Scientific Journal Impact Factor (SJIF) : 4.364
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

A STUDY SOME DATA MINING CLASSIFICATION TECHNIQUES

Kapil Panihar[1] and Vijay Kumar Verma[2]
[1] *M. Tech. CSE IV Semester, Lord Krishan College of Technology Indore M.P*
[2]*Assistance Professor CSE, Lord Krishan College of Technology Indore M.P*

Abstract- Data mining techniques are useful in medical science to analysis medical data and diseases contents. New and efficient data mining techniques are developed and used to discover various hidden and useful pattern form historical database. New Models are developed from these techniques and used in medical practitioners to take successful decision. Heart attack is one of most important task in medical science. The term Heart attack includes the various diseases that involve the heart attack problem. Predicting and identifying heart attack problem from different symptoms is an important problem. In this research paper present some traditional techniques the study of some common classification techniques like Decision tree induction, Bayesian Classification, Support Vector Machines (SVM) Rule-based classification, Neural Network as a Classifier The k-Nearest Neighbor as and Genetic Algorithms (GA) .

Keyword- Data Mining, Diagnosis, Heart Attack, Symptoms, Classification, Prediction

I. INTRODUCTION

Classification is a process which classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data. Classification predicts categorical class labels. Classification is divided into two-step.

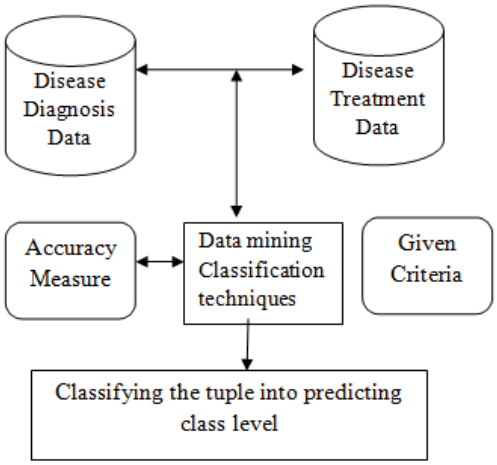

Figure 1 Data Mining classification process

1. Constructing a Model: This step describing a set of predetermined classes. Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute. The set of tuples used for model construction is training set. The model is represented as classification rules, decision trees, or mathematical formulae

2. Usages constructed Model: In this we use unknown tupl or unknown objects. Figure 1 show the system architecture for classification process in which sample data are used to construct model and then tested this model for accuracy to the a given tuple predicted the class level.

II. CLASSIFICATION PARAMETERS

There are some parameters which are used to evaluate classification methods. These parameters are

1. Correctness: - Correctness defines accuracy of the classifier in term of predicting the class label, Guessing value of predicted attributes. Accuracy can be estimated using one or more test sets that are independent of the training set.

2. Prediction time : This include the required time to construct the model (training time) and time to use the model (classification/prediction time). In other word this refers to the computational costs.

3. Strength: This is the ability of the classifier or predictor to make correct predictions given noisy data or data with missing values.

4. Scalability: Efficiency in term of database size.

5. Interpretability:-Understanding and insight provided by the model. Interpretability is subjective and therefore more difficult to assess.

6. Other measures: Includes goodness of rules, such as decision tree size or compactness of classification rules.

III. VARIOUS CLASSIFIERS

3.1 Classification by Decision Tree

Decision tree is A flow-chart-like tree structure Leaf nodes represent class labels or class distribution. Decision tree is a classifier in which each non-terminal node represents either a test or decision for the given data item. Which branch to be select next is depends upon the outcome of the test. To classify a given data item, need to from start at the root node and follow the assertions down until we reach a terminal node or leaf node.

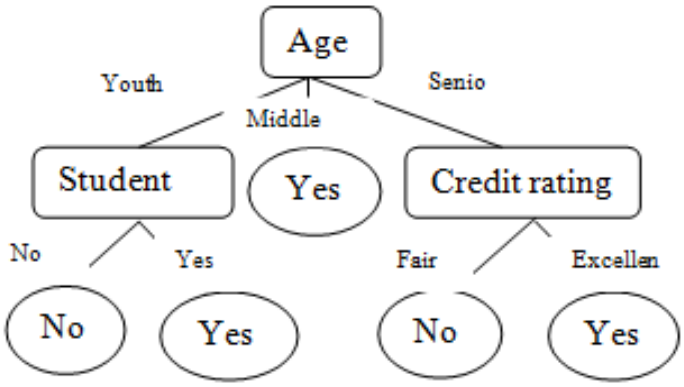

Figure2 simple decision tree classification

Decision is made when a terminal node is approached. Decision trees use recursive data partitioning. The important things in decision tree are attribute selection measure. There is important parameter used for attribute selection. The attribute with highest information gain is used to be selected as a root.

