BEST: International Journal of Management, Information
Technology and Engineering (BEST: IJMITE)
ISSN (P): 2348-0513, ISSN (E): 2454-471X,
Vol. 4, Issue 6, Jun 2016, 107-110
© BEST Journals

# A SURVEY ON RISK ESTIMATION OF DIABETES MELLITUS USING ASSOCIATION RULE MINING

**ANOOP.S & ARUN P.S**

Department of Computer Science and Engineering, Sree Buddha College of Engineering,
Pattoor, Kerala, India

## ABSTRACT

Diabetes Mellitus is considered to be one of the dangerous and life-long health conditions that affect billions of people all around the globe. Early detection and prevention of this disease is of utmost importance in order to lessen the risk of getting affected by it. An Electronic Medical Record [EMR] with various health parameters, for diabetes mellitus, of various patients, uploaded by a registered medical laboratory is subjected to association rule mining, for that an efficient Association Rule Mining algorithm known as Apriori algorithm is used. The obtained association rules are subjected to distributional association rule mining, which separates affected and unaffected sub-population in order to further reduce the number of rules. Then among that statistically significant rules are selected based on a significant threshold. Then, for summarizing the obtained rules, four summarizations techniques named BUS algorithm, top-k algorithm, APRx Collection and RP Global Clustering is used.

**KEYWORDS**: Association Rule Mining, Apriori Algorith, BUS Algorithm

## INTRODUCTION

In the world where we are living in, we may get affected with numerous diseases. Among all such diseases, diabetes mellitus takes a lion’s share. Billions of people are affected with this disease. Early detection and prevention are of utmost importance for not being a victim to this killer disease as persists life-long. One has to be aware of the risk of them getting affected with this disease based on their age, location body frame and which one with their respective health parameters may get affected by diabetes mellitus is estimated through data mining techniques like association rule mining followed by an efficient summarization technique. A registered medical laboratory, or in other words, a service provider provides numerous Electronic Medical Record [EMR] corresponding to diabetes mellitus. A set of patient details of a particular location is available from the EMR; the details of the original EMR available to the service provider are uploaded to the system as per a specified format. The EMR contains various patient id, age, location and various health parameters like hdl, stab. glu, chol , height, weight etc. With the EMR as such it is not possible to estimate the risk of the one getting affected with diabetes mellitus. Association Rule Mining is the first step towards this. A data mining algorithm known as Apriori algorithm is used for that. Association Rules are obtained as a result of Apriori algorithm. Distributional Association Rule Mining is the next step and which separates affected and unaffected sub-population. The rules belonging the affected sub-population are eliminated thereby reducing the number of rules. From those rules, statistically significant rules are selected based on a significance threshold. Summarization is the next step of the process. A number of successful association rule set summarization techniques have been proposed but no clear guidance present regarding the applicability, strengths and weaknesses of these techniques. The focus of this manuscript is to review and organize four existing association rule summarization techniques and provide guidance to practitioners in selecting the most suitable one. A common shortcoming of these techniques is their inability to take diabetes risk continuous outcome into account.

## LITERATURE SURVEY

[1] consider the issue of finding association rules between items in a large database of sales transactions. We present two new algorithms for fixing this problem that are different from the known algorithms. Empirical evaluation reveals that these algorithms outperform the known algorithms by factors ranging from three for small problems to more than an order of magnitude for large problems. A hybrid algorithm called Apriori Hybrid is used here. Apriori Hybrid proves to scale linearly with number of transactions.

In the paper[2], they presented an efficient algorithm named Fuzzy Cluster-Based Association Rules(FCBAR). Cluster table formation by scanning the database is the main method involved in FCBAR. It is followed by subjecting the transaction records to clustering till the nth cluster table. ’n’ corresponds to the length of the record. Contrasting with the partial cluster tables is the mechanism by which fuzzy large item sets are formed. Pruning of fairly large amount of data is achieved through this thereby decreasing the time required for performing data scans. FCBAR performs much better than Fuzzy Apriori Algorithm as per experiments done with real-life data base.

In this paper[3], a novel framework is proposed for designing an IDS based on data mining techniques. In this framework, Association Based Classification (ABC) is what the classification engine uses. The proposed classification algorithm uses Fuzzy association rules are used by the classification algorithms for building classifiers. Particularly, the fuzzy association rule-sets are exploited as the descriptive models of different classes makes use of fuzzy association rule sets. A new sample with different class rule sets’ compatibility is assessed by applying some matching measures and the class corresponding to the best matched rule-set is tagged up as the label of the sample

A generalized fuzzy data mining algorithm for extracting interesting patterns is shown in paper [4]. The proposed algorithm does fuzzification of the quantitative Web usage data along with predefined membership function. They also use predefined support and confidence. The whole database is partitioned based on hours. The fuzzy mining algorithm to extract association rules is applied separately on each partition. The combination of all hours association rules is used to declare total number of rules for given database.

