Original Article
http://mjiri.iums.ac.ir
Medical Journal of the Islamic Republic of Iran (MJIRI)

Med J Islam Repub Iran. 2018(12 Sep);32.85. https://doi.org/10.14196/mjiri.32.85

A novel method for fuzzy diagnostic system design

Mostafa Langarizadeh[1], Azam Orooji[2*]

Received: 16 Feb 2017 Published: 12 Sep 2018

Abstract

Background: In recent years, liver disorders have been continuously increased. Proper performance of data mining techniques in decision-making and forecasting caused to use them commonly in designing of automatic medical diagnostic systems. The main aim of this paper is to introduce a classifier for diagnosis of liver disease that not only has high precision but also is understandable and has been created without expert knowledge.

Methods: In regards to this purpose, fuzzy association rules have been extracted from dataset according to fuzzy membership functions which determined by fuzzy C-means clustering method; while each time, extracting fuzzy association rules, one of the five quality measures including confidence, coverage, reliability, comprehensibility and interestingness is used and five fuzzy rule-bases extracted based on them. Then, five fuzzy inference systems are designed on the basis of obtained rule-bases and evaluated in order to choose the best model in terms of diagnostic accuracy.

Results: The proposed diagnostic method was examined using data set of Indian liver patients available at UCI repository. Results showed that among considered quality measures, interestingness, reliability and truth outperformed respectively, and yielded precision, sensitivity, specificity and accuracy of more than 90%.

Conclusion: In this paper, a classification method was developed to predict liver disease which in addition to high classification accuracy, it has been created without expert knowledge and provided an understandable explanation of data. This method is convenient, user friendly, efficient and requires no expertise.

Keywords: Fuzzy association rule mining, Membership function extraction, Liver disease, Fuzzy diagnostic system, Rule quality measures

Cite this article as: Langarizadeh M, Orooji A. A novel method for fuzzy diagnostic system design. *Med J Islam Repub Iran. 2018* (12 Sep);32:85. https://doi.org/10.14196/mjiri.32.85

Introduction

Diagnosis of liver diseases is not easy in the early stage, since it works well for a long time even when a great part of the liver is damaged. Early diagnosis of liver disorders can increase patients' survival rate (1). Data mining techniques has been increasingly applied successfully on medical data in the past decade (2). Such techniques are divided into two predictive and descriptive categories. Both groups try to discover hidden patterns in data; but descriptive methods such as clustering and Association Rule Mining (ARM) do not need labeled data while predictive methods such as classification require training data (3).

Classification techniques have been commonly used for automatic diagnosis in different medical areas (4-7). Automatic diagnosis could be helpful in order to reduce physicians' work load significantly (1). However, few works have used data mining to diagnose liver disorders (8). To make a model for liver diseases diagnosis, classification methods such as Bayesian network (9, 10), Support Vector Machine (SVM) (10-12) and Artificial Neural Network (ANN) (2,4) have been mostly used,

Corresponding author: Azam Orooji, orooji.a@tak.iums.ac.ir

[1] Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran
[2] Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran

↑What is "already known" in this topic:
Disease prediction is a vibrant research area. In data mining, classification techniques are much popular in medical diagnosis and predicting diseases. However, application of classification techniques in medical diagnosis has several deficiencies and shortages.

→What this article adds:
A weighted-fuzzy association rule-based classifier was used for predicting liver disease which in addition to high classification accuracy, it created without expert knowledge and provided an understandable explanation of data. This method is convenient, efficient, and requires no expert.

[DOI: 10.14196/mjiri.32.85]

Table 1. Descriptive statistics for ILPD

Attribute name		Characteristics	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	Age	Integer	4.00	32.00	43.00	43.47	57.00	90.00
2	Total Bilirubin	Real	0.400	0.700	0.900	2.513	1.800	75.000
3	Direct Bilirubin	Real	0.100	0.200	0.200	1.089	0.800	19.700
4	Alkaline Phosphotase	Integer	63.0	168.0	198.0	264.8	279.0	2110.0
5	Sgpt Alamine Aminotransferase	Integer	10.00	22.00	31.00	63.57	52.00	2000.00
6	Sgot Aspartate Aminotransferase	Integer	10.0	23.0	35.0	84.7	66.0	4929.0
7	Total Proteins	Real	2.700	5.800	6.600	6.505	7.200	9.600
8	Albumin	Real	0.900	2.700	3.200	3.216	3.900	5.500
9	Albumin and Globulin Ratio	Real	0.3000	0.8000	1.0000	0.9767	1.1100	2.8000
10	Gender	Binary	Male : 441, Female : 142					
11	Selector (Class Label)	Binary	Liver patient : 416 , Non liver patient : 167					

however they all called black box since the predicting model is not understandable for humans, while in many cases it is important for people to know how the system works (13). Hence, data mining approaches such as ARM are important to summarize large volume of data with an understandable format (14).

ARM discovers hidden associations in huge volume of data (15). The obtained associations are in the form of if-then rule; hence ARM plays a significant role in terms of creating decision models (16). In recent years, ARM method has been considered because of its good function for classification (17, 18). Indeed, association rules can be used for classification, if the consequent part includes only the class labels. ARM has been worked only with categorical data (19). Therefore, discretization algorithms were introduced to change a continuous interval attribute into several segments and mapping data based on them. But in this method interval borders are crisp. The problem was solved using Fuzzy ARM extracting rules which can be used for classification (20). This method is understandable and human-friendly providing what happened inside the classifier.

Although several papers have addressed the subject of extracting fuzzy rules from databases and construction of fuzzy rule-based classification system (21, 22), few papers have considered on extracting weighted fuzzy rules (14, 23, 24) and defining appropriate membership functions (20,25). While the weight of rules, the number of parameters of fuzzy membership functions play a significant role in the prediction of system efficiency. Different Quality Measures (QMs) have been introduced for association rules which here five measures including truth (or confidence), coverage, reliability, comprehensibility and interestingness have been examined (14, 25). Fuzzy membership functions could be conducted on the basis of expert knowledge, but since it is not always possible, other methods were introduced for extracting functions using data (26). Evolutionary algorithms such as Genetic Algorithm (GA) (27, 28), Particle Swarm Optimization (PSO) (25,29), Ant Colony Optimization (ACO) (30) are the most commonly used methods in order to infer fuzzy membership functions based on existed data. Since such methods are time-consuming, they were replaced with clustering methods as well as fuzzy c-means (FCM) which outperformed other techniques (31,32).

The main aim of this paper was to propose a classifier for diagnosis of liver disease. For this purpose, a known ARM algorithm called *a priori* (33) has been modified based on membership functions which calculated using FCM clustering method to extract fuzzy classification rules. In addition, for weighting rules, five different quality measures were considered. Finally, effect of each factor on the accuracy of proposed method was examined.

The construction of proposed method is addressed in the second section of this paper. The findings are reported in the third section. In the fourth section, the proposed method is compared with algorithms presented in previous works and finally conclusion prepared.

Methods

In this work, data set of patients with liver disease available in UCI Machine Learning Repository, University of California (http://archive.ics.uci.edu/ml) was used. This data set includes information about 416 patients with liver disease and 167 non-liver patients. Descriptive statistics is showed in Table 1.

The proposed method is prepared in six steps as shown in Fig. 1. First step was pre-processing and next steps addressed clustering and extraction of fuzzy association

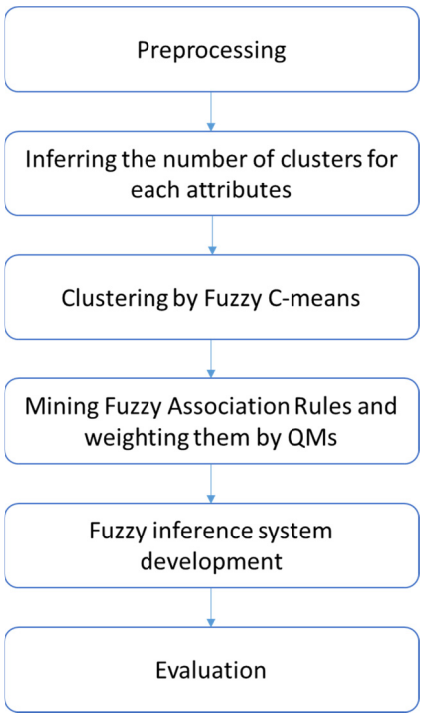

Fig. 1. Steps of the proposed classification method

[DOI: 10.14196/mjiri.32.85]

rules. Each time, in extracting fuzzy association rules, one of the QMs was used and five fuzzy rules extracted based on five measures. Then, FISs were designed on the basis of output of previous steps and evaluated in order to choose the best model.

Pre-processing

In this step, the average of each column was used to fill its missed values. To reduce the variety of baseline between variables, all data were normalized in (0, 1) interval by using uniform normalization method. In addition to the mentioned issues, SMOTE as an over-sampling technique was used to balance two classes of samples. SMOTE is a known algorithm which generates synthetic examples from every minority classes on the basis of the nearest neighbors in order to increase generalization performance of classifier over the minority classes.

