Risk Assessment for Diabetes Mellitus using Association Rule Mining

X.Rexeena, B.Suganya Devi, S.Saranya

Abstract

Diabetes is part of the growing epidemic of non communicable diseases, with a high burden for the society on developing countries in future.For suppressing the development of diabetes mellitus and the onset of complications to manage their healthcare or personal data. We aim to apply association rule mining to electronic medical records to discover sets of risk factors. The four methods summaries the high risk of diabetes. Our extension to the bottom up summarization algorithm produced the most suitable summary.

Keyword-Data mining, Association rule mining, Distribution association rule,Fuzzy clustering means.

1 .Introduction

Diabetes is a group of diseases characterized by high blood glucose (blood sugar). When a person has diabetes, the body either does not produce enough insulin or is unable to use its own insulin effectively. 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 is a major cause of heart disease and stroke. Death rates for heart disease and the risk of stroke are about 2–4 times higher among adults with diabetes than among those without diabetes. In addition, 67% of U.S. adults who report having diabetes also report having high blood pressure.1 For people with diabetes, high blood pressure levels, high cholesterol levels, and smoking increase the risk of heart disease and stroke. This risk can be reduced by controlling blood pressure and cholesterol levels and stopping smoking. 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 [15]) gained acceptance in clinical practice and are used as guidance in treatment. 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.Association rule mining

Association rule mining, one of the most important and well researched techniques of data mining. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases .

Let an item be a binary indicator signifying whether a patient possesses the corresponding risk factor. Eg. The item htn indicates whether the patient has been diagnosed with hypertension. Let X denote 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→ 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 applies. 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 [13],[14] 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 [12] and quantitative association rules [3] further expanded this paradigm allowing for a continuous outcome variable *y* to serve as the "consequent" of a rule.

3. Existing system

Exiting systems aim to apply association rule mining to electronic medical records 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 we need to summarize for easy clinical use. We reviewed four association rule set summarization techniques and conducted a comparative evaluation to provide guidance regarding their applicability, strengths and weaknesses.

Technique:

- Four association rule
- Sequential coverage

4.Proposed architecture

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.

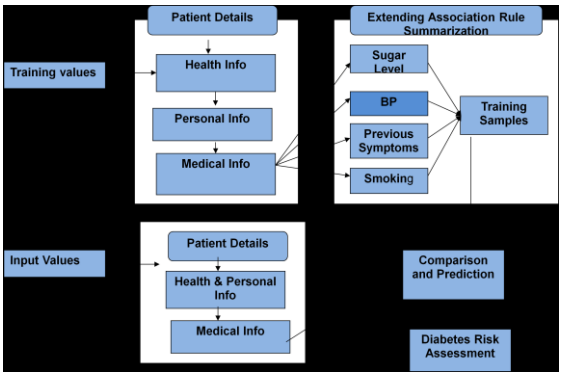

Fig1:overall description for risk assessment.

Technique:

- Ten association rule
- Fuzzy clustering means(FCM)

5. Distribution association rules

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.

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

$ER = OR - i \in DRyi,$.

where yi is the martingale residual for patient i. The **relative risk** of a set of risk factors that define R is

$RR = OR/ER$.

Table1: Description of the risk factors that appeared in any of the summarized rules.

Parameter	Weightage	Values
Male& Female	Age<30 >30to<50 Age>50&Age<70 Age>70	0.1 0.3 0.7 0.8
Smoking	Never Past Current	0.1 0.3 0.6
Overweight	Yes	0.8

	No	0.1
Alchol Intake	Never Past Current	0.1 0.3 0.6
Heart Rate	Low(<60bpm) Norma(60to100bpm) High(>100bpm)	0.9 0.1 0.9
Blood Sugar	High(>120&<400) Normal(>90&<120) Low(<90)	0.5 0.1 0.4
Bad Cholesterol	Very High>200 High(160to200) Normal<160	0.9 0.8 0.1

.6.METHOD

Many of these rules are slight variants of each other leading to the obfuscation of the clinical patterns underlying the ruleset. One remedy 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. We first review the existing rule set and database summarization methods, then propose a generic framework that these methods fit into and finally, we extend these methods so that they can take a continuous outcome variable (the martingale residual in our case) into account.

