Fuzzy C-Means with Apriori and ID3 for Predicting Heart Stroke Risk Level

MOHD ABDUL HAMEED[1]
FETENECH MESKELE[2]
O.JAMSHEELA[3]

[1,2]University College of Engineering,
Osmania University
Department of Computer Sciences and Engineering
India
[3] EMEA College of arts and Science
Department of Computer Science
India
[1]professor.hameed@gmail.com
[2]fatwne4g@gmail.com
[3]ojamshi@gmail.com

Abstract. The past decades have brought many remarkable researches in diagnosis of disease. The interpretation of the problems in medicine is a significant and tedious task. The detection of heart problem from various factors or symptoms is an issue which is not free from false presumptions often accompanied by unpredictable effects. Thus the effort to utilize knowledge and experience of numerous specialists and clinical data of patients collected earlier to facilitate the interpretation process is considered as a valuable asset. This paper introduces an efficient approach to predict heart stroke risk levels from the heart problem dataset by using machine learning technique. Earlier researchers have used k-means based mafia algorithm and the accuracy was 74%. When modifying the algorithm with fuzzy c-means, the accuracy is increased to 89%. There is a 15% improvement while comparing to the earlier algorithm.

Keywords: Machine learnin: Heart Problem: Fuzzy C-means: Apriori Algorithm and ID3 Algorithm

(Received October 13th, 2019 / Accepted November 23rd, 2019)

1 Introduction

Heart problem or cardiovascular problem is a frequently happening problem and a kind of serious health imperiling. The world health organization has estimated that every year 12 million deaths occur worldwide due to the cardiovascular problem. Advances in the field of medicine over the past few decades enabled the identification of risk factors of cardiovascular problem. Narrowing or blockage of the coronary arteries, the blood vessels that supply blood to the heart, is considered as the most common cause of heart problem. This problem happens slowly over time and is called coronary artery problem. It's the major reason of heart strokes.

Unfortunately the specialized skilled doctors are a scarce resource in many places. Automation would be very useful in this kind of situations. The automated medical interpretation system can enhance medical care and also reduce costs. Only a highly skilled and experienced physician can diagnose heart stroke in a patient. Thus the effort to utilize knowledge and experience of various specialists and clinical data of patients collected earlier to facilitate the diagnosis process is considered a value system that is the integration of clinical decision support with computer technology. Past patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome [1]. The main objective of this research is to develop an Intelligent Heart Problem Prediction System with Apriori algorithm, Decision tree and Fuzzy C means clustering.

2 RELATED WORK

Various supervised machine learning algorithms such as Naive Bayes, Neural Network along with association rule mining Apriori algorithm, have been used for analyzing the dataset in [2]. Weka 3.6.6, the machine learning tool, is used for the experiment. Weka is a collection of Machine learning algorithms for machine learning tasks. The algorithms can either be called from your own Java code or applied directly to a dataset. Weka includes tools for regression, data pre-processing, clustering, Classification, visualization and association rules. The algorithm is also well-suited to form new machine learning algorithms. By using the heart patients' data set Chaurasia and Pal [16] has conducted a study to predict the heart stroke risk levels. The algorithm significantly uses 11 important attributes with basic machine learning technique like J48 decision tree, NaÃ¯ve Bayes and Bagging approaches. The result proves that bagging techniques is more accurate than J48 and Bayesian classification. It shows that the bagging prediction system is capable to predict the heart stroke effectively [3].

A simple probabilistic naive Bayes classifier is used for classification in [3]. In naive Bayesian classifier, the occurrence (or nonoccurrence) of a particular feature of a class is considered as independent to the presence (or absence) of any other feature. The chief NaÃ¯ve Bayes Classifier technique is applicable if more efficient result is expecting and the dimension of the inputs is also high [4, 5, 6]. NaÃ¯ve Bayes model identifies the features and physical characteristics of heart patients. It gives the possibility for each input of attribute for the expectable state. Many new methods have been proposed in health care based on data mining. Nedzved et al proposed a new methodology of an intelligent software development for medical image analysis[21]. Silva et al suggested another method in this area, they have presented a paper with the results of an experiment to treat uncertainties information in human health monitoring system.[22].

