Parkinson Disease Prediction Using Machine Learning Algorithm

Richa Mathur, Vibhakar Pathak and Devesh Bandil

Abstract Parkinson disease, the second most common neurological disorder that causes significant disability, reduces the quality of life and has no cure. Approximately, 90% affected people with Parkinson have speech disorders. The medical dataset contains heterogeneous data in the form of text, numbers, and images that can be mined. Big Data has the potential to give valuable information after processing that can be discovered through deep analysis and efficient processing of data by decision-makers. Data mining is the process of selecting, extracting, and modeling the unknown hidden patterns from large datasets. Machine learning algorithm (MLA) can be used for early detection of disease to increase the chances of elderly people's lifespan and improved lifestyle with Parkinson. In this paper, we use various MLAs that can help in improving the performance of datasets and play a vital role in making the early prediction of disease at right time. After comparison of these algorithms, we choose the most effective one in terms of accuracy. From our experimental results, it is analyzed that the accuracy obtained from the combined effect of KNN algorithm with ANN is better as compared to other algorithms.

Keywords Parkinson disease · Predictive analytics · Voice datasets
SVM · KNN · ANN

R. Mathur (✉)
Department of CS & Applications, SGVU, Jaipur, Rajasthan, India
e-mail: richa0058@gmail.com

V. Pathak
CS/IT, Arya College of Engineering and IT, Jaipur, Rajasthan, India
e-mail: vibhakarp@rediffmail.com

D. Bandil
Computer Applications, Suresh Gyan Vihar University, Jaipur, Rajasthan, India
e-mail: mcagwailor@gmail.com

V. S. Rathore et al. (eds.), *Emerging Trends in Expert Applications and Security*,
Advances in Intelligent Systems and Computing 841,
https://doi.org/10.1007/978-981-13-2285-3_42

1 Introduction

Parkinson disease (PD), a neurodegenerative disorder of CNS system, also produces movement disorder. In the primary stage when nerve cells or neurons of the brain become impaired, people may begin to notice symptoms like tremor, stiffness in limb or trunk of the body, movement issues, or impaired balance. With the progression of the disease, people may have difficulty in walking, talking, or completing other simple tasks. PD has no cure but several treatments are known to provide relief from the symptoms. PD affects brain cells that produce Dopamine, a neurotransmitter, which is responsible for coordinate and control muscle activity [1].

Typically, PD occurs in people over the age of 60, among which 1% of people are affected. It is called as young onset PD when it is seen in the people before age 50 [2]. According to estimates, PD affected 6.2 million people and about 117400 deaths globally in 2015 [3].

Nowadays, data is becoming more valuable but how to handle data and finding hidden facts from it is more important. The term “Big Data” describes a large amount of datasets that are so complex that it is not possible to process them via conventional methods and technologies. To extract valuable insights from such varied and rapid growing datasets, various tools and techniques of Big Data analytics can be used that may lead to better decision-making and strategic planning.

The rising population generates a large amount of data related to patients clinical and laboratory tests. With this dataset, doctors can detect and diagnose the disease at their early stages. Early prediction of motor symptoms of PD can get a proper treatment at right time to a patient.

2 Related Work

Shamli and Sathiyabhama [4] proposed multi-classifier system, i.e., based on Big Data analytics to improve predictive performance and efficient time to answer cost-effective actions. The author introduced Big Data with its characteristics and Big Data analytics with their types as Descriptive, Predictive and Prescriptive in healthcare industries. Dopamine, a neurotransmitter, generated by brain cells, is responsible to send signals to other brain cells to control muscle activity. The degeneration of dopamine-producing brain cells causes PD. For analysis purposes, voice dataset of PD is collected from UCI machine learning library. By implementing multiple predictive models to disease datasets, multiple accuracies and results of different classifiers are acquired. C4.5, SVM, and ANN give better results than other machine learning algorithms. After comparing the results of these classifiers, best results are chosen for the final decision. This approach helps organizations to analyze their large datasets quickly and efficiently with maximum accuracy.

Azad et al. [5] explored a predictive model for PD that is based on decision tree algorithm. They introduced PD, a second most common neurodegenerative disease with its symptoms, possible complications, and risk factors associated with it. Various applications of data mining are used for classification purposes that are decision tree, attribute selection measures, ID3 and decision stumps. Their dataset (have 197 instances) is taken from UCI repository and built up from the data of 31 people. For performance analysis, two parameters accuracy and classification error are used. For validation, 10-fold cross-validation technique is used that gives the unbiased outcome. They found that decision tree algorithm performs best and gives the best accuracy and less classification error than other algorithms in their experimental results.

Sriram et al. [6] proposed a method for diagnosis of PD using its voice dataset. This voice dataset is built up from the voice of 31 people among which 23 people are affected by PD. This dataset contains 5875 instances and 26 attributes. In their experiment for statistical analysis, classification, evaluation, visualization, and unsupervised methods Weka V3.4.10 and Orange V2.0b software are used. They achieved the best accuracy 90.2% from Random Forest algorithm.

