CANCER GENOMICS & PROTEOMICS *3*: 63-70 (2006)

Improving Gene Expression Sample Classification Using Support Vector Machine Ensembles Aggregated by Boosting

ANDREI DRAGOMIR and ANASTASIOS BEZERIANOS

Department of Medical Physics, Medical School, University of Patras, Rio 26500, Greece

Abstract. *The molecular characterization of different tumor types using gene expression profiling is expected to uncover fundamental aspects related to cancer diagnosis and drug discovery. There is, therefore, a need for reliable, accurate sample classification tools, as well as methods for efficient identification of genes informative for the class discrimination. We propose a method based on Support Vector Machine (SVM) ensembles, trained within a boosting framework. The approach allows sequential training of classifiers on different data subsets, their aggregate yielding results superior to single SVM. Results from binary and multiclass classification experiments performed on several data sets are presented.*

Gene expression profiling using microarray technology creates a molecular picture of a cell's internal state and microenvironment by measuring the expression levels of thousands of genes in a single microarray experiment (1, 2). By performing such experiments on samples of distinct pathogenetic disease type, as well as on samples from healthy tissues, a broader understanding of the molecular variations between healthy and disease tissues could be obtained, based on the quantitative measure of thousands of parameters (the gene expression levels). The data mining research community is thus offered the challenge of analyzing and interpreting the wealth of data produced by this technology that has revolutionized biological research. Novel or adapted machine learning and statistical techniques are required. Some of the opportunities opened by employing the analysis of gene expression data as a phenotyping tool include diagnostic categorization of cancer *versus* non-cancer tissues (2, 3), discrimination among different subtypes of tumors (4-6), as well as drug response prediction or cancer prognosis (7). Obtaining a reliable distinction between normal and tumor tissues or between tumor tissue subtypes has recently attracted significant research efforts, since conventional clinical diagnostic methods are based on subjective evaluation, which may negatively influence the efficacy of subsequent therapy.

Different classification methods have been employed to date in diagnostic applications, such as classic nearest neighbor classifiers, linear discriminant analysis or more recent approaches, including bayesian networks, support vector machines (SVMs) and other machine learning tools (8-14). Classification is a supervised learning task, having as a goal the development of an efficient model for predicting the class membership of the data. The learning system is given the training data, consisting of data points chosen from the input data space and their respective class labels. The model derived from the training data is expected not only to produce the correct label on the training data, but to correctly predict labels of unseen data (also referred to as test set). In the cases when the classification task is dichotomous, we deal with a binary classification problem, while in the cases when the data has at least three classes, we are confronted with a multiclass classification problem.

Whatever classification method is used for gene expression data analysis, the questions to be solved are either classic classification issues (such as the curse of dimensionality – referring to the case when the dimension of the feature space is much larger than the number of available observations, a fact that leads to a drastic rise in computational complexity and classification errors (15)), or specific problems related to gene expression data (noise and large variability of data among samples (16)). Feature selection is another classic classification task, closely related to the data dimensionality issue. It refers to data dimensionality reduction by keeping only features that are significant for the class discrimination. This is of great importance, since, using only a subset of genes reduces computational complexity, is more convenient for developing diagnostic tests and for obtaining interesting biological insights into the molecular mechanisms triggering diseases. Since the prediction of diagnostic categories is such a sensitive task, it is crucial to deal with the above issues

Correspondence to: Andrei Dragomir, Department of Medical Physics, Medical School, University of Patras, Rio 26500, Greece. Tel: +30-2610-996115, Fax: +30-2610-992496, e-mail: adragomir@heart.med.upatras.gr

Key Words: Gene expression, classification, support vector machines, boosting, ensemble learning, feature selection.

accordingly, in order to obtain optimal classification performance, while confidence in the results should make the analysis suitable for further interpretation by clinicians.

The current study presents an approach that aims at improving the classification performance and, at the same time, increasing the results confidence. Current publicly available gene expression data sets obtained from microarray experiments have huge dimensionality, with only a few tens to a hundred experimental samples (that correspond to the input vectors in our algorithms, which from here on will be referred to as profiles) and with thousands of gene expression measurements (the variables, or features in our analysis). Working on such data, there is always the risk of overfitting when trying to find a suitable classifying model. Overfitting refers to the case when the model estimated may very accurately fit the samples in the training set, but be very inaccurate in assigning the label of a new sample. Therefore, we employed SVMs, which reduce the overfitting risk to some extent (17). Although SVMs theoretically provide near optimal classification results, their training on large data sets is enormously time-consuming and, therefore, approximate implementations are normally used in order to reduce the computation time, with a direct result in degrading the classification performance (18, 19). By using such sub-optimal implementations of SVMs aggregated in an ensemble scheme, classification indices superior to single SVMs and to other classic methods should be obtained. The supposition is based on results from the theory of ensembles, which prove that a combination of individual classifiers, each performing better than average and having negatively associated errors, result in an aggregate with improved classification performance (20). Our choice of aggregating single SVMs in an ensemble is that of a boosting framework, since boosting has been proved to be an efficient class prediction tool with remarkable success in a wide variety of applications, especially in those dealing with high dimensionality data (13, 19).

