

A Comparative study of Classification techniques in data mining for distinctive Neurological disorders

D.Karthiga
Ph.D, Research scholar

Dr.P.Sumitra
Assistant Professor, Dept of Computer Science,
Vivekanandha College of Arts and Sciences for Women (Autonomous)
Elayampalayam, Namakkal (DT)-637205

Abstract: Data Mining includes a set of techniques and methodologies directed to excision of data from the vast amount of database. Classification is the most important technique to classify several kinds of data used in day-to-day life in medical field. The early detection of the heterogeneous neurological disorders through its symptoms and treatment can be essential in increasing the life expectancy of the patient affected with a particular disease. The aim is to contemplate the prognosis of several disorders through its symptoms for the practitioners to detect and provide correct treatment in the primordial stage. According to this review, different methods for classification techniques such as Decision tree induction, Rule-Base method, Naive Bayes, Support Vector Machine have been briefly discussed.

Keywords: Decision Tree Induction, Rule-Base Method, Naïve Bayes, Support Vector Machine

I.INTRODUCTION

Data Mining is the method of finding out patterns from vast data sets. These patterns are used to make some effective decision for the development in Medical field. It is also referred as a Knowledge discovery process, Knowledge mining from data, Knowledge extraction or data pattern analysis. Data Mining is the multidisciplinary subfield of computer science and statistics with an inclusive objective to remove data from a dataset and transmit the data into possible patterns for future use. The massive importance is given in the medical field for using data mining to prognosis the neurological disorder in primordial stage, avoiding of hospital mistakes, and detection of fraudulent insurance claims. Early treatment helps patients to increase the span of years to live in this world.

