Analyzing the Suitability of Relevant Classification Techniques on Medical Data Set for Better Prediction

Mrs.R.Sarala,
AssistantProfessor ,
Computer Science Department,
Velammal College of Engineering and Technology, Madurai, India.

F.Dane Vijay Naveen
G.Balasubramani
Velammal College of Engineering and Technology, Madurai, India.

M.S.Bharathi Kannan
J.C.Krishna Moorthy
Velammal College of Engineering and Technology, Madurai, India.

Abstract

With the upcoming trend of online transaction in all sectors amount of newly created information increases every year. The huge amount of information has made it unfeasible for data analysts to achieve a deeper understanding of their data without at least some form of computer-aid. Data mining can be used to mechanize the process of knowledge discovery from databases. One of the techniques used in data mining is classification. Data mining classification algorithms can follow three special learning approaches: semi-supervised learning, supervised learning and unsupervised learning. In this paper we apply and analyze the commonly used classification algorithms on medical data set that helps to predict heart disease that accounts to be the primary cause of death worldwide. It is complex for medical practitioners to envisage the heart attack as it requires experience and knowledge. The health sector today contains concealed yet significant information for making decisions. Experiments conducted reveal that algorithm such as J48, SIMPLE CART and REPTREE provides more predictive accuracy than other algorithms.

Keywords *— Data mining, Classification techniques, Diseases.*

I. INTRODUCTION

Data mining is a knowledge discovery technique to examine data & summarize it Nearest neighbour, Artificial neural network, Support Vector machine etc.

Present research intends to predict probability of getting heart disease given patient data set [2].

Predictions and descriptions are principal goals of data mining, put into practice. Prediction in data mining involves attributes or variables in the data set to find unknown or future state values of other attributes [3]. Description gives emphasis on discovering patterns that explains the data to be interpreted by humans [4]. The intention of predictions in data mining in medical field is to help patients to find out trends to get improve in their health [1].

Modern science and engineering based on first-principle models to describe physical, biological and social system. Such an loom starts with a basic scientific model, such as Newton's laws of motion or Maxwell's equation in electromagnetism. Data mining is an iterative process within which advancement is defined by discovering through either automatic or manual methods. The goals of prediction are achieved by using data mining techniques. There is some primary task in data mining such as:

- Classification
- Regression
- Clustering.
- Summarization.
- Dependency
- Modelling.
- Change and Deviation Detection.

There are some of the experimental procedure adapted to data mining problems involves the following steps as in fig 1:

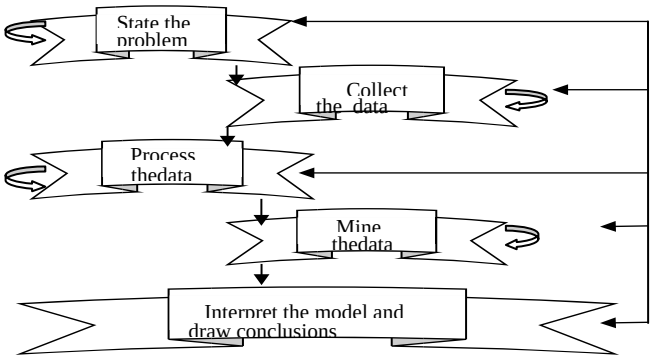


Fig 1: The data mining process

II. RELATED WORK

Quite a number of research work have been carried out in recent decades using data mining techniques on medical data. S. A. Pattekari et al., [5] has developed the web based intelligent system using naïve bayes algorithm to answer difficult queries for diagnosing heart disease and assist medical practitioners with clinical decisions. The prototype using naïve bayes and weighted associative classifier (WAC) to predict the probability of patients receiving heart attacks has been discussed in [6] N. A. Sundar et al.,

The work of M. Jabbar et al., [7] also has proposed a new approach for association rule mining found on sequence number and clustering transactional data set for heart disease predictions. The achievement of the proposed approach was implemented in C programming language and reduced main memory requisite by considering a small cluster at a time in order to be considered scalable and efficient .

S. U. Amin et al.,[8] have implemented a hybrid system that uses global optimization improvement of genetic algorithm for initialization of neural network weights. The prediction of the heart disease is based on risk factors such as age, family history, diabetes, hypertension, high cholesterol, smoking, alcohol intake and obesity.