SVM ensembles have been proposed by (17) and their aggregation by a boosting scheme has been applied to various classification tasks. Although boosting was originally designed for the aggregation of the so-called weak classifiers (performing only slightly better than average), previous studies (and our current results) confirm the supposition that the SVM performance may also be enhanced by boosting: Yan *et al.* (22) used such a method for inferring high-level semantic scene categories from low-level visual features, while Kim *et al.* (23) applied it in fraud detection in a mobile telecommunication payment system, as well as on several benchmark classification data sets. For the purpose of gene expression data analysis, boosting has only been used to date in conjunction with weaker classifiers, such as decision stumps (13).

A detailed explanation of the methods used and the ideas underlying our approach are presented, together with the results of binary and multiclass classification on several benchmark data sets. Finally, the results are discussed and the direction of future research is outlined.

Materials and Methods

A simple representation of gene expression data in the context of the current work would be that of a $n \times m$ matrix $X=(x_{ij})$, with x_{ij} representing the expression level of gene i in sample j. Gene expression profiles $x_j=(x_{j1}, \ldots, x_{jn})$ were assigned labels yi in the case when the respective sample belongs to a known diagnostic class. Since SVMs, which are binary classification tools, $y_j \in \{+1; -1\}$, an extension for the multiclass classification is presented.

Based on a learning set of k profiles, $LS=\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_k, y_k)\}$, which are known to belong to certain classes, the learning algorithm must build a classifier C that is able to predict correct class labels for a new set of expression profiles, called the test set, with unknown class labels. The classifier must be understood as a discriminant (or decision) function f, such that $y=f(\mathbf{x})$. Therefore, the supervised learning task consists, in fact, of finding suitable discriminant functions or their best approximations.

Support vector machines. SVMs (17) find hyperplanes $<\mathbf{w},b>$ that optimally separate the classes by maximizing the width of the separating band between the data points and the hyperplane. In the linearly separable problem, the discriminant function is of the form $f(\mathbf{x})=\mathbf{w}\cdot\mathbf{x}+b$, which has an associated decision function:

$$f_d(\mathbf{x})=sign(\mathbf{w}\cdot\mathbf{x}+b) \quad \textbf{(1)}$$

The optimal hyperplane must obey the following constraints:

$$\underset{\mathbf{w},b}{\text{minimize}} \quad \tau(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{subject to} \quad y_i(<\mathbf{w}, \mathbf{x}_i>+b) \geq 1 \text{ for all } i=1, \ldots, m \quad \textbf{(2)}$$

The constraints ensure that $f(\mathbf{x}_i)$ will be $+1$ for $y_i=+1$ and -1 for $y_i=-1$. In the case of linearly non-separable data, the constraints are adapted to allow misclassification data points but penalize them by means of some slack variables $\xi_i \geq 0$. The constraints for the optimal hyperplane are modified in this case as follows:

$$\text{minimize} \quad \tau(\mathbf{w},\xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{subject to} \quad y_i\left(\mathbf{w}^{\mathrm{T}}\cdot\Phi(\mathbf{x}_i)+b\right) \geq 1-\xi_i \quad \text{for all } i=1, \ldots, m \quad \textbf{(3)}$$

where $C>0$ determines the trade-off between margin maximization and training error minimization, while $\Phi(\cdot)$ is a non-linear function which maps the input space into a higher dimensional space.

To solve the constraint problem of eq. **(2),** a Lagrangian method (24) is used, which eliminates the primal variables $\mathbf{w}$ and b and introduces instead the Lagrangian multipliers. Thus, from the problem of eq. **(2),** performing some mathematical manipulations, one has to deal with:

$$\underset{\alpha\in\Re^m}{\text{maximize}} \quad W(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{subject to} \quad \alpha_i \geq 0 \text{ for all } i=1, \ldots, \mathrm{m} \text{ and } \sum_{i=1}^{m}\alpha_i y_i = 0 \quad \textbf{(4)}$$

which is solvable in practice. Note that above $k(\mathbf{x}_i, \mathbf{x}_j) = <\Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j)>$ is the so-called kernel function. The decision function of eq. **(1)** now takes the following form:

$$f_d(\mathbf{x}) = sign\left(\sum_{i=1}^{m} y_i \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b\right) \quad \textbf{(5)}$$

As previously stated, approximate algorithms (19) are used in practical implementations of SVMs in order to be able to deal with the size of the quadratic problem of eq. **(4)** for large scale data.

