International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org Volume 3 Issue 4 || April 2014 || PP.23-28

A Prototype of Parkinson's and Primary Tumor Diseases Prediction Using Data Mining Techniques

Tawseef Ayoub Shaikh
Department Of Computer Science And Engineering , Guru Nanak Dev University Amritsar,Punjab India
shaikhtawseef@yahoo.com

ABSTRACT: *In the absence of medical diagnosis evidences, it is difficult for the experts to opine about the grade of disease with affirmation. Generally many tests are done that involve classification or clustering of large scale data. However , many tests can complicate the main diagnosis process and lead to the difficulty in obtaining end results, particularly in the case where many tests are performed. This could be solved by the aid of machine learning techniques. In this paper Psychiatric datasets of Parkinson's & Primary tumor Diseases are modeled and used it to predict their probability in the patients. The performance of Artificial Neural Network , Decision Trees Algorithm and Naive Bayes Algorithm on these medical data were measured. The results showed that Artificial Neural Network performed best with accuracy of 90.7692 % ,then Decision trees with accuracy of 80.5128 % and finally NaiveBayes with accuracy 69.2308 % in case of Parkinson's while as in primary tumor NavieBayes performs best with an accuracy of 49.1176%, then Artificial Neural Network with an accuracy of 42.0588% and lastly Decision trees with accuracy 32.3529%.*

KEYWORDS **:** *Artificial Neural Network,Decision Trees,NaiveBayes,Parkinson's Dataset,Primary Tumor*

I. INTRODUCTION

Ever since Parkinson Diseases(PD) was first described in 1817, scientists have pursued the causes and treatment of the disease. In the early 1960s, scientists identified the primary problem underlying the disease: the loss of brain cells that produce a chemical called dopamine, which helps to coordinate and control muscle activity. This discovery led to the first successful treatment for PD and suggested ways of devising new and even more effective therapies. Parkinson's research continues to be a very active field, with new and intriguing findings reported every day. Research suggests that PD affects at least 500,000 people in the United States, and some estimates are much higher. Society pays an enormous price for PD. The total cost to the nation is estimated to exceed $6 billion annually. The financial and public health impact of this disease is expected to increase as the population ages[1].

People in the later stages of Parkinson's disease cannot be sure what they will be able to do from one moment to the next. The unpredictably or major fluctuations in mobility creates anxiety and adds uncertainty to daily life. A simple movement, such as turning over in bed, may require major effort and concentration. Parkinson's disease is not fatal, although health complications may occur due to problems with immobility, falls, coexisting illness or drug therapy. Parkinson's disease (PD) is a common neurodegenerative disorder with a cumulative effect on patients, their families and the healthcare and social care systems. In Scotland, there are between 120 and 230 patients with PD per 100,000 people.1-3[2],While the population of Scotland remains stable, the age related incidence of PD means that the number of cases will increase by 25–30% over the next 25 years.4 There are a wide range of drug treatments for Parkinson's disease. However, it is not always clear which is the most appropriate treatment for the patient and whether the choice should be affected by age, clinical condition, or other factors. Brain tumors are also not rare. Thousands of people are diagnosed every year with tumors of the brain and the rest of the nervous system. The diagnosis and treatment of the brain tumor depends on the type of tumor ,tumor grade and where it started[3]. With the increased use of computers powered with automated tools, storage and retrieval of large volumes of medical data are being collected and are being made available to the medical research community who has been interested in developing prediction models for survivability[4][5]. As a result, new research avenues such as knowledge discovery in databases (KDD), which includes data mining techniques, has become a popular research tool for medical researchers who seek to identify and exploit patterns and relationships among large number of variables, and be able to predict the outcome of a disease using the historical cases stored within datasets [6]. Based on surveyed data, we used three different types of classification models: NaiveBayes,artificial neural network (ANN) and decision tree along with a 10-fold cross validation technique to compare the accuracy of these classification models. The amount of data being collected and stored in databases (both in medical and in other fields) has increased dramatically due to the advancements in software capabilities and hardware tools that enabled the automated data collection (along with the decreasing trend of hardware and software cost) in the last decade.As a result, traditional data analysis techniques have become inadequate for processing such volumes of data, and new techniques have been developed. A major area of development is called KDD. KDD encompasses variety of statistical analysis, pattern recognition and machine learning techniques. In essence, KDD is a formal process whereby the steps of understanding the domain, understanding the data, data preparation, gathering and formulating knowledge from pattern extraction, and ''post-processing of the knowledge'' are employed to exploit the knowledge from large amount of recorded data [6]. The step of gathering and formulating knowledge from data using pattern extraction methods is commonly referred to as data mining [7]. Applications of data mining have already been proven to provide benefits to many areas of medicine, including diagnosis.