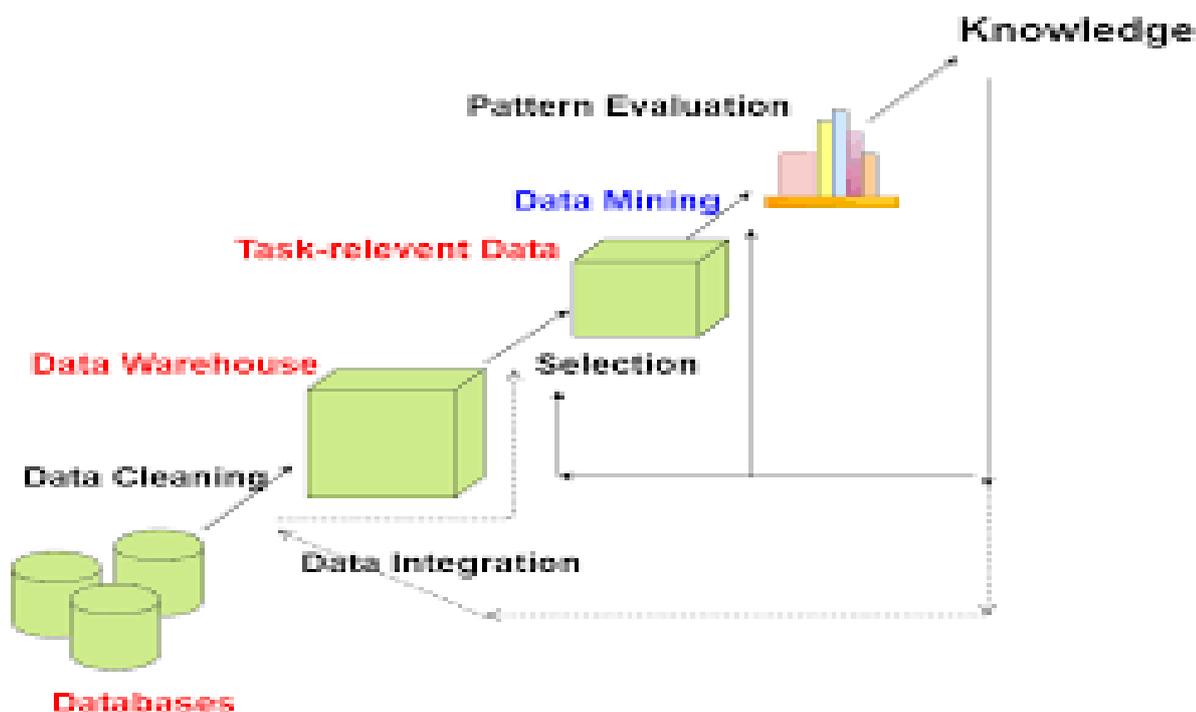


Fig 1.1 Knowledge Process

II. CLASSIFICATION

The Classification is one of the major techniques in Data Mining. Data Mining Techniques widely classified into two kinds. They are Predictive and Descriptive. These two kinds of methods are used to excision the hidden patterns from the vast amount of data. Classification are used to classify each item in a data set into a predetermined set of classes. In this technique, data will be split as a training data set. It uses the prediction to predict the category labels. The goal of classifying is to accurately predict the target class for each case in the data. A Classification task commence with a dataset in which class labels are known. It is categorized into Supervised and Unsupervised Classification. The training data are associated with labels indicating the analysis of the class is known as Supervised Classification. The Class labels of training data are unidentified are known as Unsupervised Classification.

The working condition of the Data Classification process includes two steps. i) Learning of a mapping ii) Using Classifier for Classification. The first step of Classification process can be reviewed as learning of a mapping. The Mapping process that separates the data classes. Classification Rules, Decision Tree or Mathematical formulae are used to represent the mapping process. The mapping is represented as Classification rules that used to identify the patients affected with which type of neurological disorder as being either first, mid or last phases. The second step is to estimate the predictive accuracy of the classifier. The training dataset is used to measure the accuracy of the classifier; this estimate would be the best, because the classifier tends to fit the data. For example, Classification model can be used to categorize the stage of the patients affected with neurological disorders as first, mid and last stages through symptoms and suitable treatment can be given at the right time.

