Measuring gene similarity by means of the classification distance

Under consideration for publication in Knowledge and Information Systems

Measuring gene similarity by means of the classification distance

Elena Baralis[1], Giulia Bruno[1] and Alessandro Fiori[1]

[1]Dipartimento di Automatica e Informatica, Politecnico di Torino, Torino, Italy

Abstract. Microarray technology provides a simple way for collecting huge amounts of data on the expression level of thousands of genes. Detecting similarities among genes is a fundamental task, both to discover previously unknown gene functions, and to focus the analysis on a limited set of genes rather than on thousands of genes. Similarity between genes is usually evaluated by analyzing their expression values. However, when additional information is available (e.g., clinical information) it may be beneficial to exploit it. In this paper, we present a new similarity measure for genes, based on their classification power, i.e., on their capability to separate samples belonging to different classes. Our method exploits a new gene representation which measures the classification power of each gene and defines the classification distance as the distance between gene classification powers. The classification distance measure has been integrated in a hierarchical clustering algorithm, but it may be adopted also by other clustering algorithms. The result of experiments run on different microarray datasets supports the intuition of the proposed approach.

Keywords: Similarity measure; Microarray; Clustering; Data mining

1. Introduction

Genome wide expression analysis with DNA microarray technology has become a fundamental tool in genomic research (El Akadi et al, 2010; Golub et al, 1999; Thompson et al, 2007; Jiang et al, 2004). An important goal of bioinformatics is the development of algorithms that can accurately analyze microarray data sets. Clustering algorithms are often used to detect functionally related genes by grouping together genes with similar patterns of expression (Datta and Datta, 2006). Many works consider the application or the adaptation of conventional clustering algorithms to gene expression data (see Jiang et al, 2004 and Thalamuthu et al, 2006 for a review) and new algorithms have recently been proposed (Bouguessa and Wang, 2009; Chu et al, 2010; Fu and Medico, 2007; Fu

Received xxx
Revised xxx
Accepted xxx

and Banerjee, 2008; Gu and Liu, 2008; Jiang et al, 2006; Wang et al, 2009). All clustering algorithms need to define the notion of similarity between elements.

Since microarray data are continuous values, several classical distance measures (such as Euclidean, Manhattan, Chebyshev, etc.) have been exploited to compute the distance between pairs of genes. However, such distance functions are not always adequate, because strong correlations may exist among genes even if they are far from each other as measured by these distance functions. The overall gene expression profile may be more interesting than the individual magnitude of each feature and traditional distance measures do not score well for shifting or scaled patterns (Zhao et al, 2006).

Other widely used schemes for determining the similarity between genes use the Pearson or Spearman correlation coefficients, which measure the similarity between two expression profiles. They have proved effective as similarity measures for gene expression data, but they are not robust with respect to outliers. Furthermore, they are a macroscopic metric and strong correlation may only exist on a subset of conditions (Zhao et al, 2006). The cosine correlation is more robust to outliers, because it computes the cosine of the angle between the expression gene value vectors. A comparison of several distance and correlation measures is provided in Zapala and Schork (2006).

Other kinds of similarity measures include pattern based (Wang et al, 2002) (which considers also simple linear transformation relationships) or tendency based (Liu and Wang, 2003) (which considers synchronous rise and fall of expression levels in a subset of conditions). In Zhao et al (2006) the authors focus on the problem of grouping also negative co-regulation patterns, while in Mitra and Majumder (2004) a maximal information compression index is used to measure dissimilarity between the expression levels of genes.

The common characteristics of these approaches is that they cluster genes only by analyzing their continuous expression values. These approaches are appropriate when there is no information about sample classes and the aim of clustering is to identify a small number of similar expression patterns among samples. However, when additional information is available (e.g., biological knowledge or clinical information), it may be beneficial to exploit it to improve cluster quality (Huang and Pan, 2006).

In this work, we address the problem of measuring gene similarity by combining the gene expression values and the sample class information. To this aim, we define the concept of *classification power* of a gene, that specifies which samples are correctly classified by a gene. A gene classifies correctly a sample if, by considering the sample expression level, it assigns the sample unambiguously to the correct class. Thus, instead of discovering genes with similar expression profiles, we identify genes which play an equivalent role for the classification task (i.e., genes that give a similar contribution for sample classification). Two genes are considered equivalent if they classify correctly the same samples. The classification power of a gene is represented by a string of 0 and 1, that denotes which samples are correctly classified. This string is named *gene mask*.

To measure gene similarity, we define a novel distance measure between genes, the *classification distance*, which computes the distance between gene masks. The classification distance has been integrated in a hierarchical clustering algorithm, which iteratively groups genes or gene clusters through a bottom up strategy (Everitt et al, 2009). To allow the computation of inter-cluster distance by means of the classification distance, the concept of *cluster mask* (i.e., the total classification power of genes in a cluster) was also defined. Besides hierarchical clustering, the classification distance measure may be integrated in clustering algorithms based on different approaches (e.g., DBSCAN Ester et al, 1996, or PAM Kaufman and Rousseeuw, 2005).

To our knowledge, there are no works which address the issue of measuring the similarity between genes by considering both their expression values and the informa-

tion about each sample class. Some works address the complementary problem, i.e., grouping samples by analyzing their gene expression values (Bushel et al, 2007; Song et al, 2008), or combining clinical and microarray data to build a model for tumor classification (Gevaert et al, 2006). Differently from sample clustering, gene clustering does not provide an easy validation procedure, because the gene class labels are unknown, and clustering accuracy cannot be computed by counting the genes correctly assigned to each cluster.

Since gene expression data is typically affected by outliers, we also introduce a new density based approach to reduce the influence of values far from the concentration core (i.e., outlier values). A popular procedure specifically used in microarray data analysis (Yang et al, 2002) for removing outliers is the Hampel identifier (Davies and Gather, 1993), also called the median absolute deviation (MAD) method. The MAD estimator smooths the effect of values far from the median value, independently of their density.

To take into account also the density distribution of values, we propose the *weighted mean deviation* (or WMD) method to reduce the influence of outliers in the definition of the gene expression intervals. In particular, mean and standard deviation are replaced by their weighted versions. A weight is assigned to each data value by considering the number of its neighbors belonging to the same class. Thus, a higher weight is assigned to values with many neighbors and a lower weight to isolated values.

We validated our method on different microarray datasets by comparing our distance measure with the widely used Euclidean distance, Pearson correlation and cosine distance measures. The experimental results confirm the intuition of the proposed approach and show the effectiveness of our distance measure in clustering genes with similar classification behavior.

The paper is organized as follows. Section 2 describes the steps to compute the classification distance between gene (or cluster) masks. Section 3 presents the integration of our distance measure in a hierarchical clustering approach. Section 4 discusses the experimental evaluation of the proposed approach and finally Section 5 draws conclusions and presents future works.

2. Measuring gene similarity

When all the samples whose gene expression value is in a given range belong to a single class, the gene can assign unambiguously these samples to the correct class. We propose a method to define the similarity between genes by measuring their classification power (i.e., their capability to correctly classify samples), which performs the following steps.

- **Core expression interval definition.** Definition of the range of expression values for a given gene in a given class. To address the problem of outliers, a density based weight is exploited in the core expression interval definition.
- **Gene mask and cluster mask generation.** Definition of the *gene mask* and the *cluster mask* as representatives of gene and cluster classification power. The gene mask is generated by analyzing the gene core expression intervals, while the cluster mask is generated by analyzing the gene masks of genes in the cluster.
- **Classification distance computation.** Definition of the *classification distance* measure to evaluate the dissimilarity between the classification power of genes (or clusters). The Hamming distance is exploited to measure the distance between masks.

