Journal of Computational Science 13 (2016) 10–25

Contents lists available at ScienceDirect

Journal of Computational Science

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

HMV: A medical decision support framework using multi-layer classifiers for disease prediction

Saba Bashir[a], Usman Qamar[a,*], Farhan Hassan Khan[a], Lubna Naseem[b]

[a] Computer Engineering Department, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan
[b] Shaheed Zulfiqar Ali Bhutto Medical University, PIMS, Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 4 August 2015
Received in revised form 2 January 2016
Accepted 4 January 2016
Available online 6 January 2016

Keywords:
Data mining
Prediction
Majority voting
Disease classification
Multi-layer
Ensemble technique

ABSTRACT

Decision support is a crucial function for decision makers in many industries. Typically, Decision Support Systems (DSS) help decision-makers to gather and interpret information and build a foundation for decision-making. Medical Decision Support Systems (MDSS) play an increasingly important role in medical practice. By assisting doctors with making clinical decisions, DSS are expected to improve the quality of medical care. Conventional clinical decision support systems are based on individual classifiers or a simple combination of these classifiers which tend to show moderate performance. In this research, a multi-layer classifier ensemble framework is proposed based on the optimal combination of heterogeneous classifiers. The proposed model named "HMV" overcomes the limitations of conventional performance bottlenecks by utilizing an ensemble of seven heterogeneous classifiers. The framework is evaluated on two different heart disease datasets, two breast cancer datasets, two diabetes datasets, two liver disease datasets, one Parkinson's disease dataset and one hepatitis dataset obtained from public repositories. Effectiveness of the proposed ensemble is investigated by comparison of results with several well-known classifiers as well as ensemble techniques. The experimental evaluation shows that the proposed framework dealt with all types of attributes and achieved high diagnosis accuracy. A case study is also presented based on a real time medical dataset in order to show the high performance and effectiveness of the proposed model.

1. Introduction

Data mining in medical domain, is a process of discovering hidden patterns and information from large medical datasets; analyzes them and uses them for disease prediction [1]. The basic goal of data mining process is to extract hidden information from medical datasets and transform it into an understandable structure for future use [2]. A large number of predictive models can be developed from data mining techniques which enable classification and prediction tasks. After discovering knowledge from data, learning phase starts; where a scientific model is built. This learning method evolves the concept of machine learning and can be formally defined as "the complex computation process of automatic pattern recognition and intelligent decision making based on training sample data" [3]. The machine learning classifiers are further categorized into supervised learning and unsupervised learning depending on the availability of data. In supervised learning, labeled training data is available and a learning model is trained. Some examples include Artificial Neural Network (ANN), Support Vector Machine (SVM), and Decision Trees (DT). In unsupervised learning, there is no class label field in sample data. Examples include, *K*-mean clustering and Self-Organization Map (SOM). An ensemble approach performs better than individual machine learning techniques by combining the results of individual classifiers [4,5]. There are multiple techniques that can be utilized for constructing the ensemble model and each result in different diagnosis accuracy. Most common ensemble approaches are bagging [6], boosting [7] and stacking [8].

1.1. Contribution

Significant amount of work has already been done on disease classification and prediction. However, there is no single methodology which shows highest performance for all datasets or diseases,

* Corresponding author.
E-mail addresses: saba.bashir@ceme.nust.edu.pk (S. Bashir), usmanq@ceme.nust.edu.pk (U. Qamar), farhan.hassan@ceme.nust.edu.pk (F.H. Khan), doctorlubna@hotmail.com (L. Naseem).

http://dx.doi.org/10.1016/j.jocs.2016.01.001
1877-7503/

while one classifier shows good performance in a given dataset, another approach outperforms the others for other dataset or disease. The proposed research focuses on a novel combination of heterogeneous classifiers for disease classification and prediction, thus overcoming the limitations of individual classifiers. The novel combination of heterogeneous classifiers is presented which is Naïve Bayes, Linear Regression, Quadratic Discriminant Analysis, *K*-Nearest Neighbor, Support Vector Machine, Decision tree using Information Gain and Decision tree using the Gini Index. The multiple classifiers are used at multiple layers to further enhance disease prediction accuracy. An application has also been developed for disease prediction. It is based on the proposed HMV ensemble framework. The proposed application can help both doctors and patients in terms of data management and disease prediction.

The rest of the paper is organized as follows: Section 2 relates to literature review. Section 3 presents the proposed ensemble framework. Section 4 provides the results and discussion from the experiments carried out. Section 5 provides a case study about proposed ensemble model, whereas Section 6 is related to the discussion. Medical application for disease diagnosis is detailed in Section 7 and finally the conclusion is provided in Section 8.

