Volume 5, Issue 3, March 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Efficient Prediction of Heart Disease using Fuzzy Clustering on Medical Data

Asst. Prof. Kanchan Jadhav*, Asst. Prof. Rajesh Phursule
Imperial College of Engg. & Research, Dept. of Computer Engg,
University of Pune, Maharashtra, India

Abstract— This paper involves data mining techniques on patient database. We are applying association rule mining on the available data set from a patient data repository. After applying the Apriori algorithm certain rules are generated. To get better results we are going to modify and transform the data. For this purpose we are going to use the fuzzy c means clustering algorithm. The fuzzy mining algorithm uses linguistic terms for data modification. The Apriori algorithm is then again applied in this data. With the help of this we get more complete set rules and helps in better prediction of heart attack in the patients.

Keywords— Data mining, heart disease, Association rule mining, apriori, clustering

I. INTRODUCTION

Data mining is a very well known technique and process for extraction of desirable knowledge or patterns from large databases for some specific purpose. It is also a process for merging together statistical analysis, machine learning and databases to extract hidden rules and relationships [1]. There are many data mining techniques like classification, association, clustering, etc. These techniques are used according to their application in particular domains.

Cardiovascular or heart diseases are one of many reasons which cause a high mortality rate all over the world. Today, due to the changing lifestyle and habits there is an adverse of on health of individuals which leads to life threatening diseases. As a result there are conditions like high blood pressure, high sugar in human bloody causing heart diseases and diabetes, etc. Due to this heart disease has received much attention for research and analysis from researchers all over the world. Specifically heart attack is a commonly observed heart disease which has increased the death rate. Mostly narrowing or blockage of the coronary arteries and the blood vessels that supply blood to the heart itself is the most common cause of heart disease. This is called coronary artery disease and happens slowly over time. Medical diagnosis is a very difficult but necessary task which should be performed by the doctors in order to save the life of the patient. At the same time various factors have to be considered while diagnosing the patient, his current body test results and previous similar records have to be considered. Unfortunately it is not an easy task and can be risky at times because the life of a patient is involved.

With high mortality rate, it is necessary to gain a clearer understanding of the risk and prevention factors for this disease, as well as improving the accuracy of diagnosis. So, this research has considered factor determinations of coronary disease as the subject for computational diagnostics. For heart disease, diagnostic systems are time consuming, costly and prone to errors. Patients suffering from heart disease need to be under constant observation as improper treatment can be fatal. Moreover, proper identification of the disease and early treatment are essential. Recently computational intelligence has been used to discover and develops relationships between different diseases and patient attributes [11].

II. LITERATURE SURVEY

A. Brief overview of heart disease

Considering the current researches, it has been seen that there is not only a single reason or condition for causing heart disease but can be many condition or their combination which causes injuries to heart and the blood vessels and if they do not carry out their function properly. In the medical field, gender is seen to be an important factor affecting heart disease [3].

The symptoms of heart disease differ from person to person and in many cases early symptoms are not detected. [6][7]. However there are some noticeable symptoms which are seen and are as follows[4]:

- chest pain (Angina pectoris);
- strong compressing or flaming sensation in the chest, neck or shoulders;
- discomforts in chest area;
- sweating, light-headedness, dizziness, shortness of breath;
- pain spanning from the chest to arm and neck, and that amplifying with exertion;
- cough;
- palpitations;
- Fluid retention.

B. Association rule mining on Heart Disease

The medical data available in UCI repository, Cleveland database is viewed as a classification problem in many researches. But in our research we have exploring and considering the use of association rule mining for knowledge extraction. Our research is going to make use of association learning [8][12]. Association learning is a greatly used tool in many fields along with its use in the medical domain. Association rule mining is a well known data mining technique. We are going to perform experiment in which we are going to find out the association rules with use of apriori algorithm by giving the minimum support and confidence.

III. IMPLEMENTATION DETAILS

A. Existing systems

In the existing system, one of the famous and widely used association rule mining algorithm which is the Apriori algorithm is used to generate rules. This algorithm begins with transaction dataset. A frequent item set is constructed, having at least a user specified threshold. In the Apriori algorithm, an item set X of length k is frequent if and only if every subset of X, having length k =1, are also frequent. Due to this there is substantial reduction of search space and rule discovery in a computationally feasible time is possible. In Apriori algorithm confidence is mainly used for the accuracy of the rule [1].

