Applied Machine Learning for Biological Problems

Project 2 Report

Edina Marica, Răzvan-Andrei Morariu & Karl Hendrik Tamkivi

January 2024

Introduction

This project focuses on DNA segments shared by a subset of *Neisseria gonorrhoeae* strains that are statistically associated with resistance to three different antibiotics. The aim of the analysis is to develop a classifier of potential antibiotic-resistant strains in samples, providing guidance for ongoing medication therapy.

Selection of the Dataset

At the stat of the project our group was able to choose between two different datasets for the project: one concerning antibiotic resistance of different genetic strains and the other about orthologous groups concerning secretion systems. After reading both datasets' descriptions, we decided to choose the antibiotic resistance dataset. We chose this as two people out of the three group members don't have a background in biology, and the antibiotic resistance dataset required a less domain knowledge to understand the data. Moreover, for the Paper Club presentation, we also had to present a paper about antibiotic resistance.

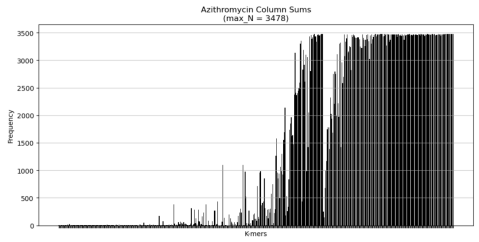
1 Exploratory Data Analysis

We started our project by importing the first data file "resist.csv", that contained information about different genetic k-mer samples and each sample was assigned the resistance to three different antibiotics: **azithromycin, cefixime** and **ciprofloxacin**. The resistance was given as a binary variable, with value 1 if the sample is resistant to the antibiotic drug, and 0 otherwise.

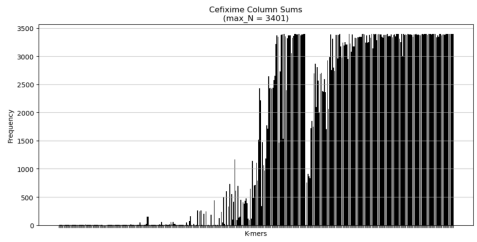
The dataset contained **missing values**, therefore, the first step was to decide how to handle them. Although the share of missing values per antibiotic was not significantly high (8.1%, 18.4% and 10.2% for azithromycin, cefixime and ciprofloxacin respectively), one could have still aimed to keep as much data points as possible for the following prediction analysis. Thus imputation of missing values would have been a possible step but in this case we had no information about the possible biological similarity of the samples. Not being certain if Sample_ID string similarity would also translate into biological similarity between the samples, we decided not to impute the missing values without additional similarity measures and omit the samples that have missing values from the analysis entirely.

After this, we divided the dataset into three different datasets, to have a separate dataset for each antibiotic. For each dataset, we included only the samples that didn't have missing values considering resistance to that particular antibiotic. Next, we loaded in the k-mer datasets for each antibiotic, and in order to merge the information of the resistance data with k-mer data, the k-mer tables were transposed (sample IDs in rows and patterns in columns). After transposing the k-mer datasets, we merged the resistance datasets with the corresponding antibiotic's k-mer dataset by the Sample_ID.

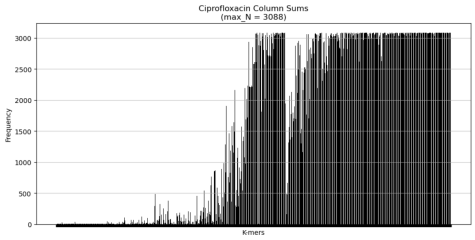
After the merging steps we were able to plot some informative visualizations to get a more in-depth understanding of the data. We started by visualizing the frequency of the k-mers in the three antibiotics' datasets separately, and we observed that there were k-mers that appeared in almost all samples while we also had k-mers that were present in almost none of them (Fig 1). We would later use this finding for the feature selection part of our predictive modelling.