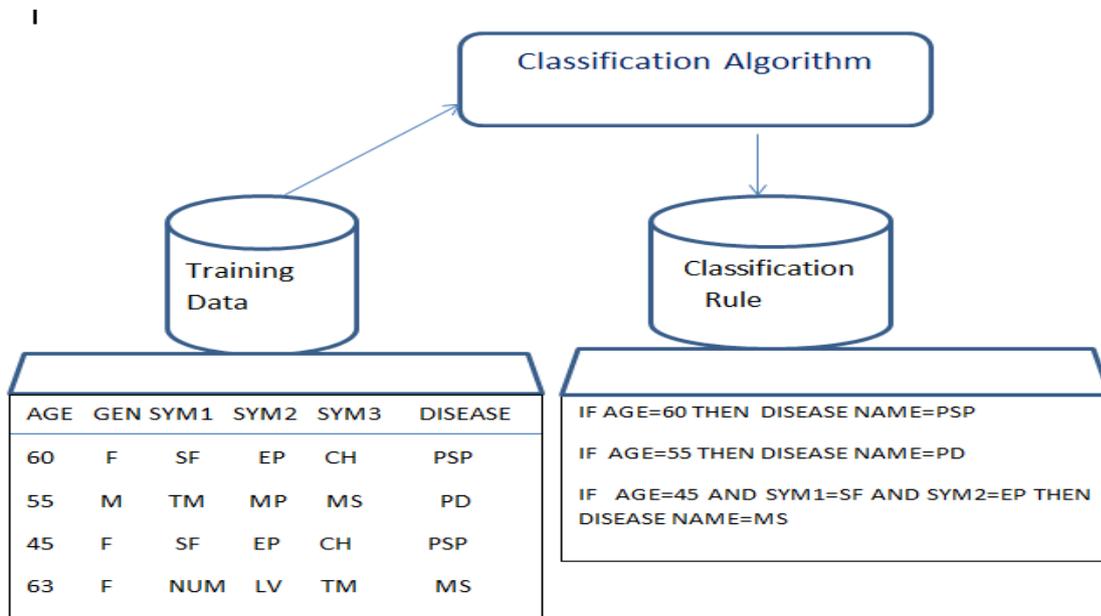


Fig 1.2 (i) Learning of a Mapping

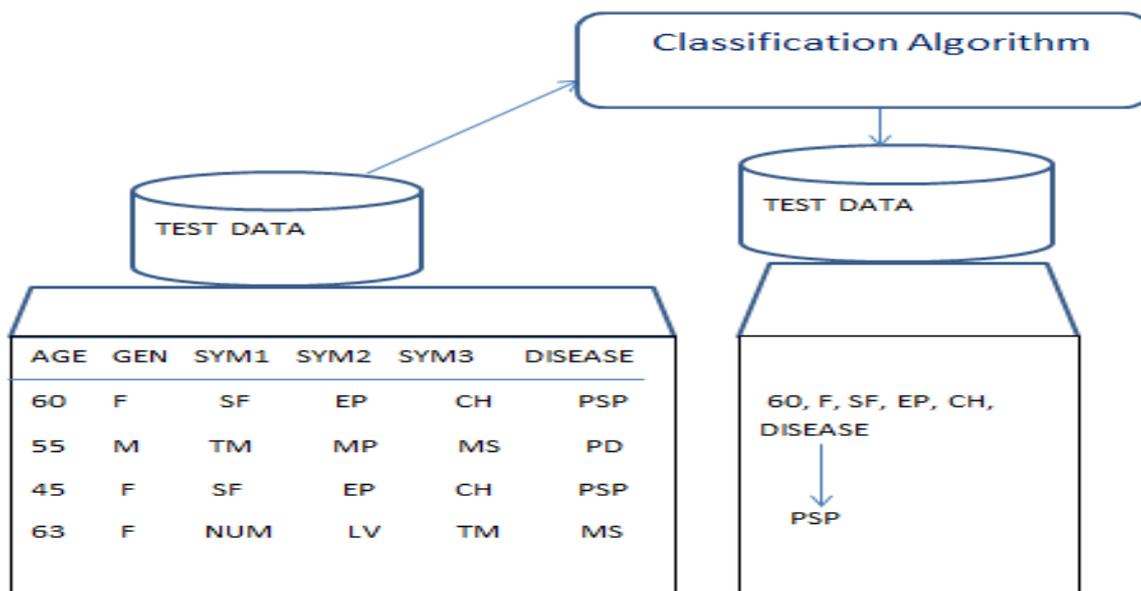


Fig 1.3 (ii) Using Classifier for Classification

III. LITERATURE SURVEY:

Neurological disorders assigns to a large number of medical conditions related to brain. This type of several neurological disorders is a major health problem in today’s life. This paper aims at analyzing the various data mining techniques used in the diagnosis of brain related problems. Abdullah.H.Wahbeh et.al [1] used several data mining classification models to show the accuracy measures by using different types of tools. G.Kesavaraj et.al [2] representing the various types of classification techniques and the steps to be followed in each algorithm. Sudhir M.Gorade et.al [3] described about the several classification models, advantages and disadvantages of different models. V.Ramamohan et.al [4] explained about the concept of Data Mining and the tools. GeetaYadav et.al [5] utilized the accompanying strategy, for example, Tree classifiers,