These steps are described in details in the following subsections.

In general, microarray data E are represented in the form of a gene expression matrix, in which each row represents a gene and each column represents a sample. For each sample, the expression level of all the genes under consideration is measured. Element e_{is} in E is the measurement of the expression level of gene i for sample s, where $i = 1, \ldots, N$ and $s = 1, ..., S$. Each sample is also characterized by a class label, representing the clinical situation of the patient or tissue being analyzed. The domain of class labels is characterized by C different values and label k_s of sample s takes a single value in this domain.

2.1. Core expression interval definition

The core expression interval of a gene in a class represents the range of gene expression values taken by samples of the considered class. Since microarray data may be noisy, we propose a density based approach to reduce the effect of outliers on the core expression interval definition, the *Weighted Mean Deviation* (or WMD). WMD is a variation of the MAD estimator (Hampel, 1974; Daszykowski et al, 2007). The MAD estimator first computes the median of the data and defines the set of absolute values of differences between each data value and the median. Then, the median of this set is computed. By multiplying this value by 1.4826 (i.e., the scale factor for normally distributed data), the MAD unbiased estimate of the standard deviation for Gaussian data is obtained. The MAD estimator smooths the effect of values far from the median value, independently of their density. In WMD the mean is replaced by the weighted mean and the standard deviation by the weighted standard deviation. The weights are computed by means of a density estimation. A higher weight is assigned to expression values with many neighbors belonging to the same class and a lower weight to isolated values. A comparison between WMD and MAD is presented in Section 4.2.

Consider an arbitrary sample s belonging to class k and its expression value e_{is} for an arbitrary gene i. Let the expression values be independent and identically distributed (i.i.d) random variables and $\sigma_{i,k}$ be the standard deviation for the expression values of gene i in class k. The density weight w_{is} measures, for a given expression value e_{is}, the number of expression values of samples of the same class which belong to the interval $\pm\sigma_{i,k}$ centered in e_{is}.

The density weight for the expression value e_{is} for a gene i and a sample s belonging to class k is defined as

$$w_{is} = \sum_{m=1, m \neq s}^{S} \delta_{im} \tag{1}$$

where δ_{im} is a function defined as

$$\delta_{im} = \begin{cases} 1 & \text{if sample } m \text{ belongs to class } k \wedge \\ & e_{im} \in [e_{is} - \sigma_{i,k}; e_{is} + \sigma_{i,k}] \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

If an expression value is characterized by many neighboring values belonging to the same class, its density weight is higher. For example, in Figure 1 the expression values of an arbitrary gene i with four samples of class 1 (labeled as *w*, *x*, *y*, and *z*) and seven of class 2 (labeled as *a*, *b*, *c*, *d*, *e*, *f*, and *g*) are shown. For sample *a*, the expression level (denoted as e_{ia} in Figure 1) is characterized by a density weight w_{ia}

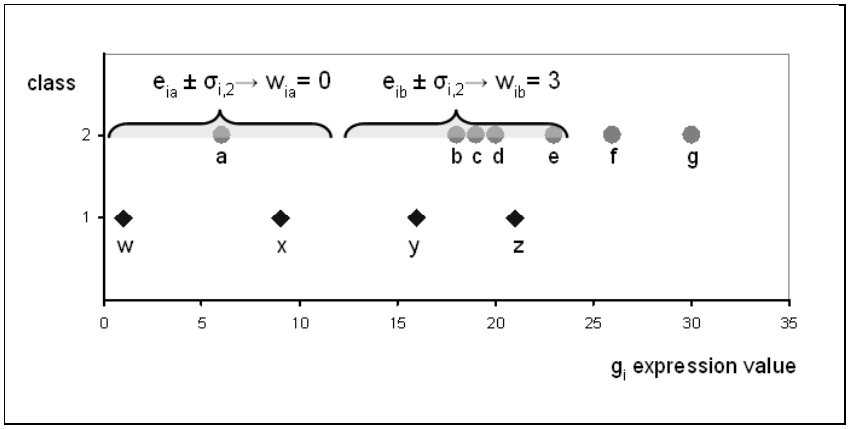

Fig. 1. Gene *i*: Density weight computation for samples *a* and *b*.

equal to 0, because for gene i there are no other expression values of class 2 in the interval $e_{ia} \pm \sigma_{i,2}$ (represented by a curly bracket). For sample *b*, the expression value (e_{ib}) is characterized instead by a density weight w_{ib} equal to 3, because three other samples of class 2 belong to the interval $e_{ib} \pm \sigma_{i,2}$.

The core expression interval of an arbitrary gene i in class k is given by

$$I_{i,k} = \hat{\mu}_{i,k} \pm (2 \cdot \hat{\sigma}_{i,k}) \tag{3}$$

where the weighted mean $\hat{\mu}_{i,k}$ and the weighted standard deviation $\hat{\sigma}_{i,k}$ are based on the density weights and are computed as follows[1].

The weighted mean $\hat{\mu}_{i,k}$ is defined as

$$\hat{\mu}_{i,k} = \frac{1}{W_{i,k}} \sum_{s=1}^{S} \delta_{is} \cdot w_{is} \cdot e_{is} \tag{4}$$

where δ_{is} is a function defined as

$$\delta_{is} = \begin{cases} 1 & \text{if sample } s \text{ belongs to class } k \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

and $W_{i,k}$ is the sum of density weights for gene i in class k (i.e., $\sum_{s=1}^{S} \delta_{is} \cdot w_{is}$).

The weighted standard deviation $\hat{\sigma}_{i,k}$ is given by

$$\hat{\sigma}_{i,k} = \sqrt{\frac{1}{W_{i,k}} \sum_{s=1}^{S} \delta_{is} \cdot w_{is} \cdot (e_{is} - \hat{\mu}_{i,k})^2} \tag{6}$$

In the upper part of Figure 2, an example of the core expression intervals for a gene

[1] The term $2 * \hat{\sigma}_{i,k}$ covers about 95% of expression values. Higher (or lower) values of the weighted standard deviation multiplicative factor may increase (or decrease) the number of included values.

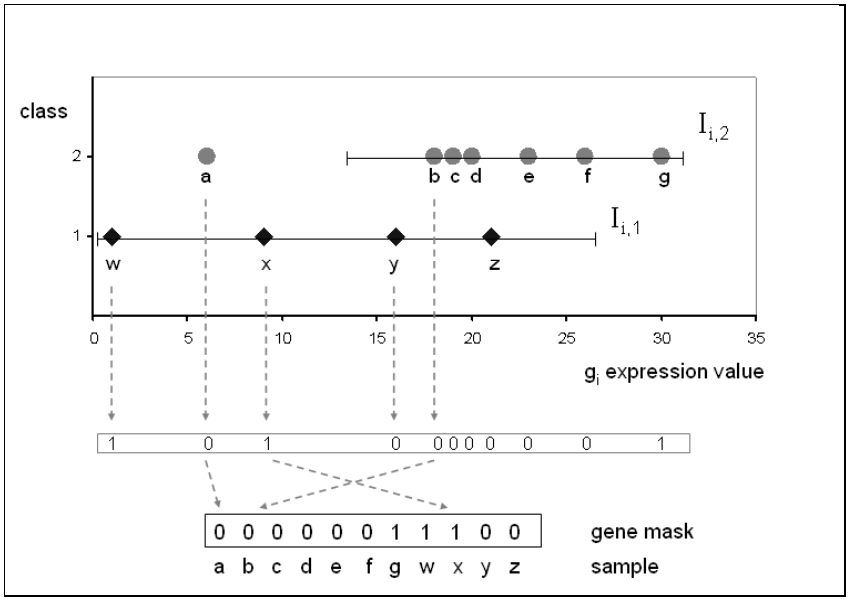

Fig. 2. Core expression interval computation for classes 1 and 2 and gene mask computation for gene g_i.

with samples belonging to two classes is shown. Since the first sample of class 2 (i.e., sample *a*) has a low density weight (equal to zero), its value provides no contribution to the weighted mean and standard deviation computation. Thus, the class 2 core expression interval is less affected by outliers

2.2. Gene mask and cluster mask generation

For each gene we define a gene mask, which is an array of S bits, where S is the number of samples. It represents the capability of the gene to classify correctly each sample, i.e., its classification power. Consider an arbitrary gene i and two arbitrary classes $c_1, c_2 \in \{1, \ldots, C\}$. Bit s of its mask is set to 1 if the corresponding expression value e_{is} belongs only to the core expression interval of a single class (e.g., I_{i,c_1}) and does not belong to the core expression interval of any other class (e.g., I_{i,c_2} with $c_1 \neq c_2$). Otherwise it is set to 0. Formally, bit s of the gene mask is computed as follows.

$$mask_{is} = \begin{cases} 1 & \text{if } (e_{is} \in I_{i,c_1}) \wedge \nexists c_2 \neq c_1 \mid e_{is} \in I_{i,c_2} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

A sample might not belong to any core expression interval (i.e., it is an outlier). In this case, the value of the corresponding bit is set to 0 according to (7).

Figure 2 shows the gene mask associated to an arbitrary gene i after the computation of its core expression intervals $I_{i,1}$ and $I_{i,2}$. The samples *g*, *w*, and *x* belong to the expression interval of a single class, thus their corresponding mask bits are set to 1. The bits corresponding to the other samples are set to 0.

The notion of classification power may be extended to clusters of genes. Given an arbitrary gene cluster, its *cluster mask* is the logical OR between the masks of the genes

in the cluster. It represents the total classification power of the cluster, i.e., the samples that can be correctly classified by considering all the genes in the cluster.

2.3. Classification distance computation

The classification distance measure captures the dissimilarity between genes (or clusters) by analyzing their masks. It evaluates the classification power of each object, represented by its mask, and allows the identification of objects which provide similar information for classification.

Given a pair of objects (i, j), the classification distance between them is defined as follows

$$d_{ij} = \frac{1}{S} \sum_{s=1}^{S} mask_{is} \oplus mask_{js} \tag{8}$$

where S is the number of samples (bits) of the mask, $mask_{is}$ is bit s of mask i, and $\oplus$ is the EX-OR operator which yields 1 if and only if the two operands are different. Hence, the classification distance is given by the Hamming distance between masks.

When two genes (or clusters) classify in the same way the same samples, their distance is equal to 0 because their masks are identical. On the other extreme, if two objects have complementary masks, their distance d_{ij} is maximum and equal to 1, because the sum of complementary bits is equal to the number of samples S.

The classification distance is a symmetric measure that assesses gene similarity by considering both correct and uncertain classification of samples. We also considered, as an alternative, an asymmetric distance measure similar to the Jaccard coefficient (Cox and Cox, 2001). This asymmetric measure considered the contribution of correctly classified samples (i.e., both 1 in the mask) and disregarded the contribution of samples for which classification is uncertain, due to interval overlap (i.e., both 0 in the mask). An experimental evaluation (not reported in the paper) of this alternative showed a worse performance, thus highlighting that also the similarity for uncertain classifications is important to group genes with similar behavior.

3. Integration in clustering algorithms

The classification distance measure may be integrated in various clustering approaches. To validate its effectiveness, we integrated it into a hierarchical clustering algorithm (Everitt et al, 2009). Agglomerative hierarchical clustering iteratively analyzes and updates a distance matrix to group genes or gene clusters through a bottom up strategy.

Consider an arbitrary set G of N genes. The triangular distance matrix D can be computed on G by means of the classification distance measure defined in (9). An arbitrary element d_{ij} in D represents the distance between two objects i and j, which may be either genes or gene clusters. Matrix D is iteratively updated each time a new cluster is created by merging genes or gene clusters. The process is repeated $N - 1$ times, until only one single element remains.

At each iteration, the two objects to be merged are selected by identifying in D the element with the lowest value d_{ij}, which represents the most similar pair of objects (genes or clusters) i and j. If more object pairs are characterized by the same minimum distance, the element with the maximum average variance is selected, because variance

Table 1. Dataset characteristics: name, number of samples, number of genes, and number of classes

Dataset	*Samples*	*Genes*	*Classes*
Tumor9	60	5726	9
Brain1	90	5920	5
Lung	203	12600	5
Leuk1	72	5327	3
Leuk2	72	11225	3
Colon	62	2000	2
Prostate	102	10509	2
SRBCT	83	2308	2
DLBCL	77	5469	2

is the simplest unsupervised evaluation method for gene ranking (He et al, 2006). In particular, genes with high variance are usually ranked higher because their expression values significantly change over conditions (He et al, 2006). Average variance of an element is given by the average over the variance of the expression levels of all genes belonging to the two objects i and j concurring to the new (cluster) element.

The classification distance measure may be integrated in other clustering approaches. For example, density-based clustering methods, such as DBSCAN (Ester et al, 1996), consider the Euclidean distance among elements to compute the reachability relationship needed to define each element neighborhood. The proposed distance measure may replace the Euclidean distance, while ϵ may be defined in terms of the maximum number of mismatching bits between the two masks (i.e., the maximum number of bits set to 1 after the EX-OR computation). Similar considerations hold for partition-based clustering algorithms (e.g., PAM (Kaufman and Rousseeuw, 2005)).

4. Experimental results

We validated our method on 9 microarray datasets, publicly available on (Statnikov et al, 2005) and (Alon et al, 1993). Table 1 summarizes their characteristics. The data distribution and cardinality of these datasets are rather diverse and allowed us to validate our approach under different experimental conditions.

We performed a set of experiments addressing the following issues.

- **Classification distance evaluation.** To evaluate the effectiveness of the classification distance in measuring the classification power of genes we compared the accuracy and the sensitivity provided by neighboring genes. Furthermore, the biological relevance of our results has been assessed by verifying if neighboring genes are reported with similar biological meaning in tumor literature.
- **Core expression interval comparison.** The Weighted Mean Deviation (WMD) and the Hampel identifier (MAD) for detecting the core expression intervals have been compared in terms of both accuracy and interval characteristics.
- **Cluster characterization.** The characteristics of the clusters yielded by hierarchical clustering exploiting the classification distance have been investigated.

4.1. Classification distance evaluation

Accuracy and sensitivity.

Accuracy is defined as the number of samples correctly associated to their class over the total number of samples. It provides an overall classification performance measure. We also analyzed the classification performance separately for each class by computing, for each class, the true positive rate (i.e., the rate of correctly assigned samples over the total number of samples belonging to the class). The true positive rate is also called sensitivity or recall.

In the context of tumor classification, to which the datasets in Table 1 are devoted, the most interesting genes are those which play a role in the disease. We focused our analysis on these genes, which are commonly selected by means of feature selection techniques (Mukkamala et al, 2006). In our experiments, we computed the accuracy provided by the set of top ranked genes selected by means of a supervised feature selection technique. Then, we substituted in turn a single gene with the most similar gene according to various distance metrics. We computed the new accuracies and we compared the obtained results to the previous accuracy value.

