GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Methodology

Multiclass microarray data classification based on confidence evaluation

H.L. Yu[1], S. Gao[1], B. Qin[1] and J. Zhao[2]

[1]School of Computer Science and Engineering,
Jiangsu University of Science and Technology, Zhenjiang, China
[2]College of Computer Science and Technology,
Harbin Engineering University, Harbin, China

Corresponding author: H.L. Yu
E-mail: yuhualong@just.edu.cn

Genet. Mol. Res. 11 (2): 1357-1369 (2012)
Received September 5, 2011
Accepted February 7, 2012
Published May 15, 2012
DOI http://dx.doi.org/10.4238/2012.May.15.6

ABSTRACT. Microarray technology is becoming a powerful tool for clinical diagnosis, as it has potential to discover gene expression patterns that are characteristic for a particular disease. To date, this possibility has received much attention in the context of cancer research, especially in tumor classification. However, most published articles have concentrated on the development of binary classification methods while neglected ubiquitous multiclass problems. Unfortunately, only a few multiclass classification approaches have had poor predictive accuracy. In an effort to improve classification accuracy, we developed a novel multiclass microarray data classification method. First, we applied a "one versus rest-support vector machine" to classify the samples. Then the classification confidence of each testing sample was evaluated according to its distribution in feature space and some with poor confidence were extracted. Next, a novel strategy, which we named as "class priority estimation method based on centroid distance", was used to make decisions about categories for those poor confidence samples. This approach was tested on seven benchmark multiclass microarray datasets, with encouraging results, demonstrating effectiveness and feasibility.

Key words: Microarray data; Multiclass classification; OVR-SVM; Support vector machine; Confidence evaluation; Priority estimation

INTRODUCTION

The advent of DNA microarray, one of the most important technological advances in the post-genomic era, has allowed the simultaneous measurement of the expression levels of thousands of genes in a single experiment. It has also facilitated the diagnosis of diseases, especially tumors, at the molecular level (Alon et al., 1999; Golub et al., 1999). In recent years, increasing researchers from different research fields, such as biology, medicine, computer science, and even statistics, have been interested in this field and proposed mass useful microarray data mining approaches and tools on the basis of their domain knowledge. These approaches and tools were developed for 3 main purposes: 1) extracting feature genes that have a close relationship with a particular disease to help doctors improve the clinical diagnostic accuracy, biologists to determine the genetic nature of a disease (Guypn et al., 2002), and medical experts to rapidly discover new medicines (Evans and Guy, 2004); 2) clustering microarray data to identify new subtypes of a particular disease to improve the efficacy of clinical treatment (Armstrong et al., 2002); 3) constructing classification models for making accurate diagnosis of diseases. The purposes mentioned above clarify that classification is one of the most attractive issues in the field of microarray data mining.

In recent years, many classification methods and tools have been applied to stratify microarray data. These approaches include K nearest neighbors classifier (Li et al., 2001), support vector machine (SVM) (Furey et al., 2000), C4.5 decision tree (Horng et al., 2009), Bayesian classifier (Asyali, 2007), and some ensemble classification methods (Chen and Zhao, 2008; Kim and Cho, 2008; Yu et al., 2010). However, most of these approaches can only be used in binary classification tasks (tumor versus normal/2 subtypes of a tumor) but are not appropriate for multiclass data (multiple tumors/multiple subtypes of a tumor). Classification of multiclass samples is more difficult than that of binary class samples. To effectively resolve this issue, researchers have proposed some flexible strategies to transform multiclass classifiers to multiple binary classifiers. For example, Yeang et al. (2001) combined weighted voting, K nearest neighbors classifier, and SVM with one versus one (OVO) and one versus rest (OVR) decomposition schemes to recognize multiclass microarray samples; similarly, Shen and Tan (2006) compared the performance of several output coding and decoding functions for multiclass classifications. In addition, some direct approaches are available to develop multiclass extension of traditional binary classifiers, such as multicategory SVMs, which was developed by Lee and Lee (2003) and often lead to a complex optimization problem. Tan et al. (2004) applied discriminant partial least squares to predict the categories for multiclass samples and Berrar et al. (2006) constructed an instance-based multiclass microarray data classification approach. Both Li et al. (2004) and Statnikov et al. (2005) performed systematic and comprehensive evaluation of several major multiclass classification methods for microarray data and concluded that OVR-SVM outperforms other approaches because it has averagely higher classification accuracy.

To achieve better predictive accuracy, we developed a novel multiclass microarray data classification approach. First, we applied OVR-SVM to classify testing samples. Next, we used the approach proposed by Yeang et al. (2001) and developed a novel method, confidence evaluation, and used it to estimate classification confidence of each sample and extract some unconfident ones. Finally, we used a strategy, class priority estimation method based on centroid distance, to mediate conflicting SVMs and categorize the unconfident samples. Since SVMs provide better results with high-dimensional data, the original dataset was directly used to construct SVMs, whereas the class priority estimation method based on centroid distance was used with a dataset consisting of only a few feature genes in order to guarantee estimation accuracy. The proposed approach was validated on 7 benchmark multiclass microarray datasets, and experimental results have proved its effectiveness and feasibility.

MATERIAL AND METHODS

Multiclass SVM

SVM introduced by Vapnik (1998) is a valuable tool for solving pattern classification problem. Compared with traditional classification methods, SVM possesses several prominent advantages as follows: 1) high generalization capability; 2) absence of local minima, and 3) suitability for high-dimensional and small-sample datasets. SVM was initially designed to solve the binary classification problem. The main idea of binary SVM was to implicitly map data to a higher dimensional space by using a kernel function and solve an optimization problem to identify the maximum-margin hyperplane that separates the 2 class training instances. New instances were classified according to the side of the hyperplane they fall into (see Figure 1).

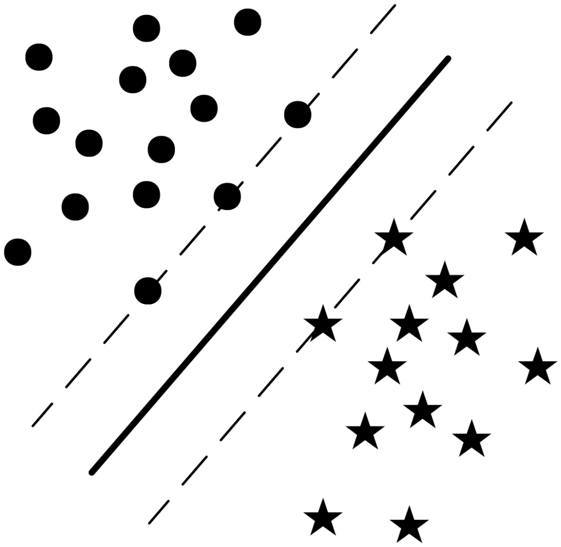

Figure 1. A binary support vector machine (SVM) classifier constructs a hyperplane (bold line) to maximize the margin between two classes (circles and pentagrams). The samples emerged on the dashed lines are called as support vectors. New instances will be classified into the side of the hyperplane they fall into.

Given dataset $S=\{(x_i,y_i)\,|\,x_i\in R^d, y_i\in\{-1,+1\}, i=1,...N\}$, where x_i is a *d*-dimension sample, y_i is the corresponding class label, and *N* is the number of samples. The discriminant function of SVM can be described as follows:

$$g(x)=\text{sgn}(\sum_{i=1}^{sv}\alpha_i y_i K(x,x_i)+b) \qquad \text{(Equation 1)}$$

In Equation 1, *sv* is the number of support vectors, α_i is a Lagrange multiplier, *b* is the bias of optimum classification hyperplane, and $K(x, x_i)$ denotes the kernel function. In our experiments, radial basis kernel function (RBF) has been used:

$$K(x_i,x_j)=\exp\left\{-\frac{|x_i-x_j|^2}{2\sigma^2}\right\} \qquad \text{(Equation 2)}$$

A complete description of the SVM theory for pattern recognition is provided by Vapnik (1998). To apply SVM to a multiclass classification problem, a decomposition strategy such as OVO and OVR is required. As 2 representative schemas, OVO strategy trains *C**(*C*-1)/2 (*C* is the number of class) SVMs, where each SVM is constructed on the basis of the samples of any 2 different classes, whereas the OVR method only needs to train *C* SVM classifiers, where each classifier selects the samples of the corresponding class as positive instances and all other samples as negative instances. When 1 testing sample *x*' is classified, the outputs of all SVMs will be combined by some reconstruction strategies such as OVO strategy that outputs the class with the most votes and OVR strategy that assigns the category whose corresponding SVM has produced the highest value. The reconstruction strategy of OVR method is as follows:

$$f(x')=\arg\max_{i=1,2,...,C} f_i(x') \qquad \text{(Equation 3)}$$

Unlike OVR, OVO generally constructs more classifiers and has lower classification accuracy owing to only a few samples used to train each SVM. Therefore, the OVR strategy seems to be more suitable for practical use than the OVO schema. In this study, OVR decomposition strategy combined with SVM classifier were applied to initially classify multiclass microarray data.

Confidence evaluation

Figure 2 shows the schema of OVR decomposition strategy combined with the SVM classifier. Without losing generality, 3-class samples are assumed. As shown in Figure 2, OVR-SVM constructs 3 hyperplanes for the 3-class training samples: A, B, and C, whereas A'~C' are real classification hyperplanes for the 3 classes based on the OVR-SVM strategy.

These 6 hyperplanes divide the feature space into 12 areas. When a testing sample is classified, the area into which it falls is studied. If it falls into Area X (1~3), that means only the SVM corresponding with class X has produced a positive value, while all other SVMs have output negative values. On the other hand, if it lies in Area X/Y, then that means both SVMs for class X and Y have regarded it as a positive instance, but it is assigned to class X considering it is closer with hyperplane X. Otherwise, the testing sample will emerge in Area X' (1'~3'), which means it has been excluded by all SVMs and allocated to class X since it is nearer to the training samples of class X. Samples x and y existing in Area 3/1 have been classified into the third class; however, classification confidence of sample x is higher than that of sample y because x is farther from the real classification hyperplane.

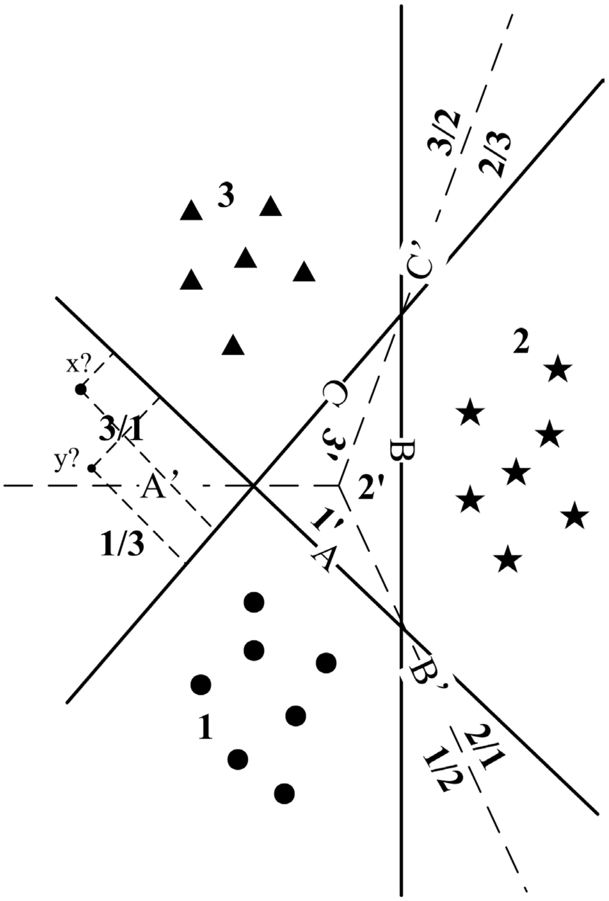

Figure 2. Schema of one versus rest-support vector machine (OVR-SVM) for three-class samples. Circles, pentagrams and triangles represent three-class samples, respectively.

The above descriptions suggest that the farther the distance between the testing sample and any real hyperplane, higher would be the classification confidence. After classification of testing samples by OVR-SVM, the confidence of each sample was evaluated, and the samples were divided into 3 groups as follows:

- High confidence: only 1 SVM produces positive output value for the testing sample, i.e., the sample is doubtlessly assigned into 1 specific category and also excluded by all other classes (see Figure 3A).
- Medium confidence: multiple or none SVMs produce positive values, and the distances between the highest output value and any other output values are higher than a given threshold *T* (see Figure 3B and C). This sample may be considered as a member of the categories probably corresponding to the highest output value.
- Low confidence: multiple or none SVMs produce positive values, and the distances between the highest output value and other output values are lower than a given threshold *T* (see Figure 3D and E). In this situation, the real class of the sample is uncertain. The SVM with the highest output value and those for whom distances between their outputs and the highest output value are lower than the given threshold *T* are known as conflicting classifiers, such as the third and fourth classifier in Figure 3.

