machine learning & knowledge extraction

MDPI

Review

Why Topology for Machine Learning and Knowledge Extraction?

Massimo Ferri [1,2,†]

[1] Department of Mathematics, University of Bologna, 40126 Bologna, Italy; massimo.ferri@unibo.it
[2] ARCES, University of Bologna, 40125 Bologna, Italy
† Current address: Department of Mathematics, University of Bologna, 40126 Bologna, Italy.

Received: 10 March 2018; Accepted: 30 April 2018; Published: 2 May 2018

Abstract: Data has shape, and shape is the domain of geometry and in particular of its "free" part, called topology. The aim of this paper is twofold. First, it provides a brief overview of applications of topology to machine learning and knowledge extraction, as well as the motivations thereof. Furthermore, this paper is aimed at promoting cross-talk between the theoretical and applied domains of topology and machine learning research. Such interactions can be beneficial for both the generation of novel theoretical tools and finding cutting-edge practical applications.

Keywords: shape; geometry; topological data analysis; persistence

1. Introduction

Data has shape in more than one sense: Macroscopically, a dataset may itself have a shape, and this is generally of capital importance for analyzing it. At a lower granularity level, each element of a dataset may have a shape, often in a scarcely formalized way. Both machine learning and knowledge extraction, then, need to understand and take advantage of either type of data shape. Moreover, in the human-machine interaction, visualization plays an important role. All these issues require a smart use of geometry and, more and more often, of that "free" branch of it which is topology. There are literally thousands of texts (articles, proceedings, books) on Topological Data Analysis (TDA). Here I will mention very few keystones, examples and comprehensive surveys.

2. The Shape of Datasets and Networks

There are different ways of thinking of the shape of a dataset. For a long time we have used Mahalanobis and Bhattacharyya distances to take into account the shape of clusters present in data. All the remarkable success of statistical learning by Support Vector Machines (SVM) is due to the consideration that, in many practical situations, if a separator of different populations in a dataset exists, it does not have a simple (i.e., linear) shape [1] (see [2], co-authored by the topologist and Fields medalist S. Smale, for a different viewpoint on similar themes).

How do relevant data embed into the space of all possible occurrences? This problem is raised in a paper co-authored by Fields medalist D. Mumford [3] and solved in a surprising way [4]: In the space of all possible 3×3 pixel patches in a digital image, consider the ones with high contrast; this is a finite subset of a finite set, so it seems impossible to sensibly talk about its shape. Still, the authors show that it is the discretization of a Klein bottle; this is made possible by persistent homology (see Section 3).

It has been well known for a long time that the topology of a space is strictly connected with the behaviour of a continuous real function defined on it (i.e., it is connected with the indices of its critical points). This is the core of classical *Morse Theory* [5] and is a key idea in TDA, for instance in `Mapper` [6] and in persistent homology (see Section 3).

A common problem is that data can be described by a high number of variables, but the dataset X can be intrinsically low-dimensional. The "true" dimension of X is easily discovered by *Principal Component Analysis*, provided that data lie in a linear subspace. Otherwise `Mapper` can come to help. Its ingredients are: A *parameter space* Z, an open covering $\mathcal{U}$ of it, a continuous function f (called *filter*) from the metric space where the data is represented to Z. Then `Mapper` builds a simplicial complex as follows: Vertices are the data clusters contained in $f^{-1}(U)$ for each open set $U \in \mathcal{U}$, and there is a k-simplex with vertices $v_0, \ldots, v_k$ whenever the intersection of the clusters represented by these vertices is nonempty. Then you get a "topological summary" of the dataset which makes it much easier to deal with it. Of course, the choices of $Z, \mathcal{U}$ and f influence the dimension of the resulting complex and the resolution of the representation. `Mapper` is widely used in companies working at data analysis, and has a lot of scientific applications: From network security to RNA sequencing to epidemiology, to mention a few [7–13].

UMAP [14] is another, recent system for reducing the dimension of datasets.

While the idea of the shape, in a topological sense, of a dataset is fairly new, topology is commonly associated with networks.

