Performance Validation of Microarray Analysis Methods

M. Zervakis, M. E. Blazadonakis, A. Banti, D. Kafetzopoulos, V. Danilatou, and M. Tsiknakis

***Abstract*—Following the rapid development of gene selection methods, several comparison studies have been reported for ranking methods on various datasets. In order to reduce bias in performance measures, most studies use an evaluation scheme based on cross-validation. In this paper we focus on the methodology of evaluation itself and address methodological problems using three representative algorithms on two public datasets. More specifically, the paper discusses the need of an independent test-set to reduce bias associated with cross-validation, the use of case specific considerations for generalization, as well as other measures that reflect stability and consistency of the result. Such measures reflect the influence of the actual dataset distribution on the performance of gene selection methods.**

I. Introduction

DNA microarray technology along with the release of the human genome working draft [1] has open a new era in the field of prognosis, diagnosis, prevention or even discovering the biological mechanisms involved in the development of cancer. Many statistically based algorithms have been developed to address the specification of a characteristic set of genes that can efficiently and effectively describe the population dataset. Nevertheless, tested algorithms derive different solutions to the same problem, so that genomic information can not yet be trusted and used in diagnostic or prognostic decision support systems. From a statistical point of view, the problem of gene selection suffers from the unbalanced sizes of features (genes) and available exams (cases), which are generally referred to as "curse of dimensionality". Another problem is the small sample random correlation effect, where a small number of features can be easily found to be randomly correlated with the outcome. Furthermore, the result of any prediction algorithm is bounded by the random measurement error on patients, due to the low quality of microarray images, where each patient may be measured with different (unknown) confidence intervals on each gene expression. Thus, there is a crucial need to develop an validation platform for data mining approaches in this field, so that the information conveyed by the "distilled" set of genes, could be trusted and used by an expert to search, discover and understand the hidden biological mechanisms involved in the development of complex diseases, such as cancer.

Two general approaches have been proposed for gene selection, namely filter and wrapper methods. A fundamental difference between these two "philosophies" is the way that gene weights are ranked as significant. Filter methods focus on intrinsic data characteristics neglecting gene interactions, while wrapper methods consider gene interactions useful for classification, neglecting intrinsic data characteristics [2]. The integration of these two approaches has been recently addressed [3], [4] by embedding filter criteria in the iterative ranking and elimination of genes performed by wrapper techniques. Several comparison studies regarding these approaches have addressed the evaluation of algorithmic performances on various public datasets, based on cross-validation schemes such as [5]. Cross-validation, however, has received the criticism of introducing bias on the estimated results, not only through its internal but also its external application on the dataset [6].

The aim of this work is to address such issues related to the evaluation framework and reveal potential bias sources, rather than to the actual comparison of algorithms. In our study two evaluation methodologies are explored, one based on cross-validation and the other based on a completely independent dataset. Through these evaluation schemes, we study issues of cross-validation from an overall population and a case-specific perspective, aiming to reduce estimation bias due to effects of small sample random correlation, bias induced by cross-validation re-sampling and uncertainties involved in the measurement process. We compare the performance measures of three representative algorithms from filter, wrapper and integrated approaches, respectively. The accuracy measures are presented along with confidence intervals on the prediction power of each methodology.

II. Materials and Methods

A. Methods

In this section we provide a brief overview of the gene selection methods studied. Filter methods are based on a direct ranking of genes, where wrapper and/or integrated methods employ a classifier in order to assess the importance of genes in decision making. The latter class of algorithms proceeds on the basis of recursive feature elimination (RFE) for the elimination of genes from the initial list. In the RFE methodology a classifier is used to assign weights to features (genes), which are ranked according to the absolute values of the assigned weights. Then, the features with the lowest weights are eliminated and the process continues recursively. Note that in such an

M. Zervakis, M. E. Blazadonakis and A. Banti are with the Technical University of Crete, Department of Electronic and Computer Engineering.

D. Kafetzopoulos, V. Danilatou are with the Institute of Molecular Biology and Biotechnology, Foundation of Research and Technology Hellas (FORTH).

M.Tsiknakis, is with the Institute of Computer Science, Biomedical Informatics Laboratory, Foundation of Research and Technology Hellas (FORTH).

approach feature weights are dynamically updated, so that the weight of a feature is adjusted continuously through out the iterations.

