Multi-view Based AdaBoost Classifier Ensemble For Class Prediction from Gene Expression Profiles

Le Li[1], Zhiwen Yu[1], Jiming Liu[2], Jane You[3], Hau-San Wong[4], Guoqiang Han[1]
[1]*School of Computer Science and Engineering, South China University of Technology*
[2]*Department of Computer Science, Hong Kong Baptist University*
[3]*Department of Computing, Hong Kong Polytechnic University*
[4]*Department of Computer Science, City University of Hong Kong*
zhwyu@scut.edu.cn

Abstract

Multi-view learning, one of the important sub-fields in the area of machine learning, has gained more and more attention in class prediction of gene expression datasets. In this paper, we propose a new classifier ensemble framework, named as multi-view based Adaboost classifier ensemble framework (MV-ACE), which not only utilizes a random view generation technique to regulate different views and applies adaboost to adjust the training set, but also designs an adaptive process which explores the feasible combination of multiple views through an optimization process. Traditional multi-view learning focuses on exploring diverse views and the best integration of multiple views in a straightforwardmanner, such as the linear combination of different views. Our proposed model, however, additionally applies a progressive training approach to improve the accuracies of the base classifiers. Moreover, we investigate the assembly of views at the model level, and employ an adaptive process to optimize the multi-view learning model to improve its performance. Our experiments on 12 cancer gene data sets for the classification task show that (i) MV-ACE works well on a diverse class of cancer gene expression profiles. (ii) It outperforms most of the state-of-the-art classifier ensemble approaches on these datasets.

1. Introduction

Class discovery and prediction from gene expression data provides a new way to enhance the diagnostic process of cancer and other complex diseases, and the recommendation of subsequent treatment [1].However, each gene expression profile only contains a small number of samples, while each sample is associated with hundreds of thousands of genes, which raises a challenge to the existing pattern recognition approaches.

In the early years, a large number of researchers adopt single classification algorithms for cancer prediction [2]-[4]. For example, Statnikov et al. investigated the performance of the support vector machines (SVM) and the random forests for cancer classification on gene expression profiles [2]. Wang and Gotoh applied their previous works on feature selection to improve the performance of different single classifier, including the K nearest neighbor classifier, the decision tree classifier, the SVM classifier and the Naive Bayes classifier, thus provided a robust classification framework for cancer classification [3]. Zhang et al. proposed the Binary Matrix Shuffling Filter (BMSF), which considered gene interactions at the gene selection stage, to address the challenges associated with over-fitting in high-dimensional search problems [4].Nowadays, instead of single classifiers , ensemble methods are becoming popular for class prediction of gene datasets [5]-[7]. For example, Ghorai et al. proposed a nonparallel plane proximal classifier (NPPC) ensemble to classify the tissue samples in different gene expression data sets, in which a subspace and model selection strategy was applied to train a group of NPPC models and, consequently, to select the suitable members of the ensemble [5]. Nanni integrated multiple feature reduction methods to train a set of SVMs and constructed an SVM classifier ensemble combining the best feature extraction approaches [6]. Li et al. applied an ensemble classifier based on KNN, SVM and the voting strategy to predict the subcellular localization of eukaryotic proteins [7].

In this paper, we propose a new classifier ensemble framework for class prediction of gene expression da-

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.40

ta, which integrates the multi-view learning approach [8], the Adaboost algorithm [9] and the optimization strategy into the ensemble framework. Multi-view technique in our framework is exploited to deal with the difficulties due to high dimensionality of gene datasets. It reduces the challenging task of classifying a high-dimensional data set to a number of simpler classification tasks on multiple lower-dimensional sets. Then we apply adaboost to optimize the parameters classifiers by performing training on different views. These two approaches offer mutual benefits and achieve common improvement: the adoption of a multi-view perspective makes the high-dimensional set tractable for adaboost and, in turn, adaboost produces the base classifiers which more closely reflect the connection between the matching view and the class label. To obtain the base classifier set, we apply a novel optimization process to search for the best multi-view model. Experiments on the cancer gene expression profiles show the efficiency, robustness and superiority of MV-ACE.

The contribution of the paper is threefold. First, a novel classifier ensemble framework is proposed for class prediction of gene expression data. Second, we utilize data transformation techniques of multi-view learning and adaboost algorithm, which not only consider different feature sets, but also adapt the weight of training samples to improve the classification performance. Third, a classifier ensemble optimization process based on local information is proposed.

