www.ijecs.in

*International Journal Of Engineering And Computer Science ISSN:2319-7242*
*Volume 2 Issue 8 August, 2013 Page No. 2481-2485*

# Association Rule Mining on Type 2 Diabetes using FP-growth association rule

***Nandita Rane[1], Madhuri Rao[2]***

[1]Thadomal Shahani Engineering College
Department of Computer Engineering, India
*hinanditarane@gmail.com*

[2]Thadomal Shahani Engineering College

Department of Information Technology, India
*my_rao@yahoo.com*

**Abstract:** T***o diagnose diabetes disease at an early stage is quite a challenging task due to complex inter dependence on various factors. There is a critical need to develop medical diagnostic decision support systems which can aid medical practitioners in the diagnostic process. For detection of diabetes at an early stage, data mining association algorithm is being used. Detecting the cause of diabetes is very important in order to stop diabetes. The data set is taken from Nirmay Diabetes Super Speciality Center repository containing total instances 900 and approximately 30 attributes of type 2 diabetes mellitus. Proposed method explores step-by-step approach to help the health doctors to explore their data and to understand the discovered rules better. The discovery of knowledge from medical databases is important in order to make effective medical diagnosis. Most of the rules focus on the improving mining efficiency but, for the medical research, mining efficiency is not the most important factor, and there is a need to find more useful association rule and abandon more useless association rule.***

***We proposed a modified equal width binning interval approach to discretizing continuous valued attributes. The approximate width of desired intervals is chosen based on the opinion of medical expert and is provided as input parameter to the system. First we have converted numeric attributes into categorical form based on above techniques. FP-growth association mining is used to generate rules which are useful to identify general association in the data.***

**Keywords:** Apriori Association rule, Diabetes, FP-growth association mining**.**

## 1. Introduction

When For a long time, the cause of diabetes has been an unsolved problem and with the deepening of diabetes research, much information has been collected. There is lot of interesting information within the collected medical information. But, the traditional medical methods can only study the human organ from the microscopic view, and it lacks the capability to process large medical data set. To diagnose diabetes disease at an early stage is quite a challenging task due to complex interdependence on various factors. The aim of data mining is to extract the information from database and generate clear and understandable description of patterns. Advanced and reliable data mining techniques are used throughout this paper for discovery of unseen and useful information. Data mining is a process of extracting valid, previously unknown and actionable information from large data set. We can find the relations which are hided in the large medical data set using the data mining technology. But, the researches on the using of data mining method into the medical studies are not too much [4, 5]. According to the limitation of traditional medical methods, data mining methods is explained here, and hope it can pick up the lost key to open the door of finding the diabetes etiology.

The purpose is to find the factors that cause diabetes and their relations with diabetes. First step is to gain sufficient diabetes information, which includes the patients' basic condition (including the weight, age, sex etc.), medical history, and their relatives' medical history and examination results. In second step, make use of correlation algorithm which can find the frequent item sets from large scale data set. In [1] authors did the experiment on medical databases and generated the association rules between different attributes. In the first stage, data preprocessing is done in order to handle missing values and equal interval binning is applied with approximate values based on medical expert advice to Pima Indian Diabetes Data. In the second stage, a well-designed association rule mining method Apriori association rule mining that describe item sets that satisfy a minimum support criterion. These item sets to generate were used rules that satisfy a minimum confidence criterion. In this algorithm transactions were executed iteratively in a level-wise approach. i.e item sets containing one items were processed first, then item sets with two items were processed, and the process was repeated, continuously adding one item each time, until prescribed criteria were met. This method also considers two important measures coverage and confidence.

Basically association rule works in two steps:

(1) Generating item sets that pass a minimum support threshold.

(2) Generating rules that pass a minimum confidence threshold.

In [2] author performed a research to extract undiscovered information of diabetes in two phases. The first phase includes association rule generation and second phase uses classification technique. This research was based on 10,000 diabetes patients' record which was gathered from General Hospital Diabetes clinic, Sri Lanka.

