iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** V **Month of publication:** May 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Diabetes Mellitus Estimation Risk Using Association rule Mining

Prita Parshwanath Dongaonkar[1], Ms. Ashwini Gaikwad[2]
[1]M.E. Final Year, [2]Asst. Professor, Computer Science & Engg Dept., D.I.E.M.S., Aurangabad

Abstract:- Diabetes is part of the growing epidemic of non communicable diseases. Early detection of patients with elevated risk of developing 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 of EMRs, association rule mining generates a very large set of rules which we need to summarize for easy clinical use. The four methods summaries the high risk of diabetes.
Keywords:- Data Mining, Association Rule, Distribution, Association Rule.

I. INTRODUCTION

Diabetes Mellitus, or simply diabetes is a group of metabolic diseases in which a person has high blood sugar (blood glucose), either because the pancreas does not produce enough insulin, or because cells do not respond to the insulin that is produced. Glucose builds up in the blood and causes a condition that, if not controlled, can lead to serious health complications and even death. The risk of death for a person with diabetes is twice the risk of a person of similar age who does not have diabetes. Diabetes mellitus is a growing epidemic that affects 25.8 million people in the U.S. (8% of the population), and approximately 7 million of them do not know they have the disease. 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, prediabetes (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, hyperlipidemia, etc. Comprehensive medical management of this large portion of the population to prevent diabetes represents an unbearable burden to the healthcare system. In response to the pressing need to identify patients at high risk of diabetes early, numerous diabetes risk indices (risk scores) have been developed. Some of these indices (e.g. the Framingham score) gained acceptance in clinical practice and are used as guidance in treatment: patients presenting high risk scores are treated more aggressively. These scores only provide a quantification of the risk, they are not suggestive of the factors that may have caused the elevation of the risk. Moreover, these scores utilize individual risk factors in an additive fashion without taking interactions among them into account. 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.

II. LITERATURE SURVEY

The goal of data mining is to extract higher level information from an abundance of raw data. Association rules are a key tool used for this purpose. An association rule is a rule of the form $X \Rightarrow Y$, where X and Y are events. 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 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. Let an **item** be a binary indicator signifying whether a patient possesses the corresponding risk factor. E.g. the item htn indicates whether the patient has been diagnosed with hypertension. Let X denotes the item matrix, which is a binary covariate matrix with rows representing patients and the columns representing items. An itemset is a set of items: it 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).
An association rule is of form $I \rightarrow J$, where I and J are both itemsets. The rule represents an implication that if J is likely to

apply to a patient given that *I* apply. The itemset *I* is the antecedent and *J* is the consequent of the rule. The strength and "significance" of the association is traditionally quantified through the support and confidence measures. In association rule mining, items do not play particular roles: there are no designated predictor variables or outcome variables. In other words, any item can appear in the antecedent of one rule and in the consequent of another. Predictive association rule mining represented the first departure from this paradigm by designating a specific item as an outcome. The consequent of the predictive association rules is always the designated outcome item. Regressive association rules and quantitative association rules further expanded this paradigm allowing for a continuous outcome variable *y* to serve as the "consequent" of a rule.

III. DISTRIBUTION ASSOCIATION RULE

A distributional association rule is defined by an itemset *I* and is an implication that for a continuous outcome *y*, its distribution between the affected and the unaffected subpopulations is *statistically significantly* different. For example, the rule *{*htn, fibra*}* indicates that the patients both presenting hypertension (high blood pressure) and taking statins (cholesterol drugs) have a significantly higher chance of progression to diabetes than the patients who are either not hypertensive or do not have statins prescribed. Since each rule is defined by an itemset, we use the words 'rule' and 'itemset' interchangeably. The discovery of distributional association rules consists of two steps. In the first step, a suitable set of itemsets is discovered and in the second step, the set of itemsets is filtered so that only the statistically significant ones are returned as distributional association rules.

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

Testing Statistical Significance. For each discovered itemset, we have to test whether the outcome distribution in the affected 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. The relative risk of a set of risk factors that define R is

$$R_R = O_R/E_R.$$

IV. METHOD

Applying method of distribution association rule mining to produce a large number of rules. Many of these rules are slight variants of each other leading to the obfuscation of the clinical patterns underlying the ruleset

A. Rule Set Summarization

The goal of *rule set summarization* is to represent 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. Since a rule is defined by a single itemset, we will use 'itemset' in place of 'rule' meaning the 'itemset that defines the rule'.

B. Extension to Account for Outcom

In this section, we discuss how we extended these techniques to incorporate the risk y of diabetes as manifested by the martingale residual. Since we are particularly interested in rules that predict high risk of diabetes, we can add $\bar{y}(I)$ the subpopulation mean risk of diabetes to the criterion with a weight λ that controls how much importance is assigned to the risk and how much to the other components of the criterion. Let L∗(I) be the resulting criterion and L(I) the original criterion

$$L*(I) = -\lambda \bar{y}(I) + (1-\lambda)L(I)$$

C. Relative Patient Coverage

A rule *A* **covers** a rule *I* if $I \subset A$. Let the set S_A denote the rules in *I* that are covered by rule *A*
and let set D_A denote the patients in *D* that are covered by rule *A*. We also define a similarity function between two itemsets *A* and *I* in terms of their **relative patient coverage** (RPC) as

$$RPC(A,I) = \frac{|DA \cap DI|}{|DA \cup DI|}$$

V. CONCLUSION

This study was significant because it was based on a large amount of data generated using electronic medical records in clinical use, a constructed data mart, and analysis of the comorbidity of DM 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 other medical facilities should be collected and analyzed to demonstrate the relevance of the program and its results. Furthermore, the Apriori algorithm is limited in determining precedence or causality of disease. Therefore, future studies to identify the temporal complications of diseases considering chronology (e.g., the sequential pattern of disease occurrence) should be conducted.

