International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 2 (2019) pp. 452-458

Performance Analysis of Classification algorithms on Parkinson's Dataset with Voice Attributes

T.Swapna

Department of Computer Science and Engineering,
G. Narayanamma Institute of Technology & Science, Hyderabad, India.

Y.Sravani Devi

Department of Computer Science and Engineering,
G. Narayanamma Institute of Technology & Science, Hyderabad, India.

Abstract

Parkinson's disease or Parkinsonism is degenerative disease of the brain that causes tremors, particularly in the elderly. In this disorder of nervous system, there is an increasing effect on body movements. It may begin as a hardly noticeable tremor initially in just one hand and rapidly progress to severe shaking of hands, making it difficult to hold a glass in the hand. Though the disease is characterized by tremors, it may also cause slow movements or stiffness.

Machine learning is a method or a field used to conceive complicated models and algorithms for predictive analysis. The different analysis models help researchers, data scientists, engineers, and analysts to produce proper results and decisions, has numerous applications such as building predictive models which can be extremely beneficial in the healthcare industry. In order to help in the diagnosis of Parkinson, classification techniques can be applied to classify real time Parkinson data based on an established training set. In this paper, a comparative study on different classification methods is carried out to this dataset and the Accuracy Analysis to come up with the best classification rule. Also the intention is to sieve the data such that the healthy and people with Parkinson will be correctly classified.

Keywords: Machine Learning, Parkinson, Speech Signal, Decision Trees, Linear support vector, Naïve Bayes, KNN, Accuracy, Recall,Support,Precision,f1-score.

INTRODUCTION

Parkinson is a neurological disease and occurs due to lack of dopamine neurons. These dopamine neurons manage all body movements. Parkinson patients have difficulty in doing all daily routine activities, and also have disturbed vocal fold movements. Using voice analysis disease can be diagnosed remotely at an early stage with more reliability and in an economic way. Parkinson's disease is a progressive nervous system disorder that affects movement. Symptoms start gradually, sometimes starting with a barely noticeable tremor in just one hand. Tremors are common, but the disorder also commonly causes stiffness or slowing of movement. In the early stages of Parkinson's disease, face may show little or no expression, arms may not swing while walking. Speech may become soft or slurred. Parkinson's disease symptoms worsen the person's condition progresses over time. Although Parkinson's disease can't be cured, medications might significantly improve symptoms. Occasionally, doctor may suggest surgery to regulate certain regions of brain to improve symptoms. Parkinson's disease signs and symptoms can be different for everyone. Early signs may be mild and go unnoticed. Symptoms often begin on one side of the body and usually remain worse on that side, even after symptoms begin to affect both sides.

Parkinson's signs and symptoms may include:

- **Tremor:** A tremor, or shaking, usually begins in a limb, then hand or fingers.
- **Slowed movement (bradykinesia):** Over time, Parkinson's disease may slow body movement, making simple tasks difficult and time-consuming.
- **Rigid muscles:** Muscle stiffness may occur in any part of the body. The stiff muscles can be painful and limit range of motion.
- **Impaired posture and balance:** Posture may become stooped, or may have balance problems as a result of Parkinson's disease.
- **Loss of automatic movements:** Person may have a decreased ability to perform unconscious movements, including blinking, smiling or swinging arms while walking.
- **Speech changes:** may speak softly, quickly, slur or hesitate before talking. Speech may be more of a monotone rather than with the usual inflections.
- **Writing changes:** It may become hard to write, and your writing may appear small.

In Parkinson's disease, certain nerve cells (neurons) in the brain gradually break down or die. Many of the symptoms are due to a loss of neurons that produce a chemical messenger in your brain called dopamine. When dopamine levels decrease,

it causes abnormal brain activity, leading to symptoms of Parkinson's disease.

The cause of Parkinson's disease is unknown, but several factors appear to play a role, including:

- **Your genes.** Researchers have identified specific genetic mutations that can cause Parkinson's disease. But these are uncommon except in rare cases with many family members affected by Parkinson's disease.

  However, certain gene variations appear to increase the risk of Parkinson's disease but with a relatively small risk of Parkinson's disease for each of these genetic markers.

