Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 55–61

ISBN 978-83-60810-22-4

Competitive and self-contained gene set analysis methods applied for class prediction

Henryk Maciejewski
Institute of Computer Engineering, Control and Robotics,
Wroclaw University of Technology,
ul. Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
Email: Henryk.Maciejewski@pwr.wroc.pl

***Abstract*—This paper compares two methodologically different approaches to gene set analysis applied for selection of features for sample classification based on microarray studies. We analyze *competitive* and *self-contained* methods in terms of predictive performance of features generated from most differentially expressed gene sets (pathways) identified with these approaches. We also observe stability of features returned. We use the features to train several classifiers (e.g., SVM, random forest, nearest shrunken centroids, etc.) We generally observe smaller classification errors and better stability of features produced by the self-contained algorithm. This comparative study is based on the leukemia data set published in [3].**

I. Introduction

BUILDING diagnostic or prognostic classifiers based on profiles of gene expression from microarray or similar massive throughput experiments seems one the most challenging tasks in bioinformatics. The problem of class prediction can be regarded as ill-formulated as the number of samples (e.g., patients) in a typical microarray study does not exceed a few hundred with the number of features (gene expression values) recorded for a sample usually exceeding 20 thousand. Many different approaches to class prediction based on massive throughput data were proposed (e.g., [7], [10], [13], [14]). One of the most challenging problems is related to feature selection based on high dimensional data. Standard approaches start with identification of sets of differentially expressed genes to be used as features for class prediction. These methods focus on features with individual strong predictive power, however they treat the features as unrelated and they do not take into account potential (biological) relationships among features. This explains why most feature selection methods produce very unstable features, ie. small changes in training data result in different feature sets, [25], [15], [16]. This further explains why classifiers built from microarray studies are very sensitive to the selection of parameters (such as the number of features, etc.), and generally demonstrate unstable estimates of prediction error.

In our previous work [17], we proposed an enhanced procedure of feature selection based on domain knowledge about potential relationships among features (genes). Such knowledge of groups of functionally related genes is available in databases e.g., KEGG, Gene Ontology or Biocarta, and is now being actively developed. The method proposed in [17] derives features for class prediction from the most strongly activated pathways. In [17] we compare this approach to the standard method and empirically show that although individually weaker, the new features seem more stable and bring improved performance.

In [17] the global test algorithm, developed in [11], was applied to identify activated pathways to be used as features for class prediction. Recently different approaches to gene set analysis were proposed, e.g., [8], [23], [11], [4], [25]. They can be broadly categorized as *competitive* or *self-contained*, and they fundamentally differ methodologically, [12]. It is not clear whether these two approaches produce similar feature sets in terms of their predictive performance and in terms of stability. The purpose of this work is to investigate this issue. We compare predictive performance of features generated with a selected competitive method (Gene set analysis (GSA) algorithm) and a selected self-contained method (global test). We also analyze whether the feature sets differ in terms of stability.

The organization of the paper is as follows. First, competitive and self-contained methods of gene set analysis are described in detail. Then an algorithm of sample classification based on activation of gene sets is presented. The algorithm is later used to evaluate predictive performance and stability of feature sets returned by the two gene set analysis methods. Finally, results of a comparative study are elaborated based on a real microarray assay.

II. Gene set analysis methods

Many different approaches to gene set analysis have been recently proposed. An overview of the most important methods is available in [25], and the statistical issues related to these different methods were analyzed by Goeman and Buehlmann in [12]. The earliest developed and probably simplest methods compare the list of genes in the set of interest with the list of differentially expressed genes. An example of such methods is the over-representation analysis (ORA) proposed in [6], which compares these two lists of genes by means of contingency tables. The chi-square test is used in order to verify the null hypothesis that the differentially expressed genes are not over-represented in the gene set of interest. Rejection of the null hypothesis indicates that the gene set is differentially expressed

This work was sponsored by the grant MNiSzW N516 510239

(or *activated*). An extension of the ORA method is the Gene Set Enrichment Analysis (GSEA), proposed in [20], [23], which does not require that genes are potentially arbitrarily declared as differentially expressed by using a fixed threshold. The method ranks the genes by some measure of differential expression and then tests the null hypothesis that the members of the gene set of interest are uniformly distributed along the ranking list. The null is tested against the alternative that the gene set appears at the top or bottom of the ranking list, ie. can be regarded as activated. GSEA uses a modified Kolmogorov-Smirnov statistic to test the null hypothesis. Another method, based on GSEA is the Get Set Analysis (GSA) algorithm developed in [8]. It uses a maxmean statistic in place of the Kolmogorov-Smirnov test which leads to slightly better power.

