IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

Application of Association Rule Mining in Risk Analysis for Diabetes Mellitus

Umang Soni[1], Sushma Behara[2], Karthik Unni Krishnan[3], Ramniwas Kumar[4]

B.Tech Student, Computer Science and Engineering, SRM University, Chennai, India[1, 2, 3, 4]

Abstract: Diabetes is a growing epidemic and non communicable disease that affects a major portion of the population in the developing countries. Prevention and overall clinical management of patients with elevated risk of developing diabetes mellitus can be aided greatly by early detection of risk of getting diabetes among patients. This is done by using Association Rule Mining of the data present in the medical records of patients. Association rule is used to discover sets of risk factors for patients at particularly high risk of developing diabetes, which are called item sets of all the interacting conditions. Association rule mining generates a very large set of rules which needs to be summarized for easy clinical use. A comparative evaluation is performed to provide guidance regarding their applicability, strengths and weaknesses.

Keywords: diabetes mellitus, association rule mining, apriori algorithm, discretization.

I. INTRODUCTION

Diabetes is a disorder which characterized by High blood glucose level (blood sugar). When a person has diabetes mellitus, either the body is unable to use its own insulin or manufacture enough hypoglycemic agent. Which causes Aldohexose to build up within the blood and cause a condition that will result in serious health complications such as stroke or even death, if not controlled.

Diabetes may be the major reason for heart attack or stroke, but the death values for a heart attack and stroke is reported to be 2-4 times higher to those having diabetes mellitus than those that do not have diabetes mellitus. United Nations reported that 67% of adults of USA who have diabetes mellitus additionally report to having High blood pressure. Also it was observed that people with diabetes mellitus, high cholesterol level, high blood glucose level and smoking increases the chance of heart attack and stroke. This risk may be reduced by reducing the blood glucose level, cholesterol level or reducing smoking.

In response to the pressing need to tackle diabetes mellitus early in patients, numerous risk indices (risk values) have been developed such as the Framingham score[index] was widely accepted in clinical areas as a remedy for detecting the illness.

There are three types of diabetes mellitus. In Type I diabetes mellitus, the body doesn't manufacture hypoglycemic agent. It is the most common type of diabetes mellitus. In Type II diabetes mellitus, the body doesn't manufacture enough of hypoglycemic agent for body use. This is the second most common type of diabetes mellitus. Type III diabetes mellitus is otherwise called as gestational diabetes mellitus which affects the females during maternity. The common symptoms of diabetes mellitus are- intense thirst and hunger, weight gain or unusual weight loss, frequent voiding, fatigue, cuts and cruises that don't heal sexual pathology in males, and tingling sensation in hands and feet.

The goal of data mining is to obtain higher level information or meta data from an abundance of raw data. Association rules are a very essential for this purpose. An association rule is a rule of the form A -> B , where A and B are events. The rule implies that with a certain probability, known as the confidence factor of the rule, when X occurs in the given database, then Y also occurs.

Association rules are used to relate a set of conditions which can potentially interact with each other and pose an imminent risk (like body mass index and hypertension of an individual). Association rules are favorable because in addition to detecting the risk of diabetes mellitus, they also provide the physician the cause for the disease in an individual which are the set of conditions (e.g. co-morbid sickness, medications, lab results and other demographic information that is available in the Electronic Medical Records (EMR). When we have comprehensive factors, the combinative rule set becomes so large that it hinders the interpretation. In order to overcome this, we use Summarization techniques for the data set to compress the original rule of data set into a compact version for easy interpretation. Association Rule Mining is an approach for discovering association of items in a data set. It can be used to detect and study the etiological pathways in the populations as they suggest interconnections of various risk factors responsible for a disease and are easily interpretable. The primary advantage of using Association Rule Mining is that it presents the rules on the data, which can be used in the clinical decision support system. One disadvantage of association rules is that though they do provide the risk associated with a particular disease in a subpopulation, they do not provide a remedy for the same.

