https://github.com/isha-git/Alzheimers-Disease

PREDICTING SUSCEPTIBILITY TO ALZHEIMER'S DISEASE

APRIL 19, 2019

https://github.com/isha-git/Alzheimers-Disease

Title

Predicting Susceptibility to Alzheimer’s Disease using Traditional Machine Learning Algorithms

Abstract

The study aims to seek out risk-disease genes using machine learning approaches. Various algorithms like Support Vector Machines (SVM) and Decision Trees have been used along to find a suitable hypothesis for classification of candidate genes as Alzheimer’s Disease (AD) associated or unassociated.

Introduction

Alzheimer’s disease (AD) is a widespread, irreversible, progressive neurodegenerative disease, with complex genetic architecture. There is a genetic component to some cases of early-onset (before age 65) Alzheimer’s disease. Late-onset (after age 65) Alzheimer's arises from a complex series of brain changes that occur over decades. A key goal of biomedical research is to seek out disease risk genes and to elucidate the function of these risk genes in the development of the disease.

Background and Problem Description

AD is characterized by impaired memory, cognitive functioning, and changed behaviour. Other common symptoms include agitation, restlessness, withdrawal, and loss of language skills. People with this disease usually require total care during the advanced stages of the disease. Affected individuals usually survive 8 to 10 years after the appearance of symptoms, but the course of the disease can range from 1 to 25 years. Its research is based on pedigree analysis (studying the inheritance of genes), rather than candidate pathway exploration (finding possible successors who might be affected by the disease). Therefore, the understanding of AD is limited by sample size and quality, making it a challenge to have an overall insight into AD.

Related Work

So far, methods based on different data-types and different strategies have been applied in predicting AD-associated genes. Prediction methods can be roughly divided into five types

1. Methods integrating protein-protein interaction networks with information such as protein subcellular localization (predicting where a protein resides inside a cell)
2. Gene expression data (the process by which information from a gene is used in the synthesis of a functional gene product)
3. Patterns of sequence-based features shared by disease genes
4. Machine learning and network topological features
5. Information about tissue-specific networks

In past research, these methods have been applied to predict associated genes or biomarkers. But there are few reports on the predictions based on the brain gene expression data.

PROPOSED SOLUTION

AD is not caused by the role of a single gene. So, its development mechanism needs to be studied from the global point of view. The AD dataset was obtained from Alzgene archive. It has been used to classify the genes into multiple categories based on their strength of supporting evidence (the number of positive evidences of family-based studies and case-control studies). These were labelled as follows

C1: probable pathogenic genes

C2: high confidence genes

C3: related genes, and

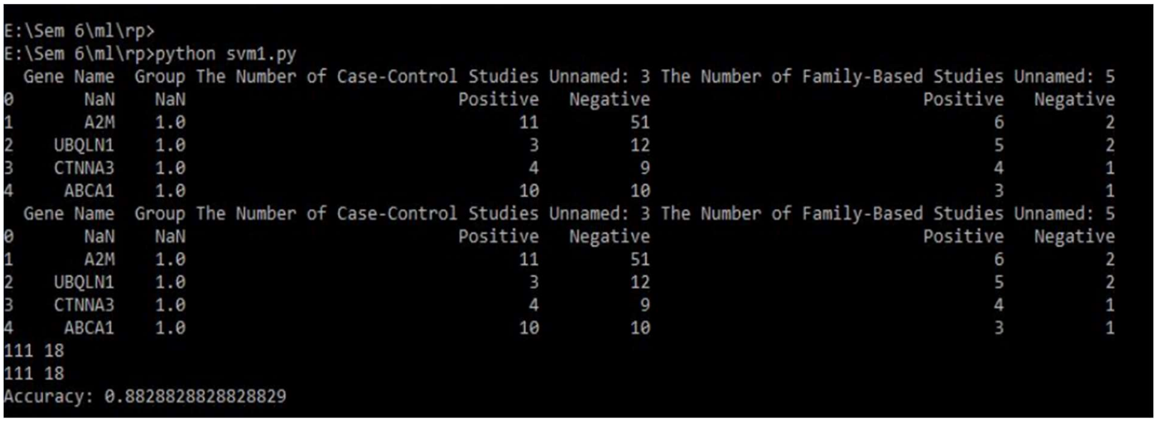
C4: possibly associated genes.

