ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 14, March 2016

Diagnosis and Classification of Parkinsons Disease Using Data Mining Techniques

P. Durga
Department of Computer Science and Engineering
Kamaraj College of Engineering and Technology
Virudhunagar, India
durupaulraj@gmail.com

V. Sutha Jebakumari
Department of Computer Science and Engineering
Kamaraj College of Engineering and Technology
Virudhunagar, India
suthaphd@gmail.com

D. Shanthi
Department of Computer Science and Engineering
PSNA College of Engineering and Technology
Dindugal, India
dshan71@gmail.com

***Abstract*— Parkinson's disease is a degenerative disorder of central nervous system that mainly affects the motor system due to the loss of dopamine, a chemical that sends the message to the part of the brain for controlling the movements. Early diagnosis of Parkinson's disease is very difficult. The Data mining techniques can be used for the classification of medical data by using WEKA. WEKA has an extensive collection of different machine learning and data mining algorithms. In this paper, various data mining techniques like Naive Bayes, Sequential Minimal Optimization (SMO), J48, Bayesian Network and Multilayer Perceptron are used to develop classifier for diagnosis of Parkinson's disease. The main objective of this paper is to identify the suitable algorithm for classification of normal and abnormal person and to find out the performance measures. The dataset was taken from UCI machine learning repository and it consists of 22 voice features such as jitter, shimmer, NHR and so on.**

Keywords—WEKA, Parkinson's Disease, Dopamine, Data mining techniques, Classification.

I. INTRODUCTION

Parkinson's disease is caused due to the deficiency of dopamine chemicals in the brain that control the movements. When the dopamine level is low, it leads to dysfunction of basal ganglia and other motor deficits. Nowadays PD affects a large part of population in the world. Researchers are trying to learn more about this disease and the methodologies for limiting its symptoms. Mostly the Parkinson disease occurs in the age of 50-60 and it is difficult to identify in the earlier stage. It is primarily characterized by the cardinal motor impairments like rigidity, tremor, and bradykinesia. Tremor is the well-known symptom of PD which results in an automatic shaking movement of the body, arm, lip and hands. The other symptom is rigidity i.e., the stiffness of muscle. Some other symptoms are the problem with walking and slowness of movements. The drug levodopa is used at the early stages of the disease. When the disease worsens the side effects also increases enormously. When the drugs become ineffective surgical treatment is the only alternative [1]. Therefore, the diagnosis of PD at an earlier stage is important for improving the quality of patient's life [2]. For diagnosing the Parkinson Disease, various signals like Electromyography (EMG), speech and gait are used. According to previous research, approximately 70%–90% of patients with PD show some form of vocal impairment [3]. Vocal impairment is found in the majority of the PD patients and it is found to be the major hindrance with this disease [4]. The symptoms of speech dysfunction in PD are weak voice, variable speech rate, and short rushes of speech, imprecise consonants, breathy and harsh voice and monotonous pitch [5]. It is found that PD patients have reduced intensity compared with healthy people and it is found in the earlier stages of PD. The voice signal is the simplest and easiest method for identifying the disease because most of the PD people suffer from Speech disorders. Speech can be recorded and analyzed automatically at much lower cost than clinical assessment. Moreover, assessment based on speech can monitor changes more accurately over time [6].

Paper Overview: The rest of this paper is organized as follows. Section II describes the tool used for the classification and its features. Section III gives the related work done elsewhere on weka and anfis. Section IV discusses the proposed methodology of our work. In section V, the experimental results are analyzed and they are presented.

II. WEKA

WEKA stands for Waikato Environment for Knowledge Analysis developed at the University of Waikato in New Zealand and was implemented in 1997.The software is open source and it was written in Java language [7]. It is issued under the GNU General Public License. It supports the .csv(Comma Separated Values) and .arff(Attribute-Relation File Format) file format. WEKA has 49 tools for processing, 76 algorithms for classification and regression, and 8

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 14, March 2016

algorithms for clustering and 3 algorithms for finding the association rules. In this paper, the WEKA version 3.7.12 is used for the diagnosis and classification of Parkinson's disease. The dataset are used to classify both normal and abnormal person based on the attributes NHR, HNR, Jitter, Shimmer, RPDE, PPE, APQ, DFA, DDA, RAP and Frequency. WEKA is the powerful Data mining tool for performing classification.

III. Related work

Tobias Bocklet et al have employed acoustic features, prosodic features and features derived from a two-mass model of the vocal folds on different kinds of speech tests like sustained phonation's, syllable repetitions, read texts and monologues. Support Vector machine is used for classification [8].

The parallel neural network method is used for the prediction of Parkinson disease. The drawback is the imbalanced datasets and the prediction is not improved after a certain number of parallel networks.