3.2 Classification by Bayesian classifiers

The Naive Bayesian classifier, or simple Bayesian classifier are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem. The Naive Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Naive Bayes Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with

predictive capabilities. It provides new ways of exploring and understanding data. Figure 3 shows the working of Naive Bayesian classifiers.

3.3 Classification by Neural Network

Neural network approach has been widely adopted as classifiers. The neural network provides several advantages, like arbitrary decision its nonparametric nature, boundary capability, easy adaptation to different types of data. Neural networks are of particular interest because they offer a means of efficiently modeling large and complex problems in which there may be hundreds of predictor variables that have many interactions. Neural nets may used in classification problems where the output is a categorical variable. Neural nets consist of three layers such as input layer, hidden layer and output layer. The nodes in the input layer linked with a number of nodes in the hidden layer. Each input node joined to each node in the hidden layer. The nodes in the hidden layer may connect to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response variables. There is numerous advantages of ANN some of these include

1) High Accuracy. 2) Independent from prior assumptions about the distribution of the data.
3) Noise tolerance. 4) ANN can be implemented in parallel hardware.

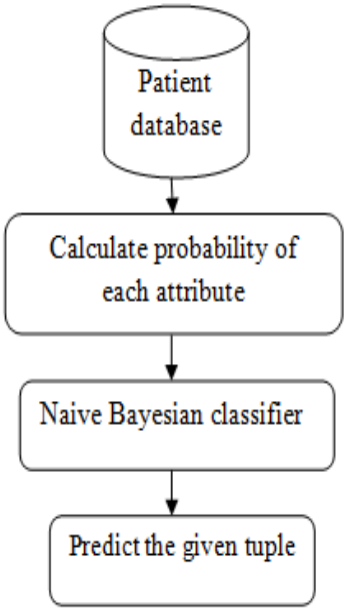

Figure 3 simple modelfor Naive Bayesian classifiers

3.4 Classification by IF-THEN Rules

Rule-based classifier uses a set of IF-THEN rules for classification.
An IF-THEN rule is an expression of the form
IF *condition* THEN *conclusion*.
An example is rule *R*1,
R1: IF *age* = *youth* AND *student* = *yes* THEN *buys computer* = *yes*.
The "IF" part of a rule is known as the rule antecedent or precondition. The "THEN" part is the rule consequent. In the rule antecedent, the condition consists of one or more *attribute tests* (such as *age* = *youth*, and *student* = *yes*)
that are logically ANDed. The rule's consequent contains a class prediction (in this case, we are predicting whether a customer will buy a computer). *R*1 can also be written as
R1: (*age* = *youth*) ^ (*student* = *yes*)) ⟶ (*buys computer* = *yes*).

If the condition (that is, all of the attribute tests) in a rule antecedent holds true for a given tuple, we say that the rule antecedent is satisfied and that the rule covers the tuple.

IV. LITERATURE REVIEW

In 2010 O.P.V Yas and Sunita Soni proposed "Using Associative Classifiers for Predictive Analysis in Health Care Data Mining". They describe that analysis technique to discover a small set of rule in the database to forms an accurate classifier Association rule mining is important. They introduce the combined approach that integrates association rule mining and classification rule mining. This is new classification approach is implemented by focusing on mining a special subset of association rules called classification association rule, then classification is being performed using rules. The associative classifiers are especially fit to applications were the model may assist domain experts in their decisions There are many associative classification approaches that have been proposed recently such as CBA, CMAR, CPAR and MCAR and MMAC.

In 2011 Mai Shouman, Tim Turner, Rob Stocker proposed "Using Decision Tree for Diagnosing Heart Disease Patients ". They show that Decision Tree is one of the successful data mining techniques used in the diagnosis of heart disease. Yet its accuracy is not perfect. The proposed work systematically tested combinations of discretization, decision tree type and voting to identify a more robust, more accurate method. They investigate a range of techniques to different types of Decision Trees seeking better performance in heart disease diagnosis and proposed a model that outperforms.

In 2012 M.Akhil jabbar , Dr.Priti Chandrab , Dr.B.L Deekshatulu Proposed Heart Disease Prediction System using Associative Classification and Genetic Algorithm". They proposed efficient associative classification algorithm using genetic approach for heart disease prediction. The main advantage of genetic algorithm is the discovery of high level prediction rules is that the discovered rules are highly comprehensible, having high predictive accuracy and of high interestingness values. The proposed method helps in the best prediction of heart disease which even helps doctors in their diagnosis decisions

In 2013 M. Akhil Jabbar, B.L Deekshatulu and Priti Chandra proposed "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection". They have proposed a new feature selection method using ANN for heart disease classification. For rank the attributes which contribute more towards classification of heart disease they applied different feature selection methods, and indirectly reduce the no. of diagnosis tests to be taken by a patient. The proposed method eliminates useless and distortive data. The proposed method will contribute reliable and faster automatic heart disease diagnosis system, where easy diagnosis of heart disease will saves lives.

