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

Aleksei Netšunajev, 177031IAPM

Sentence Writing Test for Parkinson’s Disease Modeling: Comparing Predictive Ability of Classifiers

Master’s Thesis

Supervisors: Sven Nõmm, PhD
Aaro Toomela, PhD
Kadri Medijainen, MD

Tallinn 2019

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond
Tarkvarateaduse instituut

Aleksei Netšunajev, 177031IAPM

Parkinsoni tõve diagnoosimisel kasutatava lause kirjutamise testi modellerimine: klassifikaatorite ennustava võime võrdlus

Magistritöö

Juhendajad: Sven Nõmm, PhD
Aaro Toomela, PhD
Kadri Medijainen, MD

Tallinn 2019

Authors declaration of originality

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been presented for examination or submitted for defense anywhere else. All used materials, references to the literature and work of others have been cited.

Olen koostanud antud töö iseseisvalt. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on viidatud. Käesolevat tööd ei ole varem esitatud kaitsmisele kusagil mujal.

Author: Aleksei Netšunajev

May 7, 2019

Abstract

Primary goal of the thesis is to analyze digitally conducted sentence writing test that is used for Parkinson's disease diagnostics. To achieve the goal a set of features that have high predictive power is needed. The features should have interpretation and should be informative enough to be used in classifier models. The hypothesis formulated in the thesis states that features based on individual letters allow to set up a classifier with high predictive power.

The thesis extends the literature on the analysis of digital handwriting tests in several directions. First, an algorithm of letter extraction for the samples of the handwritten sentence is developed. It is worth noting that the approach taken in the thesis is language agnostic and is based on the properties of the data that are recorded while conducting digital test. Second, in the feature engineering phase interpretation to most important features is given. Finally, classification analysis that is based on selected features is performed.

For full sentence features describing whether the person is maintaining a straight line are calculated. Features that represent size, kinematics, duration and fluency of writing are calculated for each individual letter. The feature selection is performed based on Fisher's score. Sentence based features do not seem to be important based on the score. On the contrary, kinematic features are found to be of primary importance. There is no letter that would give rise to more than four features that are selected by Fisher's score into classification models meaning that all parts of the handwritten test are equally important for the analysis. A horse race of seven classifiers is performed for the set of features that have high predictive power. The random forest classifier performs the best with the accuracy of 88.5% on 3-fold cross-validation. All other classifiers tend to produce a considerable amount of false positive results.

The result of letter extraction, feature engineering and classification analysis confirm the hypothesis set up in the thesis. The achieved classification accuracy is on a similar level with the studies in this area. The hard task for classifiers seems to be tagging the healthy control subjects.

The present thesis is written in English and is 36 pages long, including 5 chapters, 14 figures and 11 tables.

Annotatsioon

Magistritöö põhieesmärk on analüüsida Parkinsoni tõve diagnostikas kasutatavat lause kirjutamise testi, mis on digitaalselt teostatud. Erinevalt varasemast kirjandusest analüüsitakse lauset tähthaaval. Eesmärgi saavutamiseks tuleb leida tunnused, mille abil on võimalik konstrueerida klassifikaatoreid kõrge ennustusvõimega. Tunnused peaksid olema tõlgendatavad ja peaksid olema piisavalt informatiivsed, et neid saaks kasutada statistilistes mudelites. Magistritöös püstitatud hüpoteesi kohaselt võib üksikute tähtede põhjal arvutada tunnused, mis võimaldavad luua kõrge prognoosimise võimega klassifikaatori.

Lõputöö laiendab digitaalsete meditsiiniliste käekirja testide uurimisega seotud kirjandust kolmes suunas. Esiteks töötatakse välja käsitsi kirjutatud lausest üksikute tähtede ekstraheerimise algoritm. Tasub märkida, et magistritöös kasutatud lähenemine ei ole keele spetsiifiline vaid põhineb testi sooritamise ajal salvestatud andmete omadustel. Teiseks, tunnuste loomise faasis tõlgendatakse kõige olulisemad nendest. Kolmandaks, valitud tunnuste põhjal hinnatakse klassifitseerimise mudelid.

Täislause põhjal arvutatakse tunnused, mis kirjeldavad, kas inimene säilitab kirjutamisel sirget joont. Iga tähe põhjal arvutatakse käekirja suuruse, kinemaatika, kestuse ja sujuvusega seotud tunnused. Tunnuste valik toimub Fisher skoori alusel. Täislausel põhinevad tunnused ei ole Fisher skoori järgi olulised. Kinemaatilised tunnused on aga esmatähtsad. Kõik käsitsi kirjutatud testi osad on analüüsi jaoks võrdselt olulised, kuna puudub täht, mille põhjal oleks arvutatud rohkem kui neli Fisher skoori poolt oluliseks peetud tunnust. Kõrge ennustusvõimega tunnuste põhjal viiakse läbi seitsme klassifikaatorite hindamine ja analüüs. Otsustusmets täpsusega $88,5\%$ ristvalideerimisel saavutab parima tulemuse. Kõik teised klassifikaatorid kipuvad tekitama märkimisväärse hulga vale positiivseid otsuseid.

Üksikute tähtede ekstraheerimine, nende põhjal tunnuste arvutamine ja klassifikaatorite võrdusanalüüs kinnitavad töös esitatud hüpoteesi. Parima mudeli täpsus on antud valdkonnas tehtud uuringutega samal tasemel. Tervete kontrollgrupi subjektide tuvastamine näib klassifikaatorite jaoks raskeks ülesandeks.

Magistritöö on kirjutatud inglise keeles ning sisaldab teksti 36 leheküljel, 5 peatükki, 14 joonist ja 11 tabelit.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Parkinson's disease (PD) is known to be a wide-spread neurodegenerative disorder. While the cure for it is not available at the moment, timely diagnose is very important as it allows to start treatment at an early stage.

The disease affects motions. The patients exhibit rigid and slow fine motor movements meaning that handwriting and drawing are affected at the first stage (Moustafa, Chakravarthy, Phillips, Gupta, Keri, Polner, Frank and Jahanshahi (2016)). For that reason a handwriting test is one of the diagnostic tools used by doctors (Nackaerts, Heremans, Smits-Engelsman, Broeder, Vandenberghe, Bergmans and Nieuwboer (2017)). Traditionally the test is conducted using pen and paper with the medical personnel evaluating the outcome.

During the recent years numerous possibilities have arisen to conduct the test by the means of a tablet and a digital pen. The adoption of the digital tests is very important as it allows to use the information that was not previously available for researchers and practitioners (see Thomas, Lenka and Kumar Pal (2017)). Various tests have been digitalized, for example Luria's alternating series test (Nõmm, Toomela, Kozhenkina and Toomsoo (2016)), clock drawing test (Nõmm, Masharov, Toomela and Medijainen (2018)) and various sentence writing tests (Smits, Tolonen, Cluitmans, Gils, Conway, Zietsma, Leenders and Maurits (2014)). Some of the tests were analyzed quantitatively, see for instance Nõmm et al. (2016), Stepien, Kawa, Wieczorek, Dabrowska, Slawek and E.J. (2019). These studies contribute to feature analysis and predictive modeling of such tests. Various papers show that it is possible to set up a classifier that discriminates reasonably well between healthy control subjects (HC) and PD patients based on fine motor drawing tests (Nõmm, Bardõs, Toomela, Medijainen and Taba, 2018).

In the present thesis specific type of the handwriting test is analyzed. The test is based written full sentence. The thesis uses the samples collected with the help of an iPad. The software for collecting the data was developed by Mašarov (2017).

Even though several studies analyzed digital data of fine motor movement tests, data originating from handwriting tests where a full sentence is written are not fully explored in the literature. For example, Smits et al. (2014) consider only a very limited amount of features and Drotar, Mekyska, Rektorov, Masarov, Smekal and Faundez-Zanuy (2016) is mainly interested in how new pressure measure improves the predictive ability. While it may be tempting to limit with words to extract features as in Toodo (2018) it is possible to extract individual letters and construct features based on the letters. This will provide more information for the model and may improve its predictive ability.

The discussion above allows to set up the research hypothesis:

Features that are extracted from the individual letters of the fine motor movements test where the person is asked to write a full sentence allow to estimate a classifier with strong predictive power.

To test the hypothesis the following steps are performed:

1. The samples of motor movement tests are processed and the individual words and letters are extracted.
2. Features are extracted from the full sentence and from individual letters.
3. Feature selection is performed.
4. Classifier models are trained based on the selected features.
5. Predictive power and decisions of the models are compared.

Figure 1.1 describes the tasks that are performed in the thesis.

