Efficient Implementation of FP Growth Algorithm-Data Mining on Medical Data

Abdullah Saad Almalaise Alghamdi

Chairman of CIT
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah
Kingdom of Saudi Arabia

Abstract
Data mining plays an essential role for extracting knowledge from large databases from enterprises operational databases. Everyday organizations collect huge amount of data from several resource. Whereas medical data is consider most famous application to mine that data to provide interesting patterns or rules for the future perspective. Data mining implementation on medical data to generate rules and patterns using Frequent Pattern (FP)-Growth algorithm is the major concern of this research study. We presented in this paper how data mining can apply on medical data. Therefore, empirical data and result presented in this paper to provide more guidance to the doctors as well as more understanding about the relation of a doctor and a patient. For this, we first discussed about the importance of data mining using medical data then discussion of general data mining techniques has been presented. Furthermore, methodology describes the conceptual model for the extraction of rules on medical databases finally result can guide the relationship between the different attributes presented in the data.

Key Words
Medical Data Mining, Association Mining, FP-Growth Algorithm

1. Introduction

Medical data has more complexities to use for data mining implementation because of its multi dimensional attributes. This type of data can include text, images, and videos also. Therefore, observation using text, numerical, images and videos type data provide the complete judgment about the human's disease. Krzysztof (2002) stated that human's data can have several perspectives such as, visual and auditory sensations, the perception about the pain, discomfort, and observation lead towards the possible occurrence of disease in the future [1].
Medical data have huge attributes includes; images, values, signals like ECG etc. Every
patient has hundreds of attributes to define his condition. This clinical data have been submitted by different department. Therefore the approach that uses these data to find and extract knowledge from it is required. In this regard, we applied FP growth algorithm for extracting rules from the medical data.
Mohammad Saraee (1998) used EasyMiner for generating the rules, where with the help of those rules, values of attributes are generalized at multiple levels [7]. It shows the importance of generating rules from wide data, in this project mainly using the medical data for generating the rules for the future association between the different attributes and suggesting the probable disease. Moreover, another work for association mining in the same way done by Kemper et al [8], performed the analysis on data where the attributes were, title, sex, marital status, etc. That analysis showed the customer limits for granting him/her loan amount. Several researcher has presented the different view by using association mining. Whereas the purpose behind is to find out hidden patterns from the large databases which can guide the organization under the consideration of future decisions.

1.1 Data Mining Approaches

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process [5]. Furthermore, Abdullah et. Al described the data mining in the sense of decision support systems(DSS) that, in decision support management terminology, data mining can be consider as a decision support process in which decision maker is searching to generate rule for the help in decision making.
Mainly, data mining tasks has been divided into descriptive and predictive methods. Classification, clustering and rule association mining are most common techniques use for predictive and descriptive analysis [10]. Thereofre, mainly scholars describe data mining in three major tasks. As Zaine [5] stated in his book chapter about major techniques of data mining as follows:

Manuscript received December 5, 2011
Manuscript revised December 20, 2011

Classification – Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection.
Classification consider as an important task of data mining. Using this approach data must be already defined a class label (target) attribute. Firstly we divide the classified data into two sets; training and testing data [11]. Where each datasets contains others atrributes also but one of the attributed must be defined as class lable attribute. Jiawei Han [11] described classification task in two steps process; first is model construction and the second is model usage. The main target of this task is to build the model by using training dataset and then assign unseen records into a class by using the trained model as accurately as possible. While training data set is use to build the model on the other hand testing data set is use to validate the model [10].
Clustering – Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification.
Clustering is one of the major task has been applying for data mining, work on unsupervised data (no predefined classes) [12]. Clustering is a collection of data objects, clustered by taking similar object to one another within the same cluster, and dissimilar to the objects related in other clusters. Cluster differentiate by using similarities between data according to the characteristics found in the data and grouping similar data objects into clusters [11].
Association – Association analysis is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets.
Data can be use to find association between several attributes, generate rules from data sets, this task is known as association rule mining [12]. Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction. The goal of association rule mining is to find all rules having support $\geq$ minsup (minimum support) threshold and confidence $\geq$ minconf (minimum confidence) threshold [10].
Moreover, association rule mining can be viewed as a two-step process, first, find all frequent itemsets: items satisfying minimum support. Second, generate strong association rules from the frequent itemsets: these rules must satisfy minimum support and minimum confidence [11].
Continue with the association mining as this technique we use in this research for generating the association rules by using medical data. [18] discussed the knowledge discovery view by using apriori algorithm, the model we can use for extracting rules and pattern form the data. Furthermore, data mining can be apply in medical data also where association between the several attributes of patient provide the result of future prediction of diagnose and disease.

