BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 2 2008, pages 258–264
doi:10.1093/bioinformatics/btm550

Data and text mining

Algebraic stability indicators for ranked lists in molecular profiling

Giuseppe Jurman[1], Stefano Merler[1], Annalisa Barla[1,2], Silvano Paoli[1,3], Antonio Galea[1] and Cesare Furlanello[1,*]

[1]FBK, via Sommarive 18, I-38100 Povo (Trento), [2]DISI, University of Genova, via Dodecaneso 35, I-16146 Genova and [3]DIT, University of Trento, via Sommarive 14, I-38100 Povo (Trento), Italy

Received on May 7, 2007; revised on October 12, 2007; accepted on October 31, 2007

Advance Access publication November 16, 2007

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: We propose a method for studying the stability of biomarker lists obtained from functional genomics studies. It is common to adopt resampling methods to tune and evaluate marker-based diagnostic and prognostic systems in order to prevent selection bias. Such caution promotes honest estimation of class prediction, but leads to alternative sets of solutions. In microarray studies, the difference in lists may be bewildering, also due to the presence of modules of functionally related genes. Methods for assessing stability understand the dependency of the markers on the data or on the predictor's type and help selecting solutions.

Results: A computational framework for comparing sets of ranked biomarker lists is presented. Notions and algorithms are based on concepts from permutation group theory. We introduce several algebraic indicators and metric methods for symmetric groups, including the Canberra distance, a weighted version of Spearman's footrule. We also consider distances between partial lists and an aggregation of sets of lists into an optimal list based on voting theory (Borda count). The stability indicators are applied in practical situations to several synthetic, cancer microarray and proteomics datasets. The addressed issues are predictive classification, presence of modules, comparison of alternative biomarker lists, outlier removal, control of selection bias by randomization techniques and enrichment analysis.

Availability: Supplementary Material and software are available at the address http://biodcv.fbk.eu/listspy.html

Contact: furlan@fbk.eu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

We discuss the problem of measuring the stability of a ranked feature list in functional genomics. Stability (measure of robustness to data perturbations) has only recently arisen as a major issue when applied to feature selection and ranking processes. In fact, selecting a reliable set of biomarkers is as important as achieving predictive classification (Baker and Kramer, 2006), and the two goals cannot always be pursued together (Marchiori, 2006). The concept of stability is indeed central in machine learning and in other workfields (see Wichmann and Kamholz, 2006 for an example in linguistics). Discussions of the feature selection stability problem can be found in Davis *et al.* (2006) for the microarray case and in Kalousis *et al.* (2005) for mass spectrometry data. Many attempts ran into the same difficulty: the resulting lists of biomarkers are very unstable with the slightest perturbation of the training set producing drastically different markers. Such issue arises when using methods indicated for error assessment and model selection on high-throughput data. Designed to reduce variance and avoid selection bias, they are structured in schemes of replicated experiments on versions of the original dataset obtained by stratified resampling and partitions in training and test subsets (Ambroise and McLachlan, 2002; Molinaro *et al.*, 2005). In a binary classification task on microarray data (e.g. disease versus control), such a scheme provides average of errors on test sets for models trained on k genes from the corresponding training sets. The no-information case (e.g. the maximal instability situation characterized by random lists) can be used as baseline by running the same analysis with randomized class labels. The effects of such schemes on feature ranking are hard to treat (Simon, 2006): in practice, if the system includes a feature ranking method, a different list of best k biomarkers will be obtained at each resampling.

The task we address in this article consists in defining a measure of similarity among the output ranked lists and devising a family of computational tools for applications. This measure reflects indeed the stability of all the steps in the profiling process by gauging the variability of feature rank due to admissible perturbations of the sample data. We will consider some specific aspects of genome-wide data. In the microarrays data case, a large number of spots are typically not relevant for the classification problem, while many others are highly correlated or possibly clones. Furthermore, some genes can be grouped together in functionally related modules. The stability methods we present in this article are designed to include feature modules. Moreover, to focus on the most informative features, similarity among partial sublists (called top-k lists) will be modeled.

The problem of comparing ranked lists is actively studied in different contexts, e.g. in voting theory or in web document

*To whom correspondence should be addressed.

 © The Author 2007. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/24/2/258/226884 by University of Toronto user on 08 March 2019

retrieval, and it is emerging in bioinformatics. The methods employed to introduce similarity measures are heterogeneous. We mention harmonic and spectral analysis (Lawson *et al.*, 2006; Saari, 2001), rank aggregations (DeConde *et al.*, 2006; Dwork *et al.*, 2001; Fagin *et al.*, 2003), string theory (Cormode *et al.*, 2001; Gusfield, 1997) and set-theory (Lottaz *et al.*, 2006; Yang *et al.*, 2006). Not all methods are suitable for modeling stability of biomarker lists, either because of computational inefficiency on high-throughput data (e.g. the harmonic analysis methods) or because they are aimed at handling different data types (as in the case of string algorithms for sequence matching).

We base our proposal on the algebraic theory of symmetric groups and adopt the Canberra distance as a disarray measure (see Section 2). The embedding of a distance problem into the framework of permutation group theory was first introduced in Critchlow (1985) and then refined in Diaconis (1988). Once the distance is defined, we consider the symmetric matrix of all mutual distances between pairs of lists and derive stability indicators from its distribution.