The problem of discovering association rules has attained considerable research attention and several fast algorithms for mining association rules have been formed. Subset of association rules are most preferred by the users. While such conditions are applied as a post-processing step, integrating them into the mining algorithm can reduce the time for execution. We take into consideration the problem of integrating constraints that are the Boolean expressions of items into the association discovery algorithm. We introduce three integrated algorithms for mining association rules with item constraints and discuss their tradeoffs.

There has been increased interest in discovering combi- nations of single-nucleotide polymorphisms (SNPs) that are are based on a phenotype even if each SNP has little impact. Since statistics with high degrees of freedom is present, the existing approaches are devoid of statistical power. In high-order combinations, because of these obstacles, the functional interactions are not explored properly. In [6] high-order combinations in case-control datasets are searched by suitable pattern-mining algorithms. The remarkably improved efficiency and scalability shown on synthetic as well as real datasets with several thousands of SNPs allows the study of numerous important mathematical and statistical features of SNP combinations with order up to eleven. They then explore the functional interactions in high-order combinations and bring out a link between the increase in discriminative power of a combination over its subsets and the functional coherence between the genes comprising the combination, supported by multiple datasets. Finally, we study numerous significant high-order combinations obtained from a lung-cancer dataset and a kidney-transplant-rejection dataset in detail to provide insights on the complex diseases. Many of these associations involve combinations of common variations present in small fractions of population. Thus, our approach is an alternative methodology which explores the genetics of rare diseases for which the current attention is on individually rare variations.

Detecting patients with elevated risk of developing diabetes mellitus is critical to the improved prevention and overall clinical management of these patients. Association Rule Mining is applied to EMR for getting risk factors. As association rule mining produces a large set of rules, it has to be summarized for clinical use. [7] reviewed four summarization techniques and their strengths and weaknesses are evaluated. Incorporating the risk of diabetes into the inference making process is done here. A real-world pre-diabetic patient group is subjected to evaluate these.

In this paper, [8] extracting redundancy-aware top-k patterns from a huge collection of frequent patterns is considered. Here, Maximal Marginal Significance (MMS) is considered, and it acts as the problem formulation.NP-hard is the name by which that problem is known. A greedy algorithm is further presented; its purpose is to approximate the optimal solution with O (log k) bound. Disk block prefetch and document theme extraction are the two applications illustrated through redundancy-aware top-k queries.

## CONCLUSIONS

The e-data generated by the use of EMRs in routine clinical practice has the potential to facilitate the discovery of new knowledge. Association rule mining in conjunction with summarization technique provides a critical tool for clinical research. It can reveal hidden clinical relationships and can propose new patterns of conditions to redirect prevention, man-agreement, and treatment approaches. APRX-COLLECTION and RPGlobal primarily operate on the expression of the rules with a primary objective of maximizing compression. Representative rules, each of which represents a number of original rules, are used. Such representative rules offer very high compression, but dilute the risk of diabetes over the typically large subpopulation they cover. TopK and BUS algorithm operate mainly 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 attain good compression. The converse is not true: high compression rate does not result in low redundancy.

## REFERENCES

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

2. An Algorithm For Mining Fuzzy Association Rules Reza Sheibani , Amir Ebrahimzadeh ,Member, IAUM

3. Intrusion detection using fuzzy association rules Arman Tajbakhsh, Mohammad Rahmati, Abdolreza Mirzqei

4. Mining Fuzzy Association Rules from Web Usage Quantitative Data Ujwala Manoj Patil and Prof. Dr. J. B. Patil Department of Computer Engineering, R.C.P.I.T., Shirpur, and Maharashtra, India.

5. B. Liu, W. Hsu, and Y. Ma, Integrating classification and association rule mining, in Proc. ACM Int. Conf. KDD,

New York, NY, USA, 1998.

6. High-Order SNP Combinations Associated with Complex Diseases: Efficient Discovery, Statistical Power and Functional Interactions Gang Fang Majda Haznadar, Wen Wang, Haoyu Yu, Michael Steinbach, Timothy R. Church, William S. Oetting, Brian Van Ness
7. Predicting Relative Risk for Diabetes Mellitus using Association Rule Summarization Technique in EMR K. Thulasi, S.Sowmiyaa, P. Prema International Journal of Innovative Research in Science, Engineering and Technology
8. D. Xin, H. Cheng, X. Yan, and J. Han, Extracting redundancy aware top-k patterns, in Proc. ACM Int. Conf. KDD, Philadelphia, PA, USA, 2006.