

# Gene expression-based prediction of malignancies

*Giorgio Valentini*

DISI - Dipartimento di Informatica e Scienze  
dell'Informazione  
Università di Genova, (Italy)  
INFN - Istituto Nazionale di Fisica della Materia  
Unità di Genova  
e-mail: valenti@disi.unige.it

## ABSTRACT

Molecular classification of malignancies can potentially stratify patients into distinct subclasses not detectable using traditional classification of tumors, opening new perspectives on the diagnosis and personalized therapy of polygenic diseases. In this paper we present a brief overview of our work on gene expression based prediction of malignancies, starting from the dichotomic classification problem of normal versus tumoural tissues, to multiclass cancer diagnosis and to functional class discovery and gene selection problems. The last part of this work present preliminary results about the application of ensembles of SVMs based on bias-variance decomposition of the error to the analysis of gene expression data of malignant tissues.

## 1 Introduction

DNA microarray technology [14] opens new prospects to the traditional classification of human malignancies based on morphological and clinical parameters. In fact, information obtained by gene expression data gives a snapshot of the overall functional status of a cell, offering new insights into potential different types of malignant tumours, based on functional and molecular discrimination. Although this technology does not take into account the information available at the translational and post-translational levels, it provides fundamental insights into the mRNA levels of all the genes, offering in such a way an approximate picture of the proteins of a cell or tissue at one time.

The large amount of gene expression data require statistical and machine learning methods to analyze and extract significant knowledge from DNA microarray data. After image analysis and preprocessing steps have been performed, data are usually collected in matrices: each row corresponds to a different microarray experiment, (or to an average between replicated microarray experiments), and columns to the expression levels of the genes involved in the experiment.

Typical problems arising from gene expression data analysis range from prediction of malignancies [28, 23]

(a classification problem from a machine learning point of view) to functional discovery of new classes or subclasses of diseases [1] (an unsupervised machine learning problem), to the identification of groups of genes responsible or correlated with malignancies or polygenic diseases [12] (a feature selection machine learning problem).

In a typical unsupervised approach, no or limited a priori knowledge is available: only gene expression levels of the genes across different microarray experiments are available. In this case we can group together sets of genes (columns of the data matrix), or different cells or different functional status of the same tissue (rows of the data matrix), in order to discover similar patterns between gene expression levels of sets of genes or groups of samples (tissues) considering the overall expression of all the genes involved in the experimentation. Clustering algorithms are used to group together similar expression patterns: sets of genes, or different cells or different functional status of the cell are grouped together. Typical ensemble methods used in the literature are hierarchical clustering [9], k-means [22], self-organizing maps [20], graph-based algorithms [19] and biclustering methods [21].

On the other hand, supervised methods exploit a priori biological and medical knowledge on the problem domain. For instance, learning algorithms with labeled examples are used to associate gene expression data with classes, in order to separate normal from cancerous tissues or to classify different classes of cells on functional basis, or to predict the functional class of unknown genes. Several supervised methods have been applied to the analysis of cDNA microarrays and high density oligonucleotide chips. These methods include decision trees, Fisher linear discriminant, Multi-Layer Perceptrons (MLP), Nearest-Neighbours classifiers, linear discriminant analysis, Parzen windows and others [11, 8, 3, 17, 13]. In particular Support Vector Machines (SVM) have been recently applied to the analysis of DNA microarray gene expression data in order to classify functional groups of genes, normal and malignant tissues and multiple tumor types [3, 10, 28].

In this paper we summarize our research on the prediction of malignancies using supervised machine learning methods (Sect. 2), including ECOC ensembles of learning machines for multiple tumor types prediction (Sect. 3). Then we outline our present research items about the application of SVM ensembles to gene expression-based prediction of malignancies, and in particular we summarize our present work on bias-variance decomposition based methods for gene expression data analysis (Sect. 4).

