Penerbit Akademia Baru

Journal of Advanced Research Design

Journal homepage: www.akademiabaru.com/ard.html
ISSN: 2289-7984

Liver Fibrosis Diagnosis with Mamdani FIS

Sara Sweidan[1,*], Shaker El-Sappagh[2], Hazem Elbakry[1], S. Sabbeh[2]

[1] Department of Information System, Faculty of computers & information, Mansoura University,33516 Mansoura, Egypt
[2] Department of Information System, Faculty of computers & informatics, Benha University, Benha, Egypt

ARTICLE INFO

Article history:
Received 5 July 2017
Received in revised form 10 October 2017
Accepted 14 December 2017
Available online 21 March 2018

Keywords:
Liver fibrosis, fuzzy system, clinical decision support, inference system

ABSTRACT

Nowadays, clinical decision support system become a part of daily life. Accurate diagnosis of liver cirrhosis helps in avoiding medical problems which may lead to death. The aim of the study is to build a fuzzy expert system for the diagnosis of liver fibrosis-stage (DLFS). The system uses machine learning tools and data mining statics to discover fuzzy rules, which help physicians to provide a fast and accurate diagnosis. The experimental have been performed on real dataset from clinical data sheets for 119 patients infected by chronic HCV. The evaluation results showed that the system identify liver fibrosis-stage with high degree of accuracy 95.7% and may decrease the need for liver biopsy.

1. Introduction

Viral hepatitis C (HCV) currently infects nearly 2% of the world's population. In Egypt the situation is very critical. In all other countries, the prevalence is between 1% to 3%. Egypt has the highest prevalence of HCV in the world reaching 14.7 % of the population [1]. There are viral logical features of HCV used as indicators to predicate staging of histological liver damage such as serum ALT/AST levels, direct and total serum bilirubin, albumin, platelet count, and INR. Non-invasive methods utilize serum markers to detect the degree of fibrosis into five stages: f_0: negative fibrosis, f_1: mild fibrosis, f_2: significant fibrosis, f_3: cirrhosis, and f_4:significant cirrhosis[2]. In developing countries as Egypt, the availability of medical experts to cover the large number of patients cannot be achieved. Automated systems such as clinical decision support systems (CDSSs) can help to overcome this issue [2]. The goal of this work is to build a CDSS to help physicians to estimate the liver fibrosis of HCV patients. Our methodology uses a combination of some techniques including a set of machine learning techniques to prepare the medical data set, fuzzy decision tree for fuzzy rules generation, and Mamdani fuzzy inference system (FIS). The proposed system will be tested by a data set of 119 patients infected by HCV. Demographics, standard serum markers, and other conditions are utilized as indicators to predict the fibrosis stage.

[*] *Corresponding author.*
E-mail address: sweera20@yahoo.com (S. Sweidan)

The number of health applications with CDSS has been increased during the last few years. Accurate diagnosis is one of the most important problem of medicine. The relationship between diagnosis and clinical treatment protocols is effected in healthcare as we have been explained in survey [3]. In [4], authors proposed a methodology to avoid the risk of liver biopsy for cirrhosis patients. Sartakhti *et al.*, [5] proposed a methodology to solve the hepatitis diagnosis problem by hybrid model of support vector machine (SVM) and simulated annealing (SA) with accuracy 96.25%. In [6], researchers used Adaptive Neuro Fuzzy Inference System (ANFIS) to diagnose liver disease to reach better and accurate diagnosis with 83% accuracy. In [7] authors presented a machine-learning model based on fuzzy rule based reasoning to estimate diabetes retinopathy risks. The system achieved 80% classification accuracy. In this study, we build DFLS system. It is a CDSS to predicate the liver fibrosis stage of HCV patients. We have a labeled training data set of patients from Liver Institute, Mansoura University, Egypt. The utilized dataset, generated rules, and used fuzzy linguistic variables and fuzzy sets are all build based mainly on the domain expert knowledge and most recent clinical practice guidelines. The fuzzy knowledge base is generated by a trained decision tree on a preprocessed and high quality data set. The proposed DFLS is based on Mamdani inference system in order to diagnose new cases.

