Journal of Biomedical Informatics 48 (2014) 114–121

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A novel artificial neural network method for biomedical prediction based on matrix pseudo-inversion

Binghuang Cai *, Xia Jiang

Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15206-3701, USA

ARTICLE INFO

Article history:
Received 9 August 2013
Accepted 11 December 2013
Available online 18 December 2013

Keywords:
Biomedical prediction and classification
Neural networks
Matrix pseudo-inversion
Least Absolute Shrinkage and Selection Operator (LASSO)
Single Nucleotide Polymorphism (SNP)
Cancer

ABSTRACT

Biomedical prediction based on clinical and genome-wide data has become increasingly important in disease diagnosis and classification. To solve the prediction problem in an effective manner for the improvement of clinical care, we develop a novel Artificial Neural Network (ANN) method based on Matrix Pseudo-Inversion (MPI) for use in biomedical applications. The MPI-ANN is constructed as a three-layer (i.e., input, hidden, and output layers) feed-forward neural network, and the weights connecting the hidden and output layers are directly determined based on MPI without a lengthy learning iteration. The LASSO (Least Absolute Shrinkage and Selection Operator) method is also presented for comparative purposes. Single Nucleotide Polymorphism (SNP) simulated data and real breast cancer data are employed to validate the performance of the MPI-ANN method via 5-fold cross validation. Experimental results demonstrate the efficacy of the developed MPI-ANN for disease classification and prediction, in view of the significantly superior accuracy (i.e., the rate of correct predictions), as compared with LASSO. The results based on the real breast cancer data also show that the MPI-ANN has better performance than other machine learning methods (including support vector machine (SVM), logistic regression (LR), and an iterative ANN). In addition, experiments demonstrate that our MPI-ANN could be used for bio-marker selection as well.

1. Introduction and background

In the biomedical area, predicting clinical outcomes and diagnosing disease from available information, such as clinical and genetic evidence, is an important task for patient care and disease cure, especially for cancer applications [1,2]. Biomedical prediction problems are widely encountered in clinical applications including prognosis, diagnosis, and prediction of response to therapy. As information technology and medical equipment rapidly develop, more and more data, including clinical and genetic information, can be collected for medical utilization, which can increase the accuracy of biomedical prediction. However, as data become very large, especially genome-wide data, the processing of the data and the computation time for biomedical prediction is time-consuming and difficult. The biomedical prediction problem has been increasingly investigated by biomedical and informatics researchers [1–4]. This paper will address this challenging problem via developing an effective method and its algorithm.

The biomedical prediction problem can be mathematically described as follows. Given an N-sample training data set with elements defined as $\{\mathbf{X}_i, \mathbf{Y}_i\}_{i=1}^{N}$, where $\mathbf{X}_i = [x_{i1}, x_{i2}, \ldots, x_{im}]^{\mathrm{T}} \in \mathbf{R}^m$ (with m being the number of features/attributes) and $\mathbf{Y}_i = [y_{i1}, y_{i2}, \ldots, y_{in}]^{\mathrm{T}} \in \mathbf{R}^n$ (with n being the number of targets), the prediction problem is to discover the relation between $\mathbf{X}_i$ and $\mathbf{Y}_i$ and develop a model to describe such a relation, so that the output of the model, $\widehat{\mathbf{Y}}_i$, can be as close to the actual targets $\mathbf{Y}_i$ as possible. The problem can be described as

$$\mathbf{X}_i \rightarrow \mathbf{M} \rightarrow Y_i, \ \text{s.t.} |\mathbf{Y}_i - \widehat{\mathbf{Y}}_i| \rightarrow 0, \ i \in \{1, 2, \ldots, N\}. \tag{1}$$

The discovered model $\mathbf{M}$ can then be used for the outcome prediction of a new observation $\widetilde{\mathbf{X}}$. This is important and useful for genetic prediction, clinical diagnosis, and disease classification. The problem is difficult to solve because biomedical data are often discontinuous, incomplete (values missing) and large-scale.

Different models and methods have been developed for the biomedical prediction problem [3,5–11]. For instance, a Bayesian approach using the logistic regression model was presented in [5] for cancer classification and prediction. The risk prediction of prostate cancer recurrence was investigated in [6] through regularized rank estimation in partly linear AFT (Accelerated Failure Time) models using high-dimensional gene and clinical data. An automatically derived class predictor was presented in [3] to determine the

* Corresponding author. Address: Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, 5607 Baum Blvd., Pittsburgh, PA 15206-3701, USA. Fax: +1 4126245310.

E-mail addresses: bic9@pitt.edu, bhcai8@gmail.com (B. Cai), xij6@pitt.edu (X. Jiang).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.12.009

class of new leukemia cases based on gene expression monitoring by DNA micro-arrays. An effective hybrid approach for selecting marker genes was developed in [7] for phenotype classification using micro-array gene expression data. A Bayesian network model for disease outbreak prediction was developed in [8].

Artificial Neural Networks (ANNs) are widely used in science and information technology due to their notable properties including parallelism, distributed storage, and adaptive self-learning capability [12–15]. They have also been utilized to solve biomedical problems, especially in the areas of classification and prediction [9–11]. For example, an artificial neural network, which was developed in [9] to determine whether breast cancer is present based on the age of the patient, mass shape, mass border, and mass density, achieved high predictive rates. A noise-injected neural network was designed in [10] for the classification of small-sample expression data for breast cancer patients. It demonstrated superior performance compared to the other methods tested. Another artificial neural network approach was used to reduce the number of gene signatures for the classification of breast cancer patients and the prediction of clinical outcomes, including the capability to accurately predict distant metastases [11]. However, all these ANN methods become very time-consuming as data become bigger, because the traditional learning method based on back-propagation algorithm is employed, and therefore they may not be applicable to practical biomedical prediction. A hybrid neural network and genetic algorithm method was applied to breast cancer detection in [16]; it used a genetic algorithm to determine the weights of the neural network (i.e., a multi-layer perceptron (MLP)). However, time-consuming iterations are still needed to get the weights for this hybrid ANN method.

Based on our previous work on weight determination of neural networks [13,14] and related work on ANN learning [17], we develop a Matrix-Pseudo-Inversion based Artificial Neural Network (MPI-ANN) for biomedical prediction. MPI-ANN is a feed-forward neural network with one input layer, one hidden layer and one output layer. Most importantly, MPI-ANN can directly determine the weights of the neural network in one step using pseudo-inversion without a traditional weight-updating iteration. For comparison purposes, LASSO (Least Absolute Shrinkage and Selection Operator) [18,19] is also presented for the biomedical prediction problem. Experimental results based on a set of simulated data sets and a real data set demonstrate the effectiveness and efficiency of the developed MPI-ANN method. Not only does MPI-ANN significantly outperform LASSO in terms of prediction accuracy for the simulated datasets, but it also demonstrates better performance in terms of statistical measurements and efficiency than other machine learning methods including support vector machine (SVM), logistic regression (LR) and an iterative ANN when analyzing a real breast cancer data.

The remainder of this paper is organized in four sections. Section 2 presents the MPI-ANN and the LASSO methods, and also introduces the experiment data sets and methods. In Section 3, experimental results are described and analyzed. A discussion appears in Section 4, and Section 5 concludes the paper with final remarks.

2. Methods

In this section, the MPI-ANN method is presented and developed for the biomedical prediction problem. The comparative method, LASSO, is also presented. Experimental data sets and the experimental method are finally introduced.

2.1. MPI-ANN

To solve the biomedical prediction problem, a feed-forward neural network is constructed according to the structure diagram shown in Fig. 1. The constructed MPI-ANN has three layers, i.e., the input layer, the hidden layer, and the output layer. The inputs to the neural network are the observed values of the features in the data set, while the outputs are the targets.

Specifically, in the input layer of MPI-ANN, the kth ($k = 1,2,\ldots,m$) input of the MPI-ANN is the observed value of the kth feature. Each neuron of the input-layer uses a linear activation function $f(x) = x$; i.e., the values input into the neural network are directly passed to the hidden layer. Moreover, the hidden layer has p neurons, which employs a group of activation functions f_l ($l = 1,2,\ldots,p$); i.e., the lth hidden-layer neuron adopts f_l as its activation function. Different kinds of non-regular functions [14,17] can be employed as the activation function of the hidden nodes. In this paper, based on a universal approximation theorem [20], the sigmoid function (i.e., $f(x) = 1/(1+e^{-x})$) is employed for the MPI-ANN, since it is continuous (and thus differentiable), its derivative can be computed quickly, and it has a limited range (from 0 to 1, exclusive) [20]. The connecting weights between the input and hidden layers, $u_{kl} \in \mathbf{R}$ ($k = 1,2,\ldots,m$; $l = 1,2,\ldots,p$), and the biases of the neurons in the hidden layer, $b_l \in \mathbf{R}$($l = 1,2,\ldots,p$), are randomly generated in any intervals of $\mathbf{R}$, since random choice of the input weights and hidden layer biases can exactly learn the training observations, make learning extremely fast, and produce good generalization performance according to [17,21]. Furthermore, the neurons in the output layer also use a linear activation function, and the inputs from the hidden layer are summed as the outputs of the neural network. The MPI-ANN can be considered as a kind of BP (Back Propagation) neural network; so it could use an error back-propagation algorithm [13,15,22] as its training law to determine the weights, $w_{lj} \in \mathbf{R}$ ($l = 1,2,\ldots,p$; $j = 1,2,\ldots,n$), between the hidden and output layers. To avoid the lengthy learning iteration of the traditional error back-propagation (BP) algorithm based on the gradient-descent method [13,15,22], we develop a matrix pseudo-inversion based weight direct determination method to determine such weights for biomedical prediction.

Mathematically, the MPI-ANN model can be formulated as

$$y_{ij} = \sum_{l=1}^{p} h_{il} w_{lj}, \tag{2}$$

where $i = 1,2,\ldots,N$; $j = 1,2,\ldots,n$ and

$$h_{il} = f_l\left(\sum_{k=1}^{m} u_{kl} x_{ik} + b_l\right), \tag{3}$$

is the output of the lth node of the hidden layer for the ith sample.

Based on matrix theory [23], the MPI-ANN model (2) can be expressed as the following matrix form.

$$\mathbf{Y} = \mathbf{H}\mathbf{W}, \tag{4}$$

where

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{Nn} \end{bmatrix} \in \mathbf{R}^{N\times n}, \quad \mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1p} \\ h_{21} & h_{22} & \cdots & h_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ h_{N1} & h_{N2} & \cdots & h_{Np} \end{bmatrix} \in \mathbf{R}^{N\times p},$$

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{p1} & w_{p2} & \cdots & w_{pn} \end{bmatrix} \in \mathbf{R}^{p\times n}.$$

The matrix-form error-function for MPI-ANN is expressed as follows:

$$E = \|\mathbf{Y} - \mathbf{Y}\|^2 = \|\mathbf{Y} - \mathbf{H}\mathbf{W}\|^2, \tag{5}$$