Prediction of Significant Heart Attack Patterns Using Clustering Algorithm

Aye Mya Han (5MCS-163)
University of Computer Studies, Yangon
AyeMyaHan23@gmail.com

Abstract

This system presents an efficient approach for discovering significant patterns from the heart disease database for heart attack prediction. The heart disease data warehouse is clustered using K-means clustering algorithm to extract related data. The primary intent of the system is to design and develop an efficient approach for extracting patterns, which are significant to heart attack, from the heart disease database. The diagnosis of diseases is a significant and tedious task in medicine. The detection of heart disease from various factors or symptoms is a multi-layered issue which is not free from false presumptions often accompanied by unpredictable effects. Thus the effort to utilize knowledge and experience of numerous specialists and clinical screening data of patients collected in databases to facilitate the diagnosis process is considered a valuable option. The proposed system aims to utilize the data mining techniques: clustering and frequent pattern mining.

Keywords : Clustering Algorithm, Heart Attack, Frequent pattern mining, Data Mining, Disease Diagnosis, Heart Disease, Pre-processing, Frequent Patterns, MAFIA (MAximal Frequent Itemset Algorithm), Clustering, K-Means, Significant Patterns.

1. Introduction

Data mining is an emerging area of computational intelligence that offers the new theories, technique and tools for analysis of large data sets. It is applied in many areas in order to extract useful patterns out of large amount of data. There are several applications of data mining in the real world. They include association, sequence or path analysis, classification, clustering and forecasting. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. Clustering algorithm depends both on the type of data available and on the particular purpose and application. Major clustering methods can be classified as: partitioning methods, hierarchical methods, density-based methods, grid-based methods and model-based methods. Among these, K-means clustering is one of the partitioning methods.

The diagnosis of diseases is a significant and tedious task in medical field. The detection of heart disease from various factors or symptoms is a quite complex and it is not free from false presumptions. It can also be accompanied by unpredictable effects. Therefore to apply the knowledge and experience of numerous specialists and clinical screening data of heart disease patients should be collected in the databases. Based on those data, making a diagnosis process is very much valuable. This system applies frequent pattern mining to the heart disease database and generate prediction of heart disease patterns.

2.Related Work

A novel technique to develop the multi-parametric feature with linear and nonlinear characteristics of HRV (Heart Rate Variability) was proposed by Heon Gyu Lee et al. [2]. Statistical and classification techniques were utilized to develop the multi-parametric feature of HRV. Besides, they have assessed the linear and the non-linear properties of HRV for three recumbent positions, to be precise the supine, left lateral and right lateral position. Numerous experiments were conducted by them on linear and nonlinear characteristics of HRV indices to assess several classifiers, e.g., Bayesian classifiers [5], CMAR (Classification based on Multiple Association Rules) [4], C4.5 (Decision Tree) [6] and SVM (Support Vector Machine) [3]. SVM surmounted the other classifiers.

A model Intelligent Heart Disease Prediction System (IHDPS) built with the aid of data mining techniques like Decision Trees, Naïve Bayes and Neural Network was proposed by Sellappan Palaniappan et al. [1]. The results illustrated the peculiar strength of each of the methodologies in comprehending the objectives of the specified mining objectives. IHDPS was capable of answering queries that the conventional decision support systems were not able to. It facilitated the establishment of vital knowledge, e.g. patterns, relationships amid medical factors connected with heart disease. IHDPS subsists well being web-based, user-friendly, scalable, reliable and expandable.

A novel heuristic for efficient computation of sparse kernel in SUPANOVA was proposed by Boleslaw Szymanski et al. [6]. It was applied to a

benchmark Boston housing market dataset and to socially significant issue of enhancing the detection of heart diseases in the population with the aid of a novel, non-invasive measurement of the heart activities on basis of magnetic field generated by the human heart. 83.7% predictions on the results were correct thereby outperforming the results obtained through Support Vector Machine and equivalent kernels. The spline kernel yielded equally good results on the benchmark Boston housing market dataset.