•Rule set and database 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'.

•Fuzzy Clustering Means

In fuzzy clustering data elements can belong to more than one cluster.The strength of the association between the data elements and a particular cluster.

In fuzzy clustering every point has a degree of belonging to as in fuzzy logic rather than belonging completely to just one cluster.

•Extension to Account for Outcome

. 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

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

•Patient Coverage

Patient coverage is simply the number of patients (or alternatively, cases) who are covered by any of the rules in the summary set A. The sum squared prediction error, coverage and restoration error (respectively) for each method as a function of the size of the summary rule subset . A summary rule subset A of size k consists of the first k rules in the summary rule set A. In each figure, the left pane corresponds to measurements only on cases, the right pane corresponds to measurements on all patients.

•Sum squared prediction error

We aim to assess how accurately a set of rules can predict the excess risk of diabetes for the patients (or only for the cases) relative to the full rule set. Towards this end, we need to first compute a "gold standard" estimate of each patient's risk $\tilde{y}i$ based on the entire original rule set I and then compare the estimate $\hat{y}i$ obtained using the summary rule set to $\tilde{y}$. We compute the "gold standard" estimate through a boosted linear regression model using cross-validation. The predictors of the model are rules in the original rule set I and the outcome is the martingale residual y. Given a summary rule set A, which is an ordered set of rules, we make a prediction for patient i through the first rule Ai that covers patient i. The predicted value is the subpopulation mean outcome on the training set.

$\hat{y}i = \bar{y}(Ai) = \text{mean}j \in Dai\ yj$.

The sum squared prediction error (SSPE) is the summed square difference between the risk predicted by the summary rule set $\hat{y}i$ and the gold standard estimate $\tilde{y}i$

$SSPE = I\ (\hat{y}i - \tilde{y}i)2.$

7.Summarized Rule Set

we present the rule sets generated by the extended summarization algorithms. For each algorithm, we used the parameter settings that provided the best results. For APRXCOLLECTION, we used $\alpha = .1$, $\lambda = 1$;

for RPGlobal, we used $\delta = .5$, $\sigma = .2$, $\lambda = .98$; for TopK, we used $\lambda = .2$; and for BUS, we used $\lambda = 1$. Note that λ significantly differs from 1 only for TopK, which already takes the risk of diabetes into account in the original loss criterion.

•APRX-COLLECTION

The APRX-COLLECTION algorithm finds supersets of the conditions (risk factors) in the rule such that most subsets of the summary rule will be valid rules in the original(unsummarized) set and these subset rules imply similar risk of diabetes.

Table2: Rule set summarized by apprx-collection

R	RR	ER	OR	RULE
1	1.96	36.24	71	fibra
20	1.34	271.71	363	bmi trigal acerab Statin aspirin htn
16	1.19	426.78	506	hdl trigl acearb aspirin htn
15	1.31	348.92	457	bmi trigal statin aspirin ihd
10	1.23	534.58	660	Bmi sbp ccb htn

•RPGLOBAL

The main drawbacks of APRX-COLLECTION were the redundancy in the rule set and the dilution of the risk. The RPGlobal summarization is similar to APRXCOLLECTION in that it is chiefly concerned with the expression of the rule, and hence it performs a very aggressive compression. However, it addresses the two drawbacks by taking patient coverage into account and by constructing the summary from rules in the original rule set (as opposed to an extended set). The summary created by RPGlobal is displayed in Table3.

Table3: Top 10 rules of the summarized rule set created by RPGlobal.

RR	ER	OR	RULE
1.69	32	55	bmi trigal acearb diuret htn
1.23	52	65	acearb bb diuret aspirin htn
1.29	42	55	sbp tchol acearb diuret htn
2.10	25	54	hdl trigal diuret aspirin htn
1.28	42	54	bmi tchol hdl trigl tobacco

•TOP-K

The Redundancy-Aware Top K (TopK) algorithm further reduces the redundancy in the rule set which was possible through operating on patients rather than the expressions of the rules. TopK still achieves high compression rate.

