Extraction of Frequent Patterns from Diabetes Cluster

Myint Swe Lai Win, Win Lelt Lelt Phyu
University of Computer Studies, Yangon
glitteryuki@gmail.com, winlei@gmail.com

Abstract

Data mining is the process of discovering interesting knowledge, such as patterns, associations, changes, anomalies and significant structures, from large amounts of data stored in databases, data warehouses, or other information repositories. In this paper, we have proposed an efficient approach for the extraction of significant patterns from the patients database for diabetes prediction. The diagnosis of diseases is a significant and tedious task in medicine. To facilitate the diagnosis process, the effort to utilize knowledge and experience of numerous specialists and clinical screening data of patients collected in databases is considered a valuable option. The patients database is clustered using the KMIX clustering algorithm, which will extract the data relevant to diabetes from the database. Subsequently the frequent patterns are mined from the extracted data, relevant to diabetes, using the MAFIA algorithm. Then the significant patterns to diabetes diagnosis are chosen from these frequent patterns. These patterns can be used to apply in the healthcare system.

Keywords: Data Mining, Diabetes Diagnosis, Clustering, KMIX, Frequent Pattern Mining, MAFIA, Significant Pattern.

1. Introduction

Data mining is the non trivial extraction of implicit previously unknown and potentially useful information about data [5]. Diverse fields such as marketing, customer relationship management, engineering, medicine, crime analysis, expert prediction, Web mining, and mobile computing, besides others utilize Data mining. 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 [4]. Medical diagnosis is considered to a significant yet intricate task that needs to be carried out precisely and efficiently. The researchers in the medical field diagnose and predict the diseases in addition to providing effective care for patients 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.

The primary intent of the research is to design and develop an efficient approach for extracting patterns, which are significant to diabetes, from the patient database. The approach aims to utilize the data mining techniques: clustering and frequent pattern mining. The patient database consists of mixed attributes containing both the numerical and categorical data. Clustering is performed using KMIX clustering algorithm [1] with K value so as to extract data relevant to diabetes. Subsequently the maximal frequent patterns significant to diabetes diagnosis are mined from the extracted data using the MAFIA algorithm [3]. Further, the patterns significant to diabetes prediction are chosen based on the calculated significant weightage [2]. These significant patterns can be used in the development of diabetes prediction system.

2. Related Work

A clustering algorithm, KMIX, which was proposed by Thuy. T.T. Nguyen et al. [1] has the advantages over K-means for comparison on the cardiovascular data. It performs as well as other published results for the Soy Bean, Votes, and Wisconsin Breast Cancer data sets. These algorithm is compared to published results and compares favourably.

Doug Burdick et al. [3] have presented a new algorithm for mining maximal frequent itemsets from a transactional database. The search strategy of the algorithm integrates a depth-first traversal of the itemset lattice with effective pruning mechanisms. The performance numbers show that the algorithm outperforms previous work by a factor of three to five.

In [2], Mr. Shantakumar Patil et al. have proposed for the extraction of significant patterns from the heart disease warehouses for heart attack prediction. They presented the data preprocessing, K-means clustering algorithm, MAFIA (MAximal Frequent

Itemset Algorithm) and significant patterns of heart attack prediction.

3. Clustering Using KMIX Algorithm

Clustering can be defined as the process of organizing objects in a database into clusters/groups such that objects within the same cluster have a high degree of similarity, while objects belonging to different clusters have a high degree of dissimilarity [7]. It can be applied in many applications. 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. Numerous methods are available for clustering. In this paper, we propose an algorithm, KMIX, which is improved from K-means in order to cluster mixed numerical and categorical data values [1]. In the KMIX algorithm, a dissimilarity measure is defined that takes into account both numeric and categorical attributes via the Euclidean distance for numerical features and the number of mismatches of categorical values for discrete feature. For example, assume that $d^N(X,Y)$ is the squared Euclidean distance between two objects X and Y over continuous features; and $d^C(X,Y)$ is the dissimilarity measure on categorical features in X, Y. The dissimilarity between two objects X, Y is given by the distance

$$d(X,Y) = d^N(X,Y) + d^C(X,Y).$$

The clustering process of the KMIX algorithm is similar to the K-means algorithm except that a new method is used to update the categorical attribute values of cluster. The motivation for proposing KMIX based on K-means is that KMIX can be used for large data sets, where hierarchical clustering methods are not efficient.

