Proceedings of the 9th International Conference on Information Technology and Applications in Biomedicine, ITAB 2009, Larnaca, Cyprus, 5-7 November 2009

Gene Expression Classifiers and Out-Of-Class Samples Detection

Alfredo Benso, Stefano Di Carlo, and Gianfranco Politano

Abstract— **The proper application of statistics, machine learning, and data-mining techniques in routine clinical diagnostics to classify diseases using their genetic expression profile is still a challenge. One critical issue is the overall inability of most state-of-the-art classifiers to identify out-of-class samples, i.e., samples that do not belong to any of the available classes. This paper shows a possible explanation for this problem and suggests how, by analyzing the distribution of the class probability estimates generated by a classifier, it is possible to build decision rules able to significantly improve its performances.**

Index Terms— **gene expression, classification, clinical diagnostics.**

I. Introduction

Scientists are using DNA microarrays to investigate several phenomena, including the response of genes to a disease or to a particular treatment [1]. Nevertheless, using DNA microarrays to classify diseases at a genetic level is still a challenging research problem.

In general, *classification* is used to group individual items according to structural or quantitative characteristics derived from a training set of previously labeled items. In the case of microarrays, it involves assessing gene expression levels from different experiments, determining genes whose expression is relevant, and then applying a rule to group experiments that show a similar gene expression profile (*classification*) [2]. Designing accurate classifiers for microarray data poses several challenges. Most of them, like the *"small N, large P"* problem of statistical learning where the number P of variables (gene expressions) is typically much larger than the number N of available samples, impact major aspects of the classifier design: the classification rule, the error estimation, and the feature selection.

In a realistic application for medical diagnostics, the classification of a gene expression profile would have to be performed by comparing the profile against a large set of classes, each identifying a possible pathology. An optimal classification algorithm should also be able to recognize samples that do not belong to any of the available classes (out-of-class), either because the pathology was not considered in the training set used to build the classifier, or because the sample is of a healthy specimen. Unfortunately, the overall inability of most classifiers to recognize out-of-class samples, makes their application to clinical diagnostic scenarios very critical. The main problem lies in the algorithms themselves, since, even in the case of out-of-class samples, most classifiers generate class probability estimates, i.e., a measure of the probability of a sample to belong to a certain class, very similar to the ones computed for correctly matched profiles. Therefore, neither the absolute or the relative value of the class probability estimates can be used to discriminate between classifiable and out-of-class samples.

This paper shows how this problem is closely related to the distribution of the probability estimates computed by the classifier, and proposes a methodology to design classifier-dependent decision rules able to significantly increase the performance of some state-of-the-art classification algorithms. The paper presents an experimental comparison between several classification algorithms on a set of microarray experiments for fifteen well known and documented diseases. Experimental results show that for some classifiers the application of decision rules based on the class probability estimates distribution strongly improves their performance.

The paper is organized as follows: Section II overviews the basic steps required to perform a classification process, while Section III introduces the experimental setup. Section IV analyzes the class probability estimates distribution of the considered classifiers, and Section V proposes a methodology to build decision rules for microarray classifiers based on the class probability estimate distribution. Section VI shows the result of the application of these decision rules in our experimental conditions, and finally Section VII summarizes the main contributions of the paper and outlines possible future activities.

II. The Classification Process

Building prediction algorithms for classifying diseases based on gene expression profiles involves several steps including *signal pre-processing* of raw microarray scans, *data modeling*, *prediction* (e.g., classification), and *validation*.

