Explicit Control of Feature Relevance and Selection Stability Through Pareto Optimality

Victor Hamer[1*] and Pierre Dupont[2]

UCLouvain - ICTEAM/INGI/Machine Learning Group, Place Sainte-Barbe 2,
B-1348 Louvain-la-Neuve, Belgium.
[1]victor.hamer@uclouvain.be [2]pierre.dupont@uclouvain.be

Abstract. Feature selection is an important issue when one deals with large amounts of *omics* data. Feature selection offers interpretability of the predictive model to the domain expert. Such interpretability is strongly affected by the typical instability of current feature selection methods. Instability here refers to the fact that the selected features may be drastically different even after marginal modification of the data. In this paper, we investigate the possibility of a tradeoff between the classification performance and the stability of a standard feature selection method: the Recursive Feature Elimination algorithm (RFE). The compromise is done by explicitly favoring the selection of some features through differential shrinkage. Such an approach allows the domain expert to control **explicitly** the trade-off between selection stability and predictive accuracy. Domain experts can thus select particular Pareto-optimal compromises, based on their personal preferences. As a secondary contribution, we propose the use of the hypervolume metric to assess the performance of methods realizing such a compromise and we define a corresponding confidence interval. Our approach is evaluated on prostate cancer diagnosis from microarray data and handwritten digit recognition tasks. Results show that the aforementioned tradeoff is effectively possible and that prior knowledge is an efficient way of stabilizing the selection.

Keywords: Feature selection · Selection stability · Classification performance · Transfer learning · Bi-objective optimization · Multitask learning

1 Introduction

Feature selection, *i.e.* the selection of a small subset of informative and relevant features to be included in a predictive model, has become compulsory for a wide variety of applications due to the appearance of very high dimensional datasets, notably in the biomedical data domain [20]. Filtering noisy and irrelevant features can avoid overfitting the data and potentially improve predictive performance. Feature selection also allows for the learning of fast and compact

* Corresponding author

models, which are easier to interpret. Such models can then be analyzed by domain experts and are easier to validate. Getting more interpretable models is also a key concern nowadays and even considered by many as a requirement when deployed in the medical domain.

Feature selection has been already largely studied. Yet, current methods are still widely unsatisfactory mainly because of the typical instability they exhibit. Instability here refers to the fact that the selected features may be drastically different for similar data, even though the true underlying processes (explaining the target variable) are essentially constant. Such instability is a key issue as it reduces the interpretability of the predictive models as well as the trust of domain experts towards the selected feature subsets. We address this problem here by designing methods balancing between the classification performance and the selection stability of the well-known Recursive Feature Elimination (RFE) algorithm. Our approach allows domain experts to explicitly control the trade-off and to select Pareto-optimal compromises based on their personal preferences.

In the rest of this section, two distinct stability problems that are tackled in this paper are introduced.

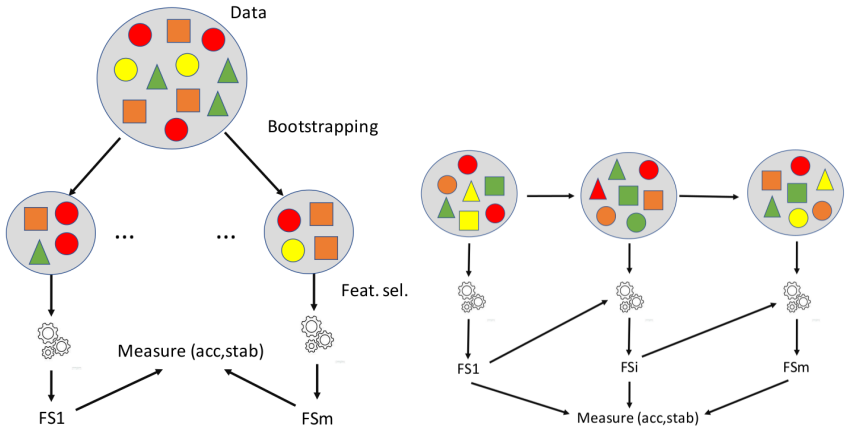
1.1 The Stability Problems

Single Task Stability (1) Feature selection methods are often inherently unstable, *i.e.* they return highly different feature sets when the training data is slightly modified. Figure 1a illustrates such an instability. The initial dataset is perturbed[1] to form different datasets. Instability arises when little overlap of the selected features occurs. This prevents a correct and sound interpretation of the selected features and strongly impacts their further validation by domain experts. Unlike optimizing the accuracy of predictive models, optimizing selection stability may look trivial since an algorithm always returning an arbitrary but fixed set of features would be stable by design. Yet, such an algorithm is not expected to select informative and predictive features. This illustrates that optimizing stability is only well-posed jointly with predictive accuracy, and possibly additional criteria such as minimal model size or sparsity.