In the experiment as we have mentioned earlier, rules with confidence levels above 70% are selected.
There are some drawbacks seen in the applied Apriori algorithm[13] [18]:

- Large Number of in-frequent item set are generated which increases the space complexity.
- Too many database scans are required because of large number of item sets generated.
- As the number of database scans are more the time complexity increases as the database increases.

B. Implemented system

In our system we are modifying the available data. Fuzzy logic is unique because it can handle numerical data and linguistic knowledge simultaneously. We are using the Fuzzy C Means clustering algorithm to cluster the data. This mining approach integrates fuzzy sets concept with apriori mining algorithm[9][10].

The FCM algorithm attempts to partition a finite collection of n elements $X = \{x_1, ..., x_n\}$ into a collection of c fuzzy clusters with respect to some given criterion. Given a finite set of data, the algorithm returns a list of c cluster centers $C = \{c_1, ..., c_c\}$ and a partition matrix $W = w_{i,j} \in [0,1],\ i = 1, ..., n,\ j = 1, ..., c$, where each element w_{ij} tells the degree to which element x_i belongs to cluster c_j . Like the k-means algorithm, the FCM aims to minimize an objective function. The standard function is:

$$w_k(x) = \frac{1}{\sum_j \left(\frac{d(\text{center}_k, x)}{d(\text{center}_j, x)}\right)^{2/(m-1)}}.$$

which differs from the k-means objective function by the addition of the membership values u_{ij} and the fuzzifier m. The fuzzifier 'm' determines the level of cluster fuzziness. A large 'm' results in smaller memberships w_{ij} and hence, fuzzier clusters. In the limit m = 1, the memberships w_{ij} converge to 0 or 1, which implies a crisp partitioning. In the absence of experimentation or domain knowledge, m is commonly set to 2. The basic FCM Algorithm, given n data points (x_1, . . ., x_n) to be clustered, a number of c clusters with (c_1, . . ., c_c) the center of the clusters, and m the level of cluster fuzziness.

After applying FCM to the dataset it is transformed, features are extracted from it. The modified and transformed data set is then used for the Apriori algorithm. The support and confidence is then used to rank the rules. Both the results, before clustering and after clustering are compared and analyzed for rule generation.

1) Fuzzy C Means Clustering Algorithm

FCM is one of the well known clustering techniques, it was proposed by Dunn [5] in 1973 and later on modified by Bezdek [4] in 1981. It this approach, the data points have their membership values with the cluster centers, which is iteratively updated. The steps for FCM algorithm are as follows:

• Step 1: Let us suppose that M-dimensional N data points represented by xi (i = 1, 2, . . . ,N), are to be clustered.

• Step 2: Assume the number of clusters to be made, that is, C, where $2 \le C \le N$.

• Step 3: Choose an appropriate level of cluster fuzziness $f > 1$.

• Step 4: Initialize the N × C ×M sized membership matrix U, at random, such that $U_{ijm} \in [0, 1]$ and $\Sigma^{c}_{j=1} U_{ijm} = 1.0$, for each i and a fixed value of m.

• Step 5: Determine the cluster centers CC_{jm}, for j^{th} cluster and its m^{th} dimension by using the expression given below:

$$CC_{jm} = \frac{\sum_{i=1}^{N} U_{ijm}^{f} x_{im}}{\sum_{i=1}^{N} U_{ijm}^{f}}$$

• Step 6: Calculate the Euclidean distance between i^{th} data point and j^{th} cluster center with respect to, say m^{th} dimension like the following:

$$D_{ijm} = \|(x_{im} - CC_{jm})\|$$

A. Data Set

We are going to use the UCI heart disease dataset which is publicly available. The dataset consists of a total of 76 attributes, however a maximum of 14 attributes are used [16][17] as these are considerably linked to the heart disease.

All the attributes are given below of which we will be using required values according our requirement These 14 attributes are as follows [16][17].

