DA 5030 INTRO DATA MINING/ MACHINE LEARNING
SPRING 2017 SEMESTER
FINAL PROJECT

SUBMITTED BY:
SUPAN KAUR

DIAGNOSIS OF PARKINSON'S DISEASE USING MULTIPLE CLASSIFICATION METHODS

ABSTRACT

Data classification techniques are proposed extensively in literature with different perspectives and understanding. One potential area of application is in exploring the out of sight patterns in the respective medical data sets. This paper deals with the application of some classification techniques to discriminate between healthy people and people with Parkinson's disease. The dataset for the disease is acquired from UCI, an online repository of large data sets. A comparative study of different five classification methods k-NN, LDA, SVM, Logistic Regression and Decision Trees is carried out to this dataset. According to confusion matrices, k-NN yields the best results with accuracy of 92%.

INTRODUCTION

Parkinson's disease (PD) is a chronic and progressive movement disorder, meaning that symptoms continue and worsen over time. Nearly one million people in the US are living with Parkinson's disease. It is a progressive disorder of the CNS that causes loss of control of movements. Early symptoms include tremors in hands, and slowed, slurred speech with symptoms developing into more uncontrollable full-body tremors. Changes in memory and cognition also occur. There is no cure for Parkinson's disease but it may be possible to slowdown or even prevent the progress of this disease by detecting it early. There are various methods of brain imaging for early Parkinson's detection test but they are expensive, there is a need for inexpensive and scalable diagnostic techniques. Recent developments include tracking postural gait and identifying anomalies in speech recordings of patients. The second technique may prove sufficiently scalable.

In the last few decades computational tools have been designed to improve the experiences and abilities of doctors and medical specialists in making decisions about their patients. With rapid changes taking place in the field of health care, decision support systems play an increasingly important role. Due to symptom overlap with other neurological diseases, only 75% of clinical diagnosis of Parkinson disease are confirmed to be Idiopathic Parkinson disease at autopsy. Classification systems can help in increasing accuracy and reliability of diagnosis and minimize the possible error as well as making the diagnosis more time efficient. The classification work on Parkinson Disease is carried out by several authors. Some of the related works are:

David Gil and Magnus Johnson [1] evaluated the performance of a classifier constructed by means of ANN and SVM. They obtained a high precision level of the confusion matrix regarding the different measurement parameters like accuracy, sensitivity, specificity, positive predictive value and negative predictive value. Another paper published by Max A. Little, et.al [2] generates an overall classification performance of 91.4%. using kernel support vector machine. Marius Ene [3] applied the Probabilistic Neural Networks (PNN) types for the classification purpose. The concrete application has provided diagnosis accuracies ranging between 79% and 81% for new, undiagnosed patients. The work done by Resul Das [4] emphasize on comparing the four classification methods. Various classifiers have been applied to recognize the Parkinson Disease by using SAS based software. Regression, DMNeural, Neural Network and Decision. As a result, Neural Network yields the best classification rate of 92.9%. R. Geetha Ramani, G. Sivagami et al.,[6] a comparative study of several algorithms on the dataset is performed. This is done by first doing the feature relevance on the dataset. Then the classifiers are implemented upon the dataset and the tree classifier yields the 100% accuracy.

In this paper, I intend to find the patterns formed by different classification algorithms to retrieve the best classifier for this specific problem and dataset. In the previous works related to this dataset, the relevant features selection using correlation, PCA or other algorithms were performed prior to implementing different classification techniques expecting to get good classification performance. However, I am applying classification algorithms without using variables selection or dimensional reduction methods hoping to achieve the optimum results. The analysis can be done by the confusion matrix which is a useful tool to detect how well the classifier recognizes the tuples of different classes. It shows the value of true positives, true negatives, false positives and false negatives. Based on the error rates the classifier accuracy can be calculated and a comparative study can be done to retrieve the best classifier.

The next section deals with the Parkinson dataset used for this research paper and the proposed classification techniques to classify patients as PD or non-PD, followed by the results and comparison of all classifier algorithms.

METHODS

The dataset chosen for work holds the features that are used to characterize healthy people from Parkinson Disease. It is available at UCI repository and was created by Max Little of the University of Oxford, in collaboration with the national Center of Voice and Speech, Denver, Colorado. The dataset is composed of a range of biomedical voice measurements from 31 people out of which 23 are diagnosed with Parkinson's disease. Each column addresses a particular voice measure and each row corresponds to one of 195 voice recordings from these individuals. The main aim of the data is to discriminate healthy people from those with Parkinson's disease, according to status column which is set to "0" for healthy and "1" for PD. It is a two decision classification problem.

