ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 128-132
doi: 10.18178/wcse.2017.06.022

Machine Learning Based Gene Data Classification Method Research

Lei Shi [+], Zeqi Xie, Jianfeng Ren, Yueyun Du, Hao Yuan, and Qing Zhang

School of Electronics and Information Engineering, Sias International University, Zhengzhou, China

Abstract. Gene expression data classification aims to automatically assign categories or classes to unseen gene expression data by using the existing history gene expression data. Artificial neural networks are computational models inspired by the structure of biological neural networks. Artificial neural networks have the advantages of self-adapting, self-organizing, and self-learning, and have the ability to possess robustness, parallelism, and generalization. Ensemble learning is one of the major advances in machine learning, and it is a method to employ multiple learners and then combine their predictions to output the final decision. In this paper, we propose to combine the artificial neural networks and ensemble learning technique to do classification for gene expression data. Experimental evaluation of different methods is performed on public gene expression dataset and the results showed the proposed method achieves significant performance improvement.

Keywords: artificial neural networks, ensemble learning, gene expression data

1. Introduction

In recent years, microarray technology has been developed to generate large datasets with expression values for thousands of genes and measure expression levels of these genes simultaneously. Scientists can monitor the expression level of thousands of genes with a single experiment. There are more and more gene expression data in real world applications, which can be used to mine useful information. Automated classification of gene expression data has attracted more and more attention to determine the functionality of known or unknown genes. An important task for gene expression data classification is to build a classifier by using the existing history gene expression data firstly and then apply the classifier to classify new unknown gene expression data. Many classification algorithms in the machine learning and data mining communities have been applied on gene expression data, such as the C4.5 decision tree algorithm [1], support vector machine (SVM) [2, 3, 4], bayesian networks [5, 6].

Artificial neural networks (ANNs) are famous machine learning algorithm, and they are regression devices which contain layers of computing neurons with remarkable information processing characteristics. ANNs are complex networks that consist of many highly interconnected, interacting processing neurons to model on the human brain roughly, and they deal with information through interactions among neurons, approximating the mapping between inputs and outputs by compositions of nonlinear functions. ANNs have the advantages of self-adapting, self-organizing, and self-learning, and they are able to detect nonlinearities that are not explicitly formulated as inputs. ANNs have the ability to possess robustness, parallelism, and generalization. They do not require any design of mathematical models and can learn solely based on experience. Thus, ANNs can be applied in the fields of classifying, function approximation, forecasting and association. Recently, ANNs are increasingly found to be powerful in modeling non-stationary processes because of its outstanding generalization capability, and have been widely used in many fields, including of image processing [7], financial data analysis [8], etc.

[+] Corresponding author. Tel.: +86-0371-66431911; fax: +86-0371-66431922.
E-mail address: cnretrieval@126.com.

Ensemble learning has attracted more and more attention from both industrial practitioners and academic researchers in the recent years. As one of the major advances in machine learning, ensemble learning is a technique to employ multiple learners and then combine their predictions to output the final decision [9]. Because of its good generalization, ensemble learning has been applied successfully in many applications such as text classification, agricultural data classification, intrusion detection], credit scoring, etc. Boosting, formulated by Freund and Schapire [10], is a popular ensemble learning algorithm. In boosting algorithm, training instances which are wrongly classified by former component classifiers will play more important effects to train later ones. Thus, the subsequent classifiers in boosting can be built in favor of those instances misclassified by previous component classifiers. There are many boosting algorithms and the version of boosting investigated in this paper is AdaBoost.M1.

To classify the gene expression data accurately is difficult but often crucial for successful diagnosis and treatment. In this paper, an ANNs ensemble is proposed to classify the gene expression data effectively. Boosting technique is employed to build multiple ANNs models and then combine their predictions to classify the gene expression data. An experimental evaluation of different methods is conducted on the public gene expression dataset. The experimental results indicate that the proposed method achieves significant performance improvement.

The rest of this paper is structured as follows: Section 2 introduces the basic theory of the ANNs and ensemble learning briefly, and then describes the ANNs ensemble for gene expression data classification. Section 3 reports the experimental results based on benchmark. Section 4 concludes the paper and presents directions of our future work.

2. ANNs Ensemble

2.1. ANNs

Radial Basis Function (RBF) network (briefly referred to as RBF in the following) is typed of ANNs, and it is introduced by Broomhead and Lowe [11]. RBF uses radial basis functions as activation functions and has attracted a lot of attention due to their learning and generalization abilities. RBF is one of the most important ANNs because of its better approximation capabilities, faster learning algorithms and simpler network structures. Typically, RBF has three layered feed-forward and fully connected network, and it uses radial basis function in the hidden layer neurons. The connections between the input and the hidden layer are not weighted. RBF can provide state-of-the-art solutions for many interesting problems in different areas, and has been successfully applied to many applications, such as medical diagnosis, Sales Forecasting, etc. In this research, RBF is used as component classifier for classifying the gene expression data.

