A Prototype of Heart Disease Risk Level Prediction Model Using An Improved Data Mining Algorithm

Malini K
P.G Scholar
SCMS School of Technology and Management
Cochin, Kerala, India

Rekha Sunny T
Assistant Professor
SCMS School of Technology and Management
Cochin, Kerala, India

Abstract

Cardiovascular diseases have become one of the serious health problems around the world. Data mining techniques can be applied for efficiently predicting heart disease risk levels. One of the major Data mining algorithms that can be used is K-means clustering. K-means clustering suffers from lots of database scan and results in huge cost. In this paper we introduce a prediction model which combines both association rule mining and clustering for the diagnosis of a heart disease which reduces the huge cost incurred in K-means clustering. The maximal frequent term set forms the description for the cluster and hence the number of attributes is reduced. Thus the numbers of tests that are required to be taken by the patients are also reduced.

1. Introduction

Cardiac disease has become one of the major causes of death in the world. This is mainly due to the lack of an efficient heart disease prediction system. Presently doctors are making predictions based on their learning and experience. But mere human intelligence cannot provide an effective prediction result. Due to the developments in information technology, large amounts of medicinal data regarding patients, disease diagnosis, and electronic patient records are available. Medicinal data mining methods can be used to analyze these data and make predictions[5][6]. An efficient prediction model which makes use of the data mining methods such as association rule mining and clustering is introduced herewith.

Data mining is the process of analyzing data from different perspectives and extracting useful knowledge from it. One of the most popular data mining approaches is to find frequent term sets from a transaction dataset and derive association rules. Finding frequent term sets (term sets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent term sets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. Apriori[3] is a seminal algorithm for finding frequent term sets using candidate generation[1].

Clustering is another data mining approach which groups a set of abstract objects into class of similar objects. One of the major clustering methods are K-means clustering method where each cluster is represented by the mean value of the objects in the cluster[2]. K-means algorithm can be applied to medicinal data for predicting diseases in an efficient way. Our proposal is to combine these two data mining methods for an efficient and less expensive prediction model for heart diseases.

The remaining sections of the paper are organized as follows: In Section 2, a brief review of some of the related works is presented. The proposed algorithm is described in Section 3. The experimental results and a prototype for prediction are given in Section 4. The conclusions and future directions are summed up in Section 5.

2. Literature Review

2.1. Apriori

Apriori is an algorithm proposed by R. Agrawal and R Srikant [4] for mining frequent term sets for boolean association rule. The algorithm uses prior knowledge of frequent term set properties. Apriori employs an iterative approach known as level-wise search, where k term set are used to explore (k+1) term sets. There are two steps in each iteration. The first step generates a set of candidate term sets. Then, in the second step we count the occurrence of each candidate set in database and prune all disqualified candidates (i.e. all infrequent term sets). Apriori uses two pruning techniques, first on the bases of support count and second for a term set to

be frequent, all its subset should be in previous frequent term set. The iterations begin with 2 term sets (size as 2) and the size is incremented after each iteration. The algorithm is based on the closure property of frequent term sets: if a set of items is frequent, then all its proper subsets are also frequent [3].

```
Initialize: k := 1, C1 = all the 1- item sets;
read the database to count the support of C1 to
determine L1.
L1 := {frequent 1- item sets};
k:=2; //k represents the pass number//
while (Lk-1 ≠ Φ) do
begin
Ck := gen_candidate_itemsets with the given Lk-1
prune(Ck)
for all transactions t ЄT do
increment the count of all candidates in CK that are
contained in t;
Lk := All candidates in Ck with minimum support ;
k := k + 1;
end
Answer := Ụ k Lk
```

Fig 1. Apriori algorithm [9]

2.2. K-means clustering

K-means algorithm takes the input parameter, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is high but the intercluster similarity is low.cluster similarity is measured with regard to the mean value of the objects in a cluster. K-means algorithm randomly select k of the objects, each of which initially represents a cluster mean or center.For each of the remaining objects, an object is assigned to the cluster to which it is most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster.This process continues until the criterion function converges [2][4].

