IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

A Comparative Study of Data Mining Tools on Parkinson's disease

Mr. Santosh Chauhan [1], Dr. Maitreyee Dutta [2], Dr. Arvind Tiwari [3]

Asst. Professor, Kashi Institute of Technology Varanasi, India[1]

Prof., NITTTR, Chandigarh, India[2]

Director, GGS IT, Chandigarh, India[3]

Abstract: The data mining techniques are a more popular in many field of medical, business, railway, science...etc; these are most commonly used for medical diagnosis and disease prediction. In this paper I am modifying the existing method of mining from the large dataset. By this method we will retrieve the similar objects each of which includes an image sequences having the similar properties or behaviours. The data mining is used for retrieving the relevant information in medical and health areas of the most important factors in medical societies. The current paper is to provide an analysis of data mining techniques to be used in Parkinson's disease.

Keywords: Parkinson disease, data mining, MCC, ROC.

I. INTRODUCTION

1.0 Background

In the current scenario , there are lots of neuro-degenerative[1] diseases that have been recognized such as Alzheimer's disease, Parkinson disease, Arthritic disease, Dementia with Lewy bodies, Corticobasal degeneration, Progressive supranuclear palsy, Prion disorders, and so on. Among all of these neurodegenerative and coordinating the body movement's diseases, Parkinson's disease is second most common disease after Alzheimer's. James Parkinson described the core clinical feature of the Parkinson's disease. James Parkinson's [2], described the Parkinson's disease as paralysis agitans. But later on surname of James Parkinson i.e., Parkinson's was adopted as disease name. Parkinson's disease affects the neuron cell in the brain of live organism.

In clinical research [3], medical information is essential for diagnosis and patient care. For clinical research, it also provides useful information to facilitate therapeutic improvement and conduct medical researchers. The medical knowledge management in the realm of medical information can be shown as the cycle among the clinical re-search, guidelines, quality indicators, performance measures, outcomes and the concept. In order to integrate clinical information management, medical data analysis, and application development, clinical decision intelligence (CDI) is emerged in the new area to streamline the data management from clinical practice, nursing, health-care management, health-care administration. As for the CDI, data mining is used in the knowledge acquisition and the evidence-based research stage to analyze the information extracted from research reports, reports, evidence tables, flow charts, guidelines that include evidence contents, sources and quality scores.

1.1 PREVIOUS RESEARCH

Many researchers classified the Parkinson's disease by several methods. Fiona O'reilly et al[4]., determined social, psychological, and physical aspects of a person whose partner was affected by Parkinson's disease and result of his study showed that person has slightly worst social, psychological, and physical profiles. Samyra H.J. Keus et al[9]. This method were tested on activities of daily living and motor examination of the Unified Parkinson's Disease Rating Scale and this method leads to decreases symptoms of medication and levodopa dose can be reduced, with a consequent reduction in dyskinesias. Yunfeng Wu et al[7]., focused on the statistical analysis of gait rhythm in patients affected with Parkinson's disease. Kenneth Revett et al., discussed rough set approach in feature selection for Parkinson's disease. They focused on the impairment in voice production and used rough set approach to identify the person with Parkinson's disease.

1.2 PREDICTION MODELS

In this paper, three different types of classification methods are used i.e., decision stump (tree classifier), logistic regression (statistical classifier), and sequential minimization optimization (support vector machine).

1.2.1 Tree classifiers

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes).

In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value. Instances are classified by

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path.

It is a predictive model that uses a set of binary rules applied to calculate a target value. It can be used for classification (categorical variables) or regression (continuous variables) applications. Rules are developed using software available in many statistics packages. Different algorithms are used to determine the "best" split at a node.

Uses training data to build model and the tree generator determines:
i) Which variable to split at a node and the value of the split.
ii) Decision to stop (make a terminal note) or split again.
iii) Assign terminal nodes to a class.

