Constructing Support Vector Machine Ensembles for Cancer Classification Based on Proteomic Profiling

Yong Mao[1]*, Xiao-Bo Zhou[2], Dao-Ying Pi[1], and You-Xian Sun[1]

[1] *National Laboratory of Industrial Control Technology, Institute of Modern Control Engineering, Zhejiang University, Hangzhou 310027, China;* [2] *Harvard Center for Neurodegeneration and Repair, Harvard Medical School and Brigham and Women's Hospital, Harvard Medical School, Harvard University, Boston, MA 02115, USA.*

In this study, we present a constructive algorithm for training cooperative support vector machine ensembles (CSVMEs). CSVME combines ensemble architecture design with cooperative training for individual SVMs in ensembles. Unlike most previous studies on training ensembles, CSVME puts emphasis on both accuracy and collaboration among individual SVMs in an ensemble. A group of SVMs selected on the basis of recursive classifier elimination is used in CSVME, and the number of the individual SVMs selected to construct CSVME is determined by 10-fold cross-validation. This kind of SVME has been tested on two ovarian cancer datasets previously obtained by proteomic mass spectrometry. By combining several individual SVMs, the proposed method achieves better performance than the SVME of all base SVMs.

Key words: support vector machine ensemble (SVME) design, constructive approach, proteomic profiling, cancer diagnosis

Introduction

Biomarker expression data are usually characterized by a small number of sample vectors of high dimension, which makes it very difficult to be treated with many kinds of single classifiers. Up to now, a possible approach to reduce the dimensionality consists in applying straightforward statistical feature selection operation. Ensemble methods based on re-sampling technique are addressed to solve problems arising from small samples and biological variability of the data (*1*). Ensembles consisting of a certain number of single classifiers outperform a single classifier greatly in terms of classification accuracy (*1*, *2*). In recent studies (*3–5*), the generation performance of an ensemble classifier mainly depends on its base learners' classification accuracy and relativity. Therefore, how to choose a group of single classifiers to compose a high-powered ensemble is a hot topic in this field.

In the present study, we propose a constructive ensemble algorithm based on double-layer hierarchical fusion strategy, that is, the outputs of all base learners are combined together by a specific single classifier. Upon the fusion strategy, the support vector machine recursive feature elimination (SVM-RFE) method (*6*) is adopted to produce a rank of base learners by their contributions to the upper-layer decision machine. The first several base learners in this rank are selected and the ensemble size (or the number of selected base learners) is given by 10-fold cross-validation. Ensembles constructed by our method not only have relatively simpler structures but also have better performance than those constructed by bagging. This algorithm has been tested on two ovarian cancer datasets previously obtained by proteomic mass spectrometry (MS).

*** Corresponding author.**
E-mail: ymao@iipc.zju.edu.cn

Algorithm

The training set of biomarker expression is represented as $Gtr = \{(x_i; y_i) | \, i = 1, 2, \ldots, l\}$, where $x_i \in R^d$ is a d-dimensional vector, in which every dimension corresponds to the expression data from a specific biomarker, and $y_i \in \{-1, +1\}$ represents the class label, that is, which class the sample belongs to. A total of K replicate training sets $\{Gtr_{bootstrap-k} | \, k = 1, 2, \ldots, K\}$ are produced independently by bootstrap technique. Each replicate training set is used to train

a certain SVM; the base learners used to constitute the lower layer of ensemble will be selected from these K SVMs. A new SVM is trained to fuse the output of these K SVMs, and these $(K+1)$ SVMs form a hierarchical structure. Using f_k to be the decision function of the k^{th} SVM in the lower layer and F to be the decision function of the SVM in the upper layer, the final decision value of a given sample x_i is determined as:

$$D(x_i) = F\big(f_1(x_i), f_1(x_i), \ldots, f_K(x_i)\big) \qquad (1)$$

The experimental results from previous studies (*3*, *4*) indicate that most gain of ensemble's performance comes from an optimal combination of several specific base learners. In our method, the optimal combination is realized by the SVM in the upper layer, where the focus is how to select these specific base learners. According to Equation (1), $f_1(x_i), f_1(x_i), \ldots, f_K(x_i)$ are regarded as the K features or inputs of the upper-layer decision machine. Therefore, to choose a group of base learners for constructing ensembles means to choose a group of most discriminative features for the upper-layer decision machine.

The SVM-RFE method proposed by Guyon *et al* (*6*) has shown sound performance on bio-feature selection in bioinformatics (*7*, *8*) and key variable identification in chemical industrial process (*9*). In the present study, this method is adopted to rank the lower-layer decision machines according to their importance to the upper-layer decision machine. In brief, RFE is a circulation procedure for eliminating features by a criterion. It consists of three steps: (1) train the classifier; (2) compute the ranking criterion; (3) remove the features with the smallest ranking scores. The ranking criterion is relative to the realization of classifier, that is to say, RFE is a wrapper algorithm. When the linear kernel SVM $f(x) = \langle w, x\rangle + b = \sum_{i=1}^{l} a_i\, y_i < x_i, x > +b$ is used as a classifier in RFE, the contribution of each feature to the discriminative function, $J(i)$, lies on its weight value, namely $J(i) = (w_i)^2$, where $\mathbf{w} = (w_i) = \sum_{i=1}^{l} a_i\, y_i\, x_i$, and the decision coefficients $\mathbf{a} = (a_i)$ and b are obtained by training of SVM (*10*). SVM is retrained after each elimination operation, because a feature of medium-low importance may be promoted by removing a correlated feature. Finally, 10-fold cross-validation is used to determine the number of classifiers in ensemble.

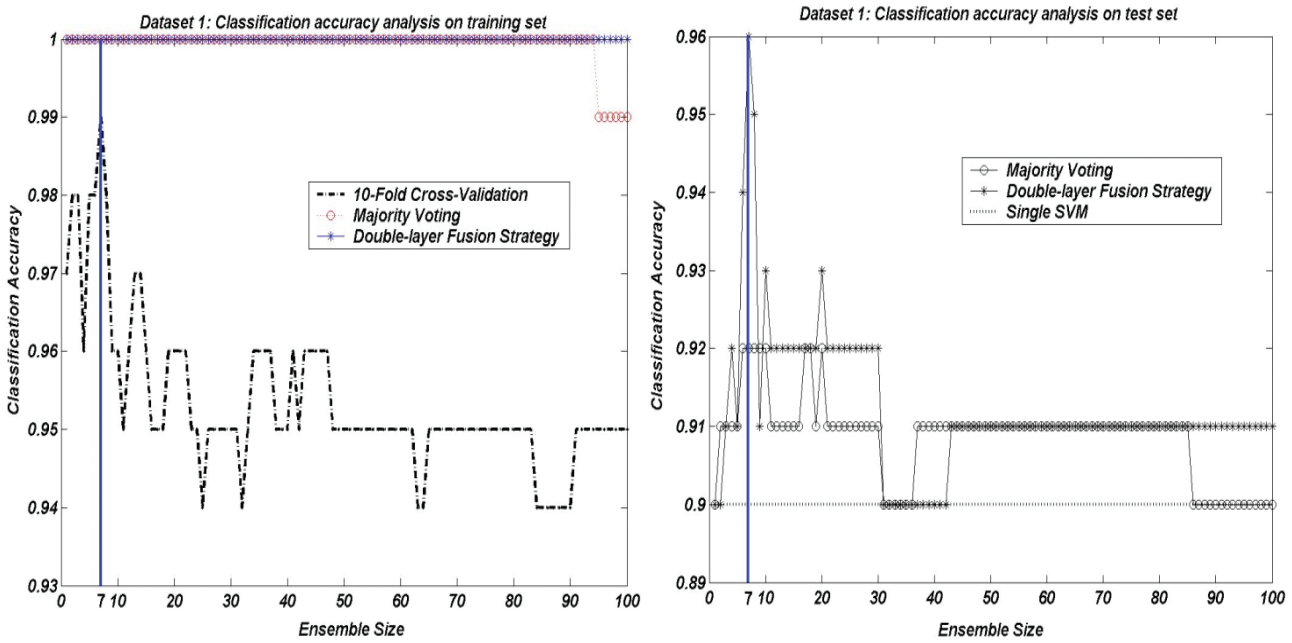
Results and Discussion

The linear kernel SVM was used throughout our experiments. To avoid the noise resulted from the oversize number of features, the Fisher criterion score $F(j) = (\mu_j^+ - \mu_j^-)^2 / \big((\sigma_j^+)^2 + (\sigma_j^-)^2\big)$ was used to preselect 100 biomarkers, where μ_j^+ and μ_j^- denote the mean value of the j^{th} biomarker for Classes 1 and 2, respectively, while σ_j^+ and σ_j^- denote the standard deviation of the j^{th} biomarker for Classes 1 and 2, respectively. In bootstrapping, the size of each resampled dataset was set as 50% of the original training set. The initialized number of base learners in the lower layer of ensemble was set as 100.

