Ensemble Methods for Robust Feature Selection

Benjamin Rabier
Alexander Goscinski

Motivation

Feature selection is a preprocessing step used in machine learning application to find a small subset of features in order to build a more accurate, simpler and faster model

However domain experts would prefer a more stable algorithm to have more confidence in the selected features

One approach for more robust results are ensemble methods

Feature Selection Methods

Symmetrical Uncertainty Measures the normalized mutual information between single features and their class

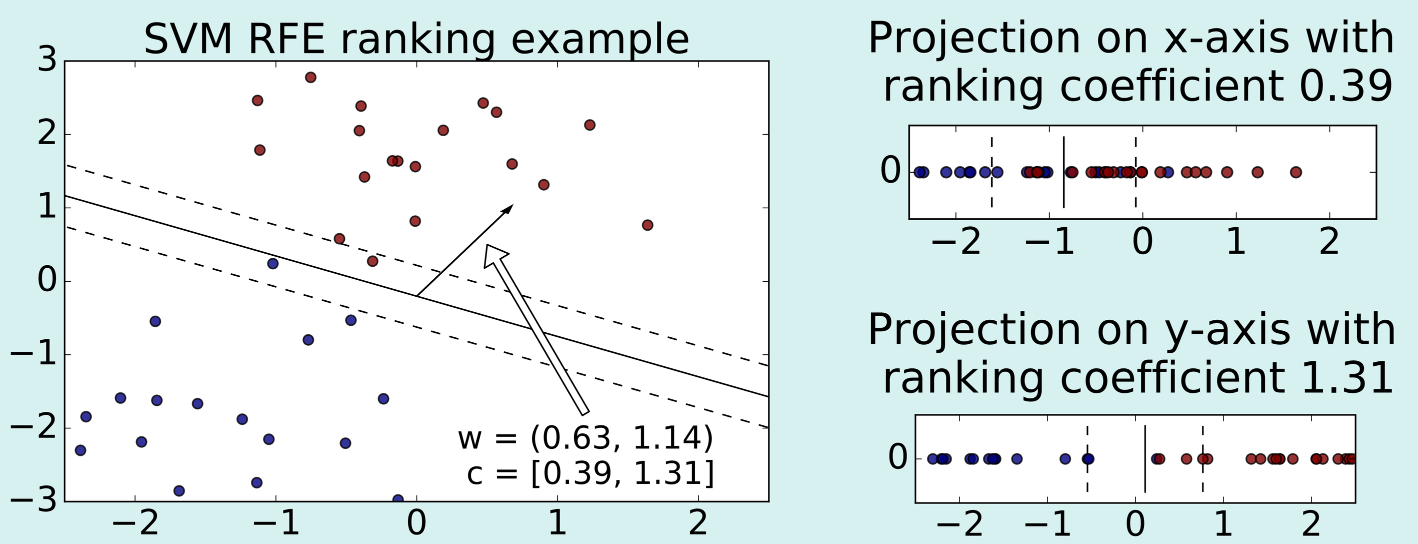
$$SU(F,C) = 2\frac{H(F)-H(F|C)}{H(F)+H(C)}, \text{ where } H(\cdot) \text{ is entropy}$$

RELIEF Samples are taken randomly and compared in terms of similarity to their nearest neighboring sample

$$W_i = W_i - \|x_i - \text{Near-hit}_i\|_2^2 + \|x_i - \text{Near-miss}_i\|_2^2$$

SVM RFE Recursive feature elimination fits a SVM and ranks the feature according to their according to their importance in the model

