ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 4, Issue 1, pp: (37-41), Month: April 2017 – September 2017, Available at: www.paperpublications.org

Diabetes Mellitus Prediction System Using Data Mining

Yamini Amrale[1], Arti Shedge[2], Sonal Singh[3], Anjum Shaikh[4]

Savitribai Pune University

Abstract: **Now a days detection of patients with elevated risk of diabetes mellitus is developing critical to the improved prevention and overall health management of these patients. We aim to apply association rule mining to electronic medical records (EMR) to invent sets of risk factors and their corresponding subpopulations that represent patients which have high risk of developing diabetes. With the high linearity of EMRs, association rule mining generates a very large set of rules which we need to summarize for easy medical use. We reviewed four association rule set summarization techniques and conducted a comparative evaluation to provide guidance regarding their applicability, advantages and drawbacks. We proposed extensions to incorporate risk of diabetes into the process of finding an optimum summary. We evaluated these modified techniques on a real-world border line diabetes patient associate. We found that all four methods gives summaries that described subpopulations at high risk of diabetes with every method having its clear strength. In this extension to the Bottom-Up Summarization (BUS) algorithm produced the most suitable summary. The subpopulations identified by this summary covered most high-risk patients, had low overlap and were at very high risk of diabetes.**

Keywords: **Agile model, Association rules, Association rule summarization, Data mining, Survival analysis ,Fuzzy Clustering.**

I. INTRODUCTION

Diabetes mellitus may be a increasing outbreak that affects 25.8 million individuals within the U.S. (8% of population), and just about seven million of them don't hold they have the sickness polygenic disease results in vital medical problems as well as anaemia heart disease ,heart condition , cardiopathie, cardiovascular sickness, stroke, renal disorder, retinopathy, pathology and peripheral vascular disease[2]. Early prediction of patients at risk of increasing polygenic disease may be a major health care demand. Proper management of patients in danger with manner changes and or medications will decrease the chance of developing diabetes by upcoming future. Multiple risk factors have been known touching an out sized rate of the population. as an example, pre diabetes (blood sugar levels above traditional ratio however below the extent of criteria for diabetes) is gift in just about thirty fifth of a adult population and will increase absolutely the risk of polygenic disease three to tenfold counting on the presence of further associated risk factors, like avoirdupois, idiopathic, hyperlipidemia etc. Association rules area unit implications that associate a group of potentially interacting cons (e.g. BMI and therefore the presence of cardiovascular disease diagnosis) with elevated risk. The use of association rules is predominantly worthwhile, because in addition to quantifying the polygenic disorder risk, they conjointly promptly supply the MD with a rationale, namely the associated set of conditions. This set of conditions is used to give proper guidance treatment towards a additional customized and targeted preventive care or polygenic disorder management. A number of winning association rule set report techniques are planned however no clear steering exists concerning the concernment, strengths and weaknesses of those techniques. the main target of this palimpsest is to review and characterize four existing association rule report techniques and supply steering to practitioners in selecting the for most appropriate one. A common defect of those techniques is their inability to take polygenic disorder risk–a progressive

outcome–into account. In order to form these a lot of applicable, we had to minimize the change we incline to expand them to include information regarding progressive outcome variables.

II. LITERATURE SURVEY

1]A polygenic disease index is in essence a prophetic model that assigns a score to a patient supported his calculable risk of polygenic disease. Collins conducted an intensive survey of polygenic disease indices describes the dangerous factors and also the modeling technique that these evidence used.

2] They found that most indices were characterized by in nature and none of the surveyed indices have taken communications among the adverse factors into consideration. While we have a tendency to don't seem to be awake to any new polygenic disease index disclose after the survey, a current study specializing in the metabolic syndrome (of that polygenic disease may be a component) represents a big progress .

3] Used association rule mining to continuosly explore incident of guessable codes. The ensuring association rules dont constitute a polygenic disease index as a result of the study doesn't designate a specific outcome of interest and that they dont evaluate or predict the adverse of polygenic disease in patients, but they Invented some important associations between predictable codes.

4]We have at present a polygenic disease study wherever we target to get the relationships among diseases in the metabolic syndrome. We have a tendency to used identical outfit as this current study, however, we have a Aim to enclosed solely eight identification codes and age as predictors.

5]We invent association rules by considering a number of the eight belief codes, assessed the risk of polygenic disease that the rules confer on patients and presented the principles as a increasable graph described however patients towards from a healthy state to polygenic disease.

6]We indisputable that a approach found clinical meaningful association rules that square measure in step with our medical expectation. With simply eight predictor variables, the dimensions of the invented rule set was moderate thirteen important rules– and outcome, disturbance was simplest. Naturally, there is no rule-set account was mandatory.

III. TECHNOLOGIES

Data mining:

Data mining is the analysis step of Knowledge Discovery in Databases or KDD. It is an interdisciplinary subfield of computer science. This is the process of discovering patterns in large datasets ("big data") involving methods at the intersection of artificial intelligence, machine learning, database systems and statistics. The final goal of the data mining process is to extract relevant information from a dataset and transform it into an understandable format for further use.

Apriori Hybrid Algorithm

Apriori and Apriori Transaction ID algorithms can be combined into a hybrid algorithm, called Apriori Hybrid. It scales linearly with the number of transactions. In addition, the execution time reduce a little as the number of items in the database increases. As the average transaction size increases (while keeping the database size constant), the execution time increases only gradually. These experiments demonstrate the feasibility of using Apriori Hybrid in real applications involving very large databases.

Association rule summarization techniques

We apply rule for data set summarization techniques namely APRX-Collection, RP Global, BUS to evaluate the Risk of Diabetes mellitus. Predicition of Diabetes mellitus depends on Body condition, Tablets .Morbidity of the Observed patients in dataset subpopulation.

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 4, Issue 1, pp: (37-41), Month: April 2017 – September 2017, Available at: www.paperpublications.org

IV. ARCHITECTURE

System architecture is described below:

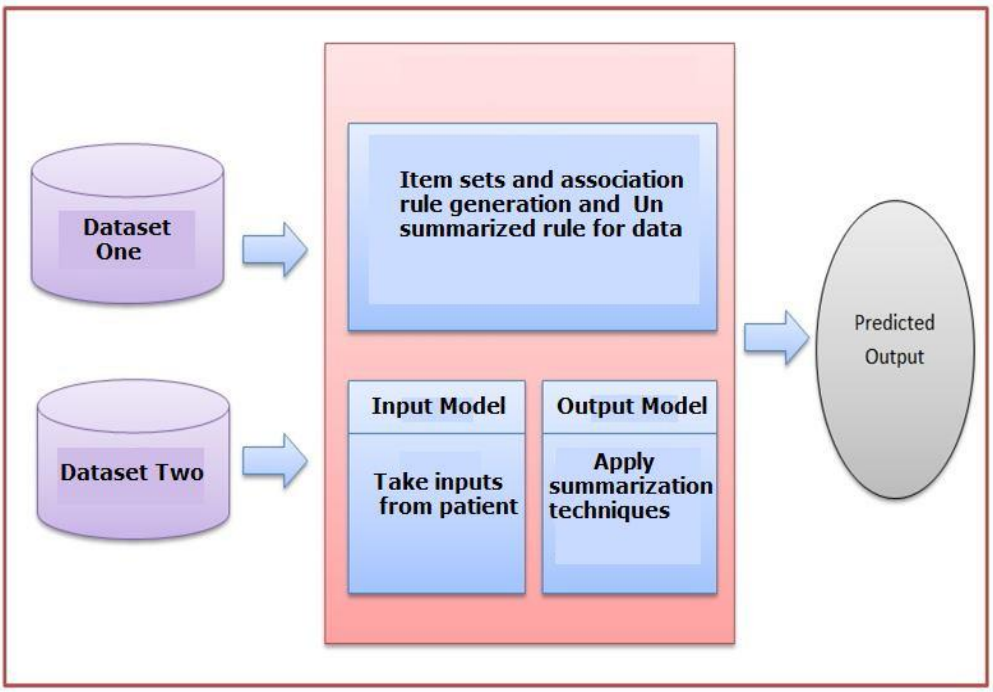

Figure 1. Architecture

This system consists of two datasets. First is big dataset and the other one is upload dataset.

Initially this application there is no Database Records. Summarization techniques are implemented in a Distributed Database not only in a single database.So have to ask permission to access the database of every medical Center Administrator .

- **Discover Item sets and association rule:**

Finding of association rules by using the apriori hybrid algorithm. The apriori hybrid algorithm, a variant of the well-known Apriori algorithm that discovers candidate set of items that contain specific items the item corresponding to the diabetes prediction results at final in this case.

- **Un summarized rule for data:**

It consists of the simillar risk and high risk of Observe patients in datasets. These values are calculated depends on the sugar level, BP, BMI, Tablets etc.

- **Apply summarization techniques:**

Rule for data set summarization techniques is used to evaluate the Risk of Diabetes mellitus. Predicition of Diabetes mellitus depends on Body condition, Tablets and Co., Morbites of the Observed patients in dataset subpopulation.

V. RESULT

The proposed technique intend to analyze the risk of diabetes mellitus. Here four association rule are used .Summarization techniques such as APRX-COLLECTION, RP Global, Top K and BUS. All these methods have its own advantage but BUS algorithm is the most efficient.

1. Age in Year.

2. Sex- (value 1: Male; value 0: Female).

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 4, Issue 1, pp: (37-41), Month: April 2017 – September 2017, Available at: www.paperpublications.org

3. Thal - (value 3: normal; value 6: fixed defect; value 7: reversible defect).

4. CA – number of major vessels colored by fluoroscopy (value 0-3).

5. Old peak – ST depression induced by exercise.

6. Exang - exercise induced angina (value 1: yes; value 0: no).

7. Chest Pain Type -(value 1:typical type1 angina, value 2: typical type 2 angina, value 3:non-angina pain; value 4: asymptomatic).

8. Restecg – resting electrographic results (value 0: normal; value 1: having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy).

9. Serum Cholestrol (mg/dl).

10 .Fasting Blood Sugar- (value 0: <120 mg/dl:value 1: >120 mg/dl;).

11. Trest Blood Pressure (mm Hg on admission to the hospital).

12. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: down sloping).

13. Thalach – maximum heart rate achieved.

14.Diabetes Disease Present - 0:No 1: Yes.