Logistic Regression (LR), and Support Vector Machine (SVM) with the assistance of k-fold cross validation technique to create predication model for Parkinson's disease identification. . The dataset has taken from UCI repository for this paper. When looking at all the calculations, the orderly precision for SVM-76% and Tree classifiers - 75%. Ismail Saritas et.al [6] utilized the following techniques, for example, Rotation Forest, Random Forest(RF), Multilayer Perceptron(MLP), Classification by means of Regression, Bagging, Java Repeated Incremental Pruning , Sequential Minimal Optimization, OneR, Naïve Bayes , Decision Tree table, J48, Bayes Net and Naive Bayes. In this study, the performances of 16 distinctive classification strategies were assessed as far as grouping exactness on Parkinson's disease dataset. When looking at the performance of calculations, it's been discovered that IB1 have higher accuracy (96.4103%). Tarigoppula V.S. Sriram et.al [7] used Bayes net, Naïve Bayes, logistics, KStar, ADTree, J48, Random Forest algorithm. Logistics represents 100%. The Author has given a comparison analysis between two datasets and given a prediction for PD by executing the data between PD and non-PD voice. Shianghau Wu et.al [8] suggests the algorithm such as logistic regression, decision tree, and neural net analysis used voice measurement variables as a diagnostic tool for Parkinson's disease. This study shows to use the data mining method to analysis the data about Parkinson's disease, to discriminate PD patients from healthy people and to create awareness about the early symptoms and early treatment for PD. Peyman Mohammadi et.al [9] compares eleven data mining algorithm for classification performance using Parkinson's Telemonitoring dataset. Chandrasekhar Azad et.al [10] used decision tree, LD3, decision stumps classification algorithms. The Parkinson's disease dataset is taken from the UCI repository. Decision tree has highest classification accuracy of 85.08%. The Author has analyzed speech articulation as one of the symptom for people affected with Parkinson's disease.

IV.VARIOUS CLASSIFICATION TECHNIQUES

The main objective of a Classification algorithm is to inflate the predictive accuracy acquired in the Classification techniques. There are several techniques used in the Classification are

- Decision Tree Induction
- Rule Based Method
- Naïve Bayesian Classification
- Neural Networks
- Support Vector Machine

4.1 DECISION TREE INDUCTION

A decision tree is mostly used classification techniques. A structure of decision tree includes leaf nodes, branches and root node. The top most nodes in the tree are the root node. An internal node is used to denote test on an attribute. The outcome of a test is denoted by branches. Every leaf node has a class label. For example, to predict patients suffering from progressive supranuclear palsy with its appropriate symptoms or not.

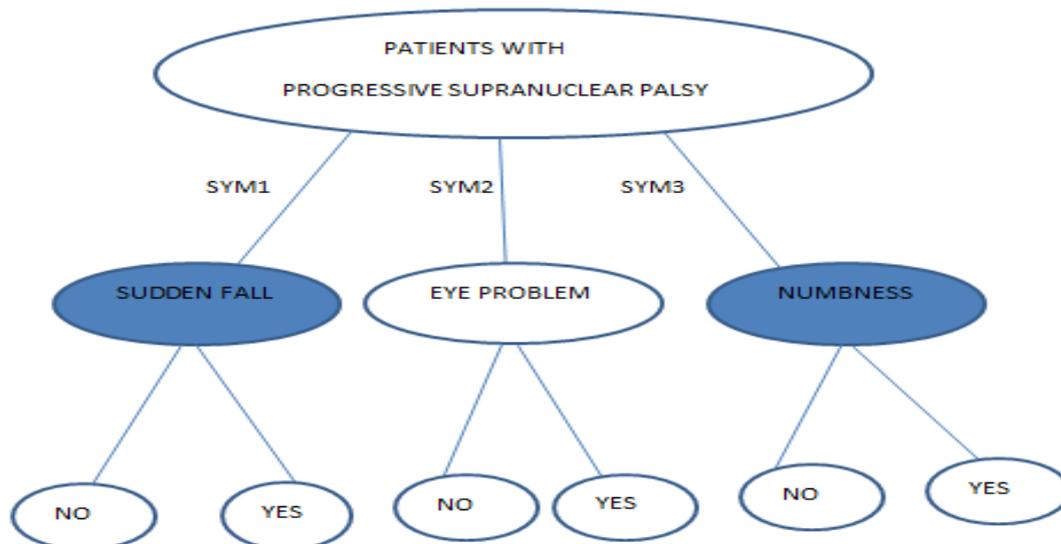


Fig 1.4 Decision tree of Patient with Progressive Supranuclear Palsy

4.2 RULE BASED METHOD:

Rule –based classifier makes use of a set of IF-THEN rules for Classification. We can specially review a rule in the following form

IF Condition THEN Conclusion

Let us consider a rule R1

**R1:IF SYMPTOM1=YES AND SYMPTOM2=YES
THEN DISEASE=PSP**

Condition statement with IF part is known as rule antecedent (or) Precondition. Condition statement with THEN part is called rule consequent. The antecedent part of the condition contains of one or more attribute tests & these tests are logically AND. The Consequent part consists of Class Prediction.

