Journal of Basic and Applied Engineering Research
p-ISSN: 2350-0077; e-ISSN: 2350-0255; Volume 3, Issue 10; July-September, 2016, pp. 838-841

http://www.krishisanskriti.org/Publication.html

Diabetes Mellitus Risk Estimation using Fuzzy Association Rule Mining

Anoop S[1] and Arun P.S[2]

[1,2]Sree Buddha College of Engineering, Pattoor
E-mail: [1]anoopsulaiman777@gmail.com, [2]arunpstec@gmail.com

Abstract—*Diabetes Mellitus is considered to be one of the dangerous and life-long health condition that affects billions of people all around the globe. Early detection and prevention of this disease is of utmost importance inorder 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 inorder 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, Fuzzy Association rule Mining is used, which combines multiple rules, and brings it into a form where a precise inference can be made. By looking at the inference, a health seeker can be aware of the risk scale of diabetes mellitus, based on his/her medical parameters.*

1. 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 various health parameters. The extent of the risk by 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 otherwords , 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 is eliminated thereby reducing the number of rules. From those rules, statistically significant rules are selected based on a significance threshold.

On the rules obtained so far, fuzzy association rule mining is applied. First step involved in this process is setting of grades [very high, high, medium, low, very low] to interval of values to all the health parameters .The fuzzy support for all the rules are calculated, unlike the normal support, fuzzy support considers data values of combination of attributes, for eg: attribute A &B with or A with B & C. Fuzzy confidence is the combination of one or more attributes with another one or more attributes over one among either of the two. For eg: (A,B) with C/ |A,B|. Fuzzy confidence here is expressed in terms of predefined grades. Then, a parameter called E-lift is calculated, which is nothing but Fuzzy confidence over Fuzzy Support , if E-lift of a rule is beyond a threshold, then rule strictly obeys. That threshold is named a Fuzzy lift, E-lift usually centers around 1.Fuzzy lift is normally set as value 1. Based on these fuzzy association rules are generated, and from that set of inferences are formed ,i.e, A person is 60 % risky of diabetes if age is between 35 and 40 and hdl is 4.5

2. RELATED WORKS

In the paper [1], risk estimation of diabetes mellitus is performed by the making use of Association Rule Mining followed by various summarization techniques. Apriori Algorithm is the technique used to obtain Association rules.The obtained rules are then subjected to distributional association rule mining to reduce the obtained number of rules. Then various summarization techniques like Top-K Algorithm,BUS algorithm,APRx algorithm and RP Global Clustering are used. All the summarization techniques are compared with each other.Rules can be summarised to the point where an inference can be made from it

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 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 itemsets 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 the paper [3], features of composite attributes are made use of in developing a framework for mining fuzzy Association Rules. The partitioning of property values into fuzzy property sets for the application of fuzzy association rule mining. The paper mainly deals with the process of extracting fuzzy sets as well as developing an efficient ARM algorithm based on correlation factor as interestingness measure and thus presents a new way for extracting association rules from items with properties.

A novel fuzzy methodology making use of fuzzy association rule mining method for biological knowledge extraction is shown in paper [4] . A yeast genome dataset is the base for this,which contains structural and functional genome informations.A number of association rules have been found, many of them agreeing with previous research in the area. In addition, a comparison between crisp and fuzzy results proves the fuzzy associations to be more reliable than crisp ones.An integrative approach as the one carried out in this work can unleash significant knowledge currently hidden and dispersed through the existing biological databases. It is shown that fuzzy association rules can model this knowledge in an intuitive way by using linguistic labels and few easy-understandable parameters.

In this paper[5], 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's 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 [6]. 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.

3. PROPOSED SYSTEM

The input to the system is an Electronic medical record [EMR] uploaded by an registered user. The EMR is available from the EMR corpus uploaded by authentic registered medical laboratories or in otherwords service providers. A specified format is available with columns corresponding to various medical and other parameter .In the system developed ,there are 19 columns for the medical data set, including location, age, hdl, stab.glu, glyhb etc. Only in that specified format , the EMR has to be uploaded. The steps involved in the processing of the EMR resulting the estimation of the risk of diabetes mellitus based on his/her various medical parameters are the following:

i). Pre –processing

In this step, the uploaded EMR for the estimation of the risk of diabetes mellitus is subjected to data cleansing. It includes removal of inconsistent, incomplete, reduandant and invalid records from the medical data set. For eg , if age columns by chance has a value 161.88,then record gets eliminated.Only if the data cleansing is done, the system returns the required accurate result. It is a very important step in the whole process involved.

ii). Association Rule Mining

The EMR, after the data cleansing , is subjected to association rule mining. An efficient technique known as Apriori algorithm is used for that. When Apriori algorithm is applied to the medical data set, it returns a set of association rules. The rules are generated based on a minimum support and minimum confidence value. Support refers to the franction of transactions that contain an itemset. Confidence refers to the measure of how often two items come together w.r.t one among those items. When the minimum support and confidence is set very high the number of rules generated will be very low. The minimum support and confidence must be set optimally. All the 19 health parameters can be considered for the association rule mining. But, that is usually not preferred. Only very crucial among those considered as it requires a lot of computational time

