Comparative Evaluation of Set-Level Techniques in Microarray Classification

Jiri Klema[1], Matej Holec[1], Filip Zelezny[1], and Jakub Tolar[2]

[1] Faculty of Electrical Engineering, Czech Technical University in Prague
[2] Department of Pediatrics, University of Minnesota, Minneapolis

Abstract. *Analysis of gene expression data in terms of a priori-defined gene sets typically yields more compact and interpretable results than those produced by traditional methods that rely on individual genes. The set-level strategy can also be adopted in predictive classification tasks accomplished with machine learning algorithms. Here, sample features originally corresponding to genes are replaced by a much smaller number of features, each corresponding to a gene set and aggregating expressions of its members into a single real value. Classifiers learned from such transformed features promise better interpretability in that they derive class predictions from overall expressions of selected gene sets (e.g. corresponding to pathways) rather than expressions of specific genes. In a large collection of experiments we test how accurate such classifiers are compared to traditional classifiers based on genes. Furthermore, we translate some recently published gene set analysis techniques to the above proposed machine learning setting and assess their contributions to the classification accuracies.*

Keywords: gene set, classifer, learning, predictive accuracy.

1 Introduction

Set-level techniques have recently attracted significant attention in the area of gene expression data analysis [20, 9, 13, 18, 14, 23]. Whereas in traditional analysis approaches one typically seeks individual genes differentially expressed across sample classes (e.g. cancerous vs. control), the set-level approach aims to identify entire sets of genes that are 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. The main advantage brought by set-level analysis is the improved interpretability of analysis results. Indeed, the long lists of differentially expressed genes characteristic of traditional expression analysis are replaced by shorter and more informative lists of actual biological processes.

Predictive classification [11] 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 are assembled into formal models prescribing how to classify new samples whose class labels are not yet known. Predictive classification techniques are thus especially relevant to diagnostic tasks and

as such have been explored since very early studies on microarray data analysis [10]. Predictive models are usually constructed by supervised machine learning algorithms [11] that automatically discover patterns among samples whose labels are already available (so-called *training samples*). Learned classifiers may take diverse forms ranging from geometrically conceived models such as *Support Vector Machines* [24], which have been especially popular in the gene expression domain, to symbolic models such as logical rules or decision trees that have also been applied in this area [27, 15].

The main motivation for extending the set-level framework to the machine learning setting is again the interpretability of results. Informally, classifiers learned using set-level features acquire forms such as "predict cancer if pathway P1 is active and pathway P2 is not" (where *activity* refers to aggregated expressions of the member genes). In contrast, classifiers learned in the standard setting derive predictions from expressions of individual genes; it is usually difficult to find relationships among the genes involved in such a classifier and to interpret the latter in terms of biological processes.

The described feature transformation incurs a significant compression of the training data since the number of considered gene sets is typically much smaller than the number of interrogated genes. This raises the natural question whether relevant information is lost in the transformation, and whether the augmented interpretability will be traded off for decreased predictive accuracy. The main objective of this study is to address this question experimentally.

A further important objective is to evaluate—from the machine learning perspective—statistical techniques proposed recently in the research on set-level gene expression analysis. These are namely the Gene Set Enrichment Analysis (GSEA) method [20], the SAM-GS algorithm [7] and a technique known as the Global test [9]. Informally, they rank a given collection of gene sets according to their correlation with phenotype classes. The methods naturally translate into the machine learning context in that they facilitate feature selection [17], i.e. they are used to determine which gene sets should be provided as sample features to the learning algorithm. We experimentally verify whether these methods work reasonably in the classification setting, i.e. whether learning algorithms produce better classifiers from gene sets ranked high by the mentioned methods than from those ranking lower. We investigate classification conducted with a single selected gene set as well as with a batch of high ranking sets.

To use a machine learning algorithm, a unique value for each feature of each training sample must be established. Set-level features correspond to multiple expressions and these must therefore be aggregated. We comparatively evaluate two aggregatation options. The first simply averages the expressions of the involved genes, whereas the second relies on the more sophisticated method proposed by [23] and based on singular value decomposition.