1.2.2 Artificial Neural Network (ANN)

An artificial neural network (ANN), usually called neural network (NN), is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks.
ANNs have been applied to many geotechnical engineering problems such as in pile capacity prediction, modeling soil behavior, site characterization, earth retaining structures, settlement of structures, slope stability, design of tunnels and underground openings, liquefaction, soil permeability and hydraulic conductivity, soil compaction, soil swelling and classification of soils.

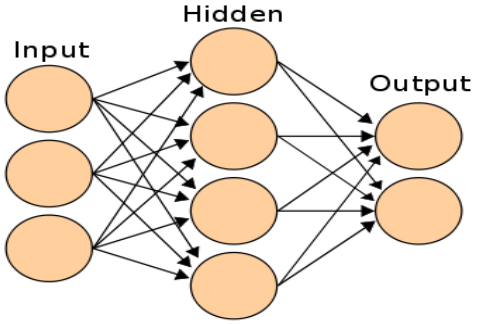

Fig.1 Three layer neural network

There are several types of architecture of NNs. However, the two most widely used NNs **Feed forward networks and recurrent networks.** In a feed forward network, information flows in one direction along connecting pathways, from the input layer via the hidden layers to the final output layer.

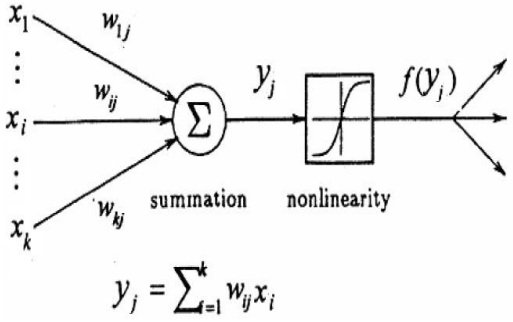

Fig.2 Graphical Representation of neuron in ANN

There is no feedback (loops) i.e., the output of any layer does not affect that same or preceding layer. **Feed-forward neural networks**, where the data from input to output units is strictly feed-forward. The data processing can extend over multiple (layers of) units, but no feedback connections are present.

1.2.3 Support vector machine (SVM)

A support vector machine (SVM) performs classification by constructing an N-dimensional hyper plane that optimally separates the data into two categories. SVM models are closely related to neural network. Support Vector Machine models are a close cousin to classical multilayer perceptron neural network. The goal of SVM modeling is to find the optimal hyper plane that separates clusters of vector in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other size of the plane.

Using a kernel function, SVM's are an alternative training method for polynomial, radial basis function and multi-layer perceptron classifiers in which the weights of the network are found by solving a quadratic programming problem with linear constraints, rather than by solving a non-convex, unconstrained minimization problem as in standard neural network training.

In the parlance of SVM literature, a predictor variable is called an attribute, and a transformed attribute that is used to define the hyper plane is called a feature.
The task of choosing the most suitable representation is known as feature selection. A set of features that describes one case (i.e., a row of predictor values) is called a vector.

Support Vector Machine: Algorithm:
1. Choose the Kernel function.
2. Choose the value for C.
3. Solve the quadratic programming problem.
4. Construct the discriminant function from the support vectors.

1.2.4 Ada Boost

Boosting refers to a general and provably effective method of producing a very accurate prediction rule by combining rough and moderately inaccurate rules of thumb. It is based on the observation that finding many rough rules of thumb can be a lot easier than finding a single, highly accurate classifier. To begin, we define an algorithm for finding the rules of thumb, which we call a weak learner. The boosting algorithm repeatedly calls this weak learner, each time feeding it a different distribution over the training data (in Ada-boost). Each call generates a weak classifier and we must combine all of these into a single classifier that, hopefully, is much more accurate than any one of the rules.
The Ada-Boost algorithm, introduced in 1995 by Freund and Schapire, which solved many of the practical difficulties of the earlier boosting algorithms (which came up with the first provable polynomial-time in 1989). For example, if we want to predict which person has Parkinson disease or not based on the symptoms, we can get a good prediction using Ada-Boost classifier.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

The algorithm takes as input (□1,1),….(□□,□□) where each xi belongs to some domain or instance space X and each level □□ in some level set Y. In most cases, we assume Y = {-1, +1}. Ada-Boost calls a given weak or base learning algorithm repeatedly in a series of rounds t = 1… T. The algorithm will maintain a distribution or set of weights over the training set. The weight of this distribution on training example i on round t is denoted by (i). At first stage all weights set equally, but on each round, the weights of misclassified examples are increased so that the weak learners is forced to focus on hard examples in the training set.

