MACHINE LEARNING BASED APPROACHES FOR PREDICTION OF PARKINSON'S DISEASE

Arvind Kumar Tiwari

GGS College of Modern Technology, SAS Nagar, Punjab, India

Abstract

The prediction of Parkinson's disease is most important and challenging problem for biomedical engineering researchers and doctors. The symptoms of disease are investigated in middle and late middle age. In this paper, minimum redundancy maximum relevance feature selection algorithms is used to select the most important feature among all the features to predict the Parkinson diseases. Here, it is observed that the random forest with 20 number of features selected by minimum redundancy maximum relevance feature selection algorithms provide the overall accuracy 90.3%, precision 90.2%, Mathews correlation coefficient values of 0.73 and ROC values 0.96 which is better in comparison to all other machine learning based approaches such as bagging, boosting, random forest, rotation forest, random subspace, support vector machine, multilayer perceptron, and decision tree based methods.

Keywords

Parkinson disease, machine learning, bagging, random forest, minimum redundancy maximum relevance, k-fold cross validation

1. INTRODUCTION

Now a days, there are variousneurodegenerativediseases that have been recognized such as Alzheimer's disease, Parkinson disease, Arthritic disease, Dementia with Lewy bodies, Corticobasal degeneration, Progressive supranuclear palsy and Prion disorders [1]. Among all of these neurodegenerative and coordinating the body movement's diseases, Parkinson's disease is second most common disease after Alzheimer's. The core clinical feature of the Parkinson's disease is described by the authors of the paper [2]. The medical information is essential for diagnosis and patient care [3]. For clinical research, it also provides useful information to facilitate therapeutic improvement and conduct medical researches. The medical knowledge management in the realm of medical information can be shown as the cycle among the clinical research, 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 is emerged in the new area to streamline the data management from clinical practice, nursing, health-care management, health-care administration. As for the clinical decision intelligence, machine learning based methods are 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.

There are various researchers classified the Parkinson's disease by several methods. The authors of the paper [3] have been used various data mining methods for the prediction of Parkinson diseases. The authors of the paper [4] also used various data mining methods with the data set

DOI:10.5121/mlaij.2016.3203

consisting various vocal attribute of Parkinson disease affected persons. The authors of the paper [5] are developed by the voice measurements of disease mainly focuses the speech signals. The Parkinson dataset is range of biomedical voice measurement from 31 people 23 characteristic features in Parkinson's disease. The authors of the paper [6] are also presented by three models to analysis the Parkinson's disease for error probability calculated by, logistic regression analysis, decision tree analysis and neural net analysis. The authors of the paper [7] are presented by speech of vocal sound test for the Parkinson's disease patients to compare the health control (HC) people. The authors of the paper [8] are also evaluated Artificial Neural Networks (ANN) and Support Vector Machines (SVM) for the vocal datasets. The authors of the paper [9] have been proposed the Multi-Layer Perceptron (MLP) with back- propagation learning algorithm and Radial Basis Function (RBF) to predict the Parkinson diseases.Here, in this paper various machine learning based methods such as bagging, boosting, random forest, rotation forest, random subspace, support vector machine, multilayer perceptron, and decision tree based methods are used with minimum redundancy maximum relevance feature selection algorithms to select the most important feature among all the features from the speech articulation difficulty symptoms of Parkinson's disease affected person to predict the Parkinson disease. The authors of the paper [10] have been proposed an ensemble method that includes sparse multinomial logistic regression, rotation forest ensemble with support vector machines and principal components analysis for the prediction of Parkinson disease. The authors of the paper [11] have been studied and adopted a novel metaheuristic data mining algorithm for the detection and classification of Parkinson's disease.The authors of the paper [12] have been proposed a fuzzy neural system (FNS) based method for the classification of Parkinson diseases.The authors of the paper [13] have been proposed a fuzzy k-nearest neighbour based methodfor the classification of Parkinson diseases. The authors of the paper [14] have been studied and proposed support vector machine based method for the prediction of Parkinson disease.

2.MATERIAL AND METHODS

2.1. Dataset Description

Here, the dataset was created by the authors of the paper [15] Max little University Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals, is used. 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. There are various attributed extracted that are defined as follows:

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 SeveralMeasures 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.

