Data Mining Classification Techniques: A Recent Survey

Abhale Babasaheb Annasaheb
M. Tech. CSE IV Sem., Lord Krishna College of Technology Indore M.P. India.

Vijay Kumar Verma
Asst. Prof. CSE Dept., Lord Krishna College of Technology Indore M.P. India.

Abstract – Prediction of heart attack is an important task in medical science. There are several factors are responsible for heart attack problem. Prediction of heart attack problem from different responsible factor is a difficult task. Data mining classification algorithm plays a vital role in several real life applications. In this research we paper present the study of various classification techniques including Decision Tree Induction, Bayesian Classification, Support Vector Machines, Rule-based classification, Neural Network Classifier and K-Nearest Neighbor Classifier. There are three important which are always considering for classifiers accuracy, Speed and Scalability.

Index Terms – Prediction, Classification Diagnosis, Heart Attack, Symptoms.

1. INTRODUCTION

In data mining classification is a process which classifies a given data set based on the training set and values of class labels. Constructing fast and accurate classifiers for large data sets is an important task in data mining and knowledge discovery. Classification is divided into two-step.

Constructing a Model:

In this step a model is constructed on the basis of set of predetermined classes. Each tuple/sample is assumed to a predefined class, as determined by the class label attribute .The set of tuples used for model construction is training set.

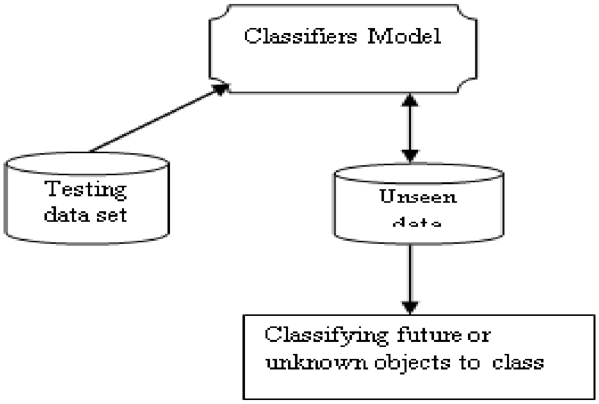

Figure 1 Constructing classifier

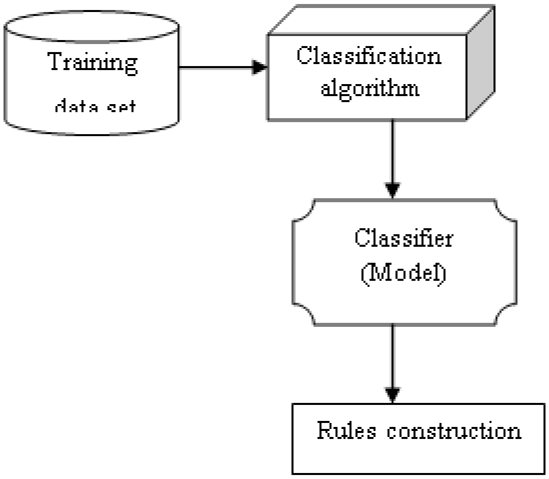

Figure 2 Used classifier model to classifying unknown object

Usages constructed Model:

In he seconds step the constructed model is now apply on the unknown set data to classify corrected. Accuracy rate is the percentage of test set samples that are correctly classified by the model.

2. CLASSIFIERS EVALUATION PARAMETERS

Some important parameters decide which classifiers are suitable for a particular data set. These parameters are

Accuracy:-

This include accuracy of the classifier in term of predicting the class label . Accuracy can be estimated using one or more test sets. Accuracy is the percentage that a classifier classifies the tuple correctly.

Speed:

How much time is required to construct the model? This also includes the time to use by the model to classify then number of tuple (prediction time). In other word this refers to the computational costs.

Robustness:

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

Scalability:

Number of tuple are independent form the classifiers. Efficiency is calculated in term of database size.

Interpretability:-

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

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

3. BASIC CLASSIFICATION TECHNIQUES

3.1 Decision Tree Classifier

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.

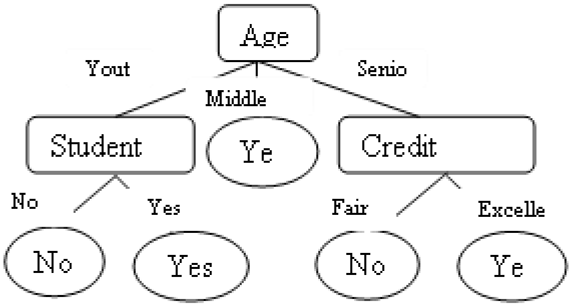

Figure 3 simple decision tree classification

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 Naive 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 1.3 shows the working of Naive Bayesian classifiers

3.3 Neural Network as a Classifier

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 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.

3.4 Using IF-THEN Rules as Classifier

A 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.

3.5 Support Vector Machines (SVMs)

Support Vector Machine (SVM) is primarily a classier method that performs classification tasks by constructing hyper planes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables. Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships.

4. LITERATURE REVIEW

In 2012 Qasem A. Al-Radaideh & Eman Nagi "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance". They proposed a classification model by using CRISP-DM data mining methodology. In proposed model decision tree was the main data mining tool used. To validate the proposed model they used real data from several companies. The model is intended to be used for predicting new applicants' performance [1].

