INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

ARS: ASSOCIATION RULE SUMMARIZATION TECHNIQUES TO DETECT RISK OF DIABETES MELLITUS

V. Kavitha[1], R. Mohan[2]

[1]*Department of Computer Science, Mailam Engineering College, Mailam, Tamil Nadu, India, ssuganyacollegemylam@gmail.com*
[2] *Department of Computer Science, Mailam Engineering College, Mailam, Tamil Nadu, India*

Abstract

The main aim of this project is to divine the excess risk of diabetes for the patients and summarize their subpopulation by using Association Rule Mining. In Data Mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. To Apply Association Rule Mining to electronic medical records (EMR) to discover sets of risk factors and their corresponding subpopulations that represent patients at particularly high risk of developing diabetes. An Electronic Medical Record (EMR) is an progressing concept defined as a routine collection of electronic health information about individual patients or population. The high dimensionality of EMR's, association rule mining generates a very large set of rules which we need to summarize for easy clinical use. Applied four association rule set summarization techniques and conducted a comparative evaluation to provide guidance regarding their applicability, strengths and weaknesses. We found that all four methods produced summaries that described subpopulations at high risk of diabetes with each method having its clear strength. For our purpose, our extension to the Bottom-Up Summarization (BUS) algorithm produced the most suitable summary.

Keywords*:* Data mining, Electronic medical records, Bottom-up summarization

1. Introduction

Diabetes mellitus is a growing outbreak that affects 25.8 million people in the U.S. (8% of the population), and imprecise 7 million of them do not know they have the disease [1]. Diabetes leads to significant medical complications including ischemic heart disease, hit, nephropathy, retinopathy, neuropathy and outer 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% [2] [3]. Multiple risk factors have been identified affecting a large amount 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 existence of additional associated risk factors, such as obesity, hypertension, hyperlipidaemia, etc. [4]. Comprehensive medical management of this large portion of the population to prevent diabetes represents an intolerable 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 scores [5]) 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 hyperlipidaemia (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 modelling these interactions is important.

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 related set of conditions. This set of conditions can be used to pilot treatment towards a more personalized and targeted preventive care or diabetes management.

While association rules themselves can be easily interpreted, the resulting rule sets can sometimes be very large, eroding the interpretability of the rule set as a full. Especially, in this work, we consider a rich set of risk factors, namely co-morbid sickness, laboratory results, medications and demographic information that are commonly available in electronic medical record (EMR) systems. With such an extensive set of risk factors, the set of discovered rules grows combinatorially large, to a size that severely hinders interpretation. To overcome this challenge, we applied rule set summarization techniques to compress the original rule set into a more compact set that can be interpreted with ease.

A number of successful association rule set summarization techniques have been proposed [6] [7] but no clear guidance exists regarding the applicability, strengths and weaknesses of these techniques. The focus of this manuscript is to review and characterize four existing association rule summarization techniques and provide guidance to practitioners in choosing the most suitable one. A common shortcoming of these techniques is their inability to take diabetes risk–a continuous outcome–into account. In order to make these techniques more appropriate, we had to minimally modify them: we extend them to incorporate information about continuous outcome variables.

2. Related Works

In an existing system, a statistical modelling technique that constructs predictive models on time-to-event data under censoring the patient records manually. Censoring takes place when we fail to obtain full information about a patient. For example, if a patient drops out of the study, we may know that he did not develop diabetes during the time period we could observe him, but we do not know whether he ultimately developed diabetes by the end of the study. The ability to use such partial information and the ability to take time into account are the key characteristics of survival analysis making it a mainstay technique in clinical research [8] [9].

To apply rule set summarization techniques namely APRX-COLLECTION, RPGlobal, TopK, BUS to compress the original rule set commonly available in electronic medical record (EMR) systems to predict the Relative Risk of Diabetics Milletus of patients in the subpopulation. Association rule set summarization techniques have been proposed but no clear guidance exists regarding the applicability, strengths and weaknesses of these techniques. The focus of this manuscript is to review and characterize four association rule summarization techniques and provide guidance to practitioners in choosing the most acceptable one. To present a clinical application of association rule mining to identify sets of Body conditions, Medications and Co morbidities. To analyze these Factors by applying summarization techniques to predict the Risk of Diabetes. Between TopK and BUS, we found that BUS retained slightly more redundancy than TopK, which allowed it to have better patient coverage and better ability to reconstruct the original data base. This advantage made BUS the best suited algorithm for these purpose [10].

3. Methods

Initially in our application there is no Database Patient Records. We are going to implement summarization techniques in a Distributed Database not only in a single database. So we have to ask permission to access the database of each Health Center Administrator [11-15].

Collect those patients Records and Fetch in our application with privacy preservation. Fetching only Patient details which are not relevant to any personal information which comes under privacy preserving The Specific Patient can be identified by means of their ID itself [16-20].

