International Journal for Modern Trends in Science and Technology
Volume: 03, Special Issue No: 02, March 2017
ISSN: 2455-3778
http://www.ijmtst.com

Heart Anomaly Forecasting Applying a Series of Data Mining Methods

Ch.Sai Chaitanya[1] | Ch.Nagur Vali[2] | V.Sai Vikas[3] | K.Sowmya[4]

[1,2,3,4]Department of IT, Andhra Loyola Institute of Engineering & Technology, Vijayawada, Andhra Pradesh, India.

To Cite this Article
Ch.Sai Chaitanya, Ch.Nagur Vali, V.Sai Vikas and K.Sowmya, "Heart Anomaly Forecasting Applying a Series of Data Mining Methods", *International Journal for Modern Trends in Science and Technology*, Vol. 03, Special Issue 02, 2017, pp. 69-73.

ABSTRACT

Many medicinal fields comprise lots of data regarding the health care. There may be some data with which we can derive some interesting patterns which frequently go exploited. Data mining is the process of analyzing huge data sets and derive the relevant information. It becomes more prominence in case of heart disease which is the major concern behind the deaths all over the world. The term heart disease refers to various types of conditions that can affect the heart functions. This medicinal condition defines the unpredicted health conditions that directly control all the parts of the heart. It makes use of different parameters included in data mining include clustering, association rule mining and prediction which are used to predict the heart disease. Initially the heart disease data set is preprocessed to eliminate any irrelevant data. The preprocessed data is now clustered using K-means clustering algorithm. Maximal Frequent Item set Algorithm (MAFIA) is used to derive the maximal frequent patterns in data set. These results can be classified using C 4.5 training algorithm to build a decision tree. The calculated metrics proved that our designed prediction system is capable of predicting the heart disease successfully.

KEYWORDS: *Data mining; K-means clustering MAFIA (Maximal Frequent Item set Algorithm; C4.5 algorithm*

I. Introduction

Data mining is process of digging significant information from huge amount of data sets. The methodologies of data mining have been applied enormously in various fields including business, bio-medical informatics, science etc. There are countless diseases, which if predicted in advance may save many lives. Heart disease is one among them.

The data mining approaches are valuable for foretelling the probability of occurrence of various diseases in the medical field. Disease prediction plays a vital role in the sectors where data mining is being carried out. There are several diseases in our real world scenario, which if predicted in advance may yield some productive information.

This paper evaluates the prediction system by employing classification algorithms. Data mining technology afford an effective analytical approach for detecting unknown and valuable information in health care industries. This identified information can be used by the healthcare analysts serve for better applications. Heart disease was the most important reason of victims in the countries like India, as it is currently witnessing nearly two million heart attacks a year and majority of the victims are

Clustering, Classification algorithms such as Decision tree, C4.5 algorithm, Neural Networks, Naive Bayes, are used to explore the different kinds of heart based problems. Data mining techniques like C4.5 algorithm and K-means clustering are used for legitimizing the accuracy of data informatics. These algorithms can be used to enhance the data storage for practical and legal purposes.

II. RELATEDWORK

The data mining techniques are intensively used to inspect a variety of diseases such as Cancer, Diabetes, Heart diseases. Coronary Heart disease is the most common type in United States and England. Heart disease kills one person every 33 seconds in India. India is witnessing two million heart attacks a year. Several Data Mining algorithms like Naïve Bayes, K-nearest neighbor, and Decision tree are used for better prediction. Data Mining tools such as Weka and Tanagra are considered as best fit for classifying the medical data. Naive Bayes algorithm fulfills the best of the classification problems [3].

Genetic algorithm have applied in [6], to reduce the definite data size to obtain the best possible attribute which is essential for heart disease prediction. Classification is supervised learning method which accurately predicts the target class for each case in data. Decision Tree, Naïve Bayes and Classification via clustering are the effective classifiers used to evaluate the heart disease data sets and project the effective risk level. Shekar et al proposed new algorithm for the extraction of association rules from health care data based on digit sequence and clustering. For heart disease prediction the entire data base is split into partitions of equal size, each partition is considered as cluster. This approach trim the main memory requirement since it consider only a single cluster at a time and it is adaptable and efficient [5].

III. MATERIALSANDMETHODS

A.Data Preprocessing

If any data mining algorithm is being employed, cleaning and filtering of data should be mandatorily carried out in order to eliminate any misleading patterns and irrelevant rules. The preprocessing mechanism initially selects an attribute to handle all missing values and explores each outcome. If an attribute is said to be having 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.

B .K-means algorithm

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. Next it 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 will continue 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 Kmeans 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 Re-compute the new cluster midpoints.
- Iterate the above two steps until various convergence condition is met.

The preprocessed data set is clustered utilizing k-means to group relevant information in the heart disease database appeared in Table I. K-means calculation creates a particular number of independent, nonhierarchical clusters. It is a least complex calculation and performs quicker than other clustering methodologies. It permits keep running on enormous databases. It is primarily appropriate for globular clusters.

C.*MAFIA*

The association rule mining issue is a primary quandary in the data mining field with different practical applications, for example, client medicinal information examination, intrusionrecognition.

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 effectively stores the value-based database as a progression of vertical bitmaps, where every bitmap speaks to a item set in the database and a bit in every bitmap speaks to whether or a given client has the corresponding item set. At first, every bitmap speaks to a item set in database. The item sets that are checked for recurrence in the database turn out to be recursively more and the vertical bitmap portrayal works flawlessly in conjunction with this item set augmentation.

D .C 4.5 Algorithm

Classification is an unsupervised learning used to predict the class of items whose class label is obscure. It is utilized for making association rules by methods for decision trees from a given informational collection.

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 theinformation entropy. It is a statistical classifier. It outputs can be in the form of if then rules

A.Sample Algorithm

```
- Test for base cases.
- For each element n, determine the
    normalized information gain from
    separating on n.
        OLet n best be the element with the
            highest normalized information gain
- Construct a decision node that breaks on a
    best.
- Recur on the sublists found by separating on
    n best, and attach these nodes as children
    of node.
```

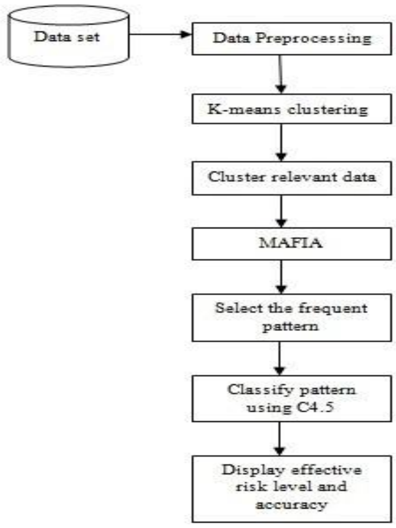
TABLE I. HEART DISEASE SET

Id	Attribute
1	Age
2	Sex(value 1: Male; value 0: Female)
3	Cp: (Chest Pain Value 1: Typical angina Value 2: Atypical angina Value 3: Non-anginal pain Value 4: Asymptomatic)
4	Trestbps: Resting blood pressure (in mm Hg on admission to the hospital)
5	Chol: serum cholesterol in mg/dl
6	Fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7	RestEcg: resting electrocardiographic results -- Value 0: normal -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8	Thalach: maximum heart rate achieved
9	exang: exercise induced angina (1 = yes; 0 = no)
10	Oldpeak = ST depression induced by exercise relative to rest
11	slope: the slope of the peak exercise ST segment(Value-1:upsloping, Value-2: flat, Value-3: downsloping)
12	Ca: number of major vessels (0-3) colored by flourosopy
13	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
14	Class: { Value-1: Positive, Value-2: Negative }

C4.5 Decision Tree Structure

If Age=<30 and Overweight=no and Alcohol Intake=never
then
Heart attack level is low
If Age=>70 and Blood pressure=High and Smoking=current then
Heart attack level is low.

