Gene selection using Random Voronoi Ensembles

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

[1] INFM-Istituto Nazionale per la Fisica della Materia
Via Dodecaneso 33, I-16146 Genova, Italy
[2] DI-Dipartimento di Informatica, Università di Pisa
Via F. Buonarroti 2, 56127 Pisa,Italy
masulli@di.unipi.it
[3] DISI-Dipartimento di Informatica e Scienze dell'Informazione
Università di Genova, Via Dodecaneso 35, 16146 Genova, Italy
rovetta@disi.unige.it

Abstract. 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 classification 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 method is extended by introducing a resampling technique and by clustering the obtained results for stability of the estimate. The method was preliminarly validated on the data published by T.R. Golub et al. on a study, at the molecular level, of two kinds of leukemia: Acute Myeloid Leukemia and Acute Lymphoblastic Leukemia (Science 5439-286, 531-537, 1999). 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.

1 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.

The problem statement is as follows. We are given a two-class labeled training sample $\left\{\mathbf{x} \in \Re^d\right\}$ of n observations. On the basis of the analysis of the decision surfaces, we want to assign an importance ranking to each individual input variable x_i with the aim of pointing out which input variables contribute most to the classification performance. This ranking can be used for the actual selection step.

2 Linear case

We assume that the normalization parameters for the data are known with sufficient statistical confidence. This is not always true, although in the case of microarray data accurate normalization is part of the standard preparation of data [3].

Let $r = g(\mathbf{x}) \in \Re$ be the discriminant or decision function, the discrimination criterion being $y = \text{sign}(r)$. We assume a classifier $r = g()$ capable of good generalization performance. We adopted Support Vector Machines [5], which provide optimal solutions with a minimum of parameter tuning.

To analyze what input variables have the largest influence over the output function, we evaluate the derivatives of r with respect to each variable, to point out which one is responsible, for a given perturbation, of the largest contribution to sign inversion (which denotes switching from one class to another). This is the so-called *derivative-based saliency*. It is a way to assess the sensitivity of the output to variations in individual inputs, and has been used in many contexts.

Since we are interested in zero crossings, the analysis should be done in a neighborhood of the locus $\{\mathbf{x}|g(\mathbf{x}) = 0\}$, and of course requires $g()$ to be locally differentiable. The latter assumption is reasonable (obviously, on a local basis) since smoothing is required by the discrete sampling of data. However, the more complex the decision surface $\{\mathbf{x}|g(\mathbf{x}) = 0\}$, the smaller the regions in which this assumption holds around any given point.

Standard input selection criteria [14] justify the application of the above technique to linear classifiers, although some small-sample issues, such as the previous consideration on normalization, are often overlooked. This technique is described for instance in [4] and [16]. In the linear case, $r = g(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$ and $\nabla r = \mathbf{w}$. The single feature r discriminates between the two classes ($r > 0$ and $r < 0$). This feature is given by a linear combination of inputs, with weights $\mathbf{w}$. Thus, by sorting the inputs according to their weights, the "importance" ranking is directly obtained. In the analysis, we examine relative importances, $\mathbf{t} = \mathbf{w} / \max_i \{w_i\}$ (w_i components of $\mathbf{w}$). The approach can be justified from many perspectives: statistical, geometrical, or in terms of classification margin.

3 Non-linear case

In the general, non-linear case, it is not possible to define a single ranking which holds in any region of the input space. A global approach employing statistical saliency evaluation based on data [13] requires large datasets which are not generally affordable, especially with DNA microarrays. Our approach involves partitioning the decision function $g()$, and performing local saliency estimates in sub-regions where $g()$ can be approximated with a linear decision function. To this end we apply a Voronoi tessellation [1], defined by drawing a set of points (termed Voronoi sites). Each Voronoi site defines a localized region in the data space, that is the locus of points in the data space for which that site is the nearest of all sites.

We can identify *empty regions* (with no data points); *homogeneous regions* (with points from one class only); *general regions* (with points from both classes).

