies and adjusts the impact of different statistical paradoxes.
2. Current frameworks are often designed to address specific types of bias, such as gender or racial bias.

3. Bias mitigation frameworks are not developed for large and complex multidimensional datasets.
4. Current bias mitigation frameworks have limited capabilities to support classical DSTs effectively.

Table 5: Popular frameworks used to mitigate bias in AI systems as compared to our proposed framework.

Framework	Description	Bias Mitigation Techniques
Google What-If Tool[70]	Web-based tool for visualizing and analyzing machine learning models	Counterfactual analysis, What-If scenarios
Microsoft Fairlearn [72]	Python package for bias mitigation	Preprocessing, in-processing, and post-processing techniques
IBM AI Fairness 360[71]	Comprehensive toolkit for bias mitigation	Metrics for measuring bias, preprocessing, in-processing, and post-processing techniques
Themis-ML [73]	Python library for bias and fairness in machine learning	Fairness metrics, bias mitigation techniques for classifiers
Aequitas [74]	Python library for bias audit and mitigation	Fairness metrics, bias mitigation techniques for classifiers

Observing paradoxical outcomes and handling statistical paradoxes is challenging for a bias mitigation framework, especially when working with different DSTs with large and complex datasets. It requires combining technical expertise, domain knowledge, and careful consideration of trade-offs between fairness and accuracy. Therefore, improving current frameworks requires a continued focus on expanding their coverage, improving flexibility, promoting collaboration with domain experts, providing guidance on balancing trade-offs, and increasing transparency.

Thus, we suggest a framework to address these challenges in the current bias mitigation frameworks. A graphical presentation of the proposed framework is given in Fig. 3. The following are the three main and two sub-components of the proposed framework.

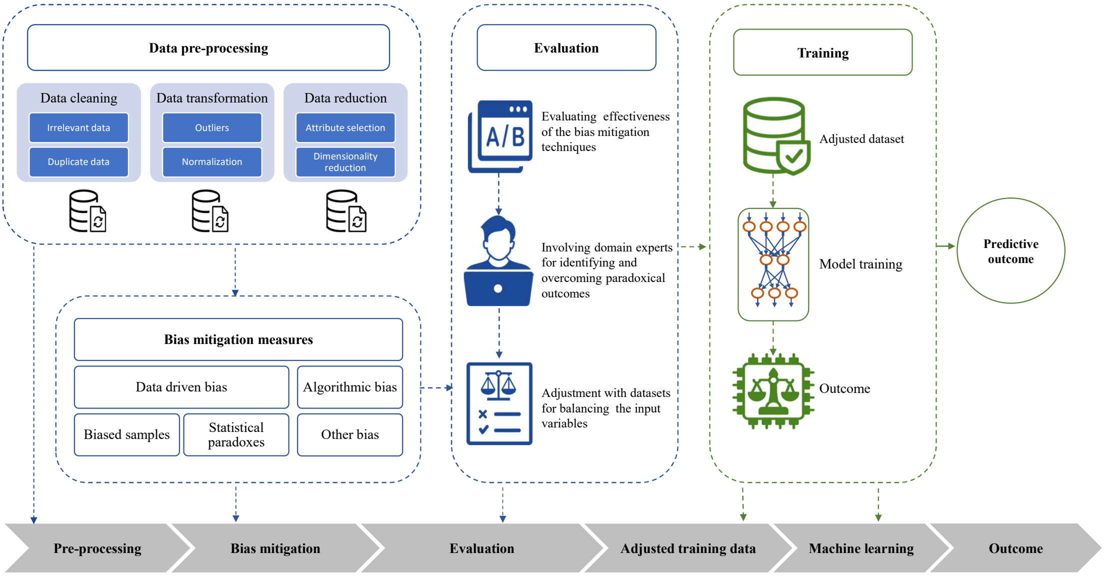


Figure 3: Framework for mitigating the impact of bias resulting from statistical paradoxes

1. Data pre-processing: This step involves identifying and addressing flaws or inconsistencies in the data. This encompasses various procedures like data cleansing, normalization, handling missing values, and outlier detection.
2. Bias mitigation techniques: To mitigate bias in the dataset, the second phase involves employing a range of bias mitigation techniques. This involves various data adjustment and augmentation methods to balance the representation of different underrepresented classes or categories in the dataset.
3. Evaluation: The evaluation aims to validate the effectiveness of bias reduction strategies in enhancing fairness and ensuring reliable outcomes. This can involve employing various metrics or comparing the outcome of DSTs on biased

and unbiased datasets.

(a) Domain knowledge integration: The step involves incorporating domain expertise into the data analysis process. This includes employing adjustment strategies and leveraging professional knowledge to guide the selection of relevant variables and features. Understanding the underlying causes of statistical paradoxes and implementing appropriate measures to mitigate their impact can further enhance the quality and utility of the dataset.

(b) Adjustments in datasets: uneven distribution of data between two or more groups is one of the reasons for bias. Therefore, by balancing the input variables across different groups in the data, an decision support tool is less likely to make biased decisions. Balancing the dataset ensures that a decision support tool is equally exposed to all groups.

The suggested framework is designed to show the following advantages:

- The framework specifically includes means to address bias resulting from undetected statistical paradoxes, which arise when statistical relationships between different groups in the data lead to unexpected and potentially biased results. This particular aspect has not yet been incorporated into existing bias mitigation frameworks.

