International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

Improved Classification Techniques to Predict the Co-disease in Diabetic Mellitus Patients using Discretization and Apriori Algorithm

Shahebaz Ahmed Khan,.M A Jabbar

Abstract*: *The demand for data mining is now unavoidable in the medical industry due to its various applications and uses in predicting the diseases at the early stage. The methods available in the data mining theories are easy to extract the useful patterns and speed to recognize the task based outcomes. In data mining the classification models are really useful in building the classes for the medical data sets for future analysis in an accurate way. Besides these facilities, Association rules in data mining are a promising technique to find hidden patterns in a medical data set and have been successfully applied with market basket data, census data and financial data. Apriori algorithm, is considered to be a classic algorithm, is useful in mining frequent item sets on a database containing a large number of transactions and it also predicts the relevant association rules. Association rules capture the relationship of items that are present in data sets and when the data set contains continuous attributes, the existing algorithms may not work due to this, discretization can be applied to the association rules in order to find the relation between various patterns in data set. In this paper of our research, using Discretized Apriori the research work is done to predict the by-disease in people who are found with diabetic syndrome; also the rules extracted are analyzed. In the discretization step, numerical data is discretized and fed to the Apriori algorithm for better association rules to predict the diseases.

Keywords*: *Hidden patterns, association rules, Apriori, co-disease, discretization, classifiers, Co-disease.

I. INTRODUCTION

Data mining techniques are widely used for medical decision making and have many applications where a large volumes of clinical data needs to be predicted and analyzed[4]. Data mining is the process of analyzing data from various perspectives by applying intelligent methods and combining it by summarizing it into useful information by considering the patterns, associations, or relationships among all this data and information. In medical diagnosis, to extract the meaningful patterns or rules from historical data mining can be efficiently used to determine possible condition of old and new patients, such as whether found with any health issues or not. Recently, due to the growing capacity and need of data mining, a supervised learning method called association rules and classification have been significantly applied to medical diagnosis and prognosis[5]. The concept or idea here is to find the frequent patterns and classify the data for future prediction. Prior to generating association rules, data discretization is one of the important steps of preprocessing step which can be used to identify a finite number of categories of data from continuous feature values of data. For example, the risk status of patients with diabetic disease can be categorized into different levels: high, very high, middle, normal and low by applying thresholds to numerical blood pressure and sugar levels. There are also other advantages for using discrete features which means categorical features instead of continuous features when solving large-scale complex problems on data sets. There are many supervised or unsupervised discretization approaches used and proposed in the recent time. When it comes to this point, equal width interval and equal frequency interval binning methods are the simplest methods available in unsupervised discretization. The equal width interval method in this context divides the feature range into a number of equal-sized intervals which are also called bins to make the task accurate and fast. Also, the performance and accuracy of machine learning algorithms can be improved by using discrete datasets. The research work done here is to predict and find the chance or the probability of other forms of health problems like poor vision, heart attacks, brain damage etc in diabetic patients. In data mining, Apriori algorithm is constituted with the discretization features so that the predication can be more accurate and efficient. Different data sets have been selected in this context to analyze the occurrence of the other diseases along with the diabetes in patients. All the experimental results in our work have been compared for various aspects of prediction possibilities of the disease in diabetic patients. With the application of discrete Apriori satisfied and estimated results have been derived. In our work, different data sets with changing attributes have been selected for discretized Apriori application on those data sets.

II. LITERATURE SURVEY

In data mining and knowledge discovery process, Association Rule Mining plays an important role in finding interesting measures, relational associations among various data items and it also derives other correlations in a large group of datasets which makes it a crucial part of market basket analysis. The implementation of associative classification is done by dividing the data sets into two blocks. One part of the data is used for the purpose of training which constitutes nearly seventy percentages from the whole and the rest thirty percentage of the data block is applied for testing of the classifier accuracy.

Manuscript published on 30 September 2019.
*Correspondence Author(s)
Shahebaz Ahmed Khan,. Department of CSE, JJTU, Jhunjhunu, Rajasthan, India
M A Jabbar,. Professor at Department of CSE, Vardhaman College of Engineering, Hyderabad

Associative classification for data mining task is considered as mostly used techniques in medical and health industry due to its ease of applicable performance [1]. Rafel Rak et al has took the data mining to new scope using associative classification, now a days a new form of associative classification with multi-label is applied on medicinal records for performing classification on medical repository which is unstructured and complex[8].It is used for retrieving the rules on medical data records of patients. It considers multi label features and dependent on associative classification. Akhil Jabbar et al have proposed a method for using the idea and theme of classification combined with association rule mining to predict the heart diseases and problems. Here machine learning was also applied using genetic algorithmic techniques.[7].

