Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210185

Machine Learning Algorithms Reveals Country-Specific Metagenomic Taxa from American Gut Project Data

Jose LIÑARES-BLANCO[a], Carlos FERNANDEZ-LOZANO[a], Jose A. SEOANE[b] and Guillermo LOPEZ-CAMPOS[c,1]
[a] *Department of Computer Science and Information Technologies, Faculty of Computer Science, University of A Coruña, CITIC, Campus Elviña s/n, A Coruña, 15071, Spain*
[b] *Stanford Cancer Institute, Stanford University School of Medicine, Stanford CA, USA*
[c] *Wellcome-Wolfson Institute for Experimental Medicine, Queen´s University Belfast, Belfast, UK*

Abstract. In recent years, microbiota has become an increasingly relevant factor for the understanding and potential treatment of diseases. In this work, based on the data reported by the largest study of microbioma in the world, a classification model has been developed based on Machine Learning (ML) capable of predicting the country of origin (United Kingdom vs United States) according to metagenomic data. The data were used for the training of a glmnet algorithm and a Random Forest algorithm. Both algorithms obtained similar results (0.698 and 0.672 in AUC, respectively). Furthermore, thanks to the application of a multivariate feature selection algorithm, eleven metagenomic genres highly correlated with the country of origin were obtained. An in-depth study of the variables used in each model is shown in the present work.

Keywords. Metagenomics, Machine-Learning, Feature Selection

1. Introduction

The gut microbiome is a highly complex ecosystem with high variability across individuals. The factors that influence this variability are both genetic and environmental. Among the environmental factors, diet, climate, environment and even social factors can influence the composition of the intestinal microbiome in individuals.

Prior studies have been conducted to explore the plausible links between diseases and the microbiome and efforts are being made to understand how microbiome varies with host lifestyle, genetics, age, nutrition, medication and environment. Also, several studies have characterized microbiome diversity through countries identifying enriched taxa associated with geographical areas [1].

Recent advancement of culture-independent, high throughput next-generation sequencing technologies has enhanced our ability to characterize the human microbiome at various states of health and disease. In addition, open large-scale initiatives such as the Human Microbiome Project (HMP) or American Gut Project

[1] Corresponding Author e-mail: g.lopezcampos@qub.ac.uk

(AGP) allow us to study and compare metagenomes under different conditions with great statistical significance.

Considering the complexity of these data, it requires the scientific community to offer new methods of analysis and prediction. At the same time, the rise of Machine Learning (ML) for the computational analysis of complex and high-capacity biomedical data makes use of models [2] with great predictive capacity developed specifically for problems with large amounts of data and with noise. Once trained, the ML algorithm is intended to have the highest possible generalization capacity so that the model works not only with the data it has learned, but also with the data we will obtain in the future.

As a proof of concept for the application of machine learning techniques in the study of the microbiome and the generation of predictive models we have studied the possibility of predicting the country of origin for microbiome samples. This model only has information from metagenomic data, so the presented results provide information about the country-specific taxa, which may be important for further work in this field of research.

2. Methods

2.1. Dataset

The data used in this work was downloaded from the American Gut Project (AGP) (ftp://ftp.microbio.me/AmericanGut/). Raw data of Operational Taxonomic Unit (OTUs) counts from AGP were downloaded. Phyloseq R package [3] was used to manage this data and filtered by feces samples. We obtained a total of 36.405 OTUs from 12.189 individuals. The first step was agglomerate all OTUs at the taxonomic rank of Genus. After this step, the dataset was simplified to 2.082 OTUs. Those OTUs that have an unknown gender (labelled as "g__") were eliminated. Then, an analysis of oultiers using the Isolation Forest technique from H20 R package [4]. With this technique we were able to eliminate a total of 1.219 individuals. The remaining individuals were labeled according to their country of origin. For our analyses we focused only in the two countries with the largest number of participants (USA and UK). The dataset was labeled and balanced to the minority class. The final dataset presented 2502 individuals from UK and 2502 from USA.

After preprocessing, whole dataset was splitted into 85% train and 15% test set. Train set was the input to feature selection algorithm. From all remained OTUs, we found a subgroup of 11 OTUs that had a high consistency and correlation. Consistency attempts to find a minimum number of features that separate classes as consistently as the full set of features can. A good feature subset is one that contains features highly correlates to the class, yet uncorrelated to each other. These features were the input to the ML algorithms.

