Ensemble learning and hierarchical data representation for microarray classification

Mattia Bosio, Pau Bellot, Philippe Salembier, Albert Oliveras Vergés

Abstract— **The microarray data classification is an open and active research field. The development of more accurate algorithms is of great interest and many of the developed techniques can be straightforwardly applied in analyzing different kinds of omics data. In this work, an ensemble learning algorithm is applied within a classification framework that already got good predictive results. Ensemble techniques take individual experts, (i.e. classifiers), to combine them to improve the individual expert results with a voting scheme. In this case, a thinning algorithm is proposed which starts by using all the available experts and removes them one by one focusing on improving the ensemble vote. Two versions of a state of the art ensemble thinning algorithm have been tested and three key elements have been introduced to work with microarray data: the ensemble cohort definition, the nonexpert notion, which defines a set of excluded expert from the thinning process, and a rule to break ties in the thinning process. Experiments have been done on seven public datasets from the Microarray Quality Control study, MAQC. The proposed key elements have shown to be useful for the prediction performance and the studied ensemble technique shown to improve the state of the art results by producing classifiers with better predictions.**

I. Introduction

Microarrays are a powerful and consolidated technology in the biomedical research, offering the ability to simultaneously measure thousands of gene expression values, thereby providing a significant amount of multivariate data with which it is possible to produce classifiers. The typical microarray analysis setting constitutes an extreme case of high-dimensionality as there is a very large number of available features with respect to the sample number. In such circumstances, it is of primal importance the application of a feature selection algorithm, as stated for example in [1], [2].

Ensemble learning combines multiple learning algorithms, called experts, to improve the overall prediction accuracy [3]. A plethora of ensemble methods has been developed to analyze biological data and there exist many alternatives reviewed for example in [3], [4]. They became popular because they allow to improve the classification by aggregating multiple experts to make decision over unseen data in a consensus way. In order to effectively improve the ensemble performances the experts should be accurate, (i.e. better than random), and diverse from each other [3].

An approach to ensemble learning called overproduce and select is described in [4] as a method to obtain good ensemble learners. It consists in producing a big set of experts and then select a subset which will be used for classification via majority voting. Several criteria of expert selection algorithms are studied and compared in [4]. Among the considered algorithms, the one called *Accuracy in diversity*, AID, [5] was able to reach the best prediction accuracy.

In this work we decided to implement two versions of the AID algorithm from [5]. One is the original AID implementation and the other is a simplified version from Kuncheva's book [4], that will be named Kun. To produce a huge and diverse set of experts, we decided to use the microarray dat feature overabundance. For each one of the available features a Linear Discriminant Analysis classifier, LDA, is built and used as an expert. The available feature set is not only composed by the genes, but also by a new set of features built as in [6] with a hierarchical clustering process. The output is a hierarchical binary tree with new features called metagenes, one for each node. These metagenes showed to be helpful in prediction by reducing the noise and merging related genes [6]. In this work we consider as features both the original gene expressions and the newly generated metagenes.

The proposed modifications to the original algorithm take into account the microarray characteristics of small sample size and big feature number. In addition to including metagenes as experts, the notion of *nonexperts* that represent a set of experts excluded from the thinning process due to they poor properties has been introduced as well as a rule to break ties in the thinning process.

The prediction ability of the ensemble algorithm is evaluated on seven datasets, available at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16716 which are from the *Micro Array Quality Control study phase II* (MAQC) [7]. Results from this work are compared to results obtained in MAQC study and from [6] following the same evaluation procedure. More than 30.000 models were built in MAQC [7] using many combination of analytical methods and classifiers, from random forests to boosting and from knn classifiers to SVM or neural networks.

This paper is organized as follows: In Section II the ensemble generation procedure is presented by describing the original *AID* algorithm [5] and the simplified Kun version [4]. Moreover the adaptations for the microarray case are detailed. In Section III, the experimental protocol is described. The classification results are presented in Section IV, compared to state of the art alternatives. A discussion about the usefulness and efficiency of the proposed method

This work has been partially financed by "Fundació privada CELLEX"; and the "Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya".

Department of Signal Theory and Communications, Technical University of Catalonia UPC,Campus Diagonal Nord, building D4 Jordi Girona 1-3 08034 Barcelona mattia.bosio@upc.edu

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

```
Samples $S = s_1 \dots s_n$
Experts $E = E_1 \dots E_p$

while $\#E > 1$
    Calculate $S_d = \{s_i\} : 0.1 \leq f(s_i) \leq 0.9$
    where $f(s_i)$ fraction of experts in the ensemble correctly classifying $i^{th}$ sample.
    Calculate $d = \frac{\#S_d}{n}$
    Lower Bound $l_b = \mu \cdot d + \frac{1-d}{n}$
    Upper Bound $U_b = \alpha \cdot d + \mu(1-d)$
    Define the set of relevant samples.
    $S_p = \{s_i\} : l_b \leq f(s_i) \leq U_b$
    $E_i$ = expert with lowest accuracy over the $S_p$ set.
    $E := E - E_i$
    Remove $E_i$ from $E$
end

$\mu$ = Mean experts accuracy
$\alpha = 0.9$
```

Fig. 1. Pseudocode for the AID algorithm.

is presented in Section V.