2.1. Background Theory

The k-means algorithm [7] is one of the widely recognized clustering tools that are applied in a variety of scientific and industrial applications. K-means groups the data in accordance with their characteristic values into K distinct clusters. Data categorized into the same cluster have identical feature values. K, the positive integer denoting the number of clusters, needs to be provided in advance.

2.2. Clustering Using K-means Algorithm

Clustering is unsupervised learning, that is the machine/software will learn on its own, using the data(learning set) and will classify the objects into a particular class.The steps involved in a K-means algorithm are given subsequently:

1. K points denoting the data to be clustering are placed into the space. These points denote the primary group centroids.
2. The data are assigned to the group that is adjacent to the centroid.
3. The position of all the K centroids are recalculated as soon as all the data are assigned.
4. Step 2 and 3 are reiterated until the centroids stop moving any further.
5. The preprocessed data warehouse is clustered with k value as 2.
6. One cluster consists of the data relevant to its application and the other contains the remaining data.

2.3. Frequent Pattern Mining

Frequent Item Mining is considered to be one of the elemental data mining problems that intends to discover groups of items or patterns that occur frequently in a dataset. It is of vital significance in a variety of Data Mining tasks that aim to mine interesting patterns from databases. The cluster that contains data most relevant to heart attack is fed as input to mine the frequent patterns present in it.

2.4. Data Preprocessing

Cleaning and filtering of the data is carried out to avoid the creation of deceptive or inappropriate rules or patterns.[8] The actions comprised in the pre-processing are

1. the removal of duplicate records,
2. normalizing the values used to represent information in the database,
3. accounting for missing data points
4. removing unneeded data fields and
5. appropriate formatting data for clustering.

3. Overview of the proposed system

The system will be used to predict significant heart attack patterns using clustering and maximal frequent itemset mining. The heart disease database contains the screening clinical data of heart patients that is shown in the heart disease dataset description. Initially, the database is preprocessed to make the mining process more efficient. In the first step, the preprocessed database is then clustered using the K-means clustering algorithm with K=2. This result in two clusters, one contains the data that are relevant to heart attack and the other contains the remaining data. In the second step, the frequent patterns are mined from the data, relevant to heart attack, using the MAFIA (MAximal Frequent Itemset Algorithm) algorithm. Finally, the significant weightage is calculated for all frequent patterns with the aid of the approach proposed. The frequent patterns with significant weightage greater than a predefined threshold are chosen. These chosen significant patterns can be used in the design and development of heart attack prediction system.

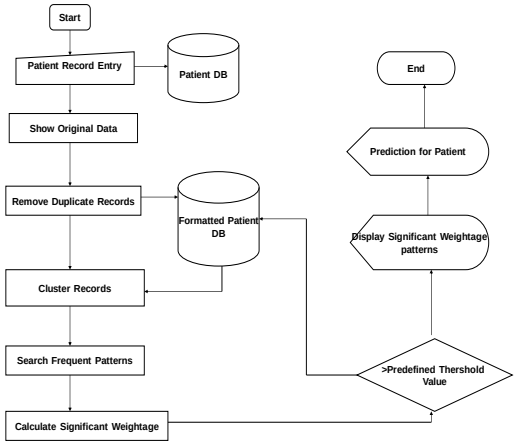
3.1. Main Algorithm

1.Begin
2.Collect the data
3.Cluster the data using K-mean
4.Search Frequent Patterns using MAFIA
5.Input *Threshold* value
6.For each *pattern* in Frequent Patterns
7.Calculate the *Significant weightage of the pattern*
8.If *Significant_Weightage_of_the_pattern* > *Threshold* Then
9.Display the *pattern*
10.End if
11.End For
12.End

Pseudo Code For the Main Algorithm

The system firstly collect the patient records and preprocess the data to invoke clustering algorithm. After preprocessing, k-means clustering

algorithm is used to divide two groups, the important cluster and the non-important cluster. The important cluster is used to search the frequent item attributes that occur frequently in the heart disease database. Then, the frequent patterns are calculated according to their weightage. Finally, the significant patterns that are greater than the predefined threshold value are generated and display the user.



