Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 616358, 9 pages
doi:10.1155/2010/616358

Research Article

Stability of Ranked Gene Lists in Large Microarray Analysis Studies

Gregor Stiglic[1] and Peter Kokol[1, 2]

[1] *Faculty of Health Sciences, Research Institute, University of Maribor, Zitna ulica 15, 2000 Maribor, Slovenia*
[2] *Laboratory for System Design, Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova ulica 17, 2000 Maribor, Slovenia*

Correspondence should be addressed to Gregor Stiglic, gregor.stiglic@uni-mb.si

Received 28 January 2010; Accepted 17 May 2010

Academic Editor: Nick Grishin

This paper presents an empirical study that aims to explain the relationship between the number of samples and stability of different gene selection techniques for microarray datasets. Unlike other similar studies where number of genes in a ranked gene list is variable, this study uses an alternative approach where stability is observed at different number of samples that are used for gene selection. Three different metrics of stability, including a novel metric in bioinformatics, were used to estimate the stability of the ranked gene lists. Results of this study demonstrate that the univariate selection methods produce significantly more stable ranked gene lists than the multivariate selection methods used in this study. More specifically, thousands of samples are needed for these multivariate selection methods to achieve the same level of stability any given univariate selection method can achieve with only hundreds.

1. Introduction

In the past decade, microarray-based gene expression profiling has become a widely studied field of research. Additionally, with the introduction of commercial gene expression-based diagnostic tests, microarrays are finally coming into practical use [1]. To construct a reliable set of genes or gene expression signature, one can use different gene selection techniques. The initially used methods for gene selection were based on simple statistical tests and could evaluate only a single gene at a time. Examples of such univariate gene selection methods include gene selection based on t-test [2] (with multiple variants), signal-to-noise (SN) ratio [3], correlation tests [4], information gain [5], and so forth. Nowadays, novel and more complex multivariate gene selection methods, which can evaluate groups of genes and detect interactions among multiple genes simultaneously, are introduced on a frequent basis.

One of the most serious problems in the detection and evaluation of reliable gene expression signatures is the lack of studies that use large number of samples. Therefore, different approaches have been proposed. A recent study by Xu et al. [6] suggests the integration of multiple microarray studies to increase the sample size, and demonstrates a novel method of microarray integration on breast-cancer gene expression signatures. However, such gene expression signatures are very unstable owing to different protocols and sources of data used in different laboratories. On the other hand, there are much more data sets with relatively high number of samples available today than a few years ago. Our study exploits the advantage of one of the largest publicly available repositories of gene expression measurements that were collected by the International Genomics Consortium.

This paper presents an empirical study on different supervised gene selection techniques from the perspective of the stability of the ranked gene lists. The usual way of estimating the "goodness" of the ranked gene lists is by using them for the classification of samples and measuring different accuracy metrics like percentage of correctly classified samples or area under ROC curve, shortly AUC. In this study, three different metrics for ranked gene lists similarity were used to estimate the stability of the gene selection method.

Another important factor is the stability of the stability metrics or metastability that can be of significant importance when comparing the performance of the feature selection on data sets with different number of samples. Meta-stability in this study was measured as the standard deviation of the stability metric. Stability and meta-stability of the gene selection methods was observed on data sets with 50 and up to 600 samples. Our study demonstrated that one should be careful in selecting a particular feature selection method when more and more gene expression data sets with up to 500 or even 1000 patients are available.

A paper by Ma [7] was one of the first papers in gene expression analysis comparing different supervised gene selection methods by bootstrapping the samples of initial data set. Ma measured the concordance and reproducibility of the supervised gene screening based on eight different gene selection methods. The measurements of concordance were done by overlapping the selected genes with different settings for n top genes. Among other conclusions, this empirical study once again explained that genes passed through different gene selection methods may be considerably different. Unfortunately, experimentation using different number of samples was not possible owing to small number of samples present in the three data sets that were used in study by Ma.

Another similar study was conducted by Qiu et al. [8] in which the stability of differentially expressed genes was evaluated using the measurement of frequency, by which a given gene is selected across the subsamples. They showed that resampling can be an appropriate technique to determine a set of genes with sufficiently high frequency. Furthermore, they recommended using resampling techniques to assess the variability of different performance indicators. In contrast to our research, their approach was based on measuring the stability of single genes and not the ranked lists of genes produced by the gene selection methods. Once again, their method was not tested on data sets with large number of samples. Their experiments were carried out on simulated data consisting of 30 (small data set) and 86 (large data set) samples, which is much lesser than the number of samples used in this study.

However, Ein-Dor et al. [9] determined that the gene lists obtained for the same clinical types of patients by different laboratories often differ significantly and have very few common genes. They evaluated the robustness of gene lists by employing the probably approximately correct (PAC) sorting, and pointed out that an achievement of a typical overlap of 50% between two lists of genes would require expression profiles of over thousand patients with early detection of breast cancer. Using microarray data sets that are available today, our study tried to confirm the conclusions of the study by Ein-Dor et al.

Although most of the above-mentioned studies used only a simple overlap metric, there are many other metrics available, which had been used in similar earlier researches. One of the fields in which such metrics were initially used in empirical studies is the information retrieval (IR), where the metrics used to compare the stability of the results returned by different web search engines. A study by Bar-Ilan et al. [10] used the metrics for the stability of ranked lists to monitor the consistency of the web search results for the most popular search engines through time. The metrics that they used were principally proposed much earlier by Fagin et al. [11] who also theoretically defined them. In our study, we initially evaluated three ranked lists stability metrics, that is, simple overlap, weighted overlap, and extended Spearman stability metric. The latter two metrics were proposed by Fagin et al. to define the metrics that would award the subjects (genes) similarly ranked at the top of the list in comparison with those that were ranked lower.