3 Fuzzy C- Means Clustering

The process of grouping or partitioning a set of data objects into a number of different clusters is called clustering. By the process of clustering the similar patterns are grouped in to one cluster. Fuzzy C-means clustering (FCM) is introduced by Dunn in 1973[20]. In 1981 Bezdek improved the algorithm and the algorithm is frequently used in pattern recognition [21]. Fuzzy C-means algorithm permits a data point belonging to one or more clusters by using membership value concept [7]. The fuzzy c means clustering algorithm works as follows.

1. Initialization: Cluster centers are randomly initialized.

2. Distance matrix creation: After initialization, next step is the calculation of the distance between the data point xi to each of the cluster center by using Euclidean distance measure.

$$dij = \sqrt{\sum (xi - ci)2} \tag{1}$$

3. Creation of membership function. The fractional distance from the point to the cluster center is considered to calculate the fuzzy measurement. This measurement increased the fraction to the inverse fuzzification parameter. The sum of all fractional distances is used to divide the fuzzification parameter and ensure that the sum of all membership is 1.

$$\mu j(xi) = \frac{i}{dj} 1(m-1) / \sum \frac{i}{dj} 1(m-1) \tag{2}$$

-verify the total membership is equal to 1

$$\sum_{j-1}^{p} \mu j(xi) = 1 \tag{3}$$

4. A new centroid is generated for each cluster by using the given formula.

$$Cj = \sum [\mu j(xi)] mxi / \sum [\mu j(xi)] m \tag{4}$$

5. To generate optimized cluster centers, the above steps are repeated.

In this experiment to cluster the patients, Fuzzy C-Means (FCM) and Hard C- Means (HCM) algorithms are used as an unsupervised clustering method. The FCM employs fuzzy partitioning such that a data point can belong to more than one group(may be to all group) with different membership grades between 0 and 1. FCM is an iterative algorithm. To find the cluster centroids(centers) which minimize a dissimilarity function FCM is applied. The use of Fuzzy C-Means leads to the class membership to become a relative one and an object can belong to more than one class at the same time but with different degrees. To increase the sensitivity of medical diagnostic systems, the above feature has an important role. Fuzzy C-means algorithm gives better result while comparing with hard-k- means algorithm in medical diagnostic systems. For the medical experts, during diagnostic, Fuzzy C-means methods can be an important supportive tool.

4 Apriori algorithm to find the frequent pattern from the dataset.

R.Agrawal and R.Srikant proposed Apriori algorithm in 1994[17]. For mining frequent itemsets, Apriori algorithm is one of the most classical and important algorithms. Apriori is an efficient algorithm to find frequent item sets from large datasets[19]. Apriori algorithm applies multiple passes over the dataset. It uses an iterative level-wise search(breadth-first search) through the dataset, where to find frequent (k+1)-item sets frequent k-item sets are used. Apriori algorithm works based on the Apriori property which states that" All nonempty subsets of a frequent item set must be frequent". Another important property called anti-monotone property is also applied in Apriori. Anti-monotone property is based on that if an itemset cannot pass the minimum support test, all its supersets also will fail to pass the test. Therefore if any itemset is infrequent then all its supersets are also infrequent and vice versa. In Apriori to prune the infrequent candidate elements this property is used. As the first step, Apriori finds the set of frequent 1-itemsets. The set of frequent 1-itemsets contains item names with support count, which satisfies the support threshold and is denoted by L. Each subsequent nth pass starts with the set of itemsets which is collected in the previous n-1th pass and used to find larger frequent itemsets. At the end of each pass k, a set of frequent k-itemsets are collected and they become the inputs for the next pass k+1. Therefore, L is used to find L1, the set of frequent 2-item-sets, which is used to find L2, and so on, until no more frequent k- item sets can be found.

5 Iterative dichotomize3 (ID3)

ID3 (Iterative dichotomiz 3) [18] is one of the most widely used decision tree algorithms. ID3 is used here as the information gain measure to select among the candidate attributes at each step while growing the tree. Information gain is simply the expected reduction in entropy. The reduction is caused by portioning the examples according to this attribute [7].The information gain, Gain (S,A) of an attribute A, relative to a collection of example S, is defined as,

$$\begin{gathered} Gain(S, A) = Entropy(S) - \\ \sum vevalues(A) \frac{sv}{s} Entropy(Sv) \end{gathered} \tag{5}$$

$$Entropy(S) = \sum_{i-1}^{c} pi\, log2pi \tag{6}$$

Where pi is the proportion of S belonging to class i.

