ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

Joint optimization of predictive performance and selection stability

Victor Hamer and Pierre Dupont

UCLouvain - ICTEAM/INGI/Machine Learning Group, Place Sainte-Barbe 2, B-1348 Louvain-la-Neuve, Belgium.

Abstract. Current feature selection methods, especially applied to high dimensional data, tend to suffer from instability since marginal modifications in the data may result in largely distinct selected feature sets. Such instability strongly limits a sound interpretation of the selected variables by domain experts. We address this issue by optimizing jointly the predictive accuracy and selection stability and by deriving Pareto-optimal trajectories. Our approach extends the Recursive Feature Elimination algorithm by enforcing the selection of some features based on a stable, univariate criterion. Experiments conducted on several high dimensional microarray datasets illustrate that large stability gains are obtained with no significant drop of accuracy.

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 domain [1]. 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 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 studied in depth [2]. 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 drastically change even after marginal modifications of the data. Domain experts would often prefer a more stable feature selection algorithm over an unstable and slightly more accurate one, as selection instability reduces their trust towards the selected features [3, 4]. We address this problem here by deriving Pareto-optimal compromises in the (accuracy, stability) objective space using an extension of the well-known Recursive Feature Elimination (RFE) algorithm. Domain experts can then choose a particular trade-off based on their preferences.

2 Related Work

Looking for a stable feature selection first requires a proper way to quantify stability itself. Many measures have already been proposed: the Kuncheva in-

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

dex [5], the Jaccard index [6], the POG [7] and nPOG [8] 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. In the hope of fixing this issue, a recent work [9] lists and analyzes 15 different stability measures. They are compared based on the satisfaction of 5 different properties that a stability measure should comply with. They also propose a novel and unifying index. This index, used throughout this paper, measures the stability across M selected subsets of features. It can be computed according to equation (1):

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

with $\bar{k}$ the mean number of features selected from the original d features and $s_f^2 = \frac{M}{M-1}\hat{p}_f(1-\hat{p}_f)$ the estimator of the variance of the selection of the f_{th} feature over the M selected subsets, where $\hat{p}_f$ is the fraction of times feature f has been selected among them. These subsets are typically obtained by resampling M times the learning data. This measure is equivalent to the Kuncheva Index (KI)[5] when the number of selected features k is constant across the M selected subsets, but it can be computed in $O(Md)$ whereas KI requires $O(M^2d)$.

Several authors have proposed different approaches to increase stability. For instance, instance-weighting for variance reduction [10] and ensemble methods for feature selection have been proposed [3] and generally increase feature stability. Stability selection [11] is a particular ensemble method which selects features with a selection frequency p_f higher that a threshold π_{thr} for at least one regularisation parameter $\lambda \in \Lambda$. While these methods have been shown to increase selection stability, the gain they offer is still limited as they were not designed to search explicitly through a bi-dimensional (accuracy, stability) objective space.

3 Hybrid Univariate-RFE

We propose to use the methodology illustrated by algorithm 1 to tune the trade-off under study. First, we find a set of *stable features*, S_N, as the top-N features based on an univariate criterion (lines 3,4). Univariate filters tend to be more stable than multivariate methods as they do not take feature interdependencies into account. These N features are then forced to be selected at each iteration of the RFE, which selects, in a multivariate fashion, the most appropriate complementary features. It does so by iteratively minimizing the logistic loss[1] (line 7), ranking every feature based on the absolute value of their weight $\mathbf{w}$ in the learned decision function (line 8) and dropping the one feature with minimal weight[2] (line 9), until the desired number of features k is reached. Finally, it

[1]The original RFE algorithm optimizes the hinge loss but we opt here for the logistic loss allowing for a smoother control of the selected features.

[2]For computational reasons, it is common to drop a fraction of the remaining features instead of a single one at each iteration. We opt here for 20%.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

learns the final decision function by minimizing the logistic loss on the k selected features (line 10), possibly with another regularization constant (λ_f). The difference between this approach and the classic RFE is that the features in S_N are never dropped and are thus always present in the final model. To take advantage of this knowledge, one can apply differential shrinkage on these features to increase their importance in the multivariate selection (line 7, with $\odot$ the element-wise product). The intensity of this differential shrinkage is dictated by the meta-parameter $\epsilon \leq 1$ defined on line 5.

If the set of *stable features*, S_N, is robust, then increasing N, the number of features selected beforehand, is expected to increase the overall selection stability at the possible cost of some predictive accuracy. If $N = 0$, this hybrid RFE is equivalent to the classical RFE, where no feature are pre-selected. When $N = k$, our approach becomes equivalent to a purely univariate filter.

Algorithm 1 Hybrid RFE.

