2011 IEEE International Workshop on Genomic Signal Processing and Statistics
December 4-6, 2011, San Antonio, Texas, USA

Feature Set Enhancement via Hierarchical Clustering for Microarray Classification

Mattia Bosio, Pau Bellot Pujalte, Philippe Salembier, Albert Oliveras-Vergés
Department of Signal Theory and Communications, Technical University of Catalonia,
UPC-Campus Nord, C/ Jordi Girona, 1-3, 08034, Barcelona, Spain,
E-mail: mattia.bosio@upc.edu

***Abstract*—A new method for gene expression classification is proposed in this paper. In a first step, the original feature set is enriched by including new features, called metagenes, produced via hierarchical clustering. In a second step, a reliable classifier is built from a wrapper feature selection process. The selection relies on two criteria: the classical classification error rate and a new reliability measure. As a result, a classifier with good predictive ability using as few features as possible to reduce the risk of overfitting is obtained. This method has been tested on three public cancer datasets: leukemia, lymphoma and colon. The proposed method has obtained interesting classification results and the experiments have confirmed the utility of both metagenes and feature ranking criterion to improve the final classifier.**

***Keywords*—hierarchical clustering; feature selection; cancer microarray classification; Treelet.**

I. Introduction

Microarrays are a powerful high-throughput technology which is often used for classification purposes because it allows the simultaneous measurement of thousands of gene expression values. A microarray dataset is typically composed of only tens of observations due to the cost of the experiments. This characteristic of sample scarcity makes necessary a feature selection process to produce reliable classifiers [1].

The aim of this work is to propose a method for microarray classification able to reach small prediction error and using as few features as possible to reduce the risk of overfitting. Many different approaches for microarary classification exist in the literature, among which evolutionary algorithms have obtained good results [2], but [3] mentions how their performances decrease when the feature set dimension grows. Furthermore, algorithms like *Tree Harvesting* highlight the usefulness of feature set expansion via hierarchical clustering [4]. The possibility to summarize groups of correlated genes in a single feature as input for the classifier has various advantages. First, the interpretability of the selected feature as a combination of correlated genes that may be involved in the same biological process. Second, robustness to noise because a group of correlated genes useful for classification is less likely to be affected by noise than individual genes. Third and last, classifying with cluster representing features can help to discover linear (or non linear) relations among groups of correlated genes.

In this paper, the benefits of an expanded feature set and of a flexible feature selection algorithm are pursued through a two-step process. At first, the original feature set of gene expression values is enriched with new features that are linear combinations of the original genes. These new features are called metagenes and are produced by hierarchical clustering. In the second step, a wrapper feature selection process is performed [3]; its aim is to find a reduced set of features on which to apply a classification algorithm. The main contributions of this paper are the way the metagenes are created and the use of a reliability criterion in the feature selection process.

This paper is organized as follows: in Section II, the metagenes creation is explained. In Section III, the feature selection algorithm is detailed. Section IV contains the adopted experimental protocol. Results and conclusions are discussed respectively in Section V and VI.

II. Metagenes Creation

In this section, the expansion of the original feature set by the introduction of metagenes obtained through hierarchical clustering is explained. The clustering process, here, is not used to find a structure of the original feature set, but to generate a new set of features, each of which summarizes in itself a cluster of genes. The objective is to create metagenes that are linear combinations of genes with common characteristics and, as a result, to reduce the noise thanks to the filtering effect of the linear combination.

Many hierarchical clustering techniques exist, depending on the chosen clustering similarity measure and the metagene generation rule (i.e. how the linear combination of the elements in a metagene is defined). The used approach is a bottom up, pairwise hierarchical clustering described in Figure 1.

The hierarchical clustering is an iterative algorithm that, at each step, merges the pair of most similar features (initially only genes). The merging creates a metagene expressed as linear combination of its features. The newly created metagene is then added to the feature set whereas the pair of most similar features are removed from it. At the end of the process, the initial feature set of p genes is expanded with p-1 metagenes. The key points in the metagene creation are the similarity metric: $d(\cdot,\cdot)$, and the generation rule: $g(\cdot,\cdot)$. Changing either one of these two functions implies the creation of a different metagene set.

Two different metagene generation methods have been studied in this work. The first technique is based on Lee's work [5], where an adaptive method for multi-scale representation and eigen-analysis of data called *Treelets* is presented. With

978-1-4673-0490-0/11/$26.00 ©2011 IEEE

Original feature set $\underline{\underline{G}}_0 = \{\underline{g}_1, \ldots, \underline{g}_p\}$
Active feature set $\underline{\underline{F}} = \underline{\underline{G}}_0$
Metagene set $\underline{\underline{M}} = \emptyset$
For *i* = 1 : *p-1*

1) *Calculate pairwise similarity metric* $d(\underline{f}_a, \underline{f}_b)$ *for all features in* $\underline{\underline{F}}$

2) *Find a,b* : $d(\underline{f}_a, \underline{f}_b) = \max(d(\cdot,\cdot))$

3) *New metagene* $\underline{m}_i = g(\underline{f}_a, \underline{f}_b)$ *generation:*
$$\underline{m}_i = \alpha_a \underline{f}_a + \alpha_b \underline{f}_b = \sum_{i=1}^{p} \beta_i \underline{g}_i;$$
Each metagene is a linear combination of all original features $\underline{g}_i$

4) $\underline{\underline{F}} := \underline{\underline{F}} \cup \{\underline{m}_i\}$: *add new metagene to active feature set*

5) $\underline{\underline{F}} := \underline{\underline{F}} \backslash \{\underline{f}_a, \underline{f}_b\}$: *remove the two features* $\underline{f}_a, \underline{f}_b$ *from the active feature set*
$\underline{\underline{M}} := \underline{\underline{M}} \cup \{\underline{m}_i\}$: *join metagene* $\underline{m}_i$ *to metagene set*

end
Define the new expanded feature set: $\underline{\underline{F}} = \underline{\underline{G}}_0 \cup \underline{\underline{M}}$ *as the union of metagenes and original gene expression profiles.*

Fig. 1. General clustering algorithm.

Feature set $\underline{\underline{F}}_0 = \{\underline{f}_1, \underline{f}_2, \underline{f}_3\}$ *with* $\underline{f}_1 = \underline{f}_2 = \underline{f}_3$
Two metagenes are created

1) *metagene* $\underline{m}_1$ *joining* $\underline{f}_1$ *and* $\underline{f}_2$
$$\underline{m}_1 = \sqrt{1/2} \cdot \underline{f}_1 + \sqrt{1/2} \cdot \underline{f}_2$$
$$\underline{m}_{1scaled} = 1/2 \cdot \underline{f}_1 + 1/2 \cdot \underline{f}_2$$

2) *metagene* $\underline{m}_2$ *joining* $\underline{m}_1$ *and* $\underline{f}_3$
$$\underline{m}_2 = \sqrt{2/3} \cdot \underline{m}_1 + \sqrt{1/3} \cdot \underline{f}_3$$
$$\underline{m}_2 = \sqrt{1/3} \cdot \underline{f}_1 + \sqrt{1/3} \cdot \underline{f}_2 + \sqrt{1/3} \cdot \underline{f}_3$$
$$\underline{m}_{2scaled} = 1/3 \cdot \underline{f}_1 + 1/3 \cdot \underline{f}_2 + 1/3 \cdot \underline{f}_3$$

Scaled versions $\underline{m}_{1scaled}$ *and* $\underline{m}_{2scaled}$ *are used for Euclidean clustering because they preserve the components dynamics. The scaled versions will expand the feature set.*
Non scaled versions $\underline{m}_1$ *and* $\underline{m}_2$ *are used in the construction phase with PCA as they preserve the energy distribution among the components*

Fig. 2. Example of metagene creation with *Euclidean* clustering.

Treelets, a clustering tree is built, in which at each level of the tree, the two most similar features are grouped and replaced by a coarse-grained approximation feature and a residual detail feature. The similarity measure in *Treelets* is the normalized correlation: $d(\underline{f}_a, \underline{f}_b) = \langle \underline{f}_a, \underline{f}_b \rangle / (\|\underline{f}_a\| \cdot \|\underline{f}_b\|)$, where $\underline{f}_a$ is the sequence of expression values for all samples. The two replacing features, approximation and detail, are obtained through a local Principal Component Analysis (i.e. PCA) on two dimensions: the approximation is the first local principal component and the detail is the second one. Here, *Treelets* has been applied on microarray data and the approximation feature in each level has been selected as metagene. The metagene corresponds to the approximation because the metagene summarizes the genes contained in the cluster. Note that, as the linear combination obtained through PCA $\underline{m}_i = \alpha_a \underline{f}_a + \alpha_b \underline{f}_b$ can be seen as an unitary transform, we have $\alpha_a^2 + \alpha_b^2 = 1$.

The second technique, called *Euclidean* clustering, adopts a similar iterative procedure but using the negative Euclidean distance, $d(\underline{f}_a, \underline{f}_b) = -\|\underline{f}_a - \underline{f}_b\|_2$, as similarity measure, so that the maximum is zero when the two features are equal. The Euclidean distance has been selected because it measures the point-wise similarity rather than the profile shape similarity as done by the normalized correlation. This choice implies a modification of the whole clustering process because choosing the first PCA component as metagene implies that a metagene is a scaled weighted average of the genes. An illustrative example is presented in Figure 2, in which all features are equal and using the first PCA component produces metagenes which are not pure weighted average of genes. There is a scaling factor that moreover depends on the number of genes in the cluster. This phenomenon is irrelevant in the *Treelets* case, as scaling does not affect the normalized correlation, but it is important when Euclidean distance is considered because, if we want to be able to compare genes and metagenes, the metagenes should be a non scaled weighted average of genes. To do so, when a metagene $\underline{m}_x$ is created, two versions of it are used, one is the same as in the *Treelets* case, and the second one is a scaled version: $\underline{m}_{xscaled} = \underline{m}_x / \|\underline{\beta}\|_1$. The scaled version $\underline{m}_{xscaled}$ is a weighted average of the corresponding genes, thus it is used to calculate the pairwise distance and considered as a metagene. The original non scaled version, instead, is used when a new metagene is built from $\underline{m}_x$ to preserve the energy distribution among the basic components as showed in Figure 2.

III. Feature selection process

In this section, the adopted feature selection process is presented. The objective is to select from the *expanded* feature set (initial genes and metagenes obtained through clustering) the smallest subset to build a good classifier. To avoid evolutionary algorithms problems when the feature set dimension grows, a deterministic approach has been used here. It tries to preserve the advantages of an evolutionary search allowing mutations of previous choices. Figure 3 illustrates the algorithm flowchart. It is a modification of the Sequential Floating Forward Selection algorithm (SFFS) with the introduction of a replacing step when backtracking does not work. It is called *Improved sequential Floating Forward Selection* (IFFS) [3].

A. Feature selection algorithm

The search process starts with an empty set and ends when a threshold value is reached. The threshold is either the maximum accepted number of features or a maximum number of iterations in case the algorithm has entered in an infinite loop. After the initialization, the selection process enters in a loop of tasks. At first, there is the *add* phase, where all the features that have not yet been selected are tested one by one. For each one, the current feature set is expanded by adding the new feature, a classifier is trained and the corresponding classification score $J(\cdot)$ is calculated. The feature obtaining the best $J(\cdot)$ is added to the current set. Then, if the threshold

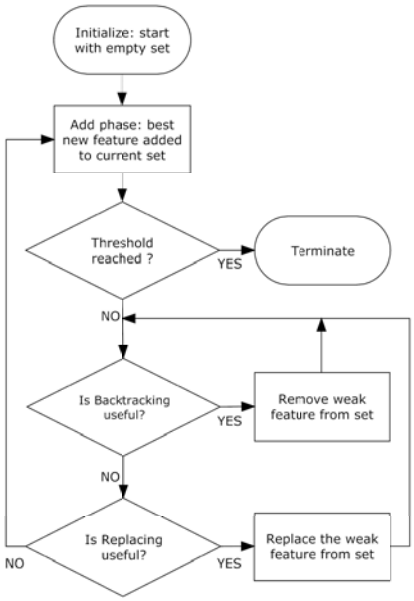

Fig. 3. IFFS feature selection algorithm.

has not been reached, the algorithm starts a *backtracking* phase. In this step, the weakest feature in the subset (i.e. the feature whose elimination implies the minimum performance loss or the maximum performance gain) is candidate for elimination. If the elimination improves $J(\cdot)$, the weakest feature is actually removed and a new *backtracking* phase is performed. Otherwise, the algorithm looks for substituting one feature in the *replacing* phase. For each feature in the current set, a substitute is chosen via an analysis like in the *add* phase. If the best substitution has proven useful, (i.e. the $J(\cdot)$ value with the substitution is better than without), the current set is updated and a new *backtracking* phase takes place. Otherwise, the algorithm goes back to the *add* phase.

B. Feature ranking criterion

The search algorithm is a wrapper feature selection process, so a classifier is applied inside the selection phase. A Linear Discriminant Analysis (LDA) classifier has been used in this study for its simplicity and to obtain a classifier robust to overfitting [6]. In Section III-A, the criterion $J(\cdot)$ is involved in the feature selection process. Its task is to rank features and to decide which one is the best. $J(\cdot)$ is a measure of how good a classifier is to predict new samples. To obtain a reliable value estimation, a 10-fold cross validation process, in which the original sample set is split ten times into a training set (approx. 90% of samples) and a test set (approx. 10% of samples); the $J(\cdot)$ value is then calculated on the test set. Due to the microarray data characteristic involving few samples and many dimensions, a $J(\cdot)$ criterion based only on error rate may not be enough in ranking features. Indeed, it is common to have a group of features with the same error rate from which only one feature has to be selected. To overcome this limitation, a two level criterion is introduced. In the first level, features are sorted for the mean misclassification rate on the test set along the cross validation iterations (error rate).

In the second level, a reliability parameter r is calculated to quantify the estimation goodness as a weighted sum of sample distances from the classification boundary. It is calculated on the test set samples and the final reliability value is the mean of the cross validation iterations. The reliability is calculated inside a cross-validation iteration for a two-class problem. It is defined in (1), where n_{test} is the test set dimension, c_l is the class of sample l (it can be 1 or 2), and $p(c_l)$ is the probability of class c_l in the test set. The value d_l is the Euclidean distance of sample l from the classifier boundary with positive sign in case of correct classification or negative sign otherwise.

$$r = \frac{1}{n_{test} \cdot \hat{\sigma}_d} \sum_{l=1}^{n_{test}} \frac{d_l}{p(c_l)} \tag{1}$$

Finally, $\hat{\sigma}_d = \sqrt{\frac{\hat{\sigma}_1}{n_1} + \frac{\hat{\sigma}_2}{n_2}}$, is an estimation of intra class variance of the sample distances from the classification boundary. In order to get a reliable estimation, the intra-class variance is estimated using the training and the test samples; n_1 and n_2 are the number of samples in class 1 and 2 respectively. It is obtained like in the independent two-sample t-test with classes of different size and variance, as it is the most general case for a two-class problem. In detail $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are the estimated variances of sample distance from boundary for all samples of class 1 and 2 respectively. Dividing by $\hat{\sigma}_d$ guarantees that r is invariant through scaling of the feature set. Dividing by $p(c_l)$ assigns to each class the same relative weight and it is useful when the test set distribution is not uniform among the classes. Reliability value, $r \in [\ \infty\, \infty]$, is positively influenced by large mean class separation in the perpendicular direction to the classifier boundary, and by small intra class data variance. It is penalized by a factor proportional to error intensity so that greater errors produce greater penalties, allowing discrimination among features with equal error rates.

The final $J(\cdot)$ value is then composed of the mean error rate and the mean reliability parameter along the ten iterations of the cross validation process. A feature is ranked to be better than another if its error rate is lower or if its reliability value is higher, in the case they have the same error rate value. This ranking criterion is used in all phases of the selection process.

IV. Experimental Protocol

In this section, the experimental protocol to evaluate the proposed methods is presented. Three public, two-class, microarray datasets have been analyzed: Leukemia, Lymphoma and Colon. The Leukemia dataset is a collection of gene expression measurements from 72 Leukemia samples (47 acute lymphoblastic Leukemia (ALL) and 25 acute myeloblastic Leukemia (AML)) as reported in [7]. Each sample contains 7129 probe set values, reduced to 3859 after filtering. The Lymphoma dataset is a collection of 96 normal and malignant lymphocyte samples [8]. It contains 42 samples of diffused large B-cell Lymphoma (DLBCL) and 54 samples of other types, each one including 4026 genes. The Colon dataset is a collection of 62 expression measurements from Colon biopsy

TABLE I
CLASSIFICATION ERROR RATES ON THREE DATASETS.

	Colon	Lymphoma	Leukemia
Euclidean: 10 CV	0.00% {5}	0.00% {4}	2.80% {15}
bResub	4.51%	0.94%	5.83%
Treelets: 10 CV	0.00% {6}	0.00% {5}	0.00% {8}
bResub	10.16%	1.66%	3.75%
Original: 10 CV	0.00% {6}	0.00% {5}	0.00% {16}
bResub	6.19%	1.45%	6.86%
NSGAA II [2]	0.03% {12}	0.00% {12}	0.00% {4}
TSP [11]	5.40% {2}	·	10.60% {2}
U, SNR, NMC [11]	12.90%	·	4.80%

samples reported in [9]. It contains 22 normal and 40 colon cancer samples, each one with 2000 gene expression values.

All datasets pass through a preprocessing phase before the metagene creation. The preprocessing consists in applying a base two logarithmic transformation to the original data, followed by a mean removal so that all genes have zero mean across the samples. For the Leukemia dataset, it has been necessary to include a filtering operation before the preprocessing because there are negative values in the original data, making the logarithmic transformation impossible in the real domain. The filtering process is the same as in [2], that led to a feature number reduction from 7129 to 3859 genes. It relies on a threshold operation, in which the minimum value is set to 20 and the maximum to 16000 for all the original data, followed by the exclusion of all genes with a fold change smaller than 5 or a dynamic range smaller than 500.

The feature selection algorithm has been applied on the expanded feature sets produced by *Treelets* and *Euclidean* clustering. It has also been applied on the original gene dataset to evaluate the possible classification benefit from the use of metagenes. The experimental results include the error rates obtained applying the best LDA classifier in each case. Error rate is estimated both with 10-fold cross validation, to have a comparable value with results in the literature, and also with bolstered resubstitution, because this approach has showed to be the best error estimator for LDA classifier in a small sample context like microarray classification [10].

V. RESULTS

The collected experimental results are presented in Table I. Columns in Table I contain the error rates for the classification of the three different datasets. In the first three rows, the obtained error rate estimations with cross validation (10 CV) and bolstered resubstitution (bResub) for the proposed method are presented. In the last rows, results from the literature are included as reference: error rates have all been estimated with cross validation. The number of used features to classify is reported in braces $\{\cdot\}$ when available.

The proposed method is able to reach 0% error rate for each one of the analyzed datasets if 10 CV is utilized, and obtains very low error rates even with bolstered resubstitution estimator, confirming the reliability of the produced classifier. Furthermore, the introduction of metagenes proves beneficial for classification: in all cases, a classifier including metagenes is better than with original gene dataset. Using metagenes, it is possible to reach smaller or equal error rates with fewer features. This is an interesting point, in particular, when the sample number is small, like in the analyzed datasets, because it reduces the risk of overfitting and increases the generalization of the obtained results [6].

Comparing the obtained results with the state of the art, it can be observed how the proposed method produces better classifiers for Lymphoma database and Colon database, while for the Leukemia dataset, the NSGAA II [2] is the best choice. Further studies are necessary to fully understand this set of results on this dataset.

VI. CONCLUSIONS

In this paper, efficient techniques for microarray classification have been studied. The key points of the proposed approach are the feature selection process using IFFS with a two-level ranking criterion, the introduction of the reliability parameter, and the feature set enhancement by metagenes resulting from hierarchical clustering. The feature set expansion with metagenes has proven beneficial for classification since it made possible to obtain smaller or equal error rates with fewer features than using original features only.

This method has obtained the best classifier for two out of three databases when compared to state of the art alternatives with cross validation error estimation.

Analyzing error estimation with bolstered resubstitution, it can be observed how the proposed method still obtains very low error rates, thus underlining the good generalization properties of the obtained classifiers that make the proposed method an interesting technique for microarray classification.

ACKNOWLEDGMENTS

This work has been partially financed by the "Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya" and from Cellex foundation.

REFERENCES

[1] S. Dudoit and J. Fridlyand, "Classification in microarray experiments," *Statistical analysis of gene expression microarray data*, pp. 93–158, 2003.

[2] K. Deb and A. Reddy, "Reliable classification of two-class cancer data using evolutionary algorithms," *BioSystems*, 2003.

[3] S. Nakariyakul and D. Casasent, "An improvement on floating search algorithms for feature subset selection," *Pattern Recogn.*, 2009.

[4] T. Hastie *et al.*, "Supervised harvesting of expression trees," *Genome Biology*, vol. 2, no. 1, 2001.

[5] A. B. Lee, B. Nadler, and L. Wasserman, "Treelets - an adaptive multi-scale basis for sparse unordered data," *Annals of Applied Statistics*, vol. 2, no. 2, pp. 435–471, 2008.

[6] U. Braga-Neto, "Fads and fallacies in the name of small-sample microarray classification - a highlight of misunderstanding and erroneous usage in the applications of genomic signal processing," *Signal Processing Magazine, IEEE*, vol. 24, no. 1, pp. 91 –99, jan. 2007.

[7] T. Golub *et al.*, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, 1999.

[8] A. Alizadeh *et al.*, "Distinct types of diffuse large b-cell lymphoma identified by gene expression," *Nature*, 2000.

[9] U. Alon *et al.*, "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*, 1999.

[10] U. Braga-Neto and E. Dougherty, "Bolstered error estimation," *Pattern Recognition*, vol. 37], no. 6, pp. 1267 – 1281, 2004.

[11] C. Lai *et al.*, "A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets," *BMC Bioinformatics*, vol. 7, no. 1, p. 235, 2006.