This has been implemented using Python libraries like scikit, pandas and matplotlib

EVALUATION STUDY

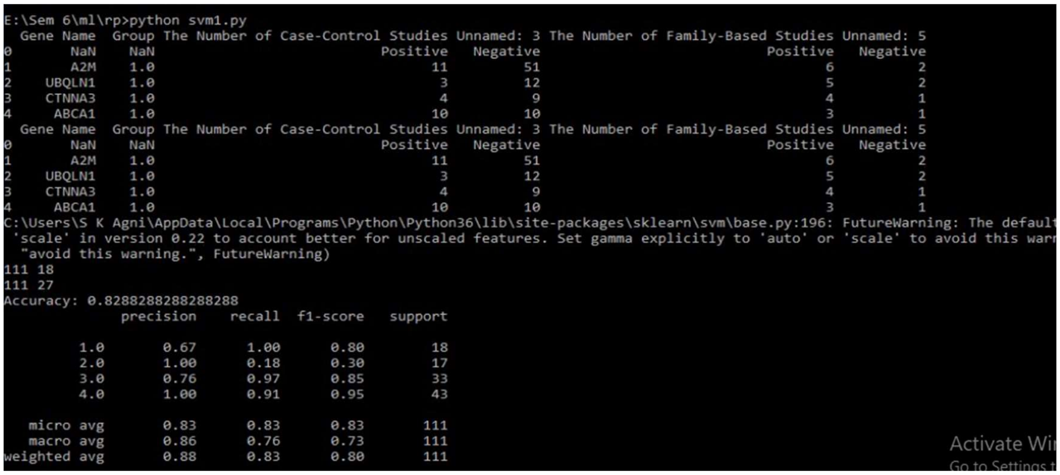
The dataset was evaluated using different kernels as well as other parameters. For instance, while applying the SVM method, Gaussian (Radial), Linear, Polynomial kernels were used to evaluate the best possible case. Cross-validation has been used to obtain optimal results. The results for 2-cross, 5-cross, 10-cross-validation have been displayed below. The Receiver Operating Characteristic (ROC) curve has also been generated for better visualization. Parameter C(=1/λ) has also been adjusted to avoid overfitting.

RESULTS



```
E:\Sem 6\ml\rp>
E:\Sem 6\ml\rp>python svm1.py
  Gene Name  Group The Number of Case-Control Studies Unnamed: 3 The Number of Family-Based Studies Unnamed: 5
0       NaN    NaN                           Positive   Negative                           Positive   Negative
1       A2M    1.0                                 11         51                                  6          2
2    UBQLN1    1.0                                  3         12                                  5          2
3    CTNNA3    1.0                                  4          9                                  4          1
4     ABCA1    1.0                                 10         10                                  3          1
  Gene Name  Group The Number of Case-Control Studies Unnamed: 3 The Number of Family-Based Studies Unnamed: 5
0       NaN    NaN                           Positive   Negative                           Positive   Negative
1       A2M    1.0                                 11         51                                  6          2
2    UBQLN1    1.0                                  3         12                                  5          2
3    CTNNA3    1.0                                  4          9                                  4          1
4     ABCA1    1.0                                 10         10                                  3          1
111 18
111 18
Accuracy: 0.8828828828828829
```

