Data Mining based Fragmentation and Prediction of Medical Data

Hnin Wint Khaing
University of Computer Studies
Mandalay, Myanmar
snow.hwk@gmail.com

Abstract— **Data mining concerns theories, methodologies, and in particular, computer systems for knowledge extraction or mining from large amounts of data. Association rule mining is a general purpose rule discovery scheme. It has been widely used for discovering rules in medical applications. The diagnosis of diseases is a significant and tedious task in medicine. The detection of heart disease from various factors or symptoms is an 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. In this paper, we presented an efficient approach for the prediction of heart attack risk levels from the heart disease database. Firstly, the heart disease database is clustered using the K-means clustering algorithm, which will extract the data relevant to heart attack from the database. This approach allows mastering the number of fragments through its k parameter. Subsequently the frequent patterns are mined from the extracted data, relevant to heart disease, using the MAFIA (Maximal Frequent Itemset Algorithm) algorithm. The machine learning algorithm is trained with the selected significant patterns for the effective prediction of heart attack. We have employed the ID3 algorithm as the training algorithm to show level of heart attack with the decision tree. The results showed that the designed prediction system is capable of predicting the heart attack effectively.**

Keywords-Data mining; Heart Disease; Frequent Patterns; MAFIA(Maximal Frequent Itemset Algorithm); ID3 Algorithm

I. INTRODUCTION

Hospitals and clinics accumulate a huge amount of patient data over the years. These data provide a basis for the analysis of risk factors for many diseases. For example, we can predict the level of heart attack to find patterns associated with heart disease. One of the major topics in data mining research is the discovery of interesting patterns in data. From the introduction of frequent itemset mining and association rules, the pattern explosion was acknowledged: at high frequency thresholds only common knowledge is revealed, while at low thresholds prohibitively many patterns are returned. A majority of areas related to medical services such as prediction of effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among clinical and diagnosis data also make use of Data Mining methodologies [1]. The effectiveness of medical treatments can be estimated by developing the data mining applications. Data mining is capable of delivering an analysis of which courses of action prove effective [2], achieved by comparing and contrasting causes, symptoms, and courses of treatments. In this paper, we presented prediction of the risk levels of heart attack from the heart disease database. The heart disease database consists of mixed attributes containing both the numerical and categorical data. These records are cleaned and filtered with the intention that the irrelevant data from the database would be removed before mining process occurs. Then clustering is performed on the preprocessed data warehouse using K-means clustering algorithm with K value so as to extract data relevant to heart attack. Subsequently the frequent patterns significant to heart disease diagnosis are mined from the extracted data using the MAFIA algorithm. Finally, we used the ID3 as training algorithm to show the effective risk level with decision tree. The remaining sections of the paper are organized as follows: In Section 2, a brief review of some of the works on heart disease diagnosis is presented. The fragmentation of risk level from heart disease database and extraction of significant patterns from heart disease database is presented detailed in Section 3. Architecture of the system is depicted in section 4. Experimental results are specified in section 5.The conclusion and future works are described in Section 6.

II. RELATED WORK

Tremendous works in literature related with heart disease diagnosis using data mining techniques have motivated our work. The researchers in the medical field diagnose and predict the diseases in addition to providing effective care for patients [3] by employing the data mining techniques. The data mining techniques have been employed by numerous works in the literature to diagnose diverse diseases, for instance: Diabetes, Hepatitis, Cancer, Heart diseases and more [4]. 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. [5].The problem of identifying constrained association rules for heart disease prediction was studied by Carlos Ordonez [6]. The assessed data set encompassed medical records of people having heart disease with attributes for risk factors, heart perfusion measurements and artery narrowing.

978-1-61284-840-2/11/$26.00 ©2011 IEEE

Association rule mining is a major data mining technique, and is a most commonly used pattern discovery method. It retrieves all frequent patterns in a data set and forms interesting rules among frequent patterns. Most frequently used association rule mining methods are Apriority and FP-growth [7]. Frequent Itemset Mining (FIM) is considered to be one of the elemental data mining problems that intends to discover groups of items or values or patterns that co-occur frequently in a dataset [8].The term Heart disease encompasses the diverse diseases that affect the heart. Heart disease was then major cause of casualties in the United States, England, Canada and Wales as in 2007. Heart disease kills one person every 34 seconds in the United States [9]. 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 [10].