The thesis extends the literature on the analysis of digital handwriting tests in three directions. First, an algorithm for letter extraction for the samples of the handwritten sentence is developed. I use own custom letter extraction system as standard approaches may not be useful due to specific handwriting of PD patients. Second, in the feature engineering phase interpretation to most important features is given. Finally, classification analysis that is based on selected features is performed. I find that kinematic features have the highest explanatory power in this task with the features that describe fluency of writing following. Random forest with a prediction accuracy of 88.5% is the model that outperform the other selected classifiers based on 3-fold cross-validation.

1.2 Related work

Handwriting of people affected by PD has been studied substantially. Several handwriting impairments are known to characterize the patients. One of them is *micrograhia* that

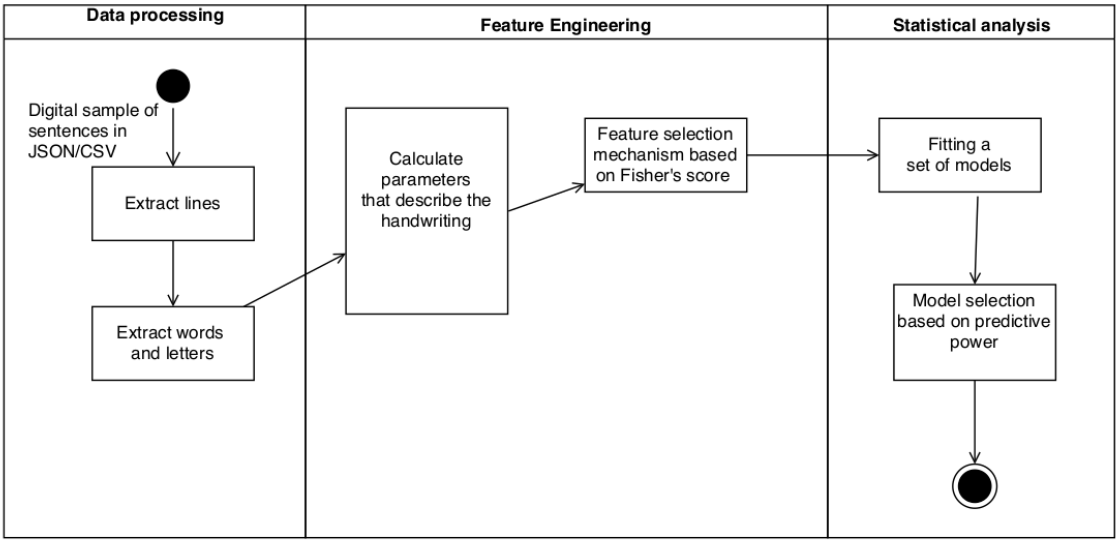

Data processing
Feature Engineering
Statistical analysis
Digital sample of sentences in JSON/CSV
Extract lines
Extract words and letters
Calculate parameters that describe the handwriting
Feature selection mechanism based on Fisher's score
Fitting a set of models
Model selection based on predictive power

Figure 1.1: Tasks completed in the thesis

is reductions in writing size. A reduction in letter size is fairly simple to detect with conventional paper-and-pen tools.

Even though Wagle Shukla, Ounpraseuth, Okun, Gray and Schwankhaus (2012) report micrographia to be found in nearly half of the PD cohort, its exact prevalence is not clear and may vary substantially. Van Gemmert, Hans-Leo and George (2001) argue that PD patients reduce the size of their handwriting strokes when concurrent processing load increases. Micrographia may result in consistent reduction in the size of letters as well as disability to keep the fixed size of letters for consecutive characters (Letanneux, Danna, Velay, Viallet and Pinto, 2014).

With the adoption of tablets with stylus new features describing the writing style of a person arise. For example iPad brings in altitude and azimuth angles of the Apple Pencil (see Nõmm et al. (2018)), that may give additional information on drawing ability. Thus researchers started questioning whether writing size is the most important predictor for PD. Over the time the focus has shifted from the analysis of only letter size to the analysis of a set of kinematic features of handwriting. Velocity and acceleration are the main kinematic properties studied in the literature.

Letanneux et al. (2014) document that kinematic features differentiate better between control participants and PD patients than the traditional measure of static writing size. Furthermore overview made by Thomas et al. (2017) convince that studies based on the kinematic analysis of handwriting have revealed that patients with PD may have abnormalities in velocity, fluency, and acceleration in addition to micrographia.

Letanneux et al. (2014) proposed the term PD dysgraphia, encompassing duration, velocity, and fluency in addition to size of handwriting to study graphomotor impairment in PD. The abnormalities in those characteristics of handwriting is now referred to as *dysgraphia* and a journey from micrographia to dysgraphia is a shift in paradigm.

Apart from kinematic features Drotar et al. (2016) proposes to utilize pressure measurements provided by the Wacom tablet, which were analyzed along with speed, duration and acceleration and showed significant discrimination power.

Jerk is the derivative of acceleration with respect to time. Although jerk may be regarded as a kinematic feature, it can also be a measure of fluency, because the value of jerk is sensitive to small changes in acceleration that affect the smoothness of writing (Thomas et al. (2017)). Other measures of fluency are the number of local minima and maxima for velocity and acceleration. The idea behind using the number of extrema for a single stroke lies in capturing smoothness of writing. The greater the number of extrema, the more disoriented the writing is.

To sum up, the studies of digitized handwriting of PD patients focus on the following features of the handwriting: size, kinematics, duration, fluency.

The current thesis is a continuation of research conducted at the Tallinn University of Technology. Research papers published by the staff include Nõmm, Toomela and Borushko (2013), Nõmm and Toomela (2013), Nõmm et al. (2016), Nõmm et al. (2018). While the earlier work is related to new measures of the smoothness of the human limb motions, the later ones discuss infrastructure for collecting digital data for various tests and statistical analysis of the digital tests.

The most important recent thesis written on the topic at the university are Kozhenkina (2016), Mašarov (2017), Bardõs (2018) and Toodo (2018). While the first two works are mostly related to digital infrastructure development for data collection, the latter two works analyze digital fine motor movement tests with quantitative methods.

Nõmm et al. (2018) exploits 34 observations and proposes a novel method to analyze Lurias alternating series patterns drawn during fine motor test. Main outcome of the work is a classifier model, capable of differentiating Parkinsons disease patients from healthy controls and providing prediction performance around 90%. Kinematic features proved to be important predictors for the PD.

Toodo (2018) possesses 26 observations and analyses the same test as is under investigation in the current thesis. He creates a word recognition system to determine and construct features based on words. Motion mass, kinematic and geometrical parameters of the sentence are extracted and analyzed. The results of this study indicate that there is a difference in the writing speeds between the PD patients and the healthy controls.

The biggest room for improvement in the work of Toodo (2018) is the handwriting recognition part. Current thesis extends the work of Toodo (2018) along that dimension. An algorithm for letter extraction for the samples of handwritten sentence is developed and its output is used in the feature engineering and statistical modeling phase.

