International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Prediction of By-Diseases in Diabetic Patients Using Associative Classification with Improved Classifier Accuracy for Decision Support System

Shahebaz Ahmed Khan, M A Jabbar

Abstract: ***Associative Classification in data mining technique formulates more and more simple methods and processes to find and predict the health problems like diabetes, tumors, heart problems, thyroid, cancer, malaria etc. The methods of classification combined with association rule mining gradually helps to predict large amount of data and also builds the accurate classification models for the future analysis. The data in medical area is sometimes vast and containss the information that relates to different diseases. It becomes difficult to estimate and analyze the disease problems that change from period to period based on severity. In this research paper, the use and need of associative classification for the medical data sets and the application of associative classification on the data in order to predict the by-diseases has been put front. The association rules in this context developed in training phase of data have predicted the chance of occurrence of other diseases in persons suffering with diabetes mellitus using Predictive Apriori. The associative classification algorithms like CAR is deployed in the context of accuracy measures.***

Keywords : Associative classification, Diabetic disease, classification model, by-disease and association mining, Predictive Apriori.

I. INTRODUCTION

Data mining is referred as the concept of extracting and finding the hidden patterns from various large size sources in order to examine and analyze the data for decision making. Data mining is the techniques that enables us to analyze large amount of data from various corners by applying intelligent methods and combining it by aggregating it into meaningful information by considering patterns, associations, or relationships among all this datasets and information. Data Mining methods and techniques are widely applied in medical decision making and have many applications where a large volumes of clinical data needs to be predicted and analyzed. Generally, in data mining there are two types of learning methods called supervised learning and the unsupervised learning.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Shahebaz Ahmed Khan*, Research Scholar at Shri Jagdish Prasad Jhabarmal Tibrewala University of Jhunjhunu.
M A Jabbar, Professor and Centre Head at the Computer Science and Engineering Department, Vardhaman College of Engineering, Hyderabad, Telangana, India.

Based on these two learning methods we can execute various data mining operations like preprocessing, classification, clustering, transformation, pattern recognition etc to reach knowledge and its discovery. In supervised learning, the class is already determined and known in the training data set at earlier stage. In unsupervised learning there is no concept of class attribute
and the search for the pattern is not limited to a single attribute. Association rule mining concept is important and most common example for the supervised learning. In theory of data mining, the combination of the associative rule mining and the classification methods is referred as Associative Classification. For prediction of the future health issues and the dangerous health symptoms in medical field, associative classification mining can be used in an efficient way when compared to other methods such as classification, association rule mining, decision tree, clustering etc. The research work done here is to predict the probability of the Co-diseases or By-Diseases in the diabetic patients. It is known fact that diabetic patients suffer with many other side effects relating to other severe diseases. In this work, the Apriori algorithm is constituted with the discretization features [9] so that the prediction can be more accurate and efficient in rule deriving process. In our work, different data sets have been selected to analyze the occurrence of the other diseases along with the diabetes in patients. All the experimental results have been compared for various aspects of accuracy and prediction possibilities of the other disease in diabetic patients. Different data sets with changing attributes have been selected for prediction of the by-disease in diabetic patients with improved efficiency by deploying the discretization idea in preprocessing feature of data mining. During Associative classification Predictive Apriori has been used by selecting CARs. A few researchers have tried to diagnose the diabetic patients. In this attempt our work constitutes its help to the health care industry in sooner prediction of the By-diseases in the patients of diabetes.

II. LITERATURE SURVEY AND PREVIOUS WORK

To find the interesting patterns in large data sets in the form of associations and correlations Association Rule Mining can be used as a strong idea of analyzing mechanism for medical data analysis.

Li et al[6] promoted and defined a hybrid form of method by combing the intelligence methods of the mining association rules and the classification algorithms which is now referred as Associative Classification. It has several advantages in the context of various properties over the traditional algorithms like decision tree and association among various data mining techniques available in today's world, associative classification serves as a better form of application and user need on health sector in order to predict the future problems. Akhil Jabbar et al[1] in their

research have proposed a method for heart disease classification system with the help of the idea of associative classification combined with genetic algorithms. The prediction method with associative classification generates higher order rules by achieving accuracy. At the same time, Rafel Rak et al[2] have proposed classification methods for medical documents using this approach with multi label classification combined with association rules from Medical Literature Analysis and Retrieval System Online, this MLARSO system also performs the task of accurate classification on health care data cubes or repositories which are purely unstructured and complex. Ya-Han Hu and Yen-Liang Chen [3] have implemented a new mining algorithm with Multiple Minimum Supports and support timing algorithm using Association Rules. not possible.

Diabetes which is not a disease but a syndrome is now considered a great threat to human health. More than 450 million people suffer with this syndrome all over the world.Many researches are being in progress for medical data diagnosis of diabetic syndrome in patients. As per the medical diagnosis type 1 or type 2 diabetes is possible in human body. It is classed as a metabolism disorder. This diabetes is considered as a metabolic problem due to which polyurea, polydipsia and polyphagia are found to be the common symptoms. In the year 1675, Thomas Wills added the word 'Milletus' to the term called 'diabeinein' due to which it is now called diabetes Mellitus.

Anyhow, the severity and need of extracting and predicting constrained association rules on medical data sets for cardiac problems was studied by Carlos. Most of associative classification algorithms in data mining follow the exhaustive search method which is found in Apriori algorithm. This method requires multiple passes over the data base to find and extract the rules. The target of discovering the knowledge is not already determined in the concept of association rule extraction, instead of this, one predetermined target is fixed for mining of association rules. The minimum subset of association rules are taken into count for the fusion of combining the association rules. This subset is called Class Association Rules (CAR). Higher accuracy levels can be targeted with this new form of approach [4].Yin and Jiawei[5] Han proposed a classification based on Predictive Association Rules (CPAR), that combines the positives of both associative classification and rule based classification. But it is to be noted that in terms of large training data CPAR adopts greedy method to generate rules.

III. ASSOCIATIVE CLASSIFICATION TECHNIQUE

To know the correlation between two or more data items in a given data set the method of association rule mining can be effectively used in data mining. It can be used to define the complete rules present in a training data set by defining certain confidence and support on it. which are later called association rules. To find the set of items in the data set that are occurring continuously or frequently the idea of association rule mining can be usedt. Association rule mining can be used identify frequent if-then associations in the data set, which are later called association rules. Classification in data mining is also an important technique which can be used to construct the classifiers [10]. Many fast and accurate classification algorithms like Naive Bayesian classifier, ID3 Algorithm, Nearest Neighbors Algorithm, C4.5 etc are used for this classifier accuracy purpose. Though the classification techniques in data mining define small set of rules, even for large data sets, there is a possibility of building accurate classifiers. The integration approach or the idea of association rule mining and classification technique is called Association Classification. The minimum subset of association rules are taken into account for this integration which are later known as Class Association Rules. Different algorithms [11] can be deployed for the purpose of associative classification; because of its easy understanding Class Association Rules (CARs) are used. Experimental results in this context conducted by researchers have shown that, whenever the classification of data is based on association rule mining it can construct strong predictive and accurate classifiers when compared to the traditional classification methods. In associative classification, the given data set is divided into two parts and from this70% of data is used for training and 30% of the data for testing the accuracy of the classifier. Classification with association rule mining is also effective than simple classification.