In the methodological paper [12], these and similar approaches were named as *competitive* methods. These methods test whether a gene set is differentially expressed (or associated with the target) by comparing expression of genes in the set with expression of genes not in the set. A fundamentally different approach is realized by *self-contained* methods [12], which directly analyze association of genes in the set of interest with the target and do not take the remaining genes into account. Examples of self-contained methods are the Global Test [11], Global Ancova, [18], PLAGE, [24] or SAM-GS proposed in [4].

In this work we use gene set analysis methods to identify pathways which will be used to generate features for classification of samples. We identify the most differentially expressed (or activated) pathways and then use genes in these pathways as features for class prediction. In this study, we use one self-contained approach (global test) and one competitive method (GSA algorithm) and experimentally compare these methods in terms of (a) predictive power of features returned, and (b) stability of features in the presence of small modifications of data. The gene set analysis methods used in the study are now explained in detail.

A. Competitive methods – GSA algorithm

The competitive methods compare differential expression of genes in the gene set of interest G with expression of genes not in G. More specifically, they aim to verify the null hypothesis:

H_0 : *The genes in G are at most as often differentially expressed as the genes not in G.*

The GSEA method and its more powerful version GSA test the specific null hypothesis that the genes in the set G are uniformly distributed over the list of all genes ranked by some measure of differential expression.

In order to test the hypothesis, the GSA algorithm uses the maxmean statistic [25]:

$$M = max\left\{\left|\frac{\sum_{i=1}^{m} I\left(t_i > 0\right) t_i}{m}\right|, \left|\frac{\sum_{i=1}^{m} I\left(t_i < 0\right) t_i}{m}\right|\right\} \quad (1)$$

where m is the number of genes in G and t_i is the measure of differential expression of the i-th gene in G.

Significance of the M statistic is evaluated using the permutation test (permutation involves both genes and class labels). In the algorithm proposed in the next section we will use the permutation based p-values to select the top differentially expressed pathways whose member genes will be used as features for class prediction.

B. Self-contained methods – global test

The global test method aims to verify the null hypothesis of no association of the set G with the target, namely:

H_0 : *No genes in a set G are associated with the target (ie. no genes in G are differentially expressed).*

In order to test the hypothesis, global test uses generalized linear models to express the relationship between expression of genes in the set G and the target, such that

$$g\left(E\left(Y_i|\beta\right)\right) = \alpha + \sum_{j=1}^{m} x_{ij}\beta_j \quad (2)$$

where g is link function in generalized linear models (e.g., the logit function for binary target), $x_{i\cdot}$ denotes vector of expression of m genes in the gene set G for sample i, with class label Y_i, and β_j is the coefficient for gene j.

The assumption that no genes in G are associated with the target is equivalent to testing the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_m = 0 \quad (3)$$

Global test assumes that the coefficients $\beta_1 \ldots \beta_m$ are iid with mean 0 which simplifies the hypothesis and makes the test feasible given small number of samples relative to the number of genes in G. In the algorithm presented in the next section we will use pathways with the smallest global test p-value as features for class prediction.

III. Algorithm of class prediction

In [17] we proposed an algorithm for class prediction based on activation of pathways and compared it with the commonly used approach where top most differentially expressed unrelated genes are used as features. Here we will use a slightly modified version of this algorithm, with two different methods of feature selection: GSA-based and global test (GT)-based. The estimates of the expected prediction error (*EPE*) returned by the algorithm will be used as a measure of quality of features produced by these competing gene set analysis methods.

