The Journal of Open Source Software

popsom: A Very Efficient Implementation of Self-Organizing Maps with Starburst Visualizations for R

Lutz Hamel[1]

1 Department of Computer Science and Statistics, University of Rhode Island, Kingston, RI 02881

DOI: 10.21105/joss.03524

Software

- Review
- Repository
- Archive

Editor: Fabian Scheipl

Reviewers:

- @rcannood
- @HerrMo

Submitted: 14 July 2021
Published: 22 July 2021

Background

The self-organizing map (SOM) was developed by Teuvo Kohonen for data exploration and data visualization in the 1980s (Kohonen, 2001). It is an artificial neural network designed for unsupervised learning. What makes the self-organizing map so attractive are the intuitive mathematical underpinnings and the straightforward visualization of computational results (Ultsch, 1990). Self-organizing maps have been applied in virtually every scientific discipline where some sort of data exploration or analysis is necessary, e.g., (Liakos et al., 2018; Mathys et al., 2019; Matić et al., 2018; Miller & Coe, 1996). A number of R-packages exist that implement self-organizing maps including (Wehrens & Kruisselbrink, 2018) and (Yan, 2016).

Statement of need

Training a self-organizing map is time consuming. Here we introduce an R-package called `popsom` (Hamel et al., 2021) that implements a training algorithm for self-organizing maps based on vector and matrix operations inspired by tensor algebra (Hamel, 2018). We have measured speed ups of the training phase of a SOM of up to 60 times using our implementation over traditional implementations of the training algorithm such as (Yan, 2016). This speedup enables researchers to look at much larger data sets or to improve their throughput with a given data size.

Visualization of a trained map is at the core of using self-organizing maps. The `popsom` package improves on the standard u-matrix visualization for self-organizing maps (Ultsch, 1990) by superimposing starbursts in order to highlight cluster structures (Hamel & Brown, 2011).

Description

At a slightly more detailed level, our `popsom` package implements Kohonen's self-organizing maps with a number of distinguishing features:

1. A very efficient, single threaded, stochastic training algorithm based on ideas from tensor algebra. Up to 60x faster than traditional single-threaded training algorithms. No special accelerator hardware is required. The speedup results from the fact that the vector and matrix structures exposed by our algorithm map neatly into vector and matrix operations available on today's CPUs. Our Fortran 90 implementation insures that these vector and matrix operations are mapped onto the hardware as efficiently as possible (Hamel, 2018).

2. Automatic centroid visualization and detection using starbursts (Hamel & Brown, 2011). Not only does `popsom` display clusters and centroids on the map using starbursts but it also computes a cluster model similar to a k-means model based on the starbursts.
3. The `popsom` package maintains two models of the given training data: (a) a self-organizing map model where elements of the map model are available to the user for analysis, and (b) a centroid based clustering model similar to a k-means model where centroid and cluster information is available to the user. Having these two perspectives of a dataset is often helpful during a data analysis.
4. The package provides a number of easily accessible quality metrics for the self-organizing map and the centroid based cluster models (Hamel, 2016; Tatoian & Hamel, 2018). In particular, the package computes the `convergence` of a map which is a linear combination of the variance captured and the topographic fidelity of the map. A value close to 1 of this metric indicates a converged map. Furthermore, `popsom` also computes the `separation` of the clusters in a model. This is computed by the formula $1 - wcss/bcss$. In general, a value close to 1 here means well separated clusters.

Usage

popsom is available on CRAN and can be installed and loaded into an R session using,

```
> install.packages("popsom")
> library(popsom)
```

Binary packages for popsom are available from CRAN for macOS, Linux, and Windows. If you are on a system that is not supported by CRAN, you can download and compile the package from GitHub.

The following is a simple use-case for `popsom` exercising some of the functionality it has to offer,

```
> ## load a data set
> data(iris)
>
> ## set data frame and labels
> df <- subset(iris,select=-Species)
> labels <- subset(iris,select=Species)
>
> ## build a self-organizing map
> m <- map(df,labels,xdim=15,ydim=10,train=100000)
>
> ## compute a summary and display it
> summary(m)

Training Parameters:
  xdim ydim alpha  train normalize seed instances
    15   10   0.3 100000      TRUE NULL       150

Quality Assessments:
  convergence separation clusters
         0.99       0.98        5
>
> ## display a starburst plot of the map model
> starburst(m)
```

The Journal of Open Source Software

The last line of the R script generates the starburst visualization shown in Figure 1. A more involved usage example can be found on Kaggle.

