International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume III, Issue IV, April- 2016

HEART DISEASE FORECASTING SYSTEM USING K-MEAN CLUSTERING ALGORITHM WITH PSO AND OTHER DATA MINING METHODS

Shilna S[1], Navya EK[2]
Mtech Student, Assistant Professor,
Department of Computer Science,
Malabar Institute of Technology
Kannur, Kerala, India
shilna94@gmail.com[1], navya.06rimaan@gmail.com[2]

Abstract

As huge amount of information is produced in medical associations, yet this information is not properly utilized. The health care system is "data rich" however "knowledge poor ". This healthcare data can be used to extract knowledge for further disease prediction. Currently data mining techniques are widely used in clinical expert systems for prediction of various diseases. These techniques discover the hidden relationships and patterns of the healthcare data. Heart disease is a term for defining a huge amount of healthcare conditions that are related to 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 is an unique combination of two most popular clustering algorithms Particle Swarm Optimization (PSO) and K-Means to achieve better clustering result. 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 successfully. Such systems will warn the people about the presence of their disease even before he concerns the doctor. This can even help doctors to carry out specific tests of the patients and target out the disease.

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

1. 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. The data mining techniques are useful for predicting the various diseases in the medical field. Cardiovascular diseases are one of the highest- flying diseases of the modern world1. According to world health organization about more than 12 million deaths occurs worldwide, every year due to heart problems. It is also one of the fatal diseases in India which causes maximum casualties. The diagnosis of this disease is intricate process. It should be diagnosed accurately and correctly. Due to limitation of the potential of the medical experts and their unavailability at certain places put their patients at high risk. Normally, it is diagnosed using intuition of the medical specialist. It would be highly advantageous if the techniques will be integrated with the medical information system

Disease prediction plays an important role in data mining. Healthcare organizations can reduce costs by accomplishment of computer based data and/or decision support systems. Healthcare services data is very huge as it incorporates patient records, resource management information and updated information. Human services associations must have capacity to break down information. Treatment records of many patients can be stored away in computerized way; furthermore data mining methods may help in finding out a few vital and basic inquiries related with healthcare organizations.

There are various reasons for the occurrence of Heart Diseases, which can be frequently investigated through the Attribute Set related to different test results of Patients. The different sources of medical data are Medical Analysis, Diagnostic Centres, past Case Sheets, Doctor Prescriptions. Heart diseases can be predicted through the analysis made on some attributes like age, sex, chest pain type, blood pressure, cholesterol, fasting blood sugar,

resting electro cardio graphic (ECG) result, Maximum heart rate achieved, Exercise Induced Angina, Old Peak, Slope of ECG reading, Number of vessels colored, thal. Based on the values of the attributes, we make indexes for all associated frequent itemsets. The presence of these itemsets depends on the threshold value specified. 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.

2. 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].

3. PROPOSED SYSTEM

A. DATA PREPROCESSING

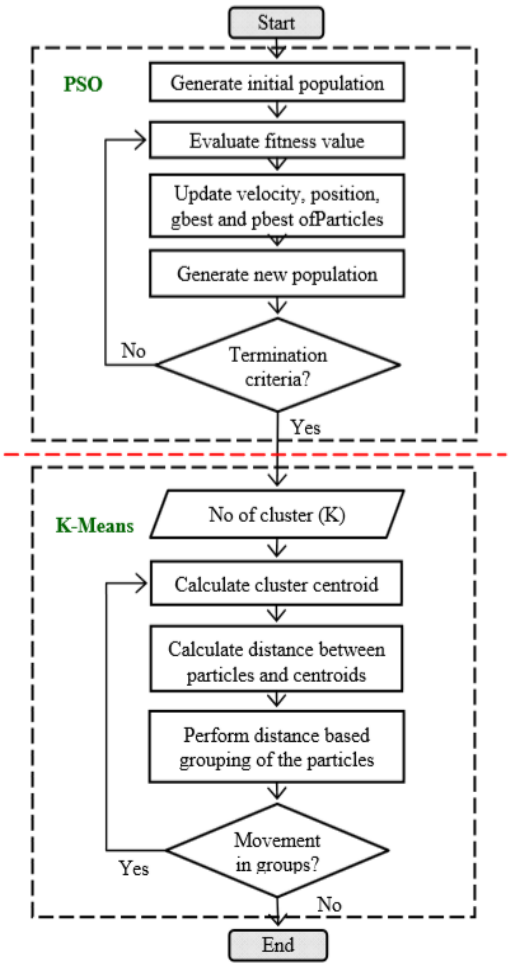
Cleaning and filtering of the data might be necessarily carried out with respect to the data and data mining algorithm employed so as to avoid the creation of deceptive or inappropriate rules or patterns. The steps involved in the pre-processing of a dataset are the removal of duplicate records, normalizing the values used to represent information in the database, accounting for missing data points and removing unneeded data fields. To make data appropriate for the mining process it needs to be transformed. The raw data is changed into data sets with a few appropriate characteristics. Moreover it might be essential to combine the data so as to reduce the number of data sets besides minimizing the memory and processing resources required by the data mining algorithm . This leads to removal of duplicate records and supplying the missing values in the heart disease data warehouse. In addition, it is also transformed to a new form which is appropriate for clustering. Clinical databases have accumulated large quantities of information about patients and their medical conditions. Heart disease is the major cause of casualties in the world. The term Heart disease encompasses the diverse disease that affects the heart. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term ―cardiovascular disease‖ includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Cardiovascular disease (CVD) results in severe illness, disability, and death. 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.