Transfer Learning Selection Stability (2) Multi-task feature selection aims at discovering variables that are relevant for several similar, yet distinct, classification tasks. Different feature subsets can be returned for each task. In this paper, we focus on the case where all learning tasks are not directly available. Information from the tasks arising first can be propagated to subsequent tasks, via *transfer learning*. Stability has to be encouraged from the domain expert point of view as features that are relevant for different data sources are likely to be particularly interesting to study. The accuracy-stability trade-off on such a learning problem (represented in Figure 1b) can take two extreme values. With complete disregard to stability, each feature set could be selected on a given task

[1] Here by bootstrapping which is often used to measure such instability, but it could be any small perturbation.

independently of the others, with no control on the across task stability. On the contrary, maximum stability can trivially be reached by returning the feature set computed for the first task, for all subsequent tasks. However, this is expected to reduce the accuracy of the models built on these subsequent tasks as previously learned features might turn out to be less informative for them. This would be the case if the different tasks are obtained by gradually enriching or correcting the data as features learned on the error-corrected data are expected to be more relevant.



(a) Single-task **(b)** Transfer learning

Fig. 1. Illustration of two stability problems. For both problems, the outcome is a measure of the trade-off between prediction accuracy and selection stability. Methods allowing domain experts to control this trade-off are proposed in the subsequent parts of this paper.

In section 2, feature selection methods and propositions to increase stability are reviewed. Section 3 introduces a metric to assess the performance of methods compromising between feature selection stability and classification performance. Then a biased variant of the RFE algorithm is proposed in section 4. Section 5 demonstrates the ability of this biased RFE to tackle the previously mentioned stability problems.

2 Related Work

Feature selection techniques are generally split into three categories: filters, wrappers and embedded methods. *Filters* evaluate the relevance of features independently of the final model, most commonly a classifier, and remove low ranked features. Simple filters (*e.g.* t-test or ANOVA) are univariate, which is computationally efficient and tends to produce a relatively stable selection but they

plainly ignore the possible dependencies between various features. Information theoretic methods, such as MRMR [7] and many others, are based on mutual information between features or with the response, but a robust estimation of these quantities in high dimensional spaces remains difficult. *Wrappers* look for the feature subset that will yield the best predictive performance on a validation set. They are classifier dependent and very often multivariate. However, they can be very computationally intensive and an optimal feature subset can rarely be found. *Embedded methods* select features by determining which features are more important in the decisions of a predictive model. Prominent examples include SVM-RFE [10] and logistic regression with a LASSO [24] or Elastic Net penalty [30]. These methods tend to be more computationally demanding than filters but they integrate into a single procedure the feature selection and the estimation of a predictive model. Yet, they also tend to produce much less stable models.

Some works specifically study the causes of selection instability. Results show that it is mostly caused by the small sample/feature ratio [2], noise in the data or imbalanced target variable [5] and feature redundancy [23]. While all of these reasons clearly play a role, the first one is likely the most important one in a biomedical domain with typically several thousands, if not millions, of features for only a few dozens or hundreds of samples. This is likely why stable feature selection is intrinsically hard in this domain and why existing techniques are still largely unsatisfactory.

Looking for a stable feature selection also requires a proper way to quantify stability itself and lots of measures have already been proposed: the Kuncheva index [15], the Jaccard index [14], the POG [21] and nPOG [26] indices among others. Under such a profusion of different measures, it becomes difficult to justify the choice of a particular index and even more to compare results of works based on different metrics. Furthermore, the large number of available measures can lead to publication bias (researchers may select the index that makes their algorithm look the most stable) [6]. In the hope of fixing this issue, a recent work [17] lists and analyzes 15 different stability measures. They are compared based on the satisfaction of 5 different properties that a stability measure should comply. A novel and unifying index has been proposed in this regard. This index, used throughout this paper, measures the stability across M selected subsets of features. It can be computed according to equation (1).

$$\phi = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} s_f^2}{\frac{k}{d} * (1 - \frac{k}{d})} \tag{1}$$

with $s_f^2 = \frac{M}{M-1}\hat{p}_f(1-\hat{p}_f)$ the estimator of the variance of the selection of the f_{th} feature over the M selected subsets and k the mean number of features selected from the original d features.[2] This measure is the only existing measure satisfying the 5 (good) properties described in [17], namely *fully defined*, *strict monotonicity*, *bounds*, *maximum stability* $\Leftrightarrow$ *deterministic selection* and *correc-*

[2] $\hat{p}_f$ is the fraction of times feature f has been selected among the M subsets.

tion for chance. It is formally bounded by -1 and 1 but is asymptotically lower bounded by 0 as $M \to \infty$. It is also equivalent to the Kuncheva Index (KI)[15] when the number of selected features k is constant across the M selected subsets but can be computed in $O(M * d)$ time, whereas KI can only be computed in $O(M^2 * d)$.

Several authors already proposed different methods to increase stability. For instance, instance-weighting for variance reduction [11] which tends to increase feature stability while keeping a comparable predictive performance. Ensemble methods for feature selection have also been proposed [1] and generally increase feature stability. Nonetheless, the gain in stability offered by existing methods is still limited and, maybe more importantly, the stability of the selection cannot be controlled explicitly, which is the main goal of this paper.

