Indian Journal of Science and Technology, *Vol 9(45), DOI: 10.17485/ijst/2016/v9i45/100943, December 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Assessing Risk of Diabetes Mellitus

S. Malati* and E. Poovammal

Department of Computer Science and Engineering, SRM University, Kattankulathur – 603203, Tamil Nadu, India;
malatisakthi@gmail.com, poovammals@gmail.com

Abstract

Diabetes is a non-communicable disease which is affecting the growth of developing countries. Our aim is to prevent diabetes by deducting it in the earlier stage so that people take treatment according to it, this can be done by examine the electronic medical record of a patient to discover set of risk factors by applying association rule mining methods. The Electronic Medical Record is very large which provides many rule set as result when association rule mining is used, so in order to summarize rules we go for Bottom-up-summarization algorithm. Rule set summarization techniques such as RPC, APRX-collection and BUS are applied to compress original rule set commonly available in Electronic Medical Record (EMR) system, then to predict the relative risk of diabetes millets as high risk, medium risk and low risk by using the K-Nearest Neighbor. RPC is a Relative patient coverage which can be extracted from status and follow-up patient record. K-Nearest Neighbor is a non-parametric method and they are mainly used for both classification and regression but here we use it for classification where the input will be training data.

Keywords: APRX Collection, Association Rule, BUS Algorithm-Nearest Neighbor, Diabetes Mellitus, Electronic Medical Record, RPC

1. Introduction

Diabetes mellitus is a chronic illness that requires proper continuing medical treatment and patient self-management of their health to prevent acute complication and to reduce the risk .Treatment of diabetes requires risk reduction strategies and many people suffer in prediabetes where blood glucose level goes above normal but below certain threshold for the diagnosis of diabetes. Prediabetes is also accompanied by diseases such as hyperlipidemia, obesity and hypertension which include treatment such as use of drugs. Statin therapy is prescribed for hyperlipidemia use of statin will reduce the cholesterol level, and risk of cardiovascular mortality. The research n recent studies indicates an risk of incident diabetes associated with their use which is increased.

The Association rule are particularly used to quantify the diabetes risk, they also provide proper proof with associated set of conditions. To guide the treatment and for preventive care these set of conditions can be used. In order to quantify the effect of statin on diabetes defined by the association pattern with combination of hypertension, renal failure association rule mining technique is to directly compare the diabetes among those who consume statins and those who do not consume statins, among the patient presenting with hypertension and renal failure.

Association rule mining which shows all the relationship in huge database which are interesting. All the rules in the database that has some minimum support and minimum confidence threshold are determined using Association rule mining technique. Support measure is important to find the frequent item set based on total number of item. Confidence is measure of frequent item set to the total number of particular item.

*Author for correspondence

In order to find risk factor sets Electronic Medical Record (EMR) is subjected to Association rule mining and their corresponding subpopulation that represents Patients at particular high risk of developing diabetes. Due to the high dimensionality of EMR a very large set of rules are generated by association rule mining which needs to be summarized. This association rules mining summarization method are used to deduct diabetes in order to prevent from risk. The summarization techniques used are RPC, APRX collection, Bus algorithm. In order to classify based on level of risk K-Nearest Neighbor method is used, where the level of risk is classified as low, medium and high.

2. Related Works

The issue which is related to the size of the output which is large is rectified by solving it using k sets which identify the collection of frequent item sets. The identification f k sets is used for solving the above problem[1]. The problem is in selecting and defining the k sets which approximate a collection of frequent item sets. The solution is given by polynomial-time approximation algorithm. In practice as per analysis the approximation methods suits accurate when compared to other methods.

Even though the Algorithm for computing frequent itemset are proper and accurate but due to large output size of the frequent pattern collection it is difficult to manipulate the patterns which are frequent and which are non-frequent. Restricting the frequent itemset collection output to the border doesn't help much in solving the problem. The difficulties faced such as border computation in polynomial time to exact size and the next problem is exponential size of the border in the database, therefore one cannot rely on search capacity so better to go for sampling methods.

