ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 IEEE International Conference on Innovations in Engineering and Technology (ICIET'14)
On 21st & 22nd March Organized by

K.L.N. College of Engineering and Technology, Madurai, Tamil Nadu, India

Heart Disease Analysis System Using Data Mining Techniques

G.Karthiga[1], C.Preethi[2], R.Delshi Howsalya Devi[3]

[1]Department of Computer Science and Engineering , KLN College of Engineering, Sivagangai, India

[2]Department of Computer Science and Engineering, KLN College of Engineering, Sivagangai, India

[3]Assistant Professor, Department of Computer Science and Engineering, KLN College of Engineering, Sivagangai, India

ABSTRACT-Data mining is an iterative progress in which evolution is defined by detection, through usual or manual methods. Knowledge discovery and data mining have found various applications in scientific domain Heart disease is a term for defining a huge amount of healthcare conditions that are related to the heart. This medicinal condition defines the unpredicted health conditions that directly control all the parts of the heart. Different data mining techniques such as association rule mining, classification, clustering are used to predict the heart disease in health care industry .The heart disease database is preprocessed to make the mining process more efficient. The preprocessed data is clustered using clustering algorithms like K-means to cluster relevant data in database. Maximal Frequent Itemset Algorithm (MAFIA) is used for mining maximal frequent patterns in heart disease database. The frequent patterns can be classified using C4.5 algorithm as training algorithm using the concept of information entropy. The results showed that the designed prediction system is capable of predicting the heart attack with good accuracy.

KEYWORDS—Data mining; K-means clustering; MAFIA (Maximal Frequent Itemset Algorithm; C4.5 algorithm

I. INTRODUCTION

Data mining is process of extracting useful information from large amount of databases.Data mining is most useful in an exploratory analysis because of nontrivial information in large volumes of data. Data mining is the process of extracting data for finding buried patterns which can be transformed into significant. Data mining knowledge afford a user-oriented approach to new and concealed patterns in the data.

The knowledge which is exposed can be used by the healthcare practitioners to get better quality of service and to reduce the extent of adverse medicine effect. Hospitals have to reduce the charge of medical tests. They can attain these consequences by employing suitable decision support systems. Health care data is enormous. It consists of patient centric data, resource organization data and altered data. Medical care organizations must have capability to explore data. Treatment records of millions of patients can be hoarded and data mining techniques will aid in answering numerous essential and decisive questions interrelated to health care.Data mining techniques has been performed in healthcare domain. This realization is in the arouse of explosion of difficult medical data. Medicinal data mining can utilize the veiled patterns present in huge medical data which otherwise is left undiscovered. Data mining techniques which are useful to medical data include association rule mining for finding frequent patterns, prediction, classification and clustering. Data mining techniques are more useful in predicting heart diseases, lung cancer, and diabetes.

This paper analyzes the heart disease predictions using classification algorithms. These hidden patterns can be used for health diagnosis in medicinal data. Data mining technology afford an effective approach to latest and indefinite patterns in the data. The information which is identified can be used by the healthcare administrators to get better services. Heart disease was the most important reason of victims in the countries like India, United States. Data mining techniques like Association Rule Mining, Clustering, Classification algorithms such as Decision tree [7], C4.5 algorithm, Neural Network [8], Naive Bayes [9] are used to explore the different kinds of heart based problems [1]. Data mining techniques like C4.5 algorithm and K-means clustering are used for validating the accuracy of medicinal data. These algorithms can be used to optimize the data storage for practical and legal purposes.

II. RELATED WORK

The data mining techniques includes different works to explore a variety of diseases such as Cancer, Diabetes, Heart diseases. Heart disease is the most important reason of fatality in the UK, USA, Canada, and England [2]. Heart disease kills individual in each 32 seconds in the world. Jyoti Soni et al proposed three different supervised machine learning algorithms for heart disease prediction. They are Naïve Bayes, K-nearest neighbor, and Decision tree. These algorithms have been used for analyzing the heart disease. Tanagra is the data mining tool used for classifying these medical data and these data are calculated using 10 fold cross validation. Naive Bayes algorithm performs well when compared to other algorithms [3].

Genetic algorithm have been used in [6], to reduce the definite data size to obtain the best possible subset of attribute which is essential for heart disease prediction. Classification is supervised learning method to extract models relating main classes of data. Decision Tree, Naïve Bayes and Classification via clustering are the three classifiers used to analyze the occurrence of heart disease for the patients. Shekar et al proposed new algorithm to mine association rules from medical data based on digit sequence and clustering for heart attack prediction the entire data base is divided into partitions of equal size, each partition will be called cluster. This approach reduces main memory requirement since it consider only a small cluster at a time and it is scalable and efficient [5].

III. MATERIALS AND METHODS

The extraction of significant patterns from the heart disease data warehouse is presented in this section. The heart disease data warehouse contains the screening clinical data of heart patients. Initially, the data warehouse is preprocessed to make the mining process more efficient. Four steps for the Disease Prediction Process are as follows:

1. Preprocessing is done in order to handle missing values.
2. Cluster relevant data using K-means Clustering.
3. Selecting the frequent pattern using MAFIA
4. Applying C4.5 algorithm to classify the pattern.