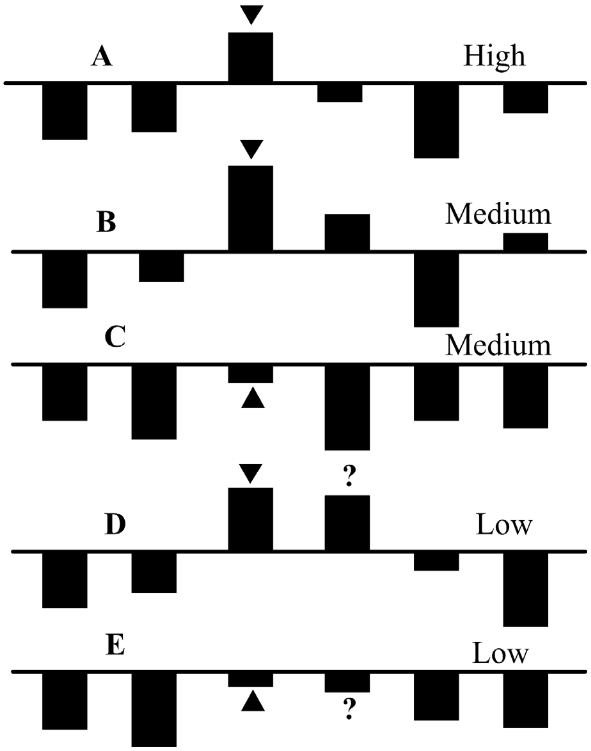

Figure 3. Schema of the classification confidence evaluation based on one versus rest-support vector machine (OVR-SVM) strategy. **A** = High confidence; **B** and **C** = medium confidence; **D** and **E** = low confidence.

The principle of the classification confidence evaluation has been described above. In this study, only low confidence samples have been studied for 2 purposes: improving predictive accuracy and saving recognition time.

Class priority estimation based on centroid distance

Classification based on microarray data is considerably different from previous classification problems in that the number of genes (typically tens of thousands) greatly exceeds the number of samples (typically less than a few hundreds). This results in the well-known problem of "curse of dimensionality" and over-fitting of the training data (Dougherty, 2001). Thus, for successful disease diagnosis, selection of a small number of discriminative genes from thousands of genes is important (Debnath and Kurita, 2010).

In recent years, various feature gene selection methods have been proposed. Most of them have been found to be helpful for improving the predictive accuracy of disease and providing useful information for biologists and medical experts. These feature gene selection methods can be grouped into 2 teams: filter, which is also called as gene ranking approach, and wrapper, which is also entitled as gene subset selection approach (Inza et al., 2004). In the filter approach, each gene is evaluated individually and assigned a score reflecting its correlation with the class according to certain criteria. Genes are then ranked by their scores, and some top-ranked ones are selected. In the wrapper approach, the space of genes is searched to evaluate the goodness of each found gene subset by estimating the accuracy percentage of the specific classifier to be used, training the classifier only with the found genes. Compared with the filter approach, wrapper approach generally obtains 1 gene subset with better classification performance but at a considerable computational cost.

Considering time complexity, the filter approach seemed to be more suitable for our study. In this study, we modified the well-known signal-noise ratio feature gene selection method proposed by Golub et al. (1999) and applied it to extract feature genes in multiclass microarray datasets. The modified method is described as follows:

$$S(g)=\sum_{i=1}^{C}\frac{|m_i-m'_i|}{s_i+s'_i} \qquad \text{(Equation 4)}$$

where m_i and m'_i denote mean values of the samples in the *ith* category and all other classes, and s_i and s'_i represent standard deviations of these values, respectively. In addition, C is the number of class. After S value of all genes are computed, they are sorted in descending order, and some top-ranked ones are selected as feature genes.

