2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing

A comparative study of ensemble learning approaches in the classification of breast cancer metastasis

Wangshu Zhang, Feng Zeng, Xuebing Wu, Xuegong Zhang, Rui Jiang
MOE Key Laboratory of Bioinformatics and
Bioinformatics Division, TNLIST/Department of Automation
Tsinghua University, Beijing 100084, China
Email: ruijiang@tsinghua.edu.cn

Abstract—**The combined use of gene expression profiles and protein-protein interaction (PPI) networks has recently shed light on breast cancer research by selecting a small number of subnetworks as disease markers and then using them for the classification of metastasis. Based on previously identified subnetwork markers, we compare three ensemble learning approaches (AdaBoost, LogitBoost and random forest) with two widely used classifiers (logistic regression and support vector machine) in the classification of breast cancer metastasis. In leave-one-out cross-validation experiments on two breast cancer data sets, the ensemble learning methods can lead logistic regression and support vector machine by 22.4% and 4.8% respectively in terms of the classification accuracy. In cross data set validation experiments, the ensemble learning methods also demonstrate superior reproducibility over the other two methods. With these results, we infer that the ensemble learning approaches with subnetwork markers might be more suitable in handling the classification problem of breast cancer metastasis, and we recommend the use of these approaches in similar classification problems.**

Keywords—breast cancer metastasis; classification; subnetwork markers; ensemble learning;

I. INTRODUCTION

Breast cancer has been reported to be the most common cancer and the second most common cause of cancer death (after lung cancer) among women in the United States [1]. The main cause of death among the breast cancer patients were distant metastases [2], in which breast cancer cells spread to locations other than the breast and lymph nodes. When the cancer is found in bones, it has usually spread to more than one site. At this stage, the disease is treatable, often for many years, but is not curable [3]. Therefore, the appropriate classification of breast cancer metastasis as early as possible is of great importance for the prognosis of this disease.

The classification of breast cancer metastasis with the use of genome-wide gene expression profiles has undergone three periods according to predictive makers (features) used. In the beginning, markers were selected as individual genes that were differentially expressed between different classes of the disease (i.e., metastasis/non-metastasis) [4, 5]. Later, due to the limited power and unobvious effects of individual genes on the prediction of metastasis, pathway analysis was applied, and combinations of genes that belong to common pathways were identified as markers [6]. Recently, benefited from the availability of large-scale protein-protein interaction networks, Chuang *et al* proposed a network-based approach for the classification of breast cancer metastasis, in which subnetworks of interacting proteins within a larger human protein-protein interaction network were used as markers [7]. The combined use of gene expression profiles and protein-protein interaction networks has several advantages over the previous methods for improved prediction power and reproducibility as well as the clear biological meaning [7].

Nevertheless, given the subnetwork makers, the classification accuracies were still not high enough for real applications. It is likely that the classification performances could be improved with the use of some other machine learning approaches. In our work, we employed three ensemble learning approaches, AdaBoost (AB) [8], LogitBoost (LB) [8], and random forest (RF) [9, 10] to take the place of logistic regression (LR) and support vector machine (SVM) [11] used by Chuang *et al* to pursue the objective of improving the performance for the classification of breast cancer metastasis. We use classification and regression trees (CART) [12] as base learners in all of these three ensemble learning approaches. We compared the performances of the five classification approaches, in terms of the classification accuracy (ACC), balanced error rate (BER), Matthew's correlation coefficient (MCC), and area under the receiver operating characteristic (ROC) curve (AUC). We implemented the comparisons using subnetwork markers selected from one cohort of breast cancer patients as predictors of the classification on the same cohort (we called this case "inner data set comparison") and using subnetwork markers selected from one cohort of breast cancer patients as predictors of the classification on the other cohort (we referred to this case as "cross data set comparison"). Both results demonstrated that the ensemble learning approaches might be more suitable to be used with subnetwork markers in dealing with the classification of breast cancer metastasis problem, and thus we suggest the application of these approaches to similar classification problems in bioinformatics.

II. MATERIALS AND METHODS

A. Pretreatment of microarray data

The two data sets involved in our analysis were obtained from the study of Chuang *et al* [7]. In their study, two previous published gene expression profiles by van de Vijver *et al* [4] and Wang *et al* [5] were considered. A subset of 8,141

978-0-7695-3739-9/09 $25.00 © 2009 IEEE
DOI 10.1109/IJCBS.2009.23

genes that present in both data sets and a protein-protein interaction (PPI) network [7] were initially screened out before other analysis. In the van de Vijver data set, 295 breast cancer patients were studied, and 78 of them were reported as metastatic. In the Wang data set, 286 patients were studied, and 106 of them were reported as metastatic.

Integrated with the gene expression and PPI network data sets, the subnetwork markers were identified via two steps. First, to assess the gene expression activity for each patient, a candidate subnetwork was scored to obtain a "subnetwork activity." Then, the discriminative potential of a candidate subnetwork was computed based on the mutual information between its activity score and the metastatic/non-metastatic disease status over all patients [7].

With this method, 149 discriminative subnetworks were identified in van de Vijver data set (including 618 genes), and 243 subnetworks were found in the Wang data set (containing 906 genes). These already identified subnetwork markers were utilized in our work as predictors (features) in the classification of breast cancer metastasis.

B. AdaBoost

AdaBoost, short for Adaptive Boosting, is one of the most popular boosting methods formulated by Freund and Schapire [8]. The hallmark of this algorithm mostly lies in its adaptive adjusting of sample weights during the boosting process, according to the weighted classification error derived from the last training. Denoting the training data as $\{(\mathbf{x}_i, y_i)\}, i = 1, \cdots, N$ and $y_i \in \{-1, 1\}$, we define

$$F^{(m)}(\mathbf{x}_i) = \sum_{m=1}^{M} c^{(m)} f^{(m)}(\mathbf{x}_i),$$

where $F^{(m)}(\mathbf{x}_i)$ is the weighted ensemble of M weak classifiers, and $f^{(m)}(\mathbf{x}_i)$ and $c^{(m)}$ are constant coefficients. We set the same initial weights $w_i^{(0)} = 1/N$ for all samples. In each round, we calculate the ensemble coefficients $c^{(m)}$ and update the sample weights $w_i^{(m)}$ according to weighted errors of the previous classifier. We estimate the post class probability of $y = 1$, $p(\mathbf{x}_i)$, by

$$p(x_i) = \frac{\exp\left(F(\mathbf{x}_i)\right)}{\exp\left(F(\mathbf{x}_i)\right) + \exp\left(-F(\mathbf{x}_i)\right)}$$

and also update it in each rounds till outputting the final class probability. Finally we output the classifier $\text{sign}(F_M(\mathbf{x}))$ and the class probability $p_M(x)$.

C. LogitBoost

Like most other classifiers, AdaBoost might also suffer from the over-fitting problem when dealing with very noisy data. To cope with this situation, Friedman *et al* suggests the use of LogitBoost, which could greatly reduce training errors and hence yield better generalization [8].

For the same training data as above, we transform the class label by letting y^* be $(y+1)/2$ (taking values from $\{0,1\}$) and $c^{(m)}$ be a constant 1/2. The LogitBoost algorithm uses adaptive Newton steps for fitting an additive symmetric logistic model by maximizing the Bernoulli log-likelihood, which optimizes $E(-y(F(x)+f(x)))$ with regard to f at each iteration [8].

D. Random forest

Random forest is an ensemble learning method implemented by growing many classification trees and having them "vote" for a final decision according to a majority role. A random forest generates a number of M decision trees according to the following rule:

(1) Assuming that the number of cases in the training set is N, sample N cases at random with replacement from the original data (bootstrap). This sample will be the training set for growing a tree.

(2) Let the number of features be M. A small number of m ($<<M$) features are selected at random, and the best split within these features is used to split the node. The value of m is held constant during the growth of the forest.

(3) Each tree is grown to the largest extent possible. There is no pruning.

Repeating the creation of a decision tree a number of L times, we obtain L distinct decision trees, forming a randomly generated "forest" [9, 10].

III. Results

We applied the above three ensemble learning approaches (AdaBoost, LogitBoost, random forest) to the problem of classification for breast cancer metastasis and compared their performance with two published methods (logistic regression and SVM). Four criteria, ACC, BER, MCC, and AUC, were adopted to evaluate the performance of these methods. These criteria were defined as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN},$$

$$BER = \frac{1}{2}\left(\frac{FP}{FP + TN} + \frac{FN}{FN + TP}\right),$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(FP + TN)(TN + FN)(FN + TP)}},$$

where TP, FP, FN, and TN stand for true positive, false positive, false negative and true negative, respectively. It is generally regarded that the higher the ACC, MCC, and AUC, and the lower the BER, the better a classifier.

A. Inner data set comparison

We first performed inner data set comparison of the above approaches using leave-one-out cross-validation experiments. We calculated the ACC, BER, MCC, and AUC for each approach and listed the results in Tables I and II. From the tables we can see that the three ensemble learning methods (AB, LB, and RF) in general achieve superior performance over the other two approaches (LR and SVM). Specifically, random forest can lead logistic regression and SVM by 22.4% and 4.8% respectively, in terms of the classification accuracy (see Table I), and LogitBoost can lead logistic

TABLE I. INNER DATA SET COMPARISON (VAN DE VIJVER DATA SET)

	ACC	BER	MCC	AUC
Logistic Regression	0.668	0.382	0.220	0.637
SVM	0.844	0.229	0.580	0.886
AdaBoost	0.868	0.176	0.656	0.893
LogitBoost	0.881	**0.167**	0.687	**0.923**
Random Forest	**0.892**	0.176	**0.709**	0.898

TABLE II. INNER DATA SET COMPARISON (WANG DATA SET)

	ACC	BER	MCC	AUC
Logistic Regression	0.636	0.380	0.236	0.632
SVM	0.801	0.230	0.563	0.874
AdaBoost	0.836	0.181	0.645	0.899
LogitBoost	0.843	**0.175**	0.659	0.925
Random Forest	**0.846**	0.178	**0.665**	**0.932**

regression and SVM by 29.3% and 5.1% respectively in terms of the AUC score (see Table II). We also plot the ROC curves of the five classification methods, as shown in Fig.1 (a and b). As we can see from the figure, the ROC curves of the random forest method climb much faster toward the upper left corner then do those of the other methods in the van de Vijver data set, suggesting that the random forest has better prediction power than the other methods. For the Wang data set, LogitBoost has the best prediction power.

To achieve a fair comparison with the original network-based method [8], we also performed 5-fold cross-validation experiments and calculated the four evaluation criteria at the fixed sensitivity of 90%. As shown in Fig.2 (a), the ensemble learning approaches in general outperform the original logistic regression and SVM in terms of ACC, BER, MCC, and AUC.

B. Cross data set comparison

To evaluate the reproducibility of the ensemble learning methods, we also performed cross data set comparison and calculate the evaluation criteria as before. The results were listed in Table III and IV. From the results we can see that the ensemble learning methods also achieve remarkable performance over logistic regression and SVM, though the improvements are not as high as in the inner data set comparison.

IV. CONCLUSIONS AND DISCUSSIONS

We applied three ensemble learning approaches, AdaBoost, LogitBoost, and random forest to the classification of breast cancer metastasis. Results show that these ensemble learning approaches significantly outperform logistic regression and SVM in both inner and cross data set validations.

As a comparative study of machine learning approaches, we did not explore the problem of obtaining subnetwork markers in this paper. This question is more fundamental and worth further investigating. Besides the information entropy method proposed by Chuang *et al* [7], multivariate statistical methods might also be considered as alternatives in dealing with this problem. Besides, it might also be helpful to apply some feature selection strategies with the ensemble learning approaches to further improve the classification performance. For this purpose, the permutation method based on the out-of-bag estimation of feature importance in the random forest approach might be extended and used with other ensemble learning approaches.

Certainly, the most important problem beyond the classification of breast cancer metastasis is to explore the underlying biological meaning of the subnetwork markers. For this purpose, functional enrichment of genes in a subnetwork can be analyzed with the use of gene ontology (GO) and statistical methods. In our future work, we will explore the possibility of incorporating the above improvements into the ensemble learning approaches and apply the improved methods to a wide range of classification problems in bioinformatics.

