ERDC/ITL SR-21-5

US Army Corps
of Engineers®
Engineer Research and
Development Center

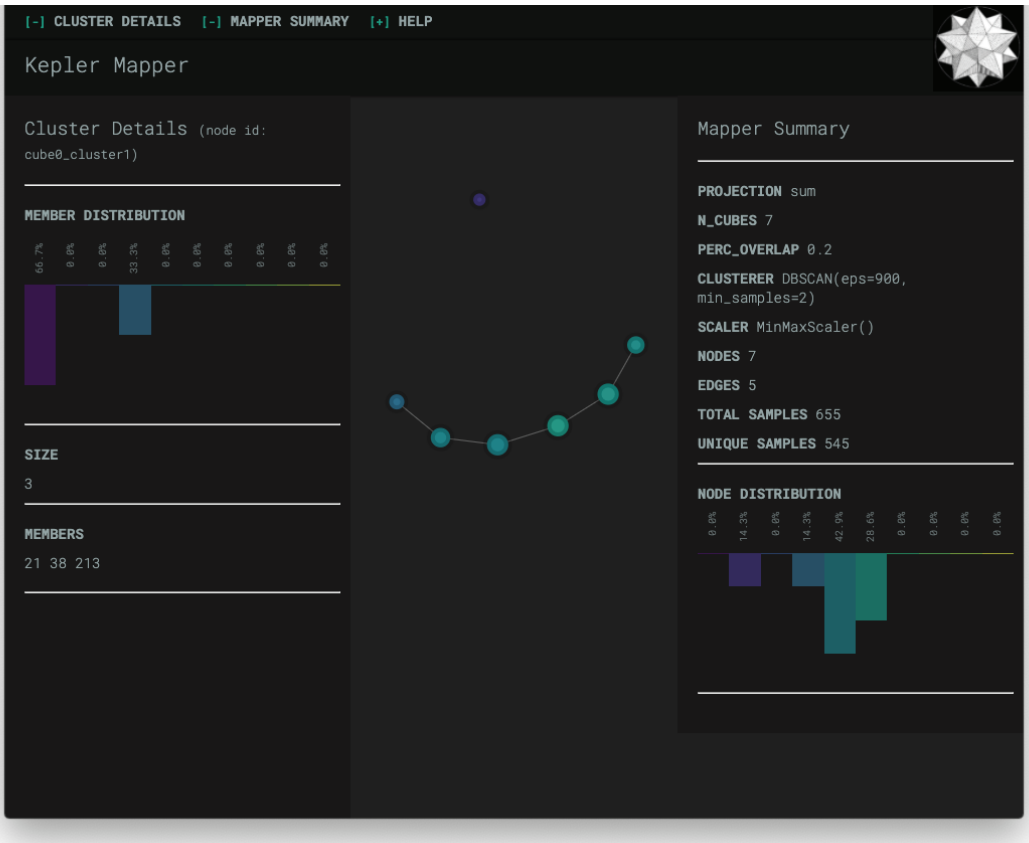
ERDC
ENGINEER RESEARCH & DEVELOPMENT CENTER

Topological Data Analysis

An Overview

Amy E. W. Bednar

June 2021



Information Technology Laboratory

The US Army Engineer Research and Development Center (ERDC) solves the nation's toughest engineering and environmental challenges. ERDC develops innovative solutions in civil and military engineering, geospatial sciences, water resources, and environmental sciences for the Army, the Department of Defense, civilian agencies, and our nation's public good. Find out more at www.erdc.usace.army.mil.

To search for other technical reports published by ERDC, visit the ERDC online library at http://acwc.sdp.sirsi.net/client/default.

ERDC/ITL SR-21-5
June 2021

Topological Data Analysis

An Overview

Amy E. W. Bednar

Information Technology Laboratory
US Army Engineer Research and Development Center
3909 Halls Ferry Road
Vicksburg, MS 39180-6199

Final report

Prepared for Office of the Secretary of Defense

Under Project 403 – Engineered Resilient Systems
Program Element: 0603833D8Z - Engineering Science and Technology (S&T)

Abstract

A growing area of mathematics topological data analysis (TDA) uses fundamental concepts of topology to analyze complex, high-dimensional data. A topological network represents the data, and the TDA uses the network to analyze the shape of the data and identify features in the network that correspond to patterns in the data. These patterns extract knowledge from the data. TDA provides a framework to advance machine learning's ability to understand and analyze large, complex data. This paper provides background information about TDA, TDA applications for large data sets, and details related to the investigation and implementation of existing tools and environments.

Contents

Preface

This study was conducted for the Office of the Secretary of Defense under Project Program Element: 0603833D8Z, “Engineering Science and Technology (S&T).” The technical monitor was Dr. Owen Eslinger.