The steps involved in a KMIX algorithm are given as follows:

Step-1: K points denoting the data to be clustered are placed into the space. These points denote the primary group centre vectors.

Step-2: The data are assigned to the group that is adjacent to the centre vector.

Step-3: The positions of all the K centre vectors are recalculated as soon as all the data are assigned.

Step-4: Repeat steps-2 and 3 until the centre vectors stop moving any further.

3.1. Notation

Assume that X is a pattern. X typically consists of m components, represented in multidimensional space as:

$$X = (x_1,x_2,x_3,...,x_m) = (x_j)_{j=1...m}.$$

Each component in multidimensional space is called a feature (attribute). A data set includes n patterns X_i where $i \in [1,n]$ and $X_i = (x_{i,1},x_{i,2},...,x_{i,m})$. Hence, we have a $n*m$ pattern matrix.

3.2. Dissimilarity measurement of patterns

For continuous features, the dissimilarity measure of two "continuous" patterns using Euclidean distance is given as:

$$dissim(x_i,x_j) = [D(x_i,x_j)]^2 = \sum_{k=1}^{m} (x_{ik} - x_{jk})^2 \ , i,j \in [1,n_1],\ n_1 \le n \quad (1)$$

where D is Euclidean distance.

The dissimilarity of two patterns x_i and x_j is the square of the Euclidean distance between them.

For discrete features, the dissimilarity will be the number of different values of two considering pattern in each categorical feature. We can represent this dissimilarity in the following formula:

$$dissim(x_i,x_j) = d(x_i,x_j) = \sum_{k=1}^{m} \theta(x_{ik},x_{jk}) \ , i,j \in [1,n_2],\ n_2 \le n \quad (2)$$

where $\theta(x_{ik},x_{jk}) = \begin{cases} 0 \text{ if } x_{ik}=x_{jk} \\ 1 \text{ if } x_{ik}\neq x_{jk} \end{cases}$, $k=1,2,...,m;\ i,j \in [1,n_2]$

3.3. Centre vectors

As the data feature set includes both continuous and discrete features, the centre vector will include 2 group components of continuous and discrete. Assume that the data features set includes m features, where the p first features are continuous features and the m-p remaining features are discrete. This means each pattern X in the space can be seen as

$$X = (x_{i1},x_{i2},...,x_{ip},x_{ip+1},x_{ip+2},...,x_{im}).$$

Assume that Q is a centre vector for the data set C (C is a sub set of whole data set). So Q can be represented as

$$Q = (q_{j1},q_{j2},...,q_{jp},q_{jp+1},q_{jp+2},...,q_{jm}).$$

The task is to find p "continuous" component values, and m-p "discrete" component values for vector Q_j.

For continuous component values, $\{q_{jk}\}_{k=1,...,p}$ are the means of the k^{th} feature in C.

For discrete component values, $\{q_{jk}\}_{k=p+1,...,m}$ are the set of $mode_k$, where $mode_k$ is the mode of the k^{th} feature.

4. Frequent Pattern Mining by MAFIA

Mining frequent itemsets in large datasets is an important problem in the data mining field since it enables essential data mining tasks such as discovering association rules, data correlations, sequential patterns, etc [6].

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 [3]. This algorithm is applied for the extraction of patterns from the clustered dataset besides performing efficiently when the database consists of very long itemsets specially. The depth-first traversal of the itemset lattice and effective pruning mechanisms are incorporated in the search strategy of the proposed algorithm. The MAFIA algorithm is as below.