P.Chandra et al., [9] created class association rules using feature subset selection to predict a model for heart disease. Association rule determine relations between attributes values and classification predicts the class in the patient dataset. into useful information [1]. There are different classification techniques in data mining such as rule based classifiers, Bayesian Network, Decision tree, Feature selection measures such as genetic search determines attributes which contribute towards the prediction of heart diseases.

S.B Patil et al., [10] use K-means clustering algorithm on a heart disease warehouse to haul out data relevant to heart disease, and relate MAFIA (Maximal Frequent Item set Algorithm) algorithm to analyze weight age of the frequent patterns significant to heart attack predictions.

III.CLASSIFICATION TECHNIQUES – SUITING THE SCENARIO

Classification consists of predicting certain result based on a given input. In order to predict the result, the algorithm processes a training set contain set of attributes and individual outcome, usually called prediction attribute. Data classification is the process of organizing data into categories for its most effective and efficient use. There is some algorithm in classification which helps to analyse work is J48, Naïve Bayes, Bayes Net, Reptree.

A.J48 DECISION TREE

Decision tree J48 is the implementation of algorithm ID3 (Iterative Dichotomiser 3) developed by the WEKA project team. J48 classifier is a easy C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most helpful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and outcome in classification for that tuple[1][3].

While building a tree, J48 ignores the omitted values i.e. the value for that item can be predicted based on what is known about the attribute values for the other account. The basic idea is to segregate the data into range based on the attribute values for that item that are found in the training sample. J48 allows classification via either decision trees or rules generated from them [13]14].

B.NAIVE-BAYES CLASSIFICATION ALGORITHM

The Bayesian Classification represents a supervised learning system as well as a statistical method for classification. Assumes concealed probabilistic replica and it allows us to confine uncertainty about the model in the principled way by determining probabilities of the outcomes. It can answer diagnostic and predictive problems. This Classification is named after Thomas Bayes (1 7021761), who proposed Bayes Theorem. Bayesian classification provides practical learning algorithms and prior knowledge and experimental data can be combined.

It is based on the Bayesian theorem. It is particularly suited when the dimensionality of the inputs is high. Parameter inference for naive Bayes models uses the process of maximum likelihood. In spite over-simplified assumptions, it often performs better in many complex real- world situations.

C.BAYESIAN NETWORK

A Bayesian network is a graphical model that encodes probabilistic dealings among variables of interest. When used in conjunction with statistical techniques, the graphical model has numerous advantages for data modelling. One, because the model encodes dependencies among all variables, it readily handles situations where some data entries are missing. Two, a Bayesian network can be used to find out causal

relationships, and hence can be used to achieve understanding about a difficulty domain and to predict the consequences of intervention. Three, because the model has mutually a causal and probabilistic semantics, it is a perfect representation for combining prior knowledge (which often comes in causal form) and data. Four, Bayesian statistical methods in conjunction with Bayesian networks present an efficient and principled loom for avoiding the over fitting of data.

D.REP Tree Classifier

Reduces Error Pruning (REP) Tree Classifier is a quick decision tree learning algorithm and is based on the principle of computing the information increase with entropy and minimizing the mistake arising from variance [15]. This algorithm is first recommended in [16]. REP Tree applies regression tree logic and generates multiple trees in altered iterations. Afterwards it picks best one from all spawned trees. This algorithm constructs the regression/decision tree using variance and information put on. Also, this algorithm prunes the tree using reduced-error pruning with back fitting technique. At the beginning of the model preparation, it sorts the values of numeric attributes one time. As in C4.5 Algorithm, this algorithm also deals the missing values by splitting the corresponding instances into pieces. [17].