2. The Proposed DLFS Framework

This section discusses the proposed DLFS system in details. It is based on fuzzy decision tree FDT [7], and it uses fuzzy inference system FIS to predict patients states [8]. Figure 1 illustrates the components of the proposed system and their relationships. It presents inputs and output of DLFS as raw patient description data and the fibrosis-stage prediction, respectively.

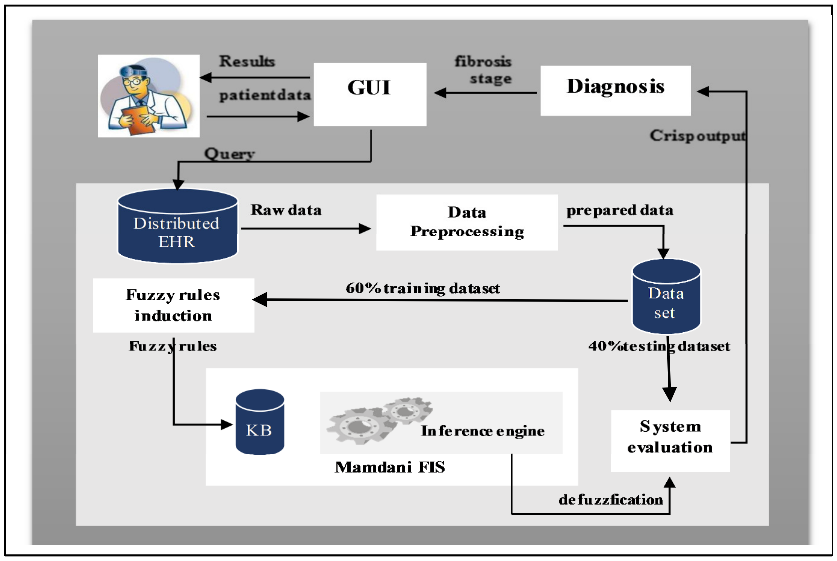

Fig. 1. The proposed DLFS framework

2.1 Problem Description

Liver fibrosis is one of the leading causes of mortality because it changes the architecture of certain organs and disrupts normal function. Liver fibrosis is resulting from chronic liver diseases such as viral hepatitis [9]. The degree of fibrosis divides into five stages: f_0: negative fibrosis, f_1: mild fibrosis, f_2: significant fibrosis, f_3: cirrhosis, and f_4:significant cirrhosis[2]. Poor diagnosis will lead to significant liver fibrosis, cirrhosis or to the liver end-stage failure subsequently lead to premature death. In this study, we develop knowledge based system to assist physicians in liver fibrosis-stage prediction.

The dataset for our present work is obtained from Liver Institute, Mansoura University, Egypt. Data are extracted from clinical data sheets of 119 patients all infected by chronic HCV. Data concern the analysis of patient's demographics data (age, gender, address), laboratory tests (ALT, SB, AST, PLT, WBC), and symptoms (diarrhea, fatigue, jaundice, dyspnea). The studied dataset comprises $n = 27$ features that are relevant to HCV disease. They are selected according to domain expert. Almost participated patients are in the range of 16 and 65 years and distributed in 80 male cases (67.5%) and 39 female cases (32.5%). This dataset includes the clinical features considered to be the most significant to the clinical decision process. The patients have been categorized according to liver biopsy test as 32 absent fibrosis patients, 30 Mild fibrosis patients, 29 significant fibrosis patients, and 28 cirrhosis patients. Liver biopsy is an accurate procedure but it has limitations include internal bleeding, pain, costs, and delay results. Our dataset is a heterogeneous dataset and need to preprocess before training.

2.2 The Data Preprocessing

The data quality has a large implication for the quality of diagnosis results. Preprocessing of data set is necessary to remove problems associated with medical data like redundant, noise, and unnecessary variables.

2.2.1 Anonymization, UoM, and normalization

Anonymization is the process of removing any data that can identify specific patient. All personal data as name, address, ID, etc. have been removed for privacy purpose. We have selected a unified Unit of Measurement (UoM) for each lab test.