II. THREE DATA MINING PREDICTION MODELS

In this paper three different types of classification models were used: Naye Bayes, artificial neural networks, decision trees. These models were selected for inclusions in this study due to their popularity in the recently published literature as well as their better than average performance in our preliminary comparative studies. What follows is a short description of these classification model types and their specific implementations for this research.

2.1 Naive Bayes

The Naïve Bayes [8] classifier provides a simple approach ,with clear semantics,representing and learning probablistic knowledge.It is termed naïve because it relies on two important simplfying assumes that the predictive attributes are conditionaly independent given the class,and it assumes that no hidden or latent attributes influnce the prediction process.

2.2 Artificial neural networks

Artificial neural networks (ANNs) are commonly known as biologically inspired, highly sophisticated analytical techniques, capable of modeling extremely complex non-linear functions. Formally defined, ANNs are analytic techniques modeled after the processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so-called learning from existing data[9].We used a popular ANN architecture called multi-layer perceptron (MLP) with back-propagation (a supervised learning algorithm). The MLP is known to be a powerful function approximator for prediction and classification problems. It is arguably the most commonly used and well-studied ANN architecture. Our experimental runs also proved the notion that for this type of classification problems MLP performs better than other ANN architectures such as radial basis function (RBF), recurrent neural network (RNN), and self-organizing map (SOM). In fact, Hornik et al. [10] empirically showed that given the right size and the structure, MLP is capable of learning arbitrarily complex nonlinear functions to arbitrary accuracy levels. The MLP is essentially the collection of nonlinear neurons (a.k.a. perceptrons) organized and connected to each other in a feedforward multi-layer structure.

2.3 Decision trees

Decision trees are powerful classification algorithms that are becoming increasingly more popular with the growth of data mining in the field of information systems. Popular decision tree algorithms include Quinlan's ID3, C4.5, C5 [11], and Breiman et al.'s CART . As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy. In doing so, they use mathematical algorithms (e.g.,information gain, Gini index, and Chi-squared test) to identify a variable and corresponding threshold for the variable that splits the input observation into two or more subgroups.This step is repeated at each leaf node until the complete tree is constructed.The objective of the splitting algorithm is to find a variable-threshold pair that maximizes the homogeneity (order) of the resulting two or more subgroups of samples. The most commonly used mathematical algorithm for splitting includes Entropy based information gain (used in ID3, C4.5,J48), Gini index (used in CART), and Chi-squared test (used in CHAID). Based on the favorable prediction results we have obtained from the preliminary runs, in this study we chose to use J48 algorithm as our decision tree method, which is an improved version of C4.5 and ID3 algorithms [11].

III. MATERIALS AND METHODS

The data used for this research were collected from UCI repository , which is open for research purposes.The Parkinson's disease contains 195 instances with 23 attributes (attributes f1 ….f22 are all numeric and the f23 is the binary Nominal attribute).The Primary Tumor contains 340 instances and 18 attributes as follows:

Table 1 is the variables included in the survey in Primary Tumor

Variables	Type
Age Nominal	String
Sex Nominal	String
Histologic type Nominal	String
Degree of Diff Nominal	String
Bone Nominal	String
Bone marrow Nominal	String
Lung Nominal	String
Pleura Nominal	String
Peritoneum Nominal	String
Liver Nominal	String
Brain Nominal	String
Skin Nominal	String
Neck Nominal	String
Supraclavicular Nominal	String
Axilar Nominal	String
Mediastinum Nominal	String
Abdominal Nominal	String
Class Nominal	String

3.1 Data Preprocessing

An important step in the data mining process is data preprocessing. One of the challenges that face the knowledge discovery process in medical database is poor data quality. For this reason much effort was laid in preparing data carefully to obtain accurate and correct results. First the the most related attributes to our mining task were chosen.

3.2 Data Mining Stages

The data mining stage was divided into three phases. At each phase all the algorithms were used to analyze the health datasets. The testing method adopted for this research was parentage split that train on a percentage of the dataset, cross validate on it and test on the remaining percentage. Sixty six percent (66%) of the health dataset which were randomly selected was used to train the dataset using all the classifiers. The validation was carried out using ten folds of the training sets. The models were now applied to unseen or new dataset which was made up of thirty four percent (34%) of randomly selected records of the datasets. Thereafter interesting patterns representing knowledge were identified.

