The Journal of Open Source Software

partition: A fast and flexible framework for data reduction in R

Malcolm Barrett[1] **and Joshua Millstein**[1]

1 Department of Preventive Medicine, University of Southern California

DOI: 10.21105/joss.01991

Software

- Review
- Repository
- Archive

Editor: Charlotte Soneson

Reviewers:

- @lmweber
- @clauswilke

Submitted: 20 December 2019
Published: 03 January 2020

Summary

Data are increasingly wider; in modern genomics, for example, high-resolution genetic data captures much more information than it did just a decade ago. While improved measurements contribute immensely to science, they also increase computational burden and complicate interpretability (Karczewski & Snyder, 2018). Data reduction techniques, such as principal components analysis and k-means clustering, are vital tools used to address these issues, particularly for noise and redundancy. However, these techniques may lead to problems in scalability, information loss, and interpretability (Malod-Dognin, Petschnigg, & Pržulj, 2018). The Partition framework is an approach to data reduction that is flexible, scalable, and interpretable, developed to address information loss while maintaining speed(Millstein et al., 2019).

The R(R Core Team, 2019) package partition is a fast and flexible data reduction tool that implements the Partition framework. partition enforces a minimum level of information, specified by the user, that reduced features must capture. Each feature begins as an independent cluster. Then, a bottom-up (agglomerative) approach grows clusters as much as possible, subject to the information loss constraint. Then, partition summarizes each feature in the subset into a single new feature. The reduced features are highly interpretable: original features map to one and only one feature in the reduced data set. The partition software is flexible, as well: how features are agglomerated, how information loss is measured, and the way data are reduced can all be customized.

partition uses an approach we call Direct-Measure-Reduce to modularize the Partition framework, facilitating the speed and flexibility of the data reduction process. These three components (directors, metrics, and reducers), collectively called partitioners, tell the partition algorithm (1) how to partition data, (2) how to measure information loss when reducing data, and (3) how to summarize partition subsets into reduced features, respectively. partition has several pre-specified partitioners for agglomerative data reduction, but this approach is also quite flexible.

The default partitioner uses a correlation-based distance matrix to find the pair of features with the smallest distance between them; intraclass correlation (ICC) to measure information explained by the reduced variable; and scaled row means to reduce variables with a sufficient minimum ICC. This approach is fast and reliable, but partition also implements other strategies for directing, measuring, and reducing. For instance, the user can use principal components analysis to reduce variables instead of scaled row means. As the framework is agnostic to how partitions are being directed, measured, or reduced, custom partitioners are easy to implement. Users may apply tools from base R or other R packages at any of the three stages of the partition.

While many tools exist in R for data reduction, including many functions in base R, partition offers a fast and flexible tool that addresses the need for interpretability and information

retention in reduced variables. To our knowledge, it is the first package of its kind. Wide data, such as high-resolution genetic data, can be quickly reduced without sacrificing information. Additionally, as each feature of the data maps to a single cluster, it is easier to make inferences from the reduced data set. The flexibility of partition makes it adaptable to the needs of many data reduction strategies.

Funding and Support

This work is supported by the National Cancer Institute (NCI), Award Number 5P01CA196569, and the National Institute on Aging (NIA), Award Number P01AG055367.

References

Karczewski, K. J., & Snyder, M. P. (2018). Integrative omics for health and disease. *Nature reviews. Genetics*, *19*(5), 299–310. doi:10.1038/nrg.2018.4

Malod-Dognin, N., Petschnigg, J., & Pržulj, N. (2018). Precision medicine a promising, yet challenging road lies ahead. *Current Opinion in Systems Biology*, *7*, 1–7. doi:https://doi.org/10.1016/j.coisb.2017.10.003

Millstein, J., Battaglin, F., Barrett, M., Cao, S., Zhang, W., Stintzing, S., Heinemann, V., et al. (2019). Partition: A surjective mapping approach for dimensionality reduction. *Bioinformatics*. doi:10.1093/bioinformatics/btz661

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/