In particular, to avoid biasing our analysis by considering a single feature selection technique, we performed supervised feature selection by means of the following popular techniques (Statnikov et al, 2005): (i) Analysis of variance (ANOVA), (ii) signal to noise ratio in one-versus-one fashion (OVO), (iii) signal to noise ratio in one-versus-rest fashion (OVR), (iv) ratio of variables between categories to within categories sum of squares (BW). New feature selection techniques have been recently developed (Liu and Motoda, 2007), but since the selection of a feature selection algorithm is not very critical and it is done only to avoid biasing the analysis by using only one of them, we limit the analysis to these four methods. Feature selection has been performed separately for each dataset. We considered the first ten genes ranked by each feature selection technique. These small gene subsets only contain genes which are relevant for discriminating among sample classes.

In each of the 10-gene sets obtained from feature selection, we substituted in turn a single gene with the most similar gene according to a distance measure. In particular, we considered the Euclidean distance, the Pearson correlation, the cosine correlation, and the classification distance. Thus, for each 10-gene set and for each distance measure, we created ten new different gene sets, each of which with one substituted gene. The accuracy and the sensitivity provided by these new sets have finally been computed and compared.

Classification has been performed by means of the LibSVM classifier (Chang and Lin, 2001), with parameters optimized by using the grid search in the scripts downloaded with the LibSVM package. Ten fold cross-validation has been exploited to avoid selection bias. The reported accuracy is the overall value computed on all the splits. The considered feature selection methods are available in the GEMS software (Statnikov et al, 2005).

Table 2 shows the accuracy results of the experiments on the Brain1 dataset. Similar results hold for the other datasets. The accuracy of the original setting (i.e., the ten original genes selected by the feature selection methods) is reported in the first column. For each feature selection method, rows labeled 1-10 report the accuracy difference between the original set and each of the modified sets (each one with a different substituted gene), while the last two rows report the average value over the 10 modified settings and the standard deviation. For three out of four feature selection methods the classification distance selects the best substituted gene with respect to the other distance measures. In the case of OVO and ANOVA, the substitution even improves accuracy with respect to the original setting (i.e., it selects a better gene with respect to that selected by the supervised feature selection method).

The different overall accuracy increase/decrease depends on the intrinsic nature of

Table 2. Differences between the accuracy of the original subset and the modified ones on the Brain1 dataset for different feature selection methods and distance measures

Method	*Gene*	*Euclidean*	*Pearson*	*Cosine*	*Classification*
ANOVA 81.11	1	-1.11	0.00	1.11	-2.22
	2	2.22	1.11	-1.11	4.44
	3	2.22	-1.11	-2.22	-1.11
	4	3.33	2.22	3.33	2.22
	5	-2.22	-3.33	-2.22	1.11
	6	-1.11	2.22	-1.11	1.11
	7	2.22	1.11	1.11	3.33
	8	-1.11	0.00	1.11	1.11
	9	-2.22	-3.33	-3.33	-2.22
	10	1.11	-2.22	-1.11	-2.22
	Mean	0.33	-0.33	-0.44	**0.56**
	Std	2.10	2.04	1.34	2.41
BW 74.45	1	2.22	-8.89	-3.33	-1.11
	2	-2.22	-3.33	-3.33	-1.11
	3	-4.44	-3.33	-1.11	-5.56
	4	7.78	-4.45	0.00	-1.11
	5	-2.22	-5.56	-3.33	-3.33
	6	-4.44	-6.67	-4.44	-5.56
	7	-5.56	-5.56	-3.33	-4.45
	8	-5.56	-5.56	-3.33	-1.11
	9	-3.33	-3.33	-3.33	-2.22
	10	-2.22	-7.78	-5.56	-3.33
	Mean	-3.56	-5.44	-3.11	**-2.89**
	Std	2.71	1.55	2.20	1.83
OVO 74.45	1	2.22	2.22	1.11	0.00
	2	0.00	-1.11	0.00	3.33
	3	3.33	5.56	6.67	2.22
	4	-4.45	5.55	4.44	5.56
	5	3.33	1.11	0.00	3.33
	6	-1.11	1.11	1.11	1.11
	7	1.11	0.00	1.11	0.00
	8	3.33	2.22	2.22	-1.11
	9	-2.22	-1.11	-1.11	-3.33
	10	2.22	2.22	3.33	5.56
	Mean	0.78	1.78	**1.89**	1.67
	Std	2.67	2.35	2.69	2.88
OVR 73.34	1	-6.67	-6.67	-7.78	-4.44
	2	-10.00	-6.67	-7.78	-5.56
	3	-5.56	-3.33	-5.56	0.00
	4	-3.33	-4.45	-2.22	-3.33
	5	-3.33	-4.45	-4.45	-2.22
	6	-5.56	-3.33	0.00	-4.45
	7	-1.11	1.11	1.11	0.00
	8	-7.78	-4.45	-3.33	-2.22
	9	-5.56	-2.22	-5.56	-2.22
	10	-1.11	-5.56	-5.56	-8.89
	Mean	-5.00	-4.00	-4.11	**-3.33**
	Std	2.83	3.01	2.29	2.67

Table 3. Average sensitivity (i.e., true positive rate) in percentage over the ten substitutions for each class (1 to 5) and the total accuracy (row All) on the Brain1 dataset for different feature selection methods and distance measures

Method	*Class*	*Euclidean*	*Pearson*	*Cosine*	*Classification*
ANOVA	1	94.17	94.50	94.00	*94.83*
	2	72.00	73.00	*74.00*	73.00
	3	*64.00*	58.00	59.00	61.00
	4	65.00	62.50	60.00	*67.50*
	5	*8.33*	6.67	6.67	*8.33*
	All	81.44	80.78	80.67	**81.67**
BW	1	*93.17*	91.17	91.00	91.67
	2	30.00	26.00	27.00	*33.00*
	3	21.00	19.00	19.00	*31.00*
	4	67.50	*70.00*	*70.00*	67.50
	5	1.67	1.67	1.67	*5.00*
	All	70.89	69.01	71.34	**71.56**
OVO	1	95.33	96.83	*97.00*	95.67
	2	57.00	56.00	58.00	*59.00*
	3	11.00	12.00	12.00	*16.00*
	4	92.50	92.50	*95.00*	90.00
	5	*0.00*	*0.00*	*0.00*	*0.00*
	All	75.23	76.23	**76.34**	76.12
OVR	1	88.67	89.50	*90.17*	89.67
	2	46.00	*50.00*	45.00	47.00
	3	8.00	8.00	9.00	*14.00*
	4	*72.50*	67.50	70.00	*72.50*
	5	0.00	*3.33*	1.67	*3.33*
	All	68.34	69.34	69.23	**70.01**

each feature selection method. For the ANOVA and OVO methods, the original gene masks are characterized by more bits set to 1 (on average 20 over 90 samples) than the other two methods (on average 8). The highly selective genes (i.e., with few 1 in their mask) chosen by BW and OVR may be more difficult to replace appropriately. In this context, the classification distance selects a gene with a classification behavior more similar to the gene to be substituted than the other distance measures. Finally note that highly selective genes do not necessarily imply high accuracy.

Table 3 provides details on the percentage of correctly classified samples for each class (1 to 5) in the Brain 1 dataset. The average sensitivity (i.e., true positive rate) in percentage over the ten substitutions for each class and the total accuracy (row All) for different feature selection methods and distance measures is reported. The cardinality of the classes are 60, 10, 10, 4, and 6 samples respectively. The sensitivity of the classification distance is typically higher than the sensitivity provided by the other distances. In particular, the classification distance provides the best sensitivity for at least three classes for all feature selection methods. Furthermore, the highest sensitivity usually characterizes the classes with low cardinality. Thus, our method is particularly suited to rare classes (i.e., classes with a low cardinality).

