EFS: An ensemble of feature selection methods for binary classification

Ursula Neumann and Dominik Heider
Department of Bioinformatics, Straubing Center of Science
u.neumann@wz-straubing.de

Background
Feature selection methods are essential to identify a subset of features that improve the prediction performance of subsequent classification models. Preceding studies showed the defectiveness of single feature selection methods, whereas an ensemble feature selection has the advantage to alleviate and compensate for such biases. By using an ensemble of feature selection methods, a quantification of the importance of features can be obtained. The key features of our EFS method are:

1. The combination of widely known and extensively tested feature selection methods.
2. The balance of biases by using an ensemble.
3. The evaluation of EFS via logistic regression.

Results
With the development of the EFS method we take advantage of the benefits of multiple feature selection methods and combine their normalized outputs to a quantitative ensemble importance. Eight different feature selection methods have been used for the EFS approach. Since random forests have drawn increased attention in the field of predictive medicine [DRH+11, RSN+16], four of the chosen feature selection methods are embedded in a random forest algorithm. Further, we considered the outcome of a logistic regression (i.e., the coefficients) as another embedded method as well as the filter methods median, Pearson-, and Spearman-correlation [YL04].
Conclusion
For feature selection purposes in binary classification our novel `ensemble_fs` function gained an improved performance evaluated by the `logreg_test` function. Both functions can be found in our *EFS* R-package. Additionally, a web application is provided for researchers which are not accustomed to R on http://EFS.heiderlab.de.

References

[DRH+11] J. Nikolaj Dybowski, Mona Riemenschneider, Sascha Hauke, Martin Pyka, Jens Verheyen, Daniel Hoffmann, and Dominik Heider. Improved Bevirimat resistance prediction by combination of structural and sequence-based classifiers. *BioData Mining*, 4:26, 2011.

[RSN+16] Mona Riemenschneider, Robin Senge, Ursula Neumann, Eyke Hüllermeier, and Dominik Heider. Exploiting HIV-1 protease and reverse transcriptase cross-resistance information for improved drug resistance prediction by means of multi-label classification. *BioData Mining*, 9:10, 2016.

[YL04] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res.*, 5:1205–1224, 2004.