To decide which conflicting classifier should be classified into one low confident testing sample, we developed a novel strategy, namely, class priority estimation method based on centroid distance. In the feature gene space, the centroid of each class in the *ith* gene was computed as follows:

$$\overline{m_{ik}}=1/n_k\sum_{j\in C_k}g_{ij} \qquad \text{(Equation 5)}$$

where n_k is the number of the samples in the *kth* class, j represents 1 sample in the *kth* class, and g_{ij} is the gene expression value of the *jth* sample in the *ith* gene. After the centroids of all feature genes were computed, the centroid of each class was acquired.

Figure 4 shows that when the class label of a testing sample "?" is inquired, the distances from the sample to centroids of each class are first calculated, and class priorities are then obtained by sorting these distances in an ascending order. In Figure 4, the class priorities of the 3 categories are 1 > 2 > 3. Obviously, the higher priority a class acquires, the more is the probability for a testing sample to be classified in that class.

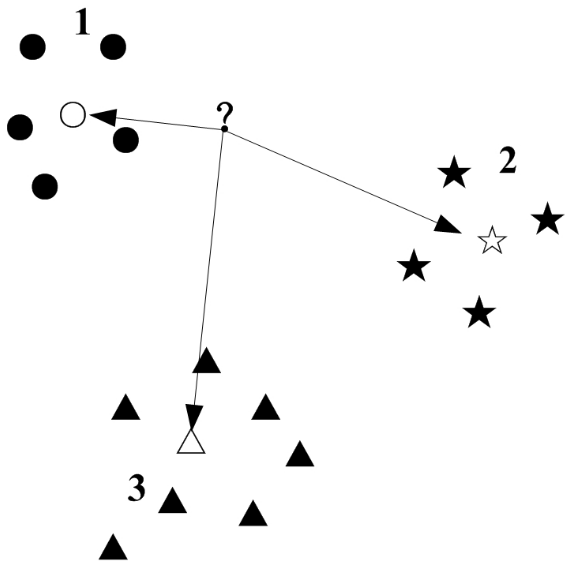

Figure 4. Schema of the class priority estimation based on centroid distance. Circles, pentagrams and triangles represent centroids of three-class samples, respectively. Class priorities are sorted based on distance from the testing sample "?" to each centroid in ascending order, i.e., 1 > 2 > 3.

Proposed approach

Combining OVR-SVM based on confidence evaluation with the class priority estimation method based on centroid distance, we proposed a novel multiclass microarray data classification approach. When this approach is used to classify a testing sample *x*, the computation procedure is as follows:

- Input: testing sample *x*
- Output: class label *y* for sample *x*
- Classification procedure:

Step 1: Apply OVR-SVM to classify testing sample *x*;

Step 2: Evaluate classification confidence, if classification result is confident, then go to Step 5, otherwise, continue to Step 3;

Step 3: Compute the distances from *x* to the centroids of each class in the feature gene space and rank priority by sorting these distances in an ascending order;

Step 4: Compare priorities of each conflicting classifier and then select the class with the highest priority from these conflicting classifiers as a class label for sample *x*;

Step 5: Output class label *y* for sample *x*.

Datasets used in this study

Seven benchmark multiclass microarray datasets (Golub et al., 1999; Khan et al., 2001; Staunton et al., 2001; Su et al., 2001; Armstrong et al., 2002; Pomeroy et al., 2002; Nutt et al., 2003) used in this study are described in Table 1. The 7 datasets have 3-11 distinct categories, 50-174 samples, and 2308-12533 genes. All datasets are available at http://www.gems-system.org.

Table 1. Datasets used in this study.

Datasets	Task descriptions	Samples	Genes	Categories	Literature
11_Tumors	11 various human tumor types	174	12,533	11	Su et al., 2001
NCI60	9 various human tumor types	60	5,726	9	Staunton et al., 2001
SRBCT	Small, round blue cell tumors of childhood	83	2,308	4	Khan et al., 2001
Brain1	5 human brain tumor types	90	5,920	5	Pomeroy et al., 2002
Brain2	4 malignant glioma types	50	10,367	4	Nutt et al., 2003
Leukemia1	Acute myelogenous leukemia (AML), acute lymphoblastic leukemia (ALL) B-cell and ALL T-cell	72	5,327	3	Golub et al., 1999
Leukemia2	AML, ALL and mixed-lineage leukemia (MLL)	72	11,225	3	Amstrong et al., 2002

Data are reported as numbers for samples, genes and catagories.