II. PROPOSED ALGORITHM

In the clinical application of association rule data mining, we find item sets of health conditions that show the consequent quantity of increased risk amount of diabetes

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

mellitus. The data extracting process using Association Rule in data mining used to describe an extensive variable set of items that result in an exponentially huge set of association rules formed. The main contribution is a comparative calculation of these summarization techniques that gives guidance to practitioners about Observed patients in datasets for choosing a relevant algorithm for a similar problem in the domain.

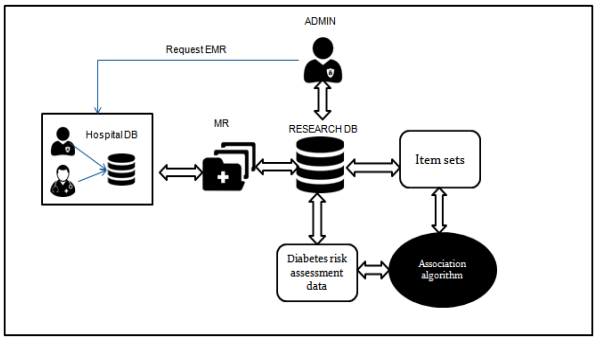

Fig. 1: System Architecture

A. Association Rule Mining

Let an item be a binary indicator to signify if a patient possesses a particular risk factor. E.g. the item ihd indicates whether the patient has been diagnosed with ischemic heart disease. Let X denote the item matrix, which is a binary covariate matrix with the columns representing items and rows representing patients . An itemset is a set of such items and it is an indication of whether the corresponding risk factors are all present in the patient. If they are, the patient is said to cover the itemset (or the itemset applies to a patient).

An association rule is of form

$I \rightarrow J$, where I and J are both item sets.

The rule represents an implication that if J is likely to apply to a patient given that I also applies. The item set I is said to be the antecedent and J is the consequent of the rule.

One feature of association rule mining is that, items do not play any particular roles, i.e. there are no designated predictor variables or outcome variables. In other words, any item can appear in the antecedent of one rule and in the consequent of another.

B. Distributional Association Rule

The distributional association rule can be defined as for a continuous outcome y, an item set I with a distribution between the affected and the unaffected subpopulations are statistically significantly different from each other. For example, the rule {htn, bmi } indicates that the patients both presenting hypertension (high blood pressure) and a high body mass index count have a significantly higher chance of progression to diabetes than the patients who are either not having high blood pressure or do not have a high body mass index count. Since each rule is defined by an itemset we use the words 'itemset' and 'rule' interchangeably.

The distributional association rules discovery consists of two steps, in the first step suitable itemsets are discovered and in the second step the already discovered itemsets are filtered to return only the statistically significant rules.

C. Data Pre-processing

Raw data that is obtained from the source is usually noisy, inconsistent and may be missing values. The quality of data affects the data mining results. Pre-processing of raw data is performed so as to aid in improving mining results as well as the quality of data. It also improves the ease of the mining process. Transformation and processing of raw data can significantly influence the efficiency of the complete data.

D. Data Cleaning

Data cleansing is the technique of identifying and rectifying noisy or incorrect data in datasets. It involves finding incorrect, incomplete or irrelevant parts of the dataset and then performing operations such as modifying, replacing or deleting the noisy data so that it becomes consistent. Once data cleaning has been performed, the data becomes easier to process since it is free from inaccuracies and inconsistencies.

Table I: Symbol Specification

Symbol	Description
preg	Number of times pregnant
Pgc	Plasma glucose concentration
Dbp	Diastolic blood pressure
tsf	Triceps skin fold thickness
insul	2-Hour serum insulin
Bmi	Body mass index
Dpf	Diabetes pedigree function