4.3 NAÏVE BAYESIAN CLASSIFICATION

The fundamentals of Naïve Bayes Classifier are Bayes theorem. A hypothesis is formed from the given set of classes. The way used to access in Naïve Bayes classifier is very simple. Based on the goal value, the attribute values are chosen which is independent. It is likely to classify the given instances with small amount of training data. Complex circumstances can be solved by using this method. This model is easy to construct for large datasets as it doesn't have a complex iterative parameter for estimation. Conditional Probability serves as a fundamental for Naïve Bayes algorithm which is used to find out the present and past frequency instances.

$$P(A/B) = P(B/A) * P(A) / P(B)$$

Where P (A) is the prior probability of A.It counts only the instances of A.

P (A/B) is the relative probability of A, given B.It is also called as posterior probability which means A is derived from B.

P(B/A) is the relative probability of B,given A

P(B) is the prior probability of B.

4.4 NEURAL NETWORKS:

Neural networks also referred as Artificial Neural Networks. It is expressed in terms of the biological neuron system; it consists of a number of separate units. The individual units are communicating to each other by sending signals. It is similar to the brain composed of many processing components. It is organized as directed graph which contains nodes and the edges connecting each node. The edges are the interconnections between each node. It receives 'm' inputs from 'n' nodes. Each edge connecting the node contains weight. The sum of the weight is calculated. The Threshold value is assigned to each neuron. If the weighted sum is greater than the threshold value it produces the output 1 otherwise 0.

4.5 SUPPORT VECTOR MACHINE:

Support Vector machine uses supervised machine learning model with associated learning algorithms that analyze data and recognize patterns. The support vector machine has been developed as robust tools for classification and regression in noisy, complex domains. The two key features of support vector machines are generalization theory, which leads to a principled way to choose a hypothesis and kernel functions, which introduce nonlinearity in the hypothesis space without explicitly requiring a non-linear algorithm.

V. ADVANTAGES AND DISADVANTAGES:

Method	Advantage	Disadvantage
Decision Tree Induction	It doesn't require any domain Knowledge. It is easy to understand. Simple and fast.	It takes more time to search. It occupies a large amount of memory to store tree.
Rule Based Method	Easy to analysis.	The rule-based methods are usually not the best specification in terms of prediction quality.
Naïve Bayesian Classification	Simple to implement. It predicts accurate results for classification and prediction problems.	If the amount of data is less, then the accuracy decreases. It requires a vast number of records to obtain good results.
Neural Networks	It is easy to use, with few parameters. Easy to implement.	If Neural network is large, then it requires high processing time. The process of learning is slow
Support Vector Machine	Works well with the data which is not linearly separable. High Accuracy.	It requires a vast space requirement in many cases. High Complexity

VI. COMPARATIVE STUDY ON THE PERFORMANCE ANALYSIS OF VARIOUS NEUROLOGICAL DISORDERS THROUGH ITS SYMPTOMS:

The performance analysis is compared with the various papers for the prognosis of various neurological disorders through its symptoms. From the below figure, the x-axis represents the various symptoms of neurological disorders and the y-axis depicts the percentage of similarity by means of symptoms.