1.2 Data Mining on Medical Data

A medical data consist of multiple attributes of patient data related with diagnostic and treatment process. This may include text, numerical, images, and spatial data. Groselj (2002) stated that, "mainly of diagnostic results are images. All on-line remarks, the majority of diagnostic, therapeutic and final results are described qualitatively. Certainly the final reports guide the patient about the status of the problem or disease. Whereas for the needs of epidemiologists the final diagnosis is encoded" [13].
In the same way, Hai Wang (2008) discussed that, in medical data mining applications for medical diagnosis or drug development can work for future prediction by using result generated by data mining application. The forecast or suggestions such as diagnose disease, suggest medicine prescription, precautions for the future, etc. Finally, the procedure to follow is to apply the computer generated result and medical knowledge for the final result [15].
In the medical centers sometimes there is a group of panel work in joint collaboration while dealing with the patients for their diagnose and cure. Indeed it need more study in the form of medical test, discussions, interview session with the patients for diagnose the patient correctly. Wang (2004), "In the medical fields such as medical diagnosis and drug development, knowledge workers attend to do one type of work at their best performance and play roles of joint collaboration" [16]. As [15] suggested that, in the real life it is difficult to find the people who suppose to be an expert from medical field as well as an expert on data mining issues.
Therefore, medical data mining requires more knowledge in several fields as discussed above in medical and in computer data mining too. Moreover, in medical literature high dimensionality, complex relationship, and privacy issues cannot be avoid. Houston et-al [6] described in his research, "by choosing medical literature as our application area for this prototype. Specifically, selected NCI's CancerLit collection because, while it has many of the data mining challenges mentioned above (e..g, high dimensionality, complex relationships, HCI – human computer interaction issues, privacy issue and the meaningfulness of detected patterns), it has a structure and each document has a major topic or theme, making it easier than Internet personal homepages to apply our data mining tools."
Ezquerra (1993), analyzed the medical data for identifying possibilities and causes of heart disease in particular by

using several attributed and experienced knowledge [17]. Association between several attributes using medical data can guide the expected disease a patient can have currently or in future. The result will help the doctors and patient for proper precaution and medicine must be start at the current level of disease. The major task of data mining is to extract unusual patterns from the data. In the same pattern [15] has proved the result in his study that "patients who are male, smoking, and older than 70, are likely to have more than 80% diseased right coronary arteries."

Another view of this study is to apply data mining for medical quality management. Although the approach is same as implementation of data mining on the patient data to ensure medical data quality management. [19] discussed about the tasks included in medical quality management that can be explained as optimization of medical center processes in the form of medical, management and cost benefits analysis. However, the major issues regarding medical data processes are the standards, strategic plans, treatment, diagnoses, medical tests, and finally quality of data. Those qualities can be measured by using several parameters. This process requires a comprehensive approach and expertise on several domains to discover quality knowledge. Therefore, [20] described this all process of medical quality management must be consider intelligent data mining if have a complete corporation and expertise between two domains; medical quality managers, and data mining experts. And process consists of data driven as well as interest driven analyses. They produced a tool of knowledge discovery assistant for supporting the above statement.

The rapid growth of hospital information systems (HIS) are cooperating in the collection of large amount of data of hospitals. In the consequence creating large amount of data repositories in the form of operational databases and data warehouses. Indeed, it open the gate for implementation of data mining tasks on the large amount of data of HIS to generate pattern and discover hidden knowledge. [21] stated in his research that HIS stores the vast amount of laboratory examinations as databases. The length of data is unlimited with number of attributes consider that hundreds of patients visiting to the doctors in commercial cities. For example, [22] discussed about university hospitals data, where more than 1000 patients visit from Monday to Friday. And a database system stores more than 1 GB numerical data of laboratory examinations for every year. Moreover, a combination of images and several other types of data is a compulsory part of medical data in current hospital information systems.

2. Methodology

The conceptual model of virtuous data mining cycle presented in figure-1, presented by [15]. Every time data mining start with the problem identification followed by data preprocessing and finally implementation of data mining task. In the figure-1 showed the basic steps involved in data mining complete process which can also consider a life cycle of data mining process may have several iterations. According to four boxes in the figures representing the four basic steps, may includes some other sub-tasks in each of the steps. For example, in the second step data transformation consider a comprehensive steps must include in the data mining cycle.

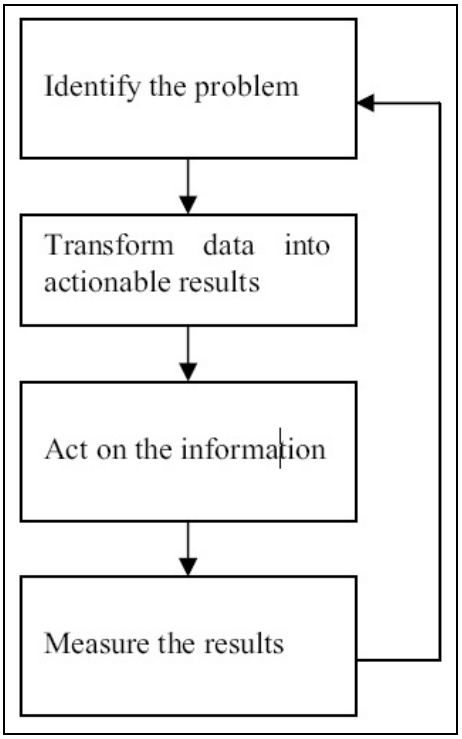


Figure-1Virtuous Cycle of Data Mining [15]

Reference to the figure-1 where data mining cycle presented to measure the results. In this section we proposed an elaborative version of that model presented in figure-2. Followed by description of our conceptual model by using empirical medical data.

2.1 Explanation of the Model – Figure-2

The data mining engine (DME) model presented in the figure-2, is the core part of this paper. The DME model has been tested also by using surveyed data of patient in the subsequent sections. The complete process inside the model with six major steps will generate new patterns and rules from the data. The first step about input data can be

taken from online direct survey or from any operational database. Hence, the basic idea here is to finding association between several attributes of patient data. Data discritization and transformation is the sub-steps of data preprocessing. Before data mining task implementation the major work to perform is called data preprocessing as data play major role for any model. Ultimately the result is based on input data we entered in the model at beginning. If the data has inconsistencies or some cleaning issues then result will not generate appropriately. As stated by [22] that in every project, problem identification and data preprocessing consume 85% of the total project time.