2.2.2 Handling missing data

When dealing with real life medical data, missing or unknown values are unavoidable [10]. We have not patient cases with missing feature values more than or equal to 50%. All features with more than or equal 25% were dropped down. Our dataset contains a set of 27 features, we drop down 5 features: {Residence, occupation, indigestion, vision problem, skin pigmentation}. The overall missing data decreased from 13% to 3.6% of the whole dataset. The complete cases arise from 9 cases to 58 complete cases fields or 48% of the data set and 52% missing cases. Hot deck imputation is a method for handling missing data [11].

2.2.3 Feature selection

Feature selection is frequently used as a data preprocessing step. Its goal is to improve the prediction accuracy by filtering the relevant features and ignore non relevant ones without decreasing the final result [12]. In our study, the input features are related to HCV medical domain, and The target output is Y_f , where $Y_f = \{f_0, f_1, f_2, f_3\}$, these features were determined by medical domain expert . We need to select the significant relevant features to the target class from each cluster to construct the final feature subset. We use different feature selection algorithms; the best performing algorithm is chi-square. So, we drop down features with weights equal zero (PCR, SA, Serum ferritin, Hemoglobin) as a result of the feature selected technique. The ratio of irrelevant features is 20% from all features. Finally, we have preprocessed dataset with subset features $n = 17$, The dataset divides into two datasets, 60% of the dataset has been used in the training, the otherwise 40% are used as test data in the later evaluation step, the selection data is a random manner.

2.3 Fuzzy Rules Induction

Fuzzy knowledge is defined in the form of fuzzy rules (IF condition THEN conclusions) to express the relationships between fuzzy parameters. Our method is to utilize a machine-learning algorithm to induce a rule set based on input/output data. We used fuzzy decision tree FDT as a rule induction algorithm [13]. We implement fuzzy decision tree with gain ratio technique for splitting nodes in order to generate the prediction rules Eq. 1

$$I^N = -\sum_{v_c \in c} \frac{P_{v_c}^N}{P^N} \log(\frac{P_{v_c}^N}{P^N}) \qquad (1)$$

where P^N and I^N denote the total example count and information measure for node N (Entropy technique) respectively, and v_c is the fuzzy attribute to split N. The system generates set of 74 fuzzy rules from training dataset. These rules have been stored in the DFLS as its knowledge base (KB). Table 1 presents sample of these rules. Each fuzzy set is presented by an appropriate notation will discussed later.

Table 1
Sample of generated fuzzy rules

Rule #	Age	Gender	PLT	WBC	ALT	AST	SB	Ascites	Spleen	Portal-V	Lesion	Appetite	Dyspnea	Diarrhea	fatigue	Vomiting	Jaundice	Fibro Stage
	$if\ A^1\ is\ a_i^1\ and\ A^2\ is\ a_i^2\ and\ \dots and\ A^n\ is\ a_i^n\ then\ C\ is\ c_i$																	
$\boldsymbol{R_1}$	O	-	L	L	VH	-	L	-	-	NO	-	-	B	B	-	A	A	$\boldsymbol{F_1}$
$\boldsymbol{R_2}$	-	M	-	L	-	VH	H	M	N	-	NO	B	-	-	-	-	-	$\boldsymbol{F_2}$
$\boldsymbol{R_3}$	O	M	VL	L	VH	VH	H	NO	N	-	-	-	-	-	-	-	-	$\boldsymbol{F_3}$
$\boldsymbol{R_4}$	A	M	-	L	-	VH	H	NO	N	NO	NO	-	-	-	-	B	-	$\boldsymbol{F_3}$
$\boldsymbol{R_5}$	A	M	-	L	-	VH	H	NO	N	NO	NO	-	-	R	-	A	-	$\boldsymbol{F_2}$

2.4 Fuzzy Sets for Input and Output Variables

According to fuzzy set theory [14], we used trapezoidal and triangle functions to define membership functions of input and output variables as shown in table 2 and table 3. And graphical representation of membership functions is shown in fig. 2 and fig.3, respectively. For output attribute X_O is fuzzified into linguistic variables $C_k = \{c_{k,1}, \dots, c_{k,i}, \dots., c_{K,Ik}\}$. As shown in table 3, the target output fuzzy sets with its parameters are identified by domain expert according to APRI index scale [2].

