

Predicting Contaminated PCR Plates in a Metagenomic Sequencing Study

Maximilian Joas¹

Universität Leipzig
Machine Learning Group
Leipzig, Germany
`mj13body@studserv.uni-leipzig.de`

Zusammenfassung. Geben Sie Ihrer Ausarbeitung einen möglichst aussagekräftigen Titel und fassen Sie Ihre Arbeit *nach Fertigstellung des eigentlichen Berichts* noch einmal an dieser Stelle so kurz wie möglich zusammen. Versuchen Sie dabei folgende Aspekte jeweils mit einem Satz (maximal zwei Sätze) zusammenzufassen: Fragestellung, Methodik, Ergebnisse, Schlussfolgerungen. Die Zusammenfassung muss nicht vollständig sein, sondern sollte in erste Linie so klar wie möglich herausstellen, warum es sich lohnt, die vorliegende Arbeit vollständig zu lesen.

Schlüsselwörter: Maschinelles Lernen, Empirische Daten, Wissenschaftliches Arbeiten

1 Research Question

Advances and cost reduction of sequencing experiments had an vast impact on the field of Metagenomics (research that focuses on many genomes at the same time). The vast amount of data obtained from whole genome sequencing experiments in metagenomic leads to opportunities in pathogen detection, ecological studies and drug development [?]. On the other hand the abundance of data poses challenges to the analysis of metagenomic data. Consequently, classical statistical methods reach their limits and other approaches such as machine learning are often used [?].

Not only the amount of data, but also the type of data are challenging, on top of that the technical process of sequencing can lead to problems. In particular, contamination and noise are important problems when studying metagenomics [?]. Noised up and or contaminated data can lead to false conclusions in experiments. Therefore it would be valuable to predict contamination on PCR plates and find factors that are distinctive to contaminated PCR plates. In this article I will investigate if it is possible to predict contaminated PCR plates based on OTU counts as well as meta data. Furthermore I will present factors that distinguish contaminated from not contaminated PCR plates with the help of machine learning.

2 State of the Art

Machine learning methods have been successfully used in the study of metagenomics data [?]. However, specifically on my research question there are, to the best of my knowledge no publications. Nevertheless, this section aims to give an overview of the state of the art of the implemented methods. In general, machine learning methods are distinguished in supervised and unsupervised methods. In the case of supervised learning there is a known mapping between input and target variable and the methods learn the association and can predict the output variable for new data. In this case the target variable is if the PCR plate is contaminated. TODO

3 Methodik

3.1 Conception

The research question is divided into two parts: (I) What did change on the contaminated PCR plates? (II) Is it possible to predict the contaminated PCR plate? In order to solve the first question in a machine learning context, the most important features for the supervised learning method were retrieved. Therefore, I had to use methods that not only predict the contaminated plate, but also determine the most important features for the prediction. The second question was more suited for classical supervised learning approach: I used two methods to predict the PCR Plate, one classical machine learning approach (Random Forest Classifier) and a neuro-inspired approach (Multi Layer Perceptron). Before the methods can be described in detail, it is of the utmost importance to know the data. MISSING TODO METADATA EXPERIMENT

3.2 Data and Preprocessing

The used data came from the ETH Zurich. The data consisted of a count table of microorganisms sequencing reads from mice from different laboratories. The data set did not come normalized. Additionally, to the count table a data set containing metadata was included in the analysis. The metadata contained a variety of technical information, such as the PCR plate, the date of the extraction run or the name of the researchers that executed the experiments. In total, the data set consisted of 199 samples and 1564 OTUs and 61 different metadata entries per sample. OTU TAXONOMY DATA TODO The median sequencing depth of the samples was 80275. Note that the ETH Zurich provided another dataset with three replicas of each sample, but for the analyses the dataset without the replicas was used.

Before the data set can be used for machine learning, preprocessing is necessary. The first step was to match the metadata of the samples, specifically the information of the PCR plate to the features of the samples (OTU counts).

