A Comparison of SVM-based Evolutionary Methods for Multicategory Cancer Diagnosis using Microarray Gene Expression Data

Rameswar Debnath, Haruhisa Takahashi
Department of Informatics
The University of Electro-Communications
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan
and
Takio Kurita
Department of Information Engineering
Hiroshima University, 1-4-1 Kagamiyama
Higashi Hiroshima, Hiroshima 739-8527, Japan

Abstract

Selection of relevant genes that will give higher accuracy for sample classification (for example, to distinguish cancerous from normal tissues) is a common task in most microarray data studies. An evolutionary method based on generalization error bound theory of support vector machine (SVM) can select a subset of potentially informative genes for SVM classifier very efficiently. The bound theories are developed for binary SVM, however multiclass SVMs do not have established bounds on the generalization error. Several multiclass SVMs have been proposed where multiclass SVMs are typically constructed by combining several binary SVMs. We evaluate an estimate of a generalization error bound for a multiclass SVM by combining the error bound of binary SVMs which are used to construct the multiclass SVM. In this paper our aims are to compare the performance of several multiclass SVMs in the SVM-based evolutionary method and then find the best multiclass SVM classifier in the SVM-based evolutionary method for multicategory cancer diagnosis using microarray gene expression data.

Keywords: support vector machine, multiclass classifier, generalization error bound, evolutionary algorithm, microarray data

1 Introduction

Patient samples for bioinformatic analyses are fairly small in number compared to the number of genes investigated such as microarray datasets [1]. The vast amount of raw gene expression data leads to statistical and analytical challenges including the classification of datasets into correct classes [2]. In machine-learning terminology, these data sets have high dimension and small sample size. However, the data management system allows researchers to gather a number of genes of ever-increasing size, many of which are irrelevant to the distinction of samples. These irrelevant genes have negative effect on the accuracy of a classifier. The microarray data also contain technical and biological noise. Selection of relevant genes that will give higher accuracy for sample classification (for example, to distinguish cancerous from normal tissues) is a common task in most microarray data studies. There exist several ranking based and evolutionary computation methods for gene selection in the microarray data [1]-[5, 9, 11]. Gene selection can moderately or significantly improve the performance of classifiers [4]. Gene selection by evolutionary method and a good choice of a classifier can outperform others [1].

Recently, Debnath and Kurita have proposed an evolutionary method which selects and recombines gene features based on SVM error bound values and an SVM evaluates the fitness function [5]. Thus, selected genes directly reflect to some extent the performance of the SVM classifier unlike the conventional methods which select and recombine genes using genetic algorithm (GA) based approaches that are independent of the algorithm to be used to construct the classifiers. The bound theories are developed for binary SVM, however multiclass SVMs do not have established bounds on the generalization error. Several multiclass SVMs have been proposed where multiclass SVMs are typically constructed by combining several binary SVMs [3, 7, 12, 13]. For multicategory diagnosis, gene features are selected by combining error bounds of binary SVMs which are used to construct the multiclass SVM and the multiclass SVM evaluates the fitness function in the SVM-based evolutionary method. Multiclass SVMs such as one-versus-one and DAGSVM methods construct $k(k-1)/2$ binary SVM classifiers for a k-class problem where each binary SVM classifier is constructed using data from two classes out of k classes whereas the other methods such as one-versus-rest, alltogether, and Crammer and Singer methods construct

k binary SVM classifiers where each binary SVM classifier is constructed using all the training data, and the classifier separates one class from the others [7, 12]. Thus, multiclass SVMs are basically two types in the case of constructing classifiers, however their decision functions are different. In [4], Statnikov et al. have compared various multiclass SVMs (MC-SVMs) with microarray gene expression datasets and empirically found that the one-versus-rest method has a superior classification performance (on an average) than the others. On the other hand, the one-versus-one method fits perfectly the known characteristics of the binary SVM, where the borderlines between classes are computed directly.