In our third module to apply rule set summarization techniques namely APRX-COLLECTION, RPGlobal, TopK, BUS to predict the Risk of Diabetics Millets. Predication of Diabetics Millets based on Body condition, Medication and Co., Morbites of the patient subpopulation. While all four methods created reasonable summaries, each method had its clear strength. We found that the most important differentiator between the algorithms is whether they use a selection criterion to include a rule in the summary based on the expression of the order or based on the patient subpopulation that the rule covers. APRX-COLLECTION and RPGlobal primarily operate on the expression of the rules with a primary objective of maximizing compression. TopK and BUS operate primarily on the patients and their objective–especially in case of TopK–can be thought of as minimizing redundancy. Between TopK and BUS, we found that BUS retained slightly more redundancy than TopK, which allowed it to have better patient coverage and better ability to reconstruct the original data base. This advantage made BUS the best suited algorithm for our purpose [21] [22].

4. Conclusion

The electronic data generated by the use of EMRs in routine clinical practice has the potential to facilitate the discovery of new knowledge. Association rule mining coupled to a summarization technique provides a critical tool for clinical research. It can uncover hidden clinical relationships and can propose new patterns of conditions to redirect prevention, management, and treatment approaches.
In our specific example, we used distributional 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 needed to be reduced to a level where clinical interpretation is feasible.
To this end, we studied four methods to summarize these rules into sets of 10-20 rules that clinical investigators can evaluate [23] [24].
While all four methods created reasonable summaries, each method had its clear strength. However, not all of these strengths are necessarily beneficial to our application.
We found that the most important differentiator between the algorithms is whether they use a selection criterion to include a rule in the summary based on the expression of the rule or based on the patient subpopulation that the rule covers.
APRX-COLLECTION and RPGlobal primarily operate on the expression of the rules with a primary objective of maximizing compression. They use *representative rules*, each of which represents a number of original rules. Such representative rules achieve very high compression, but dilute the risk of diabetes over the typically large subpopulation they cover. TopK and BUS operate primarily on the patients and their objective–especially in case of TopK–can be thought of as minimizing redundancy. They produced good summaries because a beneficial side effect of reducing redundancy is to achieve good compression. The converse is not true: high compression rate does not result in low redundancy. Between TopK and BUS, we found that BUS retained slightly more redundancy than TopK, which allowed it to have better patient coverage and better ability to reconstruct the original data base. This advantage made BUS the best suited algorithm for our purpose.

REFERENCES

[1] 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.
[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] 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.

[5] Centers for Disease Control and Prevention. National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the united states, 2011. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention http://www.cdc.gov/diabetes/pubs/factsheet11.htm, 2011.
[6] Varun Chandola and Vipin Kumar. Summarization – compressing data into an informative representation. *Knowledge and Information Systems*, 2006.
[7] 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.
[8] 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.
[9] 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.
[10] Mohammad Al Hasan. Summarization in pattern mining. In *Encyclopedia of Data Warehousing and Mining, (2nd Ed)*. Information Science Reference, 2008.
[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] 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.
[13] Bing Liu, Wynne Hsu, and Yiming Ma. Integrating classification and association rule mining. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 1998.
[14] Bing Liu, Wynne Hsu, and Yiming Ma. Pruning and summarizing the discovered associations. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 1999.
[15] 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.
[16] R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. In *American Association for Artificial Intelligence (AAAI)*, 1997.
[17] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer, 2010.
[18] 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.
[19] Chao Wang and Srinivasan Parthasarathy. Summarizing itemset patterns using probabilistic models. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.
[20] 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.
[21] Peter W F Wilson, Ralph B D'Agostino, Helen Parise, Lisa Sullivan, and James B Meigs. Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus. *Circulation*, 112(20):3066–72, Nov 2005.
[22] Dong Xin, Hong Cheng, Xifeng Yan, and Jiawei Han. Extracting redundancy-aware top-k patterns. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.
[23] Dong Xin, Jiawei Han, Xifeng Yan, and Hong Cheng. Mining compressed frequent-pattern sets. In *International Conference on Very Large Databases (VLDB)*, 2005.
[24] Xiaoxin Yin and Jiawei Han. CPAR: Classification based on predictive association rules. In *SIAM International Conference on Data Mining (SDM)*, 2003.

A Brief Author Biography

V. Kavitha has received his B.E (CSE) degree in the year 2010. At present she is pursuing M.E. (CSE) in Mailam Engineering College, Villupuram, Tamil Nadu, India. She has published 1 paper in National conferences. Her research interests lies in the areas of Data Mining, Networking and Bigdata.

R. Mohan completed his B.Tech (IT) degree in the year 2007, MBA(SYSTEMS) in the year 2009, M. Tech (IT) degree in the year 2012.Currently he is working as a Assistant professor in Computer Science and Engineering at Mailam Engineering college, Villupuram, Tamil Nadu, India. He has published 2 papers in International conferences and 5 papers in National conferences. He attended many workshops & National level seminars in various technologies and also attended Faculty Development Programme.