We naturally derive the definition of a distance measure among partial lists from the one on global lists. The group theoretical framework for the partial lists (Critchlow, 1985; Fagin *et al.*, 2003) is then used to model a higher stability weight for top-ranked features in lists and to hatch the presence of irrelevant features. We then generalize the approach to reduce the effect of rank switches between functionally related genes. We test the new indicators on different datasets, synthetic, microarrays and proteomic mass spectra (see Sections 3–4).

The proposed indicators can be applied to other contexts in which ranked lists are commonly used, such as information retrieval systems or for enrichment analysis (see Supplementary Material). In Section 4 we test our indicators in combination with GSEA (Subramanian *et al.*, 2005), one of the most used application for gene-set enrichment analysis. Details on theory, software and additional examples are given in Supplementary Material.

2 SYSTEM AND METHODS

2.1 Notation

Let D be a dataset consisting of n samples described by a set $\mathcal{F}$ of p features indexed by integers from 1 to p: $\mathcal{F}=\{F_j\}_j=1\dots P$. In the microarray data case, features are gene expression values and p ranges between 5000 and 50 000, while in proteomics mass spectrometry the F_j describe intensities at m/z mass-to-charge ratios, typically with $50\leq p\leq 1000$ after preprocessing. With B, we indicate the number of replicated experiments required by complete validation of profiling. They consist of instances of classification and feature ranking: we also call them runs. At least $B=100$ runs are used in the experiments described in Section 4. At each replicate $i=1\dots B$, we assume that a ranking process sorts the features according to their importance in building the i-th classifier and an ordered list L_i is produced. Let $\mathcal{L}=\{L_i\}_{i=1}^{B}$ be the set of all lists. Let L_i^k be its top-k list, i.e. the sublist consisting of the first k ranked elements from L_i. Let $\tau_i(j)$ be the rank (position) of feature F_j in L_i: we call *dual list* of L_i the permutation $\tau_i=(\tau_i(j))_{j=1}^{p}$ (see Fig. 1). We consider S_p, the set of all the $p!$ dual lists obtained by ranking p elements, also known as the symmetric group on p objects. Based on these concepts, in this article we will use methods from permutation group theory, confining the mathematical details to Supplementary Material 1–5. We first introduce two indicators of stability for single features (see Subsection 2.2), which are used to derive a proposal for optimal list construction (see Subsection 2.3). The union number indicator is then defined (see Subsection 2.4). Finally, the Canberra distance (see Subsection 2.5) and a derived metric indicator (see Subsection 2.6) are presented.

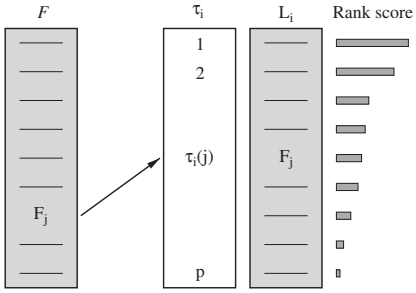

Fig. 1. The ranked list and its dual.

2.2 Extraction and position numbers

Within this framework, some basic information can be inferred by counting on $\mathcal{L}$. For each feature F_j, we define its top-k extraction set

$$E_k(j)=\{i\in\{1\dots B\}:\tau_i(j)\leq k\}\ ,$$

i.e. the indices of the runs ranking the feature in the top-k sublist. We can count how many top-k lists include a feature by defining the *extraction number* of F_j as the number of elements in the set $E_k(j)$

$$e_k(j)=\mathrm{Card}(E_k(j))$$

and estimate its (average) *position number*:

$$a_k(j)=\frac{1}{e_k(j)}\sum_{i\in E_k(j)}\tau_i(j).$$

For a given k, the numbers $e_k(j)$ and $a_k(j)$ induce a ranking of the features. We can use it as a first stability criterion: high e_k and low a_k indicate features extracted often in top positions, an element of stability for the study described by the set of lists $\mathcal{L}$.

2.3 The optimal list

We combine the definitions in Subsection 2.2 and propose a simple solution to encode the ranking information coming from all the lists in $\mathcal{L}$ into a single optimal list (possibly not in $\mathcal{L}$).

For each k, consider as the ranking criterion the extraction number $e_k(j)$ in decreasing order, and, when ties occur, the position number $a_k(j)$ in increasing order. The criterion defines a dual list τ_O^k, which correspond to a list of features called the *optimal top*-k *list* of $\mathcal{L}$. When $k=p$, the complete lists are considered and trivially it is $e_p(j)=B$ for all F_j; thus $\tau_O=\tau_O^p$ is determined only by the $a_p(j)$ average values.

The optimal list definition can be linked to the *Borda count*, an algorithm for optimally merging a set of lists that is well known in voting theory (Borda, 1781; Saari, 2001). Given a set of B ranked lists on p candidates, the Borda count associates to each candidate F_j a score $s(j)$ given by the total number of candidates with higher position over all lists. The Borda optimal list is then derived by ranking candidates with higher scores. In Supplementary Material 2, it is shown that $s(j)=B(p-a_p(j))$, i.e. ranking according to such decreasing score is equivalent to ranking for increasing average position.

2.4 The union number

For each k, a first natural measure of disarray for the set $\mathcal{L}$ is given by the number Σ_k of different features occurring in all the B top-k lists.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/24/2/258/226884 by University of Toronto user on 08 March 2019

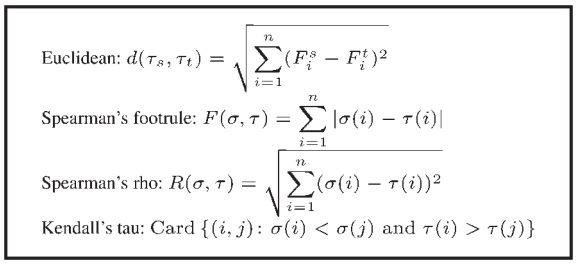

Fig. 2. Statistical distances.

