Data Mining Apriori Algorithm for Heart Disease Prediction

Mirpouya Mirmozaffari[1], Alireza Alinezhad[2], and Azadeh Gilanpour[3]

***Abstract*—**Heart disease is a major cause of morbidity and mortality in the modern society. Almost 60% of the world population fall victim to the heart disease. Although significant progress has been made in the diagnosis and treatment of coronary heart disease, further investigation is still needed. Data mining, as a solution to extract hidden pattern from the clinical dataset are applied to a database in this research. The database consists of 209 instances and 8 attributes. The system was implemented in WEKA and MATLAB software and prediction accuracy within Apriori algorithm in 3 steps, are compared. MATLAB is introduced as better performance software.

***Keywords*—** Data mining, Apriori, MATLAB, WEKA.

I. INTRODUCTION

CARDIOVASCULAR diseases, such as coronary heart disease and arrhythmia, are among diseases which endanger human life [1]. Medical practitioners conduct different surveys on heart diseases and gather information of heart patients, their symptoms and disease progression. Increasingly are reported about patients with common diseases who have typical symptoms.

Data Mining is the process of extracting hidden knowledge from large volumes of raw data. [2]. It has been defined as "the nontrivial extraction of previously unknown, implicit and potentially useful information from data. Data mining is the science of extracting useful information from large databases.

To find the unknown trends in heart disease, Apriori algorithm in association rule are applied to a unique dataset and their accuracy are compared in two different software. A dataset of 209 instances and 8 attributes (7 inputs and 1 output) are used to test and justify the algorithm. To further enhance accuracy and achieve more reliable variables, the dataset is purified by Discretization unsupervised filter. Finally, better performance software for Apriori algorithm with better accuracy is introduced.

Mirpouya Mirmozaffari[1], Msc. student, Faculty of Industrial and Mechanical Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran

Alireza Alinezhad[2], Associate Professor, Faculty of Industrial and Mechanical Engineering, Qazvin Branch, Islamic Azad University, Qazvin,Iran .

Azadeh Gilanpour[3], Islamic Azad University (IAU).

II. BACKGROUND AND LITERATURE REVIEW

Growing number of heart patients worldwide have motivated researchers to do comprehensive research to reveal hidden patterns in clinical datasets. This section provides an overview of previous computational studies on pattern recognition in heart disease. Not only are different techniques addressed, but also various heart disease datasets are covered to have a fair comparison. Finally, the gap in existing literature, which was the main motivation of this study is also provided. Some of the key studies are as follows:

- Das et al. introduced a neural network classifier for diagnosing of the valvular heart disease. The ensemble-based methods create new models by combining the posterior probabilities or the predicted values from multiple predecessor models. An effective model has been created and experimentally tested. A classification accuracy of 97.4% from the experiment on a dataset containing 215 samples is achieved [3].
- Pandey et al. proposed the performance of clustering algorithm using heart disease dataset. They evaluated the performance and prediction accuracy of some clustering algorithms. The performance of clusters will be calculated using the mode of classes to clusters evaluation. Finally, they proposed Make Density Based Cluster with the prediction accuracy of 85.8086%, as the most versatile algorithm for heart disease diagnosis [4].
- Karaolis et al. developed a data mining system using association analysis based on the Apriori algorithm for the assessment of heart-related risk factors with WEKA tools. A total of 369 cases were collected from the Paphos CHD Survey, most of them with more than one event. Selected rules were evaluated according to the importance of each rule. Each extracted rule was further evaluated by inspection of the number of cases within the database [5].

Therefore, pattern recognition in heart disease can be addressed through different computational techniques. In regard to association rule algorithms, other respected works, focused on diverse aspects of heart disease on different datasets can be mentioned: Danapana et al., 2011 [6]; Ordonez 2006 [7]; Han et al., 2011 [8]; Deekshatulu 2012 [9]; Deepika 2011 [10]; Lakshmi et al., 2013 [11]. Also, different computational techniques for other health care issues have been reported in the literature [12-13].