A way to think of the shape of a network is to analyze whether it has similar features to a small-world network [15]. A different problem is the coverage of a given domain by a network of sensors [16]; this already provides an example of use of topological persistence, which I will treat later. There is a sort of nonvisible shape in a neural network: The datum of its weights. While learning, a neural network modifies its own "shape" up to a stable one. This shape stabilization interacts with the (more) visible topological structure of the network itself as a graph, which may be—so to say—non-Euclidean; this is the subject of "geometric deep learning" which makes use of Fourier theory, spectral graph theory and much more [17].

3. The Shape of a Data Element

What does a machine learn? How can one extract knowledge from a dataset? Quite often it is a matter of estimating similarity in a broad sense: Classifying a music piece, retrieving images, ascribing a text to an author require to determine either a distance or a relation among objects, or both. In this task, topology is extremely powerful. By its very nature, it is a method for formalizing qualitative aspects of reality. So it is no wonder that TDA enjoyed a terrific growth in recent years.

When speaking of TDA, one often refers to Persistent Homology. Homology is a branch of algebraic topology, a discipline which assigns algebraic invariants to topological spaces in such a way that two topologically equivalent (*homeomorphic*) spaces get the same invariants (the converse is unfortunately not true in general); in particular, homology is the class of invariants which can best be computed for practical uses [18]. What about *persistent* homology? It comes from the following class of problems.

Imagine that there is an object of interest, subset of a metric (e.g., Euclidean) space, of which you only have a finite cloud of samples. This is the common situation of a digitalized image, of a 3D mesh, but also what you have in the aforementioned cases of a network of sensors, or of the pixel patches. How can you guess the (algebraic) topology of the original object out of these samples? One first idea is to build a continuous object out of the point cloud, hoping that what you get is a good approximation of the original object. One way to do that is by centering a ball of fixed radius on each sample point. There are some smart techniques based on this idea: e.g., the construction of Vietoris-Rips, Čech, alpha complexes [19]. Out of these constructions we can compute topological invariants; typically they are the dimensions of the homology modules at various dimensions k, called *Betti numbers*. Substantially, they count the numbers of *k-cycles*, i.e. connected components and voids in the object. There is a problem: The Betti numbers that one obtains depend on the radius of the balls. While varying the radius, there may be k-cycles which *persist*, and a good guess is that those may correspond to the true k-cycles of the sampled object [20,21].

In a more formal and general way, instead of studying only a topological space X, persistence studies pairs (X, f), where f is a continuous real function defined on X (*filtering function*). Then, for each pair of reals $u < v$, the *k-th persistent Betti number* function counts how many k-cycles present in the sublevel set under u (i.e., in the set $\{x \in X \,|\, f(x) \leq u\}$) survive in the sublevel set under v [22,23]. These functions are usually summarized by *persistence diagrams* or *barcodes* and yield lower bounds for a *natural pseudodistance* between such pairs [24]. There are several derived structures—e.g., *zigzag diagrams, landscapes, vineyards, extended persistent diagrams*—and a compelling algebraic setting, *persistence modules*. An important development is the use of a different range than $\mathbb{R}$ for the filtering function; in particular, $\mathbb{R}^n$ as a range is very tempting for applications but poses hard problems [25,26]. The stability of these representations is the object of particular attention [27–29]; this (and much more) is thoroughly covered for the 1-dimensional range in [30].

The freedom of choice of the filtering function and the generality of the setting grant a great modularity to this tool; it was clear already when applying the historical predecessor of persistent homology, *Size Functions*, to classification problems [31,32]. This flexibility has been widely exploited in the analysis of data of natural origin [33]. The multiset nature of persistence diagrams and barcodes represents a problem and a challenge for them to be input to a machine learning system. A solution, roughly said, is to substitute cornerpoints with Gaussian kernels [34–37]; this idea was proposed long before, in rudimentary form, for size functions [38,39]. The idea of associating different pairs (X, f) to the same data is part of a general philosophy of inserting the observer into the observed phenomenon, well expressed in [40].

Implementations of persistent homology algorithms are covered in [41]. Python implementations of TDA are surveyed in [42]. A *multiscale mapper* combining `Mapper` with the ideas of persistent homology has been recently developed [43]. Exporting the structure of persistence diagrams beyond topology is the main goal of [44].

