International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-5, Issue-8)

Research Article | **August 2016**

Estimate the Risk of Diabetes Mellitus Using Association Rule Mining

Prita Parshwanath Dongaonkar
M.E. Final Year Computer Science & Engineering, D.I.E.M.S., Aurangabad, Maharashtra, India

Ashwini Sanjay Gaikwad
Asst. Professor, Computer Science & Engineering Department, D.I.E.M.S., Aurangabad, Maharashtra, India

Abstract:

Diabetes mellitus is part of the growing epidemic of noninfectious diseases. Early detection of female patients with elevated risk of generating diabetes mellitus is critical to the improved prevention and overall clinical management of these patients. Aim to apply association rule mining to electronic medical records (EMR) to discover sets of risk factors. Given the high dimensionality ofelectronic medical records, association rule mining generates a very large set of rules for that it will be necessary to summarize for easy clinical use

Keywords: Data Mining, Association Rule, Distribution, Association Rule.

I. INTRODUCTION

Diabetes, or diabetes mellitus[1],[2] is a group of complex metabolic diseases which is characterized by high blood sugar (blood glucose) level in a person, either because inadequate production of insulin (insulin is made by the pancreas and lowers blood glucose), or inadequate sensitivity of cells to the action of insulin.In simple words, diabetes as an illness which occurs due to the problem in production and supply of insulin in the body. Level of sugar increased in the blood and causes a critical condition , if not controlled, can phase to very serious health complications and even also death. The risk of death for a person with the disease of diabetes is twice the risk of a person who have similar age who does not have diabetes.

Worldwide disease of diabetes is increasing quickly and in some countries it is reaching epidemic proportions. Diabetes mellitus is a growing epidemic that affect millions people in the Country. (8% of the population), and approximately 7 million of them do not know they have the disease. Diabetes leads to significant medical complications including heart disease, peripheral vascular disease. Early identification of patients at risk of generating diabetes is a major healthcare need. Appropriate management of patients at risk with changes in lifestyle and/or medications can decrease the risk of developing diabetes by 30% to 60%. Number of risk factors have been identified affecting a large proportion of the population. Prediabetes is good example, prediabetes (blood sugar levels above normal range level but below the level of criteria for diabetes) is present in near about 30% to 35% of the adult population and increases the absolute risk of diabetes 3 to 10 fold depending on the presence of any additional associated risk factors, such as obesity, hypertension, etc.

Diabetes is part of the metabolic syndrome [3], which is a constellation of diseases including hypertension (high blood pressure) and obesity (withbody mass index exceeding 30 kg/m2). These type of diseases interact with each other, with vascular and cardiac diseases and thus understanding these interactions is important.

II. LITERATURE SURVEY

A diabetes index is in essence a predictive model that assigns a score to a patient based on his estimated risk of diabetes. N. Kanagavalli, A .Vijayaraj and B. Sakthi Saravanan [4] analyses the risk factors and co morbidity conditions to detect diabetes early for that they use PubMed and EMBASE databases to identify and extract key information which are describes that aspects of developing a prediction model.

To find useful association rule in large data set it is a big challenge. A Krishnakumar and his co-worker[5] proposed several mining algorithms for finding association rules in transaction data based on the concept of large itemsets. They proposes a different interactive approach to prune and filterdiscovered rules for that they use ontology**.** The ontologies are described asdata schemas, providing a controlled concept of vocabulary, each with an explicitly defined and machineprocess able semantics.

Most of the mining technique focused on discovering association rules between items in a large database of sales transactions. R Agrawal and his co-workers[6]present two different algorithm to discovering association rule by using Apriori Algorithm and second is AprioriTid Algorithm. The Apriori and AprioriTid algorithms generatethe candidate itemsets to be counted in a pass byusing only the itemsets found large in the previous pass without considering the transactions in the database.

III. ASSOCIATION RULE

The aim of data mining is to extract higher level information from lower raw data. Association rules are a very important key tool used for this purpose. An association rule [7] is a rule of the form $X \Rightarrow Y$, where X and Y are rule. The rule states that with a certain probability, called the confidence of the rule, when X occurs in the given database so does Y.