3.1 Data Preprocessing

Cleaning and filtering of the data might be necessarily carried out with respect to the data and data mining algorithm employed so as to avoid the creation of deceptive or inappropriate rules or patterns. In preprocessing first it selects an attribute for selecting a subset of attributes with good predicting capability. It handles all missing values and investigates each possibility. If an attribute has more than 5% missing values then the records should not be deleted and it is advisable to impute values where data is missing, using a suitable method.

3.2 K-means Clustering

Grouping a set of objects in such a way that objects in the same group is more similar to each other than to those in other groups. Clustering is an unsupervised learning. The algorithm clusters information's into k groups, where k is considered as an input parameter. It then assigns each information's to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then more computed and the process begins again.The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging and related fields. The steps involved in K-means algorithms are as follows:

- Choose the number of clusters, k.

- Randomly create k clusters and find the cluster midpoint.
- Consign each point to the closest cluster midpoint

- Recompute
the new cluster midpoints.

- Iterate the above two steps until various convergence condition is met.

The preprocessed data is clustered using k-means to cluster relevant data in the heart disease database shown in Table I.K-means algorithm produces a specific number of separate, non-hierarchical clusters. It is a simplest algorithm and performs faster than other clustering algorithms. It allows run on huge databases. It is mainly suitable for globular clusters.

3.3 MAFIA

The association rule mining problem is a main predicament in the data mining field with various realistic applications such as user medical data analysis, intrusion detection.

MAFIA is used for mining maximal frequent item sets from a transactional database [4]. This algorithm is mainly efficient when the item sets in the database are very long. The search strategy of this algorithm integrates a depth-first traversal of the item set lattice with efficient pruning mechanisms. A ⊆ I an item set, and call A a k-item set if the cardinality of item set A is k. Let database X be a multiset of subsets of I, and let support (A) be the fraction of item sets B in X such that A ⊆ B.. If support(A)=minSup, then A is a frequent item set, and indicate the set of all Frequent Item sets(MFI) by FI. If A is recurrent and no superset of A is frequent, then A is a Maximally Frequent Item set, and denotes the set of all Maximally Frequent Item sets by MFI.

MAFIA efficiently stores the transactional database as a series of vertical bitmaps, where each bitmap represents an item set in the database and a bit in each bitmap represents whether or a given customer has the corresponding item set. Initially, each bitmap represents an item set in database. The item sets that are checked for frequency in the database become recursively longer and the vertical bitmap representation works perfectly in conjunction with this item set extension

TABLE I. HEART DISEASE DATABASE

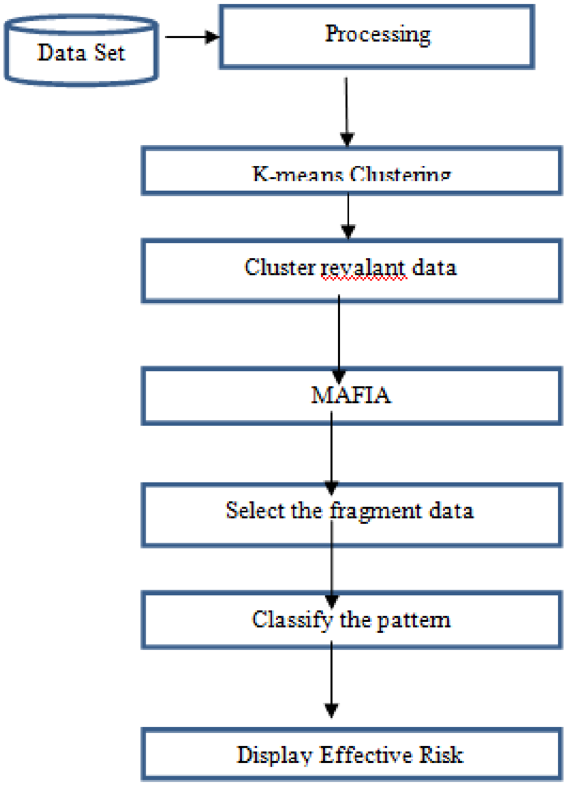
ID	ATTRIBUTE
1	Patient Id
2	Age
3	Sex(value 1: Male; value 0: Female)
4	Slope: the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat;value 3: down sloping)
5	famhist: family history of coronary artery disease (value 1 :yes; value 0 : no)
6	Fasting Blood Sugar (value 1: >120 mg/dl; value 0: <120 mg/dl
7	painloc: chest pain location (value 1:substernal; value 0: otherwise),
8	Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
9	chol: serum cholesterol
10	trestbps: resting blood pressure

11	Exang : exercise induced angina (value 1: yes; value 0: no)
12	Maximum Heart rate achieved :Value(0,0)

3.4 C4.5 Algorithm

Classification is an unsupervised learning used to predict the class of objects whose class label is unknown. It is used for creating classification rules by means of decision trees from a given data set.

