Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 387673, 10 pages
http://dx.doi.org/10.1155/2013/387673

Research Article

A Comparative Analysis of Biomarker Selection Techniques

Nicoletta Dessì, Emanuele Pascariello, and Barbara Pes

Dipartimento di Matematica e Informatica, Università degli Studi di Cagliari, Via Ospedale 72, 09124 Cagliari, Italy

Correspondence should be addressed to Barbara Pes; pes@unica.it

Received 23 April 2013; Revised 22 September 2013; Accepted 23 September 2013

Academic Editor: Eugénio Ferreira

Feature selection has become the essential step in biomarker discovery from high-dimensional genomics data. It is recognized that different feature selection techniques may result in different set of biomarkers, that is, different groups of genes highly correlated to a given pathological condition, but few direct comparisons exist which quantify these differences in a systematic way. In this paper, we propose a general methodology for comparing the outcomes of different selection techniques in the context of biomarker discovery. The comparison is carried out along two dimensions: (i) measuring the similarity/dissimilarity of selected gene sets; (ii) evaluating the implications of these differences in terms of both predictive performance and stability of selected gene sets. As a case study, we considered three benchmarks deriving from DNA microarray experiments and conducted a comparative analysis among eight selection methods, representatives of different classes of feature selection techniques. Our results show that the proposed approach can provide useful insight about the pattern of agreement of biomarker discovery techniques.

1. Introduction

Biomarker discovery from high-dimensional genomics data is a critical problem with numerous applications in biology and medicine, such as diagnosis and treatment of complex diseases at the molecular level. As reported in [1], a biomarker can be defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention."

The discovery of biomarkers is typically modeled as a feature selection problem, where the aim is to identify the most discriminating features (i.e., genes) for a given classification task, for example, distinguishing between healthy and tumor tissues or between different tumor stages. While many feature selection techniques have been proposed [2], they do not necessarily identify the same feature subsets in the biomarker discovery process: indeed, even for the same data, different techniques can result in different groups of genes, raising questions about the biological significance of the discovered markers [3].

Surprisingly very few works in the literature have investigated, in a systematic way, the degree of similarity/dissimilarity between the outputs of different feature selection techniques in the context of biomarker discovery. Existing studies mostly focus on comparing the outcomes of different techniques in terms of predictive performance (see, e.g., [4, 5]), and, only recently, researchers have investigated the issue of stability of feature selection techniques with respect to sample variation [6, 7].

In this paper, we propose a general methodology for comparing different approaches to biomarker selection. The comparison is carried out along two dimensions: (i) measuring the similarity/dissimilarity of selected gene sets; (ii) evaluating the implications of these differences in terms of both predictive performance and stability of selected gene sets. As regards the similarity analysis, our methodology incorporates two ways of evaluating the degree of consistency among the gene sets: similarity in terms of *gene overlapping* and *functional similarity*. This twofold evaluation aims to investigate in what measure the biological functions captured by different gene sets can be similar, despite a limited overlapping among these sets. As regards the analysis of predictive performance and stability of selected biomarkers, our approach leverages on best practices from the literature [8, 9] and incorporates them into a unified comparative framework.

As a case study, we considered three benchmarks deriving from DNA microarray experiments, that is, the *Colon Tumor* dataset [10], the *Leukemia* dataset [11], and the *Prostate*

dataset [12]. In the empirical analysis, eight selection methods were included as representative of different classes of feature selection techniques. Specifically, we considered both univariate approaches that evaluate the relevance of each single gene independently from the others and multivariate approaches that take into account interdependencies among genes. Our results show that the adopted methodology can provide useful insight about the pattern of agreement of different biomarker selection methods.

The paper is organized as follows. Section 2 details the methodology, motivating it in the context of the underlying background. Section 3 illustrates the considered case study, describing the datasets, the selection methods, and the settings used in the experiments. The experimental results are presented and discussed in Section 4. Finally, Section 5 contains some final remarks and future research directions.

2. Background and Methodology

In this study we focus on feature selection methods that produce a ranking of features based on their relevance for the predictive task at hand. Referred in the following as *rankers*, these methods assign a weight to each feature according to some scoring criterion that evaluates the degree of correlation between that feature and the target class. This weighting process can be carried out in two ways [2]: evaluating each feature independently from the others (univariate methods) or taking into account feature dependencies (multivariate methods). Once each feature has been weighted, a ranked list is produced where features appear in descending order of relevance: this list can be cut at a proper threshold point in order to obtain a subset of highly predictive features.

In the context of gene selection, the resulting feature subset can be interpreted as a signature that captures significant knowledge for a given diagnostic task. Our aim here is to compare, in a systematic way, the signatures produced by different rankers; this comparison is carried out along two dimensions, as detailed in the following.

2.1. Evaluating Similarity of Selected Gene Sets. Given a dataset D with Z instances and N features (i.e., genes), a number M of rankers R_i $(i = 1, 2, \ldots, M)$ are applied to D: each R_i produces a ranked list L_i $(i = 1, 2, \ldots, M)$ where the N features appear in descending order of relevance. As illustrated in Figure 1, this results in M distinct ranked lists each expressing a different ordering of the N genes.

When two lists L_i and L_i $(i, j = 1, 2, \ldots, M)$ are cut at a given threshold point t, the resulting gene sets S_i and S_j $(i, j = 1, 2, \ldots, M)$, of the same size t, can be compared according to some similarity index I. In particular, our methodology incorporates two approaches to measure the degree of similarity/dissimilarity among selected gene sets: similarity in terms of gene overlapping (I-overlap) and functional similarity(I-functional).

The similarity I-overlap between two sets S_i and S_j can be expressed as the number of genes that are present in both sets, that is, $|S_i \cap S_j|$, properly normalized in the range $[0, 1]$, with 0 meaning that there is no overlap between the two sets and 1 that the two sets are identical. As normalization factor, we use $|S_i \cup S_j|$, as in [8]. After computing the I-overlap value for each pair of gene sets, we average over all pairwise similarity comparisons to obtain an overall evaluation of the degree of similarity between the M gene sets.

It has been observed, however, that the biological functions captured by different gene sets can be similar, despite a little degree of overlapping between these sets [13–15]. To compare two gene sets in functional terms, we exploit gene annotations from the Gene Ontology (GO) database [16], which provides a set of controlled vocabularies (biological or biochemical terms) describing gene products based on their functions in the cell. Specifically, for each gene set S_i $(i = 1, 2, \ldots, M)$, we extract the list of *molecular function* GO terms that annotate the t genes in the set. The resulting M lists of GO terms are then compared, in pairs, using the similarity measure (I-functional) proposed in [17] which considers not only the overlap between the lists but also the semantic relationships between GO terms.

2.2. Evaluating Predictive Performance and Stability of Selected Gene Sets. The predictive performance of a candidate gene set, that is, its capacity of discriminating a given target class (e.g., a pathological state), can be measured inducing a classification model on that set and using some test instances to evaluate this model in terms of metrics such as accuracy or ROC area [18]. This is usually done in a cross-validation setting, though it has been observed that it can produce overoptimistic results on small sample size domains [19].

Instead, no well-established evaluation protocol exists for measuring the stability of a biomarker selection algorithm, that is, its robustness with respect to sample variation: small changes in the original dataset should not affect the outcome of the selection process in a significant way. Research work on designing a suitable experimental procedure for testing stability in high-dimensional/small sample size domains is still ongoing [7], and in most cases stability is not evaluated in conjunction with predictive performance but in independent experiments.

The methodology we adopt involves a single experimental setup to jointly evaluate both stability and predictive performance in the context of biomarker discovery. As illustrated in Figure 2, we extract from the original dataset D, with Z instances and N features (i.e., genes), a number P of reduced dataset D_k $(k = 1, 2, \ldots, P)$, each containing $f \cdot Z$ (with $f \in (0, 1)$) instances randomly drawn from D.

Each of the previously considered rankers R_i $(i = 1, 2, \ldots, M)$ is then applied to each reduced datasets D_k $(k = 1, 2, \ldots, P)$ in order to obtain a ranked list L_{ik} and, after cutting the list at threshold t, a gene subset S_{ik}. The P subsets selected by the ranker R_i from the P reduced datasets are then compared in terms of overlapping: the more similar (in average) these subsets are, the more stable the ranker R_i is.

We observe that the I-overlap measure is used in our approach in a twofold way: to compare the subsets produced by different rankers on the same dataset D and to compare the subsets produced by the same ranker on different reduced datasets drawn from D. Moreover, it should be observed

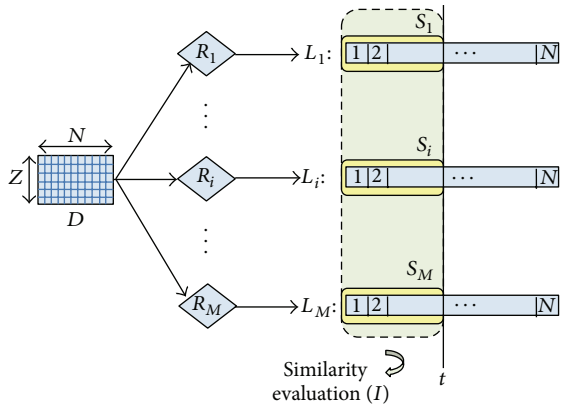


Figure 1: Similarity evaluation.

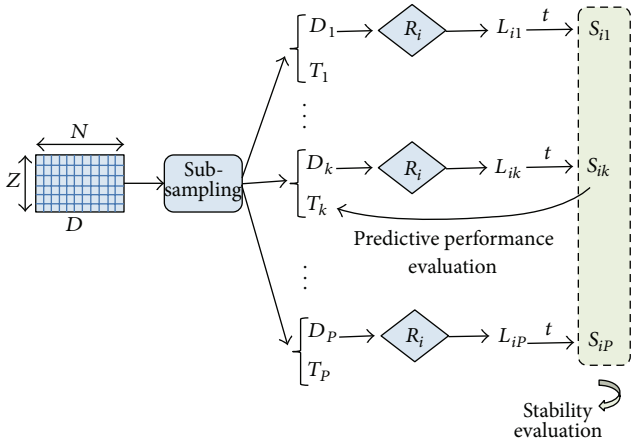


Figure 2: Joint evaluation of stability and predictive performance.

that the recent literature suggests the Kuncheva index [20] as a more suitable similarity measure in the context of stability evaluation: it considers the degree of overlapping between two feature subsets and introduces a correction term that takes into account the probability that a feature is included in those subsets purely by chance. This correction term, however, does not affect the similarity value for feature subsets of small size, as the ones considered in the context of biomarker discovery [21].

To incorporate predictive performance evaluation in the above experimental protocol, we build on each reduced dataset D_k ($k = 1, 2, \ldots, P$) a classification model based on the gene set S_{ik} selected by the ith ranker: the model performance is then estimated on a test set T_k containing the fraction of instances of D not included in D_k (i.e., $(1 - f) \cdot Z$ instances). As performance metric we use the AUC (area under the ROC curve), as it synthesizes the information of sensitivity and specificity and provides a more reliable estimate in the case of unbalanced class distribution [22]. By averaging the AUC performance of the P classification models induced on the P gene subsets selected by the ranker R_i, we can evaluate the effectiveness of that ranker in identifying highly predictive gene sets. This approach overcomes the risk of selection bias [23] since the test instances are not considered in any way in the gene selection stage.

The above methodology ensures a joint evaluation of two fundamental requirements of any biomarker selection technique, that is, stability with respect to sample variation and effectiveness in terms of classification results, enabling the comparison of different techniques in a unified framework.

3. Case Study: Datasets and Settings

Consistently with the methodology described in Section 2, we conducted an empirical analysis on three benchmarks deriving from DNA microarray experiments.

(i) *Colon Tumor* dataset [10], containing 62 biological samples distinguished between tumor *colon* tissues (40 samples) and normal *colon* tissues (22 samples); each sample is described by the expression level of 2000 genes.

(ii) *Leukemia* dataset [11], containing 72 samples belonging to patients suffering from acute myeloid *leukemia* (25 samples) and acute lymphoblastic *leukemia* (47 samples); each sample is described by the expression level of 7129 genes.

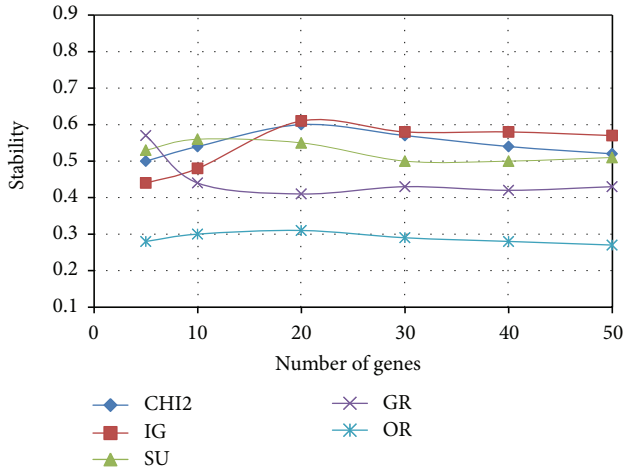
(iii) *Prostate* dataset [12], containing 102 samples differed between healthy and tumor *prostate* tissues (50 and 52 samples, resp.); each sample is described by the expression level of 12600 genes.

The task, in terms of feature selection, is to identify the genes most useful in discriminating between cancerous and normal tissues (*Colon* and *Prostate* datasets) or between different tumor types (*Leukemia* dataset).

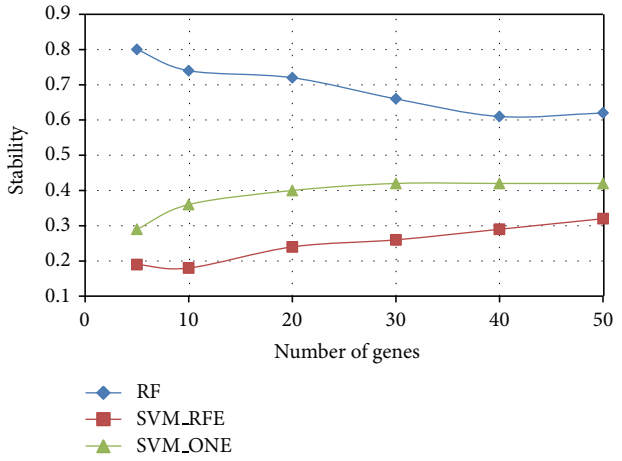
In our experiments, we compared $M = 8$ rankers that are representative of different classes of selection methods. In particular, we considered both univariate approaches, where each feature is ranked individually, and multivariate approaches that take into account feature dependencies.

Among the univariate techniques, we chose: chi Squared (χ^2) [24] as representative of statistic methods; information gain (IG) [25], symmetrical uncertainty (SU) [26], and gain ratio (GR) [27] as representatives of entropic methods; and finally OneR (OR) [28] as representative of methods that incorporate a classification technique (in this case, a simple rule-based classifier).

Among the multivariate techniques, we considered *ReliefF* (RF) [29] and *SVM-embedded feature selection* [30]. The basic idea of RF is to estimate the relevance of features based on their ability to distinguish between instances that are near to each other. Instead, SVM-embedded feature selection uses a linear SVM classifier to derive a weight for each feature. Then, based on their weights, the features can be ordered from the most important to the less important (SVM_ONE approach). Moreover, a backward elimination strategy can be adopted which iteratively removes the features with the lowest weights and repeats the overall weighting process on the remaining features (SVM_RFE approach). The fraction

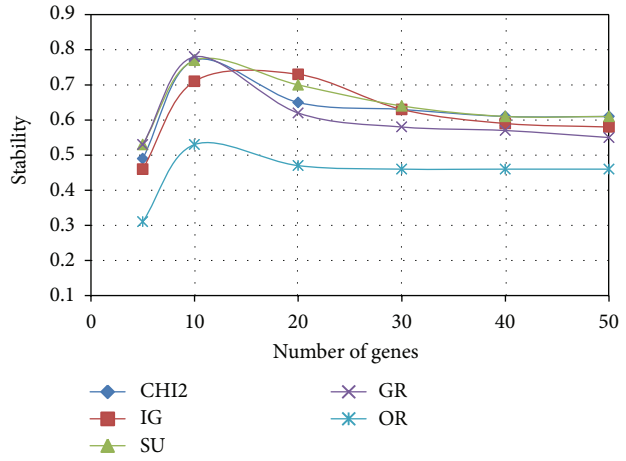


(a) Univariate ranking methods

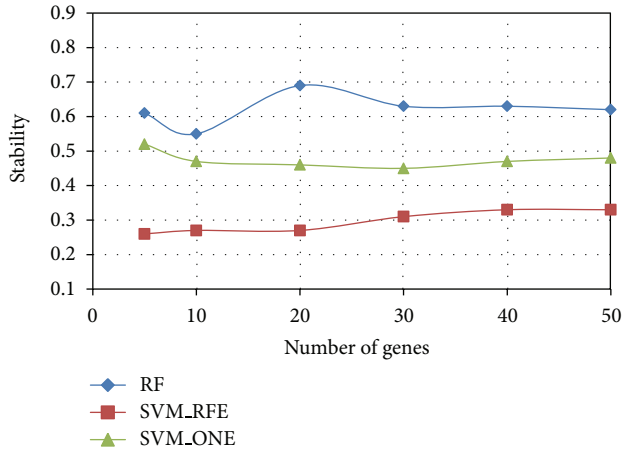


(b) Multivariate ranking methods

FIGURE 3: *Colon* dataset: stability versus number of genes.



(a) Univariate ranking methods



(b) Multivariate ranking methods

FIGURE 4: *Leukemia* dataset: stability versus number of genes.

of features removed at each iteration, 10% in our experiments, greatly influences the computational complexity of the method.

For all the above feature selection techniques we used the implementation provided by the WEKA machine learning environment [31]. To systematically evaluate the degree of overlapping between the gene subsets selected by the considered rankers, we developed a software module that interfaces with WEKA. In what concerns the functional aspects, the similarity analysis was performed by the online tools available at [32].

As regards the evaluation of predictive performance and stability of selected gene subsets, the parameters of our methodology were set as follows: a number $P = 20$ of reduced datasets were extracted from the original dataset, and each reduced dataset contains a fraction $f = 0.9$ of the original samples. The ranked lists produced on these datasets by the $M = 8$ rankers were cut at different threshold points ($t = 5$, $t = 10$, $t = 20$, $t = 30$, etc.) as to evaluate stability and predictive performance for gene subsets of increasing size. In evaluating the predictive performance, we used as induction algorithm a linear SVM classifier, which is widely considered the "best of class" method in the context of microarray data analysis; specifically, we employed the SVM implementation provided by WEKA.

4. Experimental Results and Discussion

In this section we present and discuss the most significant experimental results. First, we concentrate on findings from the similarity analysis among the gene subsets selected by different rankers (see Section 2.1); then we examine the results of the joint evaluation of stability and predictive performance of selected gene subsets (see Section 2.2).

4.1. Results of Similarity Analysis. The similarity analysis was first performed in terms of gene overlapping; that is, we used the I-overlap index to compare, in pairs, the subsets produced by the ranking methods presented in Section 3 (χ^2, IG, SU, GR, OR, RF, SVM_RFE, and SVM_ONE). Table 1 shows the results of this comparison for gene subsets of size $t = 10$. Specifically, for each of the considered datasets,

Table 1: Similarity in terms of gene overlapping.

(a) *Colon* dataset

	CHI2	IG	SU	GR	OR	RF	SVM_RFE	SVM_ONE
CHI2		0.67	0.67	0.43	0.54	0.43	0.18	0.00
IG	0.67		0.67	0.43	0.43	0.25	0.18	0.00
SU	0.67	0.67		0.54	0.33	0.43	0.18	0.00
GR	0.43	0.43	0.54		0.33	0.33	0.11	0.00
OR	0.54	0.43	0.33	0.33		0.33	0.05	0.00
RF	0.43	0.25	0.43	0.33	0.33		0.11	0.00
SVM_RFE	0.18	0.18	0.18	0.11	0.05	0.11		0.11
SVM_ONE	0.00	0.00	0.00	0.00	0.00	0.00	0.11	

(b) *Leukemia* dataset

	CHI2	IG	SU	GR	OR	RF	SVM_RFE	SVM_ONE
CHI2		0.82	1	1	1	0.33	0.18	0.11
IG	0.82		0.82	0.82	0.82	0.43	0.25	0.18
SU	1	0.82		1	1	0.33	0.18	0.11
GR	1	0.82	1		1	0.33	0.18	0.11
OR	1	0.82	1	1		0.33	0.18	0.11
RF	0.33	0.43	0.33	0.33	0.33		0.25	0.43
SVM_RFE	0.18	0.25	0.18	0.18	0.18	0.25		0.33
SVM_ONE	0.11	0.18	0.11	0.11	0.11	0.43	0.33	

(c) *Prostate* dataset

	CHI2	IG	SU	GR	OR	RF	SVM_RFE	SVM_ONE
CHI2		0.82	0.67	0.54	0.54	0.18	0.05	0.05
IG	0.82		0.67	0.54	0.43	0.18	0.05	0.05
SU	0.67	0.67		0.82	0.54	0.25	0.05	0.05
GR	0.54	0.54	0.82		0.43	0.18	0.00	0.05
OR	0.54	0.43	0.54	0.43		0.18	0.05	0.05
RF	0.18	0.18	0.25	0.18	0.18		0.11	0.25
SVM_RFE	0.05	0.05	0.05	0.00	0.05	0.11		0.25
SVM_ONE	0.05	0.05	0.05	0.05	0.05	0.25	0.25	

that is, (a) *Colon*, (b) *Leukemia*, and (c) *Prostate*, the results are represented in a matricial form: each cell contains the I-overlap value for the pair of subsets selected by the rankers in the corresponding row and column. Different shades of gray are used to highlight different similarity ranges: the darker the gray, the higher the similarity values. The average similarity over all pairwise comparisons is 0.28 for *Colon*, 0.49 for *Leukemia*; and 0.29 for *Prostate* (excluding the cells in the main diagonal where each subset is compared with itself).

Results in Table 1 give useful insight about the pattern of agreement of the considered methods. As regards the univariate approaches (i.e., χ^2, IG, SU, GR, and OR), first evidence is that the χ^2 statistic produces results quite similar to entropic methods IG and SU (I-overlap ≥ 0.67 for all the considered benchmarks). The other entropic method, that is, GR, turns out very similar to both IG and SU in the *Leukemia* dataset but exhibits a somewhat different behavior in the other datasets, especially *Colon* which is recognized as a more noisy benchmark. Globally, the univariate methods are more similar to each other than to the multivariate approaches (i.e., RF, SVM_RFE, and SVM_ONE). In particular, the SVM-embedded feature selection produces feature subsets that overlap to a small extent (or do not overlap at all) with the subsets selected by other methods.

As a further step, the same gene subsets of size $t = 10$ were compared in functional terms based on the molecular function GO annotations of genes in each subset. Results of this comparison are shown in Table 2. Again, for each dataset, results are reported in a matricial form: each cell contains here the I-functional value for a pair of subsets selected by the considered ranking methods. The average similarity is 0.78 for *Colon*, 0.86 for *Leukemia*, and 0.79 for *Prostate*.

Though similarity values in Tables 1 and 2 are not directly comparable, due to the different similarity measures, the ontological analysis shows that the functions captured by different gene subsets can be similar, despite a little degree of overlapping between these subsets. Interestingly, even two subsets with no genes in common may exhibit a fairly high level of functional similarity. Hence, there may be common functions shared across different subsets that are not

Table 2: Functional similarity.

(a) *Colon* dataset

	CHI2	IG	SU	GR	OR	RF	SVM_RFE	SVM_ONE
CHI2		0.99	0.92	0.83	0.96	0.76	0.76	0.65
IG	0.99		0.93	0.84	0.95	0.74	0.76	0.66
SU	0.92	0.93		0.87	0.87	0.79	0.73	0.66
GR	0.83	0.84	0.87		0.83	0.73	0.69	0.63
OR	0.96	0.95	0.87	0.83		0.77	0.74	0.69
RF	0.76	0.74	0.79	0.73	0.77		0.63	0.63
SVM_RFE	0.76	0.76	0.73	0.69	0.74	0.63		0.75
SVM_ONE	0.65	0.66	0.66	0.63	0.69	0.63	0.75	

(b) *Leukemia* dataset

	CHI2	IG	SU	GR	OR	RF	SVM_RFE	SVM_ONE
CHI2		0.99	1	1	1	0.82	0.77	0.76
IG	0.99		0.99	0.99	0.99	0.83	0.78	0.77
SU	1	0.99		1	1	0.82	0.77	0.76
GR	1	0.99	1		1	0.82	0.77	0.76
OR	1	0.99	1	1		0.82	0.77	0.76
RF	0.82	0.83	0.82	0.82	0.82		0.80	0.83
SVM_RFE	0.77	0.78	0.77	0.77	0.77	0.80		0.86
SVM_ONE	0.76	0.77	0.76	0.76	0.76	0.83	0.86	

(c) *Prostate* dataset

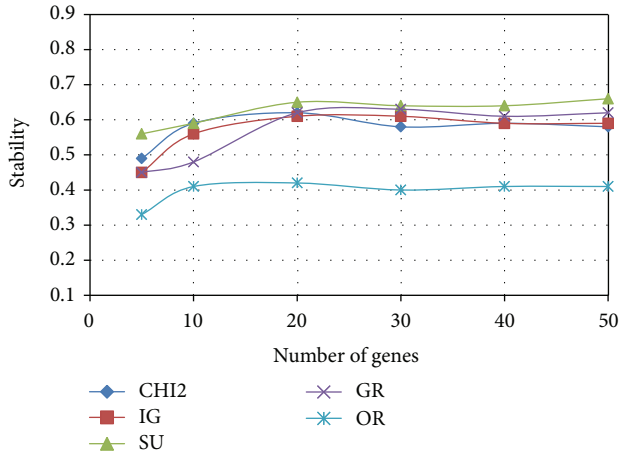
	CHI2	IG	SU	GR	OR	RF	SVM_RFE	SVM_ONE
CHI2		0.98	0.96	0.90	0.97	0.77	0.71	0.72
IG	0.98		0.96	0.90	0.94	0.77	0.69	0.69
SU	0.96	0.96		0.95	0.95	0.79	0.70	0.70
GR	0.90	0.90	0.95		0.89	0.69	0.58	0.66
OR	0.97	0.94	0.95	0.89		0.79	0.72	0.72
RF	0.77	0.77	0.79	0.69	0.79		0.70	0.77
SVM_RFE	0.71	0.69	0.70	0.58	0.72	0.70		0.68
SVM_ONE	0.72	0.69	0.70	0.66	0.72	0.77	0.68	

apparent on the individual gene level. This helps explain why different selection methods can produce different biological signatures: these signatures may be in some way "consistent," even if they do not contain the same genes.

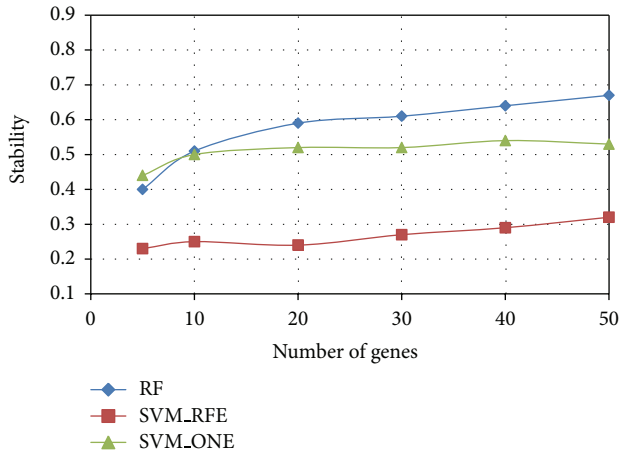
4.2. Results about Stability and Predictive Performance. After evaluating the degree of similarity/dissimilarity among the outcomes of different ranking methods, we empirically examined the implications of these differences in terms of both stability and predictive performance of selected gene subsets. In Figures 3, 4, and 5 we summarize, respectively, for *Colon*, *Leukemia*, and *Prostate* datasets, the results of stability analysis on gene subsets of increasing size. As explained in Section 2.2, the stability value was obtained, for a given ranking method, as the average similarity (I-overlap) among the gene subsets selected by this method from a number $P = 20$ of reduced datasets randomly drawn from the original dataset.

Among the univariate approaches (χ^2, IG, SU, GR, and OR), χ^2 and the entropic methods IG and SU exhibit, in each dataset, a similar trend in terms of stability, while GR slightly deviates from the other entropic methods in the *Colon* dataset. The worst performing univariate method is OR, which always results in a poor stability irrespective of the number of genes included in the subset. Among the multivariate approaches (RF, SVM_RFE, and SVM_ONE), RF outperforms the SVM-embedded feature selection in each of the benchmarks here considered. In particular, though SVM_RFE is known in the literature [30, 33] as a very effective feature selection technique, it exhibits the worst behavior in terms of stability.

As regards the evaluation of predictive performance, we trained a linear SVM classifier on each of the $P = 20$ gene subsets (of a given size) selected by a given ranking method from the reduced datasets randomly drawn from the original dataset: these reduced datasets serve at this stage as training sets. The average AUC performance, measured on the independent test sets (see Section 2.2), is shown in Figure 6 (*Colon*), Figure 7 (*Leukemia*), and Figure 8 (*Prostate*) for both univariate (χ^2, IG, SU, GR, and OR) and multivariate

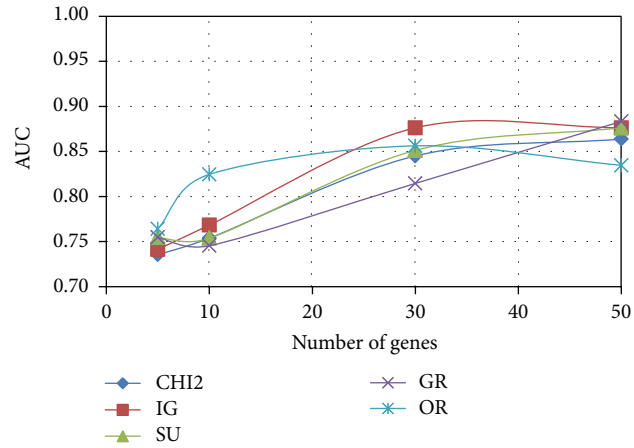


(a) Univariate ranking methods

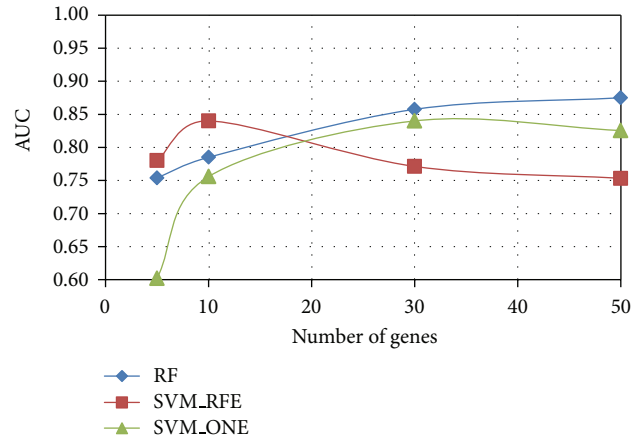


(b) Multivariate ranking methods

Figure 5: *Prostate* dataset: stability versus number of genes.



(a) Univariate ranking methods



(b) Multivariate ranking methods

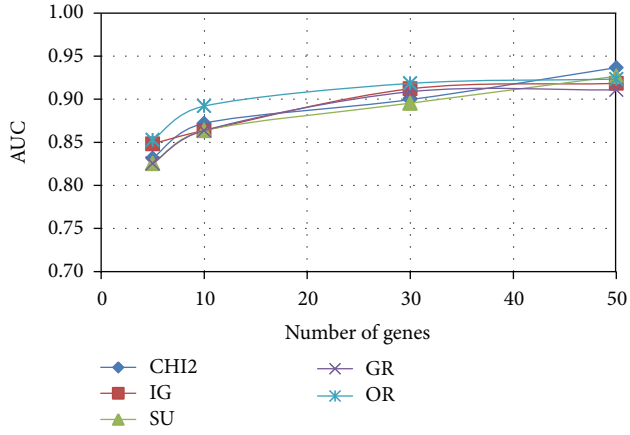
Figure 6: *Colon* dataset: AUC versus number of genes.

methods (RF, SVM_RFE, and SVM_ONE); in each figure, the AUC trend is reported for gene subsets of increasing size.

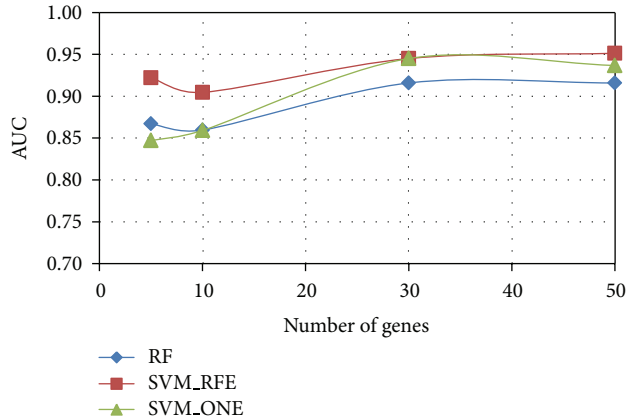
As we can see, χ^2 and entropic methods globally exhibit a similar behavior, almost coincident in *Leukemia* and *Prostate* datasets, with a slight superiority of IG in the more problematic *Colon* dataset: here GR turns out to be, for subsets of small-moderate size (<40), the worst performing univariate method. Interestingly, the OR method performs well in terms of AUC (even better than other univariate approaches, for subsets of small size, in both *Colon* and *Leukemia* datasets), though its behavior in terms of stability is quite poor.

As regards the AUC performance of multivariate approaches, there is no method that univocally outperforms the others, contrary to what is observed in the stability analysis. In the *Prostate* dataset, indeed, the three multivariate methods are almost equivalent, while greater differences can be observed in the *Leukemia* dataset and, even more, in the *Colon* dataset. However, it is worth remarking that SVM_RFE, in all the considered benchmarks, is very effective in identifying small subsets of highly predictive genes, despite its very low stability. We also observe that RF, which is the more stable multivariate method, has globally a good performance also in terms of AUC.

To conclude, a number of observations can be drawn from the joint analysis of stability and AUC patterns of the eight ranking methods considered in this study. As a first point, a high level of agreement exists between the behavior of the statistical approach χ^2 and the behavior of entropic approaches, especially SU and IG. However, in the *Colon* dataset (which is recognized as a more challenging benchmark), the entropic method GR performs worse, probably due to its higher sensitivity to noise [34]. Moreover, it is interesting to highlight that the less stable methods, that is, OR in the univariate category and SVM_RFE in the multivariate category, are both capable of selecting small-sized subsets of highly predictive genes. Such cases of instability coupled with high predictive performance could be explained in terms of redundancy within the full set of genes: the dataset may contain various markers that are highly correlated which might lead the algorithm to select different genes on different samples [7]. Globally, χ^2, SU, and IG, representatives of univariate approaches, and RF, representative of multivariate approaches, seem to best satisfy

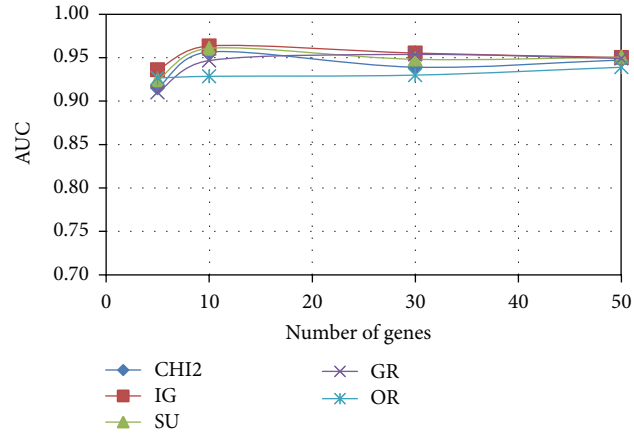


(a) Univariate ranking methods

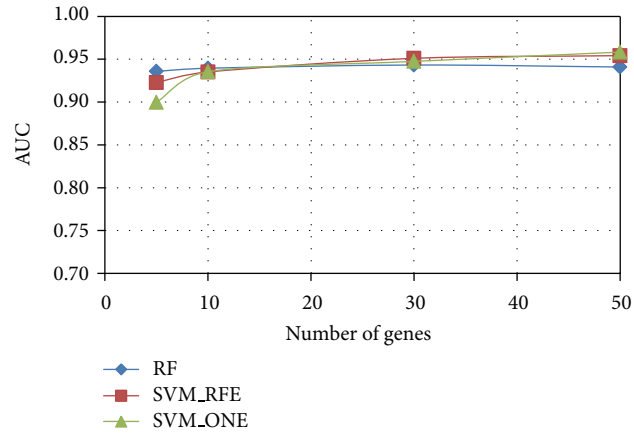


(b) Multivariate ranking methods

Figure 7: *Leukemia* dataset: AUC versus number of genes.



(a) Univariate ranking methods



(b) Multivariate ranking methods

Figure 8: *Prostate* dataset: AUC versus number of genes.

the objective of jointly optimized stability and effectiveness of selected biomarkers.

5. Concluding Remarks and Future Research Directions

A methodology has been presented for comparing the outcomes of different feature selection techniques in the context of biomarker discovery. Leveraging on best practices from the literature, the proposed approach enables a multifaceted evaluation of the degree of consistency among the genetic signatures selected by different techniques.

As a case study, three public benchmarks have been used to empirically evaluate the pattern of agreement of some popular biomarker discovery methods. For future work, further experiments will be performed using more datasets as well as different selection methods. Moreover, different similarity measures could be incorporated in our methodology, especially in what concerns the evaluation of the functional similarity among signatures, which is recognized as a controversial research problem [15].

We also observe that the approach presented in this paper can be a starting point for defining a suitable "ensemble" strategy for biomarker selection. Indeed, recent research efforts attempt to combine multiple feature selection techniques, instead of using a single one, in order to overcome the intrinsic limitations of each technique and obtain a more reliable "consensus" result (e.g., a consensus ranking or a consensus subset containing the most frequently selected features). However, this combination is often made on an "ad hoc" basis [35–39], depending on the specific problem at hand, without considering the degree of diversity/similarity of the involved methods. In our opinion, instead, this important aspect should not be neglected: it would not be beneficial, indeed, to combine two or more techniques that give almost identical results. On the contrary, in an ensemble perspective, the aim should be to reach a consensus result among methods that are capable of giving different and complementary representations of the considered domain. On this premise, our future research will explore suitable ways of combining biomarker selection techniques on the basis of their degree of diversity/similarity, as assessed according to the approach here discussed.

Acknowledgment

This research was supported by Regione Autonoma della Sardegna (RAS) (Legge regionale 7 agosto 2007, no. 7 "Promozione della ricerca scientifica e dell'innovazione tecnologica in Sardegna") in the project "*DENIS: Dataspaces Enhancing the Next Internet in Sardinia.*"

References

[1] J. A. Arthur, W. A. Colburn, V. G. DeGruttola et al., "Biomarkers and surrogate endpoints: preferred definitions and conceptual framework," *Clinical Pharmacology and Therapeutics*, vol. 69, no. 3, pp. 89–95, 2001.

[2] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[3] J. P. A. Ioannidis, "Microarrays and molecular research: noise discovery?" *The Lancet*, vol. 365, no. 9458, pp. 454–455, 2005.

[4] C. Lai, M. J. T. Reinders, L. J. van't Veer, and L. F. A. Wessels, "A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets," *BMC Bioinformatics*, vol. 7, article 235, 2006.

[5] I. B. Jeffery, D. G. Higgins, and A. C. Culhane, "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data," *BMC Bioinformatics*, vol. 7, article 359, 2006.

[6] Z. He and W. Yu, "Stable feature selection for biomarker discovery," *Computational Biology and Chemistry*, vol. 34, no. 4, pp. 215–225, 2010.

[7] W. Awada, T. M. Khoshgoftaar, D. J. Dittman, R. Wald, and A. Napolitano, "A review of the stability of feature selection techniques for bioinformatics data," in *Proceedings of the IEEE 13th International Conference on Information Reuse and Integration*, pp. 356–363, 2012.

[8] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowledge and Information Systems*, vol. 12, no. 1, pp. 95–116, 2007.

[9] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust Feature Selection Using Ensemble Feature Selection Techniques," in *Proceedings of the European Conference ECML (PKDD '08)*, vol. 5212 of *Lecture Notes in Artificial Intelligence*, pp. 313–325, Springer, 2008.

[10] U. Alon, N. Barka, D. A. Notterman et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 12, pp. 6745–6750, 1999.

[11] 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. 531–527, 1999.

[12] D. Singh, P. G. Febbo, K. Ross et al., "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.

[13] R. Shen, A. Chinnaiyan, and D. Ghosh, "Pathway analysis reveals functional convergence of gene expression profiles in breast cancer," *BMC Medical Genomics*, vol. 1, no. 1, article 28, 2008.

[14] F. Reyal, M. H. van Vliet, N. J. Armstrong et al., "A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the Proliferation, Immune response and RNA splicing modules in breast cancer," *Breast Cancer Research*, vol. 10, no. 6, article R93, 2008.

[15] P. Wirapati, C. Sotiriou, S. Kunkel et al., "Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures," *Breast Cancer Research*, vol. 10, no. 4, article R65, 2008.

[16] http://www.geneontology.org.

[17] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.

[18] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Elsevier, 3rd edition, 2011.

[19] U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification?" *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.

[20] L. I. Kuncheva, "A Stability Index for Feature Selection," in *Proceedings of the International Multi-Conference Artificial Intelligence and Applications*, ACTA Press, Anaheim, Calif, USA, 2007.

[21] L. M. Cannas, N. Dessì, and B. Pes, "Assessing similarity of feature selection techniques in high-dimensional domains," *Pattern Recognition Letters*, vol. 34, no. 12, pp. 1446–1453, 2013.

[22] T. Fawcett, "ROC graphs: notes and practical considerations for researchers," Tech. Rep. HPL-2003-4, HP Laboratories, 2003.

[23] 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.

[24] H. Liu and R. Setiono, "Chi2: feature selection and discretization of numeric attributes," in *Proceedings of the IEEE 7th International Conference on Tools with Artificial Intelligence*, pp. 388–391, November 1995.

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

[26] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C*, 1998.

[27] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Francisco, Calif, USA, 1993.

[28] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, no. 1, pp. 63–91, 1993.

[29] I. Kononenko, "Estimating attributes: analysis and extensions of RELIEF," in *Proceedings of the European Conference on Machine Learning*, pp. 171–182, 1994.

[30] 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.

[31] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[32] http://bioinformatics.clemson.edu/G-SESAME.

[33] X. Zhou and D. P. Tuck, "MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data," *Bioinformatics*, vol. 23, no. 9, pp. 1106–1114, 2007.

[34] W. Altidor, T. M. Khoshgoftaar, and J. Van Hulse, "Robustness of filter-based feature ranking: a case study," in *Proceedings*

of the 24th International Florida Artificial Intelligence Research Society (FLAIRS '11), pp. 453–458, May 2011.

[35] N. Dessì and B. Pes, "An evolutionary method for combining different feature selection criteria in microarray data classification," *Journal of Artificial Evolution and Applications*, vol. 2009, Article ID 803973, 10 pages, 2009.

[36] J. Dutkowski and A. Gambin, "On consensus biomarker selection," *BMC Bioinformatics*, vol. 8, supplement 5, article S5, 2007.

[37] Y. Leung and Y. Hung, "A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 1, pp. 108–117, 2010.

[38] T. Feng, F. Xuezheng, Z. Yanqing, and A. G. Bourgeois, "Improving feature subset selection using a genetic algorithm for microarray gene expression data," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC '06)*, pp. 2529–2534, Vancouver, Canada, July 2006.

[39] P. Yang, B. B. Zhou, Z. Zhang, and A. Y. Zomaya, "A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data," *BMC Bioinformatics*, vol. 11, supplement 1, article S5, 2010.