Tele-monitoring of the disease using voice measurement [10] plays a vital role in early diagnosis of Parkinson's disease. The support vector machine is used in this paper for classification.

The Classification algorithm like Random Forest and Naive Bayes are used to identify the PD [11] using the voice datasets. The Random Forest shows the best accuracy 90.26% comparing to the Naive Bayes.

Mohammad Shahbakhi has used the Genetic algorithm for feature selection and used support vector machine for classification. An accuracy of 93.66 per 7 optimized features and the accuracy of 94.22 percent per 9 optimized features are obtained [12].

According to previous research [13], the neurological disorder is identified using the Mel-Frequency Cepstral Coefficient and Support Vector Machine. The MFCC is used for the feature extraction and the SVM is used for the classification of healthy and abnormal one. The SVM shows the 93% accuracy.

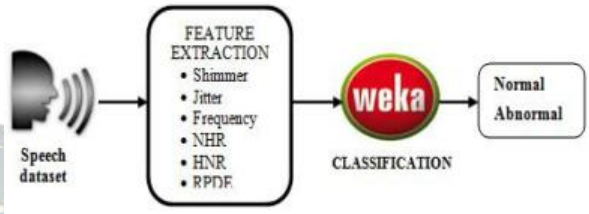
The various data mining techniques have been used for the prediction of hypothyroid disease [14]. The CART and c4.5 algorithm show the best accuracy 99.54%. WEKA is used for classification.

The Neuro-fuzzy inference system is used for the investigation of fluctuations in the Parkinson's disease [15]. The rules are generated from the statistical model and the parameters of membership functions of fuzzy values were changed.

Hybrid learning for Adaptive neuro-fuzzy inference system is used for the system identification based on the data. The training algorithms like back propagation, descent learning algorithm and Runge-Kutta algorithm are used. The performance of the training algorithm is compared as the major measures. Among them, the Runge-Kutta shows the best estimation performance.

IV. Methodology

The block diagram of the proposed work for Parkinson's disease detection is shown in Fig. 1. It consists of three steps namely Dataset collection, Feature Extraction and Classification using WEKA tool. The various voice features like jitter, shimmer, HNR, NHR, frequency etc are extracted



from the given voice for both normal and abnormal person. These features are given as input to WEKA and classified using various data mining algorithms such as J48, Multilayer Perceptron, SMO, Naive Bayes and Bayesian Network.

Fig. 1. Block Diagram of our Analysis

A. *Dataset*

The voice dataset is collected from the UCI Machine learning repository from the centre of machine learning and intelligent systems which were created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado. This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals [17]. The last column represents the status of the people i.e., "0" for normal and "1" for PD.

Weka accepts the data in ARFF format and CSV format. . For example

```
@RELATION sampleee

@ATTRIBUTE FO NUMERIC
@ATTRIBUTE Jitter NUMERIC
….
@ATTRIBUTE class {0, 1}
@DATA
116.848, 0.00531…0
198.383, 0.00212….0
…

152.845, 0.00294…1
156.405, 0.00768…1
```

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 14, March 2016

B. Feature Extraction

From the dataset, both linear and non-linear features are extracted. Table I contains all the features and its description. The features are based on frequency, shimmer, jitter, noise to harmonic ratio and harmonic to noise ratio are used for analysis. The changes in the factors are noticeable in people with Parkinson's disease and healthy.

The MDVP stands for the Multidimensional Voice Program Analysis and it is the useful tool for diagnosing the vocal cord dysfunction.

TABLE I. FEATURES AND DESCRIPTION

FEATURES	DESCRIPTION
MDVP Fo(Hz)	Average vocal fundamental Frequency
Shimmer, Jitter	Variations in amplitude
NHR, HNR	Two measures of ratio of noise to tonal components in the voice
RPDE	Non linear dynamic complexity measure
DFA	Signal fractal scaling exponent
PPE	Non linear measure of fundamental frequency

C. Classification

The goal of the classification is to accurately predict the target class from the data. Christo Ananth et al. [9] discussed about an eye blinking sensor. Nowadays heart attack patients are increasing day by day."Though it is tough to save the heart attack patients, we can increase the statistics of saving the life of patients & the life of others whom they are responsible for. The main design of this project is to track the heart attack of patients who are suffering from any attacks during driving and send them a medical need & thereby to stop the vehicle to ensure that the persons along them are safe from accident. Here, an eye blinking sensor is used to sense the blinking of the eye. spO2 sensor checks the pulse rate of the patient. Both are connected to micro controller.If eye blinking gets stopped then the signal is sent to the controller to make an alarm through the buffer. If spO2 sensor senses a variation in pulse or low oxygen content in blood, it may results in heart failure and therefore the controller stops the motor of the vehicle. Then Tarang F4 transmitter is used to send the vehicle number & the mobile number of the patient to a nearest medical station within 25 km for medical aid. The pulse rate monitored via LCD .The Tarang F4 receiver receives the signal and passes through controller and the number gets displayed in the LCD screen and an alarm is produced through a buzzer as soon the signal is received.