Experiments performed with larger gene sets (i.e., 50 genes) showed a similar behavior. The original accuracy is higher (for example, it is 77.78% for BW when a set of 50 genes is considered) and the average difference in accuracy is lower (about 0.5% for the classification distance and -0.3% for the cosine distance). When the number of considered genes increases, the effect of a single gene on the classification performance becomes less evident. Hence, these experiments are less effective in evaluating the char-

acteristics of the classification distance.

Biological investigation.
To assess the biological meaning of similar genes, we focused on the Colon and Prostate datasets, which have been widely studied in previous works. Two genes that are known to play a role in the colon tumor progression are J02854 (Myosin regulatory light chain 2, smooth muscle isoform) and M76378 (Cysteine-rich protein gene). According to the classification distance, the genes nearest to J02854 are M63391, T92451, R78934, and T60155. Gene M63391 is listed in the top relevant genes for colon cancer in (Chen et al, 2007; Yu et al, 2004; Bo and Jonassen, 2002; Ben-Dor et al, 2000), while gene T60155 is cited in (Ben-Dor et al, 2000) and (Yu et al, 2004). Furthermore, the genes nearest to M76378 are M63391 and J02854, both relevant for colon cancer. We also analyzed the performance of other distance measures on the Colon dataset. The cosine correlation shows a similar behavior. For example, in the case of gene J02854, it detects as nearest three of the genes detected by the classification distance (R78934, T60155, T92451). On the contrary, there is no intersection between the nearest genes yielded by the classification and Euclidean distances. For example, for the Euclidean distance, the nearest to gene J02854 are genes R87126, X12369, R46753 and R67358. Among them, only gene X12369 shows a correlation to the colon cancer (Yang and Zhang, 2007).

In the prostate cancer the ETS-related gene (ERG), a member of the ETS transcription factor family, is the most frequently overexpressed proto-oncogene in the transcriptome of malignant prostate epithelial cells (Petrovics et al, 2005; Gregg et al, 2010). The classification distance detects as the most similar genes the Lys-Asp-Glu-Leu endoplasmic reticulum protein retention receptor 3 (KDELR3), the fibroblast growth factor binding protein 1 (FGFBP1), the TNF receptor-associated factor 2 (TRAF2) and the annexin A7 (ANXA7) which show an overexpression and play an important role in the prostate cancer proliferation as reported in (Aicha et al, 2007; Royuela et al, 2008; Rosini et al, 2002; Torosyan et al, 2002).

These results show that our distance metric groups genes with both comparable classification accuracy and similar biological meaning. Hence, our method can effectively support further investigation in biological correlation analysis.

4.2. Core expression interval comparison

Recall from Section 2.1 that the MAD estimator smooths the effect of values far from the median value, independently of their density. Instead, WMD takes into account the density of values and smooths the effects of isolated values. The core expression intervals defined by MAD are usually narrower than those defined by WMD. Thus, the number of ones in the masks is generally larger for MAD, because the intervals are less overlapped. Figure 3 reports the boxplots of the distributions of the number of ones in the masks corresponding to intervals generated by means of WMD and MAD.

For each gene i, we computed the similarity between the masks generated by the two approaches (both characterized by S bits) by means of the following formula:

$$Similarity(mask_{i,MAD}, mask_{i,WMD}) = \frac{1}{S}\sum_{j=1}^{S}\overline{mask_{ij,MAD} \oplus mask_{ij,WMD}} \quad (9)$$

Figure 4 shows the boxplot of the distribution of the similarity values. The masks agree in roughly 90% of cases (i.e., gene/class pairs).

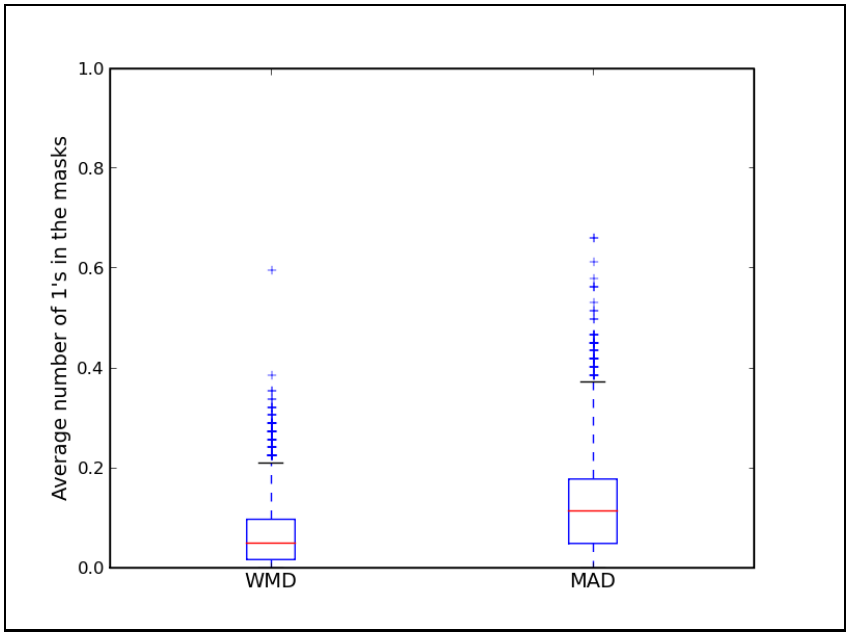

Fig. 3. Boxplots of the distributions of ones in the gene masks created by using the WMD (left) and MAD (right) methods for outlier detection.

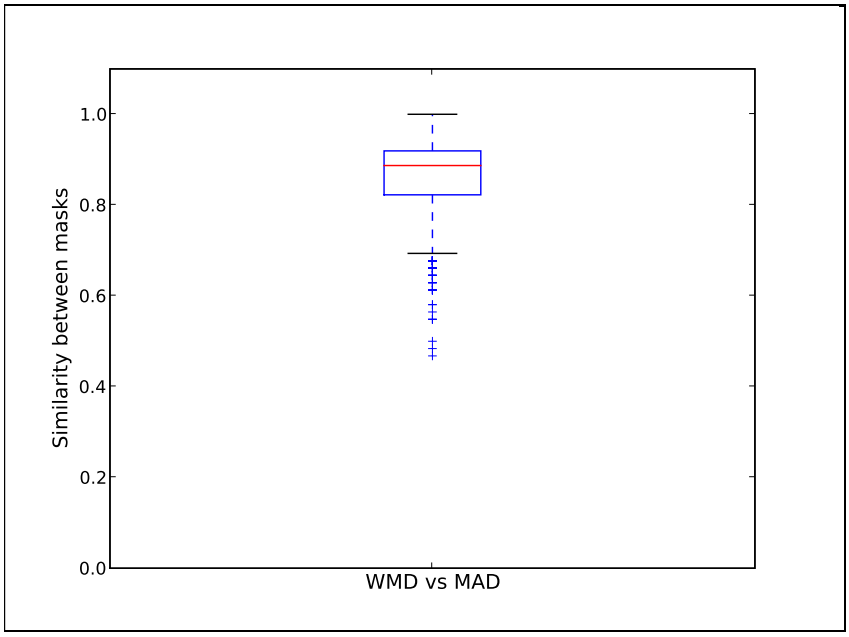

Fig. 4. Boxplot of the similarity between the gene masks created by using the WMD and MAD methods for outlier detection.