IV. SYSTEMARCHITECTURE

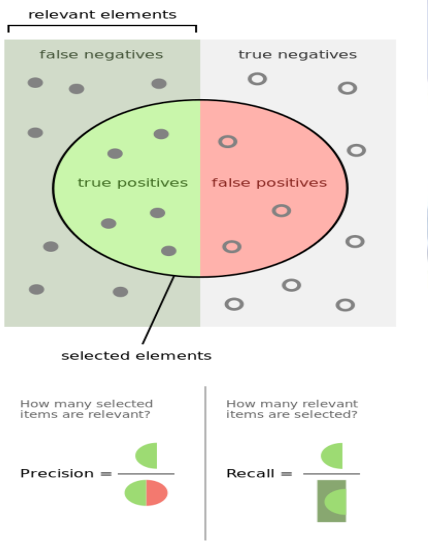

Data set
Data Preprocessing
K-means clustering
Cluster relevant data
MAFIA
Select the frequent pattern
Classify pattern using C4.5
Display effective risk level and accuracy

V.EXPERIMENTALRESULTS

The consequence of test investigation in distinguishing essential examples for anticipating heart illnesses are exhibited in this segment. The heart disease database is preprocessed viably by evacuating related records and given that missing qualities .The very much mannered heart disease data set[10], coming about because of preprocessing, is then made by K-means calculation with the K estimation of 2.Then the incessant structures are mined proficiently from the set appropriate to heart disease ,utilizing the MAFIA .

The experimental results of this approach as presented in Table II.The objective is to have greater accuracy, as high precision and recall metrics. These metrics can be converted into True-Positive (TP), True-Negative (TN), False-Negative (FN) and False-positive (FP) metrics.

In an arrangement undertaking, the accuracy for a class is the quantity of genuine positives (i.e. the quantity of things effectively marked as having a place with the positive class) isolated by the aggregate number of components named as having a place with the positive class (i.e. the aggregate of genuine positives and false positives, which are things erroneously marked as having a place with the class). Review in this setting is characterized as the quantity of genuine positives partitioned by the aggregate number of components that really have a place with the positive class (i.e. the whole of genuine positives and false negatives, which are things which were not named as having a place with the positive class but rather ought to have been).



$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

- ○ True Positive (TP): Total fraction of 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 oTrue 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).

TABLE II. PREDICTION ACCURACY BETWEEN SIMPLE MAFIA AND PROPOSED K-MEAN BASED MAFIA

Data mining Technique	Precision	Recall	Accuracy (%)
K-mean based MAFIA	0.78	0.64	74%
K-mean based MAFIA with ID3	0.80	0.84	84%
K-mean based MAFIA with ID3 and C4.5	0.82	0.89	89%

VI.CONCLUSIONANDFUTUREWORK

Medical related data's are colossal in nature and it can be gotten from various origin which are not by any stretch of the imagination appropriate in highlight. In this work, coronary illness expectation framework was produced utilizing clustering and classification techniques to anticipate the viable hazard level and precision of the patients. In future work, we have wanted to propose a compelling malady forecast framework to foresee the heart disease with better exactness utilizing diverse datamining methods and contrast the execution of algorithm and other related data mining algorithms.

Also integration of Data mining algorithms with R-programming may serve the best purpose in visual representation of the analysis. Our future work will involve the amalgamation of the various specified algorithms to augment the accuracy so that the diagnosis can develop into more accurate in case of imperceptibly identified data sets. Ongoing efforts are geared towards increasing the size of data set. The research work is of great help for analyzing various factors for booming situation of heart diseases. The system is of great relevance to the user in detection of the various factors related to various dangerous diseases which will help in providing the correct medication about him/her and will help in saving his precious human live.

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.4348, 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," A Survey 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.