IV. EXPERIMENTAL DESIGN

The Artificial Neural Networks , Decision Tree algorithms and Naie Bayes algorithms were used to analyse the health data. The ANN algorithms used were Multilayer Perceptron , the Decision Tree Algorithms used is J48 and NaiveBayes. The ANN models were trained with 500 epochs to minimize the root mean square and mean absolute error. Different numbers of hidden neurons were experimented with and the models with highest classification accuracy for the correctly classified instances were recorded. For the Decision Tree models, each class was trained with entropy of fit measure, the prior class probabilities parameter was set to equal, the stopping option for pruning was misclassification error, the minimum n per node was set to 5, the fraction of objects was 0.05, surrogates was 5, 10 fold cross-validation was used, and generated comprehensive results.

V. RESULTS

In this paper, the performance measures which are used for comparison are : accuracy, sensitivity and specificity.A distinguished confusion matrix is obtained to calculate the three measures. Confusion matrix is a matrix representation of the classification results. the upper left cell denotes the number of samples classifies as true while they were true (i.e., true positives), and lower right cell denotes the number of samples classified as false while they were actually false (i.e., true false). The other two cells (lower left cell and upper right cell) denote the number of samples misclassified. Specifically, the lower left cell denoting the number of samples classified as false while they actually were true (i.e., false negatives), and the upper right cell denoting the number of samples classified as true while they actually were false (i.e., false positives).Once the confusion matrixes were constructed, the accuracy, sensitivity and specificity are easily calculated as: sensitivity = TP/(TP

+ FN); specificity =TN/(TN + FP). Accuracy = (TP + TN)/(TP + FP + TN + FN); where TP, TN, FP and FN denotes true positives, true negatives, false positives and false negatives, respectively.

10-fold cross validation is used here to minimize the bias produced by random sampling of the training and test data samples. Extensive tests on numerous datasets, with different learning strategies, have shown that 10 is about the right number of folds to get the best estimate of error, and there is also some theoretical evidence that backs this up [12]. More Matrices include used are as :

- **Time:** This is referred to as the time required to complete training or modeling of a dataset. It is represented in seconds
- **Kappa Statistic:** A measure of the degree of nonrandom agreement between observers or measurements of the same categorical variable.
- **Mean Absolute Error:** Mean absolute error is the average of the difference between predicted and the actual value in all test cases; it is the average prediction error.
- **Mean Squared Error:** Mean-squared error is one of the most commonly used measures of success for numeric prediction. This value is computed by taking the average of the squared differences between each computed value and its corresponding correct value. The mean-squared error is simply the square root of the mean- squared-error. The mean-squared error gives the error value the same dimensionality as the actual and predicted values.
- **Root relative squared error:** Relative squared error is the total squared error made relative to what the error would have been if the prediction had been the average of the absolute value. As with the root mean-squared error, the square root of the relative squared error is taken to give it the same dimensions as the predicted value.
- **Relative Absolute Error:** Relative Absolute Error is the total absolute error made relative to what the error would have been if the prediction simply had been the average of the actual values.

Every model was evaluated based on the above measures discussed above . The results were achieved using average value of 10 fold cross-validation for each algorithm. We found that in case of Parkinson's case the MLP achieved classification accuracy of 90.7693% with a sensitivity of 85.4167% and a specificity of 92.5170% while as J48 achieved a classification accuracy of 80.5128% with a sensitivity of 58.33% and a specificity of 87.76% and NaiveBayes achieved a classification accuracy of 69.2308% with a sensitivity of 91.6667% and a specificity of 61.9048% .In the same way in case of Primary tumor MLP got an accuracy of 42.0588% , sensitivity of 42.1101% and specificity of 95.00%.Similarly Decision Tress got accuracy of 32.3529%, sensitivity of 42.4204% and specificity of 35.7206% and finally NavieBayes got an accuracy of 49.1176%, sensitivity of 49.1011% and specificity of 95.00%. Table 2 shows the complete set of results in a tabular format. The detailed prediction results of the validation datasets are presented in form of confusion matrixes.

