Diagnosis of Parkinson Disease Using Machine Learning and Data Mining Systems from Voice Dataset

Tarigoppula V.S. Sriram[1], M. Venkateswara Rao[2], G.V. Satya Narayana[3], and D.S.V.G.K. Kaladhar[4]

[1] MCA, Raghu Engineering College, Visakhapatnam, India
[2] IT, GITAM University, Rushikonda, Visakhapatnam, India
[3] IT, Raghu Institute of Technology, Visakhapatnam, India
[4] Dept. of Bioinformatics, GITAM University, Visakhapatnam, India
{ramjeesis,dkaladhar}@gmail.com
mandapati_venkat@yahoo.co.in,
gv_satyanarayana@yahoo.com,

Abstract. Parkinson's disease is one of the most painful, dangerous and non curable diseases which occurs at older ages (mostly above 50 years) in humans. The data-set for the disease is retrieved from UCI repository. A relative study on feature relevance analysis and the accuracy using different classification methods was carried out on Parkinson data-set. Sieve multigram data and Survey graph provide the statistical analysis on the voice data so that the healthy and Parkinson patients would be correctly classified. KStar and NNge present good accuracy based classification methods. Sieve multigram shows the edges between the nodes such as Fhi, Flo, Jitter, JitterAb, RAP and PPQ. KStar and NNge have connections with Shimmer and ShimmerDB . ADTree shows 21 leaves with 31 leaves and SimpleCART shows 13 leaves and 7 leaves. Most of the clusters vary with DBScan and SimpleKMeans with 25% and 38% towards Parkinson disease.

1 Introduction

In a data warehouse, the user usually has a view of multiple data-sets collected from different data sources and understanding their relationships is an extremely important part of the KDD process. Data mining is a rapidly evolving advanced technology that used in bio-medical sciences and research in order to predict and analyze large volumes of medical data such as Parkinson's disease [4]. Data mining in neuro-degenerative diseases like Parkinson disease is an emerging field with enormous importance for deeper understanding of mechanisms relevant to disease, providing prognosis and complete treatment [1].

A relatively high prediction performance has been achieved with classification accuracy. Classification algorithms predict accuracy for discrete variables that are relevant on the other attributes present in the data-set. The continual increase of the number of features can significantly contribute with clustering-based feature

S.C. Satapathy et al. (eds.), *Proc. of the 3rd Int. Conf. on Front. of Intell. Comput. (FICTA) 2014 – Vol. 1*, Advances in Intelligent Systems and Computing 327, DOI: 10.1007/978-3-319-11933-5_17

analysis [2]. Clustering is a datamining technique that is mostly studied for predicting unstructured data to informal data. Clustering algorithms divide data into segments with clusters predicting similar branches and properties. Association algorithms find the correlations between different attributes in a data-set to conduct polynomial associations for better hypothesis.

The data mining intelligence systems can analyze voice fluctuations in patients suffering from Parkinson's disease (PD) by using data obtained as signals from bio-medical equipment like surface electromyographic (EMG) and accelerometer (ACC) [3]. Specific clinical applications requiring advanced analysis techniques, such as data mining and other techniques can be obtained from sensors to analyze large data sets. Medical data mining is a new emerging field that has great influence for exploring patterns from respective medical data sets in clinical research [5].

PD is the 2nd most common type of Neurodegenerative disorder that causes disturbances in speech and accurate voice [6]. No genetic component is evident that a disease begins after age 50 years [7]. Research and investigations on efficacy of intensive voice therapy are underway to improve the functional communication in the patients with idiopathic PD [8]. Parkinson's disease occurs between the ages of 50 and 95 years [9].

2 Methodology

The data-set for the PD was obtained from an online data repository UCI. Comparative classification studies on different data-sets in an entity have been applied for accuracy analysis and the time taken to execute the data-set in order to find the best classification rule. The data of healthy people and those with Parkinson can be correctly classified by using machine learning and data mining systems.

Bayesian theory is a mathematical model in calculus of degrees that predicts proposition of interest. BayesNet and NaiveBayes are the most practical learning methods that have a random sequence model within each class. System classification summarizes a sequence of classification methods using algorithms. Logistic function can be derived from simple classification problems, measuring from minus infinity to plus infinity. Simple logistic regression is used to explore associations between dichotomous outcome with continuous, ordinal, or categorical exposure variable.