Filter Method

The filter method tested is based on the Fisher's metric (Fisher 1936) for gene ranking. For each gene this metric considers its discriminative power by means of:

$$k = \frac{\left|\mu_+(g_i) - \mu_-(g_i)\right|}{\sigma_+(g_i) + \sigma_-(g_i)} \quad (1)$$

where $\sigma_+(g_i)$, $\sigma_-(g_i)$, $\mu_+(g_i)$ and $\mu_-(g_i)$ correspond to the standard deviation and means of the two classes of interest, for the specific gene g_i.

Recusive Feature Elimination with Support Vector Machines (RFE-SVM)

This approach [7] employs an SVM classifier [8] for assigning gene weights. SVM searches for the best separating hyperplane to distinguish between the two classes of interest. Towards the solution of this problem, we obtain the following expression for the direction vector $\boldsymbol{w}$:

$$\boldsymbol{w} = \sum_{j=1}^{n} \lambda_j y_j \boldsymbol{x}_j \quad (2)$$

where $1 \le \lambda_j \le C$ which is actually an expansion of those training samples with non-zero λ_j, i.e. the support vectors. λ_j s correspond to Lagrange multipliers, y_j corresponds to the label associated with the sample $\boldsymbol{x}_j$.

RFE with Fisher's metric on Support Vectors (RFE-FSV)

In this approach [4] the learning process is appropriately enriched with a filter criterion, which actually yields the hybridization between the wrapper and filter approaches. More specifically, a variation of the Fisher's ratio is appropriately integrated to the weight vector equation (3) of an SVM as follows. Let *SVs* be the set of support vectors and *S* be the set of indices defined as $S = \{k : \boldsymbol{x_k} \in SVs\}$. Then based on (3) a new direction vector $\boldsymbol{w'}$ is defined as follows:

$$w_i' = \sum_{j \in S} sign(\lambda_j) \cdot y_j \cdot (x_{ij}) \cdot \frac{\left|\mu_+(g_i) - \mu_-(g_i)\right|}{\sigma_+(g_i) + \sigma_-(g_i)} \quad (3)$$

We also point out that the weighting scheme can be expanded using nonlinear kernels in the SVM operation, such as polynomials of high degree. In this work we use a seven-degree polynomial kernel.

B. Datasets

Two datasets are considered in this study, i.e. the leukemia and breast cancer (BC) datasets published in [9] and [10], respectively. Both datasets consist of a training set and an independent test-set. The leukemia dataset consisting of 7129 array elements, representing 6817 genes, with the training set consists of 38 samples (27 ALL and 11 AML), all normalized to a zero mean and standard deviation one, as suggested in the original publication [9]. The BC dataset contains 24481 genes and 78 samples on the training set, 44 of which are characterized negative and correspond to patients that remain disease-free for a period of at least five years, whereas the remaining 34 are characterized positive and correspond to patients that developed a relapse within a period of five years. 293 genes expressing missing information for all 78 patients were removed and the remaining 13604 missing values were substituted using Expectation Maximization (EM) imputation [12]. An advantage of these datasets is their supplement with an independent test-set directly derived from the design of the experiments. The independent test-set consists of 19 samples (7 negative and 12 positive) in the case of breast cancer, while it contains 34 samples (20 AML and 14 ALL) in the leukaemia dataset. Thus, we can also consider algorithmic performance on these data, without mingling with the design of the study. In order to reduce random correlation effects on the prediction estimates and increase the confidence on performance measures, we use multiple random splits for the derivation of various training sets by means of 10-fold cross-validation. The process of cross-validation has been reported to induce bias due to the mixing of samples in the training and testing phases [5]. To address this bias on domain specific datasets we exploit the truly independent test-sets aiming to a more objective evaluation of the prediction and generalization abilities of each model. Thus, for each run of the CV process we test the performance not only on the corresponding portion of data that has been assigned for CV testing, but also on the entire independent test set. Besides the performance on the overall population, we also resort to case specific considerations as to address the influence of measurement errors on the estimation of the prediction power of each method. All measures are presented along with appropriate confidence intervals derived from the cross-validation process.