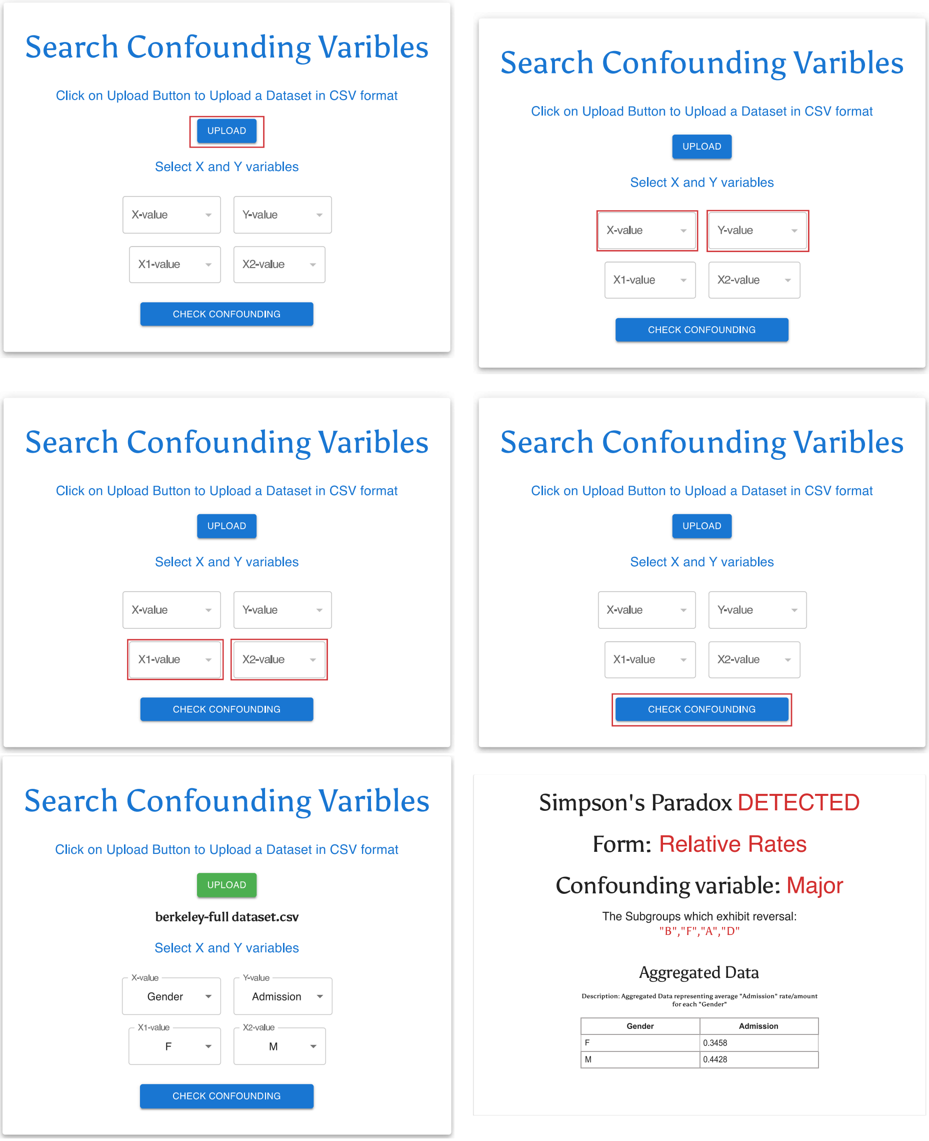
- Existing frameworks focus more heavily on one or a few specific strategies or techniques for mitigating bias. However, the proposed framework takes a comprehensive approach to support different types of DSTs and address different types of biases by supporting multiple bias mitigation techniques, evaluation, balancing, and incorporation of domain knowledge.

- The proposed framework strongly emphasises balancing and adjustment of variables in training datasets to minimize the risk of statistical paradoxes.

- The proposed framework emphasizes the importance of incorporating domain experts into developing and deploying AI systems. This helps in making informed decisions aligned with social and ethical values.

7. The Web-based Application

Based on the proposed framework, a web-based application is further developed to identify the impacts of confounding variables and to deal with the statistical paradoxes. Currently, the application systematically identifies the impact of confounding



Search Confounding Varibles

Click on Upload Button to Upload a Dataset in CSV format

Select X and Y variables

Search Confounding Varibles

Click on Upload Button to Upload a Dataset in CSV format

Select X and Y variables

Search Confounding Varibles

Click on Upload Button to Upload a Dataset in CSV format

Select X and Y variables

Search Confounding Varibles

Click on Upload Button to Upload a Dataset in CSV format

Select X and Y variables

Search Confounding Varibles

Click on Upload Button to Upload a Dataset in CSV format

berkeley-full dataset.csv

Select X and Y variables

X-value: Gender

Y-value: Admission

X1-value: F

X2-value: M

Simpson's Paradox DETECTED

Form: Relative Rates

Confounding variable: Major

The Subgroups which exhibit reversal:
"B","F","A","D"

Aggregated Data

Description: Aggregated Data representing average "Admission" rate/amount for each "Gender"

Gender	Admission
F	0.3458
M	0.4428

Figure 4: The application's graphical user interface offers a simple and straightforward experience. With just a few simple steps, users can upload a dataset, specify input parameters (X, Y), select the X1-Value and X2-Value variables, and identify confounding variables and instances of Simpson's paradox by clicking the "check confounding" button.

variables in categorical datasets. Further, it also detects the existence of Simpson's paradox within the dataset.

The tool has been developed using Python 3.8 programming language and the FastAPIframework, leveraging the benefits of its fast development and high-performance capabilities. The tool comprises two endpoints, "confounder" and "dropdown," which receive post requests containing corresponding parameters. These parameters are extracted from the user's form submission on the web interface and sent to the endpoint as requested. The backend service has been deployed on the *Deta platform*, providing a highly convenient solution for deploying microservices without needing server configuration or permission management. The deployment and management of the API are facilitated through the *Deta CLI tool*.

Figure 4 demonstrates the graphical user interface of the tool. The interface itself is simple and user-friendly. It requires just a few straightforward steps; importing a dataset, inputting parameters, and a few clicks to detect the presence of confounding variables and identify the instances of Simpson's paradox. The programming code and usage guidelines for the proposed tool can be found in the GitHub repository [2].

8. Experiments

To identify the existence of confounding variables in machine learning datasets, we conducted experiments on a range of both real-life and benchmark datasets that included both categorical and continuous values. In this article, we use four popular datasets, two of which are popular real-life case studies and the other two are benchmark datasets for machine learning. A piece of brief information about the datasets is given in Table 6, and their usage and results are discussed in Sects. 8.1, 8.2, 8.3 and 8.4.

Table 6: Information about the datasets used in the experimentation.

Dataset	Author	Year	Class
Iris Dataset	Ronald Fisher [75]	1936	Continuous
Auto MPG	Ross Quinlan [57]	1993	Continuous
UC Berkeley Admissions	PC Bicel [60]	1973	Categorical
Kidney Stone	Charig et al. [76]	1986	Categorical

[2] https://https://github.com/rahul-sharmaa/SimpsonP/

Table 7: Correlation analysis between variable X and variable Y with respect to a subgroups in Iris dataset

SubGroup	CatAttr	Variable X	Variable Y	Corr.	AggCorr.
Iris-Setosa	class	Sepal Width	Sepal Length	0.7467	-0.1093
Iris-versicolor	class	Sepal Width	Sepal Length	0.5259	-0.1093
Iris-virginica	class	Sepal Width	Sepal Length	0.4572	-0.1093
Iris-setosa	class	Petal Length	Sepal Width	0.1766	-0.4205
Iris-versicolor	class	Petal Length	Sepal Width	0.5605	-0.4205
Iris-virginica	class	Petal Length	Sepal Width	0.4010	-0.4205
Iris-setosa	class	Petal Width	Sepal Width	0.2799	-0.3565
Iris-versicolor	class	Petal Width	Sepal Width	0.6639	-0.3565
Iris-virginica	class	Petal Width	Sepal Width	0.5377	-0.3565

8.1. Iris dataset

The Iris dataset is one of the well-known benchmark datasets used in machine learning. Ronald Fisher introduced the dataset in a research paper [75]. It consists of three types of iris species, i.e., Setosa, Versicolor, and Virginicare, each with 50 data samples. The species names are categorical, and length and width are continuous attributes. To identify the existence of confounding variables and statistical paradoxes in the dataset, first, we check the type of variables and visualise the relationship between the length and width of each pair of candidate attributes.