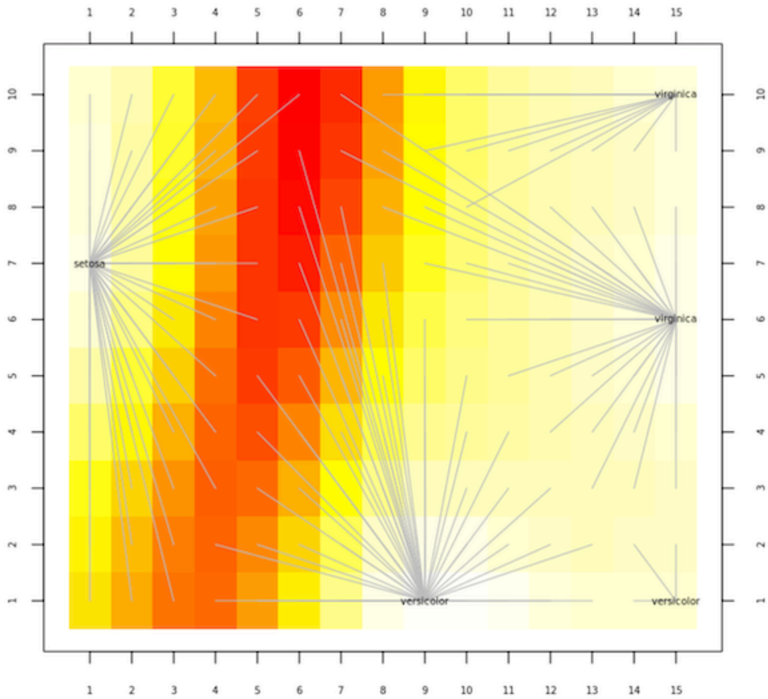

Figure 1: Starburst visualization of a self-organizing map.

Acknowledgements

We would like to thank the University of Rhode Island for financial support of this project through their 'Project Completion Grants.' A special thanks to the following people for their contributions to the project in no particular order: Benjamin Ott, Gregory Breard, Robert Tatoian, Michael Eiger, and Vishakh Gopu.

References

Hamel, L. (2018). VSOM: Efficient, stochastic self-organizing map training. *Proceedings of SAI Intelligent Systems Conference*, 805–821. https://doi.org/10.1007/978-3-030-01057-7_60

Hamel, L. (2016). Som quality measures: An efficient statistical approach. In *Advances in self-organizing maps and learning vector quantization* (pp. 49–59). Springer. https://doi.org/10.1007/978-3-319-28518-4_4

Hamel, L., & Brown, C. W. (2011). Improved interpretability of the unified distance matrix with connected components. *Proceedings of the International Conference on Data Science (ICDATA)*, 1.

Hamel, L., Ott, B., Breard, G., Tatoian, R., Eiger, M., & Gopu, V. (2021). *Popsom: A very efficient implementation of kohonen's self-organizing maps (SOMs) with starburst visualizations*. https://CRAN.R-project.org/package=popsom

Kohonen, T. (2001). *Self-organizing maps*. Springer Berlin.

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8), 2674.

Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., Menon, M., He, L., Abdurrob, F., Jiang, X., & others. (2019). Single-cell transcriptomic analysis of alzheimer's disease. *Nature*, *570*(7761), 332–337.

Matić, F., Kalinić, H., & Vilibić, I. (2018). Interpreting self-organizing map errors in the classification of ocean patterns. *Computers & Geosciences*, *119*, 9–17. https://doi.org/10.1016/j.cageo.2018.06.006

Miller, A., & Coe, M. (1996). Star/galaxy classification using kohonen self-organizing maps. *Monthly Notices of the Royal Astronomical Society*, *279*(1), 293–300. https://doi.org/10.1093/mnras/279.1.293

Tatoian, R., & Hamel, L. (2018). Self-organizing map convergence. *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET)*, *9*(2), 61–84.

Ultsch, A. (1990). Self-organizing feature maps for exploratory data analysis. *Proc. Of the International Neural Network Conference (INNC), 1990*.

Wehrens, R., & Kruisselbrink, J. (2018). Flexible self-organizing maps in kohonen 3.0. *Journal of Statistical Software*, *87*(7), 1–18. https://doi.org/10.18637/jss.v087.i07

Yan, J. (2016). *Som: Self-organizing map*. https://CRAN.R-project.org/package=som