## 2 Gene expression-based prediction of malignancies with SVMs and MLPs

We tried to separate malignant from normal lymphoid cells using data of a specialized DNA microarray, named

”Lymphochip”, developed at Stanford University School of Medicine [1], specifically designed to study lymphoid and malignant development related genes. These data are very challenging from a machine learning standpoint, considering that they are constituted by a small number (96) of 4026-dimensional samples.

Considering the high dimensional input space we applied SVMs with  $L_1$ -norm soft-margin as they are well-suited to work with high dimensional data. In fact they represent an implementation of the Vapnik’s structural risk minimization induction principle [27], or can be equivalently interpreted as regularized learning machines that address the trade-off between the accuracy on the learning set through the minimization of the number of misclassified patterns and the generalization capabilities through the maximization of the margin separating the classes.

Three different types of SVMs, with linear, polynomial and radial basis kernel functions have been applied to the separation of malignant from normal lymphoid cells. Moreover we used also as baseline classifiers a Multi-Layer Perceptrons (MLP) with one hidden layer and a simple perceptron (LP) [25].

MLP with 10 hidden neurons showed an estimated generalization error of about 2% (using 10-fold cross validation), while SVM-linear achieved the best results (about 1% error) for a large range of values of the regularization parameter ( $1 \leq C \leq 1000$ ). Comparing our results with those obtained in [1] using hierarchical clustering, we achieved, as expected, a significant improvement of the classification accuracy.

In several classification tasks, the accuracy is not a sufficient criterion to evaluate the performance of a classifier. For instance, in this case, considering the detection of seriously diseased patients, it is preferable to avoid false negative results rather than false positive ones. In order to address this problem the Receiver Operating Characteristic (*ROC*) analysis offers a suitable tool to jointly evaluate the relationships between the rate of true negative and false positive examples. Performing a ROC analysis we concluded that supervised machine learning methods correctly separate malignant from normal lymphoid tissues, but only linear SVM and MLP can be used to build classifiers with a high-sensitivity and a low rate of false positives. Indeed Fig. 1 shows that the ROC curve of SVM-linear is ideal (that is the area under the curve is 1), and nearly optimal is also the MLP ROC curve, while the linear perceptron shows a worse ROC curve, but with reasonable values lying on the highest and leftmost part of the ROC plane. However, for medical applications, the sensitivity of the LP is surely too low.

We proposed also a two stage unsupervised/supervised approach [24, 23] performing an experimental analysis that supports the hypothesis of Alizadeh et al. [1] about the existence of two distinct subgroups of *Diffuse Large B-Cell Lymphoma* (DLBCL) using subsets of functionally correlated genes (expression signatures)

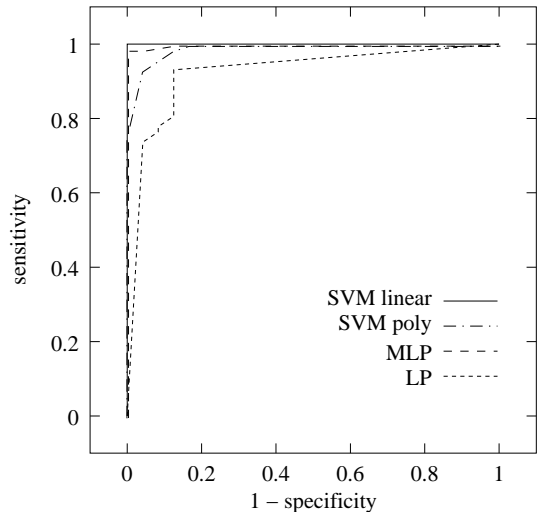


Figure 1: Comparison of ROC curves between SVM, LP and MLP for the classification problem of separating malignant from normal tissues.

individuated by clustering algorithms and by the a priori biological knowledge of their functions. Among the selected expression signatures, we identified a group of coordinately expressed genes specifically related to this separation using a supervised approach [23].

