Supervised gene expression data analysis using Support Vector Machines and Multi-Layer Perceptrons

Giorgio Valentini
INFM, Istituto Nazionale di Fisica della Materia,
DISI - Dipartimento di Informatica e Scienze dell'Informazione
Università di Genova, Italy
e-mail: valenti@disi.unige.it

Abstract. We apply supervised machine learning methods to the recognition of human lymphoma, using gene expression data from "Lymphochip", a specialized DNA microarray developed at Stanford University School of Medicine. We show that Multi-Layer Perceptrons and Support Vector Machines can correctly separate cancerous from normal tissues, giving also insights into the role of sets of coordinately expressed genes in carcinogenic processes of lymphoid tissues. Moreover our experimental results show that two subgroups of cells can be distinguished inside the class of *Diffuse Large B-cell Lymphoma*, confirming the existence of two distinct tumoral diseases inside this type of aggressive malignancy of mature B lymphocytes.

1 Introduction

DNA hybridization microarrays [8] supply information about gene expression through measurements of mRNA levels of large amounts of genes in a cell, offering a wide picture of its functional status. The large amount of data produced by this powerful analytic technique can be processed through machine learning methods, using both unsupervised and supervised approaches. In a typical unsupervised approach, expression patterns of several hundreds or thousands of genes are obtained both from cancerous and non cancerous tissues; then clustering algorithms [4, 2] are used to group together similar expression patterns corresponding to different cells, in order to correctly separate cancerous from normal samples.

Anyway, unsupervised methods cannot always correctly separate classes, because they use unlabeled data to indirectly identify classes through clusters of gene expression data. Supervised methods can overcome this problem, exploiting "a priori" biological and medical knowledge on the problem domain, using labeled data to directly identify and separate classes. Recently, Support Vector Machines (SVM) and other supervised machine learning methods have been applied to the analysis of DNA microarray gene expression data in order to classify functional groups of genes and multiple tumor types [3, 5, 11].

In this paper we applied SVM and Multi-Layer Perceptrons (MLP) for tasks related to human lymphoma classification using DNA gene expression data. We tackled two types of classification problems related to the analysis of DNA microarray data: classification of cancerous and non-cancerous lymphoid tissues, and the identification of *Diffuse Large B-cell*

Lymphoma subgroups, exploiting "a priori" biological knowledge about sets of genes and information provided by clusters of coordinately expressed genes.

2 Classification of lymphoma using DNA microarray data

2.1 Data

We used data of a specialized DNA microarray, named "Lymphochip", developed at Stanford University School of Medicine [1]. Data used in our experimentation consist in 96 tissue samples from normal and malignant populations of human lymphocytes, considering for each sample 4026 different genes preferentially expressed in lymphoid cells or with known roles in processes important in immunology or cancer. Gene expression data are expressed as fluorescence ratios normalized subtracting for each value the median between all the values. Missing gene expression data (about 6% of all the data) have been replaced with zeros. We considered three main classes of lymphoma: *Diffuse Large B-cell Lymphoma* (DLBCL), *Follicular Lymphoma* (FL) and *Chronic Lymphocytic Leukemia* (CLL) together with *Transformed Cell Lines* (TCL) and normal lymphoid tissues [1]. The total number of cancerous samples is 72, while the number of different non-cancerous samples amounts to 24.

2.2 Classification tasks and methods

In our experimentation we faced two problems: (i) *Separating normal from cancerous lymphoid cells:* we tried to distinguish malignant from normal tissues using the overall information available, i.e. the expression data relative to all the 4026 genes; (ii) *Identifying and separating two different subclasses of lymphoma inside Diffuse Large B-cell Lymphoma using subsets of coordinately expressed genes:* in this second task we tried to validate the hypothesis of Alizadeh et al. [1] about the existence of two distinct functional types of lymphoma inside DLBCL. Two subgroups of DLBCL, that they named *Germinal Centre B-like DLBCL* (*GCB-like*) and *Activated B-like DLBCL* (*AB-like*) can be separated using hierarchical clustering algorithms. We tried to support Alizadeh's hypothesis using supervised methods to separate *GCB-like* from *AB-like* cells.

In order to solve the above classification problems, we applied three different types of SVMs, i.e. linear, polynomial and radial basis kernel functions, varying kernel and regularization parameters, and MLPs with one hidden layer. The generalization error of the learning machines had been evaluated through 10-fold cross validation techniques, and through the Joachims' estimator $\xi\alpha$ [7] of the leave-one-out error (for the SVMs only).

In all learning tasks we used *NEURObjects* [10], a set of C++ library classes for neural networks development, and *SVMlight* [6], a set of C applications implementing dichotomic SVM for classification tasks.

[1] The original "Lymphochip" DNA microarray data are available at `http://llmpp.nih.gov/lymphoma` and the data directly suitable for SVM and MLP analysis are available at `http://ftp.disi.unige.it/person/ValentiniG/Data/Lymphoma`.