(a) azithromycin



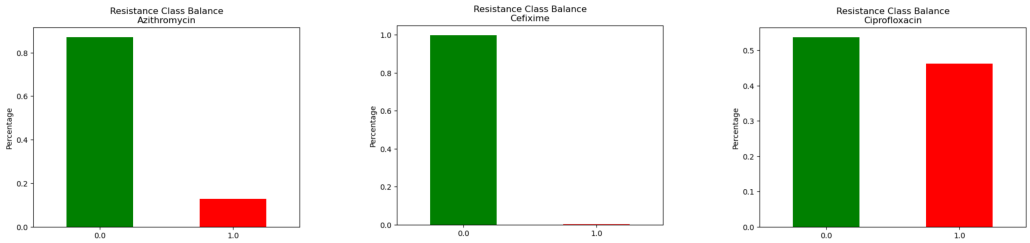
(b) cefixime



(c) ciprofloxacin

Figure 1: Frequency of k-mers

After this we checked the resistance class balance for each antibiotic and noted that the data was very **unbalanced** for azithromycin and cefixime. The former had more than 90% of the samples with target value 0, meaning that more than 90% of the samples in the data were non-resistant, while the latter only had 1 sample that was resistant (Fig. 2a, 2b). The ciprofloxacin data however, was more balanced (Fig 2c). This was an important finding that we needed to consider for the following model training and selection phase.



(a) azithromycin (b) cefixime (c) ciprofloxacin

Figure 2: Resistance class balance

Remark. *We note that we worked with the three datasets separately, each of them corresponding to one antibiotic. One could argue that it would be better to treat the three antibiotics together and to train one model on the whole dataset. This would result in a generalised approach, that could maybe later be used for other antibiotics, not only the given ones. However, we decided to train models separately on the three antibiotics' datasets, as this way, we could possibly expect more accurate predictions of resistance to the given antibiotics for new samples.*

2 Feature Selection

Each dataset had 517 features, meaning 517 k-mers, and we wanted to eliminate the ones that would not have a high influence on the target variable. For this, we took into consideration the findings from the Explanatory Data Analysis part, where we plotted the frequencies of the k-mers. The ones that appeared in almost all or in almost none of the samples would not give us too much novel information about the target variable and therefore, we removed them. The rest of the k-mers were kept for the further analysis.

3 Model Selection

Before selecting between possible model types, we focused on handling the class imbalances:

1) For the ciprofloxacin data we did not need to apply any additoinal class balancing methods since the data was already quite well balanced.

2) For the azithromycin data we decided to use **SMOTE-N**. This is a modification on the classic oversampling algorithm SMOTE, that is intended for categorical data.

3) For the cefixime we decided to drop that antibiotic from the further analysis entirely because of the severe class inbalance. Firstly, we cannot split the dataset in training and test and having samples from that class in both training and test. In addition to this, we can not be confident in a model that predicts well for one single sample. Moreover, the SMOTE-N technique needs more samples from a specific class to work properly.

For the model selection, we compared the performance of three different models:

- Random Forest
- SVC
- KNN

We created a function **"find_best_model"** that uses **grid search** to find the best model out of the three mentioned above together with the best parameters. We used this function on the training data for ciprofloxacin, azithromycin, and also azithromycin modified with SMOTE-N. The scoring metric used was balanced accuracy because it is a good scoring metric for imbalanced datasets. The parameters that we tried in grid search were the following:

Model	Parameters	Values
Random Forest	n_estimators max_depth	{10, 50} {10, 20}
SVC	C kernel	{1, 10, 20, 30, 50} {linear, rbf}
KNN	n_neighbors weights	{3, 5, 7} {uniform, distance}

Consequently, we chose the best model for every dataset and computed the performance on the training set. The best performing model selected by the grid search was a **SVC** in all three cases.

4 Performance Evaluation

For the evaluation part, we measured the final balanced accuracy on the test set for the best model. The balanced accuracy scores for the three tested datasets were the following:

1) azithromycin: **0.919**
2) azithromycin with SMOTE-N: **0.922**
3) ciprofloxacin: **0.969**

The models were able to predict the resistance class surprisingly well, having a balanced accuracy higher than 90% in all three cases. It is also worth mentioning that using SMOTE-N on the azithromycin data resulted in increase of the balanced accuracy score, although a minor one.

Remark. *For different splits or different SMOTE-N behavior the scores might vary slightly.*