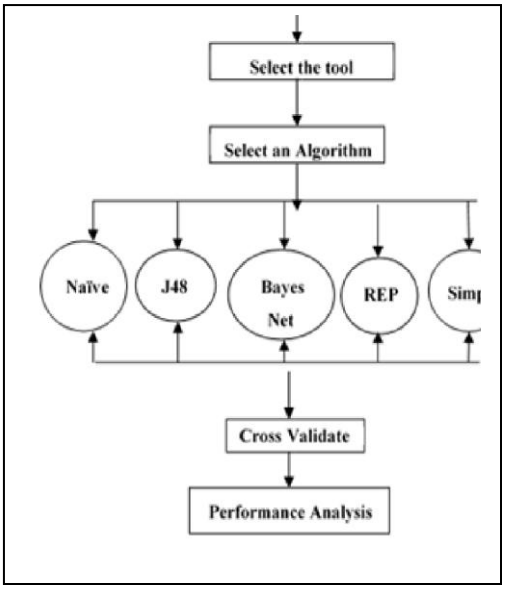
E.SIMPLE CART

Simple Cart is a classification technique that produces the binary decision tree. Since output is a binary tree, it produces only two children. Entropy is used to select the best splitting attribute. Simple Cart handles the misplaced data by ignoring that record. This algorithm is most excellent for the training data. Classification and regression trees (CART) decision tree is a learning technique, which gives the outcome as either classification or regression trees, depending on categorical or numeric data set.

IV. METHODOLOGY

To evaluate the performance of our approach, patient data set is loaded into the WEKA tool. Naïve Bayes, J48, REP tree, Simple Cart and BayesNet are selected. Data is then cross validated using performance classifier measure, the results and performance of each algorithm is then compared to each other. Figure 2 reveals the working of WEKA tool

The patient data set is compiled from data collected from medical practitioners in India. Only 11 attributes from the database are considered for the predictions required for the heart disease. The following attributes with nominal values are considered: Patient Identification Number (replaced with dummy values), Gender, Cardiogram, Age, Chest Pain, Blood Pressure Level, Heart Rate, Cholesterol, Smoking, Alcohol consumption and Blood Sugar Level.



V. EXPERIMENTAL RESULTS

The algorithms are applied on the data set using stratified 10-fold cross-validation in order to assess the performance of classification techniques for analyzing the patient data set. The confusion matrix of each algorithms are listed below:

Confusion Matrix of SIMPLE CART Algorithm

=== Confusion Matrix ===
a b <-- classified as
89 1 | a = TRUE
0 18 | b = FALSE

Confusion Matrix of REPTREE Algorithm

=== Confusion Matrix ===
a b <-- classified as
89 1 | a = TRUE
0 18 | b = FALSE

Confusion Matrix of NAÏVE BAYES Algorithm

```
=== Confusion Matrix ===
a   b   <-- classified as
88  2  | a = TRUE
1  17  | b = FALSE
```

Confusion Matrix of BAYESNET Algorithm

```
=== Confusion Matrix ===
a   b   <-- classified as
88  2  | a = TRUE
018   | b = FALSE
```

Confusion Matrix of J48 Algorithm

```
=== Confusion Matrix
=== a   b   <-- classified
as  89  1  | a = TRUE
0  18  | b = FALSE
```

Table 1 Description of the Data Set

Attributes	Description	Possible Values
PatientId	Dummy Identification of the patient	Patient Id
Gender	Sex of the patient	Male, Female
Age	Youth = 30-39, Young Adult =40-49 Adult =50-59 Old People =6069	Youth Young Adult Adult Old
Chest Pain Type	Stable Angina – Predictable Chest Pain Unstable Angina –Chest pain that signal impending heart attack Prinzmetal's Angina – have coronary artery disease	Stable angina Non-angina Unstable angina Prinzmetal's angina Asymptomatic
Heart Rate	No of heart beats per unit of time.	Low pulse rate High pulse rate
Cholesterol	Low-density lipoproteins (LDL) (Bad Cholesterol), Highdensity lipoproteins (HDL) (Good Cholesterol)	LDL HDL
Smoking	Coronary heart disease and stroke	Yes, No
Blood Sugar	If Blood Sugar level is > 120 mg/dl -Increase the risk	True, False
Blood Pressure	Normal- (systolic140 mmHg), High – (systolic > 160 mmHg)	Normal Pre hypertension High
Electro cardio graph(ECG)	Normal - ST_T wave Abnormality, Left Ventricular Hypertrophy (LVH) {Electrocardiographic results }	Normal Abnormal
Diet	Nourishment	Healthy, Unhealthy
Alcohol	Drug	True, False

The confusion matrix [18] obtained reveals parameters such as accuracy, sensitivity and specific measures etc. The matrix denotes samples classifications as true and false. The matrix validates the effectiveness of the model. The confusion matrix clearly classifies the accuracy of the mode. Evaluation of the confusion matrix shows that REPTREE, J48 and SIMPLE CART prove a prediction model of 89 cases with a hazard factor positive for heart attacks.