1.2.5 Random Forest

Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. The term came from random decision forests that was first proposed by Tin Kam Ho of Bell Labs in 1995.The method combines Breiman's "bagging" idea and the random selection of features.

Splits are chosen according to a purity measure:
E.g. squared error (regression), Gini index or devinace (classification)
i)How to select N?
Build trees until the error no longer decreases.

ii)How to select M?
Try to recommend defaults, half of them and twice of them and pick the best.

1.3 Dataset information and performance matrices

The dataset was created by Max little University Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. The original study published the feature extraction methods for general voice disorders.

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 (name column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD. The data are in ASCII CSV format.

The rows of the CSV file contain an instance corresponding to one voice recording. There are around six recordings per patient; the name of the patient is identified in the first column.

1.4 Attribute information

Matrix column entries (attributes):
Name: ASCII subject name and recording
Number MDVP: Fo (Hz) - Average vocal fundamental frequency
MDVP: Fhi (Hz) - Maximum vocal fundamental frequency
MDVP: Flo (Hz) - Minimum vocal fundamental frequency
MDVP: Jitter(%), MDVP: Jitter (Abs), MDVP:RAP, MDVP: PPQ,
Jitter: DDP - Several
Measures of variation in fundamental frequency MDVP: Shimmer, MDVP: Shimmer (dB),
Shimmer: APQ3, Shimmer: APQ5, MDVP:APQ, Shimmer: DDA - Several measures of variation in amplitude
NHR, HNR: Two measures of ratio of noise to tonal components in the voice
Status:
Health status of the subject (one) -Parkinson′s, (zero) - healthy RPDE,
D2: Two nonlinear dynamical complexity measures DFA: Signal fractal scaling exponent Spread1, Spread2, PPE: Three nonlinear measures of fundamental frequency variation.

Performance matrices

In this paper, three performance matrices are used to evaluate the performance of the discussed classifiers/methods. Performance of a diagnostic method is usually described in terms of classification accuracy, sensitivity and specificity.

• Accuracy: Accuracy defines relationship between predicted value and actual value i.e., how closes a predicted value to actual value. Accuracy can be defined as:

$$\text{Accuracy} \rightarrow (TP + TN)/(TP + TN + FP + FN)$$

• Sensitivity: It can be defined as predicted the value of output model with respect to change in the input of model. It is used to determine which attribute is more important to obtained correct output value. Now, it can be defined as:

$$\text{Sensitivity} \rightarrow TP/(TP + FN)$$

• Specificity: Specificity can be defined as high degree of confidence. It can be calculated as:
Specificity → TN/(TN + FP)

1.5 K-FOLD CROSS VALIDATION METHOD

To obtain desired result, in this paper[1], k-fold cross validation method is used. This method is also known as rotational method. Basically, k-fold cross method is derived from cross-validation method that is used to measure and compare the learning algorithm.In cross-validation method data are divided in two segments in which one segment is used to learn/train a model and another used to validate model. S. Larsen defined the idea of cross-validation in 1930. Mosteller F. et al[1]., defined the statement for cross-validation that is similar to current k-fold cross method. Cross-validation is appropriate for accuracy predication and selection of model. In k-fold cross-validation dataset is divided into k disjoint folds of equal size d and learning algorithm runs K times. Each iteration of k-fold cross method, , different fold of data is used for validation while k-1 folds are used for learning. Kohavi provides comparative study of several approaches to estimate accuracy that includes cross-validation (including regular cross-validation, leave-one-out cross-

validation, and stratified cross-validation) and bootstrap (sample with replacement) and recommended stratified 10-fold cross-validation the best model selection method. In this paper, the result is based on the comparative study from the various ensemble classifiers with the existing machine learning tools.