Using the same notation as in [17], we denote results of a microarray study as a matrix $X_{n,p}$ with p features (gene expressions) measured for n samples, with class designation for a sample i given in Y_i, $i = 1, 2, \ldots, n$ (which represent e.g., tumor and control samples). Also let PWDB be the set of d subsets (denoted $PW_i, i = 1, 2, \ldots, d$); these represent a priori domain knowledge of groups of related features (e.g., genes in a signaling pathway or genes with a common GO term, etc). The purpose of the algorithm proposed in [17] is

to (a) build the sample classifier given X, Y and PWDB, and (b) estimate the expected prediction error for new samples. Since the number of samples n is small relative to the number of features p (the $p \gg n$ problem), the *EPE* has to be estimated by data reuse techniques. We use *internal* cross validation (CV) where the data are repeatedly split into training and test partitions, with the *EPE* calculated as the average misclassification rate over all the test partitions. *Internal* cross validation places the feature selection step within subsequent iterations of cross validation, which is mandatory to obtain a reliable measure of classifier performance, as argued e.g. in [19]. (Note that the commonly used and computationally cheaper *external* cross validation realizes the feature selection step once, prior to the CV loop, and based on the complete training data.)

It should be noted that by using internal cross validation we can simultaneously observe *stability* of features generated under slight modifications of the training data. Namely, with the leave-one-out (LOO) internal cross validation scheme the data used for feature selection differs by one sample in consecutive iterations of CV. Hence by observing how the features change during CV steps we will be able to compare the GSA and GT-based procedures in terms of stability of features generated. This is the main justification of our choice to use the LOO cross validation loop in the algorithm proposed. It should noted that LOO cross validation (realizing low bias but high variance of the estimate of prediction error) is often used in similar studies, e.g., [9], [21], [22].

It should be noted that poor stability of features produced by standard methods (ie. by selecting most differentially expressed unrelated genes) may account for unstable behavior of classifiers built from microarray data [16].

The class prediction algorithm can be summarized in the following steps.

1) Leave out sample i, $i = 1, 2, \ldots, n$ for model testing, ie., remove row i from X and element i from vector Y and denote the remaining matrix and vector as X^i and Y^i.
2) Using the training data (X^i, Y^i) calculate the p-value with the GT or GSA for each of the PWs in PWDB. Order the PWs by increasing p-value: $PW_{(1)}, PW_{(2)}, \ldots, PW_{(d)}$.
3) Remove columns from X^i related to features not present in $PW_{(1)} \cup \ldots \cup PW_{(k)}$, denote this matrix as X^i_{tr}.
4) Using the training data (X^i_{tr}, Y^i) fit a predictive model f and classify the sample Y_i as $\hat{Y}_i = f(Y_i)$.
5) In the list of counters $c\left(PW_j\right), j = 1, \ldots, d$, corresponding to the d elements of PWDB, increment the counters $c\left(PW_{(j)}\right), j = 1, \ldots, k$, which correspond to the PWs selected in the current step.
6) Repeat steps 1 through 5 for $i = 1, 2, \ldots, n$.
7) Calculate the expected misclassification rate as $EPE = \frac{1}{n}\sum_{i=1}^{n} I(\hat{Y}_i \neq Y_i)$.

In the following section, we will compare performance of several classifiers based on GSA or GT features in terms of the *EPE*. For the purposes of this numerical example, the following classifiers were used:

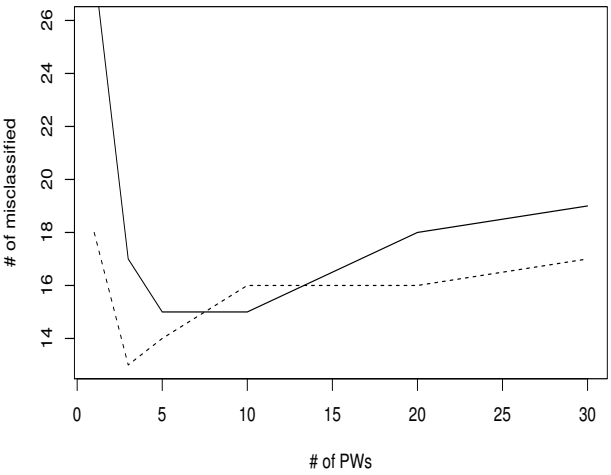

Fig. 1. Number of misclassifications as a function of the number of PWs selected. SVM classifier. Solid line - GSA method, dashed line - global test.

- support vector machine (SVM),
- logistic regression with L2 (Ridge) penalty,
- nearest neighbors,
- nearest shrunken centroid algorithm,
- random forests.

The cost parameter of the SVM classifier as well as the lambda (shrinkage) parameter of the logistic regression were tuned using a simple grid search.

We will also compare the methods in terms of stability of features selected for varying number of gene sets used as features: $k \in \{1, 3, 5, 10, 20, 30\}$. Note that the counters in step 5 of the algorithm are maintained to facilitate stability analysis.