Chapter 2

Data

2.1 Data description

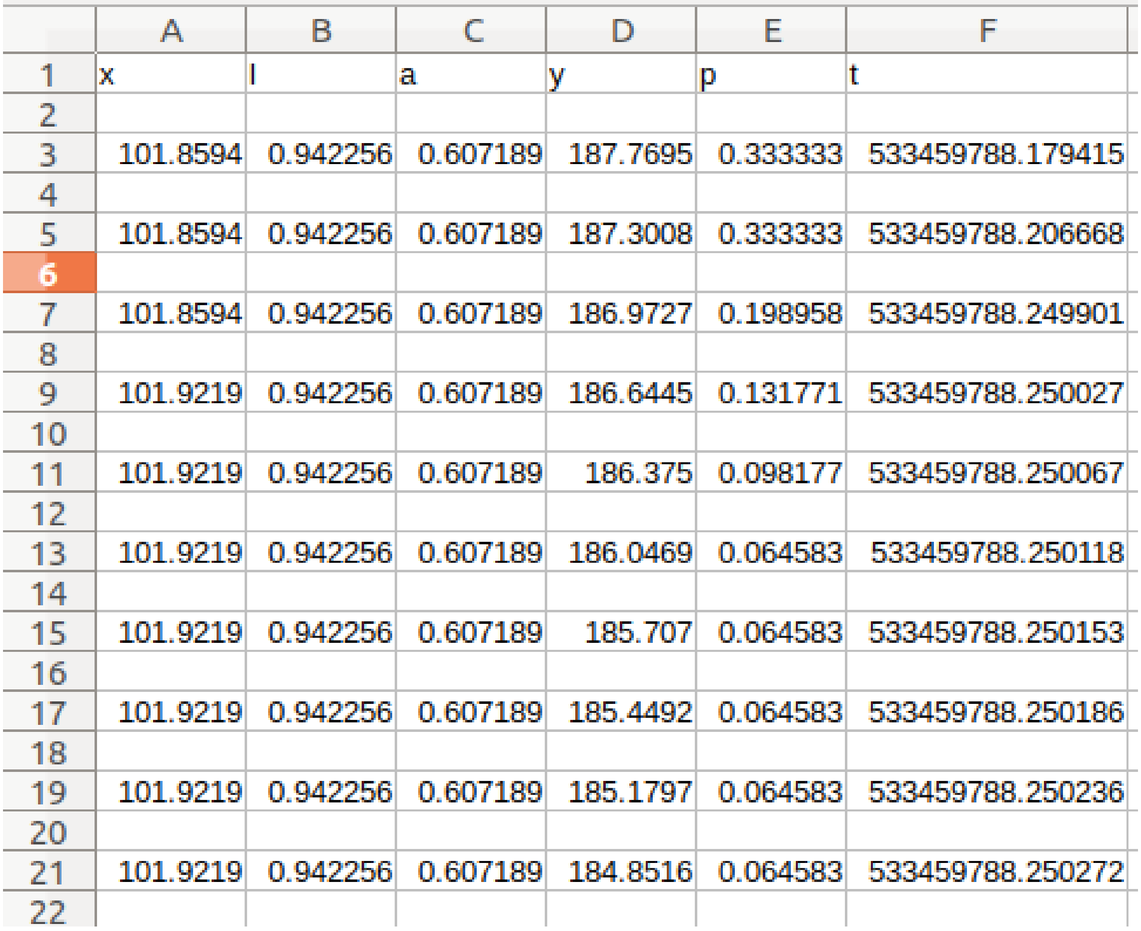
Digital version of the sentence writing test requires one to write given sentence on the screen of tablet PC using the stylus pen. The sentence is chosen form the literature that is usually taught during the first years of school education. This guarantees that all the subjects with the same native language know the sentence very well. In the frameworks of the present research subjects who's native language is Estonian were tested. The sentence reads "Kui Arno isaga koolimajja jõudsid, olid tunnid juba alanud" which means "When Arno with his father arrived to the school lessons have already started".

The data is collected using an iPad application that records the state of Apple Pencil at certain discretization level. The observations include floating point numbers that describe the way a person is performing the test. Acquired data is stored in the form of numeric array where the rows correspond to different time instances and columns to the data attributes. Figure 2.2 shows an example of the full sentence. The dataset consists of 11 PD patients and 8 healthy controls of approximately the same age of 68 years.

The sentence in Figure 2.2 is written on two lines. In the analyzed sample the sentence is written on three lines for most of the cases. That makes the number of lines an additional parameter to be estimated in the letter extraction phase. No cases of a sentence fitting on a single line are observed.

The sentence is recorded as an array of point that is saved in JSON or CVS format with predefined properties available for each recorded point. Table 2.1 present the properties that are used for the analysis and figure 2.1 displays an example of the raw data.

The most important feature of the data is that the points may be chronologically ordered and thus it is possible to calculate time changes for a sequence of points. By the same token, sorting the points based on time gives the sequence of points in the order the person has written them. This is very important feature of the data as it will be used to construct the letters extraction algorithm.



	A	B	C	D	E	F
1	x	l	a	y	p	t
2						
3	101.8594	0.942256	0.607189	187.7695	0.333333	533459788.179415
4						
5	101.8594	0.942256	0.607189	187.3008	0.333333	533459788.206668
6						
7	101.8594	0.942256	0.607189	186.9727	0.198958	533459788.249901
8						
9	101.9219	0.942256	0.607189	186.6445	0.131771	533459788.250027
10						
11	101.9219	0.942256	0.607189	186.375	0.098177	533459788.250067
12						
13	101.9219	0.942256	0.607189	186.0469	0.064583	533459788.250118
14						
15	101.9219	0.942256	0.607189	185.707	0.064583	533459788.250153
16						
17	101.9219	0.942256	0.607189	185.4492	0.064583	533459788.250186
18						
19	101.9219	0.942256	0.607189	185.1797	0.064583	533459788.250236
20						
21	101.9219	0.942256	0.607189	184.8516	0.064583	533459788.250272
22						

Figure 2.1: An example of the CSV file with data

2.2 Difficult cases

While the data is collected for subjects with PD it is quite likely that the recorded tests will be of different quality. The algorithm for letters extraction should take these peculiarities into account.

It may happen that the person is unable to maintain a straight line and the lines become tilted. The line may form either positive or negative angle with imaginary x-axis. Figure 2.3 presents an example of the sentence written in the tilted way.

One of the worst examples is shown on Figure 2.4. In this case the sentence is misspelled having ”kooli” instead of ”koolimajja”. The beginning of the second line is hardly readable even for a human. Some of the letters in the example are not readable, finally capital and small letters are used interchangeably. All these abnormalities make the processing of the sentences hardly feasible for standard algorithms.

It is important to notice that the sentence may have been written on various number of lines. While Figure 2.2 shows a two-line example, writing the sentence on three lines is also common in the analyzed sample.

Table 2.1: Data description

Property	Description
x	x-axis coordinate
y	y-axis coordinate
t	timestamp, start time is 00:00 UTC 01.01.2001
a	altitude angle
l	latitude angle
p	pressure