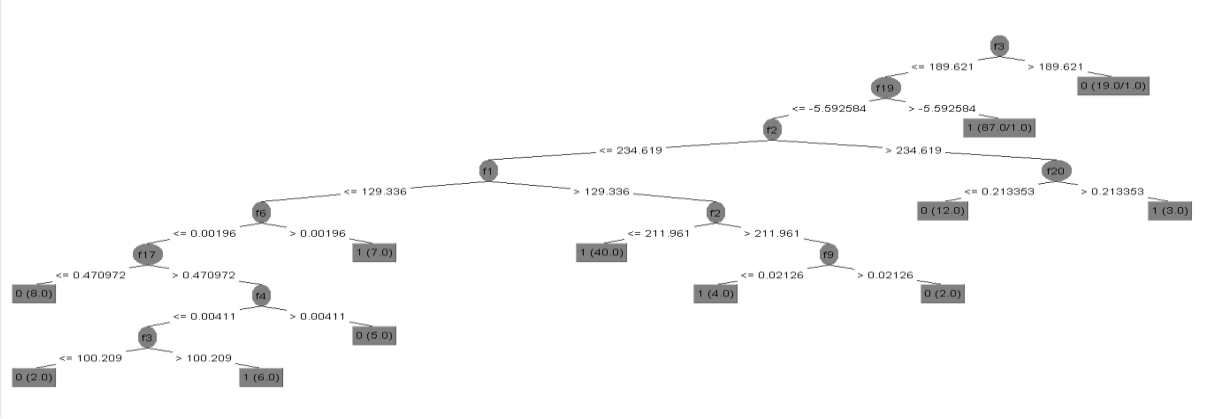


Figure 1:Visualization tree for the Parkinson's Prediction

Table 2: Performance of Artificial Neural Network, Decision Tree and NaiveBayes Algorithms

Performance Matrices	Artificial Neural Network (Mulilayerperceptron)		Decision Trees(J48)		Bayes(NaiveBayes)	
	Parkinson's Disease	Primary Tumor	Parkinson's Disease	Primary Tumor	Parkinson's Disease	Primary Tumor
Time	1.13	18.6	0.01	0.07	0	0.04
Kappa Statistics	0.7581	0.3426	0.4674	0.3353	0.3925	0.4146
MAE	0.1173	0.0569	0.2019	0.0623	0.3068	0.054
RMSE	0.2925	0.2041	0.4265	0.1931	0.5438	0.1759
RAE(%)	31.4756%	69.921%	54.1685%	76.5509%	82.3371%	66.4034%
RRSE(%)	67.8805%	101.4136%	98.9968%	95.9566%	126.2181%	87.9412%
Accuracy=(TP+TN)/ (TP+FP+TN+FN)	90.7693%	42.0588%	80.5128%	32.3529%	69.2308%	49.1176%
Sensitivity=TP/TP+FN	85.4167%	42.1101%	58.33%	42.4204%	91.6667%	49.1011%
Specificity=TN/TN+FP	92.5170%	95.00%	87.76%	35.7206%	61.9048%	95.00%