Pseudo code for MAFIA:

```
MAFIA(C, MFI, Boolean IsHUT){
name HUT = C.head U C.tail;
if HUT is in MFI
stop generation of children and return
Count all children, use PEP to trim the tail, and
reorder by increasing support,
For each item i in C.trimmed_tail{
IsHUT= whether i is the first item in the  tail
newNode = C U 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
}
```

5. Significance Weightage Calculation

The significance weightage of each pattern is calculated based on the weightage of each attribute present in the pattern and the frequency of each pattern [2].

The formula used to determine the significant weightage (S_w) is as follows:

$$S_{wi} = \sum_{i=1}^{n} W_i f_i$$

where W_i represents the weightage of each attribute and f_i denotes the frequency of each rule. Subsequently the patterns having significant weightage greater than a predefined threshold are chosen to aid the prediction of diabetes

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

where *SFP* represents the significant frequent patterns and Φ represents the predefined threshold. This *SFP* can be used in the diabetes prediction system.

6. Overview of the System

The patients database contains the screening clinical data of patients. Initially, the database is clustered using the KMIX clustering algorithm with K=2. This result in two clusters, one contains the data that are most relevant to diabetes and the other contains the remaining data. The maximal frequent patterns are mined from the data, relevant to diabetes, using the MAFIA algorithm. The significant weightage is calculated for all maximal 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 development of diabetes prediction system. The overview design of these system is as shown in Figure 1.

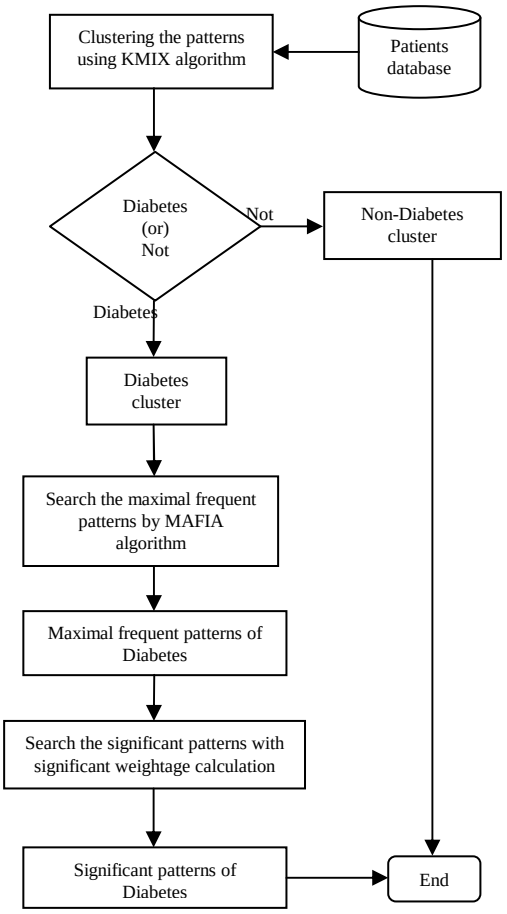

Figure 1. Overview design of the system

7. Experimental Results

The results of our experimental analysis in finding significant patterns for diabetes prediction are presented in this section. The patients dataset was taken from the Yangon General Hospital. It contains about three thousand patients' data records, each being described by 16 attributes. Among them, four attributes (FBS, BP, BMI, Age) are numeric and all others are categorical ones. The measurement units of each numerical attributes are set as in the following: FBS is mg/dl, BP is mmHg, BMI is kg/m^2 where kg is weight and m is height, and Age represents

years.The detailed description of the dataset is given in Table 1.

