© 2016 IJSRSET | Volume 2 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus

Prof. Dr. B. Murugeshwari, Jannathul Firdous A, Venmathi V

ABSTRACT

The main aim of this project is to predict 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 evolving concept defined as a systematic 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) is the best practice in the entire above summary.

Keywords: Data Mining, Association rule Mining, Survival Analysis, Summarization Technique

I. INTRODUCTION

Association rule learning is a method for discovering interesting relations between variables in large databases. Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, catalog design and store layout. Programmers use association rules to build programs capable of machine learning. Machine learning is a type of artificial intelligence (AI) that seeks to build programs with the ability to become more efficient without being explicitly programmed.

II. METHODS AND MATERIAL

Existing System:

In an existing system, a statistical modeling 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.

Problem Definition

1. 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 whole.
2. With such an extensive set of risk factors, the set of discovered rules grows combinatorial large, to a size that severely hinders interpretation.

Proposed System

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 suitable 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.

Modules:

- Permitting Health Center Database
- Fetching Database Collection in EMR
- APRX and RPGlobal Summarization
- Topk and BUS Summarization

Permitting Health Center Database

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.

Fetching Database Collection in EMR

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.

APRX and RPGlobal Summarization

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. More specifically, for example, the second rule having 6 conditions represents a set of 21 rules with 4, 5 or 6 conditions. Out of these 21 rules, 20 are actually present in the original rule set. Since the summary rules represents 20 original rules, we define the subpopulation covered by the summary rule as the union of the subpopulations covered by the 20 original rules. 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.

Topk and BUS Summarization

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 While this approach forfeited the outstanding compression rates of the previous two algorithm, TopK still achieves high compression rate (as we will show in the next section) and it successfully identified rules with high risk and low redundancy. 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.

III. RESULTS AND DISCUSSION

Enhancement

In this project we are going to perform enhancement on developing hospital application. From that application we performed database permitting and checking risk level from research centers

Algorithms Used

1. Association Rule Mining
2. Summarization Techniques

Architecture Diagram

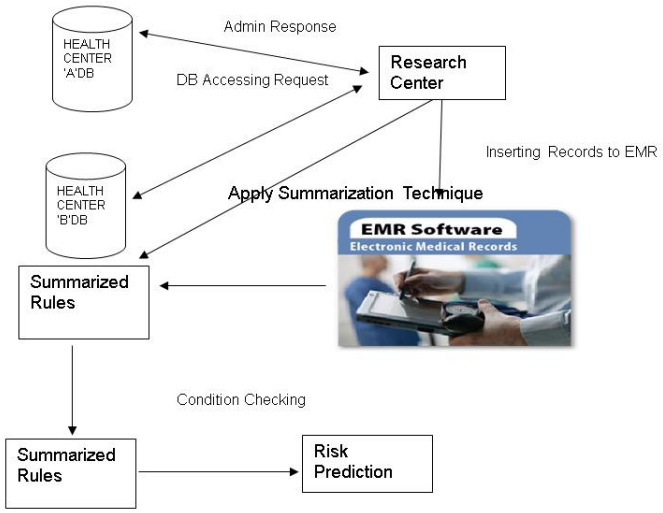

Figure 1 : Architecture Diagram

IV. CONCLUSION

Thus we designed and developing to predict the excess risk of diabetes for the patients and summarize their subpopulation by using Association Rule Mining.

V. REFERENCES

[1] F. Afrati, A. Gionis, and H. Mannila, "Approximating a collection of frequent sets," in Proc. ACM Int. Conf. KDD, Washington, DC, USA, 2004.

[2] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th VLDB, Santiago, Chile, 1994.

[3] Y. Aumann and Y. Lindell, "A statistical theory for quantitative association rules," in Proc. 5th KDD, New York, NY, USA, 1999.

[4] P. J. Caraballo, M. R. Castro, S. S. Cha, P. W. Li, and G. J. Simon, "Use of association rule mining to assess diabetes risk in patients with impared fasting glucose," in Proc. AMIA Annu. Symp., 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," U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2011 .

[6] V. Chandola and V. Kumar, "Summarization – Compressing data into an informative representation," Knowl. Inform. Syst., vol. 12, no. 3, pp. 355–378, 2006.

[7] G. S Collins, S. Mallett, O. Omar, and L.-M. Yu, "Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting," BMC Med., 9:103, Sept. 2011.

[8] Diabetes Prevention Program Research Group, "Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin," N. Engl. J. Med., vol. 346, no. 6, pp. 393–403, Feb. 2002.