2.2. Ensemble Learning

Ensemble learning is one of the main advances in supervised machine learning since the early 1990's. The main idea of ensemble learning is that no single classifier is uniformly superior to any other. Because of the integration of several single classifiers enhancing the accuracy and reliability of the final classifier, ensemble learning proceeds by building a population of diverse component classifiers and can has overall better performance than the individual component classifiers. The performance of ensemble learning is strongly reliant on the accuracy and the diversity of the component classifiers. To construct the ensemble, a weak learner is used as component classifier, a diversification heuristics is used to extract sufficiently diverse classifiers and a voting mechanism is used to aggregate the constructed diverse classifiers. In the last decades, ensemble learning has become one of the principal current directions in machine learning.

2.3. ANNs Ensemble

The proposed approach uses ANNs as the basic classifiers and adopts the Boosting method to produce a series of component classifier for constructing the ensemble. Initially the weights are uniform for all the training gene expression instances. Then these weights are adjusted after the training of each ANNs classifier is completed in the boosting procedure. The weights will be increased for misclassified instances, and they also will be decreased for correctly classified instances. Finally, according to their own accuracies the final ensemble is constructed by combining individual ANNs classifiers. The detailed algorithm is described as Alg.1.

Algorithm 1: The ANNs ensemble algorithm for gene expression data classification

Input: gene expression dataset $S = \{(x_1, y_1), \cdots, (x_N, y_N)\}$,

Number of iterations T

1. Initialize: $d_n^{(1)} = 1/N$ for all $n = 1, \cdots, N$

2. Do for $t = 1, \cdots, T$

(a)Train ANNs classifier with respect to the weighted sample set $\{S, d^{(t)}\}$ and obtain hypothesis $h_t : x \to \{-1, +1\}$, i.e. $h_t = L(S, d^{(t)})$

(b)Calculate the weighted training error ε_t of h_t:

$$\varepsilon_t = \sum_{n=1}^{N} d_n^{(t)} I(y_n \neq h_t(x_n)),$$

(c)Set

$$\alpha_t = \frac{1}{2} \log \frac{1-\varepsilon_t}{\varepsilon_t}$$

(d)Update weights:

$$d_n^{(t+1)} = d_n^{(t)} \exp\{-\alpha_t y_n h_t(x_n)\} / Z_t,$$

where Z_t is a normalization constant, such that $\sum_{n=1}^{N} d_n^{(t+1)} = 1$.

3. Break if $\varepsilon_t = 0$ or $\varepsilon_t \geq \frac{1}{2}$ and set $T = t - 1$.

4. Output: $f_T(x) = \sum_{t=1}^{T} \frac{\alpha_t}{\sum_{r=1}^{T} \alpha_r} h_t(x)$

3. Experimental Setting

3.1. Datasets

In this section, experiments are conducted on one publicly available real dataset Leukemia [12]. Leukemia dataset includes of 72 leukemia patients using Affymetrix HuGeneFL array. It contains 47 cases of acute lymphoblastic leukemia (ALL) and 25 cases of acute myeloid leukemia (AML) with the expression levels of 7,129 genes.

3.2. Experimental Setting and Results

The statistics of classification performance of each algorithm is measured by 10-fold cross-validation approach to reduce the bias and variance of classification results.

To evaluate the performance of the proposed algorithm for gene expression data classification, popular machine learning techniques SVM, decision tree, RBF network and zeroR are implemented and the classification results of them are included for comparison in the experiments. In the research, the LIBSVM [13] is used for SVM implementation and set linear kernel as default kernel function of SVM. The radial basis function network is used for artificial neural network implementation. The C4.5 algorithm is used for decision tree implementation. To analyze the performance of classification, two famous performance metrics Accuracy and F1 are used in the experiments. The confusion matrix is shown in Table 1.

Table 1: A sample of confusion matrix

Class C		Result of classifier	
		Belong	Not belong
Real classification	Belong	TP	FN
	Not belong	FP	TN

Then the Accuracy, Precision, Recall and F_1 measure can be defined to evaluate the performance of the classification:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The Accuracy represents the proportion of correctly classified instances. Precision for a class is the number of true positive divided by the total number of instances labeled as belonging to the positive class. Recall is defined as the number of true positive divided by the total number of instances that actually belong to the positive class. F_1 measure is the harmonic mean of Precision and Recall.

Fig. 1 shows the weight average values of Accuracy of the proposed technique with several benchmarks on the dataset.