In this paper, we compare the performance of one-versus-one and one-versus-rest MC-SVM methods in the SVM-based evolutionary method proposed in [5]. Both linear and nonlinear kernels are investigated to evaluate the performance of multiclass SVMs. We perform experiments with binary SVM's Opper-Winther bound [10], Zhou-Tuck bound [11], and radius-margin bound [13] for feature selection. Numerous experiments on several data present that a small number of genes can linearly separate datasets into classes with highest classification accuracies. We also find that the one-versus-one MC-SVM shows slightly better results than the one-versus-rest MC-SVM in the evolutionary method.

2 Methods and Materials

2.1 SVM Classifier

The SVM is arguably the single most important development in supervising learning area. Theoretically, the SVM approximately implements the structural risk minimization principle, thus the SVM is situated on a strong theoretical foundation. It has no local minima, i.e., it solves a convex optimization problem. The algorithm can automatically determine a network architecture. It is less sensitive to the curse-of-dimensionality and more robust to a small number of high dimensional samples than other non-SVM classifiers. For these reasons, it is much more attractive in application areas than the other neural networks. Investigation of numerous experiments on gene expression data using various models of SVMs and non-SVMs for cancer diagnosis were performed previously [4]. The best results were obtained using the SVM methods. Basically SVM is designed for binary classification problems, and several algorithms exist that allow multiclass classification with SVMs. In this section, we outline the principles behind SVM algorithms used in this study. Detailed review of binary SVM, exact mathematical formulations of both binary and multiclass SVM algorithms are given in [3, 7, 12, 13].

2.1.1 Binary SVM

Binary SVM is a linear classifier that maximizes the margin between the separating hyperplane and the training data points [13]. The hyperplane is based on a set of training data which lie closest to the boundary and are called support vectors. The algorithm implicitly maps the input data in the feature space, and an inner-product induced in the algorithm is calculated by kernel functions without considering the feature space itself. The SVM problem is expressed by a quadratic programming (QP) optimization problem with linear constraints. For this reason, the SVM always produces global solution for classifiers. The SVM is an unique supervised learning algorithm that often achieves superior generalization performance compared to other learning algorithms across most domains and tasks.

2.1.2 Multiclass SVM: One-Versus-Rest

For a k-class problem, the one-versus-rest method constructs k SVM models [3, 7, 12]. The ith SVM is trained with all of the training examples in the ith class with positive labels, and all other examples with negative labels. The final output of the one-against-rest method is the class that corresponds to the SVM with the largest output value. The method is computationally expensive, since we need to solve k QP optimization problems where each problem size is the same as the training data set size. This technique does not have theoretical justification such as the analysis of generalization, which is a relevant property of a robust learning algorithm.

2.1.3 Multiclass SVM: One-Versus-One

The one-versus-one method constructs all possible pairwise binary classifiers, where each classifier is constructed using the training examples from two classes chosen out of k classes [3, 7, 12]. There exist $k(k-1)/2$ different decision functions for a k-class problem. The most popular method for the class identification of the one-versus-one method is the "Max Wins" algorithm. In the "Max Wins" algorithm, each classifier casts one vote for its preferred class, and the final result is the class with the most votes. When more than one class have the same number of votes, i.e., a tie situation arises, each point in the unclassifiable (tie) region is assigned to the closest class using the real valued decision functions. Unlike the one-versus-rest method, tie-breaking plays only a minor role and does not affect the decision boundaries significantly. One of the benefits of this algorithm is that for every pair of classes we deal with a much smaller optimization problem. Although we need to solve $k(k-1)/2$ QP optimization problems, the computational complexity is polynomial to the training data set size. Similar to one-versus-rest method, one-versus-one method does not have an established bound on the generalization error. However, the one-versus-one method fits perfectly the known characteristics of the binary SVM, where the border-

lines between classes are computed directly. Moreover, some researchers postulate that even if the entire multicategory problem is non-separable, while some of the binary subproblems are separable, then one-versus-one method can lead to improvement of classification accuracy compared to one-versus-rest method [4, 12].

