The Department of Computer and Information Science
University of Genova, Via Dodecaneso 35 - 16146 Genova Italy
DISI Technical report no. DISI-TR-03-04

An ensemble approach to variable selection for classification of DNA microarray data

Francesco Masulli [1] Stefano Rovetta [2]

February 14, 2003

[1]Department of Computer Science, University of Pisa, Italy, and INFM, the National Institute for the Physics of Matter. E-mail: masulli@di.unipi.it

[2]Department of Computer and Information Sciences, University of Genoa, Italy, and INFM, the National Institute for the Physics of Matter. E-mail: rovetta@disi.unige.it

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. The work is preliminary, and many properties and options are to be investigated in future research.

1 Introduction and problem statement

We are given a labeled training sample $\mathbf{x} = X \subset \mathbb{R}^d$ of n observations. Labels define a dichotomy on X, i.e., the task to be learned is two-class classification. We refer to the problem of assigning an importance ranking to each individual input variable x_i with respect to the classification task, with the aim of pointing out which input variables contribute most to the classification performance.

This problem is properly called *input variable selection*, although it is commonly termed also "feature selection" or even "feature extraction" (which is, more correctly, the task of optimal pre-processing and combining the raw inputs into more significant composite variables).

Variable selection has always been a central problem in pattern recognition. The traditional emphasis has always been on technological issues (enhancing performance of automated recognition methods, lowering computational requirements, reducing the cost of data acquisition, e.g. [1]). However, in relatively recent years, the problem of assessing the relevance of variables has found many applications in basic science.

A clear example of this type of task arises from DNA microarray data. This technology provides high volumes of data for each single experiment, yielding measurements for hundreds of genes simultaneously. When inspecting for instance the outcome of a gene expression experiment to identify the "signature" corresponding to a given pathology, the procedure involves almost invariably the application of an automated classification method and the subsequent analysis of the results in seek of the most significant input variables. In this case, input variable selection is a tool supporting scientific inquiry.

The method we propose aims at pinpointing the variables which have the largest influence on the classification performance, also providing a relevance ranking. We are not necessarily interested in finding a good (or optimal) set of variables on which to build a better classifier. We address a so-called *wrapper* approach [2] to supervised variable selection. Wrapper techniques are those relying on the performance of a given learning machine (thus "wrapping" around the learning task). The alternative *filter* approach is based on extracting intrinsic knowledge from the data, by evaluation of some measure of influence of inputs over output such as mutual information [3] or simple correlation [4][5]. Finally, we focus on dichotomic (two-class) classification problems.

Given this problem setting, we are interested in obtaining an indication on the possible causes to be included in a more refined model. Therefore in a sense the "selection" phase itself is not even strictly necessary, and we focus on the phase of assessing "input saliency rankings".

The method has been designed for use in typical tasks of analysis of gene expression data (a well known instance of which is represented by [5]), and has been preliminarily validated on actual microarray data.

2 Derivative-based ranking of input variables

2.1 General approach

Let the input variables x_i be standardized, i.e., $E\{x_i\} = 0\ \forall i$ and $E\left\{x_i^2\right\} = 1\ \forall i$. These assumptions can be easily satisfied by pre-processing the input space based on the

training set, as in the standard practice. This is especially true of microarray data, where all measurements are made on the same scale and accurate normalization is viewed as a standard part of the preparation of data [6]. Inferring normalization parameters from data with sufficient statistical confidence is not so immediate in general cases where variables are not homogeneous in nature and scale.

Let $r = g(\mathbf{x})$ be the discriminant or decision function, defined on the d-dimensional input vector $\mathbf{x} \in \mathbb{R}^d$ and taking values in $\mathbb{R}$, the discrimination criterion being the value of $y = \text{sign}(r)$. We assume that a good classifier $r = g()$ is given. This is an important assumption. However current classification methods (support vector machines [7]) provide optimal solutions with a minimum of parameter tuning, so that, given a data set, a good classifier is readily obtained.

If we want to analyze what input variables have the largest influence over the output function, we should evaluate the derivatives of r with respect to each variable. This should be done in a neighborhood of the locus $\{\mathbf{x}|g(\mathbf{x}) = 0\}$, and of course requires $g()$ to be locally differentiable (which is a reasonable assumption since smoothing is required by the discrete sampling of data).

This is the so-called *derivative-based saliency*. It is a way to assess the sensitivity of the output to variations in individual inputs. This approach has been used in many contexts and has been experimentally shown to be quite efficient [8].