4. Visualization for the Human-In-The-Loop Paradigm

There are several reasons for an automatic system to be—at least for the moment—just a smart assistant of a human operator in a number of tasks. This is particularly the case of biomedical applications, where technology offers invaluable support, but the final word is still the competence of the physician. If a machine has the advantage of speed and tirelessness, the human expert has the unbeaten capability of learning concepts with few examples and of evaluating problems and solutions in non-formalized environments. These different skills risk becoming drawbacks if they remain separated; they can, on the contrary, enhance each other if they integrate together in the human-in-the-loop paradigm [45]. This is what happens, e.g., in *Active Learning*, where continual interaction between human and machine optimizes the trade-off between what an algorithm can offer and what the human actually looks for [46]. This can take the simple form of relevance feedback in data retrieval, or of smarter systems where the machine poses queries to the operator, or even exploits peculiarities of human psychology to minimize time and cognitive burden. Interaction then becomes unavoidable when data exploration goals are ill-defined or evolve over time.

A structural difficulty consists in the fact that a learning machine usually works in a very high-dimensional space, what is hardly something a human can deal with. On the other hand, human experts are highly skilled in extracting knowledge from datasets of dimension ≤ 3. This is why visualization is a keystone of human-machine interaction, bringing information down to the sensorial domain of a human user. Data exploration and analysis, but also data presentation can greatly benefit from it. A typical example is the representation of relations by a graph, where additional information can be conveyed by size, color and shape of vertices and edges. (A peculiar example is the one of *crystallizations*, edge-colored graphs by which all piecewise-linear manifolds of any dimension can be completely represented, with applications in pure topology and, recently, in theoretical physics [47].) Selection, modification, and hypothesis testing thereof then become easier. In general, it is necessary to build algorithms that are able to provide representations of data interpretable by humans (see

the already quoted [9,10,14] for some). This is part of a large project [48,49] in which geometry and particularly topology play a relevant role.

5. Conclusions

A lot of work has been done, applying topology to machine learning and knowledge extraction, but much more awaits the competence and imagination of experts from both sides. I passionately hope that more topologists discover the challenges, suggestions, and application chances coming from this domain, but I also invite researchers from computer science, artificial intelligence and even robotics [50] to add topology to their toolboxes.

Acknowledgments: Work performed under the auspices of INdAM-GNSAGA. I wish to thank Mattia G. Bergomi for the very helpful discussions and the Reviewers for the many helpful corrections and suggestions.

Conflicts of Interest: The author declares no conflict of interest. The funding sponsor had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Byun, H.; Lee, S.W. A survey on pattern recognition applications of support vector machines. *Int. J. Pattern Recognit. Artif. Intell.* **2003**, *17*, 459–486. [CrossRef]
2. Cucker, F.; Smale, S. On the mathematical foundations of learning. *Bull. Am. Math. Soc.* **2002**, *39*, 1–49. [CrossRef]
3. Lee, A.B.; Pedersen, K.S.; Mumford, D. The nonlinear statistics of high-contrast patches in natural images. *Int. J. Comput. Vis.* **2003**, *54*, 83–103. [CrossRef]
4. Carlsson, G.; Ishkhanov, T.; de Silva, V.; Zomorodian, A. On the local behavior of spaces of natural images. *Int. J. Comput. Vis.* **2008**, *76*, 1–12. [CrossRef]
5. Knudson, K.P. *Morse Theory: Smooth and Discrete*; World Scientific Publishing Company: Singpore, 2015.
6. Singh, G.; Mémoli, F.; Carlsson, G.E. Topological methods for the analysis of high dimensional data sets and 3D object recognition. In *SPBG 2017*; The Eurographics Association: Geneva, Switzerland, 2007; pp. 91–100.
7. Nicolau, M.; Levine, A.J.; Carlsson, G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 7265–7270. [CrossRef] [PubMed]
8. Lum, P.Y.; Singh, G.; Lehman, A.; Ishkanov, T.; Vejdemo-Johansson, M.; Alagappan, M.; Carlsson, J.; Carlsson, G. Extracting insights from the shape of complex data using topology. *Sci. Rep.* **2013**, *3*, 1236. [CrossRef] [PubMed]
9. Coudriau, M.; Lahmadi, A.; François, J. Topological analysis and visualisation of network monitoring data: Darknet case study. In Proceedings of the 2016 IEEE International Workshop on Information Forensics and Security (WIFS), Abu Dhabi, UAE, 4–7 December 2016; pp. 1–6.
10. Guo, W.; Banerjee, A.G. Toward automated prediction of manufacturing productivity based on feature selection using topological data analysis. In Proceedings of the 2016 IEEE International Symposium on Assembly and Manufacturing (ISAM), Fort Worth, TX, USA, 21–22 August 2016; pp. 31–36.
11. Rizvi, A.H.; Camara, P.G.; Kandror, E.K.; Roberts, T.J.; Schieren, I.; Maniatis, T.; Rabadan, R. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nat. Biotechnol.* **2017**, *35*, 551–560. [CrossRef] [PubMed]
12. Feged-Rivadeneira, A.; Angel, A.; González-Casabianca, F.; Rivera, C. Malaria intensity in Colombia by regions and populations. *arXiv* **2017**, arXiv:1710.00317.
13. Saggar, M.; Sporns, O.; Gonzalez-Castillo, J.; Bandettini, P.A.; Carlsson, G.; Glover, G.; Reiss, A.L. Towards a new approach to reveal dynamical organization of the brain using topological data analysis. *Nat. Commun.* **2018**, *9*, 1399. [CrossRef] [PubMed]
14. McInnes, L.; Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2018**, arXiv:1802.03426.
15. Strogatz, S.H. Exploring complex networks. *Nature* **2001**, *410*, 268–276. [CrossRef] [PubMed]
16. De Silva, V.; Ghrist, R. Homological sensor networks. *Not. Am. Math. Soc.* **2007**, *54*, 1.