6 EXPERIMENTAL RESULTS

The details of the experimental analysis are presented in this section. The objective of this paper is to find significant patterns efficiently for heart stroke prediction. The patterns, which are relevant to the heart stroke prediction, are extracted using the above mentioned approaches. The patients' data, which is already recorded, is preprocessed successfully by adding missing values and removing duplicate records as shown in Table-2. After applying the preprocessing methods on the dataset, the filtered dataset is then clustered by using fuzzy C-means algorithm with C value as 2. Once the cluster of the data is constituted, by using the Apriori algorithm the frequent patterns are mined from the cluster. Table-3 shows the sample combinations of heart stroke parameters along with their levels and values for risk level and normal level. The other tables contain the remaining data. The sample combination is analyzed as follows:-

1. If the value is less than or equal to 0.4 the prediction indicates that it is in normal level .

2. If the value is greater than or equal to 0.4, then it shows that it is not in normal level and is in risk level.

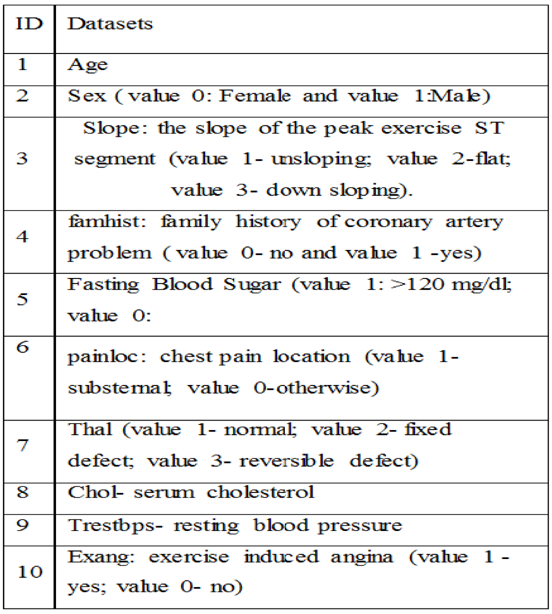
3. And if the value is (0.8) or (0.9) then it shows the higher risk level.

After completing the above procedure, frequent patterns, which are relevant to heart problem, are mined efficiently with and orderly from the cluster by using the Apriori algorithm. Finally the new algorithm is compared with the existing algorithm and the accuracy, precision, efficiency and recall are shown by graphs.

If Age=<30 and Overweight=no and Alcohol Intake=never Then Heart stroke level is Low (Or) If Age=>70 and Blood pressure=High and Smoking=current Then Heart stroke level is High

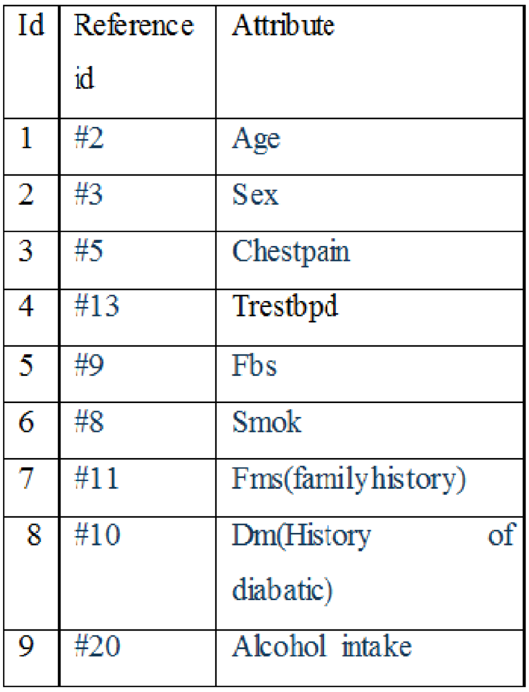
7 Performance Measures

For calculating other aggregate performance measures, the performance measures like SPECIFICITY, RECALL (SENSITIVITY) and F-measure are also used. Besides high precision and recall metrics, the goal is



ID	Datasets
1	Age
2	Sex (value 0: Female and value 1:Male)
3	Slope: the slope of the peak exercise ST segment (value 1- unsloping; value 2-flat; value 3- down sloping).
4	famhist: family history of coronary artery problem (value 0- no and value 1 -yes)
5	Fasting Blood Sugar (value 1: >120 mg/dl; value 0:
6	painloc: chest pain location (value 1- substernal; value 0-otherwise)
7	Thal (value 1- normal; value 2- fixed defect; value 3- reversible defect)
8	Chol- serum cholesterol
9	Trestbps- resting blood pressure
10	Exang: exercise induced angina (value 1 - yes; value 0- no)

Table-1 Dataset of Heart Patients



Id	Reference id	Attribute
1	#2	Age
2	#3	Sex
3	#5	Chestpain
4	#13	Trestbpd
5	#9	Fbs
6	#8	Smok
7	#11	Fms(familyhistory)
8	#10	Dm(History of diabatic)
9	#20	Alcohol intake

Table-2 Clustered relevant data

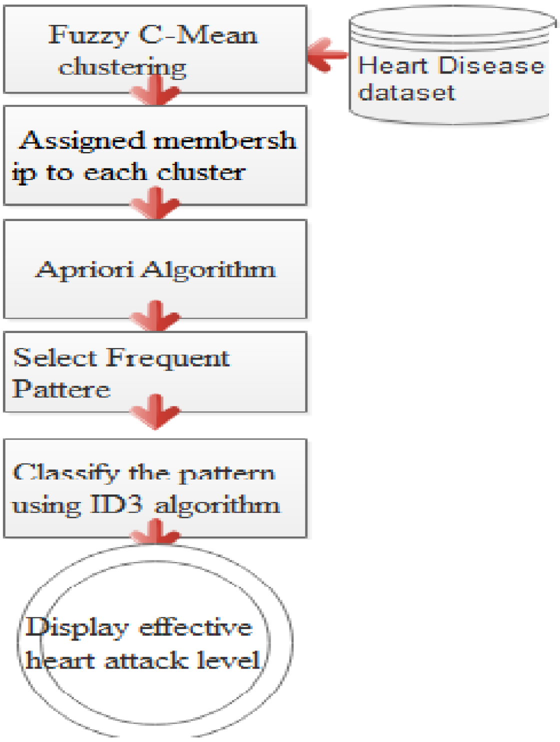

Fuzzy C-Mean clustering
Heart Disease dataset
Assigned membership to each cluster
Apriori Algorithm
Select Frequent Pattere
Classify the pattern using ID3 algorithm
Display effective heart attack level

Fig.1.SYSTEM ARCHITECTURE

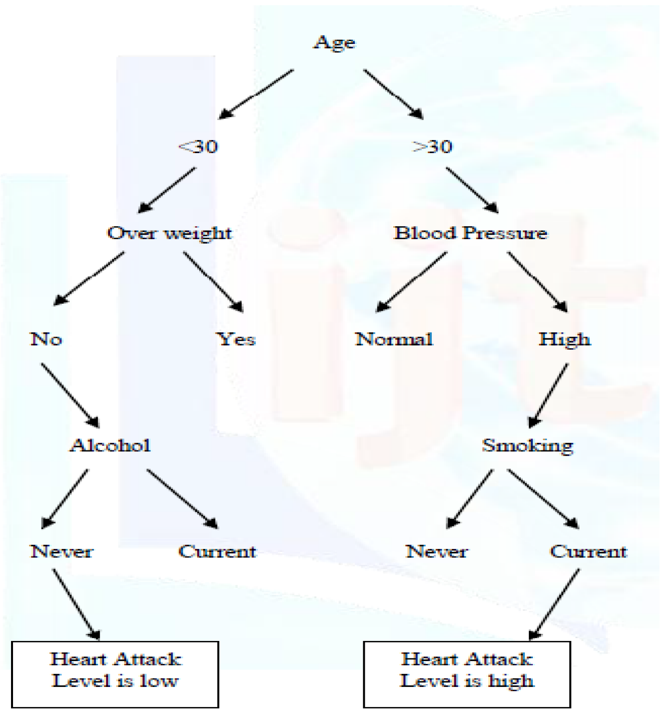
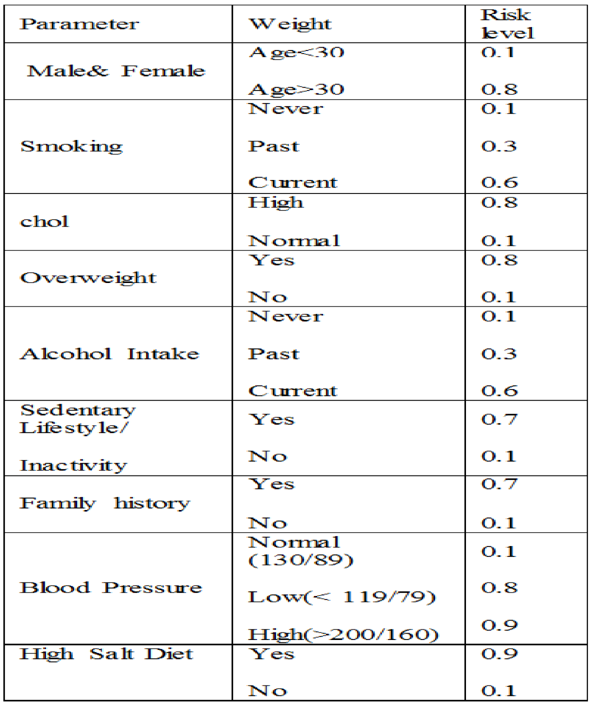
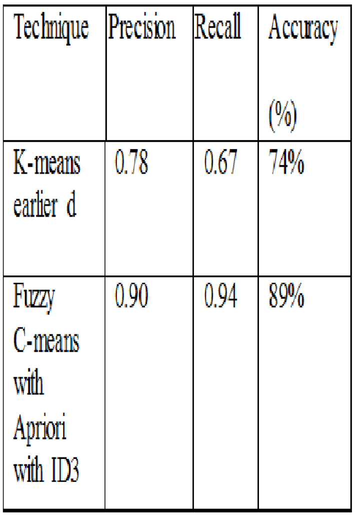


Fig-2 Decision tree to predict heart stroke level



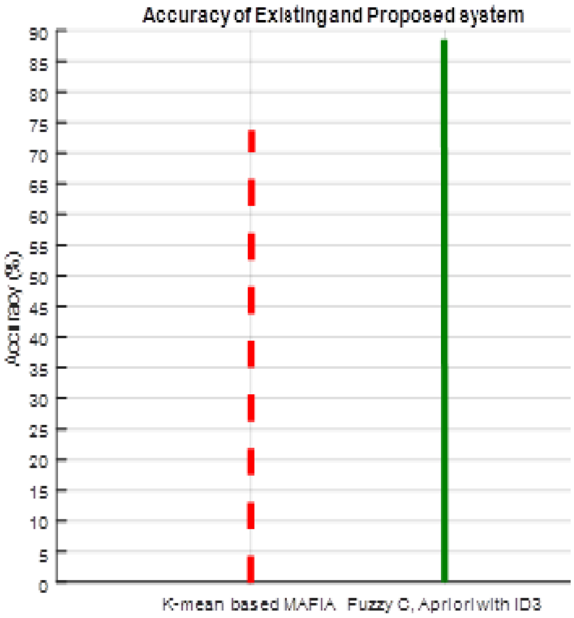
Parameter	Weight	Risk level
Male& Female	Age<30	0.1
	Age>30	0.8
Smoking	Never	0.1
	Past	0.3
	Current	0.6
chol	High	0.8
	Normal	0.1
Overweight	Yes	0.8
	No	0.1
Alcohol Intake	Never	0.1
	Past	0.3
	Current	0.6
Sedentary Lifestyle/ Inactivity	Yes	0.7
	No	0.1
Family history	Yes	0.7
	No	0.1
Blood Pressure	Normal (130/89)	0.1
	Low(< 119/79)	0.8
	High(>200/160)	0.9
High Salt Diet	Yes	0.9
	No	0.1

Table-3 heart stroke parameters with corresponding values and their weight



Technique	Precision	Recall	Accuracy (%)
K-means earlier d	0.78	0.67	74%
Fuzzy C-means with Apriori with ID3	0.90	0.94	89%

Table-4 comparison between simple mafia with k means and proposed fuzzy C- means with Apriori and ID3 algorithm



also to have high accuracy. These metrics can be derived from the confusion matrix and can be easily converted true-positive (TP) and false-positive (FP) metrics.

$$Precision = \frac{TP}{TP+FP}, recall = Recall = \frac{TP}{TP+FN}$$
$$and Accuracy = TP+TN/TP+FP+FN+TN \quad (7)$$

1.True Positive (TP): Total percentage of members classified as Class A belongs to Class A.

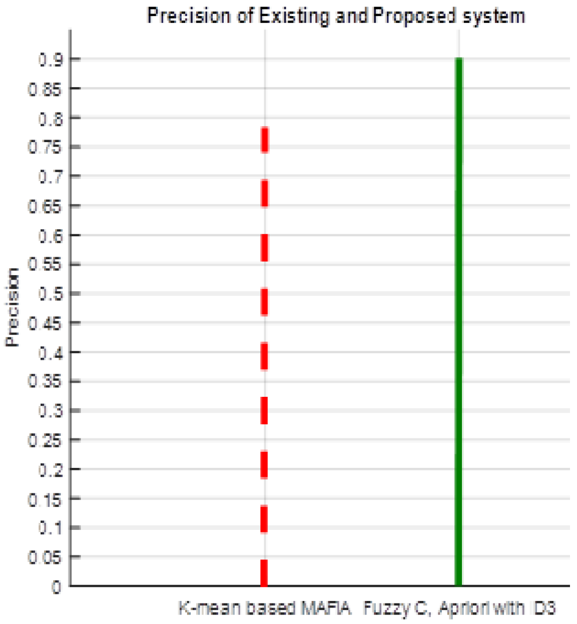
2. False Positive (FP): Total percentage of members of Class A but does not belong to Class A.

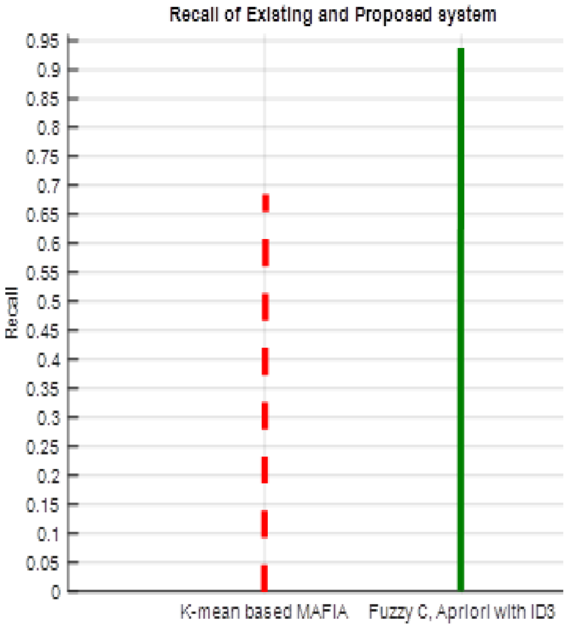
3. False Negative (FN): Total percentage of members of Class A incorrectly classified as not belonging to Class A.

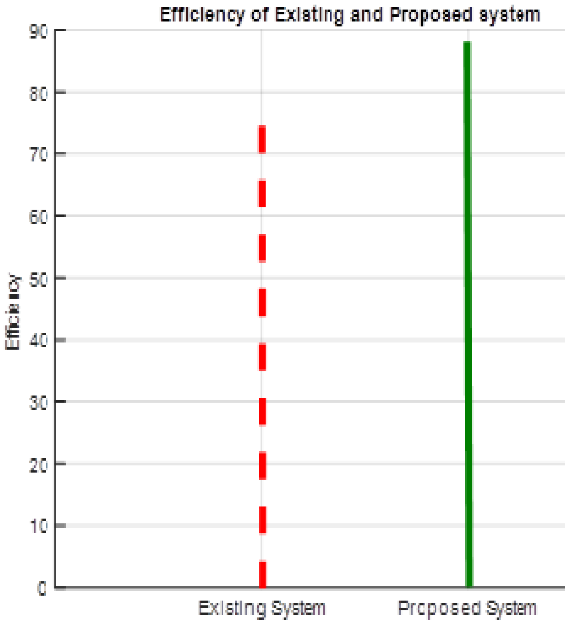
4. True Negative (TN): Total percentage of members which do not belong to Class A, are classified not a part of Class A. It can also be given as (100

8 Conclusion and future work

Heart problem is one of the leading causes of death worldwide. So the early prediction of heart problem is very important. The computer-aided heart prob-







lem prediction system helps the physician as a tool for heart problem interpretation. This paper has presented a Heart Problem Interpretation System using machine learning techniques. Apriori algorithm is used to find the frequent term sets and the maximal frequent term set is generated. Clustering is performed using k-means clustering algorithm. Lastly, the ID3 algorism is applied to show the classification. Defining the clusters earlier d on maximal frequent item sets provided improved accuracy. As a future work, an efficient heart stroke prediction system by using python programming language and with Oracle Dataset is planned to develop.

