The EFS-Server: A Web-Application for Feature Selection in Binary Classification

Ursula Neumann and Dominik Heider

Straubing Center of Science,
Petersgasse 18, 94315 Straubing, Germany
u.neumann@wz-straubing.de,d.heider@wz-straubing.de
http://efs.heiderlab.de/

Abstract. Feature selection methods are essential to identify a subset of features that improve the prediction performance of subsequent classification models and thereby also simplify their interpretability. Preceding studies showed the defectiveness in terms of specific biases of single feature selection methods, whereas an ensemble of feature selection techniques has the advantage to alleviate and compensate for such biases. With the development of the ensemble feature selection (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. We evaluated the EFS method on a testset and it turned out that the subset of features retrieved by the EFS method showed a significantly improved performance in a subsequent logistic regression (LR) model compared to a model using all available features.
EFS can be downloaded as an R-package or used in a websever at http://EFS.heiderlab.de.

Introduction

Machine learning models have been widely used for classification of biomedical problems, e.g., in drug resistance [1] or prediction of the severity of diseases [2]. However, in these areas one is frequently faced with high-dimensional data and small-n-large-p problems, thus the need for simplification of datasets with many parameters frequently emerges.

Therefore, a great variety of feature selection (FS) techniques already exists. However, different feature selection methods provide different subsets of features. There are several factors that can cause instability and unreliability of the feature selection, e.g., the complexity of multiple relevant features, a small-n-large-p-problem, or when the algorithm simply ignores stability [3, 4]. To counteract instability and therewith unreliability of feature selection methods, we developed an ensemble feature selection (EFS) method, which compensates biases of single FS. The idea of ensemble methods is already widely used in learning algorithms [5]. By using an ensemble of feature selection methods, a quantification of the importance of features can be obtained and the method-specific biases can be compensated.

Methods

The EFS method provides eight different techniques for feature selection in binary classification: Since random forests [6] have been shown to give highly accurate predictions on biological [7, 8] and biomedical data [1, 9], four of the chosen feature selection methods are embedded in a random forest algorithm. Further, we considered the outcome of an LR (i.e., the coefficients) as another embedded method as well as the filter methods median, Pearson-, and Spearman-correlation [10]. 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 LR.

We normalized all individual outputs to a common scale, an interval from 0 to 1. Thereby we ensure the comparability between different FS methods and conserve the distances of importance between one feature to another. This normalization is achieved in two different ways: For all feature selections, except for the median, the absolute value of the FS method output is a value which illustrates the increase of importance. By dividing through the maximum value we get values between 0 and 1:

$$imp_{X_i} = \frac{\beta_i}{\max(\beta_m)_{m \in M}}.$$

In the case of the median FS we receive a $p-$value for each feature X_i, which is normalized as follows:

$$imp_{X_i} = 1 - p_i + \min(p_i).$$

By dividing the calculated importances through the number of selected methods (1 to 8) and summing up all individual importances, we get an EFS importance between 0 and 1. The EFS system selects those parameter that have a higher importance than the mean importance:

$$imp_{X_i} > \overline{imp_{X_M}},$$

where $\overline{imp_{X_M}}$ symbolizes the mean of all variable importances.

Results

In order to evaluate our EFS method, we used an LR model with leave-one-out cross-validation (LOOCV). For comparison purposes, we also trained an LR model without feature selection and examined both AUC-values of the ROC curves with ROCR [11]. The dataset *SPECTF* has been obtained from the UCI Machine Learning Repository [12]. It describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. The class-variable is distinguishing between normal (= 0) and abnormal (= 1). In panel **A)** of Figure 1 the resulting ROC curves are shown. Additionally, the p-value ($p < 0.001$)

is located in the bottom right corner of the diagram. The p-value clearly shows that there is a significant improvement in terms of AUC of the LR with features selected by the EFS method compared the LR model without feature selection. The calculation of the p-value is based on the method of DeLong et al. [13]