III. THEORITICAL BACKGROND

A. Fragmentation of heart attack data using K_mean clustering

The wide acceptance of the relational approach in data-processing applications and the continuous improvement of commercially existing relational databases has increased the interest for using database in non-conventional applications, such as computer-aided design (CAD), geographic information systems, image, and graphic database systems [10]. The relational approach distinguishes two kinds of fragmentation: horizontal and vertical. There are many algorithms developed for horizontal and vertical fragmentation [11]. In this paper, we proposed k-mean clustering based fragmentation. Clustering is the process to divide a data set into several classes or clusters, and the same data objects within a group are of higher similarity while data objects in different groups are of lower similarity. Clustering medical data into small yet meaningful clusters can aid in the discovery of patterns by supporting the extraction of numerous appropriate features from each of the clusters thereby introducing structure into the data and aiding the application of conventional data mining techniques. In this paper, we proposed the most popular distance measure, ***Euclidean distance***, which is defined as

$$\text{dist}(i, j) = \| x_i - x_j \|$$

$$= \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots (x_{ip} - x_{jp})^2}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p_dimensional data objects.

The steps involved in a K-means algorithm:

1. x_i points denoting the data to be clustered 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 positions of all the x_j centroids are recalculated as soon as all the data are assigned.

Steps 2 and 3 are reiterated until the centroids stop moving any further. This results in the separation of data into groups from which the metric to be minimized can be deliberated. The preprocessed heart disease database is clustered using the K-means algorithm with K value as 2. One cluster consists of the data relevant to the heart disease and the other contains the remaining data. Later on, the frequent patterns are mined from the cluster relevant to heart disease, using the MAFIA algorithm.

B. Frequent pattern mining using MAFIA

Frequent Itemset Mining (FIM) is considered to be one of the elemental data mining problems that intends to discover groups of items or values or patterns that co-occur frequently in a dataset. The extraction of significant patterns from the heart disease data warehouse is presented in this section. The heart disease database contains the screening clinical data of heart patients. Frequent Itemset Mining (FIM) is considered to be one of the elemental data mining problems that intends to discover groups of items or values or patterns that co-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, like association rules, correlations, sequences, episodes, classifiers, clusters and the like. The proposed approach utilizes an efficient algorithm called MAFIA (Maximal Frequent Itemset Algorithm) which combines diverse old and new algorithmic ideas to form a practical algorithm [8]. In this paper, we presented the MFPA (Maximal Frequent Pattern Algorithm) algorithm for the frequent patterns applicable to heart disease are mined from the data extracted.

MFPA algorithm:

Pseudo code for MAFIA :

MAFIA(**C**, **MFI**, Boolean **IsHUT**) {
name **HUT** = **C**.head $\cup$ **C**.tail;
if **HUT** is in **MFI**
Stop generation of children and return
Count all children; use **PEP** to trim the tail, and recorder by increasing support,
For each item **i** in **C**, trimmed_tail {
IsHUT = whether **i** is the first item in the tail
newNode = **C I**
MAFIA (newNode, **MFI, IsHUT**)}
if (**IsHUT** and all extensions are frequent)
Stop search and go back up subtree
If (**C** is a leaf and **C.**head is not in **MFI**)

Add **C**.head to **MFI**
}
The cluster that contains data most relevant to heart attack is fed as input to MAFIA algorithm to mine the frequent patterns present in it.

C. Decision Tree Representation

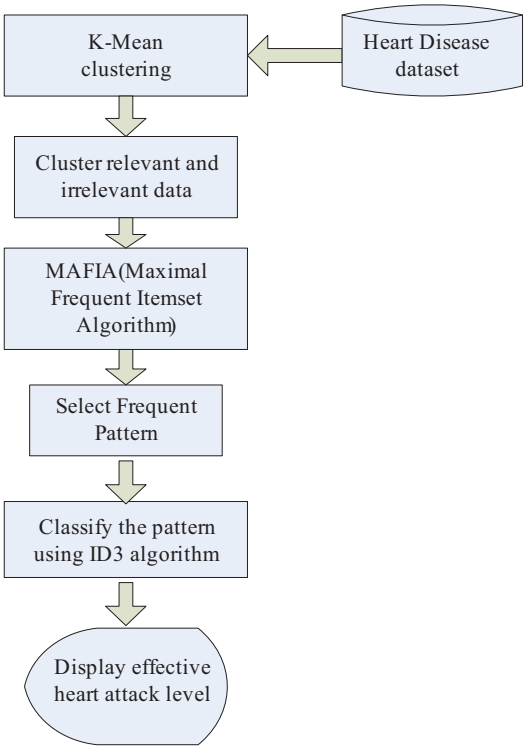
Decision trees are a popular logical method for classification. A decision tree is a hierarchical structure that partitions data into some disjoint groups based on their different attribute values. Leafs of a decision tree contain records of one or nearly one class, and so it has been used for classification. An advantage of decision tree methods is that decision trees can be converted into understandable rules. A most widely used decision tree system is C4.5, its ancestor ID3, and a commercial version C5.0 .Decision trees have been mainly used to build diagnosis models for medical data. When it is used for exploring patterns in medical data, work in shows that it is inadequate for such exploration [1]. In this paper, we used the ID3 algorithm, uses the information gain measure to select among the candidate attributes at each step while growing the tree. Information gain, is simply the expected reduction in entropy caused by portioning the examples according to this attribute. The information gain, *Gain (S , A)* of an attribute A, relative to a collection of example S, is defined as,

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \qquad (1)$$

$$Entropy(S) = \sum_{i=1}^{c} p_i \log_2 p_i \qquad (2)$$

where p_i is the proportion of S belonging to class i.

