The influence of smoking on methylation

A research with methylation rates in CpG islands

Marcel Setz

2021-11-10

Samenvatting

This research came to be with the thought of developing a machine learning technique to determine whether a person smokes or not. On the other side it may be interesting to see if the person is male or female. The provided data exists of methylation rates on certain CpG islands. Also given is the fact whether a person currently smokes or never has. Finally the gender and age are also given.

Contents

Introduction

A CpG island or CpG site is a part of the DNA where the GC content is greater than 50%. In this dataset methylation values of certain CpG sites are displayed with also the age, gender and smoking status for 671 people.

Objective

The objective of this project is to find out if it is possible to determine whether a person smokes or not according to the methylation rates. The research question is: Is it possible to identify a person's gender, age or status of smoking given their methylation values on CpG islands?

Materials & Methods

This projects exists of two parts: the analysis of the data and the development of a Java wrapper which processes the data. All processed results and the analysis are in the next to repositories: https://github.com/marcelsetz/AnalysisT9 en https://github.com/marcelsetz/JavaWrapper.

The data was analysed using different plotting methods in Rstudio. the main plotting library used here was ggplot2. The analysis was mainly focused on checking the data: was there any missing data, were there any outliers, does the data seem logical. The provided data had a lot of missing values with only gender data, age. There where no extreme outliers in the data.

The cleaned data was tested with multiple machine learning algorithms in Weka, like ZeroR, OneR, J48, RandomForest, SMO, SimpleLogistics and NaiveBayes. These algorithms were performed multiple times with 10-fold cross-validation.

The Java wrapper was created with two of these algorithms: J48 and RandomForest. This program takes arguments from a user and an input file to classify the unknown labels using J48 and RandomForest. The new classified file will then be put in an output file.

Results

These PCA plots show the differences between smoking status and gender. Obvious here is that there is a slight difference in smoking status where the current smoking group is slightly to the left. This can also be, because there is a huge difference in the number of current smokers and non-smokers. In the second graph, all males are left and all females are right with no exceptions. The graph is a bit scaled to make it more readable.

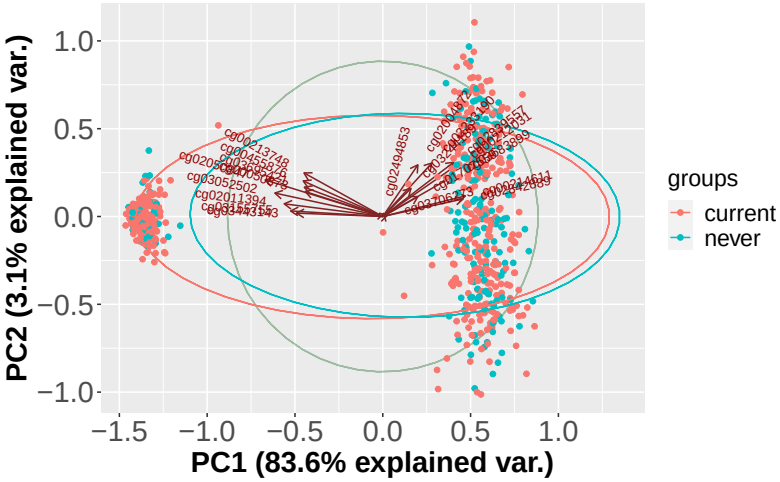

Figure 4: PCA graph displaying smoking status as different groups

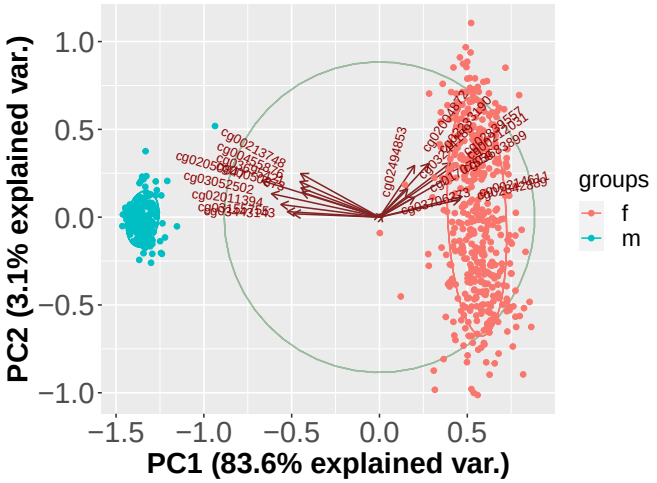

Figure 5: PCA graph displaying gender as different groups

Conclusion and Discussion

So far we don't have a concrete answer to the research question, because the algorithms aren't extremely overwhelming. So for now it isn't really possible to determine whether someone is smoking or not, because the data is too much spreaded out and not really divided into groups.

It's quite obvious that the percentage of correctly classified instances of every algorithm except for NaiveBayes are very close to eachother, the only significant lower scores according to weka are NaiveBayes and OneR, the best (not really significant) algorithm are surprisingly ZeroR, J48, SimpleLogistics and SMO. However it's quite remarkable that the ROC-area and the precision are the highest on NaiveBayes. So even though the results are disappointing, this is the best algorithm according to these quality metrics.

The Data with the gender as class is quite more interesting, these algorithms have a way higher area under the curve and correctly classified instance percentage.

One possible purpose for the use of the gender class may be for forensic research. If some suspect need to be found, it's a way to find out if the suspect is male or female.

After some research I found out what the reason for this remarkable difference is. These differences have everything to do with the diversity of the transcriptomic and proteomic profiles in the two sexes. [1]

Literature

[1] Yusipov, I., Bacalini, M. G., Kalyakulina, A., Krivonosov, M., Pirazzini, C., Gensous, N., Ravaioli, F., Milazzo, M., Giuliani, C., Vedunova, M., Fiorito, G., Gagliardi, A., Polidoro, S., Garagnani, P., Ivanchenko, M., & Franceschi, C. (2020). Age-related DNA methylation changes are sex-specific: a comprehensive assessment. Aging, 12(23), 24057–24080. https://doi.org/10.18632/aging.202251

Appendix

Appendix 1: Nederlandse Samenvatting

Dit onderzoek is tot stand gekomen met de gedachte op het ontwikkelen van een machine learning techniek om uit te vinden of iemand rookt of niet. Anderzijds kan het interessant zijn om uit te zoeken of iemand man of vrouw is. De gegeven data bestaat uit methylatie waardes van bepaalde CpG eilandjes. Ook in de data te vinden zijn gender, leeftijd en of een persoon rookt of niet.