Multi-task feature selection has already been largely studied [27]. Encouraging the selection of common predictors across tasks can be done by using the ℓ_1/ℓ_p regularization scheme. The cost of selecting different predictors for different tasks can be controlled by using different norms ℓ_1/ℓ_p, as $p \to \infty$ favors the selection of common features. As with the differential shrinkage approach proposed here, penalties caused by selecting several times the same feature are reduced. Notably, the ℓ_1/ℓ_∞ [16] and ℓ_1/ℓ_2 [18,19] penalties have been studied in details. Efficient projected gradient algorithms, for general p, are proposed and the effect of p on the shared sparsity pattern and the classification performance is analyzed [25]. The main goal of [25] is to find adequate feature-sharing degrees such as to maximize the prediction performance of the models, which is different from the objective of explicit control of the accuracy-stability trade-off that is pursued in the present paper. Although this approach has been originally introduced for standard multi-task feature selection, it can trivially be adapted to the transfer learning setting [25]. Other similar approaches have also been proposed [3,4,8] (see [27] for a complete survey).

3 A Multi-Objective Evaluation Framework Through Pareto Optimality

In this section, we propose to use a classical evaluation framework in multi-objective optimization to assess the efficiency of methods balancing between classification performance and selection stability. An (accuracy,stability) pair[3] (a_1, ϕ_1) dominates another pair (a_2, ϕ_2) iff $a_1 \geq a_2 \wedge \phi_1 \geq \phi_2$ and at least one of the inequalities is strict ($>$). A given method m is able to generate some pairs P_m in the space of all possible pairs[4] $P = \{(a, \phi) : 0 \leq a, \phi \leq 1\}$. From the set of generated pairs P_m, the set of pairs that are not dominated by any other pair,

[3] Common alternatives to the classification accuracy, such as specificity/sensitivity or AUC, can also be used.

[4] The careful reader may remember that the stability measure ϕ formally lies in the $[-1, 1]$ interval. However, as $\phi = 0$ corresponds to the stability of a uniformly random selection, we argue that the only interesting part of the stability spectrum is in fact $[0, 1]$.

Pa_m, can be found. This set, called the Pareto set, defines a subspace where no point dominates any other point. A domain expert would then choose his favorite pair based on his personal preference towards classification performance and feature selection stability.

As performance metric, we propose the widely used hypervolume measure [29], also known as S-metric. This volume represents the space containing the sets of accuracy-stability pairs that are dominated by at least one point of the Pareto set Pa_m. The hypervolume measure has the convenient property that whenever a Pareto set dominates another, the hypervolume of the former is greater. As our objective space is bidimensional, the hypervolume measure is referred to as the Dominance Area (DA) in the rest of this work.

An example of the DA metric can be seen in Figure 2. The blue method starts from the left with a higher accuracy. It thus gains some area over the red method. Nonetheless, the red method can reach higher stabilities without dropping the accuracy as much as the blue one. Overall, the red method has a larger DA. Note that this DA is also equal to the fraction of the total area that is dominated by the method, or 1 minus the fraction of area that dominates the method. Its value thus lies in the $[0, 1]$ interval.

As noticed by [28], this DA measure is biased towards convex, inner parts of the objective space. [28] tackles this problem by giving different weights to different portions of the objective space. This weighted DA can be computed via the weighted integral

$$DA_P = \int_0^1 \int_0^1 w(a, \phi) f_P(a, \phi) da d\phi \tag{2}$$

with w the weighting function and f_P the attainment function which is equal to 1 if (a, ϕ) is dominated and 0 otherwise. To preserve the $[0 - 1]$ bounds, w has to be normalized such that its integral over the objective space is 1. For example, the normalized weighting function $w_a(a, \phi) \triangleq \frac{e^{A*a}}{e^{A}-1}$ gives a higher weight to the portions of the space where the accuracy is high. In the example of Figure 2, the blue method actually outperforms the red one for $A > 2.5$. For the sake of generality, our methods are evaluated with $w(a, \phi) = 1$ but the proposed evaluation framework allows for more, for instance, if domain experts are particularly interested in some parts of the objective space.

In order to evaluate the pair $(a, \phi) \in P$ corresponding, for instance, to a set of meta-parameters, some data has to be used to learn the features and some data to evaluate them on independent examples. This can be done via standard cross-validation or by bootstrapping. Each set of meta-parameters produces different pairs in P and their average value is reported. Concretely, each point in Figure 2 comes with an uncertainty linked to the sampling of the data. In the following, we define a confidence interval on the true value of DA based on the derivation of confidence *regions* for each Pareto-optimal pair.

Let A be the random variable representing the accuracy value measured on a given subsampling of the data and Φ be the corresponding stability value. Let $\mathcal{P} = (A, \Phi)$ be the multivariate random variable with the accuracy and

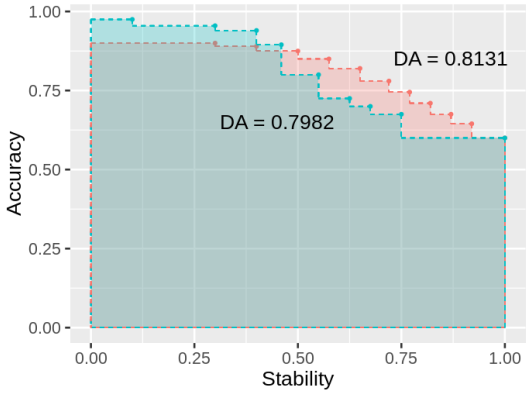


Fig. 2. Dominace Area (DA) toy example. The DA metric represents the area that is dominated by at least one point generated by the method.

