Shape of data matters for non-Euclidean Self-Organizing Maps

Dorota Celińska-Kopczyńska[1]

1. Institute of Informatics, University of Warsaw

Self-Organizing Maps (SOMs, also known as Kohonen networks) belong to neural network models of the unsupervised class allowing for dimension reduction in data without a significant loss of information. SOMs preserve the underlying topology of high-dimensional input and transform the information into one or two-dimensional layer of neurons. The projection is nonlinear, and in the display, the clustering of the data space and the metric-topological relations of the data items are visible [Koh97]. In comparison to other techniques of reducing dimensionality, SOMs have many advantages. They do not impose any assumptions regarding the distributions of the variables and do not require independence among variables. They allow for solving non-linear problems; their applications are numerous, e.g., in pattern recognition (see, e.g., [GC91]), brain studies [BHC+93, RGC+97, PA13] or biological modeling [MVJB03, BNT+12]. At the same time, they are relatively easy to implement and modify [Koh97, AE12].

A typical setup for SOM assumes usage of a region of Euclidean plane. Most data analysts take it for granted to use some subregions of a flat space as their data model; however, by the assumption that the underlying geometry is non-Euclidean we obtain a new degree of freedom for the techniques that translate the similarities into spatial neighborhood relationships. Moreover, non-Euclidean geometries are steadily gaining attention of the data scientists [Was18, CM17]. In particular, hyperbolic geometry has been proven useful in data visualization [Mun98] and the modeling of scale-free networks [KPK+10, PKS+12]. Such a usefulness comes from the exponential growth property of hyperbolic geometry, which makes it much more appropriate than Euclidean for modeling and visualizing hierarchical data. Since the idea of SOM roots in geometry, we expect to gain new insights from non-Euclidean SOM setups. Surprisingly, there are nearly no attempts to do so. Even if there have been propositions to use hyperbolic geometry in SOMs [Rit99, OR01], other possibilites of inclusion of non-Euclidean geometries and different topologies (e.g., spherical geometry, quotient spaces) have been neglected. There is also no research on characteristics of data that affect the quality of Self-Organizing Maps.

Although we have numerous technical contributions on the algorithm itself, the presentation will focus on the quantitative analysis of the experimental data. We represent the data inside $\mathbb{R}^k$ as usual. Our algorithm aims to find a lower-dimensional representation, e.g., the data actually lies on a two-dimensional sphere in $\mathbb{R}^k$ and we want to recover this shape. The embedding manifold E represents this shape. In real-life applications, we usually do not know how the manifold E is embedded in $\mathbb{R}^k$, so finding out what SOM's template we should use is non-trival. Even if our core algorithm is used to find such a mapping, a wrong choice might come with a high cost of obtaining irrelevant results. That is why we analyze factors that might affect the quality of topology preservation. Our general experimental setup is as follows.

- We construct the original manifold O. Let T_O be the set of tiles and E_O be the set of edges between the tiles.
- We map all the tiles into the Euclidean space $m : T_O \to \mathbb{R}^d$, where d is the number of dimensions.
- We construct the target embedding manifold E. Let T_E be the set of tiles and E_E be the set of edges between the tiles.
- We apply our algorithm to the data given by m, This effectively yields an embedding $e : T_O \to E_O$.
- We measure the quality of the embedding.

To limit the effects of randomness (random initial weight of neurons, random ordering of data) we apply this process independently 100 times for every pair of manifolds E and O. Our quantitative analyses based on OLS and Tobit regressions show that the shape of data matters for Self-Organizing Maps. We use measures of topology preservation from the literature, as well as our own measures.

Joint work with Eryk Kopczyński. Work supported by the National Science Centre, Poland, grant UMO-2019/35/B/ST6/04456.

References

[AE12] Umut Asan and Secil Ercan. An introduction to self-organizing maps. In C. Kahraman, editor, *Computational Intelligence Systems in Industrial Engineering: with Recent Theory and Applications*, pages 299–319. Atlantis Press, 2012.

[BHC+93] James C Bezdek, LO Hall, L_P Clarke, et al. Review of mr image segmentation techniques using pattern recognition. *MEDICAL PHYSICS-LANCASTER PA-*, 20:1033–1033, 1993.

[BNT+12] P. Boniecki, K. Nowakowski, R. Tomczak, S. Kujawa, and H. Piekarska-Boniecka. The application of the Kohonen neural network in the nonparametric-quality-based classification of tomatoes. In Mohamed Othman, Sukumar Senthilkumar, and Xie Yi, editors, *Fourth International Conference on Digital Image Processing (ICDIP 2012)*, volume 8334, pages 440 – 444. International Society for Optics and Photonics, SPIE, 2012.

[CM17] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data sceintists. Technical report, 2017.

[GC91] Stephen Grossberg and Gail A. Carpenter. *Pattern recognition by self-organizing neural networks / edited by Gail A. Carpenter and Stephen Grossberg.* MIT Press Cambridge, Mass, 1991.

[HHG20] Allison Marie Horst, Alison Presmanes Hill, and Kristen B Gorman. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*, 2020. R package version 0.1.0.

[Koh97] Teuvo Kohonen, editor. *Self-organizing Maps.* Springer-Verlag, Berlin, Heidelberg, 1997.

[KPK+10] Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguñá. Hyperbolic geometry of complex networks. *Phys. Rev. E*, 82:036106, Sep 2010.

[Mun98] Tamara Munzner. Exploring large graphs in 3d hyperbolic space. *IEEE Computer Graphics and Applications*, 18(4):18–23, 1998.

[MVJB03] Paolo Mazzatorta, Marjan Vracko, Aneta Jezierska, and Emilio Benfenati. Modeling toxicity by using supervised kohonen neural networks. *Journal of chemical information and computer sciences*, 43(2):485–492, 2003.

[OR01] Jörg Ontrup and Helge Ritter. Hyperbolic self-organizing maps for semantic navigation. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pages 1417–1424, Cambridge, MA, USA, 2001. MIT Press.

[PA13] Ricardo Pérez-Aguila. Enhancing brain tissue segmentation and image classification via 1d kohonen networks and discrete compactness: An experimental study. *Engineering Letters*, 21(4), 2013.

[PKS+12] Fragkiskos Papadopoulos, Maksim Kitsak, M. Angeles Serrano, Marian Boguñá, and Dmitri Krioukov. Popularity versus Similarity in Growing Networks. *Nature*, 489:537–540, Sep 2012.

[RGC+97] W. E. Reddick, J. O. Glass, E. N. Cook, T. D. Elkin, and R. J. Deaton. Automated segmentation and classification of multispectral magnetic resonance images of brain using artificial neural networks. *IEEE Transactions on Medical Imaging*, 16(6):911–918, 1997.

[Rit99] Helge Ritter. Self-organizing maps on non-euclidean spaces. In E. Oja and S. Kaski, editors, *Kohonen Maps*, pages 97–108. Elsevier, 1999.

[Was18] Larry Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5(1):501–532, 2018.