In 2012 M. Akhil jabbar & Dr.Priti Chandrab 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 [2]

In 2012 K. Rajesh, V. Sangeetha "Application of Data Mining Methods and Techniques for Diabetes Diagnosis". The proposed project helps for mining the relationship in Diabetes data for efficient classification. The data mining methods and techniques will be explored to identify the suitable methods and techniques for efficient classification of Diabetes dataset and in mining useful patterns [3].

In 2013 M. Akhil Jabbar, B.L Deekshatulu & Priti Chandra proposed "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection". They 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 [4].

In 2013 V. Krishnaiah Dr. G. Narsimha & Dr. N. Subhash Chandra "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques". The proposed approach is an extension of naïve Bayes to imprecise probabilities that aims at delivering robust classifications also when dealing with small or incomplete data sets. Diagnosis of Lung Cancer Disease can answer complex "what if" queries which traditional decision support systems cannot. The proposed model is used for early detection and correct diagnosis of the disease which will help the doctor in saving the life of the patient [5].

In 2013 Divya Tomar & Sonali Agarwal "A survey on Data Mining approaches for Healthcare". They explore the utility of various Data Mining techniques such as classification, clustering, association, regression in health domain. They present a brief introduction of these techniques and their advantages and disadvantages. They also highlights applications, challenges and future issues of Data Mining in healthcare. Recommendation regarding the suitable choice of available Data Mining technique is also discussed [6].

In 2014 N. S. Nithya and K. Duraiswamy proposed "Gain ratio based fuzzy weighted association rule mining classifier for medical diagnostic interface". They showed that earlier model based on information gain and fuzzy association rule mining algorithm for extracting both association rules and membership functions are not feasible. They used large number of distinct values. They modify gain ratio based fuzzy weighted association rule mining and improve the classifier accuracy[7].

In 2015 S. Olalekan Akinola, O. Jephthar Oyabugbe proposed "Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study". They proposed study was designed to determine how data mining classification algorithm perform with increase in input data sizes. They used three data mining classification algorithms Decision Tree, Multi-Layer Perceptron (MLP) Neural Network and Naïve Bayes were subjected to varying simulated data sizes. The time taken by the algorithms for trainings and accuracies of their classifications were analyzed for the different data sizes. [8]

In 2015 Jaimini Majali, Rishikesh Niranjan & Vinamra Phatak proposed "Data Mining Techniques for Diagnosis and Prognosis of Cancer". They used data mining techniques for diagnosis and prognosis of cancer. They presented a system for diagnosis and prognosis of cancer using Classification and Association approach in Data Mining. They used FP algorithm in Association Rule Mining to conclude the patterns frequently found in benign and malignant patients [9].

In 2016 Nikhil N. Salvithal & R.B. Kulkarni proposed "Appraisal Management System using Data mining Classification Technique". The proposed assorted classifier algorithms applied on Talent dataset to spot the talent set so as to judge the performance of the individual. Finally counting on accuracy one best suited classifier is chosen this method has been used to construct classification rules to predict the potential talent that for promotion or not[10].

In 2016 Tanvi Sharma & Anand Sharma proposed "Performance Analysis of Data Mining Classification Techniques on Public Health Care". The proposed study focused on the application of various data mining classification techniques using different machine learning tools such as WEKA and Rapid miner over the public healthcare dataset for analyzing the health care system. The percentage of accuracy of every applied data mining classification technique is used as a standard for performance measure. The best technique for particular data set is chosen based on highest accuracy [11].

5. CONCLUSION

Decision tree classifiers, Bayesian classifiers, classification by back propagation, support vector machines, eager learners in that they use training tuples to construct a generalization model. In contrasts nearest-neighbor classifiers and case-based reasoning classifiers are lazy learners which store all of the training tuples in pattern space and wait until presented with a test tuple before performing generalization. Hence, lazy learners require efficient indexing techniques. 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.

REFERENCES

[1] Qasem A. Al-Radaideh & Eman Al Nagi "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance " (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, 2012

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

[3] K. Rajesh, V. Sangeetha "Application of Data Mining Methods and Techniques for Diabetes Diagnosis" International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012ISSN: 2277-3754

[4] 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.

[5] V. Krishnaiah, Dr. G. Narsimha & 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.

[6] Divya Tomar & Sonali Agarwal " A survey on Data Mining approaches for Healthcare" International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013) 241-266.

[7] N S Nithya & K Duraiswamy "Gain ratio based fuzzy weighted association rule mining classifier for medical diagnostic interface" Sadhana Vol. 39, Part 1, February 2014, Indian Academy of Sciences

[8] S. Olalekan Akinola & O. Jephthar Oyabugbe "Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study" Journal of Software Engineering and Applications, 2015, Published Online September 2015

[9] Jaimini Majali, Rishikesh & Niranjan, Vinamra Phatak "Data Mining Techniques For Diagnosis And Prognosis Of Cancer" International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 3, March 2015

[10] Nikhil N. Salvithal & R.B. Kulkarni, "Appraisal Management System using Data mining Classification Technique" International Journal of Computer Applications (0975 – 8887) Volume 135 – No.12, February 2016

[11] Tanvi Sharma, Anand Sharma & Vibhakar Mansotra "Performance Analysis of Data Mining Classification Techniques on Public Health Care Data" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 6, June 2016.