In the analysis, the following quantities are used:

- The (local) discriminant feature at data point $\overline{\mathbf{x}}$

$$\mathbf{w} = \nabla g(\mathbf{x})|_{\mathbf{x}=\overline{\mathbf{x}}} \tag{1}$$

- The saliency vector

$$\mathbf{t} = \frac{\mathbf{w}}{\max_i \{w_i\}}, \tag{2}$$

  where w_i are the individual components of vector $\mathbf{w}$

- The saliency rank vector

$$\mathbf{s} : s_i = \text{rank}\,(t_i, \mathbf{t}), \tag{3}$$

  where s_i and t_i are the individual components of vectors $\mathbf{s}$ and $\mathbf{t}$ respectively, and $\text{rank}\,(t_i)$ is the rank of component t_i among the set of component values of vector $\mathbf{t}$.

Given the ranking provided by $\nabla g()$, a variable selection procedure can then be based on a criterion similar to one of the following:

- Fix a number m of input variables and select the first m variables in the ordered list
- Fix a percentage of the total weights and select the inputs which account for that percentage
- Fix a maximum allowed increase in classification error and select the minimum number of variables in the ordered list (starting from the top) for which the error threshold is not exceeded.

The appropriate variable selection strategy depends on the availability of ad-hoc metrics for the applicative problem at hand and also on the problem perspective, since input space reduction aims at the minimum loss of information, while model selection aims at explaining in the clearest way the observed experiments. As a consequence, in the former case bounds on the error will be preferred, while in the latter case the constraint will rather be on the number of inputs.

2.2 The linear case

The popularity of linear classifiers is vast. Early work on classification [9][10] had concentrated on linear classifiers mainly due to computational constraints. In the recent past, linear classifiers have received renewed attention because of their relevance in kernel-based classifier theory and the support vector approach. This justifies the interest of the linear case by itself. Moreover, the linear case can be used to approach nonlinear situations as well, as explained in the following.

In the linear case, $g(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$ and

$$\nabla r = \left[\frac{\partial r}{\partial x_1}, \ldots, \frac{\partial r}{\partial x_d} \right] = \mathbf{w} \tag{4}$$

In this case, the derivative-based saliency measure can be justified in terms of "percentage of variance explained". The covariance of the input $\mathbf{x}$ has been assumed to be the unit matrix $\Sigma_{\mathbf{x}} = I$. The variance of the output r is therefore $\sigma_r^2 = \mathbf{w}^{\mathrm{T}} \Sigma_{\mathbf{x}} \mathbf{w} = ||\mathbf{w}||_2$. It is clear that, under the assumptions made, the input which gives the largest contribution to the variance of r is the one with the largest coefficient in the vector $\mathbf{w}$. (The assumption above, especially that $\Sigma_{\mathbf{x}} = I$, can be relaxed.)

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 relative weights $\mathbf{w}$. Thus, by sorting the inputs according to their weights, the "importance" ranking is directly obtained.

The mapping from the input space to the discriminant feature r is an orthogonal projection, therefore the selection of the best input variables by evaluation of output sensitivity yields also the projection with minimal error in terms of Euclidean distance (by Luenberger's projection theorem [11]). This justifies the derivative-based approach also from a vector approximation perspective.

2.3 The general nonlinear case

In the non-linear case, it is not possible to define a single clear ranking which holds in any region of the input space. A global approach can employ statistical evaluation of saliency based on data [8], but this requires large datasets which are not generally affordable, and especially so in the case of the DNA microarray methodology.

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. This local linearization is likely to introduce small errors, due to the local sparsity of data introduced by subsampling.

We can identify three kinds of region: *empty regions* contain no data points; *homogeneous regions* contain points from one class only; *mixed regions* contain points from both classes.

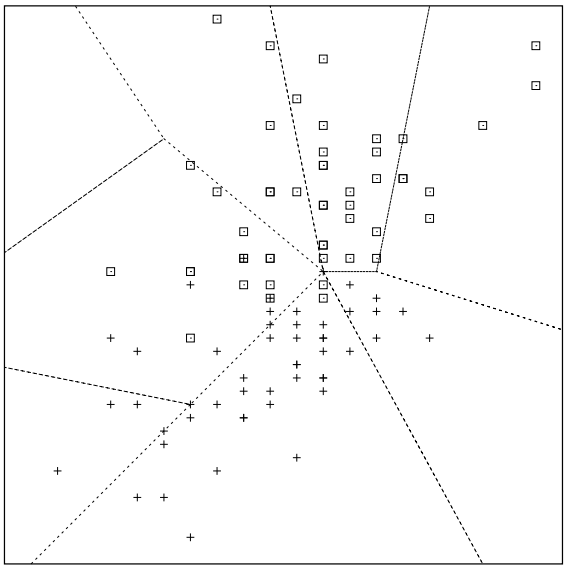

Figure 1: An arbitrarily partitioned dataset showing empty regions, regions with few samples, homogeneous neighboring regions of different classes.

In the simplest approach, local linearization is made on the basis of an arbitrarily selected partitioning of the data space. Homogeneous and empty regions are discarded. General regions, containing points from both classes, may be crossed by the true decision surface, and in any case a classifier can be built within them; thus they are retained for saliency analysis.

This basic method has several drawbacks:

- subsampling reduces the cardinality of data (sub)sets, lowering the confidence of classifiers induced on each localized region;
- if the correct decision surface lies between two different localized regions, each of which is homogeneous and has a different class, both regions are discarded and the analysis is distorted by this artifact
- the number of regions is to be selected a priori, but there is no clear way to decide it
- the saliency rankings obtained in one region may or may not be in agreement with those in neighboring or other regions, but in most cases they will agree only in part, and there is no way to decide whether several rankings should be combined or kept distinct.

The proposed method addresses all these issues, and will be presented in the remainder of the paper.

3 The Random Voronoi resampling method

3.1 Outline

We start with an exposition of the overall method; then the steps will be detailed in the following.

The method is summarized below:

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

3.2 Random Voronoi sampling

A Voronoi partition is induced by drawing a *Voronoi diagram* [12] in the data space. A Voronoi diagram is a tessellation defined by a set of reference points (*sites*); for each site, the corresponding *cell* is the locus of all points in the data space which are closer to that site than to any other site.

Voronoi tessellations are a very common tool in surface reconstruction for 3D graphics, and have also applications in the physics of matter. In particular, random Voronoi diagrams can efficiently model complex, collective properties of physical systems. Higher dimensional Voronoi tessellations are at the core of vector quantization methods.

A random Voronoi partition is obtained by throwing a set of random points in the data space. Since this is likely to generate many empty regions, the random diagram is 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.

3.3 Local linear classification

Within each Voronoi region, a linear classification is performed. There are many options for analyzing linear separability within a region. The state-of-the-art method is Support Vector Machines (SVM) [7] with a linear kernel. SVMs do not suffer from initialization and parameter sensitivity as other more traditional learning classifiers (e.g. perceptrons), and they provide a single parameter to be tuned for trading off strict separation with robust classification (and generalization).

Since the present approach is based on subsampling, the computational complexity of SVM training is small.

3.4 Saliency vectors

Saliency vectors, as computed in (2), are stored along with their respective sites. This retains the locality information associated with each saliency vector.

The whole set of saliency vectors stored during the iterations of the procedure are analyzed, at the end of the run, by applying a clustering step.

3.5 Building the ensemble: the resampling step

Resampling is one of the techniques used to obtain an *Ensemble method* [13]. Ensemble methods work by combining the outcome of many learning machines or many different instances of a learning machine (as in the present case). The subsequent clustering step acts as the integrator, or arbiter: its role is to integrate the individual outcomes and to output a global response.

In this research, we are interested in partitioning the data space and in obtaining localized "experts". One peculiarity of this approach is that the integrator may output a single response, but it may also output a set of combined responses, each specialized on a given region of the data space. The method can be thus viewed as a sort of "ensemble of ensembles", where the learning machine which is replicated by resampling is in turn a committee of local experts.

Resampling is the key step of the method. It ensures that the data set is smoothly covered and contributes to the stability of the outcomes, by averaging the strong statistical fluctuations. In the proposed approach, the random Voronoi subsampling is replicated by randomly perturbating the initial sites. In our experiments, we applied uniform perturbations with amplitude related to the pairwise distances between data points (e.g. by setting the amplitude equal to the maximum distance).

Unfortunately, it is difficult to obtain theoretical guidelines on how many replications are required as a function of the dimension of the data space and on how to compute the perturbations. This is because theoretical results on stability of Voronoi neighbors are available only for low dimensions [14], and typically rely on assumptions related to the dimension (so that they cannot be generalized to other dimensions).