In 2014 N. S. Nithya and K. Duraiswamy proposed "Gain ratio based fuzzy weighted association rule mining classifier for medical diagnostic interface". Earlier model based on information gain and fuzzy association rule mining algorithm for extracting both association rules and membership functions are not feasible.

When taking a large number of distinct values. So they modify gain ratio based fuzzy weighted association rule mining and improve the classifier accuracy.

V. ADVANTAGES AND LIMITATIONS

5.1 Decision Tree

Decision Tree has following advantage

No requirements of domain knowledge in the construction of decision tree. High dimension data can easily process.

Decision Tree has following limitations

It generates categorical output. Classifier is depend upon the type of dataset It is restricted to one output attribute

5.2 Bayesian Classification

Bayesian Classification has following advantages
Naive Bayesian classifiers make computational process easy.
Naive Bayesian classifiers provides better speed and accuracy for huge datasets

Bayesian Classification has following limitation advantages
Bayesian classifiers are a probability based methods. It does not give accurate results
Class conditional independence, therefore loss of accuracy

VI. CONCLUSION

There are various classification techniques that can be used for the identification and prevention of heart disease. The performance of classification techniques depends on the type of dataset that we have taken for doing experiment. Classification techniques provide benefit to all the people such as doctor, healthcare insurers, patients and organizations who are engaged in healthcare industry. Decision tree, Bays Naive classification, Support Vector Machine, Rule based classification, Neural Network as a classifier etc. These techniques are compared on basis of Sensitivity, Specificity, Accuracy, Error Rate, True Positive Rate and False Positive Rate. The objective of each technique is to predict more accurately the presence of heart disease with reduced number of attributes.

REFERENCE

[1] Divya Tomar and Sonali Agarwal " A survey on Data Mining approaches for Healthcare" International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013).
[2] Bangaru Veera Balaji and Vedula Venkateswara Rao "Improved Classification Based Association Rule Mining" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 5, May 2013.
[3] V. Krishnaiah, Dr. G. Narsimha and Dr. N. Subhash Chandra" Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1) 2013.
[4] Shamsher Bahadur Patel, Pramod Kumar Yadav and Dr. D. P.Shukla " Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques" IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS) e-ISSN: 2319-2380, p-ISSN: 2319-2372. Volume 4, Issue 2 (Jul. - Aug. 2013),
[5] N. Suneetha ,Ch.V.M.K.Hari and Sunil Kumar " Modified Gini Index Classification: A Case Study Of Heart Disease Dataset" (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 06, 2010, 1959-1965
[6] Jyoti Soni, Uzma Ansari, Dipesh Sharma and Sunita Soni "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers**"** Jyoti Soni et al. / International Journal on Computer Science and Engineering (IJCSE) ISSN : 0975-3397 Vol. 3 No. 6 June 2011
[7] Sunita Soni and O.P.Vyas Using Associative Classifiers for Predictive Analysis in Health Care Data Mining "International Journal of Computer Applications (0975 – 8887) Volume 4 – No.5, July 2010
[8] Mai Shouman, Tim Turner, Rob Stocker "Using Decision Tree for Diagnosing Heart Disease Patients" Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia
[9] M. Akhil jabbar, Dr B.L Deekshatulu and Dr Priti Chandra " Heart Disease Classification Using Nearest Neighbor Classifier With Feature Subset Selection" Computer Science and Telecommunications 2013|No.3(39)
[10] Sunita Soni and O.P.Vyas "Fuzzy Weighted Associative Classifier: A Predictive Technique For Health Care Data Mining" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.1, February 2012
[11] Chaitrali S. Dangare and Sulabha S. Apte, PhD. "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012
[12] M. Akhil Jabbar, B.L Deekshatulu & Priti Chandra "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection" Global Journal of Computer Science and Technology Neural & Artificial Intelligence Volume 13 Issue 3 Version 1.0 Year 2013 Type: Double Blind Peer Reviewed International Research Journal
[13] Publisher: Global Journals Inc. (USA) Online ISSN: 0975-4172 & Print ISSN: 0975-4350
[14] N S Nithya and K Duraiswamy Gain ratio based fuzzy weighted association rule mining classifier for medical diagnostic interface Vol. 39, Part 1, February 2014, pp. 39–52. Indian Academy of Sciences

[15] M.Akhil jabbar, Dr.Priti Chandrab , Dr.B.L Deekshatuluc "Heart Disease Prediction System using Associative Classification and Genetic Algorithm" International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012

[16] A. Anushya and A. Pethalakshmi "A Comparative Study of Fuzzy Classifiers With Genetic On Heart Data" International Conference on Advancement in Engineering Studies & Technology, ISBN : 978-93-81693-72-8, 15th JULY, 2012, Puducherry