Informatics in Medicine Unlocked 2 (2016) 1–11

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

Hybrid approach using fuzzy sets and extreme learning machine for classifying clinical datasets

Kindie Biredagn Nahato [a], Khanna H. Nehemiah [a,*], A. Kannan [b]

[a] Ramanujan Computing Centre, Anna University, Chennai 600025, India
[b] Department of Information Science and Technology, Anna University, Chennai 600025, India

ARTICLE INFO

Article history:
Received 31 October 2015
Received in revised form
18 December 2015
Accepted 9 January 2016
Available online 15 February 2016

Keywords:
Extreme learning machine
Fuzzification
Fuzzy set
Classification
Euclidean distance
Membership function

ABSTRACT

Data mining techniques play a major role in developing computer aided diagnosis systems and expert systems that will aid a physician in clinical decision making. In this work, a classifier that combines the relative merits of fuzzy sets and extreme learning machine (FELM) for clinical datasets is proposed. The three major subsystems in the FELM framework are preprocessing subsystem, fuzzification subsystem and classification subsystem. Missing value imputation and outlier elimination are handled by the preprocessing subsystem. The fuzzification subsystem maps each feature to a fuzzy set and the classification subsystem uses extreme learning machine for classification.

Cleveland heart disease (CHD), Statlog heart disease (SHD) and Pima Indian diabetes (PID) datasets from the University of California Irvine (UCI) machine learning repository have been used for experimentation. The CHD and SHD datasets have been experimented with two class labels one indicating the absence and the other indicating the presence of heart disease. The CHD dataset has also been experimented with five class labels, one class label indicating the absence of heart disease and the other four class labels indicating the severity of heart disease namely low risk, medium risk, high risk and serious. The PID data set has been experimented with two class labels one indicating the absence and the other indicating the presence of gestational diabetes.

The classifier has achieved an accuracy of 93.55% for CHD data set with two class labels; 73.77% for CHD data set with five class labels; 94.44% for SHD data set and 92.54% for PID dataset.

1. Introduction

The major silent killer diseases are heart disease and diabetes [1]. Cardiovascular diseases (CVD) refer a group of disorders of the heart and blood vessels. Diabetes is one of the risk factor for CVD [2]. Diabetes is a chronic disease that is a consequence when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin produced by the pancreas [3].

"Clinical decision support systems (CDSS) are computer systems designed to impact clinician decision making about individual patients at the point in time that these decisions are made" [4]. CDSS focuses on increasing the accuracy of decision making and decreases the processing time and cost. Data mining algorithms can be used for developing CDSS. Data mining encompasses statistical analysis, machine learning techniques to discover useful and previously unknown patterns from voluminous amount of data from databases [5,6]. The major data mining functionalities are association rule mining, classification and clustering [5,7]. Association rule mining discovers interesting relationship between items. The interestingness of the relationship is measured using two metrics, namely, support and confidence [7]. Classification is the process of developing a model that describes for the purpose of being able to use the developed model to distinguish or predict the class of objects whose class label is unknown [7]. Clustering is performed on a dataset to categorize into a group by maximizing the similarity and minimizing the difference in the group [7]. The learning technique used in classification is supervised whereas in clustering it is unsupervised. In this work, first each clinical dataset is preprocessed for handling missing values and outliers. Missing values are imputed and instances with outlier value(s) are eliminated from the clinical dataset. Second fuzzification is performed on the preprocessed data and third the classifier is modeled using extreme learning machine (ELM).

2. Background of fuzzy extreme learning machine

Fuzzy extreme learning machine (FELM) combines the advantage of ELM and fuzzy set theory. ELM, a learning algorithm

* Corresponding author. Tel.: +91 44 22358013.
E-mail address: nehemiah@annauniv.edu (K.H. Nehemiah).

http://dx.doi.org/10.1016/j.imu.2016.01.001

developed by Huang et al. [8], is applied in a single layer feed-forward neural network (SLFNN) where the weights between the input layer neurons and hidden layer neurons are randomly generated and the weights between hidden layer neurons and output layer neurons are analytically determined through simple generalized inverse operations [8,9]. ELM overcomes the limitation of the popularly used backpropagation learning algorithm in SLFNN. Backpropagation neural network (BPNN) learning algorithm is either very slow due to improper learning rate or easily converges to local minima. Besides this, BPNN needs many iterative learning steps to accomplish the learning task. ELM has better generalization performance compared to BPNN and it tends to reach the solutions without using parameters like, learning rate, momentum rate as that of backpropagation learning algorithms. In ELM once the weights between the input layer neurons and the hidden layer neurons are randomly generated, the weights will not be iteratively tuned or adjusted as in BPNN [9]. This significantly reduces the time taken to train the network.

