Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

An Ensemble Approach for Phenotype Classification Based on Fuzzy Partitioning of Gene Expression Data

A. Dragomir, I. Maraziotis, and A. Bezerianos, *Member, IEEE*

***Abstract*—We focus on developing a pattern recognition method suitable for performing supervised analysis tasks on molecular data resulting from microarray experiments. Molecular characterization of tissue samples using microarray gene expression profiling is expected to uncover fundamental aspects related to cancer diagnosis and drug discovery. There is therefore a need for reliable, accurate classification methods. With this study we propose a framework for constructing an ensemble of individually trained SVM classifiers, each of them specialized on subsets of the input space. The fuzzy approach used for partitioning the data produces overlapping subsets of the input space that facilitates subsequent classification tasks.**

I. Introduction

Advances in microarray technology currently allows effective insights into the cellular mechanism and the nature of complex biological processes. Expression profiling using microarrays produces a molecular picture of a cell's internal state by measuring the expression levels of thousands of genes in a single microarray experiment [1,2]. Performing such experiments on samples of distinct pathogenetic disease type, as well as on samples from healthy tissues, and subsequent computational analysis of the resulting expression data, allows extracting valuable information that can be used for clinical support in tissue diagnosis [3], prediction of outcomes in response to treatment [4], or identification of disease markers [5].

The use of gene expression data analysis as a reliable phenotyping tool has attracted significant research efforts recently [6]. Given the large-scale nature of the data at hand, accurate and highly performing machine learning techniques are required. Developing a diagnostic or prognostic tool from gene expression data involves a supervised learning task, having as goal the identification of an efficient model for predicting the class membership of the data. The learning system (classifier) is given the training data, consisting of patterns chosen from the input data space and their respective class labels. The model derived from a training data set is expected not only to recognize the correct labels of the training data, but also to successfully predict labels of the unseen data (also referred to as test set). When the classification task is dichotomous we deal with binary classification problems, while in the cases when there are at least three classes within the data, we are confronted with a multi-class classification problem. Several classification methods have been applied recently in diagnostic applications based on gene expression data, such as nearest neighbor classifiers, linear discriminant analysis, Bayesian networks, support vector machines (SVM), as well as other machine learning techniques (see [6,7] for extensive reviews).

Recently, there has been increasing interest in the machine learning community regarding the classifiers combination, broadly defined as ensemble methods [8]. The basic idea behind classifier ensembles is the construction of a set of classifiers whose individual decisions are combined in some way such as their committee decision on the class of new data patterns outperforms the accuracy of individual classifiers. A principled explanation for this is given in [8].

Having as task the class prediction of tissue samples, the challenges to be surpassed are either classic classification issues (the curse of dimensionality – referring to the case when the dimension of feature space is much larger than the number of available observations – resulting in a drastic rise in computational complexity and classification errors), or specific problems related to gene expression data analysis (noisy measurements and large variability of data among samples). Since sample diagnosis is a sensitive task, it is crucial that the above issues are accordingly dealt with, in order to obtain optimal classification performance, while confidence in the results should make the analysis suitable for further interpretation by clinicians.

The rationale behind our approach is to split the learning task into multiple less complex tasks by specializing a series of base classifiers (in our case SVMs) on specific subsets of the input data space and subsequently make use of their combined expertise for predicting the class of new data patterns, in a probabilistic decision manner. As a consequence, we split the input data space using a fuzzy clustering approach. The training set is partitioned into several non-disjoint subsets, the fuzzy clustering allowing training patterns to be assigned to multiple clusters with different degrees of membership.

II. Methods

A. Tissue Classification Using Gene Expression Data

In the class prediction context, we can represent the gene

This work was supported by a grant from the General Secretariat for Research and Technology, Ministry of Development of Greece (013/PENED03) to A.B.

A.D., I.M. and A.B. are with the Department of Medical Physics, University of Patras, 26500 Rio, Greece (emails: adragomir@heart.med.upatras.gr, imarazi@heart.med.upatras.gr and bezer@ patreas.upatras.gr).

expression data as a matrix $X \in \mathbb{R}^{n \times m}$, with x_{ij} representing the expression level of gene i in sample j. Gene expression profiles $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})$ have assigned labels y_j in the case when the respective sample belongs to a known diagnostic class. For binary classification tasks $y_j \in \{+1;-1\}$. Based on a training set of p gene profiles $TS = \{(\mathbf{x}_1,y_1), \ldots, (\mathbf{x}_p,y_p)\}$, known to belong to certain classes, the learning algorithm must build a classifier C that is able to predict correct class labels for a new set of expression profiles of unknown class labels (the training set). The classifier must be understood as a discriminant (or decision) function f_d such that $y = f_d(\mathbf{x})$. Therefore, the supervised learning task consists in finding suitable discriminant functions, or their best approximations.

Currently available gene expression data obtained from microarray experiments are of high dimensionality, with only few tens to a hundred experimental samples (corresponding to the m expression profiles, or patterns, in our analysis) and with thousands or tens of thousands of gene expression measurements (the n variables, or features, $n>>m$). The risk of overfitting is extremely high when trying to find a suitable classifying model from such data. Overfitting refers to the case when the model estimated may very accurately fit the samples in the training set, but be very inaccurate in assigning the label of a new sample. Consequently, we have employed SVM, which is known to reduce the risk of overfitting to some extent [9], as our base classifier. By combining several SVMs aggregated in an ensemble scheme, classification indices superior to single SVMs and to other classic methods should be obtained. The supposition is based on results from theory of ensembles, which prove that a combination of individual classifiers, each performing better than average and having negatively associated errors, result in a composite model with improved classification performance [8].

B. Ensemble of classifiers

Our approach for creating the classifier ensemble grows on a method proposed by [10] and is based on the philosophy of manipulating the input data space in such a manner that different classifiers are offered training examples as different as possible, and thus produce highly accurate predictions for certain areas of the input data space. To that goal we perform a fuzzy partitioning, using Fuzzy C-Means clustering [11]. Considering a training set TS as above, new training subsets are generated by dividing it into K overlapping clusters $TS_1, TS_2, \ldots, TS_k$. Clustering is based on computing the Euclidian distance $\|\mathbf{x}_j - \boldsymbol{\mu}_j\|^2$ from the input patterns to cluster center vectors $\boldsymbol{\mu}_j \in \mathbb{R}^n$. Due to the fuzzy partitioning, each $\mathbf{x}_j$ of the training set TS belongs to cluster k (with center $\boldsymbol{\mu}_k$) with a membership degree $u_{jk} \in [0,1]$.

The membership degrees of each expression profile are normalized, such that $\sum_{k=1}^{K} u_{jk} = 1$. The cluster center vectors are arbitrary initialized and subsequently updated, together with the membership functions, using Eq. (1) and (2) bellow:

$$\boldsymbol{\mu}_k = \frac{\sum_{j=1}^{m} u_{jk}^{\alpha} \mathbf{x}_j}{\sum_{j=1}^{m} u_{jk}^{\alpha}} \tag{1}$$

$$u_{jk} = \frac{1}{\sum_{r=1}^{K} \left(\frac{\|\mathbf{x}_j - \boldsymbol{\mu}_k\|^2}{\|\mathbf{x}_j - \boldsymbol{\mu}_r\|^2} \right)^{\frac{2}{\alpha-1}}} \tag{2}$$

where α is a real number greater than 1 (degree of fuzzification). Fuzzy partitioning is carried out in an iterative manner with the task of minimizing the following objective function:

$$J(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k) = \sum_{j=1}^{p} \sum_{k=1}^{K} u_{jk}^{\alpha} \|x_j - \mu_k\|^2 \tag{3}$$

The obtained clusters correspond to the K overlapping training subsets. An input pattern $\mathbf{x}_j$ will be assigned deterministically to the training subset corresponding to the cluster where it has the largest membership value. Subsequently, for each subset k we automatically set a threshold $u_{thr,k}$ (according to the distribution of membership values within the respective cluster) and probabilistically assign the pattern $\mathbf{x}_j$ to each of the remaining training subsets if $u_{jk} > u_{thr,k}$ (with a probability equal to the degree of membership to the respective clusters). Fig.1 bellow contains a schematic representation of the ensemble procedure.

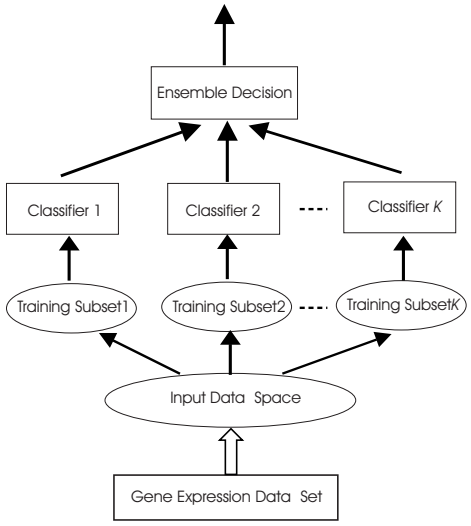

Fig. 1. Schematic representation of the classifier ensemble. Individual classifiers are trained on non-disjoint subsets resulting from fuzzy clustering of the input data space. The final class prediction is performed by aggregating in a probabilistic manner the decisions of the individual classifiers.

C. Base classifiers: SVMs

The choice for the independent base classifiers that will be trained on the resulting subsets and subsequently have their decisions aggregated into an ensemble is the SVM. SVMs find hyperplanes that optimally separate the classes by maximizing the width of the separating band between the data points and the hyperplane. The decision function associated to the retrieved hyperplane will take the form:

$$f_d(\mathbf{x}) = sign\left(\sum_{j=1}^{m} a_j y_j \mathbf{k}(\mathbf{x}_j, \mathbf{x}) + b\right) \quad (4)$$

where scalars a_j and bias b are obtained by solving a quadratic programming problem and the function $\mathbf{k}(\mathbf{x}_j,\mathbf{x})$ is a polynomial or a Gaussian kernel [9].

D. Decisions aggregation

As presented above, the original training data set is partitioned into K non-disjoint subsets and each individual SVM classifier is trained on one of the subsets. The appropriate number of clusters (and thus subsets) in the data set is established following a procedure based on the Xie-Beni validity index [12]. It consists in computing the index value for successively increasing number of clusters. The partition corresponding to the lowest Xie-Beni index is the optimal one. The minimum number of clusters is set to three, in order to ensure better intra-cluster diversity and the maximum is empirically set to ten (further increasing the number of clusters does not produce adequate results, taking into account the low number of data samples). Given a new input pattern $\mathbf{x}_j$, a class label C_{jk} ($k = 1,\ldots,K$) is produced by each SVM$_k$ and the membership degree u_{jk} to the respective cluster k is computed.

The final decision on the class of the new input pattern is taken by combining the decisions of individual SVMs in a probabilistic manner. The probability $P(q \mid \mathbf{x}_j)$ that pattern $\mathbf{x}_j$ belongs to class q is computed, with $q = 1,\ldots, c$ (c is the total number of classes). The class C is selected as final decision using the maximum $P(C \mid \mathbf{x}_j)$, following the Bayes rule. $P(q \mid \mathbf{x}_j)$ is computed as follows:

$$P(q|\mathbf{x}_j) = \sum_{k=1}^{K} u_{jk} I(C_{jk} = q) \quad (5)$$

where $I(C_{jk}=q) = 1$ if $C_{jk}=q$ and $I(C_{jk}=q) = 0$, otherwise. It is obvious from the above that $\sum_{q=1}^{c} P(q \mid \mathbf{x}_j) = 1$ and that the class probability $P(q|\mathbf{x}_j)$ results as the sum of the membership degrees u_{jk} corresponding to classifiers that suggest class q.

III. RESULTS

We assessed the performance of our approach by performing classification experiments on benchmark gene expression data sets. The expression measurements originate from microarray experiments monitoring either tumor/ healthy tissue samples or samples of different tumor subtypes.

TABLE I
COMPARATIVE CLASSIFICATION ERRORS

Dataset	*k*-NN	SVM	Boosting	FuzzyEns
Prostate Cancer	10.66%	7.81%	8.72%	6.33%
Leukemia	3.83%	1.83%	5.67%	1.62%
Breast Cancer	9.64%	7.13%	8.22%	6.11%
Lung Cancer	14.96%	11.21%	13.09%	9.36%

Misclassification rates for our approach denoted as FuzzyEns compared to those of benchmark classification methods on the four public expression data sets. The results are averaged over the 50 random splits into training sets (2/3 of the data) and test sets (1/3 of the data).

Prostate cancer data set. The dataset consists of 102 samples (52 tumor and 50 healthy tissue samples) containing expression levels of 6033 genes derived from the study of prostate tumors, which are among the most heterogeneous of cancers, both histologically and clinically. The experiments were performed by Singh *et al*. [13] and are available at: http://www-genome.wi.mit.edu/mpr/prostate.

Leukemia data set. The dataset consists of 72 microarray experiments containing 7129 genes from cancer patients with two types of leukemia (acute lymphoblastic leukemia – ALL and acute myeloid leukemia – AML). The experiments were performed by Golub *et al.*[14] and are available at: www.genome.wi.mit.edu/MPR.

Breast cancer data set. Van't Veer *et al*. [4] performed microarray experiments on breast cancer patients as well as on patients who remained healthy from the disease, after an initial diagnosis, for an interval of at least 5 years. The dataset consists of 97 samples (46 disease and 51 healthy) containing 24481 gene measurements. Data are available at: http://www.rii.com/publications/2002/vantveer.html.

Lung cancer data set. The dataset consists of 181 samples taken during microarray experiments on patients with malignant pleural mesothelioma (MPM – 31 samples) and adenocarcinoma (ADCA – 150 samples) of the lung. Experiments consist of gene expression measurements for 12533 genes and were performed by Gordon *et al.* [15] and data is available at www.chestsurg.org.

All the data sets had their missing values systematically filled-in using the weighted *K*-nearest neighbors imputation method [16]. Expression profiles were base 10-log transformed and standardized to zero-mean and unit variance, in order to prevent single arrays from dominating the analysis. For all the datasets and all the classification methods used, the data was split into training sets containing 2/3 of the data and test sets consisting of the remaining 1/3. Classification experiments were repeated for 50 times, each time with a different random splitting of the data.

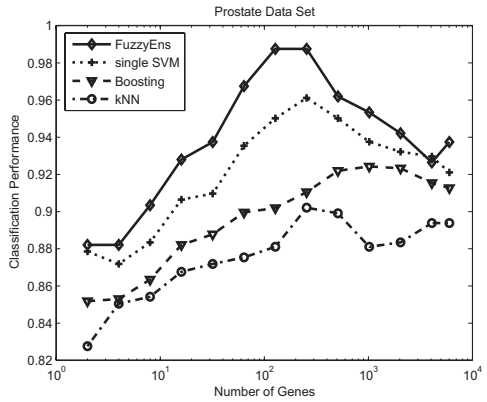

Fig. 2. Comparison between the results obtained using our approach *(FuzzyEns)* and other classification methods on the *Prostate* data set. The plots correspond to subsets of successively reduced number of features (genes).

As benchmark method we have used the *k*-Nearest Neighbor (*k*-NN), single SVMs and Boosting (using as base classifiers decision trees). The classification errors obtained using our approach (denoted as *FuzzyEns*), compared to those of the other methods, for the four data sets are presented in Table 1. As noticed, our ensemble of classifiers yields significantly less classification errors; the only data set for which the performance is only marginally better than that of single SVM being the *Leukemia*. All the methods exhibit lower performance in the case of the *Lung* and *Prostate* data sets, which may be explained by the higher rate of measurement noise produced within the respective experiments, as well as from the heterogeneous nature of the data (data represents a collection of several independent experiments).

In a separate experiment we attempted to evaluate the influence of feature selection (dimensionality reduction) on the classification performance of our approach. In feature selection, subsets of genes that are relevant for the phenotype distinction are identified. This results not only in improved classification performance (since we remove the genes that have lower influence, and thus may represent 'noise' for our task) but is also relevant from the biological point of view. Identifying genes that have high discriminant influence, and making them amenable to further biological study may help investigating the causes of the disease. We perform gene selection following the methodology in [14]. A ranking measure for each gene is computed, based on its class correlation, and genes with lowest rankings are successively removed. At each step, half of the genes are removed, yielding data of dimensionality 6033, 4096, 2048, 1024, … 8, 4 and 2 genes, respectively. Comparative results for all four classification methods under scrutiny are presented in Fig. 2, for the *Prostate* data set. The plots allow us to draw the conclusion that the optimal classification rates are obtained on subsets of 128 or 256 genes. As it can be noticed *k*-NN exhibits the lowest performance amongst the methods tested, while our fuzzy created ensembles repeatedly produce among the highest classification rates.

I. Conclusion

The paper proposes a method for improving the classification accuracy in the case of tissue diagnosis based on gene expression data. The approach first divides the original training set into overlapping subsets by fuzzy clustering. Subsequently, independent classifiers are trained on the respective subsets, creating thus an ensemble of classifiers that act in a *divide-and-conquer* manner on the input data space. The classification decision in the case of a new pattern presented to the system is taken by appropriately combining the decisions of the classifiers in a probabilistic manner. Although the present study refers only to the binary classification problems, a framework for multi-class problems can be easily imagined, by employing multi-class learners and by adapting the gene selection procedure as in [7] (one over the rest scenario).

References

[1] M.B. Eisen, P.T. Spellman, O.B. Patrick, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns", *Proc. Natl. Acad. Science*, vol. 95, pp. 14863-4868, 1998.

[2] A.W.-C. Liew, H. Yan, and M. Yang, "Pattern recognition techniques for the emerging field of bioinformatics", *Pattern Recognition,* vol. 38, pp. 2055—2073, 2005.

[3] U. Alon, N. Barkai, D.A. Notterman, et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", *Proc. Natl. Acad. Science*, vol. 96, pp. 6745-67504868, 1999.

[4] L.J Van't Veer, H. Dai, M.J. Van de Vijer, et al., "Gene expression profiling predicts clinical outcome of breasts cancer", *Nature*, vol. 415, pp. 530-536, 2002.

[5] M. Schummer, W.V. Ng, R.E. Bumgarner, et al., "Comparative hybridization of an array of 21500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas", *Gene*, vol. 238, pp. 375-385, 1999.

[6] R.R. Brentani, D.M. Carraro, S. Verjovski-Almeida, et al., "Gene expression arrays in cancer research: methods and applications", *Critical Reviews in Oncology/Hematology*, vol. 54, pp. 95-105, 2005.

[7] A. Statnikov, C.F. Aliferis, I. Tsamardinos, et al., "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis", *Bioinformatics*, vol. 21, pp. 631-643, 2005.

[8] T.G. Dietterich, "Ensemble Methods in machine learning," *Lecture Notes in Computer Science,* vol. 1857, pp. 1–15, 2000.

[9] VN. Vapnik, *Statistical Learning Theory.* New York: Wiley Interscience, 1998.

[10] L. Nanni, A. Lumini, "FuzzyBagging: A novel ensemble of classifiers", *Pattern Recognition*, vol. 39, pp. 488-490, (2006).

[11] J.C. Bezdek, *Pattern Recognition with fuzzy objective function algorithms*. New York: Plenum Press, 1981.

[12] N.R. Pal, J.C. Bezdek, "On cluster validity for the Fuzzy C-Means model", *IEEE Trans. OnFuzzy System*s, vol. 3, pp. 370-379, 1995.

[13] D. Singh, P. Febbo, K. Ross et al., "Gene expression correlates of clinical prostate cancer behavior", *Cancer Cell*, vol. 1, pp. 203-209, 2002.

[14] T.R. Golub, D.K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", *Science*, vol. 286, pp. 531-537, 1999.

[15] G.J. Gordon, R.V. Jensen, L.-L. Hsiao, et al., "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma", *Cancer Research*, vol. 62, pp. 4963-4967, 2002.

[16] O. Troyanskaya, M. Cantor, G. Sherlock et al., "Missing value estimation methods for DNA microarrays", *Bioinformatics*, vol. 17 pp. 520-525, 2001.