The remainder of the paper is organized as follows. Section 2 describes the framework of multi-view based adaboost classifier ensemble (MV-ACE). Section 3 experimentally investigates the performance of MV-ACE on real gene datasets, and Section 4 presents the conclusion.

2 Multi-View based Adaboost Classifier Ensemble (MV-ACE)

The Multi-view based adaboost classifier ensemble (MV-ACE) approach is divided into two parts: the training process of single MV-ACE model as shown in Figure 1 and the adaptive process of the MC-ACE models as illustrated in Figure 2.

Assume (1) the gene expression dataset X consists of n samples $X = (x_1, x_2, ..., x_n)$. (2) Each sample contains m genes. (3) The true class vector Y consists of n class labels $Y = (y_1, y_2, ..., y_n)$, and y_i denotes the class label of x_i (where $i \in \{1, ..., n\}$). The MV-ACE approach first randomly generate B views $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_B$, and each view is an indicator vector of which the length is m. If the gene is selected in the b-th view (where $b \in \{1, ..., B\}$), the corresponding value

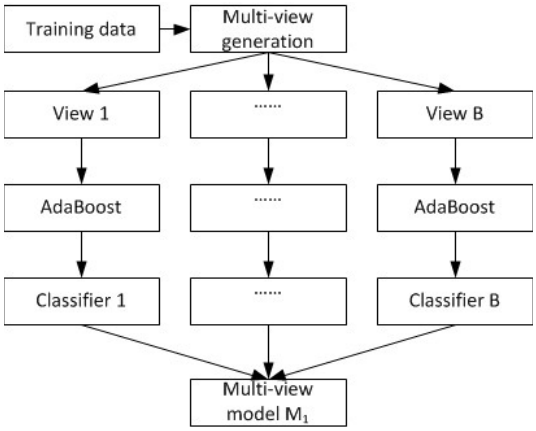

Figure 1. The training process of multi-view based Adaboost classifier ensemble model (MC-ACE)

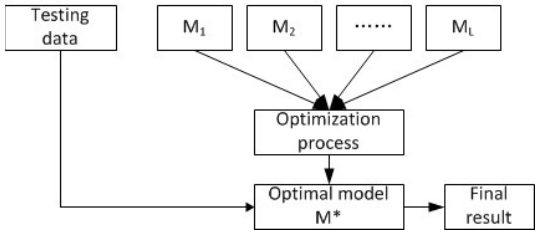

Figure 2. An adaptive process for the models (M_i denotes the i-th MC-ACE)

in $\mathbf{v}_b$ is 1. Otherwise, it is 0. Then, the new training dataset X_b can be generated from the b-th view, and $X_b = (x_1^{(v_b)}, x_2^{(v_b)}, ..., x_n^{(v_b)})$ (where $x_i^{(v_b)}$ denotes the i-th sample in the b-th view). Note that, all the new datasets share the same class vector Y. In the following, a set of the Adaboost classifiers are trained by a set of new training datasets $\{X_1, ..., X_B\}$ generated from different views. Adaboost is a meta-learning algorithm which modulates the weights of training samples iteratively in order to optimize the classifier. Specifically, it first initializes all weights of samples equally, which is as follows:

$$w_i^0 = \frac{1}{n} \tag{1}$$

where $i \in \{1, 2, ..., n\}$. Then, A weak learner ϕ_t in the t-th iteration is trained on the X_b with the weight vector w^{t-1} in the $(t-1)$-th iteration. The total classification error e_t is calculated as follows:

$$e_t = \sum_{i=1}^{n} w_i^{t-1} |c_i - y_i| \tag{2}$$

where c_i is the class label of $x_i^{(v_b)}$ predicted by ϕ_t. In the following, the weight vector w^t is updated as follows:

$$w^t = w' / \sum_k w'_i \tag{3}$$

$$w'_i = w_i^{t-1} \frac{e_t}{1-e_t}^{1-|c_i - y_i|} \tag{4}$$

It is obvious that the weights of correctly classified samples will decrease when e_t is in (0,0.5). Finally, the algorithm iteratively conducts the above process to update the weights of samples in order to train multiple weak classifiers, which could be integrated into a robust classifier. We will obtain B classifiers trained by the Adaboost algorithm with different views. There B classifiers associated B views can be viewed as a multi-view model which is denoted as M.