Many people consider diabetes as a disease but it is a syndrome in where blood glucose or levels of sugar in the human body are too high. As per the medical diagnosis any diabetic human body can become a victim of type 1 diabetes or diabetes of type 2. Type 1 diabetes the pancreas produces no insulin in the human body The problems with diabetic disorder and syndrome are very severe which can cause psychological issues like dementia, , dental diseases, neurological problems, vision complications, cardiac attacks, thrombocytopenia, and brain strokes etc. When there is more than enough levels of glucose that too much quantity of glucose levels in human body then it can lead towards unexpected and serious health problems. Due to this, there is a finite probability of damaging the vision, excretion system, nervous system and also can create cardiac arrests by causing the severe heart pain. More over, diabetes can be considered as a metabolic disorder and due to this disorder the symptoms of polyuria, polydipsia etc is found to be the common problems. More or less, 10 percentage of the total population is found with type 1 diabetes and 90 % of the people are victims of type 2 diabetes. Type 1 diabetes is a chronic disorder. It takes the roots of health to become vulnerable with other complicated and dangerous diseases.. In this regards, this makes a need to find the By-diseases at the very first stage that occur with diabetes [7].

To safeguard the health and to avoid the health problems the need and necessity for prediction of diabetic disease and its co diseases arises.. Ii is estimated that there are more than 400 million people who suffer with diabetic syndrome all over the world. The present is intended for the co-disease prediction using discretization Apriori. The four types of data sets have been taken into consideration for analysis of the medical data sets related to diabetic patients with variant attribute selection. Anyhow, the problem of identifying constrained association rules on medical data sets for heart disease prediction was studied by Carlos. Most of the associative classification algorithms follow the exhaustive search method found in Apriori algorithm which requires multiple passes over the data base to discover the rules.

III. ASSOCIATION RULES AND APRIORI

Machine learning methods are used for mining the rules that occur frequently with close relationships in a data set. It can be a way to analyze the patterns of data and frequent occurrence in it. Association rule mining can be used identify frequent if-then associations in the data set, which are later called association rules. The association of items and mining of rules is done by selecting the well-formed and well-built rules. The items in a dataset which combine two or more than two items are taken for the calculation of rules that are generated.[3].Meaningful patterns are derived by analyzing the frequently occurring data items from all deeper sense of knowledge discovery techniques. To mine the association rules support and confidence parameters which apply if and then rule basis searching are used to find and match the interesting patterns and relationships from the data[2]. Support in association rule mining indicates the frequent possibility of the items of a given data set with association match at the same or similar frequency and confidence is used to indicate the number of times the item with if then case rule is found[6]. Here, in this framework of association rule mining another important measure is considered which is known as lift, lift is a metric which is used for comparing the confidence levels with the given confidence. Association rule mining uses many algorithms like AIS, Apriori, STEM etc[9]. In Apriori algorithm, using the large item sets of the previous pass candidate item sets are generated. It can be applied to large data bases like super market data, medical data, education data etc. Later, to generate and derive the larger size item sets, large item set is selected and combined itself with the past result round. In the case of no generation of the expected large item set, deletion of it preferable. Rest of the item sets are treated as the candidate item sets. Apriori algorithm was developed and implemented by R. Agrawal and R. Srikant in the year 1994 to find frequent item sets in a dataset. The name of the algorithm is Apriori because this algorithm uses prior knowledge of frequent item set characteristics. In this approach, m-frequent occurred item sets are taken to find m+1 item sets. In this approach, almost all the sub item sets are normal item sets with frequently occurring class. If the minimum support count is less when compared to the normal count of support then the Apriori algorithm will reduce count of the candidates [12].