Table 1. Attributes and their values

No.	Attributes	Values
1	Fasting Blood Sugar (FBS)	<100(Normal)=0, 100-126(Border line)=1, >126(High)=2
2	Diastolic Blood Pressure (BP)	<75=0, ≥75=1
3	Body Mass Index (BMI)	8-25=0, >25=1
4	Age	≤40=0, >40=1
5	Family History (FH)	Absent=0, Present=1
6	Polyuria (Pou)	No=0, Yes=1
7	Urinary Incontinence (UI)	No=0, Yes=1
8	Polydipsia (Pod)	No=0, Yes=1
9	Lethargy (LG)	No=0, Yes=1
10	Weight Loss (WL)	No=0, Yes=1
11	Blurred Vision (BV)	No=0, Yes=1
12	Nocturia (Not)	No=0, Yes=1
13	Balanitis (or) Pruritis Vulva (BoP)	No=0, Yes=1
14	Polycystic ovary syndrome (or) Erectile dysfunction (PoE)	No=0, Yes=1
15	Recurrent skin infection (RSI)	No=0, Yes=1
16	Numbness (Num)	No=0, Yes=1

As an example, there are 9 patients' data records, each having their 16 attributes, are given in the following Table 2.

Table 2. Example records of patient data

Patient-no / Attributes	1	2	3	4	5	6	7	8	9
FBS	2	2	2	0	0	0	2	1	1
BP	0	0	0	0	0	0	0	0	0
BMI	1	1	1	0	0	0	0	0	0
Age	0	1	1	0	0	0	0	0	1
FH	1	0	1	1	0	1	0	1	1
Pou	1	1	0	1	0	0	1	0	1
UI	1	1	1	0	1	0	0	0	0
Pod	1	1	1	0	1	0	1	0	0
LG	1	1	1	1	0	0	0	0	0
WL	1	1	1	0	0	0	1	0	1
BV	0	0	1	0	0	0	0	0	0
Not	0	0	0	0	1	0	0	0	0
BoP	0	1	0	0	0	1	0	0	0
PoE	0	0	1	0	0	0	0	0	0
RSI	1	0	0	0	0	0	0	0	0
Num	1	0	0	1	0	0	0	0	0

These patient database is clustered using KMIX clustering algorithm with K value as 2. One cluster consists of the data relevant to the diabetes and the other contains the remaining data. At first, the centre vectors of two clusters are regarded by the knowledge of specialists as shown in Table 3.

Table 3. Centre vectors of two clusters

Attributes	Centre vector of diabetes cluster	Centre vector of non-diabetes cluster
FBS	2	1
BP	0	0
BMI	0	0
Age	0	0
FH	0	0
Pou	0	0
UI	0	0
Pod	0	0
LG	0	0
WL	0	0
BV	0	0
Not	0	0
BoP	0	0
PoE	0	0
RSI	0	0
Num	0	0

For example, two patients (patient-1 and patient-4 in Table 2) are calculated as in the given:

$d^N(1,D)=|2-2|^2+|1-0|^2=1$
$d^C(1,D)=8$
$d(1,D)=d^N(1,D)+d^C(1,D)=9$

$d^N(1,ND)=|2-1|^2+|1-0|^2=2$
$d^C(1,ND)=8$
$d(1,ND)=d^N(1,ND)+d^C(1,ND)=10$

$d^N(4,D)=|2-0|^2=4$
$d^C(4,D)=4$
$d(4,D)=d^N(4,D)+d^C(4,D)=8$

$d^N(4,ND)=|1-0|^2=1$
$d^C(4,ND)=4$
$d(4,ND)=d^N(4,ND)+d^C(4,ND)=5$

where *D* represents diabetes cluster, and *ND* also represents non-diabetes cluster; $d^N(X_i,Q_j)$ is calculated according to (1) and $d^C(X_i,Q_j)$ is calculated according to (2).
Based on the calculation, patient-1 is more similar to diabetes cluster and patient-4 is to non-diabetes one. Finally, by using KMIX algorithm, all the patients from Table 2 are clustered. The result patients of diabetes cluster and non-diabetes cluster are as shown in Table 4.