2.1.4 Bounds on Generalization Error

SVMs are provided with many statistics that allow us to estimate their generalization performance from bounds on the leave-one-out error. The leave-one-out error is an unbiased estimate for the true error rate of a classifier. Several error bound theories for binary SVMs exist and most of the bounds are developed for hard margin SVMs. Multiclass SVMs do not have established bound on generalization error. The one-versus-one and one-versus-rest MC-SVMs are constructed by combining several binary SVMs. We evaluate an estimate of a generalization error bound for a multiclass SVM by averaging the values of the generalization error bound of all binary SVMs which are used to construct the multiclass SVM. In this paper, we test our experiments with binary SVM's Opper-Winther bound [10], Zhou-Tuck bound [11], and radius-margin bound [13] for gene feature selection. Detailed description of these bound are presented in [10, 11, 13].

2.2 SVM-based Evolutionary Method

The evolutionary algorithm that we use maintains a population of predictors whose effectiveness can be determined by using them as features in an SVM classifier. The initial predictors in the population are randomly constructed from the gene features set. Instead of applying crossover and mutation operations, the method selects and recombines new features based on an estimate of the error bound value of an SVM and the frequency of occurrence of the features in the evolutionary approach. Let us denote that T_m is the bound value of an SVM where the training dataset contains m features of a predictor and T^i_{m-1} is the bound value for all m genes except gene i. Then, T^i_{m-1} for all i in each predictor are calculated. The $T^j_{m-1} < T^k_{m-1}$ means removing gene j from the predictor can reduce error bound much more than removing gene k. Thus genes j with small T^j_{m-1} should be deleted in the next generation. Again, if T^i_{m+1} is the bound value for m genes on a predictor plus a new gene i. The $T^j_{m+1} < T^k_{m+1}$ means adding gene j to the predictor can reduce error bound much more than adding gene k. The derivative of the bound values can also be used for feature selection. The minimum derivative value with respect to a gene feature means the bound is less sensitive to that gene feature and the maximum derivative value with respect to a gene feature means the bound is highly sensitive to that gene feature. We can remove a gene feature from a predictor to which derivative of error-bound is minimum and add a gene feature to a predictor to which derivative of error-bound is maximum. The k-fold cross validation is used as an estimator of the generalization performance that also measures the fitness value. The termination criteria is defined using both the maximum number of generations and the criteria of no improvement of maximum fitness value of the population. The algorithm is described below:

1. A population E_0 of n predictors $\{G_1, G_2, ..., G_n\}$ is created. A predictor G_i is a subset of m gene features $\{g_1, g_2, ..., g_m\}$ initially created randomly. Evaluate the fitness values of all predictors. The fitness value of a predictor is evaluated by the SVM applied on the k-fold cross validation data set with m gene features of that predictor.

2. Until termination criteria not satisfied do the following:

3. For each predictor $G_i \in E_k$, create a new predictor G'_i

   3.1. Delete p genes from G_i, whose error bound values or derivative of error bound values are minimum and selected in a few previous generations as briefly described above. For details, see [5].

   3.2. Add the same number of p genes from a random subset of data except those are in G_i in population E_k whose error bound values are minimum or derivative of error bound values are maximum with the rest of the genes in G_k after deletion and frequently selected in the previous generations.

   3.3. Compute the fitness function for the new predictor G'_i using SVMs.

4. Create a new population E_{k+1} by replacing all new G'_i.

5. Replace some worse predictors of the new population E_{k+1} based on classification accuracy by some best predictors from the previous generation. To do this, merge the features of some best predictors from the previous generation and then randomly split features of the merged features set into the same number of predictors. Then select some predictors for the new G'_i.