stability as dimensions. Let us assume that the evaluation protocol produces B measurements of $\mathcal{P}$ for each Pareto-optimal point[5], represented by the vector $\mathbf{p}$. The Hotelling distribution T^2 is the multivariate counterpart of the Student's t distribution, with which we can define confidence (here 2-dimensional) regions.

$$T^2 = B(\bar{\mathbf{p}} - \mu(\mathbf{p}))' C^{-1} (\bar{\mathbf{p}} - \mu(\mathbf{p})) \sim \frac{2(B-1)}{B-2} * \mathcal{F}_{2,B-2} \tag{3}$$

with C the sample covariance matrix. It can be shown that T^2 is distributed like a Fisher distribution $\mathcal{F}_{2,B-2}$. Thus,

$$P\left[(\bar{\mathbf{p}} - \mu(\mathbf{p}))' C^{-1} (\bar{\mathbf{p}} - \mu(\mathbf{p}) \leq \frac{2 * (B-1)}{B-2} * \mathcal{F}_{2,B-2}(\alpha)\right] = 1 - \alpha \tag{4}$$

The inequality defines an ellipsoidal region, that is likely to cover $\mu(\mathbf{p})$. The center of the ellipsoid is $\bar{\mathbf{p}}$. The length of the axis and their angle can be found by computing the eigenvalues and eigenvectors of the sample covariance matrix C. To compute a confidence interval on the DA, the most dominant and dominated point of each ellipse are found and used to compute the upper and lower bound of the confidence interval (see Figure 3c and 3d for concrete examples in our experiments).

4 A Biased Variant of the RFE Algorithm

In this section, we propose a simple method to balance between the classification performance and selection stability of a logistic RFE algorithm. The RFE

[5] B could be *e.g.* the number of bootstrap samples.

algorithm was originally introduced with a hinge loss. We prefer here the logistic variant for an expected smoother control of the trade-off under study. RFE iteratively drops the least relevant features until the desired number of features k is reached. We opt here to drop a fixed fraction (20%) of the features at each iteration. The loss function that a logistic RFE minimizes for a binary classification task is the following, with n the number of samples, $\mathbf{x_i}$ sample number i made of d features as dimensions, and y_i its label.

$$L = \sum_{i=1}^{n} \log(1 + e^{-y_i * (\mathbf{w} * \mathbf{x_i})}) + \lambda ||\mathbf{w}||_2 \tag{5}$$

The weight vector $\mathbf{w}$ contains a weight assigned to each feature. The features are then ranked based on the absolute value of their weight, which represents the importance of the feature in the final prediction. The term $\lambda||\mathbf{w}||_2$ of the loss function is a regularization term, preventing coefficients of the model to take too high values, which would most likely result in overfitting. In the classical approach, every feature is regularized by the same amount λ.

We propose to extend equation (5), such that the regularization term becomes $\lambda\boldsymbol{\beta}||\mathbf{w}||_2$. The function of the vector $\boldsymbol{\beta}$ is to bias the selection towards certain features via differential shrinkage. A feature f with a small β_f is less regularized and vice-versa. Its selection in the model is less penalized than a feature with a higher β_f. The search is thus *biased* towards features with small β_f. A similar differential shrinkage has already been applied to the ℓ_1-AROM and ℓ_2-AROM methods [12,13] with the objective of biasing the selection towards *a priori* relevant features or in a transfer learning context. In the remaining part of this section, three possible schemes to set the $\boldsymbol{\beta}$ vector, according to the setting of interest, are discussed.

Biased RFE for Single Task Feature Selection By varying the distribution of $\boldsymbol{\beta}$, the accuracy-stability trade-off of the biased RFE can be controlled. The biased RFE is equivalent to a standard RFE when $\boldsymbol{\beta} = \mathbf{1}$. Otherwise, the selection is biased towards features with a small β_f. This is expected to increase stability at the possible cost of some classification performance, as uninformative features could be prioritized. In this initial approach, we decide to favor some features non-uniformly at random, following a gamma distribution.

$$\beta_f \sim \Gamma(\alpha, 1)$$

with α the shape of the gamma distribution, which controls the trade-off. All β_f are post centered such that $\mu(\boldsymbol{\beta}) = 1$. As $\alpha \to \infty$, the gamma distribution tends to a Dirac delta, $\delta(\alpha)$. All features have then the same weight (equal to $\mu(\boldsymbol{\beta}) = 1$) and no bias is put in the selection. As $\alpha \to 0$, the distribution of $\boldsymbol{\beta}$ departs from $\delta(\alpha)$ which increases the bias. Domains experts can thus play with the α values and therefore explicitly tune the trade-off between selection stability and prediction accuracy.

Using Prior Knowledge The biased RFE can take advantage of available prior knowledge. If a ranking of the features is known, then the β_f can be assigned such that this ranking is respected. If the prior knowledge is meaningfull, the selection is no longer biased towards arbitrary features, but towards features that are high in the ranking, and thus potentially informative. Another type of prior knowledge could be an unordered set of features that are suspected to take part in the process of interest. The lowest β_f could then systematically be assigned to those features.