Table 1. Characteristic Features Of Parkinson Dataset

Feature Number	Feature Name	Description
1	MDVP:Fo(Hz)	Average vocal fundamental frequency
2	MDVP:Fhi(Hz)	Max vocal fundamental frequency
3	MDVP:Flo(Hz)	Min vocal fundamental frequency
4	MDVP:Jitter(%)	Kay Pentax MDVP jitter on percentage
5	MDVP:Jitter(Abs)	Kay Pentax MDVP absolute jitter in microseconds
6	MDVP:RAP	Kay Pentax MDVP Relative Amplitude Pertubation
7	MDVP:PPQ	Kay Pentax MDVP five-point Period Pertubation Quotient
8	Jitter:DDP	Average absolute difference of differences between cycles divided by average period
9	MDVP:Shimmer	Kay Pentax MDVP local shimmer
10	MDVP:Shimmer(dB)	Kay Pentax MDVP local shimmer in decibels
11	Shimmer:APQ3	3 Point Amplitude Pertubation Quotient

12	Shimmer:APQ5	5 Point Amplitude Pertubation Quotient
13	MDVP:APQ	Kay Pentax MDVP eleven point Amplitude Pertubation Quotient
14	Shimmer:DDA	Average Absolute difference between consecutive differences between the amplitude of consecutive periods
15	NHR	Noise to Hormonic ratio
16	HNR	Hormonics to Noise ratio
17	RPDE	Recurrence Period Density Entropy
18	DFA	Detrended Fluctuation Analysis
19	Spread1	Non linear measure of fundamental frequency
20	Spread2	Non linear measure of fundamental frequency
21	D2	Correlation dimension
22	PPE	Pitch Period Entropy
23	Status	Health Status: 1 – Parkinson Disease 0 - Healthy

The dataset is scaled to avoid attributes in greater numeric ranges like MDVP.Fo.Hz, MDVP.Flo.Hz etc. dominating those in smaller numeric ranges. Then it is divided into two subsets, training and testing set preceded by random shuffling. The classification algorithms are applied on the training set first, and then the model is subsequently tested on the testing sets to evaluate the prediction and accuracy. All the work has been done using R software. An overview of all the classification algorithms used on Parkinson dataset are discussed here:

1. 1.k-Nearest Neighbors (k-NN)

It is a simple supervised learning, nonparametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive

integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. Pick an odd k is preferred as the odd vote can break the ties.

2. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis is a generalization of Fisher's linear discriminant to find a linear combination of features that characterizes or separates two or more classes of objects or events. Discriminant analysis seeks to generate lines that are efficient for discrimination. LDA can also be used as a supervised technique by finding a discriminant projection that maximizing between-class distance and minimizing within-class distance. It classifies n items X = x1, ... xn to one of G groups based on measurements on p predictors. Similar to linear regression except our line act to separate groups.

3. Support Vector Machine (SVM)

Support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Given a set of data points that belong to either of two classes, an SVM finds the hyperplane that 1) Leaves the largest possible fraction of points of the same class on the same side. 2) Maximizes the distance of either class from the hyperplane. 3) Find the optimal separating hyperplane that minimizes the risk of misclassifying the training samples and unseen test samples. Kernels help us turn a linear classifier into a non-linear one or vice versa.

