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 the discretization method for continuous variables called 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 **RWeka** package (Hornik et al., 2009) for discretization and **entropy** (Hausser and Strimmer, 2014) for entropy calculations - for both 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

In modern statistical learning the biggest bottlenecks are computation times of model training procedures and the overfitting. Both are caused by the same issue - the high dimension of explanatory variables space. Researchers have encountered problems with too big sets of features used in machine learning algorithms also in terms of model interpretaion. This motivates applying feature selection algorithms before performing statisical modeling, so that on smaller set of attributes the training time will be shorter, the interpretation might be clearer and the noise from non important features can be avoided. More motivation can be found in John et al. (1994).

Many methods were developed to reduce the curse of dimensionality like Principal Component Analysis (F.R.S., 1901) or Singular Value Decomposition (Eckart and Young, 1936) which approximates the variables by smaller number of combinations of original variables, but this approach is hard to interpret in the final model.

Sophisticated methods of attribute selection as Boruta algoritm (Kursa and Rudnicki, 2010), genetic algorithms (Kuhn and Johnson, 2013; Aziz et al., 2013) or simulated annealing techniques (Khachaturyan et al., 1981) are known and broadly used but in some cases for those algorithms computations can take even days, not to mention that datasets are growing every hour.

Few classification and regression models can reduce redundand variables during the training phase of statistical learning process, e.g. Decision Trees (Rokach and Maimon, 2008; Breiman et al., 1984), LASSO Regularized Generalized Linear Models (with cross-validation) (Friedman et al., 2010) or Regularized Support Vector Machine (Xu et al., 2009), but still computations starting with full set of explanatory variables are time consuming and the understaning of the feature selection procedure in this case is not simple and those methods are sometimes used without the understanding.

In business applications there appear a need to provide a fast feature selection that is extremely easy to understand. For such demands easy methods are prefered. This motivates using simple techniques like Entropy Based Feature Selection (Largeron et al., 2011), where every feature can be checked independently so that computations can be performed in a parallel to shorter the procedure's time. For this approach we provide an R interface to Rcpp reimplementation (Dirk Eddelbuettel, 2011) of methods included in **FSelector** package which we also extended with parallel background and sparse matrix support. This has significant impact on computations time and can be used on greater datasets, comparing to **FSelector**. Additionally we avoided the Weka (Hall et al., 2009) dependency and we provided faster discretization implementations than those from **entropy** package, used originally in **FSelector**.

Discretization

In statistical modelling, the **discretization** is the process of transferring continuous explanatory variables into discrete counterparts. The problems caused by categorization of continuous variables are known and widely spread (Harrell, 2015), but in some cases there appear an algorithmic requirement for the discretization. Moreover, there exist few algorithms, like decision trees (Salzberg, 1994), where continuous attributes are discretized during the learning process. Other reason for variable discretization include increasing the speed of induction algorithms (Catlett, 1991).

Even though many categorization algorithms have been developed (Holte, 1993; Chan et al., 1991), in this chapter we focus on a recursive entropy minimization heuristic for categorization coupled with a *Minimum Description Length* critetion (Rissanen, 1986) that controls the number of intervals produced over the continuous space. This is motivated by the original usage of this algorithm in **FSelector** package, which is our baseline. As Dougherty et al. (1995) showed better performance of classification for discretized feature set on real-world datasets and states that the described method was found promising by the authors not only for the local discretization but also for global discretization (Ting, 1994).

Entropy Dased Minimum Description Length Discretization Method

Overview of the method.

- Backward compatibility with **FSelector**

```
library(RWeka)
library(FSelectorRcpp)
RWeka::Discretize(Species~Sepal.Length, data = iris)[, 1] -> Rweka_disc_out
FSelectorRcpp::discretize(iris$Sepal.Length, iris$Species) ->FSelectorRcpp_disc_out
table(Rweka_disc_out,FSelectorRcpp_disc_out)
```

```
               FSelectorRcpp_disc_out
Rweka_disc_out  (-Inf;5.550000] (5.550000;6.150000] (6.150000;Inf)
  '(-inf-5.55]'              59                   0              0
  '(5.55-6.15]'               0                  36              0
  '(6.15-inf)'                0                   0             55
```

- Time comparison on small datasets

```
library(microbenchmark)
microbenchmark(discretize(iris$Sepal.Length, iris$Species),
               Discretize(Species~Sepal.Length, data = iris))
```

```
Unit: microseconds
                                          expr      min       lq      mean    median        uq
     discretize(iris$Sepal.Length, iris$Species)  135.303  243.881  265.1694  275.3655  300.3235
 Discretize(Species ~ Sepal.Length, data = iris) 5045.496 5549.937 6388.4559 5899.8145 7037.4495
       max neval
   410.577   100
 10792.033   100
```

- Comparison on big datasets
- Plot comparison

Entropy Based Feature Selection Algorithms

In the information theory the term **entropy** (Shannon, 2001) is

Stepwise Attribute Selection Evaluators

FSelectorRcpp and FSelector Computation Times Comparison

Conclusion

Acknowledgment

Bibliography

A. S. A. Aziz, A. T. Azar, M. A. Salama, A. E. Hassanien, and S. E. O. Hanfy. Genetic algorithms with different feature selection techniques for anomaly detectors generation. In M. P. M. Ganzha, L. Maciaszek, editor, *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, pages pages 769–774. IEEE, 2013. [p1]

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]

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984. new edition **?**? [p1]

J. Catlett. *On changing continuous attributes into ordered discrete attributes*, pages 164–178. Springer Berlin Heidelberg, Berlin, Heidelberg, 1991. ISBN 978-3-540-46308-5. doi: 10.1007/BFb0017012. URL http://dx.doi.org/10.1007/BFb0017012. [p2]

C. C. Chan, C. Batur, and A. Srinivasan. Determination of quantization intervals in rule based model for dynamic systems. In *Systems, Man, and Cybernetics, 1991. 'Decision Aiding for Complex Systems, Conference Proceedings., 1991 IEEE International Conference on*, pages 1719–1723 vol.3, Oct 1991. doi: 10.1109/ICSMC.1991.169942. [p2]

R. F. Dirk Eddelbuettel. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8): 1–18, 2011. URL http://www.jstatsoft.org/v40/i08/. [p1]

J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *MACHINE LEARNING: PROCEEDINGS OF THE TWELFTH INTERNATIONAL CONFERENCE*, pages 194–202. Morgan Kaufmann, 1995. [p2]

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1 (3):211–218, 1936. doi: 10.1007/BF02288367. [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]

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL http://www.jstatsoft.org/v33/i01/. [p1]

K. P. F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572, 1901. [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]

F. Harrell. General handouts and reference material on biostatistical modeling - problems caused by categorizing continuous variables, 2015. URL http://biostat.mc.vanderbilt.edu/wiki/Main/CatContinuous. [p2]

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. C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11(1):63–90, 1993. ISSN 1573-0565. doi: 10.1023/A:1022631118932. URL http://dx.doi.org/10.1023/A:1022631118932. [p2]

K. Hornik, C. Buchta, and A. Zeileis. Open-source machine learning: R meets weka. *Computational Statistics*, 24(2):225–232, 2009. ISSN 1613-9658. doi: 10.1007/s00180-008-0119-7. URL http://dx.doi.org/10.1007/s00180-008-0119-7. [p1]

G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *MACHINE LEARNING: PROCEEDINGS OF THE ELEVENTH INTERNATIONAL*, pages 121–129. Morgan Kaufmann, 1994. [p1]

A. Khachaturyan, S. Semenovsovskaya, and B. Vainshtein. The thermodynamic approach to the structure analysis of crystals. *Acta Crystallographica Section A*, 37(5):742–754, Sep 1981. [p1]

M. Kuhn and K. Johnson. Applied predictive modeling, 2013. URL http://www.amazon.com/Applied-Predictive-Modeling-Max-Kuhn/dp/1461468485/. [p1]

M. B. Kursa and W. R. Rudnicki. Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11):1–13, 2010. URL http://www.jstatsoft.org/v36/i11/. [p1]

C. Largeron, C. Moulin, and M. Géry. Entropy based feature selection for text categorization. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, SAC '11, pages 924–928, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0113-8. doi: 10.1145/1982185.1982389. URL http://doi.acm.org/10.1145/1982185.1982389. [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]

J. Rissanen. Stochastic complexity and modeling. *Ann. Statist.*, 14(3):1080–1100, 09 1986. doi: 10.1214/aos/1176350051. URL http://dx.doi.org/10.1214/aos/1176350051. [p2]

L. Rokach and O. Maimon. *Data Mining with Decision Trees: Theroy and Applications*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2008. ISBN 9789812771711, 9812771719. [p1]

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

S. L. Salzberg. C4.5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, 16(3):235–240, 1994. ISSN 1573-0565. doi: 10.1007/BF00993309. URL http://dx.doi.org/10.1007/BF00993309. [p2]

C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, Jan. 2001. ISSN 1559-1662. doi: 10.1145/584091.584093. URL http://doi.acm.org/10.1145/584091.584093. [p2]

K. M. Ting. Discretization of continuous-valued attributes and instance-based learning. Technical report, 1994. [p2]

H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *J. Mach. Learn. Res.*, 10:1485–1510, Dec. 2009. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1577069.1755834. [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