Ranking coefficient: $c_i = w_i^2$



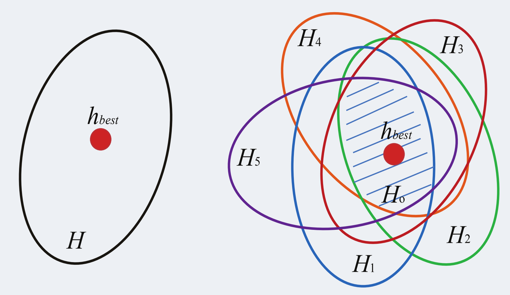
Lasso RFE Recursive feature elimination fits a Lasso model and ranks the feature according to their according to their importance in the model

$$\min_w \quad \frac{1}{2\,\#\text{features}} \|y - Xw\|_2^2 + \alpha \|w\|_1$$

An advantage of a Lasso model is that the weight vector is extremely sparse due to the L1 Norm

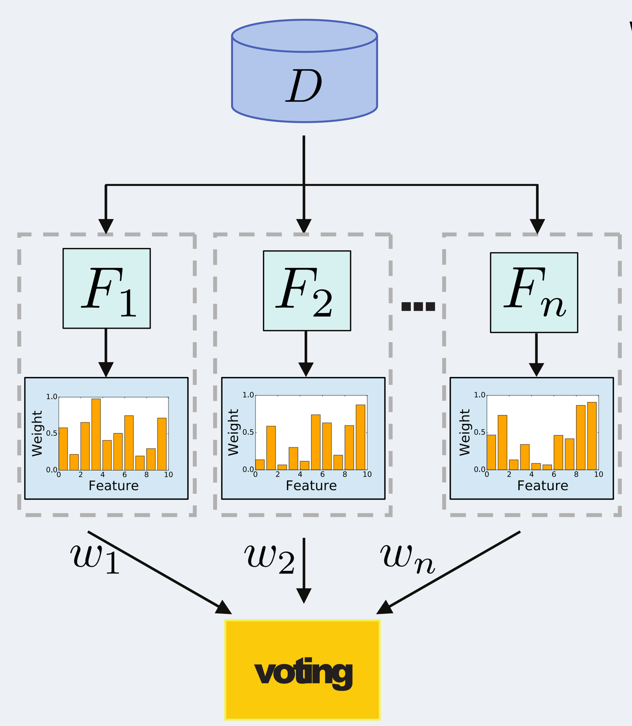
Ensemble Learning

In ensemble learning multiple algorithms are combined to obtain better performance or stability.



We build on the work of Saeys et al. (2008) by combining multiple feature selection methods with different ensemble methods

Ensemble methods The weights of different feature selection methods are first computed and then linearly aggregated to obtain the new weights



We used two methods for this work

Averaging The mean is used as the new weight

Perfomance-related The accuracy is estimated for each of the feature selectors for multiple classifiers. The feature selectors are then weighted according to their performance

Feature Selectors: F_i Aggregation weights: w_i

Stability

Jaccard index was used to measure the similarity between two feature selectons.