C. Evaluation Measures

For the effective evaluation of measures we created the so-called performance profile of an algorithm (model), which is in fact a table with columns reflecting all patients and rows indicating the cross-validation runs. For each run (row), the table captures a binary value for each patient (column) if this patient is in the test-set of the run. This value indicates the prediction success for this patient on the specific run. In this form, the average per patient over all runs reflects the per-patient accuracy of the algorithm, whereas the average per run over all patients reflects its per-run accuracy. The notation of such accuracy measures is specified in the following.

Let S be the "*performance profile"* matrix of m rows and n columns, where m is the number of runs (folds) and n the number of patients (samples). Along each row of matrix S we define the cardinality C_{Ri} of row i and the cardinality C_{Pj} of column j as the number of active entries

(one or zero) within each row and column, respectively. Along each row of S we define the mean accuracy per run:

$$acc_{R_i} = \frac{1}{C_{R_i}} \sum_{j=1}^{n} S_{ij} \quad (4)$$

which assesses the model's generalization on the test-set, while keeping the training set fixed. Based on multiple split-sample runs, Michiels et. al. [11] proposed a strategy for the estimation of confidence intervals on the true prediction power of a method, by means of a percentile on the empirical distribution of multiple run estimates. In a similar form we employ the sample mean and standard deviation to model multiple run estimates and derive measures for the mean prediction accuracy and its std (standard deviation) over all runs, denoted by the pair (acc_R, std_R). The standard deviation reflects a range of variation for the performance of the model, depending on variations of the training set. Thus, it forms one measure of stability for that model.

Alternative to row measures on the performance profile, operating on the columns of S we derive the mean per-patient P_j accuracy given by:

$$acc_{P_j} = \frac{1}{C_{P_j}} \sum_{i=1}^{m} S_{ij} \quad (5)$$

The mean accuracy acc_P over all tested samples/patients may be used as an index of the prediction power of an algorithm for individual cases. Furthermore, the standard deviation of the per-patient accuracies over all tested samples is a measure of algorithmic stability of the prediction of individual patient outcome. As such, small standard deviation of the per-patient accuracies of an algorithm indicates robustness and generalization abilities of the algorithm in changes of the sample distributions.

Besides performance measures, we also consider a gene overlap index over the cross-validation runs in order to provide a measure of the robustness of the algorithm in selecting the same set of genes under different initialization conditions induced through variations of the training set. At the end of the cross-validation process for a set size of surviving genes, let q_i determines the frequency of selection of the i^{th} gene. The average of these frequencies provides our index of gene overlap over the different iterations. We focus on a fixed number of most frequently selected genes and consider the progress of the gene overlap index as the size of surviving genes proceeds towards a minimum.