In our experiment, we calculate the Pearson correlation between the *sepal length* and the *sepal width* variable and traverse the complete list of variables to identify the possible confounders and compute the ratio of the subgroup reversals to learn about the existence of Simpson's paradox.

In the experimentation with the iris dataset, confounding effects in three pairs of measurements have been reported, i.e., (1) sepal length and width, (2) sepal width and petal length, and (3) sepal width and petal width. Table 7 and Fig. 5 demonstrate these effects via three scatter plots with regression lines. Figure 5 illustrates that each species has a positive correlation between sepal width and length (dashed line). However, the correlation between the entire population's width and sepal length is negative (solid red trend line). Similarly, the pair of petal length, width, the pair of petal width and sepal width have positive trends for each species; however, the overall trend for the length and width for the entire population is negative in both cases.

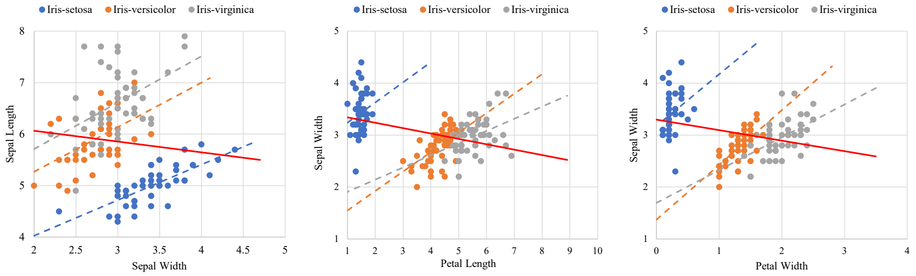


Figure 5: In the Iris dataset, there is a positive correlation between the three pairs of sepal length and petal width for the Iris-setosa, Iris-versicolor and Iris-virginicare (dashed lines); however, the overall trend for the length and width for the entire population is negative (solid red line) in all three combinations.

Table 8: Correlation analysis between variable X and variable Y with respect to a subgroups in Auto-MPG dataset

SubGroup	CatAttr	Variable X	Variable Y	Corr.	AggCorr.
3	cylinders	mpg	acceleration	-0.8190	0.423
6	cylinders	mpg	acceleration	-0.3410	0.423
3	cylinders	mpg	horsepower	0.621	-0.778
6	cylinders	mpg	horsepower	0.013	-0.778
75	model-year	mpg	acceleration	-0.0510	0.423
79	model-year	mpg	acceleration	-0.0510	0.423

8.2. The MPG dataset

Ross Quinlan used the Auto MPG dataset in 1993 [57]. The dataset contains 398 automobile records from 1970 to 1982, including the vehicle's name, MPG, number of cylinders, horsepower, and weight. The dataset includes three multi-valued discrete attributes and five continuous attributes. In MPG datasets, as provided in Table 8, we analysed the relationship between MPG, acceleration and horsepower for two categorical attributes (number of cylinders and model year). The goal of analysing the dataset is to learn about the factors that influence each car's overall fuel consumption. The dataset consists of fuel consumption in mpg, horsepower, number of cylinders, displacement, weight, and acceleration.

Similar to the Iris dataset, in the MPG dataset, the tool has reported confounding effects in the three pairs of measurements, i.e., (1) MPG with acceleration according to the engine cylinders, (2) MPG with acceleration with respect to their model year, and (3) MPG with horsepower according to the engine cylinders. To demonstrate

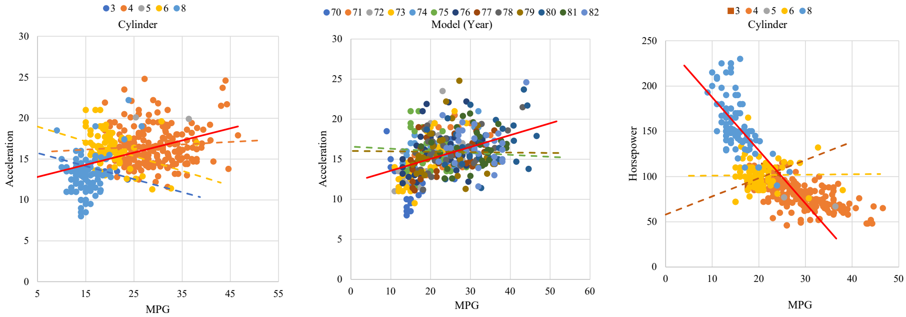


Figure 6: Scatter plot with trend lines for MPG dataset: In this dataset, Simpson's paradox has been observed in three sets of measurements; first, MPG and acceleration based on engine cylinders, second, MPG and acceleration based on model year, and third, MPG and horsepower based on engine cylinders

these effects, a scatter plot with regression lines is also given for all three pairs of measurements in Fig. 6. This Figure demonstrates a negative correlation between MPG and acceleration for three cylinders engines and six cylinders engines; however, the overall trend between MPG and acceleration is positive (solid red line). Similarly, the overall trend is the opposite for MPG with acceleration with respect to the model year and MPG with horsepower according to the engine cylinders.

