Use of machine learning to study batch effect removal methods in metabolomics data

Marc Solé Estragues, 202002465

Master's Thesis, Bioinformatics

Supervisor: Palle Villesen

June 2022

AARHUS
UNIVERSITET
CENTER FOR BIOINFORMATIK (BIRC)

Abstract

The fusion of criminology and machine learning is a fascinating field, the use of machine learning models able to identify and predict biological characteristics is changing the way criminal investigations are handled. Using blood metabolomics data it is possible to construct machine learning models able to interpret the chemical changes that happen as the sample degrades through time and are able to predict the age of a blood spot. This information can then be used to determine the time when a crime took place. One of the big limitations of metabolomics data is the presence of batch effect. The removal of such undesired effect is a necessity in order to obtain unbiased data which can later be used in comparative analyses which require of high statistical power such as the construction of machine learning models. Many different techniques to achieve this correction on the data are available, and which one should be used to obtain the best data capable of the best machine learning performance is a question that needs to be answered.

This thesis has sought to study how these methods can be applied and how they transform a specific set of blood metabolomics data obtained through HPLC-MS/MS which is being used in research to construct a machine learning model able to predict the age of dried blood spots. Of special interest has been to compare how well different methods are able to remove the batch effect from the samples, as well as how much improvement is observed in the predictive error using simple elastic-net and random forests regression models.

It was determined through principal component analysis that *ComBat* and *WaveICA2.0*, two of the most advanced batch removal methods available, as well as mean centering, perform good batch correction on the data, with other methods such as row normalisation and probabilistic quotient normalisation performing the worst in that regard. The analysis also revealed that when applying *WaveICA2.0* the correct cutoff parameter needs to be chosen, and much work went into exploring how these cutoff value changes the amount of batch correction applied. In the end the PCAs revealed that the data does contain batch effect but it is orthogonal to the biological variable of age being studied, and thus when the batch correction is applied little change in the statistical performance is observed and the root mean square error obtained from the ML model's predictions can even increase in some cases.

Over-all, the results show that some batch correction methods perform better than others, but their use will depend on the final aim of the project and the type of data that is being worked with. In the case of this specific metabolomics data, strong batch correction results in lower predictive performance overall and methods such as *WaveICA2.0* which have a tuning capability in the cutoff parameter could be considered to adjust for the best amount of batch correction while maintaining good prediction accuracy.

Aknowledgements

I dedicate this work of non fiction to my family who from the distance have always supported me, my bros Adam Petro and Lisa Berling who have always motivated me to keep going, to my supervisor Palle Villesen for his great guidance and advice, and finally I want to thank the whole country of Denmark and Aarhus University for welcoming me with open arms.

Marc Solé Estragués,
Aarhus, 12th June 2022.

Contents