Table II and Table III show the classification accuracy based on diverse techniques applied, A close observation reveal that J48, SIMPLE CART, REPTREE ALGORITHM prove the best classification techniques, while Bayes Net algorithm out-performed the Naïve Bayes algorithm. Experiments conducted show that J48, SIMPLE CART and REPTREE provide more predictive accuracy than other algorithms.

Table II Predictive performance of the classifiers

Evaluation criteria	Classifiers				
	J48	***Reptree***	***Naïve Bayes***	***Bayes Net***	***Simple Cart***
Timing to build model (in secs)	0.0	0.0	0	0.02	0.1
Correctly Classified Instances	107	107	105	106	107
Incorrectly Classified Instances	1	1	3	2	1
Predictive Accuracy	99.074	99.074	97.222	98.148	99.074

Table III Comparison of estimates

Evaluation criteria	Classifiers				
	J48	*Reptree*	*Naïve Bayes*	*BayesNet*	*Simple CART*
Kappa statistic	0.9674	0.9674	0.9022	0.9362	0.9674
Mean absolute error	0.018	0.018	0.071	0.053	0.018
Root mean squared error	0.099	0.099	0.165	0.140	0.099
Relative absolute error	6.547	6.547	25.280	18.952	6.547

Figure 1 shows the graph based on evaluation criteria such as Timely build model in secs, Correctly Classified Instances, Incorrectly Classified Instances and Predictive accuracy.

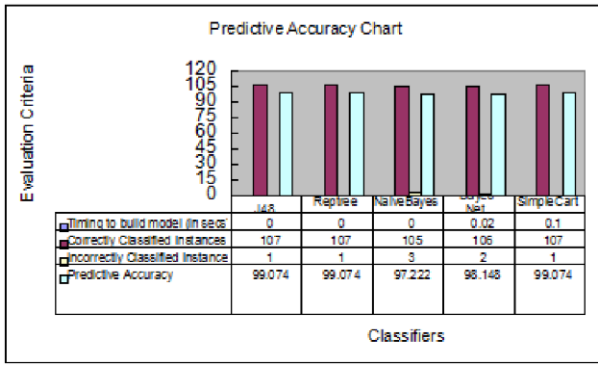


Figure 1. Predictive Accuracy Chart

Figure 1 shows the graph based on evaluation criteria such as Timely build model in secs, Correctly Classified Instances, Incorrectly Classified Instances and Predictive accuracy

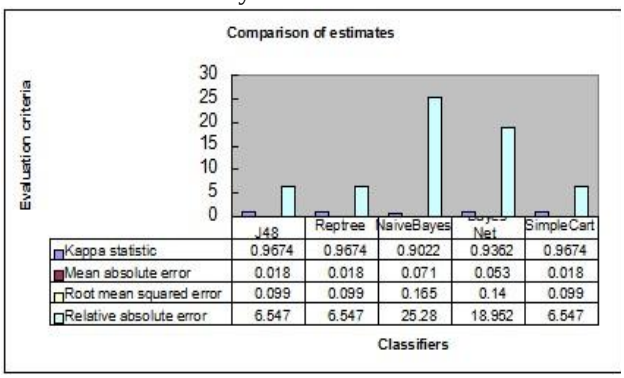


Figure 2. Comparison of estimates Chart

VI. DECISION TREE MODEL

The J48 algorithm produces an initial tree. Fig 3 shows the tree representation by using the J48 algorithm. The construction of the tree in J48 differs with the construction of the tree in several respects from REPTREE in Fig 4. These two trees show a graphical demonstration of the relations that is present in the dataset.