The size of the output which is large is rectified by computing the k sets which identify the collection of frequent itemsets the main idea is by computing minimal error on the frequent itemset support count. The way of using the collection differs by two ways such as first one is interest in single individual pattern and their frequency of occurrence then followed by next one can be interested in whole collection. The level of approximation gives a complete understanding of structure of the set of data without compressing the information though restricting the border of output does not bring solution to the problem and it is not efficient in improving the performance.

Bangaru Veera Balaji introduced to classify the data accurately Association rule minind and classification method is used combinedly[2]. The two new algorithms CPAR (Classification Based on Predictive Association Rule)and CMAR (Classification Based on Multiple-Class Association Rule) that combines the advantages of both traditional rule-based classification and associative classification. CPAR which generates and tests more rules also adopts greedy algorithm to generate rules directly from training data and has traditional rule- based classifiers to avoid missing important rules and thereby avoiding over fitting. For prediction and accuracy evaluation of each rule CPAR uses best k rules. CMAR applies tree structure for storing and retrieving mined association rule efficiently then to effectively prune rules based on confidence, correlation and database coverage. There are two phases where first phase is rule generation and second phase is classification. It is consistent highly effective at classification of various kinds of databases and has better average classification accuracy.

Yonatan Aumann introduced proper organised search of PubMed and EMBASE database to describe the model of development of two or more variables to predict the risk of prevalent or incident type 2 diabetes[3]. This is a free search engine accessing primarily the MEDLINE database of reference and abstract on life science and Biomedical topic. PubMed search automatically add field and the maintain standard searches that translate search formulation and names. Searching on PubMed can be carried out by entering key aspects of subject into PubMed search window.

The optimization problem involves objective functions such as compaction gain and information loss[4] and problem of summarization related to dataset of transaction with categorical attributes. The metrics which are proposed to characterize the output of any summarization algorithm has two Approaches. First is clustering adaptation and next is usage of frequent item sets from the association analysis.

In order to summarize network traffic into a compact and meaningful representation this technique can be effectively used. The amount of reduction done in the transformation from the actual data to summary shows the compaction gain. Information loss is total amount of information missing over all original data transaction in the summary.

Mohammad Al Hasan proposed a number of successful association rule set summarization techniques but no clear idea regarding the strengths, applicability and

weakness of these technique[5]. The disadvantages of these techniques is their drawback to take diabetes risk a continuous outcome into account.

The solution to problem of frequent pattern sets compression can be given in either two ways one is clustering can be adopted with measure of tightness as reference and the next method is selection of pre-sensitive patterns with respect to each cluster. The representative patterns can be discovered using a method NP-Hard.

3. System Architecture

System architecture is a diagrammatic representation of hoe system behaves and organization structure of the system both functional and non functional. In the Figure-1 the patient life database is generated then extraction of status and follow-up patients is done. After finding status and follow-up patient then data mining process is carried out. As a result of this frequent item set is generated to perform association rule mining then RPC process is performed to generate relative patient coverage the result of which is loaded to perform the APRX collection which generates false positive report then the outcome of APRX is processed using BUS algorithm to generate high risk patient report at the top. Then K-Nearest Neighbor algorithm is performed to cluster the patient according to high, medium, low risk

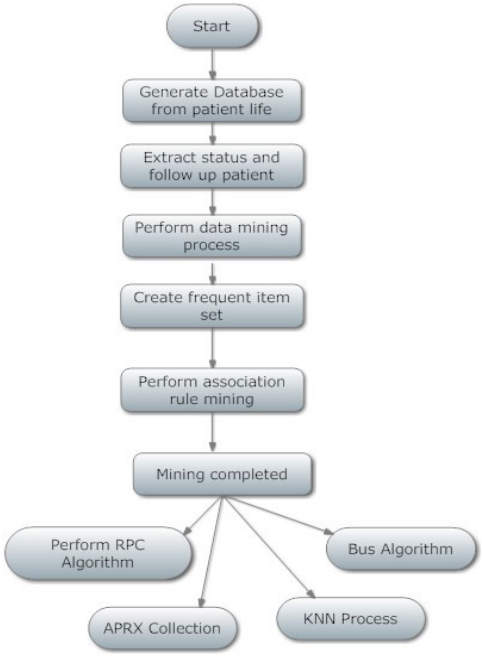

Figure 1. System Architecture.