Association rule mining is primarily focused on finding frequent co-occurring associations among a collection of items. The use of association rules is particularly beneficial, because in addition to quantifying the diabetes risk, they also readily provide the physician namely the associated set of conditions. This set of conditions can be used to guide treatment to personalized and targeted preventive care or diabetes management.

Let an item be a binary indicator signifying whether a patient possesses the corresponding risk factor. E.g. the item pres represents what exactly the pressure of patient . Let M indicate the item matrix, which is a binary covariate matrix with rows indicating patients and the columns indicating items. An itemset indicates whether the corresponding risk factors are all present in the patient. If they are, the patient is said to be covered by the itemset (or the itemset applies to a patient).

In association rule mining, items do not play particular roles: there are no designated outcome variables. In other words, any item can appear in the antecedent of one rule and in the consequent of another.

IV. DISTRIBUTION ASSOCIATION RULE

A **distributional association rule** is referred by an itemset *I* and is an implication that for a continuous outcome *y*, its distribution between the affected subpopulation and the unaffected subpopulations is *statistically significantly* different. For example, the rule *{*pres, plas*}* indicates that the patients both presenting diastolic blood pressure (high blood pressure) and plasma glucose concentration have a chance of progression to diabetes than the patients who are either not diastolic blood pressure or do not have plasma glucose concentration prescribed. Since each rule is defined

by an itemset, here use the words 'rule' and 'itemset' interchangeably.

The distributional association rules consists of two steps. In the first step, a suitable or a compatible set of itemsets is discovered and in the second step,the set of itemsets is filtered so that returned as distributional association rules only which they are the statistically significant.

Discovery of Itemset . Most if not all itemset in the outcome algorithms can be used to discover itemsets. For that in that paper, used the Apriori algorithm, a variant of the well-known Reorder algorithm that only discovers candidate itemsets that contain specific items—the item corresponding to the diabetes outcome in that case.

Testing Statistical Significance. For each and every discovered itemset, we have to test whether the outcome distribution in the affected subpopulation and the unaffected subpopulations are indeed different.

The distributional association rules are characterized by the following statistics. For rule R, let O_R denote the observed number of diabetes incidents in the subpopulation D_R covered by R. Let E_R denote the expected number of diabetes incidents in the subpopulation covered by R.

$E_R = O_R - i \in D_R y_i$,.

where y_i is the martingale residual for patient i.

he **relative risk** of a set of risk factors that define R is

$$R_R = O_R / E_R.$$

V. PROPOSED SCHEME

A. Attribute

In particular, all patients are females atleast 21 years old of Pima Indian heritage. Attributes are

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

B. Data Collection

Figure 1, shows Schematic diagram of the workflow. First, data were collected from database of diabetes patient which is of Pima Indian Heritage. This data set was obtained fromUCI Repository of Machine Learning Databases. In that personal information such as name, address, contact details was not collected.

Data mining in the proposed associative classification [8] first step is discretizing continuous attributes and second generating all the association rules.

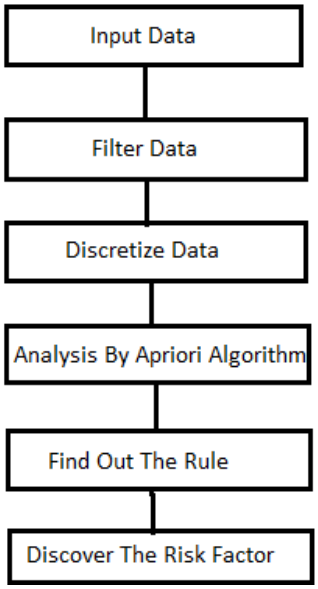

Figure 1. Schematic diagram of the workflow.

C. Analysis Method

In that paper, Apriori Algorithm use to analyze the association rule. The Apriori Algorithm is an Association Rule Mining (ARM) technique. It is very well known that mining algorithms can discover association rules; for instance, thousands of rules are extracted from a database of several numbers of attributes and several hundreds of transactions. An association rule is represent in the form of A→B, where A and B are both itemset. The rule represents implication that B itemset is apply to a patient which is also A applies. The itemset A is the antecedent and B is the consequent. An antecedent is an item found in the itemset. A consequent is an item that is found in combination with the antecedent. The rules are evaluated by Support and Confidence[9],[10].

$$Support = \frac{Number\ of\ Itemset\ A \cap B}{Total\ Number\ of\ Itemset}$$

$$Confidence = \frac{Number\ of\ Itemset\ A \cap B}{Number\ of\ Itemset\ A}$$

The support of an itemset the number of occurrences of itemset A and itemset B from all itemset and confidence of a rule the number of occurrences of itemset A co-occurring with itemset B.