4. Logistic Regression

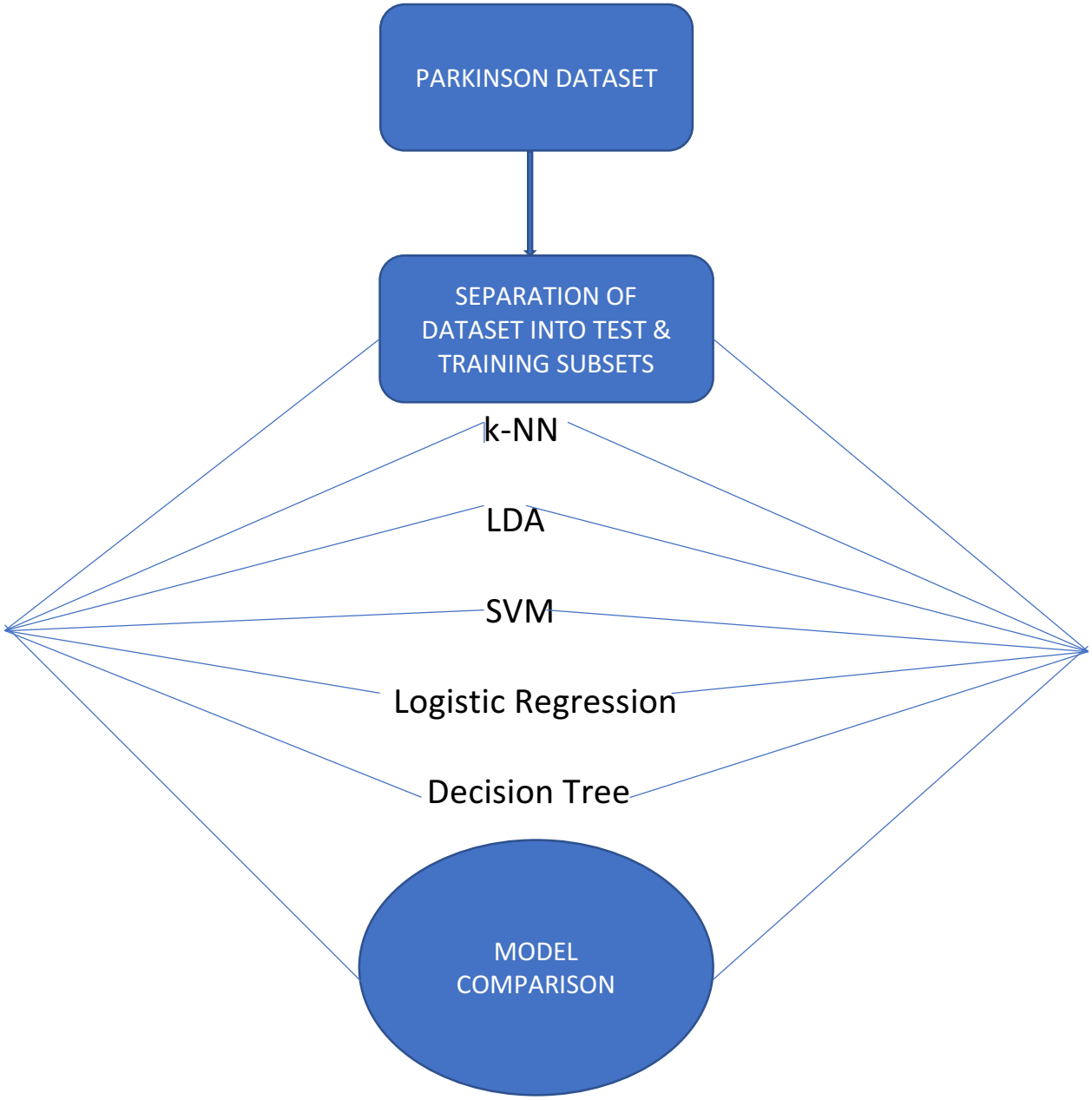
Logistic Regression is a predictive model where the class label or the target is categorical. This variable owns two categories, yes / no. For this Parkinson Disease dataset, the category will be whether the person is affected with Parkinson Disease or the person is a non-Parkinson Disease. Logistic regression can be used only with two types of target variables. 1. A categorical target variable that has exactly two categories (i.e., a binary or dichotomous variable). 2. A continuous target variable that has values in the range 0.0 to 1.0 representing probability values or proportions.

5. Decision Trees

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their outcomes. The decision tree can be linearized into decision

rules, where the outcome is the contents of the leaf node, and the conditions along the path form a conjunction in the if clause. In general, the rules have the form: If condition 1 and condition2 and condition3 then outcome. Each node in the tree is a decisions/tests. Each path from the tree root to a leaf corresponds to a conjunction of attribute decisions/tests. The tree itself corresponds to a disjunction of these conjunctions.

Fig.1 The applied methods for PD classification



RESULTS

Data was shuffled, scaled except the status attribute and then divided into two sections of 100 and 95 instances each. The algorithm was trained on one section and then tested on another section. The learner's performance on the test set was compared against the expected values, and a confusion matrix was generated consisting of the number of true positives, true negatives, false positives and false negatives.