- Age: numeric;
- Sex: nominal – 2 values: male, female;
- Chest pain type: nominal – 4 values: typical angina (angina), atypical angina (abnang), non-anginal pain (notang), sasymptomatic (asympt).1
- Trestbps: numeric, indicates resting blood pressure on admission;
- Chol:: numeric, indicates Serum cholesterol in mg/dl;
- Fbs: nominal – 2 values: True, False, indicates whether fasting blood sugar is greater than 120 mg/dl;
- Restecg: nominal – 4 values: normal (norm), abnormal (abn): ST–T wave abnormality, ventricular hypertrophy (hyp) –indicates resting electrocardiographic outcomes;
- Thalach: numeric, indicates maximum heart rate achieved;
- Exang: nominal – 2 values: yes, no – highlights existence of exercise induced angina;
- Oldpeak: numeric: ST depression induced by exercise relative to rest;
- Slope: nominal – 3 values: upsloping, flat, downsloping – the slope characteristics of the peak exercise ST segment;
- Ca: numeric – number of fluoroscopy colored major vessels(0–3);
- Thal: nominal – 3 values: normal, fixed defect, reversible defect- the heart status;
- The class attribute: value is either healthy or existence of heart disease (sick type: 1, 2, 3, and 4).

B. Block diagram

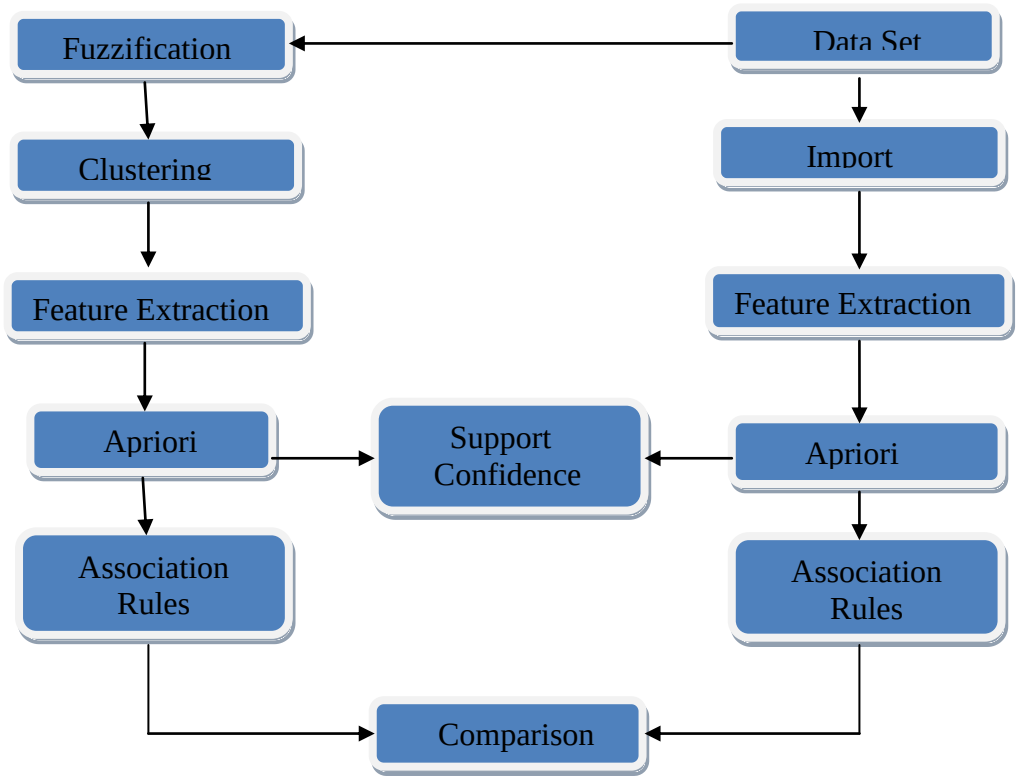

Fig 1. Block diagram of implemented system

C. Results

TABLE I NO OF RULES GENERATED WIT AND WITHOUT FUZZIFICATION

Sr. No.	Support	Confidence	Rules generated	
			Apriori	Apriori after fuzzy clustering
1	30	60	292	124
2	40	70	70	26
3	50	70	26	10

TABLE II COMPARISON OF RESULTS WITH AND WITHOUT FUZZY

Sr. No	Age	BP	Apriori	Fuzzy
1	65	110	0	1
2	65	160	2	1
3	58	150	2	1
4	48	108	0	1
5	52	172	2	1
6	63	145	2	1
7	52	172	2	1
8	41	130	0	1

