CALIFORNIA STATE SCIENCE FAIR
2017 PROJECT SUMMARY

Name(s)

Michelle Li

Project Number

S0820

Project Title

A Machine Learning Approach for the Diagnosis of Parkinson's Disease via Speech Analysis

Abstract

Objectives/Goals

There exist no single, reliable methods of diagnosis for Parkinson's Disease, which motivates a machine learning approach. The purpose of this experiment is to perform a comparative analysis of various classes of machine learning algorithms to identify the best models predicative of Parkinson's. It should be reasonable to create a model achieving at least 90% accuracy and a Matthews Correlation Coefficient (MCC) of at least 0.9, which would provide a significant improvement over current methods of diagnosis.

Methods/Materials

The following equipment were used in this project: a laptop equipped with Python 3.6, a code editor, and the University of Oxford / National Center for Voice and Speech dataset for Parkinson's. My procedure trained and validated each of the following models using 10-fold cross validation: Logistic Regression, Linear Discriminant Analysis, k Nearest Neighbors, Decision Tree, Multilayer Perceptron (MLP) Neural Network, Naive Bayes, and Gradient Boost. For each of these models, I used two versions of the Oxford Dataset: the raw dataset and a scaled version. I analyzed the accuracy and MCC of these models for 3 different train-test splits: 80-20, 75-25, and 70-30. I used Python's Sci-kit Learn package for the algorithms and data processing.

Results

The two best performing models, k Nearest Neighbors and the MLP Neural Network, both produced a validation accuracy, sensitivity, specificity, ROC, and F1-score of 0.98 (98%). KNN produced a MCC of 0.94, and the Neural Network produced a MCC of 0.96 (1.0 being a perfect classifier). These models tended to perform better on the rescaled dataset than on the raw dataset, and achieved the best results with a 75-25 train-test split.

Conclusions/Discussion

Overall, my results show that KNN and MLP NN produce very robust, promising results that far exceeded my engineering goal and most literature on this subject. This suggests that a machine learning model can be implemented to significantly improve diagnosis methods of Parkinson's Disease. Not only is my machine learning method of diagnosis more reliable and robust, it is also more cost-effective (requiring only features of the patient's voice) to implement, meaning it can be utilized in less developed countries. These results are significant because millions of Parkinson's patients would benefit from a more reliable method of diagnosis.

Summary Statement

Using a Neural Network and the k Nearest Neighbors algorithm, I achieve 98% accuracy with a cost-effective method of diagnosing Parkinson's disease.

Help Received

None