This quantity, called the *union number*, can be formally defined as follows:

$$\Sigma_k(\mathcal{L}) = \mathrm{Card} \bigcup_{i=1}^{B} \{F_j \in L_i : \tau_i(j) \leq k,\ 1 \leq j \leq p\} .$$

The union number satisfies the constraints $k \leq \Sigma_k(\mathcal{L}) \leq \min\{p, kB\}$ for each k; obviously $\Sigma_p(\mathcal{L}) = p$. By listing the Σ_k we derive the *union number indicator* defined as the sequence

$$I_U(\mathcal{L}) = \{(k, \Sigma_k) : 1 \leq k \leq p\}. \quad (1)$$

The higher the values of Σ_k, the less stable is $\mathcal{L}$: the worst possible situation is given by the no-information curve $I_U = \{(k,p) : 1 \leq k \leq p\}$. Note that a similar idea was introduced in Yang *et al.* (2006), where (a function of) the number of common genes in the top-k and the bottom-k sublists of differential gene expression lists was considered to account for strongly up-regulated and bottom-regulated genes at the opposite ends of the lists. Here, the ranking is derived from classification and thus the most discriminating features are expected to stay at the top.

The above defined e_k, a_k and I_U provide a first estimate of variability in a set of lists; however, if a finer analysis is required, more complex methods need to be employed. Metrics are proposed in Subsection 2.5. Moreover, if intersection is used instead of union, an indicator keeping track of the number of genes common to all top-k sublists is defined for each k. Intersection is commonly used, but it is less informative than union for small values of k (consider for instance a set of all but one identical lists).

2.5 Canberra distance

Metrics between ranked lists require an additional condition called right-invariance (Critchlow, 1985) to make it independent from a relabeling (see Supplementary Material 1). Among the classic statistical distances (Fig. 2), the Euclidean distance is not right-invariant and thus not a metric for S_p. Admissible metrics on S_p are the Spearman's footrule, the Spearman's rho and the Kendall's tau. A variation of the Spearman's rank correlation measure was employed in Kalousis *et al.* (2005) to study list stability. As a second constraint specific to list comparison, we require that variations in lower portions of the lists should be less relevant than those in the top. A natural solution is to consider weighting factors. In particular, we develop our theory on a weighted version of Spearman's footrule, e.g. the Canberra distance:

$$\mathrm{Ca}(\tau, \sigma) = \sum_{i=1}^{p} \frac{|\tau(i) - \sigma(i)|}{\tau(i) + \sigma(i)}$$

Finally, since the most important features are usually located in the upper part of the ranked lists, we specialize our study on top-k lists. Managing partial lists is however much harder than in the complete ($k = p$) case because different subsets of elements are involved. Here we follow (Critchlow, 1985) and consider a Hausdorff metric. This framework is equivalent to the notion of distance with location parameter: for details see Supplementary Material 3. The definition of the Canberra distance with location parameter $k+1$ is then

$$\mathrm{Ca}^{(k+1)}(\tau, \sigma) = \sum_{i=1}^{p} \frac{|\min\{\tau(i), k+1\} - \min\{\sigma(i), k+1\}|}{\min\{\tau(i), k+1\} + \min\{\sigma(i), k+1\}} ,$$

with $\mathrm{Ca}^{(p+1)} = \mathrm{Ca}$.

We provide the expected (average) value of the Canberra metric on S_n in Theorem 1 (Supplementary Material 5). It can be approximated up to terms $o(1)$ as

$$\hat{E}\{\mathrm{Ca}^{(k+1)}\}_{S_p} = \frac{(k+1)(2p-k)}{p} \log(4) - \frac{2kp + 3p - k - k^2}{p} . \quad (2)$$

The approximation gives $\hat{E}\{\mathrm{Ca}\}_{S_p} = (\log(4) - 1)p + \log(4) - 2$ for complete lists. For top-k lists, up to terms converging to zero with p, the approximation $\hat{E}\{\mathrm{Ca}^{(k+1)}\}_{S_p}$ as a function of k and fixed p is a parabola. Note that the difference between the exact and approximated values is very small: for instance, for $p \approx 10^4$ it is about 10^{-3}. The result in Equation (2) is used to characterize a set of ranked lists in the maximal instability case.

2.6 List distance indicators

Given a set of lists $\mathcal{L} = \{L_t\}_{t=1}^{B}$ of p features and an integer $k \leq p$, the computation of all mutual top-k distances leads to the construction of a distance matrix $M_k \in \mathcal{M}(B \times B, \mathbb{R}^+)$; let μ_k be the mean of all its $\frac{B(B-1)}{2}$ non-trivial values:

$$\mu_k = \frac{2}{B(B-1)} \sum_{1 \leq i < j \leq B} (M_k)_{ij} . \quad (3)$$

Then the *mean list distance indicator* is the sequence

$$I_{D_\mu}(\mathcal{L}) = \{(k, \mu_k) : 1 \leq k \leq p\} . \quad (4)$$

We also define the *no-information curve*, i.e. the sequence

$$I_{D_\mu}(S_p) = \{(k, E\{\mathrm{Ca}^{(k+1)}\}_{S_p}) : 1 \leq k \leq p\} \quad (5)$$

associated to the group S_p of all possible dual lists with p features. An indication of the stability of a list set $\mathcal{L}$ is given by comparing $I_{D_\mu}(\mathcal{L})$ with $I_{D_\mu}(S_p)$, the latter representing the situation of maximal instability.