Table4: Top 10 summarized rule created by the top-k algorithm.

RR	ER	OR	RULE
2.40	21.70	52	fibra htn
1.58	37.97	60	bmi hdl ihd
1.47	45.52	67	sbp htn tobacco
1.46	317.03	464	bmi htn
1.62	32.16	52	sbp tchol trigal statin htn

•BUS

BUS (as opposed to TopK) operates on the patients and not on the rules. Therefore, redundancy in terms of rule expression can occur. However, BUS explicitly controls the redundancy in the patient space through the parameter mandating the minimum number of new (previously uncovered) cases (patients with diabetes incident) that need to be covered by each rule. Thus the reduced variability in the rule expression does not translate into increased redundancy.

Table5: Top 10 summarized rule created by bus.

RR	ER	OR	RULE
2.34	24	57	bmi trigal acearb statin htn
2.10	25	54	hdl trigal diuret aspirin htn
1.91	56	107	bmi trigal statin htn
1.54	78	121	bmi trigal tobacco
1.37	39	54	dbp diuret htn

8. Empirical Evaluation

We evaluate these methods on the Mayo Clinic patient data obtained during the study period from 1/1999 to 12/2004 with follow-up information available until the summer of 2010. We included patients who were at risk at any time during the study; that is, they did not have diabetes before the beginning of the study and they had fasting glucose level between 100 and 125 mg/dl.

9. Conclusion

Association rule mining to identify sets of risk factors and the corresponding patient subpopulations who are at significantly increased risk of progressing to diabetes.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.

REFERNCES

[1] Pedro J. Caraballo, M. Regina Castro, Stephen S. Cha, Peter W. Li, and Gyorgy J. Simon. Use of association rule mining to assess diabetes risk in patients with impared fasting glucose. In AMIA Annual Symposium, 2011.
[2] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In VLDB Conference, 1994.
[3] Yonatan Aumann and Yehuda Lindell. A statistical theory for quantitative association rules. In Knowledge Discovery and Data Mining, 1999.
[4] Varun Chandola and Vipin Kumar. Summarization – compressing data into an informative representation. Knowledge and Information Systems, 2006.
[5] Gary S Collins, Susan Mallett, Omar Omar, and Ly-Mee Yu. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Medicine, 2011.
[6] 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.
[7] 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.
[8] Mohammad Al Hasan. Summarization in pattern mining. In Encyclopedia of Data Warehousing and Mining, (2nd Ed). Information Science Reference, 2008.
[9]R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. In American Association for Artificial Intelligence (AAAI), 1997.
[10] Terry M. Therneau and Patricia M. Grambsch. Modeling Survival Data: Extending the Cox Model. Statistics for Biology and Health. Springer, 2010.
[11] 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.
[12] Aysel Ozgur, Pang-Ning Tan, and Vipin Kumar. RBA: An integrated framework for regression based on association rules. In SIAM International [13] Bing Liu, Wynne Hsu, and Yiming Ma.Integrating classification and association rule mining. In ACM International Conference on Knowledge [14] Xiaoxin Yin and Jiawei Han. CPAR: Classification based on predictive association rules. In SIAM International Conference on Data Mining (SDM), 2003.
[15] 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. *Archives of Internal Medicine*, 167, 2007.

X.Rexeena is a PG Scholar pursuing her Master of Engineering in Computer Science and Engineering at Ranganathan Engineering College,Coimbatore and she done her Bachelor of Engineering in Computer Science from SSK College of Engineering and Technology,Coimbatore.Her research interest's are Datamining,Network Security,Cloud Computing.

B.Suganya devi is an Assistant Professor in Ranganathan Engineering College.She received her BE degree from Coimbatore Institute of Engineering and Technology,Coimbatore.She received her ME degree from Coimbatore Institute of Engineering and Technology,Coimbatore.Her research interest's are Datamining,Networking.

S.Saranya is currently a PG scholar with computer science department at Anna university,Chennai.She received the BE degree from Roever Engineering College, perambalur in 2012 and pursuing the ME degree from Ranganathan Engineering College, Coimbatore in 2014.Her area of interest include Network security, Data mining.