Ensemble framework. In order to overcome the drawback of the practical implementations of SVMs, described above, we use an ensemble framework as suggested by (23). There are two kinds of problems that such an approach solves. The first one arises when the learning algorithm is searching a space of possible discriminant functions that is too large for the amount of the available training data. The second problem is encountered when the space of possible discriminant functions does not contain any function that is a good approximation to the ideal discriminant function.

From a practical point of view, variability and noise are well-known issues to face when analyzing gene expression data (16). Either physiological variability (which describes differences in the expression characteristics of cells from macroscopically identical conditions) or sampling variability (differences in sample characteristics such as tissue heterogeneity and other host factors) may constitute impediments difficult to surpass by current analytical techniques. Given the small sample size of current microarray experiments, it is puzzling for a learning algorithm to correctly estimate the correct distribution of the data.

Taking into account the peculiarities of the gene expression data, we believe that an ensemble approach would overcome problems of this type. Ensembles construct a set of classifiers and then have those classifiers vote in a weighted manner to predict the class label of a new data point. Ensembles theory (20, 25) justifies the enhanced performance of an ensemble of classifiers over single classifiers by the following reasoning. If we have a set of classifiers and a new data case x_i, in the case when all classifiers are identical, then if one classifier is wrong, all the other classifiers will be wrong too. However, if the classifiers are different and their errors are uncorrelated, if one of the classifiers is wrong, the majority of the other may be correct, so the majority vote of the ensemble will correctly classify $\mathbf{x}_i$. A formal characterization of the problem is: if the probability of error of the individual classifiers is $p<1/2$ and the errors are independent, then the error probability p_ε of the majority voting of a set of d classifiers is:

$$p_\varepsilon = \sum_{i=d/2}^{d} p^i (1-p)^{d-i} < \sum_{i=d/2}^{d} \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{d-i} = \sum_{i=d/2}^{d} \left(\frac{1}{2}\right)^d \quad \textbf{(6)}$$

which decreases as the number of classifiers d increases.

Boosting. Our approach of constructing an ensemble is that of an additive model (21), which predicts the class label of a new data point by performing a weighted sum of a set of component classifiers, in such a way that the weighted sum fits the data well. Boosting is an efficient and flexible method based on this principle. Specifically, it incrementally adds a new classifier at a time to an ensemble. Each new classifier is constructed by a learning algorithm that tries to minimize the classification error on a weighted training data set. At each iteration step, the current classifier-weighted error is applied to update the weights of the training examples. The desired effect is to place more weight on the training examples that were misclassified and less weight on examples that were correctly classified. Therefore, in subsequent iteration, the boosting framework constructs progressively more difficult learning problems. Intuitively, we can imagine that subsequent classifiers concentrate on class boundary regions of the data space, where classification decisions are difficult to take.

The boosting algorithm we employ initializes by creating a training set TS of m labeled examples (gene expression profiles and their respective class label) $TS=\{(\mathbf{x}_i, \mathrm{y}_i)|i=1,...,m\}$. At the same time, the same weight values are assigned for all the profiles in the training set according to: $p_0(\mathbf{x}_i) = 1/m$. At the following k iteration steps training subsets $TS_k=\{(\mathbf{x}_i, \mathrm{y}_i)|i=1,...,l\}$ are built by selecting l training samples (with $l<m$) from the initial training set, according to their weight values. The weight values of the training samples are re-evaluated at each iteration step, based on their contribution to the classification error of the respective classifier. Accordingly, weights of the samples misclassified at the previous steps are increased, while samples that were correctly classified have their weights decreased. The procedure results in the construction of training subsets consisting, in an increasing manner, of samples that are difficult to classify.

The final classifier is constructed based on a weighted voting of the individual classifiers. Each classifier is weighted (by ε_k) according to its accuracy on the weighted training set that it was trained on:

$$f_d(\mathbf{x}_i) = \sum_k \varepsilon_k f_{dk}(\mathbf{x}_i) \quad \textbf{(7)}$$

Multiclass classification. Although SVMs were originally designed as binary classification tools, several manners of extension to the multiclass were proposed (26-28). The approaches we are considering split the multiclass classification problem into multiple binary problems and may be roughly divided into two types: the one-against-all and one-against-one methods. In the former, classifiers for discriminating one from all the other cases are built. Therefore for a q class problem, we would employ q different binary classifiers. In the latter case, a classifier is built for each pair of classes, so that we would have $q(q - 1)/2$ independent binary classifiers.

For practical considerations, concerning the complexity of the ensemble model, in the case of multiclass classification we use the one-against-all approach. Specifically, in the case of q class classification problems, we will have an aggregation consisting of q SVM ensembles. Each SVM ensemble will consist of M independent binary SVMs (where M is the number of classifiers resulting from the boosting framework) discriminating classes in a one-against-all manner. The final classification decision derived from the decision results of the q SVM ensembles will be taken through a maximum wins voting strategy.