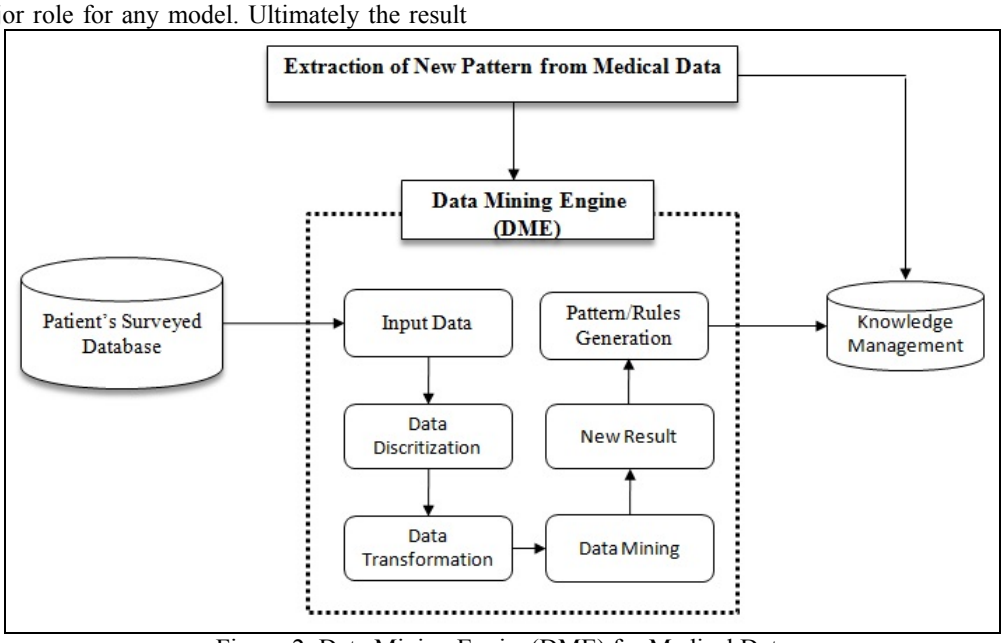


Figure-2: Data Mining Engine(DME) for Medical Data

Furthermore, after selection of data and data preprocessing the next phase is to select appropriate method for data mining task. Depend on the data itself, problem definition and target value will guide here in this phase that which technique will be more suitable here. Therefore, model showed here its simplicity and common approach can be adaptable by any type of data or organization. The model is applicable for marketing strategies, sales forecast, insurance companies, and medical data also.

Finally the last two boxes are showing the generation of the result and extraction of new patterns and rules. Knowledge management database is always updating by new result or experience data. For any problem interface is connected with both directly; knowledge management and new process. First the cursor will search the queried answer in the knowledge management, where knowledge management always updating by new result and experience queried data. Hence, searching in the knowledge management database will consider rapid result and save of times for running out new process always. It shows also the importance of knowledge management database in this model.

In the succeeding section a case study using empirical medical data by using the above mentioned model and followed by each step.

3. Case Study

The proposed model of data mining engine in figure-2 presented in above section is the key to generate association rules from the given data. For the practical implementation of that model we applied in this section by using National Bureau of Economic Research surveyed data [27]. For the testing the model we selected some of the attributes from the data to apply in this data mining engine by using association analysis task of data mining. The algorithm used in this study is Frequent pattern (FP) growth algorithm. It uses the approach based on divide and conquer method. Furthermore, the algorithm will provide the frequent item sets having minimum support. And then finally using those frequent item set the rules will be generated for further actions. Generated results in the end will be forwarded to the knowledge management database for the future queries. The discussion in the subsequent sections will provide the complete understanding of the implementation.

3.1 FP Growth algorithm

FP Growth is one of the basic algorithm use for generate association rules. FP growth is an approach based on

divide and conquers method. The main purpose of this technique is to produce frequent item sets by using the combination of data attributes. It basically works on to generate frequent item set without candidate set generation [14]. Another sister method of FP growth is known as apriori algorithm also use in association data mining. Apriori is a classic algorithm for learning association rules [23]. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation) [25, 26]. One way to construct a simpler model computed from data, easier to understand and with more predictive power is to create a set of simplified rules [24].

The major steps of FP growth are consist of the following steps:

"Step1- First condenses the database showing frequent item set in to FP-tree.

Step2: It divides the FP-tree in to a set of conditional database and mines each database separately, thus extract frequent item sets from FP-tree directly. It consist of one root labeled as null, a set of item prefix sub trees as the children of the root, and a frequent .item header table. Each node in the item prefix sub tree consists of three fields: item-name, count and node link where---item-name registers which item the node represents; count registers the number of transactions represented by the portion of path reaching this node, node link links to the next node in the FP- tree. Each item in the header table consists of two fields---item name and head of node link, which points to the first node in the FP-tree carrying the item name" [14].

3.2 FP Growth Pseudo Code

Input: constructed FP-tree
Output: complete set of frequent patterns
Method: Call FP-growth (FP-tree, null).
procedure FP-growth (Tree, α)
{
1) if Tree contains a single path P then
2) for each combination do generate pattern β
α with support = minimum support
of nodes in β.
3) Else For each header ai in the header of Tree
do {
4) Generate pattern β = ai α with support =
ai.support;
5) Construct β.s conditional pattern base and
then β.s conditional FP-tree Tree β
6) If Tree β = null
7) Then call FP-growth (Tree β, β)}
} [14]

3.3 Tool used for FP-Growth Implementation

The tool used for the implementation of FP growth in this study is RapidMiner. RapidMiner is an open source package provides good range of major tasks using in data mining; includes, regression, clustering, classification, and association with a good range of sub algorithms too [12]. RapidMiner provides an integrated development environment to build and apply the data mining tasks [28]. It provides the long list of processes, operators and data sets for the help of students, decision makers and researchers showed in figure-3.