2.2.Machine Learning Based Approaches

Here, in this paper various machine learning based methods such as bagging, boosting, random forest, rotation forest, random subspace, support vector machine, multilayer perceptron, and decision tree based methods are used with minimum redundancy maximum relevance feature selection algorithms [16] to select the most important feature among all the features from the speech articulation difficulty symptoms of Parkinson's disease affected person to predict the Parkinson disease.

2.2.1. Random Forests

Random forest classifier [17] used an ensemble of random trees. Each of the random trees is generated by using a bootstrap sample data. At each node of the tree a subset of feature with highest information gain is selected from a random subset of entire features. Thus random forest used bagging as well as feature selection to generate the trees. Once a forest is generated every tree participates in classification by voting to a class. The final classification is based on the majority voting of a particular class. It performs better in comparison with single tree classifiers such as CART and C 5.0 etc.

3. PERFORMANCE MATRICES

In this paper, 10-fold cross validation is used to measure the performance of various machine learning based methods. In this paper, accuracy (ACC), Precision, Receiver Operating Characteristics (ROC) and Matthew's correlation coefficient (MCC) is used to measure the performance.

Accuracy is measured by the following formulae.

$ACC(i) = \frac{C(i)}{T(i)}, \quad i = 1, 2, \dots$ .... where T(i) is the total number of sequences in class i, C(i) is the correctly predicted sequences of class i and n is the total number of classes.

MCC is a balanced measure that considers both true and false positives and negatives. The MCC can be obtained as

$$MCC = \frac{(\mathrm{TP})(\mathrm{TN}) - (\mathrm{FP})(\mathrm{FN})}{\sqrt{[\mathrm{TP} + \mathrm{FP}][\mathrm{TP} + \mathrm{FN}][\mathrm{TN} + \mathrm{FP}][\mathrm{TN} + \mathrm{FN}]}}$$

Where TP is the true positive, TN is the true negative, FP is the false positive, and FN is the false negative.
Precision is the proportion of instances classified as positive that are really positive. It is defined as

$$\text{Precision} = \frac{\mathrm{TP}}{(TP + FP)}$$

Area under ROC curve (AUC) of a classifier is the probability that the classifier ranks a randomly chosen positive instance higher than a randomly chosen negative instance.

4.RESULT AND COMPARATIVE ANALYSIS

In this paper, various machine learning based methods such as bagging, boosting, random forest, rotation forest, random subspace, support vector machine, multilayer perceptron, and decision tree based methods are used to predict the Parkinson disease. The minimum redundancy maximum relevance feature selection algorithms is used to select the most important feature among all the features from the speech articulation difficulty symptoms of Parkinson's disease affected person to predict the Parkinson disease.

Here, theminimum redundancy maximum relevance feature selection algorithms is used to select the 5 number of features, 8 number of features, 10 number of features, 15 number of featuresand 20 number of features among all the features. Here, the performance ofvarious machine learning based methods such as bagging, boosting, random forest, rotation forest, random subspace, support vector machine, multilayer perceptron, and decision tree based methods are computed with the different-different features selected by minimum redundancy maximum relevance feature selection algorithms (See Table-1).

Here, it is observed that the random forest with 20 numbers of features selected by minimum redundancy maximum relevance (MRMR) feature selection algorithms provide the overall accuracy 90.3%, precision 90.2%, Mathew's correlation coefficient values of 0.73 and ROC values 0.96 which is better in comparison to all other machine learning based approaches (See Table-1).

Table-1 Result analysis for prediction of Parkinson diseases with various machine learning based approaches using different features selected by MRMR

Classifiers	5 Features Selected by MRMR					8 Features Selected by MRMR			
	Class	Accuracy	Precision	MCC	ROC	Accuracy	Precision	MCC	ROC
Bagging	Parkinson	93.9	87.9	0.59	0.89	95.9	89.8	0.68	0.90
	no Parkinson	60.4	76.3	0.59	0.89	66.7	84.2	0.68	0.90
	Overall	85.6	85	0.59	0.89	88.7	88.4	0.68	0.90
Boosting	Parkinson	91.2	89.3	0.59	0.82	91.2	88.7	0.57	0.88
	no Parkinson	66.7	71.1	0.59	0.82	64.6	70.5	0.57	0.88
	Overall	85.1	84.8	0.59	0.82	84.6	84.2	0.57	0.88
Rotation Forest	Parkinson	96.6	87.1	0.62	0.90	95.9	91	0.71	0.93
	no Parkinson	56.3	84.4	0.62	0.90	70.8	85	0.71	0.93
	Overall	86.7	86.4	0.62	0.90	89.7	89.5	0.71	0.93
Random Subspace	Parkinson	94.6	84.8	0.50	0.79	95.9	86	0.57	0.90
	no Parkinson	47.9	74.2	0.50	0.79	52.1	80.6	0.57	0.90
	Overall	83.1	82.2	0.50	0.79	85.1	84.7	0.57	0.90