Feature selection. In high dimensional data analysis, as is the case with gene expression analysis, feature selection methods are essential if the researcher is to make sense of his data. In classification problems, the task translates into finding ways to reduce the dimensionality of the feature space to overcome the risk of overfitting. Although the SVM is, as stated before, a method that is not particularly vulnerable to overfitting, it was proven that it benefits from feature space dimensionality reduction (29).

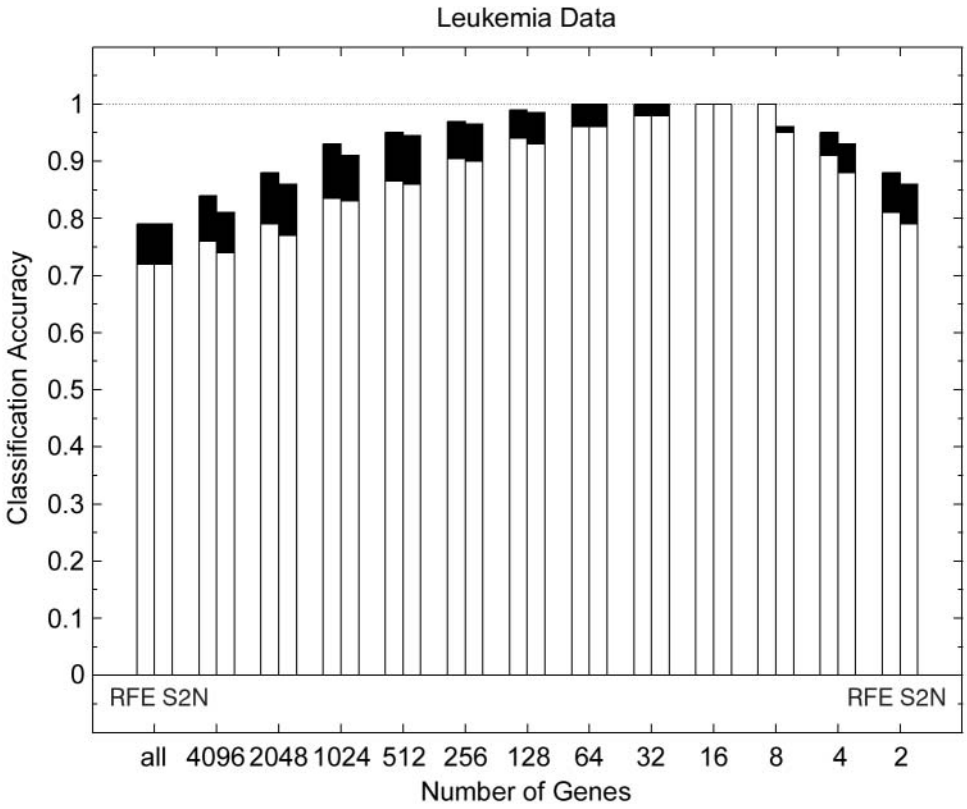

Figure 1. *Average classification accuracies over the 10 random splits of the data set for SVM ensembles (black bars) compared to those of single SVMs (white bars – overlapped), for the two feature selection strategies [recursive feature selection (RFE) and signal to noise (S2N)], obtained from experiments on the leukemia data set. Within each group of bars, to the left is the result corresponding to RFE feature selection, while to the right is the one corresponding to S2N.*

Identifying subsets of genes that are most relevant for the classification task is important not only from the statistical learning point of view, but also from the biological one. Subsets of genes important for the phenotype distinction can be further employed to investigate the biology of disease.

We use two feature selection methods to evaluate the performance of our approach on sets of subsequently reduced dimensionality: Recursive Feature Elimination - RFE (29), which is a wrapper-based method using a linear SVM to remove features based upon the absolute magnitude of the hyperplane elements, already proved to be a successful application in gene selection and the Signal-to-Noise ratio - S2N (4), which computes a ranking measure for each gene based on its class correlation. In the multiclass case, the feature selection is applied in a one-versus-all fashion.

Results

In order to assess the performance of our method, several classification experiments were performed on five publicly available data sets. The expression measurements originate from microarray experiments monitoring either tumor/healthy tissue samples or samples of different tumor subtypes.

Leukemia data set. The data consisted of 72 microarray experiments containing 7129 genes from cancer patients with two types of leukemia (acute lymphoblastic leukemia – ALL and acute myeloid leukemia – AML) and is available at http://www.genome.wi.mit.edu/MPR (4). The data was split into a training set of 38 samples (27 ALL and 11 AML) and a test set of 34 samples (20 ALL and 14 AML).

Colon cancer data set. The data set contained tissue samples from 22 normal and 40 colon cancer tissues. The expressions of 2000 genes, some of which are non-human, were provided across the 62 tissue samples by the authors of the microarray experiment Alon *et al.* (3). The data set is available at http://microarray.princeton.edu/oncology .