Biased RFE for Transfer Learning We are now interested in the across task stability that can be obtained via transfer learning. Tasks are thus ranked such that information from previous tasks can be used in the selection of features for subsequent tasks.[6] In task i, features that have been returned for tasks $0..i-1$ should be prioritized over the rest, such that the feature stability is increased. Given the definition of stability used here (equation (1)), it is actually possible to compute the drop/gain in stability that the selection of a feature would cause. Intuitively, we propose to bias the selection, through a specific choice of the vector $\boldsymbol{\beta}$, towards features that would cause the highest gain/lowest drop in stability if they were to be selected. Constant terms put aside[7], each feature influences (negatively) the total stability by its variance in the selection $s_f^2 \propto p_f(1-p_f)$. Feature f is given an attractiveness score sc_f, expressed in equation (6).

$$sc_f = \frac{(N+1)^2}{N} * (p_{f,no}(1-p_{f,no}) - p_{f,yes}(1-p_{f,yes})) \tag{6}$$

with $s^2_{f,no}$ the selection variance of feature f assuming f is not selected in the current task and $s^2_{f,yes}$ its selection variance if it were to be selected. N is the number of tasks for which feature sets have already been selected. This score is thus proportional to the difference of stability between the cases where the feature is selected for a given task and not. This is illustrated in table 1a where the current task is $T4$. For instance, the selection of the feature $F2$ in task $T4$ would make its mean selection, p_f, equal to 0.75. If it were not to be selected, p_f would be equal to 0.5. The attractiveness score of $F2$, sc_{F2} is actually positive, meaning that the selection of $F2$ in $T4$ would increase the measured stability.

The $\frac{(N+1)^2}{N}$ factor of equation (6) is there to correct a downards tendency of sc_f when the index of the considered task increases. This is illustrated on table 1b. If feature f is selected in each task, sc_f would actually decrease which would decrease the bias. It can be shown that including the correction term leads to $sc_f = 2 * p_f - 1$ with p_f the proportion of the selections of feature f in the past N tasks. Let S be the sum of the such selections. By definition, $p_f = \frac{S}{N}$,

[6] Tasks can be ranked naturally from their chronological order or by the domain expert.

[7] We purposely drop the $M/(M-1)$ term, for convenience. Also, the denominator $\frac{k}{d}(1-\frac{k}{d})$ is constant if the number k of selected features is fixed.

Table 1. Illustration of the attractiveness score (a). Need for the correction term of equation (6)(b).

(a)

	F1	F2	F3	F4	F5
T1	0	0	1	1	0
T2	0	1	1	0	1
T3	0	1	1	1	0
T4	?	?	?	?	?
$p_{f,no}$	0	0.5	0.75	0.5	0.25
$p_{f,yes}$	0.25	0.75	1	0.75	0.5
sc_f	-1	1/3	1	1/3	-1/3

(b)

	T1	T2	T3	T4	T5
f	1	1	1	1	
$s^2_{f,no}$		1/4	2/9	3/16	4/25
$s^2_{f,yes}$		0	0	0	0

$s^2_{f,no} = \frac{S}{N+1} * (1 - \frac{S}{N+1})$ and $s^2_{f,yes} = \frac{S+1}{N+1} * (1 - \frac{S+1}{N+1})$. Thus,

$$sc_f = \frac{(N+1)^2}{N} * \left(\frac{S}{N+1} - \frac{S^2}{(N+1)^2} - \frac{S+1}{N+1} + \frac{(S+1)^2}{(N+1)^2} \right) =$$

$$\frac{(N+1)^2}{N} * \left(\frac{2S+1}{(N+1)^2} - \frac{1}{N+1} \right) = \frac{(N+1)^2}{N} * \frac{2S-N}{(N+1)^2} = 2p_f - 1 \quad \square$$

This results demonstrates the intuitive idea that the selection should be biased towards features that have been selected often in previous tasks. Based on the attractiveness scores, we propose to pose

$$\beta_f = exp(-sc_f * \alpha_t) \tag{7}$$

to bias the selection towards previously selected features.[8] With $\alpha_t = 0$, features are learned independently on each task. On the contrary, an increasing α_t raises the bias towards features that were already selected in past tasks. Domain experts can thus tune the α_t values to control the accuracy-stability trade-off in such a transfer learning setting.

5 Experiments

In this section, we evaluate to what extent an actual compromise between prediction accuracy and selection stability can be made with the proposed approaches. Experiments are performed on two distinct tasks, prostate cancer diagnosis from microarray data and handwritten digit recognition. The Prostate dataset contains 12600-dimensional (microarray) gene expression data from 52 patients with prostate tumors and 50 healthy patients [22]. The Gisette dataset contains 5000-dimensional integer data, with features aimed at discerning pictures of the number 4 from the number 9. Gisette was originally constructed from the MNIST

[8] Again, $\boldsymbol{\beta}$ is post-centered such that $\mu(\boldsymbol{\beta}) = 1$ at each iteration of the RFE algorithm.

data but was extended with 2500 noisy features [9]. It consists of 6000 examples, but, in order to better illustrate the trade-off, only 100 examples are used here.