IV. SYSTEM ARCHITECTURE



V. EXPERIMENTAL RESULTS

The results of our experimental analysis in finding significant patterns for heart attack prediction are presented in this section. With the help of the database, the patterns significant to the heart attack prediction are extracted using the approach discussed. The heart disease database is preprocessed successfully by removing duplicate records and supplying missing values as shown in table I. The refined heart disease data set, resultant from preprocessing, is then clustered using K-means algorithm with K value as 2. One cluster consists of the data relevant to the heart disease as shown in table II and the other contains the remaining data. Then the frequent patterns are mined efficiently from the cluster relevant to heart disease, using the MAFIA algorithm. The sample combinations of heart attack parameters for normal and risk level along with their values and levels are detailed below. In that, lesser value (0.1) of weight comprises the normal level of prediction and higher values other than 0.1 comprise the higher risk levels. Table III show the parameters of heart attack prediction with corresponding values and their levels. Table IV show the example of training data to predict the heart attack level and then figure 2 shows the efficient heart attack level with tree using the ID3 by information gain.

TABLE I. HEART DISEASE DATA SET

ID	Attribute
1	Age
2	Sex
3	painloc: chest pain location
4	relrest
5	cp: chest pain type
6	trestbps: resting blood pressure
7	chol: serum cholesterol in mg/dl
8	smoke
9	cigs (cigarettes per day)
10	years (number of years as a smoker)
11	fbs: (fasting blood sugar > 120 mg/dl)
12	dm (1 = history of diabetes; 0 = no such history)
13	famhist: family history of coronary artery disease
. . .	. . .
38	exeref: exercise radinalid (sp?) ejection
39	exerwm: exercise wall (sp?) motion
40	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
41	Cmo: month of cardiac cath (sp?) (perhaps "call")
42	Cday: day of cardiac cath (sp?)
43	Cyr: year of cardiac cath (sp?)
44	num: diagnosis of heart disease (angiographic disease status)
. . .	. . .

TABLE II. CLUSTERED RELEVANT DATA BASED ON HEART DISEASE DATASET

ID	Reference ID	Attribute
1	#1	Age
2	#2	Sex
3	#5	overweight
4	#6	trestbps: resting blood pressure
5	#7	chol: serum cholestoral in mg/dl
6	#11	fbs: (fasting blood sugar > 120 mg/dl)
7	#14	Alchol intake
8	#27	thalach: maximum heart rate achieved
9	#32	exang: exercise induced angina
10	#34	Sedentary Lifestyle/inactivity
11	#35	slope: the slope of the peak exercise ST segment
12	#37	ca: number of major vessels (0-3) colored by flourosopy
13	#40	Hereditary
14	#44	num: diagnosis of heart disease

TABLE III. HEART ATTACK PARAMETERS WITH CORRESPONDING VALUES AND THEIR WEIGHT

Parameter	Weight	Risk level
Male and Female	Age<30 Age>30	0.1 0.8
Smoking	Never Past Current	0.1 0.3 0.6
Overweight	Yes No	0.8 0.1
Alcohol Intake	Never Past Current	0.1 0.3 0.6
High Salt Diet	Yes No	0.9 0.1
High saturated diet	Yes No	0.9 0.1
Exercise	Regular Never	0.1 0.6
Sedentary Lifestyle/inactivity	Yes No	0.7 0.1
Hereditary	Yes No	0.7 0.1
Bad cholesterol	High Normal	0.8 0.1
Blood Pressure	Normal (130/89) Low (< 119/79) High (>200/160)	0.1 0.8 0.9
Blood sugar	High (>120&<400) Normal (>90&<120) Low (<90)	0.5 0.1 0.4
Heart Rate	Low (< 60bpm) Normal (60 to 100) High (>100bpm)	0.9 0.1 0.9

