IJCSI International Journal of Computer Science Issues, Volume 14, Issue 5, September 2017
ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org https://doi.org/10.20943/01201705.3136

Parkinson's Disease Clustering and Classification with Multi Layer Perceptron using Best Ranked Voice Feature and Associative Rule Generation

Arup Kumar Bhattacharjee[1] and Soumen Mukherjee[2]

[1] Dept. of Computer Application
RCC Institute of Information Technology, Kolkata, India

[2] Dept. of Computer Application
RCC Institute of Information Technology, Kolkata, India

Abstract

In the present paper several clustering and classification work is done with Parkinson's patient and normal human voice sample available in UCI data repository. An accuracy of 89.66% is achieved using MLP classifier with all 22 raw features. An accuracy of 85% is achieved with only 1 PCA component. ReliefF algorithm is used to find the best ranked features. With 3 best ranked raw features the present work achieves 85% accuracy. Association rule are also shown using Apriori algorithm with 100% support and confidence. kMeans and Filtered Cluster are applied to the dataset with 2 cluster centers.

Keywords: *Neuro disorder, Dopamine, Vocal frequency, SVM, ROC, Apriori algorithm, PCA.*

1. Introduction

Parkinson's disease is a situation possible to happen in case of loss of nerve cells in brain [1-2]. It is the second most common neuro disorder after Alzheimer's disease. Some of the motor symptoms of Parkinson's diseases are slowness, postural instability, muscle stiffness and tremor and non-motor symptoms are cognition and psychiatric [1-2].

Soft computing is that domain of computer science which is used to compute solutions that tolerate uncertainty and works on partial truth and approximation. Soft computing is used for problems with no exact solutions available in some fixed time. Thus soft computing comes with various approaches to deal with real life scenario. Two fundamental soft computing approaches to characterized objects by one or more features are –

(a) Classification – A supervised learning approach to assign objects to groups.

(b) Clustering - An unsupervised learning approach to group objects that are similar based on some predefines rule.

These approaches can be used to extract information from a database and process it for further use.

With the technological and industrial advancements, human society is exposed to various diseases. One such disease is Parkinson's. Parkinson's is typically develops after the ages of 55 years. Around 0.3% of population is affected by this disease and a chance of men affected by this disease is more than women [3].

In this paper various clustering and classification approach are used to extract information from the Parkinson's Data set [4] available in the url archive.ics.uci.edu/ml/[5].This data contains 195 instances of data containing 23 features. Among these features, one feature is status which is set to 0 for healthy and 1 for person with Parkinson's disease.

2. Related Works

Parkinson's disease is progressive disorder that causes balance problem and the severity and pattern of the disease is highly influenced by age [6] and it is found that progress of the disease increases with the increase in age. Various studies being made with various groups of people in different age group as well as persons with and without motor disability. Many works are also done to predict the future cases of Parkinson's disease where a range of biomedical measurements are taken from healthy and effected people like vocal fundamental frequency, several measures of variation in amplitude and fundamental frequency. Huge amount of clinical features were collected which includes Gait disturbances, Sensory symptoms, Voice, speech and swallowing disorder, sleep disorder[7]. Many researchers have shown that loss of cells containing

neurotransmitter dopamine leads to the Parkinson's disease. Reduction of dopamine level causes movement disorder with a variety of motor and non-motor symptoms [8]. In the recent time many soft computing techniques are applied on the biomedical measurements of Parkinson's disease patients to infer conclusion about patients like their severity, influence of various measures.

3. Parkinson's Dataset

This dataset is downloaded from UCI machine learning repository of Center for Machine Learning and Intelligent Systems. This dataset is a multivariate dataset with 195 samples of biomedical voice measurements of 31 persons with 23 real number attributes, with no missing values. This dataset is prepared by the University of Oxford, in collaboration with National Centre for Voice and Speech, Colorado in the year 2008. The dataset is used in this work for classification of healthy and Parkinson's disease person using the attribute like maximum, average and minimum vocal fundamental frequency, different measurement of variation in fundamental frequency and amplitude, ratio of noise to tonal components in the voice, signal fractal scaling exponent and different nonlinear measures of fundamental frequency variation. For classification in this work the 22 voice sample features or attributes is used as a predictor and the healthy or Parkinson's disease is used as response [4-5].