Two ovarian cancer datasets from the surface-enhanced laser desorption ionization time-of-flight (SELDI-TOF) experiments by MS (*11*) were analyzed in this study. Dataset 1 was prepared manually, consisting of 200 samples that were separated into a training set and a test set. Each set has 100 samples, including 50 ovarian cancer samples and 50 control (normal) samples. Dataset 2 was prepared with a robotic instrument, consisting of 162 ovarian cancer samples and 91 control samples. Its training set was made up of 60 cancer samples and 40 control samples drawn out statistically, and the test set consisted of the remaining samples. Both two original datasets had 15,154 bio-features in total. Each feature corresponded to the relative intensity of a certain kind of ionized proteomic molecule with specific m/z value.

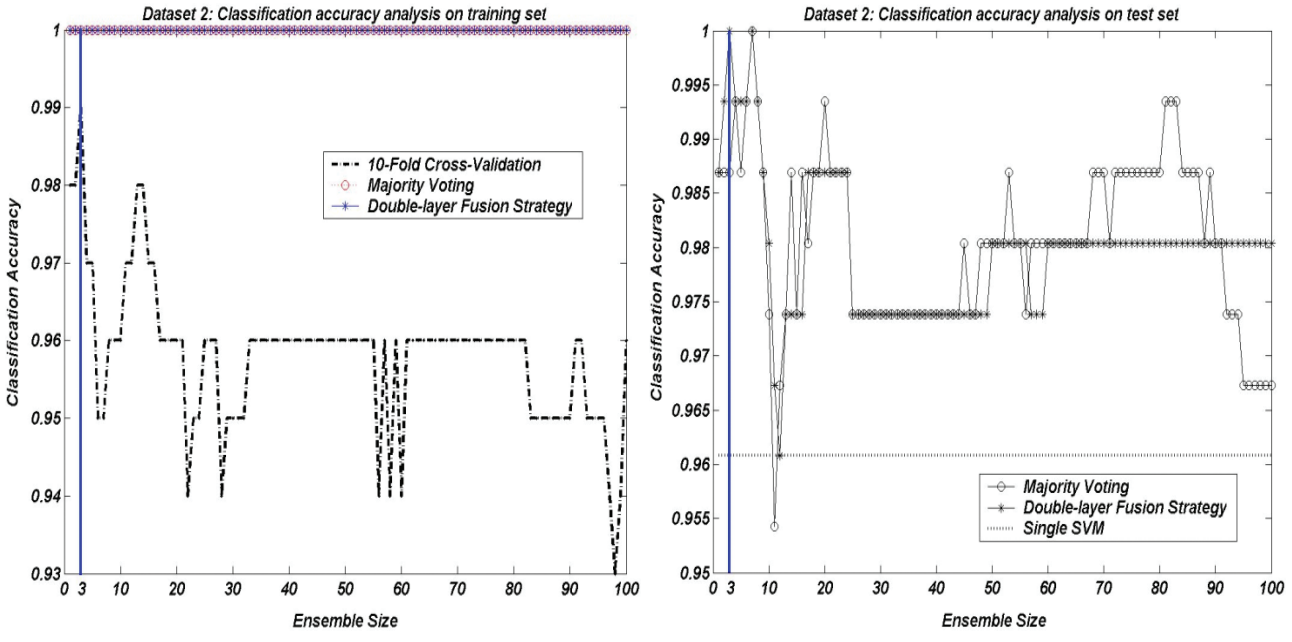
To prove the effectiveness of classifier selection, Majority Voting and double-layer fusion strategy were used as the decision method respectively on the classifier ranking results to test the ensemble accuracy. The classification accuracy analysis on Dataset 1 is shown in Figure 1. The ensemble size indicated by 10-fold cross-validation is 7, and these seven base learners are selected by SVM-RFE. The positive predictive values are 100% on the training set and 96% on the test set, with the ensemble constructed by these seven classifiers. In other cases, the classification accuracy is lower. According to Figure 1, the ensemble constructed by all base learners cannot achieve a lower error rate whether it is fused by Majority Voting or double-layer fusion strategy.