RESULTS AND DISCUSSION

Our experiments were performed in Matlab environment, where multiclass SVM was used with statistics pattern recognition toolbox written by Franc and Hlavac (2004) and the RBF was used as kernel function. The parameter σ of the RBF kernel function was assigned as 10, and the penalty factor *C* was 500. In addition, leave-one-out cross-validation (LOOCV) was performed to test the predictive accuracy because of the small sample size. In LOOCV, one of all samples is evaluated as a testing instance, and the others are used as training data. After each sample is used as testing data for once, the predictive accuracy is obtained as the ratio of the number of the correctly classified samples and the total number of samples in the dataset. The class priority estimation method based on centroid distance was conducted using 50 feature genes.

First, the predictive accuracy of the following 4 classification approaches was tested: OVO-SVM, OVR-SVM, the highest priority based on centroid distance, and the proposed combined approach. In the proposed approach, threshold *T* was assigned as 0.1, 0.3, and 0.5 in order to determine the relationship between classification accuracy and threshold. The detailed classification results are shown in Table 2.

Table 2. Classification results of four approaches in seven datasets (%).

Datasets	OVO-SVM	OVR-SVM	Highest priority	Proposed approach		
				0.1	0.3	0.5
11_Tumors	82.8	85.6	86.2	**89.1**	87.9	86.8
NCI60	53.3	63.3	58.3	63.3	**66.7**	63.3
SRBCT	98.8	98.8	98.8	**100.0**	**100.0**	98.8
Brain1	84.4	85.6	63.3	**87.8**	**87.8**	**87.8**
Brain2	72.0	**78.0**	72.0	**78.0**	74.0	72.0
Leukemia1	95.8	94.4	95.8	**97.2**	**97.2**	**97.2**
Leukemia2	90.3	87.5	88.9	90.3	**91.7**	**91.7**
Average	82.5	84.7	80.5	**86.5**	**86.5**	85.4

OVO = one versus one; OVR = one versus rest; SVM = support vector machine. Bold numbers represent the highest classification accuracies for the corresponding data set.

The results showed that.

- OVR-SVM is superior to OVO-SVM and the highest priority estimation method based on centroid distance because it yielded higher predictive accuracy in majority datasets, indicating that OVR-SVM should be used as the initial classifier.
- Irrespective of the threshold used, the proposed combined approach acquired better recognition rate than other classification methods, indicating that it is effective and feasible.
- Selecting an appropriate threshold for the proposed approach seems difficult. Both thresholds 0.1 and 0.3 produced the same average classification accuracy of 86.5% in all the 5 datasets. In contrast, the proposed approach with threshold 0.5 yielded the worst classification performance. This might be because large thresholds consider some confident samples as uncertain ones and make wrong decisions for them.

The number and error rate of samples that were classified by OVR-SVM with high confidence, medium confidence, and low confidence based on threshold 0.1 in all the 7 datasets were then counted. The statistical results are shown in Table 3. The results suggested that most errors emerged in low and medium confidence samples. Of the 214 high confidence samples, only 5 were misclassified; on the other hand, of the 43 low confidence samples, 28 were misclassified. This strongly suggests that errors tended to appear around the real classification hyperplane of multiple classes. Samples from those regions are easily confused with other classes. The experimental results shown in Table 3 also explain why only low confidence samples were estimated by the class priority estimation method.

Table 3. Number and error rate of samples based on different classification confidence.

	Classification confidence		
	High	Medium	Low
Number of samples	214	344	43
Error rate	2.3%	12.2%	65.1%

The *11_Tumors* dataset was used as an example to analyze the performance of the proposed approach in detail. The *11_Tumors* dataset contains 11 tumor types (type/size of samples): 1) bladder and ureter carcinomas/8; 2) breast carcinomas/26; 3) colorectal carcinomas/23; 4) gastroesophagus carcinomas/12; 5) kidney carcinomas/11; 6) liver carcinomas/7; 7) prostate carcinomas/26; 8) pancreas carcinomas/6; 9) lung adenocarcinomas/14; 10) lung squamous cell carcinomas/14, and 11) ovary carcinomas/27. A confusion matrix for both OVR-SVM classifier and the proposed classification approach with a threshold of 0.1 on the *11_Tumors* dataset is presented in Table 4. The data from this table shows that 100% classification accuracy was obtained for colorectal carcinomas and prostate carcinomas. For the other tumor types having many samples, such as breast and ovary carcinomas, more than 90% predictive accuracy was obtained. The more positive samples were used to train SVM, the more accurate was the results of the SVM classifier. In addition, several samples that were misclassified by OVR-SVM were modified by our proposed method.

