Using Supervised Machine Learning for Prediction of Sjogren Syndrome in Salivary Glands Based on Gene Expression

Chibly, Alejandro
Matrix and Morphogenesis Lab
National Institute of Dental and Craniofacial Research
Bethesda, MD
chiblyaa@nih.gov

***Abstract*—Current diagnosis of Sjogren Syndrome (SS) is based on a combination of histopathological features and subjective scores assigned by physicians based on symptoms. Here, a supervised machine learning approach was used to predict sjogren's syndrome based on the transcriptional profile of salivary glands. Data were obtained from GEO: GSE154926 which contains gene expression from 43 diagnosed SS patients and 7 healthy controls. PCA analysis was used to identify the most significant features, which were used to train the ML model. To identify the best ML algorith for this dataset, 6 different algorithms were tested including logistic regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Decision Tree Classifier (DTC), Gaussian Naive Bayes (NB), and Support Vector Machines (SVM) using a k-fold cross-validation approach. In the training set, all algorithms had similar performance with moderate accuracy ranging between 74 and 86%. Interestingly, DTC was the best performing method in the validation set with an accuracy of 100%. The generated DTC model was based on the expression of the genes ENSG00000166257, ENSG00000124256, and ENSG00000214262**

***Index Terms*—Machine Learning, Decision Tree Classifiers, Sjogren Syndrome**

I. Introduction

Sjogren's syndrome (SS) is an immune disorder that leads to loss of lacrimal and salivary secretions resulting in dry eye and mouth conditions. SS is often accompanied by rheumatoid arthritis and lupus, and thus also causes joint pain, swelling, and overall decreased quality of life. Currently, there is no definitive diagnostic test to confirm Sjogren Syndrome. Instead, diagnosis is based on a careful inspection of the different symptoms such as loss of saliva and tears, and requires a combined effort from multiple specialists, such as a rheumatologist, ophthalmologist or dentist and/or oral medicine specialist. Diagnosing SS requires gathering a lot of information; therefore, a combination of tests are required on the eyes and mouth, ahd samples are collected from blood, urine, and even biopsies, primarily from salivary glands. A major challenge in diagnosing SS is that it may present differently in different patients, some of which do not present the whole array of symptoms while others are more affected. This often results in incomplete or erroneous diagnosis which in turn negatively impacts the course of treatment for patients.

Machine learning algorithms have been widely applied to scientific and medical research in recent years to improve diagnosis and discover new treatments. New technologies and development of more advanced computational tools enable scientists to generate large amounts of patient data such as clinical images, sequencing data of biological samples, behavioural and demographic data, as well as social media-derived information. Analysis of these types of big data is time-consuming and often impossible to perform without computational tools. Machine Learning is a form of artificial intelligence that enables robust interrogation of complex datasets to identify otherwise hidden or undiscovered patterns and relationships in the data, and thus is extremely useful to investigate potential relationships between next-gen sequencing data and clinical outcomes.

In the current report, I explore the potential of Machine Learning algorithms to predict SS based on the transcriptional signature of minor salivary glands collected from patients. The model was train and validated in a gene expression dataset comprised of 43 salivary gland biopsies from diagnosed SS patients and 7 healthy controls. The results show that Decision Tree Classifier algorithm is able to predict SS using expression data with 80% accuracy.

II. Data Pre-Processing

A. Data Accessibility

RNA-sequencing data was obtained from Gene Omnibus (GEO) repository: GSE154926. A ready-to-use gene expression matrix and code use for analysis are also available via GitHub: https://github.com/chiblyaa/GSE154926

B. Data cleanup

Python 3.0 was used for data processing and development of the machine learning model. The dataset contained 50 samples (7 controls and 43 SS patients) organized in rows and 34476 features (genes) as columns. Low expressing genes (those present in ¡50% of samples were discarded, which reduced the matrix to 18498 features. Next, the sum of all expression values was used to calculate the library size of each sample (Figure 1).

