|| Volume 1 ||Issue 6 ||Sept 2016||ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

Survey on Assess Co-Morbid Risk of Diabetes Mellitus by using Split and Merge Association Rule Summarization Techniques

Mr. A. A. Dange[1], Prof. Saad Siddiqui[2]
Department of Computer Science & Engineering , Everest Educational Society's Group of Institutions, Aurangabad, Maharashtra, India.

***Abstract*— A Diabetes is a life-threatening issue in modern health care domain. With the use of data mining techniques, diabetes factors and co morbid risk conditions associated with diabetes has found. In order to stifle the evolution of diabetes mellitus, applies distributed association rule mining and summarization techniques to electronic medical records. This helps to discover set of risk factors and co morbid conditions in distributed medical dataset using frequent item set mining. In general, association rule mining (ARM) generates bulky volume of data sets which need to summarize certain rules over medical record. This encompasses a novel approach to find the common factors which lead to high risks of diabetes and co morbid conditions associated with diabetes. This performs both association rule mining and association rule summarization techniques with improved classification algorithms. Exiting systems aim to apply association rule mining to electronic medical records to discover sets of risk factors and their corresponding subpopulations that represent patients at particularly high risk of developing diabetes. Given the high dimensionality of EMRs (Electronic Medical Records), association rule mining generates a very large set of rules which we need to summarize for easy clinical use. The existing system reviewed four association rule set summarization techniques and conducted a comparative evaluation to provide guidance regarding the diabetes risk prediction. In the field of medical domain, the prediction of diabetes and its Co-Morbid in earlier stage is important. We propose a set of methods to perform the Co-Morbid prediction. The propose technique named as SAM (Split and Merge), which is based on fast distributed quantitative association rule mining and rule filtering for prediction co morbid conditions associated with diabetes. SAM algorithm is used to discover the frequent data item sets and summarized data sets. In performance comparison of proposed SAM with existing BUS approach based on prediction efficiency SAM is better than BUS.**

Keywords:- Association Rule Mining(ARM), Diabetes, Co-Morbid, Risk Prediction, Electronic Medical Record (EMR), Split and Merge(SAM), Data Mining, Bottom-up Summerization (BUS).

I INTRODUCTION

Diabetes mellitus is commonly referred to as diabetes of pancreas. Today diabetes is largely distributed all over the world that affects almost 29.1 million Americans. From the survey out of these 29.1 million Americans 21.0 million were diagnosed and 8.3 million were undiagnosed. Diabetes leads to significant co-morbid conditions including stroke, heat disease, retinopathy, nephropathy, hypertension etc. As of 2014, totally 387 million people have diabetes worldwide. This is equal to 8.3% of the adult population. In the years 2012 to 2014 diabetes is appraisal to have resulted in 1.5 to 4.9 million deaths per year. Diabetes doubles the risk of death. The number of people with diabetes is anticipated to rise to 592 million by 2035. Early recognition of disease and its risk prediction of patients using their EMR is a major healthcare process. Appropriate management of patients at risk with lifestyle changes and/or medications can decrease the risk of developing diabetes by 30% to 60%. Multiple risk factors have been identified affecting a large proportion of the population. For example, pre-diabetes (blood sugar levels above normal range but below the level of criteria for diabetes) is present in approximately 35% of the adult population and increases the absolute risk of diabetes 3 to 10 fold depending on the presence of additional associated risk factors, such as obesity, hypertension, hyperlipidemia, etc. Comprehensive medical management of this large portion of the population to prevent diabetes represents an unbearable burden to the healthcare system.