We remark that the indicators defined in this section are independent from the particular classifier and feature ranking algorithm, since only the set of lists is used as the input for the indicator, regardless of the ranking method. Finally, the pointwise ratios of I_U and I_{D_μ} with their no-information curves [$\Sigma_k = p$ and $I_{D_\mu}(S_p)$, respectively] are independent from p and k. The resulting normalized indicators $\hat{\mathrm{I}}_U$ and $\hat{I}_{D_\mu}$ can be then used to compare experiments on different datasets (see Supplementary Material 6 for details).

2.6.1 Feature modules It can be relevant to take into account the presence of mutually related groups of features (modules) and reduce the penalization in case of a rank variation within a module than other permutations. The interest for this additional constraint may be due to biological background or to technical reasons. In case of microarray data, it is often the case that highly correlated genes may be functionally related or just swapped for geometrical reasons (due to the nature of the employed classifier) during the ranking process. The constraint can be implemented by making the distance independent from the ordering of the features inside the module: the mathematical formulation is detailed in Supplementary Material 7. Defining one or more feature modules will take care of the intermodule instability when computing the stability indicator. Therefore, the result is always a curve describing a more stable situation than in the corresponding case without modules. Thus, the introduction of this constraint allows the computation of a more realistic value of the stability indicator, as evidenced in the experiments in Subsections 4.1 and 4.3.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/24/2/258/226884 by University of Toronto user on 08 March 2019

3 DATA DESCRIPTION

3.1 Synthetic data

A family of synthetic datasets fX/Y has been created to simulate a binary classification problem on N samples described by Y features, of which X are discriminant. The features are modeled from the Gaussian $\mathcal{N}(\mu,\sigma)$, with mean μ fixed according to the sample label for the discriminant features and sampled from the uniform $\mathcal{U}[a,b]$ for the remaining ones. The class of the i-th sample is 1 for $1 \leq i \leq \frac{N}{2}$ and -1 otherwise, while the value of its j-th feature is $fX/Y_{i,j} \in \mathcal{N}(m_{i,j}, \sigma_{i,j})$, where $m_{i,j}$ and $\sigma_{i,j}$ are defined as follows:

$$m_{i,j} = \begin{cases} \text{class}(i) & \text{if } 1 \leq j \leq X \leq Y \\ u \in \cup[-2,2] & \text{if } X < j \leq Y \end{cases}$$

$$\sigma_{i,j} \in [a,1] \text{ for } a = \begin{cases} \frac{1}{10} & \text{if } 1 \leq j \leq X \leq Y \\ 0 & \text{if } X < j \leq Y\ . \end{cases}$$

In the considered examples, we choose $N = 100$.

3.2 Proteomics ovarian cancer data

We considered a proteomic pattern dataset, produced at Keck Laboratory with a Micromass MALDI-L/R instrument on 77 controls and 93 ovarian cancer samples, as described and used in Wu *et al*. (2006). Here, we consider the linear analyzer data in the region of 3450–28000 Da; by preprocessing all spectra with the methods described in Barla *et al*. (2006), 123 peaks were identified and intensities at these peaks were used as features.

3.3 Microarray breast cancer data

A microarray dataset of samples from patients with invasive breast carcinoma (Sotiriou *et al*., 2006) was also considered for studying stability of gene signatures and of the corresponding enriched ranked lists. The platform is the Affymetrix U133A GeneChips; after preprocessing, the data matrix includes 183 cases described by 22 215 gene expression values. The profiling of Estrogen Receptor (ER) status is the task, given 149 ER+ and 34 ER− cases.

3.4 Supplementary data

In the Supplementary Material, two additional microarray datasets are used: the Cardiogenomics Mouse Model of Myocardial Infarction (MI: 37 samples, 12 488 genes) and the Huntington's Disease Data (HD: 31 samples, 5186 genes). The given task, on both datasets, is gene profiling for discriminating disease and control samples.

4 RESULTS

In this section, we compute the list distance indicators on one synthetic and four high-throughput real molecular profiling tasks described in Section 3 to demonstrate the behavior of the proposed indicators in a set of practical situations (predictive classification, presence of modules, comparison of alternative biomarker lists and gene enrichment). We also provide a few indications for an effective usage of the indicator curves as supporting tools for choices among classifiers and feature selection algorithms. Further experiments are shown in Supplementary Material 8–10.

Table 1. Synthetic data: feature rank (Number) in terms of the extraction number e_k and the position number a_k

$f10/100$						$f30/100$				
Number	Top-9			Complete		Top-10			Complete	
	Features	e_k	a_k	Feat.	a_k	Features	e_k	a_k	Features	a_k
1	G5	100	1.3	G5	1.3	G11	97	1.7	G11	2.1
2	G9	100	2.1	G9	2.1	G21	88	4.9	G10	5.8
3	G8	100	3.4	G8	3.4	G10	84	4.3	G21	6.0
4	G6	100	4.7	G6	4.7	G1	78	5.9	G1	7.6
5	G4	100	4.8	G4	4.8	G23	74	4.8	G23	8.0
6	G1	100	5.7	G1	5.7	G20	72	4.7	G20	8.6
7	G3	100	7.1	G3	7.1	G8	66	5.3	G8	8.8
8	G2	83	7.4	G2	7.5	G13	63	7.2	G13	9.7
9	G10	83	8.9	G10	8.9	G26	59	7.2	G26	9.7
10	G7	34	9.6	G7	9.6	G14	49	5.3	G12	11.1
11	G*n*	0	0	G90	18.8	G12	46	7.4	G14	12.1

Features are indicated by G*j*, with $1 \leq j \leq X$ ($X = 10$ or $X = 30$) if discriminant, or by the shortcut G*n* if non discriminant. For complete lists, $e_k = B$ constantly.