### 3 Gene expression-based prediction of multiple tumor types with ECOC ensemble methods

Molecular multiple-class classification of tumors using gene expression data is a particularly hard task, as it is characterized by a large dimensionality of the datasets, a small number of examples, a small but significant uncertainty in the original labels, the noise in the experimental and measurement processes and the intrinsic biological variation from specimen to specimen. Previous works try to use ensembles of SVMs using a One-Versus-All and an All Pairs approach [28, 18], achieving quite encouraging results, but exploiting only in part the error recovering capabilities of *Output Coding* (OC) decomposition methods [5].

We tried to directly classify different types of lymphoma (a multi-class problem) using OC ensembles. OC methods are characterized by a divide-et-impera approach to multi-class classification: a multi-class problem is decomposed in a set of  $L$  two-class subproblems, then the resulting ensemble of dichotomizers is trained and subsequently the outputs of the  $L$  dichotomizers are combined to predict the class label.

ECOC methods improve the accuracy of the multi-classifier systems as they enhance the error recovering capabilities of the system itself (e.g., even if the outcome of some classifier of the ensemble is wrong, the overall prediction of the system can be correct, espe-

cially if the codewords coding the classes are dissimilar [6]. Considering that recent studies show clearly that there is a trade-off between the error recovering capabilities of ECOC and the complexity of the induced dichotomies [15], we used MLPs as base learners for ECOC ensembles. In principle also other "strong" learners could be used, but, with SVMs for instance, output normalization and correct decoding functions must be carefully selected in order to fully exploit the error recovering capabilities of ECOC.

In particular we used One-Per-Class Parallel Non linear Dichotomizers (OPC-PND) and Error-Correcting-Output-Coding Parallel Non linear Dichotomizers (ECOC-PND) ensembles based on output coding methods [16] (Fig. 2) and a multi-class MLP as reference. Each MLP of the ensemble was independently trained

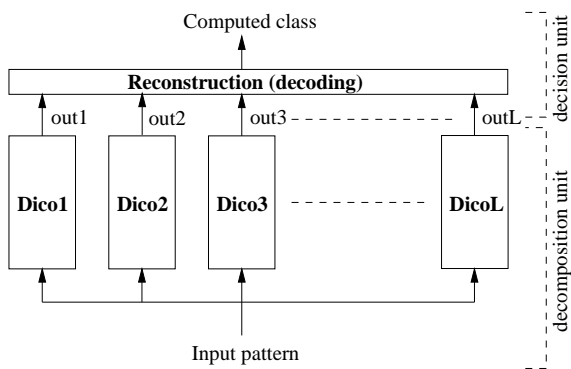


Figure 2: Schematic architecture of a *PND* ensemble of dichotomizers.

to learn each individual bit of the codeword coding the classes. The decision unit was implemented using the  $L_1$  norm distance between the outputs of the decomposition unit (that is the vector of the continuous outputs of the base dichotomic MLP learners) and the codeword of the classes.

OPC and ECOC *PND* outperformed the multiclass MLP, showing also less sensitivity to model parameters: we obtained good results varying the number of hidden units and the parameters of the learning algorithms for a relatively large range of values. For instance, with ECOC *PND* we obtained the best results using base classifiers with 8 hidden units, but varying the number of the hidden units from 3 to 15 the predicted error ranges from 0.052 to 0.062, while with multiclass MLP, varying the number of hidden units, the error ranged from 0.062 to 0.094 [23]. Interestingly enough, an estimated accuracy of about 0.95 has been achieved using all the available genes, without applying any feature selection method.

## 4 Ensembles of SVMs for gene expression data analysis

Ensembles of learning machines are well-suited for gene expression data analysis, as they can reduce variance due to the low cardinality of the available training sets [2]. Indeed, in recent works, bagging and boosting, that are ensemble methods based on resampling techniques, one-versus-all and all-pairs combinations of binary classifiers and ECOC ensembles of MLP have been applied to the analysis of DNA microarray data [8, 28, 23].