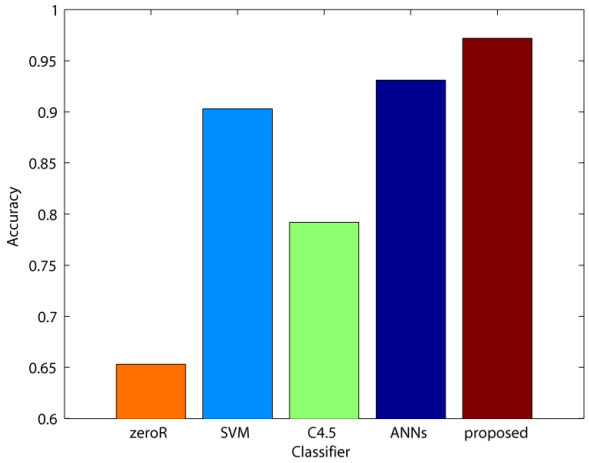

Fig. 1: Comparison of the Accuracy of the methods

From the experimental results shown in Fig. 1, it is clear that the proposed approach outperform other techniques. The weight average values of Accuracy of the proposed approach is 97.2%.The proposed approach is 31.9% better than the zeroR classifier, 6.9% better than the SVM classifier, 18.0% better than the C4.5 classifier and 4.1% better than the ANNs classifier in terms of Accuracy score.

Fig. 2 demonstrates the performance comparison of different techniques on the dataset in terms of F_1.

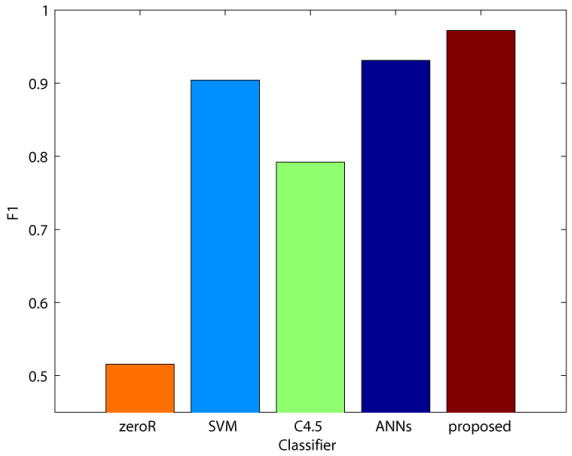

Fig. 2: Comparison of the F_1 of the methods

The F_1 of the proposed approach is 97.2%, which is approximately 45.6% higher than that of zeroR classifier, 6.8% higher than that of SVM classifier, 18.0% higher than that of C4.5 classifier and 4.1% higher than that of ANNs classifier.

4. Conclusion

Analysis and classification of gene expression data has received more and more attention due to the rapid increase in size of the biomedical databases. In this paper, two popular machine learning algorithms, i.e., ANNs and ensemble learning are combined to classify the gene expression data effectively. Comparison the proposed approach with other famous techniques is conducted on public benchmark. The experimental results indicate that the proposed approach achieves significant performance improvement. Gene expression dataset usually has a large number of gene expression values and a relatively small sample size, which poses a severe challenge for accurate classification. Rough set theory is an important and novel feature selection technique. Our future effort is to combine the proposed approach and rough set to improve the classification performance of gene expression data.

5. References

[1] J.-Y.Yeh, T.-H. Wu. Cascade of genetic algorithm and decision tree for cancer classification on gene expression data. *Expert Systems*. 2010, 27(3):201-218.

[2] TS. Furey, N. Cristianini, et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*. 2000, 16:906-914.

[3] I. Guyon, J. Weston, S. Barnhill, V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*. 2001, 46(1-3):389-422.

[4] M. Pirooznia, Y. Deng. SVM Classifier-a comprehensive java interface for support vector machine classification of microarray data. *BMC Bioinformatics*. 2006, 7(4):S25.

[5] N. Dojer, A. Gambin, A. Mizera, B. Wilczynski, J. Tiuryn. Applying dynamic Bayesian networks to perturbed gene expression data. *BMC Bioinformatics*. 2006, 7:249.

[6] N. Friedman, M. Linial, I. Nachman, D. Pe'er. Using Bayesian networks to analyze expression data. *J Comput Biol*. 2000, 7(3-4):601-620.

[7] M. Balasubramanian, S. Palanivel, V. Ramalingam. Real time face and mouth recognition using radial basis function neural networks. *Expert Systems with Applications*. 2009, 36:6879-6888.

[8] A.S. Chen, M.T. Leung, D. Hazem. Application of neural networks to an emerging financial market: Forecasting and trading the Taiwan Stock Index. *Computer and Operations Research*. 2003,30: 901-923.

[9] T. Dietterich. Ensemble methods in machine learning. In: Kittler, J., & Roli,F. (eds.), *First International Workshop on Multiple Classifier Systems*. Lecture Notes in Computer Science, pp. 1-15, 2000.

[10] Y. Freund, R.E. Schapire.A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Systems Sci*.1997,55(1):119-139.

[11] D. Broomhead, D. Lowe. Multivariable functional interpolation and adaptative networks. *Complex Systems*. 1988, 2:321-355.

[12] TR Golub, DK Slonim, P Tamayo, C Huard, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 1999, 286:531-537.

[13] C.-C. Chang, C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011, 2:27:1-27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm