Volume 2, Issue 3, March 2017 ISSN (online): 2456-0006
International Journal of Science Technology Management and Research
Available online at: www.ijstmr.com

A New Approach to Classify Tuples More Accurately

Kapil Panihar
M.Tech ,IV Sem.,
Computer Science & Engg. Dept.
LKCT College, INDORE
M.P. India

Vijay Kumar Verma
Assistant Professor
Computer Science & Engg. Dept.
LKCT College, Indore
M.P. India

Abstract:-Data mining methods are used to analyse the medical data contents. Superior data mining techniques are developed and used to discover hidden pattern form historical data. New Models are developed from these techniques will be useful for medical practitioners to take successful decision. Diagnosis of heart disease is a significant task in medical science. The term Heart disease includes the various diseases that involve the heart disease problem. The exposure of heart disease problem from different symptoms is an important issue for predicting heart disease problem. There are various classification techniques like Classification by decision tree induction, Bayesian Classification, Rule-based classification, Classification by back propagation, Support Vector Machines (SVM) Neural Network as a Classifier The k-Nearest Neighbour Algorithm and Classification using Genetic Algorithms (GA) are used for heart disease problem. In proposed work we have taken 10 attribute which are responsible for the heart disease problem. We assign a weight to every attribute suggested by the physician the attribute includes Age, Blood Pressure (BP) , Cholesterol , Fasting Blood Sugar (FBS) Resting ECG, Thalach Value Beats/Minute, Old Peak, Slope, Thal Value. We count the number of tuples for each attribute based on given threshold value. We start pairing by using joining operation and every time compare with the thresholds value. The attribute satisfy the condition will consider for higher level otherwise discards them.

Keywords: Heart disease, classification, symptoms, prediction, tuples

I. INTRODUCTION

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. Classification divides data samples into target classes. The classification technique predicts the target class for each data points. For example, patient can be classified as “high risk” or “low risk” patient on the basis of their disease pattern using data classification approach. It is a supervised learning approach having known class categories. Binary and multilevel are the two methods of classification. In binary classification, only two possible classes such as, “high” or “low” risk patient may be considered while the multiclass approach has more than two targets for example, “high”, “medium” and “low” risk patient. Data set is partitioned as training and testing dataset. Using training dataset we trained the classifier. Correctness of the classifier could be tested using test dataset. Classification is one of the most widely used methods of Data Mining in Healthcare organization. Different classification method such as decision tree, SVM and ensemble approach is used for analyzing data. Classification techniques are also used for predicting the treatment cost of healthcare services which for analyzing data. Classification techniques are also used for predicting the treatment cost of healthcare services which is increases with rapid growth every year and is becoming a main concern for everyone.

In our everyday life there are several example exit where we have to analyze the historical data for example a bank loans officer needs analysis of her data in order to learn which loan applicants are “safe” and which are “risky” for the bank. Similarly for a medical researcher it is necessary to analyze breast cancer data in order to predict specific treatments for a patient. These are some examples

where the data analysis task required before taking any decision. Classification is a data analysis process, where a classifier is constructed to predict class, for bank loan example prediction class is "yes" or "no" Similarly for a medical researcher prediction class is "treatment A," "treatment B," or "treatment C" for the medical data.

Classification process can be divided into two parts

(1)Learning: Training data are analyzed by a classification algorithm. Here, the class label attribute is loan decision, and the learned model or classifier is represented in the form of classification rules.

(2) Classification: Test data are used to estimate the accuracy of the classification rules. If the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples

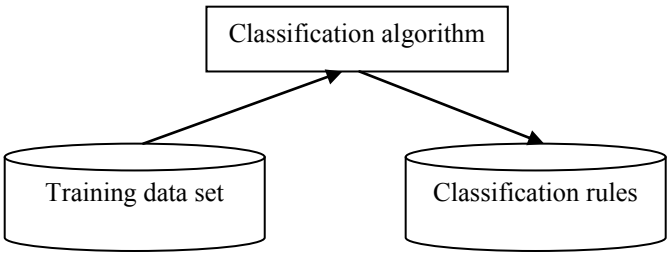

Fig.1. Rules construction with training dataset

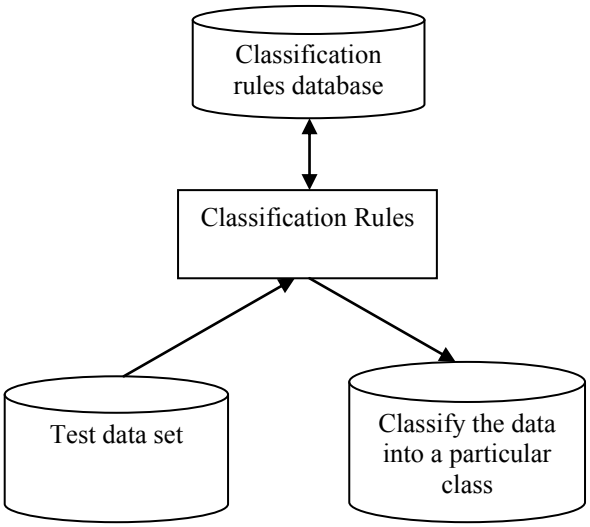

Fig. 2 Rules uses for classifying unknown tuples

II. LITRATURE SURVEY

In 2011 Mr. K. Ramesh &Mr. M. Chinna Rao proposed "Decision Support in Heart Disease Prediction System using Naive Bayes" Data Mining refers to using a variety of techniques to identify suggest of information or decision making knowledge in the database and extracting these in a way that they can put to use in areas such as decision support, predictions, forecasting and estimation. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective decision making. Discovering relations that connect variables in a database is the subject of data mining. This research has developed a Decision Support in Heart Disease Prediction System (DSHDPS) using data mining modeling technique, namely, Naïve Bayes. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It is implemented as web based questionnaire application. It can serve a training tool to train nurses and medical students to diagnose patients with heart disease.

In 2012 Qasem A. Radaideh & Eman Nagi proposed " Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance". They represent a study of data mining techniques and build a classification model to predict the performance

of employees. They build CRISP-DM model. They used Decision tree to build the classification model . They perform several experiments using real data collected from several companies. The model is intended to be used for predicting new applicants.

In 2012 M.Akhil jabbar, Dr.Priti Chandrab & Dr.B.L Deekshatuluc proposed "Heart Disease Prediction System using Associative Classification and Genetic Algorithm". They proposed efficient associative classification algorithm using genetic approach for heart disease prediction. Their main motivation for using genetic algorithm in the discovery of high level prediction rules is that the discovered rules are highly comprehensible, having high predictive accuracy and of high interestingness values. By the experimental analysis they show that most of the classifier rules help in the best prediction of heart disease which even helps doctors in their diagnosis decisions. They proposed a system for heart disease prediction using data mining techniques. In feature this work is used to reduce no. of attributes and to determine the attribute which contribute towards the diagnosis of disease using genetic algorithm.

In 2012 K. Rajesh, V. Sangeetha proposed "Application of Data Mining Methods and Techniques for Diabetes Diagnosis". Their main aim mining the relationship in Diabetes data for efficient classification. They applied many classification algorithms on Diabetes dataset and the performance of those algorithms is analyzed. In future this works enhance of improvisation of the C4.5 algorithms to improve the classification rate to achieve greater accuracy in classification .

In 2013 M. Akhil Jabbar, B.L Deekshatulu & Priti Chandra proposed "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection". They introduced a classification approach based ANN and feature subset selection. They used PCA for preprocessing and to reduce no. Of attributes which indirectly reduces the no. of diagnosis tests which are needed to be taken by a patient. We applied our approach on Andhra Pradesh heart disease data base. Our experimental results show that accuracy improved over traditional classification techniques. This system is feasible and faster and more accurate for diagnosis of heart disease.

In 2013 V. Krishnaiah, Dr. G. Narsimha, Dr. N. Subhash Chandra proposed "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" Cancer is the most important cause of death for both men and women. The early detection of cancer can be helpful in curing the disease completely. So the requirement of techniques to detect the occurrence of cancer nodule in early stage is increasing. A disease that is commonly misdiagnosed is lung cancer. Earlier diagnosis of Lung Cancer saves enormous lives, failing which may lead to other severe problems causing sudden fatal end. Its cure rate and prediction depends mainly on the early detection and diagnosis of the disease. One of the most common forms of medical malpractices globally is an error in diagnosis. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare system. In this study, we briefly examine the potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to massive volume of healthcare data. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information. For data preprocessing and effective decision making One Dependency Augmented Naïve Bayes classifier (ODANB) and naive creedal classifier 2 (NCC2) are used. This is an extension of naïve Bayes to imprecise probabilities that aims at delivering robust classifications also when dealing with small or incomplete data sets.

In 2013 Divya Tomar and Sonali Agarwal proposed "A survey on Data Mining approaches for Healthcare". Survey explores the utility of various Data Mining techniques such as classification, clustering, association, regression in health domain. They represent a brief introduction of these techniques and their advantages and disadvantages. This survey 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 .

In 2014 Dr. B Rosiline Jeetha proposed "Efficient Classification Method for Large Dataset by Assigning the Key Value in Clustering". They proposed classification method to discover data of big difference from the instances in training data, which may mean a new data type. The generalize Canberra distance for continuous numerical attributes data to mixed attributes data, and use clustering analysis technique to squash existing instances, improve the classical nearest neighbor classification method.

In 2015 S. Olalekan Akinola, O. Jephthar Oyabugbe proposed "Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study". They determine how data mining classification algorithm perform with increase in input data sizes. 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. By the result show that Naïve Bayes takes least time to train data but with least accuracy as compared to MLP and Decision Tree algorithms.

In 2016 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 proposed a system for diagnosis and prognosis of cancer using Classification and Association approach in Data Mining. They used FP algorithm in Association Rule Mining (ARM) to conclude the patterns frequently found in benign and malignant patients. They also used Decision Tree algorithm under classification to predict the possibility of cancer in context to age.

