A Comparative Evaluation of Gene Set Analysis Techniques in Predictive Classification of Expression Samples

Matej Holec, Filip Zelezny, Jiri Klema
Czech Technical University in Prague
{holecm1,klema,zelezny}@fel.cvut.cz

Jakub Tolar
University of Minnesota, Minneapolis
tolar003@umn.edu

Abstract

We demonstrate how some recently developed techniques of set-level gene expression data analysis may be exploited in the context of predictive classification of gene expression samples for the tasks of attribute selection and extraction. With four benchmark gene expression datasets, we empirically test the influence of these method on the predictive accuracy of constructed classification models in a comparative setting. Our results mainly indicate that gene set selection methods (SAM-GS and the global test) can boost the predictive accuracy if used with caution.

1. Introduction

Set-level techniques have recently prevailed in the area of gene expression data analysis [7]. Whereas in traditional analysis approaches one typically seeks individual genes differentially expressed across sample classes, the set-level approach aims to identify entire sets of genes significant e.g. in the sense that they contain an unexpectedly large number of differentially expressed genes. The gene sets considered for significance testing are defined prior to analysis, using appropriate biological background knowledge. For example, each considered gene set may contain genes acting in a given cellular pathway. The main advantage brought by set-level analysis is the improved interpretability of analysis results. Indeed, long lists of differentially expressed genes characteristic for traditional expression analysis are replaced by shorter lists of more informative units corresponding to entire biological processes.

Predictive classification is a form of data analysis going beyond the mere identification of differentially expressed units. Here, units deemed significant for the discrimination between sample classes (e.g., heathly and ill) are assembled into formal models prescribing how to determine the class of new samples whose class labels are so far unknown. Predictive classification techniques are thus especially relevant to diagnostic tasks and as such have been explored since very early studies on microarray data analysis [4]. Predictive models may take diverse forms. A class of predictive models with an especially notable track record in expression data modeling is known as *support vector machines* [9]. Another family of techniques aims at making the resulting models instantly readable by a human. Here, models acquire forms such as decision trees or logical rules [2].

The combination of set-level techniques with predictive classification has not yet been thoroughly explored. In our recent study [5] we have demonstrated that the set-level approach enables to construct predictive models applicable to expression samples assorted from diverse microarray platforms. However, the set-level strategy carries even a more direct promise for predictive classification. We hypothesize that it will reduce the risk of *overfitting*, a well known problem in expression data classification. Overfitting denotes the effect where a model classifies training samples (i.e., the samples available for guiding the construction of the model) accurately, but fails to accurately classify previously unseen samples. The risk of overfitting grows with the number of sample attributes, that is, the variables for which the sample provides values. Thus a straightforward way to mitigate the risk is to reduce the number of attributes provided that relevant information is not lost as a consequence. We speculate that replacing the original attributes corresponding to genes by attributes corresponding to gene sets (whose amount is typically much smaller) will provide exactly such a reduction. Testing this hypothesis is the first goal of the present paper.

Furthermore, recent developments in set-level analysis of gene expression data have yielded techniques which, as we show here, are directly exploitable in the process of predictive model construction. One family of such methods aims at detecting gene sets significant for the discrimination between sample classes. State of the art representatives of this family are the SAM-GS algorithm [1] and a method known as the global test [3]. Relying on these methods, the classifier construction algorithm can be forced to focus on gene sets actually relevant to the discrimination task rather than consider all available gene sets in the process of build-

ing a classifier. Here, we intend to test how the final classification accuracy is influenced by the choise of one of the two mentioned algorithms for gene set pre-selection. In addition, we test the influence of these methods also in the gene based (rather than gene set based) setting, where attributes correspond to genes occurring in the gene sets selected by the respective methods.