The symptoms of marked hyperglycemia include which includes polyuria, polydipsia, weight loss, sometimes with polyphagia, and blurred vision impairment. The growth and susceptibility to certain infections may also accompany chronic hyperglycemia in so many patients. Life-threatening consequences of uncontrolled diabetes are hyperglycemia with keto acidosis and the non ketotic hyperosmolar syndrome. Long-term complications of diabetes causes retinopathy with potential vision loss, nephropathy leading to renal failure,

peripheral neuropathy with foot ulcers, amputations, Charcot joints, and autonomic neuropathy causing gastrointestinal, genitourinary, and cardiovascular symptoms and sexual dysfunction. Patients with diabetes have an increased incidence of atherosclerotic cardiovascular, peripheral arterial and cerebrovascular disease. Hypertension and abnormalities of lipoprotein metabolism are often found in people with diabetes. The vast majority of cases of diabetes fall into two broad etiopathogenetic categories. In one category, type 1 diabetes, the cause is an absolute deficiency of insulin secretion. Individuals at increased risk of developing this type of diabetes can often be identified by serological evidence of an autoimmune pathologic process. In the other, much more prevalent category, type 2diabetes, the cause is a combination of resistance to insulin action and an inadequate compensatory insulin secretary response. In the latter category, a degree of hyperglycemia sufficient to cause pathologic and functional changes in various target tissues, but without clinical symptoms, may be present for a long period of time before diabetes is detected. During this asymptomatic period, it is possible to demonstrate an abnormality in carbohydrate metabolism by measurement of plasma glucose in the fasting state or after a challenge with an oral glucose load.

II CLASSIFICATION OF DIABETES MELLITUS AND OTHER CATEGORIES OF GLUCOSE REGULATION

Assigning a type of diabetes to an individual often depends on the circumstances present at the time of diagnosis, and many diabetic individuals do not easily fit into a single class. For example, a person with gestational diabetes mellitus may continue to be hyperglycemic after delivery and may be determined to have, in fact, type 2diabetes. Alternatively, a person who acquires diabetes because of large doses of exogenous steroids may become normoglycemic once the gluco corticoids are discontinued, but then may develop diabetes many years later after recurrent episodes of pancreatitis. Another example would be a person treated with thyroids who develops diabetes years later. Because thiazides in themselves seldom cause severe hyperglycemias, such individuals probably have type 2 diabetes that is exacerbated by the drug. Thus, for the clinician and patient, it is less important to label the particular type of diabetes than it is to understand the pathogenesis of the hyperglycemia and to treat it effectively.

A Abnormality Detection

In data mining, abnormality detection is the search for data items in a dataset which do not conform to an expected pattern. The patterns thus detected are called anomalies and often translate to critical and actionable information in several application domains. Anomalies are also referred to as outliers. A good definition of an abnormality is as follows “an abnormality is an observation that deviates so much from other observations as to arouse suspicions that it was caused by a different mechanism”. Distance-based measures have been used in algorithms to delineate outliers or abnormal records from normal records.

Abnormality detection is the process of identifying abnormal pattern from set of objects. Abnormality detection aims to identify a small group of instances which deviate remarkably from the existing data. A well-known definition of “abnormality” is given “an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism”.

- Abnormality detection refers to the problem of finding patterns in data that do not conform to expected normal behavior.
- These anomalous patterns are often referred to as outliers, anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants in different application domains.

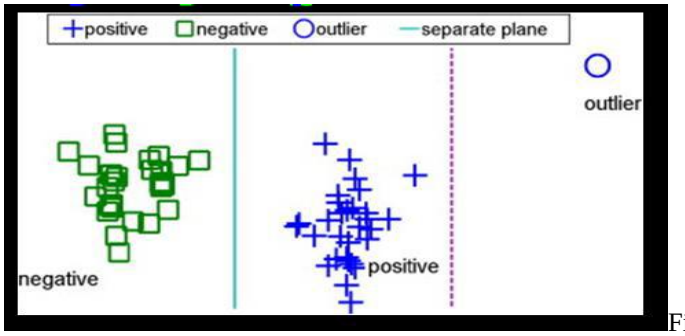

Figure 1: Abnormality Detection

Exiting systems aim to apply association rule mining to electronic medical records to discover sets of risk factors and their corresponding subpopulations that represent patients at particularly high risk of developing diabetes. Given the high dimensionality of EMRs (Electronic Medical Records), association rule mining generates a very large set of rules which we need to summarize for easy clinical use. The existing system reviewed four association rule set summarization techniques and conducted a comparative evaluation to provide guidance regarding the diabetes risk prediction.