The signal pre-processing stage elaborates the raw image obtained from the scanning process of a microarray (*sample*), and calculates the *expression level* of each DNA probe placed on it. The result of this stage is a *gene expression profile* of the sample, where each gene is associated with its expression level(s). In very general terms, at this point the classifier (or predictor) takes the gene expression profile of the sample and compares it against the profiles of a set of available classes, each representing a different disease. This step creates a proximity vector where each element is associated to one of the available classes. Each element indicates, directly or indirectly, the *Class Probability Estimate* (CPE) of the considered sample to belong to the corresponding class. Finally, based on the proximity vector, the classifier uses a *decision rule* to predict the target class. Classifiers usually

A. Benso, S. Di Carlo, and G. Politano are with the Department of Control and Computer Engineering of Politecnico di Torino, Corso Duca degli Abruzzi 24 -10129 Torino Italy. Emails: {alfredo.benso, stefano.dicarlo, gianfranco.politano}@polito.it

978-1-4244-5379-5/09/$26.00 ©2009 IEEE

use the "maximum proximity" rule, i.e., the class with highest CPE becomes the predicted class.

State-of-the-art classifiers adopting this rule are really effective when working on samples that are known to belong to one of the available classes (classifiable samples). Unfortunately, a classifier suitable for real diagnostic applications should be able to discriminate two possible situations: (i) the sample actually belongs to one of the classes the classifier has been trained for, or (ii) the sample does not belong to any class because it is either a healthy sample or a sample showing a disease not considered when the classifier was trained. Failing in distinguishing between these two situations creates a very high rate of False Positives. This dramatically affects the classifier's *specificity,* i.e., the proportion of actual negatives which are correctly identified as such, and its usability in a real scenario, where both False Positives and False Negatives have to be lowered as much as possible, if not completely removed. In this paper we are not interested in evaluating the classifiers' performance, i.e., their capability of correctly handling classifiable samples. Instead, we consider a binary classification test designed to test the ability of a classifier to discriminate between classifiable and out-of-class samples. The following definitions are used in the remaining of the paper:

- *True Positives:* are samples of class X classified in one of the available classes. This includes both samples classified in the correct class (*matches*), and samples classified in the wrong class (*mismatches*),
- *True Negatives* are out-of-class samples correctly classified as out-of-class,
- *False Positives* are out-of-class samples erroneously classified in one of the available classes, and
- *False Negatives* are samples of class X classified as out-of-class.

III. Experimental Design

The experimental setup used throughout this paper involves a number of classification experiments on a large set of microarrays performed using a collection of widely used classification algorithms. The set of considered microarrays comes from the cDNA Stanford Microarray database [3]. To provide a good diversity to the class and sample space, the data-set comprises two sub-sets, the first related to similar blood diseases, and the second related to completely different pathologies. This choice allows us to better test the characteristics of the classifiers in dealing with both very similar and very different classes.

A total of 15 pathologies is considered in this study: Diffuse Large B-Cell Lymphoma (DLBCL), Lymphocytic Leukemia Watch&Wait (CLLww), Lymphocytic Leukemia (CLL), Acute Lymphoblastic Leukemia (ALL), Core Binding Factor Acute Myeloid Leukemia (CBF-AML), Breast Cancer (BC), Cutaneous B-Cell Lymphomas (CBCL), Follicular Lymphoma (FL), Healthy Blood (HB), Hematopoietic Lymphoma (HL), Normal Lymphoid subset (NL), Solid Ovarian tumor (SOT), Solid Brain tumor (SBT), Solid Lung tumor (SLT), and Acute Myeloid Leukemia (AML). Each pathology comprises 10 to 60 samples (clinical cases) run on microarrays ranging from 9k to 45k genes (variables). From the first 9 pathologies an unfolded subset of about 10 samples per pathology has been used as training set for the classifiers, while the remaining data represent the set of classifiable samples. Data from the remaining 6 pathologies are used as out-of-class samples.

The classification algorithms used in the experiments have been implemented as an *R* script [4]. The considered classifiers are: k–Nearest Neighbors (KNN) [5], Neural Networks (NNET) [6], Linear Discriminant Analysis (LDA) [7], Partial Least Square (PLS) [8], Support Vector Machines (SVM)[9], Random Forests (RF) [10], and Differential Gene Expression Graphs (DGEG) [11], [12]. For all classifiers besides PLS and DGEG, PCA has been used to perform variable reduction. The optimization parameters of each classifiers, as well as the transformation of computed scores into class probability estimations, is performed by the *R CARET* package (Classification And REgression Training) as described in [13].

IV. Class Probability Estimates Distribution

Most of the classifiers considered in this paper (see Section III) are unable to correctly deal with out-of-class samples. The main reason is that the CPE computed for an out-of-class sample is, in most of the cases, indistinguishable from the one computed for a classifiable sample. For example, CPEs obtained in our experiments by a Support Vector Machine are of 0.1604 for a True Positive, 0.1604 for a False Negative, and 0.1849 for a False Positive.

Starting from this observation, we decided to analyze, for each classifier, the CPEs distribution for all samples in the considered data-set.

Figure 1 reports two plots for each classifier. MAX shows the distribution of the highest CPE of the proximity vector, i.e., the one corresponding to the predicted class, for all classifiable samples (True Positives - solid line) and all out-of-class samples (False Positives - dotted line), while DIFF shows the difference between the values of the highest two CPEs for each sample, again for both classifiable and out-of-class samples. Looking at the MAX plots, the first very interesting observation is that, in most classifiers (SVM, PLS, LDA, KNN), the absolute value of the CPE cannot be used to discriminate between True Positives, and False Positives since their distribution is in fact overlapping. Instead, for three of the classifiers (RF, DGEG, and to a certain extent NNET), the distributions show two distinct peaks. In this case, the absolute value of the CPE could be exploited to identify a large number of False Positives and therefore to increase the specificity of the classifier.

Another interesting consideration comes from the DIFF plot. In this case RF shows a very clear distinction between True and False Positives. True Positives (solid line) have a max around 0.8, far from the False Positives (dotted line) max which falls around 0.1. This means that, for True Positives, the difference between the top rated class and the second rated class is very high (around 0.8 in most of the cases). Instead, False Positives show a very low difference between the CPEs of the two top ranked classes, revealing the inability of the classifier to make a clear decision. Again, in classifiers that show this property, this observation could be used to discriminate between True and False Positives. KNN, NNET, and DGEG show less differences between the two distributions, but partial discrimination is still possible. Similar considerations are instead impossible for the other classifiers (SVM, PLS, LDA).

V. Decision Rule and Diagnosis

The analysis performed in Section IV clearly highlights how the simple "maximum proximity" rule, i.e., selecting the class with highest CPE, used by most state-of-the-art classifiers does not allow to consider all conditions that usually arise when performing predictions for diagnostic purposes. We therefore propose a methodology for building, whenever possible, a set of decision rules allowing a more detailed and precise understanding and use of CPEs.

The CPE space can be partitioned into three distinct areas: (i) maximum probability area, (ii) decision area, and (iii) out-of-class area, delimited by two thresholds T_{MAX} and T_{OOC} ($T_{MAX} > T_{OOC}$), specifically defined, wherever possible, for each classifiers.

Based on this partitioning, the following decision rules can be applied:

- *R1 (maximize true negatives)*: to predict a class for a sample, at least one class should exhibit a CPE higher than T_{OOC}. If all CPEs are lower than T_{OOC} (out-of-class area), the sample is considered out-of-class;
- *R2 (maximize true positives)*: if at least one class shows a Proximity Estimate higher than T_{MAX} (maximum probability area), the class with the maximum proximity score is predicted. This gives maximum confidence to predictions with high score;
- *R3:* in case neither R1 or R2 are satisfied, then at least one CPE falls between the two thresholds T_{MAX} and T_{OOC} (decision area). In this case if the two top ranked CPEs differ of at least a minimum value T_{diff}, considered as the minimum difference to discriminate between the two top predictions, the class with the maximum CPE is selected. Otherwise, if the second ranked CPE falls in the out-of-class area, the sample is classified as out-of-class. This rule avoids to provide a result if the distinction between two classes is not sufficient to take a clear decision;
- *R4*: whenever the first three rules cannot be applied, the prediction is considered *uncertain*, and the classifier does not produce any classification result. In alternative, multiple classification results can be also provided here to alert the user that the confidence in the prediction is low.

The proposed decision rules try to imitate a human cognitive process to identify the correct classification. They take into consideration not only the absolute value of the CPEs, but also their relative values. This property is very useful for clinical diagnostics. In fact, the classifier is not only able

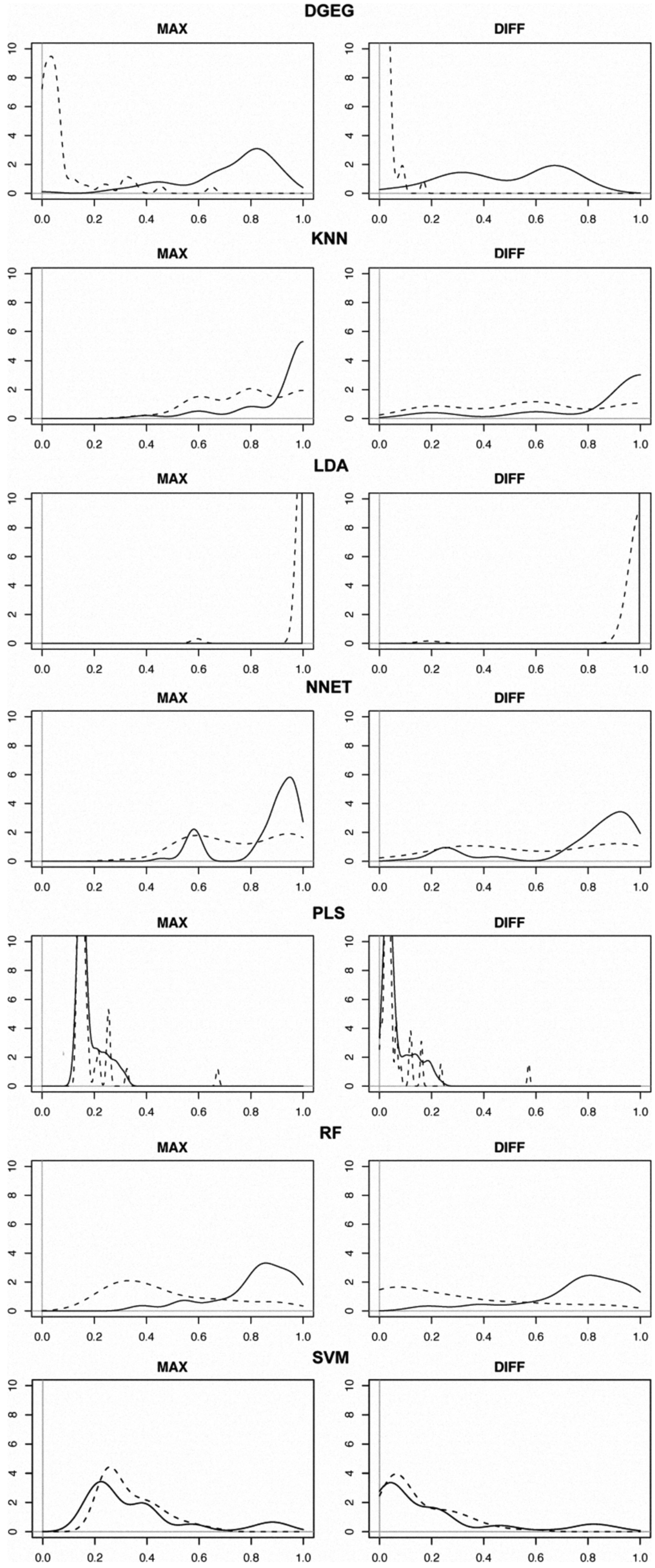

Fig. 1. CPE distributions for selected classifiers considering: (MAX) the highest CPE of the proximity vector of each sample, and (DIFF) the difference between the values of the highest two CPEs for each sample.

to recognize out-of-class samples, but, looking at the CPEs it can also provide important diagnostic information such as the identification of genetic similarities with known diseases.

The definition of the three thresholds can be done looking at the CPE distributions:

- T_{MAX}: if the MAX plot shows a clear separation between True and the False Positives distributions, T_{MAX} can be placed in such a way to have most True Positives immediately detected by R2. T_{MAX} defines the maximum probability area. Looking at the MAX plot of the RF classifier (Figure 2), a good choice for T_{MAX} is between 0.6 and 0.8. The more the threshold is placed near 1.0 the less False Negatives will appear, but also less True Positives will be detected using the maximum probability rule (R2);
- T_{OOC}: similarly to T_{MAX}, looking at the MAX plot T_{OOC} can be defined in order to correctly identify False Positives using rule R1, i.e., CPE lower than the threshold. From the MAX plot of the RF classifier (Figure 2), it is clear that a good choice falls between 0.2 and 0.5;
- T_{diff}: for all samples that fall between T_{OOC} and T_{MAX}, i.e., in the decision area, the DIFF graph can be used to define T_{diff}. A good heuristic is to consider the point where the two curves intersect. Again in the DIFF plot of the RF classifier it is obvious that in general, and therefore also in the decision area, the two top classes show very close CPEs only in the case of False Positives (dotted line). In the case of True Positives, the distinction is much higher (between 0.7 and 1.0). A threshold of about 0.6 will maximize, in this case, the number of True Positives, moving several samples from False Positives to True Negatives.

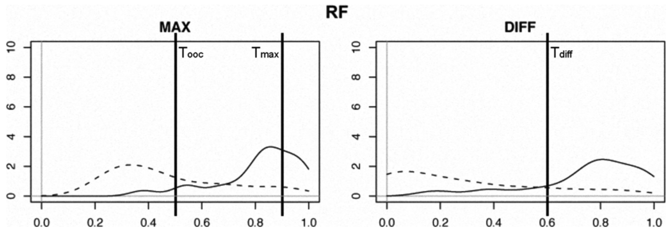

Fig. 2. Thresholds definition for the RF classifier

It is important to point out that, especially when the CPE distributions of True and False Positives are not clearly separated, the choice of the thresholds always involves a trade-off between increasing the sensitivity, and lowering the specificity of the classifier, i.e., the proportion of negatives which are correctly identified as such. An automated algorithm to choose the best thresholds is under development.

VI. Experimental Results

The proposed decision rules have been tested using the data-set presented in Section III, and all the experiments showed promising results in the rules capability of increasing both the performance and the reliability of selected classifiers.

Table I reports the results in terms of True Positives (TP), True negatives (TN), False Positives (FP), and False Negatives (FN) using both the maximum proximity rule, and the new rules defined in Section V. In the table M indicates mismatches and allows to measure the performance of the classifiers on classifiable samples, while U indicates samples classified as uncertain. A combination of two additional measures has also been used to evaluate the effect of the decision rules on the classifiers. The *sensitivity* measures True Positives, i.e. the proportion of classifiable samples that are correctly identified as such (e.g., the percentage of sick people who are identified as having a disease), while the *specificity* measures True Negatives, i.e. the proportion of out-of-class samples that are correctly identified (e.g., the percentage of healthy people who are identified as not having any disease). In a perfect classifier they should be always equal to 1.

The new decision rules have only been applied to DGEG, RF, NNET, and KNN. For the other classifiers, their application is not possible since the MAX and DIFF plots of Figure 1 do not allow a clear separation of True and False Positives.

It is interesting to note that, as expected, the application of the proposed rules allows a very significant improvement in the classifiers specificity, since several False Positive outcomes became True Negatives. Nevertheless, the KNN and NNET classifiers show a slight higher reduction of sensitivity (consequence of a decreased number of True Positives) not compensated by the same increasing of specificity as for the other two classifiers. In fact the shape of their MAX and DIFF plots do not allow the definition of optimal thresholds as it is possible for the DGEG and RF classifiers that show a well defined separation of the two distributions in the MAX and DIFF plots (Figure 1). Table I also shows that the rule does not reduce the performance of the classifiers on classifiable samples, i.e., the number of mismatches M remains the same. Moreover the number of samples classified as uncertain is really low thus having a minor impact on the classification result.

To conclude, Table I reports the different thresholds defined for each classifier based on the corresponding Probability Estimates distribution.

VII. Conclusions & Future Works

This paper analyzes the ability of state-of-the-art classifiers of identifying out-of-class samples, i.e., samples not belonging to any of the classes the classifier has been trained for. The main contributions of the paper stem in a possible explanation of the source of this this problem, and in the definition of a set of new decision rules able to significantly improve some classifiers' performances. Future activity on this topic, mainly focuses in the study of efficient heuristics to define the different thresholds that represent the key point for the correct behavior of the proposed decision rules. This work is also part of a wider effort to increase

TABLE I
RESULTS COMPARISON

	Maximum proximity						New Rules						Thresholds	Difference	
	TP	TN	FN	FP	SE	SP	TP	TN	FN	FP	SE	SP	$T_{OOC}, T_{MAX}, T_{diff}$	△ SE	△ SP
DGEG	100%(M:1)	**0**	0	**100%**	1	**0**	94% (M:1)	95%	6%	5% (U:1)	0.95	**0.95**	(0.3,0.8,0.15)	-5%	+95%
RF	100%	**0**	0	**100%**	1	**0**	83%	85%	17%	15%	0.83	**0.85**	(0.5,0.9 0.6)	-17%	+85%
KNN	100%	**0**	0	**100%**	1	**0**	77%	66%	23%	34%	0.77	**0.66**	(0.6,0.8,0.7)	-23%	+66%
NNET	100%(M:1)	**0**	0	**100%**	1	**0**	80% (M:1)	43%	20%	57%	0.80	**0.43**	(0.5,0.7,0.6)	-20%	+43%

the reliability of DNA microarray classifiers used in real diagnostic applications.

REFERENCES

[1] G. Gibson, "Microarray analysis," *PLoS Biology*, vol. 1, no. 1, pp. 28–29, Oct. 2003.

[2] E. R. Dougherty, "The fundamental role of pattern recognition for gene-expression/microarray data in bioinformatics," *Pattern Recognition*, vol. 38, no. 12, pp. 2226–2228, Dec 2005.

[3] cdna stanford's microarray database. [Online]. Available: http://genome-www.stanford.edu/

[4] The r project for statistical computing: http://www.r-project.org/.

[5] S. Buttrey and C. Karo, "Using k-nearest-neighbor classification in the leaves of a tree," *Computational Statistics and Data Analysis*, no. 40, pp. 27–37, 2002.

[6] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York: Oxford Univ. Press, 1995.

[7] V. Roth and T. Lange, "Bayesian class discovery in microarray datasets," *IEEE Trans Biomed Eng*, vol. 51, pp. 707–718, 2004.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification, 2nd ed.* New York: Wiley, 200.

[9] J. Weston and C. Watkins, "Support vector machines for multi-class pattern recognition," *In Proceedings of the Seventh European Symposium on Artificial Neural Networks*, 1999.

[10] B. L., "Random forests," *Machine Learning*, vol. 1, no. 45, pp. 5–32, 2001.

[11] A. Benso, S. Di Carlo, S. Politano, and L. Sterpone, "A graph-based representation of gene expression profiles in dna microarrays," in *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, Sept. 2008.

[12] A. Benso, S. Di Carlo, G. Politano, and L. Sterpone, "Differential gene expression graphs: A data structure for classification in dna microarrays," *BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on*, pp. 1–6, Oct. 2008.

[13] M. Kuhn, "Building predictive models in r using the caret package," *Journal of Statistical Software*, vol. 28, no. 5, pp. 1–26, August 2008.