

Random Voronoi Ensembles for gene selection

Francesco Masulli^{a,c} and Stefano Rovetta^{b,c}

^a *Dipartimento di Informatica, Università di Pisa, Italy*

^b *Dipartimento di Informatica e Scienze dell'Informazione, Università di Genova, Italy*

^c *Istituto Nazionale per la Fisica della Materia, Italy*

Abstract

The paper addresses the issue of assessing the importance of input variables 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.

Key words: Classifier combinations; Input selection; Quantization; DNA Microarrays

1 Introduction and problem statement

The problem of input variable selection, a central issue in pattern recognition, was traditionally focused on technological issues, e.g., performance enhancement, lowering computational requirements, and reduction of data acquisition costs. However, in relatively recent years, it has found many applications in basic science as a model selection and discovery technique.

There is a rich recent literature on this subject, witnessing the interest of the topic especially in the field of bioinformatics. A clear example arises from DNA microarray data. This technology 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 $\{\mathbf{x} \in \mathbb{R}^d\}$ of n observations. We want to assign an importance ranking to

Email addresses: masulli@di.unipi.it (Francesco Masulli), rovetta@disi.unige.it (Stefano Rovetta).

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.

We base 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.

2 Derivative-based ranking of input variables

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 [1].

Let $r = g(\mathbf{x}) \in \mathbb{R}$ be the discriminant or decision function, the discrimination criterion being $y = \text{sign}(r)$. We assume a classifier $r = g(\cdot)$ capable of good generalization performance. We adopted Support Vector Machines [2], 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 $g(\cdot)$ 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 [3] 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 [5]. 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.

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 [6] 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 [7], 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).

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. The proposed method addresses all these issues.

3 The Random Voronoi Ensemble method

We term our method “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 [8].

The method can be outlined as follows:

- (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

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 [9] 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 [10], and typically cannot be generalized to higher dimensions.

To integrate the outcomes of the ensemble, we use the Graded Possibilistic Clustering technique to ensure an appropriate level of outlier insensitivity. This technique is a generalization of the Possibilistic approach to fuzzy c -Means clustering of Keller and Krishnapuram [11][12] in which cluster membership can be constrained to sum to 1 (as in the standard fuzzy clustering approaches [13]), 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 [14] to the parameter β , which directly influences the width of clusters and is a measure of the “resolution” of the method (see [15] for details and the mathematical model).

4 Preliminary experimental results

The method was preliminarily validated on the data published in [16], 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).

Table 1
 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	+

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 Table 1, comparing the most important genes with those obtained by the original authors. Genes that were indicated both in [16] 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.

Acknowledgements

Work funded by the Italian National Institute for the Physics of Matter (INFN) and by the Italian Ministry of Education, University and Research (“Cofin2002”). We thank the anonymous reviewers for their constructive comments.

References

- [1] M. Bilban, L.K. Buehler, S. Head, G. Desoye, V. Quaranta, Normalizing DNA microarray data, *Curr Issues Mol Biol*, 4 (2) (2002) 57-64.
- [2] N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge Univ. Press (2000).
- [3] B.D. Ripley, *Pattern recognition and neural networks*, Cambridge Univ. Press (1996).
- [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] V. Sindhvani, 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.
- [6] 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.
- [7] F. Aurenhammer, Voronoi diagrams-a survey of a fundamental geometric data structure, *ACM Computing Surveys*, 3 (23) (September 1991), 345-405.
- [8] T.G. Dietterich, Machine-learning research: Four current directions *The AI Magazine* 4 (18) (Winter 1998) 97-136.
- [9] R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, New York (USA) 1973.
- [10] 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).
- [11] R. Krishnapuram, J.M. Keller, A possibilistic approach to clustering, *IEEE Trans. on Fuzzy Systems*, 2 (1) (May 1993) 98-110.
- [12] R. Krishnapuram, J.M. Keller, The possibilistic *c*-Means algorithm: insights and recommendations, *IEEE Trans. on Fuzzy Systems*, 3 (4) (August 1996) 385-393.
- [13] J.C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Plenum, New York (1981).
- [14] K. Rose, Deterministic annealing for clustering, compression, classification, regression, and related optimization problems, *Proceedings of IEEE*, 11 (86) (November 1998) 2210-2239.
- [15] 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>.
- [16] T.R. Golub *et al.*, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science* 5439 (286) (1999) 531-537.