II. Ensemble selection

In this section, both the AID and *Kun* algorithms are described as well as the proposed modifications to work with microarray classification.

A. Accuracy in diversity

The principle on which the AID algorithm is based is to include the most diverse and accurate classifiers by eliminating classifiers that are most often incorrect on examples that are misclassified by many experts. A pseudo code for the AID algorithm is shown in Figure 1. It is an iterative process in which, for each iteration, one expert is removed from the ensemble. For each iteration we consider to have a set of n samples and p experts [5]. To determine which expert E_i must be removed, some elements are calculated. The first one is an ensemble diversity measure called Percentage Correct Diversity Measure d [5], which is the the percentage of samples correctly classified by a percentage of individual experts between 10 and 90 %. The d measure is then combined with other parameters like the average experts accuracy μ and an α value to define two boundaries as in Figure 1 which are used to identify a set of relevant point S_p. The S_p set is composed of all samples which are correctly classified by a percentage of experts between the two calculated boundaries. Finally, the expert E_i to be removed from the ensemble is the one with lowest accuracy on the S_p set.

The rationale behind this is that the samples in S_p are those on which the ensemble is most uncertain, thus are those for which the elimination of an expert can be more relevant because it can change the ensemble majority voting. Therefore, excluding the expert that more poorly performs on these samples affects more positively the ensemble accuracy than simply excluding the expert with overall lowest accuracy. Since the ensemble changes throughout the iterations, the d value changes, as well as the boundaries, thus meaning that the set of relevant samples adapts to the ensemble changing characteristics.

In [5] it is stated how the adaptive boundaries to define the S_p set are defined by considering the known relationship between the experts mean accuracy and the ensemble diversity [4]. On the other side, in [4] it is remarked how the AID algorithm could have equivalent performances with fixed boundary values, suggesting to use the ones in the calculation of the d measure: 10% and 90%. Since we could not find any works comparing the two alternatives, we chose to apply both and keep the one with better performances.

B. Microarray adaptations for thinning

Considering the microarray data characteristics we propose some key points to obtain a good ensemble system:

*a) **Experts cohort**:* We chose to build thousands of experts by defining each expert as an LDA classifier trained on a different feature. Both genes and metagenes, obtained with the algorithm from [6] are considered as individual features since metagenes helped in finding better classifier than with genes only.

*b) **Nonexperts**:* We introduce the notion of nonexpert to remove a whole set of "experts" with poor training characteristics. We decided to exclude from the thinning process all those experts that classify all the training sample with the same label. Considering that the expert is unable to distinguish two classes, it is not considered as a useful ensemble component. The nonexpert number can vary depending on the data type and it increases when the class distribution is highly skewed. Furthermore, the idea of nonexpert responds to the microarray data characteristic of feature overabundance: the majority of the available features are useless for prediction purposes since they are not related to the classified phenomenon. Thus, we included this simple criterion in the thinning process.

*c) **Tie breaking**:* Considering the typical case of small sample number for microarrays and considering that the S_p sample is smaller or equal to the whole training sample number, there is a relevant probability to have ties when comparing experts accuracies. To reduce this problem and introduce a rule, it has been chosen to consider the metagene generation process. When ties occur, the excluded expert is the one which has been generated at a higher level in the hierarchical tree, so that metagenes composed of many sons with low similarity will be eliminated instead than another metagene with more correlated components. This because it is more likely that a metagene with more correlated sons will replicate its behavior than another one merging many different individual genes. Finally, the ties between individual genes are randomly resolved since they all are on the same level of the hierarchical tree.

The usefulness of these three elements is assessed by experiments comparing the complete algorithm with three

TABLE I

MCC results comparing the studied AID and Kunalgorithms. Each line is a different algorithm comparing AID, Kun and its variants, each one excluding one different key-element to adapt the ensemble algorithm for microarray data.

	A	C	D	E	F	G	H	MEAN
AID	0.293	0.793	0.459	0.789	0.221	0.231	0.813	0.514
Kun	0.407	0.812	0.459	0.789	0.221	0.236	0.828	0.533
Kun_{tie}	0.303	0.804	0.451	0.789	0.221	0.236	0.828	0.519
Kun_{genes}	0.346	0.781	0.366	0.773	—	0.313	0.817	0.485
Kun_{all}	—	0.792	—	0.789	—	—	0.031	0.230

modified algorithms, each of which does not use one of the proposed key elements.

