RESEARCH

Project-1

Sujan Darai (sujad96@zedat.fu-berlin.de), Matanat Mammadli (matanam94@zedat.fu-berlin.de), Samra Hamidovic (samrah96@zedat.fu-berlin.de) and Ibraimi Ibraim (ibraimii@hu-berlin.de)

Full list of author information is available at the end of the article

Abstract

Goal of the project: Analysis and result replication of eight different machine learning algorithms on metabolomics data.

Main results of the project:

Comparison and evaluation of two machine learning models with the previously published Mendez paper[2]. Identification of important metabolites like cystine and lysine related to postmenopausal hormone therapy

Personal key learning:

1 Sujan Darai: Creating a virtual environment and installing the dependencies required to run the code.
2 Samra Hamidovic: Creating a virtual environment and better knowledge about different machine learning methods
3 Matanat Mammadli: Downloading Anaconda, installing packages to run the code, and learning about different machine learning methods.
4 Ibraim Ibraimi:

Estimated working hours:

1 Sujan Darai: 10 hours
2 Matanat Mammadli: 8 hours
3 Samra Hamidovic: 8 hours

Project evaluation: 1

No. of words: 1540 words.

Keywords: Metabolics, Principle Component Regression (PCR), Linear Kernel Support Vector Machine (SVM-lin)

1 Introduction

Metabolomics is one of the many omics approaches that are used in the next-generation sequencing era in view of dissecting the molecular mechanisms behind gene, protein, or metabolite function and their implication in disease. One of the main purposes of metabolomics research is to identify the metabolites in the biological samples to facilitate their role in metabolism. Different machine-learning algorithms have been developed in recent times to analyze the metabolic data. Until a short while ago, the Partial Least Squares (PLS) machine learning algorithm was used in metabolics due to its capacity to handle high-dimensional information and give interpretable information. Due to the nonlinear structures in metabolic data, advanced nonlinear machine learning strategies like Random Forest (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANNs) are picking up considerable attention. Despite their capacity, these models have been neglected in metabolomics due to computational complexities and cost reasons. This paper compares the predictive performances of different machine-learning algorithms on metabolomics datasets. [2].

2 Goal of the project

The primary focus of this project is to reproduce the results of the 8 different machine learning algorithms on metabolomics data. It also includes the optimization of the algorithms if necessary and extracting the important features from the machine learning models. The goal also includes the reproducibility of the performance of the model with previously studied papers. The machine learning algorithms used in this report are Principal Component Regression (PCR) and Linear Kernel Support Vector Machine (SVM-Lin).

3 Data and preprocessing

Data used in the paper were of clinical origin. They were previously published. These data are publicly available at either MetaboLights or Metabolomics Workbench data repositories (https://www.ebi.ac.uk).

We had 10 datasets to choose from, we have chosen the MTBLS136 dataset from all of them, which consists of a serum LC-MS datasets. These datasets consist of 949 named metabolites associated with postmenopausal hormone use. Multiple metabolites were compared in the group of women including 332 estrogen users and 337 estrogen plus progestin users versus 667 non-users. In this research, metabolite levels were compared between estrogen-only and estrogen-plus progestin users. This means it consists of the metabolite concentration of users that is supposed to be due to estrogen-only versus estrogen-plus-progestin hormones [3].

In data preprocessing, all the required packages were imported and the dataset was loaded in the jupyter notebook. The data is then divided into two tables, namely: DataTables and PeakTables using the function $cb.utils.load\_dataXL$. The DataTable and PeakTable represent two distinct types of data tables or data structures that are commonly used in metabolomics research. The DataTable typically refers to a table or data structure containing the main metabolomics data. The PeakTable contains information about detected peaks in our metabolomics data. The PeakTable consists of null values more than 20 percent removed using the $perc\_missing$ column. The DataTable2 was created using DataTable having class values 0 and 1 only. The binary class column was then assigned to an outcome vector Y. The data is then split into train and test datasets using $train - test - split$ from DataTable2 and Y. Xtrain was introduced by taking the names of metabolites from PeakList and then converted into logarithmic value and then scaled using function $cb.utils.scale$. The missing values were then imputed using the K-Nearest Neighbor technique with the k=3 neighbors. The XTestKnn was produced using Datatest similar to the XTrainKnn dataset.

4 Methods

The machine learning algorithms used in this project are described below.

4.1 Principal component regression

Principal component regression (PCR) is a linear regression model that couples Principal component analysis (PCA) and Multiple linear regression (MLR). In PCA, the training data (X matrix) is rotated and projected into lower dimensional space depending upon the orthogonal covariance called the principal components. These

principal components are treated as independent variables and in MLR, the coefficients are calculated using the least-squares method. In PCR, independently calculated PCA and MLR coefficients can be combined, and reduced to linear regression for the prediction analysis as shown in Equation 1.