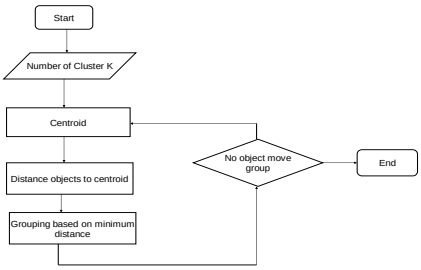
Flowchart of the system

3.3. Euclidean Distance Equation

$$d(x,y)=\sqrt{(x_i-y_i)^2+(x_{xi+1}-y_{i+1})^2+\ldots+(x_{i+n}-y_{i+n})^2} \quad (1)$$

The distance of the records from centriod is calculated by using equation(1). After the distance calculation, the mean valued is calculated using the following equation(2).

$$\text{Average mean} = \frac{\text{the total distance of all records}}{\text{the number of records}} \quad (2)$$



Flowchart for k-means Clustering Algorithm

3.4. Maximal Frequent Itemset Algorithm

MAFIA(C, MFI, Boolean IsHUT)
{
1.name HUT = C.head $\cup$ C.tail;
2.if HUT is in MFI
stop generation of children and return
3.For each item i in C, trimmed_tail{
4.IsHUT = whether i is the first item in the tail
5.newNode = C $\cup$ I
6.MAFIA (newNode, MFI, IsHUT)}
7.if (IsHUT and all extensions are frequent)
8.Stop search and go back up subtree
9.If (C is a leaf and C.head is not in MFI)
10.Add C.head to MFI
}

Pseudo code for MAFIA

3.5. Significant Weightage Calculation

The significant weightage of each pattern is calculated on each attribute present in the pattern and the frequency of each pattern.
The formula used to determine the significant weightage (Sw) is

$$S_{W2} = \sum_{t=1}^{n} W_t f_t$$

(3)
W_i = the weight of each attribute
f_i = the frequency of the pattern
n = the total number of pattern

To aid the prediction of heart attack

$$SFP = \{x : S_w(x) \geq \Phi\}$$

(4)
SFP = significant frequent patterns and represents the significant weightage

4.Implementation of the Proposed System

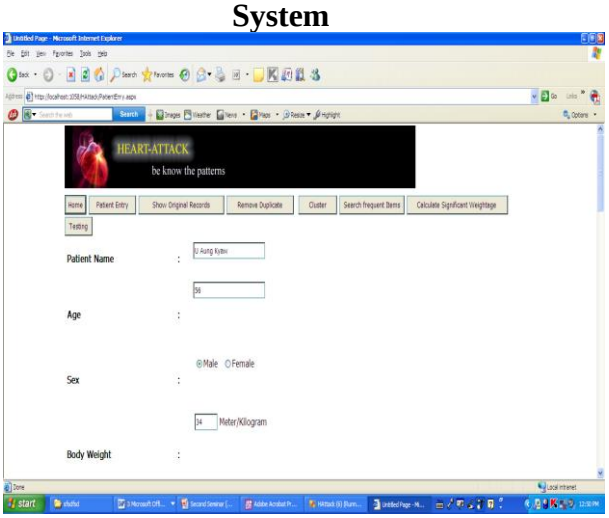


Figure.1

In the system, patient records can be inserted using

the above user interface.

Figure.2

The patient records in the heart disease database are showed in the Figure.2.

