The Journal of Open Source Software

stabm: Stability Measures for Feature Selection

Andrea Bommert[1] **and Michel Lang**[1]

1 Faculty of Statistics, TU Dortmund University, 44221 Dortmund, Germany

DOI: 10.21105/joss.03010

Software

- Review
- Repository
- Archive

Editor: Marcos Vital

Reviewers:

- @mcavs
- @clauswilke
- @osorensen

Submitted: 09 December 2020
Published: 31 March 2021

Summary

The R (R Core Team, 2020) package *stabm* provides functionality for quantifying the similarity of two or more sets. For example, consider the two sets $\{A, B, C, D\}$ and $\{A, B, C, E\}$. Intuitively, these sets are quite similar because their overlap is large compared to the cardinality of the two sets. The R package *stabm* implements functions to express the similarity of sets by a real valued score. Quantifying the similarity of sets is useful for comparing sets of selected features. But also for many other tasks like similarity analyses of gene sets or text corpora, the R package *stabm* can be employed.

In the context of feature selection, the similarity of sets of selected features is assessed in order to determine the stability of a feature selection algorithm. The stability of a feature selection algorithm is defined as the robustness of the set of selected features towards different data sets from the same data generating distribution (Kalousis et al., 2007). For stability assessment, either m data sets from the same data generating process are available or m data sets are created from one data set. The latter is often achieved with subsampling or random perturbations (Awada et al., 2012). Then, the feature selection algorithm of interest is applied to each of the m data sets, resulting in m feature sets. To quantify the stability of the feature selection algorithm, the similarity of the m sets is calculated. In the context of feature selection stability, set similarity measures are called stability measures.

The R package *stabm* provides an open-source implementation of the 20 stability measures displayed in the table below. Argument checks are performed with checkmate (Lang, 2017) to provide helpful error messages. It is publicly available on CRAN and on Github and it has only a few dependencies.

Name	Reference
stabilityDavis	Davis et al. (2006)
stabilityDice	Dice (1945)
stabilityHamming	Dunne et al. (2002)
stabilityIntersectionCount	Bommert & Rahnenführer (2020)
stabilityIntersectionGreedy	Bommert & Rahnenführer (2020)
stabilityIntersectionMBM	Bommert & Rahnenführer (2020)
stabilityIntersectionMean	Bommert & Rahnenführer (2020)
stabilityJaccard	Jaccard (1901)
stabilityKappa	Carletta (1996)
stabilityLustgarten	Lustgarten et al. (2009)
stabilityNogueira	Nogueira et al. (2018)
stabilityNovovicova	Novovičová et al. (2009)
stabilityOchiai	Ochiai (1957)
stabilityPhi	Nogueira & Brown (2016)
stabilitySechidis	Sechidis et al. (2020)
stabilitySomol	Somol & Novovičová (2008)

Name	Reference
stabilityUnadjusted	Bommert & Rahnenführer (2020)
stabilityWald	Wald et al. (2013)
stabilityYu	Yu et al. (2012)
stabilityZucknick	Zucknick et al. (2008)

Statement of Need

The R package *stabm* provides an implementation of many stability measures. For theoretical and empirical comparative studies of the stability measures implemented in *stabm*, we refer to Bommert et al. (2017), Bommert & Rahnenführer (2020), Bommert (2020), and Nogueira et al. (2018). It has been demonstrated that considering the feature selection stability when fitting a predictive model often is beneficial for obtaining models with high predictive accuracy (Bommert et al., 2017; Bommert, 2020; Schirra et al., 2016). The stability measures implemented in the R package *stabm* have been employed in Bommert et al. (2017), Bommert et al. (2020), Bommert & Rahnenführer (2020), and Bommert (2020).

Related Software

A subset of the implemented stability measures is also available in other R or Python packages. The R package *sets* (Meyer & Hornik, 2009) and the Python package *scikit-learn* (Pedregosa et al., 2011) provide an implementation of the Jaccard index (Jaccard, 1901) to assess the similarity of two sets. The Python package *GSimPy* (Zhang & Cao, 2020) implements the Jaccard index, the Dice index (Dice, 1945), and the Ochiai index (Ochiai, 1957). The source code for the publication Nogueira et al. (2018) provides an implementation of their stability measure in R, Python, and Matlab.

Acknowledgements

This work was supported by the German Research Foundation (DFG), Project RA 870/7-1, and Collaborative Research Center SFB 876, A3.