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Fig 2. K-means algorithm

2.3. Decision Trees

Decision Tree is one of the data mining techniques used in the diagnosis of heart disease. Andreeva used C4.5 Decision Tree in the diagnosis of heart disease. Sitair-Taut et al. introduced Naive Bayes and J4.8 Decision Trees for detecting coronary heart disease. Mai Shouman, Tim Turner and Rob Stocker proposed an alternative decision tree with better performance [8].

2.4. Integration of Decision Trees and K-Means

An algorithm which integrates K-means clustering with decision tree is introduced in paper [7]. Enhancements by introducing different initial centroid selection add more accuracy in diagnosing heart disease patients.

3. Proposed Algorithm

The proposed algorithm is to develop a risk level prediction model by applying data mining techniques such as frequent term set mining and clustering. Architecture of the proposed system is as depicted in the following figure.

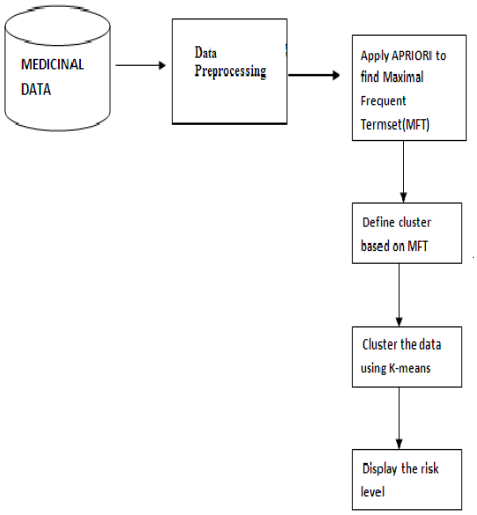

Fig3. Proposed architecture

Initially data preprocessing techniques are applied to the available medicinal dataset which includes dimensionality reduction for reducing the dimensions of the dataset.

Next step is finding the maximal frequent term set (MFT) using Apriori algorithm. A frequent term set is said to be maximal if it is not a subset of any other frequent term set. Now MFT forms the cluster description. K-means clustering algorithm is applied and the input is assigned to the closest cluster. Finally the risk level is displayed based on the cluster result.

3.1. Finding MFT

By applying Apriori algorithm to the medicinal data set, frequent term sets are found out from which the maximal frequent term set is generated. The maximal frequent term set is the superset of all other frequent term sets.

3.2. Cluster Definition based on MFT

Once the maximal frequent term set is found, it is the MFT that defines cluster. So only the attributes that are part of MFT needs to be considered while defining the cluster. Also the new data for which prediction is to be done should be represented in terms of MFT for clustering.

3.3. Risk Level Prediction

Finally we need to predict the risk level based on the cluster result. Risk levels are chosen as either low, or medium or high. Similarity with the three clusters is calculated and risk level for the input is set as that of the closest cluster.

4. Experimental Results

The results of our experimental analysis in predicting the risk level for heart disease are presented in this section. We have implemented our proposed approach in Java. The heart attack dataset is been used for our experiments.

We have used medicinal dataset of small size (100) for the experimental analysis. After preprocessing phase where dimensionality reduction was done, maximal frequent term set was found out by applying Apriori algorithm. Attributes that comprises the maximal frequent term set (generated using Apriori) are as listed in the following table.

Sl no:	Attribute	Description
1	P_age	Age in years
2	Gender	Male or Female
3	Smoking	Smoking type like past,current,never
4	Overweight	Overweight type like yes,no
5	Hereditary	Values like yes,no
6	bad_cholestrol_level	Values like very high,high,normal
7	blood_sugar_level	Values like high, normal, low

8	alcohol_intake	Values like never, past, current
9	high-salt_diet	Values as yes,no
10	bloodpressure	Values like, normal, low, high
11	sedentry_life_style	Values like yes,no
12	exercise_habit	Exercise habits like never,regular,high
13	Heart_rate	Heart rates as normal, low, high
14	high_saturated _fat_diet	Values like yes,no
15	risklevel	Low,Medium,High

Table 1.Heart Disease Parameters

The clusters were defined based on the attributes listed in the table given above. After clustering using K-means, risk levels were predicted.The samples of heart attack risk level prediction (Low, Medium, and High) are as below.