Inferring the number of clusters for each attributes

To determine fuzzy system membership functions, FCM clustering method was used. One of the factors affecting the efficiency of proposed method was the number of clusters that should be set by the user. To find the best number of clusters for each feature, two different groups of clustering quality measures called Internal and Stability were examined for 2, 3 and 4 clusters. Stability measures included Average Proportion of Non-overlap (APN), Average Distance (AD), Average Distance between Means (ADM), Figure Of Merit (FOM) and internal measures were included connectivity, Silhouette and Dunn index (34). The description of stability and internal measures is shown in Table 2.

Clustering by fuzzy C-means

After determining the best number of clusters, FCM method was applied to each feature and fuzzy membership functions prepared based on clustering output.

Mining fuzzy association rules and weighting them by QMs

According to obtained membership functions, fuzzy association rules were extracted. Each time, one of the QMs considered in extracting fuzzy association rules i.e. five fuzzy rule bases were extracted based on five measures.

Format of a fuzzy association rule was as IF x is X then y is Y where x (y) is input (output) variable and X (Y) are input (output) membership functions. Then the quality measures are defined and calculated as follows.

1. Truth /Confidence (T)

This measure is equal to means of ratio of transactions in dataset which antecedent and consequent parts of rules occur together divided to total number of transactions containing the antecedent part expressed as a percentage (Eq. 1).

$$T = \frac{\sum_{m=1}^{M} min(\mu_X(x^m).\mu_Y(y^m))}{\sum_{m=1}^{M} \mu_X(x^m)} \quad (1)$$

Where, M is the number of input data (here 917).

2. Coverage (C)

It specifies whether a rule is supported by sufficient amount of data. For calculation of C first coverage ratio is calculated as follows:

$$r_c = \frac{\sum_{m=1}^{M} t_m}{M} \quad (2)$$

Where

$$t_m = \begin{cases} 1 & \mu_X(x^m) > 0 \ and \ \mu_Y(y^m) > 0 \\ 0 & otherwise \end{cases} \quad (3)$$

Since r_c is very small (often less than 0.1), its value normalized by function f in the range of 0 and 1 i.e.,

Table 2. Description of stability and internal measures [34]

Cluster validity criteria		Description
Internal criteria	Connectivity	Connectivity indicates the degree of connectedness of the clusters, as determined by k-nearest neighbors. Connectedness corresponds to what extent items are placed in the same cluster as their nearest neighbors in the data space. The connectivity has a value between 0 and infinity and should be minimized.
	Silhouette Width	The Silhouette Width is the average of each observation's Silhouette value. The Silhouette value measures the degree of confidence in a particular clustering assignment and lies in the interval [-1, 1], with well-clustered observations having values near 1 and poorly clustered observations having values near -1.
	Dunn Index	The Dunn Index is the ratio between the smallest distances between observations not in the same cluster to the largest intra-cluster distance. It has a value between 0 and infinity and should be maximized.
Stability criteria	Average proportion of non-overlap (APN)	The APN measures the average proportion of observations not placed in the same cluster by clustering based on the full data and clustering based on the data with a single column removed. The values of APN range from 0 to 1, with smaller value corresponding with highly consistent clustering results.
	Average distance (AD)	The AD measures the average distance between observations placed in the same cluster under both cases (full dataset and removal of one column). AD has a value between 0 and infinity, and smaller values are also preferred.
	Average distance between means (ADM)	The ADM measures the average distance between cluster centers for observations placed in the same cluster under both cases. The values of ADM range from 0 to 1, with smaller value corresponding with highly consistent clustering results.
	Figure of merit (FOM)	The FOM measures the average intra-cluster variance of the deleted column, where the clustering is based on the remaining (undeleted) columns. It also has a value between zero and 1, and again smaller values are preferred. The values of FOM range from 0 to 1, with smaller value corresponding with highly consistent clustering results.

[DOI: 10.14196/mjiri.32.85]

$$f(r_c)=\begin{cases}0\ .\ r_c\le r_1\\ 2\left(\frac{r_c-r_1}{r_2-r_1}\right)^2\ .\ r_1<r_c<\frac{r_1+r_2}{2}\\ 1-2\left(\frac{r_2-r_c}{r_2-r_1}\right)^2\ .\ \frac{r_1+r_2}{2}\le r_c<r_2\\ 1\ .\ r_c\ge r_2\end{cases} \quad (4)$$

Function f has two parameters that in this paper represented as $r_1=0.02$ and $r_2=0.15$ and finally $C=f(r_c)$.