2. Literature review

Extensive amount of work has already been done on disease classification and prediction. However, most of the literature has focused on using a single classifier for a specific disease.

Pattekari and Parveen [9] presented a heart disease prediction system based on a Naïve Bayes algorithm to predict the hidden patterns in a given dataset. The proposed technique limits the use of only categorical data and uses only single classifier. Other data mining techniques such as ensembles, time series, clustering and association mining can be incorporated to improve the results. Similarly Ghumbre, Patil, and Ghatol [10] presented a heart disease prediction system using radial based function network structure and support vector machine. Again a single classifier is being utilized. Prashanth et al. [11] proposed automatic classification and prediction of Parkinson's disease. SVM and logistic regression are used for model construction. SVM classifier with RBF kernel produced high classification accuracy. Improvements can be made by incorporating ensemble classifier instead of a single classifier. Übeyli [12] used different classifiers for disease diagnosis and analyzed that support vector machine achieved the highest performance. The method limits the use of a single classifier for diagnosis and prediction. Ba-Alw et al. [13] presented a survey on data mining approached for hepatitis classification and prediction. The comparison of results indicates that Naïve Bayes attained high classification and prediction accuracy. However, again a single machine learning technique is considered.

Ensemble techniques have been for disease prediction. Zolfaghari [14] proposed a framework for diagnosis of diabetes in female patients. The proposed framework uses an ensemble classifier which is based on neural network and support vector machine. Sapna and Tamilarasi [15] proposed an algorithm that uses fuzzy systems and neural networks for learning membership functions. However, both frameworks use a single layer approach. Multiple layers of classifiers can be incorporated to further increase accuracy. Temurtas [16] introduced a neural network ensemble method for thyroid disease diagnosis in medical datasets. The proposed research focuses on using multilayer, probabilistic and learning vector quantization methods for implementing the neural networks. However the framework is tested only for thyroid disease.

The literature review shows that multiple techniques that have been utilized for disease classification and prediction. However, there is no single methodology which shows highest performance for all datasets or diseases Therefore, the proposed research focuses on multi-classifier and multi-layer ensemble framework for disease classification and prediction with high accuracy for all diseases and datasets.

3. HMV ensemble framework

The proposed ensemble framework consists of three modules, namely data acquisition and preprocessing, classifier training and HMV (Hierarchical Majority Voting) ensemble model for disease classification and prediction with three layered approach.

3.1. Data acquisition and pre-processing module

Data acquisition and pre-processing module includes feature selection, missing value imputation, noise removal and outlier detection. There are multiple methods for feature selection and some of them are given as follows. The HMV ensemble framework utilizes F-score feature selection method.

3.1.1. Feature extraction using principal component analysis (PCA)

The principal component analysis technique assumes that most interesting and useful feature in the dataset is one which has the largest variance and spread. This theory is based on the fact that the dimension with the largest variance represents the dimension which has the largest value of entropy and thus corresponds to maximum information. Eigen vector represents x and y coordinates for a given data. The smallest eigenvectors will often simply represent noise components, whereas the largest eigenvectors often correspond to the principal components that define the data. Dimensionality reduction by means of PCA is then accomplished simply by projecting the data onto the largest eigenvectors of its covariance matrix. Therefore, we obtain a linear M-dimensional subspace of the original N-dimensional data, where $M \leq N$. The Singular Value Decomposition (SVD) is a way to perform PCA analysis and is given by [17]:

$$[U, S, V] = \mathrm{SVD}(A) \tag{1}$$

Therefore,

$$A = USV^T \tag{2}$$

where A is covariance input, U and S hold the eigenvectors of A. The advantages of feature extraction using PCA comprise stability, robustness and fancy extension.

3.1.2. Feature extraction using particle swarm optimization (PSO)

The particle swarm optimization is an evolutionary computational technique where a population is termed as swarm which consists of candidate solutions that are encoded as particles in the search space. The random initialization of a population of particles is used as starting criteria having its own objective function value. The feature extraction method uses PSO as a filter and Correlation-based Feature Selection (CFS) as fitness function. A search algorithm is used by CFS for evaluation of feature subsets. The usefulness of each feature is then evaluated for predicting the class label along with level of inter-correlation between features. Highly correlated features with the class and uncorrelated with each other are considered as good feature subsets. PSO searches for the optimal solution by updating the velocity and the position of each particle because each particle flies in the search space with a velocity adjusted by its own flying memory [18].

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \tag{3}$$