One of the basic features of association rule mining is it uses Apriori algorithm for large data sets [12], in this using the large itemsets of the previous pass candidate itemsets are generated. It can be applied to large data bases like super market data, medical data, education data etc. Apriori algorithm was developed and implemented by R. Agrawal and R. Srikant in the year 1994 to find frequent itemsets in a dataset. The Apriori algorithm reduces the number of candidates whose support count is greater than the minimum support count. An association rule is considered as a frequent and strong if its support is greater than minimum support and if its confidence is greater than minimum confidence respectively. It can be as follows

Support (X-> Y) = $\sigma (X \cup Y) / N$

Confidence (X-> Y) = $\sigma (X \cup Y) / \sigma (X)$

Associative classification works using three steps, In the first step, generation of a set of association rules from training set by providing certain support and confidence threshold as candidate rules is done. In the second step, to introduce over fitting the discovered rules are pruned. The at last classification is performed to predict the test data so that the classifier's accuracy can be measured. It needs rule ranking procedure and classifier building for accuracy. Associative Classification can be used for prediction of medical constraints because it suits well on large set of data with easy building of rules with good accuracy. Among various data mining techniques available, associative classification technique has a good application in medical domain. Most of the attributes in medical data sets are associated with quantitative domains such as blood pressure,

lipid profile, sugar levels, body temperature and body mass index etc. Associative Classification uses many algorithmic approaches to mine the frequent patterns. Many rules in associative classification are classification based. Using Associative Classification high dimensional data sets can be handled easily and efficiently due to the quick building of accurate classifiers.

A. Associative Classification Methods and CAR

The associative classifier can be built by various methods like database coverage method and full and partial match rule evaluation method.Various methods of Associative Classification [8] are available in data mining context. All these methods have comparative advantages and cons in their respective usages. Some of the Associative classification methods are Classification Based Association Rules(CBA),Classification Based on Multiple Association Rules(CMAR), Classification Based on Association Rule Generated in a Bidirectional Approach(CARGBA), Hierarchical Multi-label Associative Classification (HMAC), CPAR, ACCF, GARC, CARPT, ACN etc.

Class Association Rule is a sub set of Association Rule [7]. This CAR contains non class and class attributes. In CAR class is specified as the consequence of association rules.

Let us assume that A is a set of instances in which each instance X is represented by < s1,s2,s3,….sn,C> and in this s1,s2,….sn are taken as attributes of non-class attributes. C is a class label. When an association rule is framed on this called X=C is considered as Class Association Rule. CAR has higher efficiency and accuracy than other methods.

IV. RESULTS AND DISCUSSION

Our experiments were carried out on Weka which is an open source tool. The work was deployed using Weka Tool because implementation of Apriori algorithm for class association rules will be easier in Weka. Weka contains many tools for classification, visualization, workflow, Clustering, Preprocessing, experiments etc. Various attributes selected were age, gender, different cholesterol levels in diabetic body, triglycerides, ldl and hdl cholesterols etc. in data set one. In data set two age, gender, diabetic, hypertension, co-diseases like cad, poor vision, sob, thyroid, other diseases, nil etc. were taken into account. Our approach obtained a good accuracy factor when compared to other previous and conventional methods of prediction.

The experimental results were found with a possibility of brain stroke and heart disease occurrence in diabetic patients. The medical data record sheets of the persons suffering from diabetic syndrome with different attributes in a data sets were diagnosed so as to find the other diseases or health issues that occur along with diabetic mellitus. In this, we have generated the CARs because we need to mine the class association rules than the normal association rules. We have generated 100 best rules of CAR using Predictive Apriori in data set 1 and 40 best rules in data set 2. The minimum confidence higher than 0.9 was set to retrieved. For each experiment the accuracy was recorded and in comparison with the original CARs, we have taken some 60 generated rules randomly from 100 rules originally achieved. Then, we have predicted the results by eliminating the class of those 60 rules taken aside and by cross verifying the rules by extracting another 40 from 60 rules selected in with the comparison of the class prediction so as to confirm the accuracy of the test and prediction of health problems. Different data sets fed to Predictive Apriori have made the results by predicting the Co-Diseases in diabetic patients. It was found in the results that, most of the diabetic patients are subjected to get brain strokes and heart attacks in future. In cross prediction of accuracy, for hdl cholesterol it was 100%,for vldl cholesterol it was 100%,for ttl cholesterol it was 83 %,for ldl cholesterol it was 100%,for triglycerides it was 73 % over all it was then 91.2 in manual cross prediction.