Prostate cancer data set. Singh *et al.* (30) performed a microarray experiment to determine whether the clinical behavior of prostate cancer is linked to underlying gene expression differences that are detectable at the time of diagnosis. One hundred and two samples (52 tumor and 50 healthy tissue samples) containing expression levels of 6033 genes were derived for the study of prostate tumors, which

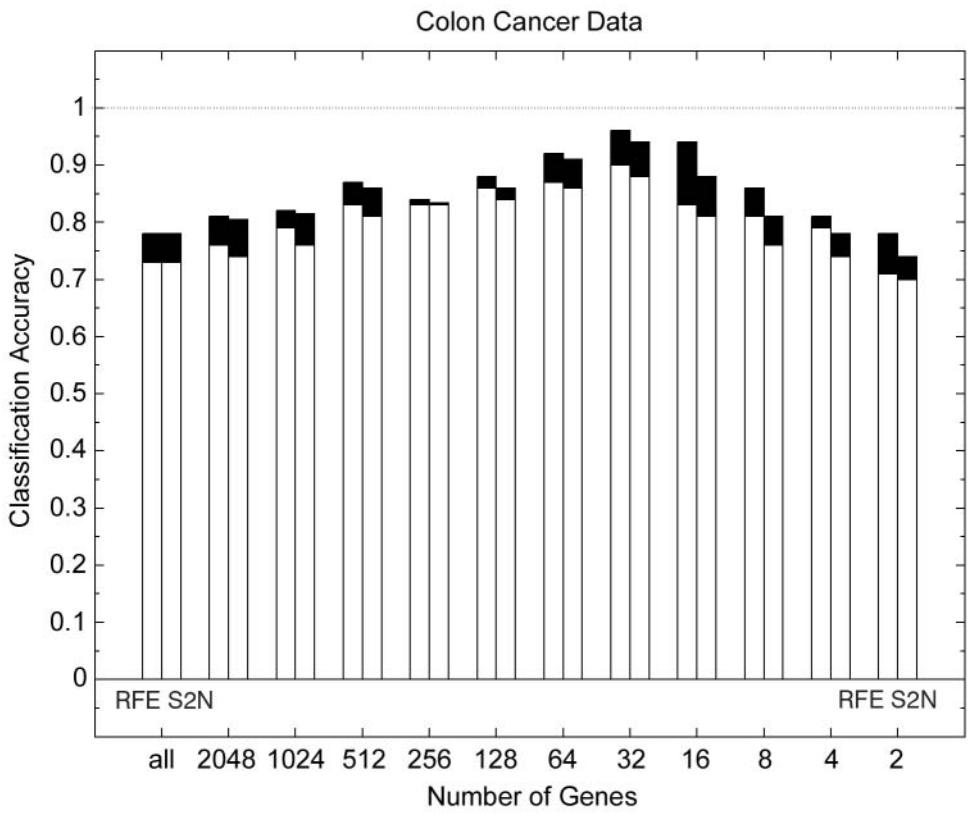

Figure 2. Average classification accuracies over the 10 random splits of the data set for SVM ensembles (black bars) compared to those of single SVMs (white bars –overlapped), for the two feature selection strategies [recursive feature selection (RFE) and signal to noise (S2N)], obtained from experiments on the Colon cancer data set. Within each group of bars, to the left is the result corresponding to RFE feature selection, while to the right is the one corresponding to S2N.

are among the most heterogeneous of cancers, both histologically and clinically. The data set is available at http://www-genome.wi.mit.edu/mpr/prostate .

Lymphoma data set. Alizadeh *et al.* (5) studied the expression levels of 4682 genes in lymphoid cells, with samples from 3 of the most prevalent adult lymphoid malignancies: diffuse large B-cell lymphoma (DLBCL, 43 samples); B-cell chronic lymphocytic leukemia (B-CLL, 29 samples) and follicular lymphoma (FL, 9 samples). The data set is available online at http://llmpp.nih.gov/lymphoma/data .

Brain tumor data set. Pomeroy *et al.* (31) studied the molecular differences of several embryonal tumors of the central nervous system. One of their data sets contained 42 patient samples of different tumor and healthy tissues: medulloblastomas (10 samples), malignant gliomas (10 samples), atypical teratoid/rhabdoid tumors (AT/RT, 10 samples), primitive neuroectodermal tumors (PNET, 6 samples) and normal cerebellum (4 samples). The expression levels of the 5597 genes were measured in the experiments and the data set is available at http://www.broad.mit.edu/mpr/CNS/ .

In order to systematically fill in the missing values from the above-described data sets the weighted K-nearest neighbors imputation method proposed in (32) was applied. In the absence of genuine test sets for 4 of the data sets (only the leukemia data set was designed by its authors such as to contain separate training and test sets), we performed random divisions of each data set into training and test sets, as described by Dudoit *et al.* in (8). Shortly, the data sets were split into a balanced training set containing two-thirds of the available samples, which were used to train the classifiers, while the class labels of the remaining third were used for comparison with the results of the classification. The splitting procedure was repeated 10 times in order to reduce the variability of the results and the classification ratios from the repeated experiments were averaged. It must be noted that, in order to avoid selection bias, the following procedure was performed: at each data splitting loop, the data set was first split into training and test subsets, then feature selection was performed using only the current training set and the performance of the classifier on the test set was assessed, using the features previously selected (see (4) and the erratum of (29)).