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.
```

In the simplest approach, local linearization is made on the basis of an arbitrary partitioning (local subsampling) of the data space; to perform random partitioning, the Voronoi sites are drawn randomly. Homogeneous and empty regions are then discarded. Within each general region, a local linear classifier is built. Thus a single random Voronoi tessellation defines a set of classifiers, each performing a local analysis.

This basic method has several drawbacks: lower confidence of classifiers (trained on sub-samples); artifacts from Voronoi borders superposed to the separating surface; lack of a criterion for the number of regions; need to combine saliency rankings obtained in different regions.

To address all these issues, we propose a method we term "Random Voronoi Ensemble" since it is based on random Voronoi partitions as described above; these partitions are replicated by resampling, so the method actually uses an ensemble of random Voronoi partitions. Ensemble methods are described for instance in [6]. The method is outlined in Tab. 1.

Since a purely random partition is likely to generate many empty regions, the Voronoi sites are initialized by a rough vector quantization step, to ensure that sites are placed within the support of the data set. Subsequent random partitions are obtained by perturbation of the initial set of points. Within each Voronoi region, a linear classification is performed using Support Vector Machines (SVM) with a linear kernel.

To build a classifier ensemble, a resampling step is applied by replicating the basic procedure. The subsequent clustering step acts as the integrator, or arbiter: its role is to integrate the individual outcomes and to output a global response. It results in a set of "prototypical" saliency patterns, corresponding to different local classification criteria. These patterns are "prototypical" in the same sense as the centroids of k-means partitions [7] are representative of the respective clusters.

Resampling helps in smoothly covering the whole data set and, by averaging, contributes to the stability of the outcomes. Unfortunately, it is difficult to obtain theoretical guidelines on how many replications are required. Theoretical results on stability of Voronoi neighbors are available only for low dimensions [17], and typically cannot be generalized to higher dimensions.

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	+

To integrate the outcomes of the ensemble, we use the Graded Possibilistic Clustering technique to ensure an appropriate level of outlier insensitivity ([12]). This technique is a generalization of the Possibilistic approach to fuzzy c-Means clustering of Keller and Krishnapuram [10][11] in which cluster membership can be constrained to sum to 1 (as in the standard fuzzy clustering approaches [2]), unconstrained (as in the Possibilistic approach), or partially constrained. Partial constraints allow the implementation of several desirable properties, among which there is a user-selectable degree of outlier insensitivity. The number of cluster centers is assessed by applying a Deterministic Annealing schedule [15] to the parameter β, which directly influences the width of clusters and is a measure of the "resolution" of the method.

4 Experimental results

The method was preliminarily validated on the data published in [8], 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 this experiment, we used only the training data to discriminate ALL from AML. Classes are in the proportion of 27 ALL and 11 AML observations. Parameters: 4 Voronoi sites; β from 0.1 down to 0.01 in 10 steps, exponential decay law; uniform perturbation of maximum amplitude 0.5, independent on each input coordinate; 100 perturbations resulting in 400 random partitions of which 61% useful (general).

Results are summarized in Tab. 2, comparing the most important genes with those obtained by the original authors. Genes that were indicated both in [8] 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.

5 Discussion and conclusions

We have described a flexible method for analyzing the relevance of input variables in high dimensional problems. The method, which is in an early phase of development, has nevertheless shown the ability to tackle dichotomic problems even in the presence on non-linear separating surfaces. Its behavior has also been validated by comparing the results obtained on a real microarray data set with those published by the original authors.

