Gene Reports 3 (2016) 31–38

Contents lists available at ScienceDirect

Gene Reports

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

Cancer classification by correntropy-based sparse compact incremental learning machine

Mojtaba Nayyeri [a,*], Hossein Sharifi Noghabi [a,b]

[a] Department of Computer Engineering, Ferdowsi University of Mashhad, Iran
[b] Center of Excellence on Soft Computing and Intelligent Information Processing, Iran

ARTICLE INFO

Article history:
Received 22 June 2015
Received in revised form 3 December 2015
Accepted 27 December 2015
Available online 4 February 2016

Keywords:
Microarray
Cancer classification
Machine Learning
Incremental Learning Machines
Correntropy

ABSTRACT

Cancer prediction is of great importance and significance and it is crucial to provide researchers and scientists with novel, accurate and robust computational tools for this issue. Recent technologies such as microarray and next-generation sequencing have paved the way for computational methods and techniques to play critical roles in this regard. Many important problems in cell biology require the dense nonlinear interactions between functional modules to be considered. The importance of computer simulation in understanding cellular processes is now widely accepted, and a variety of simulation algorithms useful for studying certain subsystems have been designed. In this article, a sparse compact incremental learning machine (SCILM) is proposed for cancer classification problem on microarray gene expression data, which take advantage of correntropy cost that makes it robust against diverse noises and outliers. Moreover, since SCILM uses l_1-norm of the weights, it has sparseness, which can be applied for gene selection purposes as well. Finally, due to compact structure, the proposed method is capable of performing classification tasks in all of the cases with only one neuron in its hidden layer. The experimental analysis is performed on 26 well-known microarray data sets regarding diverse kinds of cancers and the results show that the proposed method not only achieved significantly high accuracy but also because of its sparseness, final connectivity weights determined the value and effectivity of each gene regarding the corresponding cancer.

1. Introduction

Most of human diseases are influenced by genes, and identifying genetic landscape and profile of diseases is an undisputable fact especially when it comes to diseases such as cancer (Upstill-Goddard et al., 2012). In the quest for determination of genetic causes of diseases, new technologies such as next-generation sequencing (Morozova and Marra, 2008; Gobernado et al., 2014) or microarray expression (Muangsub et al., 2014), which are high-throughput procedures, have paved the way to quantitate and record thousands of genes expression levels simultaneously (Taylor et al., 2015; Nguyen and Rocke, 2002; Daiely). These new technologies provide computational oncologists with valuable information for cancer prediction and cancer classification (Upstill-Goddard et al., 2012; Larranaga et al., 2006; Fogel, 2008). Making the best use of these valuable information and extracting it from data sets requires advanced, accurate and robust computational techniques because these data sets most of the time follow "large-p-small-n" paradigm, which means they have high number of observed genes but low number of samples (Luque-Baena et al., 2014). Cancer classification has been studied comprehensively with diverse methods from weighted voting scheme (Golub et al., 1999) and partial least square (PLS) (Nguyen and Rocke, 2002) to support vector machines (SVM) (Furey et al., 2000) and extreme learning machines (ELM) (Huang et al., 2006a). In addition to these methods, artificial neural networks (ANNs) (Lancashire et al., 2009), probabilistic neural networks (PNNs) (Statnikov et al., 2005) and soft computing approaches (hybrid of evolutionary computation and machine learning) were also applied and developed for cancer diagnosis and cancer classification (Luque-Baena et al., 2014; Liu et al., 2005). One of the well-known types of ANNs are constructive networks whose optimum structures (number of nodes in the hidden layer) are determined automatically (Kwok and Yeung, 1997; Fahlman and Lebiere, 1989; Huang et al., 2006b). In these networks, the number of nodes and connectivity weights are gradually increased from the lowest to the optimum value and they are categorized in two types: compact (Kwok and Yeung, 1997; Fahlman and Lebiere, 1989; Huang et al., 2012) and non-compact (Huang et al., 2006b). Input parameters of the newly added node in the non-compact type are specified randomly whereas in the compact one, they are adjusted via an optimization process.

Abbreviation: SCILM, sparse compact incremental learning machine; SVM, support vector machine; PLS, partial least square; ELM, extreme learning machine; ANNs, artificial neural networks; PNNs, probabilistic neural networks; RBF, radial basis function.

* Corresponding author.
E-mail address: mojtabanayyere@gmail.com (M. Nayyeri).

http://dx.doi.org/10.1016/j.genrep.2016.01.001
2452-0144/

Most of these methods are suffering from "curse of dimensionality," which is related to high dimensions of these data sets. Another aspect of cancer classification is related to feature selection (gene selection) methods in order to prevent overfitting in the learning process (Song et al., 2007). Model et al. (2001_ applied several feature selection methods for DNA methylation based cancer classification. Another comparative study for feature selection was performed by Li et al. (2004) for tissue classification based on gene expression. Cawley and Talbot (2006) proposed a sparse logistic regression with Bayesian regularization for gene selection in cancer classification and Zhang et al. (2006) used SVM with non-convex penalty for the same problem. Piao et al. (2012) take advantage of ensemble an correlation-based gene selection method for gene expression data regarding cancer classification. Interested readers can refer to five good surveys of feature selection in (Saeys et al., 2007; Lazar et al., 2012; Hemphill et al., 2014; Ma and Huang, 2008) and (Duval and Hao, 2010) and the references therein. However, feature selection comes with certain prices such as addition of another layer of complexity to the model or information loss (Saeys et al., 2007).

