The Power of Topological Data Analysis for Machine Learning

Charles Zhang

Abstract

Topological Data Analysis (TDA) applies techniques from algebraic topology to study and extract topological and geometric information on the shape of data [1]. By considering geometric and topological features of multi-dimensional data arising from various distance metrics imposed on the data, complex relationships within the data can be preserved and jointly considered. This often leads to better results than using standard analytical tools. This project then aims at harnessing the power of TDA for machine learning. We can combine and compare a wide range of TDA techniques to extract features from images that are usually used separately. This project will introduce several topological preliminary backgrounds, especially cubical complexes and their persistent homology which are much more natural for images with 'cube' pixels. Then, for the first part, I discover the shape of the data by the mapper algorithm[2]. Visualising and understanding very high dimensional datasets is vital for further understanding more about the nature, fundamental structure and underlying relationships of the data and preparing the data for machine learning algorithms as the preprocess for classifying tasks. Second, This project presents a more general way to use TDA for machine learning tasks on grayscale images by applying various TDA techniques [3]. Specifically, this project applies persistent homology to generate a wide range of topological features using a point cloud obtained from an image, its natural grayscale filtration, and different filtrations defined on the binarized image. We show that this topological machine learning pipeline can be used as a highly relevant dimensionality reduction by applying it to the MNIST digits dataset. For the result, we can observe that the trained classifier can classify digit images while reducing the size of the feature set by more than half in comparison with the grayscale pixel value features and maintain similar even higher accuracy as 97%. The digital artifact of this project can be seen ***HERE***(version without codes ***HERE***) where the GitHub repository can be seen ***HERE***.

Keywords: Topological Data Analysis, Machine Learning, Cubical Complex, Mapper, Persistent Homology

Bibliography

[1] G. Carlsson, “Topology and data,” Bull. Amer. Math. Soc. (N.S.), vol. 46, no. 2, pp. 255–308, 2009.

[2] Singh, Gurjeet, Facundo Mémoli, and Gunnar E. Carlsson. "Topological methods for the analysis of high dimensional data sets and 3d object recognition." SPBG 91 (2007): 100.

[3] Garin, Adélie, and Guillaume Tauzin. "A Topological" Reading" Lesson: Classification of MNIST using TDA." 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). IEEE, 2019.

[4] H. Edelsbrunner and J. Harer, “Persistent homology—a survey,” in Surveys on discrete and computational geometry, ser. Contemp. Math. Amer. Math. Soc., Providence, RI, 2008, vol. 453, pp. 257–282.

[5] Ziou, Djemel, and Madjid Allili. "Generating cubical complexes from image data and computation of the Euler number." Pattern Recognition 35, no. 12 (2002): 2833-2839.

[6] Allili, Madjid, Konstantin Mischaikow, and Allen Tannenbaum. "Cubical homology and the topological classification of 2D and 3D imagery." In Proceedings 2001 international conference on image processing (Cat. No. 01CH37205), vol. 2, pp. 173-176. IEEE, 2001.

[7] V. Robins, P. Wood, and A. P Sheppard, “Theory and algorithms for constructing discrete Morse complexes from grayscale digital images,” IEEE transactions on pattern analysis and machine intelligence, vol. 33, May 2011.

[8] D. M. Boyer, S. Mukherjee, and K. Turner, “Persistent homology transform for modeling shapes and surfaces,” Information and Inference: A Journal of the IMA, vol. 3, no. 4, pp. 310–344, Dec 2014.

[9] L. Kanari, P. Dłotko, M. Scolamiero, R. Levi, J. Shillcock, K. Hess, and H. Markram, “A topological representation of branching neuronal morphologies,” Neuroinformatics, vol. 16, no. 1, pp. 3–13, Jan 2018.

[10] J. Reininghaus, S. Huber, U. Bauer, and R. Kwitt, “A stable multi-scale kernel for topological machine learning,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 4741– 4748.

[11] N. Atienza, L. M. Escudero, M. J. Jimenez, and M. Soriano-Trigueros, “Persistent entropy: a scale-invariant topological statistic for analyzing cell arrangements,” 2019.

[12] Ho, Tin Kam. "Random decision forests." In Proceedings of 3rd international conference on document analysis and recognition, vol. 1, pp. 278-282. IEEE, 1995.

[13] vneogi199, Handwritten-Digit-Recognition-Using-Random-Forest, accuracy: 94.2%, https://github.com/vneogi199/Handwritten-Digit-Recognition-Using-Random-Forest

[14] Bernard, Simon, Sébastien Adam, and Laurent Heutte. "Using random forests for handwritten digit recognition." In Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 2, pp. 1043-1047. IEEE, 2007.