Table II: Continuous dataset

	preg	pgc	dbp	tsf	insul	bmi	dpf	Age	Class variable
1	6.000	148.000	72.000	35.000	0.000	33.600	0.627	50.000	1
2	1.000	85.000	66.000	29.000	0.000	26.600	0.351	31.000	0
3	8.000	183.000	64.000	0.000	0.000	23.300	0.672	32.000	1
4	1.000	89.000	66.000	23.000	94.000	28.100	0.167	21.000	0
5	0.000	137.000	40.000	35.000	168.000	43.100	2.288	33.000	1
6	5.000	116.000	74.000	0.000	0.000	25.600	0.201	30.000	0
7	3.000	78.000	50.000	32.000	88.000	31.000	0.248	26.000	1
8	10.000	115.000	0.000	0.000	0.000	35.300	0.134	29.000	0
9	2.000	197.000	70.000	45.000	543.000	30.500	0.158	53.000	1
10	8.000	125.000	96.000	0.000	0.000	0.000	0.232	54.000	1
11	4.000	110.000	92.000	0.000	0.000	37.600	0.191	30.000	0
12	10.000	168.000	74.000	0.000	0.000	38.000	0.537	34.000	1
13	10.000	139.000	80.000	0.000	0.000	27.100	1.441	57.000	0
14	1.000	189.000	60.000	23.000	846.000	30.100	0.398	59.000	1
15	5.000	166.000	72.000	19.000	175.000	25.800	0.587	51.000	1
16	7.000	100.000	0.000	0.000	0.000	30.000	0.484	32.000	1
17	0.000	118.000	84.000	47.000	230.000	45.800	0.551	31.000	1
18	7.000	107.000	74.000	0.000	0.000	29.600	0.254	31.000	1
19	1.000	103.000	30.000	38.000	83.000	43.300	0.183	33.000	0
20	1.000	115.000	70.000	30.000	96.000	34.600	0.529	32.000	1

E. Continuous to Discrete Variable.

Datasets such as bmi, age come under continuous data which are difficult to manipulate. Any form of processing that is performed on this kind of data would lead to large amounts of overhead. A solution to this problem would be to convert the continuous data into its discrete forms. There are various techniques to preprocess data before

DOI 10.17148/IJARCCE.2016.54136

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

evaluation. Discretization is necessary for generating frequent itemsets. We can use equal-frequency discretization, equal width discretization, or Entropy-MDL discretization etc. A class variable is defined which provides training data for the rules.

Table III: Discrete dataset

	pgc	Age	bmi	insul	preg	tsf	dpf	dbp	Class variable
1	≥ 147.5	≥ 42.5	30.15 - 33.75	< 7	3.5 - 6.5	29.5 - 36.5	0.455 - 0.69	69 - 74.5	1
2	< 95.5	26.5 - 32.5	25.95 - 30.15	< 7	0.5 - 1.5	21.5 - 29.5	0.302 - 0.455	60.5 - 69	0
3	≥ 147.5	26.5 - 32.5	< 25.95	< 7	≥ 6.5	< 3.5	0.455 - 0.69	60.5 - 69	1
4	< 95.5	< 22.5	25.95 - 30.15	76.5 - 125.5	0.5 - 1.5	21.5 - 29.5	< 0.22	60.5 - 69	0
5	125.5 - 147.5	32.5 - 42.5	≥ 37.85	125.5 - 190.5	< 0.5	29.5 - 36.5	≥ 0.69	< 60.5	1
6	109.5 - 125.5	26.5 - 32.5	< 25.95	< 7	3.5 - 6.5	< 3.5	< 0.22	69 - 74.5	0
7	< 95.5	22.5 - 26.5	30.15 - 33.75	76.5 - 125.5	1.5 - 3.5	29.5 - 36.5	0.22 - 0.302	< 60.5	1
8	109.5 - 125.5	26.5 - 32.5	33.75 - 37.85	< 7	≥ 6.5	< 3.5	< 0.22	< 60.5	0
9	≥ 147.5	≥ 42.5	30.15 - 33.75	≥ 190.5	1.5 - 3.5	≥ 36.5	< 0.22	69 - 74.5	1
10	109.5 - 125.5	≥ 42.5	< 25.95	< 7	≥ 6.5	< 3.5	0.22 - 0.302	≥ 83	1
11	109.5 - 125.5	26.5 - 32.5	33.75 - 37.85	< 7	3.5 - 6.5	< 3.5	< 0.22	≥ 83	0
12	≥ 147.5	32.5 - 42.5	≥ 37.85	< 7	≥ 6.5	< 3.5	0.455 - 0.69	69 - 74.5	1
13	125.5 - 147.5	≥ 42.5	25.95 - 30.15	< 7	≥ 6.5	< 3.5	≥ 0.69	74.5 - 83	0
14	≥ 147.5	≥ 42.5	25.95 - 30.15	≥ 190.5	0.5 - 1.5	21.5 - 29.5	0.302 - 0.455	< 60.5	1
15	≥ 147.5	≥ 42.5	< 25.95	125.5 - 190.5	3.5 - 6.5	3.5 - 21.5	0.455 - 0.69	69 - 74.5	1
16	95.5 - 109.5	26.5 - 32.5	25.95 - 30.15	< 7	≥ 6.5	< 3.5	0.455 - 0.69	< 60.5	1
17	109.5 - 125.5	26.5 - 32.5	≥ 37.85	≥ 190.5	< 0.5	≥ 36.5	0.455 - 0.69	≥ 83	1
18	95.5 - 109.5	26.5 - 32.5	25.95 - 30.15	< 7	≥ 6.5	< 3.5	0.22 - 0.302	69 - 74.5	1
19	95.5 - 109.5	32.5 - 42.5	≥ 37.85	76.5 - 125.5	0.5 - 1.5	≥ 36.5	< 0.22	< 60.5	0
20	109.5 - 125.5	26.5 - 32.5	33.75 - 37.85	76.5 - 125.5	0.5 - 1.5	29.5 - 36.5	0.455 - 0.69	69 - 74.5	1