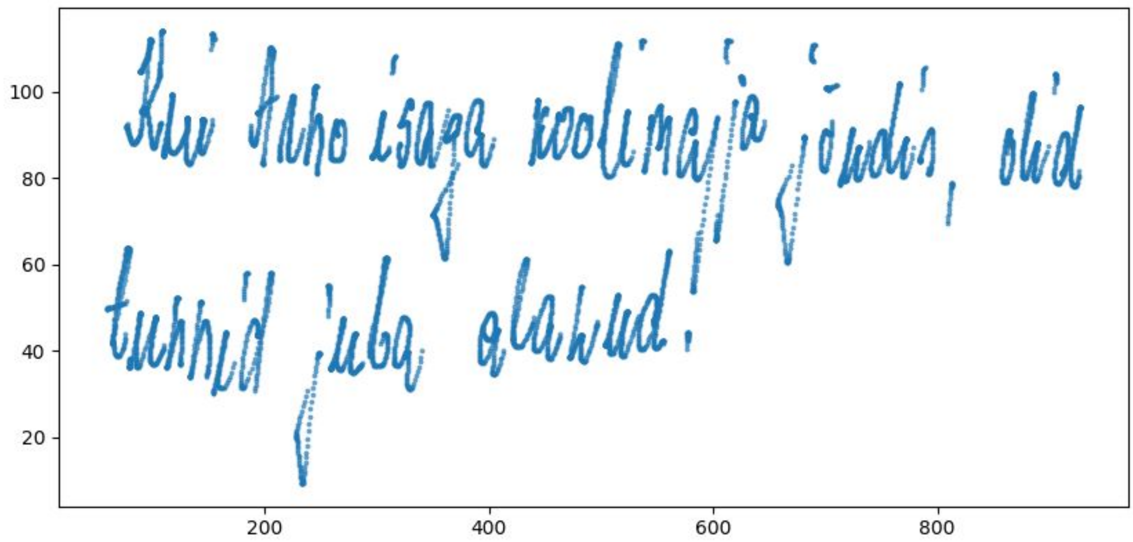


Figure 2.2: Full sentence shown as text

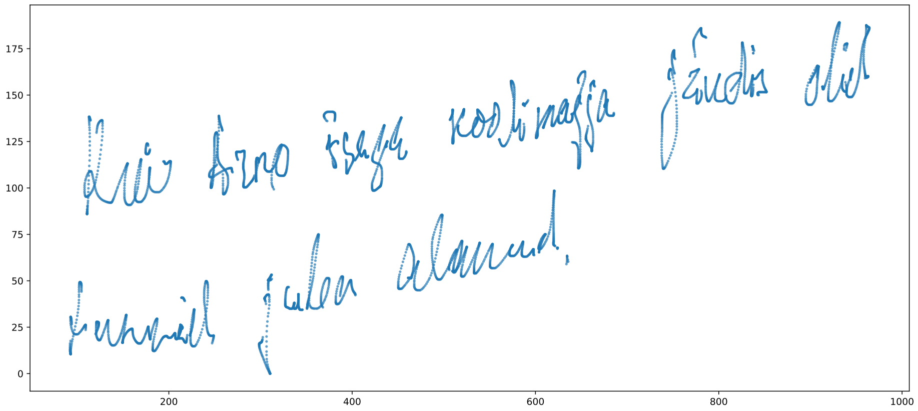


Figure 2.3: Full sentence written in a tilted way

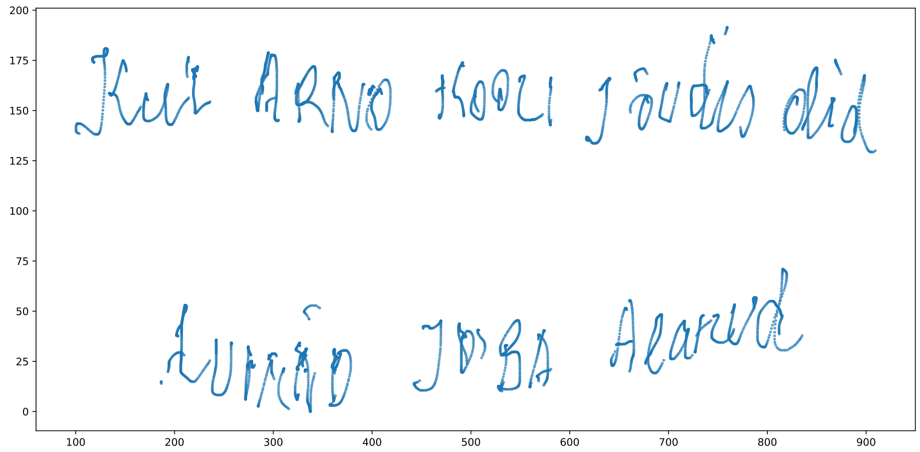


Figure 2.4: Full sentence with misspelled and missing words

Chapter 3

Handwriting recognition

3.1 Related work

Character recognition in freestyle handwritten documents has received a significant amount of attention in the past several decades. Various possible algorithms are proposed to extract individual characters.

The unsupervised clustering is a standard approach for this task. Common problems that may result in poor separation of characters are discussed in the literature. These are grouping of characters, overlapping of strokes, touch among characters, and non-linear separation boundaries between characters. All this leads to failure of standard clustering methods such as distance based k-means. Density based methods such as DBSCAN are not working reasonably well due to non-elliptic distribution of points that comprise a character. Probability based mixture model for character extraction may fail in the situations where characters are hardly separable which may often be the consequence of PD. Al-Dmour and Fraij (2004) conclude that standard clustering algorithms may be outperformed by a specialized one in word segmentation.

Tseng and Lee (1999) propose segmentation method for handwritten Chinese characters. Non-linear segmentation paths are initially located using a probabilistic Viterbi algorithm. Candidate segmentation paths are determined by verifying overlapping paths, between-character gaps, and adjacent-path distances. A segmentation graph is constructed using these candidate paths with the shortest path finally detected as the segmentation path.

Tan, Lai, Wang, Wang and Zuo (2012) first segment the entire text line into strokes, the similarity matrix of which is computed according to stroke gravities. Then, the nonlinear clustering methods are performed on this similarity matrix to obtain cluster labels for these strokes. According to the obtained cluster labels, the strokes are combined to form characters.

Most of the proposed techniques for word extraction in the literature consider a spatial measure of the gap between successive connected components, and define a threshold to classify within and between word gaps (Seni and Cohen (1994)).

Yamaguchi, Yoshikawa, Shinogi, Tsuruoka and Teramoto (2001) proposed a segmentation method for touching Japanese handwritten characters, which decreases over-segmentation by utilizing connecting condition of lines at the touching point. This method is effective when characters are linearly separable.

3.2 Algorithm for letter recognition

Clustering of individual letters is one of the most difficult part of the current task. Proposed solution uses the general idea of Tan et al. (2012) to tackle the problem. The algorithm processes the entire sentence, for example the one shown in Figure 2.2. Given a freestyle handwritten sentence segmentation of lines and individual characters has to be done. Various writing styles make both text line and single character segmentation challenging.

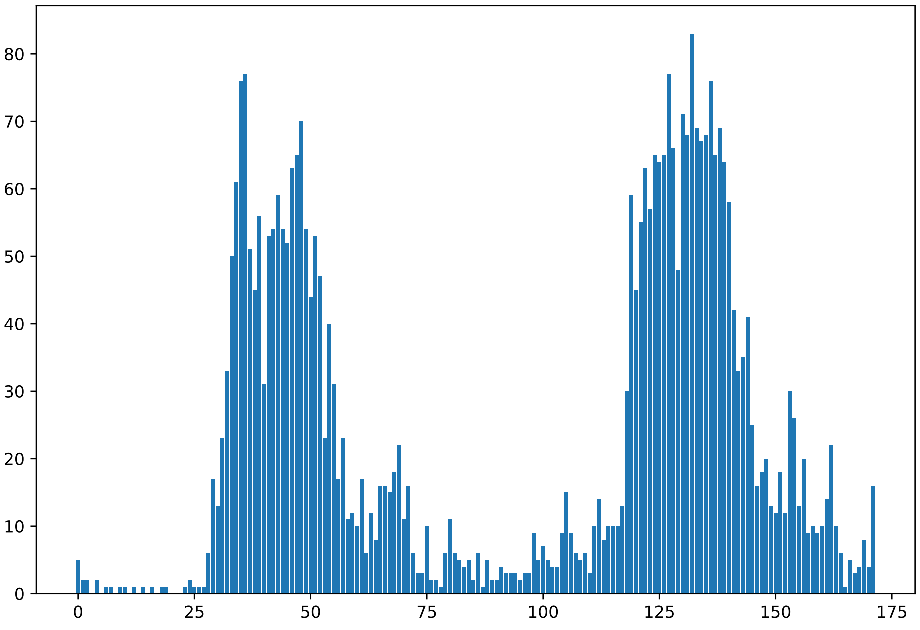


Figure 3.1: Histogram of the y-coordinates of a sentence

At first the number of lines has to be determined for the specific piece of handwriting. As discussed before some of sentences are written on two and some on three lines. To solve this problem a Gaussian mixture model is estimated for number of mixture components $K = 2, 3, 4$. It is worth noting that only y coordinates are used as observed variables in the mixture models. The idea of taking only single coordinate is based on the histogram of these coordinates. Figure 3.1 shows an example of the histogram for the y coordinates of a sentence. Two humps of points that may be modeled as Gaussians are clearly visible. The decision over K is made as an argument that maximizes mean silhouette score for given K.

Even separation boundary between the lines may be non-linear due to letters with long tails, such as g and j. The mixture model may not be taking it into account fully. To detect the parts not belonging to the line the pattern of time changes Δt, where t is time, for two consecutive lines is analyzed. If a situation where points do not appear consecutively in time is detected for a line, these abnormal points are shifted to the place where they belong to in the different line.

Processing of individual lines goes next. The separation of a line into smaller parts is based on the calculated time changes Δt of points for the whole line. An example of the series is shown on Figure 3.2. It becomes obvious that there are bigger chunks of points that are separated with longer time intervals. These larger chunks are groups of letters, sometimes full words, sometimes parts of words. This depends very much on the way how the person writes. Separating the points that belong to a common chunk applying a threshold on Δt makes up words for most lines. It is very important to pick the right separation threshold. With various candidates examined, the final value is chosen to be 20% of the maximum time change.

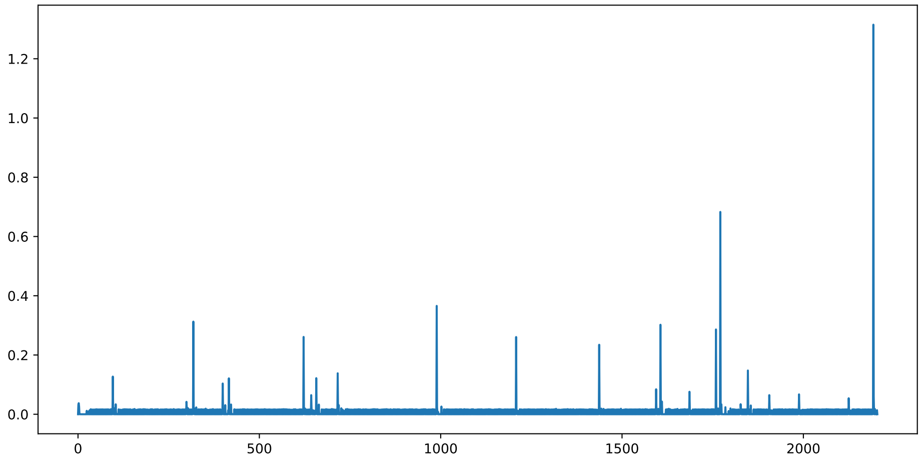


Figure 3.2: An example of the time change

Further the chunks of points extracted in the previous step are processed. These are separated into groups based on the distance of points. The idea is to capture the points where the writing person made a gap between points. This is very likely to be the gap between letters. Note that small clusters such as dots over letters i or j arise as a result of this procedure. For that reason clusters having less than 85 points are merged with the preceding cluster. An example of intermediate result is shown on Figure 3.3.

