Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus

Elakkia M

[1]Elakkia M , Address: Thanjavur
[2]Professor: V.MATHIMALAR,M.Sc.,M.Phil.,MBA.,Dept. of Computer Science, Shrimathi Indira Gandhi College, Tamil Nadu, India

Abstract - *Early detection of patients with elevated risk of developing diabetes mellitus is critical to the improved prevention and overall clinical management of the patients. The main aim to apply association rule mining to electronic medical records (EMR) to discover sets of risk factors and their corresponding subpopulations that represent patients at particularly high risk of developing diabetes. Given the high dimensionality of EMRs, association rule mining generates a very large set of rules which need to summarize for easy clinical use. The system reviewed four association rule set summarization techniques and conducted a comparative evaluation to provide guidance regarding their applicability, strengths and weaknesses. In this project propose K-Means Clustering Algorithm to resolve the above problems. User can find easily by this application and information about common basic diseases and symptoms. Main aim of this project is to develop a software application for doctors and patients for immediately diagnose disease using K-Means Algorithm. This system will used to quickly find out the disease and generate reports on about the patient status which will be useful for further understanding to deal with the case. This system don't require any interference of doctor for analyzing case , system will analyze and send reports to doctor and patient based on symptoms. After analyzing patients details are stored in database which can view any time for further analysis. Using this report doctor can provide treatment for patient and software will guide patient what type of diet patients have to take.*

Key Words: **Data Mining, Association Rules, Survival Analysis, Association Rule Summarization**

1. INTRODUCTION

Diabetes mellitus is a growing epidemic that affects large number of people, and approximately 7 million of them do not know what disease they have. Diabetes leads to significant medical complications including ischemic heart disease, stroke, nephropathy, retinopathy, neuropathy and peripheral vascular disease. Early identification of patients at risk of developing diabetes is a major healthcare need. Appropriate management of patients at risk with lifestyle changes and/or medications can decrease the risk of developing diabetes by 30% to 60%. Multiple risk factors have been identified affecting a large proportion of the population. For example, pre diabetes (blood sugar levels above normal range but below the level of criteria for diabetes) is present in approximately 35% of the adult population and increases the absolute risk of diabetes 3 to 10 fold depending on the presence of additional associated risk factors, such as obesity, hypertension, hyper lipidemia, etc. Comprehensive medical management of this large portion of the population to prevent diabetes represents an unbearable burden to the healthcare system. Diabetes is part of the metabolic syndrome, which is a constellation of diseases including hyperlipidemia (elevated triglyceride and low HDL levels), hypertension (high blood pressure) and central obesity (with body mass index exceeding 30 kg/m2). These diseases interact with each other, with cardiac and vascular diseases and thus understanding and modeling these interactions is important. Association rules are implications that associate a set of potentially interacting conditions (e.g. high BMI and the presence of hypertension diagnosis) with elevated risk. The use of association rules is particularly beneficial, because in addition to quantifying the diabetes risk, they also readily provide the physician with a "justification", namely the associated set of conditions. This set of conditions can be used to guide treatment towards a more personalized and targeted preventive care or diabetes management.

2. Existing System

A number of successful association rule set summarization techniques have been proposed but no clear guidance exists regarding the applicability, strengths and weaknesses of these techniques. The focus of this manuscript is to review and characterize four existing association rule summarization techniques and provide guidance to practitioners in choosing the most suitable one. A common shortcoming of these techniques is their inability to take diabetes risk a continuous outcome into account. In order to make these techniques more appropriate, the system had to minimally modify them: Extend them to incorporate information about continuous outcome variables. Specifically, the key contributions are as follows. A clinical application of association rule mining is to identify sets of co-morbid conditions (and the patient subpopulations who suffer from these conditions) that imply significantly increased risk of diabetes. Association rule mining on this extensive set of variables resulted in an exponentially large set of association rules. Extended four popular association rule set summarization techniques (mainly from the review by incorporating the risk of diabetes into the process of finding an optimal summary. The main contribution is a comparative evaluation of these extended summarization techniques that provides guidance to practitioners in selecting an appropriate algorithm for a similar problem.

2.1 Survey

A. A statistical theory for quantitative association rules.

Association rules are a key data-mining tool and as such have been well researched. So far, this research has focused predominantly on databases containing categorical data only. However, many real-world databases contain quantitative attributes and current solutions for this case are so far inadequate. These systems introduce a new definition of quantitative association rules based on statistical inference theory. In this definition reflects the intuition that the goal of association rules is to find extraordinary and therefore interesting phenomena in databases. Also introduce the concept of sub-rules which can be applied to any type of association rule. Rigorous experimental evaluation on real-world datasets is presented, demonstrating the usefulness and characteristics of rules mined according to our definition. Introduction Association Rules. The goal of data mining is to extract higher level information from an abundance of raw data.