Table 2
Sample of fuzzy sets description for input variables

Variable	Fuzzy set	Notation	Fit vector	Variable	Fuzzy set	Notation	Fit vector
PLT	V-low	VL	(50, 128, 175)	AST	Low	L	(0, 10, 18)
Range:	Low	L	(140,190, 240)	Range:	Normal	N	(10, 20, 30)
(50- 500)	Normal	N	(190, 240, 340, 390)	(0- 150)	High	H	(23, 33, 43)
ID: x_4				ID: x_2			
	High	H	(340, 390, 440)		V-high	VH	(35, 45, 140, 150)
WBC	V-low	VL	(0, 2, 4)	SB	Low	L	(0, 0.5, 0.8)
Range:	Low	L	(2,4, 6, 8)	Range:	Normal	N	(0.6, 0.8, 1.5, 1.8)
(0- 20)				(0- 4)			
ID: x_5	Normal	N	(6, 8, 10, 12)	ID: x_3	High	H	(1.5, 1.8, 3.5, 3.9)
	High	H	(10, 12, 14, 16)	symptoms	Absent	A	(1, 1, 1)
ALT	Low	L	(0, 11, 20)	Range:	Rare	R	(2, 2, 2)
Range:	Normal	N	(11, 20, 36, 45)	(0- 4)	Bad	B	(3, 3, 3)
(0- 150)	High	H	(34, 40, 50, 56)	ID: x_8,x_9,			
ID: x_1	V-high	VH	(50, 60, 140, 150)	x_{10}			

Table 3
Fuzzy set description for target variable

Variable	Fuzzy set	Notation	Fit vector
Fibro stage	Absent	F_0	(0, 0.23, 0.64)
Range: (0- 5)	Mild	F_1	(0.62, 1, 1.5)
Code: x_{18}	Significant	F_2	(1.5, 2.19, 2.9)
	cirrhosis	F_3	(2, 2.6, 5)

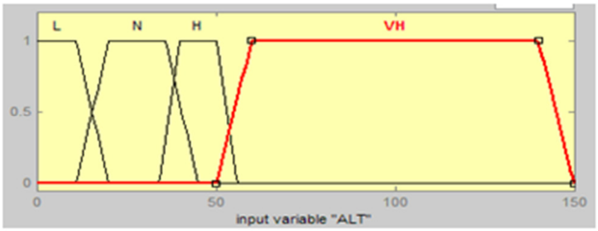

Fig. 2. MF for ALT input variable

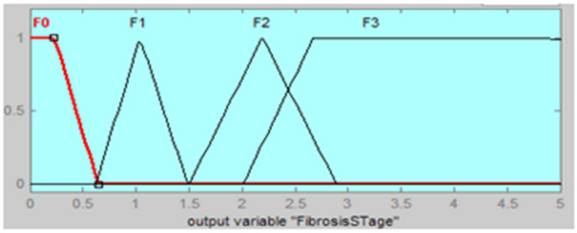

Fig. 3. MF for output variable

Finally, we have a knowledge base composed of fuzzy rules and membership functions for each variable. In FIS the fuzzy rule is inferred by input vector x using fuzzy inference engine [15].

3. Result and Discussion

The aim of this study is to solve the predication problem of liver fibrosis stage for HCV patients using a set of input parameters of HCV patients' dataset. To achieve this target, we implement DLFS. This is a new medical fuzzy inference system for fibrosis diagnosis. we constructed prediction model by discovering the fuzzy rules using fuzzy rule reasoning method from the experimental datasets and generalized the relationship input and output parameters ($y = f(x_1, x_2, \ldots.., x_{17})$) for accurate prediction of disease. The study is applied on real dataset to show how it can be utilized for real medical diagnosis. The paper develops DLFS using a set of steps including input fuzzification, generating membership function MF_s, extracting fuzzy rules, and output defuzzification. In the fuzzification step, trapezoidal, triangular MF_s are used to determine the degree of inputs/output variables that they belong to each of the appropriate fuzzy sets. In $defuzzification$ step, we used the center of gravity (COG) which calculates the center of area under the curve[16]. In our proposed system, we implement the $Mamdani$ fuzzy inference method through fuzzy logic toolbox provided in $Matlab\ R2012a$ software, based on fuzzy rule based was generated by FDT in the evaluation process. We developed the fuzzy rule-based system for predicting the class of diagnosis and considered appropriate MF_s for the inputs fuzzification and output defuzzification in FIS. We have four output classes, and totally 74 fuzzy rules are used for the prediction model. As shown in figure 1, we use 40% of the dataset in the evaluation process. The test dataset is made of 47 data samples, demonstrated that patients had all stages of fibrosis. Each test case is described by 17 patient features. All of the features used separately to build one dimensional system. The manner of data selection was random selection. The rule base is inferred by the crisp inputs to get the final crisp diagnosis.