Table 1: Overview of GEMLeR datasets used in this study.

Dataset	No. of Samples	Class 1	Class 2	Probes
AP_Breast_Colon	630	344	286	10937
AP_Breast_Kidney	604	344	260	10937
AP_Breast_Ovary	542	344	198	10937
AP_Colon_Kidney	546	286	260	10937

2. Methods and Data Sets

The original data were obtained from the Expression Project for Oncology (expO) data set from The International Genomics Consortium (http://www.intgen.org/) that was deposited at Gene Expression Omnibus (GEO) repository [12], accession number GSE2109. Samples from this collection represent an integral part of the Gene Expression Machine Learning Repository (GEMLeR), available at http://gemler.fzv.uni-mb.si/. All the samples were based on Affymetrix GeneChip U133 Plus 2.0 arrays and were normalized using MAS5.

The empirical tests for this study were conducted using four data sets from GEMLeR that were chosen by the highest number of available samples. Therefore, a collection of four data sets with more than 500 samples (Table 1) was selected for the experimental work. One can observe the lower than normal number of probes for U133 Plus 2.0 arrays which is due to prefiltering using maximal variance of signal where 20% of probes with the highest variance were retained.

2.1. Gene Selection. Gene selection methods are used to identify the gene subsets of microarray data that can be further used to effectively discriminate the different clinical conditions. This study compares the stability of four different gene selection methods. All the gene selection and classification methods, except Signal-to-noise method, are a part of Weka environment [13] that was used to conduct this study.

Signal-to-noise (SN) gene selection was introduced by Golub et al. [3] and represents a typical univariate gene selection method, that is, genes are scored individually one at a time and later ranked by their score defining their ability to separate samples based on a single gene. The SN gene ranking method is a variant of the widely used t-score-based statistical gene selection methods. It has a simple definition

$$SN_i = \frac{|\mu_1 - \mu_2|}{\sigma_1 + \sigma_2}, \quad (1)$$

where difference of mean gene expression values for classes 1 and 2 is divided by sum of σ_1 and σ_2 that represent standard deviations of gene expression values for samples belonging to class 1 and class two, respectively. Subsequently, the genes are sorted according to their SN_i scores and the n top genes are selected as members of the ranked gene list. Direct interpretation of this score suggests that it measures the degree of overlap for the ith gene distribution in both the classes. However, it can only be used in binary problems where there are only two different classes.

Information gain (IG) is another method from the family of univariate gene selection methods. Information theory-based [14] gene selection that was initially used in the decision-tree-building algorithms [15] has also been used as one of the gene selection procedures. The IG is a measure-based method, where each gene obtains an IG score. In our study an InfoGainAttributeEval implementation from WEKA environment, as defined in [13], was used for evaluation of genes by measuring their information gain with respect to the class. Discretization of numeric values was done using the Minimum Description Length (MDL)-based discretization method by [16]. Usually, n top genes are selected for the inclusion in the gene expression signature. In our study, 100 top genes were used in all the experiments.

ReliefF gene selection method is a representative of the multivariate gene selection techniques, where a group of genes is evaluated synchronously. ReliefF algorithm is based on the original Relief algorithm [17] that could only be used for the classification problems with two class values and nominal values of variables. The basic idea of Relief algorithm is ranking of features based on their ability to distinguish between instances that are near to each other. The original algorithm was extended by Kononenko [18], so that it can deal with multiclass problems and missing values. Later, it was further improved by Robnik-Sikonja and Kononenko [19] to suit the noisy data and for its usage in regression problems.

Support Vector Machines Recursive Feature Elimination (SVM-RFE) represents another multivariate gene selection method. The SVM in combination with RFE were introduced for gene selection in bioinformatics by Guyon et al. [20]. The SVM-RFE feature selection method is based on linear SVM used as the learning algorithm. In the final step of each cycle, all the genes are ranked and a preselected number of the lowest-ranked genes are eliminated. By default, a single gene is eliminated in each round. However, it is also possible to remove more than one feature per round.

2.2. Stability Metrics. To measure the stability of gene lists produced by specific gene selection methods, one should employ appropriate metrics that are able to statistically measure the differences between the two lists of ranked genes. This study uses three different stability measures, that is, overlap, weighted overlap, and Spearman ranked-list similarity metric.

Overlap is one of the simplest measures of similarity, where the similarity of the two lists (ℓ_1, ℓ_2) is not based on the ranking of the genes. The degree of similarity is calculated by simple counting of the genes that are present in both the lists and dividing them by the number of genes in each list. The top-k genes overlap can be defined as

$$\delta(k, \ell_1, \ell_2) = \frac{\sum_{i=1}^{k} \chi(\ell_1(i), \ell_2)}{k}, \tag{2}$$

where

$$\chi(\ell_1(i), \ell_2) = \begin{cases} 0, & \text{if } \ell_1(i) \bigcap \ell_2 = \varnothing, \\ 1, & \text{otherwise.} \end{cases} \tag{3}$$