4. Classification

The Parkinson's dataset is classified with different classifier to identify the health status of a person i.e. is the person has Parkinson's disease or the person is healthy. This classification work is carried out with all 22 available features and 195 samples of data using Linear SVM, Quadratic SVM, Cubic SVM Neural Network and Complex Tree and it is found that Neural Network gives best accuracy of 89.66% with 5 fold cross-validation using and neural network toolbox. The other classifications are done using MATLAB classification learner toolbox. The details of the result with 22 features are shown in the Table 1. When the same classification is done with only one Principal Component feature the Neural Network gives best accuracy of 85% in comparison with the same set of classifier. The details of the result with only 1 PCA component are shown in the Table 2.

Table 1: Accuracy found with Parkinson's dataset with all 22 features and 195 samples

Classifier	*Accuracy (%)*
Linear SVM	87.2
Quadratic SVM	88.7
Cubic SVM	88.7
Neural Network(MLP)	89.66
Complex Tree	84.6

Table 2: Accuracy found with Parkinson's dataset with only 1 PCA Component and 195 samples

Classifier	*Accuracy (%)*
Linear SVM	75.4
Quadratic SVM	71.3
Cubic SVM	68.2
Neural Network(MLP)	85
Complex Tree	75.4

To find the best result with neural network classifier i.e. Multi Layer Perceptron (MLP) is used with one hidden layer with varying number of neuron varying from 1 to 20. The accuracy versus the number of neuron in the single hidden layer graph is shown in Fig.1. It can be found from the graph that the best accuracy of 89.66% is found with 9 neurons in the hidden layer. The Fig. 2, Fig.3 and Fig.4 are the graph of sensitivity, specificity and precision respectively with varying number of neuron in the single hidden layer. It can be found from the graph that a high sensitivity of 94.03%, specificity of 79.17% and a high precision of 92.88% is found in the classification work with 9 neurons in the single hidden layer.

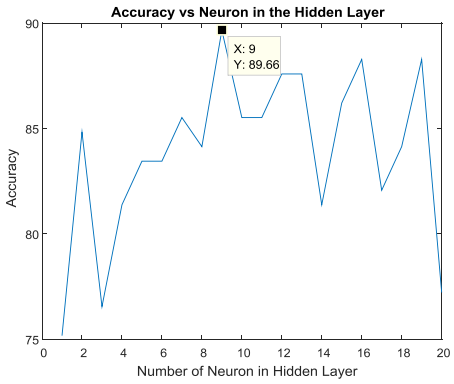

Fig 1: Accuracy versus number of neuron in hidden layer with all 22 features and 195 samples using MLP.

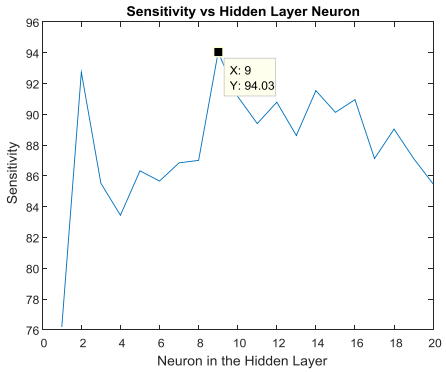

Fig 2: Sensitivity versus number of neuron in hidden layer with all 22 features and 195 samples using MLP.

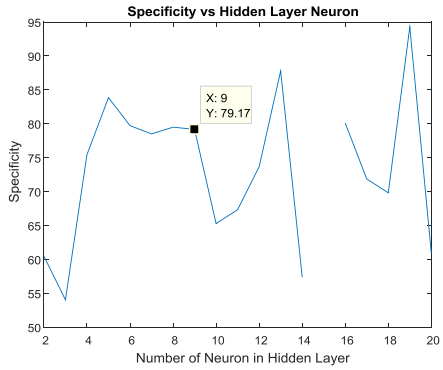

Fig 3: Specificity versus number of neuron in hidden layer with all 22 features and 195 samples using MLP.