Let us return to the initial experimental question concerned with how the final predictive accuracy is influenced by the training data compression incurred by reformulating features to the gene set level. As follows from the above, two factors contribute to this compression: selection (not every gene from the orig-

inal sample representation is a member of a gene set used in the set-level representation, i.e. some interrogated genes become ignored) and aggregation (for every gene set in the set-level representation, expressions of all its members are aggregated into a single value). We quantify the effects of these factors on predictive accuracy. Regarding selection, we experiment with set-level representations based on 10 best gene sets and 1 best gene set, respectively, and we do this for all three of the above-mentioned selection methods. We compare the obtained accuracies to the baseline case where all individual genes are provided as features to the learning algorithm. For each of the selection cases, we want to evaluate the contribution of the aggregation factor. This is done by comparing both of the above mentioned aggregation mechanisms to the control case where no aggregation is performed at all; in this case, individual genes combined from the selected gene groups act as features.

The contribution of the present study lies in the thorough experimental evaluation of a number of aspects and techniques of the gene set framework employed in the machine learning context. Our contribution is, however, also significant beyond the machine learning scope. In the general area of set-level expression analysis, it is undoubtedly important to establish a performance ranking of the various statistical techniques for the identification of significant gene sets in class-labeled expression data. This is made difficult by the lack of an unquestionable ranking criterion—there is in general no ground truth stipulating which gene sets should indeed be identified by the tested algorithms. The typical approach embraced by comparative studies (such as [7]) is thus to appeal to intuition (e.g. *the p53 pathway should be identified in p53-gene mutation data*). However legitimate such arguments are, evaluations based on them are obviously limited in generality and objectivity. We propose that the predictive classification setting supported by the cross-validation procedure for unbiased accuracy estimation, as adopted in this paper, represents exactly such a needed framework enabling objective comparative assessment of gene set selection techniques. In this framework, results of gene set selection are deemed good if the selected gene sets allow accurate classification of new samples. Through cross-validation, the accuracy can be estimated in an unbiased manner.

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 methods adopted for gene set ranking, gene expression aggregation, and for classifier learning. Next we present the data sets used as benchmarks in the comparative experiments. Lastly, we describe the protocol followed by our experiments.

2.1 Gene Set Ranking

Three methods are considered for gene set selection. As inputs, all of the methods assume a set $G = \{g_1, g_2, \ldots g_n\}$ of interrogated genes, and a set S of m expression samples where for each $s_i \in S$, $s_i = (e_{1,i}, e_{2,i}, \ldots e_{n,i}) \in \mathsf{R}^n$ where $e_{j,i}$ denotes the (normalized) expression of gene g_j in sample s_i. The sample set S is partitioned into phenotype 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 $\Gamma \in \mathcal{G}$ it holds $\Gamma \subseteq G$. In the output, each of the methods ranks all gene sets in $\mathcal{G}$ by their estimated power to discriminate samples into the predefined classes.

Next we give a brief account of the three methods and refer to the original sources for a more detailed description. In experiments, we used the original implementations of the procedures as provided by the respective authors.

GSEA [20]. Gene set enrichment analysis tests a null hypothesis that gene rankings in a gene set Γ, according to an association measure with the phenotype, are randomly distributed over the rankings of all genes. It first sorts G by correlation with binary phenotype. Then it calculates an enrichment score (ES) for each $\Gamma \in \mathcal{G}$ by walking down the sorted gene list, increasing a running-sum statistic when encountering a gene $g_i \in \Gamma$ and decreasing it otherwise. The magnitude of the change depends on the correlation of g_i with the phenotype. The enrichment score is the maximum deviation from zero encountered in the random walk. The statistical significance of the ES is estimated by an empirical phenotype-based permutation test procedure that preserves the correlation structure of the gene expression data. GSEA was one of the first specialized gene-set analysis techniques. It has been reported to attribute statistical significance to gene sets that have no gene associated with the phenotype, and to have less power than other recent test statistics [7, 9].