3.6 Integration of the results: clustering saliency vectors

We use the Graded Possibilistic Clustering technique [15] 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 [16], in which cluster membership can be constrained to sum to 1 (as in the standard fuzzy clustering approaches), can be unconstrained (as in the Possibilistic approach), or can be 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 [17] to the resolution parameter β, which is used in the algorithm implementation presented in [15]. The number of clusters is selected to be an arbitrary and abundant quantity at the start of the procedure, when β equals a suitably chosen initial value $\beta^{(i)}$. Cluster centers collapse in early iterations, but with decreasing β they start to differentiate where required by the data distribution. The annealing can stop when

Table 1: Relevant inputs for the synthetic problem

Voronoi sites	Saliency vectors	Saliency rank vectors
1	0.91 1.00 0.89 0.79	2 1 3 4
2	0.58 1.00 0.46 0.41	2 1 3 4
	1.00 0.67 0.36 0.51	1 2 4 3
4	1.00 0.41 0.33 0.34	1 2 4 3
	0.30 1.00 0.27 0.27	2 1 3 4
	0.84 0.60 1.00 0.51	2 3 1 4
8	1.00 0.21 0.12 0.19	1 2 4 3
	0.64 1.00 0.25 0.19	2 1 3 4
16	1.00 0.57 0.31 0.33	1 2 4 3
	0.51 1.00 0.44 0.11	2 1 3 4
	0.91 0.88 0.13 1.00	2 3 4 1

β reaches a predefined final value $\beta^{(f)}$, chosen according to a reasonable criterion. For instance, $\beta^{(f)}$ may be comparable to the average pairwise distance between data points.

4 Preliminary experimental results

Since the method is in an early stage of development, many design decisions are still to be evaluated and experimental results are preliminary. In particular, comparative results are needed for proper assessment of the method efficacy, and the performance on actual biological data should be more thoroughly assessed.

4.1 Results on a synthetic dataset

Let's consider now an artificial dataset. The four-dimensional data (200 points) have been generated by a mixture of 3 two-dimensional gaussian clusters, one for the first class and the other two for the second class, at the vertices of a triangle. The separating surface between the points of the two classes was therefore approximately hyperbolic. The gaussian mixture data formed the first two components of the input space; the other were generated at random.

Table 1 reports the results for varying number of Voronoi sites. The true relevant components are 1 and 2. Note that in some cases there are clusters in which the values are all close to 1, and the corresponding ranking has no significance. These may be "lost" clusters from the clustering phase, due to a value of the resolution parameter β that is too small. However, in the majority of cases, only two clusters emerge, and they indicate correctly the two most significant directions for classification.

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	$+$

4.2 Results on a gene expression dataset

The method has undergone a preliminary validation by comparing its results on the data published by Golub et al. [5]. Data refer to the study, at the molecular level, of two kinds of leukemia, Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL) The data were obtained by DNA microarray experiments (high-density oligonucleotide microarray by Affymetrics) reporting on 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). The original experiments aimed at class discovery and prediction.

In this experiment, we used only the training data for the class discovery (also known as classification) task to discriminate ALL from AML. Classes are in the proportion of 27 ALL and 11 AML observations. The parameters used are as follows: number of sites $= 4$; β decreasing from $\beta^{(i)} = 0.1$ to $\beta^{(f)} = 0.01$ in 10 steps with exponential decay law; perturbation with uniform noise of maximum amplitude 0.5, independent on each input coordinate; 100 perturbations resulting in 400 random partitions of which 61% with mixed classes (the rest being either empty or homogeneous).

The results obtained are summarized in Table 2, which compares the most important genes with those obtained by the original authors. Genes that were indicated both in [5] and by our technique are listed with the sign of the corresponding saliency value. Our technique indicates that, among the top 20 genes found by the final analysis described in Subsection 3.6, 8 of the 50 genes listed in the original work feature the maximum discriminating power. We choose to 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, among the 50 most correlated genes found by Golub et al., not all contribute to the actual discrimination to the same extent. In fact, the large number of variables compared to the small number of observations calss for a careful statistical evaluation of the significance of the results obtained.

The ALL class was encoded with $+1$ and the AML class with -1; it is possible to notice that all genes whose expression was found to be correlated with ALL have positive saliency, while those correlated with AML have consistently a negative saliency

value. Absolute values of course are not reported since they are not of interest in the present context.

5 Discussion and open topics

There are a number of design options and theoretical topics that can be investigated. Some have been touched in the body of the paper; here we add some observations.

5.1 Choice of the scale

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 [18], 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.

5.2 Enhancements to the clustering step

To make the analysis more robust with respect to variations in the actual saliency values (**t**), it is possible to analyze the saliency rank values **s** instead. Clustering can therefore be made on the space of vectors **s**.

