A Survey of Predicting Relative Risk for Diabetes Mellitus using Association Rule Summarization Techniques

K.Sinduja[1] N.Saravanan[2]
[1,2]Department of Information Technology
[1,2]K.S.R College of Engineering, Tamilnadu, India

Abstract— The detection of diabetes mellitus with elevated risk at early stage is critical in global clinical management. It aims to apply association rule mining to electronic medical records (EMR) to detect sets of risk factors and their corresponding subpopulations of patients. Association rule mining accomplishes a very large set of rules for summarizing the risk of diabetes in EMR with high dimensionality. To review the association rule set summarization techniques and conducted comparative evaluation to provide the best optimal summary based on their merits and demerits. In this paper, discuss about various methods to summarize the high risk of diabetes with accuracy.

Key words: Data Mining, Association Rule Summarization, Survival Analysis, Association Rules, Regression, Top-K, Markov Random Field

I. INTRODUCTION

Diabetes mellitus, commonly referred as diabetes is a group of metabolic diseases and its leads a medical complications including ischemic heart disease, stroke, nephropathy, retinopathy, neuropathy and peripheral vascular disease. It affects 25.8 million people in U.S. approximately 7 million of the people do not know they have the disease.

Association rule mining are created by analysing data for frequency to identify the most relationship. It implies 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 to specify the diabetes risk with set of conditions. These conditions can be used to provide treatment towards a more personalized and targeted diabetes management.

The focus of this manuscript is to review and characterize existing association rule summarization techniques and provide guidance to practitioners in choosing the most suitable one. Our main contribution is a comparative evaluation of these extended summarization techniques that provides guidance to practitioners in selecting an appropriate algorithm for a similar problem.

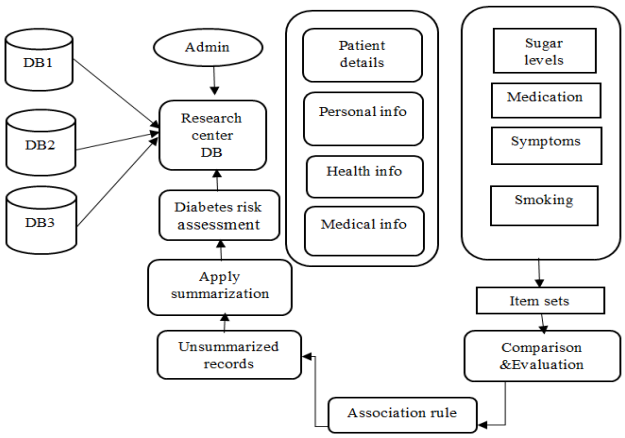

Fig. 1: Architecture

II. LITERATURE SURVEY

A. Effective and Efficient Itemset Pattern Summarization: Regression-based Approaches

Ruoming Jin, Muad Abu-Ata, Yang Xiang, Ning Ruan proposed a set of novel regression-based approaches to provide the summarization of frequent itemsets patterns effectively. It reduces the restoration error for set of itemsets. In this approaches it has two methods, k-regression and tree regression used to partition the entire collection of frequent itemsets. The k-regression approach achieves the total restoration error using K-means clustering method. The tree-regression approach employs to partition the entire collection of frequent itemsets. Our approach improves the summarization performance with high accuracy and less computational cost. The major drawbacks of this approach is difficult to partition large complex data sets.

B. Comorbidity Study on Type 2 Diabetes Mellitus Using Data Mining

In this paper, the author using association rule mining to analyse the comorbidity in patients with 2 types of diabetes mellitus (T2DM). Association rule mining describes how to relate the two data itemsets using special methods of exploring patterns. It possible to analyse the association between not only two diseases but also three or more comorbidities. So, the author aim to apply the association rule mining in electronic medical records. In this Dx Analyze Tool using the Apriori algorithm to analyse the association between clinical diagnoses. This tool helps to refines the data and extract a data between a specific disease and its related disease by association rule.

C. Summarizing Itemset Patterns Using Probabilistic Models

Chao Wang and Srinivasan Parthasarathy proposed a novel probabilistic approach to summarize the frequent set item patterns. In this approach the dataset items are random variables. To construct the Markov Random Field in these variables to provide the summary with high accuracy. It establishes the graphical model in dataset items and provide the summary of the data from set of frequent non-derivable patterns. Non-derivable itemsets are lossless form of compressing frequent itemsets. Maximal itemsets allow greater compression when it compared with closed patterns. It is important to note the number of closed, maximal or non -derivable itemsets. In this paper Top-K patterns are involved to compute the most frequent closed itemset to provide approximate summary to end-user. This approach helps to reduce the memory space and computational cost.