The Disadvantages include:

- Less accuracy
- Need more training data
- Processing overhead
- Only predicts the diabetes risk failed to perform the co-morbid conditions and its risk.

III LITERATURE SURVEY

Data mining has been actively involved in the intelligent medical system which is stated in papers [3][8][9]. The associations of disorders and the real causes of the disorders and the effects of symptoms that are impulsively seen in patients can be evaluated by the users via data mining techniques. In the application of health domain, Bulky databases can be applied as the input data to the system to find the association between attributes. The effects of associations have not been evaluated adequately in the literature. This have been

explored the relationships of hidden knowledge placed among the large Medical databases. This has been searched relevant attributes by means of finding frequent items using candidate generation.

Learning of the risk factors associated with diabetes helps health care professionals to identify patients at high risk of having diabetes disease. Statistical analysis and data mining techniques [10] helps to healthcare professionals in the diagnosis of heart oriented diseases. Such analysis has identified the disorders of the heart and blood vessels, using statistical values, and this includes cerebrovascular disease known as stroke, coronary heart disease also known as heart attacks, raised blood pressure [hypertension], heart failure, rheumatic heart disease, peripheral artery disease and congenital heart disease.

In paper [11] presented an efficient approach for the prediction of heart attack risk levels from the heart disease dataset using clustering techniques. Initially the heart disease dataset is clustered using the K-means clustering algorithm, which will extract the attributes and data relevant to heart attack from the dataset. This allows the dataset to be portioned into k fragments. This approach mines the frequent patterns subsequently from the extracted data related to heart disease. This used MAFIA a maximal frequent Item set algorithm, which is a machine learning algorithms trained with selected significant patterns. This basically predicts the heart attack. Additionally some technique from [12] resolves the prediction accuracy oriented issues. The approach utilizes the ID3 algorithm as a training algorithm. The results showed that the designed prediction system is capable of predicting the heart attack effectively. But the prediction of diabetes is slightly different from the above.

A study on the prediction of heart attack risk levels from the heart disease database with the use of bayes algorithms has conducted in [9]. This utilized the basic data mining classification techniques with 11 important attributes. Mainly that is concentrated the bagging technique. From the results of [8] bagging technique is accurate and capable than the J48 and Bayesian classification algorithms for heart attack prediction.

In a predictive model, scores will be calculated to estimate the risk of diabetes, so there is a need of diabetes index. The need of diabetes index has been recognized in [2], this conducted a survey regarding the diabetes risk factors. They found that most indices were additive in nature and none of the surveyed indices have taken interactions among the risk factors into account.

Paper used association rule mining to systematically explore associations of diagnosis codes. The resulting association rules do not constitute a diabetes index because the study does not designate a particular outcome of interest and they do not assess or predict the risk of diabetes in patients, but they discovered some significant associations between diagnosis codes. Understanding Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges as to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost , but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression.

Random forests are an effective tool in prediction. Because of the Law of Large Numbers they do not over fit. Injecting the right kind of randomness makes them accurate classifiers and repressors. Furthermore, the framework in terms of strength of the individual predictors and their correlations gives insight into the ability of the random forest to predict. Using out-of-bag estimation makes concrete the otherwise theoretical values of strength and correlation. For a while, the conventional thinking was that forests could not compete with arcing type algorithms in terms of accuracy. The results dispel this belief, but lead to interesting questions. Boosting and arcing algorithms have the ability to reduce bias as well as variance .

IV PROPOSED SYSTEM

The propose technique named as SAM (Split and Merge), which is based on fast distributed quantitative association rule mining and rule filtering for prediction co morbid conditions associated with diabetes. In the field of medical domain, the prediction of diabetes and its Co-Morbid in earlier stage is important. We propose a set of methods to perform the Co-Morbid prediction. This chapter specifies the process included in the proposed system.