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_N \quad (1)$$

Where y^* is a prediction of the model and $\beta_0, ...\beta_n$ are vectors of PCR coefficients. The only single hyperparameter in this algorithm is the number of principal components used [2].

4.2 Linear kernel support vector machine (SVM-Lin)

SVM-Lin algorithms aim to find the best hyperplane in a multi-dimensional space from the set of hyperplanes to separate N data points of a given X matrix (N $\times$ M). The direction of the hyperplane is chosen in such a way that it maximizes the margin of discrimination. It has a single tuning hyperparameter called C ("Cost" in Python library) [2].

4.3 Cross validation

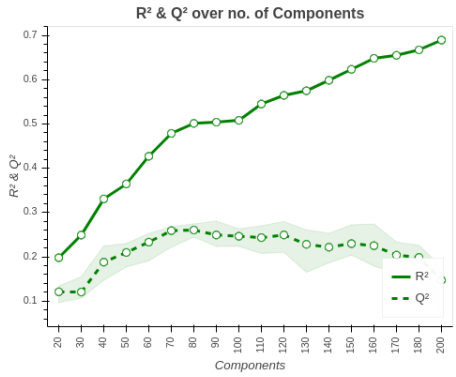
The dictionary of hyperparameters was created depending on the model used. The model is then 5-fold cross-validated using 10 Monte Carlo re-partitions, a dictionary of hyperparameters, training input data XTrainKnn, and training target data YTrain. The plots of R^2Q^2 were generated and helped to estimate the optimized hyper-parameter values where R2 is the coefficient of determination for the full data set, and Q2 is the mean coefficient of determination for cross-validated prediction data[1].

4.4 Evaluation

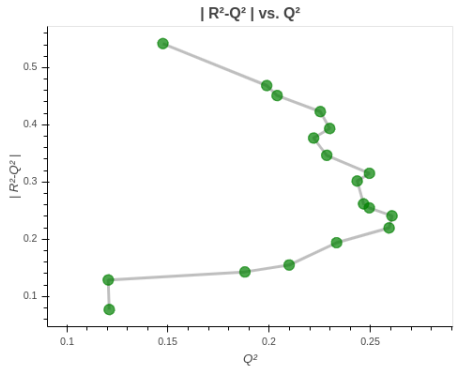
The model was then trained and tested with the training and testing datasets respectively using an optimized hyperparameter and the performance of the model was evaluated by taking the AUC values. However, the models can still suffer from sampling bias due to the small sample size and the model could provide a biased estimate of performance. For those reasons, bootstrap evaluation is done in both training and test data to measure the confidence interval in the uncertainty of prediction [2]. Before bootstrap evaluation, data was logarithmized, auto-scaled, and imputed with the k nearest neighbor technique.

4.5 Important features

Features are metabolites (bio-markers) in this report that alter or impact the estrogen and estrogen plus progestin signaling pathways. To extract the important features from each model, variables with higher coefficients were taken into account using the argsort() function which returns the indices of each variable. The top 3 metabolite indices with the highest coefficients were used to extract the important features or metabolites from the PeakTable dataset.



(a)



(b)

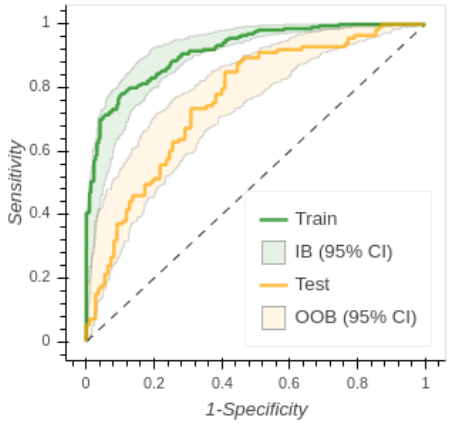
Figure 1: Optimization of hyper-parameter in PCR. a) $|R^2 - Q^2|$ vs. Q^2 curve b) R^2 and Q^2 vs. no. of components curve

5 Results and Discussions

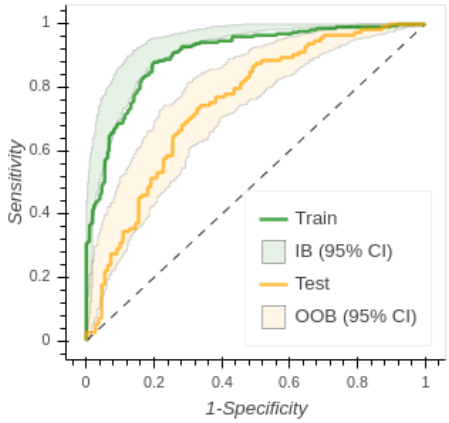
We have used PCR and SVM-lin methods only due to word limitation.