III. Experimental Protocol

A. Datasets

The analyzed data are a subset of the provided datasets by the MAQC II consortium, for more information refer to [7]. In this work, seven datasets have been used, named A, C to H [7], available at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16716. These data have been chosen because they are highly reliable, selected after a quality control process in order to provide a common test ground and because for each dataset both a training set and an independent validation set are provided [7]. Furthermore, many different laboratories have tested their algorithms on the same datasets with the same evaluation protocol (i.e. train the classifiers on the training set with performance assessment on the validation dataset) and published their final outcomes [7], [8], [9], [6]. Thus an accurate benchmark can be performed to understand how well does a proposed algorithm perform with respect to a large number of state of the art alternatives.

B. Experiments

The first experiment evaluates whether the original AID algorithm [5] or the simplified version in [4] has better performances. They will be identified by *AID* and *Kun* respectively. Both the algorithms are trained on the seven datasets. For each dataset they produce thousands of nested ensembles, one for each iteration. These ensembles are then applied on independent validation datasets and the best performing ensemble is taken as representative of the predictive potential of the algorithm as in [9], [6]. In order to avoid voting artifacts, only ensembles with an odd number of experts are considered.

The chosen performance metric is the Matthews Correlation Coefficient (MCC) [10], since, as stated in [7] it is informative when the distribution of the two classes is highly skewed, it is simple to calculate and available for all models with which the proposed method has been compared to. MCC values range from -1 (i.e. perfect inverse prediction) to 1 (perfect prediction).

The second experiment has the same setup as the first one, but it evaluates the usefulness of the introduced elements in II-B: the nonexpert notation, the metagene inclusion and the tie breaking rule. Three algorithms are compared to the original one. Each one applies two of the three elements and are identified, for the *Kun* algorithm by:

- Kun_{all} : This algorithm does not exclude the nonexperts from the thinning process.
- Kun_{genes} : This algorithm excludes the nonexperts but it does not use any metagene.
- Kun_{tie} : This algorithm resolves each tie without considering the tree structure, thus eliminating the first expert it encounters with lowest accuracy on S_p set.

Finally, the best performing algorithm is compared to state of the art alternatives from MAQC study [7] and from [6]. In this way it is also possible to compare the differences introduced by the ensemble thinning algorithm with respect to the algorithm from [6], that uses the same features.

IV. Results and discussion

In Table I are shown the MCC results for all the studied algorithms in this work. Each dataset corresponds to a column and the last column is the mean MCC value across the datasets. The comparison between the AID and the simplified Kun algorithm can be done observing the first two lines in Table I. The Kun algorithm obtains better overall MCC mean value and in every single dataset it obtains better or equal MCC values. It can be stated that the simpler Kun algorithm achieves better prediction results and it should be preferred to the AID algorithm.

In the last four rows of Table I, the main proposed innovations are analyzed by comparing the full Kun algorithm, with three algorithms, each one excluding a different aspect. They are organized by decreasing mean MCC, so that it can be straightforwardly seen which algorithm obtains the best performances and how much each of the key elements affects the final result. Globally, the Kun algorithm obtains better results with an overall MCC of 0.533 and the introduced elements have different impacts. The tie breaking rule is the least affecting factor since Kun_{tie} obtains a mean 0.519 MCC. The metagene inclusion as individual feature importantly affects the predictive ability, so that an MCC of 0.485 is obtained. Here too, the metagenes are useful for classification as in [6] and not using them can lead to undesirable MCC values since the missing values represent an undetermined MCC due to the null denominator. This is obtained when all the validation samples are assigned to one class [6]. Finally, the most important of the introduced elements is the nonexpert definition. Not including this concept leads to very poor results and, more importantly, to undetermined MCC values in many of the analyzed datasets.

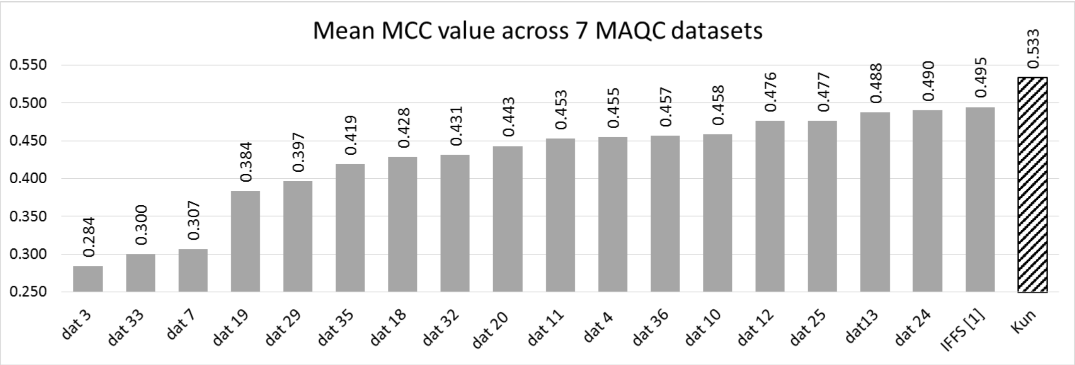

Fig. 2. Mean MCC results comparison with state of the art results from [7], [6].