References

Awada, W., Khoshgoftaar, T. M., Dittman, D., Wald, R., & Napolitano, A. (2012). A review of the stability of feature selection techniques for bioinformatics data. *2012 IEEE International Conference on Information Reuse and Integration*, 356–363. https://doi.org/10.1109/IRI.2012.6303031

Bommert, A. (2020). *Integration of feature selection stability in model fitting* [PhD thesis, TU Dortmund University, Germany]. https://doi.org/10.17877/DE290R-21906

Bommert, A., & Rahnenführer, J. (2020). Adjusted measures for feature selection stability for data sets with similar features. *Machine Learning, Optimization, and Data Science*, 203–214. https://doi.org/10.1007/978-3-030-64583-0_19

Bommert, A., Rahnenführer, J., & Lang, M. (2017). A multicriteria approach to find predictive and sparse models with stable feature selection for high-dimensional data. *Computational and Mathematical Methods in Medicine*, *2017*. https://doi.org/10.1155/2017/7907163

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, *143*. https://doi.org/10.1016/j.csda.2019.106839

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, *22*(2), 249–254.

Davis, C. A., Gerick, F., Hintermair, V., Friedel, C. C., Fundel, K., Küffner, R., & Zimmer, R. (2006). Reliable gene signatures for microarray classification: Assessment of stability and performance. *Bioinformatics*, *22*(19), 2356–2363. https://doi.org/10.1093/bioinformatics/btl400

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302. https://doi.org/10.2307/1932409

Dunne, K., Cunningham, P., & Azuaje, F. (2002). *Solutions to instability problems with sequential wrapper-based approaches to feature selection* [Technical Report]. Trinity College, Dublin, Ireland.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de La Société Vaudoise Des Sciences Naturelles*, *37*, 547–579. https://doi.org/10.5169/seals-266450

Kalousis, A., Prados, J., & Hilario, M. (2007). Stability of feature selection algorithms: A study on high-dimensional spaces. *Knowledge and Information Systems*, *12*(1), 95–116. https://doi.org/10.1007/s10115-006-0040-8

Lang, M. (2017). checkmate: Fast argument checks for defensive R programming. *The R Journal*, *9*(1), 437–445. https://doi.org/10.32614/RJ-2017-028

Lustgarten, J. L., Gopalakrishnan, V., & Visweswaran, S. (2009). Measuring stability of feature selection in biomedical datasets. *AMIA Annual Symposium Proceedings 2009*, 406–410.

Meyer, D., & Hornik, K. (2009). Generalized and customizable sets in R. *Journal of Statistical Software*, *31*(2), 1–27. https://doi.org/10.18637/jss.v031.i02

Nogueira, S., & Brown, G. (2016). Measuring the stability of feature selection. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 442–457. https://doi.org/10.1007/978-3-319-46227-1_28

Nogueira, S., Sechidis, K., & Brown, G. (2018). On the stability of feature selection algorithms. *Journal of Machine Learning Research*, *18*(174), 1–54. http://jmlr.org/papers/v18/17-514.html

Novovičová, J., Somol, P., & Pudil, P. (2009). A new measure of feature selection algorithms' stability. *2009 IEEE International Conference on Data Mining Workshops*, 382–387. https://doi.org/10.1109/icdmw.2009.32

Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bulletin of the Japanese Society for the Science of Fish*, *22*(9), 526–530. https://doi.org/10.2331/suisan.22.531

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, Édouard. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Schirra, L.-R., Lausser, L., & Kestler, H. A. (2016). Selection stability as a means of biomarker discovery in classification. In *Analysis of large and complex data* (pp. 79–89). Springer. https://doi.org/10.1007/978-3-319-25226-1_7

Sechidis, K., Papangelou, K., Nogueira, S., Weatherall, J., & Brown, G. (2020). On the stability of feature selection in the presence of feature correlations. In *Machine learning and knowledge discovery in databases* (pp. 327–342). Springer International Publishing. https://doi.org/10.1007/978-3-030-46150-8_20

Somol, P., & Novovičová, J. (2008). Evaluating the stability of feature selectors that optimize feature subset cardinality. *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition*, 956–966. https://doi.org/10.1109/tpami.2010.34

Wald, R., Khoshgoftaar, T. M., & Napolitano, A. (2013). Stability of filter- and wrapper-based feature subset selection. *2013 IEEE 25th International Conference on Tools with Artificial Intelligence*, 374–380. https://doi.org/10.1109/ictai.2013.63

Yu, L., Han, Y., & Berens, M. E. (2012). Stable gene selection from microarray data via sample weighting. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *9*(1), 262–272. https://doi.org/10.1109/tcbb.2011.47

Zhang, Y., & Cao, J. (2020). GSimPy: A python package for measuring group similarity. *SoftwareX*, *12*. https://doi.org/10.1016/j.softx.2020.100526

Zucknick, M., Richardson, S., & Stronach, E. A. (2008). Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. *Statistical Applications in Genetics and Molecular Biology*, *7*(1). https://doi.org/10.2202/1544-6115.1307