A. Discretized Apriori

In machine learning and data mining techniques the discretized data set has many applications. Association rule mining uses the discretized datasets. To find the relationship between different items in a dataset association rule suits good. The numerical data is discretized and fed to an Apriori algorithm for rule induction [10]. Then a new observation is classified using this approach. There will be both supervised and unsupervised discretization. In unsupervised, discretization we can find equal width interval and equal frequency interval binning methods which are of simplest forms [11]. The equal width interval method divides the feature range into a number of equal-sized bins and the equal frequency interval method divides the feature range into k bins. Here, one should notice that both contain the same number of observation points. There is disadvantage in the unsupervised methods that they do not take into account the class information at the time of binning process. Even re-sampling based bagging technique have been introduced in the recent scenarios. The quality of the discretized data set is improved by using bagging to produce the sets of candidate points so that it can reduce the variance.

IV. PROPOSED WORK AND DERIVED RESULTS

The health record sheets of persons suffering with diabetes mellitus were selected and analyzed for data prediction. This was done with different attributes which were diagnosed to know and predict the other form of side effects and diseases in the persons with diabetes. More than 8 attributes like age, sex, different cholesterol levels in diabetic body, triglycerides, different physical imbalances and blood pressure levels were taken into count from the data set one. In data set two age, gender, diabetic, hypertension, co-disease were taken as attributes. In third data set, diabetic, thyroid, platelet count, hemoglobin, age, genders were the attributes. In the fourth the average of all these were taken into account. All the datasets were trained on Weka tool. The implementation of the work was carried out in favorable conditions of Weka tool for rule derivation and prediction. The proposed association rules can accurately and efficiently predict the data in the various context of disease analysis by improving other conventional existing frameworks.

For Data set one the total instances were 114 and the attributes were 10.

Instances	114
Attributes	10

Table 1 Instance and Attributes-ds1

The Minimum support is 0.5 (51 instances),Minimum metric <confidence>: 0.9. For which the Number of cycles performed were 10. The derived rules found in this are as follows

disease=no 75 ==> ttlchol='(-inf-51]' 75 conf:(1)
hdlcholesterol=normal 70 ==> ttlchol='(-inf-51]' 70 conf:(1)
ldlcholesterol=good 68 ==> ttlchol='(-inf-51]' 68 conf:(1)
gender=male 60 ==> ttlchol='(-inf-51]' 60 conf:(1)
hdlcholesterol=normal disease=no 60 ==> ttlchol='(-inf-51]' 60 conf:(1)
gender=male hdlcholesterol=normal 58 ==> ttlchol='(-inf-51]' 58 conf:(1)
cholesterol=normal ldlcholesterol=good 57 ==> ttlchol='(-inf-51]' 57 conf:(1)
cholesterol=normal disease=no 56 ==> ldlcholesterol=good 56 conf:(1)
cholesterol=normal disease=no 56 ==> ttlchol='(-inf-51]' 56 conf:(1) For Data Set 2

Instances	114
Attributes	5

Table 2 Instance and attributes-ds2

Minimum support: 0.1 (11 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18
Some rules found were
gender=male 67 ==> diabetic=yes 67 conf:(1)
hypertension=yes 54 ==> diabetic=yes 54 conf:(1)
gender=male hypertension=yes 41 ==> diabetic=yes 41 conf:(1)
codisease=nil 38 ==> diabetic=yes 38 conf:(1)
hypertension=no 36 ==> diabetic=yes 36 conf:(1)

For data set three the following rules were derived
Minimum support: 0.3 (8 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 14
diabetes=no 9 ==> thyroid=yes 9 conf:(1)
diabetes=no 9 ==> platelet count= normal 9 conf:(1)
diabetes=no platelet count= normal 9 ==> thyroid=yes 9 conf:(1)
diabetes=no thyroid=yes 9 ==> platelet count= normal 9 conf:(1)
diabetes=no 9 ==> thyroid=yes platelet count= normal 9 conf:(1)
heamoglobin= less disease=yes 8 ==> gender=female 8 conf:(1)

