Identification of Gene Signature Profiles of Asthma Using Machine Learning

William Huang, Stanford University

Objective

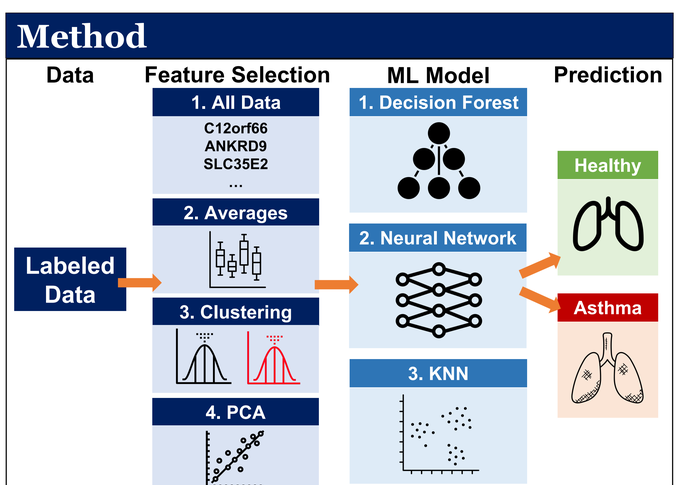
Asthma is a prevalent disease that affects over 26 million people in the US alone. Asthma's intermittent symptoms makes the disease underdiagnosed. Gene expression can be easily obtained through methods such as sampling cells in the nasal cavity or mouth, and thus a cheap and reliable classifier of asthma based on gene expression is highly desirable. We set the following goals for our research:

1. Apply multiple machine learning algorithms to identify the unique gene signature profile(s) of asthma and effectively diagnose asthma.
2. Implement a novel machine learning algorithm to improve the successful rate of prediction.
3. Genes found to be significant in the classification of asthma represent possible pathways involved in asthma that can be used in further studies, as well as the development of new treatments for asthmatic patients.

Abstract

Asthma is a prevalent disease, but it is hard to diagnose effectively. We proposed an improved machine learning algorithm as a classifier of asthma. To test our classifier, 2 datasets from the GEO (Gene Expression Omnibus) containing gene expression information on both asthmatic (80+) and control (20+) patients were used. 4 gene selection algorithms and 3 machine learning algorithms were implemented, compared, and contrasted. Machine learning algorithms such as Deep Neural Networks, Decision Forests and KNN were used to classify patients. To reduce noise in gene expressions with over ten thousand genes, a novel gene clustering algorithm was used to separate important, influential genes from unrelated genes. Machine learning models with up to 95.8% accuracy on average were created as opposed to 50-70% obtained without the new gene clustering algorithm.

Method



Method (cont.)

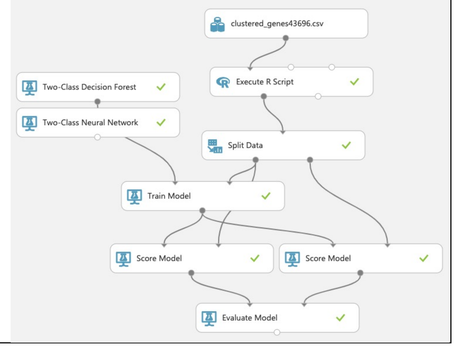
Features were selected in one of four ways:

1) Using All Genes
2) Principal Component Analysis
3) Averaged Expressions
4) A Novel Gene Clustering Algorithm

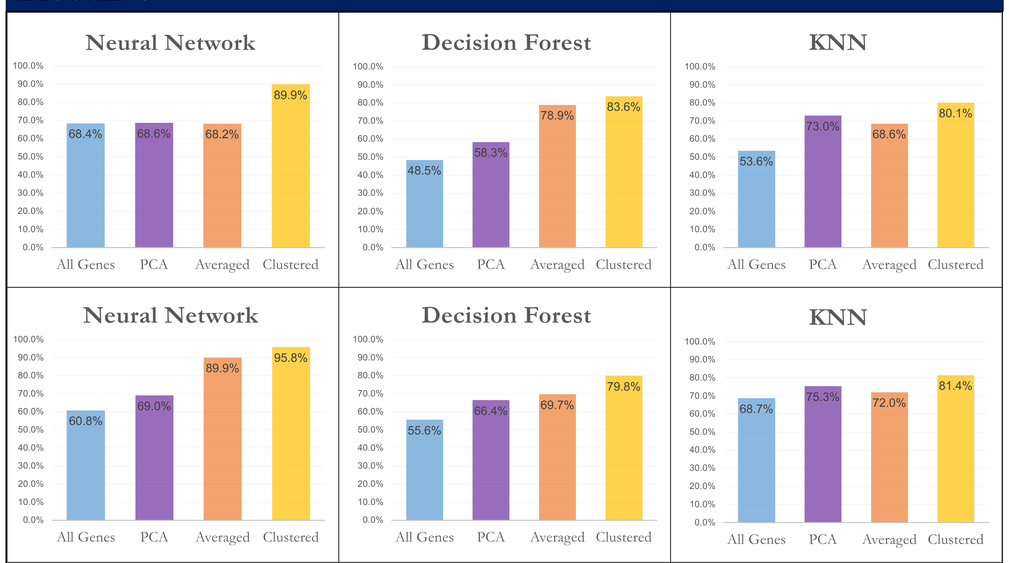
The following ML algorithms were trained:

1) Neural Network (Azure)
2) Decision Forest (Azure)
3) K-Nearest-Neighbors (Java)

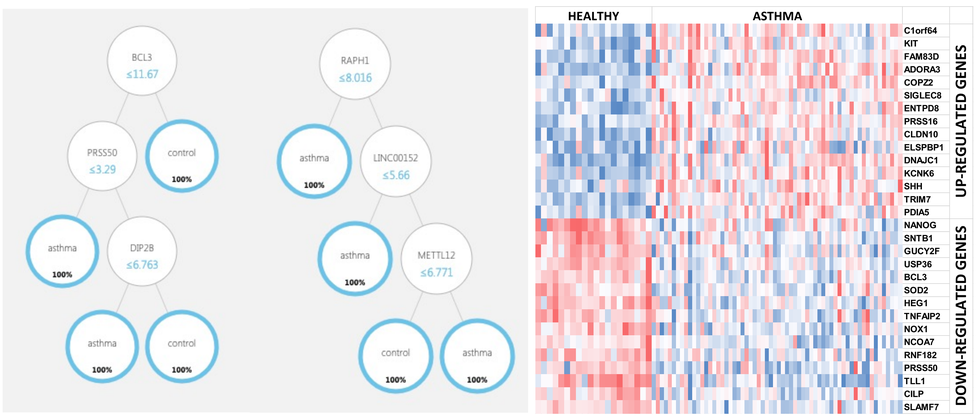
These two steps classified asthmatic patients.



Results

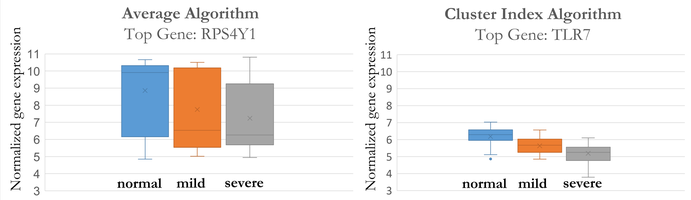


GSE 43696 (*top 3 figures*) and GSE 76262 (*bottom 3 figures*) classification results shown above. Our novel gene clustering algorithm outperforms other types of feature selection across all models.



Left: an example decision forest. *Right:* gene expression heat map; genes found by gene cluster algorithm in GSE 43696.

Interpretation



Our novel gene clustering algorithm is an efficient method to filter out significant genes. The algorithm makes a gene a higher priority if normal patients and asthma patients are clustered together (see right figure) rather than having only the average expression further apart (see left figure). In contrast, principal component analysis fails to achieve the very same effect. There are roughly 19,000 gene expressions, so choosing all genes as feature genes results in over-fitting. In addition, PCA works best with linear relations, and the relation for asthma is likely non-linear. In practice, we see PCA is unable to recover which genes are related to asthma.

Conclusion/Biotechnology Applications

The Decision Forest, Neural Network, and KNN algorithms were able to accurately classify asthma patients when presented with around 100 clustered genes. **A Neural Network machine learning classifier was able to get 95.8% accuracy using the novel gene clustering algorithm, an improvement from 60-70%.** The effectiveness of the novel gene clustering algorithm is illustrated by our heat map: a clear distinction between asthma and healthy patients is visible. Our work supports other scholarly research on asthma: the selected genes our new gene clustering algorithm match 10+ genes selected in 2018 Nature paper "A Nasal Brush-based Classifier of Asthma Identified by Machine Learning Analysis of Nasal RNA Sequence Data." We find it is not sufficient to present the entire data to machine learning; effective data processing that loses minimal information is necessary for an effective model. Clustering genes is more effective because it considers the distribution of gene expressions as well as the average difference in expression. Given the insight on possible genes related to asthma, further research can be done to explore the role of genes in asthma.

Future Directions

Further work should be done to explore how genes found by the cluster algorithm and machine learning models relate to asthma. Additionally, research into unsupervised models may perhaps find a classification system better than the current classification system based on severity of symptoms.

Acknowledgements

We thank Michele Quindipan for providing mathematical insight and helpful discussion along the way, as well as providing general research advice, guidance, and support.