This was necessary, because the sample id was different for the two datasets and in a different order. Subsequently, the data was checked for missing value. The counttable did not contain any missing values, the metadata had multiple features with missing values. Since the information on the PCR plate was complete for all samples, no sample was excluded for the main analyses. When the metadata was used for the prediction, I excluded samples with missing values. 26 samples were excluded for this experiment. Subsequently, the datatypes were checked, non numeric data types were transformed to a suitable format. Since, I wanted to use machine learning not only with the counttable, but also with the metadata, this step was necessary (the count data is numeric by default). Each category of the non-numeric metadata was encoded by a number from one to n (number of different categories). In contrast to one-hot encoding this method does increase the feature space. However, it can have an influence on the predictions. TODO EXPLAIN BETTER. Lastly, I checked for outliers in the metadata via Boxplots, but did not find too extreme values that had to be excluded. For the counttable a z-transformation was performed, in order to use it for the multilayer perceptron.

3.3 Supervised Learning Methods

4 Ergebnisse

Nach Möglichkeit sollten alle *relevanten* Ergebnisse der Arbeit in einem eigenen Ergebnis-Abschnitt gebündelt wiedergegeben werden. Nutzen Sie hierfür ggf. Subsections. Die Darstellung der Ergebnisse kann je nach Zusammenhang oder Zielsetzung mittels Tabellen (siehe z.B. Tabelle 1), Diagrammen oder anderer Abbildungen erfolgen, die zusätzlich kurz so neutral und objektiv wie möglich in Worten beschrieben werden. Beschränken Sie sich dabei auf die wesentlichsten Ergebnisse - ohne jedoch zu beschönigen oder unliebsame Daten auszulassen.

Tabelle 1. Beispieltabelle für erzielte Ergebnisse.

Year	World population
8000 B.C.	5,000,000
50 A.D.	200,000,000
1650 A.D.	500,000,000
1945 A.D.	2,300,000,000
1980 A.D.	4,400,000,000

5 Diskussion

Wichtiges Kennzeichen wissenschaftlichen Arbeitens ist die Trennung von Ergebnissen und deren Bewertung. Während im Abschnitt “Ergebnisse” entsprechend

weitgehend auf Bewertungen verzichtet wird, findet die eigentliche Auswertung bzw. Interpretation der Ergebnisse in diesem Abschnitt statt. Ein weiteres Ziel ist eine kritische Reflexion der verwendeten Methodik.

Bei der **Interpretation der Ergebnisse** sollten Sie regelmäßig Bezug auf die im Abschnitt “Ergebnisse” aufgeführten Tabellen oder Abbildungen nehmen. Eine bloße Zusammenfassung der erzielten Ergebnisse ist allerdings nicht ausreichend. Entscheidend ist vielmehr, die Ergebnisse einzuordnen und hinsichtlich ihrer Plausibilität sowie möglicher Fehler zu hinterfragen.

Bei der **Reflexion der Methoden** stellen Sie dar, inwieweit die von Ihnen gewählte Methodik dazu geeignet ist, die gewählte Fragestellung zu beantworten. Ein wichtiges wissenschaftliches Instrument dafür ist der Vergleich Ihrer Methode(n) mit einer oder mehreren Referenzmethoden. Es sollen sowohl Stärken bzw. Nutzen der Arbeit als auch deren Schwächen bzw. Grenzen dargestellt werden.

6 Schlussfolgerungen

Der letzte Abschnitt rundet die Arbeit mit einer zusammenfassenden Einordnung der erzielten Ergebnisse und der gewonnenen Erkenntnisse ab. Darin sollte insbesondere die im ersten Abschnitt eingeführte Forschungsfrage aufgegriffen und beantwortet werden. Im Idealfall wird darüber hinaus ein (kurzer) Ausblick auf potenzielle oder tatsächlich geplante weiterführende Arbeiten oder neue Fragestellungen gegeben, die sich aus der vorliegenden Arbeit ergeben.