Methods for Multi-Category Cancer Diagnosis from Gene Expression Data: A Comprehensive Evaluation to Inform Decision Support System Development

Alexander Statnikov, Constantin F. Aliferis, Ioannis Tsamardinos

Discovery Systems Laboratory, Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA

Abstract

Cancer diagnosis is a major clinical applications area of gene expression microarray technology. We are seeking to develop a system for cancer diagnostic model creation based on microarray data. In order to equip the system with the optimal combination of data modeling methods, we performed a comprehensive evaluation of several major classification algorithms, gene selection methods, and cross-validation designs using 11 datasets spanning 74 diagnostic categories (41 cancer types and 12 normal tissue types). The Multi-Category Support Vector Machine techniques by Crammer and Singer, Weston and Watkins, and one-versus-rest were found to be the best methods and they outperform other learning algorithms such as K-Nearest Neighbors and Neural Networks often to a remarkable degree. Gene selection techniques are shown to significantly improve classification performance. These results guided the development of a software system that fully automates cancer diagnostic model construction with quality on par with or better than previously published results derived by expert human analysts.

Keywords:

Artificial Intelligence, Support Vector Machines, Diagnosis, Computer-Assisted, Oligonucleotide Array Sequence Analysis.

Introduction

An important emerging medical application domain for microarray gene expression profiling technology is clinical decision support in the form of diagnosis of disease as well as prediction of clinical outcomes in response to treatment. The two areas in medicine that currently attract the greatest attention are management of cancer and infectious diseases [1,2]. A necessary prerequisite for the creation of clinically successful microarray-based diagnostic models is a solid understanding of the relative strengths and weaknesses of available classification and related (i.e., gene selection and cross-validation) methods. While prior research has established the feasibility of creating accurate models for cancer diagnosis, the corresponding studies conducted limited experiments in terms of the number of classifiers, gene selection algorithms, number of datasets, and types of cancer involved (e.g., [3,4]). In addition, the results of these studies cannot be combined into a comprehensive comparative meta-analysis because each study follows different experimental protocols and applies learning algorithms differently. Thus, it is not clear from the literature which classifier (if any) performs best among the many available alternatives. It is also currently poorly understood what the best combinations of classification and gene selection algorithms are across most array-based cancer datasets. Another major methodological concern is the problem of *overfitting*; that is creating diagnostic models that may not generalize well to new data (from the same cancer types and data distribution) despite excellent performance on the training set. Since many algorithms are highly parametric and datasets consist of a relatively small number of high-dimensional samples, it is easy to overfit both the classifiers and the gene selection procedures especially when using intensive model search and powerful learners. Indeed recently, a number of reports appeared in the literature raising doubts about the generalization ability of classifiers produced by major studies in the field [5,6]. In recent meta-analytic assessment of 84 published microarray cancer outcome predictive studies [2] it was found that 74% of studies did not perform independent validation or cross-validation of proposed findings, 13% applied cross-validation in an incomplete fashion, and only 13% performed cross-validation correctly.

The major motivation of this research is to build *a fully-automated software system* that generates *high-quality* (ideally optimal) and *robust* (i.e., non-overfitted) diagnostic and prognostic models for use in clinical applications. For such a system to be successful, it must implement the best possible classification and gene selection algorithms for the domain and guide model selection by enforcing sound principles of data analysis. Hence, to inform the development of such a system, the goals of the present work are to: (a) investigate which one among the many powerful classifiers currently available for gene expression diagnosis performs the best *across many cancer types*; (b) how classifiers interact with existing gene selection methods in datasets with varying sample size, number of genes, and cancer types; (c) how to parameterize the classifiers and gene selection procedures so as to *avoid overfitting*.

Methods and Materials

Classification algorithms

To maintain a focus on realistic medical applications, we consider only classification algorithms that can handle multiple classes and a large number of variables. We also consider only algorithms that are fairly insensitive to noise and large variable-to-sample ratios (as is appropriate in gene expression array analysis). Given the above and based on the results of prior studies in this domain, we selected for our experiments Multi-Category Support Vector Machines (MC-SVMs), Neural Networks (NNs), and K-Nearest Neighbors (KNN). We also conducted additional experiments (reported only in part

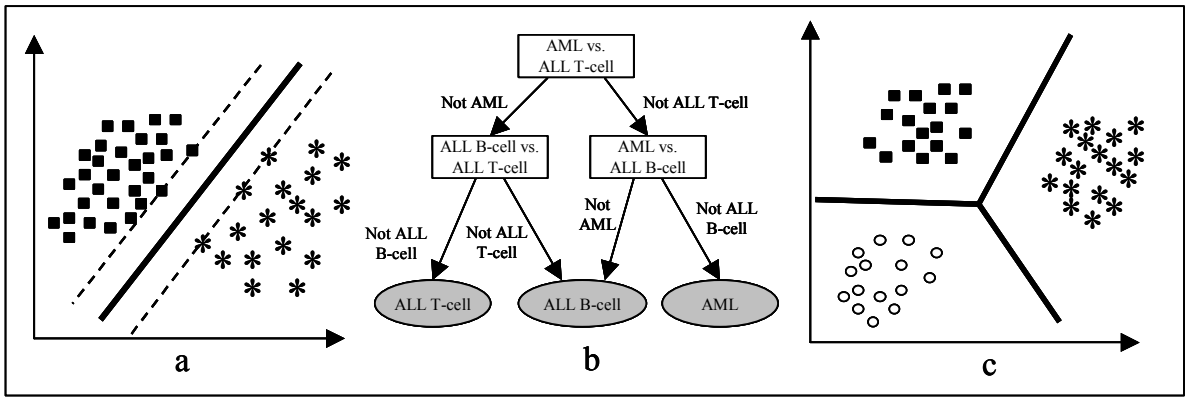

Figure 1 – (a) A binary SVM selects a hyperplane (bold line) that maximizes the width of the "gap" (margin) between the two classes. New cases are classified according to the side of the hyperplane they fall into. (b) Multi-category SVM algorithm DAGSVM construct a decision tree on the basis of one-versus-one binary SVM classifiers. (c) MC-SVM algorithms WW and CS maximize margin between all classes simultaneously.

here due to space limitations) with Decision Tree (DT) Induction, Weighted Voting (WV), and various flavors of Ensemble Classification (EC) methods (the detailed results for the latter three learning algorithm families can be found in the online supplement of the present paper [7]).

The main idea of Support Vector Machines **(SVMs)** [8] is to map the data (implicitly) to a higher dimensional space via a kernel function and then identify the maximum-margin hyperplane that separates training instances. New instances are classified according to the side of the hyperplane they fall into (Figure 1a). The optimization problem is most often formulated in a way that allows for non-separable data by penalizing misclassifications.

Multi-Category SVMs (**MC-SVMs**) extend the idea behind binary SVMs to multiple categories (classes). Two algorithmic families of MC-SVMs were used in the present study. The first family includes algorithms based on the solution of binary SVM problems: one-versus-rest (**OVR**) [9] which separates each class from all others and constructs a combined classifier incorporating tie-resolution strategies; one-versus-one (**OVO**) [9] which separates all classes pairwise and constructs a combined classifier using voting schemes and tie-breaking strategies; and **DAGSVM** [10] which constructs a decision tree on the basis of all binary OVO classifiers (Figure 1b). The second family includes methods based on consideration of all classes at once: the method by Weston and Watkins (**WW**) [11] and the one by Crammer and Singer (**CS**) [12] (Figure 1c). Detailed descriptions of the algorithms with intuitive pictorial examples and exact mathematical formulations can be found in [7].

The K-Nearest Neighbors, Neural Networks, Decision Trees, Ensemble Classifiers, and Weighted Voting methods are fairly standard in informatics so we omit corresponding introductory descriptions (but we include all details in [7]).

Parameters for the classification algorithms

Parameters for the classification algorithms were chosen by nested cross-validation procedures (details presented in the experimental design subsection) to optimize performance while avoiding overfitting. For all five MC-SVM methods we used a polynomial kernel and performed classifier optimization over degrees $=\{1,2,3\}$ and costs $=\{0.0001,0.01,1,100\}$. We optimized the KNN classifier over all values of K (number of neighbors) ranging from 1 to total number of instances in the training dataset. We used feed-forward NNs with one hidden layer. The number of units was chosen heuristically from the set $\{2,5,10,30,50\}$. We employed gradient descent with adaptive learning rate backpropagation, mean squared error performance goal set to 10^{-8} (an arbitrary value very close to zero), fixed momentum of 10^{-3}, and an optimal number of epochs chosen from the range [100,10000] based on the error on the validation set.

Datasets and data preparatory steps

The 11 datasets used in this work are described in Table 1. They had 2-26 distinct diagnostic categories, 50-308 samples (patients), and 2308-15009 variables (genes). In total the 11

Table 1 – Cancer-related human gene expression datasets used in this study. In addition to 9 multicategory datasets, 2 datasets with two diagnoses were included to empirically confirm that MC-SVM methods behave as well as binary SVMs in binary classification tasks (as theoretically expected).

Dataset name	Diagnostic Task	Number of				Max. prior
		Samples	Variables (genes)	Categories	Variables / Samples	
11_Tumors	11 various human tumor types	174	12533	11	72	15.5%
14_Tumors	14 various human tumor types and 12 normal tissue types	308	15009	26	49	9.7%
9_Tumors	9 various human tumor types	60	5726	9	95	15.0%
Brain_Tumor1	5 human brain tumor types	90	5920	5	66	66.7%
Brain_Tumor2	4 malignant glioma types	50	10367	4	207	30.0%
Leukemia1	Acute myelogenous leukemia (AML), acute lympboblastic leukemia (ALL) B-cell, and ALL T-cell	72	5327	3	74	52.8%
Leukemia2	AML, ALL, and mixed-lineage leukemia (MLL)	72	11225	3	156	38.9%
Lung_Cancer	4 lung cancer types and normal tissues	203	12600	5	62	68.5%
SRBCT	Small, round blue cell tumors (SRBCT) of childhood	83	2308	4	28	34.9%
Prostate_Tumor	Prostate tumor and normal tissues	102	10509	2	103	51.0%
DLBCL	Diffuse large b-cell lymphomas (DLBCL) and follicular lymphomas	77	5469	2	71	75.3%

datasets span 74 diagnostic categories (41 cancer types and 12 normal tissue types) and are available for download from [7].

We note that no new methods to preprocess gene expression data were invented. We relied instead on standard normalization and data preparatory steps performed by the authors of the primary dataset studies. In addition, we performed a simple rescaling of gene expression values (to be between 0 and 1) to speed up training of SVMs. The rescaling was performed based on the training set in order to avoid overfitting.

Experimental design for model selection and evaluation

Two experimental designs were employed to obtain reliable performance estimates and avoid overfitting. Both experimental designs are based on two loops. The inner loop is used to determine the best parameters of the classifier (i.e. values of parameters yielding the best performance on the validation dataset). The outer loop is used for estimating the performance of the classifier built using the previously found best parameters by testing on the *independent set of patients*. Design I uses a stratified 10-fold cross-validation in the outer loop and a stratified 9-fold cross-validation in the inner loop [13]. It is often referred to as "nested stratified 10-fold cross-validation." See Figure 2 for a simplified pictorial example of a 3-fold Design I applied to 3 patient groups (P1, P2, P3) with optimization of parameter *C* (which takes values "1" or "2") of some classifier (Note, that in reality we do not optimize just one parameter but, rather, a large set of combined parameters). Design II uses leave-one-out cross-validation (LOOCV) in the outer loop and a stratified 10-fold cross-validation in the inner loop. We chose to employ both designs because there exists contradictory evidence in the machine learning literature regarding whether *N*-fold cross-validation provides more accurate performance estimates than LOOCV and vice-versa for zero-one loss classification [14].

Gene selection

To study how dimensionality reduction can improve classification performance, we applied all classifiers with subsets of 25, 50, 100, 500, and 1000 top-ranked genes (following the example set by [15]). Genes were selected according to four gene selection methods/metrics: (1) ratio of genes between-categories to within-category sums of squares (BW) [4]; (2-3) signal-to-noise (S2N) scores [3] applied in a one-versus-rest (S2N-OVR) and one-versus-one (S2N-OVO) fashion; and (4) Kruskal-Wallis nonparametric one-way ANOVA (KW). *The ranking of the genes was performed based on the training set of samples to avoid overfitting.*

Performance metrics

We used two classification performance metrics. The first metric is accuracy since we wanted to compare our results with the previously published studies that also used this performance metric. Accuracy is easy to interpret and simplifies statistical testing. On the other hand, accuracy is sensitive to the prior class probabilities and does not fully describe the actual difficulty of the decision problem for highly unbalanced distributions. For example, it is more difficult to achieve an accuracy of 50% for a 26-class dataset *14_Tumors* (with prior probability of the major class = 9.7%) compared to an accuracy of 75% for a binary dataset *DLBCL* (with prior of the major class = 75.3%).

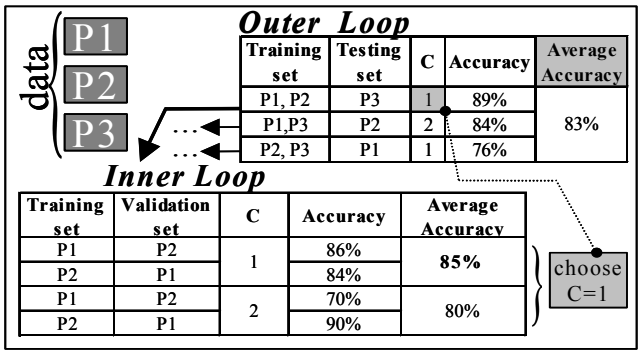

Figure 2 - Pictorial simplified example of Design I. The data are split into mutually exclusive sets P1, P2 and P3. The performance is estimated in the outer loop by training on all splits but one, using the remaining one for testing. The average performance over testing sets is reported. The inner loop is used to discover the optimal value of parameter C (in a cross-validated fashion) for training in the outer loop.

The second metric is *relative classifier information* (RCI), which corrects for differences in base-rates of diagnostic categories, as well as the number of categories. RCI is an entropy-based measure that quantifies *how much the uncertainty of a decision problem is reduced by a classifier relative to classifying using the priors* [16]. We also considered using (1) generally-accepted misclassification functions, but could not identify any for the cancer domain, and (2) emerging multi-class extensions of the area under ROC curve [17,18], but we decided against them because currently such methods are not well-suited to our problem characteristics.

Overall research design

To maintain the feasibility of this study, we pursued a *staged factorial design*: in stage I, we conducted a fully factorial design involving datasets and classifiers without gene selection; in stage II, we focused on the datasets for which the full gene sets yielded poor performance and applied (in a factorial fashion) gene selection. In addition, we optimized algorithms using accuracy only and limited the possible cardinalities of selected gene sets to only five choices (see subsection on gene selection).

While the above choices restricted the number of models generated, *the resulting analyses still generated a total of ~$2.6 \cdot 10^6$ diagnostic models*. The total time required was 4 single-CPU months (platform used: Intel Xeon 2.4 GHz). Out of this set of models, only one model was selected for each combination of algorithm and dataset.

Notice that, despite the very large number of examined models, the final performance estimates are not overfitted. This is because only one model is selected per split for the estimation of the final performance and it is applied to previously *unseen cases*. Thus, regardless of how much performance is overestimated in the inner loop (which, in the worst case, may result in not choosing the best possible parameters' combination), the outer loop guarantees proper estimation of performance.