3.1 Performance

In this section, we evaluate the performance “goodness” of the proposed fuzzy system by both classification error index (CE) Eq. 2, and squared classification error index (SCE) Eq. 3 [17],

$$CE = \frac{1}{N} * \sum_{i=1}^{N} y_i \qquad (2)$$

$$SCE = \frac{1}{N} * \frac{1}{K} * \sum_{i=1}^{N} \sum_{k=1}^{K} (\alpha_i^k - \delta_i^k)^2 \qquad (3)$$

SCE can be used to evaluate the classification uncertainly calculated by the Eq. 3 and gets result 0.125, the low degree (high confidence assigned to right solutions). While CE is simply fraction of wrong classified i^{th} data sample the system performs error ratio (0.085) Eq.2. In order to measure the classification accuracy $acc = 1 - CE$, the system performs accuracy 91.4%.

3.2 Evaluation by Measured Terms

Next, the experimental results of DLFS for risk prediction have been evaluated using measured terms [18]. These terms have been implemented in in the proposed system comparing with APRI index (Noninvasive serum biomarkers) which is friendly test [2]. The test has limitations in the diagnosis the levels of significant fibrosis and cirrhosis, figure 4 represents the comparison results between DLFS and APRI medical test with accuracies 95.7% and 86%, respectively.

3.3 Evaluation by 10-fold Cross Validation

In this section, we evaluate the system by all the data set, training and test datasets. We use the 10-fold cross validation technique [19]. we calculate the average of estimated accuracy by 10-fold cross validation, the system obtained accuracy (94%1).

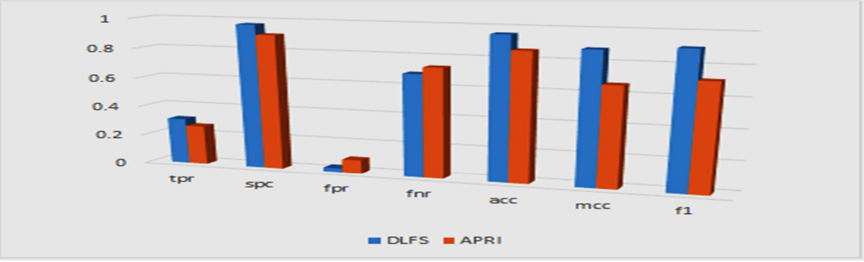

Fig. 4. Comparison results between DLFS and APRI

4. Conclusion

In this paper, we proposed a new knowledge-based system for liver fibrosis stage prediction using fuzzy reasoning technique. We used entropy to generate the fuzzy rules to be used in the knowledge based system of fuzzy rule based reasoning method for the disease classification. We have evaluated the knowledge based system on real test dataset of 119 HCV patient cases. The results of our experiments on the dataset indicated that the proposed system achieved good prediction accuracy 95.7% for liver fibrosis stage. In future studies, we will solve the same problem by adding the semantic dimension in the picture. This will be achieved by using the fuzzy ontology techniques.

References

[1] Amer, Fatma A., Maha Gohar, and Monkez Yousef. "Epidemiology of Hepatitis C Virus Infection in Egypt." (2015).

[2] Badria, Farid, and Sami Gabr. "Prediction of Liver Fibrosis and Cirrhosis Among Egyptians Using Noninvasive Index." *J. Pure & Appl Microbiol* 1, no. 1 (2007): 45-50.