3. Reliability (R)

The reliability can be viewed as measuring the computed validity of a rule using equation 5. A rule is valid if and only has high degree of truth (T) and coverage (C).

$$R=min(T.C) \quad (5)$$

4. Comprehensibility (Com)

The measure considers the length of each rule. If the number of antecedent variables is l_a and the number of consequent variables of a rule is l_c., then Com is as follows:

$$Com=\frac{log(1+l_c)}{log(1+l_c+l_a)} \quad (6)$$

5. Interestingness (I)

The measure had a high value for rules that comparatively was a less occurrence in the whole of dataset and possibly had a specific innovation and is calculated as follows.

$$I=\frac{\sum_{m=1}^{M}min(\mu_X(x^m).\mu_Y(y))}{\sum_{m=1}^{M}\mu_X(x^m)}\times\frac{\sum_{m=1}^{M}min(\mu_X(x^m).\mu_Y(y))}{\sum_{m=1}^{M}\mu_Y(y)}\times\left(1-\frac{\sum_{m=1}^{M}min(\mu_X(x^m).\mu_Y(y))}{M}\right) \quad (7)$$

Fuzzy inference system development

Five Mamdani product Fuzzy Inference Systems (FISs) were designed based on membership functions and rules which obtained in steps 2 and 3 that each of which contained one of the five QMs as weight of rules.

Evaluation

In order to evaluate FIS, measures such as precision, specificity, sensitivity and accuracy were used. The calculations of measures are given in Table 3.

Results

The proposed classification method was examined using data set of Indian liver patients available at UCI repository and programmed using R3.2.3 and MATLAB R2014a. In pre-processing step, first of all, in 4 cases the rate of A/G was filled using average of column, and then, whole data set normalized to (0, 1) to reduce the variability of baseline between variables. Finally, since in given data set, less than one-third of records were assigned to class 2 (non-liver patients), number of records in this class increased up to three times using SMOTE technique. Total number of samples increased to 917.

To find fuzzy membership functions, FCM clustering was used by receiving the number of clusters as input. Furthermore, to determine the best number of clusters two different groups of clustering quality measures including stability and internal was used. Table 4 is clearly shown the values of these measures in terms of the number of clusters for each attribute. For two binary attributes i.e. gender and selector, two single fuzzy membership functions were defined and tuned on 0 and 1.

In this step, fuzzy association rules were mined according to calculated membership functions. FARM algorithm has two parameters: 1) min-support that specifies the minimum support for finding frequent item sets and 2) min-confidence that puts only rules in output that have confidence higher than threshold. However, in this paper, four other QMs have been introduced in addition to confidence value, that each of which represents a specific aspect of the rule quality. In order to make a fuzzy rule base, FARM algorithm was implemented five times with min-support= 0.02 and min-QM= 0.7. Obtained rules in each of the FISs were weighted according to one of the QMs. The performance of the proposed classification method using weighted rule bases is shown in Table 5.

Discussion

To predict liver disease, Jin et al. used six classification algorithms including Naïve Bayes, Decision tree, K-Nearest neighbors (KNN), Multi-Layered Perceptron (MLP), Logistic and Random Forest (RF)and compared precision, accuracy, sensitivity, specificity, the area under ROC curve, Kappa and Root Mean Square Error (RMSE) using the ILPD. The method of cross validation with 10 fold was used to evaluate and compare classification algorithms. The results showed that the logistic had the best performance in terms of sensitivity (= 91.3%), Accuracy (= 72.7%), ROC and RMSE (= 0.42) and Naïve Bayes in terms of precision (= 95.1%), specificity (= 95.2%) and Kappa (35).

Gulia et al. used some classification algorithms including J-48 classifier, MLP, SVM, Bayesian network and RF classified ILPD. In second phase, most important features were selected using greedy step wise approach and then classification algorithms applied on obtained significant subset of features. Finally, the results of two examinations, with and without feature selection, were compared based on accuracy and mean absolute error. All steps were performed using WEKA data mining tool. The

Table 3. Statistical Meaning for Sensitivity, Specificity, Precision and Accuracy where

Statistical measures of the performance	Definitions
Precision or Positive Predictive Value (PPV)	PPV= TP/(TP+FP)
Specificity (SPC) or true negative rate	SPC = TN/N = TN/(TN+FP)
Sensitivity or True Positive Rate (TPR) or recall	TPR=TP/P = TP/ (TP+FN)
Accuracy (ACC)	ACC=(TP+TN)/(TP+TN+FP+FN)

TP is true positive, FN false negative, FP false positive and TN true negative

[DOI: 10.14196/mjiri.32.85]

Table 4. Values of internal and stability measures in terms of the number of clusters for each attribute

Attribute	#Clusters: 2		#Clusters: 3		#Clusters: 4	
	Internal criteria	Stability criteria	Internal criteria	Stability criteria	Internal criteria	Stability criteria
Age	Conn: 3.5369 Dunn: 0.0222 SW: 0.5799	APN: 0.0000 AD: 0.1672 ADM: 0.0774 FOM: 0.0942	Conn: 3.7000 Dunn: 0.0286 SW: 0.5462	APN: 0.0000 AD: 0.1487 ADM: 0.0706 FOM: 0.0943	Conn: 13.1179 Dunn: 0.0294 SW: 0.5473	APN: 0.0000 AD: 0.1389 ADM: 0.0772 FOM: 0.0944
Total Bilirubin (TB)	Conn: 2.5762 Dunn: 0.0044 SW: 0.8302	APN: 0.0000 AD: 0.0428 ADM: 0.0217 FOM: 0.0417	Conn: 9.6683 Dunn: 0.0016 SW: 0.6769	APN: 0.0000 AD: 0.0385 ADM: 0.0218 FOM: 0.0417	Conn: 5.5905 Dunn: 0.0016 SW: 0.6757	APN: 0.0000 AD: 0.0360 ADM: 0.0226 FOM: 0.0417
Direct Bilirubin (DB)	Conn: 5.5956 Dunn: 0.0060 SW: 0.8179	APN: 0.0000 AD: 0.0810 ADM: 0.0413 FOM: 0.0717	Conn: 10.1587 Dunn: 0.0069 SW: 0.6945	APN: 0.0000 AD: 0.0716 ADM: 0.0418 FOM: 0.0718	Conn: 10.4147 Dunn: 0.0079 SW: 0.6924	APN: 0.0000 AD: 0.0669 ADM: 0.0427 FOM: 0.0718
Alkphos Alkaline Phosphotase	Conn: 2.3290 Dunn: 0.0034 SW: 0.7008	APN: 0.0000 AD: 0.0752 ADM: 0.0341 FOM: 0.0594	Conn: 5.2746 Dunn: 0.0006 SW: 0.5734	APN: 0.0000 AD: 0.0672 ADM: 0.0327 FOM: 0.0594	Conn: 14.0056 Dunn: 0.0006 SW: 0.4650	APN: 0.0000 AD: 0.0648 ADM: 0.0336 FOM: 0.0595
Sgpt Alamine Aminotransferase	Conn: 5.2579 Dunn: 0.0005 SW: 0.7301	APN: 0.0000 AD: 0.0413 ADM: 0.0192 FOM: 0.0459	Conn: 5.6563 Dunn: 0.0006 SW: 0.5551	APN: 0.0000 AD: 0.0361 ADM: 0.0172 FOM: 0.0460	Conn: 10.1750 Dunn: 0.0007 SW: 0.5492	APN: 0.0000 AD: 0.0329 ADM: 0.0190 FOM: 0.0460
Sgot Aspartate Aminotransferase	Conn: 3.3377 Dunn: 0.0008 SW: 0.7322	APN: 0.0000 AD: 0.0240 ADM: 0.0112 FOM: 0.0294	Conn: 13.2794 Dunn: 0.0002 SW: 0.6140	APN: 0.0000 AD: 0.0212 ADM: 0.0105 FOM: 0.0294	Conn: 15.3611 Dunn: 0.0002 SW: 0.5464	APN: 0.0000 AD: 0.0201 ADM: 0.0113 FOM: 0.0294
Total Protiens (TP)	Conn: 0.0000 Dunn: 0.0270 SW: 0.5769	APN: 0.0000 AD: 0.1384 ADM: 0.0631 FOM: 0.0787	Conn: 0.0000 Dunn: 0.0294 SW: 0.5088	APN: 0.0000 AD: 0.1258 ADM: 0.0604 FOM: 0.0788	Conn: 0.0000 Dunn: 0.0345 SW: 0.5438	APN: 0.0000 AD: 0.1156 ADM: 0.0630 FOM: 0.0789
Albumin (ALB)	Conn: 0.0000 Dunn: 0.0435 SW: 0.5773	APN: 0.0000 AD: 0.1534 ADM: 0.0699 FOM: 0.0865	Conn: 0.0000 Dunn: 0.0526 SW: 0.5715	APN: 0.0000 AD: 0.1353 ADM: 0.0636 FOM: 0.0866	Conn: 0.0000 Dunn: 0.0588 SW: 0.5296	APN: 0.0000 AD: 0.1276 ADM: 0.0704 FOM: 0.0867
Albumin and Globulin Ratio (A/G ratio)	Conn: 0.6722 Dunn: 0.0106 SW: 0.5313	APN: 0.0000 AD: 0.1104 ADM: 0.0476 FOM: 0.0638	Conn: 0.5000 Dunn: 0.0244 SW: 0.5637	APN: 0.0000 AD: 0.0966 ADM: 0.0455 FOM: 0.0638	Conn: 4.6012 Dunn: 0.0061 SW: 0.5444	APN: 0.0000 AD: 0.0906 ADM: 0.0476 FOM: 0.0639

Table 5. Performance of the proposed classification method using weighted rule bases

	Interestingness>0.7 #Rules:126	Comprehensibility>0.7 #Rules: 69	Reliability>0.7 #Rules: 121	Coverage>0.7 #Rules: 148	Confidence>0.7 #Rules: 133
Precision	0.9184	0.8405	0.9182	0.8147	0.9180
Specificity	0.9281	0.8517	0.9281	0.8184	0.9281
Sensitivity	0.9736	0.9375	0.9712	0.9615	0.9688
Accuracy	0.9487	0.8907	0.9477	0.8833	0.9466

results revealed that SVM with Accuracy of 71.36% had the best performance for the whole of database and RF reached to accuracy of 71.87% after feature selection (4).

SVM has been used to classify two data sets available in UCI repository consisting of ILPD and BUPA by Hashem et al. In this paper, features have been ranked. The classification results have been evaluated based on different sets of most ranked features. MATLAB has been used to implement SVM and feature ranking algorithm. Applying SVM to 4, 6 and 8 most significant features of ILPD showed that this algorithm yielded better results for 8 (6) first features, with an error rate of 26.8 (27) percent, sensitivity of 90 (96.6%), Prevalence 71 (71%), accuracy 73.2 (73%) and specificity 30 (12%) respectively (36).

Liang et al. have proposed a combination of GA and artificial immune to diagnose liver disease. Two data sets (ILPD and Liver Disorder) from UCI repository and 20-flod cross-validation have been used to evaluate the proposed method. Accuracy, sensitivity, specificity, precision and F-measure measured as 98.1%, 98.9%, 96%, 98.5% and 98.7% respectively. The results showed that the proposed method for ILPD obtained higher accuracy than C4.5 and Bayes methods (37).

Vijayarani et al. have used two classification algorithms, SVM and Naïve Bayes, to predict liver disease in ILPD. Two classifiers were implemented using MATLAB and compared based on precision, F-score and execution time. Results indicated that although SVM yielded precision of 76.6% and F-Score of 33.1% was better than Naïve Bayes but its execution time (3210.00 ms) was twice in comparison with Bayes (1670.00 ms) (10).

Ramana et al. have used two data sets, BUPA and ILPD, for evaluation of algorithms that has been implemented using WEKA. First significant features were selected by 4 different feature selection algorithms including Principle Component Analysis (PCA), Correlation-based Feature Selection (CFS), random projection and random subset. Then a number of 10 algorithms from 5 different categories of classification algorithms including tree-, statistical-, MLP-, rule-based and lazy learners were considered as liver disease

[DOI: 10.14196/mjiri.32.85]

prediction models. Results showed that the combination of K-Star method with CFS feature selection algorithm had the best accuracy 73.07% in terms of predicting liver disease (38).

Kiruba et al. have trained a set of 22 classification algorithms using a data set consists of 900 records which has been obtained from merging of two data sets of liver patients known as BUPA and ILPD. After training, performance of classifiers was tested on two mentioned data sets separately. Results showed that the classification accuracy of random tree and C4.5 were 100%, while C4.5 had lower execution time than random tree (39).

Tiwari et al. have examined the performance of ANN based classification algorithms. For this purpose, ILPD data set divided into two groups of men and women and people younger than 18 years were excluded. Then, significant features of two subsets were extracted using univariate analysis of variance and CFS. The performance of 4 ANN-based classification algorithms including SVM, self-organization map and Radial Basis Function (RBF) were compared based on the 5 classification quality factors including accuracy, mean absolute error, RMSE, relative absolute error and root relative squared error. They concluded that SVM outperformed other techniques. Results showed that accuracy of SVM was equal to 99.76% and 97.7% for men and women data sets respectively with a low error rate (2).

Sarojini has addressed reducing data dimension by excluding unimportant features and improving the performance of classification algorithms at the same time. First most significant attributes of ILPD were selected using wrapper based feature subset selection approach. Then the proposed classification algorithm was implemented before and after removing unimportant features. Results showed that the proposed method caused to reduce data dimension by 70% and increase classification accuracy from 66.038 to 73.413 (~ 7%) (40).