Statistical comparison among classifiers

To test if differences in accuracy between the best method (i.e. one with the largest average accuracy) and all remaining algorithms are non-random, we need a statistical comparison of observed differences in accuracies. We used random permuta-

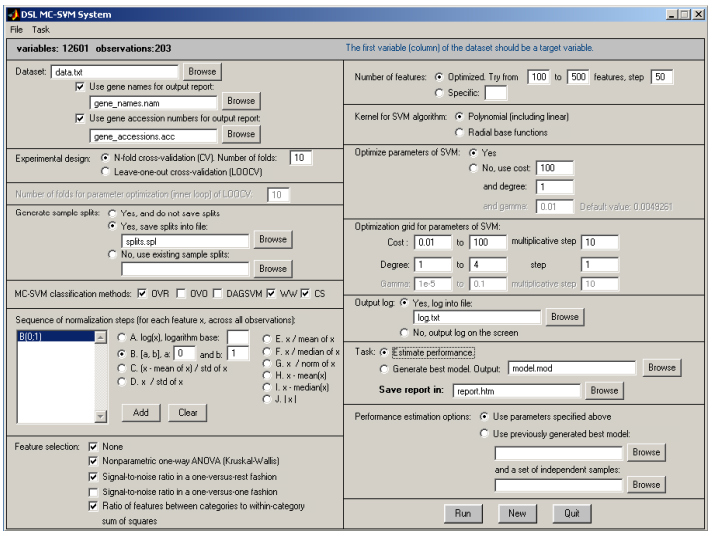

Figure 3 - Screenshot of the GEMS system. Most fields are automatically filled out with default values. All experiments in this study can be replicated using the system with a few clicks of the mouse.

tion testing (with an alpha level of 0.05), which does not rely on independence assumptions (that are clearly violated in cross-validated datasets) and can be applied to several datasets [19]. The specific details of how this test works can be found in [7].

Implementation

We used the MC-SVM algorithms implemented by the LibSVM team [20], since they use state-of-the-art optimization methods for the solution of MC-SVM problems. The implementations of NN and DT classifiers were based on the Matlab Neural Networks and Statistics toolboxes, respectively. We used our own implementations of KNN, WV, and EC algorithms, as well as gene selection and statistical comparison.

The prototype analysis system ***GEMS*** (Gene Expression Model Selector) based, on the results and analyses reported here, was built using Matlab R13 and MS Visual C++ 6. *GEMS* has a graphics user interface consisting of a single form (Figure 3) and is freely available for download from [7].

Results

The performance results (accuracies and entropy-based measure RCI) of experiments using Design I without gene selection are summarized in Table 2. Results for Design II are very similar and due to space limitations are provided only in [7].

Table 2 – Performance results (accuracies and RCI) without gene selection obtained using a nested stratified 10-fold cross-validation design. These results are further improved by gene selection (see Figure 4).

		Accuracy		RCI	
Method		Average	Range	Average	Range
MC-SVM	OVR	**88.46%**	65.1% - 100%	**87.54%**	71.14% - 100%
	OVO	80.64%	47.07% - 100%	84.72%	64.99% - 100%
	DAGSVM	80.71%	47.35% - 100%	84.55%	65.64% - 100%
	WW	**86.52%**	62.24% - 100%	**86.29%**	71.14% - 100%
	CS	**88.69%**	65.33% -100%	**87.17%**	71.14% - 100%
non-SVM	KNN	74.44%	43.9% - 89.64%	69.74%	51.09% - 83.93%
	NN	59.49%	11.12% - 91.03%	58.97%	16.24% - 87.50%

The fact that we obtained similar results with two different experimental designs is evidence in favor of the reliability of the performance estimates.

In 8 out of 11 datasets, *MC-SVMs perform cancer diagnoses with accuracies > 90%* and in 7 datasets *with RCI > 90%. Overall, all MC-SVMs outperform KNN and NNs significantly. MC-SVM methods CS, OVR, and WW yield the best results* (and are not statistically significantly different from each other). On the other hand, OVO, DAGSVM, KNN, and NNs have poorer performance than the above three methods to a statistically significant degree.

The summary of classification with four gene selection methods (BW, S2N-OVR, S2N-OVO, and KW) applied to the most challenging datasets (*9_Tumors*, *14_Tumors*, *Brain_Tumor1*, and *Brain_Tumor2*) is presented in Figure 4.