Out of all those attributes, selected attributes for the flat table were; age, gender, diabetes type, education, occupation, monthly income, FBS (Fasting Blood Sugar), BMI (Body Mass Index), Potassium level, Cholesterol level, Sodium level, Diastolic blood pressure, Systolic blood pressure, Edema, and Wheezes. After the flat table preparation, diabetes patients' data table was available with more than 10,000 records. To remove the missing values in this created flat table, appropriate missing data handling mechanisms were carried out and a table with no missing values could be created. Next performed activity was the domain analysis. Before start analyzing the data with data mining techniques, domain analysis was used to gain an initial knowledge about the overall project domain. Filtering process was used to do the domain analysis. Advanced techniques needed to be used for further analysis to identify the validity of the assumptions made during the domain analysis. So, in the first phase of this paper, association rules were used and in this phase classification is used.

(3) Using the Association rules generating feature of WEKA, 10 rules, 50 rules, 100 rules, 500 rules and 1000 rules were generated in different times. After the analysis of the finally generated 1000 rules, three of the best rules that affects towards diabetes were selected considering the attribute combinations, confidence level, etc.

(4) The research in [3] used a new algorithm called as DiaApriori association algorithm for generating association rules and comparing the result with Apriori Association algorithm.

An information gathering system was established for the experiment, and the medical data were got from 3 first-level hospital of Handan, China. Proposed system used the data structure which includes 40 attributes. A data set of 57291 records was got with the time span form april, 2007 to april 2009. And among all the data, 13748 records are from diabetes patients, and the value of DIABETES is 1 for them. Among the 13748 record, 12956 records are valid for algorithm 1. We call DiaAprior with DIABETES=1, N=2. A total 174 direct association rules, and 563 two level indirect association rules are found. It can be seen that the rule number is much smaller than traditional apriori algorithm, because author has just choose the rule with the pattern *=>diabetes=1. And some of the meaningful results are shown below from the attributes which are listed in the given

Age=3 => diabetes=1;

DME=3 => diabetes=1;

AGE=3, dme=3 => diabetes=1;

Author evaluated the cause of diabetes is most likely be living habit including food habit and job condition, and seems to have nothing to do with height, heart disease, Nephropathy history, Liver disease and sex.

In this paper, two phase approach is introduced to generate association rules for detection of diabetes. First phase includes modified equal width binning interval approach to discretizing continuous valued attributes. Second phase consist of implementation of FP-growth association rule mining. Purpose of this paper is to generate strong association rules in order to detect cause of diabetes. The rest of the paper is organized as follows: In section 2, we briefly discuss the proposed method. We present the results in Section 3. Conclusion is highlighted in section 4.

## 2. Proposed Method

In first stage of our paper, we used preprocessing in order to handle missing values. Later we applied equal interval binning with approximate values based on medical expert advice to 'Nirmay Diabetes Super Speciality Center' diabetes data. Lastly we applied FP-growth association rule mining to generate the rules. We also consider two important measures coverage and confidence and lift. Strong association rule are generated using constraint based mining such as knowledge type constraints, data constraints, dimension/level constraints (direct and indirect rule), interestingness constraints (antecedent=>consequent) and rule constraints.
The framework is shown in Fig.1