S V M	Parkinson	95.2	80.9	0.36	0.63	97.3	83.1	0.49	0.68
	no Parkinson	31.3	68.2	0.36	0.63	39.6	82.6	0.49	0.68
	Overall	79.5	77.8	0.36	0.63	83.1	83	0.49	0.68
MLP	Parkinson	93.9	80.2	0.31	0.78	95.9	86.5	0.58	0.89
	no Parkinson	29.2	60.9	0.31	0.78	54.2	81.3	0.58	0.89
	Overall	77.9	75.5	0.31	0.78	85.6	85.2	0.58	0.89
Decision Tree	Parkinson	88.4	87.2	0.50	0.72	94.6	87.4	0.59	0.80
	no Parkinson	60.4	63	0.50	0.72	58.3	77.8	0.59	0.80
	Overall	81.5	81.3	0.50	0.72	85.6	85	0.59	0.80
Random Forest	**Parkinson**	**94.6**	**91.4**	**0.70**	**0.92**	**95.9**	**91.6**	**0.73**	**0.90**
	no Parkinson	**72.9**	**81.4**	**0.70**	**0.92**	**72.9**	**85.4**	**0.73**	**0.90**
	Overall	**89.2**	**89**	**0.70**	**0.92**	**90.3**	**90**	**0.73**	**0.90**

Classifiers	**10 Features Selected by MRMR**					**15 Features Selected by MRMR**			
	Class	**Accuracy**	**Precision**	**MCC**	**ROC**	**Accuracy**	**Precision**	**MCC**	**ROC**
Bagging	Parkinson	95.2	89.2	0.65	0.90	95.2	89.2	0.65	0.90
	no Parkinson	64.6	81.6	0.65	0.90	64.6	81.6	0.65	0.90
	Overall	87.7	87.3	0.65	0.90	87.7	87.3	0.65	0.90
Boosting	Parkinson	93.2	91.9	0.69	0.92	95.2	89.2	0.65	0.90
	no Parkinson	75	78.3	0.69	0.92	64.6	81.6	0.65	0.90
	Overall	88.7	88.6	0.69	0.92	87.7	87.3	0.65	0.90
Random Forest	**Parkinson**	**95.9**	**91.6**	**0.73**	**0.93**	**95.9**	**90.4**	**0.70**	**0.92**
	no Parkinson	**72.9**	**85.4**	**0.73**	**0.93**	**68.8**	**84.6**	**0.70**	**0.92**
	Overall	**90.3**	**90**	**0.73**	**0.93**	**89.2**	**89**	**0.70**	**0.92**
Rotation Forest	Parkinson	95.2	90.9	0.70	0.94	95.2	89.7	0.67	0.92
	no Parkinson	70.8	82.9	0.70	0.94	66.7	82.1	0.67	0.92
	Overall	89.2	88.9	0.70	0.94	88.2	87.9	0.67	0.92
Random Subspace	Parkinson	97.3	85.1	0.56	0.88	96.6	86.1	0.58	0.89
	no Parkinson	47.9	85.2	0.56	0.88	52.1	83.3	0.58	0.89
	Overall	85.1	85.1	0.56	0.88	85.6	85.4	0.58	0.89
S V M	Parkinson	97.3	82.7	0.47	0.67	98.6	78.8	0.33	0.59
	no Parkinson	37.5	81.8	0.47	0.67	18.8	81.8	0.33	0.59
	Overall	82.6	82.5	0.47	0.67	79	79.5	0.33	0.59
MLP	Parkinson	96.6	88.2	0.65	0.89	93.2	89	0.61	0.92
	no Parkinson	60.4	85.3	0.65	0.89	64.6	75.6	0.61	0.92
	Overall	87.7	87.5	0.65	0.89	86.2	85.7	0.61	0.92
Decision Tree	Parkinson	94.6	87.4	0.59	0.78	93.2	91.3	0.68	0.86
	no Parkinson	58.3	77.8	0.59	0.78	72.9	77.8	0.68	0.86
	Overall	85.6	85	0.59	0.78	88.2	88	0.68	0.86