This is due to the fact that all nonexperts agree on every sample, thus strongly biasing the ensemble vote.

From the results in Table I, the best performing algorithm is full Kun and all the introduced tweaks helped in obtaining such results. In Figure 2, the mean MCC value of Kun algorithm is compared with state of the art alternatives. The vast majority, all the dat_{xx} columns, are the mean MCC value from the MAQC study [7]. In addition to them, the column labeled as *IFFS [6]* is the mean MCC value from [6], which makes use of the same features, genes and metagenes, but adopts a feature selection algorithm called IFFS [6]. The state of the art algorithms are represented as solid gray columns, while the Kun mean MCC value is represented with a black and white lines pattern.

It can be observed how the Kun algorithm obtains a remarkable improvement when compared to state of the art alternatives and, comparing the shown results with the mean values in Table I, it can be observed how various of the ensemble algorithms would have obtained better than state of the art results. This confirms the goodness of ensemble thinning as approach to combine multiple experts for classification [4].

V. Conclusion

In this work, an ensemble thinning algorithm has been applied to improve the classification of microarray data. The chosen base algorithm is the AID algorithm from [5] and its simplified version from [4] called Kun. Some key elements have been introduced to consider the data characteristics in the thinning process: the expert set composition, the nonexpert concept and the tie breaking rule.

The proposed algorithm has been tested over seven publicly available datasets and it has been compared with state of the art results from the MAQC study [7] and from [6]. Both AID and Kun versions have been compared, showing how the simpler Kun achieves better performances. An additional study showed how the introduced key elements are beneficial to obtain improved predictions. The most important element is the nonexpert concept introduction, followed by the metagene utilization and the tie breaking rule.

The proposed algorithm has proven to be a very good alternative for classification obtaining significantly higher mean MCC values in the comparison with state of the art alternatives. Further work will be dedicated to the study improvements by using classifiers other than LDA and in exploring additional rules to define the *nonexpert* subset, so to encompass a more complete scenario.

References

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

[2] J. Hua, W. Tembe, and E. R. Dougherty, “Performance of feature-selection methods in the classification of high-dimension data,” *Pattern Recognition*, vol. 42, no. 3, pp. 409–424, 2009.

[3] P. Yang, Y. H. Yang, B. B. Zhou, and A. Y. Zomaya, “A review of ensemble methods in bioinformatics,” *Current Bioinformatics*, no. 5, pp. 296–308, 2010.

[4] L. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, 2004. [Online]. Available: http://books.google.es/books?id=B4TZtp7X82oC

[5] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, “A new ensemble diversity measure applied to thinning ensembles.” in *Multiple Classifier Systems*, ser. Lecture Notes in Computer Science, T. Windeatt and F. Roli, Eds., vol. 2709. Springer, 2003, pp. 306–316. [Online]. Available: http://dblp.uni-trier.de/db/conf/mcs/mcs2003.html#BanfieldHBK03

[6] M. Bosio, P. Bellot, P. Salembier, and A. Oliveras-Vergés, “Gene expression data classification combining hierarchical representation and efficient feature selection,” *Journal of Biological Systems*, vol. 20, no. 04, pp. 349–375, 2012.

[7] L. Shi, G. Campbell, W. D. Jones, F. Campagne, and Z. Wen, “The microarray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models.” *Nature biotechnology*, vol. 28, pp. 827–38, 2010 Aug 2010.

[8] R. Parry, W. Jones, T. Stokes, J. Phan, R. Moffitt, H. Fang, L. Shi, A. Oberthuer, M. Fischer, W. Tong, and M. Wang, “k-nearest neighbor models for microarray gene expression analysis and clinical outcome prediction.” *Pharmacogenomics J*, vol. 10, no. 4, pp. 292–309, 2010.

[9] Q. Liu, A. H. Sung, Z. Chen, J. Liu, X. Huang, and Y. Deng, “Feature selection and classification of maqc-ii breast cancer and multiple myeloma microarray gene expression data,” *PLoS ONE*, vol. 4, no. 12, p. e8250, 12 2009.

[10] B. W. Matthews, “Comparison of the predicted and observed secondary structure of t4 phage lysozyme.” *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, 1975.