SAM-GS [7]. This method tests a null hypothesis that the mean vectors of the expressions of genes in a gene set do not differ by phenotype. Each sample s_i is viewed as a point in an n-dimensional Euclidean space. Each gene set $\Gamma \in \mathcal{G}$ defines its $|\Gamma|$-dimensional subspace in which projections s_i^Γ of samples s_i are given by coordinates corresponding to genes in Γ. The method judges a given Γ by how distinctly the clusters of points $\{s_i^\Gamma | s_i \in C_1\}$ and $\{s_j^\Gamma | s_j \in C_2\}$ are separated from each other in the subspace induced by Γ. 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.

Global Test [9]. The global test, analogically to SAM-GS, projects the expression samples into subspaces defined by gene sets $\Gamma \in \mathcal{G}$. In contrast to the Euclidean distance applied in SAM-GS, it proceeds instead 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 Expression Aggregation

Two methods are considered for assigning a value to a given gene set Γ for a given sample s_i by aggregation of expressions of genes in Γ.

Averaging. The first method simply produces the arithmetic average of the expressions of all Γ genes in sample s_i. The value assigned to the pair (s_i, Γ) is thus independent of samples s_j, $i \neq j$.

Singular Value Decomposition. A more sophisticated approach was employed by [23]. Here, the value assigned to (s_i, Γ) depends on other samples s_j. In particular, all samples in the sample set S are viewed as points in the $|\Gamma|$-dimensional Euclidean space induced by Γ 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, Γ) is the real-valued position of the projection of s_i on the maximum-variance vector in the space induced by Γ. We refer to the paper [23] for detailed explanation.

2.3 Machine Learning

We experimented with five diverse machine learning algorithms to avoid dependence of experimental results on a specific choice of a learning method, namely Support Vector Machine, 1-Nearest Neighbor, 3-Nearest Neighbors, Naive Bayes and Decision Tree. These algorithms are explained in depth for example by [11]. In experiments, we used the implementations available in the WEKA software due to [25], using the default settings. None of the methods below is in principle superior to the others, although the first one prevails in predictive modeling of gene expression data and is usually associated with high resistance to noise.

2.4 Expression and Gene Sets

We conducted our experiments using 20 public gene expression datasets, each containing samples pertaining to two classes. Table 1 shows for each dataset the number of samples in each class, the number of interrogated genes and the reference for further details. Some of the two-class datasets were derived from the three-class problems (Colitis and Crohn, Parkinson).

Besides expression datasets, we utilized a gene set database consisting of 1685 manually curated sets of genes obtained from the Molecular Signatures Database (MSigDB v2.0) [20]. These gene sets have been compiled from various online databases (e.g. KEGG, GenMAPP, BioCarta).

2.5 Experimental Protocol

Classifier learning in the set-level framework follows a simple workflow whose performance is influenced by several factors, each corresponding to a particular

Dataset	*Genes*	*Class 1*	*Class 2*	*Reference*
ALL/AML	10056	24	24	[1]
Brain/muscle	13380	41	20	[13]
Colitis and Crohn 1	14902	42	26	[4]
Colitis and Crohn 2	14902	42	59	[4]
Colitis and Crohn 3	14902	26	59	[4]
Diabetes	13380	17	17	[18]
Heme/stroma	13380	18	33	[13]
Gastric cancer	5664	8	22	[12]
Gender	15056	15	17	[20]
Gliomas	14902	26	59	[8]
Lung Cancer Boston	5217	31	31	[3]
Lung Cancer Michigan	5217	24	62	[2]
Melanoma	14902	18	45	[21]
p53	10101	33	17	[20]
Parkinson 1	14902	22	33	[19]
Parkinson 2	14902	22	50	[19]
Parkinson 3	14902	33	50	[19]
Pollution	37804	88	41	[16]
Sarcoma and hypoxia	14902	15	39	[26]
Smoking	5664	18	26	[5]

Table 1. Number of genes interrogated and number of samples in each of the two classes of each dataset.

choice from a class of techniques (such as for gene set ranking). We evaluate the contribution of these factors to the predictive accuracy of the resulting classifiers through repeated executions of the learning workflow, varying the factors.