II. RESULTS

Classifiers	**20 Features Selected by MRMR**				
	Class	**Accuracy**	**Precision**	**MCC**	**ROC**
Bagging	Parkinson	93.9	90.2	0.66	0.91
	no Parkinson	68.8	78.6	0.66	0.91
	Overall	87.7	87.3	0.66	0.91
Boosting	Parkinson	94.6	93.9	0.76	0.96
	no Parkinson	81.3	83	0.76	0.96
	Overall	91.3	91.2	0.76	0.96
Random Forest	**Parkinson**	**97.3**	**90.5**	**0.73**	**0.96**
	no Parkinson	**68.8**	**89.2**	**0.73**	**0.96**
	Overall	**90.3**	**90.2**	**0.73**	**0.96**
Rotation Forest	Parkinson	96.6	94.7	0.82	0.97
	no Parkinson	83.3	88.9	0.82	0.97
	Overall	93.3	93.2	0.82	0.97
Random Subspace	Parkinson	98	90	0.73	0.95
	no Parkinson	66.7	91.4	0.73	0.95
	Overall	90.3	90.4	0.73	0.95
S V M	Parkinson	100	78.2	0.34	0.57
	no Parkinson	14.6	100	0.34	0.57
	Overall	79	83.6	0.34	0.57
MLP	Parkinson	91.2	93.7	0.71	0.96
	no Parkinson	81.3	75	0.71	0.96
	Overall	88.7	89.1	0.71	0.96
Decision Tree	Parkinson	90.5	90.5	0.61	0.80
	no Parkinson	70.8	70.8	0.61	0.80
	Overall	85.6	85.6	0.61	0.80

III.CONCLUSIONS

The result was compared with many other classifiers as mentioned in the comparison table. The classifier was used on different-2 attributes among the whole data. The different classifiers likes Bagging, Boosting, Rotation forest, Random Subspace, Multi-layer Perceptron, SVM, Decision Tree and Random Forest are used on the different attributes of data set for 5 features, 8 features,10 features ,15 features and 20 features. The result from different attributes showed that Random forest gives the highest accuracy from 5 features is 89%, from 8 features 90.3%, from 15 features 89.2% and from the whole data accuracy is 90.3%, which is always more than the other classifier for the same feature selection scheme.

REFERENCES

[1] Geeta Yadav, Yugal kumar, Gadadhar Sahoo, "Prediction of Parkinson's disease using data mining methods: A Comparative analysis of tree, Statistical, and Support Vector Machine Classifiers" Sunday, October 26, 2014, Indian Journal of Medical Sciences, Vol. 65, No. 6, June 2011

[2] Parkinson J. Neuropsychiatry ClinNeurosci, "An essay on shaking palsy" 2002; ISSN14:223‐36.

[3] Dr. R. Geetha Ramani, G.Sivagami, Shomona Gracia Jacob "Feature Relevance Analysis and Classification of Parkinson's Disease Tele-Monitoring data Through Data Mining" , International Journal of Advanced Research in Computer Science and SoftwareEngineering,vol-2,Issue 3, March 2012.

[4] Farhad Soleimanian Gharehehopogh, Peymen Mohammadi, "A Case Study of Parkinson's Disease Diagnosis Using Artifical Neural Networks" , International Journal of Computer Applications, Vol-73,No.19, July 2013.

[5] JinxinGao, David B. Hitchcock James-Stein "Shrinkage to Improve K-means Cluster Analysis" University of South Carolina, Department of Statistics Nov, 2009

[6] Maria-Luiza Antonie, Osmar R. Zaiane, Alexandru Coman, "Application of Data Mining Techniques for Medical Image Classification" Proceedings of the Second International Workshop on Multimedia Data Mining,2001.

[7] R. Krishnamurthy, Y. Li, S. Raghavan, F. Reiss, S.Vaithyanathan, and H. Zhu, "System T: A System for Declarative Information Extraction," Proceedings of Association for Computing Machinery Special Interest Group on Management of Data Record, vol. 37, 2009.