By definition, each attribute must carry an unambigous value for each data sample. When attributes are genes, as in the traditional approach, the value is simply the (normalized) measured expression of the gene corresponding to the attribute in the given sample. When attributes are gene sets, as in the set-level approach, a question arises as to how to establish a value for each gene set and sample. In [5] we simply assigned the average of expressions of the genes contained in the given set. A more sophisticated method based on the statistical technique known as singular value decomposition was proposed by [8]. The final contribution of this study is in testing the whether the utilization of the latter technique improves the ultimate classification performance when compared to the former approach based on simple averaging.

The rest of the paper is organized as follows. The next section describes the specific methods and data sets used in our experiments. In Section 3 we expose the experimental results. Section 4 summarizes the main conclusions and proposes directions for follow-up research.

2. Methods and Data

Here we first describe the techniques adopted initial processing of expression sample, i.e., for selecting significant gene sets to act as sample attributes, and for determining the values of these attributes. When attributes are selected and instantiated, predictive models can be constructed; the methods employed for this sake are reviewed subsequently. Next we describe the data sets used as benchmarks in the comparative experiments. Lastly, we explain the protocol followed by our experiments.

2.1 Gene Set Selection

Two methods are considered for gene set selection, both of which have been proposed recently to avoid some substantial deficiencies of the previously popular technique of *gene set enrichment analysis* [7]. As inputs, both of the methods assume a set G of n interrogated genes, and a set S of m expression samples where for each $s_i \in S$, $s_i = (e_{i,1}, e_{i,1}, \ldots e_{i,n}) \in \mathsf{R}^n$ where $e_{i,j}$ denotes the (normalized) expression of gene j in sample i. The sample set S is partitioned into classes $S = C_1 \cup C_2 \cup \ldots \cup C_o$ so that $C_i \cap C_j = \{\}$ for $i \neq j$. For simplicity in this paper we assume binary classification, i.e. $o = 2$. A further input is a collection of gene sets $\mathcal{G}$ such that for each $GS \in \mathcal{G}$ it holds $GS \subseteq G$. As the output, each of the two methods ranks all gene sets in $\mathcal{G}$ by their estimated power to discriminate samples into classes. Subsequently we take the first k top-ranking gene sets (we namely consider the cases $k = 1$ and $k = 10$).

The specific methods used to obtain the ranking are the global test [3] and the SAM-GS technique [1]. Here we give a brief informal description of these methods and refer to the original sources for a rigorous treatment. Each sample s_i is viewed as a point in an n-dimensional Euclidean space. Each gene set $GS \in \mathcal{G}$ defines its $|GS|$-dimensional subspace in which projections s_i^{GS} of samples s_i are given by coordinates corresponding to genes in GS. Both methods judge a given GS by how distinctly the clusters of points $\{s_i^{GS} | s_i \in C_1\}$ and $\{s_j^{GS} | s_j \in C_2\}$ are separated from each other in the subspace induced by GS. SAM-GS measures the Euclidean distance between the centroids of the respective clusters and applies a permutation test to determine whether, and how significantly this distance is larger than one obtained if samples were assigned to classes randomly. The global test rather proceeds by fitting a regression function in the subspace, such that the function value acts as the class indicator. The degree to which the two clusters are separated then corresponds to the magnitude of the coefficients of the regression function.

2.2 Gene Set Value Assignment

Two methods are considered for the sake of assigning a value to a given gene set GS for a given sample s_i. The first is a baseline method adopted from [5] which simply produces the average of the expressions of all GS genes in sample s_i. The value assigned to the pair (s_i, GS) is thus independent of samples s_j, $i \neq j$.

A more sophisticated approach was proposed in [8]. Here, the value assigned to (s_i, GS) depends on other samples s_j. In particular, all samples in the sample set S are viewed as points in the $|GS|$-dimensional Euclidean space induced by GS the same way as explained in Section 2.1. Subsequently, the specific vector in the space is identified, along which the sample points exhibit maximum variance. Each point $s_k \in S$ is then projected onto this vector. Finally, the value assigned to (s_i, GS) is the real-valued position of the projection of s_i on the maximum-variance vector in the space induced by GS. Again, we refer to [8] for detailed explanation.

