Volume 3, Issue 12, December 2018 ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

Decision Support System for Disease Prediction Using Clustering

Seema Choudhary
M Tech IV Semester CSE Department
Lord Krishna College of Technology
Indore M P. India
seemachoudhari94@gmail.com

Vijay Kumar Verma
Asst Professor C.S.E. Department
Lord Krishna College of Technology
Indore M P. India
vijayvermaonline@gmail.com

Abstract: ***Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data. The derived model may be represented in various forms, such as classification IF-THEN rules, decision trees, mathematical formulae, or neural networks. Classification methods can handle both numerical and categorical attributes. Constructing fast and accurate classifiers for large data sets is an important task in data mining and knowledge discovery. Classification predicts categorical class labels and classifies data based on the training set. Classification is two step processes. In this paper we proposed a new approach to classify heart attack using clustering.***

Keywords: Classification, Clustering, Accuracy Disease , Prediction

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 is increases with rapid growth every year and is becoming a main concern for everyone.

II. CLASSIFICATION AND PREDICTION

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.

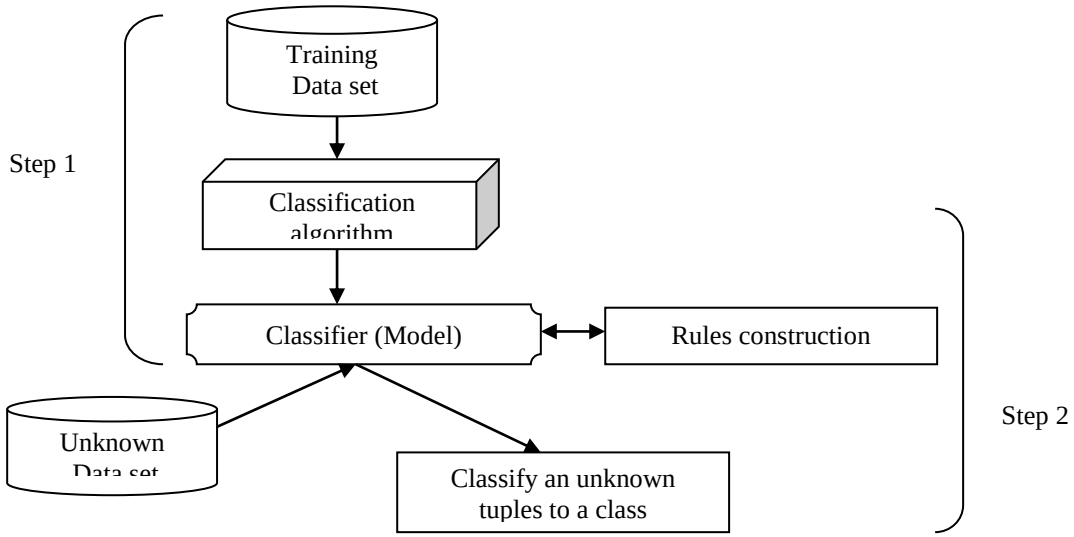


Figure 1.Classification steps

The derived model is based on the analysis of a set of training data. The derived model may be represented in various forms, such as classification IF-THEN rules, decision trees, mathematical formulae, or neural networks. Classification methods can handle both numerical and categorical attributes. Constructing fast and accurate classifiers for large data sets is an important task in data mining and knowledge discovery. Classification predicts categorical class labels and classifies data based on the training set. Classification is two step process.

Model construction: describing a set of predetermined classes. Each tuple/sample is assumed to long 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 formula.
Model usage: for classifying future or unknown objects. Estimate accuracy of the model. The known label of test sample is compared with the classified result from the model .Accuracy rate is the percentage of test set samples that are correctly classified by the model .Test set is independent of training set.

III. CLASSIFICATION TECHNIQUES

Classification is the task of generalizing known structure to apply to new data. The classification task can be seen as a supervised technique where each instance belongs to a class, which is indicated by the value of a special goal attribute or simply the class attribute. The goal attribute can take on categorical values, each of them corresponding to a class. One of the major goals of a Classification algorithm is to maximize the predictive accuracy obtained by the classification model when classifying examples in the test set unseen during training Three are several techniques are used for classification some of them are[22,23,24].

