ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

Prediction of Co-Morbid Conditions Associated with Diabetes using Split and Merge Algorithm

Priya.R[1], Roshma.R[2]

Head, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India[1]

M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India[2]

ABSTRACT: Diabetes is a life-threatening issue in modern health care domain. With the use of data mining techniques, diabetes risk factors and co morbid conditions associated with diabetes has found. In order to stifle the evolution of diabetes mellitus, we applied distributed quantitative association rule mining and summarization techniques in Electronic Medical Records (EMR). This is a hybrid summarization technique, which leverages rule criteria using the genetic algorithm and SAM (Split and Merge) algorithm for high diabetes risk prediction. This helps to discover a set of risk factors and co morbid conditions in distributed medical dataset. 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 risk of diabetes and co morbid conditions associated with diabetes.

KEYWORDS: Data mining, association rule mining, association rule summarization, Genetic algorithm, Diabetes risk prediction, Electronic Medical Records (EMR).

I. INTRODUCTION

Diabetes mellitus is commonly referred to as diabetes, which is a disease of the pancreas, an organ behind the human stomach that produces the hormone insulin. Diabetes is a widespread disease around the world that affects 29.1 million Americans [4]. From the survey 21.0 million were diagnosed and 8.3 million were undiagnosed. Diabetes leads to significant medical complications or Co-Morbid Conditions including stroke, heart disease, retinopathy, nephropathy, neuropathy, Dyslipidemia, peripheral vascular disease and hypertension. Early disease recognition and its risk finding 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 35% to 68% [5], [6]. 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 which are obesity, hyperlipidemia and hypertension etc. [7]

With the use of association rules, the risk of diabetes will be identified. The rule discovered by the implications that associate a set of potentially interacting conditions with important risk. For example the presence of hypertension and BMI are the important conditions of diabetes. The use of association rules is particularly valuable in distributed database, because in distributed databases should perform local and global pruning.

In addition this helps to enumerate the risk of diabetes; they also readily provide the physician with a “justification", namely the associated set of conditions. This set of conditions can be used to guide treatment towards a more personalized and targeted preventive care or diabetes management. In general Association rules are if-then statements that help to expose associations between unconnected data in a distributed database, relational database or other information repository. ARM is used to find the associations between the conditions which are frequently come together.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

Our work presents a clinical application using enhanced association rule mining to identify sets of co-morbid conditions (and the patient subpopulations who suffer from these conditions) that imply significantly increased risk of diabetes.

Association rule mining on this extensive set of variables resulted in an exponentially large set of association rules. This extended four popular association rule set summarization techniques by incorporating the risk of diabetes into the process of finding an optimal summary. The main contribution is a comparative evaluation of these extended summarization techniques that provides guidance to practitioners in selecting an appropriate algorithm for a similar problem.

To handle the existing problem, machine learning techniques have been developed to gain knowledge automatically from examples or raw data. Here, a weighted Association rule-based clinical decision support system is presented for the diagnosis of heart disease and diabetes, automatically obtaining knowledge from the patient's clinical data.
The proposed system for the risk prediction of heart and diabetes patients consists of two phases:

(1) Automated approach for the generation of distributed association rules and summarization using quantitative rule mining method.

(2) Developing a Genetic-based decision support system to improve prediction accuracy.

(3) Implementing a new association rule summarization algorithm named as SAM (Split and Merge algorithm)

In the first phase, we have used the attribute selection and attribute weightage method to obtain the weighted association rules. Then, the proposed system is constructed in accordance with the weighted rules and chosen attributes. The experimentation is carried out on the proposed system using the datasets obtained from the UCI repository.

II. BACKGROUND OR RELATED WORK

Data mining has been participated an imperative role in the intelligent medical systems which is stated in paper [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 [13]. This utilized the basic data mining classification techniques with 11 important attributes. Mainly that is concentrated the bagging technique. From the results of [13] 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],

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

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 [14] 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.

III.PROBLEM SPECIFICATION

Several authors in literature aim to apply association rule mining in 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 need to summarize for easy clinical use. The system [1] reviewed four association rule set summarization techniques and conducted a comparative evaluation to provide guidance regarding the diabetes risk prediction. That extended the bottom up summarization technique and compared with the TOP-k methods. Finally this paper concluded BUS is optimal than the others. Even though the paper results with optimal technique, that system suffers from time constraints. This needs the candidate generation every time. So we proposed a frequent itemset mining technique with time and memory constraint.