F. Apriori Algorithm

The Apriori algorithm is a groundbreaking algorithm developed by Srikant et al in 1994[2]. The association rule mining is an implication of the form A=>B where A and B both belongs to an itemset I where A∩B = □ . The rule A=>B holds in the transaction set Z with support s, where s is the fraction of transactions in D that contain A∪B i.e. P(A∪B). The rule A=>B has confidence c in the transaction set D, where c is the fraction of transactions in Z containing A that also contain B i.e. P(B|A). Rules that satisfy both a minimum support threshold and minimum confidence threshold are called strong. The set of frequent k-itemsets is commonly denoted L_k. Association rule mining can be performed in the following steps:

1. Finding all frequent itemsets: All of these itemsets will occur at least as frequently as a preset minimum support count, min_sup.
2. Generating strong association rules from the frequent itemsets: The generated rules must satisfy minimum support and minimum confidence.

Join Step: To find L_k, Generate k-itemsets by joining L_{k-1} with itself.

Prune Step: Any (k-1)-itemset that are not frequent cannot be a subset of a frequent k-itemset.

Pseudocode:

C_k: Candidate itemset of size k
L_k: frequent itemset of size k
L_1=find_frequent_1_itemsets;
for(k=2; L_k-1!=null; k++)
begin
C_k = Candidates generated from L_{k-1};
for each transaction t in database do
increment the count each candidates in C_k that are contained in t
L_{k+1} =candidates in C_{k+1} with min_supp
end
return $U_k L_k$

III. EXPERIMENTAL RESULTS

Once the data pre-processing is completed and the dataset is available in a discrete form, the association rules can be mined. In order to ensure that only the rules worth considering are generated, the minimum support threshold is set to 0.1 and minimum confidence threshold is set to 0.6. Using these parameters the apriori algorithm was applied on the dataset leading to the generation of 6 association rules. The rules are in descending order of their support factor.