All experiments are carried out in a complete validation methodology aimed at avoiding the selection bias effect (Ambroise and McLachlan, 2002; Furlanello *et al*., 2003). In brief, B different training-test splits of the original dataset are prepared and, for each instance, a feature selection process is performed on the training part. The resulting ranked list is used to build classification models of different feature set sizes and accuracy is evaluated on the left-out set of test samples. Global accuracy on the whole dataset is obtained by averaging over the B runs. Different types of support vector machine (SVM) classifiers and of recursive feature elimination (RFE) procedures were used with metaparameters chosen by cross-validation or bootstrap.

4.1 Predictive profiling on synthetic data

In this experiment, we study the basic properties of the list indicators in the classification of datasets $f10/100$ and $f30/100$. In both tasks, we consider $B = 100$ runs, a linear SVM classifier with regularizer $C = 100$, and single-step RFE as the feature ranking algorithm. On both datasets less than five features are sufficient to reach perfect classification. In Table 1, features are ranked in terms of the extraction (e_k) and position (a_k) numbers (see Subsection 2.2).

We compare the results for the top-k sublist and the complete list in both tasks. All discriminative features of the $f10/100$ dataset are the best ranked on the top-9 sublist and on the complete lists; 7 out of 10 features occur in all the top-10 lists, i.e. they have $e_k = 100$. For the $f30/100$ dataset, the profiling process selects discriminant features at the top of the list both for the partial top-10 and for the complete case, with high e_k.

In general, a large number of features with high e_k indicate that many genes are consistently among the best ranked throughout the whole list set, thus forming a subset of well discriminating features for the employed classifier. When considering complete lists, the above consideration still holds for features with low values of a_k.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/24/2/258/226884 by University of Toronto user on 08 March 2019

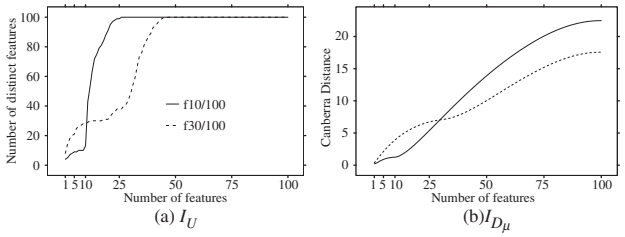

Fig. 3. Response of two stability indicators in profiling synthetic data (solid lines: $f10/100$; dotted lines $f30/100$). In (**a**): the union number indicator I_U; in (**b**): the mean Canberra distance indicator I_{D_μ}.

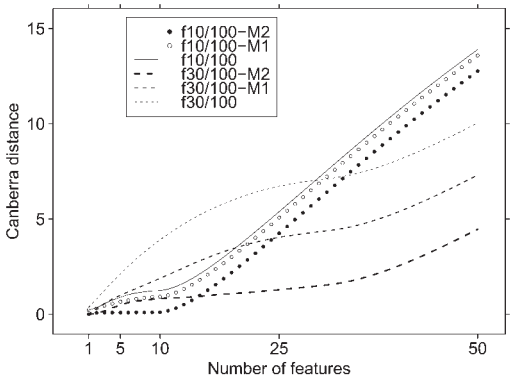

Fig. 4. Factoring out feature modules from the mean stability indicator. For the 50 top-k sublists, I_{D_μ} curves are compared to their corrected versions for the M1 and M2 modules on $f10/100$ and $f30/100$.

In Figure 3, the curves defined by the two indicator on $f10/100$ and $f30/100$ are displayed for increasing top-k list sizes. Both for I_U and I_{D_μ}, the curves show a sudden change of slope near the abscissa value X (the number of discriminant features in fX/Y). Note that the union number indicator attains the upper bound Y soon after the critical value, while the mean indicator does not saturate. The latter may thus be used for finer analyses, such as the comparison of lists produced by different classifiers. As shown in Figure 3b, for the complete lists ($p = 100$), the mean indicator values are $I_{D_\mu}(100) = 22$ for $f10/100$, and $I_{D_\mu}(100) = 18$ for $f30/100$; both values are much smaller than the $E_{S_{100}}\{\mathrm{Ca}\} \approx 38$, discussed in the consistency experiment Supplementary Material 8. Difference in stability with respect to the baseline case can be used to detect no-information cases and compare methods.

We therefore consider the response of the mean stability indicator I_{D_μ} in presence of feature modules (Fig. 4). With the modules M1 and M2, we group together features of the fX/Y synthetic datasets having absolute Pearson correlation greater than 0.8 (resp. 0.7). The resulting modules are $f10/100$-M1, $f10/100$-M2 $f30/100$-M1 and $f30/100$-M2 of 9, 12, 23 and 33 features, respectively. We note that the module-based correction has a greater effect for larger modules. Factoring out the modules from the distance computation can strongly affect the relation between sets of lists: in the displayed case, the difference between the stability of the sets of lists for the two datasets increases when the modules are taken into account.