17. Bronstein, M.M.; Bruna, J.; LeCun, Y.; Szlam, A.; Vandergheynst, P. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Process. Mag.* **2017**, *34*, 18–42. [CrossRef]
18. Kaczynski, T.; Mischaikow, K.; Mrozek, M. *Computational Homology*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006; Volume 157.
19. Edelsbrunner, H.; Harer, J. *Computational Topology: An Introduction*; American Mathematical Society: Providence, RI, USA, 2010.
20. Carlsson, G. Topology and data. *Bull. Am. Math. Soc.* **2009**, *46*, 255–308. [CrossRef]
21. Carlsson, G. Topological pattern recognition for point cloud data. *Acta Numer.* **2014**, *23*, 289–368. [CrossRef]
22. Edelsbrunner, H.; Harer, J. Persistent homology-a survey. *Contemp. Math.* **2008**, *453*, 257–282.
23. Ghrist, R. Barcodes: the persistent topology of data. *Bull. Am. Math. Soc.* **2008**, *45*, 61–75. [CrossRef]
24. Donatini, P.; Frosini, P. Natural pseudodistances between closed manifolds. *Forum Math.* **2004**, *16*, 695–715. [CrossRef]
25. Carlsson, G.; Zomorodian, A. The theory of multidimensional persistence. *Discret. Comput. Geom.* **2009**, *42*, 71–93. [CrossRef]
26. Cagliari, F.; Di Fabio, B.; Ferri, M. One-dimensional reduction of multidimensional persistent homology. *Proc. Am. Math. Soc.* **2010**, *138*, 3003–3017. [CrossRef]
27. Cohen-Steiner, D.; Edelsbrunner, H.; Harer, J. Stability of persistence diagrams. In Proceedings of the 21th Annual Symposium on Computational Geometry, Pisa, Italy, 6–8 June 2005; Mitchell, J.S.B., Rote, G., Eds.; ACM: New York, NY, USA, 2005; pp. 263–271.
28. Chazal, F.; Cohen-Steiner, D.; Glisse, M.; Guibas, L.J.; Oudot, S.Y. Proximity of persistence modules and their diagrams. In Proceedings of the 25th Annual Symposium on Computational Geometry, Aarhus, Denmark, 8–10 June 2009; ACM: New York, NY, USA, 2009; pp. 237–246.
29. Cerri, A.; Di Fabio, B.; Ferri, M.; Frosini, P.; Landi, C. Betti numbers in multidimensional persistent homology are stable functions. *Math. Methods Appl. Sci.* **2013**, *36*, 1543–1557. [CrossRef]
30. Chazal, F.; de Silva, V.; Glisse, M.; Oudot, S. *The Structure and Stability of Persistence Modules*; Springer: Berlin/Heidelberg, Germany, 2016.
31. Verri, A.; Uras, C.; Frosini, P.; Ferri, M. On the use of size functions for shape analysis. *Biol. Cybern.* **1993**, *70*, 99–107. [CrossRef]
32. Biasotti, S.; Cerri, A.; Frosini, P.; Giorgi, D.; Landi, C. Multidimensional size functions for shape comparison. *J. Math. Imaging Vis.* **2008**, *32*, 161–179. [CrossRef]
33. Ferri, M. Persistent topology for natural data analysis—A survey. In *Towards Integrative Machine Learning and Knowledge Extraction*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 117–133.
34. Bubenik, P. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* **2015**, 16, 77–102.
35. Reininghaus, J.; Huber, S.; Bauer, U.; Kwitt, R. A stable multi-scale kernel for topological machine learning. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4741–4748.
36. Kusano, G.; Hiraoka, Y.; Fukumizu, K. Persistence weighted Gaussian kernel for topological data analysis. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 2004–2013.
37. Hofer, C.; Kwitt, R.; Niethammer, M.; Uhl, A. Deep Learning with Topological Signatures. In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 1633–1643.
38. Ferri, M.; Frosini, P.; Lovato, A.; Zambelli, C. Point selection: A new comparison scheme for size functions (With an application to monogram recognition). In *Computer Vision — ACCV'98. ACCV 1998*; Chin, R., Pong, T.C., Eds.; Springer: Berlin/Heidelberg, Germany, 1998; pp. 329–337.
39. Donatini, P.; Frosini, P.; Lovato, A. Size functions for signature recognition. In Proceedings of the International Symposium on Optical Science, Engineering, and Instrumentation, San Diego, CA, USA, 2 October 1998; Vision Geometry VII, Volume 3454, pp. 178–184.
40. Frosini, P. G-invariant persistent homology. *Math. Methods Appl. Sci.* **2015**, *38*, 1190–1199. [CrossRef]
41. Otter, N.; Porter, M.A.; Tillmann, U.; Grindrod, P.; Harrington, H.A. A roadmap for the computation of persistent homology. *EPJ Data Sci.* **2017**, *6*, 17. [CrossRef]