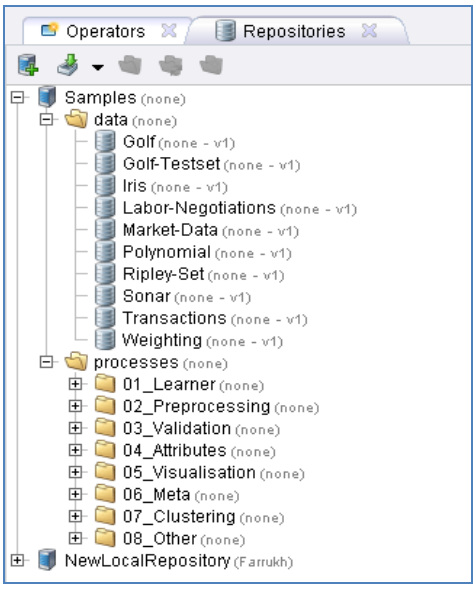


Figure-3 List of Operators [29]

3.4 Attributes Selection

Presented in figure-4, we selected some attributes for the implementation of FP growth algorithm using RapidMiner. There were more than hundred (100) questions asked in the survey done by National Bureau of Economic Research [27]. In this test fifteen (15) attributes selected for better understanding of result.

Attribute #	Description	Code
1	Race	DMARACER
2	Sex	HSSEX
3	Family size (persons in family)	HSFSIZER
4	Marital status	HFA12
5	Mother's weight calculated in pounds	HFC2S
6	Father's weight calculated in pounds	HFC4S
7	Parent high blood pres/stroke before 50	HFC6A
8	Parent heart attack/angina before 50	HFC6B
9	Parent high blood cholesterol at any age	HFC6C
10	Parent asthma or hay fever at any age	HFC6D
11	Parent diabetes at any age	HFC6E
12	Country mother born	HFC10
13	Country father born	HFC11
14	Doctor say had arm muscle weak/paralysis	HYE1E
15	Doctor ever say SP had asthma	HYE1G

Figure-4 Attributes Selection

3.5 Data Discritization and Transformation

Data discritization has been performed on the data to make the data more applicable for the implementation of FP growth algorithm. Originally, the attributes has the numerical data types ranges from 1 to 1000, we discritized them into two and three different categories to make easy to understand and readable data. And finally made transformation on data into binary data as binary data is applicable for FP growth algorithm. Where the 0 is denote that attribute has not selected for the patient and 1 means that attribute has selected for the patient.

Figure-5 represent the data look after making categorical data where figure -6 is representing the final data selected after transformation of the data from numerical values to binominal values. Where we converted the each attributes into two or three categories, therefore in figure no. 6 each column is representing the each category of particular attribute. For example attribute no-1 in figure 6 has three categories named Race1, Race2, and Race3. Where data values inside each category will guide us about selection of the category, i.e (the first patient data values for the attribute-1 are 0,1,0) it showed that the patient has the Race2 data value selection for this attribute.

Race		Sex		Family size (persons in family)		Marital status		Mother's weight calculated in pounds		Father's weight calculated in pounds		Parent high blood pres/stroke before		Parent heart attack/angina		Parent high blood cholesterol at any
2	0,1,0	2	0,1	10	0,0,1	2	0,1,0,0	160	0,1,0	182	0,1,0	2	1,0,0	2	1,0,0	2
1	1,0,0	2	0,1	4	0,1,0	2	0,1,0,0	122	0,1,0	167	0,1,0	2	1,0,0	2	1,0,0	2
1	1,0,0	1	1,0	3	1,0,0	2	0,1,0,0	125	0,1,0	200	0,1,0	2	1,0,0	2	1,0,0	2
1	1,0,0	2	0,1	3	1,0,0	1	1,0,0,0	165	0,1,0	195	0,1,0	1	1,0,0	2	1,0,0	2
1	1,0,0	1	1,0	10	0,0,1	1	1,0,0,0	133	0,1,0	182	0,1,0	2	1,0,0	2	1,0,0	2
1	1,0,0	2	0,1	6	0,1,0	2	0,1,0,0	160	0,1,0	210	0,0,1	2	1,0,0	2	1,0,0	2
2	0,1,0	2	0,1	3	1,0,0	2	0,1,0,0	110	1,0,0	165	0,1,0	2	1,0,0	2	1,0,0	2
2	0,1,0	2	0,1	3	1,0,0	2	0,1,0,0	105	1,0,0	160	0,1,0	2	1,0,0	2	1,0,0	2
1	1,0,0	2	0,1	4	0,1,0	2	0,1,0,0	120	1,0,0	160	0,1,0	2	1,0,0	2	1,0,0	2
1	1,0,0	1	1,0	7	0,0,1	2	0,1,0,0	133	0,1,0	182	0,1,0	1	1,0,0	2	1,0,0	1
2	0,1,0	1	1,0	3	1,0,0	2	0,1,0,0	125	0,1,0	180	0,1,0	2	1,0,0	2	1,0,0	2
1	1,0,0	2	0,1	5	0,1,0	2	0,1,0,0	123	0,1,0	173	0,1,0	2	1,0,0	2	1,0,0	1
1	1,0,0	1	1,0	3	1,0,0	2	0,1,0,0	115	1,0,0	140	1,0,0	2	1,0,0	2	1,0,0	2
1	1,0,0	2	0,1	4	0,1,0	1	1,0,0,0	120	1,0,0	230	0,0,1	2	1,0,0	2	1,0,0	2
1	1,0,0	1	1,0	4	0,1,0	2	0,1,0,0	150	0,1,0	172	0,1,0	2	1,0,0	2	1,0,0	2
1	1,0,0	1	1,0	3	1,0,0	2	0,1,0,0	107	1,0,0	162	0,1,0	2	1,0,0	2	1,0,0	2
2	0,1,0	2	0,1	4	0,1,0	2	0,1,0,0	162	0,1,0	190	0,1,0	2	1,0,0	2	1,0,0	2
2	0,1,0	2	0,1	3	1,0,0	2	0,1,0,0	170	0,1,0	210	0,0,1	1	1,0,0	2	1,0,0	1
1	1,0,0	1	1,0	6	0,1,0	2	0,1,0,0	120	1,0,0	135	1,0,0	2	1,0,0	2	1,0,0	2
1	1,0,0	2	0,1	4	0,1,0	2	0,1,0,0	150	0,1,0	115	1,0,0	2	1,0,0	2	1,0,0	2
2	0,1,0	1	1,0	5	0,1,0	2	0,1,0,0	130	0,1,0	230	0,0,1	2	1,0,0	2	1,0,0	2
1	1,0,0	1	1,0	3	1,0,0	2	0,1,0,0	155	0,1,0	275	0,0,1	2	1,0,0	2	1,0,0	2
1	1,0,0	1	1,0	3	1,0,0	2	0,1,0,0	150	0,1,0	173	0,1,0	2	1,0,0	2	1,0,0	2
1	1,0,0	2	0,1	5	0,1,0	2	0,1,0,0	140	0,1,0	200	0,1,0	2	1,0,0	2	1,0,0	1
1	1,0,0	1	1,0	5	0,1,0	2	0,1,0,0	133	0,1,0	180	0,1,0	8	0,0,1	8	0,0,1	8
2	0,1,0	2	0,1	5	0,1,0	2	0,1,0,0	124	0,1,0	148	0,1,0	2	1,0,0	2	1,0,0	2
1	1,0,0	1	1,0	5	0,1,0	2	0,1,0,0	180	0,1,0	214	0,0,1	2	1,0,0	2	1,0,0	1