8.3. UC Berkeley Admissions Dataset Fall 1973

UC Berkeley admissions dataset is a categorical dataset; It is a classic example of demonstrating confounding effects and explaining Simpson's paradox. The dataset was provided by UC Berkeley researchers to investigate any possible cases of gender bias in admissions. The dataset contains 12763 records with four attributes *Student_id, Gender, Major, Admission.*

As per the aggregate data given in Table 9 and demonstrated by the bar chart given in Fig. 7, the overall number of women applicants is significantly less than the total men applicants. However, their rejection rate is high compared to the male applicants. Statistically, it clearly shows significant bias in the admission percentage toward the male gender when we look at the data by gender. However, on the other end, adding a third variable in the analysis reversed the results in most of the departments. Fig. 8 and Table 10 display disaggregated data for each department and demonstrates the percentage of admissions by gender and department: Here, the information conditioned by the departments demonstrates the reason for bias, and

Table 9: UC Berkeley admission dataset fall 1973: Aggregate information for both men and women applicants. The overall number of women applicants is significantly less than the total men applicants. However, their rejection rate is high compared to male applicants; this indicates a significant bias towards Male applicants

	Applications	Admitted	Rejected	Admission %
Men	8442	3738	4704	44%
Women	4321	1494	2827	35%

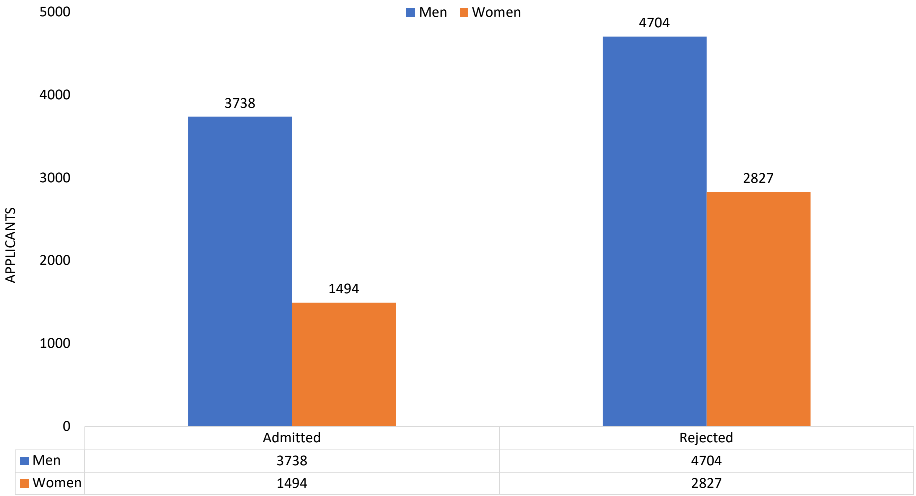


Figure 7: UC Berkeley admission dataset fall 1973: This figure demonstrates that the overall number of women applicants is significantly less than the total men applicants. However, their rejection rate is high compared to the male applicants; this indicates a significant bias towards Male applicants

it reveals an opposite story and bias in favour of Female applicants. Notably, in our analysis, we observed that female applicants tend to apply more for very selective majors while males for the less selective ones, creating an unbalanced distribution of males and females in applicants in the departments.

The dataset has been experimented with the web-based tool. In the tool, *Gender*

Table 10: UC-Berkeley admission dataset (fall 1973): % of acceptance rate for both *men* and *women* applicants in different departments

Gender	Departments					
	A	B	C	D	E	F
Men	72.49%	63.03%	36.92%	33.09%	27.74%	5.89%
Women	82.40%	68%	33.89%	34.93%	23.91%	7.33%

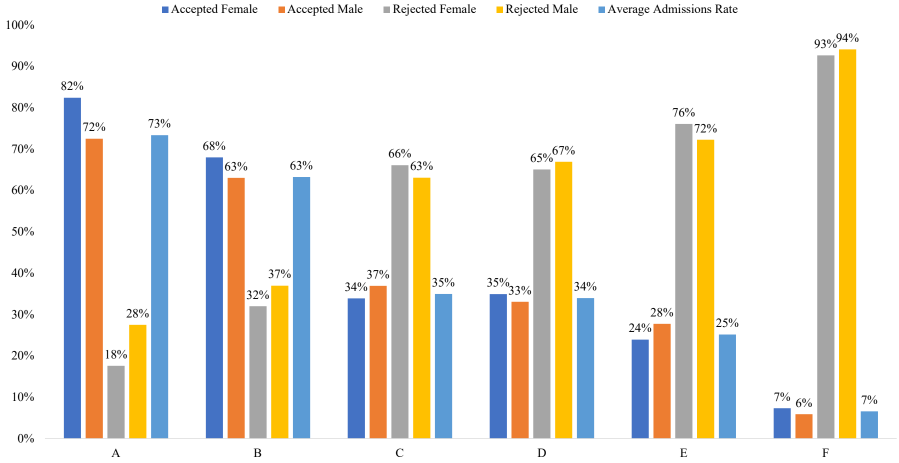


Figure 8: This figure shows the percentage of admissions by gender and department, along with the percentage of accepted and rejected female and male applicants in each department. Analysis of the data reveals that the admission rate for female applicants was generally higher than that of male applicants in most departments

attribute is set as X variable and *Admission* attribute is set as Y variable. Next, in the prepossessing step, the values of *gender* variable, i.e., *Female* and *Male* are categorised by the binary values 1 and 0, similarly, the values of *admission* variable, i.e., *Failure* and *Success* are categorised by the binary values 0 and 1, respectively.

The application first calculates the Pearson correlation between *Gender* and *Admission* variables and traverses the complete list of variables to identify the possible confounding variable and compute the ratio of the subgroup reversals to know about Simpson's paradox. The computed correlation coefficient between *Gender* and *Ad-*

mission variables indicate a negative correlation for majors *A, B, D, and F* but a positive correlation for the entire population.

8.4. Kidney Stone Treatment Dataset

We use another dataset with categorical values from a medical case study published by Charig et al. [76] in “The British Medical Journal” in 1986. This study compares the success rate of two different types of treatments to remove large and small kidney stones.

In Fig. 9, the distribution of both of the treatments is given for small and large kidney stones. At first, as demonstrated in Table 11, for both small kidney stones and large kidney stones groups, treatment A, performs better than treatment B; however, when the data for both treatments are combined, the treatment B (*Success Rate:* 83%) outperforms the treatment A (*Success Rate:* 78%).

Upon experimenting with this dataset using the web-based tool, the tool reported the presence of a confounding variable and a potential case of Simpson’s paradox.