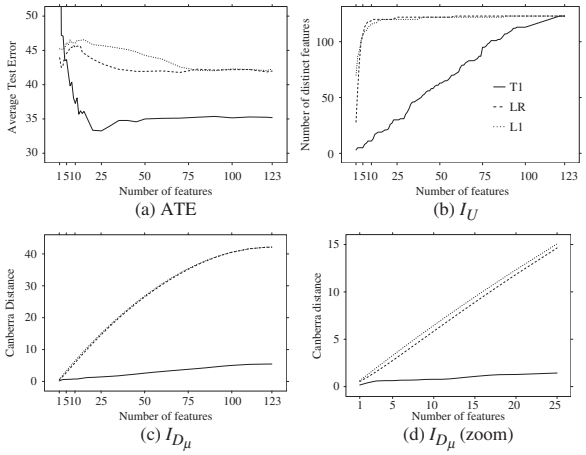

Fig. 5. Ovarian cancer data: (**a**) accuracy (ATE); (**b**) Union number I_U; (**c**) mean distance indicator I_{D_μ} for three profiling choices: LinearSVM/1RFE (L1: dotted line), LinearSVM/RFE (LR: dashed line) and TRSVM/1RFE (T1: solid line); (**d**) zoom up to 25 features on the leftmost part of (c).

4.2 High-throughput data applications

4.2.1 Comparing classifiers The stability indicators can be used to compare list sets produced by different profiling methods. Here, we consider the effect of classifiers and of ranking procedures on lists of features for the proteomics ovarian cancer dataset (see Subsection 3.2). A stability analysis is applied to an experiment introduced in Merler *et al.* (2007), where different SVM and RFE methods are studied. In particular, we compare linear SVM and the Terminated Ramp kernel (TR-SVM), coupled with RFE (ranking recomputed at each step) and 1RFE (ranking computed only once). Three profiling studies were performed in complete validation with $B = 400$ runs, with regularizer $C = 1000$ and $C = 1$ for the linear SVM and TR-SVM, respectively. The results are displayed in Figure 5. The average test error (ATE) computed by complete validation is not far from the no-information error rate (45.29%) for this problem. TR-SVM with 1RFE is more accurate, confirming that choosing a suitable kernel is a good strategy in hard classification tasks. The new list stability indicators are also computed. As a baseline for stability, from Equation (5) we consider the expected values for $E\{\mathrm{Ca}^{(k+1)}\}_{S_{123}} \approx 47$ for the complete list of 123 features.

Stability with TR-SVM is consistently better than with SVM, either with the 1RFE or the full RFE versions. Moreover, the I_U reaches its maximum value at a much slower rate with the more complex kernel. The analysis indicates that the TR-SVM method is providing a set of more accurate and stable candidate peaks in this profiling task.

4.2.2 Accuracy and stability Although they share a common trend in many cases, accuracy and stability are independent measures. In easy classification tasks, alternative methods can give negligible gains in accuracy but more stable lists. As an example, the Breast Cancer dataset in Subsection 3.3 is analyzed with a linear SVM ($C = 0.1$) with Entropy based RFE (E–RFE; reverting to RFE for the last 100 steps) as in Furlanello *et al.* (2003), and with a TR-SVM ($C = 0.01$)

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/24/2/258/226884 by University of Toronto user on 08 March 2019

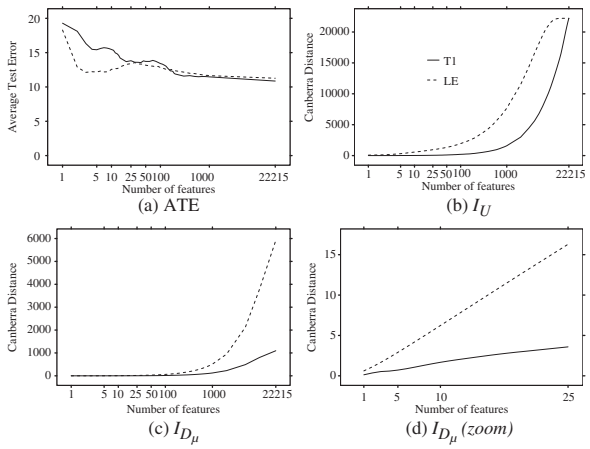

Fig. 6. Breast cancer data: comparison of LinearSVM/E–RFE (LE: dashed lines) and TRSVM/1RFE (T1: solid lines) for (**a**) average test error (ATE); (**b**) union number (I_U); (**c**) mean distance indicator (I_{D_μ}) with zoom up to 25 features in (**d**).

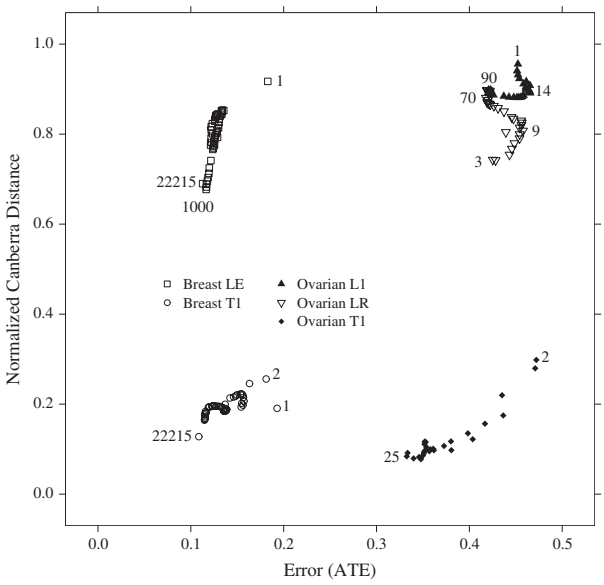

Fig. 7. Plot of accuracy (ATE) versus normalized stability ($\hat{I}_{D_\mu}$) for different profiling methods and cancer datasets; each point corresponds to a feature subset size (indicated for the extremal models).

with 1RFE. With both methods, the mean indicator is distant from the no-information curve $I_{D_\mu}(S_{22215})$ defined by Equation (5), whose value on the complete list is about 8581.