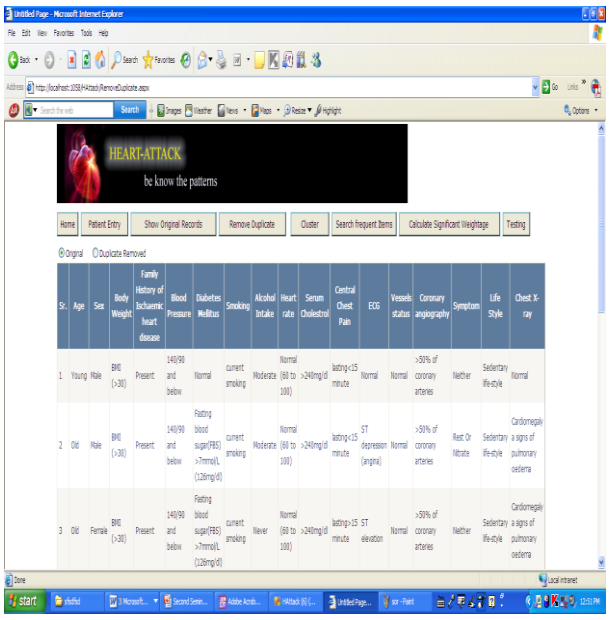


Figure.3

Before applying the data mining algorithm, the data set are preprocessed and shown in Figure.3.

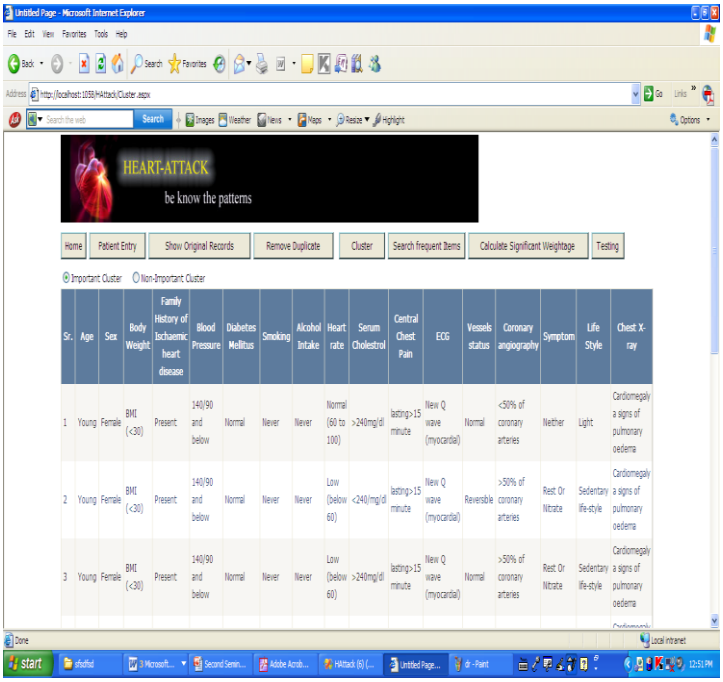


Figure.4

In the Figure.4, the dataset is clustered using the k-means clustering algorithm and the system can display important and non-important clusters.

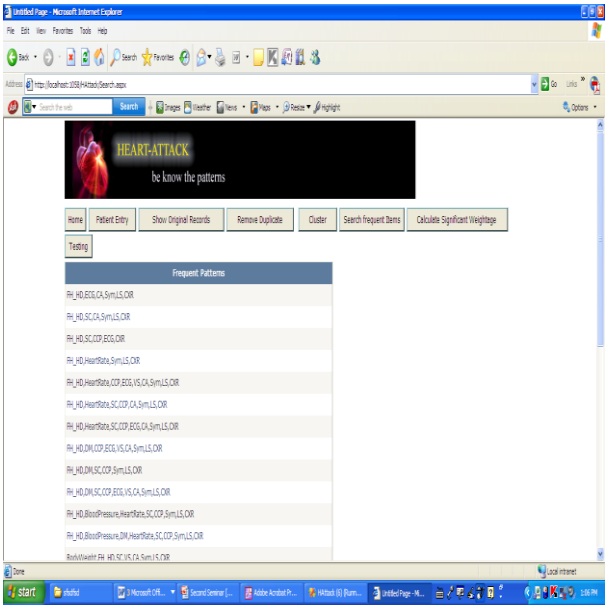


Figure.5

From the important cluster, the frequent items sets are searched using the maximal frequent itemset algorithm.