Figure-5 Data Discritization

Race1	Race2	Race3	Sex1	Sex2	Family size (persons in	Family size (persons in	Family size (persons in	Marital status1	Marital status2	Mothers weight calculated in	Mothers weight calculated in	Mothers weight calculated in	Fathers weight calculated in	Fathers weight calculated in	Fathers weight calculated in	Parent high blood pres/stro	Parent high blood pres/stro	Parent high blood pres/stro
0	1	0	0	1	0	0	1	0	1	0	1	0	0	1	0	1	0	0
1	0	0	0	1	0	1	0	0	1	0	1	0	0	1	0	1	0	0
1	0	0	1	0	1	0	0	0	1	0	1	0	0	1	0	1	0	0
1	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0	1	0	0
1	0	0	1	0	0	0	1	1	0	0	1	0	0	1	0	1	0	0
1	0	0	0	1	0	1	0	0	1	0	1	0	0	0	1	1	0	0
0	1	0	0	1	1	0	0	0	1	1	0	0	0	1	0	1	0	0
0	1	0	0	1	1	0	0	0	1	1	0	0	0	1	0	1	0	0
1	0	0	0	1	0	1	0	0	1	1	0	0	0	1	0	1	0	0
1	0	0	1	0	0	0	1	0	1	0	1	0	0	1	0	1	0	0
0	1	0	1	0	1	0	0	0	1	0	1	0	0	1	0	1	0	0
1	0	0	0	1	0	1	0	0	1	0	1	0	0	1	0	1	0	0
1	0	0	1	0	1	0	0	0	1	1	0	0	1	0	0	1	0	0
1	0	0	0	1	0	1	0	1	0	1	0	0	0	0	1	1	0	0
1	0	0	1	0	0	1	0	0	1	0	1	0	0	1	0	1	0	0
1	0	0	1	0	1	0	0	0	1	1	0	0	0	1	0	1	0	0
0	1	0	0	1	0	1	0	0	1	0	1	0	0	1	0	1	0	0
0	1	0	0	1	1	0	0	0	1	0	1	0	0	0	1	1	0	0
1	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0
1	0	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1	0	0

Figure-6 Data Transformation

3.6 FP Growth Process

Now, the data has been selected, checked, and modified using data pre processing techniques available inside RapidMiner. All steps in data pre processing are depending on the applying of association mining algorithm FP growth. The process of FP growth algorithm using rapid miner tool is presented in figure-7.

The explanation of the entire operators used in rapid miner and steps for the complete process of generate association rules are presented below:

Operator No.1: is the Read CSV data file from the hard disk. It use for import data from .csv (comma delimited) file. The input data file figure-6 has been imported for the using in the process.

Operator No.2: is a Type Converter of numerical data into binominal data. Using this operator we convert all the data values into binominal values before connecting it with the FP growth.

Operator No.3: is a FP growth algorithm receiving the data file of patient information. Given support value is 60% for extracting frequent item set. Generated frequent item set will be used by the next operator for create association rules.

Operator No.4: Finally this operator is a Rules Generator. After extraction of frequent item set by using input data now only frequent item set will be forwarded for the generation of association rules from the given data having confidence value 95%.