TABLE1.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) :>0.0 and <=80, value (1.0) : >81 and <119,
13	painloc: chest pain location (value 1:substernal; value 0:otherwise)

B. K-MEANS CLUSTERING WITH PSO

The proposed hybridization technique includes two clustering algorithms; first one is PSO and second one is K-Means. Although PSO is a good clustering algorithm, it does not perform well when the dataset is large or complex. PSO is efficient in global search but its local search ability is poor. While K-Means is a good option (fast, robust and easier to understand) for local search ability but it didn't work well with global clusters [11]. Even its performance is un-consistent at different initial partitions, it produce different results at different initial partitions. At the initial stage, the PSO clustering algorithm is executed to search for the location of clusters' centroid. These locations are used as initial centroid for K-means clustering algorithm for refining and generating the optimal clustering solution. This arrangement is not only resolving the limitations of these algorithms but multiplying the advantages of both algorithms as well [12].



This unique combination of PSO and K-Means algorithm will generate the better result compared to the result of both individual algorithms.This algorithm can be better understood by a flowchart given above. PSO algorithm is a probabilistic approach to find the optimal solution. 10 runs are suggested for the termination criteria for PSO, it generates a new optimal solution near around global optimal point at every run. 10 runs are enough for further processing with K-Means to obtain better result . PSO algorithm is used at the initial stage to discover the optimal solution by a global search. The result from PSO is proximity of global solution and it will be used as the initial seed to the K-Means data clustering algorithm for refining and generating the final optimal solution [11]

Steps of our proposed approach is given below-

Step 1. Randomly generate particles(or pick particles from a given dataset) and form a population by grouping these particles.

Step 2. Initialize the position and velocity of particles using equations (2) and (3).

Step 3. Calculate the fitness value based on equation (6).

Step 4. Update the position, velocity, gbest and pbest of particles using equations (4) and (5).

Step 5. Repeat step 3and 4 until one of following termination conditions is satisfied. a. The maximum number of iterations is exceeded. b. The average change in centroid vectors is less than a predefined value.

Step 6. Input the number of clusters (K) to be generated.

Step 7. Initialize cluster centroids for K-Means using the K best position particles ofPSO.

Step 8. Assign each particle of the population to the closest centroid cluster of K-Means.

Step 9. Recalculate the cluster centroid of K-Means using equation (7).

Step 10. Repeating step 8 and 9 until the centroids no longer move.

C. 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.

D. C4.5 Algorithm

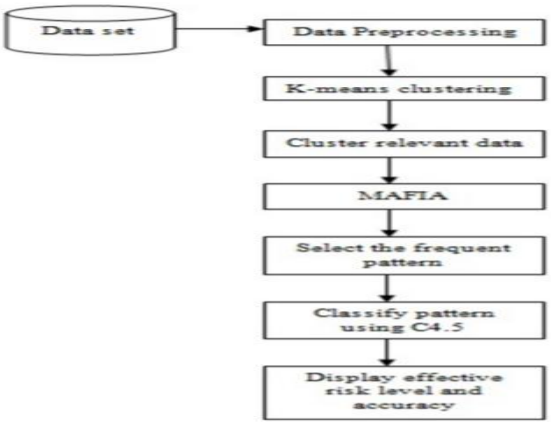
Classification is an unsupervised learning used to predict the class of objects whose class label is unknown. It is used for creating classification rules by means of decision trees from a given data set. Decision tree is used as a prognostic model. C4.5, C5.0, CART, ID3 are methods for building decision 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

A. Sample Algorithm

- Test for base cases.
- For each element n, determine the normalized information gain from separating on n.
  - o Let n_best be the element with the highest normalized information gain
- Construct a decision node that breaks on a best.
- Rescuer on the sub lists found by separating on n_best, and attach these nodes as children of node.

C4.5 DECISION TREE STRUCTURE

If Age=<30 a n d 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 high



Flowchart1: SYSTEM ARCHITECTURE

4. EXPERIMENTAL RESULTS

The result of experimental analysis in identifying important patterns for predicting heart diseases are presented in this section. 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 and PSO with the K value of 2.Then the frequent forms are mined efficiently from the set appropriate to heart disease, using the MAFIA. .

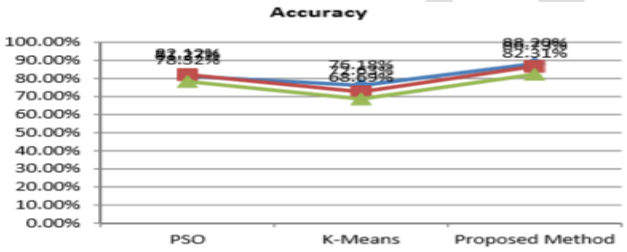


Fig1:Cluster analysis

5. CONCLUSION

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.. A sequential hybridization of two popular data clustering approach (PSO and K-Means) has been proposed drawbacks of K-Means clustering algorithm can be minimized by using PSO over it

ACKNOWLEDGMENT

Expressing my sincere gratitude to great almighty, members of computer science department and family.

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," 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.

[11] Hai Shen ; Li Jin ; Yunlong Zhu ; Zhu Zhu, "Hybridization of particle swarm optimization with the K-Means algorithm for clustering analysis", IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA), 2010.

[12] Yucheng Kao; Szu-Yuan Lee, "Combining K-means and particle swarm optimization for dynamic data clustering problems", IEEE International Conference on Intelligent Computing and Intelligent Systems, 2009.