[8] Tien Fu-Ming," Fractal-based Image Database Retrieval", Master dissertation, Shan University, Taiwan, 2001

[9] Tripti Kapoor, R.K.Sharma, "Parkinson's Disease Diagnosis Using Mel-Frequency Cepstral Coefficients and Vector Quantization", International Journal of Computer Applications, Vol-4, No.3, Jan2011. .

[10] Fayyad, U.M., Djorgovski, S.G., and Weir, N. "Automating the Analysis and Cataloging of Sky Surveys. Advances in Knowledge Discovery and Data Mining",1996.

[11] Kitamoto, A. "Data Mining for Typhoon Image Collection" proceedings of Second International Workshop on Multimedia Data Mining ,2001.

[12] Yuni Xia, Bowei Xi "Conceptual Clustering Categorical Data with Uncertainty", Indiana University – Purdue University Indianapolis, IN 46202, vol 1, 20007.

[13] Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E., Milios, E.: "Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web" In: 7th ACM Intern. Workshop on Web Information and Data Management,2005.
[14] P. K. Singh. "Unsupervised segmentation of medical images using dct coefficients". In VIP '05: Proceedings of the Pan-Sydney area workshop on Visual information processing, pages75–81, 2004.
[15] Manjunath, B.S. & Ma, W.Y. "Texture Features for Browsing and Retrieval of Image Data", IEEE explore Pattern Analysis and Machine Intelligence, vol.18, no. 8, pp. 837-842,1996.
[16] Bagirov, AM., Mardaneh, K. Modified global K-means algorithm for clustering in gene expression data sets workshop on Intelligent Systems for Bioinformatics, Darlinghurst, Australia.
[17] Bohm, C., Kailing, K., Kriegel, H., Kroger, "Density connected clustering with local subspace" preferences 4th, IEEE.International Conference on Data Mining (ICDM '04). Washington, DC, USA ,P 2004..
[18] R. Aggrawal, J. Gehrke , D. Gunopulos and P. Raghavan , " Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications" , In Proceeding of the ACM-SIGMOD Conference On the Management of Data, pp.94-105, 1998.
[19] Aptoula, E. and S. Lefèvre, "Morphological description color images for content-based imageretrieval" published in international journal ISSN 2505-2517.DOI: 10.1109/TIP.2009.
[20] Arangasamy, R., J. Sundararajan, I. Vennila and G.Shankar, "Automatic glucose insulin regulation system-comparison of embedded control design and Harr wavelet method for type-1 diabetes" American Journal of Applied Sciences,11: 433-447. DOI:10.3844/ajassp.2014.433.447
[21] Bencharef, O., B. Jarmouni and A. Souissi, "Research similar images based global descriptors multiple clustering". International Journal of Engineering and Technology, 5: 3142-315, 2013.
[22] Huang, M., Bian, F. "An improved density-based spatial clustering algorithm based on key factors of object's distribution" International Joint Conference on Artificial Intelligence (IJCAI '09),2009.
[23] Calcagno, S., F.L. Foresta and M. Versaci, "Independent component analysis and discrete wavelet transform for artifact removal in biomedical signal processing". Am. J. Applied Sci., 11: 57-68.DOI: 10.3844/ajassp.2014.57.68, 2014.
[24] Chun, Y.D., S.Y. Seo and N.C. Kim, "Image retrieval using BDIP and BVLC moments" proceedings of IEEE transaction on Circuits System Video Technology, 9:951-957.DOI:10.1109/TCSVT.2003. 816507, 2003.
[25] M. A. Little, P. E. McSharry, E. J. Hunter and L. O. Ramig, "Suitability of Dysphonia Measurements for Tele monitoring of Parkinson's Disease," IEEE Transactions on Biomedical Engineering, 2008.
[26] Udaya kumar and Magesh kumar. "Classification of Parkinson's disease using Multipass Lvq, Logistic Model Tree, K-star for Audio Dataset, classification of Parkinson Disease using Audio Dataset". Dalarna University, School of Technology and Business Studies, Computer Engineering, 2011