2.2. Machine Learning

Machine learning algorithms are capable of transforming data into intelligent actions, extracting specific knowledge from a set of data that humans would not be able to achieve. In this work we carried out our experiments using two supervised

classification algorithms, Random Forest (RF) [5] and Generalized Linear Model (glmnet) [6]. Both algorithms have a particular set of hyperparameters that should be tuned to find the best possible combination to achieve the best performance. Below, we provide a description of each algorithm and how the hyperparameters were tuned.

2.2.1. Random Forest

The RF algorithm consists of a set of independent decision trees based on the random resampling of the variables for the construction of each tree. A search was made for the appropriate values for the hyperparameters *mtry* (number of variables randomly sampled in each data division), *nodesize* (minimum size of the terminal nodes) and number of trees. The range for the number of variables was set between 1 and, as an upper limit, the square root of the number of variables of the data set. The minimum size of the terminal nodes was set between 1 and 3. Low values of this parameter provide high growth and depth of each tree, which improves the accuracy of the predictions. In addition, the number of trees was 1000. A large number of trees ensures that each observation is predicted at least several times.

2.2.2. Generalized Linear Model

The *glmnet* algorithm is a rapid regularization algorithm that fits a generalized linear model with elastic network penalties. The network penalty depends on two terms: the *ridge* penalty, which aims to reduce the coefficients of the predictors correlated with each other, and the *lasso* penalty, which tends to choose one of them and discard the others. A search was carried out for the appropriate values for *alpha* (controls the penalty of the elastic network) and *lambda* (controls the total strength of the penalty). The values of *alpha* ranged from 0.0001, 0.001, 0.01, 0.1 and 1, while those of *lambda* were 0, 0.15, 0.25, 0.35, 0.5, 0.65, 0.75, 0.85, 1.

2.3. Feature Selection

In this work, a predominant correlation analysis [7] was used to evaluate feature (taxa) correlation in metagenomic data and to filter the most informative features, reducing the dimensionality of the analysis. This approach is basically a multivariate filtering method, which uses the measure of entropy (H) and the Information Gain (IG) for the search of the subgroup of dominant features for a specific condition. The action of these two measures is encapsulated in the Symmetrical Uncertainty (SU) [8].

Initially, the SU value was calculated for each feature, keeping relevant features based on a threshold (0.025) and sorting them in descending order according to this value. Secondly, features providing redundant information were removed. For a better understanding of this methodology, see [7]. Thus, we selected features in a model-independent manner, selecting features with high correlation with patient country origin, but little correlation with other non-informative features (predominant correlation). In our study, this approach was run on the entire set of features, after preprocessing, around 2.000 different features. Out of these, the algorithm extracted 11 that satisfied the defined requirements.

3. Results

Due to the complexity of the data, two types of algorithms that have been widely used in the field of omics analysis were chosen (Random Forest and *glmnet*). Since no ML algorithm has obtained great results in the field of metagenomics, mainly due to the sparse characteristics of these data, we have chosen to work with two algorithms that are simple to implement, of low computational cost and mainly that can be explained later. In this way, it will be possible to extract information on the characteristics used in each model and their respective importance. The code and data generated from this work is available in the following GitHub repository (https://github.com/jlinaresb/MIE_Metagenomics). Results of in train and test sets are presented in Table 1. Measures of Accuracy (Acc) and Area under ROC curve (AUC) are presented. The models were validated with a two-level cross validation (CV) in training phase. This type of validation consists of two CV processes, an independent internal level (2/3 holdout for training and 1/3 for validation) for the selection of the best hyperparameters of each algorithm and an independent external level (in this case, 5 repetitions of a 10-fold CV) to evaluate the model's capacity for generalization and ensure that there are no biases in the data. It is shown in Table 1 we found an algorithm-independent signature, since both ML algorithms achieved similar performances scores. To ensure that there is no over-fitting in the model, the data were validated in a test set.