42. Ray, J.; Trovati, M. A Survey of Topological Data Analysis (TDA) Methods Implemented in Python. In *Advances in Intelligent Networking and Collaborative Systems. INCoS 2017*; Barolli, L., Woungang, I., Hussain, O., Eds.; Springer: Cham, Switzerland, 2017; pp. 594–600.
43. Dey, T.K.; Mémoli, F.; Wang, Y. Multiscale mapper: Topological summarization via codomain covers. In Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, Arlington, VA, USA, 10–12 January 2016; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2016; pp. 997–1013.
44. Bergomi, M.G.; Ferri, M.; Zuffi, L. Graph persistence. *arXiv* **2017**, arXiv:1707.09670.
45. Holzinger, A. Human-Computer Interaction and Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In *Availability, Reliability, and Security in Information Systems and HCI. CD-ARES 2013;* Cuzzocrea, A., Kittl, C., Simos, D.E., Weippl, E., Xu, L., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 319–328.
46. Holzinger, A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Inform.* **2016**, *3*, 119–131. [CrossRef] [PubMed]
47. Casali, M.R.; Cristofori, P.; Dartois, S.; Grasselli, L. Topology in colored tensor models via crystallization theory. *J. Geom. Phys.* **2018**, *129*, 142–167. [CrossRef]
48. Otasek, D.; Pastrello, C.; Holzinger, A.; Jurisica, I. Visual data mining: effective exploration of the biological universe. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*; Holzinger, A., Jurisica, I., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 19–33.
49. Turkay, C.; Jeanquartier, F.; Holzinger, A.; Hauser, H. On computationally-enhanced visual analysis of heterogeneous data and its application in biomedical informatics. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*; Holzinger, A., Jurisica, I., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 117–140.
50. Farber, M. *Invitation to Topological Robotics*; European Mathematical Society: Zurich, Switzerland, 2008; Volume 8.