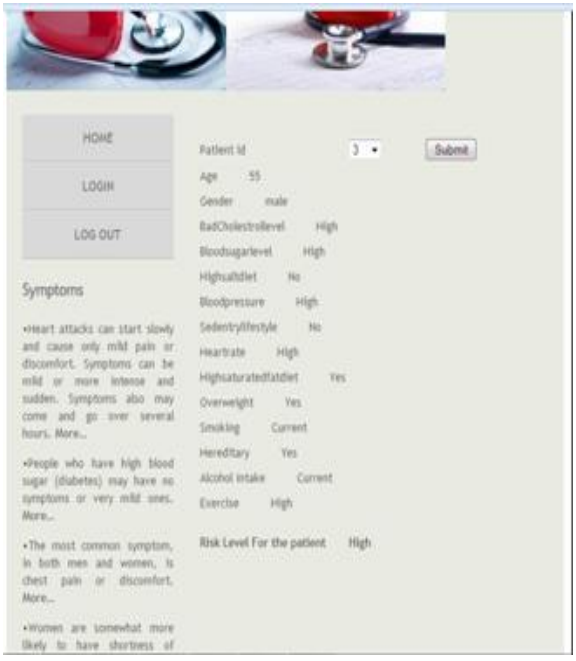

Fig. 4. Sample of prediction for risk level: high

Fig. 5. Sample of prediction for risk level: medium

Fig. 6. Sample of prediction for risk level: low

5. Conclusion and Future Directions

In this paper, we have presented a Heart Disease Diagnosis System using data mining techniques. Apriori algorithm is used to find the frequent term sets and the maximal frequent term set is generated. Clustering is performed using K-means clustering algorithm. Defining the clusters based on maximal frequent term sets provided improved accuracy and

less diagnosis cost. Incorporation of optimization techniques further improves the accuracy.

6. Acknowledgements

The authors gratefully acknowledge the insights from all the supporters and reviewers of this paper.

7. References

[1] Lakshmi, K.R, M. Veera Krishna, and S. Prem Kumar, "Performance Comparison of data Mining Techniques for Predicting of Heart Disease Survivability". *International Journal of Scientific and Research Publications*, Volume 3,Issue 6,June 2013 1 ISSN 2250-3153

[2] Han,Jiawei,Micheline Kamber, *Data Mining:Concepts and Techniques*, The Morgan Kaufmann Series,2001.

[3] Bodon.F., "A Fast Apriori Implementation",*FIMI'03*,November 2003.

[4] Agarwal. R,Imielimski T, Swami A, "Mining Association Rules between sets of Items in Large databases",*SIGMOD'93*,pp.207-216,1993.

[5] Jyoti Soni,Sunitha Soni , "Predictive data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction"; *International Journal of Computer Applications*(0975-8887) Volume 17-No.8,March 2011.

[6] Fariba Shadabi,Dharmendra Sharma, "Artificial Intelligence and Data Mining Techniques in Medicine-Success Stories", *International Conference on BioMedical Engineering and Informatics*, vol. 1, pp.235-239, 2008

[7] Shouman,Mai, Tim Turner, and Rob Stocker. "Using decision tree for diagnosing heart disease patients".*Proceedings of the Ninth Australasian Data Mining Conference*-Volume 121.Australia Computer Society, Inc 2011.

[8] Asha rajkumar, G.Sophia Reena,-"Diagnosis of Heart Disease using Datamining Algorithm*",Global Journal of Computer Science and Technology*,page 38 Vol 10 Issue 10 ver.1.0 September,2010.

[9] Anshu,Chaturvedi,and C.S. Raghuvanshi. "An Algorithm for Frequent pattern Mining based On Apriori".*International Journal on Computer Science and Engineering* Vol.02,No 04,2010,942-947.