	Truth: PD	Truth: Healthy
Predicted: PD	True positives	False positives
Predicted: Healthy	False negatives	True negatives

Table 2. A sample confusion matrix

These values were used to calculate accuracy, recall, precision and F-score as shown in equations 1.1 through 1.4

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \quad 1.1$$

$$\text{Recall} = TP/(TP + FN) \quad 1.2$$

$$\text{Precision} = TP/(TP + FP) \quad 1.3$$

$$\text{F-score} = 2*\text{precision}*\text{recall}/(\text{precision} + \text{recall}) \quad 1.4$$

Table 3. Confusion Matrix comparing the results of all classifiers

ALGORITHM	CONFUSION MATRIX		
k-NN		Parkinson	Healthy
	Parkinson	69	1
	Healthy	7	18
LDA		Parkinson	Healthy
	Parkinson	66	9
	Healthy	4	16

SVM		Parkinson	Healthy
	Parkinson	69	15
	Healthy	1	10
Logistic Regression		Parkinson	Healthy
	Parkinson	57	12
	Healthy	13	13
Decision Tree (C4.5)		Parkinson	Healthy
	Parkinson	62	8
	Healthy	3	22

Table 4. Performance of all classifiers

Algorithm	Accuracy	Recall	Precision	F-score
k-NN	0.92	0.91	0.99	0.95
LDA	0.86	0.94	0.88	0.91
SVM	0.83	0.99	0.82	0.90
Logistic Regression	0.74	0.81	0.83	0.82
Decision Trees	0.88	0.95	0.89	0.92

DISCUSSION

Parkinson's Disease is often difficult to diagnosis, but even at early stages, small vocal differences may be machine-detectable. Using this information, it becomes possible to predict PD using voice recordings from potential patients. Several machine learning approaches are effective for this task, with F-Scores above 0.90.

The confusion matrices and overall performance of each classifier are present in Table 3 and Table 4 respectively. The top three algorithms that did good job with highest accuracy and F-score are k-NN, Decision Trees and LDA. The k-NN with k value as 5, 10, 15 and 20 were tried and the least error rate was with k = 5. For SVM, the best model was selected with the best kernel and cost before running it on the test data. Among all classifiers, Logistic Regression classifies the dataset with the least accuracy of 74% and F-score of 0.82. However, the most effective

approach is k-NN means algorithm, with an F-Score of 0.95 and accuracy of 92%. It has the highest precision too.

The paper is intended to verify the effectiveness of the application of various classifiers to the Parkinson Dataset. A comparative study of five algorithms on the dataset is performed. In this research paper, k-NN classifier yields the 92% accuracy. Since other works included the attributes selection or dimensional reduction before applying classifier algorithms, this work was done with the understanding that past results would not be exactly replicated.

The future work might extend to combine the different datasets with same features and then apply classifier algorithms. Or use of the algorithms to select the best attributes prior to classifying the dataset.

REFERENCES

[1] Gil, D., Manuel, D. 2009 Diagnosing Parkinson by using Artificial Neural Networks and Support Vector Machines.

[2]'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection', Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)

[3] Marius Ene. 2008. Neural Network based approach to discriminate healthy people from those with Parkinson's disease.

[4] Resul Das. 2010. A comparison of multiple classification methods for diagnosis of Parkinson disease.

[5] Parkinson's Disease Foundation, http://www.pdf.org/en/about_pd, (2011.09.16).

[6] Dr. R.Geetha Ramani and G.Sivagami "Parkinson Disease Classification using Data Mining Algorithms", International Journal of Computer Applications (IJCA),Vol-32,No.9, October 2011

[7] National Institute of Neurological Disorders and Stroke, http://www.ninds.nih.gov/disorders/parkinsons_disease/detail_parkinsons_disease.htm, (2011.09.16).

[8] Gil, D. and Johnson, M. (2009) Diagnosing Parkinson by Using Artificial Neural Networks and Support Vector Machines. Global Journal of Compute Science and Technology, 9.

[9] Marius Ene. 2008. Neural Network based approach to discriminate healthy people from those with Parkinson's disease.

[10] Mehmet Fatih CAGLAR, Bayram CETISLI, Inayet Burcu TOPRAK. 2010. Automatic Recognition of Parkinson 's disease from Sustained Phonation Tests using ANN and Adaptive Neuro-Fuzzy Classifier .

[11] UCI Machine Learning Repository- Center for Machine Learning and Intelligent System. 'http://archive.ics.uci.edu/ml/machine-learning-databases/parkinsons/parkinsons.data'