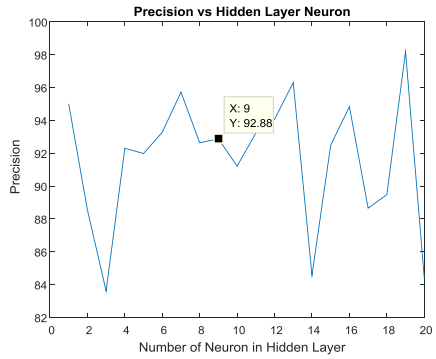

Fig 4: Precision versus number of neuron in hidden layer with all 22 features and 195 samples using MLP.

In this Parkinson's disease classification task 70% (137) samples is used for training the neural network and 15% (29) samples are used separately for both validation and testing. The confusion matrix is shown in Fig. 5.

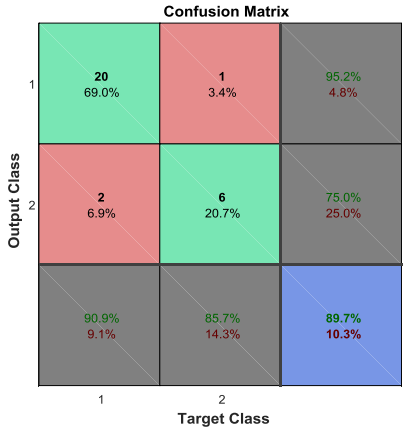

Fig 5: The confusion matrix of the neural network classification task.

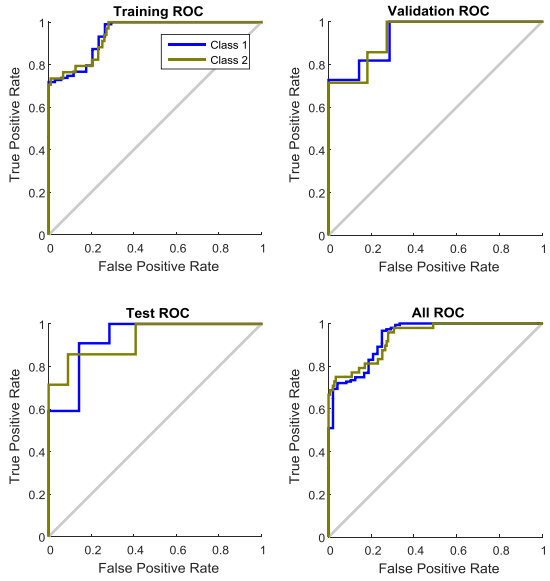

Fig 6: The ROC (Receiver Operating Characteristics) curve of training, validation, testing and all work

Fig. 6 shows the ROC (Receiver Operating Characteristics) curve of training, validation, testing and all work together. It is found that all ROC curve gives AUC (Area Under Curve) nearly to 1.

ReliefF algorithm is used to rank the best features among all the 22 features used in this work. The Fig. 7 shows the variations of weight of each feature with varying sample size. It can be found from the graph the weight stabilizes with around 60 sample size.

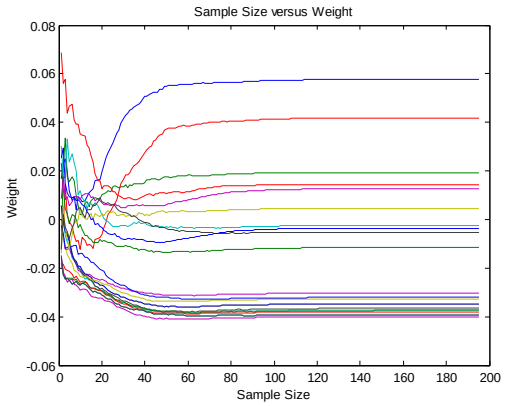

Fig 7: Weight versus sample size in ReliefF feature ranking algorithm

In Fig. 8 the change of accuracy with varying best ranked feature found using ReliefF is shown. Best accuracy result of 89.7% is found with 13 best ranked raw features. With 3 best ranked raw features the present work achieves 85% accuracy