We also analyzed the classification accuracy yielded by the gene mask representations provided by the MAD and the WMD methods. The same experimental design described in Section 4.1 has been used for these experiments. In most cases WMD provided a better accuracy than MAD. For example on the Brain1 dataset, the difference in accuracy between the original subset and the modified subset obtained by exploiting the MAD technique is -0.22$\pm$1.74 with ANOVA, 3$\pm$3.07 with BW, 1.56$\pm$2.24 with OVO, and -6.33$\pm$1.74 with OVR. Thus, for ANOVA, OVO and OVR, WMD accuracy (see Table 2) is higher than MAD accuracy. Furthermore, the standard deviation of the accuracy difference of MAD is, on average, larger than the standard deviation of WMD, thus showing a less stable behavior. Similar results are obtained for the other datasets.

This behavior may be due to an overestimation of the gene classification power when intervals are defined by means of MAD. In particular, since the core expression intervals defined by MAD are narrower, they are also less overlapped. Hence, the resulting masks are characterized by a larger number of ones, which represent a higher gene discriminating capability.

4.3. Cluster characterization

We evaluated the characteristics of the hierarchical clustering algorithm presented in Section 3, which integrates the classification distance measure. Since sample class labels are available, but gene class labels are unknown, the result of gene clustering cannot be straightforwardly validated. To evaluate the characteristics of our approach, we (i) compared by means of the Rand Index (Rand, 1971) the clustering results obtained by using our measure, the cosine, and the Euclidean metrics, (ii) analyzed the variation of the cluster size when varying the cluster number, and (iii) evaluated the homogeneity of the clusters by analyzing the classification behavior of genes included into the same cluster. Clustering results, together with a tool to navigate the dendrogram and explore the clusters, are available on our website. [2]

Rand Index

To measure the agreement between the clustering results obtained with different metrics, we computed the Rand Index (Rand, 1971). It measures the number of pairwise agreements between a clustering K and a set of class labels C over the same set of objects. It is computed as follows

$$R\left(C,K\right)=\frac{a+b}{\binom{N}{2}} \tag{10}$$

where a denotes the number of object pairs with the same label in C and assigned to the same cluster in K, b denotes the number of pairs with a different label in C that were assigned to a different cluster in K and N is the number of objects. The values of the index are in the range 0 (totally distinct clusters) to 1 (exactly coincident clusters). The Rand Index is meaningful for a number of clusters in the range $[2; N-1]$, where N is the number of objects. Clusters composed by a single element provide no contribution to the Rand Index evaluation (Rand, 1971).

To perform a pairwise comparison of the clustering results obtained by different distance metrics, we selected one metric to generate the clustering K and used as labels

[2] `https://dbdmg.polito.it/twiki/bin/view/Public/ClassificationDistance`

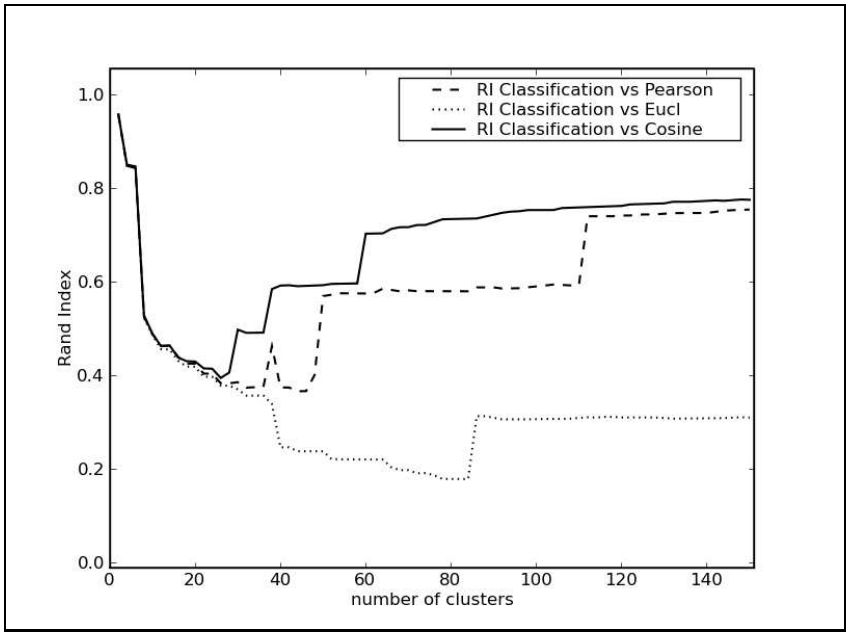

Fig. 5. Pairwise Rand index evaluation between classification, Euclidean, and cosine distance metrics on the Colon dataset.

C the cluster identifiers obtained by clustering with the same hierarchical algorithm and a different distance metric. We repeated the process to perform the pairwise comparison of all three metrics. The results for the Colon dataset are shown in Figure 5. Similar results are obtained on the other datasets. Hierarchical clustering based on the classification distance shows a good agreement (ca. 70%) with cosine correlation clustering. Instead, the Rand Index between classification distance clustering and Euclidean distance clustering is very low. This last behavior is similar to that between Euclidean distance clustering and cosine correlation clustering.

Cluster size

We evaluated the trend of the maximum cluster size when increasing the number of final clusters (from 1 to 150 clusters) for the Euclidean distance, Pearson correlation and Classification distance metrics. Figure 6 shows the results on the Brain1 dataset (characterized by 5920 genes). The other datasets showed a similar behavior.

The Euclidean distance typically yields one big cluster containing the majority of genes and a number of small clusters with few genes. The maximum size is stable around 5740 genes and remains constant until 554 clusters, where it falls to 4810. Pearson correlation also creates one very large cluster (the number of elements is roughly 4800), whose size abruptly changes to roughly half of the genes (around 2500 elements) around 60 clusters. Cosine correlation shows a behavior similar to the Pearson correlation, but the maximum size abruptly changes around 25 clusters. Classification distance yields a decrease in the cluster size until a maximum size around 1000 genes. Hence, it partitions data in smaller clusters, while the other distance measures typically yield a very large cluster, which behaves as a generic gene container.

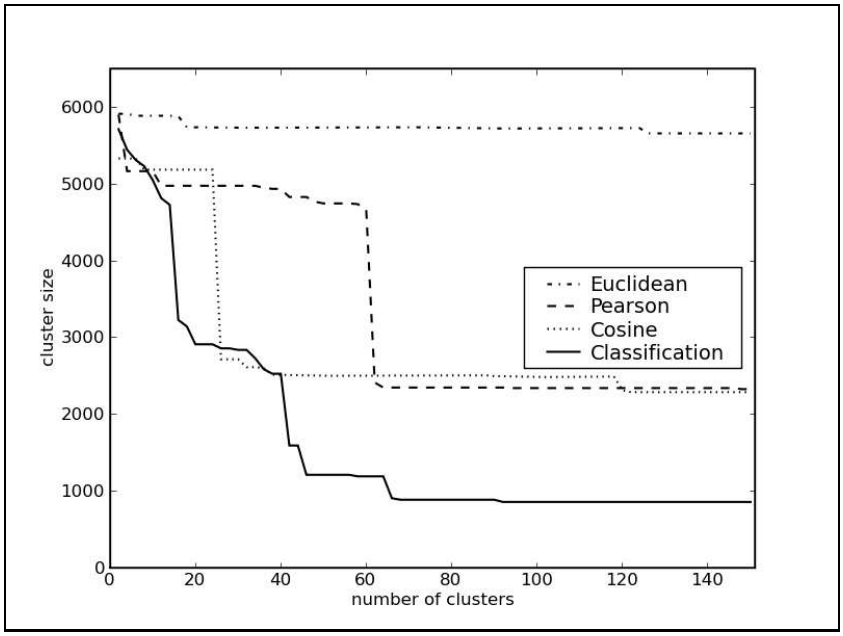

Fig. 6. Maximum cluster size for an increasing number of clusters for Euclidean distance, Pearson correlation and classication distance.