Table 4. Confusion matrix for both of one versus rest-support vector machine (OVR-SVM) classifier and the proposed classification approach on *11_Tumors* dataset.

Real class	Predictive class (OVR-SVM/proposed approach)											
	1	2	3	4	5	6	7	8	9	10	11	Total
1	4/5		1/1	1/0				1/1		1/1		8
2		25/25				1/1						26
3			23/23									23
4		1/1	1/1	8/8					2/2			12
5					9/10				2/1			11
6		4/2				3/5						7
7							26/26					26
8					1/1			4/4	1/1			6
9		1/1							10/11	3/2		14
10			1/1						1/0	12/13		14
11		2/2									25/25	27
Total	4/5	33/31	26/26	9/8	10/11	4/6	26/26	5/5	16/15	16/16	25/25	174

Table 5 shows some examples classified by the proposed approach in different situations. The classification result of the first sample by OVR-SVM showed high confidence; thus, evaluation by class priority estimation method was not required. For the 20th sample, the 2nd and 11th SVM conflicted with each other, and hence, it was classified into the second class because the priority of the second class was higher than that of the 11th class. There were 3 conflicting classifiers for the 63rd sample, and the misclassified result could be modified by the class priority estimation method. This shows that the proposed approach could be used to improve the predictive accuracy.

Table 5. Examples of analysis.

Example	1	2	3	4	5	6	7	8	9	10	11	Category
Sample 1												
OVR-SVM	-0.01	-0.49	-0.21	-0.10	-0.02	-0.23	**1.23**	-0.04	-0.06	-0.08	-0.07	7
Priority estimation	5	10	9	3	2	4	**1**	11	8	6	7	
Sample 20												
OVR-SVM	-0.24	**0.28**	0.02	0.04	-0.11	0.05	-0.07	-0.06	-0.06	-0.06	**0.21**	2
Priority estimation	1	**3**	8	10	4	6	11	9	5	2	**7**	
Sample 67												
OVR-SVM	-0.23	**0.17**	0.00	0.00	-0.09	**0.12**	-0.02	-0.05	0.02	-0.02	**0.10**	6
Priority estimation	5	**3**	9	10	2	**1**	11	8	6	4	**7**	

Bold numbers represent the classification confidences and priorities for the conflicted classifiers, respectively.

Table 6 summarizes the detailed classification results of OVR-SVM and the highest priority estimation method. The results showed that 138 and 13 samples were classified correctly and incorrectly by both the approaches, respectively. Fortunately, the proposed method modified the classification for 6 samples that were misclassified by OVR-SVM but were classified accurately by the highest priority estimation method. Meanwhile, all the 11 samples that were misclassified by the highest priority estimation method but were classified correctly by OVR-SVM did not fall into the low confidence regions of OVR-SVM.

Since OVR-SVM was constructed using the original training set, and only a few feature genes were selected to calculate the centroid of each class in highest priority estimation method, there were distinct differences between them. Thus, the possibility of making errors for classification in different regions was higher. Fortunately, the regions that tended to show errors for OVR-SVM were identified, and the class priority estimation method could obtain better recognition in these regions. This explains why the proposed approach could produce higher classification accuracy than both OVR-SVM classification method and the highest priority estimation method.

Table 6. Detailed classification results of one versus rest-support vector machine (OVR-SVM) and the highest priority estimation method.

Proposed approach (correct/incorrect)		OVR-SVM	
		√	×
Highest priority	√	138 (138/0)	12 (6/6)
	×	11 (11/0)	13 (0/13)

√ = correct classification; × = incorrect classification.

CONCLUSIONS

Herein, we proposed a novel multiclass microarray data classification approach that combined the OVR-SVM classification schema with the class priority estimation method. The main findings of this study are as follows:

1. Presenting a novel strategy, namely confidence evaluation, and applying it to extract the samples that are more possible to be predicted incorrectly by OVR-SVM.

2. Proposing class priority estimation method based on centroid distance that can be used to not only estimate priority of each category but also directly classify testing samples.