Weighted overlap represents an improved version of the overlap metric, where the higher-ranked genes present in both the lists contribute to higher degree of similarity than the lower-ranked genes present in both the lists. Weighted overlap gradually builds overlap score by starting with only the top ranked gene from both compared lists and continues to add top ranked genes one by one. It can be defined as

$$\delta^{(w)}(k, \ell_1, \ell_2) = \frac{\sum_{i=1}^{k} \delta(i, \ell_1, \ell_2)}{k}. \tag{4}$$

Spearman's ranked-lists similarity measure, also called Spearman's Footrule, requires all the genes present in both the lists for the actual implementation of this measure. In case of gene selection that would mean that both compared selection methods must select the same set of genes, that can be ranked differently. To define Spearman's Footrule, it is necessary to introduce $\sigma_1(i)$ and $\sigma_2(i)$ that represent ranks of gene i in lists 1 and 2, respectively. By introducing normalization to $[0, 1]$ interval, the value 1 is obtained when both the lists completely agree, and 0 is obtained for the genes from both the lists in the opposite order

$$\delta^{(Sp)}(k, \ell_1, \ell_2) = 1 - \frac{\sum_{i=1}^{k} |\sigma_1(i) - \sigma_2(i)|}{\max(\delta^{(Sp)}(k, \ell_1, \ell_2))}. \tag{5}$$

The original method was applied to the nonequal lists of the same size by Fagin et al. [11], and is therefore much more suitable for comparison of ranked gene lists

$$\delta^{(F)}(k, \ell_1, \ell_2) = 2(k - z)(k + 1) + \sum_{i \in Z} |\sigma_1(i) - \sigma_2(i)| - \sum_{i \in S} \sigma_1(i) - \sum_{i \in T} \sigma_1(i), \tag{6}$$

where z stands for number of genes present in both lists, Z represents a set of genes that are present in both lists, S represents set of genes that are present only in the first list of genes and vice-versa for T.

After normalization and transformation to $[0, 1]$, we get

$$\delta^{(Fn)}(k, \ell_1, \ell_2) = 1 - \frac{\delta^{(F)}(k, \ell_1, \ell_2)}{\max(\delta^{(F)}(k, \ell_1, \ell_2))}. \tag{7}$$

2.3. Classification Performance. Two classification models were used to evaluate the classification performance, after the initial set of genes was reduced to 100 highest-ranked genes.

K-Nearest Neighbors (k-NN) is one of the simplest classification algorithms, and is based on the distance metric that measures the distance to the nearest neighbors of a sample that is to be classified. The k-NN classifier is widely used in machine learning and has been applied to bioinformatics problems [21]. With the given test sample of unknown class, this model finds the k nearest neighbors in the training set using Euclidean distance (d), and assigns the label of the test sample according to the labels of those neighbors. A setting with 5 nearest neighbors was used for the estimation of classification performance in our study. To ensure a majority in the voting process, we used the weighting technique where each vote is weighted by its distance to the test sample. The weight assigned to each vote was calculated as $1/d$.

Support Vector Machines (SVMs) can handle very large-scale classification with respect to both the number of samples as well as the number of variables [22]. It takes two steps to create an SVM. The first step includes mapping of data vectors to a high-dimensional space, while the second step attempts to find a hyperplane in the newly mapped space with maximum margin separating the classes of data. The simplest example of SVM is a hyperplane that separates two classes of examples with maximum margin. The margin is defined by the distance from the hyperplane to the nearest of the data points. Our study employed an optimized version of SVM called the Sequential Minimized Optimization, proposed by Platt [23], which is implemented in the Weka environment. The linear kernel using default Weka setting was used throughout this study.

3. Results

3.1. Resampling-Based Stability Measurements. The first group of experiments is based on sampling using different number of samples from the original data. Random sampling with replacement leaves the number of samples in a new sampled data set unchanged as it was in the original, owing to some samples being picked more than once. This method of resampling is often called bootstrapping and is used for estimating the properties of an estimator (such as its stability) by measuring those properties when sampling from an approximating distribution.

Similarity measurement for each of the four gene selection methods consisted of the following steps.

(1) Randomly select n samples from original dataset, with n ranging from 50 to m in steps of 25 until m exceeds number of all patients.

(2) Create k new subsets using sampling with replacement.

(3) Each of four gene selection methods is used on each subset to construct k ranked lists of genes (each of them containing 100 top ranked genes).

(4) Stability S is then measured by averaging over all pairwise comparisons of k ranked gene lists for each gene selection method separately

$$S = \frac{2\sum_{i=1}^{k}\sum_{j=i+1}^{k}\mathrm{Sim}\left(g_i, g_j\right)}{k(k-1)}, \tag{8}$$

where g_i represents ranked list of genes from subsample k ($1 \le i \le k$), and $\mathrm{Sim}(g_i, g_j)$ represents one of three similarity measures ($\delta, \delta^{(w)}$ or $\delta^{(Fn)}$) used to estimate the stability of two ranked gene lists by calculating similarity between them.

Figure 1 shows the stability results on the largest data set, comparing 344 breast cancer tissues against 286 colon cancer tissues. Similar results were obtained for the three remaining data sets (see supplemental files in the supplementary material available online on doi:616358). Simultaneous comparison of SN and IG as well as ReliefF and SVM-RFE also represents the comparison of two families of gene selection methods, that is, univariate versus multivariate. The difference between these two groups is clearly evident from the stability measurements. Univariate methods show steady growth of stability, in contrast to the multivariate methods, where stability measures rarely (ReliefF) or never (SVM-RFE) show any growth when the number of samples gets higher. On comparing the univariate gene selection methods, it is evident that IG produces by far the most stable lists of ranked genes. By directly interpreting the overlap metric, it can be stated that when the number of samples reaches 400, IG gives more than 80 out of 100 genes in both the lists from the two bootstrapped data sets. On the other hand, SVM-RFE is capable of producing only around 20 common genes on an average.

It can be observed that all the three similarity metrics that were used in the stability measurements for all of the four gene selection methods showed very similar behavior. Weighted overlap metric produced the most optimistic results, followed by classical overlap and Spearman metric. The higher values of weighted overlap metric can be explained by the fact that this metric prefers overlap of higher-ranked genes.

Standard deviation measurements for all the three similarity metrics, as a function of the number of samples, can be obtained in the supplemental materials. From the standard deviation results, it can be observed that ReliefF is the only method where larger number of samples does not stabilize the similarity measurements. Furthermore, the IG again proved to be the most optimal method with the lowest standard deviation values, especially for large number of samples.

3.2. Measuring Stability Using Partitioning. Stability of produced gene lists becomes even more important when ranked gene lists from two different studies are compared. A lot of recent studies include a comparison of breast cancer gene signatures where different lists of genes are usually compared to referential study by Van't Veer et al. [24]. A paper by Ein-Dor et al. [9] shows that thousands of

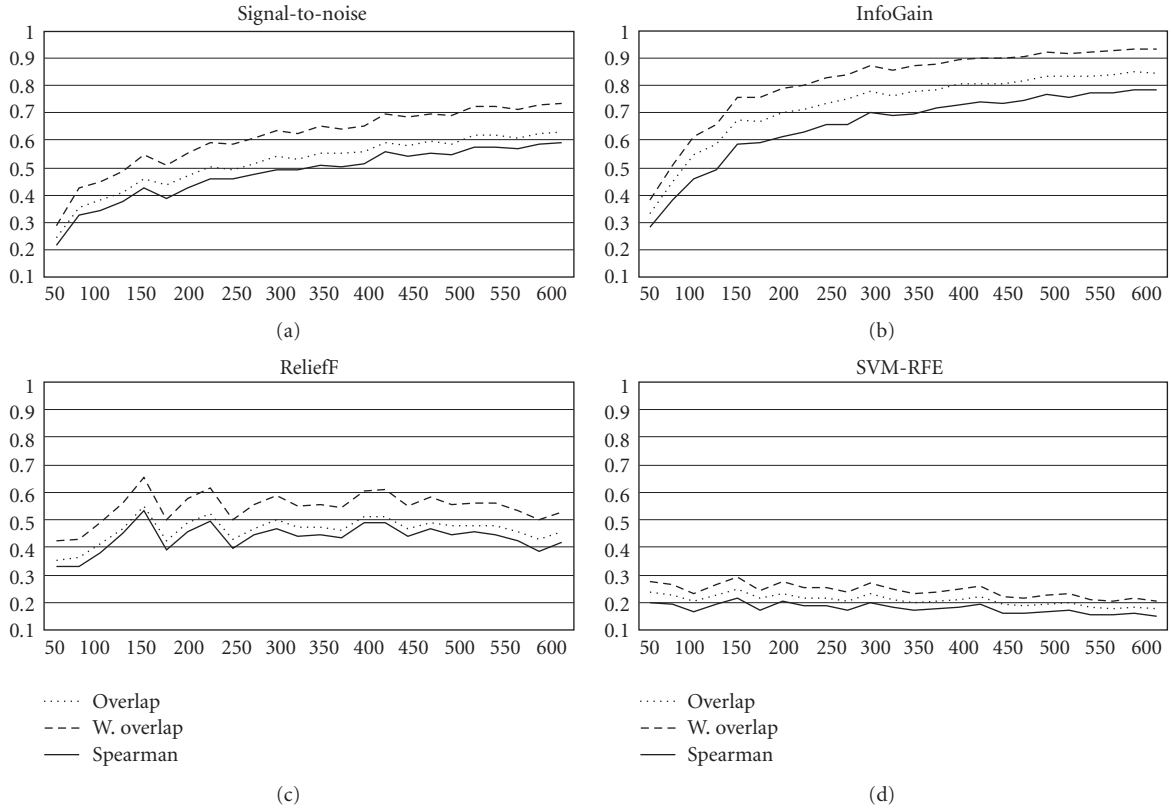

Figure 1: Relation between the number of samples and stability using resampling technique (breast versus colon cancer data set).

genes would be needed to achieve an overlap of 50% for breast cancer dataset. To confirm this pessimistic statement, another experimental setup was prepared, where partitioning of initial data set in two independent subsets was employed. The samples were randomly shuffled and equally partitioned into two groups with equal number of samples. Again, two univariate gene selection techniques were compared with two multivariate methods, to measure the similarity of the two ranked gene lists produced from each partition of data. The whole procedure was repeated 10 times and the measurements were averaged for each point, shown in Figure 2, representing the number of samples in each group.

This experiment demonstrated similar results to the former one, with an even bigger gap between the most- and the least-stable gene selection technique. The ReliefF showed good performance when compared with sampling experiment, but was still unable to grow steadily and was therefore considered very unstable. The SVM-RFE demonstrated very low stability again, while IG managed to cross 50% overlap at somewhere between 50 and 75 samples in each group. By observing the results for SN gene selection, it is evident that more than 300 samples in each group would be needed to reach an overlap of 50%.

3.3. Classification Accuracy. The final experiment compared the accuracy of the classification rates for IG and SVM-RFE-based gene selection methods to test the possible significance of SVM-RFE-based gene selection in providing better results than the classification using IG gene selection. Each feature selection method was used in combination with two classification models, that is, SVM and k-NN. The two measures of accuracy were used to evaluate the classification performance of the four classification combinations. In addition to the classical accuracy estimation (denoted as ACC), calculated as the number of correctly classified samples divided by the total number of samples, the area under ROC curve, or in short, the area under curve (AUC) was used. Classification performance estimation was carried out using 10-fold cross-validation cycle that was repeated 10 times to ensure higher reliability of the results, especially in cases where a small number of samples were used. The number of selected genes was 100 as in the previous experiments. To avoid the so-called selection bias that was exposed in earlier studies by Ambroise and McLachlan [25] and Simon et al. [26], gene selection along with classification was carried out exclusively on the training set within the cross-validation procedure.

Figure 3 presents the results of ACC and AUC, while Figure 4 shows the standard deviation values for different number of samples used in the comparison of breast and colon cancers. It can be observed that ACC levels do not vary as much as AUC levels when evaluating different combinations of gene selection and classification methods. Friedman's nonparametric test was used for all the four data sets, comparing the performance of the four gene selection and classification combinations to confirm the significance of differences in ACC and AUC levels. After Friedman's test, which confirmed significant differences between the compared gene

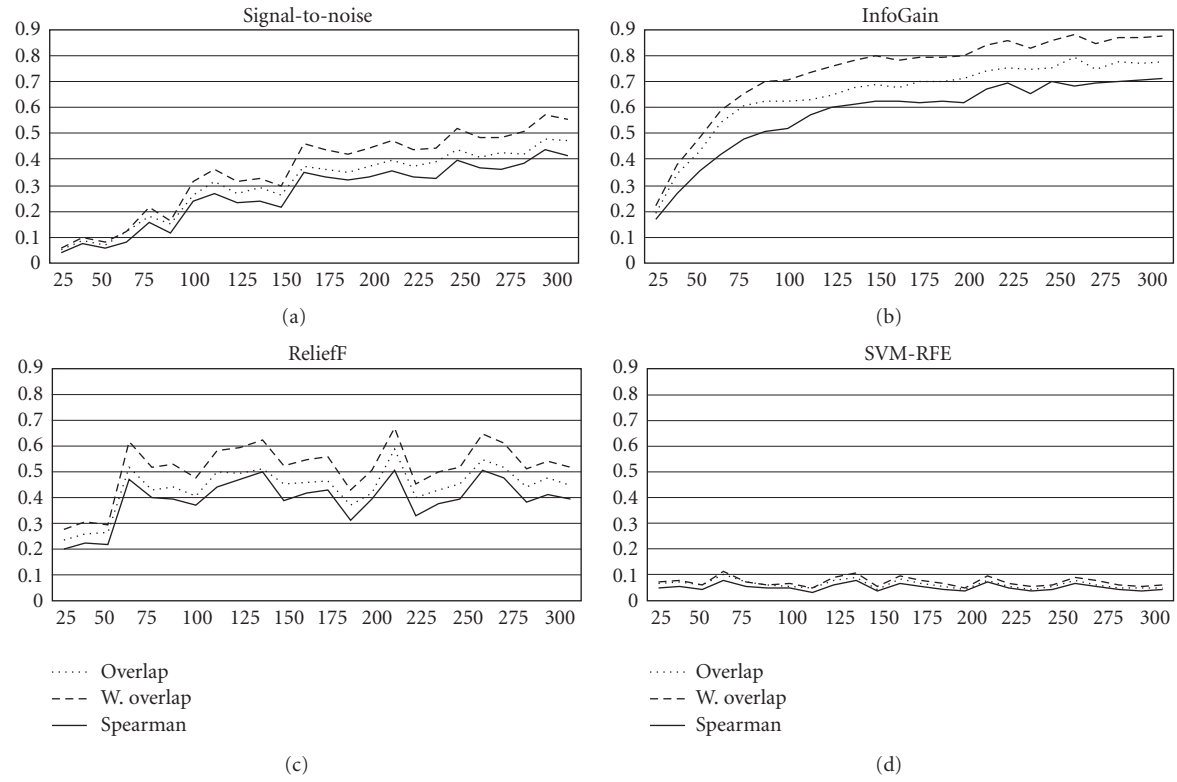

Figure 2: Relation between number of samples and stability using partitioning technique (breast versus colon cancer dataset).

selection and classification-model combinations, Wilcoxon signed ranks test was carried out.

Statistical evaluation of pairwise comparisons using Wilcoxon signed ranks test showed that in most cases, no significant difference was observed with respect to ACC, when compared with k-NN using IG or SVM-RFE as the gene selection method. The only significant difference in ACC was found in colon-kidney cancer data set, where SVM-RFE-supported k-NN classifier significantly outperformed the IG-supported k-NN classifier. There was only one additional significant difference with respect to ACC, which was confirmed using Wilcoxon test on breast-kidney data set, where IG-supported SVM significantly outperformed SVM-RFE-supported SVM classifier. However, in terms of AUC and comparison between IG- and SVM-RFE-supported classifiers, no significant differences were observed. On the whole, there were only two cases of significantly different results in 16 pairwise comparisons of ACC and AUC. The detailed statistical results of the tests described earlier as well as the additional results are available in the supplemental file (including significant difference in the performance between SVM and k-NN, shown in Figure 3).

3.4. Number of Selected Genes. All experiments described earlier used a fixed number of 100-top ranked genes to calculate stability metrics. However, changing the number of selected genes can have a significant impact on characteristics of stability performance. In case of 100 selected top genes it may happen that half or even more of them do not have any informative value on some datasets. Therefore, another experiment was designed where we observed simple overlap score at different number of selected genes for breast versus colon cancer dataset. Partitioning of dataset in two parts of equal number of samples was repeated 500 times that is more than in previous experiments to get even more reliable results. One of the reasons to introduce this kind of experiment was to clarify a very strange behavior of overlap scores for SVM-RFE selection method where stability monotonically dropped when the number of samples was getting higher. Figure 5 presents the results for two most interesting gene selection techniques—that is, Information Gain and SVM-RFE. In case of Information Gain one can observe that overlap values rise with the higher number of samples in all cases. On the other hand, SVM-RFE demonstrates very interesting pattern where only very small number of selected genes guarantees that stability will improve with higher number of samples, while in cases of larger number of selected genes, the stability drops significantly when the number of available samples increases. This fact hints at very low number of true informative genes in this dataset.

4. Discussion

Based on our results, it can be concluded that the choice of the gene selection method with respect to the aim of the study is very important. In cases that require a reproducible set of genes or single new genes for a specific disease, which still allow good classification performance, one should use

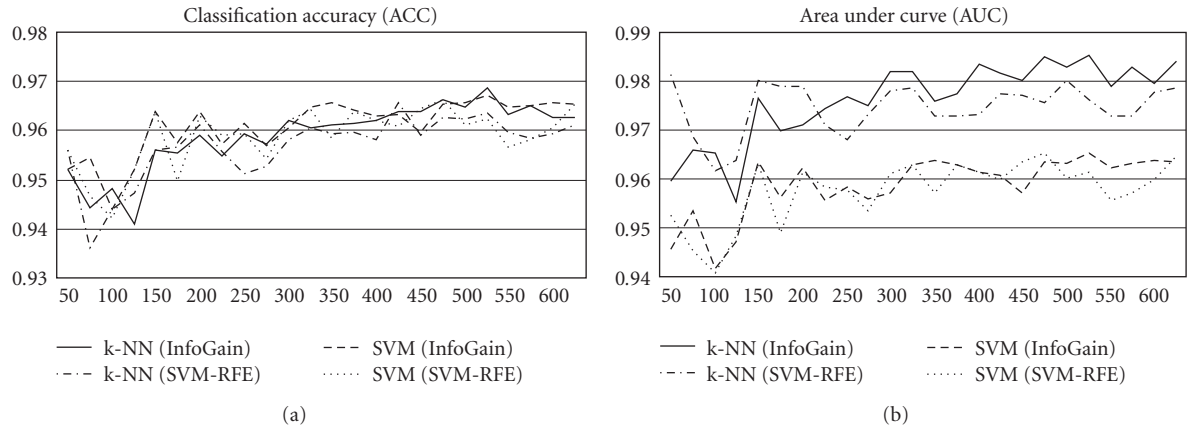

Figure 3: Classification accuracy and AUC using four different classification models (breast versus colon cancer dataset).

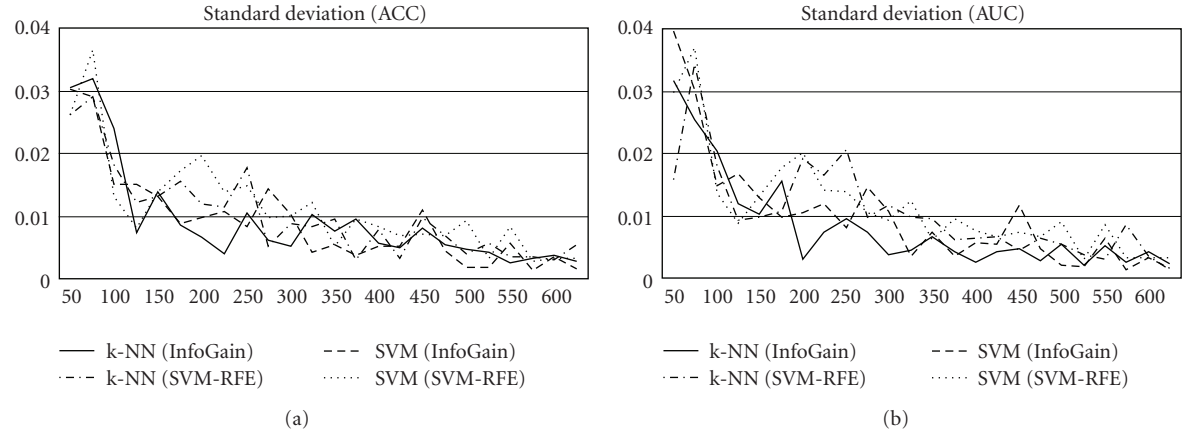

Figure 4: Standard deviation levels for classification accuracy and AUC (breast versus colon cancer dataset).

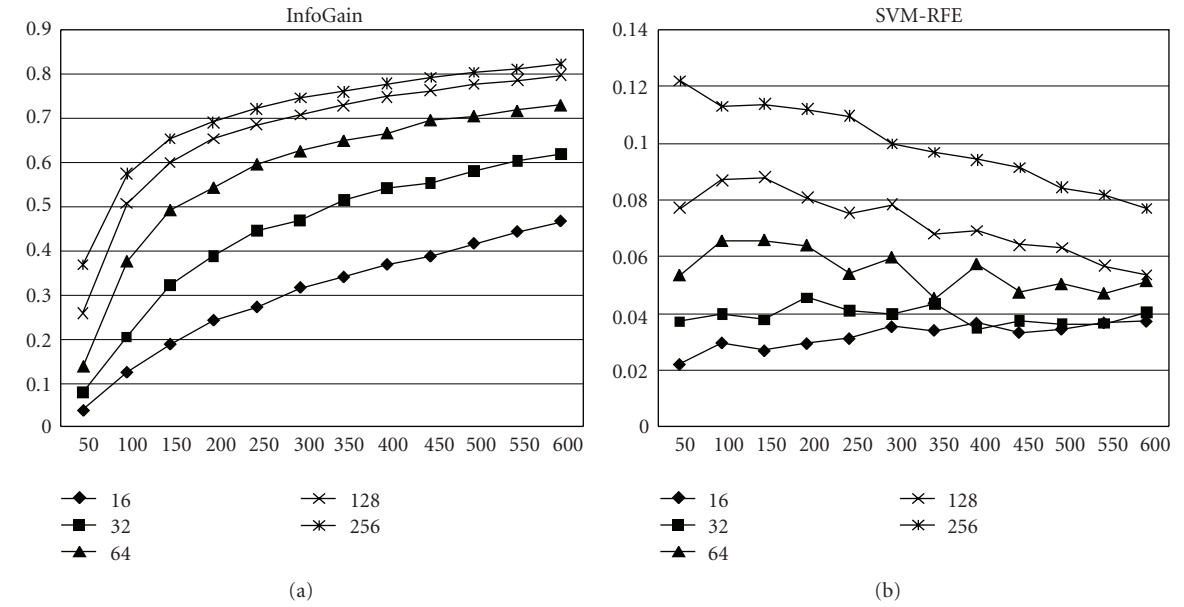

Figure 5: Overlap as a function of dataset size (number of samples) for different number of selected genes ranging from 16 to 256 (breast versus colon cancer dataset).

"good old" univariate methods. Univariate methods can offer much more reproducible and stable gene expression signatures, while multivariate methods present a possibility to discover multiple new gene expression signatures that include interactions of multiple coexpressed genes. However, it is evident that even univariate gene selection methods can perform very differently. In the case of partitioning experiment using IG, it can be observed that 50% overlap is reached with less than 75 samples, while SN ranking fails to reach that level of similarity. In addition to the observation by Ein-Dor et al., who stated that thousands of samples are needed to achieve 50% overlap between two ranked lists of genes, it should be noted that the choice of the gene ranking method used is very important. Similarity of the ranked gene lists based on SVM-RFE gene selection method demonstrates that a target of 50% overlap will hardly ever be reached, even if thousands of samples are employed in a study.

An important question to answer at this point is what makes SVM-RFE, which has been regarded as one of the most promising gene selection methods in recent years, so unstable. A part of this question was already answered by Hardin et al. [27], who claimed that a linear SVM may assign zero weights to strongly relevant variables, with weakly relevant variables getting higher weights. They compared the theoretical foundations of SVM-based feature selection with the Markov–Blanket-based gene selection techniques. In conclusion, they pointed out that gene selection algorithms that routinely miss the strongly relevant features (like SVM-RFE) do not necessarily affect the classification performance when such algorithms are used in combination with the classification models. Efficiency of unstable methods is further supported by Simon [28], who stated that one can find many correlated genes in each microarray study, and hence, it can be assumed that genes selected for inclusion in a classifier will not be stable among different studies.

Based on the assumptions of earlier studies, it is evident that SVM-RFE gene selection method works well on various different subsets of genes from the initial pool of genes. From the results presented in this study, it can be inferred that though a list of approximately 1% of genes used for classification can be very inconsistent, the SVM classifier would still be able to obtain very good classification performance using various subsets. This is very useful when we use SVM-RFE-supported classifiers exclusively for classification, but can present a serious threat to reproducibility when the classifier is used for biomarker identification. However, our last experiment hints at low number of informative genes, especially in case of SVM-RFE gene selection. It could also be said that SVM-RFE can be very helpful when searching for a number of true informative genes is our target, but one should be aware that in general SVM-RFE still lags behind in terms of stability when compared to univariate selection methods used in this study.

As already mentioned in the introductory section of this paper there are different ways of measuring the gene set stability. This opens a new question whether gene set stability represents the most appropriate way of measuring the reproducibility of microarray analysis results. There are other ways to estimate the stability of gene sets through functional enrichment of selected gene sets where it is also possible to compare overlap of enriched sets of genes. But on the other hand, we do not really know how many of gene sets used in gene set enrichment methods are true informative genes (e.g., in MSigDB [29]). Therefore, it would be interesting to assess the stability of produced gene sets by comparing the overlap of enriched groups of genes from random subsets of the original dataset. A possible approach was recently proposed by Stiglic et al. in [30].

5. Conclusions

This study compared the empirical results of four gene selection methods to calculate the stability of the gene selection lists. Three different metrics were used to compare the stability of gene selection methods to show that they mostly produced very similar results. It was observed that in most cases, simple overlap metric can be used for the estimation of gene selection stability.

The results presented also show that there is no significant difference in the accuracy between SVM-RFE-based gene selection classification and the IG-based gene selection classification. On the other hand, it was demonstrated that in terms of ranked gene lists similarity IG gene selection clearly outperforms SVM-RFE that shows the worst performance among compared gene selection methods.

Our study represents an attempt to empirically estimate the number of needed samples to achieve the desired level of stability. However, only a setting using 100 selected genes has been observed. In the future, with more computational power available, it may be possible to compare the relationships in three dimensions, consisting of the number of genes selected, number of available samples, and stability of the gene lists.

Acknowledgments

The authors acknowledge expO (Expression Project for Oncology) and International Genomics Consortium for allowing free access to the expO gene expression samples. This paper was supported by the Slovenian Research Agency through the bilateral project AGRA—Analysis of Gene Ranking Algorithms (BI-JP/09-11-002).

References

[1] S. Mook, L. J. Van't Veer, E. J. T. Rutgers, M. J. Piccart-Gebhart, and F. Cardoso, "Individualization of therapy using mammaprint: from development to the MINDACT trial," *Cancer Genomics & Proteomics*, vol. 4, no. 3, pp. 147–155, 2007.

[2] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classfication methods for tissue classification based on gene expression," *Bioinformatics*, vol. 20, no. 15, pp. 2429–2437, 2004.

[3] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[4] P. J. Park, M. Pagano, and M. Bonetti, "A nonparametric scoring algorithm for identifying informative genes from microarray data," *Pacific Symposium on Biocomputing*, pp. 52–63, 2001.

[5] E. P. Xing, M. I. Jordan, and R. M. Karp, "Feature selection for high-dimensional genomic microarray data," in *Proceedings of the 18th International Conference on Machine Learning*, pp. 601–608, 2001.

[6] L. Xu, A. C. Tan, R. L. Winslow, and D. Geman, "Merging microarray data from separate breast cancer studies provides a robust prognostic test," *BMC Bioinformatics*, vol. 9, article 125, 2008.

[7] S. Ma, "Empirical study of supervised gene screening," *BMC Bioinformatics*, vol. 7, article 537, 2006.

[8] X. Qiu, Y. Xiao, A. Gordon, and A. Yakovlev, "Assessing stability of gene selection in microarray data analysis," *BMC Bioinformatics*, vol. 7, no. 1, article 50, 2006.

[9] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 15, pp. 5923–5928, 2006.

[10] J. Bar-Ilan, M. Mat-Hassan, and M. Levene, "Methods for comparing rankings of search engine results," *Computer Networks*, vol. 50, no. 10, pp. 1448–1463, 2006.

[11] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top k lists," in *Proceedings of the 4th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '03)*, pp. 28–36, Baltimore, Md, USA, 2003.

[12] T. Barrett, D. B. Troup, S. E. Wilhite et al., "NCBI GEO: mining tens of millions of expression profiles-database and tools update," *Nucleic Acids Research*, vol. 35, no. 1, pp. D760–D765, 2007.

[13] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, San Mateo, Calif, USA, 2nd edition, 2005.

[14] C. Shannon and W. Weaver, *Mathematical Theory of Communication*, University of Illinois Press, Urbana, Ill, USA, 2002.

[15] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[16] U. M. Fayyad and K. B. Irani, "On the handling of continuous-valued attributes in decision tree generation," *Machine Learning*, vol. 8, no. 1, pp. 87–102, 1992.

[17] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proceedings of the 9th International Workshop on Machine learning*, pp. 249–256, San Mateo, Calif, USA, 1992.

[18] I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *Proceedings of the European Conference on Machine Learning*, pp. 171–181, Catania, Italy, 1994.

[19] M. R. Sikonja and I. Kononenko, "An adaptation of relief for attribute estimation in regression," in *Proceedings of the 4th International Conference on Machine Learning (ICML '97)*, pp. 296–304, Corvalis, Ore, USA, 1997.

[20] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[21] C.-H. Yeang, S. Ramaswamy, P. Tamayo et al., "Molecular classification of multiple tumor types," *Bioinformatics*, vol. 17,supplement 1, pp. S316–S322, 2001.

[22] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Information Science and Statistics, Springer, London, UK, 1999.

[23] J. Platt, "Sequential minimal optimization: a fast algorithm for training support vector machines," in *Advances in Kernel Methods—Support Vector Learning*, B. Scholkopf, J. C. Burges, and A. J. Smola, Eds., pp. 185–208, MIT Press, Cambridge, Mass, USA, 1999.

[24] L. J. Van't Veer, H. Dai, M. J. Van de Vijver et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[25] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 10, pp. 6562–6566, 2002.

[26] R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane, "Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification," *Journal of the National Cancer Institute*, vol. 95, no. 1, pp. 14–18, 2003.

[27] D. Hardin, I. Tsamardinos, and C. F. Aliferis, "A theoretical characterization of linear SVM-based feature selection," in *Proceedings of the 21ST International Conference on Machine Learning (ICML '04)*, pp. 377–384, ACM, Alberta, Canada, July 2004.

[28] R. Simon, "Development and evaluation of therapeutically relevant predictive classifiers using gene expression profiling," *Journal of the National Cancer Institute*, vol. 98, no. 17, pp. 1169–1171, 2006.

[29] A. Subramanian, P. Tamayo, V. K. Mootha et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[30] G. Stiglic, M. Bajgot, and P. Kokol, "Gene set enrichment meta-learning analysis: next- generation sequencing versus microarrays," *BMC Bioinformatics*, vol. 11, article 176, 2010.