1: **procedure** SELECTFEATURES($N, \lambda, \epsilon, \lambda_f$)
2: $\mathcal{F} \leftarrow$ the set of all features
3: $r_f \leftarrow$ univariate criterion rank of each feature (descending order)
4: $S_N \leftarrow \{f : r_f \leq N\}$
5: $\beta_f \leftarrow \epsilon$ if $f \in S_N, 1$ otherwise
6: **while** $|\mathcal{F}| > k$ **do**
7: $\mathbf{w}^* \leftarrow \text{argmin}_{\mathbf{w}} \sum_{i=1}^{n} \log(1 + \exp^{-y_i(\mathbf{w}\mathbf{x_i})}) + \lambda||\boldsymbol{\beta} \odot \mathbf{w}||_2$
8: $\mathbf{r}^* \leftarrow$ rank features $\{f \in \mathcal{F} \setminus S_N\}$ on $|w_f^*|$ in descending order
9: $\mathcal{F} \leftarrow S_N \cup \{f : r_f^* \leq (|\mathcal{F}| - N - 1)\}$
10: $\mathbf{w}^* \leftarrow \text{argmin}_{\mathbf{w}} \sum_{i=1}^{n} \log(1 + \exp^{-y_i(\mathbf{w}\mathbf{x_i})}) + \lambda_f||\mathbf{w}||_2$
11: **return** $(\mathcal{F}, \mathbf{w}^*)$

4 Experiments

In our experiments, we focus on two univariate criteria: the supervised Golub's ratio [12] and the unsupervised sample variance. The Golub's ratio measures the *signal to noise* ratio and has been often applied when analyzing gene expression data. We observed experimentally that, unlike the variance which is very stable, the Golub's ratio is generally only slightly more stable than the multivariate RFE selection, making any compromise between the two difficult. To generate a family of univariate filters with different (relevance, stability) trade-offs, we pose the following criterion

$$GV_f = \frac{|\mu_+(f) - \mu_-(f)|}{\sigma_+(f) + \sigma_-(f)} + \lambda_v \text{var}(f) \tag{2}$$

with λ_v a parameter balancing between the weight given to the Golub's ratio and the sample variance. A low λ_v generates a relevant but not very stable univariate filter while a high λ_v defines a less relevant but more stable criterion. Experiments are performed on six micro-array datasets, summarized in Table 1.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

Table 1: Information on the micro-array datasets.

author	year	n	d	disease	d after filtering
`alon`	1999	62	2000	colon cancer	2000
`borovecki`	2005	31	22283	Huntington's	1000
`singh`	2002	102	12600	prostate cancer	1250
`gravier`	2010	168	2905	breast cancer	2905
`chiaretti`	2004	111	12625	leukemia	5000
`chin`	2006	118	22215	breast cancer	100

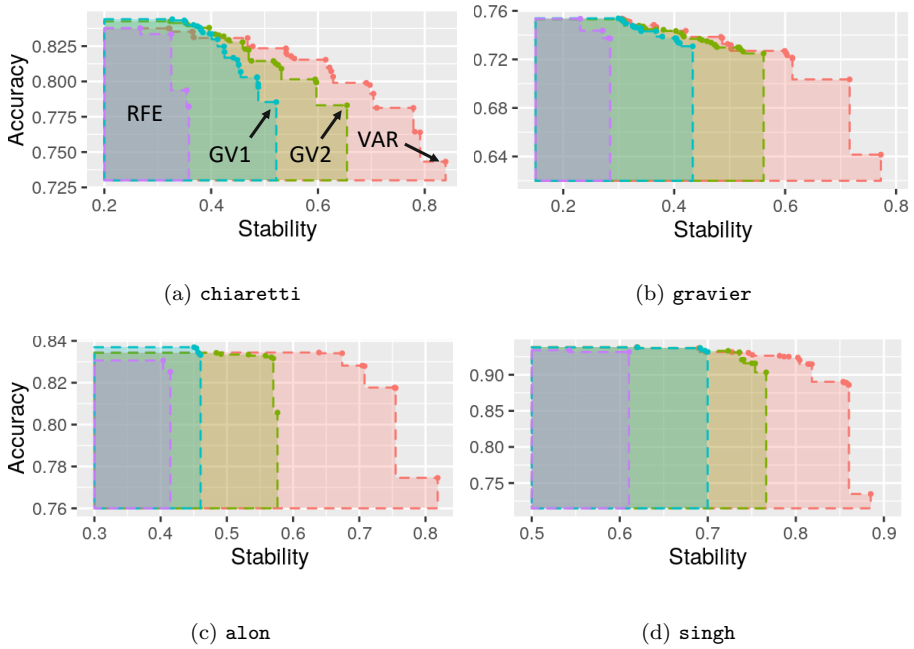
All these datasets have a small n (number of samples) to d (number of features) ratio, which generally causes feature selection methods to be particularly unstable. The learning task consists in predicting whether or not a patient is suffering from the corresponding disease. As is often done when dealing with high dimensional datasets, we first pre-filter the feature space by removing the features with lowest variance (except for `alon` and `gravier`, for which such a pre-filtering had already been performed). The amount of pre-filtering is found such as to maximize the predictive performance of the classical RFE ($N = 0$) and is kept constant for all values of N, λ, ϵ and λ_f. To measure the accuracy and stability obtained with a given set of meta-parameters, we use the classic bootstrap protocol which draws with replacement M samples of the same size as the original dataset. Each model is evaluated on the out-of-bag examples and the mean accuracy is reported. The selection stability is evaluated using equation (1) over the M resamplings.