[3] Sweidan, Sara, Hazem El-Bakry, Shaker El-Sappagh, Sahar Sabah, and Nikos Mastorakis. "Viral Hepatitis Diagnosis: A Survey of Artificial Intelligent Techniques." *International Journal of Biology and Biomedicine* 1, (2016): 106-115

[4] Hashem, Ahmed M., M. Emad M. Rasmy, Khaled M. Wahba, and Olfat G. Shaker. "Prediction of the degree of liver fibrosis using different pattern recognition techniques." In *Biomedical Engineering Conference (CIBEC), 2010 5th Cairo International*, pp. 210-214. IEEE, 2010.

[5] Sartakhti, Javad Salimi, Mohammad Hossein Zangooei, and Kourosh Mozafari. "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)." *Computer methods and programs in biomedicine* 108, no. 2 (2012): 570-579.

[6] Farokhzad, M., and Ebrahimi, L. “a novel adaptive neuro fuzzy inference system for the diagnosis of liver disease”, International journal of academic research in computer engineering IJARCE, 1, no.1, (2016): 61-66.

[7] Saleh, Emran, Aida Valls, Antonio Moreno, Pedro Romero-Aroca, Sofia de la Riva-Fernandez, and Ramon Sagarra-Alamo. "Diabetic retinopathy risk estimation using fuzzy rules on electronic health record data." In *Modeling Decisions for Artificial Intelligence*, pp. 263-274. Springer, Cham, 2016.

[8] Lee, Chuen-Chien. "Fuzzy logic in control systems: fuzzy logic controller. I." *IEEE Transactions on systems, man, and cybernetics* 20, no. 2 (1990): 404-418.

[9] Suk, Ki Tae, and Dong Joon Kim. "Staging of liver fibrosis or cirrhosis: The role of hepatic venous pressure gradient measurement." *World journal of hepatology* 7, no. 3 (2015): 607.

[10] Almuhaideb, Sarab, and Mohamed El Bachir Menai. "Impact of preprocessing on medical data classification." *Frontiers of Computer Science* 10, no. 6 (2016): 1082-1102.

[11] Andridge, Rebecca R., and Roderick JA Little. "A review of hot deck imputation for survey non-response." *International statistical review* 78, no. 1 (2010): 40-64.

[12] Mohsin, Mohamad Farhan Mohamad, Abdul Razak Hamdan, and Azuraliza Abu Bakar. "An evaluation of feature selection technique for dendrite cell algorithm." In *IT Convergence and Security (ICITCS), 2014 International Conference on*, pp. 1-5. IEEE, 2014.

[13] Guillaume, Serge, and Brigitte Charnomordic. "Learning interpretable fuzzy inference systems with FisPro." *Information Sciences* 181, no. 20 (2011): 4409-4427.

[14] Zadeh, Lotfi A. "Fuzzy sets." In *Fuzzy Sets, Fuzzy Logic, And Fuzzy Systems: Selected Papers by Lotfi A Zadeh*, pp. 394-432. 1996.

[15] Kaur, A., and Kaur, A. “comparison of Mamdani-type and sugeno-type fuzzy inference systems for air conditioning System”, International journal of soft computing and engineering IJSCE, 2, Issue2, (2012): 323-325.

[16] Van Broekhoven, Ester, and Bernard De Baets. "Fast and accurate center of gravity defuzzification of fuzzy system outputs defined on trapezoidal fuzzy partitions." *Fuzzy Sets and Systems* 157, no. 7 (2006): 904-918.

[17] Pota, Marco, Massimo Esposito, and Giuseppe De Pietro. "Designing rule-based fuzzy systems for classification in medicine." *Knowledge-Based Systems* 124 (2017): 105-132.

[18] Anooj, P. K. "Implementing decision tree fuzzy rules in clinical decision support system after comparing with fuzzy based and neural network based systems." In *IT Convergence and Security (ICITCS), 2013 International Conference on*, pp. 1-6. IEEE, 2013.

[19] Tsipouras, Markos G., Themis P. Exarchos, Dimitrios I. Fotiadis, Anna P. Kotsia, Konstantinos V. Vakalis, Katerina K. Naka, and Lampros K. Michalis. "Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling." *IEEE Transactions on Information Technology in Biomedicine* 12, no. 4 (2008): 447-458.