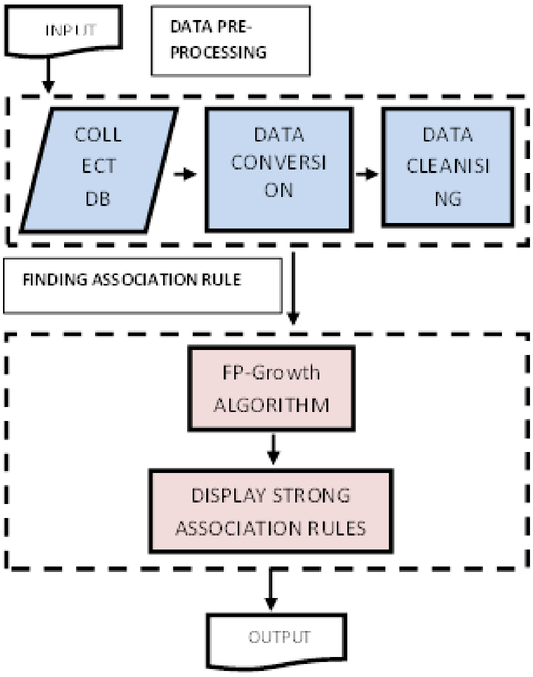

**Fig.1** System Architecture

### 2.1 Dataset

Out of 30 attributes approximately 12 attributes are being considered for finding association rules. The selected set of attributes contents patient id, age, family history, diabetes since those many number of years, height, weight, Systolic blood pressure, Diastolic blood pressure, fasting blood sugar, blood sugar post food, thyroid test, body mass index, urine alb creat ratio and investigation result.

## 2.2 Data Preprocessing

Missing values records are being deleted from the database. In this proposed method, we used preprocessing to improve the quality of data. Next we applied equal interval binning with approximate values based on medical expert advice to Nirmay Diabetes Super Specialty Center diabetes data. We applied FP-growth association rule algorithm to generate the rules. The generalization of rule may be further improved by considering the factors which influence the diabetes. In this paper we have included only type-2 diabetic patients.

Sort the attribute values and partition them into bins then smooth by bin means, bin median, or bin boundaries as shown in Table 2. Following are some examples of Bin boundaries.

Categories: Bin1:T2DM
Bin2:T1DM
Bin3:GDM
Bin4: Secondary Diabetic
Bin5:LADA
Age: Bin1:0-25
Bin2:26-35
Bin3:36-45
Bin4:46-55
Bin5:55 onwards
Sex: Bin1:M
Bin2:F
BMI: Bin1:0-18.4
Bin2:18.5-23.5
Bin3:23.6 Onwards
Family History: Bin1:1
Bin2:2
Bin3:3
Bin4: >4
Bin5: Nil

Blood Pressure: Bin1: Low
Bin2: Normal

Bin3: Low etc.

On the basis of Expert advised [1] selected attributes are converted to categorical form as shown in Table 2. Following are the range given by medical experts for attribute blood pressure [6].

- **Normal Blood Pressure** - Blood pressure reading below 120/80 is considered normal.
- **High Blood Pressure** - Blood pressure of 140/90 or higher is considered high blood pressure. If one or both numbers are usually high, you have high blood pressure.
- **Low Blood Pressure** - Blood pressure that is too low is known as hypotension. The similarity in pronunciation with hypertension can cause confusion.

## 2.3 Attribute/feature construction

Attribute/feature construction is done for attribute Body Mass Index (BMI).BMI is calculated from weight and height of the patient by the given formula

$$\text{BMI} = \frac{\text{Mass (Kg)}}{\text{(Height(m))}}$$

**Table 1**: Diabetes Test Attributes

| Sr. No | Test | Attributes |
|---|---|---|
| 1 | Heart Problem | CABG |
| | | H/o Cardiac Event |
| | | Stent |
| 2 | Liver Problem | SGPT (fatty liver disease) |
| 3 | Kidney Problem | Urine Alb Cretinine |
| | | Cretenine |
| 4 | Joint Problem | S.Uric Acid |
| 5 | Bad Cholesterol | LDL (Low Density Lipoprotein) |
| | | HDL (High Density Lipoprotein) |
| | | TC (Total Cholesterol) |
| | | TG (Trygylceride) |
| 6 | Sugar Problem | F1 |
| | | HbALc test (cost very high) |
| | | Average Sugar test |
| | | PP |
| 7 | Thyroid | TSH |