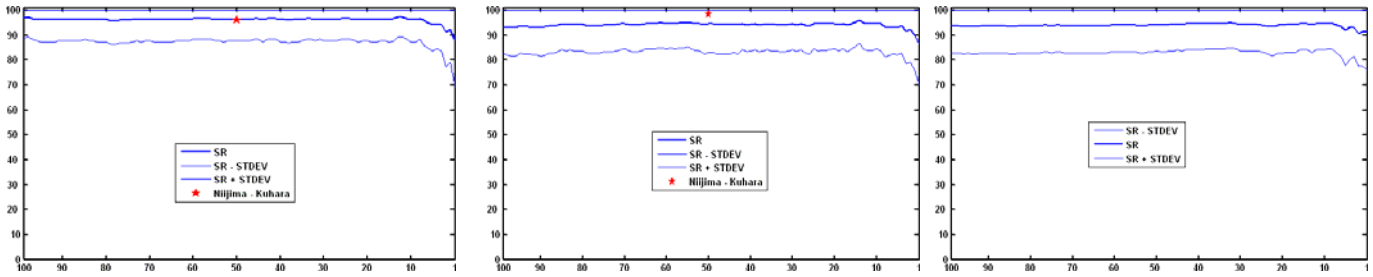
III. Experimental results

The discussed experimental scenarios are evaluated on three representative (filter, wrapper and integrated) methodologies and the results are presented in Fig. 1 and Fig. 2 for the leukemia and BC datasets, respectively. At each cross-validation run, the algorithm is tested on the test subset of the cross-validation iteration (denoted by "cv"), on the independent test-set (denoted by "t"), and on all samples (t + cv) available for testing (denoted by "all"). The latter case is considered as a more unbiased estimate of algorithmic performance, since it involves a large number of testing cases, either linked with the distribution of the training set(s) or being completely independent. Each figure includes three columns of plots, one for each representative algorithm tested. Each plot presents average accuracy estimates over the cross-validation runs along with stds, plotted versus the number of selected genes. The top plot presents the accuracy estimate computed from cross-validation $\left(acc_R^{cv}, std_R^{cv}\right)$, i.e. when testing is performed on subsets of cross-validation splits. The middle plot is the accuracy estimate $\left(acc_R^{all}, std_R^{all}\right)$ derived from all tested samples (testing subsets of cross-validation splits plus the independent test-set). As mentioned before, this plot is considered as less biased and is used as a "golden standard" for comparison. The bottom plot in each column depicts the performance measures obtained on a per-case consideration of samples in the independent test-set, i.e. $\left(acc_p^t, std_p^t\right)$. Finally, the last (common) plot depicts the gene overlap index for the three algorithms at specific cut-off points on the number of selected genes (size of gene signature).

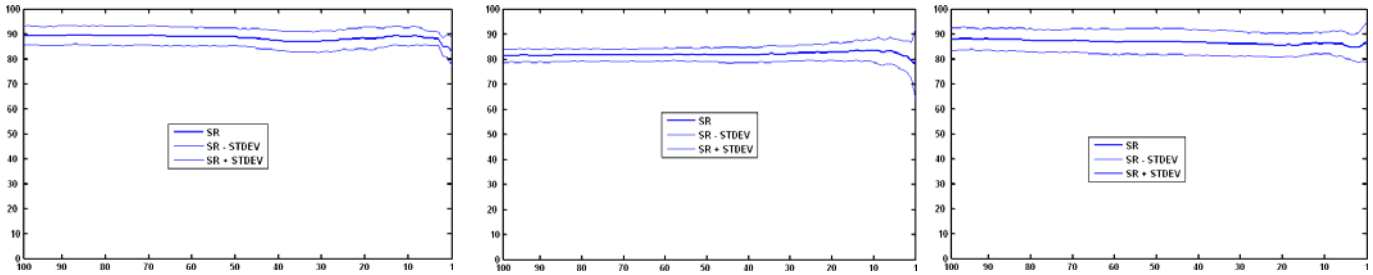
Results from other studies [5],[9] and [10] are also superimposed on these figures with appropriate symbols. Similar to the comparative study of Michiels et al. [11] some of them are outside or in the boundaries of estimation limits. Notice the result of the Van't Veer's study [10], which is outside our limits, in exact agreement to [11].

Considering both diseases and all prediction models, we observe that the performance over the entire testing set (2nd row) is quite different from the result of cross-validation (1st row). The bias induced by cross-validation is more significant in the case of leukemia, in which it overestimates the algorithmic performance. Better proximity to the golden standard than cross-validation is achieved by using only the independent test set on per patient basis (3rd row). Thus, the independent test set appears to have an important role in reducing the performance bias. Furthermore, the variance of performance results in individual iterations appears to be high for the cross-validation scheme, leading to larger std intervals. This large variance of potential estimates, which is highly affected by the design of random splits, allows for and partially explains the over-optimistic results that have been reported in the literature [11] using cross-validation. Proceeding with the gene overlap index, which has been computed based on the frequencies of the 20 most often selected genes over the cross-validation iterations, we observe that it drops significantly as the size of selected genes decreases. Thus, the consistency of algorithms in selecting the same genes decreases with the progress of iterations, raising concerns that at the stages that we consider (fewer than 100 genes), the selection of some genes may be random.The best index is achieved by the filter method, indicating good stability in selecting consistent gene signatures.

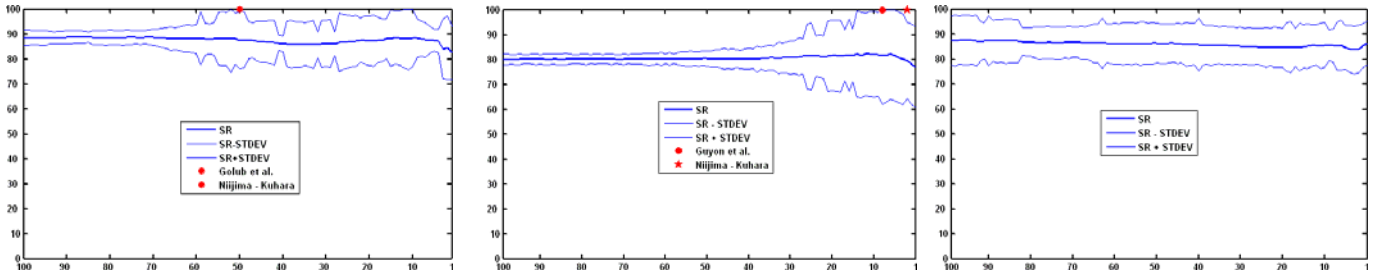
Filter Model **RFE-SVM Model** **RFE-FSVs Model**