5.1 Evaluation Methodology

To obtain the results presented in the next sections, the following methodology has been used. Each (a, ϕ) pair is obtained with a different α (problem 1) or α_t (problem 2). For the single task stability problem, the β_f are first sampled from the gamma distribution. Then, M bootstrap samples are built. k features are then selected using the proposed biased RFE on each bootstrap sample. For the transfer learning stability problem, a single bootstrap sample for each task is created. Features are selected from it, then $\boldsymbol{\beta}$ for the next task is computed according to equation (7). The final prediction model is learned by minimizing the classical, unbiased, logistic loss with a L2 regularization (see equation (5)) with a non-strongly fitted[9] regularization parameter λ. Every model is evaluated on its out-of-bag examples. The mean accuracy as well as the stability of the selected features are computed. As these values are obviously dependent on the sampling of $\boldsymbol{\beta}$ (problem 1) or the features learned on the first few tasks (problem 2), this procedure is repeated B times and the mean values are reported. The 95% confidence *regions* of the expected value of the accuracy-stability trade-off are computed as well as the confidence interval on DA described in section 3. Stability of the feature selection (x-axis on Figures 2, 3 and 4) has not to be confused with its corresponding uncertainty which is the width of the confidence regions along the x-axis.

5.2 Single Task Selection Stability

The λ meta-parameter of the RFE formulation (equation (5)) has not been strongly optimized. A value of 0.1 which provides a good accuracy has been used for both tasks. To obtain the below graphs, the methodology detailed in section 5.1 has been used with $M = 30, k = 20$ and $B = 100$.

Results on both data sets can be seen in Figure 3. The blue curves are obtained without any prior knowledge. The top-left point of each subgraph corresponds to the (accuracy,stability) trade-off obtained with the classical logistic RFE method. Following Pareto lines from left to right, the shape α of the gamma distribution decreases. This makes the biased logistic RFE departs from its unbiased version which raises stability but reduces classification performance. As the method fails to reach maximum stability, it was extended with the trivial point $(a_{rand}, 1)$, obtained by always returning the same arbitrary feature subset.[10] It

[9] Values used for λ are 0.1 for problem 1 and 1 for problem 2. The final classification algorithm does not influence the selection stability. It can thus be optimized to maximize the predictive accuracy only.

[10] It is actually impossible to reach a maximum stability of 1 for a finite regularization parameter λ. In such a case, even with no regularization, a feature is not guaranteed to be always selected.

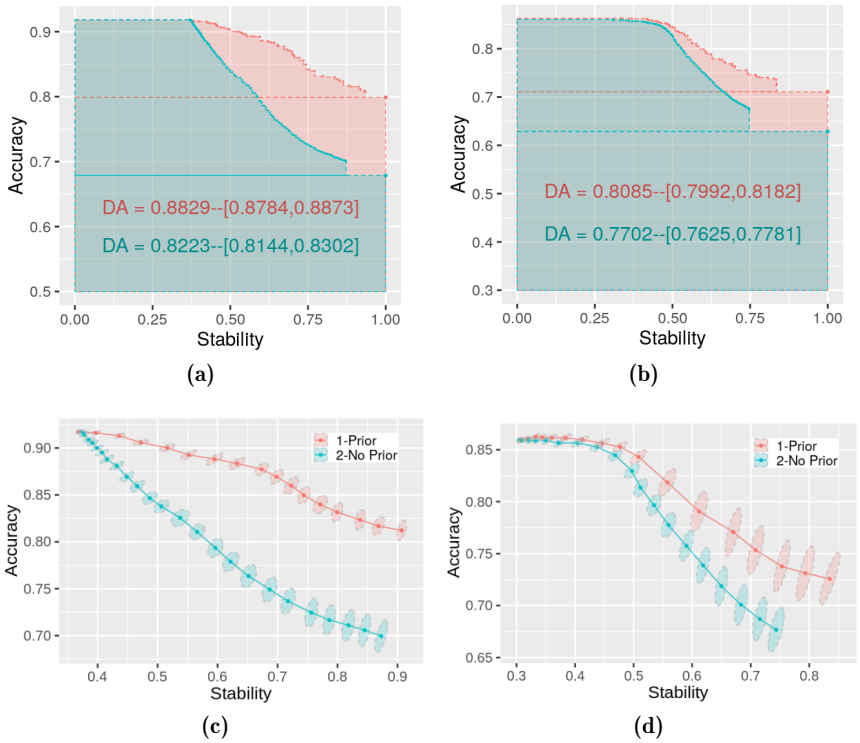


Fig. 3. Performance evaluation of the single task stability problem. DA obtained on Prostate (a,c) and Gisette (b,d) with or without prior knowledge and the corresponding confidence regions.

seems that, while it is possible to increase significantly the stability without degrading too much the accuracy on the Gisette dataset (Figure 3b), it is not the case for the Prostate dataset (Figure 3a) where the accuracy drops directly.

To measure the effect of prior knowledge, $N = 10$ examples are sampled randomly. The 100 features with the highest variance are selected as part of the prior knowledge, here representing a set of potentially relevant features. As can be seen on Figure 3a and 3b, even such a small prior knowledge improves the Dominance Area considerably.