**Table 2:** Approximate equal Binning based on expert advice

| No | SEX | AGE | DM since | F/H | BMI | - |
|---|---|---|---|---|---|---|
| 1 | F | 4 | 15 | 1 | 3 | - |
| 2 | M | 4 | 10 | 2 | 3 | - |
| 3 | F | 4 | 1.5 | 1 | 2 | - |
| 4 | F | 4 | 0 | 0 | 3 | - |
| 5 | M | 4 | 0 | 0 | 2 | - |
| 6 | M | 4 | 10 | 2 | 3 | - |
| 7 | F | 4 | 1.5 | 1 | 3 | - |
| 8 | F | 4 | 4 | 1 | 3 | - |
| 9 | M | 4 | 0 | 0 | 3 | - |
| 10 | F | 4 | 3 | 2 | 3 | - |
| 11 | M | 4 | 3 | 1 | 3 | - |

## 2.4 Association Rule Mining Algorithm

**FP-growth structure**

FP-Growth: allows frequent itemset discovery without candidate itemset generation. Two step approach [5]:
Step 1: Build a compact data structure called the FP-tree. Built using 2 passes over the data-set.
Step 2: Extracts frequent itemsets directly from the FP-tree

**Benefits of the FP-tree Structure**

- □ Completeness
  - ■ Preserve complete information for frequent pattern mining
  - ■ Never break a long pattern of any transaction
- □ Compactness
  - ■ Reduce irrelevant info—infrequent items are gone
  - ■ Items in frequency descending order: the more frequently occurring, the more likely to be shared
  - ■ Never be larger than the original database (not count node-links and the *count* field)
- ➢ FP-Tree is constructed using 2 passes over the data-set:

Pass 1:

1. Scan data and find support for each item.
2. Discard infrequent items.
3. Sort frequent items in decreasing order based on their support.

Use this order when building the FP-Tree, so common prefixes can be shared.

Pass 2:

Nodes correspond to items and have a counter

1. FP-Growth reads 1 transaction at a time and maps it to a path
2. Fixed order is used, so paths can overlap when transactions share items (when they have the same prefix ).
   - – In this case, counters are incremented
3. Pointers are maintained between nodes containing the same item, creating singly linked lists (dotted lines)
   - – The more paths that overlap, the higher the compression. FP-tree may fit in memory.
4. Frequent item sets extracted from the FP-Tree.

## 3. Result

The result set of strong association rules is generated using object constraints and subjective constraints. Object constraints implemented are support, lift and confidence. Subjective constraints are as follows:
*Knowledge type constraints:* These specify the type of knowledge to be mined, such as association or correlation.
Data constraints: These specify the set of task-relevant data. Here we have used the diabetes database which includes various attributes such as body mass index, thyroid, blood pressure etc.
*Dimension/level constraints:* These specify the desired dimensions (or attributes) of the data, or levels of the concept hierarchies, to be used in mining. This constraint is implemented using selective attributes and selective patient records.
*Interestingness constraints:* These specify thresholds on statistical measures of rule interestingness, such as support, confidence, and correlation. In this paper FP-growth association mining uses support, confidence and lift as interesting measure.
*Rule constraints:* These specify the form of rules to be mined. Such constraints may be expressed as meta rules (rule templates), as the maximum or minimum number of predicates that can occur in the rule antecedent or consequent, or as relationships among attributes, attribute values, and/or aggregates.
Here we have defined rule as

LHS => RHS

LHS strictly contains following attribute set:{age, BMI, systolic blood pressure, diastolic blood pressure, thyroid, heart, HDL, LDL, TG, ......ETC} RHS strictly contains following attribute set:{diabetes: yes//no}

Following are some of the rules generated using FP-growth mining which help to detect cause of diabetes.