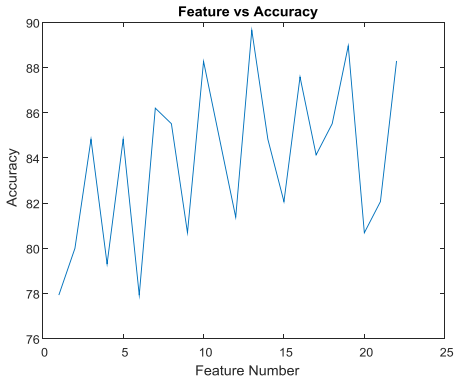

Fig 8: Accuracy versus best ranked feature size found by ReliefF algorithm

The accuracy versus PCA component with all 22 features and 195 samples using MLP with 9 neurons in the hidden layer is shown in the Fig. 9. It is found that the best accuracy of 91% is achieved with only 20 PCA components.

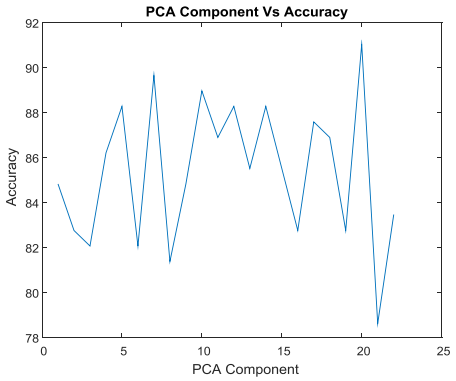

Fig 9: Accuracy versus PCA component with all 22 features and 195 samples using MLP with 9 neurons in the hidden layer.

Table 3: Best accuracies found with MLP classifier with different feature type.

MLP classification with feature type	***Best Accuracy (%)***
MLP with all 22 Feature	89.66
MLP with 1 PCA component	85
MLP with 20 PCA component	91
MLP with 13 best ranked feature found by ReliefF	89.7

There are many numbers of researches going on in the field of classification of Parkinson's disease using soft computing and machine learning. Long et al. [9] worked on automatic classification of early Parkinson's disease with multi-modal MR imaging. In Table 3 all the best accuracies found in present work with MLP classifier with different feature type is given. The best accuracy of 91% is found with 20 PCA components in the present work. This result (91%) is better than the result found by Khan [10] of 90.26% with K-NN classifier.

Another approach of classification on the same Parkinson's data set with 195 instances and each instance having 22 features are applied. In this method Weka software [11] is used for classification and clustering. Different classifier model used in Weka software are M5 Pruned model, ZeroR and Decision Table as shown in Table 4. Here 10 fold cross-validations are used. Various results obtained are –

Table 4: Various classification models and their performance

Classifier Model	*Correlation coefficient*	*Mean absolute error*	*Root mean squared error*	*Relative absolute error*	*Root relative squared error*
M5 pruned model (using smoothed linear models)	0.8943	0.0165	0.0465	23.09%	51.49 %
ZeroR	-0.1965	0.0716	0.0903	100%	100%
Decision Table	0.9475	0.0184	0.0288	25.73%	31.85 %

5. Association Rule Mining and Clustering

Association rules are also created by analyzing data for frequent pattern using the support and confidence to find the relationships between the status and various attributes. Apriori Association rules with 100% support and 100% confidence are applied and it is found that all attributes (antecedent) can determine status (consequence). Some of the highest support and confidence rules are –

(a) MDVPFo(Hz), MDVPFhi(Hz) -> Status

Here MDVPFo measures average vocal fundamental frequency and MDVPFhi measures maximum vocal fundamental frequency, both in Hertz.

(b) MDVPFo(Hz), MDVPFlo(Hz) -> Status

Here MDVPFo measures average vocal fundamental frequency and MDVPFlo measures Minimum vocal fundamental frequency, both in Hertz.

(c) MDVPFo(Hz), MDVPJitter(%) -> Status

Here MDVPFo measures average vocal fundamental frequency (in Hertz) and MDVPJitter(%) gives percentage measures of variation in fundamental frequency.

(d) MDVPFo(Hz), MDVPJitter(Abs) -> Status

Here MDVPFo measures average vocal fundamental frequency (in Hertz) and MDVPJitter(Abs) gives absolute measures of variation in fundamental frequency.

(e) MDVPFo(Hz), MDVPRAP -> Status

Here MDVPFo measures average vocal fundamental frequency and MDVPRAP measures the variation in fundamental frequency.