- **Environmental triggers.** Exposure to certain toxins or environmental factors may increase the risk of later Parkinson's disease, but the risk is relatively small.

Researchers have also noted that many changes occur in the brains of people with Parkinson's disease, although it's not clear why these changes occur. These changes include:

- **The presence of Lewy bodies.** Clumps of specific substances within brain cells are microscopic markers of Parkinson's disease. These are called Lewy bodies, and researchers believe these Lewy bodies hold an important clue to the cause of Parkinson's disease.

- **Alpha-synuclein is found within Lewy bodies.** Although many substances are found within Lewy bodies, scientists believe an important one is the natural and widespread protein called alpha-synuclein (a-synuclein). It's found in all Lewy bodies in a clumped form that cells can't break down. This is currently an important focus among Parkinson's disease researchers.

Risk factors for Parkinson's disease include:

- **Age.** Young adults rarely experience Parkinson's disease. It ordinarily begins in middle or late life, and the risk increases with age. People usually develop the disease around age 60 or older.

- **Heredity.** Having a close relative with Parkinson's disease increases the chances that you'll develop the disease. However, your risks are still small unless you have many relatives in your family with Parkinson's disease.

- **Sex.** Men are more likely to develop Parkinson's disease than are women.

- **Exposure to toxins.** Ongoing exposure to herbicides and pesticides may slightly increase your risk of Parkinson's disease.

Parkinson's disease is often accompanied by these additional problems, which may be treatable:

- **Thinking difficulties.** One may experience cognitive problems (dementia) and thinking difficulties. These usually occur in the later stages of Parkinson's disease. Such cognitive problems aren't very responsive to medications.

- **Depression and emotional changes.** One may experience depression, sometimes in the very early stages. Receiving treatment for depression can make it easier to handle the other challenges of Parkinson's disease. One may also experience other emotional changes, such as fear, anxiety or loss of motivation. Doctors may give you medications to treat these symptoms.

- **Swallowing problems.** Difficulties with swallowing as the condition progresses. Saliva may accumulate in the mouth due to slowed swallowing, leading to drooling.

- **Chewing and eating problems.** Late-stage Parkinson's disease affects the muscles in your mouth, making chewing difficult. This can lead to choking and poor nutrition.

- **Sleep problems and sleep disorders.** People with Parkinson's disease often have sleep problems, including waking up frequently throughout the night, waking up early or falling asleep during the day.

People may also experience rapid eye movement sleep behavior disorder, which involves acting out your dreams. Medications may help your sleep problems.

- **Bladder problems.** Parkinson's disease may cause bladder problems, including being unable to control urine or having difficulty urinating.

- **Constipation.** Many people with Parkinson's disease develop constipation, mainly due to a slower digestive tract.

The effected may also has to experience:

- **Blood pressure changes.**
- **Smell dysfunction.**
- **Fatigue.**
- **Pain.**

Because the cause of Parkinson's is unknown, proven ways to prevent the disease also remain a mystery. Some research has shown that regular aerobic exercise might reduce the risk of Parkinson's disease.

Some other research has shown that people who drink caffeine — which is found in coffee, tea and cola — get Parkinson's disease less often than those who don't drink it. However, it is still not known whether caffeine actually protects against getting Parkinson's, or is related in some other way. Currently there is not enough evidence to suggest drinking caffeinated beverages to protect against Parkinson's. Green tea is also related to a reduced risk of developing Parkinson's disease.

DATA COLLECTION

The Parkinson data set has been referenced from UCI Repository [5]. The database consists of 756 samples. It has 754 attributes including the Class label attribute. The data used in this study were gathered from 188 patients with PD (107 men and 81 women) with ages ranging from 33 to 87 at the Department of Neurology in CerrahpaÅŸa Faculty of Medicine, Istanbul University. The control group consists of 64 healthy individuals (23 men and 41 women) with ages varying between 41 and 82.