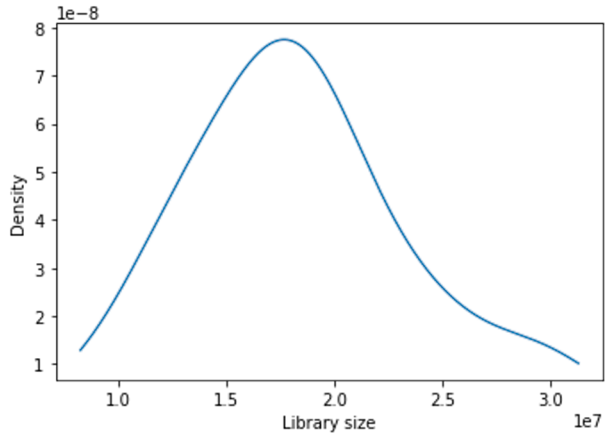

Fig. 1. Distribution of library sizes. Min: 8225201, Mean: 18367311.38, Max: 31304488

Because of a large size difference between the smallest and largest libraries, normalization was performed by dividing the raw gene expression value by the library size (sum of all expression values for each individual sample) for each sample and multiplying by 1 million. Data was then log2-transformed to ensure all genes had similar weights in the subsequent scaling and data reduction.

III. Scaling (Normalization)

To ensure the dataset was in a smaller scale and normal distribution, which are more optimal for machine learning. Data were normalized using the `MinMaxScaler()` function from scikit-learn library in Python. All expression values in the resulting matrix were in a 0-1 range (Figure 2).

	0	1	2	3	4
count	50.000000	50.000000	50.000000	50.000000	50.000000
mean	0.769855	0.402743	0.583159	0.663752	0.688069
std	0.215439	0.237416	0.219820	0.252177	0.190597
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.717316	0.243558	0.451404	0.496930	0.667824
50%	0.845107	0.344781	0.557632	0.749358	0.725365
75%	0.918337	0.565949	0.736944	0.841395	0.801783
max	1.000000	1.000000	1.000000	1.000000	1.000000

Fig. 2. Representative features post-normalization.

IV. Dimensionality Reduction

The goal of dimensionality reduction for this dataset was to obtain the 10 most relevant features for each of the identified components to generate a reduced expression matrix for the machine learning model. To be able to obtain these features from the data, dimensionality reduction was performed with PCA, which allows for interpretation. PCA was performed using `PCA(n_components=10)` to obtain only the top feature for each component.

The top identified features were:

- `#0: ENSG00000214262`
- `#1: ENSG00000120306`
- `#2: ENSG00000226054`
- `#3: ENSG00000077454`
- `#4: ENSG00000165300`
- `#5: ENSG00000124256`
- `#6: ENSG00000166257`
- `#7: ENSG00000258790`
- `#8: ENSG00000255819`
- `#9: ENSG00000111700`

K-means clustering was performed with PCA-reduced and non-reduced data to evaluate performance of data reduction. The accuracy of generated clusters was 58% before PCA and it improved to 64% after PCA.

V. K-fold cross validation of ML methods

Next, K-Fold cross-validation was used to test the performance of different machine learning algorithms to predict SS in the reduced dataset containing only the 10 relevant features determined by PCA. In order to do this, a column containing diagnosis information was was added to the reduced data, and a `target` array was created to identify the class (diagnosis) each sample belonged to. Next, the dataset was split 70:30 into training:validation subsets. The 70% split was performed with stratification based on diagnosis because of the low number of samples in the Control group (n=7).

K-fold cross validation was performed for 6 different algorithms which included both linear and non-linear methods. These were logistic regression (LR), Linear Discriminant Analysis (LDA), K-nearest Neighbor (KNN), Decision Trees (DTC), Gaussian Bayes (NB), and Support Vector Machines (SVM). These methods were selected because they represent both linear and non-linear methods. Of note, the distribution of gene expression data is assumed to be normally distributed after pre-processing and transformation; nonetheless, KNN and SVM are included in the performance test as they work well with non-normally distributed data. K-fold cross validation was performed with 3 splits of the training set. Again, one caveat of this analysis was the suboptimal number of samples in the control group, which prevented the use of a larger number of splits for cross-validation.