TABLE III. CROSS DATA SET COMPARISON (TRAINED ON VAN DE VIJVER DATA SET, TESTED ON WANG DATA SET)

	ACC	BER	MCC	AUC
Logistic Regression	0.619	0.398	0.201	0.608
SVM	0.640	0.441	0.151	0.652
AdaBoost	0.657	0.387	0.239	0.633
LogitBoost	**0.668**	**0.372**	**0.266**	**0.677**
Random Forest	0.654	0.436	0.180	0.645

TABLE IV. CROSS DATA SET COMPARISON (TRAINED ON WANG DATA SET, TESTED ON VAN DE VIJVER DATA SET)

	ACC	BER	MCC	AUC
Logistic Regression	0.576	0.473	0.049	0.532
SVM	0.742	0.446	0.181	0.708
AdaBoost	0.715	**0.399**	**0.219**	0.661
LogitBoost	0.698	0.423	0.168	0.613
Random Forest	**0.746**	0.431	0.210	**0.731**

ACKNOWLEDGMENT

We thank Dr. Trey Ideker for the generosity of providing us the identified subnetwork data. This work was partially supported by Natural Science Foundation of China grants 60805010, 30625012, and 60721003, Hi-tech Research and Development Program of China (863 program) grant 2006AA02Z325, National Basic Research Program of China (973 program) grant 2004CB518605, Tsinghua National Laboratory for Information Science and Technology (TNLIST) Cross-discipline Foundation, Research Fund for the Doctoral Program of Higher Education of China, Scientific Research Foundation for Returned Overseas Chinese Scholars, Research Fund from Intel China Research Center, and a starting up supporting plan at Tsinghua University.

REFERENCES

[1] D. K. Espey, X. C. Wu, J. Swan, C. Wiggins, M. A. Jim *et al*., "Annual report to the nation on the status of cancer, 1975-2004, featuring cancer in American Indians and Alaska Natives" Cancer, vol. 110, Oct. 2007, pp. 2119-2152.

[2] B. Weigelt, J. L. Peterse, L. J. vant't Veer, "Breast cancer metastasis: markers and models" Nature Reviews Cancer, vol. 5, Aug. 2005, pp. 591-602.

[3] M. Lacroix, "Significance, detection and markers of disseminated breast cancer cells" Endocr Relat Cancer, vol. 13, Apr. 2006, pp. 1033-1067.

[4] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart *et al*., "A gene-expression signature as a predictor of survival in breast cancer" The New England journal of medicine, vol.347, Dec. 2002, pp. 1999-2009.

[5] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look *et al*., "Gene-expression profiles to predict distant metastasis of lymph-node-negative promary breast cancer" Lancet, vol. 365, Jul. 2005, pp. 671-679.

[6] Z. Wei and H. Li, "A Markov random field model for network-based analysis of genomic data" Bioinformatics, vol. 23, Mar. 2007, pp. 1537-1544.

[7] H. Chuang, E. Lee, Y. Liu, D. Lee and T.Ideker, "Network-based classification of breast cancer metastasis" Molecular Systems Biology, vol. 3, Aug. 2007, pp. 1-10.

[8] J. Friedman, T. Hastie and R. Tibshirani, "Additive Logistic Regression: a Statistical View of Boosting" Annals of Statistics, vol. 28, Aug. 2000, pp. 337-407.

[9] R. Jiang, H. Yang, F. Sun, T. Chen, "Searching for interpretable rules for disease mutations: a simulated annealing strategy" BMC Bioinformatics, vol. 7:417, Sept. 2006.

[10] R. Jiang, W. Tang, X. Wu and W. Fu, "A random forest approach to the detection of epistatic interactions in case-control studies" BMC Bioinformatics, vol. 10 (Suppl 1): S65, Jan. 2009.

[11] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, 1995. 1-212.

[12] L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and Regression Trees, Belmont CA: Wadsworth International Group, 1984, pp. 1-368.

