IUPUI
SCHOOL OF INFORMATICS AND COMPUTING
DEPARTMENT OF BIOHEALTH INFORMATICS
Indiana University–Purdue University Indianapolis

Assessment of Parkinson's Disease Progression By Feature Relevance Analysis and Regression Techniques Using Machine Learning Algorithms

Rakesh Gullapelli, MS[1]; Josette Jones, RN, PhD[2]; Lai, Patrick T.S., MS, (PhD)[2]
School of Informatics and Computing
Indiana University Purdue University Indianapolis

ABSTRACT

Objective: To extract the relevant features from speech signal variables that contribute to UPDRS scores and to build a predictive model using machine learning algorithms to assess the Parkinson's disease progression in early stages.

Method: Feature extraction is performed using Multiple Linear Regression, Regression Decision Trees and prediction models were built using Random Forest (RF) and Support Vector Regression (SVR) techniques. K-fold cross validation is applied to test the effectives of each prediction model. Eight key predictor variables were extracted from seventeen predictor variables that contribute to UPDRS scores.

Results: The results of experimental analysis demonstrate that the proposed models were effective in predicting the UPDRS and assessing PD disease progression. Random Forest algorithm was found to be a more effective predictive model among others tested with a correlation accuracy between predicted and actual motor_UPDRS was 97.5%. Further, the RF results were compared with SVR, an advanced regression technique but RF outperformed with a squared correlation coefficient (R^2) value of 83.5.

Conclusion: This study provides an evidence of support that feature extraction and regression using machine learning techniques serves as best approach for predicting PD disease progression in early stages with non-invasive methods.

CONTACT

Rakesh Gullapelli, MS
School of Informatics,
Indiana University Purdue University Indianapolis
rakegull@iu.edu
(646)209-1016

INTRODUCTION

Remote patient tracking has been gaining increased attention due to its low-cost non-invasive methods. Unified Parkinson's Disease Rating Scale (UPDRS) is used often to track Parkinson's Disease (PD) symptoms which requires the patient's visit to the clinic and time consuming medical tests that may not be feasible for most of the elderly PD patients. One of the major concerns to predict the PD in early stages is that PD symptoms overlap with the symptoms of other diseases such as Multiple Sclerosis, Alzheimer's disease. Moreover, most of the current methods used for tracking PD rely on expert clinical raters, from which PD symptoms assessment may be difficult due to inter-individual variability.
Predicting relevant features using machine learning algorithms is helpful in providing the scientific decision-making classification rules necessary to assess the disease progression in early stages.

METHODS AND MATERIALS

Parkinson's telemonitoring dataset containing UPDRS scores was collected from UCI machine learning archives which have total voice recordings of 5875 for 42 subjects (28 women and 14 men) with early-stage PD. Dataset was preprocessed and separated into one dataset each for motor_UPDRS and one for total_UPDRS prediction. Feature extraction and regression techniques were performed on the normalized dataset using RStudio software. The detailed methodology is depicted in Figure 1.

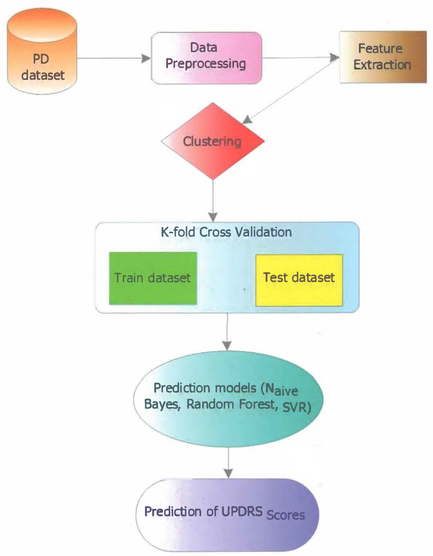

Fig 1. Methodology

RESULTS AND DISCUSSION

Multiple Linear Regression: Since the dataset has continuous variables, multi linear regression was performed on both the extracted datasets to find the relevant features (Charts 1 and 2) and coefficient matrix (Figure.2). QQ-Plots (Figure.3) and the relevance of age with total_UPDRS was studied using box plot Figure.4. The p-values were <0.05 which confirms the correlation is significant.
Regression Trees: Predictor variables from linear regression were compared by performing regression decision trees (Figure.5 and Figure.6) describing each predictor variable value for outcome variable. HNR variable has most highest value for motor_UPDRS.
Random Forest: Random forest models also extracted same variables (Figure.6 and Figure.7) like other algorithms but the correlation between predicted and actual motor_UPDRS was found to be 97.5%.
Support Vector Regression: SVR is performed only on relevant features extracted by previous algorithms to reduce the redundancies. Squared correlation coefficient (R^2) value was found to be 83.5 for motor_UPDRS score.

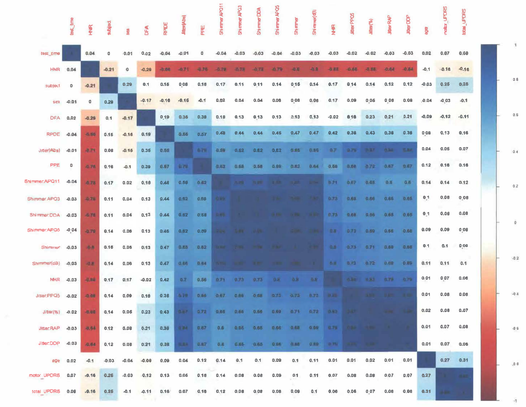

Fig 2. Correlation Matrix.

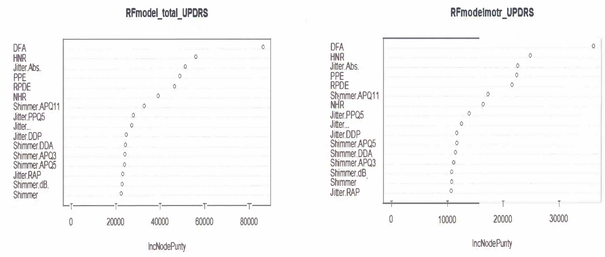

Fig 7. Random Forest features

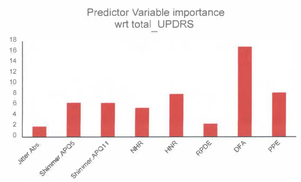

Chart 1. Predictor variables motor_UPDRS.

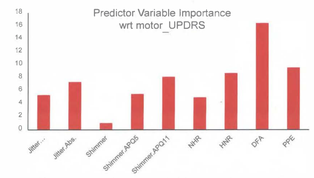

Chart 1. Predictor variables total_UPDRS.

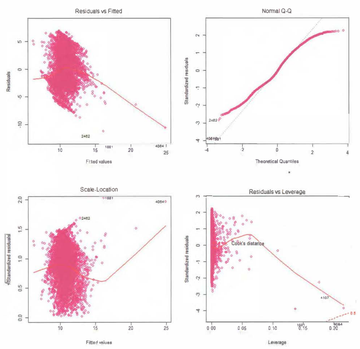

Fig 3. QQ Plot motor_UPDRS

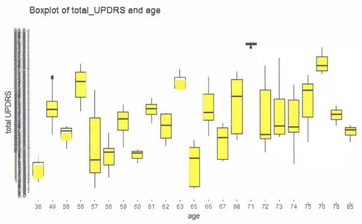

Fig 4. BoxPlot total_UPDRS vs age

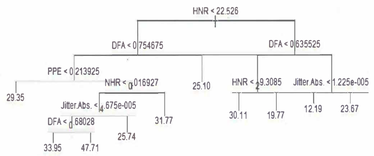

Fig 5. Regression Tree total_UPDRS.

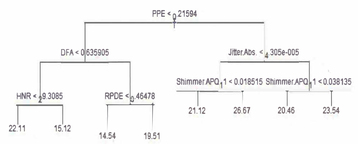

Fig 6. Regression Tree motor_UPDRS

CONCLUSION

Random Forest algorithm is found to be a more effective predictive model among others tested with a correlation accuracy between predicted and actual motor_UPDRS was 97.5%. Further, the RF results were compared with SVR, an advanced regression technique but RF outperformed SVR where a squared correlation coefficient (R^2) value was found to be 83.5.
This study provides an evidence of support that remote tracking of PD using voice variables through machine learning algorithms would enhance the clinical monitoring of elderly people and increase the chances of early diagnosis of PD.
with non-invasive methods.

REFERENCES

1. Nilashi, M., Ibrahim, O., & Ahani, A. (2016). Accuracy improvement for predicting Parkinson's disease progression. *Scientific reports*, *6*, 34181
2. Eskidere, Ö., Ertaş, F., & Hanilçi, C. (2012). A comparison of regression methods for remote tracking of Parkinson's disease progression. *Expert Systems with Applications*, *39*(5), 5523-5528.