Attributes are derived using Various speech signal processing algorithms including Time Frequency Features, Mel Frequency Cepstral Coefficients (MFCCs), Wavelet Transform based Features, Vocal Fold Features and TWQT features have been applied to the speech recordings of

Parkinson's disease (PD) patients to extract clinically useful information for PD assessment. Machine Learning Algorithms have been applied on this dataset to classify the records into two categories: Parkinson effected or not which are the values of the Class Label 1 or 0.

The attribute set of Parkinson diseases is as follows.

Attribute information is depicted in Table 1.

Table 1

Feature	Attribute	Description
1. Base line features	Detrended Fluctuation Analysis(DFA)	It is a method for determining the statistical self-affinity of a signal.
	PPE	It is robust to many uncontrollable confounding effects including noisy acoustic environments and normal, healthy variations in voice frequency.
	RPDA	Recurrence Period Density Entropy determines the periodicity, or repetitiveness of a signal.
	Jitter (%) Jitter (Abs)	Several measures of variation in fundamental frequency
	Shimmer : Shimmer (dB) Shimmer: APQ3 Shimmer: APQ5	Several measures of variation in amplitude
2. Intensity Parameters	Min intensity, Max intensity, Mean Intensity	To know the loudness effect of *speech signal*.
3. Formant Frequencies	f1, f2,f3,f4 f1: first formant 500 Hz f2: second formant1500 Hz f3 :third formant 2500 Hz Etc.	Each of the preferred resonating frequencies of the vocal tract (each bump in the frequency response curve) is known as a formant.
4. Band Width Parameters	b1,b2,b3,b4	Bandwidth is a measure of frequency band of a sound, especially a resonance. Bandwidth is determined at the half-power (3 dB down) points of the frequency response curve
5. Vocal fold	22 attributes measured based on vocal fold.	Experiments show that wavelet transform can improve the frequency characteristics of signal, and compress the dimension of characteristics space, and it has very good classification effect of speech signal.
6. MFCC	Mel-frequency cepstral coefficients (MFCC) have traditionally been used in speaker identification applications. MFCC 12 coefficients are calculated from the input signal	calculation of MFCC does not require pitch detection and these parameters have been shown to be fairly robust against some kinds of voice distortion
7. TQWT features	431 different parameters are calculated from TQWT.	TQWT decomposes EMG signal into sub-bands and these sub-bands are used for extraction of statistical features namely mean absolute deviation (MAD), interquartile range (IQR), kurtosis, mode, and entropy.

CLASSIFICATION

In machine learning and statistics, classification algorithms classify tuples into a set of categories [6]. It is a supervised learning approach in which the computer program learns from the data input (i.e. trained example) given to it and then uses this learning to classify new observation based on classification rules. The input data set may be either a bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Classification algorithms have various practical applications in different fields such as bioinformatics, natural language processing, market segmentation and text categorization. It is used for speech recognition, facial detection, filtering spam messages, handwriting recognition, understanding spoken language, bio metric identification, document classification etc.

In this work the following algorithms are implemented on Parkinson dataset.

Naive Bayes Classifier

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors [8]. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Random Forest

Random forests or random decision forests are an ensemble learning method. It is used mainly used to solve classification, regression problems and also other problems. Random forest is one of the accurate learning algorithm. The basic concept of the algorithm is to build many small decision-tree and then merging them to form a forest. It is computationally easy and cheap process to build many such small and weak decision trees. So such decision trees can be formed in parallel and then it can be combined to form a single and strong forest. The algorithm for random forests uses the common technique of bootstrap bagging.

Neural networks

In this work, back propagation neural network (Mandal and Sairam 2011; Kuo and Yang 2012) is trained with a sigmoid function Z(t) ¼ 1/p(1 þ e t) along with fuzzy logic. For training purpose, dynamic learning rate is used. Dynamic learning used along with momentum is applied to the weights while updating, which increases the efficiency of the network as discussed in Mandal (2010b).

The proposed algorithm is as follows:

1) Provide training sample to NN.
2) Compute error from difference between network output and expected output.
3) For every neuron, compute how much to modify the weight (local error).
4) Optimize weight to reduce local error.
5) Compute 'blame' on each error.
6) Repeat from 3.