A. Tables and Figures

Table IV: Risk Factors Associated with Distribution

PARAMETER	WEIGHTAGE	VALUES
Blood pressure	<60.5 60.5-69 69-74.5 74.5-83 >83	1 3 5 7 9
Triceps skin fold thickness	21.5-29.5 29.5-36.5 >36.5	1 5 9
Age	<22.5 22.5-26.5 26.5-32.5 33.75-37.85 >37.85	1 3 5 7 9
Body mass index	<25.95 25.95-30.15 30.15-33.75 33.75-37.85 >37.85	1 3 5 7 9
Plasma glucose concentration	<95.5 95.5-109.5 109.5-125.5 125.5-147.5 >147.5	1 3 5 7 9

Table V: Association Rules

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent	
0.30	0.61	0.49	0.61	2.05	0.15	insul=< 7 →	tsf=< 3.5
0.30	1.00	0.30	1.65	2.05	0.15	tsf=< 3.5 →	insul=< 7
0.14	0.62	0.23	2.14	1.27	0.03	preg=3.5 - 6.5 →	insul=< 7
0.13	0.60	0.22	2.21	1.24	0.03	preg=≥ 6.5 →	insul=< 7
0.13	0.65	0.20	2.43	1.33	0.03	dpf=< 0.22 →	insul=< 7
0.12	0.62	0.20	2.43	1.28	0.03	Age=≥ 42.5 →	insul=< 7

DOI 10.17148/IJARCCE.2016.54136

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

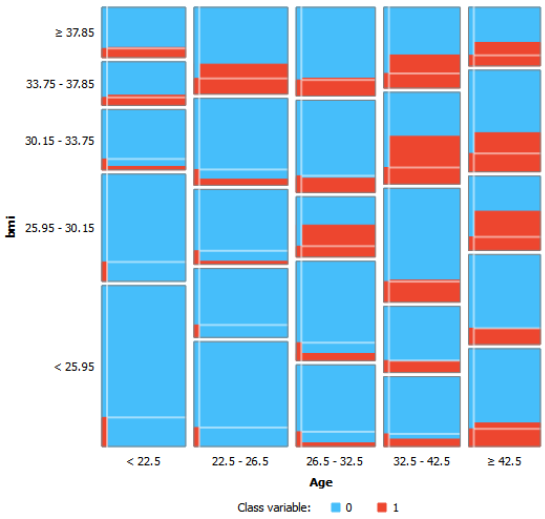

Fig 2: Distribution of Age and BMI factor corresponding to expected Frequency

These rules vary in confidence and support values as well as their length. A rule set which consist of only class variable at the consequent side is also denoted which eventually constitute to diabetic risk

Table VI: Association Rules with Class Variable Consequent

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		
0.31	0.63	0.49	1.34	0.97	-0.01	insul=< 7	→	Class variable=0
0.19	0.91	0.21	3.12	1.40	0.05	pgc=< 95.5	→	Class variable=0
0.18	0.61	0.30	2.20	0.94	-0.01	tsf=< 3.5	→	Class variable=0
0.18	0.61	0.30	2.20	0.94	-0.01	insul=< 7, tsf=< 3.5	→	Class variable=0
0.18	0.91	0.20	3.23	1.40	0.05	bmi=< 25.95	→	Class variable=0
0.17	0.74	0.23	2.81	1.14	0.02	preg=1.5 - 3.5	→	Class variable=0
0.17	0.78	0.21	3.03	1.19	0.03	Age=22.5 - 26.5	→	Class variable=0
0.16	0.81	0.20	3.25	1.24	0.03	pgc=95.5 - 109.5	→	Class variable=0
0.16	0.78	0.20	3.25	1.20	0.03	dpf=< 0.22	→	Class variable=0
0.15	0.66	0.23	2.86	1.01	0.00	preg=3.5 - 6.5	→	Class variable=0
0.15	0.71	0.21	3.12	1.08	0.01	dbp=60.5 - 69	→	Class variable=0
0.15	0.75	0.21	3.16	1.16	0.02	dbp=< 60.5	→	Class variable=0
0.15	0.73	0.20	3.18	1.13	0.02	bmi=25.95 - 30.15	→	Class variable=0
0.15	0.88	0.18	3.70	1.35	0.04	Age=< 22.5	→	Class variable=0
0.15	0.86	0.17	3.73	1.32	0.04	tsf=3.5 - 21.5	→	Class variable=0
0.14	0.69	0.20	3.18	1.07	0.01	pgc=109.5 - 125.5	→	Class variable=0
0.14	0.74	0.19	1.81	2.13	0.08	pgc=≥ 147.5	→	Class variable=1
0.14	0.79	0.18	3.70	1.21	0.02	preg=0.5 - 1.5	→	Class variable=0
0.13	0.62	0.21	3.09	0.95	-0.01	dbp=74.5 - 83	→	Class variable=0
0.13	0.62	0.20	3.18	0.96	-0.01	Age=26.5 - 32.5	→	Class variable=0
0.13	0.66	0.20	3.25	1.02	0.00	dpf=0.302 - 0.455	→	Class variable=0
0.13	0.64	0.20	3.25	0.98	-0.00	dpf=0.455 - 0.69	→	Class variable=0
0.13	0.67	0.20	3.27	1.03	0.00	dpf=0.22 - 0.302	→	Class variable=0
0.13	0.63	0.20	3.27	0.97	-0.00	dbp=69 - 74.5	→	Class variable=0
0.13	0.69	0.18	3.52	1.06	0.01	tsf=21.5 - 29.5	→	Class variable=0
0.12	0.92	0.13	5.05	1.41	0.03	insul=7 - 76.5	→	Class variable=0
0.10	0.90	0.11	5.81	1.38	0.03	bmi=< 25.95, insul=< 7	→	Class variable=0

