International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Predicting Relative Risk for Diabetes Mellitus Using Association Rule Summarization in EMR

R. Sangeetha[1], M. Vivekanantha Moorthy[2]

[1]Department of Information Technology, Department of Information Technology
SRM University, Chennai.SRM University, Chennai, India

[2]Assistant Professor, Department of Information Technology, Department of Information Technology
SRM University, Chennai.SRM University, Chennai, India

Abstract: ***Diabetes is a growing epidemic of non-communicable disease which affects most of the people in the world. In order to suppress the growth of diabetes mellitus we use association rule summarization to electronic medical records to discover set of risk factors and the corresponding sub-population which represents patients at particularly high risk of developing diabetes. Usually association rule mining generates large volume of data sets which we need to summarize for any medical record or any clinical use. We incorporate four methods to find the common factors which lead to high risk of diabetes all these four methods produced summaries that described sub populations at high risk of diabetes with each method having its clear strength. According to our purpose we use bottom up summarization (BUS) algorithm which produces more suitable summary.***

Keyword: data mining, association rule mining, survival analysis, association rule summarization

1. Introduction

Diabetes mellitus is growing epidemic disease which affects more than 25.8 million people and approximately 7 million of them do not know they have this disease. Usually diabetes is a group of diseases characterized by high blood sugar (blood glucose). When a person has diabetes the body either produces enough insulin or unable to use its own insulin effectively. When glucose gets build up in our blood, that glucose should be controlled or must be effectively used else it may to lead death. The risk of death of a person who has diabetes is twice as the person who does not have diabetes of same age.

The major complications of diabetes are heart disease and stroke. Adults with diabetes have heart disease death rates about 2 to 4 times higher than adults without diabetes the risk of stroke is 2 to 4 times higher among people with diabetes. It also leads to hypertension and 67% of diabetic patients have blood pressure greater than or equal to 140/90 millimetres of mercury or used prescription medication for hypertension. Diabetes is a leading cause of blindness among adults aged 20-74 years. About 60% 70% of people with diabetes have mild to severe forms of nervous system damage. The result of such damage include impaired sensation or pain in the feet or hands showed digestion of food in the stomach carpal tunnel syndrome or other nerve problem.

Almost 30% of people with diabetes aged 40 years or older have impaired sensation in the feet. Diabetes may also lead to complication during pregnancy, poorly controlled diabetes before conception and during the first trimesters of pregnancy among women with type 1.

Diabetes can cause major birth defects in 5% to 10% of pregnancy and spontaneous abortions in 15% to 20% of pregnancies. On other hand for a women pre-existing diabetes optimizing blood glucose levels before and during early pregnancy can be reduce the risk of birth defects in their infants. Poorly controlled diabetes during the second and third trimesters of pregnancy can result in excessively large babies posing a risk to both mother and child.

Association rule are implication that associate a set potentially interacting conditions (eg: high BMI and the presence of hypertension diagnosis). The use of association rules is particularly beneficial because in addition to quantifying the diabetes risk, the also readily provide the physician with a "justification" namely the associated set of conditions. These 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 was proposed by Rakesh Agrawal in 1994, this association rule mining was initially used for market basket analysis to find how items purchased by customers are related. Association rule mining mainly aims to extract interesting correlations frequent patterns associations or casual structures among sets of items in the transaction database[2].

Let an item be a binary indicator signifying whether possesses the corresponding risk factor. Eg: the item htn indicates whether the patient has been diagnosed with hypertension. Let X denotes the item matrix which is binary co-variant matrix with row representing patients and the columns representing items. An item set is a set of items which indicates whether the corresponding risk factors are all present in the patient, if they are the patient is said to be covered by item set.

An association rule is of form I→J whether I and J are both item set. The rule represents an implication that if J is likely to apply to a patient given that I applies, the item set I is the antecedent and J is the consequent of the rule. The support

of an item set is the number of patients covered by that item set and that confidence of a rule R:T→J, is the fraction of patients covered by J among those who covered by I.

In association rule mining items do not play particular roles this means there are no designated predicator variable or outcome variable. Association rule mining makes it possible to analyse the association between not only two diseases, but also among there are more comorbidities that can be calculated from existing statics. Predictive association rule mining [13] are also used for the purpose to find the relations, regressive association rule [15] and quantitative association rule [3] are used for further expanded paradigm.