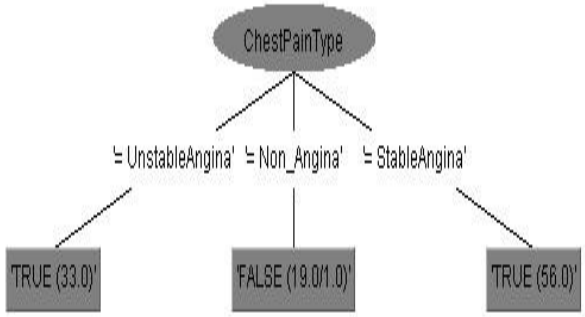


Figure 3. J48 Tree

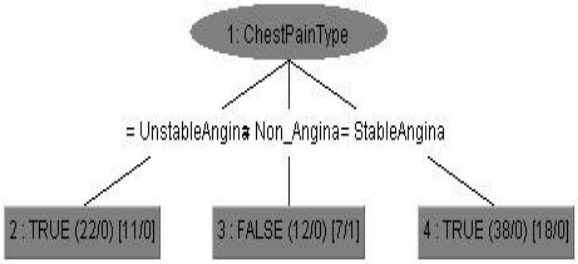


Figure 4. REP Tree

VII. CONCLUSION

This research work assumed an experiment on application of various data mining algorithms to predict the heart attacks and to compare the best method of prediction. The research results do not presents a remarkable difference in the prediction when using dissimilar classification algorithms in data mining. The experiment can serve as an significant tool for physicians to predict dangerous cases in practice and counsel accordingly. The representation given will be able to respond more difficult queries in forecasting the heart attack diseases. The predictive accuracy determined by REPTREE, J48 and BayesNet algorithms propose that parameters used are consistent indicator to predict the heart diseases. In the future, more parameters can be considered for better prediction.

REFERENCES

[1] Margaret H. Danham,S. Sridhar, " Data mining, Introductory and Advanced Topics", Person education , 1st ed., 2006.

[2] M. Jabbar, P. Chandra, and B. Deekshatulu, "CLUSTER BASED ASSOCIATION RULE MINING," Journal of Theoretical & Applied Information Technology, vol. 32, no. 2, pp. 196–201, 2011.

[3] R. Rao, "SURVEY ON PREDICTION OF HEART MORBIDITY USING DATA MINING TECHNIQUES," International Journal of Data Mining & Knowledge Management Process (IJDKP), vol. 1, no. 3, pp. 14–34, 2011.

[4] S. Vijiyarani and S. Sudha, "Disease Prediction in Data Mining Technique – A Survey," International Journal of Computer Applications & Information Technology, vol. II, no. I, pp. 17–21, 2013.

[5] S. A. Pattekari and A. Parveen, "PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES," International journal of Advanced Computer and Mathematical Sciences, vol. 3, no. 3, pp. 290–294, 2012.

[6] N. A. Sundar, P. P. Latha, and M. R. Chandra, "PERFORMANCE ANALYSIS OF CLASSIFICATION DATA

[7] MINING TECHNIQUES OVER HEART DISEASE DATA BASE," International Journal of Engineering Science & Advanced Technology, vol. 2, no. 3, pp. 470– 478, 2012.

[8] M. Jabbar, P. Chandra, and B. Deekshatulu, "CLUSTER BASED ASSOCIATION RULE MINING FOR," Journal of Theoretical & Applied Information Technology, vol. 32, no. 2, pp. 196–201, 2011.

[9] S. U. Amin, K. Agarwal, and R. Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors," in Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT 2013), 2013, no. Ict, pp. 1227– 1231.

[10] P. Chandra, M. .Jabbar, and B. .Deekshatulu, "Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection," in 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012, pp. 628– 634.

[11] S. B. Patil and Y. S. Kumaraswamy, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction," International Journal of Computer Science and Network Security (IJCSNS), vol. 9, no. 2, pp. 228–235, 2009.

[12] Wenke Lee, Salvatore J. Stolfo, Kui W. Mok, "A Data Mining Framework for Building Intrusion Detection Models"

[13] Aman Kumar Sharma, SuruchiSahni, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis", IJCSE, Vol. 3, No. 5, 2011, pp. 1890-1895. [13]http://www.jstor.org/discover/10.2307/40398417?uid=373825 6&uid=2134&uid=368470121&uid=2&u

[14] id=70&uid=3&uid=368470111&uid=60&sid=21101751936641 [14]http://stackoverflow.com/questions/10317885/decision-treevs-naive-bayes-classifier.

[15] Witten IH, and Frank E .2005 Data mining: practical machine learning tools and techniques 2nd ed. the United States of America, Morgan Kaufmann series in data management systems.

[16] Quinlan J (1987) Simplifying decision trees, International Journal of Man Machine Studies, 27(3), 221–234.

[17] S.K. Jayanthi and S.Sasikala . 2013.REPTree Classifier for indentifying Link Spam in Web Search Engines. IJSC, Volume 3, Issue 2, (Jan 2013), 498 –505.

[18] Y. Xing, J. Wang, Z. Zhao, and A. Gao, "Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease," in 2007 International Conference on Convergence Information Technology (ICCIT 2007), 2007, pp. 868–872.