According to studies done on ILPD, it is revealed that most of them used supervised classification methods for prediction, while all considered as black box except decision trees. Moreover, several studies (2, 4, 36, 38, 40) applied feature selection algorithms and classified a subset of important features. Selecting features, caused to not consider all relationships between data, while in many cases the purpose of the researchers, was gaining a clear insight of predictive model and hidden associations between attributes, in addition to obtaining high accuracy in predicting. For this reason, despite the fact that some previous approaches (37,39) have achieved higher accuracy than the proposed approach, in this study fuzzy association rule-based classifier was used for predicting liver disease. Of course, it should be noted that the proposed method outperformed 31 from 34 methods applied in previous studies and this means that this model despite good performance in predicting, is also understandable for humans.

In addition, this research using fuzzy sets to handle the effect of uncertainty which has been considered only in Sarojini's work (40), however, their method was not based on rules. As a result, it did not provide an understandable model for humans. Moreover, in this paper the number and parameters of fuzzy membership functions were obtained using FCM (i.e. this method constructs a data-fitted prediction model without the need for expert knowledge).

Weighting of rules has not been addressed in previous studies, while in this study, 5 QMs were conducted. Also QMs determined to ensure that the proposed model not only requires no expert knowledge but also has the best fit to data set.

In the evaluation step, it became clear that among the QMs intended interestingness, reliability and confidence outperformed respectively and precision, sensitivity, specificity and accuracy are over 90%. According to the results of weighting with comprehensibility and coverage measures, it is found that the majority of rules belonging to the class of non-liver patient had less support, therefore less weight assigned to them. For this reason, (FN/ TP) was less than (FP/TN), thus the sensitivity was more than specificity.

Conclusion

In this paper, a classification method was developed to predict liver disease which in addition to high classification accuracy, it was created without expert knowledge and provided an understandable explanation of data. This method is convenient and efficient specially when there is no access to experts. Future works may be applying this method on the other data sets or using different methods for pruning the rule base in order to make a more understandable description of data set.

Conflict of Interests

The authors declare that they have no competing interests.

References

1. Brock G, Pihur V, Datta S, Datta S. clValid, an R package for cluster validation. J Stat Softw. 2011.
2. Ramana BV, Babu MSP, Venkateswarlu N. A critical study of selected classification algorithms for liver disease diagnosis. Int J Database Manag Syst. 2011;3(2):101-14.
3. Tiwari AK, Sharma LK, Krishna GR. Comparative Study of Artificial Neural Network based Classification for Liver Patient. J Inform Eng Appl. 2013;3(4):2225-0506.
4. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform. 2008;77(2):81-97.
5. Gulia A, Vohra R, Rani P. Liver patient classification using intelligent techniques. Int J Comput Sci Inform Tech. 2014;5(4):5110-5.
6. Anunciaçao O, Gomes BC, Vinga S, Gaspar J, Oliveira AL, Rueff J. A data mining approach for the detection of high-risk breast cancer groups. Advances in Bioinformatics: Adv Intel Soft Compu. 2010;74:43-51.
7. Pradhan M, Sahu RK. Predict the onset of diabetes disease using Artificial Neural Network (ANN). Int J Comput Sci Emerg Tech. 2011;2(2).
8. Suneetha N, Hari V, Kumar VS. Modified gini index classification: a case study of heart disease dataset. Int J Comput Sci Eng. 2010;2(6):1959-65.
9. Bahramirad S, Mustapha A, Eshraghi M, editors. Classification of liver disease diagnosis: A comparative study. Informatics and Applications (ICIA), 2013 Second International Conference on; 2013: IEEE.
10. Dhamodharan S, editor Liver Disease Prediction Using Bayesian Classification2014: Special Issues, 4th National Conference on Advance Computing, Application Technologies.