The similar result achieved on Dataset 2 is shown in Figure 2. The ensemble size indicated by 10-fold cross-validation is 3. The positive predictive values are 100% on both the training set and the test set, with the ensemble constructed by these three classifiers. In other cases, the classification accuracy is



A **B**

Fig. 1 Performance comparisons of Majority Voting and double-layer fusion strategy with ranked classifiers on training set and test set from Dataset 1. **A**. Classification analysis on training set. The optimal ensemble size is indicated by 10-fold cross-validation. **B**. Classification analysis on test set. The performance of ensemble with optimal ensemble size is indicated.



A **B**

Fig. 2 Performance comparisons of Majority Voting and double-layer fusion strategy with ranked classifiers on training set and test set from Dataset 2. **A**. Classification analysis on training set. The optimal ensemble size is indicated by 10-fold cross-validation. **B**. Classification analysis on test set. The performance of ensemble with optimal ensemble size is indicated.

lower. According to Figure 2, it is also concluded that the ensemble constructed by all base learners cannot achieve a lower error rate whether it is fused by Majority Voting or double-layer fusion strategy.

By all the above analysis, it can be seen that Majority Voting and double-layer fusion strategy can depress the error rate of cancer diagnosis resulted from single SVM. In most cases, ensembles constructed by double-layer fusion strategy with classifier selection achieve better performance and simpler structure.

Conclusion

We have presented a constructive algorithm for training cooperative support vector machine ensembles (CSVMEs). CSVME combines ensemble architecture design with cooperative training for individual SVMs in ensembles and puts emphasis on both base learner's accuracy and collaboration among individual SVMs. This kind of SVME has been tested on two ovarian cancer datasets previously obtained by proteomic MS. By combining several optimally selected individual SVMs, the proposed method achieves better performance and simpler structure than the SVME of all base SVMs. In addition, CSVME performs better than a single SVM trained on the whole training set.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (No. 60574019 and 60474045), the National Basic Research Program (973 Program) of China (No. 2002CB312200), the Key Technologies R&D Program of Zhejiang Province (No. 2005C21087), the Academician Foundation of Zhejiang Province (No. 2005A1001-13), and the Center for Bioinformatics Program Grant of Harvard Center of Neurodegeneration and Repair, Harvard Medical School, Boston, USA.

References

1. Valentini, G., *et al.* 2004. Cancer recognition with bagged ensembles of support vector machines. *Neurocomputing* 56: 461-466.
2. Bertoni, A., *et al.* 2005. Bio-molecular cancer prediction with random subspace ensembles of support vector machines. *Neurocomputing* 63: 535-539.
3. Kuncheva, L.I., *et al.* 2002. An experimental study on diversity for bagging and boosting with linear classifiers. *Information Fusion* 3: 245-258.
4. Windeatt, T. 2005. Diversity measures for multiple classifier system analysis and design. *Information Fusion* 6: 21-36.
5. Zhou, Z.H., *et al.* 2002. Ensembling neural networks: many could be better than all. *Arti. Intell.* 137: 239-263.
6. Guyon, I., *et al.* 2002. Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46: 389-422.
7. Mao, Y., *et al.* 2005. Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection. *J. Biomed. Biotechnol.* 2005: 160-171.
8. Mao, Y., *et al.* 2005. Parameters selection in gene selection using Gaussian kernel support vector machines by genetic algorithm. *J. Zhejiang Univ. Sci. B* 6: 961-973.
9. Mao, Y., *et al.* 2006. Accelerated recursive feature elimination based on support vector machine for key variable identification. *Chin. J. Chem. Eng.* 14: 65-72.
10. Vapnik, V.N. 1999. *The Nature of Statistical Learning Theory.* Springer, New York, USA.
11. Lilien, R.H., *et al.* 2003. Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. *J. Comput. Biol.* 10: 925-946.