As shown by Figure 6, the resulting ATE curves are quite similar for the two methods but the improvement obtained by the TR-SVM classifier in terms of list stability is relevant even for small k.

4.2.3 A diagnostic plot We compare simultaneously the five experiments above in terms of accuracy and normalized stability in the ATE versus $\hat{I}_{D_\mu}$ plot (Fig. 7). Each point represents the performances of a model built with top-k features. This graph allows the comparison of different datasets, different profiling methods (classifiers/feature ranking algorithm) and different models, resulting in a new effective tool based on the stability indicator. In general, the lower the point is in the graph, the higher its stability and, similarly, the closer the point to the left border the higher its accuracy. Observing Figure 7, the first consideration that can be done is that the Breast Cancer dataset with TR-SVM as the classifier and 1-RFE as the feature ranking algorithm (Breast T1 for short) is in average the best performing combination. In fact Breast T1 points have better stability than Breast LE and Ovarian L1 and LR. They are comparable to the Ovarian T1 but they have a much better accuracy. Moreover, the Breast T1 points are mutually closer than those of the other combinations. This means that models for Breast T1 at different feature set sizes are more similar than in the other cases; for instance note the wide accuracy difference between the models with 2 and 25 features for the Ovarian T1 dataset, and the difference in stability between the models with 1 and 1000 features for the Breast LE dataset. Combinations Ovarian L1 and LR also have quite mutually closer points, but all of them have both (relative) poor stability and accuracy.

The definition of $\hat{I}_{D_\mu}$ is given in Supplementary Material 6. As further examples of procedures of practical interest in gene profiling studies, the list stability approach is used to analyze data shaving (outlier analysis and removal) in Supplementary Material 9 and label randomization in Supplementary Material 10.

4.3 Applications to enrichment

For enrichment analysis, we considered the GSEA algorithm, using the Normalized Enrichment Scores (NES) as the ranking value. We applied the stability indicator to enriched ranked lists of gene sets as computed with GSEA from ranked lists of genes. To prove the consistency of the procedure, we randomly extracted a set G of n_g genes included in the gene sets MSigDB collection, and constructed the set $\mathcal{L}$ of l random permutations L_i of G. Then, we tested $\mathcal{L}$ for enrichment against a set G_S of n_{g_s} gene sets from those in the entire collection in MSigDB including at least one element of G. Thus, we obtained a set $\mathcal{G}_S$ of l ranked gene sets. For this experiment, we set $n_g = 1000$, $l = 100$ and $n_{g_s} = 100$.

As shown in Figure 8a, both the obtained ranked lists and the enriched lists are highly unstable ($I_{D_\mu} > 0.8$), also after factoring out the gene sets as feature modules and with different weighting parameters p as in Subramanian *et al.* (2005) (correlation value for the gene ranked j-th fixed to $\frac{1}{j}$). Then we constrained the admissible permutations of genes so that positions s,t in L_i satisfies $|s-t| \leq z$, and compared to the resulting sets of ranked enriched lists ($z = 20$ and same parameters as above). The constrain produces more stable gene lists, and consistently stable enriched lists (Fig. 8b).

Finally, in the same setup, we considered gene set invariant permutations, i.e. those exchanging genes within the same gene set. Here, we used $n_g = 100$ and the $n_{g_s} = 124$ gene sets sharing at least two gene with the ranked gene lists. Results are shown in Figure 9. While the permuted gene lists are unstable ($I_{D_\mu} > 0.6$), the module correction detects the structure at gene level but still shows limited stability. On the contrary, the enriched lists are very stable for both values of p. In this case the stability analysis shows that apparently diverse ranked gene lists actually share a very similar pathway profile, thus confirming the enrichment information.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/24/2/258/226884 by University of Toronto user on 08 March 2019

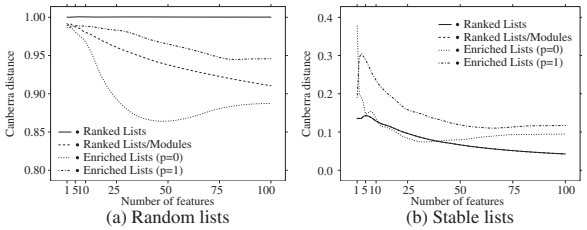

Fig. 8. Stability curves for ranked gene lists (solid line), ranked gene lists with the gene sets modules factored out (dashed line) and the two ranked lists of enrichment gene sets (dashed and dashed-dotted line). (**a**) Totally random permutations; (**b**) position-constrained permutations.

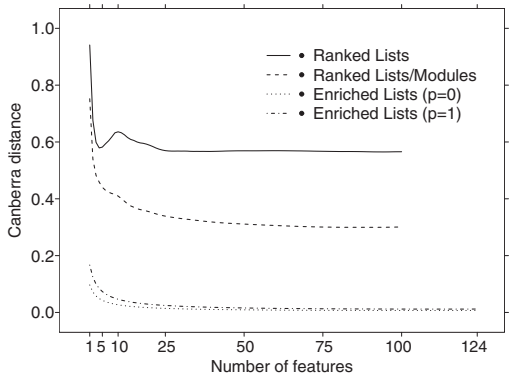

Fig. 9. Stability curves for the ranked gene lists (solid line), the ranked gene lists with the gene sets modules factored out (dashed line) and the two ranked lists of enrichment gene sets (dashed and dashed-dotted line).