Decision tree is used as a prognostic model. C4.5, C5.0, CART, ID3 are methods for building decision trees. It is an extension of the basic ID3 algorithm. By using C4.5, decision trees can be building from a set of training data with the information entropy. It is a statistical classifier. It outputs can be in the form of if then rules



IV. EXPERIMENTAL RESULTS

The result of experimental analysis in identifying important patterns for predicting heart diseases are presented in Figure 1. The heart disease database is preprocessed effectively by removing related records and given that missing values .The well mannered heart disease data set[10], resulting from preprocessing, is then composed by K-means algorithm with the K value of 2.Then the frequent forms are mined efficiently from the set appropriate to heart disease, using the MAFIA .

The experimental results of this approach as presented in Figure1.Confusion matrix is a matrix representation of the classification results. From the confusion matrix to explore the performance criterion for the classifiers in disease prediction accuracy have been calculated for the medical datasets.

These can be transformed into True-Positive (TP), True-Negative (TN), False- Negative (FN) and False-Positive (FP) metrics.

(5.1) TNFNFP+TPTN TP =Accurac

o True Positive (TP): Total fraction of Data Pre members classified as Class A belongs to Class A

- False Positive (FP): Total fraction of members of Class A but does not belong to Class A.
- False Negative (FN): Total fraction of members of Class A incorrectly classified as not belonging to Class A

- True Negative (TN): Total fraction of members which does not belong to Class A are classified not a part of Class A .It can also be given as(100%-FP).

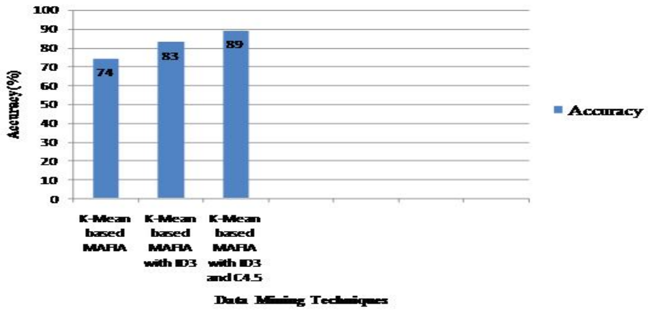


Fig 1. Prediction Accuracy between Simple MAFIA and Proposed K-Mean Based MAFIA

V. CONCLUSIONS AND FUTURE WORK

Medical related information's are huge in nature and it can be derived from different birthplaces which are not entirely applicable in feature. In this work, heart disease prediction system was developed using clustering and classification algorithms to predict the effective risk level and accuracy of the patients. In future work, we have planned to propose an effective disease prediction system to predict the heart disease with better accuracy using different data mining techniques and compare the performance of algorithm with other related data mining algorithms.

ACKNOWLEDGEMENT

The authors would like to thank the reviewers for their valuable comments that would help to improve this

REFERENCES

[1] V. Manikantan and S. Latha, "Predicting the analysis of heart disease symptoms using medicinal data mining methods", International Journal of Advanced Computer Theory and Engineering, vol. 2, pp.46-51, 2013.

[2] Shadab Adam Pattekari and Alma Parveen,"Prediction system for heart disease using Naïve Bayes", International Journal of Advanced Computer and Mathematical Sciences, vol.3,pp 290-294,2012.

[3] Jyoti Soni, Ujma Ansari, Dipesh Sharma and Sunita Soni, "Predictive data mining for medical diagnosis: an overview of heart disease prediction", International Journal of Computer Science and Engineering, vol. 3, pp.43-48, 2011.

[4] Hnin Wint Khaing, "Data Mining based fragmentation and prediction of medical data", International Conference on Computer Research and Development, ISBN: 978-1-61284-840-2, 2011.

[5] K.Shekar, N.Deepika and D.Sujatha,"Association rule for classification of heart-attack patients", International Journal of Advanced Engineering Sciences and Technologies, vol.11, no. 2, pp.253-257, 2011.

[6] M. Anbarasi, E. Anupriya and N.Iyengar, "Enhanced prediction of heart disease with feature subset selection using Genetic algorithm", International Journal of Engineering Science and Technology vol.2, pp.5370- 5376, 2010.

[7] Sellappan Palaniappan and Rafiah Awang, "Intelligent heart disease prediction system using data mining techniques", International Journal of Computer Science and Network Security, vol.8, no.8, pp. 343-350,2008.

[8] K.Srinivas, Dr.G.Ragavendra and Dr. A. Govardhan,"ASurvey 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, May 2011.

[9] G.Subbalakshmi, K.Ramesh and N.Chinna Rao,"Decision support in heart disease prediction system using Naïve Bayes", ISSN: 0976-5166, vol. 2, no. 2.pp.170-176, 2011

[10] Cleveland dataset from http://archive.ics.uci.edu.