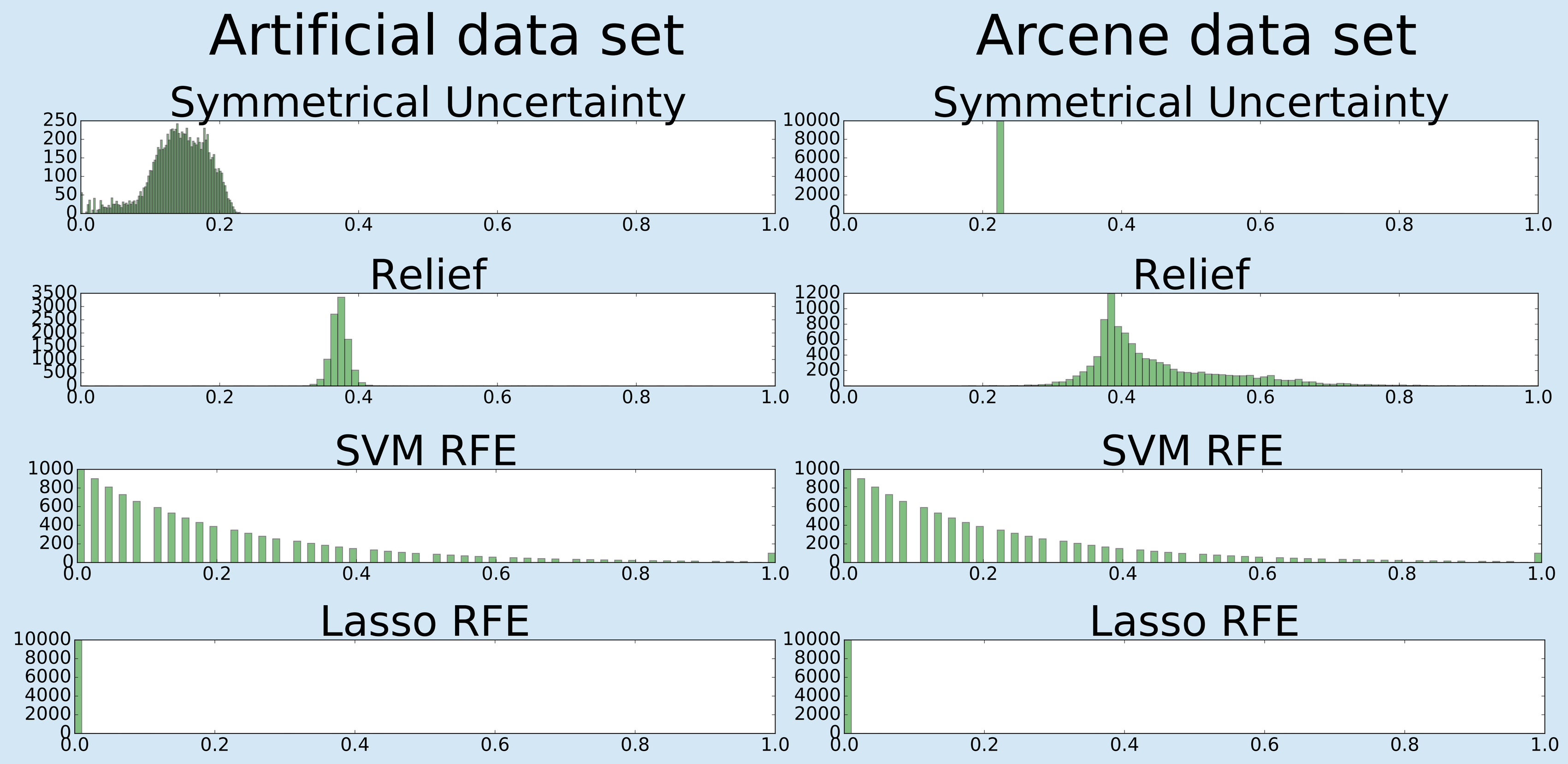
$$S(\mathbf{f}_i, \mathbf{f}_j) = \frac{|\mathbf{f}_i \cap \mathbf{f}_j|}{|\mathbf{f}_i \cup \mathbf{f}_j|} \qquad S_{\text{tot}} = \frac{2\sum_{i=1}^{k}\sum_{j=i+1}^{k} S(\mathbf{f}_i, \mathbf{f}_j)}{k(k-1)}$$

The set of features chosen by feature selector $i : \mathbf{f}_i$

Stability measurement

- Took 10 subsets of the data
- Weighted each subset with the feature selection method
- Calculated the similarity between the subsets with Jaccard index for the 1% best features and normalized it

Weight distributions



Discussion

The ensemble method using the mean is always one of the best, if not the best, in terms of performance. However, while the stability of the selected features was higher than the Lasso and the SVM RFE, it was often considerably lower with RELIEF and SU. On some data sets (e. g. GISETTE) SU was the best trade-off between stability and performance. Moreover weighting the feature selectors according to their performance did not change considerably the result. It was at best as good as the mean, and sometimes even worse.

References

Yvan Saeys, Thomas Abeel, and Yves Van de Peer. Robust Feature Selection Using Ensemble Feature Selection Techniques, pages 313–325. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

Graphics: Yang, Pengyi, et al. "A review of ensemble methods in bioinformatics." Current Bioinformatics 5.4 (2010): 296-308.