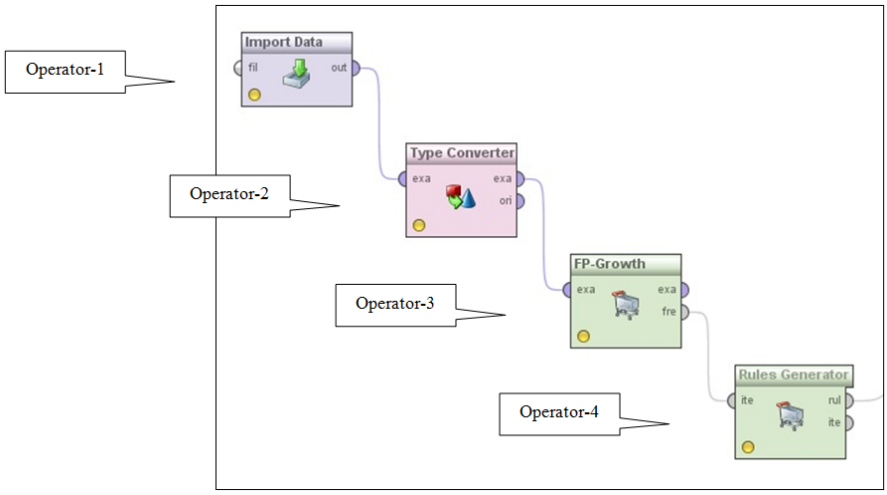
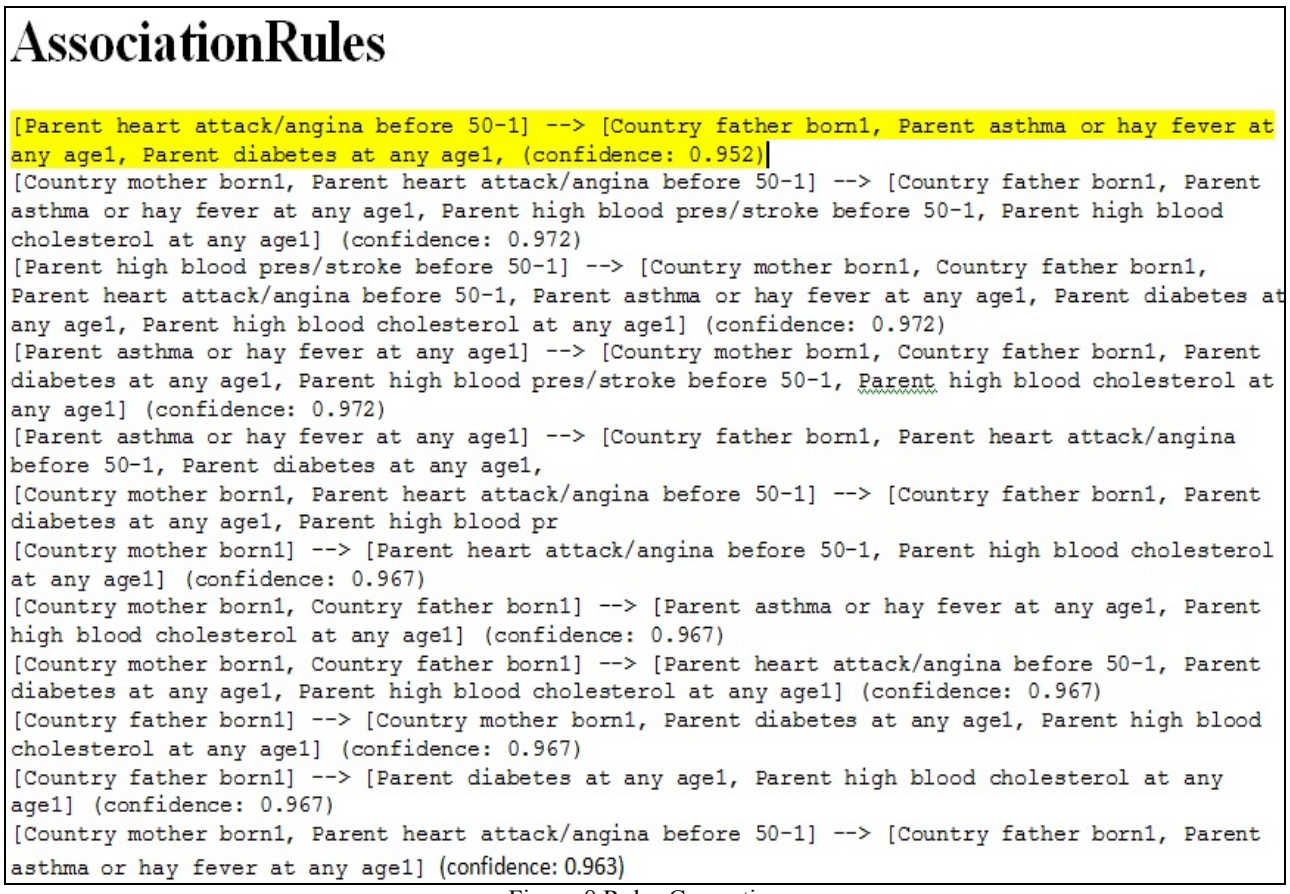


Figure-7 FP Growth Process

3.7 Rules Generation

After running process the output generated in the form of association between the attributes selected in the input file presented in figure.8. There are several decisions a decision maker can extract from these rules. For example, parent heart attack/angina before 50 age is depending on where the father born, if parent has asthma before at any age, and if parent has diabetes at any of the age with confidence value 0.952.