2.3 Predictive classification

In all experiments, the *support vector machine* [9] classifier type was used. This choise was motivated by the prevailing use of this classifier type in the area of gene expres-

Dataset	Genes	Class 1	Class 2
heme/stroma	13380	18	33
brain/muscle	13380	41	20
diabetes	13380	17	17
p53	10101	33	17

Table 1. Number of genes interogated and number of samples in each of two classes of each benchmark dataset.

sion data modeling. In each experiment, a particular classifier was constructed from training data through the SMO algorithm implemented in the public machine-learning software suite WEKA [10].

2.4 Datasets

We conduct our experiments using four public datasets, each containing gene expression samples pertaining to two classes. To avoid bias, we deliberately combined two 'easy' datasets (the first two) where phenotype classes are very distinct with two 'difficult' dataset where the separation is less straightforward. Table 1 shows, for each dataset, the number of samples in each class and the number of interogated genes.

Classes in the first dataset (*heme/stroma*) correspond to blood-forming (hematopoietic) and supportive (stromal) cellular compartments in the bone marrow, respectively. In the second dataset (*brain/muscle*) contains samples from skeletal muscle and brain. All samples in the first two datasets were manually collected from the NCBI gene expression omnibus database.

The third dataset (*diabetes*) come from [6]. From the original dataset, we extracted a two-class subset. The first class corresponds to patients with diabetes mellitus 2 and the second class pertains to healthy patients. The last dataset (*p53*) is adopted from [1] and contains expressions for 50 cell lines from the NCI-60 collection of cancer cell lines, for which mutational status of the p53 gene has been reported, divided into the wild-type class and a class containing cell lines carrying mutations in the gene.

In all experiments we work with a single family of gene sets. All of them are sets of genes acting in respective cellular pathways. The gene sets are fully taken from the study [1].

2.5 Experimental Protocol

Our criterion to evaluate a particular combination of gene set selection and value assignment method is the predictive accuracy (i.e., the proportion of correctly classified samples) achieved on the benchmark datasets by that combination. To estimate predictive accuracy, we use the standard procedure of 10-fold stratified cross-validation. The specific steps conducted to estimate the accuracy for a single benchmark dataset is shown in Fig. 1. To preserve methodological correctness, the process of gene set selection is embedded inside the cross-validation loop along with the classifier construction step. In other words, gene sets are selected only on the training splits rather than on all data. The final accuracies used to rank the individual combinations of methods are further averaged over all 4 benchmark domains, that is, they are averages over 40 experiments.

Partition the input sample set S into 10 folds $f_1, f_2, \ldots f_n$ of equal size such that the class-proportion is (approximately) equal in all folds.
for $i = 1, 2, \ldots 10$ **do**
  $S^{train} \leftarrow S \setminus f_i$
  $S^{test} \leftarrow f_i$
  Select n top-ranking gene sets on S^{train}
  Construct a classifier C with S^{train} using the selected gene sets
  A_i = classification accuracy of C on S^{test}
end for
return Average of $A_1, A_2, \ldots A_{10}$

Figure 1. A skeleton of the stratified cross-validation procedure used to obtain accuracy estimates for on a single sample set.

The algorithm in Fig. 1 has 4 degrees of freedom accomodating the various combinations of methods to be tested. In particular, the gene set selection step is either conducted by the global test method or the SAM-GS method, as described earlier. Variable n is either instantiated to 1 or 10. The classifier construction step is either performed with attributes corresponding to all genes found in the selected gene sets, or with attributes corresponding to the gene sets themselves. In the latter case, the attributes are assigned values by one of the two earlier described methods of value assignment. Alltogether, we test 12 different combinations of methods. These combinations are enumerated in Table 2 (exclude columns 1 and 6, and line 8).

Additionally, we also estimate the accuracy of the baseline method, where a classifier is constructed with attributes corresponding to all genes present in the original sample representation, i.e. no gene set selection is performed at all. This accuracy is also estimated via 10-fold stratified cross-validation and futher averaged of all four benchmark domains.

3. Results

A preliminary question that should be addressed before we investigate the particular ranking of methods is whether at all there is a clear influence of the gene set selection method on the final classification accuracy. To this end, the diagram in Fig. 2 plots the cross-validated accuracies achieved with a single selected gene set as a function of the rank of the gene set produced by the gene set selection method. To obtain such an accuracy estimate for gene set rank r, the step "Select n top-ranking gene sets" in Fig. 1 is replaced by "Select one gene set ranking r-th, and the algorithm is run for all combinations of methods and all 4 benchmark datasets, averaging the resulting estimates. The diagram shows a clear trend of accuracy falling as the gene set rank increases, i.e. the estimated class-discimination power of that gene set drops.

In Fig. 3 we also show the average number of genes in a gene set in dependence on the rank of the set. These quantities are also averaged over cross-validation folds, all combinations of methods and all four datasets. This figure is shown to illustrate i) the degree of attribute compression incurred by changing sample representation from gene attributes to gene set attributes, and ii) the number of attributes used for classifier construction in cases where attributes correspond to genes. Any trends possibly observed in the shown dependency are not of interest in our study.

Finally, Table 2 provides the ranking of all 13 tested combinations of methods according to the estimated predictive accuracy. The **Attrib** column indicates whether selected gene sets ('sets') or genes extracted from the selected gene sets ('genes') were used as attributes to describe samples. The **# Sets** column denotes the choice of the n parameter in Fig. 1. The **# Select** column captures the employed gene set selection method. In Column **Assign**, the method for gene set value assignment is listed, where AVG corresponds to simple averaging used in [5] and SVD stands for the method proposed by [8]. The final column shows the estimated predictive accuracy.

The principal trends observed are as follows.

- Methods based on the selection of 10 best gene sets systematically outperform the baseline method, indicating the positive influence of the gene set selection process performed prior to classifier induction.
- Conversely, relying only on the 1 top-ranking gene set leads to poor predictive accuracies, indicating that a single selected gene set does not capture enough information for to induce a reliable classifier.

The ranking obtained is however inconclusive in terms of the following comparisons:

- the two considered methods for gene set selection

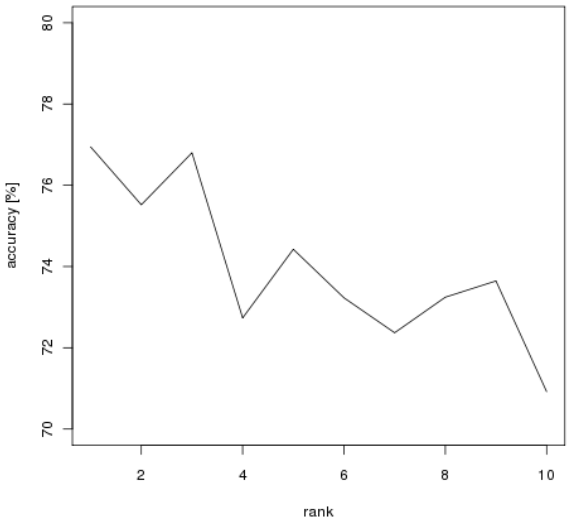

Figure 2. Average predictive accuracy tends to fall as lower-ranking gene sets are used to constitute attributes (see text for details).

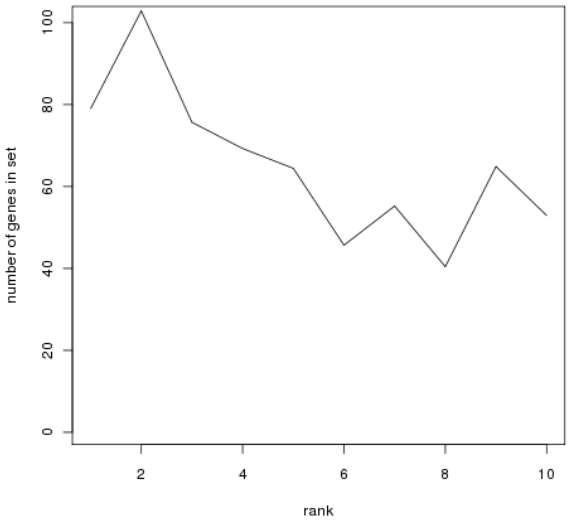

Figure 3. Average gene set size as a function of gene set rank.

Rank	**Attrib**	**# Sets**	**Select**	**Assign**	**Acc**
1	sets	10	SAM-GS	SVD	90.45
2	genes	1	GLOBAL	n/a	84.74
3	sets	10	SAM-GS	AVG	84.18
4	sets	10	GLOBAL	AVG	84.08
5	sets	10	GLOBAL	SVD	83.21
6	genes	10	GLOBAL	n/a	82.25
7	sets	1	GLOBAL	SVD	81.08
8	baseline	n/a	n/a	n/a	80.3
9	genes	10	SAM-GS	n/a	80.08
10	sets	1	SAM-GS	AVG	78.91
11	genes	1	SAM-GS	n/a	78.35
12	sets	1	SAM-GS	SVD	67.12
13	sets	1	SAM-GS	AVG	60.87

Table 2. Final ranking of all combinations of methods in terms of average predictive accuracy (see text for details).

- the two considered methods for gene set value assignment
- whether selected gene sets, or genes extracted from the selected gene sets are used as attributes

4. Conclusions and Future Work

Our experimental findings support our initial hypothesis that methods recently developed for gene set selection can be used with benefits to improve predictive accuracy of gene expression sample classification by providing a relevant attribute set to which the classifier constructor is constrained. This conclusion is however not valid in the extreme case where one only relies on a single top ranking selected gene set. In this case, the predictive accuracy actually drops, indicating that a single selected gene set does not capture enough information for to induce a reliable classifier.

Our experiment however did not provide conclusive answers to our further questions regarding the mutual ranking of the two considered gene set selection methods (SAM-GS [1] and the global test [3]) and the two gene set value assignment methods (averaging [5] and singular value decomposition [8]).

A further survey performed on a significantly larger collection of gene expression benchmarks is needed to answer the questions left open. Such a study consitutes our next research steps.

We believe that our present study, albeit preliminary, represents the first steps towards the important goal of determining how the recently developed methods of set-level based gene expression data analysis can contribute to tasks of predictive classification. Getting such an insight is significant mainly due to the direct relevance of predictive data modeling tasks to clinical diagnosis procedures, and thus - in the longer term - to the proliferation of personalized medicine.

Acknowledgements MH is supported by the Czech Science Foundation through project 201/09/1665. FZ is supported by the Czech Ministry of Education through project ME910.

References

[1] I. Dinu. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, 2007.

[2] D. Gamberger, N. Lavrac, F. Zelezny, and J. Tolar. Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. *Journal of Biomedical Informatics*, 34(4):269–284.

[3] J. J. Goeman and P. Buhlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 2007.

[4] T. R. Golub, D. K. Slonim, P. Tamayo, M. G. C. Huard, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[5] M. Holec, F. Zelezny, J. Klema, and J. Tolar. Integrating multiple-platform expression data through gene set features. In *The 5th International Symposium on Bioinformatics Research and Applications (ISBRA 2009)*. Springer, 2009.

[6] V. Mootha, C. Lindgren, and S. L. et al. Pgc-1-alpha-responsive genes involved in oxidative phosphorylation are coorinately down regulated in human diabetes. *Nature Genetics*, 34:267–273, 2003.

[7] A. Subramanian. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 2005.

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

[9] V. N. Vapnik. *The Nature of Statistical Learning*. Springer, 2000.

[10] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco, 2005.