Figures 3c and 3d have been obtained with a small subset of the Pareto points. The ellipses are the 95% confidence *regions* of the expected value (on the $\boldsymbol{\beta}$ sampling) of the accuracy-stability trade-off. For large α values, the importance of $\boldsymbol{\beta}$ is reduced, and thus the uncertainty limited. As α decreases, this confidence region grows. The ellipses are also all inclined towards the right. This represents the covariance between the accuracy and stability for a single $\boldsymbol{\beta}$ sampling. If the sampling appears to be bad, *.i.e.* poor features are prioritized, poor accuracy

and poor stability are obtained. The opposite is true for a good sampling. By using the top-right and bottom-left point of each ellipse, it is possible to derive a confidence interval on the true DA of the method on these datasets.

5.3 Multi-task Selection Stability via Transfer Learning

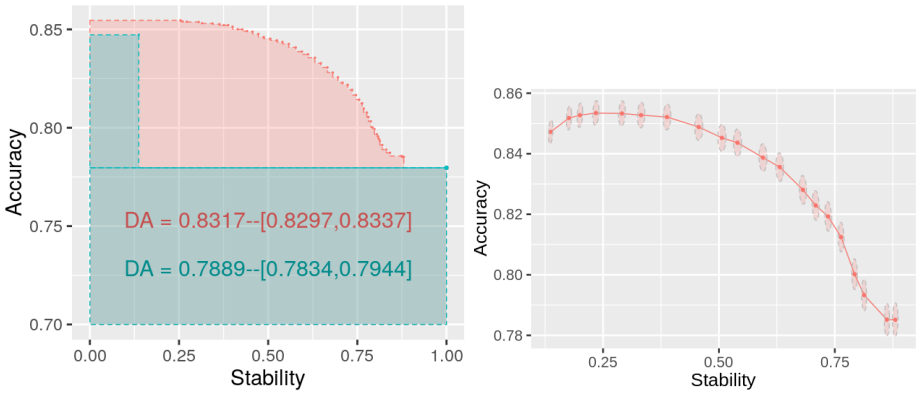


Fig. 4. Accuracy-stability trade-off in a transfer learning setting evaluated on Prostate (a). In red is displayed the DA obtained with the proposed biased RFE. In blue is the DA obtained by combining the two trivial options: either select features on each task independently, or always return the features selected for the first task. Confidence regions of a few points computed by the biased RFE in the transfer learning setting (b).

To generate different, yet similar, classification tasks, normally distributed noise has been added to the Prostate dataset. This noise is centered on 0 and has a specific standard deviation for every couple of feature and task, such that features relevant in some task, could be irrelevant in others. Yet, tasks are expected to share common informative features. 8 tasks are considered here, with an arbitrary order between them. Results with $k = 10, B = 500, \lambda = 1$ are shown on Figure 4a. The blue area is obtained by combining two trivial options. First, the features learned on the first task can be selected for all subsequent tasks, achieving a stability of 1. Or features can be learned independently from each other (equivalent to $\alpha_t = 0$). This strategy offers poor selection stability, but also a sub-optimal classification performance. Knowledge from previous tasks can be used to guide the search towards potentially good features for subsequent tasks. This increases both the accuracy and stability at first. Then, the accuracy starts to decrease, as the selection of features is forced too much. This tendency is

better illustrated in Figure 4b, which contains some non-Pareto optimal points. This result is consistent with the conclusion drawn by the analysis of the Group-Lasso with ℓ_1/ℓ_p regularization [25], *i.e.* that weak coupling norms ($1.5 \leq p \leq 2$) outperforms no and strong coupling norms. Unlike for single task feature selection, the confidence regions are similar for all compromises, meaning that differential shrinkage does not increase the uncertainty of the obtained accuracy-stability pair. Furthermore, as the ellipses are straight, the measured accuracy and stability are uncorrelated.

6 Conclusion and Perspectives

The typical instability of standard feature selection methods is a key concern nowadays as it reduces the interpretability of the predictive models as well as the trust of domain experts towards the selected feature subsets. Such experts would often prefer a more stable feature selection algorithm over an unstable and slightly more accurate one. In this paper, the compromise between feature relevance and selection stability is made explicit by biasing the selection towards some features through differential shrinkage of the Recursive Feature Elimination algorithm. Domain experts are given the opportunity to select any Pareto-optimal trade-off of accuracy and selection stability based on their preferences. We propose the use of the hypervolume metric to assess the performance of methods realizing such a compromise. An associated confidence interval, based on the derivation of confidence *regions* of the accuracy-stability trade-off, is derived.

Results on prostate cancer diagnosis and handwritten recognition tasks show that the selection stability can be increased at will, often with a cost of classification performance. When some prior knowledge is available, far better compromises can be made. The design and evaluation of hybrid methods, learning the prior knowledge from the data, and using it to stabilize the selection is part of our future work.