3. Combining the advantages of the 2 methods to predict class labels for testing samples in order to improve the classification performance.

Extensive experiments on the 7 datasets were conducted, and the experimental results showed that the proposed approach was more effective than the traditional methods since it had resulted in higher classification accuracy.

ACKNOWLEDGMENTS

Research partially supported by the National Natural Science Foundation of China under grant #60873036, the Ph.D Programs Foundation for young researchers of the Ministry of Education of China under Grant #20070217051, and the Ph.D Foundation of Jiangsu University of Science and Technology under Grant #35301002.

REFERENCES

Alon U, Barkai N, Notterman DA, Gish K, et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U. S. A.* 96: 6745-6750.

Armstrong SA, Staunton JE, Silverman LB, Pieters R, et al. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.* 30: 41-47.

Asyali MH (2007). Gene expression profile class prediction using linear Bayesian classifiers. *Comput. Biol. Med.* 37: 1690-1699.

Berrar D, Bradbury I and Dubitzky W (2006). Instance-based concept learning from multiclass DNA microarray data. *BMC Bioinformatics* 7: 73.

Chen YH and Zhao YO (2008). A novel ensemble of classifiers for microarray data classification. *Appl. Soft Comp.* 8: 1664-1669.

Debnath R and Kurita T (2010). An evolutionary approach for gene selection and classification of microarray data based on SVM error-bound theories. *Biosystems* 100: 39-46.

Dougherty ER (2001). Small sample issues for microarray-based classification. *Comp. Funct. Genomics* 2: 28-34.

Evans WE and Guy RK (2004). Gene expression as a drug discovery tool. *Nat. Genet.* 36: 214-215.

Franc V and Hlavac V (2004). Statistical Pattern Recognition Toolbox for Matlab User's Guide. Available at [http://www.geneticsmr.com/sites/all/files/Example_of_reference_style.pdf]. Accessed August 26, 2011.

Furey TS, Cristianini N, Duffy N, Bednarski DW, et al. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16: 906-914.

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

Guypn I, Weston J, Barnhill S and Vapnik V (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46: 389-422.

Horng JT, Wu LC, Liu BJ and Kuo JL (2009). An expert system to classify microarray gene expression data using gene selection by decision tree. *Expert. Syst. Appl.* 36: 9072-9081.

Inza I, Larranaga P, Blanco R and Cerrolaza AJ (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif. Intell. Med.* 31: 91-103.

Khan J, Wei JS, Ringner M, Saal LH, et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.* 7: 673-679.

Kim KJ and Cho SB (2008). An evolutionary algorithm approach to optimal ensemble classifiers for DNA microarray data analysis. *IEEE Trans. Evol. Comput.* 12: 377-388.

Lee YK and Lee C-K (2003). Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics* 19: 1132-1139.

Li LP, Weinberg CR, Darden TA and Pedersen LG (2001). Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics* 17: 1131-1142.

Li T, Zhang CL and Ogihara M (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics* 20: 2429-2437.

Nutt CL, Mani DR, Betensky RA, Tamayo P, et al. (2003). Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res.* 63: 1602-1607.

Pomeroy SL, Tamayo P, Gaasenbeek M and Sturla LM (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415: 436-442.

Shen L and Tan EC (2006). Reducing multiclass cancer classification to binary by output coding and SVM. *Comput. Biol. Chem.* 30: 63-71.

Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, et al. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* 21: 631-643.

Staunton JE, Slonim DK, Coller HA, Tamayo P, et al. (2001). Chemosensitivity prediction by transcriptional profiling. *Proc. Natl. Acad. Sci. U. S. A.* 98: 10787-10792.

Su AI, Welsh JB, Sapinoso LM, Kern SG, et al. (2001). Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res.* 61: 7388-7393.

Tan Y, Shi L, Tong W, Hwang GT, et al. (2004). Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models. *Comput. Biol. Chem.* 28: 235-244.

Vapnik V (1998). Statistical Learning Theory. Wiley Publishers, New York.

Yeang CH, Ramaswamy S, Tamayo P, Mukherjee S, et al. (2001). Molecular classification of multiple tumor types. *Bioinformatics* 17 (Suppl 1): S316-S322.

Yu HL, Gu GC, Liu HB and Shen J (2010). Feature subspace ensemble classifiers for microarray data. *ICIC Express Lett.* 4: 143-147.