Applied Soft Computing 13 (2013) 3429–3438

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease

Yılmaz Kaya[a,*], Murat Uyar[b]

[a] Siirt University, Engineering and Architecture Faculty, Computer Engineering, 56100 Siirt, Turkey
[b] Siirt University, Engineering and Architecture Faculty, Electrical and Electronics Engineering, 56100 Siirt, Turkey

ARTICLE INFO

Article history:
Received 1 March 2012
Received in revised form 8 October 2012
Accepted 12 March 2013
Available online 3 May 2013

Keywords:
Hepatitis disease
Rough set
Dimensionality reduction
Extreme learning machine

ABSTRACT

Hepatitis is a disease which is seen at all levels of age. Hepatitis disease solely does not have a lethal effect, but the early diagnosis and treatment of hepatitis is crucial as it triggers other diseases. In this study, a new hybrid medical decision support system based on rough set (RS) and extreme learning machine (ELM) has been proposed for the diagnosis of hepatitis disease. RS-ELM consists of two stages. In the first one, redundant features have been removed from the data set through RS approach. In the second one, classification process has been implemented through ELM by using remaining features. Hepatitis data set, taken from UCI machine learning repository has been used to test the proposed hybrid model. A major part of the data set (48.3%) includes missing values. As removal of missing values from the data set leads to data loss, feature selection has been done in the first stage without deleting missing values. In the second stage, the classification process has been performed through ELM after the removal of missing values from sub-featured data sets that were reduced in different dimensions. The results showed that the highest 100.00% classification accuracy has been achieved through RS-ELM and it has been observed that RS-ELM model has been considerably successful compared to the other methods in the literature. Furthermore in this study, the most significant features have been determined for the diagnosis of the hepatitis. It is considered that proposed method is to be useful in similar medical applications.

1. Introduction

Liver is an organ which has a wide range of functions, including digestion, energy production, glycogen storage, detoxification and regulation of blood glucose. Various diseases or microorganisms such as virus, bacteria prevents liver from functioning by damaging it [1,2]. One of these viruses, hepatitis locates in the cells in the liver tissue leading to the loss of the functioning of these cells. Healthy cells with virus infected become dysfunctional. Recently, types of hepatitis disease have been widespread throughout the world [3]. Hepatitis diseases has five different types and there are termed as Hepatitis A, Hepatitis B, Hepatitis C, Hepatitis D and Hepatitis E [1,4]. The target organ is liver in all types of Hepatitis. Each type has specific symptoms and most of them are treated succesfully. Hepatitis A is generally seen in children and named as infection hepatitits. Hepatitis A is prevented generally by going through an illness or vaccination. Hepatitis B and C are considered as carriers and there are no signs or symptoms of these diseases. Hepatitis B, caused by hepatitis B virus which damages liver by attacking it, is the most common liver infection in the world. Hepatitis diseases can be infected through blood, unprotected sex, shared or reused syringes. Moreover, during pregnancy or postpartum, it can be transmitted to the infant from mother with hepatitis [2]. For early diagnosis of these diseases, a blood test is required once a year.

Besides clinic tests, machine learning and pattern recognition methods have been widely used for early diagnosis of hepatitis diseases in medicine by specialists. With the help of diagnostic systems, the possible errors in experts made in the stage of diagnosis can be decreased, and the medical data can be analysed in shorter time and more detailed as well [5]. The classical major steps for an automatic pattern recognition system are feature extraction and classification. Feature extraction is one of the most considerable steps in the area of pattern recognition because it can directly influence the result of the diagnosis system. Thus, there is a need to extract the most significant features from hepatitis dataset for the diagnosis of the hepatitis diseases. The feature reduction from hepatitis database can be implemented by using stochastic techniques, such as genetic algorithm (GA) [6] and simulated annealing (SA) [7] or statistical techniques, such as linear discriminant analysis (LDA) [8], principal component analysis (PCA) [2,9], the fisher discriminant analysis (FDA) [10] and the local fisher discriminant analysis (LFDA) [5]. In the classification step, feature vectors that

* Corresponding author. Tel.: +90 506488 49 90; fax: +90 484 223 20 43.
E-mail addresses: yilmazkaya1977@gmail.com (Y. Kaya), muratuyar1@gmail.com (M. Uyar).

1568-4946/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2013.03.008

are obtained from the feature reduction process is applied as input to a classifier algorithm, such as artificial neural network (ANN) [11,12], artificial immune system (AIS) [1,9], probabilistic neural network (PNN) [13], support vector machine (SVM) [5,7] and fuzzy inference system [8]. The machine learning methods widely used and highly successful for the diagnosis of hepatitis diseases are to be discussed in the next section.