Lazy classification scheme uses Hierarchical SVMs to select a subset of candidate classes for each test instance, in order to determine the overall best performer. K-Star is an Instance-Based learner. The class of a test instance is based upon the class of those that is determined by similarity function with similar training instances. It is different from the other instance-based machine learners using entropy-based distance function.

Meta-classification makes its binary decision by classifying synthesized feature vectors and one meta-classifier for each class is built for each. Bagging (or Bootstrap aggregating) is an algorithm to improve machine learning of statistical classification

and regression models in terms of stability and classification accuracy. A statistical method of distance based classification with that of best matching rule can be explained by NNGE (Non-Nested Generalized Exemplars).

The classification is used to follow the path dictated by the successive test placed along the tree until it finds a leaf containing the class to assign a new attribute. ADTree (An alternating decision tree), J48, Random Forest and Simple CART (A simple Classification And Regression Tree) are some of the trees being used in classification of Parkinson's disease.

A survey graph and sieve multigram is used to construct the statistical analysis on the multiple data-sets of the voice data-set from UCI.

3 Results and Discussions

The present experimentation is conducted to predict attributes that are associated and classified from voice data. Table 1 shows the classification of the voice data-sets, accuracy and time taken to execute the classified data-sets. KStar and NNge have presented good accuracy based analysis on the classification. The execution time taken is also less in providing the output for the submitted data-sets (<2 seconds). The classification methods i.e. Support Vector Machine method (SVM), can be used to distinguish people with Parkinson's disease from the healthy people [10]. UCI data set on PD was composed of a voice measurement from bio-medical instrumentation with 195 samples with 16 attributes. Two training algorithms like Scaled Conjugate Gradient (SCG) and Levenberg-Marquardt (LM) using Multilayer Perceptrons (MLPs) and Neural Network had performed with high accuracy. LM performed with an accuracy rate of 92.95% while SCG obtained 78.21% accuracy [11].

Table 1. Classification of voice data-sets

Classification type	Classification model	Correctly classified	Incorrectly classified	Time taken (in seconds)
Bayes	BayesNet	84.6	15.4	0.11
	NaiveBayes	70.26	29.74	0.02
Functions	Simple logistic	85.13	14.87	0.98
Lazy	KStar	100	0	0
Meta	Bagging	92.31	7.69	0.27
Rules	NNge	100	0	0.1
Trees	ADTree	2.05	97.95	0.22
	J48	98.97	1.03	0.47
	RandomForest	99.49	0.51	0.17
	SimpleCART	96.41	3.59	0.42

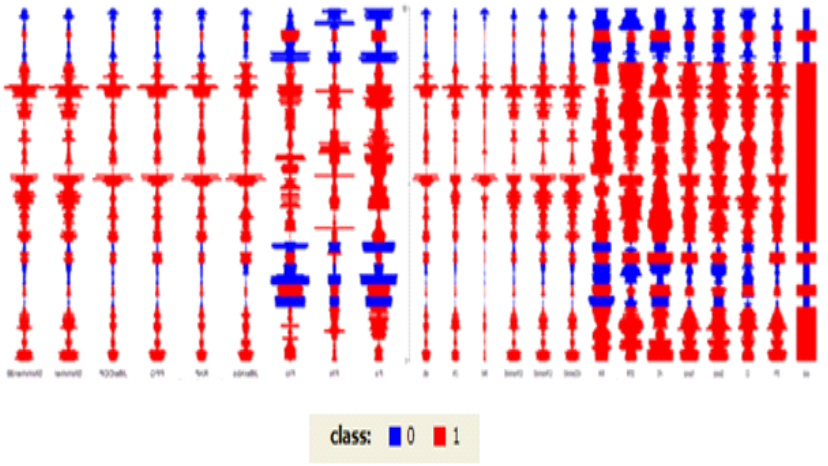

Fig. 1. Survey graph

Figure 1 and 2 presents the survey path and sieve multigram for the complete dataset. The survey graph shows high number of variances with healthy and diseased voice data-sets. Most of the healthy data-sets showed constant levels. Sieve multigram showed the edges between the nodes such as Fhi, Flo, Jitter, JitterAb, RAP and PPQ. Connections with edges are also present between Shimmer and ShimmerDB. The important observation by Ramani and Sivagami, 2011, shows the feature relevance analysis for better classification purpose and showed three important features like spread1, PPE and spread2 based on PD dataset [5].