Decision Tree Classifier

Decision tree [7] builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

KNN Classifier

KNN classifier is an instance-based learning Algorithm which is based on a distance function for pairs of observations, such as the Euclidean distance or Cosine. In this paradigm, k nearest neighbors of a training data is computed first. Then the similarities of one sample from testing data to the k nearest neighbors are aggregated according to the class of the neighbors, and the testing sample is assigned to the most similar class.[10]

AdaBoost Classifier

AdaBoost or Adaptive Boosting is one of ensemble boosting classifier proposed by Yoav Freund and Robert Schapire in 1996. It combines multiple classifiers to increase the accuracy of classifiers. AdaBoost is an iterative ensemble method. AdaBoost classifier builds a strong classifier by combining multiple poorly performing classifiers so that you will get high accuracy strong classifier. The basic concept behind Adaboost is to set the weights of classifiers and training the data sample in each iteration such that it ensures the accurate predictions of unusual observations.

Linear Support Vector Machine

A Support Vector Machine (SVM) [11] is a classifier generally represented by separating the hyper plane. The algorithm classifies the output from the given set of training data into a hyper plane categorizing the data. This hyper plane is a line which divides the plane into two parts where each class lay on either side of the plane. Linear algebra can be used for transforming the given problem into the hyper plane

of linear SVM. Kernel plays a very important role here. For Linear Kernel new input can be predicted using dot product between the input (x) and each support vector (xi), which can be calculated as follows

f(x) = B(0) + sum(ai * (x,xi))

Since the considered Parkinson's dataset has 754 features, as SVM involves lot of computations, PCA a dimensionality reduction algorithm is used on the original dataset for dimensionality reduction.In this work PCA with two components has given maximum accuracy.

CLASSIFICATION PERFORMANCE MEASURES

The performance of the above-mentioned classification technique can be calculated by the following metrics [9]:

A Confusion Matrix:

The Confusion matrix is one of the easiest metrics used for finding the correctness and accuracy of the model. It is used for Classification problem where the output can be of two or more types of classes. The Confusion matrix in itself is not a performance measure as such, but almost all of the performance metrics are based on Confusion Matrix and the numbers inside it.

The following terms are associated with the confusion matrix.

True Positives (TP): True positives are the cases when the actual class of the data point was 1(True) and the predicted is also 1(True)

True Negatives (TN): True negatives are the cases when the actual class of the data point was 0(False) and the predicted is also 0(False)

False Positives (FP): False positives are the cases when the actual class of the data point was 0(False) and the predicted is 1(True). False is because the model has predicted incorrectly and positive because the class predicted was a positive one. (1)

False Negatives (FN): False negatives are the cases when the actual class of the data point was 1(True) and the predicted is 0(False). False is because the model has predicted incorrectly and negative because the class predicted was a negative one.

B. Precision:

It is the measure in which the fraction of true positives in contrary to all positive results is calculated.

$$Precision = \frac{frequency\ of\ true\ \ positives}{frequency\ of\ true\ positives + frequency\ of\ \ false\ positives}$$

C. Accuracy

The percentage of the test tuples that are properly classified by the classifiers is nothing but the accuracy of the particular algorithm in hand.

$$Accuracy = \frac{frequency\ of\ true\ \ positives + frequency\ of\ true\ negatives}{frequency\ of\ true\ positives + \ false\ negatives + \ false\ positives + true\ negatives}$$

D. Recall

It is the ratio of the number of relevant tuples obtained to the total number of relevant tuples in the data set. It is usually expressed as a percentage.

Recall = 100*X/(X+Y)

X= number of relevant tuples obtained

Y=number of relevant tuples not obtained

E. F1 Score

It is the weighted average of precision and recall. In other words, it conveys the balance between precision and recall.

F1 Score =$(2*Precision*Recall)/(Precision+Recall)$

EXPERIMENTAL RESULTS:

The following table depicts the confusion matrix and the other classification performance measures of various classification algorithms on the Parkinson's dataset respectively.