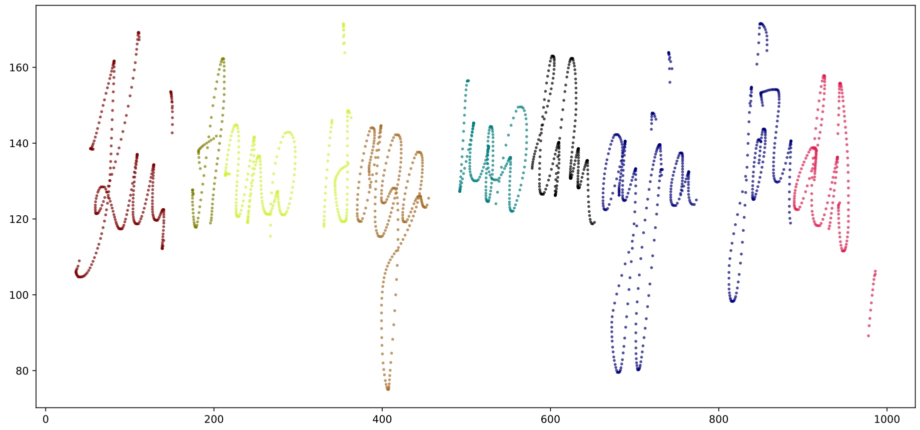


Figure 3.3: Intermediate result

It may be observed that the groups are rather big and mostly represent the sequences of letters that are written without removing the pen from the iPad. For that reason separation of large groups into several smaller ones is performed next. It is done based on the size of the group, from bigger groups more letters are extracted. An example of clusters obtained after this step is shown in Figure 3.4. This step differs for the sentences that are written on two and three lines. It turns out that sentences occupying two lines tend to have letters that are more narrow. Thus more letters should be extracted from a group that is obtained from such sentence if compared to a chunk of similar size derived from a three-lined sentence.

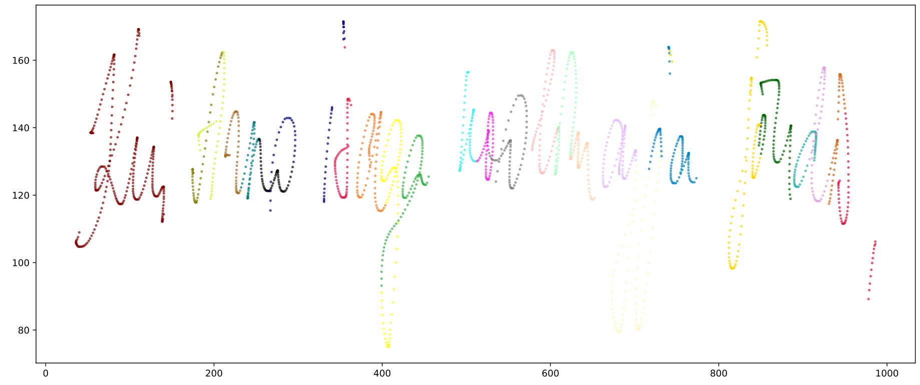
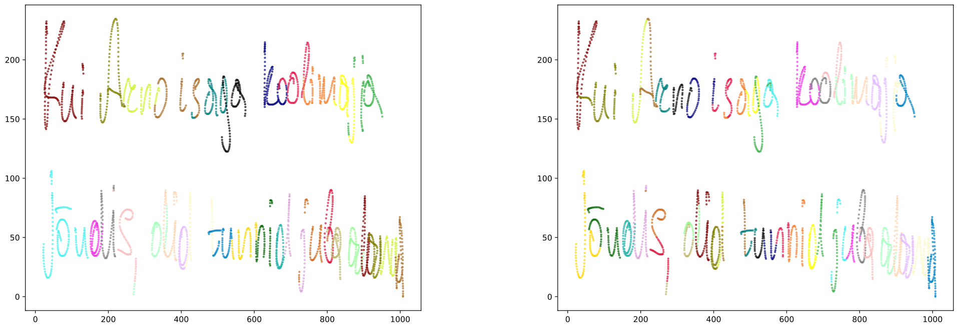


Figure 3.4: From larger clusters to smaller clusters

As the final step the number of clusters determined in the sentence should be optimized. As the features will be extracted from the clusters, those have to be consistent



(a) Sentence before achieving target number of clusters

(b) Sentence with target number of clusters

Figure 3.5: Example of the full sentence

through the analyzed sentences and the number of clusters determined in handwriting samples has to be constant. This target number of clusters in the full sentence is 48 and it is determined by the number of letters in the test sentence. Cases when the actual number of clusters is less than or larger than the target have to be considered separately. When the number of determined clusters is larger than it should be, the largest clusters get separated. In the opposite case smallest clusters get merged to the preceding clusters. An example of full sentence prior to having target number of clusters and with the target number of clusters is shown in Figure 3.5. The steps performed to extract the individual letters are illustrated by the Figure 3.6.

With the letters extracted from the handwriting samples feature extraction and statistical modeling may be performed next.

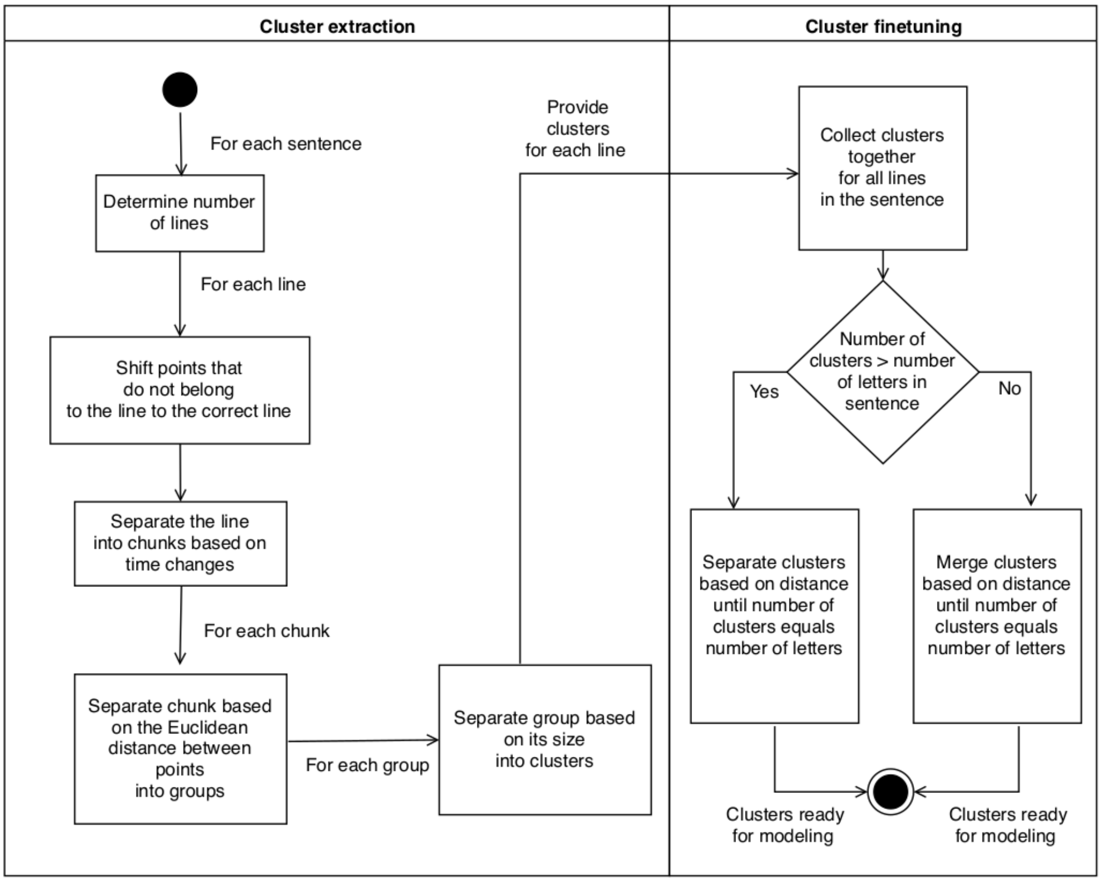

Cluster extraction
Cluster finetuning
For each sentence
Determine number of lines
For each line
Shift points that do not belong to the line to the correct line
Separate the line into chunks based on time changes
For each chunk
Separate chunk based on the Euclidean distance between points into groups
For each group
Separate group based on its size into clusters
Provide clusters for each line
Collect clusters together for all lines in the sentence
Number of clusters > number of letters in sentence
Yes
No
Separate clusters based on distance until number of clusters equals number of letters
Merge clusters based on distance until number of clusters equals number of letters
Clusters ready for modeling
Clusters ready for modeling