3. Existing System

When association rule mining is applied to censoring data w may fail to obtain full information about the particular patient or they are chances to miss certain information about the patient. For example if a patient drops out the study, we may know that he does not develop diabetes during the time period we could observe them but we do not have whether he has developed diabetes at the end of the study. Hence the censoring technique does not update the information fully.

4. Proposed System

We try to use association rule mining to the electronic medical record (EMR); All the risk factor about a patient namely co-morbid disease and laboratory results and medications are being available in the EMR, there are less chances to miss details about a patient with the extensive set of risk factors the set of discovered risk becomes extremely large to overcome this we use rule set summarization technique which is used to compress the original rule set into a compact set. We use the following techniques 1. APRX-collection 2. RPG-global 3.TOPK 4.BUS.

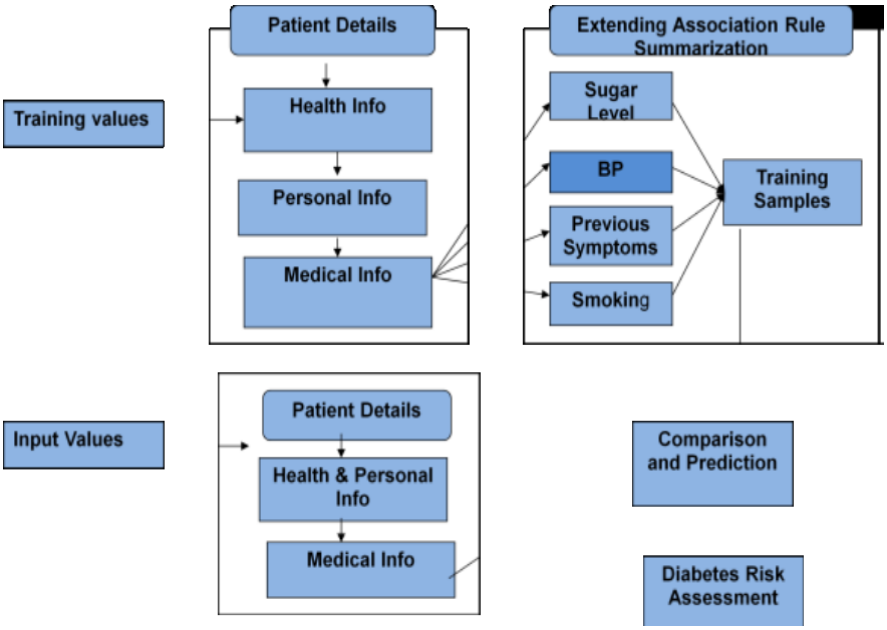

Figure 1: overall description of risk assessment.

Techniques used are
1. APRX-collection
2. RPGlobal
3. TOPK
4. BUS.
5. Navy based mining.

A. Distributional association rules

A distributional association rule is defined by an itemset I is an implication for a continuous outcome y, its distribution between the affected and unaffected subpopulation is statistically 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 hypertension or not have statins prescribed.Two steps are involved in finding the distributional association rule

Step-1 suitable set of itemset are discovered, the item sets are discovered from large items by using apirori algorithm we first count the support of individual items and determine which of them are large (ie.) we have minimum support. In each pass we start with a seed set of items found to be large in the previous pass. We use this seed set for generating new potentially large item sets called candidate key, item set and count the actual support for these candidate item set during the pass over the data.

At the end of the pass we determine which candidate item set are actually large and they become seed for not pass. This process contains until no new large item set are found. **Testing statistical significance:** for each discovered item set we have to test whether the outcome distribution in the affected and unaffected subpopulation is indeed different.

Step-2 the set of item set is filtered so that only the statistically significant ones are returned as distributional association rule, this rule is characterized by the following statistics from the number of item set collected. Let OR be the observed number of diabetes incident in the subpopulation DR covered by R. let ER denote the expected

number of diabetes incidents in the subpopulation covered by R.
ER = OR-iεDRy$_i$ where y$_i$ is the martingale for patient.

The relative risk factor is defined by R
RR = OR/ER.