This procedure will be performed for a set of SVM hyperparameters and the best hyperparameters for each predictor will be obtained. Different combinations of genes with the same high accuracy rate can be evaluated in evolutionary computations through generation of individuals of a population. From this procedure we will get n' feature sets according to the best high classification accuracies where $n' \leq n$. From the n' sets we will choose N_{best} features according to occurrence frequency and classification accuracy rate. The hyperparameters for the final learning machine (SVM) will be selected by averaging the best hyperparameters of that n' predictors. For details about the algorithm and

principles behind these, see [5]. For multicategory microarray data, we apply one-versus-one MC-SVM classifier and one-versus-rest MC-SVM classifier to evaluate the fitness function in the evolutionary method and T_m of an MC-SVM is evaluated by averaging the bound values of all binary SVMs used to construct the MC-SVM. We call the proposed evolutionary methods as evolutionary one-versus-one MC-SVM and evolutionary one-versus-rest MC-SVM throughout the paper.

3 Computational Experiments

3.1 Data Analysis

In our experiments, we use 6 cancer-related human gene expression datasets that are described in Table 1. The dataset are available on http://www-gems-system.org for non-commercial use. The studied datasets were produced primarily by oligonucleotide-based technology. Specifically, all datasets except for SRBCT, RNA were hybridized to high-density oligonucleotide Affymetrix arrays HG-U95 or Hu6800, and expression values (average difference units) were computed using Affymetrix GENECHIP analysis software. The SRBCT dataset was obtained by using two-color cDNA platform with consecutive image analysis performed by DeArray Software and filtering for a minimal level of expression. The datasets have 3-5 distinct diagnostic categories, 50-203 patient samples, and 2308-12600 variables (gene features) after preprocessing (details in [4]). We rescale gene expression values of these datasets linearly into the range [-1,1].

3.2 Parameter Setting

The number of predictors is set to 50. The size of each predictor and the numbers of deletions and additions of genes are set experimentally (usually half of the predictor is deleted and added in our experiments). In each generation, at best 10 worst predictors in the new population is replaced with the 10 best predictors of the previous population according to step 5 of the algorithm. To evaluate the performance of the proposed method we use 5-fold cross-validation on each dataset. The stopping condition of the algorithm is to use 100 generations. The SVM parameters are as: trade-off parameter $C = [2^{-2}, 2^{-1}, \ldots, 2^{9}, 2^{10}]$, and RBF kernel parameter $\gamma = [2^{-5}, 2^{-4}, \ldots, 2^{3}, 2^{4}]$.

3.3 Experimental Results

To compare the performance of evolutionary one-versus-one and one-versus-rest MC-SVMs, we first choose the linear kernel. For linear kernel, we performed experiments on Opper-Winther bound, Zhou-Tuck bound, and radius-margin bound but reported the results in the tables of the bounds that show the best performance. The results of evolutionary one-versus-one MC-SVM and evolutionary one-versus-rest MC-SVM with some other existing methods such as MC-SVM [4] and ESVM [1] are shown in Table 2. We show the results of MC-SVMs and ESVM from [4] and [1], respectively, in Table 2. From experimental results, we see that evolutionary MC-SVMs with linear kernel obtain better results than the other existing methods, and the evolutionary one-versus-one MC-SVM shows slightly better results than the evolutionary one-versus-rest MC-SVM. Table 3 shows the results of evolutionary one-versus-one MC-SVM with linear kernel where gene features are selected using derivative of the bound values. In this experiment, we also include the derivative of weight vector for feature selection. From the results we see that bound values are more stable than their derivatives for feature selection. The computational costs of derivatives are computationally more expensive than the cost of bounds themselves except the cost of the derivative of weight vector. The derivative of weight vector can perform better among the derivatives of error bounds but it is not as good as error bounds. The RBF kernel shows better generalization performance for some complex problems. We also compute the experimental results of RBF kernel on evolutionary one-versus-one and one-versus-rest MC-SVMs. The Opper-Winther bound and radius-margin bound are investigated for RBF kernel. Table 4 shows the comparison of evolutionary one-versus-one and one-versus-rest MC-SVMs using linear and RBF kernels with the same number of genes in each dataset. From the results we see that RBF kernel shows good results in the evolutionary one-versus-rest MC-SVM; however, the same results are obtained using linear kernel in the evolutionary one-versus-one MC-SVMs. The comparison of linear and nonlinear kernels reveals the necessity of using linear kernel in some situations. In the case of clinical applications such as diagnosis of disease as well as prediction of clinical outcomes in response to treatment, a small number of genes that can separate datasets into classes linearly with highest classification accuracies are more desirable. By our experiments, we see that a small number of genes can linearly separate datasets into classes with highest classification accuracies. Experiments with different error-bounds also present that the evolutionary one-versus-one MC-SVM shows slightly better results than the evolutionary one-versus-rest MC-SVM.