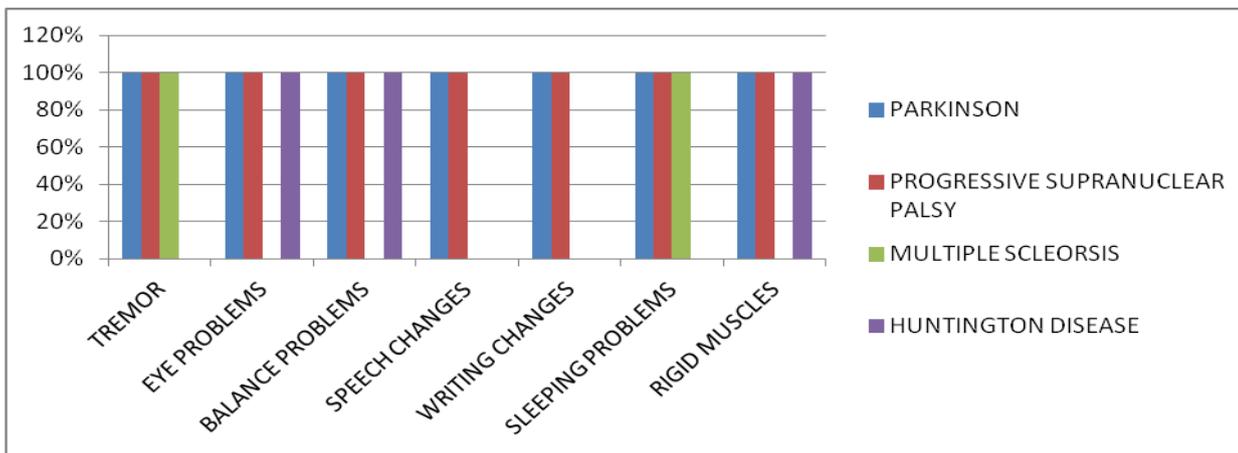


Fig 1.5 Performance analysis of various neurological disorders through its symptoms

CONCLUSION:

The main aim of this paper is to describe the study of various classification techniques in Data mining. This study has focused on various neurological disorders diagnosed through its symptoms. The future study aims to implement an efficient method that can be applied for the diagnosis of different disorders related to brain by using different classification techniques such as Decision tree induction, Rule-Base method, Naive Bayes, Support vector Machine.

REFERENCES

- [1]Abdullah H.Wahbeh,Qasem A.Al-Radaieh,Mohammed N.Al-Kabi,and Emad M.Al-Shawakfa “A Comparison study between Data Mining Tools over some Classification Methods”, International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Artificial Intelligence, 2011.
- [2]G.Kesavaraj, Dr.S.Sukumaran “A Study on Classification Techniques in Data Mining”,Institute of Electrical and Electronics Engineers, 2014.
- [3]Sudhir M.Gorade, Ankit Deo, PreeteshPurohit, “A study of some Data Mining Classification Techniques”, International Research Journal of Engineering and Technology (IRJET), Vol.4, Issue: 04, April (2017).
- [4]Y.Ramamohan, K.Vasantharao, C.Kalyana Chakravarti, A.S.K.Ratnam,”A study of Data Mining Tools in Knowledge discovery process”, International Journal of Soft Computing and Engineering (IJSCE), Volume.2, Issue: 03, July (2012).
- [5]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”, Indian Journal of Medical Sciences, Vol: 65, No: 6, June 2011.
- [6]Ismail Saritas, Murat Koklu, Kemal Tutuncu, “Performance of classification techniques on Parkinson’s disease”, Department of Electrical-Electronics and Computer Engineering, Selcuk University, Konya, Turkey.
- [7]Tarigoppula V.S. Sriram, M.Venkateswara Rao, G.V.SatyaNarayana, DSVGK Kaladhar ,“A Comparative And Prediction Analysis for the Diagnosis of Parkinson Disease using Data Mining Techniques on Voice Datasets” ,International Journal of Applied Engineering Research, Volume 11,Number 9(2016).
- [8]Shianghau Wu, Jiannjong Guo, “A Data Mining Analysis of the Parkinson’s Disease”, iBusiness, 2011, 3, 71-75.
- [9]Peyman Mohammadi, Abdoireza Hatamlou and Mohammed Masdari,“ A comparative study on remote tracking of Parkinson’s disease progression using data mining methods”, International Journal in Foundations of Computer Science & Technology (IJFCS), Vol.3,No.6, November 2013.
- [10]Chandrasekhar Azad, Sanjay Jai Vijay Kumar Jha, Design and Analysis of Data Mining Based Prediction Model for Parkinson’s disease”, International Journal of Computer science Engineering (IJCSE), Vol.3, No: 3, May 2014.