Towards robust feature selection techniques

Yvan Saeys YVAN.SAEYS@PSB.UGENT.BE
Thomas Abeel THOMAS.ABEEL@PSB.UGENT.BE
Yves Van de Peer YVES.VANDEPEER@PSB.UGENT.BE
Department of Plant Systems Biology, VIB, Technologiepark 927, 9052 Gent, Belgium,
Department of Molecular Genetics, Ghent University, Technologiepark 927, 9052 Gent, Belgium

Abstract

Robustness of feature selection techniques is a topic of recent interest, especially in high dimensional domains with small sample sizes, where selected feature subsets are subsequently analysed by domain experts to gain more insight into the problem modelled. In this work, we investigate the robustness of various feature selection techniques, and provide a general scheme to improve robustness using ensemble feature selection. We show that ensemble feature selection techniques show great promise for small sample domains, and provide more robust feature subsets than a single feature selection technique. In addition, we also investigate the effect of ensemble feature selection techniques on classification performance, giving rise to a new model selection strategy.

1. Introduction

During the past decade, the use of feature selection for knowledge discovery has become increasingly important in many domains that are characterized by a large number of features, but a small number of samples. Typical examples of such domains include text mining, computational chemistry and the bioinformatics and biomedical field, where the number of features (problem dimensionality) often exceeds the number of samples by orders of magnitude (Saeys et al., 2007). When using feature selection in these domains, not only model performance but also robustness of the feature selection process is important, as domain experts would prefer a stable feature selection algorithm over an unstable one when only small changes are made to the dataset.

Surprisingly, the robustness (stability) of feature selection techniques is an important aspect that received only relatively little attention during the past. Recent works in this area mainly focus on the stability indices to be used for feature selection, introducing measures based on Hamming distance (Dunne et al., 2002), correlation coefficients (Kalousis et al., 2007), consistency (Kuncheva, 2007) and information theory (Krízek et al., 2007). The work of Kalousis et al. (2007) also presents an extensive comparative evaluation of feature selection stability over a number of high-dimensional datasets. However, most of these recent works only focus on the stability of single feature selection techniques.

In this work, we investigate whether the use of ensemble feature selection techniques can be used to yield more robust feature selection techniques, and whether combining multiple methods has any effect on the classification performance.

2. Methods

2.1. Quantification of robustness

Depending on the outcome of a feature selection technique, the result can be either a set of weights, a ranking, or a particular feature subset. In order to assess robustness, a subsampling scheme is used that generates k subsamples containing 90% of the original data. The robustness of a technique is then measured by the average over all pairwise similarity comparisons between the different feature selectors:

$$S_{\text{tot}} = \frac{2\sum_{i=1}^{k}\sum_{j=i+1}^{k} S(\mathbf{f}_i, \mathbf{f}_j)}{k(k-1)}$$

where $\mathbf{f}_i$ represents the outcome of the feature selection method applied to subsample i $(1 \leq i \leq k)$, and $S(\mathbf{f}_i, \mathbf{f}_j)$ represents a similarity measure between $\mathbf{f}_i$ and $\mathbf{f}_j$.
Here, we focus on similarities between rankings - using the Spearman rank correlation coefficient - and subsets, using the Jaccard index (Kalousis et al., 2007) or

Table 1. Robustness of the different feature selectors across the different datasets. Spearman correlation coefficient (Sp), Jacard index (JC) and consistency index (CI) on a subset of 1% best features.

Dataset		SU		Relief		SVM_RFE		RF	
		Single	Ensemble	Single	Ensemble	Single	Ensemble	Single	Ensemble
Colon	Sp	0.61	0.76	0.62	0.85	0.7	0.81	0.91	0.99
	JC	0.3	0.55	0.45	0.56	0.44	0.5	0.01	0.64
	CI	0.45	0.7	0.61	0.71	0.6	0.65	0.01	0.77
Leukemia	Sp	0.68	0.76	0.58	0.79	0.73	0.79	0.97	0.99
	JC	0.54	0.6	0.44	0.55	0.49	0.57	0.36	0.8
	CI	0.7	0.74	0.6	0.71	0.64	0.72	0.53	0.89
Lymphoma	Sp	0.59	0.74	0.49	0.76	0.77	0.81	0.96	0.99
	JC	0.37	0.55	0.42	0.56	0.43	0.46	0.22	0.73
	CI	0.53	0.7	0.58	0.71	0.6	0.63	0.35	0.84
Average	Sp	0.63	0.75	0.56	0.8	0.73	0.80	0.95	0.99
	JC	0.40	0.57	0.44	0.56	0.45	0.51	0.2	0.72
	CI	0.56	0.71	0.6	0.71	0.61	0.67	0.3	0.83

the consistency index (Kuncheva, 2007).

2.2. Ensemble feature selection

In order to improve the robustness of feature selection, a similar idea as in ensemble learning can be used, where multiple classifiers are combined in order to improve performance. In this work, we construct an ensemble of feature selectors by bootstrapping the data, and creating a *consensus feature selector* that aggregates the results of the single feature selectors by rank summation.

3. Results

Table 1 shows the robustness of a representative sample of feature selectors, including two filter based approaches (Symmetrical Uncertainty (SU) and Relief) and two embedded approaches (recursive feature elimination using a SVM (SVM_RFE) and Random Forests (RF)). It can be observed that constructing an ensemble version of each feature selector significantly improves robustness.

However, considering only robustness of a feature selection technique is not an appropriate strategy to find good feature rankings or subsets, and also model performance should be taken into account to decide which features to select. Therefore, feature selection needs to be combined with a classification model in order to get an estimate of the performance of the feature selector-classifier combination.

Our results show that in most cases, classification performance using ensemble feature selection is comparable to the performance using conventional feature selection techniques, or performs only slightly worse (data not shown). It turns out that the best trade-off between robustness and classification performance depends on the dataset at hand, giving rise to a new model selection strategy, incorporating both classification performance as well as robustness in the evaluation strategy by taking e.g. the harmonic mean of robustness and classification performance as a combined measure.

Acknowledgments

Y.S. is funded by a post-doctoral grant from the Research Foundation Flanders (FWO-Vlaanderen). T.A. is funded by a grant from the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

References

Dunne, K., Cunningham, P., & Azuaje, F. (2002). *Solutions to instability problems with sequential wrapper-based approaches to feature selection* (Technical Report TCD-2002-28). Dept. of Computer Science, Trinity College, Dublin, Ireland.

Kalousis, A., Prados, J., & Hilario, M. (2007). Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowl. Inf. Syst.*, *12*, 95–116.

Krízek, P., Kittler, J., & Hlavác, V. (2007). Improving stability of feature selection methods. *Proceedings of the 12th International Conference on Computer Analysis of Images and Patterns* (pp. 929–936).

Kuncheva, L. (2007). A stability index for feature selection. *Proceedings of the 25th International Multi-Conference on Artificial Intelligence and Applications* (pp. 309–395).

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*, 2507–2517.