A HARS Technique

- *Association Rule in Health care*

Our propose technique named as HARS (Hybrid Association Rule Summarization), which is based on fast distributed quantitative association rule mining and rule filtering for prediction co morbid conditions associated with diabetes. In the field of medical domain, the prediction of diabetes and its risks in earlier stage is important. We propose a set of methods to perform the risk prediction. Data mining technique such as Association rule mining is applied to discover patterns or associations encoded in the data. Association rule is in the form of A ➔B where *A* is the antecedent and *B* is the consequent and A and *B* are sets of predicates. The association rule is based on concepts of support and confidence. The support is the probability of a transaction/event in the database containing both the antecedent and the consequent and the confidence is the probability that a record that contains the antecedent also contains the consequent. If $I=\{i_1,i_2,\ldots i_n\}$ is a set of items, a transaction *T* is a subset of *I*, and dataset *D* is set of transaction. Association rule then means finding rules in the form of

$$R=> i[S,C]$$

where , R belongs to I and $i \in I$, *S* is the support and *C* is the confidence. The support, *support D(X)* of an item *X* in the dataset can be defined as

$$SupportD(X)= CountD(X) / |D|$$

where *countD(X)* is the number of transactions in *D* containing *X*. The user specifies a minimum support (*min_sup*) and confidence value (*min_conf*). An itemset is said to be frequent if its support is greater than the *min_sup* value specified. Number

of algorithms are proposed for discovering association rules from large database (Agrawal et al 1994; Han, et al 2000; Berzal et al 2001).

The *apriori algorithm* is on the most popularly used algorithms for discovering association rules. The algorithm first discovers all frequent itemsets I_F belongs to I which has a value of support equal to or greater than *min_sup*. The algorithm merges all the frequent itemsets until no more I_F are found. On generation of the frequent itemsets, it is split in any possible way into a rule antecedent R belongs to I and a rule consequent $i \in I$ such that R U $i=I_F$ and $R \cap i=\Phi$. The confidence is calculated for each rule candidate and the rule is output if the confidence is above *min_conf*.

Heath attributes are the data related to disease diagnosis. In co-morbid association rule mining, it seeks to find associations among transactions that are encoded explicitly in a database, association rule mining seeks to find patterns in spatial relationships that are typically not encoded in a database but are rather embedded within the health care framework. These rules must be extracted from the data prior to the actual association rule mining. Association mining rule is applied on the data collected to discover patterns. Every node mines patterns in the following form:

$$A1 \wedge \ldots \wedge Am => E[S,C]$$

where event *E* occurs at node *n* with support *S* and confidence *C* given that antecedents A_i holds true. Antecedents are in the form of

$$A_i = (E_i, D_i, T_i, N_i)$$

- *BUS*

BUS (Bottom-Up Summarization) algorithm operates on the patients and not on the rules associated with. In BUS redundancy in terms of rule expression can occur. The algorithm controls the redundancy in the patient space by applying the rule expression handled earlier.

- *SAM Algorithm*

In this paper the genetic algorithm is applied over the rules fetched from SAM algorithm which is an extension of Apriori. The proposed method for generating association rule by SAM is as follows:

1. Initiate the process by uploading the dataset D

2. Load a sample of records from the D that fits in the memory.

3. Apply SAM algorithm to find the frequent itemsets A with the minimum support.

4. Set R= Φ where R is the rule set, which contains the association rule.

5. Perform selection criteria specification using genetic algorithm.

6. Represent each frequent item set of A as quantity data using the combination of representation.

7. Select the two members from the frequent item set and predict the risk by the genetic algorithm.

8. The next iteration is applying the crossover and mutation on the selected rule set to generate the final priority association rules.

9. Find the fitness function for each rule x→y and check the following condition.

10. If (fitness function > min confidence)

11. Set R = R U {x →y}

12. If the desired number of generations is not completed, then go to Step 5.

The above steps explain the process of HARS. The algorithm terminates the execution when the condition is met.

The criteria are defined by the genetic algorithm. It also terminates execution when the total number of generations specified by the user is reached. The support of an association pattern is the percentage of task-relevant data tuples for which the pattern is true.