In this study, a new hybrid approach based on rough set (RS) and extreme learning machine (ELM) has been proposed for diagnosis of hepatitis diseases. The main objectives of this study are (1) to investigate the feasibility of RS to extract significant information from attributes in hepatitis data set and reduce its data size, (2) to make feature selection without removing missing values in the data set through RS, (3) to improve the classification accuracy for hepatitis disease diagnosis. In the first stage of the model consisted of two stages, the sub-features that best represents data are obtained by RS. The main contribution of RS is to execute feature reduction inspite of missing values. In the next stage, classification process has been done through ELM classifier by using the reducted feature sets. ELM as a relatively new learning algorithm for single hidden-layer feedforward networks (SLFNs) was first introduced by Huang et al. [14]. There are some advantages of the ELM algorithm: (1) it is extremely fast, (2) it has better generalization performance, (3) it tends to reach the solutions straightforward without trivial issues such as local minima, learning rate, momentum rate and over-fitting encountered in traditional gradient based learning algorithm [15].

In order to test the effectiveness of the proposed hybrid model, hepatitis data set, taken from UCI machine learning repository, has been used. An important part of this data set (48.3%) consists of missing values. Removal of the missing values from the data set may lead to the data loss both during feature reduction and classification process. Within a major part of studies in the literature, this data set has been subjected to classification process after missing values were removed. In this study, feature selection through RS has been carried out without removing missing values in the data set. RS-ELM hybrid model was tested for various training-test rates and finally, when training and test data sets were selected respectively at the rates of 80% and 20%, classification success of 100.00% was achieved. The experimental results show that the proposed RS-ELM can effectively improve the classification performance. It has also shown that RS-ELM outperforms the other methods and has achieved the best predicative classification accuracy with the reduced feature subset. As a result, the proposed hybrid model can be considered as helpful tool for the specialist in making a decision on diagnosing hepatitis diseases.

The content of this study was organized as follows. In the next section, other studies performed by using hepatitis data set have been summarized. In Section 3, acquisition and introduction of data set have been done. In Section 4, theoretical information about RS and ELM has been given. In Section 5, experimental results have been presented. In the last section, results of this study have been discussed.

2. Related studies for diagnose of hepatitis diseases

In this section, the proposed methods for the diagnosis of hepatitis disease have been briefly reviewed.

It has been reported that Ster and Dobnikar obtained classification successes of 86.4%, 85.3%, and 83.2% by using LDA and FDA methods, respectively [10]. Polat and Güneş achieved a success rate of 94.14% in their studies by using AIS and PCA [9]. A classification success rate of 94.16% was achieved in the study of Doğantekin et al. by using hybrid model based on LDA and ANFIS [8]. Çalışır and Doğantekin achieved a classification success rate of 95.00% by using PCA and least square SVM (LS-SVM) [2]. Javad et al. performed a hybrid model based on SVM and SA and they reported classification success rate of 96.25% [7]. Chen et al. demonstrated that they have achieved a success rate of 96.77% by using LFDA and SVM (LFDA-SVM) [5].

In this study, RS and ELM based a new hybrid model is proposed for diagnosis hepatitis diseases. It was observed that RS-ELM achieved the best classification accuracies (10,000% for 80–20% training – testing partition) for a reduced feature subset that included four features.

Table 1
Details of attributes in hepatitis database.

Features	Domain value	Missing percent %
Age	10, 20, 30, 40, 50, 60, 70, 80	0.00
Sex	Male, female	0.00
Steroid	Yes, No	1.00
Antivirals	Yes, No	0.00
Fatigue	Yes, No	1.00
Malaise	Yes, No	1.00
Anorexia	Yes, No	1.00
Liver Big	Yes, No	6.00
Liver Firm	Yes, No	7.00
Spleen Palpable	Yes, No	3.00
Spiders	Yes, No	3.00
Ascites	Yes, No	3.00
Varices	Yes, No	3.00
Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00	4.00
Alk Phosphate	33, 80, 120, 160, 200, 250	19.00
Sgot.	13, 100, 200, 300, 400, 500	3.00
Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0	10.00
Protime	10, 20, 30, 40, 50, 60, 70, 80, 90	43.00
Histology	Yes, No	0.00
Class	Die, Alive	0.00

3. Data set

Data set related to hepatitis diseases, used in this study, has been taken from UCI machine learning repository [16]. Data set is a set that identifies whether patients suffering hepatitis are alive or not. Data set includes 155 samples and 19 features. Decision feature has been coded as 1 for those alive and 0 for those die. Decision feature consists two classes in which there are 32 (20.6%) dies and the rest 123 (79.4%) is alive. Approximately 48.30% of the data set includes missing value. Features available in data set have been shown in Table 1.

4. Methodology

4.1. Roughs set theory

RS, defined by Pawlak et al. [17], is a mathematical approach used for various purposes such as feature selection, feature extraction, feature reduction and extraction of decision rules in data, especially in the case of uncertain and incomplete data [18,19].

This section presents the basic definitions that are required to understand the application of RS into feature reduction for hepatitis disease classification problems.

4.1.1. Decision table

In RS, the data is collected in a table, called decision table. A decision table is denoted as bellow:

$$S = (U, A, C, D) \quad (1)$$

where $U = \{x_1, x_2, \ldots, x_n\}$ is a finite set of cases (the universe). $A = \{a_1, a_2, \ldots, a_m\}$ is a set of attributes, and C,$D \subset A$ are two subsets of features or attributes that are called condition (C) and decision (D) attributes, respectively. Thus, a decision table specifies the