MV-ACE repeats the above procedure L times and produces L multi-view models $\{M_1, M_2, ..., M_L\}$ as shown in the Figure 2. An adaptive process using local information and the interaction of models is designed to optimize the generated multi-view models and find the best multi-view model M^*. It first determines the local environment for each multi-view model M_l ($l \in \{1, ..., L\}$). M_l can be represented by the distribution of genes D^l in different views, which is calculated as follows:

$$D^l = \sum_{b=1}^{B} v_b^l. \tag{5}$$

The local environment of M_l is clarified as K nearest neighbours of M_l with respect to D^l.

The adaptive process includes three operators, which are the local engagement operator, the group competition operator and the random variation operator. The local engagement operator is defined as the win-win cooperation of model pairs in the same local environment. Given two multi-view models, M_i and M_j, their respective base classifier sets and view sets are $G(M_i)$ with $U(M_i)$ and $G(M_j)$ with $U(M_j)$, which are defined as

$$G(M_i) = \{C_1^i, ..., C_{B/2}^i, C_{B/2+1}^i, ..., C_B^i\} \tag{6}$$

$$U(M_i) = \{v_1^i, ..., v_{B/2}^i, v_{B/2+1}^i, ..., v_B^i\} \tag{7}$$

$$G(M_j) = \{C_1^j, ..., C_{B/2}^j, C_{B/2+1}^j, ..., C_B^j\} \tag{8}$$

$$U(M_j) = \{v_1^j, ..., v_{B/2}^j, v_{B/2+1}^j, ..., v_B^j\} \tag{9}$$

where the classifiers with smaller subscripts outperform the ones with larger subscripts in both models, e.g. C_s^r shows a better performance than C_{s+1}^r. Then these two models offer their superior half parts to combine into a new models, M_α, as follow

$$G(M_\alpha) = \{C_1^i, ..., C_{B/2}^i, C_1^j, ..., C_{B/2}^j\} \tag{10}$$

$$U(M_\alpha) = \{v_1^i, ..., v_{B/2}^i, v_1^j, ..., v_{B/2}^j\} \tag{11}$$

M_α, which inherits the merits of M_i and M_j, is more likely to be a better ensemble model. It will be added into the model population.

The group competition operator is an opposite operation compared with local competition. It randomly selects three model individuals, M_i, M_j and M_k, from the population and makes them a group. Then it defines an objective function of models, θ ,to find the poorest model M_β among the group

$$M_\beta = \arg \min_{M' \in \{M_i, M_j, M_k\}} \theta(M'). \tag{12}$$

M_β will be abandoned from the population in order to actively boost the average performance of the available models. Besides, this procedure makes model population's size unchanged and avoids its unlimited expansion due to local engagement.

The random variation operator is a simple but effective strategy to deal with the local optimal problem. In our method, the applied view set of gene datasets is limited and incomplete, which means the global optimum result may be impossibly achieved through the above two operators. Aiming at this situation, MV-ACE constructs a new adaboost classifier, C', trained on new views, v', and updates a randomly selected model, e.g. M_i, by replacing a classifier therein, e.g. C_k^i, with C'. The model is updated as follow:

$$G(M'_i) = \{C_1^i, ..., C_{k-1}^i, C', C_{k+1}^i, ..., C_B^i\} \tag{13}$$

$$U(M'_i) = \{v_1^i, ..., v_{k-1}^i, v', v_{k+1}^i, ..., v_B^i\}. \tag{14}$$

Random variation expands the search space of our optimization approach and effectively reduces the possibility of being local optimal. A optimized multi-view model set, $\widetilde{M_l} | l = 1, 2, ..., L$, is derived from iterative employment of the above three operations on the initial model set. MV-ACE treat the training accuracy as the evaluation function to select the best model, M^*, from the population

$$l^* = \arg \max_{l \in \{1,2,...,L\}} \zeta(M_l) \tag{15}$$

$$M^* = M_{l^*} \tag{16}$$

where $\zeta()$ denotes the classification accuracy of the multi-view model on the training set. M^* is subsequently applied to classify the testing samples.