IV. PROPOSED SYSTEM

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. This chapter specifies the process included in the proposed system.

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 HARM. 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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

Minimum Support Threshold

IF A => B

$$\text{Support } (A \Rightarrow B) = \frac{\text{no\_tuples\_containing\_both\_A\_and\_B}}{\text{total\_no\_of\_tuples}}$$

Minimum Confidence Threshold

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

IF A => B

$$\text{Confidence } (A \Rightarrow B) = \frac{\text{no\_tuples\_containing\_both\_A\_and\_B}}{\text{No\_of\_tuples containing A}}$$

V. EXPERIMENTS AND RESULTS

Dataset:

We perform the experiment on the Mayo Clinic patient data obtained during the study period from 1/1999 to 12/2004 with follow-up information available until the summer of 2010. Another dataset used in this study is the Cleveland Clinic Foundation, which is named as Heart disease data set available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease. The data set has 13 attributes. The experiment takes Heart disease dataset from UCI repository.

The dataset contains 13 attributes considered are: age, sex, FBS (fasting blood sugar > 120 mg/dl), chol (serum cholesterol in mg/dl), restecg (resting electrocardiographic results), trestbps (resting blood pressure), thalach (maximum heart rate achieved), exang (exercise induced angina), slope (the slope of the peak exercise ST segment), oldpeak (ST depression induced by exercise relative to rest). there are a total of 750 patient records in the database. Based on the two real world dataset, diabetes and co morbid conditions associated with diabetes were assessed.

RecordID	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

Table 1.0: Diabetes dataset (Minimum support = 40%, minimum confidence = 50%)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

The table 1.0 represents 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.

Rules (Sample)	Support	Confidence
{Age: 20..30} and {bmi >30}=>{cholesterol >5.0}	20%	100%
{Dbp >100}=> {cholesterol >5.0}	40%	66.6%

Table 2.0: Support and confidence value with quantitative association rule mining

Table 2.0 represents the quantitative rule mining with minimum support and confidence threshold. Table 2.0 shows this mapping for the attributes of the diabetes dataset table given in Figure 1.0. From the dataset, Age is segmented into two partition: 20...29 and 30...34. The other attributes, such as bmi ,Dbp,sbp and cholesterol are numeric attributes. Finally this prioritizes the rules by applying genetic based rule miner for effective disease risk prediction. This has been implemented the rule summarization technique using the Java platform that is used in the research for presenting the results.

In this chapter we evaluated our proposed work with existing rule set summarization techniques such as Top-K, BUS to predict the Risk of Diabetics Millets and co morbid conditions. Finding the risk and Predication the co morbid conditions associated with Diabetics based on genetic miner. This evaluates the patient attributes such as laboratory results, Co-Morbites of the patient subpopulation and medications.

TOP-K

The Top K (TopK) algorithm reduces the redundancy in the rule set, which was possible through operating on patients rather than the expressions of the rules. TopK still achieves high compression rate and this helps to identify rules with high risk and low redundancy.

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.

VI COMPARATIVE STUDY

The above TopK and BUS minimizing rule redundancy. While comparing with TopK and BUS, we found that the proposed HARS retained slightly more redundant than Top-K and BUS, which allowed it to have better patient coverage and better ability to reconstruct the original data base and helps for robust prediction of risk.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

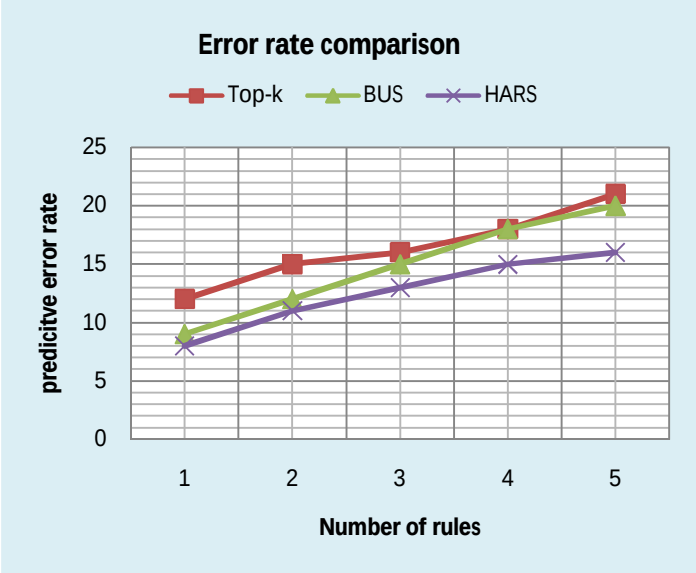

Figure 1.0 Sum squared prediction error of the summarization methods as a function of the number of rules on all patients.