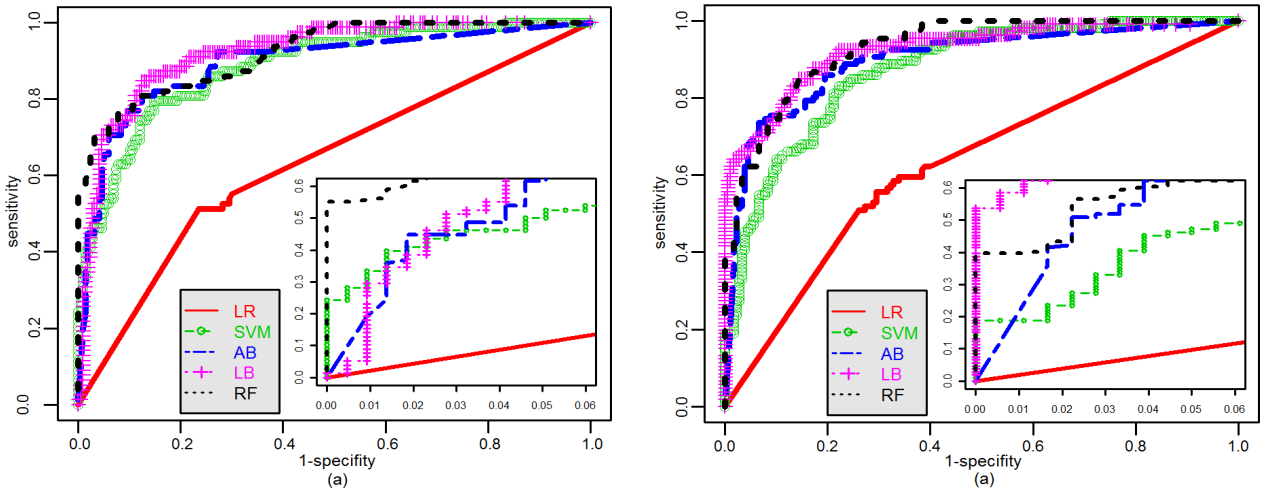

Figure 1. ROC curves. (a) trained on van de Vijver data set and tested on van de Vijver data set. (b) trained on Wang data set and tested on Wang data set. LR is short for Logistic Regression, SVM for support vector machine, AB for AdaBoost, LB for LogitBoost, and RF for Random Forest.

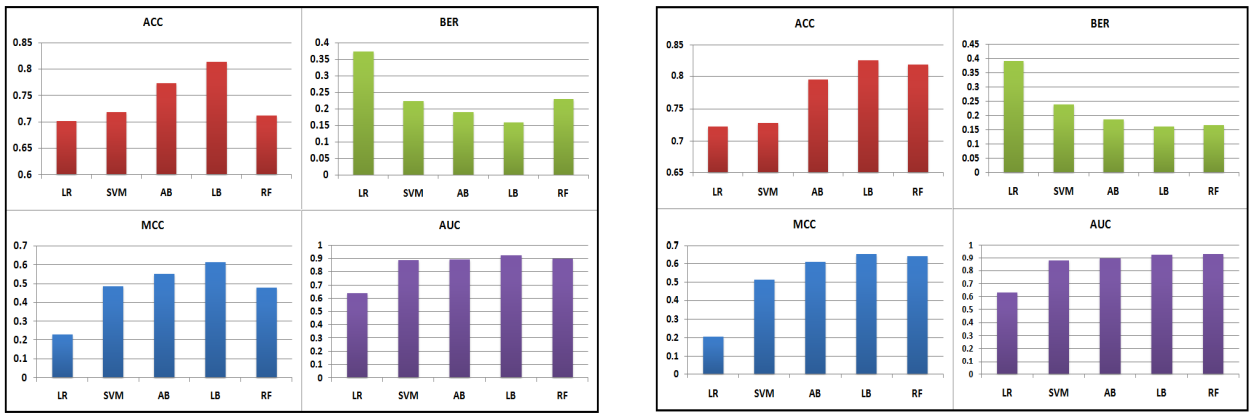

Figure 2. Comparison of the five approaches at the fixed sensitivity of 90%. (a) inner van de Vijver data set. (b) inner Wang data set. LR is short for Logistic Regression, SVM for support vector machine, AB for AdaBoost, LB for LogitBoost, and RF for Random Forest.