Computer Methods and Programs in Biomedicine 153 (2018) 1–9

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

A deep learning-based multi-model ensemble method for cancer prediction

Yawen Xiao[a], Jun Wu[b], Zongli Lin[c,*], Xiaodong Zhao[b]

[a] Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing of Ministry of Education, Shanghai 200240, China
[b] School of Biomedical Engineering Shanghai Jiao Tong University, Shanghai 200240, China
[c] Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia, P.O. Box 400743, Charlottesville, VA 22904-4743, USA

ARTICLE INFO

Article history:
Received 26 April 2017
Revised 7 August 2017
Accepted 6 September 2017

Keywords:
Multi-model ensemble
Deep learning
Gene expression
Feature selection
Cancer prediction

ABSTRACT

Background and Objective: Cancer is a complex worldwide health problem associated with high mortality. With the rapid development of the high-throughput sequencing technology and the application of various machine learning methods that have emerged in recent years, progress in cancer prediction has been increasingly made based on gene expression, providing insight into effective and accurate treatment decision making. Thus, developing machine learning methods, which can successfully distinguish cancer patients from healthy persons, is of great current interest. However, among the classification methods applied to cancer prediction so far, no one method outperforms all the others.

Methods: In this paper, we demonstrate a new strategy, which applies deep learning to an ensemble approach that incorporates multiple different machine learning models. We supply informative gene data selected by differential gene expression analysis to five different classification models. Then, a deep learning method is employed to ensemble the outputs of the five classifiers.

Results: The proposed deep learning-based multi-model ensemble method was tested on three public RNA-seq data sets of three kinds of cancers, Lung Adenocarcinoma, Stomach Adenocarcinoma and Breast Invasive Carcinoma. The test results indicate that it increases the prediction accuracy of cancer for all the tested RNA-seq data sets as compared to using a single classifier or the majority voting algorithm.

Conclusions: By taking full advantage of different classifiers, the proposed deep learning-based multi-model ensemble method is shown to be accurate and effective for cancer prediction.

1. Introduction

Cancer has been characterized as a collection of related diseases involving abnormal cell growth with the potential to divide without stopping and spread into surrounding tissues [1]. According to the GLOBOCAN project [2], in 2012 alone, about 14.1 million new cases of cancer occurred globally (not including skin cancer other than melanoma), which caused about 14.6% of the death. Since cancer is a major cause of morbidity and mortality, diagnosis and detection of cancer in its early stage is of great importance for its cure. Over the past decades, a continuous evolution of cancer research has been performed [3]. Among the diverse methods and techniques developed for cancer prediction, the utilization of gene expression level is one of the research hotspots in this field. Data analysis on gene expression level has facilitated cancer diagnosis and treatment to a great extent. Accurate prediction of cancer is one of the most critical and urgent tasks for physicians [4].

With the rapid development of computer-aided techniques in recent years, application of machine learning methods is playing an increasingly important role in the cancer diagnosis, and various prediction algorithms are being explored continuously by researchers. Sayed et al. [5] conducted a comparative study on feature selection and classification using data collected from the central database of the National Cancer Registry Program of Egypt, and three classifiers were applied, including support vector machines (SVMs), *k*-nearest neighbour (*k*NN) and Naive Bayes (NBs). The results showed that SVMs with polynomial kernel functions yielded higher classification accuracy compared with *k*NN and NBs. Statnikov et al. [6] carried a comprehensive comparison of random forests (RFs) and SVMs for cancer diagnosis. The results were obtained that SVMs outperformed RFs in fifteen data sets, RFs outperformed SVMs in four data sets, and the two algorithms performed the same in three data sets. These results were obtained by using full set of genes. Similar results were derived based on the gene se-

* Corresponding author.
E-mail addresses: foreverxyw@sjtu.edu.cn (Y. Xiao), junwu302@gmail.com (J. Wu), zl5y@virginia.edu (Z. Lin), xiaodong122@yahoo.com (X. Zhao).

http://dx.doi.org/10.1016/j.cmpb.2017.09.005
0169-2607/

lection method. From a large body of literature in cancer prediction research, none of these machine learning methods is fully accurate and each method may be lacking in different facets in the classification procedure. For instance, it is difficult for SVMs to figure out an appropriate kernel function, and although RFs have solved the over-fitting of decision trees (DTs), RFs may lead the classification result to the category with more samples.