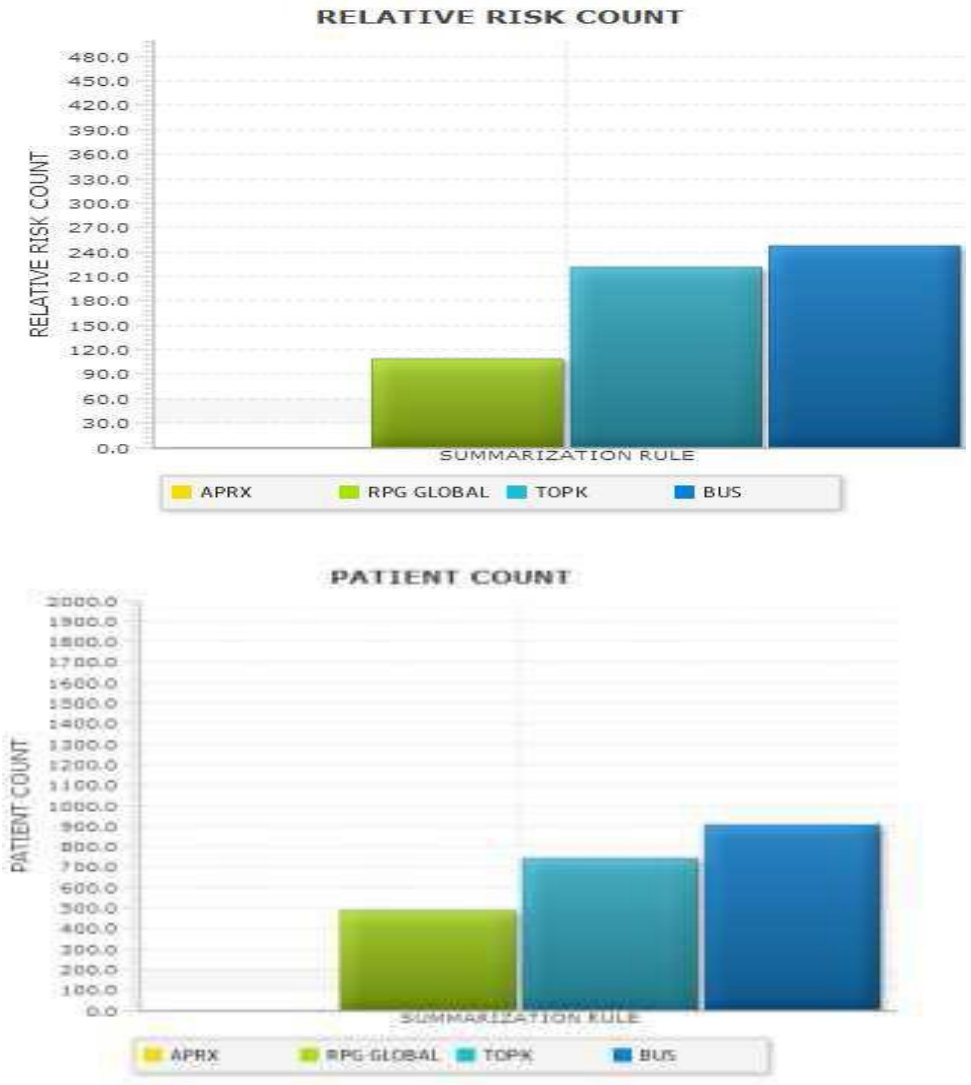

Figure 2. Relative Risk Count of Patient

VI. CONCLUSION

Here we have studied diabetes mellitus prediction system using data mining solution. And association rule mining to identify sets of risk factors and the corresponding patient subpopulations that significantly increased risk of diabetes. And many number of association rules were discovered including the clinical interpretation results. For this method, the number of rules are used for health interpretation makes feasible.

REFERENCES

[1] F. Afrati, A. Gionis, and H. Mannila, "Approximating a collection of frequent sets," in *Proc. ACM Int. Conf. KDD*, Washington, DC, USA, 2004.

[2] "Fast algorithms for mining association rules," R. Agrawal and R. Srikant, in *Proc. 20th VLDB*, Santiago, Chile, 1994.

[3] "A statistical theory for quantitative association rules," Y. Aumann and Y. Lindell, in *Proc. 5th KDD*, New York, NY, USA, 1999.

[4] "Use of association rule mining to assess diabetes risk in patients with impared fasting glucose," P. J. Caraballo, M. R. Castro, S. S. Cha, P. W. Li, and G. J. Simon, in *Proc. AMIA Annu. Symp.*, 2011.

[5] A Fuzzy Rule-Based Clustering Algorithm,proc IEEE transaction Eghbal G.Mansoori, ―FRBC:. Fuzzy sytems. Vol 19no.5, October 2011.

[6] "Mining association rules with item constraints," R. Srikant, Q. Vu, and R. Agrawal, in *Proc. AAAI*, 1997.

[7] Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. The New England Journal of Medicine, 346(6), 2002.

[8] In SIAM International Conference on Data Mining Xiaoxin Yin and Jiawei Han. CPAR: Classification based on predictive association rules (SDM), 2003.