PREDICTING RELATIVE RISK FOR DIABETESMELLITUS USING ASSOCIATION RULE SUMMARIZATION TECHNIQUES

Thanushka.M.V[1], Sangeetha.P[2], Suhasini.A[3], TamilElakkiya.P[4],

Mrs. A.Vinothini[5]

[1,2,3,4] Department of IT, Panimalar Engineering College, Tamil Nadu, (India)

[5]Assistant Professor, Department of IT, Panimalar Engineering College, Tamil Nadu, (India)

ABSTRACT

The detection of diabetes mellitus at the earlier stages is difficult in clinical management.In an existing system, apriori algorithm is used to find the item sets for association rules .But it is not efficient in finding item sets and it uses only four association rules. In this paper we aim to maintain a EMR (Electronic Medical Record) and apply association rule mining to discover sets of risk factors and their corresponding subpopulations. We reviewed four association rule summarization techniques and conducted comparative evaluation based on their advantages and disadvantages.These foursummarization methods having its fair strength but the BUS (Bottom Up Summarization) algorithm developed the best acceptable summary.

Index Terms: Data Mining, Association Rules, Survival Analysis, Association Rule Summarization Techniques

I INTRODUCTION

Diabetes mellitus, commonly referred to as diabetes is a group of metabolic diseases in which there are high blood sugar levels over a prolonged period. It affects 25.8 million people in the U.S. Approximately 7 million of the people do not know they have the disease. Serious health complications such as stroke and may occur if not controlled properly. Diabetes is the major reason for heart diseases.

As of 2014, totally 387 million people have diabetes worldwide. This is equal to 8.3% of the adult population. In the years 2012 to 2014, diabetes is appraisal to have resulted in 1.5 to 4.9 million deaths per year. Diabetes doubles the risk of death. The number of people with diabetes is anticipated to rise to 592 million by 2035.

There are three main types of diabetes mellitus. Type 1 diabetes mellitus results from the body failure to produce enough insulin. Type 2 diabetes mellitus begins with insulin resistance, a condition in which cell fail to produce insulin properly. This may result in lack of insulin. The primary cause of this type of diabetes is excessive body weight and not enough exercise. Gestational diabetes is the third form of diabetes which occurs when pregnant women develop a high blood glucose level.

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 association rules is important in order to quantify the diabetes risks which also provide the physician with a “justification”, namely the associated set of conditions. This set of conditions can be used to provide treatment towards a more personalized and targeted preventive care or diabetes management.

II EXISTING SYSTEM

In an existing system, the patient records are recorded manually. Because of this the full information of the patient cannot be obtained. This method is called as**Censoring.** If a patient drops out of the study, we may not know if hegets diabetes at the end of the study. The ability to use partial information is the key characteristics of survival analysis making it a mainstay technique in clinical research.

III PROPOSED ALGORITHM

The original rule set available in the Electronic Medical Record(EMR) are compressed using the four rule set summarization techniques namely APRX-COLLECTION, RPGlobal, TopK,BUS to predict the Relative Risk of Diabetics Mellitus of patients. The applicability and strength of the Association rule set summarization techniques have been proposed .But it cannot provide the exact results. The four summarization techniques enables the practitioners in choosing the most suitable one. Between TopK and BUS, we found that BUS retained slightly more redundancy than TopK,. Top K has better ability and patient coverage. Thus BUS has been made the best suited algorithm for these purposes.

IV ASSOCIATION RULE MINING

Association rule mining, one of the most important and well researched techniques of data mining. It aspires to extract interesting correlations, frequent patterns and associations among sets of items in the transaction databases. 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 denote the item matrix, which is a binary covariate matrix with rows representing patients and the columns representing items. An item set 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 item set (or the item set applies to a patient).An association rule is of form $I\rightarrow J$, where I and J are both item sets. The rule represents an implication that if J is likely to apply to a patient given that I apply. The item set 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.

V DISTRIBUTIONAL 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. 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.

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

Whereyi is the martingale residual for patient i.

The relative risk of a set of risk factors that define R is RR =OR/ER.

```
Input: Set $I$ of item sets, number $k$ of summary rules
Output: Set $A$ of item sets, s.t.$A$ minimizes the criterion $L$
Generate an extended set $E$ of item sets based on $I$
$A$ = $
while|$A$| <$k$ do
    $A = \text{argmin}E\in E\ L(E)$
    Add $A$ to $A$
    Remove the effect of $A$
end while
```

VI 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.