Table 2

<table>
<tr><th>Algorithm</th><th></th><th>Accuracy</th></tr>
<tr><td rowspan="2">Naive Bayes Classifier</td><td>
Confusion Matrix
<table>
<tr><td></td><td>Predicted(Yes)</td><td>Predicted(No)</td></tr>
<tr><td>Actual(Yes)</td><td>99</td><td>14</td></tr>
<tr><td>Actual(No)</td><td>24</td><td>15</td></tr>
</table>
Metrics
<table>
<tr><td></td><td>Precision</td><td>Recall</td><td>F1 Score</td><td>Support</td></tr>
<tr><td>0</td><td>0.52</td><td>0.38</td><td>0.44</td><td>39</td></tr>
<tr><td>1</td><td>0.80</td><td>0.88</td><td>0.84</td><td>113</td></tr>
<tr><td>Avg/Total</td><td>0.73</td><td>0.75</td><td>0.74</td><td>152</td></tr>
</table>
</td><td>70.82%</td></tr>
<tr><td>Random Forest</td><td>
Confusion Matrix
<table>
<tr><td></td><td>Predicted(Yes)</td><td>Predicted(No)</td></tr>
<tr><td>Actual(Yes)</td><td>107</td><td>3</td></tr>
<tr><td>Actual(No)</td><td>30</td><td>12</td></tr>
</table>
</td><td></td></tr>
</table>

Metrics

	Precision	Recall	F1 Score	Support
0	0.80	0.29	0.42	42
1	0.78	0.97	0.87	110
Avg/Total	0.79	0.78	0.74	152

78.56%

MLP Classifier

Confusion Matrix

	Predicted(Yes)	Predicted(No)
Actual(Yes)	115	0
Actual(No)	0	37

76.72%

Metrics

	Precision	Recall	F1 Score	Support
0	0..00	0..00	0..00	32
1	0.76	1.00	0.86	115
Avg/Total	0.57	0.76	0.65	152

Decision Tree

	Predicted(Yes)	Predicted(No)
Actual(Yes)	93	22
Actual(No)	12	25

Confusion Matrix

Metrics

	Precision	Recall	F1 Score	Support
0	0.53	0.68	0.60	37
1	0.89	0.81	0 0.85	115
Avg/Total	0.80	0.78	0.78	152

77.63%

KNN Classifier

Confusion Matrix

	Predicted(Yes)	Predicted(No)
Actual(Yes)	102	14
Actual(No)	25	11

Metrics

	Precision	Recall	F1 Score	Support
0	0.44	0.31	0.36	36
1	0.80	0.88	0 0.84	116
Avg/Total	0.72	0.74	0 0.73	152

74.89%

AdaBoost

Confusion Matrix

	Predicted(Yes)	Predicted(No)
Actual(Yes)	108	8
Actual(No)	12	24

Metrics

	Precision	Recall	F1 Score	Support
0	0.75	0.67	0.71	36
1	0.90	0.93	0.92	116
Avg/Total	0.86	0.87	0.87	152

76.56%

SVM

Confusion Matrix

	Predicted(Yes)	Predicted(No)
Actual(Yes)	110	6
Actual(No)	36	0

Metrics

	Precision	Recall	F1 Score	Support
0	0.00	0.00	0.00	36
1	0.75	0.95	0.84	116
Avg/Total	0.57	0.72	0 0.64	152

72.76%

RESULT ANALYSIS

This paper deals with the application of seven classification algorithms on the acquired data set and then drawing out a comparison of the results to one another and also predicting the outcome whether the person is healthy or Parkinson disease effected from the given data. The results of the selected algorithms namely Naïve Bayes, Random Forest, Neural Networks, Decision Trees, AdaBoost, SVM,KNN were compared and tabulated. According to the outputs derived with the help of python, implementing Scikit Libraries,in order to calculate the accuracy and find out the performance, a confusion matrix was constructed at first. From that matrix, the true positives, true negatives along with the false positives and false negatives were used to calculate the support, recall, f1-score and precision were calculated by implementing specified modules. Final accuracy was calculated using these parameters.

From the results, the conclusion obtained is as follows: the Random Forest algorithm gives with optimum accuracy of 78.56% which is closely followed by Decision Tree Algorithm with the optimal accuracy of 77.63%. Following the Decision Tree Algorithm is the MLP Classifier with an optimal accuracy of 76.72%, and lastly the Naïve Bayes Algorithm which has the optimal accuracy of 70.82%. Finally, these algorithms can help in classifying whether a person get effected with Parkinson's disease or not. Table 2 displays the accuracies of the various classification algorithms when applied on the Parkinson dataset.