From the figure 1, curve $|R^2 - Q^2|$ vs. Q^2 is used to measure the optimized hyper-parameters by taking the point of inflection of the outer convex hull. Each point on the plot corresponds to a specific number of components tested during cross-validation. From the figure, the optimal number of components where the curve of R^2 and Q^2 changes the shape is around 80.

Figure 2 illustrates the full picture of the ROC curves for the dataset of MT-BLS136. It is also clearly seen that the sensitivity of the training dataset for both of the models is higher than the testing dataset. However, this is to be expected because the sample size is small, and complex models learn training data too well than testing data. The ROC curves are mainly used to demonstrate the performance of the model on the test sets. Even though the models are optimized using the K-fold cross-validation, the models still suffer from over-training and are proportional to the complexity of the model. We can observe that in both plots, the area under the curve of training datasets is bigger than the testing data set, and a bigger AUC means better performance on training data. When the model performs well on training data but fails to generalize to new, unseen data, we can assume that our model is overfitted or overtrained. The accuracy of the PCR model using



(a) PCR method



(b) SVM-Lin method

Figure 2: Receiver Operator Characteristic (ROC) curve of MTBLS136 dataset using the a)PCR model and b)SVM-lin model. The green color indicates the in-bag (IB) 95 % confidence interval and the yellow color indicates the out-of-bag (OOB)95 % confidence interval

the AUC value for the train and test datasets was found to be 0.92 and 0.76 respectively which are almost identical with paper[1, 2]. For the SVM-lin model, the AUC scores for the train and test datasets were 0.9 and 0.75 respectively. The higher the coefficients are, the more the implications of the metabolites. The top 3 important features extracted from PCR and SVM-lin models are shown in Table 1.

However different metabolites such as cystine, lysine, and 2-amino heptanoate, were extracted from the two models but they make a significant impact on estrogen and estrogen plus progestin hormone function pathways. Cystine is responsible for the synthesis of glutathione and is closely related to kidney stones, neurodegenerative diseases, and colorectal cancers, and the level of certain amino acids including cystine fluctuates during the postmenopausal stage in women. Lysine also decreases in the postmenopausal stage and pregnancy period in women than in the premenopausal period. However, there is still a big gap in information about the

Number	PCR	SVM-lin
1	cystine	1-linoleoyl-GPA (18:2)*
2	isobutyrylcarnitine (C4)	lysine
3	gluconate	2-aminoheptanoate

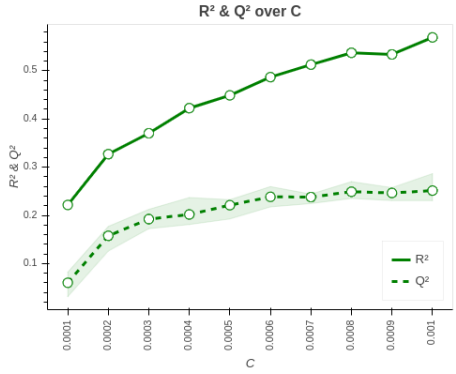
Table 1: Important features i.e. metabolites

mechanisms of those metabolites and their relations with disease treatment. both of the models have moderate performance in prediction in metabolomics datasets.

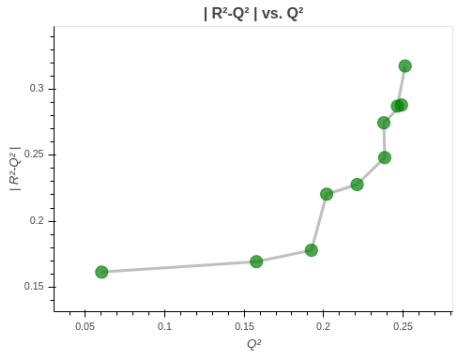
6 Contributions

Sujan Darai: Running codes and writing report.
Ibraim Ibraimi: writing and reviewing the report.
Matanat Mammadli: writing report.
Samra Hamidovic: writing report.

7 Appendix



(c)



(d)

Figure 3: Optimization of hyper-parameter in SVM-lin method. a) $|R^2 - Q^2|$ vs. Q^2 curve b) R^2 and Q^2 vs. Cost (C) curve

Author details

References

1. Kevin MMendez. CIMCB/MetabComparisonBinaryML. https://github.com/CIMCB/MetabComparisonBinaryML. [Accessed: April 24, 2024].
2. Kevin M Mendez, Stacey N Reinke, and David I Broadhurst. A comparative evaluation of the generalized predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics*, 15:1–15, 2019.
3. Victoria L Stevens, Ying Wang, Brian D Carter, Mia M Gaudet, and Susan M Gapstur. Serum metabolomic profiles associated with postmenopausal hormone use. *Metabolomics*, 14:1–14, 2018.