6.1 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‟.

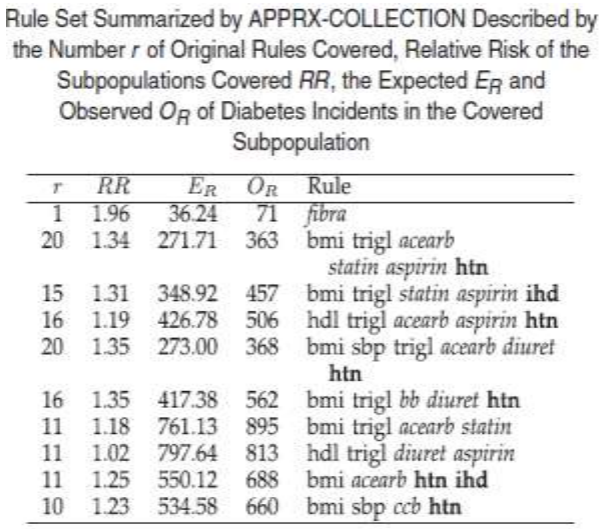
VII SUMMARIZATION TECHNIQUES AND SUMMARIZED RULE SET

Summarization is a key data mining concept which involves techniques for finding a compact description of a dataset. Simple summarization methods such as tabulating the mean and standard deviations are often applied for analysis of data, visualization of data and automated report generation.Four summarization techniques are used. we present the rule sets generated by the extended summarization algorithms. For each one algorithm, it provided the best suitable outcome because we used the parameter settings. For APRXCOLLECTION, we used $\alpha = .1$, $\lambda = 1$; for RPGlobal, we used $\delta = .5$, $\sigma = .2$, $\lambda = .98$; for Top-K, we used $\lambda = .2$; and for BUS, we used $\lambda= 1$. Note that λ notably varies from 1 single for Top-K, which previously takes the risk of diabetes into relation in the usual loss condition.

VIII SUMMARIZARTION TECHNIQUES

8.1 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.

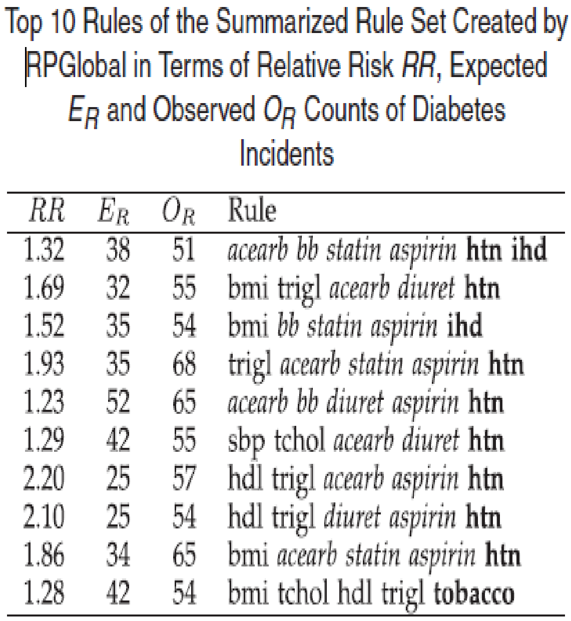


Rule Set Summarized by APPRX-COLLECTION Described by the Number *r* of Original Rules Covered, Relative Risk of the Subpopulations Covered *RR*, the Expected E_R and Observed O_R of Diabetes Incidents in the Covered Subpopulation

r	RR	E_R	O_R	Rule
1	1.96	36.24	71	*fibra*
20	1.34	271.71	363	bmi trigl *acearb* *statin aspirin* **htn**
15	1.31	348.92	457	bmi trigl *statin aspirin* **ihd**
16	1.19	426.78	506	hdl trigl *acearb aspirin* **htn**
20	1.35	273.00	368	bmi sbp trigl *acearb diuret* **htn**
16	1.35	417.38	562	bmi trigl *bb diuret* **htn**
11	1.18	761.13	895	bmi trigl *acearb statin*
11	1.02	797.64	813	hdl trigl *diuret aspirin*
11	1.25	550.12	688	bmi *acearb* **htn ihd**
10	1.23	534.58	660	bmi sbp *ccb* **htn**

8.2 RP Global

APRX-COLLECTION has some major limitations such as redundancy and intensity of risk. The RP Global mainly uses the rule expression. It also has two main drawbacks such as taking the exposure of patients into relation and creating summary from rules.

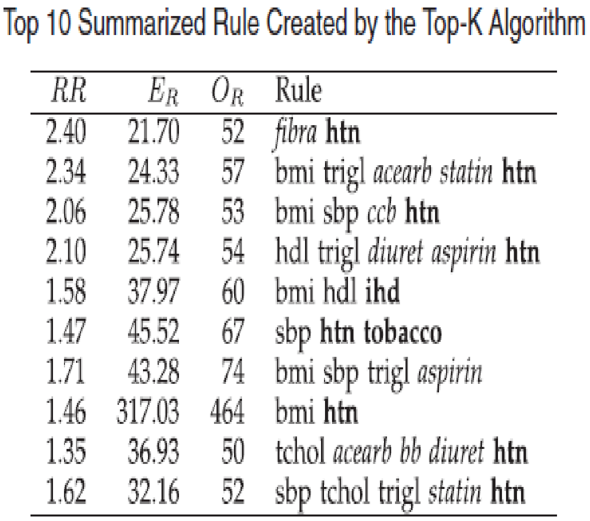


Top 10 Rules of the Summarized Rule Set Created by RPGlobal in Terms of Relative Risk *RR*, Expected E_R and Observed O_R Counts of Diabetes Incidents

RR	E_R	O_R	Rule
1.32	38	51	*acearb bb statin aspirin* **htn ihd**
1.69	32	55	bmi trigl *acearb diuret* **htn**
1.52	35	54	bmi *bb statin aspirin* **ihd**
1.93	35	68	trigl *acearb statin aspirin* **htn**
1.23	52	65	*acearb bb diuret aspirin* **htn**
1.29	42	55	sbp tchol *acearb diuret* **htn**
2.20	25	57	hdl trigl *acearb aspirin* **htn**
2.10	25	54	hdl trigl *diuret aspirin* **htn**
1.86	34	65	bmi *acearb statin aspirin* **htn**
1.28	42	54	bmi tchol hdl trigl **tobacco**

8.3 TOP-K

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

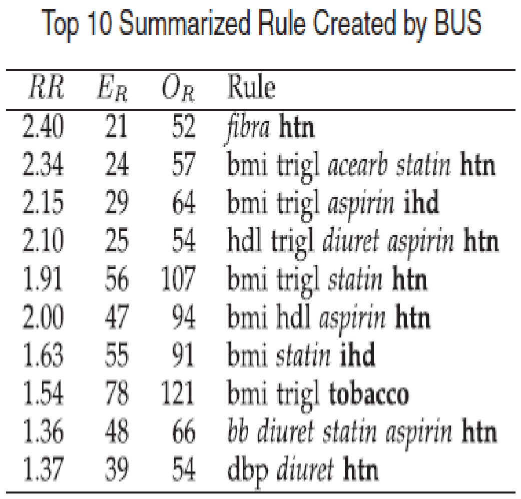


Top 10 Summarized Rule Created by the Top-K Algorithm

RR	E_R	O_R	Rule
2.40	21.70	52	*fibra* **htn**
2.34	24.33	57	bmi trigl *acearb statin* **htn**
2.06	25.78	53	bmi sbp *ccb* **htn**
2.10	25.74	54	hdl trigl *diuret aspirin* **htn**
1.58	37.97	60	bmi hdl **ihd**
1.47	45.52	67	sbp **htn tobacco**
1.71	43.28	74	bmi sbp trigl *aspirin*
1.46	317.03	464	bmi **htn**
1.35	36.93	50	tchol *acearb bb diuret* **htn**
1.62	32.16	52	sbp tchol trigl *statin* **htn**

8.4 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.