Results with $k = 20$ and $M = 1000$ can be seen on Figure 1. The plot represents the areas dominated by the Pareto-optimal curves that can be drawn by model selection on λ and λ_f of the classic RFE (purple), the hybrid variance RFE ($\lambda_v \to \infty$) (red) and the hybrid RFE with two other values of λ_v (cyan and green). The hybrid-RFE is able to increase the selection stability by considerable amounts, sometimes even without decreasing the predictive accuracy at all. Model selection is strictly dominated by our approach for all datasets. Apart from some small performance increases for low stabilities, better compromises are reached when $\lambda_v \to \infty$, making our approach mostly sensitive to the stability of the *stable set*, and less to its predictive accuracy. Results on `chin` are similar to the ones reported here. On `borovecki`, the best Pareto curve is obtained by increasing λ_v gradually and fixing N=k. In other words, it appears that there is no benefit to using multivariate selection on this dataset.

Figure 2 shows that the hybrid-RFE is able to select features that are complementary to the ones whose selection is forced. The blue area is obtained by selecting the remaining features independently of the pre-selected ones. The green (red) area is obtained with the hybrid-RFE without (with) differential shrinkage. The independent selection is dominated by our approach and differential shrinkage can improve the compromise further by 1) helping the RFE to select adequate complementary features, hence increasing the predictive accu-

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

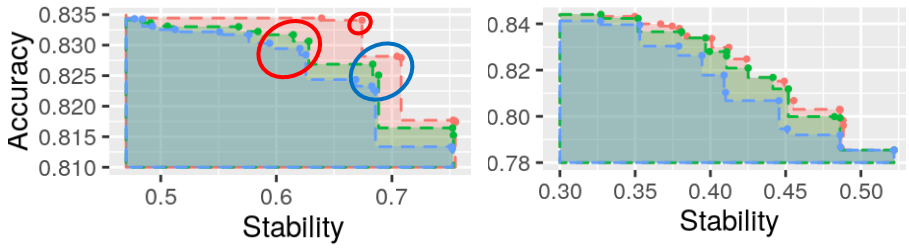
racy and 2) stabilizing the selection of the complementary features. This can be clearly seen on the Figure 2a where points inside circles of the same color corresponds to the same value of N (here 12 and 15).



(a) `chiaretti` (b) `gravier`

(c) `alon` (d) `singh`

Fig. 1: Evaluation of the classic RFE (purple), the variance-hybrid RFE (VAR) (red) and some golub-variance-hybrid RFE (GV1,GV2) (cyan and green). For the sake of readability, the axis of the different plots are not at the same scale.



(a) `alon` (b) `chiaretti`

Fig. 2: Comparison of the independent selection (blue), the hybrid RFE with no differential shrinkage (green) and with differential shrinkage (red). Points inside circles of the same color are obtained with the same value of N.

ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2. Available from http://www.i6doc.com/en/.

5 Conclusion

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, we tackle this issue by considering selection stability as an actual goal in a bi-objective framework. We derive Pareto-optimal trajectories from which domain experts can choose a particular compromise based on their personal preferences. The trajectories are obtained by pre-selecting some features based on a stable univariate criteria, before running the multivariate Recursive Feature Elmination (RFE) algorithm which then selects the most appropriate complementary features.

Results on multiple micro-array datasets show that large stability increases are obtained at small cost of classification accuracy. The performance of our approach mostly depends on the stability of the considered univariate criterion, which makes the sample variance criterion particularly appealing.

References

[1] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[2] Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data classification: Algorithms and applications*, page 37, 2014.

[3] Thomas Abeel, Thibault Helleputte, Yves Van de Peer, Pierre Dupont, and Yvan Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–398, 2010.

[4] Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, 12(1):95–116, 2007.

[5] Ludmila I Kuncheva. A stability index for feature selection. In *Artificial intelligence and applications*, pages 421–427, 2007.

[6] Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms. In *Data Mining, Fifth IEEE International Conference on*, pages 8–pp. IEEE, 2005.

[7] Leming Shi, Laura H Reid, Wendell D Jones, Richard Shippy, Janet A Warrington, Baker, et al. The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature biotechnology*, 24(9):1151, 2006.

[8] Min Zhang, Lin Zhang, Jinfeng Zou, Chen Yao, Hui Xiao, et al. Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics*, 25(13):1662–1668, 2009.

[9] Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the stability of feature selection algorithms. *The Journal of Machine Learning Research*, 18(1):6345–6398, 2017.

[10] Yue Han and Lei Yu. A variance reduction framework for stable feature selection. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):428–445, 2012.

[11] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[12] Todd R Golub et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.