Table 1: Classification of cancerous and non-cancerous lymphoid cells: generalization error, precision and sensitivity percent estimation through 10-fold cross validation. SVM-poly stands for polynomial SVM, SVM-RBF for radial basis (gaussian) SVM and LP for linear perceptron

Learning machine model	**Gen. error**	**St. dev.**	**Prec.**	**Sensitivity**
SVM-linear	1.04	3.16	98.63	100.0
SVM-poly	4.17	5.46	94.74	100.0
SVM-RBF	25.00	4.48	75.00	100.0
MLP	2.08	4.45	98.61	98.61
LP	9.38	10.24	95.65	91.66

3 Results

3.1 Classifying malignant and normal tissues

The results of the first classification task are shown in Tab. 1. SVM-linear achieved the best results, but MLP also showed an estimated generalization error of about 2% (using 10-fold cross validation). SVM showed also a very high estimation (100%) of the probability of detecting malignant lymphoid cells (sensitivity), no matter the type of kernel function used. Radial basis SVM showed an high misclassification rate, entirely due to the low precision of this type of SVM.

Receiver Operating Characteristic ROC analysis gives more insights into the behavior of the SVMs and MLPs used in this classification task. The ROC curve of the SVM-linear is almost ideal (Fig. 1 a), and the polynomial SVM also achieves a reasonably good ROC curve, lying just below the SVM linear ROC curve. The SVM-RBF registers the worst ROC curve, with values lying on the diagonal: the highest sensitivity is achieved only when it completely fails in correctly detecting normal cells. It is likely that the local nature of the radial basis

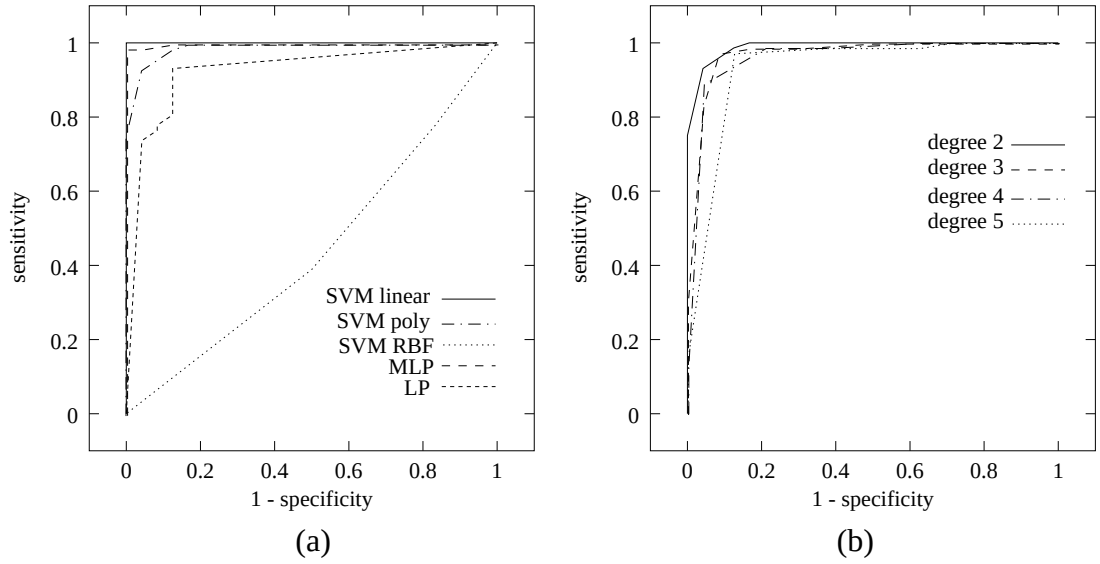

Figure 1: ROC curves for the classification problem of separating cancerous from normal tissues: (a) Comparison of ROC curves between SVM, LP and MLP; (b) Polynomial kernel SVM.

SVM in this case yields to overfitting (in all cases the accuracy on the training set is very high), considering that we have a small data set associated with a very high dimension of the input data. This type of SVM has an high estimated *Vapnik Chervonenkis* dimension, confirmed also by the fact that systematically all the input patterns are support vectors. The ROC analysis of the polynomial kernel SVM (Fig. 1 b) confirms that it is better to use not too complex SVM for this classification task. In this case the simplest (second degree) outperforms all other higher degree polynomial kernels.

Comparing our results with those obtained in [1] using hierarchical clustering, we achieved, as expected, a significant improvement of the classification accuracy, as supervised methods exploit 'a priori' biological knowledge (i.e. labeled data), while clustering methods use only unlabeled gene expression data to group together different tissues.