3 Problem Definition

The huge amount of data known as Big Data is generated everywhere, and this data can be used to perform analysis and make future predictions. According to some research, most of the healthcare data is in the unstructured form that can be stored in the centralized repository to make useful interpretation out of it. To improve the quality of patient care at low cost, this unstructured data can be analyzed further by merging it with structured datasets. The problem is to classify and discover data pattern to predict future disease, so that doctors can detect and diagnose the disease at an early stage.

4 System Architecture

The dataset used in this paper is taken from UCI library [7]. Analysis of these data will provide early diagnosis and detection of disease at reducing cost [4]. The gathered information is in unstructured format, i.e., it is not in a particular kind of format. After that, we convert this mixed type of disease data into the structured form which will ease the process.

For that, a layered Big Data framework is used which has mainly following three components:

Hadoop: A very popular, distributed processing, and storage framework that can handle large and complex unstructured data. Therefore, Hadoop is the best option for analyzing unstructured disease dataset. Hadoop has two main components: MapReduce and HDFS for processing and storing a large amount of datasets. Hadoop uses HDFS to store very large data files (in GBs to TBs) that cannot be stored on a single machine. MapReduce is a software programming paradigm used to process a large amount of datasets by distributing the work to various independent nodes.

Predictive Analytics: Is a probabilistic platform for predicting the future. It uses a variety of statistical modeling, data mining techniques, and machine learning techniques. It provides actionable insights based on the data so that organizations can identify a pattern from the data and apply statistical modeling technique and algorithms to find relationships between various datasets [4].

Prediction Models: To classify data various machine learning (ML) algorithms are used. ML (i.e., a branch of Artificial Intelligence (AI) concerned with the study of classification and pattern analysis) algorithm allows us to automatically recognize complex patterns and make intelligent decisions based on data. Some ML algorithms that we used in this paper are as follows:

Support Vector Machine **(SVM)** yields more accurate results when it is used for classifying text. It is successfully employed in text classification and various other sequence processing applications as it is a type of linear classifier.

Artificial Neural Network **(ANN)** is a type of supervised learning models and it is derived from the functionality of human brains, i.e., highly sophisticated analytical technique, capable of modeling extremely complex nonlinear functions [8]. We used a popular ANN algorithm called multilayer perceptron **(MLP)**, i.e., a type of supervised learning model used for prediction and classification problems [9].

5 Proposed Model

We use Waikato Environment for Knowledge Analysis (WEKA) to implement data mining algorithms for preprocessing, classification, clustering, and analysis of results. This environment includes java libraries that implement algorithms and provide the best environment to researchers for classifying datasets.

5.1 Data Collection

The dataset used in this paper is taken from UCI machine learning library [7]. The dataset consists of 195 instances and 24 attributes. This feature set consists of name, Fo(Hz), Fhi(Hz), Flo(Hz), Jitter(%), Jitter(Abs), RAP, PPQ, Jitter:DDP, Shimmer, Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, APQ, Shimmer:DDA, NHR, HNR, status, RPDE, DFA, spread1, spread2, D2, and PPE. We store these datasets in .CSV format and then convert it into. ARFF format for further analysis. The dataset is divided into two classes according to its "status" column which is set to 0 for healthy subjects and 1 for those PD [10].

5.2 Data Preprocessing

The poor data quality in the medical dataset is one of the big challenges that are faced by the knowledge discovery process. This process decreases the number of attributes into a better subset which can increase accuracy, and also it brings a reduction in training time. It is done using Filters and Wrappers. WEKA provides "AttributeSelection" filter to choose an attribute evaluation method. We use "cfsSubsetEval" attribute evaluator and "BestFirstSearch" method which considers the individual predictive ability of each feature to evaluate the worth of an attribute. From that, a new feature data subset is prepared which contains 11 features.

5.3 Data Mining

In this proposed framework, we used different classification techniques for analyzing PD patient's record. To evaluate performance, we apply 10-fold cross-validation technique which splits the original set into training sample to train the model and a test set to evaluate results.

An approach known as "Information Retrieval Metrics" can be used to evaluate experimental results in terms of precision, recall, f-measure, and accuracy with the use of following formulas [11]:

$$\text{Precision} : \text{TP}/(\text{TP} + \text{FP}); \quad \text{Recall} : \text{TP}/(\text{TP} + \text{FP})$$
$$\text{F} - \text{measure} : 2 * \ \text{Precision} * \text{Recall}/(\text{Precision} + \text{Recall})$$
$$\text{Accuracy} : \text{TP} + \text{TN}/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

Here, TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative [12].