Table 5: Results of various clustering approach considering 2 clusters are

Clustering Techniques	*No. of instances in Cluster 1*	*No. of instances in Cluster 2*
kMeans	148 (76%)	48(24%)
Filtered Cluster	148 (76%)	48(24%)
Density Based Cluster	101 (52%)	95 (48%)

As shown in the Table 5 both kMeans and Filtered clustering [12] with 2 clusters have same result, number of instances in cluster 1 is 148 and in cluster 2 is 48 whereas in Density Based Clustering technique number of instances in cluster 1 is 101 and in cluster 2 is 95.

6. Conclusions

It is found from the present work that Multi Layer Perceptron works better than other classifier with Parkinson's dataset. It is also found that Apriori Association rules with 100% support and 100% confidence with all feature or attributes can determine dependency of status. Also two cluster results are shown with better performance by kMeans and filtered clusters.

References

[1] Thushara Perera, Wesley Thevathasan (2014), "An Introduction to Parkinson's Disease".

[2] Aarsland, D., K. Bronnick, et al. (2007). "Neuropsychiatric symptoms in patients with Parkinson's disease and dementia: frequency, profile and associated care giver stress." J Neurol Neurosurg Psychiatry 78(1): 36-42. [2]

[3] Michael S. Okun, " Deep Brain Stimulation for Parkinson's Disease", The NEW ENGLAND JOURNAL of MEDICINE 2012, pp. 1529-38.

[4] '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)

[5] archive.ics.uci.edu/ml/

[6] Blin, J., Dubois, B., Bonet, A.M. et al., "Does ageing aggravate parkinsonian disability?", J. Neurol. Neurosurg. Psychiatry, 54: 780–782, 1991.

[7] Factor S. A. , Weiner W. J., "Parkinson's Disease Diagnosis and clinical management", Second Edition, Demos Medical Publishing 2008.

[8] Perera T., Thevathasan W., "An introduction to Parkinson's Disease", http://brainfoundation.org.au/wp-content/uploads/2015/05/Perera-Thushara-Parkinsons-Disease.pdf

[9] Dan Long, Jinwei Wang, Min Xuan, Quanquan Gu, Xiaojun Xu, Dexing Kong, Minming Zhang, "Automatic Classification of Early Parkinson's Disease with Multi-Modal MR Imaging", 2012, https://doi.org/10.1371/journal.pone.0047714.

[10] Sajid Ullah Khan, "Classification of Parkinson's Disease Using Data Mining Techniques", Avens Publishing Group, Journal of Parkinson's disease & Alzheimer's disease, July 2015 Vol.:2, Issue:1.
[11] www.cs.waikato.ac.nz/ml/weka/

[12] Sree Ram Nimmagadda, Phaneendra Kanakamedala, Vijay Bashkarreddy Yaramala, "Implementation of Clustering Through Machine Learning Tool", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 1, January 2011, ISSN (Online): 1694-0814.

Arup Kumar Bhattacharjee received his MCA Degree from University of Kalyani and his M.Tech from West Bengal University of Technology. He is an Assistant Professor of Computer Application at RCC Institute of Information Technology, Kolkata, India. He has more than 14 years teaching experience in the field of Computer Science and Application. His research interests include Soft Computing and Object Technology. He has contributed in over 20 internationally acclaimed books in the field of Computer Science and Engineering. He has also edited 2 books.

Soumen Mukherjee did his B.Sc (Physics Honours) from Calcutta University, M.C.A. from Kalyani University and ME in Information Technology from West Bengal University of Technology. He is the silver medalist for ME examination in the university. He has done his Post-Graduate Diploma in Business Management from Institute of Management Technology, Center of Distance Learning, Ghaziabad. He is now working as an Assistant Professor in RCC Institute of Information Technology, Kolkata. He has 14 years teaching experience in the field of Computer Science and Application. He has over 30 research paper published in different National and International Journal and Conferences. He has contributed in over 20 internationally acclaimed books in the field of Computer Science and Engineering. He has edited 2 books. His research fields are Image Processing and Machine Learning. He is a life member of several institutions like IETE, CSI, ISTE, FOSET etc.