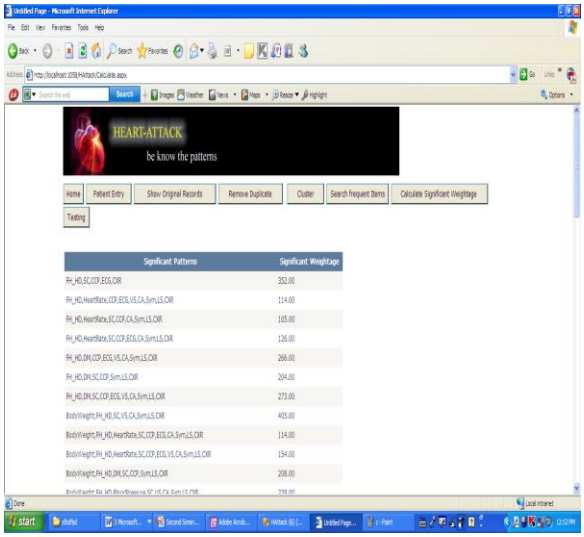


Figure.6

The frequent items are calculated using the significant weightage calculation, equation (3). Finally, the frequent patterns that are greater than the predefined threshold value are generated.

4.1 Evaluation

Accuracy can be measured by sensitivity and specificity.

$$\text{sensitivity} = \frac{t - pos}{pos} = \frac{number\ of\ true\ positive}{number\ of\ positive} \quad (5)$$

$$\text{specificity} = \frac{t - neg}{neg} \quad \frac{number\ of\ true\ negative}{number\ of\ negative} \quad (6)$$

$$\text{Accuracy} = \left(sensitivity \frac{pos}{(pos + neg)} + specificity \frac{neg}{(pos + neg)}\right) \quad (7)$$

true positive = a positive instance that is correctly classified

true negative = a positive instance that is incorrectly classified

false positive = a negative instance that is correctly classified

false negative = a negative instance that is incorrectly classified

5. Conclusion

Health care related data are voluminous in nature and they arrive from diverse sources all of them not entirely appropriate in structure or quality. The system presents an efficient approach for extracting significant patterns from the heart disease data warehouses for the efficient prediction of heart attack. Extracting the data are carried out by using K-means clustering algorithm and the frequent patterns for the valuable heart attack prediction. The system will help medical professional and patients to know significant facts that could lead to heart attack.

References

[1] Sellappan Palaniappan, Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008.

[2] Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, "Mining Biosignal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV," LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, May 2007.

[3] Cristianini, N., Shawe-Taylor, J.: An introduction to Support Vector Machines. Cambridge University Press, Cambridge , 2000.

[4] Li, W., Han, J., Pei, J.: CMAR: Accurate and Efficient Classification Based on Multiple Association Rules. In: Proc. of 2001 Interna'l Conference on Data Mining .2001.

[5] Chen, J., Greiner, R.: Comparing Bayesian Network Classifiers. In Proc. of UAI-99, pp.101–108 ,1999.

[6] Boleslaw Szymanski, Long Han, Mark Embrechts, Alexander Ross, Karsten Sternickel, Lijuan Zhu, "Using Efficient Supanova Kernel For Heart Disease Diagnosis", proc. ANNIE 06, intelligent engineering systems through artificial neural networks, vol. 16, pp:305-310, 2006.

[7] C. Ordonez, "Programming the K-Means Clustering Algorithm in SQL," Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining, pp. 823-828, 2004.

[8] Gerhard M¨unz, Sa Li, Georg Carle. "Traffic anomaly detection using k-means clustering *GI/ITG-Workshop MMBnet 2007, Hamburg, Germany*,September2007