6 Experimental Results

Various techniques are used in the analysis and prediction of PD. Methods that are based on analytics can give an appropriate prediction for a particular disease by grouping people with similar symptoms. In our experiment, obtained accuracies using SMO, KNN [13], Random Forest, AdaBoost.M1 [14], Bagging, MLP, and DT algorithms are 86.67%, 90.76%, 89.23%, 88.20%, 89.23%, and 89.74%,

Table 1 Performance measured by classifiers

	KNN + AdaBoosta.M1	KNN + Bagging	KNN + MLP
Accuracy	91.28%	90.76	91.28
Classification error	8.717	9.23	8.71
Time taken to build model	0.01	0.02	0.43
Precision	0.873	0.866	0.873
Recall	0.907	0.904	0.907
F-Measure	0.888	0.882	0.888

respectively. In our experiment, we used several ensemble methods that are capable to combine classifiers with their predictions and base estimators. For using more than one classification model, we have to "meta" option under classification tab, and then select "vote" classifier. After that, select the classifier properties and in classifiers tab, we can add multiple classifiers as per our need. KNN provides better accuracy and less execution time than SMO and random forest. So we combined ANN algorithms with KNN algorithm. Table 1 shows performance measure that is reported by our experimental result with disease dataset, after conducting 10-fold cross-validation technique.

Figure 1 shows overall precision, recall, and F-measure rate of combined approach. The accuracy acquired by AdaBoosta.M1 and MLP with KNN are same, i.e., 91.28% which is better when we use these separate algorithms. Our experimental result shows that same preprocessing methods on a different dataset affect similarly the classifiers performance. After analyzing results, it is observed that when we combine two classifiers, accuracy is increased and time taken to build model is reduced. The accuracy acquired by AdaBoosta.M1 and MLP with KNN are same, i.e., 91.28% which is better when we use these separate algorithms. And this accuracy becomes 100% when we use training dataset with selected attributes.

Time taken to build a model with AdaBoosta is 0.01 s, whereas time taken to build model with MLP is 0.43 s. So that AdaBoosta1 with KNN gives best accuracy with time and less classification error.

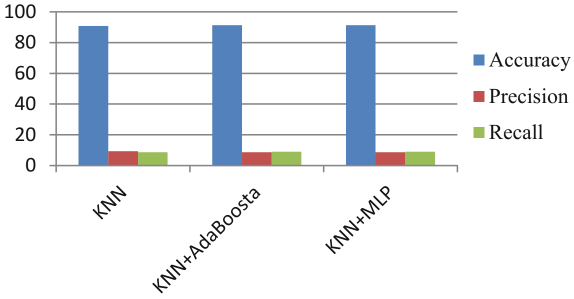

Fig. 1 Comparison of accuracy achieved by ANN algorithms with KNN

7 Conclusion and Future Scope

Big Data analytics plays a huge role in the healthcare industry, as these data are scattered everywhere, big, and complex in nature. In this paper, we discuss early stage prediction of Parkinson disease for that we presented a methodology of data mining using Weka tool for classifying disease dataset. We use various MLAs for classifying our experimental data that indicate the combined effect of ANN algorithms with KNN which is better as compared when we use other algorithms. The system detects the maximum accuracy of the multi-classifier, and their result predicts the disease at its early stage. We discuss the comparative analysis and calculate the overall performance measures in terms of precision, recall, and f-measure. In future, effective optimization techniques can be used to achieve better accuracy and cost-effective interventions for Parkinson disease. Also, limited data is available that describes the real potential of early PD treatment which requires more research to explore the real impact of early treatment.

References

1. https://en.wikipedia.org/wiki/Parkinson%27s_disease
2. PD: hope through research. https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Hope-Through-Research/Parkinsons-Disease-Hope-Through-Research
3. Saloni et al (2015) Detection of Parkinson disease using clinical voice data mining
4. Shamli N et al (2016) Parkinson's Brain disease prediction using big data analytics
5. Azad C et al (2014) Design and analysis of data mining based prediction model for Parkinson's disease
6. Sriram TVS et al (2013) Intelligent Parkinson disease prediction using machine learning algorithms
7. Shaikh, TA (2014) A prototype of Parkinson's and Primary tumor diseases prediction using data mining techniques
8. Kirubha V et al (2016) Survey on data mining algorithms in disease prediction
9. Salekin A Detection of chronic Kidney disease and selecting important predictive attributes
10. Gaur V et al (2013) A multi-objective optimization of cloud based SLA-Violation prediction and adaptation
11. PD dataset from UCI repository. https://archive.ics.uci.edu/ml/machine-learning-databases/parkinsons/
12. Boukenze B et al (2016) Performance on data mining techniques to predict in healthcare industry: Chronic Kidney failure disease
13. Rana M et al (2015) Breast Cancer diagnosis and recurrence prediction using machine learning techniques
14. Freund Y et al (1996) Experiments with a new boosting algorithm