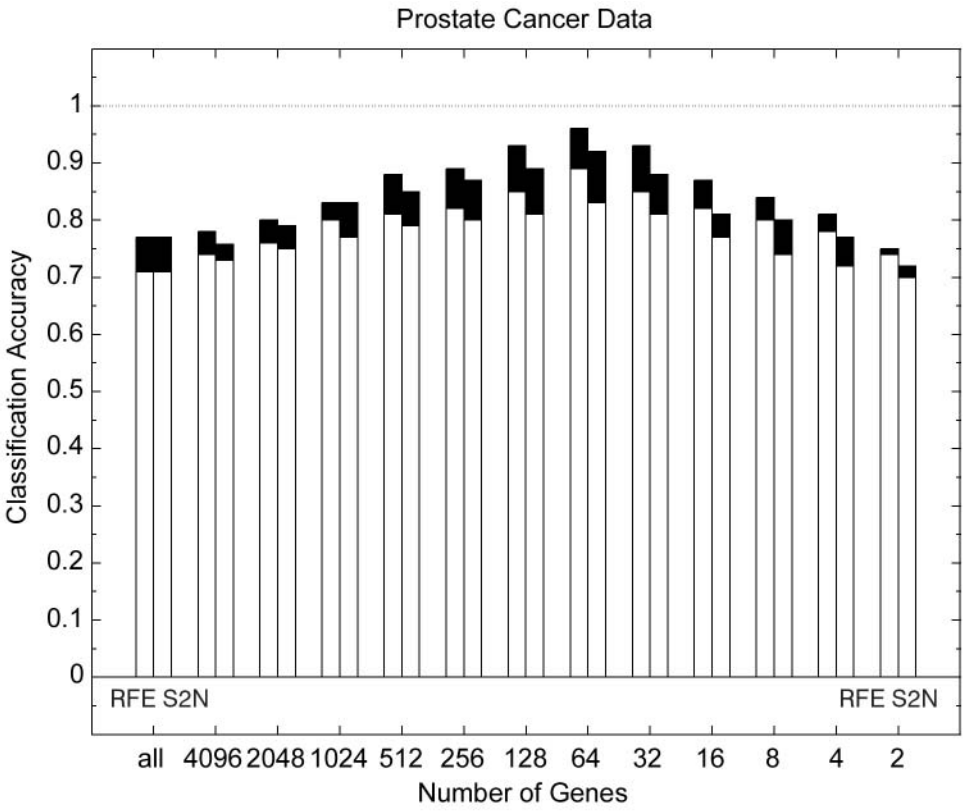

Figure 3. *Average classification accuracies over the 10 random splits of the data set for SVM ensembles (black bars) compared to those of single SVMs (white bars –overlapped), for the two feature selection strategies [recursive feature selection (RFE) and signal to noise (S2N)], obtained from experiments on the Prostate cancer data set. Within each group of bars, to the left is the result corresponding to RFE feature selection, while to the right is the one corresponding to S2N.*

Each individual SVM that is aggregated in the ensembles used a 2D polynomial kernel function; we performed a classifier optimization over the set of values of cost C=[0.0001, 0.01, 1, 100] (presented results correspond to C=100), while for the purpose of feature selection with RFE linear SVMs were employed. The total number of classifiers was empirically set in an ensemble to M=20, further increasing the classifiers number having no influence on the total classification accuracy. Our implementation grew on the LibSVM package of (33). For the experiments employing the kNN classifier k=1 was set.

Binary classification experiments. In this set of experiments, the classification performances of our approach on the leukemia, colon and prostate cancer data sets were tested. Specifically, the ensemble of classifiers was first trained on a training set with the complete number of features and the classification performance was estimated by performing the classification on the test set. The success rate:

$$T_{succ} = \frac{1}{2k}\sum_{i=1}^{k}\left|y_i + f_d(\mathbf{x}_i)\right| \quad \textbf{(9)}$$

was used to quantify the classification performance, where k was the number of samples in the test set. Subsequently, feature reduction was performed by means of RFE and S2N, respectively. The number of genes used as features of the training set was iteratively decreased by half, in order to identify a suitable subset of genes that may be used to discriminate between the 2 classes. Figures 1, 2 and 3 present the results obtained for the leukemia, colon and prostate cancer, respectively. The black bars represent classification ratios for the SVM ensembles, while the white ones (overlapped) correspond to the classification ratios of single SVMs. The results were averaged for the 10 random splits of the data sets.