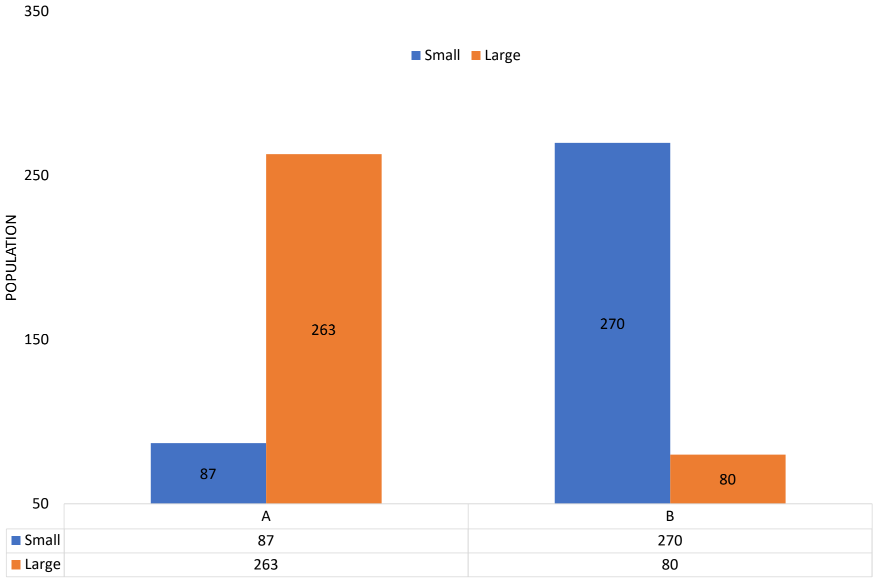
Table 11: Kidney stone treatment dataset: Treatment A outperforms treatment B for *large* and *small* kidney stones, but for both kidney stones together, treatment B exceeds treatment A

	Treatment (A)= 350			**Treatment (B)** = 350		
Stone Size	Success (S)	Failure (F)	Success Rate %	Success (S)	Failure (F)	Success Rate %
Small	81	6	$\approx$ 93%	234	36	$\approx$ 87%
Large	192	71	$\approx$ 73%	55	25	$\approx$ 69%
Both	273	77	$\approx$ 78%	289	61	$\approx$ 83%

9. Discussion

9.1. Relevance and Implications

In this paper, we presented that handling statistical paradoxes is a significant challenge towards developing fair and reliable AI applications. As discussed in Sect. 3, the existence of statistical paradoxes in benchmark datasets provides a direction to understand the vital role of confounders and statistical paradoxes in AI systems. To develop capabilities to handle statistical paradoxes in AI frameworks, it is important to take a systematic approach and incorporate diverse techniques for identifying potential sources of bias, utilising causal inference techniques to account for confounding variables and promoting transparency and open communication throughout the



	A	B
🟦 Small	87	270
🟧 Large	263	80

Figure 9: Kidney stone treatment dataset: The distribution of treatments A and B for *small* and *large* kidney stones demonstrates that treatment A was mostly given to patients with *large* kidney stones, and treatment B was mostly given to patients with *small* kidney stones

data analysis process. By integrating these components into a flexible and adaptable framework, researchers can mitigate the impact of statistical paradoxes and ensure that their results accurately reflect the true relationship between variables.

In this paper, an adaptable, multifaceted framework is presented. Additionally, an instance of the proposed framework is developed by implementing a sample web-based tool. The tool is evaluated by a series of experiments using several real and synthetic data sets. In its current use, the tool can identify and adjust the impacts of possible confounders in categorical and continuous datasets. Further, the tool also identifies the instances of Simpson's paradox and provides adjusted observations to reduce the impacts of the paradox. A series of experiments validated the framework and highlighted the importance of human experts in improving the accuracy of AI systems. The experimental results, in general, suggest that a thorough understanding of statistical concepts and paradoxes is necessary to alleviate the severe impacts of statistical paradoxes effectively and to address several other paradoxes successfully.

9.2. Limitations and future work

The proposed framework offers a comprehensive approach for mitigating bias due to statistical paradoxes. However, the complex nature of statistical paradoxes and related concepts poses a challenge in developing effective mitigation strategies. Currently, the web-based tool developed using the proposed framework is highly effective in identifying confounding variables and detecting instances of Simpson's paradox only in categorical and continuous datasets. However, further research and a deeper understanding of statistical concepts are necessary to handle several other statistical paradoxes. Moreover, mitigating statistical paradoxes in complex and high-dimensional datasets, where the relationships between variables can be highly non-linear and interactive, is another significant challenge. Therefore, developing a bias mitigation framework that addresses several statistical paradoxes is still challenging. To address these challenges, interdisciplinary research efforts are required that bring together experts in statistics, machine learning and social sciences. Overcoming these challenges can lead to the development of trustworthy AI systems that not only provide accurate and efficient decision-making but also promote fairness, transparency, and accountability, thereby enhancing their trustworthiness and societal impact.

10. Conclusion

This paper demonstrated the importance of addressing statistical paradoxes in DSTs and aimed to contribute towards developing fair and trustworthy DSTs. A framework has been suggested for mitigating the impacts of statistical paradoxes in DSTs. Furthermore, different measures for adjusting the impact of confounders are discussed. Based on the discussed measures, a web-based application has also been developed to validate the effectiveness and usefulness of the proposed framework. The application, in its current state, allows for investigating possible confounders via detecting instances of Simpson's paradox and provides a feature for adjusted observations. To provide evidence for the relevance of the framework and the application towards developing fair and trustworthy DSTs, a range of experiments have been conducted on real-world and benchmark datasets. The application serves as a valuable artefact for data scientists and researchers in their theoretical and practical endeavours. The potential of the application is not limited to its current use, for example, we plan to extend it to address various other paradoxical challenges in AI applications in the future.

References

[1] R. Binns, Algorithmic Accountability and Public Reason, Philosophy & Technology 31 (4) (2018) 543–556.

[2] J. C. Bjerring, J. Busch, Artificial Intelligence and Patient-Centered Decision-Making, Philosophy & Technology 34 (2) (2021) 349–371.

[3] K. Crawford, R. Calo, There is a blind spot in AI research, Nature 538 (7625) (2016) 311–313.

[4] P. B. de Laat, Algorithmic decision-making based on machine learning from big data: Can transparency restore accountability?, Philosophy & Technology 31 (4) (2018) 525–541.

[5] D. Helbing, B. S. Frey, G. Gigerenzer, E. Hafen, M. Hagner, Y. Hofstetter, J. Van Den Hoven, R. V. Zicari, A. Zwitter, Will democracy survive big data and artificial intelligence?, Springer, 2019.

[6] C. O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, first edition Edition, Crown, New York, 2016.

[7] A. Julia, L. Jeff, K. Lauren, M. Surya, Machine Bias – propublica.org, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, [Accessed 06-Feb-2023] (2016).

[8] M. Bertl, P. Ross, D. Draheim, A survey on AI and decision support systems in psychiatry – uncovering a dilemma, Expert Systems with Applications 202 (117464) (2022) 1–14.

[9] W. Hämäläinen, G. I. Webb, Specious rules: an efficient and effective unifying method for removing misleading and uninformative patterns in association rule mining, pp. 309–317.

[10] M. Gentzel, Biased face recognition technology used by government: A problem for liberal democracy, Philosophy and Technology 34 (4) (2021) 1639–1663.

