JMLR: Workshop and Conference Proceedings 8: 55-64 Machine Learning in Systems Biology

On utility of gene set signatures in gene expression-based cancer class prediction

Minca Mramor MINCA.MRAMOR@FRI.UNI-LJ.SI
Faculty of Computer and Information Science
University of Ljubljana, Slovenia

Marko Toplak MARKO.TOPLAK@FRI.UNI-LJ.SI
Faculty of Computer and Information Science
University of Ljubljana, Slovenia

Gregor Leban GREGOR.LEBAN@FRI.UNI-LJ.SI
Faculty of Computer and Information Science
University of Ljubljana, Slovenia

Tomaž Curk TOMAZ.CURK@FRI.UNI-LJ.SI
Faculty of Computer and Information Science
University of Ljubljana, Slovenia

Janez Demšar JANEZ.DEMSAR@FRI.UNI-LJ.SI
Faculty of Computer and Information Science
University of Ljubljana, Slovenia

Blaž Zupan BLAZ.ZUPAN@FRI.UNI-LJ.SI
Faculty of Computer and Information Science
University of Ljubljana, Slovenia and
Dept. of Human and Mol. Genetics,
Baylor College of Medicine, Houston, USA

Editor: Sašo Džeroski, Pierre Geurts, and Juho Rousu

Abstract

Machine learning methods that can use additional knowledge in their inference process are central to the development of integrative bioinformatics. Inclusion of background knowledge improves robustness, predictive accuracy and interpretability. Recently, a set of such techniques has been proposed that use information on gene sets for supervised data mining of class-labeled microarray data sets. We here present a new gene set-based supervised learning approach named SetSig and systematically investigate the predictive accuracy of this and other gene set approaches compared to the standard inference model where only gene expression information is used. Our results indicate that SetSig outperforms other gene set approaches, but contrary to earlier reports, transformation of gene expression data to the space of gene set signatures does not result in increased accuracy of predictive models when compared to those trained directly from original (not transformed) data.
Keywords: Microarrays, Classification, Gene sets

1. Introduction

Methods to incorporate additional *domain knowledge* in the model inference process have from its early ages been central to machine learning research. Also referred to as *background knowledge*, its inclusion should increase model stability, predictive accuracy and interpretability.

In systems biology the sources of domain knowledge abound. They include information on gene structure and annotation, protein interactions, tissue localization, biological pathways, literature references, and other. From the onset of high-throughput data acquisition, bioinformatics has striven to include such additional knowledge in the discovery process. Consider, for instance, genome-wide gene expression analysis. From the first reports on utility of computational techniques such as clustering, the relevance of results was confirmed using function annotations (Eisen et al., 1998). Later, the procedure was formalized in *enrichment analysis*, where knowledge on groups of related genes, called *gene sets*, was used to identify groups including either over or under-expressed genes under specific experimental conditions (Subramanian et al., 2005). Reporting enriched gene sets, rather than a list of differentially expressed genes, should yield stability, improve robustness across data sets of the same kind coming from different sources (labs), and help us in gaining a deeper understanding of the underlying processes due to identification of affected pathways (Nam and Kim, 2008).

Gene set enrichment is by definition an explorative data analysis technique. If the task in genome-wide microarray analysis is class prediction, such as tumor classification, diagnosis and prognosis, standard supervised machine learning techniques should be used instead (Simon et al., 2003). Early efforts in this domain directly applied machine learning to class-labeled expression data (Brown et al., 2000) and used gene expressions as features. Recently, a number of techniques have been proposed to incorporate the knowledge on gene sets in the model inference process, where each individual observation (*e.g.* tissue sample) should be described by features (*signatures*) that correspond to gene sets. These are computed from expression of its constituents (genes) and are then used for model inference. At present, these approaches can be classified based on whether they use class information when computing the signatures. Approaches that do not use class information include methods that compute average gene set expression (Guo et al., 2005), use principal component analysis (PCA) (Liu et al., 2007) or singular value decomposition (Tomfohr et al., 2005; Bild et al., 2006), while domain-enhanced analysis with partial least squares (Liu et al., 2007), PCA with relevant gene selection (Chen et al., 2008), activity scores based on condition-responsive genes (Lee et al., 2008), averages of expression values of genes supporting the gene set score (Efron and Tibshirani, 2007) and ASSESS (Edelman et al., 2006) do.