Motivated by the needs of domain experts, across tasks feature stability is also studied in a transfer learning setting (*i.e.* when tasks are ordered). A biasing scheme that takes the stability measure explicitly into account is proposed. For similar, yet different, tasks, we show on microarray data that some bias is at first beneficial for both the accuracy and the stability. A too strong bias continues to increase the selection stability but at the cost of some classification performance, as the most relevant features vary across tasks. Our approach is evaluated here in a simulated transfer learning setting and further experimental validations will be conducted.

Different multi-task feature selection methods have been proposed in the literature (*e.g.* Group-Lasso with ℓ_1/ℓ_p regularization [25], additive linear models [8], ...). Such methods were introduced with the primary objective of building accurate predictive models across several (this time unordered) tasks. We will study to which extent they could also be used to allow the tuning of the across task selection stability and classification performance trade-off. The biased RFE proposed here can be extended to tackle classical multi-task feature selection, for

example by prioritizing the most relevant features when all tasks are considered together. Our future work includes the evaluation of all these approaches in the proposed assessment framework.

The present paper answers the growing necessity of considering the selection stability not only as a side-effect of learning accurate predictive models but as an actual goal in a bi-objective framework. It proposes initial approaches to learn Pareto-optimal compromises in such a framework and, hopefully, opens the way to new works and improvements in this area.

References

1. Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., Saeys, Y.: Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics **26**(3), 392–398 (2010). https://doi.org/10.1093/bioinformatics/btp630
2. Alelyani, S.: On feature selection stability: A data perspective. Ph.D. thesis, Arizona State University (2013)
3. Argyriou, A., Evgeniou, T., Pontil, M.: Multi-task feature learning. In: Advances in neural information processing systems. pp. 41–48 (2007)
4. Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Machine Learning **73**(3), 243–272 (2008)
5. Awada, W., Khoshgoftaar, T.M., Dittman, D., Wald, R., Napolitano, A.: A review of the stability of feature selection techniques for bioinformatics data. In: Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on. pp. 356–363. IEEE (2012)
6. Boulesteix, A.L., Slawski, M.: Stability and aggregation of ranked gene lists. Briefings in bioinformatics **10**(5), 556–568 (2009)
7. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology **3**(02), 185–205 (2005)
8. Evgeniou, T., Pontil, M.: Regularized multi–task learning. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 109–117. ACM (2004)
9. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A.: Feature extraction: foundations and applications, vol. 207. Springer (2008)
10. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine learning **46**(1-3), 389–422 (2002)
11. Han, Y., Yu, L.: A variance reduction framework for stable feature selection. Statistical Analysis and Data Mining: The ASA Data Science Journal **5**(5), 428–445 (2012)
12. Helleputte, T., Dupont, P.: Feature selection by transfer learning with linear regularized models. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 533–547. Springer (2009)
13. Helleputte, T., Dupont, P.: Partially supervised feature selection with regularized linear models. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 409–416. ACM (2009)
14. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms. In: Data Mining, Fifth IEEE International Conference on. pp. 8–pp. IEEE (2005)
15. Kuncheva, L.I.: A stability index for feature selection. In: Artificial intelligence and applications. pp. 421–427 (2007)

16. Liu, H., Palatucci, M., Zhang, J.: Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 649–656. ACM (2009)
17. Nogueira, S., Sechidis, K., Brown, G.: On the stability of feature selection algorithms. The Journal of Machine Learning Research **18**(1), 6345–6398 (2017)
18. Obozinski, G., Taskar, B., Jordan, M.: Multi-task feature selection. Statistics Department, UC Berkeley, Tech. Rep **2** (2006)
19. Obozinski, G., Taskar, B., Jordan, M.I.: Joint covariate selection and joint subspace selection for multiple classification problems. Statistics and Computing **20**(2), 231–252 (2010)
20. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics **23**(19), 2507–2517 (2007)
21. Shi, L., Reid, L.H., Jones, W.D., Shippy, R., Warrington, J.A., Baker, S.C., Collins, P.J., De Longueville, F., Kawasaki, E.S., Lee, K.Y., et al.: The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements. Nature biotechnology **24**(9), 1151 (2006)
22. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., et al.: Gene expression correlates of clinical prostate cancer behavior. Cancer cell **1**(2), 203–209 (2002)
23. Somol, P., Novovicova, J.: Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(11), 1921–1939 (2010)
24. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288 (1996)
25. Vogt, J., Roth, V.: A complete analysis of the l_1, p group-lasso. arXiv preprint arXiv:1206.4632 (2012)
26. Zhang, M., Zhang, L., Zou, J., Yao, C., Xiao, H., Liu, Q., Wang, J., Wang, D., Wang, C., Guo, Z.: Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. Bioinformatics **25**(13), 1662–1668 (2009)
27. Zhang, Y., Yang, Q.: A survey on multi-task learning. arXiv preprint arXiv:1707.08114 (2017)
28. Zitzler, E., Brockhoff, D., Thiele, L.: The hypervolume indicator revisited: On the design of pareto-compliant indicators via weighted integration. In: International Conference on Evolutionary Multi-Criterion Optimization. pp. 862–876. Springer (2007)
29. Zitzler, E., Thiele, L.: Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. IEEE transactions on Evolutionary Computation **3**(4), 257–271 (1999)
30. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(2), 301–320 (2005)