[11] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Conference on fairness, accountability and transparency, PMLR, 2018, pp. 77–91.

[12] R. Sharma, M. Kaushik, S. A. Peious, A. Bazin, S. A. Shah, I. Fister, S. B. Yahia, D. Draheim, A novel framework for unification of association rule mining, online analytical processing and statistical reasoning, IEEE Access 10 (2022) 12792–12813.

[13] B. van Giffen, D. Herhausen, T. Fahse, Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods, Journal of Business Research 144 (2022) 93–106.

[14] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, A. Tzovara, Addressing bias in big data and ai for health care: A call for open science, Patterns 2 (10) (2021) 100347.

[15] G. U. Yule, Notes on the theory of association of attributes in statistics, Biometrika 2 (2) (1903) 121–134.

[16] J. Pearl, Causal inference without counterfactuals: Comment, Journal of the American Statistical Association 95 (450) (2000) 428–431.

[17] J. Berkson, Limitations of the application of fourfold table analysis to hospital data, Biometrics Bulletin 2 (3) (1946) 47–53.

[18] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[19] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297.

[20] E. H. Simpson, The interpretation of interaction in contingency tables, Journal of the Royal Statistical Society: Series B (Methodological) 13 (2) (1951) 238–241.

[21] J. Berkson, Limitations of the application of fourfold table analysis to hospital data, Biometrics Bulletin 2 (3) (1946) 47–53.

[22] F. M. Lord, A paradox in the interpretation of group comparisons., Psychological bulletin 68 5 (1967) 304–5.

[23] D. Kahneman, A. Tversky, On the psychology of prediction, Psychological Review 80 (1973) 237–251.

[24] M. Pia Sormani, The will rogers phenomenon: the effect of different diagnostic criteria, Journal of the Neurological Sciences 287 (2009) S46–S49, supplement title: Recent advances in multiple sclerosis: challenging paradigms.

[25] T. Afonja, Accuracy paradox, https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b (Dec 2017).

[26] D. Braess, Über ein Paradoxon aus der Verkehrsplanung, Unternehmensforschung 12 (1) (1968) 258–268.

[27] R. Otte, Probabilistic Causality and Simpson's Paradox, Philosophy of Science 52 (1) (1985) 110–125.

[28] J. Pearl, Simpson's paradox: An anatomy, UCLA Cognitive Systems Laboratory, Technical Report.

[29] M. Schield, Simpson's paradox and cornfield's conditions, ASA Proceedings of the Section on Statistical Education 1999 (1999) 106–111.

[30] B. A. Spellman, C. M. Price, J. M. Logan, How two causes are different from one: The use of (un)conditional information in Simpson's paradox, Memory & Cognition 29 (2) (2001) 193–208.

[31] M. Schaller, Sample size, aggregation, and statistical reasoning in social inference, Journal of Experimental Social Psychology 28 (1) (1992) 65–85.

[32] A. P. Dawid, Conditional independence in statistical theory, Journal of the Royal Statistical Society. Series B (Methodological) 41 (1) (1979) 1–31.

[33] N. Cartwright, Causal Laws and Effective Strategies, Noûs 13 (4) (1979) 419.

[34] K. Fiedler, Beware of samples! A cognitive-ecological sampling approach to judgment biases., Psychological Review 107 (4) (2000) 659–676.

[35] K. Fiedler, E. Walther, P. Freytag, S. Nickel, Inductive Reasoning and Judgment Interference: Experiments on Simpson's Paradox, Personality and Social Psychology Bulletin 29 (1) (2003) 14–27.

[36] K. Fiedler, The ultimate sampling dilemma in experience-based decision making., Journal of Experimental Psychology: Learning, Memory, and Cognition 34 (1) (2008) 186–203.

[37] K. Fiedler, P. Freytag, T. Meiser, Pseudocontingencies: An integrative account of an intriguing cognitive illusion., Psychological Review 116 (1) (2009) 187–206.

[38] R. Kievit, W. Frankenhuis, L. Waldorp, D. Borsboom, Simpson's paradox in psychological science: a practical guide, Frontiers in Psychology 4 (2013) 513.

[39] N. Alipourfard, P. G. Fennell, K. Lerman, Can you trust the trend? discovering simpson's paradoxes in social data, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 19–27.

[40] N. Alipourfard, P. G. Fennell, K. Lerman, Using simpson's paradox to discover interesting patterns in behavioral data, in: Proceedings of the Twelfth International AAAI Conference on Web and Social Media, AAAI Publications, 2018, pp. 2–11.

[41] A. A. Freitas, On objective measures of rule surprisingness, in: European Symposium on Principles of Data Mining and Knowledge Discovery, Springer, 1998, pp. 1–9.

[42] C. R. Blyth, On Simpson's paradox and the sure-thing principle, Journal of the American Statistical Association 67 (338) (1972) 364–366.

[43] S. P. Curley, G. J. Browne, Normative and Descriptive Analyses of Simpson's Paradox in Decision Making, Organizational Behavior and Human Decision Processes 84 (2) (2001) 308–333.

[44] S. Greenland, Simpson's paradox from adding constants in contingency tables as an example of bayesian non collapsibility, The American Statistician 64 (4) (2010) 340–344.

[45] M. A. Hernán, D. Clayton, N. Keiding, The Simpson's paradox unraveled, International Journal of Epidemiology 40 (3) (2011) 780–785.

[46] Y.-K. Tu, D. Gunnell, M. S. Gilthorpe, Simpson's paradox, lord's paradox, and suppression effects are the same phenomenon–the reversal paradox, Emerging themes in epidemiology 5 (1) (2008) 1–9.

[47] Amazon scraps secret AI recruiting tool that showed bias against women, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (Oct. 2018).

[48] R. A. Fisher, III. The influence of rainfall on the yield of wheat at Rothamsted, Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character 213 (402-410) (1925) 89–142.

[49] R. B. Cattell, P-technique factorization and the determination of individual dynamic structure., Journal of Clinical Psychology 8 (1) (1952) 5–10.

[50] A. J. Conger, A revised definition for suppressor variables: A guide to their identification and interpretation, Educational and psychological measurement 34 (1) (1974) 35–46.

[51] P. R. Rosenbaum, D. B. Rubin, The central role of the propensity score in observational studies for causal effects, Biometrika 70 (1) (1983) 41–55.

[52] J. Pearl, Understanding simpson's paradox, SSRN Electronic Journal 68.

[53] D. P. MacKinnon, A. J. Fairchild, M. S. Fritz, Mediation analysis, Annual Review of Psychology 58 (1) (2007) 593–614.

[54] W. S. Robinson, Ecological correlations and the behavior of individuals, American Sociological Review 15 (3) (1950) 351–357.