B. Summarizing Itemset Patterns Using Probabilistic Models.

In this paper, The system is to propose a novel probabilistic approach to summarize frequent itemset patterns. Such techniques are useful for summarization, post-processing, and end-user interpretation, particularly for problems where the resulting set of patterns are huge. In our approach items in the dataset are modeled as random variables. Then construct a Markov Random Fields (MRF) on these variables based on frequent itemsets and their occurrence statistics. The summarization proceeds in a level-wise iterative fashion. Occurrence statistics of itemsets at the lowest level are used to construct an initial MRF. Statistics of itemsets at the next level can then be inferred from the model. We use those patterns whose occurrence cannot be accurately inferred from the model to augment the model in an iterative manner, repeating the procedure until all frequent itemsets can be modeled. The resulting MRF model affords a concise and useful representation of the original collection of itemsets. Extensive empirical study on real datasets show that the new approach can effectively summarize a large number of itemsets and typically significantly out performs extant approaches.

C. Mining compressed frequent-pattern sets.

A major challenge in frequent-pattern mining is the sheer size of its mining results. In many cases, a high min sup threshold may discover only commonsense patterns but a low one may generate an explosive number of output patterns, which severely restricts its usage. In this paper, in these systems analyze the problem of compressing frequent-pattern sets. Typically, frequent patterns can be clustered with a tightness measure δ (called δ-cluster), and a representative pattern can be selected for each cluster. Unfortunately, finding a minimum set of representative patterns is NP-Hard. Develop two greedy methods, RPglobal and RPlocal. The former has the guaranteed compression bound but higher computational complexity. The latter sacrifices the theoretical bounds but is far more efficient. The performance study shows that the compression quality using RPlocal is very close to RPglobal, and both can reduce the number of closed frequent patterns by almost two orders of magnitude. Furthermore, RPlocal mines even faster than FPClose, a very fast closed frequent pattern mining method. This system also show that RPglobal and RPlocal can be combined together to balance the quality and efficiency.

D. CPAR: Classification based on Predictive Association Rules.

Recent studies in data mining have proposed a new classification approach, called associative classification, which, according to several reports, such as [7, 6], achieves higher classification accuracy than traditional classification approaches such as C4.5. However, the approach also suffers from two major deficiencies: (1) it generates a very large number of association rules, which leads to high processing overhead; and (2) its confidence-based rule evaluation measure may lead to over fitting. In comparison with associative classification, traditional rule-based classifiers, such as C4.5, FOIL and RIPPER, are substantially faster but their accuracy, in most cases, may not be as high. In this paper, propose a new classification approach, CPAR (Classification based on Predictive Association Rules), which combines the advantages of both associative classification and traditional rule-based classification. Instead of generating a large number of candidate rules as in associative classification, CPAR adopts a greedy algorithm to generate rules directly from training data. Moreover, CPAR generates and tests more rules than traditional rule-based classifiers to avoid missing important rules. To avoid over fitting, CPAR uses expected accuracy to evaluate each rule and uses the best k rules in prediction.

E. Simulation Studies on the Dynamics of Insulin-glucose in Diabetic Mellitus Patients.

Glucose-insulin interaction in an insulin-dependent diabetic patient has been simulated using an overall model based on pharmacokinetic diagrams of insulin and glucose. Model is capable of predicting the blood glucose and insulin levels, total glucose uptake and the renal glucose excretion. The treatment strategy is based on a four-daily dose of regular insulin, which is applied through a subcutaneous route 30 min prior to each meal.

3. Proposed System

In proposed system, introduce K-Means Algorithm. This is a simple iterative method to partition a given dataset into a user specified number of clusters, k. The algorithm operates on a set of d-dimensional vectors, D = {xi | i = 1,..., N}, where xi ∈ d denotes the ith data point. The algorithm is initialized by picking k points in d as the initial k cluster. Techniques for selecting these initial seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data k times. Then the algorithm iterates between two steps till convergence: A clinical application of association rule mining to identify sets of co-morbid conditions that imply significantly increased risk of diabetes. Association rule mining on this extensive set of variables resulted in an exponentially large set of association rules. The main contribution is a comparative evaluation of these extended summarization techniques that provides guidance to practitioners in selecting an appropriate algorithm for a similar problem. It can uncover hidden clinical relationships and can propose new patterns of conditions to redirect prevention, management, and treatment approaches.

