Ensemble Feature Selection for Regression Problems

Ursula Neumann and Dominik Heider

Department of Mathematics and Computer Science, University of Marburg

Abstract

Feature selection (FS) can be used to detect and select a relevant subset from a set features for the construction of a machine learning or statistical models. In this process, as much of irrelevant and redundant information as possible should be identified and removed from the data. Thus, FS methods cannot only be used to identify relevant features, but also to reduce noise.

Based on the assumption that combining several weak FS algorithms obtains more reliable results than individual FS approaches, ensemble feature selection gained a high level of attention in the recent years (Neumann et al., 2016). This is due to the fact that single FS methods are prone to be biased depending on the data quality and distribution. Datasets are often imbalanced and heterogeneous. Thus, different methods have different biases and benefits according to the type of features, the degree of imbalance, and the size of the dataset.

In a former study, we implemented an ensemble feature selection (EFS) for binary classification. In the current study, we introduce the extension to regression problems. Classification is the task of predicting a discrete class label, whereas regression is the problem of predicting a continuous quantity output. Some FS algorithms can be used for both classification and regression with small adaptions, such as the random forests importance estimations, however, others are only applicable for feature selection in classification problems.

In the current study, we implemented and evaluated different FS methods for quantitative feature ranking in regression problems. Moreover, the best-performing methods were then combined into a regression EFS approach, implemented into the R package EFS (Neumann et al., 2017), and evaluated using several real-world datasets.

References

U. Neumann , M. Riemenschneider,J.-P. Sowa,T. Baars, J. Kälsch, A. Canbay, D. Heider (2016). Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach. BioData Mining, 9(1):36.

U. Neumann, N. Genze, D. Heider (2017). EFS: An Ensemble Feature Selection Tool implemented as R-package and Web-Application. BioData Mining, 10:21.