It is observed various associators are frequently utilized in different studies to predict heart disease. Therefore, a comprehensive comparison of association rules algorithms practically provides an insight into associator performances.

This comparison is of great importance to medical practitioners who desire to predict heart failure at a proper step of its progression. Furthermore, except for Ref. [14], which has evaluated 4 classification techniques, there is not any other study on the current dataset. Finally, a unique coding in MATLAB software is applied in Apriori algorithm which eventually results in better performance in compare with WEKA software, covered in this study.

III. Dataset Description

The standard dataset, compiled in this study contains 209 records, which is collected from a hospital in Iran, under the supervision of National Health Ministry. Data is gathered from a single resource, so it precludes any integration operations. Eight attributes are utilized, from them, 7 are considered as inputs which predict the future state of the attribute "Diagnosis". All the attributes, along with their values and data types are discussed in Table I.

TABLE I
THE ARRANGEMENT OF CHANNELS

Attributes	Descriptions	Encoding\Values	Feature
Age	Age in years	28-66	Numeric
Chest Pain Type	It signals heart attack and has four different conditions: Asymptotic, Atypical Angina, Typical Angina, and without Angina.	Asymptotic = 1 Atypical Angina = 2 Typical Angina = 3 Non-Angina = 4	Nominal
Rest Blood Pressure	Patient's resting blood pressure in mm Hg at the time of admission to the hospital	94-200	Numeric
Blood Sugar	Below 120 mm Hg- Normal Above 120 mm Hg- High	High = 1 Normal = 0	Nominal Binary
Rest Electrocardiographic	Normal, Left Ventricular Hypertrophy (LVH) ST_T wave abnormality	Normal=1 Left Vent Hyper = 2 ST_T wave abnormality = 3	Nominal
Maximum Heart Rate	maximum heart rate attained in sport test	82-188	Numeric
Exercise Angina	It includes two conditions of positive and negative	Positive = 1 Negative = 0	Nominal Binary
Diagnosis	It includes two conditions of positive and negative	Positive = 1 Negative = 0	Nominal Binary

IV. Research Methodology

The objective of this study is to effectively predict possible heart attacks, from the patient dataset. Using a prediction methodology, a model was developed to determine the characteristics of heart disease in terms of some attributes.

Data mining in this research is utilized to build models for prediction of the class based on selected attributes. Waikato Environment for knowledge Analysis (WEKA) has been used for prediction due to its proficiency in discovering, analysis and predicting of patterns [15]. In addition, the system was implemented using MATLAB R2013a. MATLAB is a high language and interactive environment for numerical computation, visualization, and programming. The language tool, and built-in math functions enable us to explore multiple approaches and reach a solution faster than with spreadsheets of traditional programming languages, such as C/C++ of JAVA. Generally, the whole process can be split into two steps as follows:

A. filtering preprocess

The data in the real world is highly susceptible to noise, missing, and inconsistency. Therefore, preprocessing of data is very important. We apply a filter on datasets and purify them from dirty and redundant data present in the dataset. In association rules, Discretization should be applied in WEKA 2016 (version 3.9.0) and MATLAB R2013a to change numeric data into nominal data. This process is implemented.

B. Evaluation in Association Rules

Figure 1 elaborates the proposed model and different steps. We apply proposed model in WEKA and MATLAB software. Finally, to choose the better software, the accuracy of three strong rules in two different software are compared.

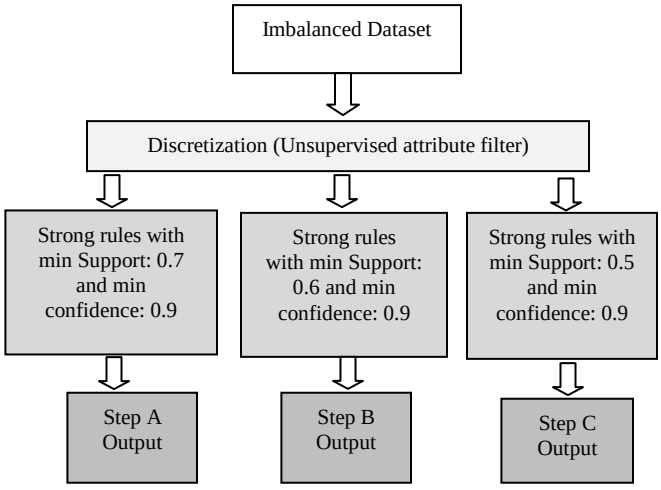

Fig 1: Implementation of Apriori Algorithm for accuracy analysis

V. Result And Discussion

The higher the support and confidence of a rule, the more it represents a regular pattern in the dataset. If these measures are relatively low, then any inconsistency would be less strong than it would be for rules with high confidence and high support. In step A, B and C, the accuracy of three strong rules in two different software, are compared.

A. Step A (strong rules with "min support": 0.7, and "min confidence": 0.9)

It should be noted, in this step, only one rule with "min support" (0.7) is evaluated. Because there is only one support higher than 0.7. In fact, it has highest support (0.7703) among all rules. This process is implemented in two software as follows:

- The First WEKA Apriori rule in Step A:

1. ECG at rest = '(-inf-1.666667]' 174 ==> Blood sugar ='(-inf-0.5]' 161 (Conf: 0.93, lift : 1)

- The First MATLAB Apriori rule in step A:

1. ECG at rest = '(-inf-1.666667]' 174 ==> Blood sugar = '(-inf-0.5]' 161 (Conf: 0.9253, lift: 1.0020, Sup: 0.7703)

B. Step B (strong rules with "min support": 0.7, and "min confidence": 0.9)

In step B, only two rules with "min support" (0.6) are evaluated. Because there is only one support between 0.6 and 0.7. The first rule is the same one in step A.

- The Second WEKA Apriori rule in Step B:

2. Exercise induced Angina = '(-inf-0.5]' 137 ==> Blood sugar = '(-inf-0.5]' 130 (Conf: 0.95, lift: 1.03)

- The Second MATLAB Apriori rule in Step B:

2. Exercise induced Angina = '(-inf-0.5]' 137 ==> Blood sugar = '(-inf-0.5]' 130 (Conf: 0.9489, lift: 1.0276, Sup: 0.6220)

C. Step C (strong rules with "min support": 0.5, and "min confidence": 0.9)

In this step, four rules with "min support" (0.5) in WEKA and three rules with "min support" (0.5) in MATLAB are evaluated. The first and the second rules are the same in step A and step B.

- The Third WEKA Apriori rule in Step C:

3. ECG at rest = '(-inf-0.666667]' Exercise induced Angina '(-inf-0.5]' 119 ==> Blood sugar ='(-inf-0.5]' 113 (Conf: 0.95, lift: 1.03)

- The Forth WEKA Apriori rule in Step C:

4. Resting blood pressure = '(128-164]' 123 ==> Blood sugar ='(-inf-0.5]' 111 (Conf: 0.9, lift: 0.98)

- The Third MATLAB Apriori rule in Step C:

3. ECG at rest = '(-inf-0.666667]' Exercise induced Angina'(-inf-0.5]' 119 ==> Blood sugar = '(-inf-0.5]' 113 (Conf: 0.9496, lift: 1.0283, Sup: 0.5407)

It is evident that in step A, B, and C, MATLAB software exhibits more appropriate performances. In a more detailed discussion, some advantages of MATLAB are thoroughly discussed below:

- All numbers with better accuracy are considered. For example, in The First WEKA Apriori rule in Step A, lift is one. In fact, there is no correlation between ECG (X) and Blood sugar (Y). But, in the first MATLAB Apriori rule in Step A, lift is 1.002. It can be observed ECG and Blood sugar have a weak positive correlation.
- Despite "min support", the exact number of supports are introduced. For instance, in The Second WEKA Apriori rule, only "min support" (0.6) is introduced. However, in The Second MATLAB Apriori rule, despite "min support" (0.6), the exact number of support (0.622) is evaluated.
- Strong rules with high support, confidence and positive correlation lift (more than one) are considered. For example, in The Forth WEKA Apriori rule in step C, lift is 0.98. It means, Resting blood pressure and Blood sugar have a negative correlation. On the other hand, MATLAB does not consider weak rules with negative correlation.

VI. Conclusion

Various Apriori algorithm's strong rules in data mining were compared to predict heart disease. A unique model consisting of one filter and evaluation methods are evolved. Three strong rules, as well as different evaluation methods, are applied to find the superior software. Apriori rules are compared regarding their exact number of support, better accuracy, and considering strong rules. The high-performance software was introduced. The experiment can serve as a practical tool for physicians to effectively predict uncertain cases and advise accordingly.

References

[1] F. Jin, J. Liu, and W. Hou, "The application of pattern recognition technology in the diagnosis and analysis on the heart disease: Current status and future," In 24th Chinese Control and Decision Conference (CCDC), pp. 1304-1307, 2012.

[2] E. Kolce, and N. Frashery, "A literature review on data mining techniques used in Healthcare data bases," ICT innovations web proceedings 2012.

[3] R. Das, I. Turkoglu, and A. Sengur, "Diagnosis of valvular heart disease through neural networks ensembles," Elsevier, 2009.

[4] A. K. Pandey, P. Pandey, K. L. Jaiswal, and A. K. Sen, "Data Mining Clustering Techniques in the Prediction of Heart Disease using Attribute Selection Method," International Journal of Science, Engineering and Technology Research (IJSETR), ISSN: 2277798, Vol 2, Issue10, October 2013.

[5] M. Karaolis, J. A. Moutiris, and C. S. Pattichis, "Association rule analysis for the assessment of the risk of coronary heart events," Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009.

https://doi.org/10.1109/iembs.2009.5334656

[6] H. Danapana, and M. S. Roy, "Effective data mining association rules for heart disease prediction system," IJCST, Vol. 2, Issue 4, Oct-Dec. 2011.

[7] C. Ordonez, "Association rule discovery with train and test approach for heart disease prediction ," IEEE transactions on information technology in biomedicine, Vol 10, No. 2, pp 334-343, April 2006.

https://doi.org/10.1109/TITB.2006.864475

[8] J. Han, M. Kamber, and J. Pay, "Data mining concepts and techniques," Elsevier 2011.

[9] B. L. Deekshatulu, M. A. Jabbar and P. Chantra, "Knowledge discovery from mining association rules for heart disease prediction," Journal of theoretical and applied information technology, Vol.41, No. 2, 2012.

[10] N. Deepika, "Association rules for classification of heart attack patients," IJAEST, Vol.11, pp. 253-257, 2011.

[11] K. R. Lakshmi, M. V. Krishna and S. P. Kumar, "Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability," International Journal of Scientific and Research Publications, ISSN 2250-3153, Vol.3, Issue.6, June 2013.

[12] A. Goodini, M. Torabi, M. Goodarzi, R. Safdari, M. Darayi, M. Tavassoli, and M. Shabani, "The simulation model of teleradiology in telemedicine project," The Health Care Manager, Vol. 34- Issue 1, p 69-78, January/March 2015.

[13] R. Isola, R. Carvalho, and A. Kumar, "Knowledge discovery in medical systems using differential diagnosis, lampstar and K-NN," conference of IEEE transactions on information technology in biomedicine, 2011.

[14] B. Bahrami, and M. H. Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques," Journal of Multidisciplinary Engineering Science and Technology (JMEST), ISSN: 3159-0040, Vol. 2, Issue 2, February 2015.

[15] I. H. Witten and E. Frank, "Data Mining Practical Machine Learning Tools and Techniques," Morgan Kaufman Publishers, 2005.