The learning workflow is shown in Fig. 1. Given a set of binary-labeled training samples from an expression dataset, the workflow starts by ranking the provided collection of a priori-defined gene sets according to their power to discriminate sample classes (see Sec. 2.1 for details). The resulting ranked list is subsequently used to select the gene sets used to form set-level sample features. Each such feature is then assigned a value for each training sample by aggregating the expressions in the gene set corresponding to the feature; an exception to this is the *none* alternative of the aggregation factor, where expressions are not aggregated, and features correspond to genes instead of gene sets. This alternative is considered for comparative purposes. Next, a machine learning algorithm produces a classifier from the reformulated training samples. Finally, the classifier's predictive accuracy is calculated as the proportion of samples correctly classifed on an independent testing sample fold. For compatibility with the learned classifier, the testing samples are also reformulated to the set level prior to testing, using the selected gene sets and aggregation as in the training phase.

Six factors along the workflow influence its result. The alternatives considered for each of them are summarized in Table 2. We want to assess the contributions of the first three factors (top in table). The remaining three auxiliary factors (bottom in table) are employed to diversify the experimental material and thus

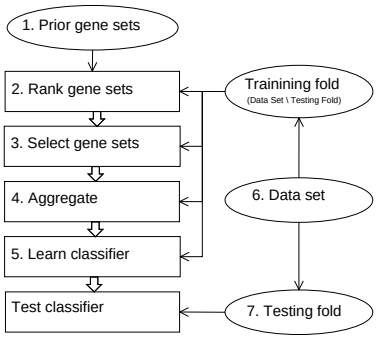

Fig. 1. The workflow of a set-level learning experiment conducted multiple times with varying alternatives in the numbered steps. For compatibility with the learned classifier, testing fold samples are also reformulated to the set level. This is done using gene sets selected in Step 3 and aggregation algorithm used in Step 4. The diagram abstracts from this operation.

increase the robustness of the findings. Factor 6 (testing fold) is involved automatically through the adoption of the 10-fold cross-validation procedure (see e.g. [11], chap. 7). We execute the workflow for each possible combination of factor alternatives, obtaining a factored sample of 198,000 predictive accuracy values.

While the measurements provided by the above protocol allow us to compare multiple variants of the set-level framework for predictive classification, we also want to compare these to the baseline gene-level alternative usually adopted in predictive classification of gene expression data. Here, each gene interrogated by a microarray represents a feature. This sample representation is passed directly to the learning algorithm without involving any of the pre-processing factors (1-3 in Table 2). The baseline results are also collected using the 5 different learning algorithms, the 20 benchmark datasets and the 10-fold crossvalidation procedure (i.e. factors 4-6 in Table 2 are employed). As a result, an additional sample of 1,000 predictive accuracy values are collected for the baseline variant.

Finally, to comply with the standard application of the cross-validation procedure, we averaged the accuracy values corresponding to the 10 cross-validation folds for each combination of the remaining factors. The subsequent statistical analysis thus deals with a sample of 19,800 and 100 measurements for the set-level and baseline experiments, described by the predictive accuracy value and the values of the relevant factors.

3 Results

All statistical tests in this section refer to the paired non-parametric Wilcoxon test (two-sided unless stated otherwise).[3] For pairing, we always related two

[3] Preliminary normality tests did not justify the application of the stronger t-test. Besides, the Wilcoxon test is argued [6] to be statistically safer than the t-test for comparing classification algorithms over multiple data sets.

Analyzed factors	*Alternatives*	*#Alts*
1. Ranking algo (Sec. 2.1)	{gsea, sam-gs, global}	3
*2. Sets forming features**	$\{1, 2, \ldots 10,$ $1676, 1677, \ldots 1685,$ 1:10, 1676:1685$\}$	22
3. Aggregation (Sec. 2.2)	{svd, avg, none}	3
Product		198

Auxiliary factors	*Alternatives*	*#Alts*
4. Learning algo (Sec. 2.3)	{svm, 1-nn, 3-nn, nb, dt}	5
5. Data set (Sec. 2.4)	$\{d_1 \ldots d_{20}\}$	20
6. Testing Fold	$\{f_1 \ldots f_{10}\}$	10
Product		1000

* identified by rank. 1685 corresponds to the lowest ranking set. i:j denotes that all of gene sets ranking i to j are used to form features.

Table 2. Alternatives considered for factors influencing the set-level learning workflow. The number left of each factor refers to the workflow step (Fig. 1) in which it acts.

measurements equal in terms of all factors except for the one investigated. All significance results are at the 0.05 level.

Using the set-level experimental sample, we first verified whether gene sets ranked high by the established set-level analysis methods (GSEA, SAM-GS and Global test) indeed lead to construction of better classifiers by machine learning algorithms, i.e. we investigated how classification accuracies depend on *Factor 3* (see Table 2). In the top panel of Fig. 2, we plot the average accuracies for Factor 3 alternatives ranging 1 to 10, and 1676 to 1685. The trend line fitted by the least squares method shows a clear decay of accuracy as lower-ranking sets are used for learning. The bottom panel corresponds to Factor 3 values 1:10 (left) and 1676:1685 (right) corresponding to the situations where the 10 highest-ranking and the 10 lowest-ranking (respectively) gene sets are combined to produce a feature set for learning. Again, the dominance of the former in terms of accuracy is obvious.

Given the above, there is no apparent reason why low-ranking gene sets should be used in experiments. Therefore, to maintain relevance of the subsequent conclusions, we conducted further analyses only with measurements where Factor 2 (gene set rank) is either 1 or 1:10.

Firstly, we assessed the difference between the remaining alternatives 1 and 1:10 corresponding to more and less (respectively) compression of training data. Not surprisingly, the 1:10 variant, where sample features capture information from the ten best gene sets exhibits significantly ($p = 0.0007$) higher accuracies than the 1 variant using only the single best gene set to consitute features (a single feature if aggregation is employed).

We further compared the three gene-set ranking methods by splitting the set-level sample according to *Factor 1*. Since three comparisons are conducted

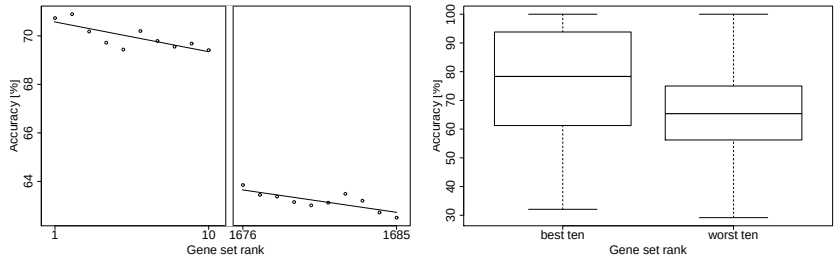

Fig. 2. Average predictive accuracy tends to fall as lower-ranking gene sets are used to constitute features (see text for details). Each point in the left panels and each box plot in the right panel follows from 16,000 learning experiments. The trend lines shown in the left panels are the ones minimizing the residual least squares.

in this case (one per pair), we used the Bonferroni-Dunn adjustment on the Wilcoxon test result. The Global test turned out to exhibit significantly higher accuracies than either SAM-GS ($p = 0.013$) or GSEA ($p = 0.027$). The difference between the latter two methods was not significant.

Concerning *Factor 3* (aggregation method), there are two questions of interest: whether one aggregation method (svd, avg) outperforms the other, and whether aggregation in general has a detrimental effect on performance. As for the first question, no significant difference between the two methods was detected. The answer to the second question turned out to depend on Factor 3 as follows. In the more compressive (1) alternative, the answer is affirmative in that both aggregation methods result in less accurate classifiers than those not incurring aggregation ($p = 0.015$ for svd, $p = 0.00052$ for avg, both after Bonferroni-Dunn adjustment). However, the detrimental effect of aggregation vanishes in the less compressive (1:10) alternative of Factor 2, where none of the two comparisons yield a significant difference.

The principle trends can also be well observed through the ranked list of methodological combinations by median classification accuracy, again generated from measurements not involving random or low-ranking gene sets. This is shown in Table 3. Position 8 refers to the baseline method where sample features capture expressions of all genes and prior gene set definitions are ignored (see Section 2.5 for details). In agreement with the statistical conclusions above, the ranked table clearly indicates the superiority of the Global test for gene-set ranking, and of using the 10 best gene sets (i.e., the 1:10 alternative) to establish features rather than relying only on the single best gene set. It is noteworthy that all three methods involving the combinations of the Global test and the 1:10 alternative (i.e., ranks 1, 2, 4) outperform the baseline method. This is especially remarkable given that the two best of them (and two best overall) involve aggregation, and the learning algorithm here receives training samples described by only 10 real-valued features. Thus, the gene-set framework allows for feature extraction characterized by vast compression of data (from the original thousands of fea-

Rank	*Methods*			*Accuracy*			
	Sets	*Rank. algo*	*Aggrgt*	*Median*	*Avg*	σ	*Iqr*
1	1:10	global	svd	86.5	79.8	17.3	32.0
2	1:10	global	avg	86.0	79.4	17.8	30.5
3	1:10	sam-gs	none	83.8	78.3	18.5	35.1
4	1:10	global	none	83.7	77.7	18.5	34.7
5	1:10	gsea	none	82.8	77.7	18.8	34.8
6	1	global	none	80.5	78.1	16.1	29.7
7	1:10	gsea	avg	79.7	76.3	17.1	28.0
8	*all genes used*			79.3	77.2	18.9	35.3
9	1	gsea	none	77.5	75.0	18.3	33.3
10	1	global	svd	77.5	74.8	15.0	25.6
11	1:10	gsea	svd	77.1	75.5	16.9	28.2
12	1:10	sam-gs	avg	74.2	75.1	16.8	28.5
13	1	sam-gs	none	73.9	74.1	15.1	26.3
14	1:10	sam-gs	svd	73.8	74.6	17.6	28.9
15	1	global	avg	72.8	72.2	14.0	22.2
16	1	gsea	avg	68.3	69.6	13.0	16.3
17	1	gsea	svd	67.4	68.5	13.2	14.4
18	1	sam-gs	avg	65.4	64.7	10.3	15.9
19	1	sam-gs	svd	64.2	65.0	12.7	13.0

Table 3. Ranking of combinations of gene set methods by median predictive accuracy achieved on 20 datasets (Table 1, Section 2.4) with 5 machine learning algorithms (Section 2.3) estimated through 10-fold cross-validation (i.e. 1,000 experiments per row). The columns indicate, respectively, the resulting rank by median accuracy, the gene sets used to form features (1 – the highest ranking set, 1:10 – the ten highest ranking sets), the gene set selection method, the expression aggregation method (see Section 2 for details on the latter 3 factors), and the median, average, standard deviation and interquartile range of the accuracy.

tures corresponding to expressions of individual genes, to 10 features) and, at the same time, by a boost in classification accuracy.

4 Conclusions and Future Work

The set-level framework can be adopted in the machine learning setting without trading off classification accuracy. To identify the best a priori-defined gene sets for classification, the *Global test* [9] significantly outperforms the *GSEA* [20] and *SAM-GS* [7] methods. To aggregate expressions of genes contained in a gene set into a value assigned to that set acting as a feature, arithmetic average could not be differentiated from the method [23] based on singular value decomposition. Using only 10 features corresponding to genuine gene sets selected by the Global test, the learned set-level classifiers systematically outperform conventional gene-level classifiers learned with access to all measured gene expressions. Data compression and increased classification accuracy thus come as additional benefits to increased interpretability of set-level classifiers.

The above-mentioned effect of data shrinkage accompanied by increased predictive accuracy could, in principle, also be achieved by generic feature extraction methods (see e.g. [17]). The advantage of our approach is that our extracted features maintain direct interpretability since they correspond to gene sets that possess a biological meaning. In future work, it would be interesting to determine whether the generic feature extraction methods could outperform the present approach at least in terms of predictive accuracy achieved with a fixed target number of extracted features. By the same mail, the optimal number of set-level features employed will vary between data domains. For our experiments, we chose the ad hoc number of 10 features for all domains. In future experiments, the optimal domain-specific number may be estimated, e.g. through internal cross-validation [11].

We applied two previously suggested general methods enabling aggregation of multiple expression values into a single value assigned to a set-level feature. The downside of this generality is that substantial information available for specific kinds of gene sets is ignored. Of relevance to pathway-based gene sets, the recent study by [22] convincingly argues that the perturbation of a pathway depends on the expressions of its member genes in a non-uniform manner. It also proposes how to quantify the impact of each member gene on the perturbation, given the graphical structure of the pathway. It seems reasonable that a pathway-specific aggregation method should also weigh member genes by their estimated impact on the pathway. Such a method would likely result in more informative pathway-level features and could outperform the two aggregagation methods we have considered, potentially giving a futher boost to the good performance of predictive classification based on a small number of set-level features.

Acknowledgement

This research was supported by the Czech Science Foundation through project No. 201/09/1665 (FZ, MH) and the Czech Ministry of Education through research programme MSM 6840770012 (JK).

References

1. Armstrong, S. A. *et al.* (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.*, **30**, 41 – 7.
2. Beer, D. G. *et al.* (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**(8), 816–824.
3. Bhattacharjee, A. *et al.* (2001). Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci.*, **98**(24), 13790–13795.
4. Burczynski, M. E. *et al.* (2006). Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. **8**(1), 51–61.

5. Carolan, B. J. *et al.* (2006). Up-regulation of expression of the ubiquitin carboxyl-terminal hydrolase L1 gene in human airway epithelium of cigarette smokers. *Cancer Res.*, **66**(22), 10729–40.
6. Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *JMRL*, **7**, 1–30.
7. Dinu, I. (2007). Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, **8**(1), 242.
8. Freije, W. A. *et al.* (2004). Gene expression profiling of gliomas strongly predicts survival. *Cancer Res.*, **64**(18), 6503–10.
9. Goeman, J. J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**(8), 980–987.
10. Golub, T. R. *et al.* (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439), 531–537.
11. Hastie, T. *et al.* (2001). *The Elements of Statistical Learning*. Springer.
12. Hippo, Y. *et al.* (2002). Global Gene Expression Analysis of Gastric Cancer by Oligonucleotide Microarrays. *Cancer Res.*, **62**(1), 233–240.
13. Holec, M. *et al.* (2009). Integrating multiple-platform expression data through gene set features. In *The 5th International Symposium on Bioinformatics Research and Applications (ISBRA 2009)*. Springer.
14. Huang, D. W. *et al.* (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*
15. Huang, J. *et al.* (2010). Decision forest for classification of gene expression data. *Comput. Biol. Med.*, **40**, 698–704.
16. Libalova, H. *et al.* (2010). Gene expression profiling in blood of asthmatic children living in polluted region of the czech republic (project airgen). In *10th International Conference on Environmental Mutagens*.
17. Liu, H. and Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer.
18. Mootha, V. K. *et al.* (2003). Pgc-1-alpha-responsive genes involved in oxidative phosphorylation are coorinately down regulated in human diabetes. *Nat. Genet.*, **34**, 267–273.
19. Scherzer, C. R. *et al.* (2007). Molecular markers of early Parkinson's disease based on gene expression in blood. *Proc. Natl. Acad. Sci.*, **104**(3), 955–60.
20. Subramanian, A. *et al.* (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.*, **102**(43), 15545–50.
21. Talantov, D. *et al.* (2005). Novel genes associated with malignant melanoma but not benign melanocytic lesions. *Clin. Cancer Res.*, **11**(20), 7234–42.
22. Tarca, A. L. *et al.* (2009). A novel signaling pathway impact analysis. *Bioinformatics*, **25**(1), 77–82.
23. Tomfohr, J. *et al.* (2005). Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, **6**, 225.
24. Vapnik, V. N. (2000). *The Nature of Statistical Learning*. Springer.
25. Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco.
26. Yoon, S. S. *et al.* (2006). Angiogenic profile of soft tissue sarcomas based on analysis of circulating factors and microarray gene expression. *J. Surg. Res.*, **135**(2), 282–90.
27. Zintzaras, E. and Kowald, A. (2010). Forest classification trees and forest support vector machines algorithms: Demonstration using microarray data. *Cell Cycle*, **40**(5), 519–24.