Figure 3.6: Steps of the clustering algorithms for letter extraction

Chapter 4

Statistical analysis

4.1 Feature engineering

In this subsection, the exhaustive list of features calculated for each sentence is described. Certain features are constructed for the sentence as a whole whereas the majority of features are derived from the individual letters. The overall feature set may be divided into three subsets. The first subset consists of features that are commonly used to conduct the test in its classic form by means of the paper and pencil. This set describes the ability of the tested individual to keep the same size of the letters during writing. The angle between the line which bounds written sentence from below or above and an imaginary horizontal line on the screen describes the ability of the patient to write in a straight line. Together with the measure of time these features constitute the first subset.

As the sentence may consist of two or three lines calculations of the angles are performed for the first two lines only. The regression lines are useful for calculating the angles that are formed by the imaginary x-axis and the regression lines. Figure 4.1 visualizes the estimated regressions for a tilted sentence. In total six values of angles in degrees are included in the set of features. These features are related to micrographia and their meaning is well known to the practitioners.

The second subset is proposed by Drotar et al. (2016) and it includes different average parameters describing fine motor motions of the writing and drawing process. Finally, the third set of features constitute integral like parameters which accumulate absolute values of the velocities, accelerations, and jerks along the tangent vector to the writing trajectory. An example of tangent vectors for a letter is shown in Figure 4.2. These features are complemented by the ratios allowing to relate their values to the particular drawing or writing task. This feature subset referred as *motion mass* parameters was initially proposed by Nõmm and Toomela (2013) to model the changes in gross motor motions, later in Nõmm et al. (2016) and Nõmm et al. (2018) the set was adopted and extended for the case of fine motor movements.

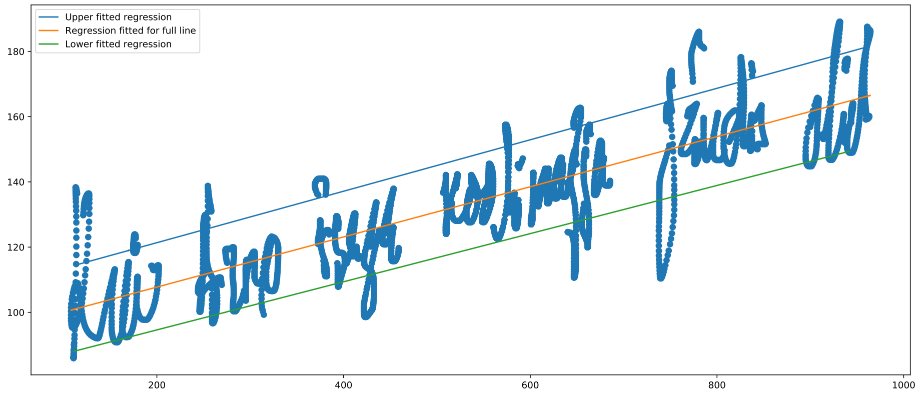


Figure 4.1: Regression fitted for a line written in tilted way

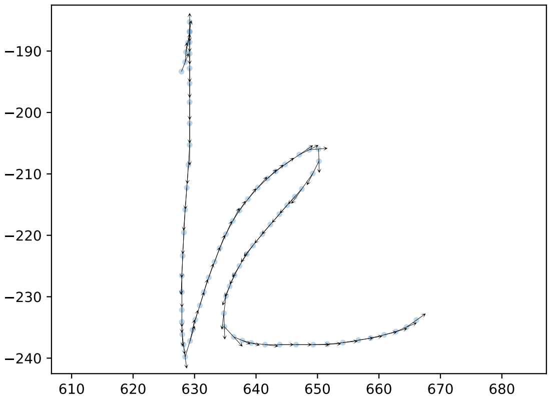


Figure 4.2: An example of tangent vectors for letter

Motion mass parameters are usually associated with a certain part of the test. This part of the test may be defined either by time interval or may be a meaningful part of the test, for example drawing or writing an element of the test. The first parameter is the trajectory length, denoted as L, it describes the entire length of the drawing observed during a certain time interval or being part of the test. It describes the amount of motion performed by the pen tip. Next three parameters describe the smoothness of the motion. The following parameter is the *velocity mass*. Let T be the time interval of interest and t are the time instances observed during the interval of T. Denote velocity along the tangent vector to the drawing curve at time instance $t \in T$ as v_t then

$$V = \sum_{t \in T} |v_t| \tag{4.1}$$

In the same manner *acceleration mass* is defined

$$A = \sum_{t \in T} |a_t| \tag{4.2}$$

where a_t is the acceleration at time t along the tangent vector. The change of acceleration is jerk j_t, jerk mass is defined by equation (4.3):

$$J = \sum_{t \in T} |j_t| \tag{4.3}$$

While a motion may be not-smooth in different ways, the necessity to have various parameters to describe smoothness of the motion arises. It was observed in Senkiv, Nõmm and Toomela (2019) that non-smoothness of different nature may be captured by various motion mass parameters. The original definition of the *motion mass* parameters also included ratios of the trajectory length to the Euclidean distance between the first and last points of the motions and ratio of the acceleration mass to the same distance. These measures are less relevant to the current work and therefore omitted. Instead of this, the logic of moving mass is applied to the pressure that is observed when the pen tip of the stylus touches the screen. The following list summarizes the features that are calculated for individual letters.

- Size features: trajectory is the sum of Euclidean distance between all points that comprise the letter; slope mass is the sum of $\Delta y/\Delta x$, slope mean, mean of first difference of slopes;

- Kinematic features: velocity mass is the sum of velocities for the letter, acceleration mass is the sum of accelerations for the letter, pressure mass is the sum of pressure values applied to the pen, velocity mean, acceleration mean, pressure mean, pressure extrema, velocity extrema, acceleration extrema;

- Duration features: duration is the time interval between the first and last point of the letter

- Fluency features: jerk mass is sum of jerk value for the letter, jerk mean, jerk extrema.

This leads to a total of 822 features that are derived from the sentence and letters. Obviously, for the majority of classical machine learning techniques, such number of features will inevitably cause the curse of dimensionality. Also interpreting decisions made on the basis of a large number of features may be difficult. Which leads the necessity of a proper feature selection.

4.2 Feature selection

Feature selection is performed by Fisher scoring. The Fisher's score is designed for numeric attributes to measure the ratio of the average interclass separation to the average intraclass separation (Aggarwal (2015)). The larger the Fisher's score, the greater the discriminatory power of the feature.

The score is defined as:

$$F = \frac{\sum_{j=1}^{N} p_j(\mu_j - \mu)}{\sum_{j=1}^{N} p_j \sigma_j^2} \tag{4.4}$$

where p_j is the fraction of data points that belong to class $j = 1,..,N$, μ_j and σ_j are respectively the mean and standard deviation of points that belong to class j and μ is the overall mean of the data points.

Features with the score above a threshold of 0.5 are included in predictive models. Table 4.1 shows the Fisher's score for the sentence-specific features. It becomes clear that none of those are included in the predictive models as the threshold is not exceeded. It is worth noting that Nõmm et al. (2018) report the angles of drawing to be of very limited importance for Luria's alternating series test with a much lower Fisher's score than those reported in Table 4.1.

Table 4.2 shows the number of features that are based on individual letters and exceed the threshold of the Fisher's score. The features that exceed the threshold most frequently are kinematic ones and are associated with the first moment of velocity and acceleration. Lange, Tucha, Walitza, Becker, Gerlach, Naumann and Tucha (2006), Drotar et al. (2016), Smits et al. (2014) and Nõmm et al. (2018) report kinematic features to be important with velocity and acceleration being the most important features in the last two papers. I find that duration and fluency features follow kinematic ones with a substantially lower number of occurrences. In contrast to Drotar et al. (2016), Rosenblum, Samuel, Zlotnik,

Table 4.1: Fisher's score of the sentence-specific features

Line	Feature	Fisher score
1	Angle of upper regression line	0.195
	Angle of middle regression line	0.116
	Angle of lower regression line	0.284
2	Angle of upper regression line	0.205
	Angle of middle regression line	0.100
	Angle of lower regression line	0.195

Erikh and Schlesinger (2013) and Stepien et al. (2019) we do not find pressure related measures to be informative which is consistent with Nõmm et al. (2018).