In view of the fact that each machine learning method may outperform others or have defects in different cases, it is thus natural to expect that a method that takes advantages of multiple machine learning methods would lead to superior performance. To this end, several studies have been reported in the literature that aim to integrate models to increase the accuracy of the prediction. For example, Breiman [7] introduced *Bagging*, which combines outputs from decision trees generated by several randomly selected subsets of the training data and votes for the final outcome. Freund and Schapire [8] introduced *Boosting*, which updates the weights of training samples after each iteration of training and combines the classification outputs by weighted votes. Wolpert [9] proposed to use linear regression to combine outputs of the neural networks, which was later known as *Stacking*. Tan and Gilbert [10] applied *Bagging* and *Boosting* on cancerous microarray data for cancer classification. Cho and Won [11] applied the majority voting algorithm to combine four classifiers using three benchmark cancer data sets. The *Stacking* and majority voting take advantages of different machine learning methods. Although the majority voting algorithm is the most common in classification tasks, it is still too simple a combination strategy to discover complex information from different classifiers. *Stacking*, through the use of a learning method in the combination stage, is a much more powerful ensemble technique. Given that the small number of deep learning studies in biomedicine have shown success with this method [12], deep learning has become a strong learning method with many advantages. Unlike the majority voting which only considers the linear relationships among classifiers and requires for manual participation, deep learning has the ability to "learn" the intricate structures, especially nonlinear structures, from the original large data sets automatically. Thus, in order to better describe the unknown relationships among different classifiers, we adopt deep learning in the *Stacking*-based ensemble learning of multiple classifiers.

In this paper, we attempt to use deep neural networks to ensemble five classification models, which are *k*NN, SVMs, DTs, RFs and gradient boosting decision trees (GBDTs), to construct a multi-model ensemble model to predict cancer in normal and tumor conditions. To avoid over-fitting, we employ the differential gene expression analysis to select important and informative genes. The selected genes are then supplied to the five classification models. After that, a deep neural network is used to ensemble the outputs of the five classification models to obtain the final prediction result. We evaluate the proposed method on three public RNA-seq data sets from lung tissues, stomach tissues and breast tissues, respectively. The final results indicate that the proposed deep learning-based multi-model ensemble method makes more effective use of the information of the limited clinical data and generates more accurate prediction than single classifiers or the majority voting algorithm.

2. Methods

The flowchart of the proposed deep learning-based ensemble strategy is shown in Fig. 1. Initially, differential expression analysis is used to select the significantly differentially expressed genes, namely the most informative features, which are then fed to the following classification process. Then, we employ the technique of *S*-fold cross validation to divide the initial data into *S* groups of training and testing data sets. After that, multiple classifiers (first-stage models) are learned from the training sets, each of which consists of $S-1$ of the *S* groups, and then applied to the corresponding test set, which is the remaining group of the *S* groups, to output the predicted class of the samples. Finally, we use a deep neural network classifier (second-stage ensemble model) to combine the predictions in the first stage with the aim of reducing the generalization error and procuring a more accurate outcome.

2.1. Feature selection

The use of gene expression data with an increasing number of features (*e.g.*, genes) and information makes it more challenging to develop classification models. In clinical practice, the number of cancer samples available is rather small in comparison with the number of features, resulting in higher risk of over-fitting and degradation of the classification performance. Feature selection is a good way to address these challenges [13]. By reducing the entire feature space to a subset of features, over-fitting of the classification model can be avoided, thus mitigating the challenges arising from a small sample size and a high data dimensionality.

In this paper, we employ the DESeq [14] method to select informative genes for the downstream classification. The DESeq method is usually used to decide whether, for a given gene, an observed difference in read count is significant, that is, whether it is greater than what would be expected just due to natural random variation [14]. In differential expression analysis, by setting the thresholds of the BH-adjusted *p*-value and the fold change level, the significantly differentially expressed genes are screened and selected.

2.2. Cross validation

For many classification models, the complexity may be governed by multiple parameters. In order to achieve the best prediction performance on new data, we wish to find appropriate values of the complexity parameters that lead to the optimal model for a particular application.

If data are plentiful, then a simple way for model selection is to divide the entire data into three subsets, the training set, the validation set and the test set. A range of models are trained on the training set, compared and selected on the validation set, and finally evaluated on the test set. Among the diverse complex models that have been trained, the one having the best predictive performance is selected, which is an effective model validated by the data in the validation set. In a practical application, however, the supply of data for training and testing is limited, leading to an increase of the generalization error. An approach to reducing the generalization error and preventing over-fitting is to use cross validation [15].

The technique of *S*-fold cross validation [15] used in this paper is illustrated in Fig. 2 for the case of $S = 4$. *S*-fold cross validation partitions the available data set *D* into *S* disjoint groups, $D_1, D_2, \ldots, D_S$, with all subsets maintaining consistency in the data distribution. After that, *S*-1 groups are used as the training set and the remaining group is used as the test set. The procedure is then repeated for all *S* possible choices of the $S-1$ groups, and the performance scores resulting from the *S* runs are then averaged. In our study, we not only utilize *S*-fold cross validation to implement model selection for every single classifier separately, but also generate new data sets for the ensemble stage by using *S*-fold cross validation on the initial data sets in order to avoid over-fitting.

2.3. Classification methods

After preprocessing of the data sets, we assess the prediction performance of five popular classification methods towards