It may be noticed that our approach yielded consistently higher classification performance that single SVMs. Only in the case of the leukemia data, did the ensembles yield identical accuracy with single SVMs, for the reduced feature data containing 16 genes (for both RFE and S2N obtained sets) and 8 genes (only for the RFE obtained set), respectively. In this case, the plots of the SVM ensembles and single SVMs completely overlapped. It may be noticed that the smallest success rates were obtained on the prostate

Table I. *Average classification accuracies of SVM ensembles and single SVMs compared with those of kNN on two multiclass tumor data sets for subsets of different number of genes (entire data set, top 1000 and top 100 selected genes, respectively). Figures represent average classification accuracies over the 10 random splits of the data set±standard error.*

Data set	Feat. selection/No. of genes		SVM ensemble	SVMs	kNN
Lymphoma		All genes	0.946±0.0076	0.913±0.0049	0.812±0.0102
	RFE	1000 genes	0.989±0.0234	0.915±0.0276	0.846±0.0197
		100 genes	0.923±0.0327	0.869±0.0398	0.803±0.0289
		All genes	0.946±0.0076	0.913±0.0049	0.812±0.0102
	S2N	1000 genes	0.965±0.0261	0.894±0.0212	0.824±0.0342
		100 genes	0.896±0.0313	0.817±0.0376	0.746±0.0313
Brain		All genes	0.836±0.0461	0.743±0.0293	0.681±0.0311
	RFE	1000 genes	0.825±0.0313	0.746±0.0365	0.695±0.0389
		100 genes	0.813±0.0489	0.724±0.0461	0.631±0.0401
		All genes	0.836±0.0461	0.743±0.0293	0.681±0.0311
	S2N	1000 genes	0.812±0.0323	0.719±0.0326	0.665±0.0368
		100 genes	0.803±0.0389	0.706±0.0414	0.624±0.0394

cancer data set, which may be due to the heterogeneity of the tissue samples, as well as to the fact that it was the largest data set of the 3.

Multiclass classification experiments. A comparative result of the classification performance of our approach, single SVMs and the kNN classsifier, is presented in Table I for 2 multiclass tumor data sets: lymphoma (3 classes) and brain (5 classes). The classification performance was computed as the percentage of correctly classified test set samples. The figures in Table I represent averages over the 10 repetitions of the experiment with random training and test sets. The splitting procedure, followed by feature selection, was performed as described above, in order to avoid bias. It may be noticed that, as in the case of all current machine learning techniques, the classification performance decreased with the number of classes involved.

Discussion

The current study proposes a method to improve the classification performance in the case of gene expression data. The idea underlying our approach is to extend the area of the data space where correct classification decisions are taken by combining several classifiers that are aggregated in an ensemble. This approach allows us to treat different regions of the data space by different means, since individual classifiers are trained on data subsets differently from each other. Successive SVM classifiers are employed for more difficult decision regions by training the SVMs with data sets constructed within a boosting framework.

We tested our method on several benchmark tumor data sets publicly available. The results prove that such an approach yields superior performance to that of single SVMs. The SVM ensemble provides optimal classification results in the case of the leukemia data set, by employing subsets of 8, 16 and 32 genes selected in a feature selection step performed by the recursive feature elimination algorithm. Also in the case of the colon cancer and prostate cancer data sets, the ensemble provides high classification performance, which is encouraging, specially in the latter case, which is known to contain heterogeneous measurements to a high degree.

Also, in multiclass classification tasks, the approach yields superior performance to single SVMs and to kNN classifiers. The multiclass ensemble is built in a one-against-all manner and offers satisfactory prediction accuracy, since it is known that multiclass classification problems are more difficult than binary ones. In our case, the prediction accuracy was very high for the lymphoma experiments, which is a success taking into account the size of this data set. The accuracy for the brain tumor data set was drastically lower, which may be explained by the sample heterogeneity and in the higher number of classes of these data sets.

A further direction for our work could be that of studying the behavior of a semi-supervised SVM, which could open the way to constructing decision functions based on both training and test data. Such an approach could yield interesting results, since it is known that, due to the high variability of the gene expression data, small training sets are not capable of offering a true model of the underlying data distribution. In the future, we hope to analyze, from a biological point of view, the small subsets of genes yielding optimal classification performance.

Acknowledgements

The authors would like to thank the European Social Fund (ESF), Operational Program for Educational and Vocational Training II (EPEAEK II) and particularly the program IRAKLEITOS for funding the above work.

References

1 Eisen MB, Spellman PT, Patrick OB and Botstein D: Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci *95*: 14863-4868, 1998.

2 Schummer M, Ng WV, Bumgarner RE, Nelson PS, Schummer B, Bednarski DW, Hassell L, Baldwin RL, Karlan BY and Hood L: Comparative hybridization of an array of 21500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. Gene *238*: 375-385, 1999.