In SLFNN the output value of the output layer neurons (O_k) can be computed using the value of the hidden layer neurons (H_j) and the connecting weights (W_{jk}^{ho}) as follows

$$O_k = f\left(\sum_{j=1}^{q}\left(H_j W_{jk}^{ho}\right)\right) \quad k = 1, 2, \ldots n \tag{1}$$

Where, f is the activation function, $q$is the number of hidden layer neurons, n is the total number of training dataset, Using the identity function, O_k becomes the summation of the product of H_j and W_{jk}^{h0} as follows

$$O_k = \sum_{j=1}^{q}\left(H_j W_{jk}^{ho}\right) \quad k = 1, 2, \ldots n \tag{2}$$

The objective of ELM neural network is to minimize the error between output value (O_k) and the target class (T_k). Using the approximate zero error mean given by $\sum_{k=1}^{n} \|O_k - T_k\| \cong 0$, hence, Eq. (2) can be written compactly as

$$T = HW \tag{3}$$

Where, T is the target class, H is the output value of the hidden layer neurons, W is the weights that connects the hidden layer neurons and the output layer neurons. Then the unknown weights (W) can be computed as

$$W = H^{\dagger}T \tag{4}$$

where $H^{\dagger}$ is Moore-Penrose's generalized inverse of H.

Fuzzy set theory was introduced by Zadeh [10] for handling uncertainty. Fuzzification refers to the process of mapping each feature in the clinical dataset to a fuzzy set with a degree of membership ranging from 0 to 1 [11,12]. Each feature is represented by two or more linguistic variables. For example the feature Diastolic blood pressure can be represented as a fuzzy set with three members namely Hypotension, Normal and Hypertensive. Clinical datasets have uncertainty; hence fuzzy set theory is used to resolve the uncertainty problem. In this study, each feature value of the instance is represented by the membership value of the corresponding linguistic variables.

FELM was used by Zhang et al. [13] for weighted classification problem. In their proposed method, fuzzy set theory has been used for weighting the instances of dataset based on the number of distinct class labels. For example the dataset with three class labels has the weight values of 0.5, 0.3 and 0.2. The summation of the given weights becomes 1. Their FELM was tested by using only a fixed number of hidden layer neurons.

FELM takes the advantage of fuzzification and ELM. Fuzzification of features of the clinical dataset helps to get higher performance accuracy and the learning process of ELM helps to obtain not only higher accuracy, but also reduces the training time.

In this research work, the FELM classifier was developed and tested with a varying number of hidden layer neurons. The first classifier has used 10 hidden layer neurons, and increase by one for the second classifier. The increment is terminated when the hidden layer neurons becomes 200. The classifier with highest performance is selected.

The rest of the paper is organized as follows. The description of related work carried out by other researchers is presented in Section 2. In Section 3, the system framework of the proposed work is discussed. Experimental results and comparison of the proposed work with works carried out by other researchers is discussed in Section 4. Conclusion and scope for future work is discussed in Section 5.

3. Related work

Related works carried out by other researchers using clinical data sets taken from the UCI machine learning repository is discussed in this section.

Aslam et al. [14] in their work have used Pima Indian Diabetes (PID) dataset for diagnosing the presence or absence of diabetes. The researchers have carried out their work in three stages. In stage one, the diabetes features have been normalized to zero mean and unit standard deviation. Student's t-test, Kolmogrov–Smirnov test, f-score selection, Kullback–Leibler divergence and genetic programming (GP) have been employed to assess the effectiveness of the normalized features. The features were arranged in decreasing order of importance based on the above tests and different subsets of features were prepared using sequential forward selection (SFS) process. In stage two they have used GP with comparative partner selection to generate new features for each subset of features prepared by SFS. In stage three they have tested the performance of the features generated in stage two using KNN and SVM classifiers. They have achieved a classification accuracy of 80.5% using GP-KNN with ten-fold cross validation and 87% using GP-SVM. The researchers have not dealt with the uncertainty and vagueness of the feature value in the dataset. Furthermore their classification approach has been tested over only one dataset which may not generalize over the other clinical dataset.

Patil et al. [15] proposed a hybrid approach by combining K-means clustering algorithm and C4.5 for classifying of Pima Indian diabetes (PID) dataset. Their proposed system has three steps. First, the data has been preprocessed by removing inappropriate and inconsistent data. Due to the 0 values associated in the PID dataset, the researchers have removed two features namely serum-insulin and triceps skin fold, and 143 instances from the dataset. After preprocessing, PID dataset is reduced from 768 to 625 instances and from 8 to 6 features. Z-score method was applied to normalize the reduced PID. Second, patterns have been extracted using K-means clustering algorithm. The incorrectly clustered patterns were removed; thereby the dataset was reduced to 433 instances. Third, a decision tree model has been constructed using the extracted patterns. They have achieved a classification accuracy of 92.38% using ten-fold cross validation. This work suffers from overfitting problem as their proposed clustering technique eliminates 192 instances (about 30% of the preprocessed dataset) that are incorrectly clustered.

Alneamy et al. [16] in their work have used teaching learning based optimization (TLBO) and fuzzy wavelet neural network (FWNN) for diagnosis of heart disease. They have used Cleveland heart disease (CHD) dataset. Gaussian membership function has been used for fuzzification. TLBO is applied to update the weight of