A given cluster can be analyzed by computing Kendall's rank concordance index W. [19]. Kendall's coefficient for N_c saliency rank vectors $\mathbf{s}^{(1)},\ldots,\mathbf{s}^{(N_c)}$ is computed as:

$$W = \frac{12\sum_{c=1}^{N_c}\left(\sum_{i=1}^{d} s_i^{(c)}\right)^2}{N_c(d^3-d)} - \frac{3(d+1)}{d-1} \tag{5}$$

and is compared to significance tables for W itself or for the related χ^2 statistics.

Clustering can also be modified to incorporate W in its cost function (W within clusters and $(1-W)$ between clusters) [20].

The experimental results indicate that a cluster validation criterion should be added to the clustering phase.

5.3 Enhancements to the algorithm

There is room for several kinds of optimizations. The technique 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 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 Conclusion

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 have also proposed several open design options and theoretical developments, which are the subject of current and future research.

Acknowledgments

This work was funded by the Italian National Institute for the Physics of Matter (INFM) and by the Italian Ministry of Education, University and Research under a "Cofin2002" grant.

References

[1] Sebastiano B. Serpico and Lorenzo Bruzzone, "A new search algorithm for feature selection in hyperspectral remote sensing images", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 7, pp. 1360–1367, July 2001.

[2] George H. John, Ron Kohavi, and Karl Pfleger, "Irrelevant features and the subset selection problem", in *Machine Learning, Proceedings of the Eleventh International Conference, Rutgers University, New Brunswick, NJ, USA*. July 1994, pp. 121–129, Morgan Kaufmann.

[3] Nojun Kwak and Chong-Ho Choi, "Input feature selection for classification problems", *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143–159, January 2002.

[4] G. Grant, E. Manduchi, and C. Stoeckert, "Using non-parametric methods in the context of multiple testing to identify differentially expressed genes", in *Methods of microarray data analysis*, S.M. Lin and K.F. Johnson, Eds., pp. 37–55. Kluwer Academic Publishers, Boston (USA), 2002.

[5] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", *Science*, vol. 286, no. 5439, pp. 531–537, October 1999.

[6] M. Bilban, L.K. Buehler, S. Head, G. Desoye, and V. Quaranta, "Normalizing DNA microarray data", *Curr Issues Mol Biol*, vol. 2, no. 4, pp. 57–64, April 2002.

[7] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines (and Other Kernel-based Learning Methods)*, Cambridge University Press, 2000.

[8] Carlo Moneta, Giancarlo Parodi, Stefano Rovetta, and Rodolfo Zunino, "Automated diagnosis and disease characterization using neural network analysis", in *Proceedings of the 1992 IEEE International Conference on Systems, Man and Cybernetics - Chicago, IL, USA*, October 1992, pp. 123–128.

[9] Ronald A. Fisher, "The use of multiple measurements in taxonomic problems", *Annual Eugenics*, vol. 7, part II, pp. 179–188, 1936.

[10] Frank Rosenblatt, *Principles of Neurodynamics*, Spartan, New York, 1962.

[11] David G. Luenberger, *Optimization by Vector Space Methods*, John Wiley and Sons, New York (USA), 1969.

[12] Franz Aurenhammer, "Voronoi diagrams-a survey of a fundamental geometric data structure", *ACM Computing Surveys*, vol. 23, no. 3, pp. 345–405, 1991.

[13] Thomas G. Dietterich, "Machine-learning research: Four current directions", *The AI Magazine*, vol. 18, no. 4, pp. 97–136, 1998.

[14] Frank Weller, "Stability of voronoi neighborhood under perturbations of the sites", in *Proceedings of Ninth Canadian Conference on Computational Geometry*, 1997.

[15] Francesco Masulli and Stefano Rovetta, "Soft transition from probabilistic to possibilistic fuzzy clustering", Tech. Rep. 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] Raghu Krishnapuram and James M. Keller, "A possibilistic approach to clustering", *IEEE Transactions on Fuzzy Systems*, vol. 1, no. 2, pp. 98–110, May 1993.

[17] Kenneth Rose, Eitan Gurewitz, and Geoffrey Fox, "A deterministic annealing approach to clustering", *Pattern Recognition Letters*, vol. 11, pp. 589–594, 1990.

[18] 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.

[19] Maurice Kendall and Jean Dickinson Gibbons, *Rank Correlation Methods*, Oxford University Press, Oxford (UK), fifth edition, 1990.

[20] R. Baumgartner, R. Somorjai, R. Summers, and W. Richter, "Assessment of cluster homogeneity in fMRI data using Kendall's coefficient of concordance", *Magnetic Resonance Imaging*, vol. 17, no. 10, pp. 1525–1532, 1999.