References

[1] Sellappan Palaniappan, Rafiah Awang, Intelligent Heart Problem Prediction System Using Machine learning Techniques, IEEE Conference, 2008, pp 108-115.

[2] Mark Hall; Eibe Frank; Georey Holmes; Bernhard Pfahringer; Peter Reutemann; Ian H. The weka machine learning software: An update. SIGKDD Explorations, 11, 2009.

[3] Shanta Kumar, B.Patil, Y.S. Kumaraswamy,Predictive machine learning for medical interpretation of heart problem prediction, IJCSE Vol .17, 2011

[4] Liu X, Lu R, Ma J, Chen L. Privacy-preserving patient-centric clinical decision support system on naive Bayesian classification. IEEE Journal of Biomedical and Health Informatics. 2016; 20(2):655-88.

[5] Patil RR. Heart problem prediction system using Naive Bayes and Jelinek-mercer smoothing. International Journal of Advanced Research in Computer Science and Communication Engineering. 2014; 3(5):6787-9

[6] Pattekari SA, Parveen A. Prediction system for heart problem using naive Bayes. International Journal of Advanced Computer and Mathematical Sciences. 2012; 3(3):290-4.

[7] Komal G, Vekariya V, Novel approach for heart problem prediction using a decision tree algorithm, International Journal of Innovative Research in Computer and Communication Engineering, 3(11):11544-1, 2015.

[8] Anand Bahety, Extension and Evaluation of ID3 - Decision Tree Algorithm. University of Maryland, College Park.

[9] S. K. Yadav and Pal S., Machine learning: A Prediction for Performance Improvement of Engineering Students using Classification, World of Computer Science and Information Technology (WCSIT), 2(2), 51-56, 2012.

[10] Quinlan J. C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann; 1993.

[11] Fayyad, P.-S. S. (1996). Advances in Knowledge Discovery and Machine learning. AAAI Press / The MIT Press, 1-34.

[12] Srinivas, K., Analysis of coronary heart problem and prediction of the heart stroke in coal mining regions using machine learning techniques, IEEE Transaction on Computer Science and Education (ICCSE), p(1344 - 1349), 2010.

[13] M. Anbarasi et. al. Enhanced Prediction of Heart Problem with Feature Subset Selection using Genetic Algorithm, International Journal of Engineering Science and Technology Vol. 2(10), 5370-5376 ,2010.

[14] Shantakumar, B.Patil, Y.S.Kumaraswamy, Intelligent and Effective Heart Stroke Prediction System Using Machine learning and Artificial Neural Network; European Journal of Scientific Research, ISSN 1450-216X Vol.31 No.4, 2009.

[15] Majali J, Niranjan R, Phatak V, Tadakhe O. Machine learning techniques for interpretation and prognosis of cancer. International Journal of Advanced Research in Computer and Communication Engineering. 2015; 4(3):613-6.

[16] Chaurasia, V. and Pal, S., 2014. Data mining approach to detect heart diseases. International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol, 2, pp.56-66.

[17] Agrawal, R. and Srikant, R., 1994, September. Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).

[18] Bhatt, R.B. and Gopal, M., 2004, July. FRID: fuzzy-rough interactive dichotomizers. In 2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No. 04CH37542) (Vol. 3, pp. 1337-1342). IEEE.

[19] Agrawal, R., Imielinski, T. and Swami, A. Mining association between sets of items in massive database. International Proceedings of the ACM-SIGMOD International Conference on Management of Data, 207-216, 1993.

[20] J.C. Dunn (1973). A fuzzy relative of the ISODATA process and its use in detecting compact wellsparated clusters. Journal of Cybernetics 3, 32-57. 21. J.C. Bezdek (1980). A convergence theorem for the fuzzy ISODATA clustering algorithms. IEEE Trans. on Pattern Anal. Machine Intel l. PAMI-2, 1-8.

[21] Nedzved, Alexander, and Valery Starovoitov. A Flexible Suite of Software Tools for Medical Image Analysis. Proc. of the First Intern. Conf. on Advanced Communications and Computation (INFOCOMP 2011). 2011.

[22] Silva, Katia Cilene Neles, and Graca Bressan. Human Health Monitoring by Sensors: Analysis of Contextual Uncertainties Through Dempster-Shafer Evidence Theory. INFOCOMP 16.1-2 (2017): 46-54.