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

5. CONCLUSIONS

The prediction of Parkinson's disease is most important and challenging problem for biomedical engineering researchers and doctors. In this paper, minimum redundancy maximum relevance feature selection algorithms was used to select the most important feature among all the features to predict the Parkinson diseases. Here, it was observed that the random forest with 20 number of features selected by minimum redundancy maximum relevance feature selection algorithms provide the overall accuracy 90.3%, precision 90.2%, Mathews correlation coefficient values of 0.73 and ROC values 0.96 which is better in comparison to all other machine learning based approachessuch as bagging, boosting, random forest, rotation forest, random subspace, support vector machine, multilayer perceptron, and decision tree based methods.

6. REFERENCES

[1]. L. Ramig, R. Sherer, I. Titze and S. Ringel, "Acoustic Analysis of Voices of Patients with Neurologic Disease: Rationale and Preliminary Data," The Annals of Otology, Rhinology, and laryngology, No. 97, pp. 164-172, 1988.

[2]. Parkinson, James. "An essay on the shaking palsy." The Journal of neuropsychiatry and clinical neurosciences ,2002.

[3]. Dr.R.GeethaRamani, G.Sivagami, ShomonaGraciajacob " Feature Relevance Analysis and Classification of Parkinson's Disease TeleMonitoring data Through Data Mining" , International Journal of Advanced Research in Computer Science and Software Engineering,vol-2,Issue 3, March 2012.
[4]. PeymanMohammadi, AbdolrezaHatamlou and Mohammed Msdaris "A Comparative Study on Remote Tracking of Parkinson's Disease Progression Using Data Mining Methods" , International Journal in Foundations of Computer Science and Technology(IJFCST),vol3,No.6, Nov 2013.
[5]. Dr.R.GeethaRamani and G.Sivagami "Parkinson Disease Classification using Data Mining Algorithms", International Journal of Computer Applications (IJCA),Vol-32,No.9, October 2011.
[6]. Shanghais Wu, JiannjongGuo "A Data Mining Analysis of the Parkinson's Disease", Scientific Research, iBusiness, 3, 71-75, 2011.
[7]. Rusz, Jan, et al. "Acoustic analysis of voice and speech characteristics in early untreated Parkinson's disease." MAVEBA. 2011.
[8]. Gil, David, and Magnus Johnson. "Diagnosing parkinson by using artificial neural networks and support vector machines." Global Journal of Computer Science and Technology 9.4: 63-71, 2009.
[9]. FarhadSoleimanianGharehehopogh, PeymenMohammadi, "A Case Study of Parkinson's Disease Diagnosis Using Artifical Neural Networks" , International Journal of Computer Applications, Vol73,No.19, July 2013
[10]. Mandal, Indrajit, and N. Sairam. "New machine-learning algorithms for prediction of Parkinson's disease." International Journal of Systems Science 45.3: 647-666, 2014.
[11]. Suganya, P., and C. P. Sumathi. "A Novel Metaheuristic Data Mining Algorithm for the Detection and Classification of Parkinson Disease." Indian Journal of Science and Technology 8.14: 1, 2015.
[12]. Abiyev, Rahib H., and SananAbizade. "Diagnosing Parkinson's Diseases Using Fuzzy Neural System." Computational and Mathematical Methods in Medicine, 2016 (2016).
[13]. Chen, Hui-Ling, et al. "An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach." Expert Systems with Applications 40.1: 263-271, 2013.
[14]. Sriram, T. V., et al. "Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms." Int J EngInnovTechnol 3.3: 212-5, 2013.
[15]. Little, Max A., et al. "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection." BioMedical Engineering OnLine6.1 : 23, 2007.
[16]. Ding, C., &Peng, H.,"Minimum redundancy feature selection from microarray gene expression data", Journal of bioinformatics and computational biology, 3(02), 185-205, 2005.
[17]. Breiman, L., "Random forests. Machine learning", 45(1), 5-32, 2001.

Author

Dr. Arvind Kumar Tiwari received the BE degree in Computer Science and Engineering from CCS university, Meerut., INDIA in 2003, and M. Tech. in Computer Science and Engineering from Uttar Pradesh Technical University, Lucknow and Ph.D. in Computer Science and Engineering from IIT (BHU), Varanasi, INDIA. He is working as and Professor and Vice Principal in GGS College of Modern Technology, Kharar, Punjab, INDIA. His research interests include computational biology and pattern recognition.