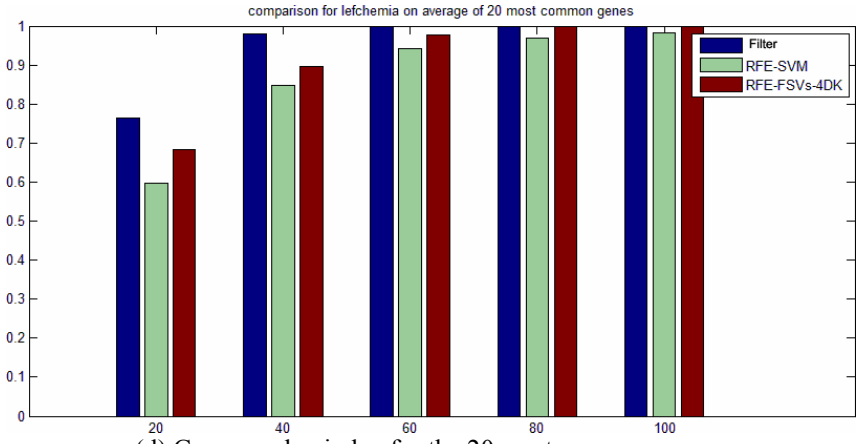
(a) Per-run Accuracies and Standard Deviations from Cross-validation $\left(acc_R^{cv}, std_R^{cv}\right)$



(b) Per-run Accuracies and Standard Deviations from Overall Set. $\left(acc_R^{all}, std_R^{all}\right)$



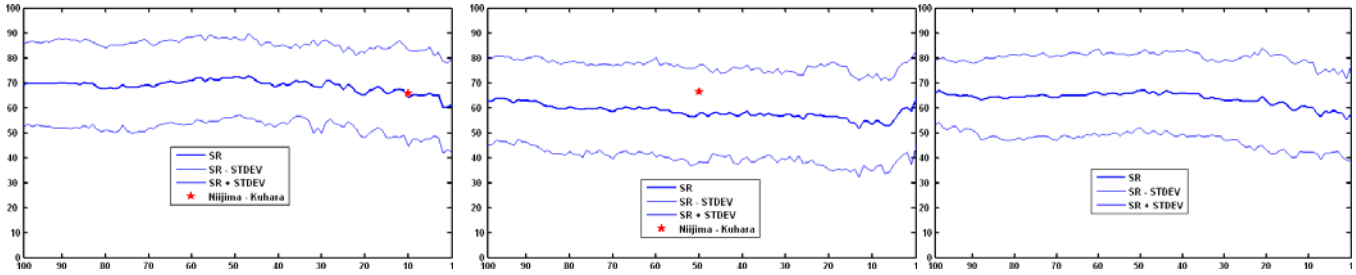
(c) Per-patient Accuracies and Standard Deviations from Independent test-set. $\left(acc_p^t, std_p^t\right)$



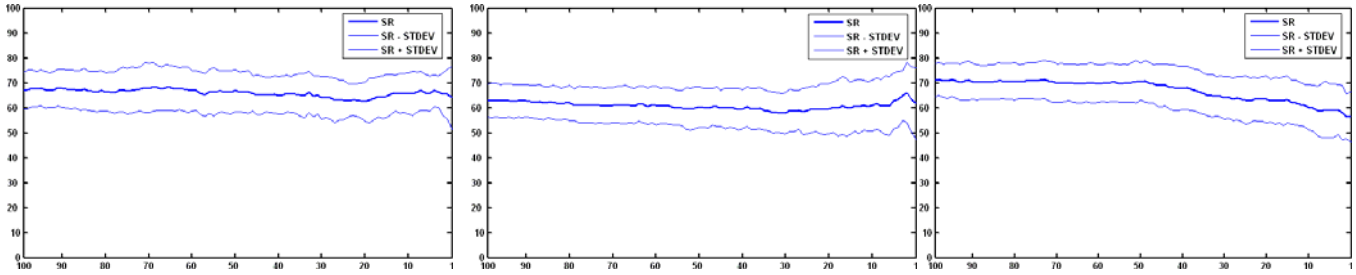
(d) Gene overlap index for the 20 most common genes.

Fig. 1:Leukemia data set

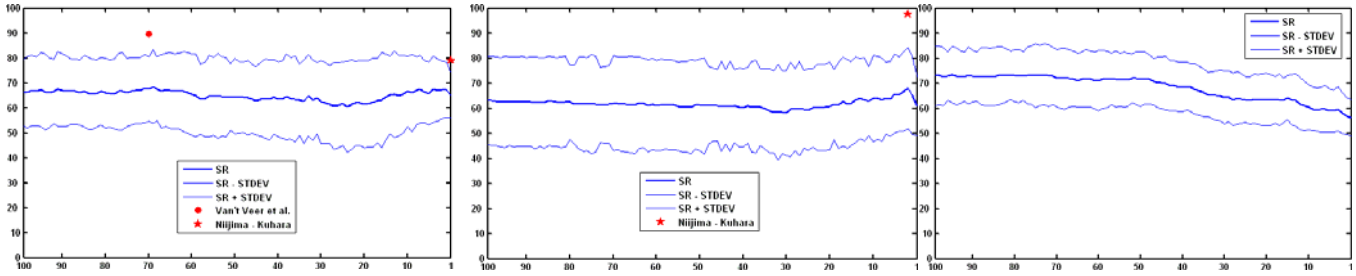
Filter Model **RFE-SVM Model** **RFE-FSVs Model**



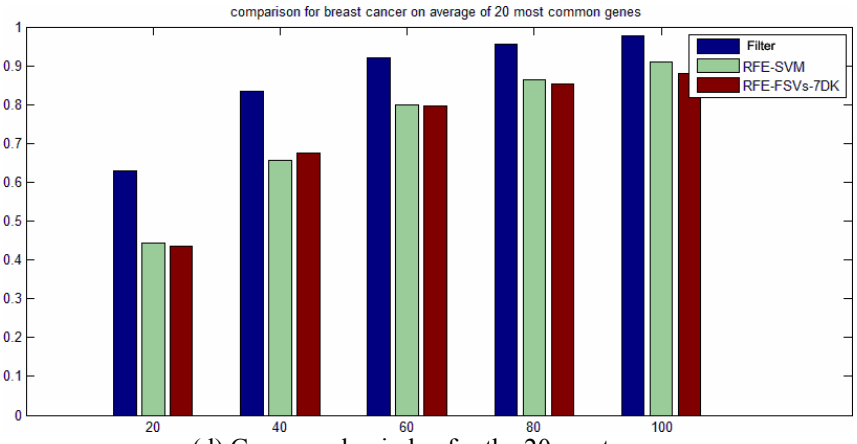
(a) Per-run Accuracies and Standard Deviations from Cross-validation $\left(acc_R^{cv}, std_R^{cv}\right)$.



(b) Per-run Accuracies and Standard Deviations from Overall Set $\left(acc_R^{all}, std_R^{all}\right)$.



(c) Per-patient Accuracies and Standard Deviations from Independent test-set. $\left(acc_p^t, std_p^t\right)$



(d) Gene overlap index for the 20 most common genes.

Fig. 2: breast cancer dataset.

This is to be expected due to the fixed ranking scheme implemented by this method for all individual genes. On the other hand, the consistency of the integrated method is higher than that of the SVM for the leukemia dataset, while both methods derive almost identical measures on the more complex dataset of breast cancer indicating similar predisposition to the small sample random correlation effect.

Attempting an overall comparison based on these results, we derive different conclusions on the two datasets. On the leukemia dataset the SVM derives the lowest performance with high variance (on per-patient trials) and low gene overlap index; similar result has been reported in [5]. The same method on breast cancer appears to achieve slightly better accuracy estimates than the filter method. For large size of gene signature the integrated method performs better than SVM, but its performance drops with reducing gene numbers. These results support the claim that the algorithmic performance depends on the distribution of the data, so that we cannot in general rank a single prediction model better than any other.