K Means Clustering

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The problem is computationally difficult (NP-hard); however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data; however, *k*-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes. The algorithm has a loose relationship to the *k*-nearest neighbor classifier, a popular machine learning technique for classification that is often confused with *k*-means because of the *k* in the name. One can apply the 1-nearest neighbor classifier on the cluster centers obtained by *k*-means to classify new data into the existing clusters. This is known as nearest centroid classifier or Rocchio algorithm.

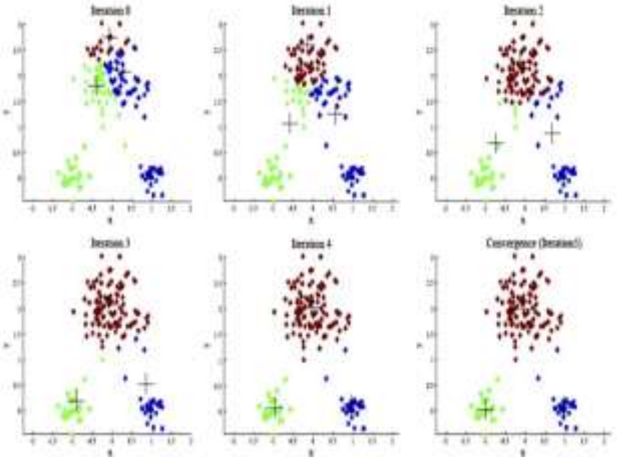


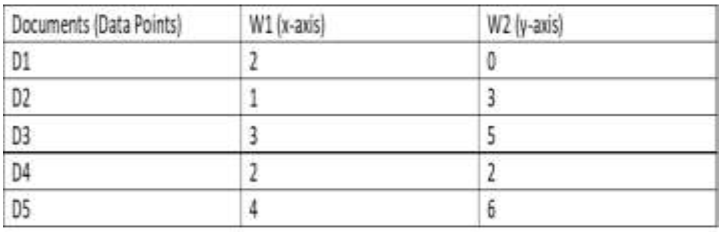
Fig1- K-Means Clustering Process

Steps to calculate centroids in cluster using K-means clustering algorithm

In this blog I will go a bit more in detail about the K-means method and explain how we can calculate the distance between centroid and data points to form a cluster.

Consider the below data set which has the values of the data points on a particular graph.

Table 1:



Documents (Data Points)	W1 (x-axis)	W2 (y-axis)
D1	2	0
D2	1	3
D3	3	5
D4	2	2
D5	4	6

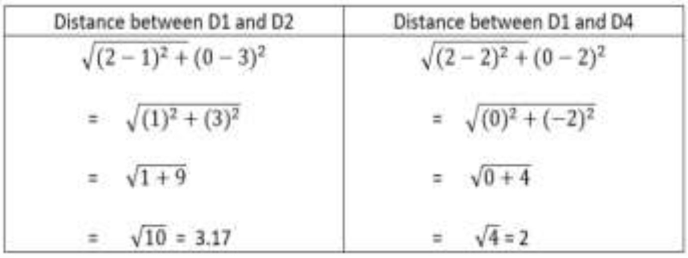
We can randomly choose two initial points as the centroids and from there we can start calculating distance of each point.

For now we will consider that D2 and D4 are the centroids. To start with we should calculate the distance with the help of Euclidean Distance which is

$$\sqrt{((x1-y1)^2 + (x2-y2)^2}$$

Iteration 1:

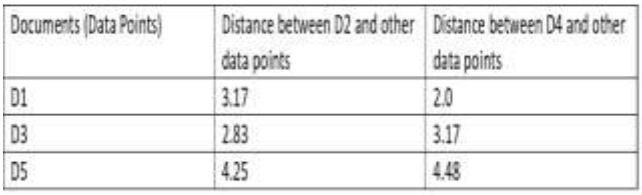
Step 1: We need to calculate the distance between the initial centroid points with other data points. Below I have shown the calculation of distance from initial centroids D2 and D4 from data point D1.



Distance between D1 and D2	Distance between D1 and D4
$\sqrt{(2-1)^2+(0-3)^2}$	$\sqrt{(2-2)^2+(0-2)^2}$
$= \sqrt{(1)^2+(3)^2}$	$= \sqrt{(0)^2+(-2)^2}$
$= \sqrt{1+9}$	$= \sqrt{0+4}$
$= \sqrt{10} = 3.17$	$= \sqrt{4} = 2$

After calculating the distance of all data points, we get the values as below.