We can underline some issues can we plan to addressed in the future work. For example, the number of Voronoi sites is an important parameter, since it is related to the scale of the tessellation (size of cells). Large cells will tend to contain segments of the separating surface which are difficult to linearize, while small cells will lead to excessively small data subset cardinality, and therefore to low generalization ability. The selection of the number of sites can be based on estimates of the problem complexity such as those proposed in [9], which are based on geometrical characterization of the data rather than the more usual statistical or information-theoretical consideration. However these must be combined with estimates of generalization to account for the trade-off outlined above.

Moreover, we have based our analysis on decision surfaces. This implies that the most natural setting of the problem is given by dichotomic (two-class) cases. Any polychotomic problem can be stated as a set of dichotomic problems, and this is what is usually done when using Support Vector Machines for classification. However a possible development of the method could imply the analysis of multi-class decision criteria, such as soft-max.

We point out that the proposed method for feature selection is especially well suited to parallel implementation at many levels, since the various steps can be pipelined, the subsamples can be processed in parallel, and the Voronoi resampling and clustering phases themselves can be implemented in parallel. All these steps involve very reduced communication. For instance, parallel resampling can be implemented by completely independent random partitions, and communication of subsamples for parallel analysis can be obtained by passing the index of selected patterns. Therefore a Beowulf-type workstation cluster may be proficiently used with limited adaptation effort.

The technique to generate the random perturbations themselves can be also optimized, to reduce the number of empty/homogeneous regions, since the data sets are expected to be extremely sparse in the data space. Perturbations can therefore be limited to a subspace, for instance by constraining them to the directions spanned by the versors of the data patterns (e.g., referring to the leukemia data, this is a basis which spans a 38-dimensional subspace of the 6817-dimensional data space).

6 Acknowledgements

Work funded by the Italian National Institute for the Physics of Matter (INFM) and by the Italian Ministry of Education, University and Research ("Prin2002").

References

[1] F. Aurenhammer, Voronoi diagrams-a survey of a fundamental geometric data structure, *ACM Computing Surveys*, 3 (23) (September 1991), 345-405.

[2] J.C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Plenum, New York (1981).

[3] M. Bilban, L.K. Buehler, S. Head, G. Desoye, V. Quaranta, Normalizing DNA microarray data, *Curr Issues Mol Biol*, 4 (2) (2002) 57-64.

[4] J. Brank, M. Grobelnik, N. Milic-Frayling, D. Mladenic, Feature selection using linear support vector machines, Tech. Rep. MSR-TR-2002-63, Microsoft Research (June 2002).

[5] N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge Univ. Press (2000).

[6] T.G. Dietterich, Machine-learning research: Four current directions *The AI Magazine* 4 (18) (Winter 1998) 97-136.

[7] R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, New York (USA) 1973.

[8] T.R. Golub *et al.*, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, Science 5439 (286) (1999) 531-537.

[9] Tin Kam Ho and Mitra Basu, "Complexity measures of supervised classification problems", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 289–300, March 2002.

[10] R. Krishnapuram, J.M. Keller, A possibilistic approach to clustering, *IEEE Trans. on Fuzzy Systems*, 2 (1) (May 1993) 98-110.

[11] R. Krishnapuram, J.M. Keller, The possibilistic *c*-Means algorithm: insights and recommendations, *IEEE Trans. on Fuzzy Systems*, 3 (4) (August 1996) 385-393.

[12] F. Masulli, S. Rovetta, Soft transition from probabilistic to possibilistic fuzzy clustering, DISI Technical Report DISI-TR-03-02, Department of Computer and Information Sciences, University of Genoa, Italy (April 2002). URL: http://www.disi.unige.it/person/RovettaS/research/techrep/DISI-TR-02-03.ps.gz.

[13] C. Moneta, G. Parodi, S. Rovetta, R. Zunino, Automated diagnosis and disease characterization using neural network analysis, in *Proc. of the 1992 IEEE Int. Conf. on Systems, Man and Cybernetics*, Chicago USA (October 1992) 123-128.

[14] B.D. Ripley, *Pattern recognition and neural networks*, Cambridge Univ. Press (1996).

[15] K. Rose, Deterministic annealing for clustering, compression, classification, regression, and related optimization problems, *Proceedings of IEEE*, 11 (86) (November 1998) 2210-2239.

[16] V. Sindhwani, P. Bhattacharya, S. Rakshit, Information theoretic feature crediting in multiclass support vector machines, in *1st SIAM Int. Conf. on Data Mining, Chicago, USA*. (April 2001) SIAM, Philadelphia.

[17] F. Weller, Stability of Voronoi neighborhood under perturbations of the sites, in *Proc. of Ninth Canadian Conf. on Computational Geometry*, Kingston, Ontario, Canada (August 1997).