IV. Comparative study

The numerical study is based on a subset of the acute leukemia microarray data, published by S. Chiaretti [3]. The dataset includes $n = 79$ samples with $p = 12625$ gene expressions; 37 samples represent patients with leukemia and 42 samples represent the control group (the samples are labeled in the original data as 'BCR/ABL' and 'NEG', respectively). In this example we use the KEGG signaling pathway databases as the collection of gene sets PWDB. The task is to classify patients as leukemia or control based on a profile of most activated pathways.

The overall performance of the different classifiers for varying number of most activated pathways used as features is summarized in Figs. 1 through 5. In the figures the *EPE* obtained for the global test and GSA feature selection is compared (note that the *EPE* is shown as the the number of misclassified items in 79 iterations of cross validation rather than the ratio).

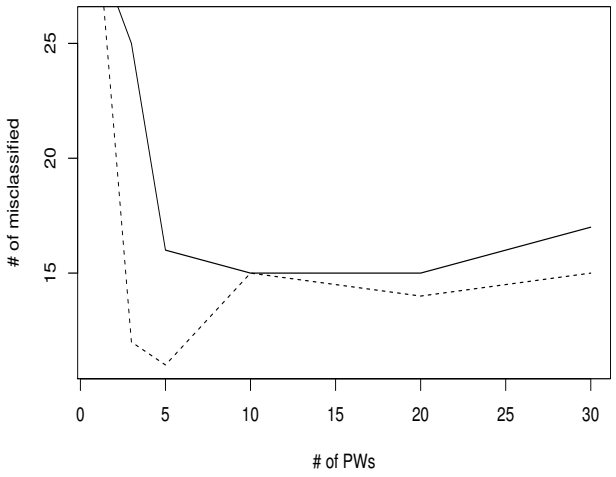

Fig. 2. Number of misclassifications as a function of the number of PWs selected. Logistic regression. Solid line - GSA method, dashed line - global test.

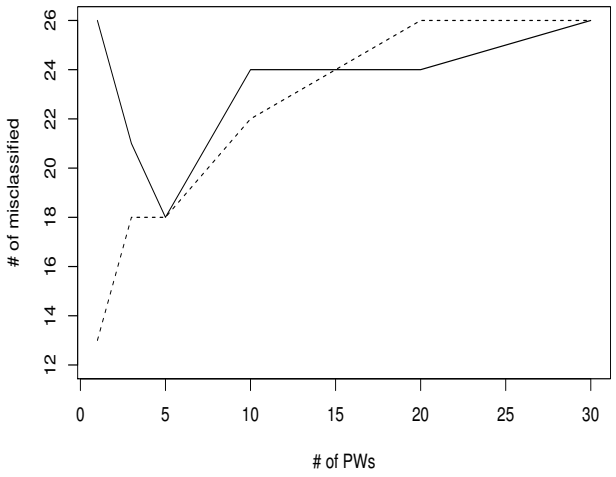

Fig. 3. Number of misclassifications as a function of the number of PWs selected. KNN classifier. Solid line - GSA method, dashed line - global test.

We observe that the overall best result was realized by the random forest classifier with $k = 3$ top PWs selected by the GT algorithm. The winning classifier realized 9 misclassifications, ie. $EPE = 11\%$. This result is better than the best result obtained with the standard feature selection method where the top ranking unrelated genes are selected as features [17]. We also observe that for all the five classifiers the smallest number of misclassifications along $k \in \{1, 3, 5, 10, 20, 30\}$ is always realized for features selected by the global test (dashed line

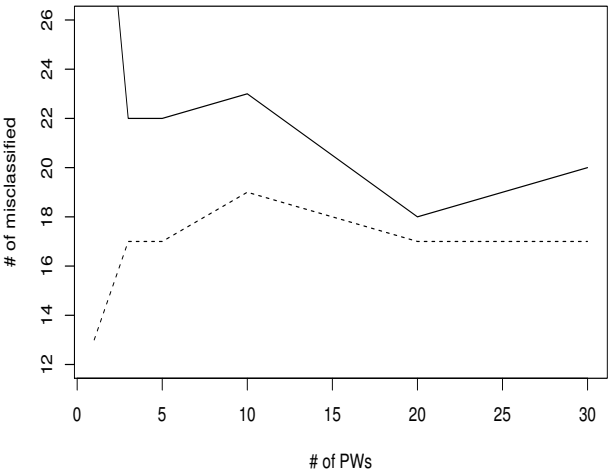

Fig. 4. Number of misclassifications as a function of the number of PWs selected. Nearest shrunken centroid algorithm. Solid line - GSA method, dashed line - global test.

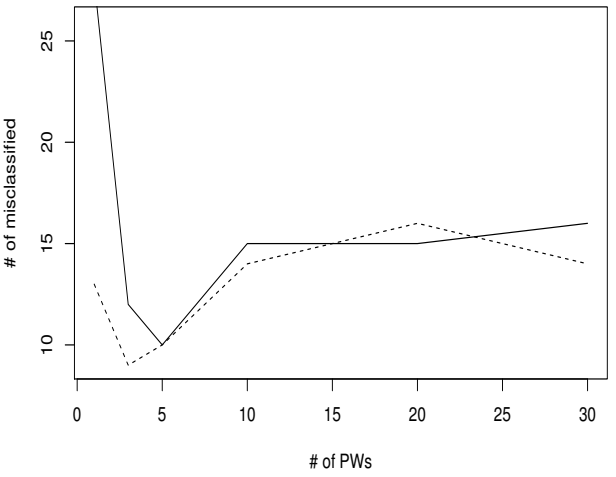

Fig. 5. Number of misclassifications as a function of the number of PWs selected. Random forest algorithm. Solid line - GSA method, dashed line - global test.

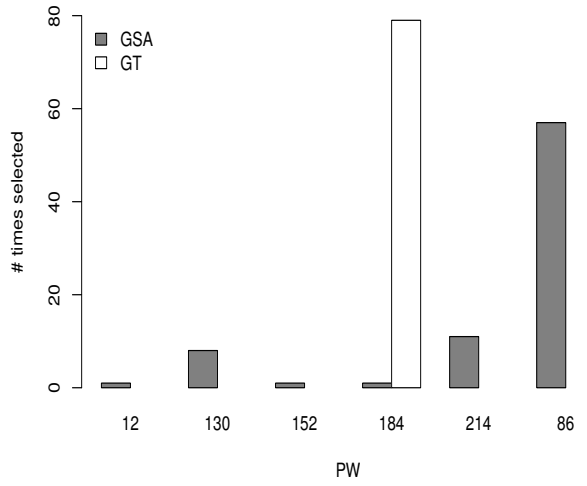

Fig. 6. Features selected in consecutive iterations of CV with corresponding frequency, 1 PW.

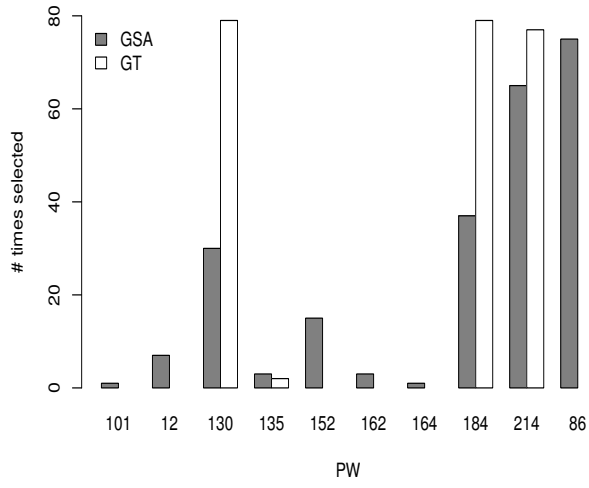

Fig. 7. Features selected in consecutive iterations of CV with corresponding frequency, 3 PWs.

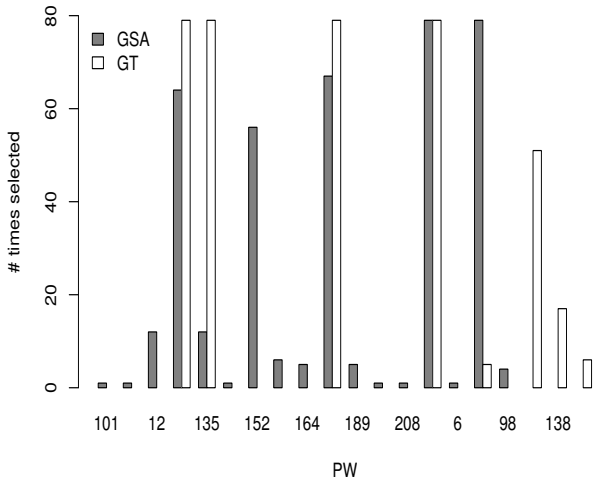

Fig. 8. Features selected in consecutive iterations of CV with corresponding frequency, 5 PWs.

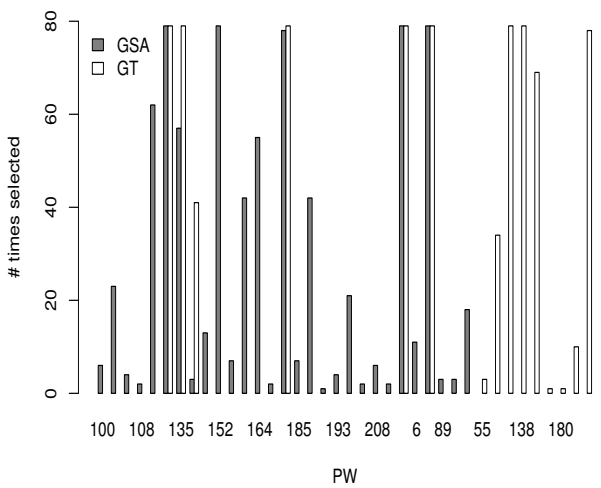

Fig. 9. Features selected in consecutive iterations of CV with corresponding frequency, 10 PWs.

in Figs. 1–5), rather than the GSA method (solid line). The smallest number of misclassifications observed for consecutive classifiers are:

- SVM: 13,
- logistic regression: 11,
- nearest neighbors: 13,
- nearest shrunken centroid algorithm: 13,
- random forests: 9.

It can be observed that with growing number of features (for $k > 3$), performance of models generally deteriorates, however this effect is strong only for the nearest neighbors classifier (Fig. 3), as the other models internally realize feature selection and therefore are more immune to overfitting.

Another important characteristic of the competing methods is stability of features observed when data changes slightly. Analysis of stability of features is doable using the table of counters $c(PW_j), j = 1, \ldots, d$ maintained in step 5 of the algorithm. The counters record how often (in all 79 iterations of cross validation) a given pathway was selected as a feature. High values of the counters indicate that the feature selection

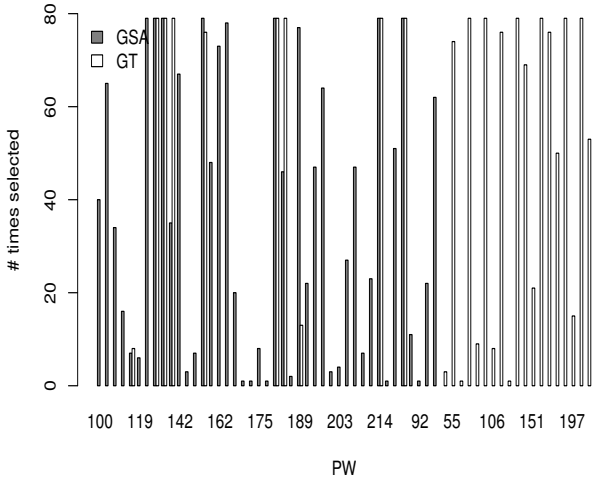

Fig. 10. Features selected in consecutive iterations of CV with corresponding frequency, 20 PWs.

method tends to produce the same features even if data is (slightly) different. On the other hand, if a change of just one sample in the training data leads to many different features, this should be regarded as a drawback of feature selection used.

To analyze stability of features, we first present values of the counters c for varying value of k (ie. the number of pathways selected as features) – Figs. 6–10. Fig. 6 shows that the global test kept selecting the same PW (PW with index = 184) over all iterations of CV, while GSA selected 6 different features in 79 iterations, with the feature with index = 86 selected most frequently. (It should be noted that pathways in Figs. 6–10 are denoted by indices to the table of KEGG pathways, as given e.g. in the Bioconductor hgu95av2.db package. For instance, the pathway index=184 corresponds to the KEGG pathway ID 05130.)

For $k = 3$ (Fig. 7) global test always selects 2 features (184 and 130) and 77 out to 79 times also feature 214. The GSA selects 10 different features (with the winner PW=86 selected 75 times). Similar observations hold for Figs. 8–10.

Although in the method proposed we do not check if p-values of the selected PWs are significant (as we rely on feature selection capabilities of the classifier used next), we observe in this study that all the top k PWs selected in step 3 are significant under multiple testing adjustment for k up to 20 (Holm-adjusted p-values below 0.05). It should be also noted that since the number of items in PWDB is much smaller than the number of genes (e.g., KEGG database includes ca. 200 pathways, as compared with ca. 10^4 genes on a typical microarray), multiple testing correction in the proposed method will be much less conservative than in standard gene selection procedures.

TABLE I
NUMBER OF DIFFERENT PWS SELECTED AS FEATURES (NF) AND MEAN FREQUENCY OF PW SELECTION IN THE CV PROCEDURE AS A FUNCTION OF THE NUMBER OF TOP PWS (K IN STEP 3 OF THE ALGORITHM)

No of top PWs (k)	GSA		GT	
	NF	mean freq (%)	NF	mean freq (%)
1	6	16.7	1	100.0
3	10	30.0	4	75.0
5	17	29.4	8	62.5
10	29	34.5	15	66.7
20	43	46.5	29	69.0
30	55	54.5	37	81.1

Analysis of stability can be summarized by using two measures calculated from the tables of counters c (see step 5 of algorithm):

$$NF = \sum_{i=1}^{d} I(c\,(PW_i) > 0) \tag{4}$$

which gives the overall number of different features selected in $n = 79$ rounds of cross validation, and

$$\textit{mean freq} = \frac{1}{n\,NF} \sum_{i=1}^{d} c\,(PW_i) = \frac{k}{NF} \tag{5}$$

which shows mean frequency of selection of features in the set of *NF* different features in n rounds of cross validation. It should be noted that for fixed value of k the second measure does not provide any more information than *NF*, however this measure is convenient to highlight the difference between the methods. Results for $k \in \{1, 3, 5, 10, 20, 30\}$ are given in Tab. 1, with the mean frequency expressed as percentage. As already observed in Fig. 6, global test always selected one feature ($NF = 1$, hence the frequency of its selection is 100%, ie. 79 times in 79 rounds of CV); the GSA selects 6 features, each with mean frequency equal ca. 17% (which translates into 13 times a feature is selected in 79 rounds of CV). For growing $k = 1, 3, 5$ we observe decreasing mean frequency, which suggests that a growing number of weaker features start getting included, which leads to less consistent feature selection when data changes. It is interesting to notice that for $k > 5$ the mean frequency again increases, however explanation of this effect requires further investigation.

The final conclusion from stability analysis is clear: the sets of features selected by global test seem more stable as compared to GSA-based features. This characteristic of the global test may account for better predictive performance of these features.

V. CONCLUSIONS

In this work two methods of gene set analysis were empirically compared in terms of predictive performance of classifiers built using most activated pathways as features. The methods realize different approaches to pathway analysis: global test is a self-contained algorithm and the GSA is a competitive method. The comparative study brings several

interesting observations. First, the self-contained method outperforms the competitive approach in terms of classification error. Second, features selected by the self-contained method appear more stable if data is modified. This is a remarkable characteristic of this feature selection procedure, as due to high dimensionality and small number of samples used for microarray class prediction, standard methods of feature selection demonstrate poor stability. Finally, the methods do select different features (pathways) for prediction (although some overlapping is observed). Based on these results a number of interesting questions and directions for further research can be raised. First, it seems interesting to investigate characteristics of the features returned by the self-contained and competitive methods: whether the activation of pathways is due to weaker effect observed consistently over a large number of member genes, or on the contrary - the method favors stronger local effect in the set of member genes. Next, a hybrid method is worth considering which would combine the strong points of these two approaches. Further research is also necessary into how features can be generated in a more sophisticated way out of the set of most activated pathways. Also, the proposed method requires more comprehensive validation based on further microarray datasets as well as simulated data. Although the purpose of this work was to compare different approaches to gene set analysis in terms of quality of feature selection, another interesting direction for further research involves comparison of these prior biological knowledge based methods with regularized learning techniques such as ridge regression, lasso or elastic net.

REFERENCES

[1] A.J. Adewale, I. Dinu, J.D. Potter, Q. Liu and Y. Yasui, "Pathway analysis of microarray data via regression," *J. Comput. Biol.,* vol. 15, 2008, pp. 269–277.

[2] D.B. Allison, X. Cui, G.P. Page, M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nature Reviews Genetics,* 7, 2006, pp. 55–65.

[3] S. Chiaretti, et al., "Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survical," *Blood,* vol. 103, 2004, pp. 2771–2778.

[4] I. Dinu, et al., "Improving gene set analysis of microarray data by SAM-GS," *BMC Bioinformatics,* 8, 242, 2007.

[5] I. Dinu, et al., "Gene-set analysis and reduction," *Briefings in Bioinformatics,* vol. 10, 2008, pp. 24–34.

[6] S. Draghici, P. Khatri, R.P. Martins, G.C. Ostermeier, S.A. Krawetz, "Global functional profiling of gene expression," *Genomics,* vol. 81, 2003, pp. 98–104.

[7] S. Dudoit, J. Fridlyand, P. Speed, "Comparison of discriminant methods for classification of tumors using gene expression data," *Journal of American Statistical Association,* vol. 192, 2005, pp. 77–87.

[8] B. Efron, R. Tibshirani, "On testing the significance of sets of genes," *Ann. Appl. Stat.,* vol. 1, 2007, pp. 107–129.

[9] K. Fujarewicz, M. Wiench, "Selecting differentially expressed genes for colon tumor classification," *Int. J. Appl. Math. Comput. Sci.,* vol. 13, 2003, pp. 327–335.

[10] A.M. Glas, "Converting a breast cancer microarray signature into a high-throughput diagnostic test," *BMC Genomics,* 7:278, 2006.

[11] J.J. Goeman, S.A. van de Geer, F. de Kort, H.C. van Houwelingen, "A global test for groups of genes: testing association with a clinical outcome," *Bioinformatics,* vol. 20, 2004, pp. 93–99.

[12] J.J. Goeman, P. Buehlmann, "Analyzing gene expression data in terms on gene sets: methodological issues," *Bioinformatics,* vol. 23, 2007, pp. 980–987.

[13] J. Khan, et al., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Med.,* vol. 7, 2001, pp. 673–679.

[14] Y.H. Lin, "Multiple gene expression classifiers from different array platforms predict poor prognosis of colorectal cancer," *Clin. Cancer Res.,* vol. 13, 2007, pp. 498–507.

[15] H. Maciejewski, "Quality of feature selection based on microarray gene expression data," in Int. Conference on Computational Science 2008, LNCS, vol. 5103, pp. 140–147.

[16] H. Maciejewski, P. Twaróg, "Model instability in microarray gene expression class prediction studies," in Eurocast 2009, LNCS, vol. 5717, pp. 745–752.

[17] H. Maciejewski, "Class prediction in microarray studies based on activation of pathways," in Hybrid Artificial Intelligence Systems 2011, LNAI, vol. 6678, pp. 321–328.

[18] U. Mansmann, R. Meister, "Testing differential gene expression in functional groups: Goeman's global test versus an ANCOVA approach," *Methods of Information in Medicine,* vol. 44, 2005, pp. 449–453.

[19] F. Markowetz, R. Spang, "Molecular diagnosis. Classification, Model Selection and Performance Evaluation," *Methods Inf. Med.,* vol. 44, 2005, pp. 438–443.

[20] V.K. Mootha, et al., "PGC-1 alpha-responsive genes involved in oxidative phoshporylation are coordinately downregulated in human diabetes," *Nature Genetics,* vol. 34, 2003, pp. 267–273.

[21] C.H. Ooi, P. Tan, "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data," *Bioinformatics* vol. 19, 2003, pp. 37–44.

[22] S. Peng, et al., "Molecular classifcation of cancer types from microarray data using the combination of genetic algorithms and support vector machines," *FEBS Letters* vol. 555, 2003, pp. 358–362.

[23] A. Subramanian, et al., "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci. USA,* vol. 102, 2005, pp. 15545–15550.

[24] J. Tomfohr, J. Lu, T.B. Kepler, "Pathway level analysis of gene expression using singular value decomposition," *BMC Bioinformatics,* 6, 225, 2005.

[25] M.C. Wu, X. Lin, "Prior biological knowledge-based approaches for the analysis of genome-wide expression profiling using gene sets and pathways," *Statistical Methods in Medical Research,* vol. 18, 2009, pp. 577–593.