Classification using Decision Trees



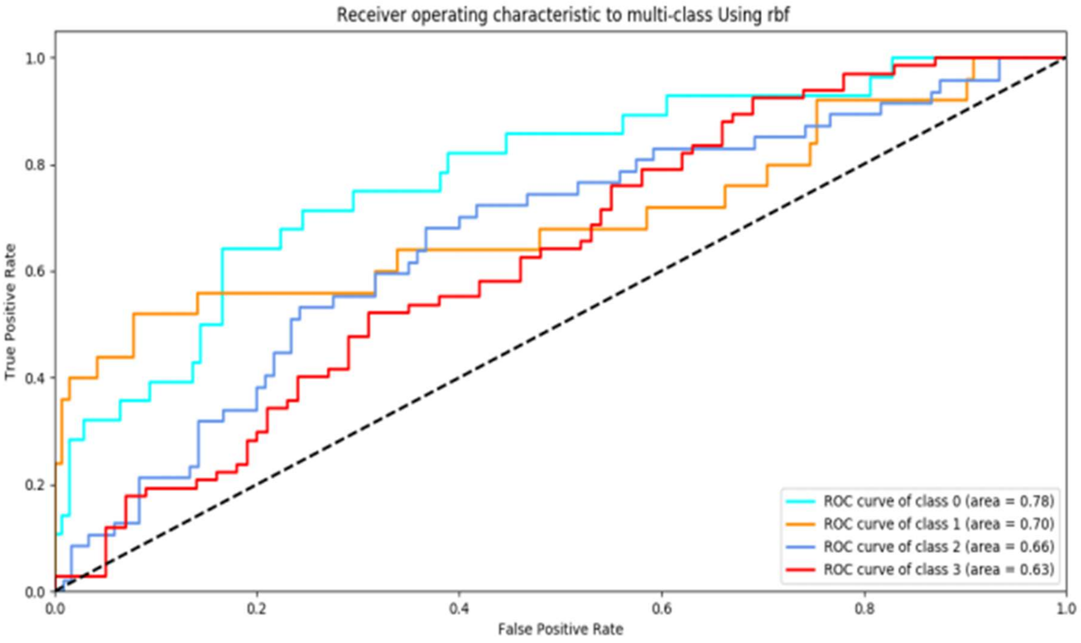
```
E:\Sem 6\ml\rp>python svm1.py
  Gene Name  Group The Number of Case-Control Studies Unnamed: 3 The Number of Family-Based Studies Unnamed: 5
0        NaN    NaN                           Positive   Negative                           Positive   Negative
1        A2M    1.0                                 11         51                                  6          2
2     UBQLN1    1.0                                  3         12                                  5          2
3     CTNNA3    1.0                                  4          9                                  4          1
4      ABCA1    1.0                                 10         10                                  3          1
  Gene Name  Group The Number of Case-Control Studies Unnamed: 3 The Number of Family-Based Studies Unnamed: 5
0        NaN    NaN                           Positive   Negative                           Positive   Negative
1        A2M    1.0                                 11         51                                  6          2
2     UBQLN1    1.0                                  3         12                                  5          2
3     CTNNA3    1.0                                  4          9                                  4          1
4      ABCA1    1.0                                 10         10                                  3          1
C:\Users\S K Agni\AppData\Local\Programs\Python\Python36\lib\site-packages\sklearn\svm\base.py:196: FutureWarning: The default
 'scale' in version 0.22 to account better for unscaled features. Set gamma explicitly to 'auto' or 'scale' to avoid this warn
  "avoid this warning.", FutureWarning)
111 18
111 27
Accuracy: 0.8288288288288288
              precision    recall  f1-score   support

         1.0       0.67      1.00      0.80        18
         2.0       1.00      0.18      0.30        17
         3.0       0.76      0.97      0.85        33
         4.0       1.00      0.91      0.95        43

   micro avg       0.83      0.83      0.83       111
   macro avg       0.86      0.76      0.73       111
weighted avg       0.88      0.83      0.80       111
```

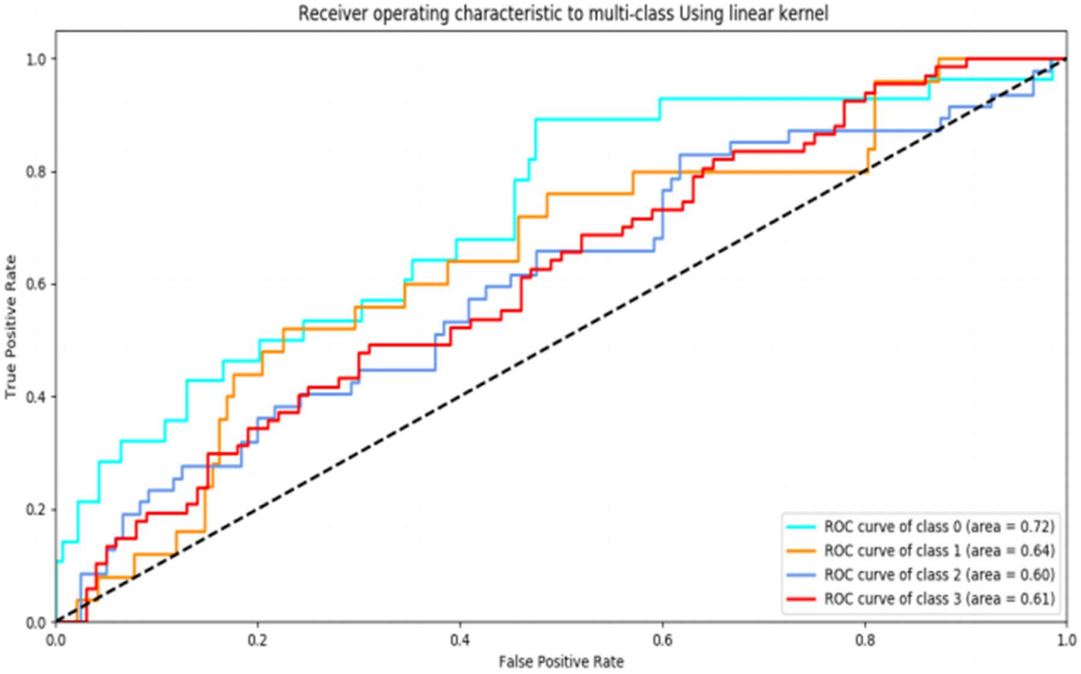
Classification using SVM (regularization parameter = 2) and 5 cross fold validation