Table 4.2: Ranking of letter specific features based on Fisher's score

Feature	Number of features with $F > 0.5$
Velocity mean	42
Acceleration mean	37
Duration	4
Jerk mass	4
Jerk mean	3
Velocity Extrema	2

It may be instructive to investigate the features that have the highest Fisher's score. For illustrative purposes scatter plot for two features with the highest score is shown in Figure 4.3. Those features happen to be slope mass and jerk mass. Interestingly none of the extracted letters may be chosen to be dominant on the basis of Fisher's score. On the contrary, all of the letters are represented on average by 2-3 features having Fisher's score above threshold.

4.3 Predictive models

Several predictive models may be used for binary classification. The purpose of the models is to estimate the probability that an observation with particular characteristics will fall into one of the categories. The model in a general form may be written as

$$y_i = f(X_i) + e_i \tag{4.5}$$

where y_i is the binary variable indicating whether the person has PD or not, X_i is a vector of features, $f(\cdot)$ is a (nonlinear) function to be estimated and e_i is the residual.

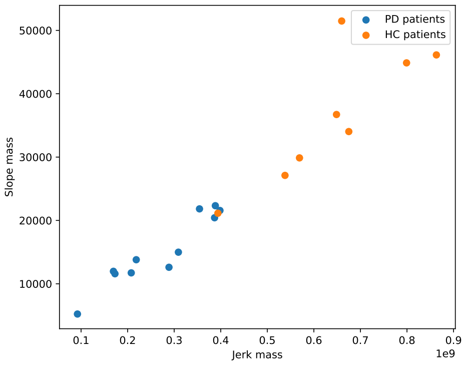


Figure 4.3: Scatter plot of two features with the highest Fisher's score

The most parsimonious model to estimate $f(\cdot)$ may be k-nearest neighbors or a logistic regression. Further decision trees and support vector machines may be used for classification. More complicated ensemble models such as random forests and adaptive boosting may be useful as well. See Aggarwal (2015) and Hastie, Tibshirani and Friedman (2002) for an overview of models and estimation techniques.

4.4 Classification analysis

The following classifiers are trained to obtain the one that performs best for the task in hand (see Aggarwal (2015) and Hastie et al. (2002)).

- Logistic regression is a model that uses a logistic function to model a binary dependent variable.

- K-nearest neighbors classifier (KNN) assigns a class to an object by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

- Decision tree classifier assigns a class to an object after a learning process that is modeled with set of hierarchical decisions based on features and represented by tree-like graph structure. Every node in the tree represents a condition on one or more features of the dataset.

- Support vector machine constructs a linearly separating maximal margin hyperplane through a dataset in order to distinguish two classes. If dataset is non-linearly separable, the observation points are mapped to higher-dimensional space by the means of a kernel function. In this work SVM with linear kerner denoted by SVM_L and radial basis kernel denoted by SVM_{RBF} are analyzed.

- Random forest classifier produces an ensemble of simple decision-tree predictors in randomly selected feature sub-space. Prediction of the class label is determined by most common class among individual decision trees. In that way the model averages out the variance producesd by individual trees and in theory should outperform non-ensemble methods.

- Adaptive boosting classifier combine weaker classifiers in order to build a stronger model. Algorithm constructs a strong classifier as a set of weak classifiers, each performing with at least 50% of accuracy.

For validation purposes, the K-fold cross-validation technique is used. The original dataset is divided into mutually exclusive K subsets that are called folds. Each fold is then used as a validation set with other $K-1$ folds being the training sets. The metric to discriminate between the models is the prediction accuracy obtained on the training set. The final accuracy of each model is the mean accuracy of the model over all folds. Given that the data is scarce and there are just 19 observations the validation is performed for $K = 3$.

Table 4.3: Accuracy of models obtained by K-fold validation with $K = 3$

Model	Accuracy
Random forest	0.885
Logit	0.838
SVM_L	0.771
Decision tree	0.714
KNN	0.742
Adaptive boosting	0.695
SVM_{RBF}	0.580

The accuracy of models is shown in Table 4.3. The model with the highest accuracy is the random forest consisting of trees with a maximum depth of 5. In the related literature Nõmm et al. (2018) use a similar set of models. Also in that study random forest is found to be the best model in terms of accuracy. To some extent, it is not surprising that the ensemble model outperforms the individual models as it averages out the variance of the prediction. Drotar et al. (2016) limit their attention to KNN, adaptive boosting and SVM and report SVM to outperform the other models with classification accuracy of 81.3%.

Tables 4.4 to 4.10 show confusion matrices for all classifiers analyzed in this section. The best classifier produces single error for each class confusing one PD subject with healthy control and vise versa. For the other classifier situation is a bit different. A common problem is classification of the healthy controls. Classifiers do a good job in determining the subjects with PD only occasionally underestimating it. Unfortunately

Table 4.4: Confusion matrix for random forest

	Healthy subjects	PD subjects
Healthy subjects	7 true HC	1 false positive
PD subjects	1 false negative	10 true PD

the models tend to underestimate the amount of healthy controls quite often making the number of false positives quite large. In the work of Toodo (2018) similar pattern is noted for decision tree and KNN.

Table 4.5: Confusion matrix for logistic regression

	Healthy subjects	PD subjects
Healthy subjects	4 true HC	4 false positive
PD subjects	1 false negative	10 true PD

Table 4.6: Confusion matrix for SVM_L

	Healthy subjects	PD subjects
Healthy subjects	5 true HC	3 false positive
PD subjects	1 false negative	10 true PD

The worst classifier is SVM_{RBF} that fails completely on healthy controls. In the related context Toodo (2018) obtained similar result for SVM classifier. Overall performance of the models leads to conclusion that the data is rather scarce and more observations are needed to improve predictive ability. With more data in hand the results may change quite substantially and another classifier may turn to be the best in terms of accuracy.

It should be noted that the classification results may be subject to omitted variable bias. Important individual characteristics may be left out of the model. The most important may be age and previous occupation of the individual. An occupation where person had to write a lot in the past, e.g. teacher or bookkeeper, may influence the way the person performs during the test. The problem of obtaining the missing information lies in the confidential nature of the medical data. Characteristics that might be used to de-anonymize the data are not revealed for current research.

The result of feature engineering and classification analysis confirm the hypothesis set up in Section 1.1. The achieved classification accuracy is on a similar level with the studies in this area. The hard task for classifiers seems to be in tagging the healthy control subjects correctly.

Table 4.7: Confusion matrix for decision tree

	Healthy subjects	PD subjects
Healthy subjects	5 true HC	3 false positive
PD subjects	3 false negative	8 true PD

Table 4.8: Confusion matrix for KNN

	Healthy subjects	PD subjects
Healthy subjects	6 true HC	2 false positive
PD subjects	1 false negative	10 true PD

Table 4.9: Confusion matrix for adaptive boosting

	Healthy subjects	PD subjects
Healthy subjects	4 true HC	4 false positive
PD subjects	2 false negative	9 true PD

Table 4.10: Confusion matrix for SVM_{RBF}

	Healthy subjects	PD subjects
Healthy subjects	0 true HC	8 false positive
PD subjects	0 false negative	11 true PD

Chapter 5

Conclusion

In the thesis a novel approach to model sentence writing test that is used for Parkinson's disease diagnostics is proposed. I analyze digitally conducted sentence writing test that was done by PD patients and healthy controls of similar age. The objective is to come up with a set of features that have high predictive power, may be interpreted and find a classifier that maximizes prediction accuracy. The hypothesis formulated in the thesis states that features based on individual letters allow to set up a classifier with high predictive power.

The dataset consists of 19 observations that represent the sentence written in Estonian language by 11 PD and 8 healthy controls. The data is collected by an iPad application and consists of coordinates of points, timestamp of points and information that describes the position of stylus.

The novelty of the thesis is in producing features on the basis of letters that are extracted from the full sentence. The letter extraction part is quite sophisticated as it requires several steps. First, the number of lines that the sentence is written on is determined. Bigger chunks of letters are extracted next on the basis of the time that had passed during writing the letters. Those sets of letters are separated into smaller parts taking into account distance between points. Finally, the number of extracted parts is optimized to meet the target number of letters the sentence consists of. It is worth noting that the approach is language agnostic and is based on the properties of the data that are recorded while writing.