1. Decision Tree,
2. K-Nearest Neighbor,
3. Support Vector Machines,
4. Naive Bayesian Classifiers,
5. Neural Networks.

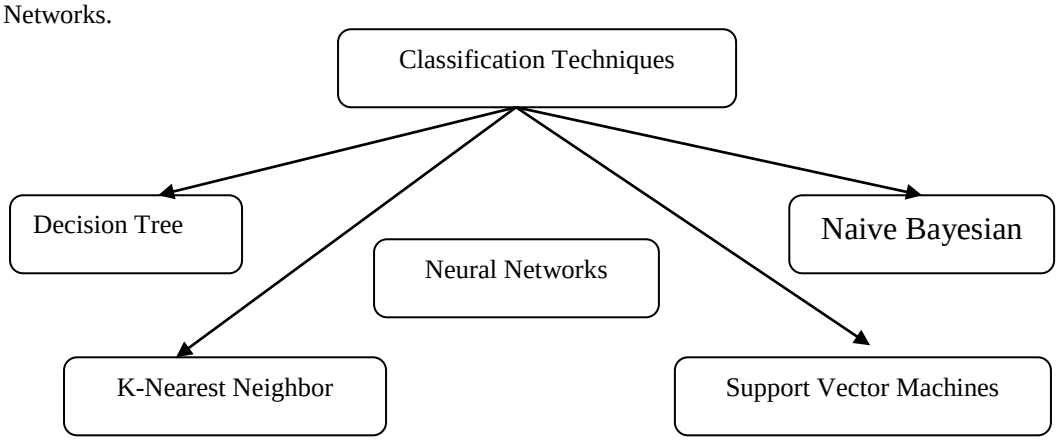


Figure-2 Classification model steps

Decision Trees

A Decision Tree Classifier consists of a decision tree generated on the basis of instances. A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces a certain discrete function of the input attributes values.

K-Nearest Neighbor Classifiers (KNN)

K-Nearest neighbor classifiers are based on learning by analogy. The training samples are described by n dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknowns ample. "Closeness" is defined in terms of Euclidean distance, where the Euclidean distance, where the Euclidean distance between two points, X=(x1,x2,……,xn) and Y=(y1,y2,….,yn) is denoted by d(X, Y).

Support Vector Machine (SVM)

SVM is a very effective method for regression, classification and general pattern recognition. It is considered a good classifier because of its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high. It is considered a good classifier because of its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high. The aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data. The metric for the concept of the "best" classification function can be realized geometrically.

Naive Bayes Classifier

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. The Naive Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naïve Bayes model identifies the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state. Figure 2.5 show working of Bayesian classifier, or simple Bayesian classifier.

Neural Networks.

Neural Network used for classification that uses gradient descent method and based on biological nervous system having multiple interrelated processing elements known as neurons, functioning in unity to solve specific problem. Rules are extracted from the trained Neural Network (NN) help to improve interoperability of the learned network. To solve a particular problem NN used neurons which are organized processing elements. Neural Network is used for classification and pattern recognition. An NN is adaptive in nature because it changes its structure and adjusts its weight in order to minimize the error. Adjustment of weight is based on the information that flows internally and externally through network during learning phase.

IV. LITERATURE REVIEW

In 2011 Mrs. G. Subbalakshmi and Mr. K. Ramesh proposed "Decision Support in Heart Disease Prediction System using Naive Bayes". They proposed 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. They implement the system by using web based questionnaire application. This system helps to train nurses and medical students to diagnose patients with heart disease. Decision Support in Heart Disease Prediction System is developed using Naive Bayesian Classification technique. The system extracts hidden knowledge from a historical heart disease database. This is the most effective model to predict patients with heart disease. This model could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy. DSHDPS can be further enhanced and expanded[1].

In 2012 M. Akhil Jabbar, Dr. Priti Chandra, Dr. B. L. Deekshatulu proposed "Heart Disease Prediction System using Associative Classification and Genetic Algorithm". They proposed an efficient associative classification algorithm using genetic approach for heart disease prediction. The 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. They proposed a system for heart disease prediction using data mining techniques. In our feature work we plan to reduce no. of attributes and to determine the attribute which contribute towards the diagnosis of disease using genetic algorithm[2].

In 2013 V. Krishnaiah , Dr. G. Narsimha, Dr. N. Subhash Chandra proposed "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques". They 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. This is an extension of Naïve Bayes to imprecise probabilities that aims at delivering robust classifications also when dealing with small or incomplete datasets. Discovery of hidden patterns and relationships often goes unexploited. They propose a model for early detection and correct diagnosis of the disease which will help the doctor in saving the life of the patient. Prototype lung cancer disease prediction system is developed using data mining classification techniques. The system extracts hidden knowledge from a historical lung cancer disease database[3].

In 2014 Mariammal D., Jayanthi S., Dr. P.S.K. Patra proposed "Major Disease Diagnosis and Treatment Suggestion System Using Data Mining Techniques". They proposed a model to systematically close those gaps to discover if applying single and multiple data mining techniques to all disease treatment data can provide as reliable performance as that achieved in diagnosing disease. Using multiple data mining techniques the accuracy also improved. Disease prediction is a major challenge in the health care industry. Instead of going for a number of tests, predicting the major disease with less number of attributes is a challenging task in Data Mining. Decision Support in Disease Prediction System is developed using all the five data mining techniques. The Disease diagnosis system extracts hidden knowledge from a historical disease database[4].