[DOI: 10.14196/mjiri.32.85]

11. Vijayarani S, Dhayanand S. Liver disease prediction using SVM and Naïve Bayes algorithms. Int J Sci Eng Technol Res(IJSETR). 2015;4(4).
12. Roslina A, Noraziah A, editors. Prediction of hepatitis prognosis using Support Vector Machines and Wrapper Method. Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on; 2010: IEEE.
13. Soliman OS, Elhamd EA. Classification of Hepatitis C Virus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine. Int J Sci Eng Res. 2014;5(3):122.
14. Fielding A. Cluster and classification techniques for the biosciences: Cambridge University Press Cambridge; 2007.
15. Wu D, Mendel JM. Linguistic summarization using IF–THEN rules and interval type-2 fuzzy sets. IEEE Trans Fuzzy Syst. 2011;19(1):136-51.
16. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. Clin Pharmacol Ther. 2012;91(6):1010-21.
17. Xiao F, Fan C. Data mining in building automation system for improving building operational performance. Energ Buildings. 2014;75:109-18.
18. Chen Y-L, Hung LT-H. Using decision trees to summarize associative classification rules. Expert Syst Appl. 2009;36(2):2338-51.
19. Baralis E, Garza P. I-prune: Item selection for associative classification. Int J Intell Syst. 2012;27(3):279-99.
20. Hong T-P, Lee Y-C. An overview of mining fuzzy association rules. Fuzzy Sets and Their Extensions: Representation, Aggregation and Models. Berlin, Heidelberg: Springer; 2008. p. 397-410.
21. Antonelli M, Ducange P, Marcelloni F, Segatori A. A novel associative classification model based on a fuzzy frequent pattern mining algorithm. Expert Syst Appl. 2015;42(4):2086-97.
22. Papageorgiou EI. A new methodology for decisions in medical informatics using fuzzy cognitive maps based on fuzzy rule-extraction techniques. Appl Soft Comput. 2011;11(1):500-13.
23. Liu X, Feng X, Pedrycz W. Extraction of fuzzy rules from fuzzy decision trees: An axiomatic fuzzy sets (AFS) approach. Data Knowl Eng. 2013;84:1-25.
24. Ishibuchi H, Yamamoto T. Rule weight specification in fuzzy rule-based classification systems. IEEE Trans Fuzzy Syst. 2005;13(4):428-35.
25. Anooj P. Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. J King Saud Univ Sci. 2012;24(1):27-40.
26. Beiranvand V, Mobasher-Kashani M, Bakar AA. Multi-objective PSO algorithm for mining numerical association rules without a priori discretization. Expert Syst Appl. 2014;41(9):4259-73.
27. Huang M-J, Tsou Y-L, Lee S-C. Integrating fuzzy data mining and fuzzy artificial neural networks for discovering implicit knowledge. Knowl Based Syst. 2006;19(6):396-403.
28. Salleb-Aouissi A, Vrain C, Nortet C, editors. QuantMiner: A Genetic Algorithm for Mining Quantitative Association Rules. IJCAI'07 Proceedings of the 20th international joint conference on Artifical intelligence 2007; Hyderabad, India.
29. Martin D, Rosete A, Alcala-Fdez J, Herrera F. A new multiobjective evolutionary algorithm for mining a reduced set of interesting positive and negative quantitative association rules. IEEE Trans Evol Comput. 2014;18(1):54-69.
30. Kuo RJ, Chao CM, Chiu Y. Application of particle swarm optimization to association rule mining. Appl Soft Comput. 2011;11(1):326-36.
31. Juang CF, Chang PH. Designing Fuzzy-Rule-Based Systems Using Continuous Ant-Colony Optimization. IEEE Trans Fuzzy Syst. 2010;18(1):138-49.
32. Sowan BI. Enhancing Fuzzy Associative Rule Mining Approaches for Improving Prediction Accuracy. Integration of Fuzzy Clustering, Apriori and Multiple Support Approaches to Develop an Associative Classification Rule Base: University of Bradford; 2012.
33. Sowan B, Dahal K, Hossain MA, Zhang L, Spencer L. Fuzzy association rule mining approaches for enhancing prediction performance. Expert Syst Appl. 2013;40(17):6928-37.
34. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. J Sigmod Rec. 1993;22(2):207-16.
35. Jin H, Kim S, Kim J. Decision factors on effective liver patient data prediction. Int J BioSci BioTechnol. 2014;6(4):167-78.
36. Hashem EM, Mabrouk MS. A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis. Am J Intell Syst. 2014;4(1):9-14.
37. Liang C, Peng L. An automated diagnosis system of liver disease using artificial immune and genetic algorithms. J Med Syst. 2013;37(2):1-10.
38. Ramana B, Babu M, Venkateswarlu N. Liver classification using modified rotation forest. Int J Eng Res Develop. 2012;1(6):17-24.
39. Kiruba HR, Arasu GT. An intelligent agent-based framework for liver disorder diagnosis using artificial intelligence techniques. J Theor Appl Inf Technol. 2014;69(1).
40. Sarojini B. A Wrapper Based Feature Subset Evaluation Using Fuzzy Rough K-NN. Int J Eng Tech. 2013;5(6):4672-6.

[DOI: 10.14196/mjiri.32.85]