In this article, we propose sparse compact incremental learning machine (SCILM), which prevents overfitting without feature selection due to its compact structure. Further, because of correntropy cost, SCILM is robust against noises and outliers. In addition to these advantages, since SCILM takes advantage of l_1-norm of the weights, it is sparse, and this sparseness determines the most effective connectivity weights corresponding to all features. Therefore, the final weights of the generated model by SCILM can be utilized for gene selection purposes as well.

SCILM is a learning method for data sets with low sample size and high dimensions. These characteristics are highly important and medical and pharmaceutical research because numbers of genes or drug compounds are significantly lower than number of features and attributes one can find for them. SCILM is proposed for such problems and microarray profiles for cancer classification have both these characteristics. The presented method prevents overfitting without feature selection due to its compact structure and also because of correntropy cost SCILM is robust against noises and outliers. Authors in (Sharifi Noghabi and Mohammadi, 2015), investigated robustness of correntropy objective function. In addition to these advantages, since SCILM takes advantage of l_1-norm of the weights, it is sparse and this sparseness determines the most effective connectivity weights corresponding to all features. Therefore, the final weights of the generated model by SCILM can be utilized for gene selection purposes as well.

The rest of the paper is organized as follows: Section 2 presents the proposed method, and Section 3 describes the results and final section concludes the paper.

2. Methods and materials

This section presents a new constructive network with sparse input side connections. The network has a single hidden layer in which the hidden nodes are added one by one until the network reaches a certain predefined performance. After the new hidden node is added and trained, its parameters are fixed and do not changed during training the next nodes. Each newly added node is trained in two phases: (a) input parameters adjustment and (b) output parameter adjustment. The input parameters of the newly added node are trained based on correntropy objective function. The output connection is adjusted by MSE objective function. In the rest of this section, some preliminaries are described followed by the description of the proposed algorithm.

2.1. Data set representation

The data set with N distinct samples is denoted by

$$\chi = \{x_j, t_j\}_{j=1}^{N},\ x_j \in R^d, t_j \in R \tag{1}$$

2.2. Network structure

Let f be a continuous mapping and f_L be the output of the network with L hidden nodes. The network is represented as

$$f_L(x) = \sum_{i=1}^{i=L} \beta_i g_i(x) \tag{2}$$

where

$$g_i(x) = g(\langle w_i, x\rangle + b_i), w_i \in R^d, b_i \in R, \tag{3}$$

where $\langle .,. \rangle$ is inner product between two elements. In this paper, g is considered as tangent hyperbolic function. The network (with L hidden nodes) error vector is defined as

$$\zeta_L = T - F \tag{4}$$

where $T = [t_1, ..., t_N]$ and $F = [f_L(x_1), ..., f_L(x_N)]$. The activation vector for the ith hidden node is

$$H_i = [H_{i1}, ..., H_{iN}], i = 1, ..., L \tag{5}$$

where $H_{ij} = g_i(x_j), j = 1, \ldots, N; i = 1, \ldots, L$.

2.3. Correntropy

Let v and u be two random variables with $\zeta = u - v$. The correntropy (Mohammadi et al., 2015) is a similarity measure between two random variables and defined as

$$V(\zeta) = E(k(\zeta)) \tag{6}$$

where $E(.)$ denotes the expectation in probability theory and k(.) denotes a kernel function which satisfy Mercer condition. In this paper, only the Gaussian kernel is used. Regard to this,

$$V(\zeta) = E\left(\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-\left(\|u-v\|^2\right)}{2\sigma^2}}\right) \tag{7}$$

2.4. Proposed method

This subsection proposes a new incremental constructive network with sparse hidden layer connections. The hidden nodes are added to the network and trained one by one. When the new node parameters are tuned, they are frozen and do not change during training the next nodes. Fig. 1 illustrates the mechanism of the proposed method.

Training of the new node performs in two stages:

2.4.1. Stage 1: input side optimization

In the previous work (Nayyeri et al., 2015), input parameters of the new node are trained based on correntropy objective function as follows:

$$\begin{aligned} V(H_L) &= \arg\max_{H_L} E\left(\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-\left(\|\zeta_{L-1} - s_L H_L\|^2\right)}{2\sigma^2}}\right) \\ &= \arg\max_{H_L} E(k(\zeta_{L-1}, s_L H_L)) = E(<\Phi(\zeta_{L-1}), \Phi(s_L H_L)>) \end{aligned} \tag{8}$$

where s_L is a real number which is obtained by trial and error and ζ_{L-1} is the residual error for the network with $L-1$ hidden nodes. Regarding Eq.(8), the new node H_L has most similarity to the residual error (regard to kernel definition). It is important to note that when the new node vector equals to the residual error vector (most similarity between the