Bias-variance decomposition of the error has been recently proposed as a tool to gain insights into the behavior of learning algorithms, in order to properly design ensemble methods well-tuned to the properties of a specific base learner [26]. For instance, analyzing how does vary the error, bias, and variance of SVMs with respect to the kernel parameters and the regularization term (see, for instance Fig. 3), we can gain insights to develop ensemble methods specific for a particular base learner. In the context of gene expression data analysis, considering that Support Vector Machines (SVM) can easily manage high dimensional data [4], a promising line of research consists in using them as base learners in ensemble methods well-tuned to their bias-variance characteristics analyzed through Domingos' unified theory on bias-variance decomposition of the error [7]. According to this general approach, we are studying ensemble methods to join the low bias properties of SVM with the low unbiased variance properties of bagging. Indeed in our analysis of the bias-variance decomposition of the error in SVMs we found, as expected, that properly tuned SVMs show a relatively low bias at the expenses of a relatively high unbiased variance, especially with low sized data sets, as DNA microarray data usually are. In order to exploit this general behaviour of the SVM learning algorithm, we proposed a basic high-level algorithm for a general *Bagged ensemble of selected low-biased SVM*:

1. Estimate bias-variance decomposition of the error for different SVM models
2. Select the SVM model with the lowest bias
3. Perform bagging using as base learner the SVM with the estimated lowest bias.

Depending on the type of kernel and parameters considered, and on the way the bias is estimated for the different SVM models, different implementations can be given. Preliminary results on a data set of 300 normal and tumor specimens spanning 14 different tumor classes, obtained from the Whitehead Institute - Massachusetts Institute of Technology Center for Genome Research, show that this approach works, outperforming both single SVMs and bagged ensembles of SVMs.

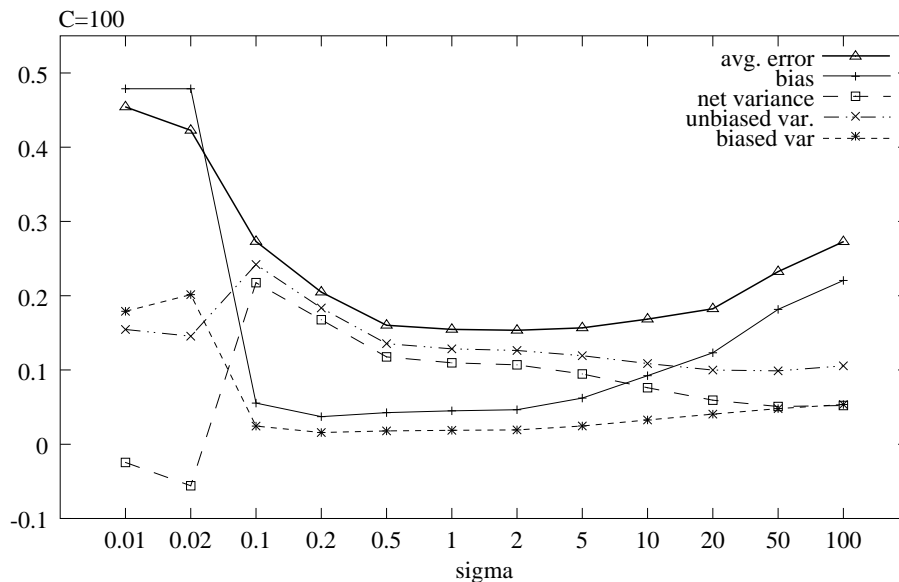


Figure 3: An example of the bias-variance decomposition of the error in SVMs with gaussian kernel, varying the spread parameter  $\sigma$  and for a fixed value of the regularization parameter  $C = 100$  for a small sized data set. The graph shows the estimated generalization error (avg. error), the bias, and the two components of the variance (unbiased and biased variance, according to the theory proposed by Domingos), and the net-variance expressed as unbiased minus biased variance. Note that the minimum of the estimated error, bias and unbiased variance occur for different values of the kernel parameter  $\sigma$ .