Pseudo code for Apriori Algorithm is given below:

C_k: Candidate itemset of size k

L_k: frequent itemset of size k

L_1= {frequent items}

for(k= 1; L_k!=; k++) do begin

C_{k+1}= candidates generated from L_k

for each transaction tin database do

increment the count of all candidates in C_{k+1} that are contained in t

L_{k+1}= candidates in $C_{k+1}$with min_support

End

iii). Distributional Association Rule Mining

In Distributional Association Rule Mining, the rules obtained after applying Apriori algorithm is further reduced by separating affected and unaffected sub-population. The affected population refers to those who are already affected with diabetes, the rules corresponding to the unaffected sub-population are eliminated and left with only the rules corresponding to affected sub-populations. The process of finding out the rules corresponding to the affected sub-population is known as item discovery. After that, among the obtained rules, statistically significant rules are found out.Statistical Significant Rules are found by eliminating rules which are having very low confidence.A threshold called significance threshold is given,and as a result of which all the rules which are lesser to it gets eliminated,thereby further reducing the number of rules.

iv). Fuzzy Association Rule Mining

The rules obtained from Apriori algorithm followed by Distributional Association Mining is subjected to Fuzzy Association Rule Mining.It summarises the rules to the point where accurate inferences can be made from it.

Steps involved in Fuzzy Association Rule Mining are:

a). Setting grades to the health parameter values

The various health parameters are divided into several intervals and they are assigned grades like very low, medium, high, very high etc. All the health parameters involved will be having itz range of values graded in this fashion for the purpose of fuzzification of obtained rules

For eg: hdl count in range of 20 to 50 very low, 50 to 80 low etc…

b). Fuzzy Association Rules

The fuzzy support for all the rules are calculated first , unlike the normal support, fuzzy support considers data values of combination of attributes, for eg: attribute A &B with or A with B & C. Fuzzy confidence is the combination of one or more attributes with another one or more attributes over one among either of the two. Fuzzy confidence here is expressed in terms of predefined grades. Then, a parameter called E-lift is calculated, which is nothing but Fuzzy confidence over Fuzzy Support , if E-lift of a rule is beyond a threshold, then rule strictly obeys. That threshold is named a Fuzzy lift, E-lift usually centers around 1.Fuzzy lift is normally set as value 1. Based on these fuzzy association rules are generated.

iv). Inference

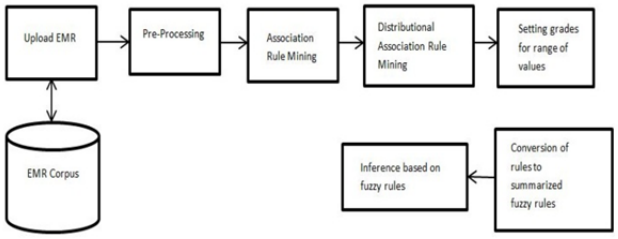
From the fuzzy association rules formed, a set of inferences are made, by taking together all the fuzzy rules formed. A set of inferences are possible, based on the obtained number of fuzzy association rules. The risk of diabetes mellitus is expressed in terms of the value of medical parameters involved.

For eg:

80 % risky of diabetes if age is between 35 and hdl is 4.3

77% risky of diabetes if hdl is 4.5 and weight is 150 pounds

30% risky of diabetes if glyd is 6 and weight is 120 pounds



Architecture of Fuzzy Association Rule Mining System for Diabetes Mellitus Risk Estimation

4. CONCLUSION

Being aware of the extend of risk foe one getting affected with a dangerous health condition like diabetes mellitus is very important for taking preventive measures against it.This paper proposed a data mining technique called fuzzy association rule mining which works on the Electronic medical record to generate normal association rules followed by fuzzy association rules therby resulting in accurate set of inferences, by referring it a person can get to know the extent of him/her getting affected with diabetes mellitus in the future

5. ACKNOWLEDGEMENT

First of all, we thank ALMIGHTY for giving us strength and courage for taking a relevant area and to do research on that. We thank everyone, who directly or indirectly helped us for doing the thesis.

REFERENCES

[1]. Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus György J. Simon, *Member, IEEE*, Pedro J. Caraballo, Terry M. Therneau, Steven S. Cha, M. Regina Castro, and Peter W. Li

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

[3]. A Framework for Mining Fuzzy Association Rules from Composite items Maybin Muyeba[1], M. Sulaiman Khan[2], Frans Coenen[3] [1]Department of Computing and Mathematics, Manchester Metropolitan University, Manchester,UK [2]Liverpool Hope University, Liverpool, L16 9JD, UK [3] Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, UK

[4]. Fuzzy association rules for biological data analysis: A case study of yeast Francisco J Lopez, Armando Blanco, Fernando Garcia1, Carlos Cano and Antonio Marin

[5]. Intrusion detection using fuzzy association rules Arman Tajbakhsh, Mohammad Rahmati ,Abdolreza Mirzqei

[6]. 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, Maharashtra, India.