3 Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D and Levine AJ: Broad patterns of gene expressions revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci *96*: 6745-6750, 1999.

4 Golub TR, Slonim DK, Tamayo P, Guard C, Gaasenbeek M, Mesirov JP, Coller HC, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD and Lander ES: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*: 531-537, 1999.

5 Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossoss IS, Rosenwald A, Boldrick JC, Sabeh H, Tran T and Yu X: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*: 503-511, 2000.

6 Khan J, Wei JS, Ringier M, Saal LH, Ladanyi M and Westermann F: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Med *7*: 673-679, 2001.

7 Van't Veer LJ, Dai H, Van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, Van der Kooy K, Marton MJ and Witteveen AT: Gene expression profiling predicts clinical outcome of breasts cancer. Nature *415*: 530-536, 2002.

8 Dudoit S, Fridlyand J and Speed T: Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc *97*: 77-87, 2002.

9 Brown MPS, Grundy VN, Lin D, Cristianini N, Sugnet C, Furey TS, Ares M and Haussler D: Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci *97*: 262-267, 1997.

10 Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E and Mesirov JP: Multiclass cancer diagnosis using tumor expression signatures. Proc Natl Acad Sci *98*: 15149-15154, 2001.

11 Nguyen DV and Rocke DM: Tumor classification by partial least squares using microarray gene expression data. Bioinformatics *18*: 39-50, 2002.

12 Li Y, Campbell C and Tipping M: Bayesian automatic relevance determination algorithms for classifying gene expression data. Bioinformatics *18*: 1332-1339, 2002.

13 Dettling M and Behlmann P: Boosting for tumor classification with gene expression data. Bioinformatics *19*: 1061-1069, 2003.

14 Dettling M: Bagboosting for tumor classification with gene expression data. Bioinformatics *20*: 3583-3593, 2004.

15 Theodoridis S and Koutroumbas K: Pattern Recognition. Academic Press, San Diego, 1999.

16 Novak JP, Sladek R and Hudson TJ: Characterization of variability in large-scale gene expression data: implication for study design. Genomics *79*: 104-114, 2002.

17 Vapnik VN: Statistical Learning Theory. Wiley Interscience, New York, 1998.

18 Burges CJC: A tutorial on support vector machines for pattern recognition. Data Mining Knowledge Discov *2*: 121-167, 1998.

19 Joachims T: Making large-scale support vector machine learning practical. Adv Kernel Methods: support vector learning, 169-185, 1999.

20 Dietterich TG: Machine learning research: four current directions. AI Magazine *18*: 97-136, 1997.

21 Friedman J, Hastie T and Tibshirani R: Additive logistic regression: a statistical view of boosting. Annal Stat *28*: 337-407 (with discussion), 2002.

22 Yan R, Liu Y, Jin R and Hauptmann A: On predicting rare classes with SVM ensembles in scene classification. IEEE ICASSP, 2003.

23 Kim HC, Pang S, Je HM, Kim D and Bang SY: Constructing support vector machines ensemble. Pattern Recog *36*: 2757-2767, 2003.

24 Schölkopf B and Smola AJ: Support vector machines and kernel algorithms, MSR-TR 2000-23. Microsoft Research, 2000.

25 Dietterich TG: An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization. Machine Learning *40*: 139-158, 2002.

26 Weston J and Watkins C: Support vector machines for multi-class pattern recognition. Proceedings of ESANN, 1999.

27 Bottou L, Cortes C, Denker J, Drucker H, Guyon I, Jackel J, LeCun Y, Muller U, Sackinger E, Simard P and Vapnik V: Comparison of classifier methods: a case study in handwritten digit recognition. Proceedings of International Conference on Pattern Recognition, 1994.

28 Lee Y and Lee CK: Classification of multiple cancer types by multicategory SVM using gene expression data. Bioinformatics *19*: 1132-1139, 2003.

29 Guyon I, Weston J, Barnhill S and Vapnik V: Gene selection for cancer classification using support vector machines. Machine Learning *46*: 389-422, 2002.

30 Singh D, Febbo P, Ross K, Jackson D, Manola J, Ladd C, Tamayo P, D' Amico A and Richie J: Gene expression correlates of clinical prostate cancer behavior. Cancer Cell *1*: 203-209, 2002.

31 Pomeroy S, Tamayo P, Gaasenbeeck M, Sturla L, Angelo M, McLaughlin M, Kim J, Goumnerova L, Black L and Lau C: Prediction of central nervous system embryonal tumor outcome based on gene expression. Nature *415*: 436-442, 2002.

32 Troyanskaya O, Cantor M, Shelock G, Brown P, Hastie T, Tibshirani R, Botstein D and Altman RB: Missing value estimation methods for DNA microarrays. Bioinformatics *17*: 520-525, 2001.

33 Chang CC and Lin CJ: LIBSVM: a library for support vector machines, 2003. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Received June 29, 2005
Accepted December 1, 2005