Table 1. The summary of real gene expression data sets

Dataset	source	n	m	k	tissue	samples per class
Alizadeh-2000-v3	[11]	62	2093	4	Blood	21, 21, 9, 11
Armstrong-2002-v2	[11]	72	2194	3	Blood	24, 20, 28
Dyrskjot-2003	[11]	40	1203	3	Bladder	9, 20, 11
Garber-2001	[11]	66	4553	4	Liver	17, 40, 4, 5
Lapointe-2004-v2	[11]	110	2496	4	Prostate	11, 39, 19, 41
Liang-2005	[11]	37	1411	3	Brain	28, 6, 3
Nutt-2003-v1	[11]	50	1377	4	Brain	14, 7, 14, 15
Pomeroy-2002-v2	[11]	42	1379	5	Brain	10, 10, 10, 4, 8
Risinger-2003	[11]	42	1771	4	Endometrium	13, 3, 19, 7
Tomlins-2006-v1	[11]	104	2315	5	Prostate	27, 20, 32, 13, 12
Tomlins-2006-v2	[11]	92	1288	4	Prostate	27, 20, 32, 13
Yeoh-2002-v2	[11]	248	2526	6	Bone Marrow	15, 27, 64, 20, 79, 43

3 Experiment

In this section, we conduct a series of experiments on twelve filtered gene expression datasets as shown in Table 1 to explore the necessities of MV-ACE and to investigate its superiority for class prediction of gene expression profiles.

The parameters of MV-ACE are pre-defined as follows: the number of views (B) in each multi-view model is set to 40, and each view is comprised by 10%-50% genes of the original dataset. The number of decision trees in the adaboost algorithm is set to 50. There are 20 multi-view models ($L = 20$) in the population during the adaptive process. The early-stop strategy is utilized here to avoid the over-fitting problem. We have done some preliminary experiments as shown in Figure 3 to determine the number of iteration. Through these experiments, we found the proper number of iterations in the adaptive process was 40. The results in the experiments below contain the average value as well as standard deviation value of 20 runs separately obtained 3-fold cross-validation prediction accuracies. All compared approaches in Figure 4 are provided by WEKA 3.6.9 [12].

3.1 The effect of the adaptive process

We investigate the effectiveness of optimization process in MV-ACE by comparing with the performance of its non-adaptive version which produces one multi-view model. All gene datasets are engaged in this experiment to make a comprehensive comparison.

Table 2 indicates that MV-ACE outperforms MV-ACE-NO. Specifically, MV-ACE has significant performance superiorities (≥ 0.05) when compared with MV-ACE-NO over the datasets, such as Alizadeh-2000-v3 (0.2386), Dyrskjot-2003 (0.0688), Nutt-2003-v1 (0.0788), Pomeroy-2002-v2 (0.0558), Tomlins-2006-v1 (0.0761) and Yeoh-200-v2 (0.1020). The good performance in this univariable-controlling experiment intuitively declares the proposed adaptive process works well and contributes to the MV-ACE approach.

3.2 The comparison of different classifier ensemble approaches

In this experiment, we compare MV-ACE with several the state-of-the-art classifier ensemble methods, such as Bagging (BAG), Multiboosting (MB), Random forests (RF), Random subspace (RSS), and Adaboost (AB), the basic clustering algorithm, on all gene expression datasets as shown in Table 1. MultiBoosting[10], as an extension of Adaboost. The results are presented in Figure 4.

It is obvious that MV-ACE has better performance than other approaches on most of datasets. For example, the accuracy values obtained by MV-ACE are 0.0542, 0.0417 and 0.0712 higher than those obtained by the second best approach on the Armstrong-2002-v2 dataset, the Garber-2001 dataset and the Pomeroy-2002-v2 dataset. In addition, MV-ACE puts up moderate steadiness among these approaches. These results demonstrate the accuracy and the superiority of MV-ACE. The possible reason are as follows: first, the multi-view technique enhances the diversity of base classifiers. Second, the adoption of the Adaboost algorithm provides more accurate for the base classifiers. Third, the adaptive process of the multi-view models improve the performance gradually.