D. Extracting Redundancy Aware Top-K Patterns

In this paper, redundancy aware Top-k patterns extracts a large collection of frequent patterns. First, to examine the

evaluation of functions to measure the combined significance of pattern set. The Greedy algorithm are involved to provide the optimal solution with better performance bound O(log k). The redundancy aware Top-K patterns are also apply in traditional database to extract the frequent patterns by queries. Top-K patterns apply association rule mining, clustering and indexing to evaluating the significance of various kinds of patterns and eliminate the redundancy among the closed patterns. Top-K patterns finds the diverse of frequent patterns and significance to answer the queries, searches and mining. Top-K patterns aims to reduce the restoration error and provide the best optimal summary.

E. Approximating a Collection of Frequent Sets

In this paper**,** collection of frequent itemsets in large transactional databases are obtained. The simple polynomial-time algorithm are involve to measure the approximating collection of sets by k sets which is defined size of the collection of covered sets. The collection of frequent patterns used two different ways: one can absorbed individual patterns and their occurrence frequencies and another can absorbed the entire collection and to determine which patterns have frequent and which patterns are not. The algorithm finding all the frequent patterns for global understanding of approximating collection sets. The collection of frequent patterns is always computed with frequency threshold to lower the limit on the occurrence probability of the pattern. To avoid oversimplification, restrict the number of false positive patterns in dataset. The collection frequent itemset is lossless manner, it helps to reduce the output size with least error.

F. Pruning and Summarizing the Discovered Associations

Bing Liu, Wynne Hsu and Yiming Ma proposed a novel technique prunes to discover associations between data. In general the data mining provides number of techniques to summarize the pattern. In this prunes technique are obtained to provide the overall structure of the large number of detailed patterns. It also removes insignificant associations and then finds a special subset of the unpruned associations and to provide the summary of the discovered associations. Association rule mining are applied to find all the patterns that satisfy the certain constraints. The subset of associations the direction setting rules are used to set the directions that are followed by the rest of the association to provide the summary. By this summary the user can able to understand the relationship of relevant data.

G. Summarization - Compressing Data into an Informative Representation

In this paper, summarization of dataset achieves compaction gain and information loss. Simple summarization methods tabulating the mean and standard deviation to analyse the data and generate the report automatically. Clustering is another technique to summarize the large database. Summarization of datasets involve two approaches: first to adaptation of clustering and second use of frequent itemsets from association analysis domain. Heuristic-based algorithm are used to generate the approximate good summary for the given set of transactions.

It provides the smaller set of individual summary for a given dataset with an objective of maximum information content. In this paper the author proposed Bottom-Up Summarization (BUS) with cluster based algorithm. This algorithm provides the best optimal summary with high accuracy and less information loss.

III. CONCLUSION

This survey paper is based in the summarization frequent itemset patterns by using association rule mining. Several methods are used to provide the better optimal summary to the datasets. Prunes techniques are used to provide the overall structure of datasets as a summary. Top-K patterns are used to evaluating the significance of different patterns to provide the summary. As the result, Bottom-Up Summarization (BUS) are the best algorithm to provide the better optimal summary with high accuracy and less information loss. The BUS algorithm are applicable at more than records and easy to predict at large complex data sets. Association rule mining are applied in diabetes management to identify the risk of diabetes. In this BUS algorithm are used to provide the summary about the diabetes with elevated risk.

REFERENCES

[1] B. Liu, W. Hsu and Y. Ma, ”Pruning and Summarizing the Discovered Associations,” in proc. ACMInt. Conf. KDD, New YORK,NY, USA,1999.
[2] F.Afarti, A.Gionos and H.Mannila, ”Approximating a Collection of Frequent Sets,” in Proc. ACM Int. Conf. KDD, Washington, DC, USA, 2004.
[3] C.Wang and S. Parathasarthy, “ Summarizing Itemset Patterns Using Probabilistic Models,” in Proc. ACMInt. Conf.KDD, New York, NY, USA, 2006**.**
[4] D. Xin, H. Cheng, X. Yan and J. Han, “ Extracting Redundancy Aware Top-K Patterns,” in proc. ACMInt. Conf. KDD, Philadelphia, PA, USA, 2006.
[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] R. Jin, M. Abu-Ata, Y. Xiang and N.Ruan, “Effective and Efficient Itemset Pattern Summarization: Regression-based Approaches,” in Proc. ACMInt. Conf. KDD, Las Vegas, NV, USA, 2008.
[7] H. S. Kim, A.M. Shin, M.K. Kim and N. Kim, “Comorbidity Study on Type 2 Diabetes Mellitus Using Data Mining,” Korean J.Intern Med., vol.27, no.2,pp. 197-202, Jun. 2012.