Table 4. Cluster results of diabetes and non-diabetes

Patients to diabetes cluster	Patients to non-diabetes cluster
Patient-1	Patient-4
Patient-2	Patient-5
Patient-3	Patient-6
Patient-7	Patient-8
	Patient-9

All the patients in the patients dataset that is used in our system are clustered to diabetes or non-diabetes cluster as calculated in the above example. Later on, the cluster that contains data most relevant to diabetes is fed as input to MAFIA algorithm to mine the maximal frequent patterns present in it. By doing many tests, the best results of maximal frequent patterns for all the patients in the diabetes cluster in the patients dataset are displayed as follows.

- FBS, BMI, FH
- FBS, BMI, Pod

- FBS, BMI, LG
- FBS, BMI, WL
- FBS, BMI, RSI
- FBS, BMI, Num
- FBS, BV
- FBS, Not
- FBS, PoE

Then the significance weightage of each pattern is calculated using the significant weightage calculation approach described in the section 5. The weight values of each attribute is shown in Table 5.

Table 5. Weight values of attributes

Attributes	Weight values
FBS	1
BP	0.5
BMI	0.5
Age	0.5
FH	0.5
Pou	0.5
UI	0.5
Pod	0.5
LG	0.5
WL	0.5
BV	0.5
Not	0.5
BoP	0.5
PoE	0.5
RSI	0.5
Num	0.5

The following is the significant weightage calculation of each pattern:

- FBS, BMI, FH=(1+0.5+0.5)*328=656
- FBS, BMI, Pod=(1+0.5+0.5)*280=560
- FBS, BMI, LG=(1+0.5+0.5)*296=592
- FBS, BMI, WL=(1+0.5+0.5)*280=560
- FBS, BMI, RSI=(1+0.5+0.5)*272=544
- FBS, BMI, Num=(1+0.5+0.5)*320=640
- FBS, BV=(1+0.5)*280=420
- FBS, Not=(1+0.5)*288=432
- FBS, PoE=(1+0.5)*272=408

Subsequently, the significant patterns for the diabetes patients in the patients dataset are extracted with the aid of the significance weightage greater than the predefined threshold. In this system, the extracted patterns from the data set significant to the diabetes prediction are given as below:

- FBS, BMI, FH
- FBS, BMI, Pod
- FBS, BMI, LG
- FBS, BMI, WL
- FBS, BMI, RSI
- FBS, BMI, Num

These patterns are the best in the results that are tested with various predefined threshold. In our experiment, there are nine maximal frequent patterns and six significant patterns for the diabetes patients in the dataset that contains three thousand patients. Those extracted significant patterns can be used to develop an efficient diabetes prediction system.

8. Conclusion

Data mining in health care management is unlike the other fields owing to the fact that the data present are heterogeneous and that certain ethical, legal, and social constraints apply to private medical information. 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 extracting significant patterns from the patient database for the efficient prediction of diabetes. And then, the significant patterns can be used to develop a diabetes prediction system to help the doctors and the diabetic patients.

9. References

[1] Thuy. T.T. Nguyen, Darryl. N. Davis, “A Clustering Algorithm for Predicting CardioVascular Risk”, In the Proceedings of the World Congress on Engineering, 2007.

[2] Shantakumar B. Patil, Dr. Y.S. Kumaraswamy, “Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction”, IJCSNS, Vol. 9, No. 2, February 2009.

[3] Douglas Burdick, Manuel Calimlim, Johannes Gehrke, “MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases”, Proceedings of the 17th International Conference on Data Engineering, pp.443-452, April 02-06, 2001.

[4] 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.

[5] Frawley and Piatetsky-Shapiro, Knowledge Discovery in Databases: An Overview. The AAAI/MIT Press, Menlo Park, C.A, 1996.

[6] Qinghua Zou, Wesley W. Chu and Baojing Lu, “SmartMiner: A depth first algorithm guided by tail information for mining maximal frequent itemsets”.

[7] Ohn Mar San, Van-Nam Huynh and Yoshiteru Nakamori, “An alternative extension of the K-means algorithm for clustering categorical data”, Int. J. Appl. Math. Comput. Sci., 2004, Vol. 14, No. 2, 241-247.