In 2016 Tanvi Sharma, Anand Sharma& Vibhakar Mansotra proposed "Performance Analysis of Data Mining Classification Techniques on Public Health Care Data". They focused on the application of various data mining classification techniques used in different machine learning tools such as WEKA and Rapid miner over the public healthcare dataset for analysing 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.

III. PROBLEM STATEMENT

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. The main problem related to classification techniques are

- ***Accuracy*: -** This includes accuracy of the classifier in term of predicting the class label, guessing value of predicted attributes.
- ***Speed:-***This include the required time to construct the model (training time) and time to use the model (classification/prediction time)
- ***Robustness*:-**This is the ability of the classifier or predictor to make correct predictions given noisy data or data with missing values.
- ***Scalability:-***Efficiency in term of database size.
- ***Interpretability:-***Understanding and insight provided by the model. Interpretability is subjective and therefore more difficult to assess.

IV. OBJECTIVES

There are several algorithms and methods have been developed to solve the problem of classification. But problem are always arises for finding a new algorithm and process for extracting knowledge for improving accuracy and efficiency. Our major objective are

- Design an efficient classification based algorithm which classify the given data set accurately.
- Design a classification based algorithm which classify the given data set using simple calculation and also reduce complexity.
- Design a classification based algorithm which reduce number of attributes for heart disease prediction.
- Design a classification based algorithm which work for both categorical as well as numerical of the attributes.

V. PROPOSED ALGORITHM

Input:

D a transaction database

FV Fitness Value

Output:

Pair of Attribute satisfies the given conations for heart attack

Method:

(1) Scan the database **D** and partition the transaction table into equal size**.**

(2) Find common attribute for each records.

(3) Consider only those attribute which satify the given minimum threshold value and delete Remaining attribute.

(4) To discover the pair of two attribute Join **L1 ⋈ L2** and perform logical AND. Number of attributes is greater than threshold value consider them for higher level otherwise prune 2attribute set sets.

(5) To determine 3 attribute set, join them and perform logical AND operation. If Number of 3attribute is greater than threshold value consider them for higher level otherwise **prune** 3 attribute set sets.

(6) The algorithm iterates to find up to pair of n- attributes item sets

(7) From each pair find out pair of n-attribute item sets. These pair of attribute are said to local attribute which satisfy the fitness value s

(8) Intersect the pair of attribute from each part to get global set of attribute which satify the fitness value

CONCLUSION

There are several algorithms and methods have been developed for classify heart attack problem accurately. But problem are always arises for finding a new algorithm and process for extracting knowledge for improving accuracy and efficiency The most popular classification methods are Decision Tree, Artificial neural networks, and Support Vector Machine and Naïve Bayes Classifier. From the experiment it clear that proposed method is more accurately classify the recodes as compared to previous method. Proposed method we try to reduce number of attributes for heart attack condition. Proposed method is also simple to under stands and calculation is also easy.

REFERENCE

1. G. Subbalakshmi & Mr. K. Ramesh "Decision Support in Heart Disease Prediction System using Naive Bayes" Indian Journal of Computer Science and Engineering (IJCSE) ISSN : 0976-5166 Vol. 2 No. 2 Apr-May 2011
2. 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
3. M.Akhil jabbar, Priti Chandrab & B.L Deekshatulu "Heart Disease Prediction System using Associative Classification and Genetic Algorithm" International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012
4. K. Rajesh & V. Sangeetha " Application of Data Mining Methods and Techniques for Diabetes Diagnosis ISSN: 2277-3754 ISO 9001:2008 Certified International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012
5. M. Akhil Jabbar, B.L Deekshatulu & Priti Chandra "Classification of Heart Disease using Artificial Neural Networkand Feature Subset Selection "Global Journal of Computer Science and TechnologyNeural & Artificial Intelligence Volume 13 Issue 3 Version 1.0 Year 2013
6. V.Krishnaiah, G.Narsimha & Dr.N.Subhash Chandra "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques V. Krishnaiah et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, 39 – 45
7. Divya Tomar & Sonali Agarwal " survey on Data Mining approaches for Healthcare International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266 http://dx.doi.org/10.14257/ijbsbt.2013.5.5.25
8. Dr. B Rosiline Jeetha "Efficient Classification Method for Large Dataset by Assigning the Key Value In Clustering IJCSMC, Vol. 3, Issue. 1, January 2014, pg.319 – 324 International Journal of Computer Science and Mobile Computing A Monthly Journal of Computer Science and Information Technology
9. 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, 8, 470-477 Published Online September.
10. 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
11. Tanvi Sharma & Anand Sharma " 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
12. 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
13. Mai Shouman, Tim Turner, Rob Stocker "Using Decision Tree for Diagnosing Heart Disease Patients Copyright © 2011, Australian Computer Society,
14. Chaitrali S. Dangare & Sulabha S. Apte, "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
15. Vikas Chaurasia & Saurabh Pal "Early Prediction of Heart Diseases Using Data Mining Techniques", Carib.j.SciTech,2013,Vol.1,208-217 © 2013. Journal of Science and Technology