In 2015 Ebenezer Obaloluwa Olaniyi and Oyebade Kayode Oyedotun proposed "Heart Diseases Diagnosis Using Neural Networks Arbitration". They proposed causes of heart diseases, the complications and the remedies for the diseases have been considered. An intelligent system which can diagnose heart diseases has been implemented. This system will prevent misdiagnosis which is the major error that may occur by medical doctors. The dataset of statlog heart disease has been used to carry out this experiment. The dataset comprises attributes of patients diagnosed for heart diseases. The diagnosis was used to confirm whether heart disease is present or absent in the patient. The datasets were obtained from the UCI Machine Learning[5].

In 2016 Isra'a Ahmed Zriqat, Ahmad Mousa Altamimi, Mohammad Azzeh proposed "A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods". They motivate is to develop an effective intelligent medical decision support system based on data mining techniques. They used five data mining classifying algorithms, with large datasets, have been utilized to assess and analyze the risk factors statistically related` to heart diseases in order to compare the performance of the implemented classifiers (e.g., Naïve Bayes, Decision Tree, Discriminate, Random Forest, and Support Vector Machine). Results of the conducted experiments showed that all classification algorithms are predictive and can give relatively correct answer. Although ensemble learning has been proved to produce superior results, but in our case the decision tree has outperformed its ensemble version[6].

In 2017 Sanjay Kumar Sen proposed "Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms". The main objective of this research is predicting the heart disease of a patient using machine learning algorithms. Comparative study of the various performances of machine learning algorithms is done through graphical representation of the results. They carried out an experiment to find the predictive performance of different classifiers. They select four popular classifiers considering their qualitative performance for the experiment. They also choose one dataset from heart available at UCI machine learning repository. Naïve base classifier is the best in performance. In order to compare the classification performance of four machine learning algorithms, classifiers are applied on same data and results are compared on the basis of misclassification and correct classification rate and according to experimental results it can be concluded that Naïve base classifier is the best as compared to Support Vector Machine, Decision Tree and K-Nearest Neighbour[7].

In 2018 Poornima V, Gladis D proposed "A novel approach for diagnosing heart disease with hybrid classifier". They proposed an Orthogonal Local Preserving Projection (OLPP) method to reduce the function dimension of the input high-dimensional data. The dimension reduction improves the prediction rate with the help of hybrid classifier i.e. Group Search Optimization Algorithm (GSO) combine with the Levenberg-Marquardt (LM) training algorithm in the neural network. The LM training algorithm is used to solve the optimization problem and it determines the best network parameters such as weights and bias that minimizes the error. The final output of the optimization technique is combined with the performance metrics as accuracy, sensitivity, and specificity. From the result, it is observed that hybrid optimization techniques increase the accuracy of the heart disease prediction system[8].

V. 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 like doctors, 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

1) Accuracy: - This includes accuracy of the classifier in term of predicting the class label, guessing value of predicted attributes.
2) Speed:-This include the required time to construct the model (training time) and time to use the model (classification/prediction time)
3) Robustness:-This is the characteristic of the classifier or predictor to make correct prediction and give correct result on noisy data.
4) Scalability:-Efficiency in term of database size

VI. PROPOSED APPROACH

First we convert data set into binary format according to the given conditions. In seconds step we divide the data set into two parts apply the fitness value condition on each attribute Find pair for each attribute which satisfy the condition .we repeat the process for grouping the attribute until no more grouping is possible. At last we find the most common attribute in both the part and calculate how much percentage data is accurately classified.

Input:
D transaction database
F Threshold value
Output: Pair of Attribute satisfy the given contains for heart attack
Method:
1. Partition database D into two parts.

2. Use most appropriate value for each attribute and compare with given condition.
3. Consider only those attribute which satisfy the given minimum threshold value and delete remaining attribute.
4. Use Join L1 ⋈L2 and perform logical AND for combination for two attribute.
5. To determine 3 attribute set ,join them and perform logical AND operation . The algorithm iterates to find up to pair of n- attribute item sets
6. 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
7. Intersect the pair of attribute from each part to get global set of attribute which satisfy the fitness value