Table 1 Rules Generated and Instances

Total instances in Data sets	600 (Separately tested by dividing the attributes)
Total Rules Generated using CARs	100
Number of Rules derived from generated rules for testing accuracy and prediction	60
Rules selected randomly from	40

Table 2 Prediction and Accuracy

Parameters	Before cross verification	After cross verification
Prediction of Disease	Heart problems and brain strokes	Heart problems and brain strokes
Accuracy	97.8	91.2

The precondition defined for By-disease prediction was if the patients were found with high or very high rate of hypertension and diabetic levels then they were considered as becoming victims of brain stroke possibilities. At the same time the diabetic patients with high cholesterol levels of hdl, vdl, ttl and bad cholesterol levels of ldl and abnormal triglycerides were taken into the account of heart diseases. Also the previously had diabetic patients with cad disorders were also considered to add the strength to the predictive results. In a separate data set with already existing By-diseases in diabetic patients, it was tested the possibility and accuracy of the effected diseases in diabetic patients. The results were irrespective of gender and age. The result part of the experiments is as follows

Table 3 Experimental Result part

hdlcholesterol=normal ldlcholesterol=good ttlchol=normal 48 ==> disease=no 48 acc:(0.99494)
gender=female diabetic=yes cholesterol=bl hdlcholesterol=abnormal vldlcholesterol=bad 8 ==> disease=yes 8 acc:(0.99223)
gender=female diabetic=yes ldlcholesterol=bad triglycerides=bl 7 ==> disease=yes 7 acc:(0.99147)
hdlcholesterol=normal ldlcholesterol=good vldlcholesterol=bad triglycerides=bl 7 ==> disease=no 7 acc:(0.99147)
age='(50.6-58.5]' hdlcholesterol=abnormal ldlcholesterol=bad 5 ==> disease=yes 5 acc:(0.98877)
etc…
diabetic=yes hypertension=yes codisease=cad 12 ==> gender=male 12 conf:(1)
gender=male

hypertension=yes codisease=cad 12 ==> diabetic=yes 12 conf:(1)
gender=male hypertension=no codisease=nil 11 ==> diabetic=yes 11 conf:(1)
gender=male hypertension=yes codisease=nil 31 ==> diabetic=yes 31 conf:(1) etc….
gender= female hypertension=high diabetes= high 41 ==> disease=yes 41 conf:(1)
hypertension=high diabetes= very high 39 ==> disease=yes 39 conf:(1)

V. CONCLUSION AND FUTURE WORK

Associative Classification can be used for effective and improved accurate prediction of the data. It aims to provide more and better accuracy than the conventional techniques in data mining. In our results it was predicted that there is a greater possibility of the occurrence of heart diseases and brain strokes in the diabetic patients. The work carried out and the results predicted are based on the attributes selected and used. When Discretization is applied to the data sets, the relationship between different items in a dataset is found with good association rules. The numerical data is discretized and fed to an Apriori algorithm for rule induction. From our results it can be concluded that a pre care can be taken to avoid the possibility of the heart diseases and brain strokes as well as sudden trauma in diabetic patients. The application of Predicted Apriori can effectively provide the accuracy to the association rules generated. Even after the cross prediction the results were found same with the earlier accuracy and predicted diseases. In future, the results can be derived by considering the constraints of age factor and gender. This can even predict the chance of other health issues in the patients of diabetic syndrome both gender wise and age.