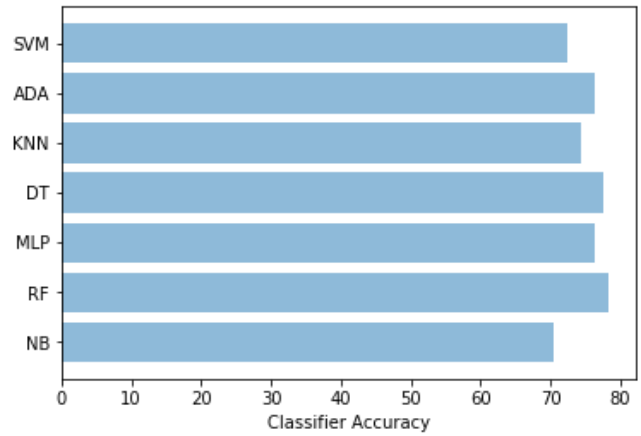

Figure 1. graphically represents the performance of the classification techniques based on their accuracy measures.

CONCLUSION

In this paper, popular Machine Learning classification algorithms were chosen to evaluate their performance in terms of classification performance measures which are accuracy, precision, recall, f1 score and support to classify if the particular person is healthy or Parkinson diseased based on the voice input parameters. It has been noticed in the result analysis that the opted algorithms under the classification technique show some well-to-do accuracy percentages, especially the Random forest and Decision Tree Algorithm. These algorithms can be preferred over the others to classify the dependent variable. The accuracy of the algorithms can be further improved by Feature Selection and Dimensionality Reduction algorithms. It is also possible to automate the process of classifying the dataset by creating a simple interface.

REFERENCES

[1] Indrajit Mandal., N.Sairam. New machine learning Algorithms for prediction of Parkinson's desease. IJSS Volume 45 Issue 3, Mar 2014, 647-666. doi:10.1080/00207721.2012.724114.

[2] George D. Magoulas, AndrianaPrentza. Machine learning in medical applications. ACAI 99,LNAI 2049. 2001, 300-307.

[3] Sarkar C.O, Serbes, G.Gunduz, A.Tunc, H.C Nizam, H.Sarkar, B.E Tutuncu, M.Aydin, T.Isenkul, M.E and Apaydin.H. A Comparative analysis of speech signal processing algorithms for Parkinson disease classification and the use of tunale Q-factor wavelet transform applied soft computing.2018.

[4] R.S. Michalski, J.G.Carbonell, T.M. Mitchell. Machine Learning: An Artificial Intelligence Approach. ISBN 978-3-662-12405-5.

[5] mith, T.F., Waterman, M.S.: Identification of common molecular subsequences. J. Mol. Biol. 147, 195197 (1981). doi:10.1016/0022-2836(81)90087-5.

[6] R.S. Michalski., J.G. Carbonell., T.M. Mitchell. Machine Learning: An Artificial Intelligence Approach.

[7] S. B. Kotsiantis.(2007). Supervised Machine Learning: A Review of Classification Techniques. Informatica 31, 249-268.

[8] Igor Kononenko. Machine Learning for Medical Diagnnosis: History, State of the Art and Perspective.

[9] Vikas B, B.S.Anuhya, K Santosh Bhargav, SipraSarangi, ManaswiniChilla.(2017, June). Application of the Apriori Algorithm for Prediction of Polycystic Ovarian Syndrome (PCOS). 4th International Conference on Information System Design And Intelligent Applications, 2017.

[10] J.Sujatha, Dr. S.P.Rajagopalan, Classification of Parkinson disease with the voice attributes using Inference System,IJPA,Vol.118,2018.

[11] Lee, M., & To, C. (2010). Comparison of Support Vector Machine and Back Propagation Neural Network in Evaluating the Enterprise Financial Distress, International Journal of Artificial Intelligence& Applications, 1(3), 31-43.