$\frac{\pi}{4}$-rotation method for feature subset selection

Ursula Neumann and Dominik Heider

Straubing Center of Science,
Petersgasse 18, 94315 Straubing, Germany
Department of Mathematics and Computer Science, University of Marburg
Hans-Meerwein-Str. 6, 35032 Marburg, Germany
u.neumann@wz-straubing.de,d.heider@wz-straubing.de

Motivation

In the field of data mining feature selection plays a crucial role as preprocessing step. Former studies revealed the importance of feature selection methods to enhance the performance and effectiveness of algorithms in pattern recognition, classification, and regression [1, 2]. The methods are designed to distinguish features which are relevant for a prediction model from those which are negligible. By that, an efficient subset selection criterion is indispensable to set a cutpoint between relevant and irrelevant features. We developed an ensemble feature selection (EFS) method, which uses the mean of importance as an integrated subset selection criterion. For small datasets, the EFS method outperforms each single method [3]. However, in large datasets, i.e., with more than 1000 features, smaller subsets seem to be less prone to overfitting and thus lead to better prediction results. Another method to find a suitable subset is detecting the cutpoint with the hightest slope, meaning the point with the highest increase of importance. For datasets with an exponential curve of importance values, this method selects only the feature with the highest importance. Based on these findings, we developed a more conservative feature subset selection method: the $\frac{\pi}{4}$-rotation.

Methods

The idea of the $\frac{\pi}{4}$-rotation method is to draw a curve of the ascending importance values received from the EFS algorithm (c.f. panel A) and B) of figure 1) and find the cutpoint, where the smoothed curve exceeds a slope of 45 degrees. Features which lie above this cutpoint are considered to be important.

It is possible, that there are several points in which the slope fulfills the requirement to be over 45 degree. Therefore, we rotate the curve by -45 degrees (i.e. $-\frac{\pi}{4}$ in radians) and seek for minima as follows: Every element of the rotated curve is tested in an interval of $\frac{1}{4}$ of the total amount of elements, if it is a global minimum in this interval. Being a global minimum means, that all elements in this interval on the left as well as on the right side have higher values than the tested element.

If the minimum is the last variable of the range ordered by ascending importance, no distinct leap at the curve of importance values exists. In this case, the $\frac{\pi}{4}$-rotation method can not be applied.

Results

We analyzed two big datasets *Ad* and *Arcene* with 1430 and 79360 features received from the UCI Machine Learning Repository [4]. Both possess an exponential curve of importance values. The evaluations via ROC curves of logistic regression models revealed that the subset selected by our $\frac{\pi}{4}$-rotation method is significantly better than by taking all features with an importance above average (*Arcene*: $p = 0.004$ resp. *Ad*: $p < 0.001$), i.e., by using the standard selection procedure of EFS. Significance was calculated by a roc-test by the method of DeLong et al. [5], by comparing the ROC curves retrieved from the mean subset with the ROC curve from the $\frac{\pi}{4}$-rotation cutpoint subset.

References

1. Kohavi R., John G.H.: Wrappers for feature subset selection. Artificial Intelligence, 97(1), 273-324 (1997)
2. Chandrashekar G., Sahin F.: A survey on feature selection methods. Computers & Electrical Engineering 40(1), 16-28 (2014)
3. Neumann U., Riemenschneider M., Sowa J.P., Baars T., Kaelsch J., Canbay A., Heider D.: Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach. Biodata Mining, 9,36 (2016)
4. Lichman, M.: UCI Machine Learning Repository (2013), http://archive.ics.uci.edu/ml
5. 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)

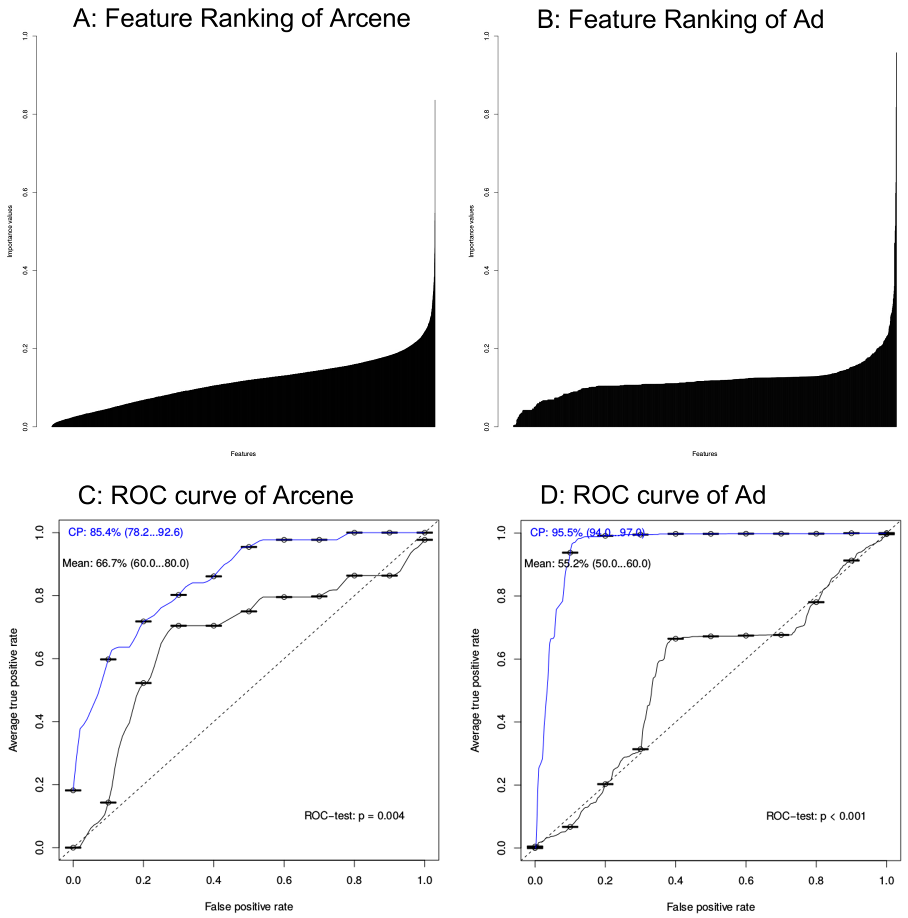

Fig. 1. Feature importance values in ascending order of A) Arcene dataset and B) Ad dataset.
ROC curves of logistic regression models with all features and features with importance values over the mean by EFS of C) Arcene dataset and D) Ad dataset.