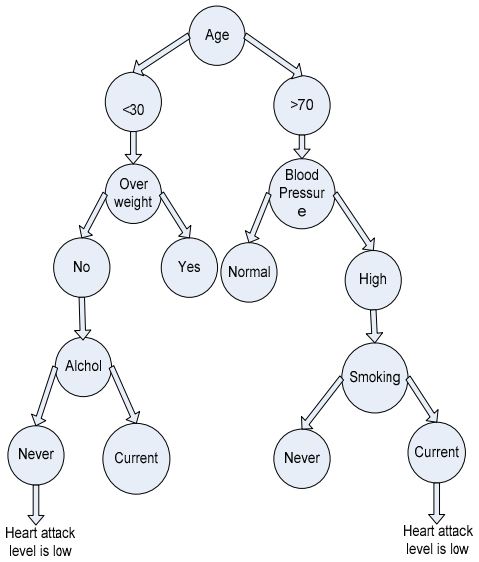


Figure 2: A decision tree for the concept heart attack level by information gain (ID3)

If

Age=<30 and Overweight=no and Alcohol Intake=never

Then

Heart attack level is Low

(Or)

If

Age=>70 and Blood pressure=High and Smoking=current

Then

Heart attack level is High

Figure 3: An example pattern of proposed system

The experimental results of our approach as presented in Table V. The goal is to have high accuracy, besides high precision and recall metrics. These metrics can be derived from the confusion matrix and can be easily converted to true-positive (TP) and false-positive (FP) metrics.

$$precision = \frac{TP}{TP + FP} \ , recall = \frac{TP}{TP + FN}$$

1. True Positive (TP): Total percentage of members classified as Class A belongs to Class A.
2. False Positive (FP): Total percentage of members of Class A but does not belong to Class A.
3. False Negative (FN): Total percentage of members of Class A incorrectly classified as not belonging to Class A.
4. True Negative (TN): Total percentage of members which do not belong to Class A are classified not a part of Class A. It can also be given as (100% - FP).

TABLE IV. CLASSIFY THE LEVELBASED ON TABLE

ID	Age	Smok-ing	Over weight	Alcohol Intake	. . .	Heart rate	Blood Pressure	Class
1	<30	Never	No	Never	. . .	Normal	Normal	Low
2	>50 & <70	Current	No	Past	. . .	Normal	Normal	Low
3	>30 & <50	Current	Yes	Current	. . .	Low	Low	High
4	>70	Never	Yes	Past	. . .	High	High	High
5	>30 & <50	Past	Yes	Past		Normal	High	Low

TABLE V. COMPARISON BETWEEN SIMPLE MAFIA AND PROPOSED K-MEAN BASED MAFIA

Technique	Precision	Recall	Accuracy (%)
K-mean based MAFIA	0.78	0.67	74%
K-mean based MAFIA with ID3	0.80	0.85	85%

VI. CONCLUSION AND FUTURE WORK

Health care related data are voluminous in nature and they arrive from diverse sources all of them not entirely appropriate in structure or quality. These days, the exploitation of knowledge and experience of numerous specialists and clinical screening data of patients gathered in a database during the diagnosis procedure, has been widely recognized. In this paper we have presented an efficient approach for fragmenting and extracting significant patterns from the heart disease data warehouses for the efficient prediction of heart attack. In our future work, we have

planned to design and develop an efficient heart attack prediction system with the aid of xml data using XParser and XQuery language.

REFERENCES

[1] Tzung-I Tang, Gang Zheng, Yalou Huang, Guangfu Shu, Pengtao Wang, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and System Reconstruction Analysis", IEMS Vol. 4, No. 1, pp. 102-108,June 2005.

[2] Hian Chye Koh and Gerald Tan ,"Data Mining Applications in Healthcare", Journal of healthcare information management, Vol. 19, Issue 2, Pages 64-72, 2005.

[3] S Stilou, P D Bamidis, N Maglaveras, C Pappas , “Mining association rules from clinical databases: an intelligent diagnostic process in healthcare” .

[4] A. Bellaachia and Erhan Guven, “Predicting Breast Cancer Survivability using Data Mining Techniques”, the Sixth SIAM International Conference on Data Mining (SDM 2006), Saturday, April 22, 2006.

[5] 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.[8] Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," Seminar Presentation at University of Tokyo, 2004.

[6] Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules,"Seminar Presentation at University of Tokyo, 2004.

[7] Agrawal, R., Imielinski, T. and Swami, A, ‘Mining association rules between sets of items in large databases’

[8] Douglas Burdick, Manuel Calimlim, Johannes Gehrke, “MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases”, Proceedings of the 17th International Conference on Data Engineering.

[9] "Heart disease" from http://en.wikipedia.org/wiki/Heart_disease.

[10] "Hospitalization for Heart Attack, Stroke, or Congestive Heart Failure among Persons with Diabetes", Special report: 2001 – 2003, New Mexico.

[11] Cornell, D, “A Vertical Partitioning Algorithm for Relational Databases”, Proceedings of the Third International Conference on Data Engineering, Los Angeles, CA, February 1987.