[55] G. King, M. Roberts, Ei: a (n r) program for ecological inference, Harvard University.

[56] L. A. Pearson Karl, B.-M. Leslie, Genetic (reproductive) selection: Inheritance of fertility in man, and of fecundity in thoroughbred racehorses, Philosophical Transactions of the Royal Society of London: Series A 192 (1899) 257–330.

[57] J. Quinlan, Combining instance-based and model-based learning, in: Machine Learning Proceedings 1993, Elsevier, 1993, pp. 236–243.

[58] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: Proceedings of VLDB'1994 – the 20th International Conference on Very Large Data Bases, Morgan Kaufmann, 1994, p. 487–499.

[59] R. Srikant, R. Agrawal, Mining quantitative association rules in large relational tables, in: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, 1996, pp. 1–12.

[60] P. J. Bickel, E. A. Hammel, J. W. O'Connell, Sex bias in graduate admissions: Data from berkeley, Science 187 (4175) (1975) 398–404.

[61] D. Draheim, DEXA'2019 Keynote presentation: Future perspectives of association rule mining based on partial conditionalization (2019), doi:10.13140/rg.2.2.17763.48163.

[62] D. B. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies., Journal of educational Psychology 66 (5) (1974) 688.

[63] J. Pearl, Causal diagrams for empirical research, Biometrika 82 (4) (1995) 669–688.

[64] J. Pearl, Understanding simpson's paradox, SSRN Electronic Journal 68.

[65] J. Aldrich, Fisher and Regression, Statistical Science 20 (4) (2005) 401 – 417.

[66] L. Li, K. Kleinman, M. W. Gillman, A comparison of confounding adjustment methods with an application to early life determinants of childhood obesity, Journal of developmental origins of health and disease 5 (6) (2014) 435–447.

[67] J. Pearl, The Book of Why, Basic Books, New York, 2018.

[68] D. Draheim, Future perspectives of association rule mining based on partial conditionalization, in: Proceedings of DEXA'2019 – the 30th International Conference on Database and Expert Systems Applications, no. 11706 in LNCS, Springer, Berlin, Heidelberg, 2019, p. xvi.

[69] D. Draheim, Generalized Jeffrey Conditionalization – A Frequentist Semantics of Partial Conditionalization, Springer, Heidelberg New York Berlin, 2017.

[70] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson, The what-if tool: Interactive probing of machine learning models, IEEE Transactions on Visualization and Computer Graphics 26 (1) (2020) 56–65.

[71] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias, IBM Journal of Research and Development 63 (4/5) (2019) 4:1–4:15.

[72] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, K. Walker, Fairlearn: A toolkit for assessing and improving fairness in ai, Tech. Rep. MSR-TR-2020-32, Microsoft (May 2020).

[73] N. Bantilan, Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation (2017).

[74] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, R. Ghani, Aequitas: A bias and fairness audit toolkit (2019).

[75] R. A. Fisher, The use of multiple measurement in taxonomic problems, Annals of Eugenics 7 (2) (1936) 179–188.

[76] C. R. Charig, D. R. Webb, S. R. Payne, J. E. Wickham, Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy., BMJ 292 (6524) (1986) 879–882.

Curriculum Vitae

Personal Data

Name	Rahul Sharma
Date and place of birth	31 December 1986, Aligarh, U.P., India
Nationality	Indian

Contact Information

Address	School of Information Technologies, Tallinn University of Technology Akadeemia tee 15a, 12618 Tallinn, Estonia
E-mail	rahul.sharma@taltech.ee

Education

2019– . . .	Tallinn University of Technology, School of Information Technologies, PhD studies
2010–2012	M.Tech, Computer Science Engineering, Dr. A.P.J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh, India
2004–2008	B.Tech, Computer Science Engineering, Dr. A.P.J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh, India

Language Competence

Hindi	Native
English	Fluent

Professional Employment

2019–2023	Early Stage Researcher, Information Systems Group Department of Software Science, Tallinn University of Technology
2014–2019	Assistant Professor Ajay Kumar Garg Engineering College, Ghaziabad, India
2012–2014	Assistant Professor Raj Kumar Goel Institute of Technology for Women, Ghaziabad, India
2009–2010	Technical Specialist, HCL Technologies Pvt. Ltd. Noida, India
2008–2009	Technical Specialist, NG Softech Pvt. Ltd. Noida, India

Fields of Research[4]

- 4.6. Computer Science
- 4.7. Information and Communications Technologies

[4] Estonian Research Information System (ETIS) fields of research

Scientific Work

1. R. Sharma, M. Kaushik, S. A. Peious, A. Bazin, S. A. Shah, I. Fister, S. B. Yahia, and D. Draheim. A novel framework for unification of association rule mining, online analytical processing and statistical reasoning. *IEEE Access*, 10:12792–12813, 2022

2. R. Sharma. On statistical paradoxes and overcoming the impact of bias in expert systems: towards fair and trustworthy decision making. *SSRN*, pages 1–37, July 2023. doi:10.2139/ssrn.4506432

3. R. Sharma, M. Kaushik, S. A. Peious, S. B. Yahia, and D. Draheim. Expected vs. unexpected: Selecting right measures of interestingness. In M. Song, I.-Y. Song, G. Kotsis, A. M. Tjoa, and I. Khalil, editors, *Proceedings of DaWaK 2020 – the 22nd International Conference on Big Data Analytics and Knowledge Discovery*, pages 38–47, Cham, 2020. Springer International Publishing

4. R. Sharma, H. Garayev, M. Kaushik, S. A. Peious, P. Tiwari, and D. Draheim. Detecting Simpson's paradox: A machine learning perspective. In C. Strauss, A. Cuzzocrea, G. Kotsis, A. M. Tjoa, and I. Khalil, editors, *Proceedings of DEXA 2022 – the 33rd International Conference on Database and Expert Systems Applications*, pages 323–335, Cham, 2022. Springer International Publishing

5. R. Sharma, M. Kaushik, S. A. Peious, M. Bertl, A. Vidyarthi, A. Kumar, and D. Draheim. Detecting Simpson's paradox: A step towards fairness in machine learning. In S. Chiusano, T. Cerquitelli, R. Wrembel, K. Nørvåg, B. Catania, G. Vargas-Solar, and E. Zumpano, editors, *Proceedings of ADBIS 2022 – the 26th International Conference on New Trends in Database and Information Systems*, pages 67–76, Cham, 2022. Springer International Publishing