The resulting accuracy for each of the tested algorithms are and as follows:

- `LR: Accuracy: 0.858586 (+/- 0.035712)`
- `LDA: Accuracy: 0.744949 (+/- 0.061959)`
- `KNN: Accuracy: 0.772727 (+/- 0.075251)`
- `DTC: Accuracy: 0.744949 (+/- 0.114447)`
- `NB: Accuracy: 0.861111 (+/- 0.103935)`
- `SVM: Accuracy: 0.858586 (+/- 0.035712)`

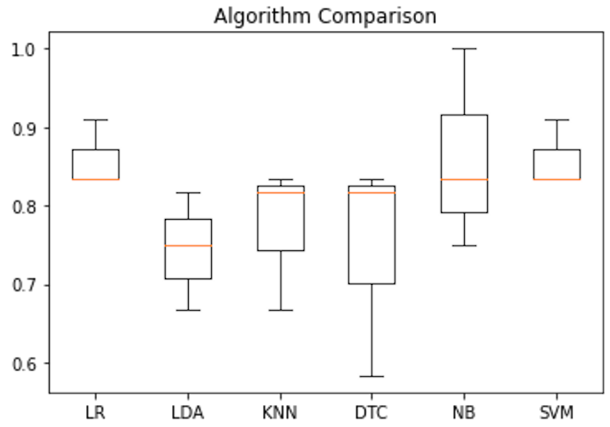

Fig. 3. Box plot of the performance of 6 machine learning algorithms. The box plot shows the median and variation from the 3 splits of k-fold cross-validation

VI. Applying SVM and DTC to predict Sjogren Syndrome

All methods had a similar performance ranging between 74 and 86% accuracy with different levels of variation in the cross-validation set (Figure 3). None of the algorithms performed reached 90% accuracy. We chose to test both SVM and DTC for comparison. Interestingly, SVM performed with 0.866 accuracy in the validation set while DTC reached 100% accuracy. Thus, it's likely that the low accuracy of DTC in the training dataset was negatively affected by the low number of samples in the control class, which was further minimized during the k-fold cross validation split.

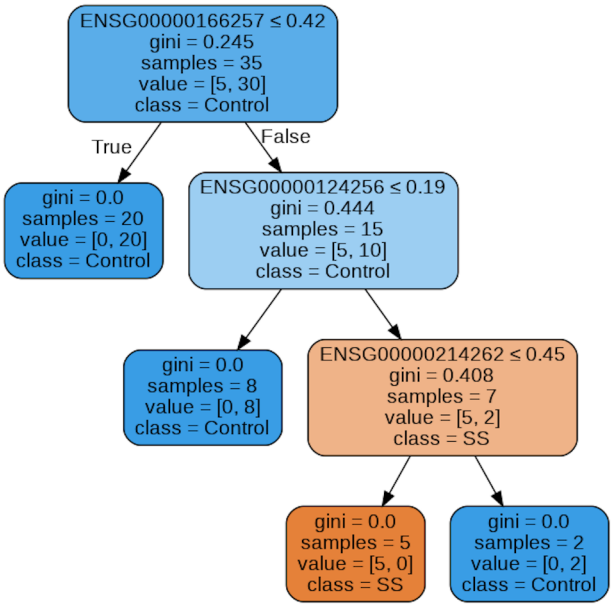

Fig. 4. Representation of the Decision Tree to predict SS based on gene expression. The prediction is based on 3 hierarchical decisions based on the expression of ENSG00000166257, ENSG00000124256, and ENSG00000214262

VII. Conclusions

All of the tested algorithms had moderate accuracy ranging between 74 and 86% in our training test and SVM and DTC were selected for validation. SVM performed similarly in the validation set while DTC reached 100% accuracy to predict SS based on the expression of 3 genes. The difference in performance in the training and validation sets suggest that we need to perform cross validation; however, the validation set is too small to accommodate an n number of splits. Nonetheless, this machine learning approach is promising and suggests that gene expression may be used to diagnose SS objectively. The features used in the decision tree correspond to the genes SCN3B, ZBP1, ANKRD26L1. Further investigation is needed to determine their role in SS.