Rule 1.{Urine Alb. < 300}{Heart problem is absent}{Ceratine is Negative}{TG < 250}{Uric Acid is Absent}{LDL is Low}-->{T2DM is present} 96.66%

Rule 2. {Ceratine is Negative}{TG < 250}{SGPT has no value}{HDL < 35}{LDL is Low}-->{T2DM is present} 96.66%

Rule 3. {Urine Alb. < 300}{Ceratine is Negative}{TG < 250}{SGPT has no value}{HDL < 35}{LDL is Low}-->{T2DM is present} 96.66%

Rule 4. {Heart problem is absent}{Ceratine is Negative}{TG < 250}{SGPT has no value}{HDL < 35}{LDL is Low}-->{T2DM is present} 96.66%

Rule 5. {Urine Alb. < 300}{Heart problem is absent}{Ceratine is Negative}{TG < 250}{SGPT has no value}{HDL < 35}{LDL is Low}-->{T2DM is present} 96.66%

Rule 6.{Urine Alb. < 300}{Ceratine is Negative}{SGPT has no value}{HDL < 35}{Thyroid problem is absent}-->{T2DM is present} 96.65%

Rule 7. {Heart problem is absent}{Ceratine is Negative}{SGPT has no value}{HDL < 35}{Thyroid problem is absent}-->{T2DM is present} 96.65%

Rule 8. {Ceratine is Negative}{TG < 250}{Uric Acid is Absent}{HDL < 35}{LDL is Low}-->{T2DM is present} 96.65%

Rule 9. {Urine Alb. < 300}{Ceratine is Negative}{TG < 250}{Uric Acid is Absent}{HDL < 35}{LDL is Low}-->{T2DM is present} 96.65%
Rule 10. {Heart problem is absent}{Ceratine is Negative}{TG < 250}{Uric Acid is Absent}{HDL < 35}{LDL is Low}-->{T2DM is present} 96.65%

Rule 11. {Urine Alb. < 300}{Heart problem is absent}{Ceratine is Negative}{TG < 250}{Uric Acid is Absent}{HDL < 35}{LDL is Low}-->{T2DM is present} 96.65%

## 4. CONCLUSION

In this paper, we used preprocessing in order to improve quality of data which includes equal interval binning based on expert advice to Nirmay Diabetes Super Speciality Center repository. In later stage we applied frequent pattern growth association mining to generate strong association rules from patient's data set. The method not only can find direct factors but also find indirect factors that cause type 2 diabetes mellitus which may help health doctors to explore their data and understand the discovered rules better.

## REFERENCES

[1] B. M. Patil, R. C. Joshi, Durga Toshniwal, "Association rule for classification of type -2 diabetic patients", ", In Proceedings of 2010 Second International Conference on Machine Learning and Computing, pp. 330–334.

[2] S.M.Nuwangi, C. R. Oruthotaarachchi, J.M.P.P. Tilakaratna & H. A. Caldera, "Usage of Association rules and Classification Techniques in Knowledge Extraction of Diabetes", pp. 372–377.

[3] Xiaofeng Zhao**,** Liyan Jiao, Jinping An, Li Wang**,** "A Data Mining Method on the Study of Medical Information**",** In Proceedings of 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), pp. V9-183 – V9-186.

[4] Milan Z, Gou M, Peter K et.al. Mining diabetes database with decision trees and association rules[C]. In: Proceedings of the 15th IEEE Symposium on Computer- Based Medical Systems, 2002, pp. 867—871.

[5] J.Han, and M.Kamber, Data mining: Concepts and techniques, San Francisco: Morgan Kaufmann Publisher, pp.47- 94, 2006.

[6] http://health.nytimes.com, September 2012.

**Author Profile**

**Nandita Rane** received the B.E. and Persuing M.E. degrees in Computer Engineering from Thadomal Shahani Engineering College.

**Madhuri Rao** working as associate professor in Thadomal Shahani Engineering College in Information Technology department since 15 years and persuing Ph.d.