VII. EXPERIMENTAL ANALYSIS

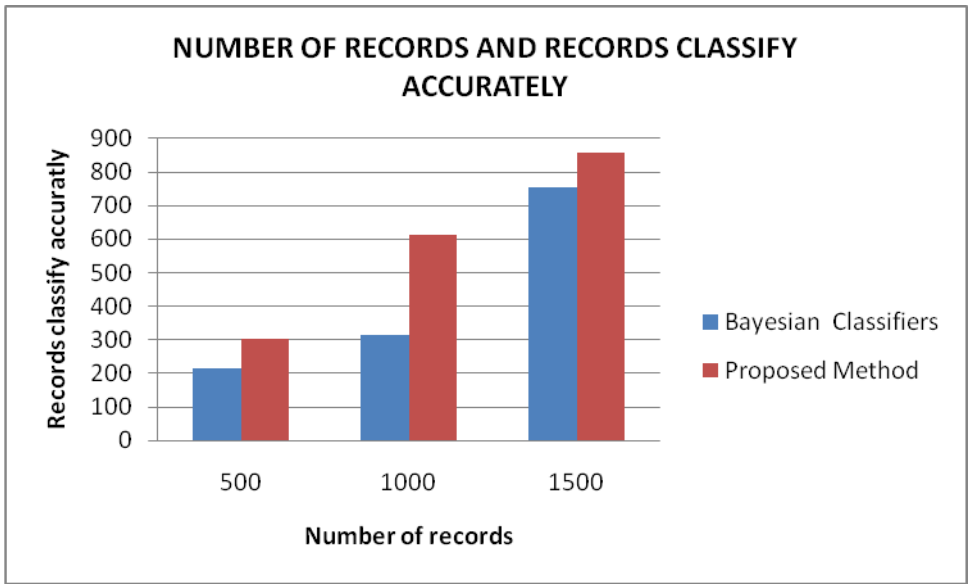
The experiments were performed on Intel Core i3processor 1GB main memory and RAM: 4GB Inbuilt HDD: 400GB OS: Windows7.The algorithms are implemented in using Dot Net Framework language version 4.0.1. Synthetic datasets are used to evaluate the performance of the algorithms. We have used SQL Server 2008 R2 for storing patient's database. We have taken 10 attribute which are mainly responsible for heart attack condition. Age, Sex, Blood pressure, Cholesterol, Fasting blood sugar, Resting ECG, Thalach value, Old peak, Slope and Thal. We have taken real life data from a laboratory. We have taken data of 1000 patient.SQL Server R2 (2008) to store our database.

COMPARISON USING NUMBER OF RECORDS AND RECORDS CLASSIFY ACCURATELY

For comparing the performance of the proposed algorithm with other methods we take records on the different number of records. In table 1 we have accuracy of the proposed method with other methods on 1000, 2000 and 3000 records of data. This comparison is based on execution time and number of objects.

Table 1 Number of Records classify accurately

Number of Records	Bayesian Classifiers	Proposed Method
1000	352	612
2000	735	1265
3000	1361	2281



CONCLUSION AND FUTURE WORK

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 Artificial neural networks, Decision Tree 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 considers all attribute given to heart attack condition. Proposed method is also simple to understand and calculation is easy.

REFERENCE

[1] Mrs. G. Subbalakshmi and Mr. K. Ramesh Decision Support in Heart Disease Prediction System using Naive Bayes Journal of Computer Science and Engineering (IJCSE) ISSN : 0976-5166 Vol. 2 No. 2 Apr-May 2011

[2] M. Akhil jabbar, Dr B.L Deekshatulu, Dr. Priti Chandra "Heart Disease Classification Using Nearest Neighbor Classifier **With** Feature Subset Selection Computer Science and Telecommunications 2013|No.3(39) ISSN 1512-1232.

[3] V. Krishnaiah, Dr. 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

[4] Mariammal. D, Jayanthi. S, Dr. P. S. K. Patra Major Disease Diagnosis and Treatment Suggestion System Using Data Mining Techniques International Journal of Advanced Research in Computer Science & Technology IJARCST All Rights Reserved 338 Vol. 2 Issue Special 1 Jan-March 2014 ISSN: 2347 - 8446 (Online) ISSN: 2347 - 9817

[5] Ebenezer Obaloluwa Olaniyi and Oyebade Kayode Oyedotun "Heart Diseases Diagnosis Using Neural Networks Arbitration" I.J. Intelligent Systems and Applications, 2015, 12, 75-82 Published Online November 2015 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijisa.2015.12.08

[6] Isra'a Ahmed Zriqat, Ahmad Mousa Altamimi, Mohammad Azzeh "A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods" International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 12, December 2016 ttps://sites.google.com/site/ijcsis/ ISSN 1947-5500

[7] Sanjay Kumar Sen "Predicting and Diagnosing of Heart Disease Using Machine Learning Algorith" International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 6 Issue 6 June 2017, Page No. 21623-21631 Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i6.14

[8] Poornima V, Gladis D A novel approach for diagnosing heart disease with hybrid Biomedical Research 2018; 29 (11): 2274-2280 ISSN 0970-938X www.biomedres.info