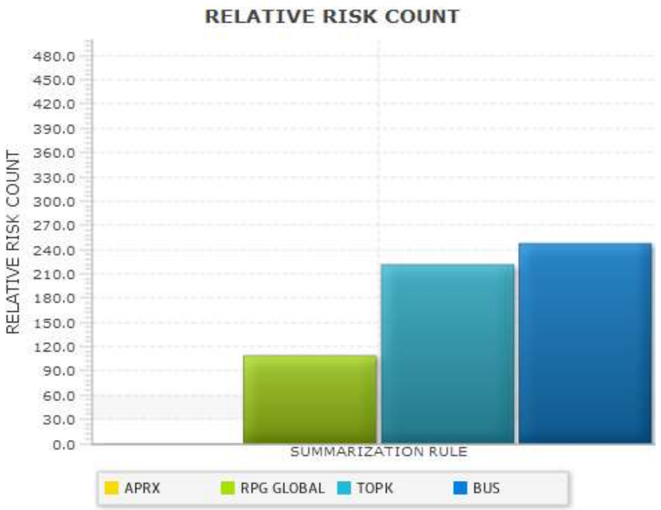


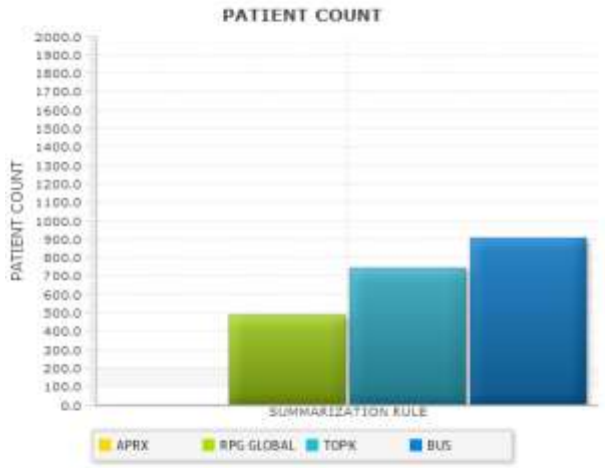
Top 10 Summarized Rule Created by BUS

RR	E_R	O_R	Rule
2.40	21	52	*fibra* **htn**
2.34	24	57	bmi trigl *acearb statin* **htn**
2.15	29	64	bmi trigl *aspirin* **ihd**
2.10	25	54	hdl trigl *diuret aspirin* **htn**
1.91	56	107	bmi trigl *statin* **htn**
2.00	47	94	bmi hdl *aspirin* **htn**
1.63	55	91	bmi *statin* **ihd**
1.54	78	121	bmi trigl **tobacco**
1.36	48	66	*bb diuret statin aspirin* **htn**
1.37	39	54	dbp *diuret* **htn**

IX RESULTS

Our proposed technique aims to predict the risk of diabetes mellitus. In this we use four association rule summarization techniques such as APRX-COLLECTION, RP Global, Top K and BUS. All these techniques have its own strength but BUS algorithm is the most efficient one.





X CONCLUSION

Association rule mining to identify sets of risk factors and the corresponding patient subpopulations that are at significantly increased risk of progressing to diabetes. An excessive number of associationrules were discovered impeding the clinical interpretation results. For this method, 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 impaired fasting glucose. In AMIA Annual Symposium, 2011.

[2] RakeshAgrawal and RamakrishnanSrikant.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] VarunChandola 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, MajdaHaznadar, 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.InEncyclopedia 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 NingRuan. Effective and efficient itemset pattern summarization: Regressionbased approach. In ACM International Conference on Knowledge Discovery and Data Mining (KDD), 2008.

[12] AyselOzgur, 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 YimingMa.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.

BIOGRAPHY

A.VINODHINI, Assistant Professor, Department of IT, Panimalar Engineering College,Poonamallee, Chennai, Tamil Nadu, India.

M.V.THANUSHKA,is Final Year student in Department of Information Technology at Panimalar Engineering College. She is the member of CSI. We presented papers in symposium and attend many workshops.

P.SANGEETHA,is Final Year student in Department of Information Technology at Panimalar Engineering College. She is the member of CSI. We presented papers in symposium and attend many workshops.

A.SUHASINI,is Final Year student in Department of Information Technology at Panimalar Engineering College. She is the member of CSI. We presented papers in symposium and attend many workshops.

P.TAMIL ELAKKIYA,is Final Year student in Department of Information Technology at Panimalar Engineering College. She is the member of CSI. We presented papers in symposium and attend many workshops.