Let us assume, A is the combination of two attributive and its quantitative measures {Age} and {bmi}. And B is {bmi} and {cholesterol}. To calculate the support and confident of A and to find the rule mining is specified below.

Minimum Support Threshold

IF A => B

$$Support\,(A => B) = \frac{\text{no\_tupples\_containing\_both\_A\_and\_B}}{\text{total\_number\_of\_tupples}}$$

Minimum Confidence Threshold

Confidence is defined as the measure of certainty or trustworthiness associated with each discovered pattern.

IF A => B

$$Confidence\,(A => B) = \frac{\text{no\_tupples\_containing\_both\_A\_and\_B}}{\text{no\_of\_tupples\_containing\_A}}$$

SAM algorithm is used to discover the frequent data item sets and classifying summarized and summarized data sets.

- *Comparative Study*

The comparative evaluation of extended summarization techniques like BUS and HARS that provides guidance to practitioners in selecting an appropriate algorithm

for a similar problem. There are many techniques for summarization of association rules to detect diabetes mellitus. Such as RPGlobal, FPApprox, TopK, BUS out of which BUS is optimal that provide accurate guidance to the practitioners in early prediction of risk of diabetes mellitus[1]. Based on the two real world dataset, diabetes and co morbid conditions associated with diabetes were assessed.

TABLE I. DIABETES DATASET
(Minimum Support=40% and Minimum Confidence=50%)

Record Id	Age	Bmi	Sbp	Dbp	Cholesterol
100	23	25	90	60	5.5
200	25	32	94	120	5.7
300	29	30	120	89	5.9
400	34	31	160	100	6.2
500	38	33	99	120	5.0

The table I represent the sample dataset used for the experiments and that contains patient record id, age, body mass index (bmi), systolic blood pressure (Sbp), and diastolic blood pressure and cholesterol values.Here we compared our proposed HARS with existing rule set summarization technique BUS to predict the risk of diabetes mellitus and co-morbid conditions. Proposed HARS retained slightly more redundant than BUS which allow us to have better patient coverage and help to correctly predict the risk.

V CONCLUSION

The study proposed a new extending association rule summarization and co-morbid risk predication technique for diabetes. During literature survey we come across two main problems, which are prediction accuracy and predication error. By applying HARS(Hybrid Association Rule Summarization) with SAM (Split and Merge) Algorithm above problems can be overcome. The SAM represents the effective rule specification criteria. The system effectively identify the co-morbid conditions of the diabetes disease and its sub types such as heart disease, retinopathy, neuropathy etc. The study shows that iterative SAM shows better predication accuracy compared to traditional techniques.

REFERENCES

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

[2] Larose, Daniel T. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014.

[3] Kotsiantis, Sotiris, and Dimitris Kanellopoulos. "Association rules mining: A recent overview." *GESTS International Transactions on Computer Science and Engineering* 32.1 (2006): 71-82.

[4] RakeshAgrawal and RamakrishnanSrikant. Fast algorithms for mining association rules. In *VLDB Conference*, 1994.

[5] Han, Jiawei, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation." *ACM SIGMOD Record*. Vol. 29. No. 2. ACM, 2000.

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

[7] Pedro J. Caraballo, M. Regina Castro, Stephen S. Cha, Peter W. Li, and Gyorgy J. Simon. Use of association rule mining to assess diabetes risk in patients with impared fasting glucose. In *AMIA Annual Symposium*, 2011.

[8] Centers for Disease Control and Prevention. National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the united states, 2011. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. http://www.cdc.gov/diabetes/pubs/factsheet11.htm, 2011.

[9] VarunChandola and Vipin Kumar. Summarization – compressing data into an informative representation. *Knowledge and Information Systems*, 2006.

[10] Gary S Collins, Susan Mallett, Omar Omar, and Ly-Mee Yu. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Medicine*, 2011.

[11] 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.

[12] Mohammad Al Hasan. Summarization in pattern mining. In *Encyclopedia of Data Warehousing and Mining, (2nd Ed)*. Information Science Reference, 2008.