3.2 Identifying DLBCL subgroups

Using clustering methods, Alizadeh et al. [1] showed that two subgroups of molecularly distinct DLBCL lymphoma can be separated. Lossos [9] and Alizadeh [1] claimed that different subsets of genes could be responsible for the distinction of these two DLBCL subgroups. The expression signatures related to proliferation, T cell, lymphnode, and genes that distinguish germinal centre B-cells from other stages in B-cell ontogeny (GCB expression signature) showed differential gene expressions between these two subgroups.

In our experimentation we employed "a priori" biological knowledge about sets of genes and information provided by clusters of coordinately expressed genes (*expression signatures*) in order to verify if we can identify an expression signature related to the DLBCL partition proposed by Alizadeh. More precisely, we performed 5 classification tasks, using SVM and leave-one-out methods for estimating the generalization error relative to the separation of germinal centre B-cells and activated B-like cells. For each classification task we used a different expression signature from the four listed above, and all the 4 signatures together. The results are shown in Fig. 2. Only with the *GCB expression signature* we achieved quite

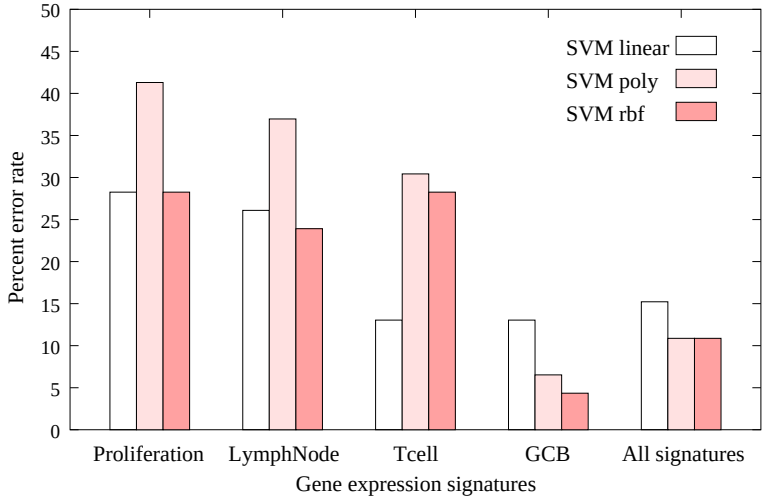

Figure 2: Estimated generalization error through leave-one-out techniques for the classification of GCB-like and AB-like subgroups of DLBCL using 4 different gene expression signatures and the 4 signatures all together. SVM poly stands for polynomial SVM and SVM-RBF stands for radial basis SVM.

good results, with an estimated generalization error (through leave-one-out techniques) of about 4% and an estimated precision of 100% (estimated sensitivity 91%) using SVM-RBF. With all the other signatures we obtained a relatively low accuracy.

The results show that the *GCB expression signature* is specifically related to the separation of *GCB-like* and *AB-like* subgroups of lymphoma inside the *DLBCL* group, supporting the hypothesis of Alizadeh about the existence of two distinct subgroups in *DLBCL*, and identify the GCB signature as a cluster of coordinately expressed genes related to the separation between the *GCB-like* and *AB-like DLBCL* subgroups.

4 Conclusion

We performed two classification tasks for the analysis of gene expression data related to diffuse large B-cell lymphoma. In the first task we showed that SVM and MLP can be successfully applied to the classification of cancerous and normal lymphoid tissues. In the second task we pointed out how to use ”a priori” biological and medical knowledge to separate two functional subclasses of DLBCL not detectable with traditional morphological classification of lymphoma, identifying a set of coordinately expressed genes related to the separation of the two DLBCL subgroups.

A planned development of this work consist in the integration of ”a priori” biological knowledge and supervised learning methods using a more structured approach based on hybrid neuro-fuzzy systems.

Acknowledgments

This work has been partially funded by INFM, Istituto Nazionale di Fisica della Materia, and University of Genova.

References

[1] A. Alizadeh et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403, 2000.

[2] A. Ben-Dor et al. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3), 1999.

[3] M. Brown et al. Knowledge-base analysis of microarray gene expression data by using Support Vector Machines. *PNAS*, 97(1), 2000.

[4] M.B. Eisen et al. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25), 1998.

[5] T.S. Furey et al. Support Vector Machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10), 2000.

[6] T. Joachims. Making large scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, MA, 1999.

[7] T. Joachims. Estimating the generalization performance of a SVM efficiently. In *ICML'2000*, 2000.

[8] D.J. Lockhart and E.A. Winzeler. Genomics, gene expression and DNA arrays. *Nature*, 405, 2000.

[9] I. Lossos et al. Ongoing immunoglobulin somatic mutation in germinal center B cell-like but not in activated B cell-like diffuse large B cell lymphomas. *PNAS*, 97(18), 2000.

[10] G. Valentini and F. Masulli. NEURObjects: an object-oriented library for neural network development. *Neurocomputing*, (to appear).

[11] C. Yeang et al. Molecular classification of multiple tumor types. In *ISMB 2001*, Copenaghen, 2001.