Method used	Accuracy
SVM Library - R	84.56%
Radial Kernel (C=1)	82%
Radial Kernel (C=2)	86%
Linear Kernel	80%
Polynomial Kernel	65%
Sigmoid Kernel	57%
Decision Tree	88.29%

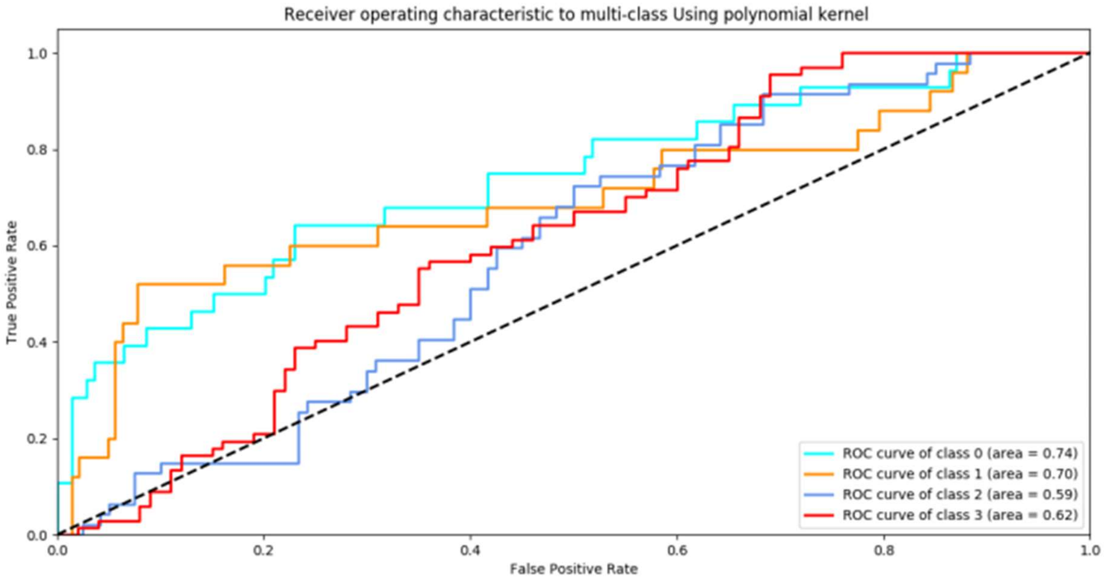
Accuracy Predicted by different hyperparameters in SVM and Decision Trees

https://github.com/isha-git/Alzheimers-Disease





https://github.com/isha-git/Alzheimers-Disease



DISCUSSIONS

The code has been implemented in Python programming language. However, the mode that the proposed research paper used was the R programming language. The coding environment has little to no impact on the final results. The dataset has been evaluated on several grounds such as testing it on polynomial, linear, Gaussian kernels. Two cross, five cross, ten cross-validation has been used.

CONCLUSION AND FUTURE DIRECTION

This study has elucidated the whole-genome spectrum of Alzheimer's Disease, using machine learning approaches. Successfully implemented the research paper in python using Scikit-learn. In the future, there can be a focus on creating a better database with more detailed information about the genes. Thus, while processing the data and performing machine learning algorithms more features can be used to have better and accurate results. Also, the increase in data points will contribute a lot in increasing the accuracy of the model.

REFERENCES

1. X. Huang, H. Liu, X. Li, L. Guan, J. Li, L. C. A. M. Tellier, H. Yang, J. Wang, and J. Zhang, "Revealing Alzheimer's disease genes spectrum in the whole-genome by machine learning", *BMC Neurology*, Vol. 18, No. 5, Jan 2018. DOI: 10.1186/s12883-017-1010-3
2. http://www.alzgene.org
3. https://ghr.nlm.nih.gov/condition/alzheimer-disease
4. https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet