Rcpp Implementation of Entropy Based Feature Selection Algorithms with Sparse Matrix Support

by Zygmunt Zawadzki, Marcin Kosiński

Abstract

Feature selection is a process of extracting valuable features that have significant influence on dependent variable. Time efficient feature selection algorithms are still an active field of research and are in the high demand in the machine learning area.

We introduce **FSelectorRcpp**, an R package (R Core Team, 2012) that includes entropy based feature selection algorithms. Methods presented in this package are not new, they were reimplemented in C++ and originally come from **FSelector** package (Romanski and Kotthoff, 2016), but we provide many technical improvements. Our reimplementation occures to have shorter computation times, it does not require earlier Java nor Weka (Hall et al., 2009) installation and provides support for sparse matrix format of data, e.g. presented in **Matrix** package (Bates and Maechler, 2016). This approach facilitates software installation and improves work with bigger datasets, in comparison to the base R implementation in **FSelector**, which is even not optimal in the sense of R code.

Additionally, we present new, C++ implementation of continuous variables Multi-Interval Discretization (MDL) method (Fayyad and Irani, 1993), which is required in entropy calculations during the feature selection process in showed methods. By default, regular **FSelector** implementation uses **entropy** package (Hausser and Strimmer, 2014), for which we also attach the computation times comparison.

Finally, we announce the full list of available functions, which are divided to 2 groups: entropy based feature selection methods and stepwise attribute selection functions that might use any evaluator to choose propoer features, e.g. presented entropy based algorithms.

Introduction and Motivation

Discretization

Entropy Based Feature Selection Algorithms

In the information theory entropy is

Stepwise Attribute Selection Evaluators

FSelectorRcpp and FSelector Computation Times Comparison

Conclusion

Acknowledgment

Bibliography

D. Bates and M. Maechler. *Matrix: Sparse and Dense Matrix Classes and Methods*, 2016. URL https://CRAN.R-project.org/package=Matrix. R package version 1.2-7. [p1]

U. M. Fayyad and K. B. Irani. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In *13th International Joint Conference on Uncertainly in Artificial Intelligence(IJCAI93)*, pages 1022–1029, 1993. [p1]

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009. ISSN 1931-0145. doi: 10.1145/1656274.1656278. URL http://doi.acm.org/10.1145/1656274.1656278. [p1]

J. Hausser and K. Strimmer. *entropy: Estimation of Entropy, Mutual Information and Related Quantities*, 2014. URL https://CRAN.R-project.org/package=entropy. R package version 1.2.1. [p1]

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL http://www.R-project.org/. ISBN 3-900051-07-0. [p1]

P. Romanski and L. Kotthoff. *FSelector: Selecting Attributes*, 2016. URL https://CRAN.R-project.org/package=FSelector. R package version 0.21. [p1]

Zygmunt Zawadzki

zygmunt@zstat.pl

Marcin Kosiński
Warsaw Univeristy of Technology
Faculty of Mathematics and Information Science
Koszykowa 75, Warsaw Poland
m.kosinski@mini.pw.edu.pl