The results show that gene selection improves classification accuracy of KNN and NNs significantly (up to 14.97% and 59.78%, in absolute terms respectively). Although KNN and NNs with gene selection performed closer to MC-SVMs than without gene selection, MC-SVM algorithms still outperformed KNN and NNs in most cases. Gene selection also improves accuracy of MC-SVMs (up to 9.53%) and, hence, it improves accuracy of the overall best classifier. Neither of the four gene selection methods performs significantly better than the other ones. The analysis of gene selection results with RCI performance measure leads to the same set of conclusions [7].

The reported classification results are equal to or better than those of previously published models on the same datasets [7].

Finally, all MC-SVM algorithms have the same accuracy for the two binary classification problems (*DLBCL* and *Pros-*

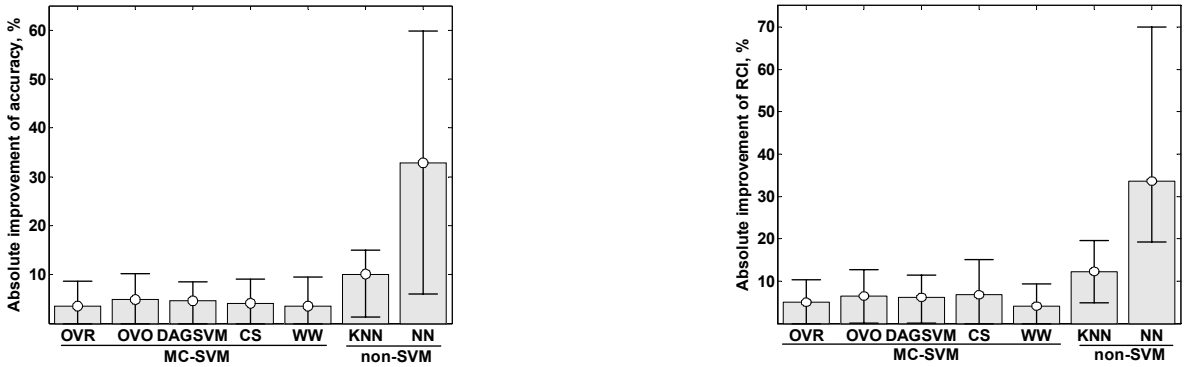

Figure 4 – Average absolute improvement of classification accuracy (left) and RCI (right) (averaged over 4 datasets) by performing gene selection. Minimum and maximum absolute improvement is shown with error-bars.

tate_Tumor) as expected. In additional classification experiments with DT, WV and EC methods, we found that both with and without gene selection, DTs perform significantly worse than MC-SVMs, worse than KNN, and similarly or worse than NNs. Similarly, WV classifiers are significantly outperformed by MC-SVMs and do not perform better than KNN and NNs. EC does not improve performance of the best non-ensemble models. The details of these additional experiments with DT, WV, and EC can be found in [7].

Conclusions and Limitations

The emergence of new cancer gene expression datasets in our institution and elsewhere will allow us to conduct a prospective evaluation of the *GEMS* system to study its ability to facilitate creation of powerful diagnostic models. We also plan to augment the preliminary version of the system with a wizard-like graphics user interface that will make *GEMS* usable by researchers with limited expertise in data-analysis.

A particularly interesting direction for future research is to improve our existing gene selection procedures by multivariate Markov blanket and local neighborhood algorithms. These techniques have been previously successfully applied to cancer gene expression domain and have the advantage of causal interpretability under fairly broad assumptions [21].

The contributions of the present study are two-fold. The *first contribution* is that we conducted the most comprehensive systematic evaluation to date of multi-category diagnosis algorithms applied to the majority of multi-category cancer-related gene expression human datasets publicly available. Based on results of this evaluation, the following conclusions can be drawn:

(a). *Multi-Category Support Vector Machines is the best family of algorithms for this type of data and medical tasks.* They outperform other popular non-SVM machine learning techniques by a large margin.

(b). Among MC-SVM methods, the ones by Crammer and Singer, Weston and Watkins, and one-versus-rest have superior classification performance.

(c). The performance of both MC-SVM and non-SVM methods can be moderately (for MC-SVMs) or significantly (for non-SVM) improved by gene selection.