REFERENCES

[1] 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.

[2] Gang Fang, Majda Haznadar, Wen Wang, Haoyu Yu, Michael Steinbach, Timothy R Church, William S Oetting, Brian Van Ness, and Vipin Kumar. High-order snp combinations associated with complex diseases: efficient discovery, statistical power and functional interactions. PLoS One, 7(4):e33531, 2012.

[3] Terry M. Therneau and Patricia M. Grambsch. Modeling Survival Data: Extending the Cox Model. Statistics for Biology and Health. Springer, 2010.

[4] Ruoming Jin, Muad Abu-Ata, Yang Xiang, and Ning Ruan. Effective and efficient itemset pattern summarization: Regressionbased approach. In ACM International Conference on Knowledge Discovery and Data Mining (KDD), 2008

[5]Varun Chandola and Vipin Kumar. Summarization – compressing data into an informative representation. Knowledge and Information Systems, 2006.

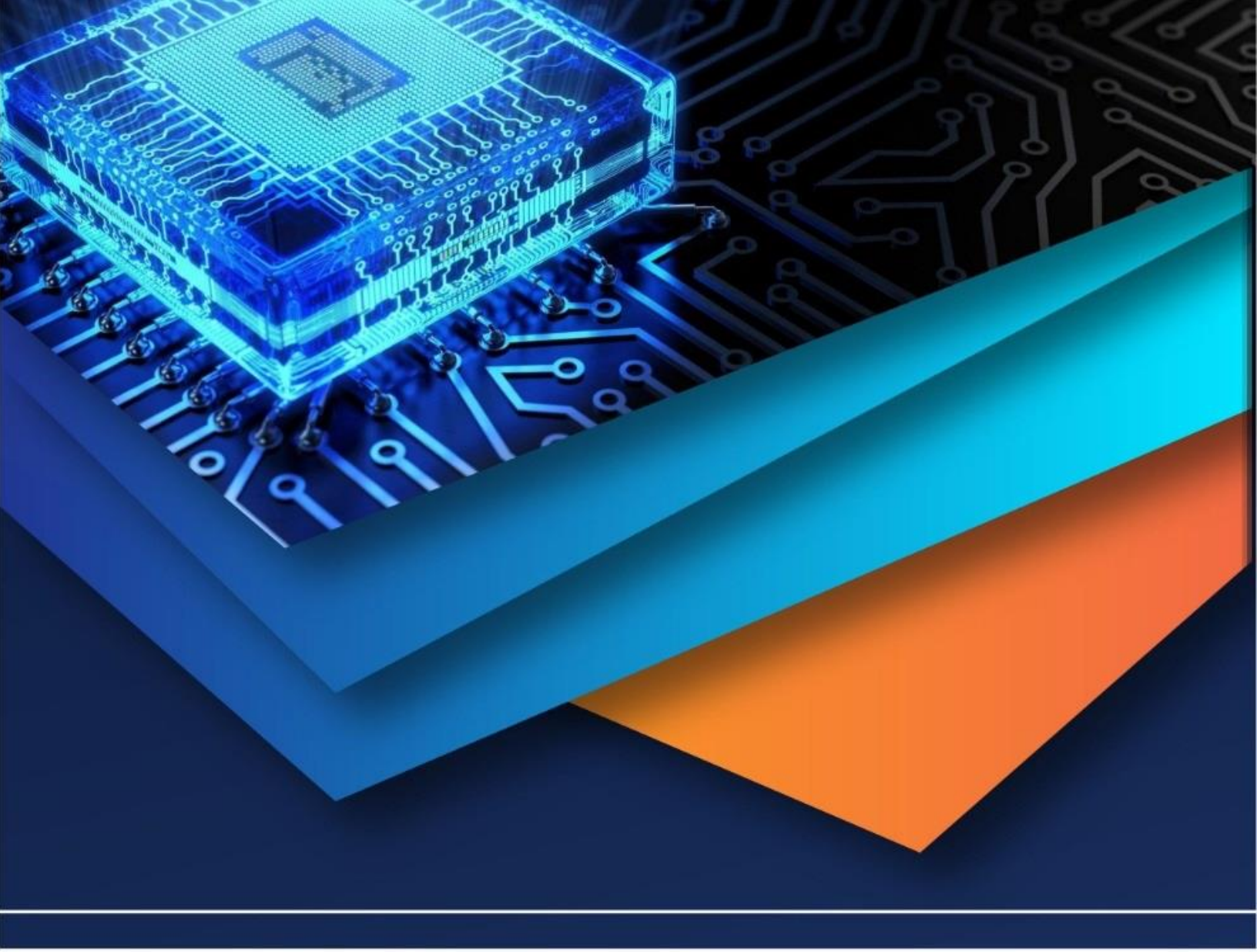
[6] Chao Wang and Srinivasan Parthasarathy. Summarizing itemset patterns using probabilistic models. In ACM International Conference on Knowledge Discovery and Data Mining (KDD), 2006.

[7]Foto Afrati, Aristides Gionis, and Heikki Mannila. Approximating a collection of frequent sets. In ACM International Conference on Knowledge Discovery and Data Mining (KDD), 2004.

[8]Aysel Ozgur, Pang-Ning Tan, and Vipin Kumar. RBA: An integrated framework for regression based on association rules. In SIAM International Conference on Data Mining (SDM), 2004.

[9] Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. The New England Journal of Medicine, 346(6), 2002.

[10] 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 England Journal of Medicine, 344(18), 2001.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY