Ensembling and Clustering Approach to Gene Selection

Francesco Masulli [1;2] and Stefano Rovetta [1;3]

[1] INFM, National Institute for the Physics of Matter, Via Dodecaneso 33 - 16146 GENOVA – Italy
[2] Department of Computer Science, University of Pisa, Via F. Buonarroti 2 - 56127 PISA - Italy
masulli@di.unipi.it
[3] Department of Computer and Information Sciences, University of Genoa, Via Dodecaneso 35 - 16146 GENOVA - Italy
rovetta@disi.unige.it

Keywords. DNA microarray data, Genomics, Clustering, Ensembles, Gene Selection

Introduction

In pattern recognition the problem of input variable selection has been traditionally focused on technological issues, e.g., performance enhancement, lowering computational requirements, and reduction of data acquisition costs. However, in the last few years, it has found many applications in basic science as a model selection and discovery technique, as shown by a rich literature on this subject, witnessing the interest of the topic especially in the field of bioinformatics. A clear example arises from DNA microarray technology that provides high volumes of data for each single experiment, yielding measurements for hundreds of genes simultaneously.
In this paper, we propose a flexible method for analyzing the relevance of input variables in high dimensional problems with respect to a given dichotomic classication problem. Both linear and non-linear cases are considered. In the linear case, the application of derivative-based saliency yields a commonly adopted ranking criterion. In the non-linear case, the approach is extended by introducing a resampling technique and by clustering the obtained results for stability of the estimate. The method we propose (seeTab. 1) is termed *Random Voronoi Ensemble* since it is based on random Voronoi partitions [1], and these partitions are replicated by resampling, so the method actually uses an ensemble of random Voronoi partitions. Within each Voronoi region, a linear classification is performed using Support Vector Machines (SVM) with a linear kernel [4], while, to integrate the outcomes of the ensemble, we use the Graded Possibilistic Clustering technique to ensure an appropriate level of outlier insensitivity [3].

Table 1. Random Voronoi Ensemble method for feature selection

1. Establish a random Voronoi partitioning of the data space;
2. Discard homogeneous and empty Voronoi cells;
3. Compute a linear classifier on each remaining Voronoi cell;
4. Store the obtained saliency vector along with the cell site;
5. Repeat steps 1-4 until a sufficient number of saliency vectors are obtained;
6. Perform joint clustering of the saliency vectors and cell centers;
7. Retrieve cluster centers and use them as estimated local saliency rankings.

Experimental Results

The method was preliminarily validated on the data published in [2], a study, at the molecular level, of two kinds of leukemia, Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). Data were obtained by an Affymetrics high-density oligonucleotide microarray, revealing the expression level of 6817 human genes plus controls. Observations refer to 38 bone marrow samples, used as a training set, and 34 samples from different tissues (the test set).
In our experiment, we used only the training data to discriminate ALL from AML. Classes are in the proportion of 27 ALL and 11 AML observations. The results are summarized in Tab. 2, comparing the most important genes with those obtained by the original authors. Genes that were indicated both in [2] and by our technique are listed with the sign of their saliency value. Our technique indicates that, among the top 20 genes found by the final cluster analysis, 8 of the 50 genes listed in the original work feature a stronger discriminating power. We restrict the analysis to few genes, since a good cluster validation step is not included in the method yet. However, the results may indicate that not all of the genes found by Golub et al. contribute to the actual discrimination to the same extent.

Table 2. Relevant inputs for the Leukemia data

Gene description	**Gene accession number**	**Correlated class**	Sign of saliency
GPX1 Glutathione peroxidase 1	Y00787	AML	-
PRG1 Proteoglycan 1, secretory granule	X17042	AML	-
CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)	M27891	AML	-
Major histocompatibility complex enhancer-binding protein mad3	M69043	AML	-
Interleukin 8 (IL8) gene	M28130	AML	-
Azurocidin gene	M96326	AML	-
MB-1 gene	U05259	ALL	+
ADA Adenosine deaminase	M13792	ALL	+

References

[1] F.Aurenhammer, Voronoi diagrams-a survey of a fundamental geometric data structure, ACM Computing Surveys, 3: 345-405, 1991.

[2] T.R. Golub et al., Molecular classication of cancer: Class discovery and class prediction by gene expression monitoring, Science 5439:531-537, 1999.

[3] F. Masulli and S. Rovetta, The Graded Possibilistic Clustering Model, Proceedings of the International Joint Conference on Neural Networks, Portland, Oregon, IEEE Neural Network Society, Piscataway, NJ, USA, p.p. 791-796, 2003.

[4] V. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.