C.1. Explorer Interface

The weka knowledge explorer is an easy graphical user interface which is used for preprocessing the data. Users can load the data file in .arff (Attribute-Relation File Format) file format. Analyze the classification accuracy by using J48, Multilayer Perceptron, SMO, Naive Bayes and Bayesian Network classification algorithms using 10 cross validation.

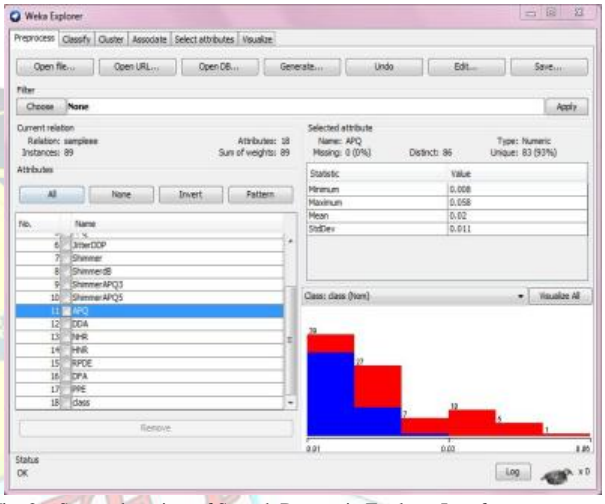


Fig. 2. Screenshot view of Speech Dataset in Explorer Interface

V. EXPERIMENTAL RESULTS

Table II shows the performance measures of various Classification algorithms. It can be seen that Naive Bayes has the least accuracy whereas the Multilayer Perceptron has the highest accuracy. Some of the performance measures [18] used for predicting the classifying accuracy is

Figure 3 shows the plot of performance measures of

Classifier Type	Performance Measures		
	Sensitivity	*Specificity*	*Accuracy*
Bayes Net	85.41%	90.47%	87.78%
Naive Bayes	80%	87.5%	83.33%
Multilayer Perceptron	95.745%	100%	97.78%
SMO	80.76%	92.10%	85.55%
J48	95.45%	93.48%	94.44%

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 14, March 2016

various Classification algorithms. It can be seen that the MLP shows the best accuracy followed by J48 and Bayes Network.

A. Sensitivity

Sensitivity is also known as True positive rate. It is used for measuring the percentage of sick people from the dataset.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (1)$$

B. Specificity

It is also known as True Negative rate. It is used for measuring the percentage of healthy people who are correctly identified from the dataset.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (2)$$

C. Accuracy

Accuracy is the measure of a predictive model that reflects the proportionate number of times that the model is correct when it is applied to data.

Accuracy = Number of samples correctly classified / Total number of samples (3)

Another way to examine the performance of the classifier is to use a ROC graph.

D. Confusion Matrix

It contains the information about actual and predicted classification done by the system. The classification is evaluated using the data in the matrix. The standard terms in the confusion matrix are True positive, True Negative, False Positive False Negative, Precision, Recall and Accuracy.

The MLP classifier uses the back propagation to classify the instances. It consists of nonlinear activation nodes and it classifies the data more accurately i.e., not linearly separable or separable by the hyper plane.

Based on the results, MLP shows the highest accuracy 97.78% with 100% Specificity and 95.475% Sensitivity. Next to MLP the J48 algorithm shows the highest accuracy 94.44% with 93.48% Specificity and 95.45% Sensitivity. The Naive Bayes algorithm shows the least accuracy 83.33% with 87.5% Specificity and 80% Sensitivity.

VI. CONCLUSION

The work has been done for the classification of speech data to identify the normal people from the Parkinson's patients. The experimental result shows that the Multilayer Perceptron (MLP) gives the highest accuracy, sensitivity and specificity compared to the other classifiers. The designed model for diagnosis of Parkinson's disease can be improved by considering the other classification algorithm and other voice features. In future, the Classification can be done using neural network classifiers, TANAGRA.

Acknowledgment

My special thanks to UCI Machine learning repository from the centre of machine learning and intelligent system for providing the dataset for my testing. I thank my Guide and my Family and my Friends for encouraging and getting me to succeed in my work.

References

[1] Defeng Wu, Kevin Warwick, Zi Ma, Jonathan G. Burgess, Song Pan, Tipu Z. Aziz " Prediction of Parkinson's disease tremor onset using radial basis function neural networks " Expert Systems with Applications 37 (2010) 2923–2928.