The results reported refer to average measures over 100 runs with different training sets, where a different gene signature is selected in each run. For the cross-validation scheme (first row) we observe large variation of performance. Similarly, even for the same size of gene signature, there is a large performance variation on a per-patient consideration. The large confidence intervals on per-case estimates is a point of caution, raising concern regarding the systematic performance of algorithms on new, unseen cases, especially for small sizes of gene signatures. The performance stability increases (smaller confidence intervals) for the outcome estimated on a per-run basis, which however may be due to random selection of correctly classified patients per iteration. As a concluding remark of this work, we stress our belief that at this stage we cannot derive safe conclusions without the correlation of statistical results with the biological meaning of selected genes.

I. Conclusion

In this work filter, wrapper and integrated methods are tested on two publicly available datasets, namely for breast cancer and leukemia. The major difference from other comparative studies is that we focus on the performance evaluation methodology rather than the algorithmic performance itself and reveal issues of bias in the most often used process of cross-validation. We believe that these issues can become useful for the design of experiments and testing scenarios. More specifically, we stress the need for an independent test-set for evaluation, as to reduce the bias induced by the cross-validation mixing of training with the testing samples. Furthermore we address the need for a gene overlap index derived for varying training sets, as to ensure the consistency of gene selection and consider stability measures of the tested algorithms. According to this framework, the ranking of methods is not always the same depending on the examined dataset, even though RFE-SVM has been claimed to outperform filter methods in application domains such as colon cancer and leukemia. This position on performance variability agrees with the conclusions of [5]. A major concern stemming from the discussion of our results is that under the current state of performance validation methodologies, we cannot derive safe conclusions regarding the ranking of algorithms, without correlating algorithmic results with the biological meaning of selected genes.

Acknowledgment

Present work was supported by Biopattern, IST EU funded project, Proposal/Contract no.: 508803, "Genotype" funded by the Greek Secretariat for Research and Technology as well as the Hellenic Ministry of Education.

References

[1] University of California Santa Cruz Genome Bioinformatics, Human Genome Working Draft, http://genome.ucsc.edu, (2001)

[2] S. G. Baker and B. S. Kramer, "Identifying genes that contribute more to good classification in microarrays", *BMC Bioinformatics*, vol 7:407 (2006).

[3] M. Blazadonakis and M. Zervakis, "The Linear Neuron as Marker Selector and Clinical Predictor in Cancer Gene Analysis", *Computer Methods and Programs in Biomedicine*, vol 91:1 pp 22-35 (2008).

[4] M. Blazadonakis and M. Zervakis, Wrapper Filtering Criteria Via a Linear Neuron and Kernel Approaches, *Computers in Biology and Medicine*, to appear.

[5] S. Niijima and S. Kuhara, "Recursive gene selection based on maximum margin criterion: a comparison with SVM-RFE", *BMC Bioinformaics* 7:543 (2006).

[6] W. Jiang, S. Varma and R. Simon, "Calculating Confidence Intervals for Prediction Error in Microarray Classification Using Resampling", *Statistical Applications in Genetics and Molecular Biology*, Vol. 7, no. 1 article 8 (2008).

[7] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene selection for cancer classification using Support vector machines, *machine learning*, vol. 36 pp. 389-422 (2002).

[8] N. V. Vapnik, "The Nature of Statistical Learning Theory" (Springer–Verlag New York), 1999.

[9] R. T. Golub, K. D. Slonim, P. Tamayo, C. Huard, et al., Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science*, vol. 286 pp. 531-536 (1999).

[10] L. J. Van't Veer, H. Dai, M. J. Van de Vijver, Y. D. He, et al., Gene expression profiling predicts clinical outcome of breast cancer. *Letters to Nature* vol. 415 pp. 530-536 (2002).

[11] S. Michiels S, Koscielny S and Hill C., "Prediction of cancer outcome with microarrays: a multiple random validation strategy", *Lancet*, vol. 365, pp. 488-492, (2005).

[12] A. Little and D. Rubin, "Statistical Analysis with Missing Data" (Wiley Series in Probability and Mathematical Statistics), 1987.

[13] R. Kohavi, "A study of Cross-Validation and Bootstrap for accuracy estimation and Model Selection", *International Joint Conference on Artificial Intelligence (IJCAI)* (1995).