Fig. 2 Comparison of Correctly Classified Parameter of Datasets

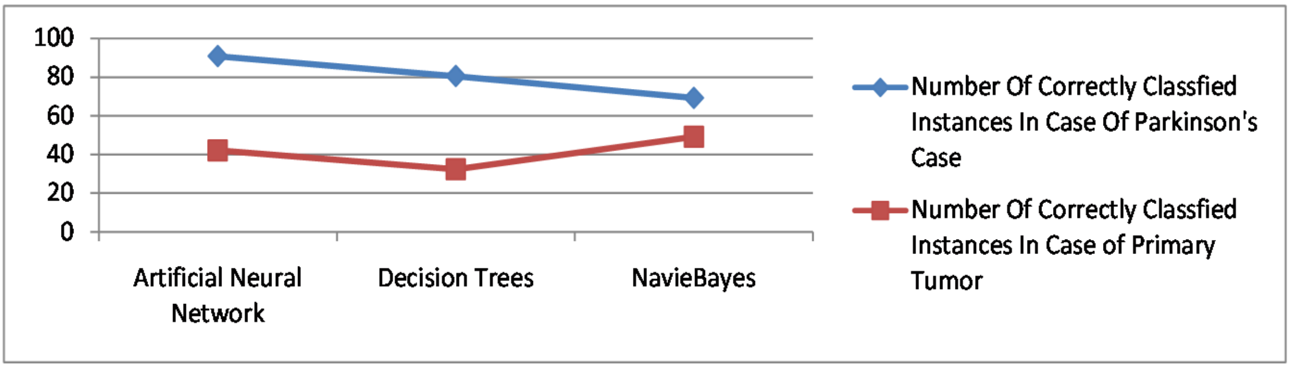
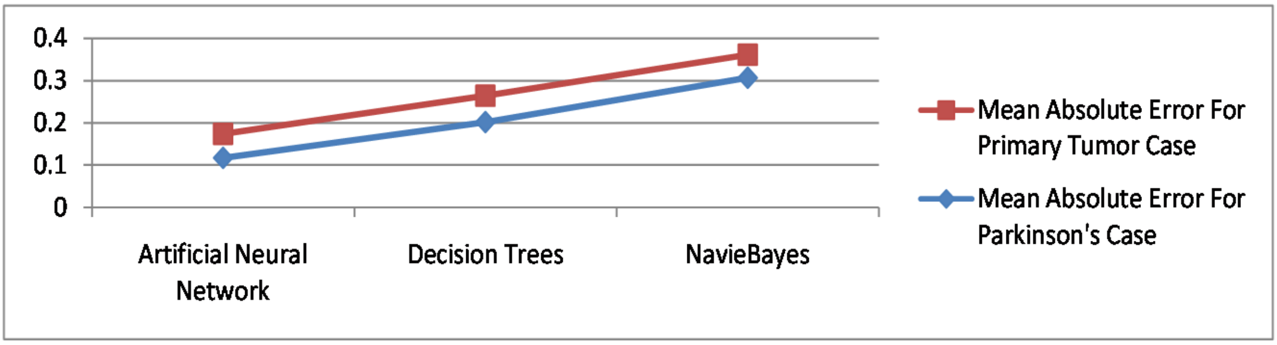


Fig. 3 Comparison of Mean Absolute Error Parameter



VI. CONCLUSION AND FUTURE WORK

On evaluating the different data mining algorithms on Psychiatric Datasets of Parkinson's and Primary tumor It came to conclusion that Artificial Neural Network (Multilayerperceptron) has got the highest classification accuracy in Parkinson's case while as NavieBayes perfoms best in case of Primary tumor.Moreover the NavieBayes classifier holds the missing values perfectely**.** Data mining techniques play an important role in finding patterns and extracting knowledge from large volume of data. It is very helpful to provide better patient care and effective diagnostic capabilities. Evidence Based Medicine (EBM) is a new direction in modern healthcare.EBM is as an important approach to make clinical decisions about the care of individual patients. This decision about patient is based on the best available Evidence. Its task is to prevent, diagnose and medicate diseases using medical evidence. It is all about providing best evidence, at right time in right manner to the clinician. External evidence-based knowledge cannot be applied directly to the patient without adjusting it to the patient’s health condition. If the rules generated by this system is approved by medical experts that can be used as evidence for further use. The conclusion can also be made that the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute. Also the hidden layer network plays an important role for detecting the relevant features.

REFERENCES

[1] Parkinson’s Disease , Challenges ,Progress And Promise ,November 2004 , National Institute Of Neurological Disorders and Stroke ,National Institutes Of Health

[2] Diagnosis And Pharmacological Management Of Parkinson’s Disease ,A National Clinical Guideline By Scottish Intercollegiate Guidelines Network

[3] 1995-2011,The Patient Education Institute ,Inc . www.X-Plain.com

[4] Cox DR. Analysis of survival data. London: Chapman & Hall; 1984.

[5] Ohno-Machado L. Modeling medical prognosis: survival analysis techniques. J Biomed Inform 2001;34:428—39

[6] Lavrac N. Selected techniques for data mining in medicine. Artif Intell Med 1999;16:3—23.

[7] Cios KJ, Moore GW. Uniqueness of medical datamining. Artif Intell Med 2002;26:1—24..

[8] G.H.John and P.Langley, “Estimating Continuous Distributions in Bayesian Classifiers,” Proceedings of the 11th Conference in University in Artificial Intelligance,San Francisco,1995,pp.338-345.

[9] Haykin S. Neural networks: a comprehensive foundation. New Jersey: Prentice Hall; 1998.

[10] Hornik K, Stinchcombe M, White H. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward network. Neural Netw 1990;3:359—66.

[11] Quinlan J. C4.5: programs for machine learning. San Mateo,CA: Morgan Kaufmann; 1993.

[12] Witten, I.H., FrankMichalewicz, E. Z. : Data Mining: Practical machine learning tools and tec hniques 2nd ed. Morgan Kaufmann, San Francisco (2005)