6. R. Sharma, M. Kaushik, S. A. Peious, M. Shahin, A. S. Yadav, and D. Draheim. Towards unification of statistical reasoning, OLAP and association rule mining: Semantics and pragmatics. In A. Bhattacharya, J. Lee Mong Li, D. Agrawal, P. K. Reddy, M. Mohania, A. Mondal, V. Goyal, and R. Uday Kiran, editors, *Proceedings of DASFAA 2022 – the 27th International Conference on Database Systems for Advanced Applications*, pages 596–603, Cham, 2022. Springer International Publishing

7. R. Sharma, M. Kaushik, S. A. Peious, M. Shahin, A. Vidyarthi, P. Tiwari, and D. Draheim. Why not to trust big data: Discussing statistical paradoxes. In U. K. Rage, V. Goyal, and P. K. Reddy, editors, *Proceedings of DASFAA 2022 International Workshops – the 27th International Conference on Database Systems for Advanced Applications*, pages 50–63, Cham, 2022. Springer International Publishing

8. R. Sharma, M. Kaushik, S. A. Peious, M. Shahin, A. Vidyarthi, and D. Draheim. Existence of the Yule-Simpson effect: An experiment with continuous data. In *Proceedings of Confluence 2022 – the 12th International Conference on Cloud Computing, Data Science & Engineering*, pages 351–355, 2022

9. D. Ghosh, M. Pandey, C. Gautam, A. Vidyarthi, R. Sharma, and D. Draheim. Utilizing continuous time markov chain for analyzing video-on-demand streaming in multimedia systems. *Expert Systems with Applications*, 223:119857, 2023

10. M. Kaushik, R. Sharma, S. A. Peious, M. Shahin, S. B. Yahia, and D. Draheim. A systematic assessment of numerical association rule mining methods. *SN Computer Science*, 2(5):1–13, 2021

11. M. Kaushik, R. Sharma, S. A. Peious, M. Shahin, S. Ben Yahia, and D. Draheim. On the potential of numerical association rule mining. In *Proceedings of FDSE'2020 – the 7th International Conference on Future Data and Security Engineering*, volume 12466 of *Lecture Notes in Computer Science*, pages 3–20. Springer Singapore, 2020

12. M. Kaushik, R. Sharma, M. Shahin, S. A. Peious, and D. Draheim. An analysis of human perception of partitions of numerical factor domains. In *Information Integration and Web Intelligence*, pages 137–144, Cham, 2022. Springer

13. M. Kaushik, R. Sharma, S. A. Peious, and D. Draheim. Impact-driven discretization of numerical factors: Case of two- and three-partitioning. In *Big Data Analytics*, pages 244–260, Cham, 2021. Springer International Publishing

14. M. Kaushik, R. Sharma, A. Vidyarthi, and D. Draheim. Discretizing numerical attributes: An analysis of human perceptions. In *New Trends in Database and Information Systems*, pages 188–197, Cham, 2022. Springer

15. M. Kaushik, R. Sharma, I. F. Jr.2, and D. Draheim. Numerical association rule mining: A systematic literature review, 2023

16. M. Kaushik. Swarm-intelligence algorithms for mining numerical association rules: An exhaustive multi-aspect analysis of performance assessment data. *SSRN Electronic Journal*, 2023

17. S. A. Peious, R. Sharma, M. Kaushik, S. A. Shah, and S. B. Yahia. Grand reports: a tool for generalizing association rule mining to numeric target values. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 28–37. Springer, 2020

18. A. Vidyarthi, R. Agarwal, D. Gupta, R. Sharma, D. Draheim, and P. Tiwari. Machine learning assisted methodology for multiclass classification of malignant brain tumors. *IEEE Access*, 10:50624–50640, 2022

19. M. Shahin, S. Arakkal Peious, R. Sharma, M. Kaushik, S. Ben Yahia, S. A. Shah, and D. Draheim. Big data analytics in association rule mining: A systematic literature review. In *International Conference on Big Data Engineering and Technology (BDET)*, page 40–49. Association for Computing Machinery, 2021

20. M. Shahin, S. Saeidi, S. A. Shah, M. Kaushik, R. Sharma, S. A. Peious, and D. Draheim. Cluster-based association rule mining for an intersection accident dataset. In *2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, pages 1–6, 2021

21. M. Shahin, M. R. Heidari Iman, M. Kaushik, R. Sharma, T. Ghasempouri, and D. Draheim. Exploring factors in a crossroad dataset using cluster-based association rule mining. *Procedia Computer Science*, 201:231–238, 2022. The 13th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 5th International Conference on Emerging Data and Industry 4.0 (EDI40)

Elulookirjeldus

1. Isikuandmed

Nimi	Rahul Sharma
Sünniaeg ja -koht	31. detsembril 1986, Delhi, India
Kodakondsus	Indian

2. Kontaktandmed

Aadress	Tallinna Tehnikaülikool, Infotehnoloogia teaduskond, Tarkvarateaduse instituut, Infosüsteemide rühm, Akadeemia tee 15a, 12618 Tallinn, Estonia
E-post	rahul.sharma@taltech.ee

3. Haridus

2019– . . .	Tallinna Tehnikaülikool, Infotehnoloogia teaduskond, doktoriõpe
2010–2012	M.Tech, arvutiteaduse tehnika, Dr. A.P.J. Abdul Kalami tehnikaülikool, Lucknow, Uttar Pradesh, India
2004–2008	B.Tech, Computer Science Engineering, Dr. A.P.J. Abdul Kalami tehnikaülikool, Lucknow, Uttar Pradesh, India

4. Keelteoskus

hindi keel	emakeel
inglise keel	kõrgtase

5. Teenistuskäik

2019–2023	Infosüsteemide uurimisrühma nooremteadur Tallinna Tehnikaülikooli tarkvarateaduse osakond
2014–2019	Abiprofessor Ajay Kumar Garg Engineering College, Ghaziabad, India
2012–2014	Abiprofessor Raj Kumar Goeli Naiste Tehnoloogiainstituut, Ghaziabad, India
2009–2010	Tehniline spetsialist, HCL Technologies Pvt. Ltd. Noida, India
2008–2009	Tehniline spetsialist, NG Softtech Pvt. Ltd. Noida, India

6. Teadustöö põhisuunad[5]

- 4.6. Arvutiteadused
- 4.7. Info- ja kommunikatsioonitehnoloogia

7. Teadustegevus

Teadusartiklite, konverentsiteeside ja konverentsiettekannete loetelu on toodud ingliskeelse elulookirjelduse juures.

[5] Eesti Teadusinfosüsteemi (ETIS) teadusvaldkondade ja -erialade klassifikaator

ISSN 2585-6901 (PDF)
ISBN 978-9916-80-056-0 (PDF)