[2] Maria C Rodriguez-Oroz, Marjan Jahanshahi, Paul Krack, Irene Litvan, Raúl Macias, Erwan Bezard, José A Obeso "Initial clinical manifestations of Parkinson's disease: features and pathophysiological mechanisms" Lancet Neurol 2009; 8: 1128–39

[3] A. K. Ho, R. Iansek, C. Marigliani, J. Bradshaw, and S. Gates, "Speech impairment in large sample of patients with Parkinson's disease," Behav. Neurol. 11, 131–137 (1998).

[4] Hartelius L. · Svensson P. "Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: a survey" FOLIA PHONIATR LOGOP. 1994;46(1):9-17.

[5] Sofie Lundgren, Thomas Saeys, Fredrik Karlsson, Katarina Olofsson, Patric Blomstedt, Jan Linder, Erik Nordh, Hamayun Zafar, and Jan van Doorn "Deep Brain Stimulation of Caudal Zona Incerta and Subthalamic Nucleus in Patients with Parkinson's Disease: Effects on Voice Intensity " SAGE-Hindawi Access to Research Parkinson's Disease Volume 2011.

[6] Alireza Bayestehtashka, Meysam Asgaria, Izhak Shafrana,□, James McNames "Fully automated assessment of the severity of Parkinson's diseasefrom speech" Computer Speech and Language (2014)

[7] Tobias Bocklet et al "Detection of Persons with Parkinson's Disease by Acoustic, Vocal, and Prosodic Analysis" IEEE Workshop on Automatic Speech Recognition and Understanding, 2011.

[8] F. Astrom and R. Koker ," A parallel neural network approach to prediction of Parkinson's disease." Expert systems with Application, vol.38,no.10,pp.12470-12474,2011

[9] Christo Ananth, S.Shafiqa Shalaysha, M.Vaishnavi, J.Sasi Rabiyathul Sabena, A.P.L.Sangeetha, M.Santhi, "Realtime Monitoring Of Cardiac Patients At Distance Using Tarang Communication", International Journal of Innovative Research in Engineering & Science (IJIRES), Volume 9, Issue 3,September 2014,pp-15-20

[10] Tarigoppula V.S Sriram, M. Venkateshwara Rao, G V Satya Narayana and DSVGK Kaladhar,"Intelligent Parkinson disease prediction using machine learning algorithms," International Journal of Engineering and Innovative Technology, vol.3, Issue 3, pp.212-215, September 2013.

[11] Shahbakhi "Speech Analysis for Diagnosis of Parkinson's Disease Using Genetic Algorithm and Support Vector Machine" Journal of Biomedical Science and Engineering, 2014.

[12] Vikram. C.M and K.Umarani, "Pathological voice analysis to detect neurological disorders using MFCC and SVM," International Journal of Advanced Electrical and Electronics Engineering, vol.2, Issue-4, pp. 87-91, 2013

[13] Shivanee Pandey, Rohit Miri and S.R. Tandan, "Diagnosis and classification of hypothyroid disease using data mining." International Journal

of Engineering Research and Technology(IJERT), ISSN: 2278-0181, Vol.2 Issue 6, June 2013
[14] C.Loganathan and K.V. Girija, "Hybrid learning for adaptive neuro fuzzy inference system." International Journal of Engineering and Science, Vol.2, Issue 11(April 2013), Pp 06-13.
[15] Yahia Alemami and Laiali Almazaydeh,"Detection of parkinson disease through voice signal features." Journal of American Science, 2014.
[16] D.P. Gaiwad and Ravindra R Thool," Intrusion detection system using cascade forward neural network with genetic algorithm based feature selection," International Journal of Advanced Computer Engineering and Communication Technology (IJACECT), vol.,3,issue-3, 2014.
[17] P.D. Wasserman,"Advanced methods in neural computing," Van Nostrand Reinhold, Newyork, 255. 1993.
[18]] D.P. Gaiwad and Ravindra R Thool," Intrusion detection system using cascade forward neural network with genetic algorithm based feature selection," International Journal of Advanced Computer Engineering and Communication Technology (IJACECT), vol.,3,issue-3, 2014.
[19] A. Araque, E.D. Martin, g. Perea, J.I.Arellano and W. Buni," Synaptically released acetylcholine evokes Ca2+ elevations in astrocytes in hippocampal slices," Journal of Neuroscience, vol.22, pp.2443-2450.
[20] P.D. Wasserman,"Advanced methods in neural computing," Van Nostrand Reinhold, Newyork, 255. 1993.
[21] LucasBernardo Nigrini, "Developing a neural network model to predict the electrical load demand in the mangaung municipal area," August 2012.
[24]http://in.mathworks.com/help/nnet/ref/trainlm.html?searchHighlight=trainlm
[22] Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2008), "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease", IEEE Transactions on Biomedical Engineering 2009 .