We believe that practitioners and software developers should take note of these results when considering construction of decision support systems in this domain, or when selecting algorithms for inclusion in related analysis software.

The *second contribution* is that we created the fully-automated software system *GEMS* that automates the experimental procedures described in this paper to (1) develop optimal classification models for the domain of cancer diagnosis with microarray gene expression data and (2) estimate their performance in future patients. The results obtained by the system in a labor-efficient manner appear to be on par with or better than previously published results in the literature on the same datasets. Although several commercial and academic software tools do exist for gene expression classification (e.g., [22]) t*o the best of our knowledge GEMS treats the task in the most comprehensive manner and is the first such system to be informed by a rigorous analysis of the available algorithms and datasets.* We hope that the methodology presented in the present paper may encourage similar principled treatment of other software development efforts in clinical bioinformatics. The system is freely available for download from [7] for non-commercial use.

Acknowledgements

This research was supported by NIH grants RO1 LM007948-01 and P20 LM 007613-01.

References

[1] Fortina P, et al. "Molecular diagnostics: hurdles for clinical implementation", Trends Mol Med. 2002; 8(6):264-6.

[2] Ntzani EE and Ioannidis JP. "Predictive ability of DNA microarrays for cancer outcomes and correlates: and empirical assessment", Lancet. 2003; 362(9394):1439-44.

[3] Ramaswamy S, et al. "Multiclass cancer diagnosis using tumor gene expression signatures", PNAS USA 2001; 98(26):15149-54.

[4] Lee Y and Lee CK. "Classification of multiple cancer types by multicategory support vector machines using gene expression data", Bioinformatics 2003; 19(9):1132-9.

[5] Schwarzer G and Vach W. "On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology", Stat Med. 2000; 19(4):541-61.

[6] Reunanen J. "Overfitting in Making Comparisons Between Variable Selection Methods", Journal of Machine Learning Research 2003; 3:1371-82.

[7] Statnikov A, Aliferis CF, and Tsamardinos I. "Online Supplement", http://discover1.mc.vanderbilt.edu/discover/public/GEMS

[8] Vapnik V. "Statistical Learning Theory", Wiley, 1998.

[9] Kressel U. "Pairwise classification and support vector machines", In Advances in Kernel Methods: Support Vector Learning (Chapter 15), MIT Press, 1999.

[10] Platt J, Cristianini N, and Shawe-Taylor J. "Large margin dags for multiclass classification", Proc. of NIPS, 2000.

[11] Weston J. and Watkins C. "Support Vector Machines for Multi-Class Pattern Recognition", Proc. of the 7th European Symposium On Artificial Neural Networks, 1999.

[12] Crammer K. and Singer Y. "On the Learnability and Design of Output Codes for Multiclass Problems", Proc. of COLT, 2000.

[13] Weiss SM and Kulikowski CA. "Computer systems that learn", Morgan Kaufmann, 1991.

[14] Kohavi R. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", IJCAI, 1995.

[15] Furey TS, et al. "Support vector machine classification and validation of cancer tissue samples using microarray expression data", Bioinformatics 2000; 16:906-14.

[16] Sindwani V, et al. "Information Theoretic Feature Crediting in Multiclass Support Vector Machines", First SIAM International Conference on Data Mining, Chicago, 2001.

[17] Mossman D. "Three-way ROCs", Medical Decision Making, 1999; 19:78-89.

[18] Ferri C, et al. "Volume under the ROC Surface for Multi-Class Problems", Proc. of 14th ECML, 2003.

[19] Noreen EW. "Computer intensive methods for testing hypotheses", Wiley, Canada, 1989.

[20] Chang C and Lin C. "LIBSVM: a library for Support Vector Machines", http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2003.

[21] Aliferis CF, et al. "HITON, A Novel Markov Blanket Algorithm for Optimal Variable Selection", AMIA, 2003.

[22] GeneCluster2, http://www.broad.mit.edu/cancer/software/genecluster2/gc2.html

Address for correspondence

Alexander Statnikov, Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37232, USA.
E-mail: alexander.statnikov@vanderbilt.edu