Table 1. Results of the ML experiment. Models were trained to predicts country origin of individuals from metagenomic profiles. Measures of accuracy (Acc) and area under ROC curve (AUC) are shown. In train results, the interval is shown over the 50 iterations.

			Glmnet		Random Forest	
	#N	**#OTUs**	**Acc**	**AUC**	**Acc**	**AUC**
Train	4253	11	0.61-0.69	0.66-0.74	0.57-0.68	0.62-0.74
Test	751	11	0.660	0.698	0.639	0.672

Variable importance of each model is shown in Figure 1. On one hand, glmnet variable importance represents the beta value of each variable. On the other hand, RF variable importance is based on the Gini impurity index.

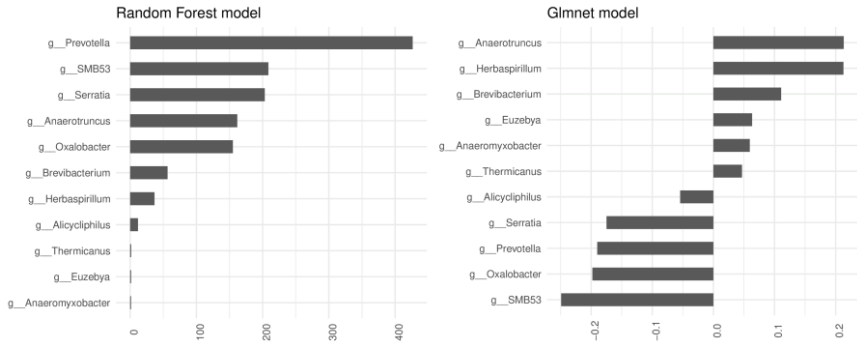

Figure 1. Importance of the features of the two carried out models.

4. Discussion

The analyses of metagenomic signatures for health-related purposes is an increasingly relevant topic in the advancement of precision medicine. Searching for geographically specific metagenomic biomarkers will avoid future problems when developing microbiome-based drugs according to the geographical characteristics of individuals. In this work, eleven genus taxa were found with high correlation with individuals origin country (UK and USA). Both algorithms, RF and glmnet give significant importance to the genus Prevotella, SMB53, Oxalobacter, Serratia, Anaerotruncus and Brevibacterium. The other genus does not seem to offer enough information to both models.

5. Conclusions

In this work we have explored the application of a feature selection combined with a machine learning approach to the study of microbiome samples focusing in the prediction of the origin of the individuals. Our results show that we have been able to apply two different ML algorithms achieving similar performance scores. Extracting information about the specific metagenomic composition of each country will open the door to microbial therapeutic strategies at a global level.

The main conclusion obtained in this proof-of-concept is that the ML algorithms, after a rigorous pre-processing of the data are able to train and achieve significant yields from metagenomic data. The complexity of the metagenomic data explains why previous work has mainly been based on conventional statistics to infer inter-individual differences. This work offers the possibility of using these algorithms in more complex problems such as the diagnosis of diseases, subtyping of patients and/or the search for specific treatments.

References

[1] Gupta VK, Paul S, Dutta C. Geography, ethnicity or subsistence-specific variationsin human microbiome composition and diversity, Frontiers in microbiology 2017;8:1162.
[2] Harrington P, Machine learning in action, Manning Publications Co., 2012.
[3] McMurdie PJ, Holmes S. Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data, PLoS ONE 2013; 8(4):e61217.
[4] LeDell E, Gill N, Aiello S, Fu A, Candel,A, Click C, Kraljevic,T, Nykodym,T, Aboyoun P, Kurka M, Malohlava M. h2o: R Interface for the 'H2O' Scalable Ma-chine Learning Platform, r package version 3.32.0.1 2020;45(1)5–32.
[5] Breiman L/ Random forests, Machine learning 2001;45(1):5–32.
[6] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linearmodels via coordinate descent, Journal of statistical software 2010;33(1):1.
[7] Yu L, Liu H. Feature selection for high-dimensional data: A fast correlation-basedfilter solution, in: Proceedings of the 20th international conference on machine learn-ing (ICML-03) 2003;856–863.
[8] Press WH, Teukolsky SA, Vetterling,WT, Flannery BP. Numerical recipes 3rdedition: The art of scientific computing, Cambridge university press, 2007.