Number of rules are slightly variant of each other rule leading in clinical patterns underlying the ruleset.Once get to this problem, which constitutes the main focus of this work, is to summarize the ruleset into a smaller set that is easier to overview. This paper first review the existing rule set and database summarization methods, then propose a generic framework that these methods fit into and finally, extend these methods so that they can take a continuous outcome variable (the martingale residual in our case) into account.

VI. SUMMARIZED RULE SET

Here present the rule sets generated by the extended summarization algorithms. For every algorithm, used the parameter setting that was provided the best results that was required.
For APRXCOLLECTION , we used $\alpha = 4$ and for RPGlobal, we used $\delta = 0.7$.

- **Aprx-Collection**

The APRX-COLLECTION [11] algorithm finds supersets of the condition in the rule such that most subsets of the summary rule will be valid rules in the original set and these subsets rule imply similar risk of diabetes.

- **Rpglobal**

The RPGlobal summarization is nearly equal to APRXCollection in that it is chiefly concerned with the expression of the rule, and hence it performs a very aggressive compression.

- **Patient Coverage**

The number of patients who are covered by any of the rules in the rule set A

- **Rule Set Summarization**

Aim of rule set summarization is to express a set I of rules with a smaller set A of rules such that I can be recovered from A with minimal loss of information.

Table 1, Rules Summarization in terms of relative risk RR, expected $E_{R,}$ observed$O_{R,}$

RR	E_R	O_R	Rule
1.17	103	88	Plas, mass, preg
2.14	210	98	Plas, pres, skin
4.34	426	98	Preg, pres, skin
1.47	130	88	Plas pres, skin, mass
0.80	83	103	Preg, skin, mass
0.91	82	90	Preg, mass, age
1.008	121	120	Preg, skin, pedi
2.95	260	88	Preg, pres, mass
0.94	90	95	Preg ,pedi, age
0.82	90	109	Preg, skin, pedi, age

- **Performance Analysis**

In order to improve performance, we must first describe it in measurable terms. So the classification accuracy is 73.82ss%. And also the model can take total time to build is up to 19 sec. In that paper use apriory algorithm to find successful result. It is a simple way to generate association rule than any other technique.

The Disadvantages include:

1] Less accuracy

2] Need more training data

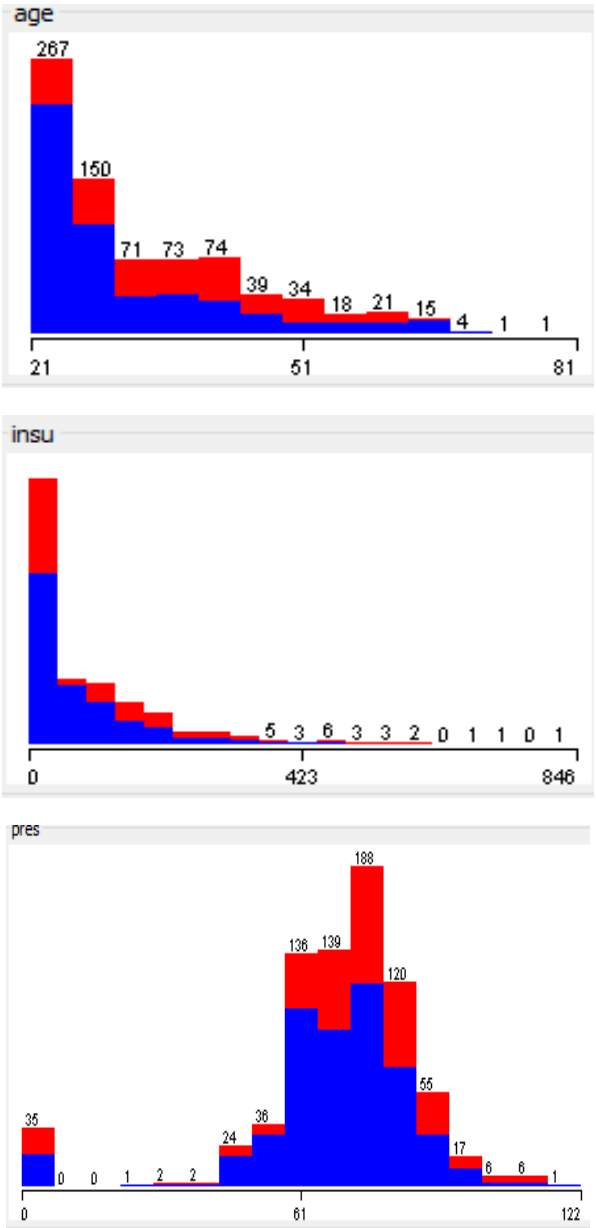

Figure 2. Visualization of three attribute between tested positive and negative.

VII. CONCLUSION

This study was significant because it was based on a large amount of data generated using electronic medical records in clinical use, aconstructed data mart, and analysis of the comorbidity of Data Mart using a program that automates the determination of the Apriori algorithm. However, a limitation of the present study is that the data came from a single medical institution. Data from many other medical facilities should be analysed and collected to demonstrate the relevance of the program and its results. Furthermore,the Apriori algorithm is limited in determiningprecedence or causality of disease.

This study indicated that data mining approach can predict the risk factors of Diabetes Mellitus.

An excessive number of association rules were discovered impeding the clinical interpretation of the results. For this method to be useful, the number of rules is used for clinical interpretation is make feasible. Future scope of studies is to identify the temporal complications of diseases considering chronology (e.g., the sequential pattern of disease occurrence) should be conducted.

REFERNCES

[1] K.Thulasi, S.Sowmiyaa, P.Prema. Predicting Relative Risk for Diabetes Mellitususing Association Rule SummarizationTechnique in EMR. International Journal of Innovative Research in Science,Engineering and Technology *(An ISO 3297: 2007 Certified Organization).*

[2] X.Rexeena, B.Suganya Devi, S.Saranya. Risk Assessment for Diabetes Mellitus using Association Rule Mining. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)

[3] J. Tuomilehto, J. Lindstr¨om, J. Eriksson, T. Valle, H. H¨am¨al¨ainen, P. Ilanne-Parikka, S. Kein¨anen-Kiukaanniemi, M. Laakso, A. Louheranta, M. Rastas, V. Salminen, and M. Uusitupa. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impared glucose tolerance. *The New EnglandJournal of Medicine*, 344(18), 2001.

[4] Ms. N. Kanagavalli , Mr. A .Vijayaraj, Mr. B. Sakthi Saravanan. Evaluation on risk factors and comorbidity conditions of diabetesusing mining algorithms and searching methods. IJAICT Volume 1, Issue 7, November 2014

[5] A.KrishnaKumar, D.Amrita, N.Swathi Priya.Mining Association Rules between Sets of Items in Large Databases. International Journal of Science and Modern Engineering (IJISME)ISSN: 2319-6386,Volume-1, Issue-5, April 2013

[6] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In VLDB Conference, 1994.

[7] Yonatan Aumann, Yehuda Lindell. A Statistical Theory for Quantitative Association Rules. In International conference on Knowledge Discovery and Data Mining, 2000

[8] Bing Liu Wynne Hsu Yiming Ma. Integrating Classification and Association Rule Mining. In In International conference on Knowledge Discovery and Data Mining 1998

[9] Hye Soon Kim, A. Mi Shin, Mi Kyung Kim, and Nyun Kim. Comorbidity study on type 2 diabetes mellitus using data mining. *Korean J Internal Medicine*, 27, 2012.

[10] Yuefeng Li and Jingtong Wu. Summarization of Association Rules in Multi-tierGranule Mining. IEEE Intelligent Informatics Bulletin December 2012 Vol.13 No.1.

[11] Foto Afrati, Aristides Gionis, and Heikki Mannila. Approximating a Collection of Frequent Sets. In International conference on Knowledge Discovery and Data Mining ,2004