For the data set four we have taken 317 instances by selecting 5 major attributes in the patients to predict the occurrence of the Co-disease. It was tested to predict whether if the diabetes levels and hypertension levels were high and more than the normal or very high level, then if there is any possibility of brain stroke or any sudden trauma in patients. The rules found were as follows

diabetes= high disease=yes 62 ==> hypertension=high 62 conf:(1)
diabetes= very high disease=yes 38 ==> hypertension=high 38 conf:(1)
gender= female diabetes= high disease=yes 38 ==> hypertension=high 38 conf:(1)
hypertension=high diabetes= very high 39 ==> disease=yes 38 conf:(0.97)
. gender= female hypertension=low 34 ==> disease=no 33 conf:(0.97)
hypertension=high diabetes= high 65 ==> disease=yes 62 conf:(0.95)
gender= male hypertension=high 42 ==> disease=yes 39 conf:(0.93)
diabetes= very high 41 ==> hypertension=high disease=yes 38 conf:(0.93)
gender= female hypertension=high diabetes= high 41 ==> disease=yes 38 conf:(0.93)
gender= female diabetes= high 41 ==> hypertension=high disease=yes 38 conf:(0.93)

The rules are sorted by considering the three attributes generally like support, confidence and length. Based on the confidence the rules are sorted and then if the two rules are of the same confidence levels, then they are sorted according to the supports and the if the same is in the case with the support consideration then they are sorted by the antecedent length.

After all the rules are searched in the data set and ranked, now the classifiers are to be built by evaluating the most significant rules for the data prediction.Different data sets fed to Discretized Apriori have made the results by predicting the Co-Diseases in diabetic patients [13]. It was found that most of the diabetic patients are subjected to get brain strokes and heart attacks in future.

V. CONCLUSION

Our results can be effectively taken for designing the decision support systems in the context of Co-disease prediction. When compared to the Simple Apriori if the filter of Discretization is applied to it to fed the data sets then 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 work and prediction, it is said that brain strokes and heart problems stand first as a form of other side effect or Co-disease in diabetic people. per the predicted results using discretized Apriori algorithm. In this aspect, .the occurrence of the heart problems and brain strokes as well as sudden trauma in such people of DM can be stopped or prevented at early stage and prediction of the diseases present can be done effectively using the various data mining techniques. The construction of the discretization with Apriori can effectively judge the capability of the association rules generated and it better suits the data sets where large sets of numerical instances are found. The rules derived can be used for future analysis of the medical records and can be helpful in finding the knowledge of diseases in people and these rules discussed are with better form of classifiers in the taken context.

REFERENCES

1. Thabtah F., Mahmood Q., McCluskey L., Abdel-jaber H (2010). A new Classification based on Association Algorithm. Journal of Information and Knowledge Management, Vol 9, No. 1, pp. 55-64. World Scientific.
2. Han, J., Pei, J., and Yin, Y. (2000) Mining frequent patterns without candidate generation. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 1-12, 2000.
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. Elveren E, Yumusak N. Tuberculosis disease diagnosis us¬ing artificial neural network trained with genetic algorithm. Journal of Medical Systems. 2011; 35(3):329–32.
5. Sellappan Palaniappan et al.: Intelligent heart disease prediction on system using data mining techniques.IJCSNS Vol 8 no 8(Aug2008)
6. Baralis, E., Chiusano, S., and Graza, P. (2004) on support thresholds inassociative classification. Proceedings of the 2004 ACM Symposium onApplied Computing, (pp. 553-558). Nicosia, Cyprus, 2004.
7. MA.Jabbar, Priti Chandra, B.L.Deekshatulu..:Cluster based association rule mining for heart attack prediction,JATIT,vol 32,no 2,(Oct 2011)
8. Carlos Ordonez.: Comparing association rules and decision trees for heart disease prediction, ACM, HICOM (2006)
9. Agrawal, R., Imieilinski, T., Swami, A.1993. MiningAssociation Rules between sets of items in large databases. SIGMOD'93, pp. 207-216.
10. Lee, C.-H.: A Hellinger-based Discretization Method for Numeric Attributes in Classification Learning. Knowledge-Based Systems 20(4), 419–425 (2007).
11. Liu, H., Hussain, F., Tan, C., Dash, M.: Discretization: An Enabling Technique. Data Mining and Knowledge Discovery 6(4), 393–423 (2002).
12. Zhuang Chen, Shibang CAI, Qiulin Song and Chonglai Zhu, "An Improved Apriori Algorithm Based on Pruning Optimization and Transaction Reduction", IEEE 2011.
13. Huan Wu, Zhigang Lu, Lin Pan, Rongsheng Xu, Wenbao Jiang," AnImproved Apriori-based Algorithm for Association Rules Mining",Sixth International Conference on Fuzzy Systems and Knowledge Discovery, IEEE Society community, 2009.

AUTHORS 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.