In Supplementary Material 12, the stability analysis is applied to enrichment gene sets derived from the Breast Cancer dataset (see Section 3.3).

5 CONCLUSIONS

We introduced a new framework of algebraic indicators to assess the stability of biomarker lists. The more refined indicators come from an adapted metric theory on permutation groups, and they can be used in diverse high-throughput studies. The indicators allow comparisons of list sets produced by different classification/ranking systems, also on different datasets, thus adding stability as a measure of interest for genomic signatures in addition to accuracy.

Results on synthetic data consistently explain test cases, and those on high-throughput data show applicability in current practice. An open source software implementation of the stability indicator is made available and it has adequate efficiency for practical tasks: in fact, computing time on a PentiumD workstation for the analysis of about 80 values of the indicator for 400 lists of 20 000 features (we recall that the algorithm is independent from sample size) amounts to ~1 h. A summary of computing performance for different list sizes can be found in Table 6 in Supplementary Material 13.

The approach can further accommodate bioinformatic knowledge on feature modules. Known relations among genes (e.g. function groups as defined in GO or KEGG or explicit correlation effects) can be injected in the algorithm by suitably defining one or more feature modules. In such cases, the rank changes within one module do not contribute to the stability indicator. This approach can be generalized by introducing weighting schemas for module-specific costs of rank change.

ACKNOWLEDGEMENTS

Part of the method was studied in two B.Sc. theses by S. Maragnoli and A. Peretti. We thank A. Caranti for hints on the Borda count and the harmonic analysis methods, C. Joachim, E. Manduchi and C. Stoeckert for reading earlier versions of the article and providing useful suggestions. Research supported by AIRC.

Conflict of Interest: none declared.

REFERENCES

Ambroise,C. and McLachlan,G. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA*, **99**, 6562–6566.

Baker,S. and Kramer,B. (2006) Identifying genes that contribute most to good classification in microarrays. *BMC Bioinformatics*, **7**, 407.

Barla,A. *et al*. (2006) Proteome profiling without selection bias. In *Proceedings of IEEE-CBMS 2006*, pp. 941–946.

Borda,J. (1781) Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, Année MDCCLXXXI.

Cormode,G. *et al*. (2001) Permutation editing and matching via embeddings. In *Proceedings of ICALP 01*. Springer, pp. 481–492.

Critchlow,D. (1985) *Metric Methods for Analyzing Partially Ranked Data. LNS 34*. Springer.

Davis,C. *et al*. (2006) Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*, **22**, 2356–2363.

DeConde,R. *et al*. (2006) Combining results of microarray experiments: a rank aggregation approach. *Stat. Appl. Genet. Mol. Biol.*, **5**. Article 15.

Diaconis,P. (1988) *Group representations in probability and statistics*. Institute of Mathematical Statistics Lecture Notes – Monograph Series 11, *IMS*.

Dwork,C. *et al*. (2001) Rank aggregation for the web. In *Proceedings of IWWWCC-WWW10*, pp. 613–622.

Fagin,R. *et al*. (2003) Comparing top-*k* lists. *SIAM J. Discrete Math*., **17**, 134–160.

Furlanello,C. *et al*. (2003) Entropy-based gene ranking without selection bias for the predictive classification of microarray data. *BMC Bioinformatics*, **4**, 54.

Gusfield,D. (1997) *Algorithms on Strings, Trees and Sequences*. CUP.

Kalousis,A. *et al*. (2005) Stability of feature selection algorithms. In *Proceedings of IEEE-ICDM 05*, pp. 218–225.

Lawson,B. *et al*. (2006) Spectral analysis of the supreme court. *Math. Mag*., **79**, 340–346.

Lottaz,C. *et al*. (2006) OrderedList – a Bioconductor package for detecting similarity in ordered gene lists. *Bioinformatics*, **22**, 2315–2316.

Marchiori,E. (2006) Feature selection, SVM-based classification and application to mass spectrometry data analysis. *Bioinformatics Data Analysis and Tools. Lecture Notes*.

Merler,S. *et al*. (2007) Deriving the kernel from training data. *Proceedings MCS 2007, LNCS*, Springer, **4472**, 72–81.

Molinaro,A. *et al*. (2005) Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, **21**, 3301–3307.

Saari,D. (2001) *Chaotic Elections! A Mathematician Looks at Voting*. AMS.

Simon,R. (2006) Development and evaluation of therapeutically relevant predictive classifiers using gene expression profiling. *J. Natl Cancer Inst*., **98**, 1169–1171.

Sotiriou,C. *et al*. (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl Cancer Inst*., **98**, 262–272.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Wichmann,S. and Kamholz,D. (2006) A stability metric for typological features. *Sprachtypologie und Universalienforschung*, in press.

Wu,B. *et al*. (2006) Ovarian cancer classification based on mass spectrometry analysis of sera. *Cancer Inform*., **2**, 123–132.

Yang,X. *et al*. (2006) Similarities of ordered gene lists. *J. Bioinform. Comput. Biol*., **4**, 693–708.

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/24/2/258/226884 by University of Toronto user on 08 March 2019