The complete list of rules showed that there are several supporting rules have generated which can be helpful for hospital information systems or general information for the patients visiting to the doctors. Especially the rules are more useful for the doctor diagnostic system and disease recognition. Also in this list some of the rules are not for use according to the situations. These rules then will be forwarded to knowledge management database for the future actions. In the next step whenever if doctor or administrator needs any information having some attributes related with the input file will be asked from knowledge management using the generated rules previously. If there will be any new data file received using survey or from data warehouse the complete process will use for the generating new rules and then update version will again forwarded to the knowledge management database. For more graphical representation figure-9 presented below is showing the rules tree according to frequent item set and rules generation.



AssociationRules

[Parent heart attack/angina before 50-1] --> [Country father born1, Parent asthma or hay fever at any age1, Parent diabetes at any age1, (confidence: 0.952)

[Country mother born1, Parent heart attack/angina before 50-1] --> [Country father born1, Parent asthma or hay fever at any age1, Parent high blood pres/stroke before 50-1, Parent high blood cholesterol at any age1] (confidence: 0.972)

[Parent high blood pres/stroke before 50-1] --> [Country mother born1, Country father born1, Parent heart attack/angina before 50-1, Parent asthma or hay fever at any age1, Parent diabetes at any age1, Parent high blood cholesterol at any age1] (confidence: 0.972)

[Parent asthma or hay fever at any age1] --> [Country mother born1, Country father born1, Parent diabetes at any age1, Parent high blood pres/stroke before 50-1, Parent high blood cholesterol at any age1] (confidence: 0.972)

[Parent asthma or hay fever at any age1] --> [Country father born1, Parent heart attack/angina before 50-1, Parent diabetes at any age1,

[Country mother born1, Parent heart attack/angina before 50-1] --> [Country father born1, Parent diabetes at any age1, Parent high blood pr

[Country mother born1] --> [Parent heart attack/angina before 50-1, Parent high blood cholesterol at any age1] (confidence: 0.967)

[Country mother born1, Country father born1] --> [Parent asthma or hay fever at any age1, Parent high blood cholesterol at any age1] (confidence: 0.967)

[Country mother born1, Country father born1] --> [Parent heart attack/angina before 50-1, Parent diabetes at any age1, Parent high blood cholesterol at any age1] (confidence: 0.967)

[Country father born1] --> [Country mother born1, Parent diabetes at any age1, Parent high blood cholesterol at any age1] (confidence: 0.967)

[Country father born1] --> [Parent diabetes at any age1, Parent high blood cholesterol at any age1] (confidence: 0.967)

[Country mother born1, Parent heart attack/angina before 50-1] --> [Country father born1, Parent asthma or hay fever at any age1] (confidence: 0.963)

Figure-8 Rules Generation

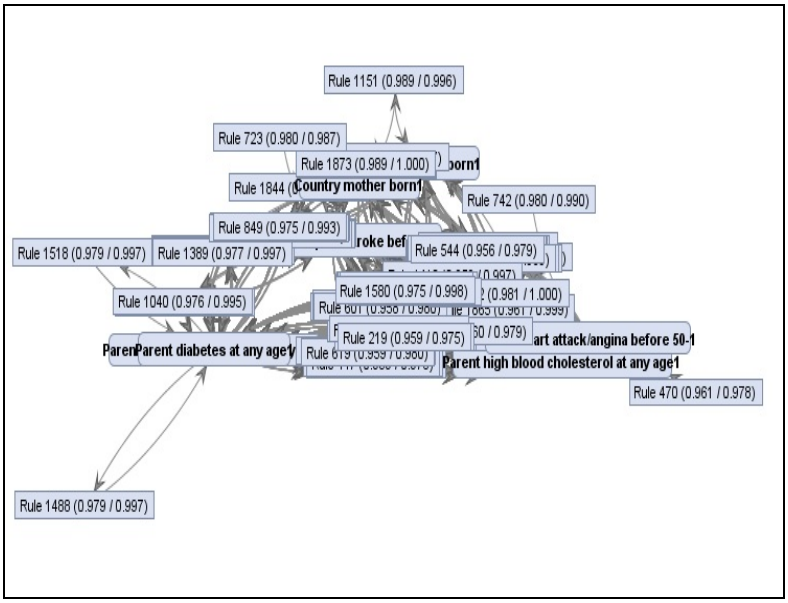


Figure-9 Rules Tree

4. Future Work

The model is applicable in future by using any other data mining task other than association mining. There is a chance for modification by using some other functions we can put in the model. The new hospital information systems while using this model can get a positive support for their doctors and research work. In future the model implementation by using some techniques will be the more beneficial for the professionals of data miner and can explore the solution in broad manner.

5. Conclusion

In today world the large amount of data has been saving regularly in the enterprises. There is a need of special treatment for the best use of these data. We applied here in this study data mining techniques which consider the best way to extract new information from the data. Started with the large amount of data in this research, after selection some attributes we presented the best use of data by creating some association rules for the medical students and doctors help. Knowledge management is providing the facility to find out these rules any time when need. In addition update version of association rules will modify the previous rules according to new data collected from several sources.

References

[1] Krzysztof J. Cios and G. William Moore, "Uniqueness of Medical Data Mining", Artificial Intelligence in Medicine Journal, 2002.

[2] Agrawal R. and Srikant R., Fast algorithms for mining association rules, VLDB (1994), 487-499.

[3] Fayyad U., Piatetsky-Shapiro G. and Smyth P., From data mining to knowledge discovery in databases, In: AAAI Magazine, 1996, (Fall), pp.37-53.