Cluster homogeneity

To evaluate cluster homogeneity, we compared the classification accuracy of genes belonging to the same cluster. To this aim, we defined two genes as representatives of each cluster, i.e., the one with the minimum (named central) and the one with the maximum (named border) classification distance to the cluster mask.

We only considered informative clusters, i.e., clusters containing relevant information for classification purposes, thus ignoring noise clusters. Informative clusters are selected by (i) identifying relevant genes, denoted as original genes in the following, by means of feature selection methods, (ii) selecting clusters such that each cluster contains a single original gene. More specifically, for the ANOVA, BW, OVO, and OVR feature selection methods, we selected the 10, 50 and 100 top ranked genes in a given dataset. For each original gene (i.e., gene in the rank), the largest cluster containing this gene and no other original gene is selected. In this way, three subsets of clusters are defined: (i) with 10 clusters, (ii) with 50 clusters, and (iii) with 100 clusters. For a larger number of clusters, the cluster size became too small and the analysis was not relevant.

Three different classification models have been built by considering (a) all original genes, (b) the substitution of each original gene with the central gene in its cluster, and (c) the substitution of each original gene with the border gene in its cluster. Classification accuracy has been computed in all three settings for each dataset, each feature selection method and each gene subset (i.e., 10, 50, and 100 genes).

Table 4 reports the original accuracy values (setting (a)) and the difference with respect to settings (b) and (c) for the OVO feature selection method on all datasets. The average size of the pool from which equivalent genes are drawn (i.e., the average cluster size) is reported in Table 5. Similar results have been obtained for the other feature selection methods.

Table 4. Differences from the original OVO rank accuracy on all datasets by using the central and the border genes

Dataset	*N*	*Original*	*Diff_central*	*Diff_border*
Brain1	10	74.45	0.00	0.00
	50	85.56	2.22	0.00
	100	84.45	2.22	1.11
Leuk1	10	94.44	0.00	-1.38
	50	97.22	0.00	2.17
	100	95.83	0.00	0.00
Lung	10	86.21	-1.97	-4.93
	50	94.09	0.00	0.98
	100	97.04	-1.47	0.00
Tumor9	10	54.89	7.72	1.54
	50	70.12	1.78	-3.33
	100	66.40	-1.11	-1.11
Leuk2	10	93.06	-1.39	-2.78
	50	94.44	0.00	0.00
	100	93.06	2.77	1.38
SRBCT	10	93.98	-1.21	-7.23
	50	100.00	0.00	0.00
	100	100.00	0.00	0.00
Prostate	10	93.14	0.00	0.00
	50	91.18	0.00	0.00
	100	92.16	0.00	0.98
DLBCL	10	85.71	2.60	1.30
	50	94.81	0.00	0.00
	100	96.10	1.30	1.30
Colon	10	81.97	0.00	0.00
	50	86.89	0.00	0.00
	100	86.89	0.00	0.00
Mean$\pm$*Std*	10		0.64n$\pm$2.96	-1.50n$\pm$2.96
	50		0.44n$\pm$0.89	-0.02n$\pm$1.45
	100		0.41n$\pm$1.42	0.41n$\pm$0.83

Differences from the original classification accuracy are low. Clusters formed by a single gene (e.g., for the Colon and Prostate datasets) are not significant, because obviously the difference in accuracy is equal to zero. For larger clusters the differences are always limited to few percentage points. For example, for the ten cluster case on the Brain1, Leuk1, Leuk2 and DLBCL (cluster size range from about 3 to 6 genes) the difference in accuracy varies from -2.78 to 2.60. Always in the ten cluster case, the bad performance of SRBCT is due to the fact that one of the selected genes is located in a big cluster (average cluster size 124.90 genes). Thus, the border gene might be very different from the original gene.

On average, the obtained clusters provide a good quality gene pool from which equivalent genes may be drawn. The substitution with the central gene usually provides better results with respect to the substitution with the border gene. This difference is more significant for the larger clusters obtained for the 10 gene subset, than for the smaller, more focused clusters obtained in the case of the 50 or 100 gene subsets.

Table 5. Average cluster size for the experiment reported in Table 4

N	*Brain1*	*Leuk1*	*Lung*	*Tumor9*	*Leuk2*	*SRBCT*	*Prostate*	*DLBCL*	*Colon*
10	4.20	6.20	17.00	20.90	3.60	124.90	1.00	6.30	1.00
50	8.00	15.10	2.06	1.92	1.58	1.90	1.00	1.00	1.00
100	1.48	1.25	1.24	1.06	7.98	1.38	1.54	5.45	1.00

5. Conclusions

In this paper we propose a new similarity measure between genes, the *classification distance*, that exploits additional information which may be available on microarray data (e.g., tumor or patient classification). The discrimination ability of each gene is represented by means of a gene mask, which describes the gene classification power, i.e., its capability to correctly classify samples. The classification distance measures gene similarity by analyzing their masks, i.e., their capability of correctly classifying the same samples.

The classification distance measure can be integrated in different clustering approaches. We have integrated it into a hierarchical clustering algorithm, by introducing the notion of cluster mask as representative of a cluster and defining as inter-cluster distance the distance between cluster masks. We validated our method on both binary and multiclass microarray datasets. The experimental results show the ability of the classification distance to group genes with similar classification power and similar biological meaning in the tumor context.

Currently, we are considering to integrate our distance metric in a (supervised) feature selection algorithm. By clustering genes which correctly classify the same samples and then selecting a single gene from each cluster, redundant genes are disregarded and both model coverage and classification accuracy may be improved.

We believe that the classification distance measure may be applied also in other application domains with the same characteristics (e.g., user profiling, hotel ranking, etc.), to improve the clustering results by exploiting additional information available on the data being clustered.

References

Aicha SB, Lessard J, Pelletier M, Fournier A, Calvo E, Labrie C (2007) Transcriptional profiling of genes that are regulated by the endoplasmic reticulum-bound transcription factor AIbZIP/CREB3L4 in prostate cells. Physiological genomics, 31(2):295

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences, 96(12):6745–6750

Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z (2000) Tissue classification with gene expression profiles. Journal of Computational Biology, 7(3-4), Mary Ann Liebert Inc., pp 559–583

Bo T, Jonassen I (2002) New feature subset selection procedures for classification of expression profiles. Genome biology, 3(4), Biomed Central, pp 17

Bouguessa M, Wang S (2009) Mining Projected Clusters in High-Dimensional Spaces. IEEE Transaction on Knowledge and Data Engineering, 21(4):507–522

Bushel PR, Wolfinger RD, Gibson G (2007) Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. BMC Systems Biology, 1(1), Biomed Central, pp 15

Chang CC, Lin CJ (2001) Training v-support vector classifiers: theory and algorithms. Neural Computation, 13(9), MIT press, pp 2119–2147

Chen JJ, Tsai CA, Tzeng SL, Chen CH (2007) Gene selection with multiple ordering criteria. BMC Bioinformatics, 8(1):74

Chu T, Huang J, Chuang K, Yang D, Chen M (2010) Density Conscious Subspace Clustering for High-Dimensional Data. IEEE Transaction on Knowledge and Data Engineering, 22(1):16–30

Cox TF, Cox MAA (2001) Multidimensional Scaling. Chapman and Hall

Daszykowski M, Kaczmarek K, Vander Heyden Y, Walczak B (2007) Robust statistics in data analysis - A review: Basic concepts. Chemometrics and Intelligent Laboratory Systems, 85(2), Elsevier, pp 203–219