4 Conclusion

In this paper, we compare the performance of one-versus-one and one-versus-rest MC-SVMs in the evolutionary algorithm where feature selection and recombination are based on the generalization error bound of SVMs. To evaluate the performance, we apply several error-bounds on each multiclass SVM-based evolutionary method. From experimental results, we see that a small number of genes can linearly separate datasets into classes with highest classification accuracies. We also find that the evolutionary one-versus-one MC-SVM shows slightly better results than the evolutionary one-versus-rest MC-SVM.

Dataset	Diagnostic Task	#Samples	#Genes	#Classes
Leukemia1	Accute myelogenous leukemia (AML), acute lympboblastic leukemia (ALL) B-cell, and ALL T-cell	72	5327	3
Leukemia2	AML, ALL, and mixed-lineage leukemia (MLL)	72	11225	3
SRBCT	Small, round blue cell tumors of childhood	83	2308	4
Brain_Tumor1	Five human brain tumor types	90	5920	5
Brain_Tumor2	Four malignant glioma types	50	10367	4
Lung_Cancer	Four lung cancer types and normal tissues	203	12600	5

Table 1: Features of microarray datasets.

Dataset	MC-SVM [4] (RBF Kernel)	ESVM [1] (RBF Kernel)		Evolutionary one-versus-one MC-SVM (Linear Kernel)			Evolutionary one-versus-rest MC-SVM (Linear Kernel)		
	Ac. Rate (%)	Ac. Rate (%)	#Genes	Ac. Rate (%)	#Genes	Bounds	Ac. Rate (%)	#Genes	Bounds
Leukemia1	97.50	**100.0**	3.4	**100.0**	**3**	OW/ZT/RM	**100.0**	**3**	OW/ZT/RM
Leukemia2	97.32	**100.0**	3.5	**100.0**	**3**	OW/ZT/RM	**100.0**	**3**	OW/ZT/RM
SRBCT	**100.0**	98.75	6.2	**100.0**	**4**	OW/ZT/RM	**100.0**	**4**	OW/ZT
Brain_Tumor1	91.67	96.67	6.1	**98.89**	8	RM	97.78	8	ZT
Brain_Tumor2	77.83	**100.0**	**4**	**100.0**	5	OW/ZT/RM	**100.0**	5	OW/RM
Lung_Cancer	96.55	95.75	**6.9**	**99.48**	10	OW	98.99	10	RM

Table 2: Mean accuracy (Ac.) rate of the evolutionary one-versus-one MC-SVM and evolutionary one-versus-rest MC-SVM, MC-SVM with all gene features in the dataset, and ESVM. The results of MC-SVM using a nested stratified 10-CV are obtained from [4] and the results of ESVM using 10-CV are obtained from [1]. The results of the evolutionary MC-SVMs are shown using 5-CV. Here 'OW', 'ZT', and 'RM' represent the Opper-Winther bound, Zhou-Tuck bound, and radius-margin bound, respectively.