[4] Petrucelli J., Nandram B. and Chen M., Applied statistics for scientists and engineers, NJ, USA, 1999.

[5] Osmar R. Zaïane, "Chapter I: Introduction to Data Mining", CMPUT690 Principles of Knowledge Discovery in Databases, 1999.

[6] Andrea L. Houston, Hsinchun Chen, Susan M. Hubbard, Bruce r. Schatz, Tobun D. Ng1, Robin R. Sewell and Kristin M. Tolle1, "Medical Data Mining on the Internet: Research on a Cancer", Information System, Artificial Intelligence Review 13: 437–466, 1999.

[7] Mohammad Saraee, George Koundourakis, Babis Theodoulidis, EASYMINER: DATA MINING IN MEDICAL DATABASES, The institution of Electrical Engineers, published by the IEE, Savoy Place, London, UK-1998.

[8] Kamber M. Winstone L, Gong W, Cheng S, Han J. Generalisation and Decision Tree Induction: Efficient Classification in Data Mining. In Proceedings of 1997 Intemational Workshop on Research Issues on Data Engineering (RTDE'97), Birmingham, England, April 1997, pp 11 1-120.

[9] Dr. Abdullah Al- Mudimigh, Farrukh Saleem, Zahid Ullah, Efficient Implementation of Data Mining: Improve Customer's Behavior", The 7th IEEE/ACS, International Conference on Computer Systems and Applications, Rabat, Morocco, to be held on May, 10-13, 2009.

[10] Pang-Ning Tan, Michael Steinbach & Vipin Kumar, "Introduction to Data Mining", Addison Wesley, 2005, ISBN 0321321367
[11] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining: Concepts and Techniques", 2nd edition, 2005, Morgan Kaufmann, ISBN 1558609016
[12] Farrukh Saleem, Areej Malibari, DATA MINING COURSE IN INFORMATION SYSTEM DEPARTMENT–CASE STUDY OF KING ABDULAZIZ UNIVERSITY, IEEE, ICEED- 3rd International Congress on Engineering Education, 7th – 8th December, 2011.
[13] C Groselj, Data Mining Problems in Medicine, Proceeding of the 15th IEEE Symposium on Computer Based Medical Systems-2002.
[14] Kuldeep Malik, Neeraj Raheja and Puneet Garg, Enhanced FP-Growth Algorithm, IJCEM Journal, April-2011.
[15] Hai Wang, Shouhong Wang, Medical Knowledge Acquisition through Data Mining, Proceedings of 2008 IEEE International Symposium on IT in Medicine and Education.
[16] Wang, S., and Ariguzo, G., Knowledge management through the development of information schema, Information & Management, 41(4), 2004, pp. 445-456.
[17] Ezquerra, N. and Mullick, R. Perfex: An expert system for interpreting myocardial perfusion, Expert Systems with Applications, 6(3), 1993, pp. 455-468
[18] Dr. Abdullah Al- Mudimigh, Farrukh Saleem, Zahid Ullah, The Role Of Data Mining In ERP-CRM Model", International Conference on Applied Computer & Applied Computational Science (ACACOS '09), Hangzhou, China, 2009.
[19] Wolf Stühlinger , Oliver Hogl , Herbert Stoyan , Michael Müller, Intelligent Data Mining for Medical Quality Management (2000), Proc. Fifth Workshop Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP2000), Workshop Notes of the 14th European Conf. Artificial Intelligence (ECAI-2000).
[20] Hogl, O.; Stoyan, H. et al.: The Knowledge Discovery Assistant: Making Data Mining Available for Business Users, in: Gunopulos, D.; Rastogi, R. (eds.): Proceedings of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD-2000), Dallas, Texas, May 2000, pp. 96-105.
[21] Van Bemme1,J. and Musen, M. A.1997. Handbook of Medical Informatics, Springer-Verlag, New York.
[22] Efraim Turban, Ramesh Sharda, Dursun Delen, Decision Support and Business Intelligence, 9th Edition, published by Pearson Education, Prentice Hall-2011.
[23] Abdullah Al- Mudimigh, Farrukh Saleem, Zahid Ullah, The Effects Of Data Mining In ERP-CRM Model – A Case Study Of MADAR, WSEAS Transaction, 2009
[24] Mar_a C. FERN_ANDEZ_, Ernestina MENASALVAS_, Óscar MARB_AN_Jos_e M. PE~NA__, Socorro MILL_AN, "MINIMAL DECISION RULES BASED ON THE APRIORI ALGORITHM y", Int. J. Appl. Math. Comput. Sc.i, Vol.11, No.3, 691704, 2001
[25] Apriori Algorithm, http://en.wikipedia.Org /wiki/Apriori algorithm #Algorithm, Accessed Date: 12th April, 2011.
[26] Agrawal R, Imielinski T, Swami AN. "Mining Association Rules between Sets of Items in Large Databases." SIGMOD. June 1993, 22(2):207-16.
[27] National Bureau of Economic Research, patient survey data Accessed date, 14th July, 2008
[28] http://rapidi.com/content/view/181/196/, accessed date, 10th May, 2011.
[29] http://rapidi.com/content/view/181/196/, accessed date, 10th May, 2011.

Abdullah Saad AL-Malaise AL-Ghamdi received his PhD from George Washington University of USA. Dr. Abdullah is an Assistant Professor of IS and CS, Information Systems Department, Faculty of Computing and Information Technology, King Abdulaziz University. He is also working as a chairman of CIT college at Jeddah Community College. Dr. Abdullah also is an Assistant Professor of MIS at MIS Department, College of Business Administration (CBA)m Jeddah, Saudi Arabia.