Summary for the Master's Thesis USING MACHINE LEARNING TO DIAGNOSE SYSTEMIC SCLEROSIS

Karl Märka
Faculty of Sciences
Institute of Chemistry and Biotechnology
Gene Technology
25.05.2015

The aim of this thesis was to provide proof of concept that machine learning techniques can be applied to human disease diagnostics development. This approach has more promise in cases where reliable and affordable diagnostics are not available such as with the autoimmune disease Scleroderma (also called Systemic Sclerosis). The current standard method for diagnosing Systemic Sclerosis relies on the very subjective pinch test where the doctor pinches the patient's skin in various locations around the body and gives a grade of how "elastic" the skin appears.

Systemic Sclerosis can also have fatal complications such as pulmonary arterial hypertension (PAH) which is currently diagnosed with right-side heart catheterization. The procedure itself is invasive, requires hospitalization and in many cases, gives a negative result which means the patient underwent an unnecessary surgical procedure with all the resulting complications and discomfort. It was hypothesized that PAH can also be diagnosed using patient gene expression data.

Several statistical and machine learning techniques were used in the development process for selecting the DNA probes with the best predictive power, reducing the dimensionality of the dataset, selecting the most optimal classifier algorithm, the most optimal settings for it etc. Eventually a relatively simple linear classifier was chosen for implementation purposes. Python and its extensive set of machine learning libraries were used in the entire process.

The resulting predictive models were implemented as a software application with a graphical user interface that reads gene expression data from a NanoString, Illumina or Agilent output file, performs the calculations and outputs the results in the interface and as a pdf file. The setup of an expression array chip containing 58 DNA probes is also suggested to reduce the cost of the final diagnostics package to less than 100$.