REFERENCES

1. MA.Jabbar, Priti Chandra, B.L.Deekshatulu..:Cluster based association rule mining for heart attack prediction,JATIT,vol 32,no 2,(Oct 2011)
2. Rafel Rak,Carlos Ordonez.: Comparing association rules and decision trees for heart disease prediction, ACM, HICOM (2006)
3. Ya-Han Hu, Yen-Liang Chen, "Mining Association Rules with Multiple Minimum Supports: A New Mining Algorithm and a Support Timing Mechanism" © 2004 Elsevier B.V.
4. D. Sasirekha1* and A. Punitha2, A Comprehensive Analysis on Associative Classification in Medical Datasets, Indian Journal of Science and Technology, Vol 8(33), DOI: 10.17485/ijst/2015/v8i33/80081, December 2015.
5. Yin and jiawei ,"CPAR: Classification based on predictive association rule. In Proceeding of the SDM pp 369-376(2003)
6. Li X, Qin D, Yu C. ACCF: Associative Classification based on Closed Frequent itemsets. Fifth International Conference on Fuzzy Systems and Knowledge Discovery FSKD'08; Shandong. 2008. p. 380–4.
7. Yu G, Li K, Shao S. Mining high utility itemsets in largehigh dimensional data. Proceedings of the 1st International Conference on Forensic Applications and Techniques in Telecommunications, Information, and Multimedia and Workshop. ICST (Institute for Computer Sciences, Social- Informatics and Telecommunications Engineering); 2008. p. 47.
8. MA.Jabbar, B.L.Deekshatulu and Priti Chandra.: An evolutionary algorithm for heart disease prediction, ICIP, CCIS 292 PP 378- 389, Springer-Verlag (2012)
9. Lee, C.-H.: A Hellinger-based Discretization Method for Numeric Attributes in Classification Learning. Knowledge-Based Systems 20(4), 419–425 (2007).
10. Kundu G. Islam MM. Munir S. Bari MF. ACN: An associative classifier with negative rules. 11th IEEE International Conference on Computational Science and Engineering, CSE'08; Sao Paulo. 2008. p. 369–75.
11. Sangsuriyun S, Marukatat S, Waiyamai K. Hierarchical Multi-label Associative Classification (HMAC) using negative rules. 9th IEEE International Conference on Cognitive Informatics (ICCI); Beijing. 2010. p. 919–24
12. Zhuang Chen, Shibang CAI, Qiulin Song and Chonglai Zhu, "An Improved Apriori Algorithm Based on Pruning Optimization and Transaction Reduction", IEEE 2011.

AUTHOR'S PROFILE

Shahebaz ahmed Khan is a Research Scholar at Shri Jagdish Prasad Jhabarmal Tibrewala University of Jhunjhunu. He has completed his B.Tech and M.Tech in CSE from Bharat Institute of Engineering and Technology, Hyderabad. He has published more than 13 research papers in various journals and conferences of both national and international levels. Shahebaz is doing his PhD in Computer Science and Engineering with Data Mining Specialization. His areas of interest are Data Mining, Machine Learning, Computer Programming, Information Retrieval Systems and Operating Systems. He has also written and published four books on various issues in the society which are of general nature. He is a member of ISTE, IAENG, CSTA, UACEE and SDIWC. The research work carried out by Shahebaz includes the prediction of diseases using data mining techniques with special reference to diabetes and thyroid.

Dr. M.A.JABBAR is a Professor and Centre Head at the Computer Science and Engineering Department, Vardhaman College of Engineering, Hyderabad, Telangana, India. He has been teaching for more than 19 years. He obtained Doctor of Philosophy (Ph.D.) from JNTUH. He published more than 50 papers in various journals and conferences. He is Reviewer for Scopus and SCI journals like Springer, Elsevier, and IEEE Transactions on Systems Man and Cybernetics, Wiley. He served as a technical committee member for more than 40 international conferences. He published 5 patents (Indian) in machine learning and allied areas**.** He is a senior member of IEEE, also a member of ACM, Life member of CSI, ISCA and currently he is Vice-chair, Computer Society Chapter of IEEE Hyderabad Section.

.