Table 1: Description of the risk factors that appeared in any of the summarized rules

Parameter	*Weightage*	*Values*
Male & Female	Age<30 >30to<50	0.1 0.3 0.7 0.8
Smoking	Never Past Current	0.1 0.3 0.6
Overweight	Yes No	0.8 0.1
Alcohol intake	Never Past Current	0.1 0.3 0.6
Heart rate	Low(<60 bpm) Normal(60 to 100bpm) 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(160 to 200) Normal<160	0.9 0.8 0.1

B. Research Method
When we try to apply distributional rule mining with our electronic medical records it produced a large number of (statistically significant) rules. Rules that were generated slightly differ from each other leading to obfuscation of clinical patters. In order to overcome the problem of this large number of rules which were generated we go for summarizing the rule set into smaller set for our easier overview. We first review the existing rule set and database summarization methods then we try to incorporate a generic framework in order to get a continuous outcome of variable into account.

C. Rule set and database summarization
The main aim of rule set summarization technique is to represent a set I of rules with smaller set A of rules such that I can be recovered from A with minimal loss of information. Data base summarization technique is used to summarize a large database into smaller set of database A of item set such that the data set can be recovered from A with minimal loss.

D. Navy based mining technique
Navy based technique is used to mine an item set (large risk factors of diabetes) from two or more number of different larger database example after collecting database of patients from two different hospitals we used naval based mining technique to extract the highest risk factor of diabetes. (ie. Major reason for getting diabetes and there symptoms)

E. Extension to account for outcome
Here we discuss how to extend technique to incorporate the risk y of diabetes manifested by the martingale residual. Since we are particularly interested in rules that predict high risk of diabetes we can add –y(I) the subpopulation mean risk of diabetes to the criterion with a weight x 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 L(I) the original criterion.
L*(I) = $-\lambda y(I)+(1-\lambda)L(I)$

F. Summarization rule set
Now we present the rule set generated by the extended summarization algorithm, for each algorithm we used the parameter setting that provided the best results for APRX-collection we used $\alpha = 0.1, \lambda = 1$ for RPG global we used $\delta = 0.5, \sigma = 0.2, \lambda = 0.98$ for top K we used $\lambda = 0.2$ and for BUS we used $\lambda = 0.1$

Note: λ differs from 1 only for top K which already takes the risk of diabetes into account in the original loss criterion.

1. APRX collection
The APRX collection algorithm is used to find the supersets of the condition (risk factor) in the rule such that most subsets of summary rule will be valid rules in the original (un summarized) set and these subset rules imply similar risk for diabetes.

The APRX collection concentrates only on expression of the rule hence it lacks information about which patients are already covered as a result patients can get covered by multiple rules leading to rules with very similar condition this method also lacks in precision and information about very high risk subgroups.

Table 2: 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

2. 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 APRX collection n in that it is chiefly concerned with the expression of the rule and hence it performs a very aggressive compression. RPG global has two drawbacks by taking Patient coverage into account and by constructing the summary from rules in the original rule set.

Table 3: Top10 rules of the summarized rules 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

3. TOP-K

TOP-K algorithm reduces the redundancy in the rule set which was possible through operating on patients rather than the expression of the rules. This approach forfeited the outstanding compression rates of previous two algorithm TOP-K still achieves high compression rate and it successfully identified rules with high risk and low redundancy.

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 tobacoo
1.46	317.03	464	Bmi htn
1.62	32.16	52	Sbp tchol trigal statin htn

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

4. BUS

This summarises which are produces by BUS and TOP-K are of similar quality. BUS operates on patient not on rules, therefore redundancy in terms of rule expression can occur. BUS explicitly controls the redundancy in the patient space through the parameter mandating the minimum number of new (previously unoccured) 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.

Table 5: 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

5. Object Evaluation

We use 2 objective measure to evaluate the four summarization techniques. These measures are sum squared prediction error, restoration error and patient coverage.

A. 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 ˜yi based on the entire original rule set I and then compare the estimate ˆyi obtained using the summary rule set to ˜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.

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

The sum squared prediction error (SSPE) is the summed square difference between the risk predicted by the summary rule set ˆyi and the gold standard estimate ˜yi SSPE =I (ˆyi − ˜yi)2.

B. 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.

6. Conclusion and Future Work

The data that are generated by electronic medical record in routine clinical practice has the potential to facilitate the discovery of high risk factors of diabetes by using all four techniques. Currently we are finding the risk factor of diabetes for current smoker and heart disease, in future we try to find the risk factor of diabetes for all disease.

References

[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.