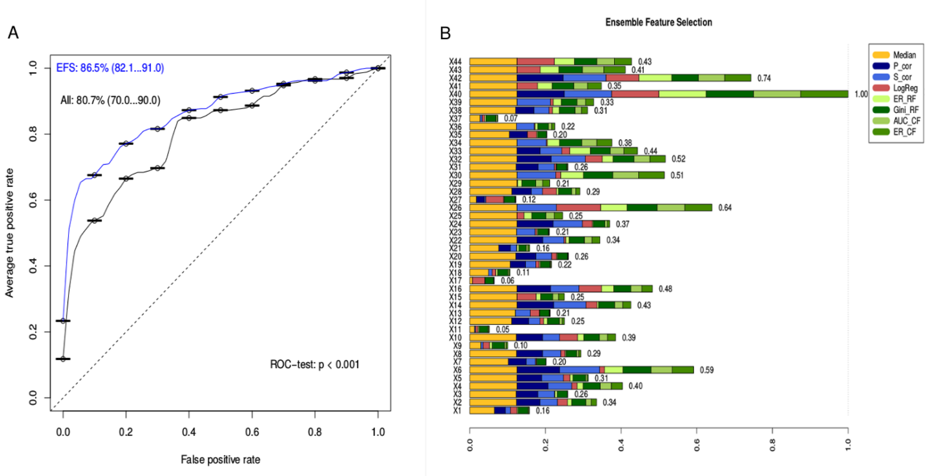

Fig. 1. A) Performance of LR model. On the y-axis the average true positive rate (i.e., sensitivity) and on the x-axis the false positive rate (i.e., 1-specificity) is shown. Two ROC curves are shown: of all features (black) and the EFS selected features (blue). The dotted line marks the performance of random guessing. **B)** Added-up-barplot output of the `barplot_fs` function of R-package EFS.

Conclusion

Besides the R-package EFS, a web application is provided for researchers which are not familiar with the use of R at `http://EFS.heiderlab.de`. The EFS-server provides a feature ranking by summing up the normalized importances of all feature selection methods. Additionally, the EFS-server produces a barplot of the importances, if the number of features does not exceed 25. If a barplot for more than 25 parameters is required, the `barplot_fs` function of R-package EFS can be used (cf. panel B in Figure 1). Moreover, the user can download all results from the feature selection methods and the EFS-method as a csv-file for further analyses. Based on the results of our EFS method, a significant improvement in prediction performance compared to all features in an LR model could be demonstrated. The EFS-server is a nice and handy tool for unexperienced users that provides a feature selection method in a simple and guided procedure. For the experienced user, the corresponding R-package can be used, which provides also deeper insights into the selection and evaluation.

References

1. Riemenschneider, M., Senge, R., Neumann, U., Hüllermeier, E., Heider, D.: 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)
2. Baars, T., Neumann, U., Jinawy, M., Hendricks, S., Sowa, J.P., Klsch, J., Riemenschneider, M., Gerken, G., Erbel, R., Heider, D., Canbay, A.: In Acute Myocardial Infarction Liver Parameters Are Associated With Stenosis Diameter. Medicine 2016, 95(6):e2807.
3. Jain, A., Zongker, D.: Feature selection: evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(2), 153–158 (1997)
4. He, Z., Yu, W.: Stable feature selection for biomarker discovery. Computational Biology and Chemistry 34, 215–225 (2010)
5. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. John Wiley & Sons (2004)
6. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
7. van den Boom, J., Heider, D., Martin, S.R., Pastore, A., Mueller, J.W.: 3-phosphoadenosine 5-phosphosulfate (paps) synthases, naturally fragile enzymes specifically stabilized by nucleotide binding. J Biol Chem. 287(21), 1764555 (2012)
8. Touw, W.G., Bayjanov, J.R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., van Hijum, S.A.: Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? Brief Bioinform. . 14(3), 31526 (2013)
9. Dybowski, J.N., Riemenschneider, M., Hauke, S., Pyka, M., Verheyen, J., Hoffmann, D., Heider, D.: Improved Bevirimat resistance prediction by combination of structural and sequence-based classifiers. BioData Mining, 4, 26 (2011)
10. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res., 5, 1205–1224 (2004)
11. Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T.: Rocr: visualizing classifier performance in r. Bioinformatics 21(20), 3940–3941 (2005)
12. Lichman, M.: UCI Machine Learning Repository (2013), http://archive.ics.uci.edu/ml
13. DeLong ,E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 44, 837–845 (1988)