4 Conclusion

In this paper, we have presented a novel multi-view based adaboost classifier ensemble framework (MV-ACE). Specifi-

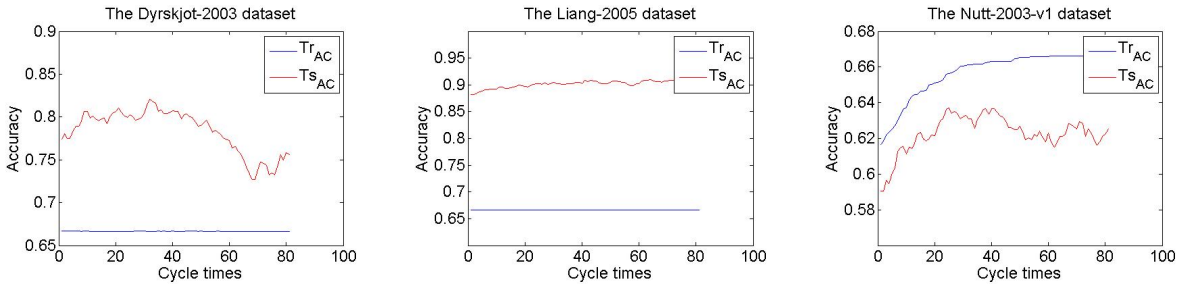

Figure 3. The variation of accuracies on training set (Tr_{AC}) and testing set (Ts_{AC}).

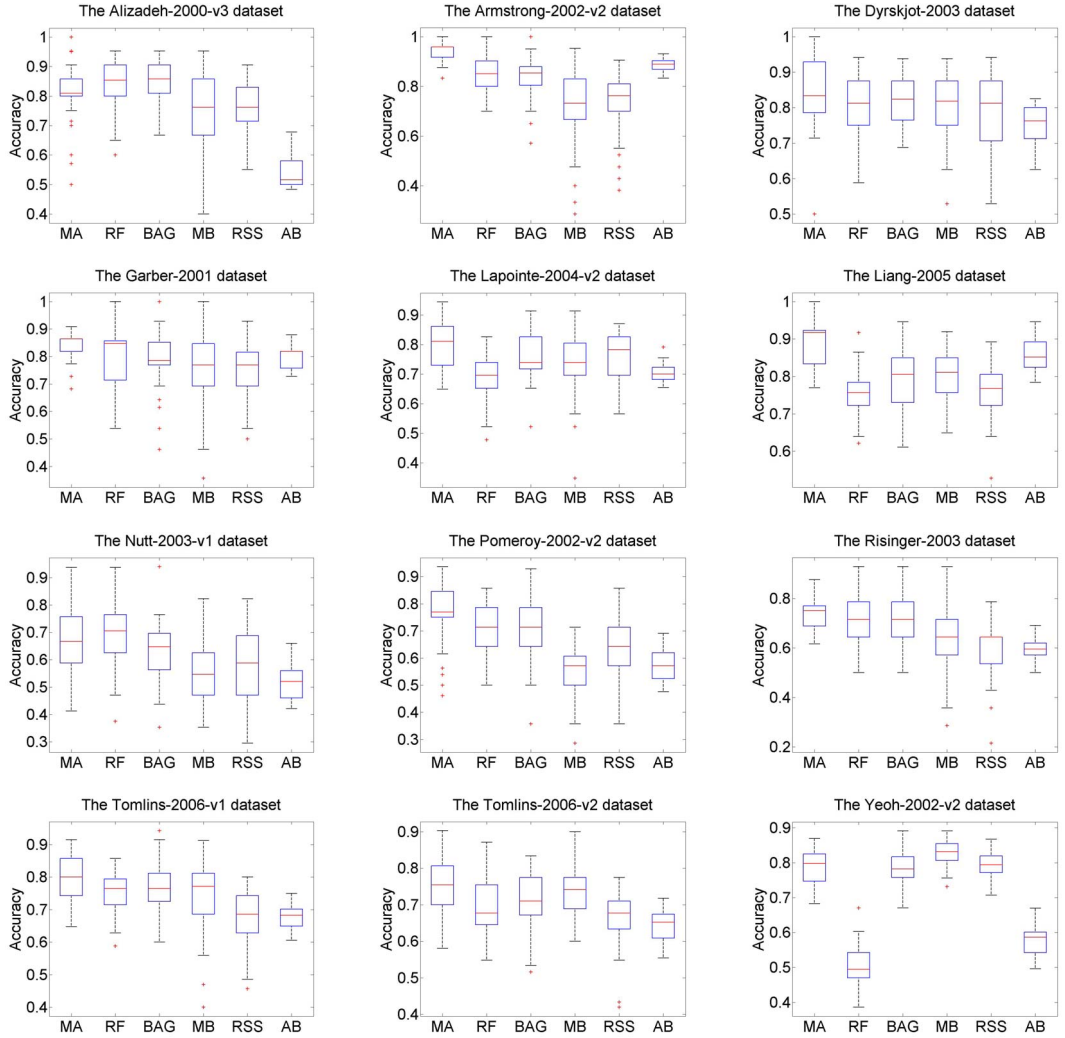

Figure 4. The comparison of classification accuracies and stabilities between MV-ACE(MA) and RF, BAG, MB, RSS, and AB on gene expression data sets.