IV. CONCLUSIONS

From the results present in the table 1, we can say that we can less number and accurate rules by using fuzzy c menas clustering algorithm on the data and using this data for apriori algorithm. Fuzzy logic makes set for age and blood pressure(BP) as high, medium and low. The results obtained for prediction of heart disease are 2 to high, 1 for medium and 0 for low. Thus by comparing the values in table 2, we can conclude that Fuzzy algorithm predicts better values then just applying apriori algorithm on the medical data set of patients. Fin ally we can conclude that It is found that FCM can be segmented into good quality and the performance of the rule based segmentation & clustering technique can achieve better and satisfactory results and good performance in prediction of heart disease or heart attack.

ACKNOWLEDGMENT

Sincere thanks to anonymous researchers for providing us such helpful opinion, findings, conclusions and recommendations. All would like to thank all my teachers and friends for their timely help and support.

REFERENCES

[1] Agrawal, R.T., Imielinski, L., & Swami, A.N. (1993). *Mining association rules between sets of items in large databases.* In International conference on. management of data (SIGMOD-93) (pp. 207–216).

[2] Andersen, L., & Haraldsdottir, J. (2009). *Tracking of cardiovascular disease risk factors including maximal oxygen uptake and physical activity from late teenage to adulthood An 8-year follow-up study.* Journal of Internal Medicine, 234, 309–315.

[3] Barrett-Connor, E., Cohn, B., Wingard, D., & Edelstein, S. (1991*). Why is diabetes mellitus a stronger risk factor for fatal ischemic heart disease in women than in men?* The Rancho Bernardo Study. JAMA, 265, 627–631.

[4] Chilnick, L. D. (2008). Heart disease: An essential guide for the newly diagnosed. Da Capo Press.

[5] Dunn, J.C.: *A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters.* J. Cybernet, Vol. 3, 1973, pp. 32–57.

[6] HEALTHS.(2010).*DefinitionofHeartDisease*.http://www.healthscout.com/ency/68/458/main.html#Definitionof HeartDisease

[7] Health, M. (2010). *Heart disease*. <http://www.mamashealth.com/Heart_disease.asp>.

[8] Huang, Z., Li, J., Su, H., Watts, G. S., & Chen, H. (2007). *Large-scale regulatory network analysis from microarray data: Modified Bayesian network learning and association rule mining.* Decision Support Systems, 43, 1207–1225.

[9] Hullermeier, E., Yi, Y.: *In Defense of Fuzzy Association Analysis.* IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, 37, 1039-1043 (2007).

[10] J. C. Bezdek, *Pattern recognition with fuzzy objective function algoritms,* vol. 2, New York: Plenum Press, 1981, pp. 1-8.

[11] Jesmin Nahar, Tasadduq Imama, Kevin S. Tickle, Yi-Ping Phoebe Chen, *Association rule mining to detect factors which contribute to heart disease in males and females,* Expert Systems with Applications 40 (2013) 1086–1093.

[12] Mangalampalli, A, Pudi,V," *Fuzzy association rule mining algorithm for fast and efficient performance on very large datasets*" Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference.

[13] R. Chang and Z. Liu, "*An Improved Apriori Algorithm*," no. Iceoe, pp. 476–478, 2011.

[14] Thuraisingham, B.: "*A Primer for Understanding and Applying Data Mining*", IT Professional, 28-31, 2000.

[15] Tzung-Pei-Hong, Chan-Sheng Kuo & Sheng-Chai Chi. (2001)"*Trade-off between computation time and number of rules for fuzzy mining from quantitative data*", International Journal of Uncertainty, Fuzziness and Knowledge based Systems, Vol 9, No. 5 (2001) 587-604

[16] UCI. 2009. Heart disease dataset <http://archive.ics.uci.edu/ml/machine-learningdatabases/heart-disease/cleve.mod>.

[17] UCI. 2010. Cleveland Heart disease data details. <http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names>.

[18] Yu Shaoqian, *A kind of improved algorithm for weighted Apriori and application to Data Mining",*, Proc. Of ACM, 2010, pp. 507-510.