The HARS has been compared with existing techniques such as Top-K and BUS and the above figure 1.0 shows, that the proposed HARS achieves minimum prediction error than others. The comparison has been made with number of rules and its prediction error rate. When the number of rule increases in HARS, then the complexity of prediction become risky. But the use of hybrid technique HARS overcomes the complexity with minimum predictive error. With the help of SAM the detection and prediction time also reduced in HARS.

VII. CONCLUSION

The study proposed a new association summarization and risk prediction scheme for diabetes and co-morbid conditions associated with diabetes. The system studied the main two problems in the literature, which are prediction accuracy and prediction error. The study overcomes the above two problem by applying the effective hybrid association rule summarization with split and merge algorithm. The HARS represents with the effective rule specification criteria which have been used by the iterative genetic algorithm. The system performs pre pruning and post pruning to eliminate irrelevant results. The system effectively identifies the risk of the diabetes disease and its sub types, the sub type which is referred as the heart disease, retinopathy, neuropathy etc., and the experimental results are evaluated using the Java. The experimental result shows that iterative HARS with genetic algorithm shows better prediction accuracy compared to traditional summarization techniques. From the experimental results, the prediction error calculated for diabetes risk assessment is almost reduced than the existing system.

REFERENCES

[1] Simon, György J., et al. "Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus." *Knowledge and Data Engineering, IEEE Transactions on* 27.1 (2015): 130-141.

[2] G. S Collins, S. Mallett, O. Omar, and L.-M. Yu, "Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting," *BMC Med.*, 9:103, Sept. 2011.

[3] A. J.T. Lee, Y.H. Liu, H.Mu Tsai, H.-Hui Lin, H-W. Wu, "Mining frequent patterns in image databases with 9DSPA representation",Journal of Systems and Software, Volume 82, Issue 4, April 2009, pp.603-618

[4] Centers for Disease Control and Prevention. "National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014." *Atlanta, ga: US Department of health and human services* (2014).

[5] Diabetes Prevention Program Research Group, "Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin," *N. Engl. J. Med.*, vol. 346, no. 6, pp. 393–403, Feb. 2002.

[6] J. Tuomilehto *et al.*, "Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impared glucose tolerance," *N. Engl. J. Med.*, vol. 344, no. 18, pp. 1343–1350, May 2001.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

[7] P. W. Wilson, R. B. D'Agostino, H. Parise, L. Sullivan, and J. B. Meigs, "Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus," *Circulation*, vol. 112, no. 20, pp. 3066–3072, Nov. 2005.

[8]. C. Aflori, M. Craus, "Grid implementation of the Apriori algorithm Advances in Engineering Software", Volume 38, Issue 5, May 2007, pp. 295-300.

. [9]. Srinivas, K., "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques", IEEE Transaction on Computer Science and Education (ICCSE), p(1344 - 1349), 2010

[10] Sa-ngasoongsong, Akkarapol, and Jongsawas Chongwatpol. "An Analysis of Diabetes Risk Factors Using Data Mining Approach." *Oklahoma state university, USA* (2012).

[11] Hnin Wint Khaing," Data Mining based Fragmentation and Prediction of Medical Data", IEEE 2011.

[12] Burdick, Doug, Manuel Calimlim, and Johannes Gehrke. "MAFIA: A maximal frequent itemset algorithm for transactional databases." *Data Engineering, 2001. Proceedings. 17th International Conference on*. IEEE, 2001.

[13]. V. Chauraisa and S. Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT),Vol. 2, No. 4,2013, pp 56-66.

[14] H. S. Kim, A. M. Shin, M. K. Kim, and N. Kim, "Comorbidity study on type 2 diabetes mellitus using data mining," *Korean J. Intern. Med.*, vol. 27, no. 2, pp. 197–202, Jun. 2012.

[15] Anbarasi, M., E. Anupriya, and N. C. S. N. Iyengar. "Enhanced prediction of heart disease with feature subset selection using genetic algorithm."*International Journal of Engineering Science and Technology* 2.10 (2010): 5370-5376.

[16] Anbarasi, M., E. Anupriya, and N. C. S. N. Iyengar. "Enhanced prediction of heart disease with feature subset selection using genetic algorithm."*International Journal of Engineering Science and Technology* 2.10 (2010): 5370-5376.