4. Methods

4.1 Data Loading

In Dataloading diabetes dataset is loaded to process. And then insert the dataset on database dynamically. After that insert the new diabetes report on database.Dataset should be loaded after preprocessing automatically and also inserted into database newely whenever process is running.

4.2 Status and Follow up Patient Report

Extract the data based on status and follow up patient. Status patient are the people who caused by diabetes in long year which based on dataset attributes. Follow Up Patient are the people who caused either diabetes at starting stage or not. Status Patient report are stored automatically to find the high risk patient report for future purpose.

4.3 Support and Confidence Measure

Next Step in this process is to find the diabetes patient based on symptoms then process with symptoms data to find the support and confidence measure. Support Measure is importantto find the frequent itemset based on itemset. To check the itemset are present in the frequent itemset if present to count theitemset. And also measure the confidence for threshold.

4.4 Association Rule Mining

After finding the result of support and confidence to mining the report based on support count. And extract the resulting itemset from overall itemset. And then extract the diabetes report based on itemset who are satisfy the condition and affected by symptoms.

4.5 RPC and Data Coverage Method

Relative patient coverage(RPC) can be extract from the status & follow up patient report who are caused by relative symptoms and affected by diabetes.This can be calculated through association rule mining and support and confidence measure Data Coverage Method is based on RPC how many dataset are related and all these process are summazation to extract the data.

4.6 APRX-Collection

APRX – Collection process is based on false positive report. False Positive is generated by identify the item set are missing in the symptoms.After calculated the false positive value and then extract the diabetes report based on false positive result.

4.7 BUS Process

The Bottom Up Approch is the BUS Process that removes the related report in the dataset of patent in bottomwise.

After calculateBUS,APRX,RPC all the three report result are merge to match the Status Patient report. And then extract the matched report based on result then finally get the High Risk Patient Report who are affected by diabetes in serious condition.

4.8 K-Nearest Neighnouring

By using BUS process, remove the related report in the dataset of patient in bottom wise. Using the K-nearest neighboring algorithm is to classify the report into three categories such as highly affected patient by diabetes mellitus, average patient and low risk patient in diabetes mellitus.

5. Conclusion

Sets of the factors identified using association rule mining and the subpopulation of related patients are at higher risk of diabetes development. The algorithm differs according to the way the selection measures are included in the summary of rules which is based on rule expression or subpopulation of the patient that is covered by the rule. On comparing both BUS and TopK we came to conclusion that Bus works better in sense of maintaining higher redundancy when compared to TopK. The main advantage of BUS is proper patient ability to rebuild the original database. These are improved measures carried out for early detection which is achieved through the process of screening. This method is used for the individual who are at high risk of developing diabetes when they approach for other checkups can be made into observation with the record and information of already observed individual who has risk of diabetes so this helps in deduction and prevention of diabetes at early stage.

6. References

1. Afrati F, Gionis A, Mannila H. Approximating a collection of frequent sets. Proc ACM Int Conf KDD, Washington, DC, USA. 2004.
2. Balaji BV, Rao VV. Improved Classification Based Association Rule Mining. Int Journal of Advanced Research in Computer and Commu Engineering. 2013 May; 5.
3. Aumann Y, Lindell Y. A statistical theory for quantitative association rules. Proc 5th KDD, New York, NY, USA. 1999.
4. Chandola V, Kumar V. Summarization – Compressing data into an informative representation. Knowl Inform Syst. 2006; 12(3):355–78.
5. Hasan MA. Summarization in pattern mining. Encyclopedia of Data Warehousing and Mining, 2nd ed. Hershey, PA, USA: Information Science Reference. 2008.
6. Xin D, Han J, Yan X, Cheng H. Mining compressed frequent-pattern sets. Proc 31st Int Conf VLDB, Trondheim, Norway. 2005.