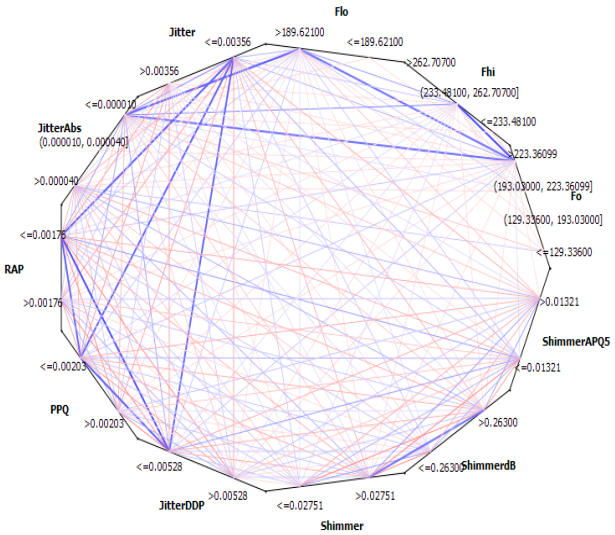

Fig. 2. Sieve multigram

ADTree and SimpleCART presented the relationships of nodes for the Parkinson data-sets. ADTree showed 21 leaves with 31 leaves and SimpleCART showed 13 leaves and 7 leaves. Fo data-sets can provide the focus in analyzing the frequencies for recognition of healthy and diseased individuals in Parkinson data-set (Figure 3).

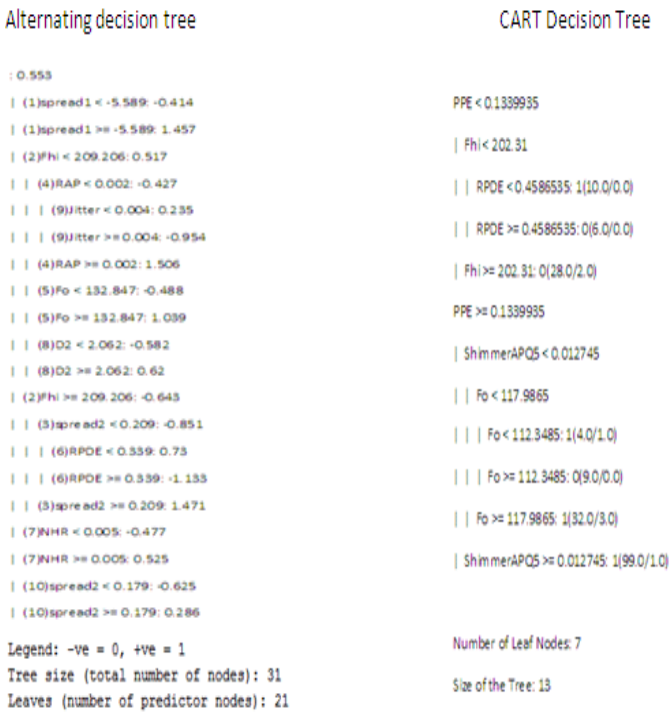

Fig. 3. Decision Tree using ADTree and CART

Table 2 provided the clustering results for the complete data-set. All the algorithms that were analyzed have shown 2 clusters. Most of the clusters vary with DBScan and SimpleKMeans with 25% and 38% towards Parkinson disease. Clustered data by SimpleKMeans has been presented in Figure 4.

Table 2. Clustering

ClusterType	Number of clusters	Cluster report
DBScan	2	0 144(75%) 1 48(25%)
Hierarchial clustering	2	0 193(99%) 1 2(1%)
SimpleKMeans	2	0 120(62%) 1 75(38%)

kMeans
======
Cluster centroids:

Attribute	Full Data (195)	Cluster# 0 (120)	1 (75)
Fo	154.2286	165.6164	136.0083
Fhi	197.1049	211.9424	173.3649
Flo	116.3246	125.6427	101.4157
Jitter	0.0062	0.0042	0.0095
JitterAbs	0	0	0.0001
RAP	0.0033	0.0021	0.0052
PPQ	0.0034	0.0022	0.0054
JitterDDP	0.0099	0.0064	0.0156
Shimmer	0.0297	0.0188	0.0472
ShimmerdB	0.2823	0.1731	0.4569
ShimmerAPQ3	0.0157	0.0099	0.0249
ShimmerAPQ5	0.0179	0.0111	0.0287
APQ	0.0241	0.0149	0.0388
ShimmerDDA	0.047	0.0297	0.0746
NHR	0.0248	0.0113	0.0466
HNR	21.886	24.2069	18.1725
RPDE	0.4985	0.4524	0.5724
DFA	0.7181	0.7068	0.7361
spread1	-5.6844	-6.2804	-4.7308
spread2	0.2265	0.1934	0.2795
D2	2.3818	2.2644	2.5696
PPE	0.2066	0.1568	0.2862
class	1	1	1

Fig. 4. Cluster centroids based on KMeans

Flo, Spread1 and APQ has shown best Attribute ranking based on RankSearch Method. Fhi, Flo, JitterDDP, APQ, NHR, spread1, spread2, D2 and PPE are the 9 locally selected attributes predicted based on CFS Subset evaluator.

4 Conclusions

Most of the work has been taken from the voice data-set. Diagnosing the voice data is useful in treating the disease by various voice exercises and also administering medicines at an early stage. Further research has to be done in these areas.

Acknowledgement. The author thanks the management and the staff of GITAM University and also RAGHU Institute of Technology, Visakhapatnam, India for the support extended in bringing out the above literature and experiment.

References

1. Joshi, S., Shenoy, D., Vibhudendra Simha, G.G., Rrashmi, P.L., Venugopal, K.R., Patnaik, L.M.: Classification of Alzheimer's Disease and Parkinson's Disease by Using Machine Learning and Neural Network Methods. In: Machine Learning and Computing (ICMLC), pp. 218–222 (2010)

2. Yang, M., Zheng, H., Wang, H., McClean, S.: Feature selection and construction for the discrimination of neurodegenerative diseases based on gait analysis. Pervasive Computing Technologies for Healthcare, 1–7 (2009)
3. Bonato, P., Sherrill, D.M., Standaert, D.G., Salles, S.S., Akay, M.: Data mining techniques to detect motor fluctuations in Parkinson's disease. In: Conf. Proc. IEEE Eng. Med. Biol. Soc., vol. 7, pp. 4766–4769 (2004)
4. Shianghau, W., Jiannjong, G.: A Data Mining Analysis of the Parkinson's Disease. IB 3(1), 71–75 (2011)
5. Geetha, R.R., Sivagami, G.: Parkinson Disease Classification using Data Mining Algorithms. International Journal of Computer Applications 32(9), 17–22 (2011)
6. Kaladhar, D.S.V.G.K., Nageswara, R.P.V., Ramesh, N.R.B.L.V.: Confusion matrix analysis for evaluation of speech on Parkinson disease using weka and matlab. International Journal of Engineering Science and Technology 2(7), 2734–2737 (2010)
7. Tanner, C.M., Ottman, R., Goldman, S.M., Ellenberg, J., Chan, P., Mayeux, R., William, L.J.: Parkinson Disease in Twins: An Etiologic Study. JAMA 281(4), 341–346 (1999)
8. Smith, M.E., Ramig, L.O., Dromey, C., Perez, K.S., Samandari, R.: Intensive voice treatment in parkinson disease: Laryngostroboscopic findings. Journal of Voice 9(4), 453–459 (1995)
9. Logemann, J. A., Gensler, G., Robbins, J., Lindblad, A. S., Brandt, D., Hind, J. A., Kosek, S., Dikeman, K., Kazandjian, M., Gramigna, G. D., Lundy, D., McGarvey-Toler, S., Miller, G. P. J.: A randomized study of three interventions for aspiration of thin liquids in patients with dementia or Parkinson's disease. J Speech Lang Hear Res. 51(1), 173–83 (2008).
10. Bhattacharya, I., Bhatia, M.P.S.: SVM classification to distinguish Parkinson disease patients. In: A2CWiC 2010, ACM, New York (2010)
11. Bakar, Z.A., Tahir, N.M., Yassin, I.M.: Classification of Parkinson's disease based on Multilayer Perceptrons Neural Network. Signal Processing and Its Applications (CSPA), 1–4 (2010)