Once the most significant association rules have been determined they can be used to form an investment plans for investors. For example consider the first association rule. When the low price drops and the open price rises, there is 100% confidence that the close price would fall. So an investor must be cautious before investing at that moment. Similarly if there is a rise in the high price and the open price then there in 97% confidence that there will be a rise in the close price. It can therefore be concluded that investing on that day would be advantageous. A study involving a dataset over larger intervals of time can give an even more comprehensive insight. By taking into consideration these rules the market state can be analyzed and investors can form their investment strategy in an accurate and efficient manner.

IV. CONCLUSION AND FUTURE WORKS

The aim of this research was to analyze diabetes screening datasets in order to aid doctors to come develop an preventive strategy. The diabetes mellitus data was taken and data pre-processing with conversion of continuous variable to discrete data was performed to for generating frequent itemsets. Once the data was pre-processed, A variation of Apriori algorithm was applied on the dataset to mine association rules. The most significant rules were extracted according to their confidence and support factor. The rules thus generated provides an insightful view to the risk of the factors.

For the purpose of this paper only a small dataset was taken. In order to arrive at a more comprehensive conclusion, a comprehensive medical records of patients must be taken into consideration. A large dataset is required for providing more accurate confidence factor and justifying the rule support.

REFERENCES

[1] Agrawal R., and Srikant R. "Fast Algorithms for Mining Association Rules in Large Databases", In Proc. 20th VLDB, PP. 478-499, Sept. 1994..

[2] Agrawal R., Imeielinski T., and Swami A. "Mining Association Rules between Sets of Items in Large Databases", Proceedings of the ACM SIGMOD Conference on Management of Data", Washington, D.C., PP. 207-216,1993

[3] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," Proceedings of the KDD, New York, pp.80-86, 1998.

[4] Gorgy J. Simon, Pedro J. Caraballo, Terry M. Therneau, Steven S. Cha, M. Regina Castro and Peter W. Li, "Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus", 2105.

[5] Yonatan Aumann and Yehuda Lindell. "A statistical theory for quantitative association rules". In Knowledge Discovery and Data Mining, 1999.

[6] Han J., and Kamber M. "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2nd edition, San Francisco, CA, 2006.

[7] Srikant R., Vu Q., and Agrawal R. "Mining Association Rules with Item Constraints", In Proc. 1997 International Conference Knowledge Discovery and Data Mining (KDD '97'), Newport Beach, CA, PP. 67-73, Aug. 1997.