Datta S, Datta S (2006) Evaluation of clustering algorithms for gene expression data. BMC Bioinformatics, 7(Suppl 4):S17

Davies L, Gather U (1993) The identification of multiple outliers. Journal of the American Statistical Association, American Statistical Association, pp 782–792

D'haeseleer P (2005) How does gene expression clustering work? Nature Biotechnology, 23(12):1499–1501

El Akadi A, Amine A, El Ouardighi A, Aboutajdine D. (2010) A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. Knowledge and Information Systems

Ester M, Kriegel H, Jörg S, Xu X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp 226–231

Everitt BS, Landau S, Leese M (2009) Cluster Analysis. Wiley, 4th Edition

Fu L, Medico E (2007) FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. BMC Bioinformatics, 8(1):3

Fu Q, Banerjee A (2008) Multiplicative Mixture Models for Overlapping Clustering. Proceedings of the Eighth IEEE International Conference on Data Mining, pp 791–796

Furlanello C, Serafini M, Merler S, Jurman G (2003) Entropy-based gene ranking without selection bias for the predictive classification of microarray data. BMC Bioinformatics, 4(1):54

Gevaert O, Smet FD, Timmerman D, Moreau Y, Moor BD (2006) Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics, 22(14)

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA and others (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286(5439), AAAS, pp 531

Gregg JL, Brown KE, Mintz EM, Piontkivska H, Fraizer GC (2010) Analysis of gene expression in prostate cancer epithelial and interstitial stromal cells using laser capture microdissection. BMC cancer, 10(1):165

Gu J, Liu J (2008) Bayesian biclustering of gene expression data. BMC Genomics, 9(Suppl 1):S4

Hampel FR (1974) The influence curve and its role in robust estimation. Journal of the American Statistical Association, 69:383–393

He X, Cai D, Niyogi P. (2006) Laplacian score for feature selection. Advances in Neural Information Processing Systems, 18:507

Huang D, Pan W (2006) Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. Bioinformatics, 22(10):1259–1268

Jiang D, Pei M, Ramanathan C, Lin C, Tang C, Zhang A (2006) Mining gene-sample-time microarray data: a coherent gene cluster discovery approach. Knowledge and Information Systems, 13(3):305–335

Jiang D, Tang C, Zhang A (2004) Cluster analysis for gene expression data: A survey. IEEE Transactions on Knowledge and Data Engineering, 16(11):1370–1386

Kaufman L, Rousseeuw PJ (1990) Finding groups in data: An introduction to cluster analysis. Wiley, New York

Liu H, Motoda H (2007) Computational Methods of Feature Selection. Chapman & Hall/CRC

Liu J, Wang W (2003) Op-cluster: Clustering by tendency in high dimensional space. Proceedings of the ICDM 2003 Conference, pp 187–194

Mitra P, Majumder DD (2004) Feature selection and gene clustering from gene expression data. Proceedings of the Pattern Recognition, 17th International Conference on, 2:343–346

Mukkamala S, Liu Q, Veeraghattamand R, Sung A (2006) Feature Selection and Ranking of Key Genes for Tumor Classification: Using Microarray Gene Expression Data. Springer Berlin/Heidelberg

Ng M, Chan L (2005) Informative gene discovery for cancer classification from microarray expression data. IEEE Workshop on Machine Learning for Signal Processing, pp 393–398

Pearson RK, Gonye GE, Schwaber JS (2003) Outliers in microarray data analysis. Methods of Microarray Data Analysis III, Springer, pp 41

Petrovics G, Liu A, Shaheduzzaman S, Furasato B, Sun C, Chen Y, Nau M, Ravindranath L, Chen Y, Dobi A and others (2005) Frequent overexpression of ETS-related gene-1 (ERG1) in prostate cancer transcriptome. Oncogene, 24(23):3847–3852

Rand WM (1971) Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, American Statistical Association, pp 846–850

Rosini P, Bonaccorsi L, Baldi E, Chiasserini C, Forti G, De Chiara G, Lucibello M, Mongiat M, Iozzo RV, Garaci E and others (2002) Androgen receptor expression induces FGF2, FGF-binding protein produc-

tion, and FGF2 release in prostate carcinoma cells: role of FGF2 in growth, survival, and androgen receptor down-modulation. The Prostate, 53(4):310–321
Royuela M, Rodríguez-Berriguete G, Fraile B, Paniagua R. (2008) TNF-alpha/IL-1/NF-kappaB transduction pathway in human cancer prostate. Histology and histopathology, 23(10):1279
Song J, Liu C, Song Y, Qu J (2008) Clustering for DNA Microarray Data analysis with a Graph Cut Based Algorithm. Seventh International Conference on Machine Learning and Applications
Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics, 21(5):631–643
Thalamuthu A, Mukhopadhyay I, Zheng X, Tseng GC (2006) Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics, 22(19):2405
Thompson RC, Deo M, Turner DL (2007) Analysis of microRNA expression by in situ hybridization with RNA oligonucleotide probes. Methods, 43(2), Elsevier, pp 153–161
Torosyan Y, Dobi A, Glasman M, Mezhevaya K, Naga S, Huang W, Paweletz C, Leighton X, Pollard HB, Srivastava M (2010) Role of multi-hnRNP nuclear complex in regulation of tumor suppressor ANXA7 in prostate cancer cells. Oncogene, 29(17):2457–2466
Wang H, Wang W, Yang J, Yu PS (2002) Clustering by pattern similarity in large data sets. Proceedings of the 2002 ACM SIGMOD international conference on Management of data, pp 394–405
Wang L, Leckie C, Ramamohanarao K, Bezdek J (2009) Automatically Determining the Number of Clusters in Unlabeled Data Sets. IEEE Transaction on Knowledge and Data Engineering, 21(3):335–350
Wang Z, Yan P, Potter D, Eng C, Huang TH, Lin S (2007) Heritable clustering and pathway discovery in breast cancer integrating epigenetic and phenotypic data. BMC Bioinformatics, 8(1):38
Wright TW (1884) A Treatise on the Adjustment of Observations. Van Nostrand, New York
Xiong M, Fang X, Zhao J (2001) Biomarker identification by feature wrappers. Genome Research, 11(11):1878–1887
Yang P, Zhang Z (2007) Hybrid Methods to Select Informative Gene Sets in Microarray Data Classification. Lecture Notes in Computer Science, 4830, Springer, pp 810
Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research, 30(4):e15
Yap YL, Zhang XW, Ling MT, Wang XH, Wong YC, Danchin A (2004) Classification between normal and tumor tissues based on the pair-wise gene expression ratio. BMC Cancer, 4(1):72
Yousef M, Jung S, Showe LC, Showe MK (2007) Recursive Cluster Elimination(RCE) for classification and feature selection from gene expression data. BMC Bioinformatics, 8(1):144
Yu LTH, Chung F, Chan SCF, Yuen SMC (2004) Using emerging pattern based projected clustering and gene expression data for cancer detection. Proceedings of the second conference on Asia-Pacific bioinformatics, 29:75–84
Zapala MA, Schork NJ (2006) Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. Proceedings of the National Academy of Sciences, 103(51):19430
Zhao Y, Wang G, Yin Y, Yu G (2006) Mining Positive and Negative Co-regulation Patterns from Microarray Data. Sixth IEEE Symposium on BioInformatics and BioEngineering, pp 86–93

Correspondence and offprint requests to: Giulia Bruno, Dipartimento di Automatica e Informatica, Politecnico di Torino, Torino, Italy. Email: giulia.bruno@polito.it