Table 2. The effect of the optimization strategy in MV-ACE (bold font denotes the significant difference)

Datasets	MV-ACE	MV-ACE-NO	superiority
Alizadeh-2000-v3	0.8149±0.1056	0.5763±0.0928	**0.2386**
Armstrong-2002-v2	0.9424±0.0461	0.9333±0.0516	0.0090
Dyrskjot-2003	0.8429±0.0953	0.7740±0.0959	**0.0688**
Garber-2001	0.8447±0.0504	0.8386±0.0600	0.0061
Lapointe-2004-v2	0.7951±0.0849	0.7743±0.0618	0.0208
Liang-2005	0.8941±0.0467	0.8824±0.0544	0.0118
Nutt-2003-v1	0.6692±0.1112	0.5904±0.1198	**0.0788**
Pomeroy-2002-v2	0.7867±0.1078	0.7309±0.1015	**0.0558**
Risinger-2003	0.7252±0.0756	0.6873±0.1145	0.0379
Tomlins-2006-v1	0.7961±0.0633	0.7200±0.0635	**0.0761**
Tomlins-2006-v2	0.7524±0.0806	0.7136±0.0659	0.0389
Yeoh-2002-v2	0.7864±0.0485	0.6843±0.0388	**0.1020**

cally, this approach adopts the multi-view technique and the adaboost algorithm to improve the base classifiers' performance. Besides, we also optimize the multi-view models with a newly designed optimization process to search for the best classifier ensemble model. The comparison between MV-ACE and several well-known ensemble classifier approaches demonstrate the effective and stability of MV-ACE.

Acknowledgment

The work described in this paper was partially funded by the grant from the Hong Kong Scholars Program (Project No. XJ2012015), the Guangdong Natural Science Funds for Distinguished Young Scholar (Project No. S2013050014677), a grant from the Fundamental Research Funds for the Central Universities (Project No. 2014G0007), the grants from NSFC (Project No. U1035004, 61273363, 61379033), the grants (Project No. 11A11080267, KC2013ZDZJ0007A, 2013M540655, NCET-11-0165, S2012010009961, 20110172120027, 2011B090400032), a grant from the City University of Hong Kong (Project No. 7004047) and the grants from the Hong Kong Polytechnic University (G-YK77 and G-YK53).

References

[1] Golub T R, Slonim D K, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, 286(5439), 531-537.

[2] Statnikov A, Wang L, Aliferis C F. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC bioinformatics, 2008, 9(1): 319.

[3] Wang X, Gotoh O. A robust gene selection method for microarray-based cancer classification. Cancer informatics, 2010, 9: 15.

[4] Zhang H, Wang H, Dai Z, et al. Improving accuracy for cancer classification with a new algorithm for genes selection. BMC bioinformatics, 2012, 13(1): 1-20.

[5] Ghorai S, Mukherjee A, et al. Cancer classification from gene expression data by NPPC ensemble. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2011, 8(3): 659-671.

[6] Nanni L, Brahnam S, Lumini A. Combining multiple approaches for gene microarray classification. Bioinformatics, 2012, 28(8): 1151-1157.

[7] Li L, Zhang Y, et al. An ensemble classifier for eukaryotic protein subcellular location prediction using gene ontology categories and amino acid hydrophobicity. PloS one, 2012, 7(1): e31057.

[8] Jones M, Viola P. Fast multi-view face detection. Mitsubishi Electric Research Lab TR-20003-96, 2003, 3: 14.

[9] Freund Y, Schapire R E. A desicion-theoretic generalization of on-line learning and an application to boosting. Computational learning theory. Springer Berlin Heidelberg, 1995: 23-37.

[10] Webb G I. Multiboosting: A technique for combining boosting and wagging. Machine learning, 2000, 40(2): 159-196.

[11] de Souto M C P, Costa I G, de Araujo D S A, et al. Clustering cancer gene expression data: a comparative study. BMC bioinformatics, 2008, 9(1): 497.

[12] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter, 2009, 11(1): 10-18.