Similarly to gains in enrichment analysis, gene sets-based inference of predictive models should improve the stability and predictive accuracy. Interestingly, however, this has not yet been systematically tested across larger collections of data sets and across different methods. Also, there is a lack of a thorough comparison of such approaches with standard machine learning from the entire set of genes.

In the paper, we demonstrate the stages of development of a gene set-based supervised learning approach in crafting our own one (SetSig), and then report on systematic investigation to determine if this and five other knowledge-based techniques produce more accurate

predictive models. Our test-bed incorporates 30 publicly available data sets, and uses standard evaluation and modelling procedures from supervised data mining. The results of our analysis are quite surprising and contradict initial reports on the superiority in accuracy of gene set-based predictive modelling (Lee et al., 2008; Efron and Tibshirani, 2007; Edelman et al., 2006).

2. Methods

2.1 Data sets

The study considered 30 cancer gene expression data sets from the Gene Expression Omnibus (GEO) (Barrett et al., 2007). All data sets have two diagnostic classes and include at least 20 samples, where each class was represented by at least 8 data instances. On average, the data sets include 44 instances (s.d.= 29.6). The GDS data sets with following ID numbers were used: 806, 971, 1059, 1062, 1209, 1210, 1220, 1221, 1282, 1329, 1375, 1390, 1562, 1618, 1650, 1667, 1714, 1887, 2113, 2201, 2250, 232, 2415, 2489, 2520, 2609, 2735, 2771, 2785 and 2842.

All data sets were preprocessed in the same manner. First, the probes measuring the expression of the same gene were joined and the average value of the expression over all probes was used. Second, in all data sets the gene expression values for each gene were normalized to zero mean ($\mu = 0$) and unit variance ($\sigma^2 = 1$).

2.2 Gene sets

We used the gene sets from the Molecular signatures data base (MSigDB v2.5) (Subramanian et al., 2005). MSigDB includes five collections of gene sets that differ in the prior knowledge or the computational method used for creating them. We have considered collections C2 and C5, where gene sets where composed based on prior biological knowledge. From these we selected gene sets that include at least five genes for which the gene expression information was provided in the explored data set. Also, large (and possibly non-specific) gene sets that included more than 200 such genes were excluded from the analysis. As a result of this filtering, we used the following gene sets:

- C2cp: 639 gene sets belonging to canonical pathways (C2 collection). These gene sets are compiled by domain experts from the pathway data bases and are usually canonical representations of a biological process.
- C2C5: gene sets from the biological process and molecular function part of gene ontology (C5 collection) in addition to gene sets from C2cp. Depending on the number of genes in the specific data set, approximately 1.600 gene sets covering up to 7.900 genes met these criteria.

2.3 SetSig: sample characterization by gene set signatures

We here describe SetSig, a new approach to summarizing gene expression data into features based on gene sets. Our primary motivation was to construct a relatively simple method that does not rely on linear transformations and on search for gene groups within gene subsets which can potentially lead to overfitting.

Gene expression data consists of a number of samples S described by gene expressions, $f_S(g)$ (where g represents a gene) and the class value. SetSig transforms the data so that samples are described by gene set signatures, $f_S(G)$ (where G is a gene set) computed from the original gene expressions. The procedure for computation of $f_S(G)$ for a particular sample S and gene set G goes as follows:

1. Let C_1 and C_2 be sets of samples belonging to the first and to the second class, respectively.

2. Calculate the Pearson correlation coefficient between the expressions of genes from gene set G in the sample S and every sample from C_1 and from C_2. For a given gene set G, let R_1 and R_2 then be the corresponding sets of correlation coefficients, that is

$$R_1 = \{r_G(S, C) : C \in C_1\}, \qquad R_2 = \{r_G(S, C) : C \in C_2\},$$

   where $r_G(S, C)$ is the correlation between $f_S(g_i)$ and $f_C(g_i)$ for $g_i \in G$.

3. The genes set G's signature for sample S, $f_S(G)$, is then computed as the Student's t-statistics for difference between R_1 and R_2:

$$f_S(G) = \frac{\overline{R_1} - \overline{R_2}}{\sqrt{s^2_{R_1}/N_1 + s^2_{R_2}/N_2}},$$

   where N_1 and N_2 are the number of samples in C_1 and C_2, respectively.

Intuitively, coefficients in R_1 are high (low) if expressions of genes from gene set G in the sample S are similar to (different from) expressions of these genes in the samples from the first class. Coefficients in R_2 describe the similarities (differences) for the second class. Student's t-test measures whether the coefficients in R_1 differ from those in R_2 that is, how important are the genes from G for distinguishing between the two classes. The sign of the t-statistic is positive (negative) if the particular sample's gene expressions are more similar to those of first (second) class.

This procedure is used on each sample and for each gene set. The result is a set of samples described with gene set-based features, instead of gene expressions. The samples without class values (the testing set) are not used to obtain R_1 and R_2. While SetSig directly addresses the data with binary class variable, it can be simply extended to multi-class prediction problems by construction of a separate classifier for each of the sample labels. In the paper we concentrate on the performance of the core method only and study only binary classification problems.

2.4 Other gene set signature transformation methods

In experiments we compared SetSig to other, previously published methods that use transformation of gene expression data sets to data sets comprising gene set scores. These transformations include:

1. Mean and Median (Guo et al., 2005), where each gene set is characterized with mean (median, respectively) expression of genes from the gene set.

2. ASSESS (Edelman et al., 2006) scores gene sets with a Kolmogorov-Smirnov like statistic on a list of ranked gene correlations, similarly to GSEA (Subramanian et al., 2005). While GSEA estimates correlations of genes with the class labels across all samples, ASSESS estimates these correlations individually for each sample. The correlations are estimated as the differential probabilities of the two classes. The parametric model (Edelman et al., 2006) was used for estimation of differential probabilities.

3. The first principal component of PCA (Liu et al., 2007) of genes in the gene set.

4. CORGs method selects a subset of genes from the gene set, named condition-responsive genes (CORGs), whose activity scores (averages across expression values) differentiate between class labels (Lee et al., 2008). Contrary to previously mentioned methods, only a subset of the gene set is used for data transformation. The CORGs are selected greedily starting with genes with the highest t-scores until the quality estimate of the subset improves.

2.5 Estimation of predictive accuracy, classification, evaluation of results

Different supervised learning methods have been used to build class prediction models in the space of gene set signatures and in the space of gene expressions. Models were built with support vector machines (SVMs) with linear kernel, a naive Bayesian classifier, a k-nearest neighbor learner, and a logistic regression learner. We report on the results for the SVM and logistic regression models, which outperformed models built with other supervised learning approaches. The results of other tested class prediction methods show similar trends.

We used leave-one-out validation to estimate the area under ROC curve (AUC) of the tested models. As some gene set transformation methods build internal data models, only the learning set was used to induce such models. The same evaluation procedure was used across the entire set of 30 data sets. For each data set, the various methods were ranked. Statistical significances of differences between average ranks of tested methods were evaluated with the Nemenyi test and were visualized with critical distance graphs (Demšar, 2006).

All supervised learning approaches were used as embedded in Orange data mining environment (Demšar et al., 2004). Orange was also used to implement SetSig and re-implement all other gene set-based supervised learning procedures investigated in this report.

3. Results

We first compared the predictive accuracy of class prediction models using SetSig transformed data sets with the C2cp and C2C5 gene set subsets with predictive accuracy of the models built with original gene expression data. For the latter, no feature selection or any additional data transformation was used. Figure 1 shows that SVM models built with original data sets perform significantly better than SetSig on the C2cp subset and better (but not significantly) for the C2C5 subset. As expected, SetSig performs better with larger number of gene sets (more biological knowledge), albeit the difference was not significant.

Figures 2 and 3 include the results for all gene set-based transformations listed in Sec. 2.4 for the SVM and logistic regression models, respectively. Gene sets in C2C5 were used as the

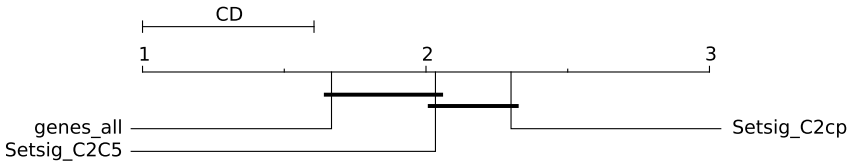

Figure 1: Critical distance graph showing the average AUC ranks of SVM models on original gene expression data sets(genes_all) and data sets transformed by SetSig (either with gene sets C2cp or C2C5). Methods connected with bold lines are not significantly different ($\alpha = 0.05$).

models built with them performed better in the experiments with SetSig reported above. Nemenyi test identifies two groups of insignificantly different methods connected with a bold line in Figure 2. Inference from gene expression without gene set transformation performs best, although the difference is only significant for two of the six gene set-based methods (PCA and ASSESS). The difference between all gene set methods is statistically insignificant. Of all the tested methods, SetSig performed best. Similar trends can be observed in Figure 3 for the models built with logistic regression. Again, models built with the original gene expression data preform best. The difference in the average ranks is significant for two of the gene set transformation methods (Median and CORGs). SetSig outperforms other gene set transformation methods and is significantly better than the Median approach.

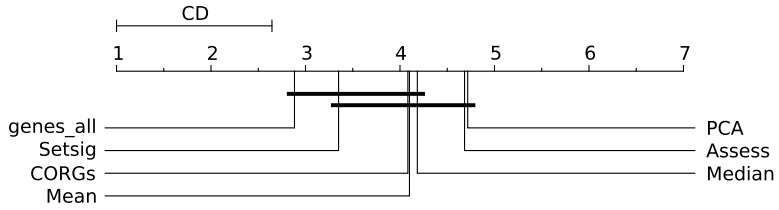

Figure 2: The average AUC ranks of SVM models on original gene expression data sets (genes_all) and transformed using a variety of gene set-based transformation methods.

Gene set-based approaches use only a subset of genes from the original expression data sets. One reason for poorer performance of these approaches could have been that some informative genes are left out. We tested this by evaluating the accuracy of predictive models built directly from gene expressions but using only a subset of genes. We have examined the following subsets in this way: (1) genes present in C2cp (genes_C2cp), (2) genes not present in C2cp (genes_notC2cp), (3) genes present in C2C5 (genes_C2C5), (4) genes not present in C2C5 (genes_notC2C5), and (5) all genes (genes_all).

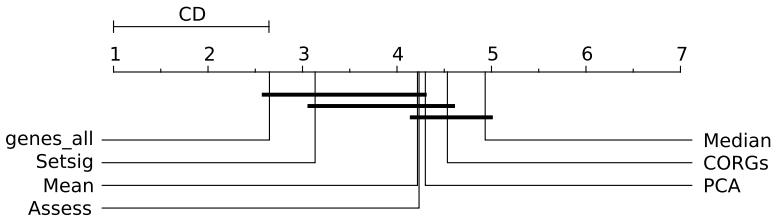

Figure 3: The average AUC ranks of logistic regression models on original gene expression data sets (genes_all) and transformed using a variety of gene set-based transformation methods.

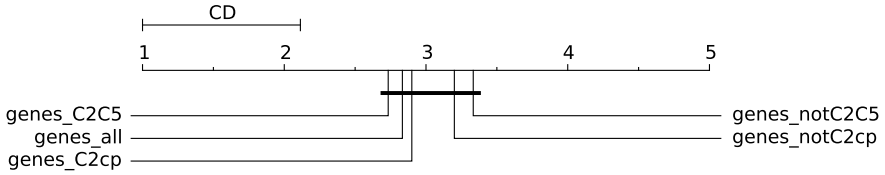

Figure 4: The average ranks of AUC-scored classifiers that use different subsets of genes. The differences are statistically insignificant.

The average ranks of the models built with the above mentioned subsets and the differences between them are shown in Figure 4. The average ranks of AUC of models built with different subsets of genes are very similar. No statistically significant differences were detected.

4. Discussion

Our experimental results indicate that transformation of gene expression data to the space of gene set signatures does not result in increased accuracy of predictive models when compared to those trained from original (not transformed) data. In fact, the latter, "gene-set free" approach consistently ranked higher in our experiments. Of all the tested gene set approaches, SetSig's performance was closest to that of using all genes.

These results come as a surprise. First, in explorative data analysis, the utility of gene sets is motivated by gains in interpretability, and also by gains in stability and robustness of results, even when compared across data sets obtained from different laboratories (Manoli et al., 2006).

Next, several recently published papers explicitly report that their gene set approaches over-perform the gene-centric approach. Closer inspection shows that these assertions are not a result of systematic study, and either used a very limited number of data sets in the study (Efron and Tibshirani, 2007; Edelman et al., 2006; Lee et al., 2008), or, as in the most recent report, are based on too restrictive gene selection (feature set selection of only a handful genes in gene-centric approach) prior to learning (Lee et al., 2008). But even

with such lack of systematic testing, all the present evidence reported votes in favor of gene set-based approaches.

Finally, we would in general (albeit naively) expect to gain with any inclusion of additional (background) knowledge in machine learning. However, in frameworks described in this paper such knowledge is used to transform, rather than complement the problem domain. We can think of a number of other reasons why the utility of gene sets with respect to predictive accuracy fails:

1. Gene sets do not include some highly class-informative genes.
2. There are too many gene sets.
3. Some gene sets are very similar to each other.
4. Gene set signature construction methods lose information.
5. Number of samples (instances) is too low to reliably estimate gene set scores.
6. Biological knowledge of the genes is incomplete. Gene sets and pathways used are not specific enough to represent biological processes that distinguish between different cancer types.

We can reject reason (1) based on results on gene-centric approach that used genes from different sets (Figure 4), where no significant differences were observed. Facts stated in (2) and (3) can hurt supervised learning, but gene-centric approaches must deal with the same kind of problems (abundance of genes, many of which are co-expressed genes). Due to (4) we have tested six different approaches, including very promising and elaborate ones such as CORGs. (5) clearly deserves further investigation. Previous studies have already shown that supervised learning methods may fail due to low sample size (Ein-Dor et al., 2005; Hanczar and Dougherty, 2008). Finally (6), despite incompleteness of biological knowledge on genes, we would expect that additional information in the form of gene sets should help us in inference of reliable classifiers, even more for the methods like CORGs which remove genes that do not contribute to class differentiation from the gene sets.

5. Conclusion

The reasons why gene set-based transformations for supervised learning from gene expression data sets fail when compared to gene-centric learning seem elusive. In fact, they do not fail, but rather – contrary to our expectations and to several recent reports – do not surpass the more standard and direct learning from gene expression profiles. Yet, predictive performance is not the only issue here, and gene set-based predictive models can significantly gain with regard to ease of interpretation and information they provide to biologists and clinicians. We have indeed observed that just like for gene-centric models (Mramor et al., 2007) we could construct very simple and highly-predictive visual models using only a few gene set signatures. We can thus conclude that knowledge on gene sets may be a useful resource for supervised microarray data analysis, but that methods for its inclusion in model inference require further studying and improvements, specifically in terms of gains in predictive accuracy.

Acknowledgements

This study was funded by the program and project grants from the Slovenian Research Agency (P2-0209, J2-9699, L2-1112).

References

Tanya Barrett, Dennis B. Troup, Stephen E. Wilhite, et al. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucl. Acids Res.*, 35:760–5, 2007.

Andrea H. Bild, Guang Yao, Jeffrey T. Chang, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, 439(7074):353–357, 2006.

Michael Brown, William Noble Grundy, David Lin, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *PNAS*, 97(1):262–7, 2000.

X. Chen, L. Wang, J. D. Smith, and B. Zhang. Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. *Bioinformatics*, 24(21):2474–81, 2008.

J Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of machine learning research*, 7(jan):1–30, 2006.

J. Demšar, B. Zupan, and G. Leban. Orange: From experimental machine learning to interactive data mining, white paper, 2004.

E. Edelman, A. Porrello, J. Guinney, et al. Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics*, 22(14):e108–16, 2006.

B Efron and R Tibshirani. On testing the significance of sets of genes. *Ann Appl Stat*, 1 (1):107–29, 2007.

Liat Ein-Dor, Itai Kela, Gad Getz, et al. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178, 2005.

Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–8, 1998.

Z. Guo, T. Zhang, X. Li, et al. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics*, 6:58, 2005.

Blaise Hanczar and Edward R. Dougherty. Classification with reject option in gene expression data. *Bioinformatics*, 24(17):1889–1895, 2008.

Eunjung Lee, Han-Yu Chuang, Jong-Won Kim, et al. Inferring pathway activity toward precise disease classification. *PLoS Comput Biol*, 4(11):e1000217, 11 2008.

J. Liu, J. M. Hughes-Oliver, and Jr. Menius, J. A. Domain-enhanced analysis of microarray data using GO annotations. *Bioinformatics*, 23(10):1225–34, 2007.

T. Manoli, N. Gretz, H. J. Grone, et al. Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics*, 22(20):2500–6, 2006.

Minca Mramor, Gregor Leban, Janez Demšar, and Blaz Zupan. Visualization-based cancer microarray data classification analysis. *Bioinformatics*, 23(16):2147–2154, 2007.

D. Nam and S. Y. Kim. Gene-set approach for expression pattern analysis. *Brief Bioinform*, 9(3):189–97, 2008.

R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane. Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst*, 95(1): 14–8, 2003.

A. Subramanian, P. Tamayo, V. K. Mootha, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102 (43):15545–50, 2005.

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