Table 2

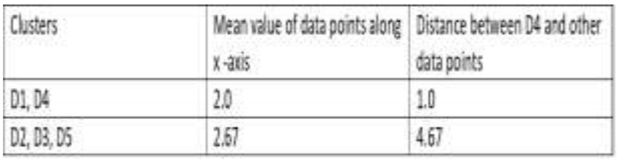


Documents (Data Points)	Distance between D2 and other data points	Distance between D4 and other data points
D1	3.17	2.0
D3	2.83	3.17
D5	4.25	4.48

Step 2: Next, we need to group the data points which are closer to centriods. Observe the above table, we can notice that D1 is closer to D4 as the distance is less. Hence we can say that D1 belongs to D4 Similarly, D3 and D5 belongs to D2. After grouping, we need to calculate the mean of grouped values from Table 1.

Cluster 1: (D1, D4) Cluster 2: (D2, D3, D5)

Step 3: Now, we calculate the mean values of the clusters created and the new centriod values will these mean values and centroid is moved along the graph.

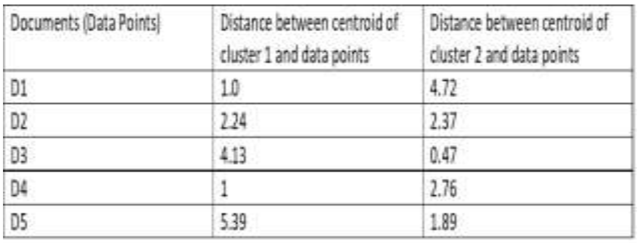


Clusters	Mean value of data points along x -axis	Distance between D4 and other data points
D1, D4	2.0	1.0
D2, D3, D5	2.67	4.67

From the above table, we can say the new centroid for cluster 1 is (2.0, 1.0) and for cluster 2 is (2.67, 4.67)

Iteration 2:

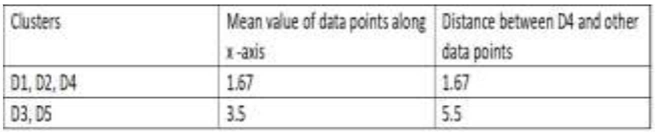
Step 4: Again the values of euclidean distance is calculated from the new centriods. Below is the table of distance between data points and new centroids.



Documents (Data Points)	Distance between centroid of cluster 1 and data points	Distance between centroid of cluster 2 and data points
D1	1.0	4.72
D2	2.24	2.37
D3	4.13	0.47
D4	1	2.76
D5	5.39	1.89

We can notice now that clusters have changed the data points. Now the cluster 1 has D1, D2 and D4 data objects. Similarly, cluster 2 has D3 and D5

Step 5: Calculate the mean values of new clustered groups from Table 1 which we followed in step 3. The below table will show the mean values



Clusters	Mean value of data points along x -axis	Distance between D4 and other data points
D1, D2, D4	1.67	1.67
D3, D5	3.5	5.5

Now we have the new centroid value as following:

cluster 1 (D1, D2, D4) - (1.67, 1.67) and cluster 2 (D3, D5) - (3.5, 5.5)

This process has to be repeated until we find a constant value for centroids and the latest cluster will be considered as the final cluster solution.

4. Conclusion

The electronic information generated by the utilization of EMRs in routine clinical follow has the potential to facilitate the invention of new information. Association rule mining coupled to a summarization technique provides an important tool for clinical analysis. It will uncover hidden clinical relationships and can propose new patterns of conditions to send determent, management, and treatment approaches. While all four strategies created cheap summaries, each methodology had its clear pith. However, not all of these strengths are essentially beneficial to this application. The founded necessary mortal between the algorithms is whether or not they use a range criterion to include a rule out the outline supported the expression of the rule or supported the patient population that the rule covers.

REFERENCES

1. Rakesh Agrawal and Ramakrishnan Srikant- Fast algorithms for mining association rules..
2. Mohammad Al Hasan- Summarization in pattern mining. In Encyclopedia of Data Warehousing and Mining.
3. Aysel Ozgur, Pang-Ning Tan, and Vipin Kumar. RBA: An integrated framework for regression based on association rules.
4. Peter W. Wilson, James B. Meigs, Lisa Sullivan, Caroline S.Fox, David M. Nathan, and Ralph B. D'Agostino- Pediction of incident diabetes mellitus in middle-aged adults–the Framingham offspring study".
5. Xiaoxin Yin and Jiawei Han- CPAR: Classification based on predictive association rules.