For the whole sentence the features describing whether the person is maintaining a straight line are calculated. Features that represent size, kinematics, duration and fluency of writing are calculated for each individual letter. The feature selection is performed based on Fisher's score. Interestingly sentence based features do not seem to be important based on the score. On the contrary, kinematic features are found to be of primary importance. This is consistent with the general paradigm shift discussed in Thomas et al. (2017) and findings in Nõmm et al. (2018). With the set of features that have high predictive power in hand, a horse race of seven classifiers is performed. I find the random forest to perform the best with the accuracy of 88.5% on 3-fold cross-validation.

The following list summarizes the main findings that apply to the analyzed sample.

- Kinematic features are important predictors of Parkinson's disease.
- Angle between the written line and an imaginary x-axis does not seem to be an important predictor.
- All of the letters appear to contribute equally to features that are important predictors.
- Random forest classifier is the model with the highest predictive accuracy.
- Classifiers, apart from random forest, tend to produce a considerable amount of false positive results hinting towards scarce data.

As future research steps accuracy of letter detection algorithm may be improved. Analysis of the dataset with a deep neural network may result in a model with higher accuracy and may be considered as an interesting extension. While some letters may be of more importance for doctors, deeper analysis of individual letters may be conducted.

Bibliography

Aggarwal, C. (2015). *Data Mining*, Springer.

Al-Dmour, A. and Fraij, F. (2004). Segmenting arabic handwritten documents into text lines and words, *International Journal of Advancements in Computing Technology* **6**(3): 109 – 119.

Bardõs, K. (2018). Analysis of interpretable anomalies and kinematic paramteres in Lurias's alternating series test for Parkinson's disease modeling, *MSc Thesis*, TUT.

Drotar, P., Mekyska, J., Rektorov, I., Masarov, L., Smekal, Z. and Faundez-Zanuy, M. (2016). Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease, *Artificial Intelligence in Medicine* **67**: 39 – 46. DOI: `https://doi.org/10.1016/j.artmed.2016.01.004`.

Hastie, T., Tibshirani, R. and Friedman, J. (2002). *The Elements of Statistical Learning. 2nd Edition*, Springer Series in Statistics, Springer.

Kozhenkina, J. (2016). A quantitative analysis of the kinematic features for the Luria's alternating series test, *MSc Thesis*, TUT.

Lange, K., Tucha, L., Walitza, S., Becker, G., Gerlach, M., Naumann, M. and Tucha, O. (2006). Brain dopamine and kinematics of graphomotor functions, *Human movement science* **25**: 492–509. DOI: `https://doi.org/10.1016/j.humov.2006.05.006`.

Letanneux, A., Danna, J., Velay, J.-L., Viallet, F. and Pinto, S. (2014). From micrographia to Parkinson's disease dysgraphia, *Movement Disorders* **29**(12): 1467–1475.

Mašarov, I. (2017). Digital clock drawing test implementation and analysis, *MSc Thesis*, TUT.

Moustafa, A. A., Chakravarthy, S., Phillips, J. R., Gupta, A., Keri, S., Polner, B., Frank, M. J. and Jahanshahi, M. (2016). Motor symptoms in Parkinson's disease: A unified framework, *Neuroscience and Biobehavioral Reviews* **68**: 727 – 740. DOI: `https://doi.org/10.1016/j.neubiorev.2016.07.010`.

Nõmm, S., Masharov, I., Toomela, A. and Medijainen, K. (2018). Interpretable quantitative description of the digital clock drawing test for Parkinson's disease modelling, *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pp. 1839–1844. DOI: https://doi.org/10.1109/ICARCV.2018.8581074.

Nõmm, S. and Toomela, A. (2013). An alternative approach to measure quantity and smoothness of the human limb motions, *Estonian Journal of Engineering* **19**(4): 298308.

Nõmm, S., Toomela, A. and Borushko, J. (2013). Alternative approach to model changes of human motor functions, *Modelling Symposium (EMS), 2013 European*, pp. 169–174. DOI: https://doi.org/10.1109/EMS.2013.30.

Nõmm, S., Toomela, A., Kozhenkina, J. and Toomsoo, T. (2016). Quantitative analysis in the digital Luria's alternating series tests, *14th International Conference on Control, Automation, Robotics and Vision*, pp. 1–6. DOI: https://doi.org/10.1109/ICARCV.2016.7838746.

Nackaerts, E., Heremans, E., Smits-Engelsman, B. C. M., Broeder, S., Vandenberghe, W., Bergmans, B. and Nieuwboer, A. (2017). Validity and reliability of a new tool to evaluate handwriting difficulties in Parkinsons disease, *PLOS ONE* **12**(3): 1–14. DOI: https://doi.org/10.1371/journal.pone.0173157.

Nõmm, S., Bardõs, K., Toomela, A., Medijainen, K. and Taba, P. (2018). Detailed analysis of the Luria's alternating series tests for Parkinson's disease diagnostics, *2018 17th IEEE International Conference on Machine Learning and Applications note = DOI: https://doi.org/10.1109/ICMLA.2018.00219*, pp. 1347–1352.

Rosenblum, S., Samuel, M., Zlotnik, S., Erikh, I. and Schlesinger, I. (2013). Handwriting as an objective tool for Parkinson's disease diagnosis, *Journal of neurology* **260**. DOI: https://doi.org/10.1007/s00415-013-6996-x.

Seni, G. and Cohen, E. (1994). External word segmentation of off-line handwritten text lines, *Pattern Recognition* **27**(1): 41 – 52. DOI: https://doi.org/10.1016/0031-3203(94)90016-7.

Senkiv, O., Nõmm, S. and Toomela, A. (2019). Applicability of spiral drawing test for mental fatigue modelling, *IFAC-PapersOnLine* **51**(34): 190 – 195. 2nd IFAC Conference on Cyber-Physical and Human Systems CPHS 2018.

Smits, E., Tolonen, A., Cluitmans, L., Gils, M., Conway, B., Zietsma, R., Leenders, K. and Maurits, N. (2014). Standardized handwriting to assess bradykinesia, micrographia and tremor in Parkinson's disease, *PloS one* **9**. DOI: https://doi.org/10.1371/journal.pone.0097614.

Stepien, P., Kawa, J., Wieczorek, D., Dabrowska, M., Slawek, J. and E.J., S. (2019). Computer aided feature extraction in the paper version of Lurias alternating series test in progressive supranuclear palsy, *in* E. Pietka, P. Badura, J. Kawa and W. Wieclawek (eds), *Information Technology in Biomedicine*, Springer. DOI: https://doi.org/10.1007/978-3-319-91211-0_49.

Tan, J., Lai, J.-H., Wang, C.-D., Wang, W.-X. and Zuo, X.-X. (2012). A new handwritten character segmentation method based on nonlinear clustering, *Neurocomputing* **89**: 213 – 219. DOI: https://doi.org/10.1016/j.neucom.2012.02.026.

Thomas, M., Lenka, A. and Kumar Pal, P. (2017). Handwriting analysis in Parkinson's disease: Current status and future directions, *Movement Disorders Clinical Practice* **4**(6): 806–818. DOI: https://doi.org/10.1002/mdc3.12552.

Toodo, T.-B. (2018). Assessment of parameters from the handwritten sentence test used to diagnose Parkinson's disease, *MSc Thesis*, TUT.

Tseng, Y.-H. and Lee, H.-J. (1999). Recognition-based handwritten chinese character segmentation using a probabilistic Viterbi algorithm, *Pattern Recognition Letters* **20**(8): 791 – 806. DOI: https://doi.org/10.1016/S0167-8655(99)00043-4.

Van Gemmert, A., Hans-Leo, T. and George, S. (2001). Parkinsonian patients reduce their stroke size with increased processing demands, *Brain and cognition* **47**(3): 504–512. DOI: https://doi.org/10.1006/brcg.2001.1328.

Wagle Shukla, A., Ounpraseuth, S., Okun, M., Gray, V. and Schwankhaus, J. (2012). Micrographia and related deficits in Parkinson's disease: a cross-sectional study, *BMJ open* **2**(3). DOI: https://doi.org/10.1136/bmjopen-2011-000628.

Yamaguchi, T., Yoshikawa, T., Shinogi, T., Tsuruoka, S. and Teramoto, M. (2001). A segmentation method for touching japanese handwritten characters based on connecting condition of lines, *Proceedings of Sixth International Conference on Document Analysis and Recognition*, pp. 837–841. DOI: https://doi.org/10.1109/ICDAR.2001.953905.