## REFERENCES

- [1] A. Alizadeh et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.
- [2] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
- [3] M. Brown et al. Knowledge-base analysis of microarray gene expression data by using support vector machines. *PNAS*, 97(1):262–267, 2000.
- [4] N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, UK, 2000.
- [5] T.G. Dietterich. Ensemble methods in machine learning. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems. First International Workshop, MCS 2000, Cagliari, Italy*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer-Verlag, 2000.
- [6] T.G. Dietterich and G. Bakiri. Solving multi-class learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, (2):263–286, 1995.
- [7] P. Domingos. A Unified Bias-Variance Decomposition for Zero-One and Squared Loss. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pages 564–569, Austin, TX, 2000. AAAI Press.
- [8] S. Dudoit, J. Fridlyand, and T. Speed. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Technical Report 576, Department of Statistics, University of California, Berkeley, 2000.
- [9] M. Eisen and P. Brown. DNA arrays for analysis of gene expression. *Methods Enzymol.*, 303:179–205, 1999.
- [10] T.S. Furey et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.
- [11] T.R. Golub et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286:531–537, 1999.
- [12] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1/3):389–422, 2002.
- [13] J. Khan et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7(6):673–679, 2001.

- [14] D.J. Lockhart and E.A. Winzeler. Genomics, gene expression and DNA arrays. *Nature*, 405:827–836, 2000.
- [15] F. Masulli and G. Valentini. Effectiveness of Error Correcting Output Coding decomposition schemes in ensemble and monolithic learning machines. *Pattern Analysis and Application*. (in press).
- [16] F. Masulli and G. Valentini. Parallel Non linear Dichotomizers. In *IJCNN2000, The IEEE-INNS-ENNS International Joint Conference on Neural Networks*, volume 2, pages 29–33, Como, Italy, 2000.
- [17] P. Pavlidis, J. Weston, J. Cai, and W.N. Grundy. Gene functional classification from heterogenous data. In *Fifth International Conference on Computational Molecular Biology*, 2001.
- [18] S. Ramaswamy et al. Multiclass cancer diagnosis using tumor gene expression signatures. *PNAS*, 98(26):15149–15154, 2001.
- [19] R. Sharan and R. Shamir. CLICK: a Clustering Algorithm with Applications to Gene Expression Analysis. In *Proc. 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'00)*, pages 307–316, Menlo Park, CA, 2000. AAAI Press.
- [20] P. Tamayo et al. Interpreting patterns of gene expression with self-organizing maps. *PNAS*, 96:2907–2912, 1999.
- [21] A. Tanay, R. Sharan, and R. Shamir. Discovering Statistically Significant Biclusters in Gene Expression Data. *Bioinformatics*, 1(1):1–9, 2002.
- [22] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nat. Genet.*, 22(3), 1999.
- [23] G. Valentini. Gene expression data analysis of human lymphoma using support vector machines and output coding ensembles. *Artificial Intelligence in Medicine*. (in press).
- [24] G. Valentini. Classification of human malignancies by machine learning methods using DNA microarray gene expression data. *Fourth International Conference Neural Networks and Expert Systems in Medicine and HealthCare*, pp. 399–408, Milos island, Greece, 2001.
- [25] G. Valentini. Supervised gene expression data analysis using Support Vector Machines and Multi-Layer perceptrons. In *Proc. of KES'2002, special session Machine Learning in Bioinformatics*, Amsterdam, the Netherlands, IOS Press. 2002.
- [26] G. Valentini and T.G. Dietterich. Bias–variance analysis and ensembles of SVM. In *Multiple Classifier Systems. Third International Workshop, MCS2002, Cagliari, Italy*, Lecture Notes in Computer Science. Springer-Verlag, 2002.
- [27] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
- [28] C. Yeang et al. Molecular classification of multiple tumor types. In *ISMB 2001, Proceedings of the 9th International Conference on Intelligent Systems for Molecular Biology*, pages 316–322, Copenhagen, Denmark, 2001. Oxford University Press.