isrjournals

International Journal of Advanced Research in Computer Science Engineering and Information Technology

Volume: 4, Issue: 3,Special Issue: 1 ,Mar,2016 ,ISSN_NO: 2321-3337

Predicting Risk of Diabetes Using Summarization Techniques

SathyaPriya.P[1], Shamili.A[2], Mrs.Revathi.M[3]

Student, Dept.of Computer Science and Engineering, Agni College of Technology, India

Asst. Professor, Dept. of Computer Science and Engineering, Agni College of Technology, Indian

ABSTRACT

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) algorithm produced the most suitable summary.

***Keywords*— Electronic Medical Record (EMR), APRX-COLLECTION, Censoring, Bottom-Up Summarization (BUS), association rules, association rule summarization.**

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 diabetes for a person with diabetes is twice the risk of a person of similar age who does not have diabetes.

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.

2. SYSTEM ANALYSIS

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.

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.

3. IMPLEMENTATION

3.1 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.

3.2 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.

3.3 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.

3.4 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.

4. SYSTEM ARCHITECTURE

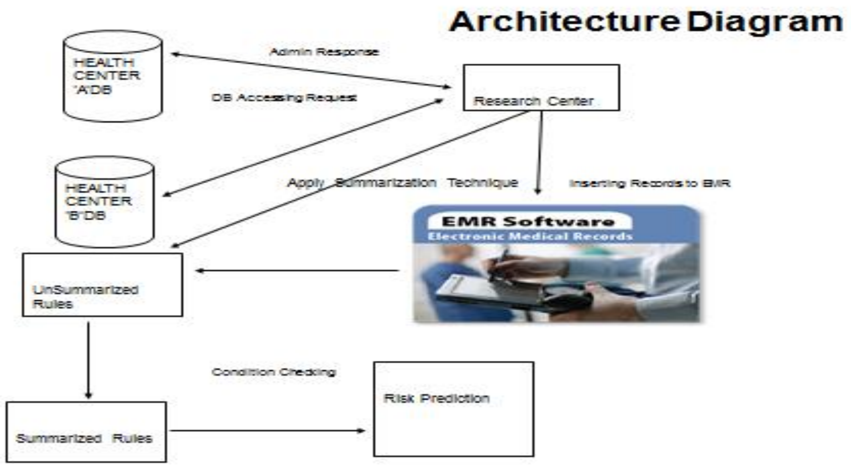

Fig 1. System Architecture

Fig 1 represents the overall System Architecture concepts.Collect the Patients Details from the Hospital database System and Permit the Patients database to the Research center.

Research center maintain the EMR. EMR is used to store the Patients records.Based on Patients Records and risk factors we form unsummarization rule. Then apply four Summarization Techniques such as Aprx Collection,RB Global,TopK and BUS. Finally predict the risk of diabetes for the Patients.

5. CONCLUSION AND FUTUREWORK

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.

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

FUTURE WORK

The future goal of the project is to performed enhancement on developing hospital application. From that application we performed database permitting and checking risk level from research centers.

For this method to be useful, the number of rules is used for clinical interpretation is make feasible.

REFERENCES:

[1] “Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus” Pedro J. Caraballo, Terry M. Therneau, Steven S. Cha, M. Regina Castro, and Peter W. Li, JANUARY 2015 .

[2] “Use of association rule mining to assess diabetes risk in patients with impared fasting glucose,” P. J. Caraballo, M. R. Castro, S. S. Cha, P. W. Li, and G. J. Simon, Symp., 2011.

[3] “Comorbidity study on type 2 diabetes mellitus using data mining,” H. S. Kim, A. M. Shin, M. K. Kim, and N. Kim,Korean J. Intern. Med., vol. 27, no. 2, pp. 197–202, Jun. 2012.

[4] “High-order SNP combinations associated with complex diseases: Efficient discovery, statistical power and func- tional interactions,” G. Fang et al., PLoS ONE, vol. 7, no. 4, Article e33531, 2012.

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

[6] 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.

[7] 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.

[8] G. Fang et al., “High-order SNP combinations associated with complex diseases: Efficient discovery, statistical power and functional interactions,” PLoS ONE, vol. 7, no. 4, Article e33531, 2012.