Dataset	#Genes (Selected)	Zeor-order Criteria		First-order Criteria	
		Ac. Rate (%)	Bounds	Ac. Rate (%)	Bound Derivatives
Leukemia1	3	**100.0**	OW/ZT/RM	**100.0**	$\nabla\|\|\mathbf{w}\|\|^2$
Leukemia2	3	**100.0**	OW/ZT/RM	**100.0**	$\nabla\|\|\mathbf{w}\|\|^2/\nabla$OW/$\nabla$ZT
SRBCT	4	**100.0**	OW/ZT/RM	**100.0**	∇OW
Brain_Tumor1	6	**97.84**	OW	93.39	$\nabla\|\|\mathbf{w}\|\|^2$
Brain_Tumor2	5	**100.0**	OW/ZT/RM	**100.0**	$\nabla\|\|\mathbf{w}\|\|^2/\nabla$ZT

Table 3: Mean accuracy (Ac.) rate of the zero-order (using error bound values) and first-order (using derivatives of error bounds) criteria in the evolutionary one-versus-one MC-SVM method.

Dataset	#Genes	Evolutionary one-versus-one MC-SVM				Evolutionary one-versus-rest MC-SVM			
		RBF Kernel		Linear Kernel		RBF Kernel		Linear Kernel	
		Ac. Rate (%)	Bounds	Ac. Rate (%)	Bounds	Ac. Rate (%)	Bounds	Ac. Rate (%)	Bounds
Brain_Tumor1	6	96.73	OW	**97.84**	OW	**97.84**	OW	91.35	ZT
Brain_Tumor2	4	**98.18**	RM	**98.18**	RM	**98.18**	OW	94.18	OW

Table 4: Mean accuracy (Ac.) rate of the evolutionary one-versus-one MC-SVM and evolutionary one-versus-rest SVM using RBF kernel. The results of the evolutionary MC-SVMs are shown using 5-CV.

Acknowledgment

This work is supported by Grant-in-Aid for JSPS Fellows from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] H.-L. Huang and F. -L. Chang, "ESVM: Evolutionary support vector machine for automatic feature selection and classification of microarray data", **Bio Systems**, Vol. 90, 2007, pp. 516-528.

[2] T. Jirapech-Umpai and S. Aitken, "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes", **BMC Bioinformatics**, Vol. 6, No. 148, 2005.

[3] A. Rakotomamonjy, "Variable selection using SVM-based criteria", **Journal of Machine Learning Research**, Vol. 3, 2003, pp. 1357-1370.

[4] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis", **Bioinformatics**, Vol. 21, No. 5, 2005, pp. 631-643.

[5] R. Debnath and T. Kurita, "An evolutionary approach for gene selection and classification of microarray data based on SVM error-bound theories", **Bio Systems**, Vol. 100, Issue 1, pp. 39-46, 2010.

[6] R. Debnath and T. Kurita, "A comparison of evolutionary methods based on SVM error-bound theories for multicategory microarray gene expression data", **Proc. International Multi-Conference on Complexity, Informatics and Cybernatics**, Orlando, Florida, USA, 2010, pp. 295-299.

[7] C. -W. Hsu and C. -J. Lin, "A comparison methods for multiclass support vector machines", **IEEE transactions on neural networks**, Vol. 13, No. 2, 2002, pp. 415-425.

[8] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio and V. Vapnik, "Feature selection for svms", **Advanced in Neural Information Processing Systems 13**, 2001.

[9] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, "Gene selection for cancer classification using support vector machines", **Machine Learning**, Vol. 46, 2002, pp. 389-422.

[10] M. Opper and O. Winther, "Gaussian process and SVM: Mean field and leave-one-out", **Advances in large margin classifiers**, A. Smola, P. Bartlett, B. Schölkopf and D. Schuurmans (Eds.), Cambridge, MA:MIT Press, 2000, pp. 297-309.

[11] X. Zhou, and D. P. Tuck, "Gene selection using a new error bound for support vector machines", **Proc. Eleventh Annual International Conference on Research in Computational Molecular Biology**, San Francisco, USA, 2007.

[12] U. H. -G. Kreßel, "Pairwise classification and support vector machines", **Advanvced in Kernel Methods: Support Vector Machine**, B. Schölkopf, C. Burges, and A. Somla (Eds.), Cambridge, MA:MIT Press, 1998, pp. 255-268.

[13] V. Vapnik, **Statistical Learning Theory**, New York:Wiley, 1998.