The work was performed by the Computational Analysis Branch of the Computational Science and Engineering Division, US Army Engineer Research and Development Center, Information Technology Laboratory (ERDC-ITL). At the time of publication, Dr. Jeffrey Hensley was Branch Chief; Dr. Benjamin Parsons was Acting Division Chief; and Dr. Robert Wallace, was Technical Director for the Engineered Resilient Systems (ERD) program. The Deputy Director of ERDC-ITL was Ms. Patti S. Duett, and the Director was Dr. David Horner.

COL Teresa Schlosser was the Commander of ERDC, and Dr. David Pittman was the Director.

1 Introduction

1.1 Background

Topological data analysis (TDA) represents data with a topological network that creates nodes, or groups of data, and edges, or relationships between the data. TDA makes sense of large arrays, and the key feature of TDA is the nature of the model or output it produces. Typically, methods ask very specific questions about large data sets. However, TDA uses the network to examine the shape of the data. This examination allows researchers to identify features in the network that correspond to patterns in the data, therefore extracting additional knowledge from the data. Thus, TDA reduces the possibility of missing critical insights by reducing the dependency on machine learning experts to choose the right algorithms, because TDA advances machine learning's native capability to analyze these large, complex data sets. It uses current machine learning techniques as input to find subtle patterns and insights.

Social networking sites, military vehicle maintenance, and even healthcare records increasingly use large, multivariable data sets. The Army has both a large amount and wide variety of data. Using TDA has a major advantage: it minimizes initial bias. It gives data scientists the ability to examine any raw data set and determine which factors actually influence the data. Normally, a subject matter expert asks a specific question about a data set, which can bias the data. TDA does not begin with a question. Rather, it records patterns in the data set without any preconceived notions. This approach provides testable hypotheses from the shape of the data set alone.

1.2 Objective

This report will first describe TDA in more detail and will provide useful definitions in topology to help the reader understand TDA techniques. The report will then describe which practical applications have successfully used TDA. Finally, the report will describe current technologies available for TDA.

1.3 Useful definitions in topology

Topology is the study of geometric properties and spatial relations unaffected by the continuous change of shape or size of figures. Coordinate invariance, deformation invariance, and compressed representation are three important topological properties. Coordinate invariance means the properties remain the same and do not depend on the coordinate system used. Deformation invariance means properties remain the same as the shape is stretched or deformed or both. Compressed representation means that a shape with infinite points can be represented by a similar shape with fewer points assuming some details can be overlooked. For example, a circle can be represented by a hexagon.

Some definitions in topology that apply to TDA can help clarify concepts described in some of the TDA applications but are not meant to be an all-inclusive list of important topological or mathematical terms for TDA. TDA divides the data into regions where we can easily notice features and then classify those features in order to gain insights about the data.

Homology or Betti numbers help describe shapes. Betti numbers distinguish connectivity of n-dimensional simplicial complexes. A simplicial complex is a set of composed points, line segments, triangles, and their n-dimensional counterparts. The Betti number counts the independent loops for each n-dimensional space. The p^{th} persistent homology groups are the images of the homomorphisms induced by inclusion, noted $H_p^{i,j}$ for $0 \leq i \leq j \leq n-1$. The corresponding p^{th} persistent Betti numbers are the ranks of these groups, $\beta_p^{i,j} = \text{rank}(H_p^{i,j})$. Critical points are associated with the birth and death of persistent homology groups and can be visually represented by a persistence diagram. (Tierny, 2006, p. 22–23).

A Reeb Graph also shows important information about data. Let $f: \mathcal{M} \to \mathbb{R}$ be a piecewise linear (PL) Morse scalar field defined on a compact PL manifold $\mathcal{M}$. Let $f^{-1}(f(p_1))_p$ be the contour of f containing the point $p \in \mathcal{M}$. The Reeb graph $\mathcal{R}(f)$ is a one-dimensional simplicial complex defined as the quotient space on $\mathcal{M} \times \mathbb{R}$ by the equivalence relation $\left(p_1, f((p_1)\right) \sim \left(p_2, f((p_2)\right)$, which holds if

$$\begin{cases} f(p_1) = f(p_2) \\ p_2 \in \left(f^{-1}\left(f(p_1)\right)\right)_{(p_1)} \end{cases} \quad (1)$$

1.4 Mapper algorithm

The mapper algorithm, developed by Gurjeet Singh, Facundo Memoli, and Gunnar Carlsson (Singh, 1991, p. 1-2), creates a topological network as a foundation for TDA analysis. The basic steps are

1. Choose a distance metric, like Euclidean, Hamming, or cosine. This metric captures similarity between data points.
2. Compute the lens or filter functions. This computation maps data points to single values on the real number line and is based on, for example, raw features, statistics, geometry, and machine learning algorithm outputs.
3. Apply cover and overlap to determine the connections for the network. The cover is the resolution, which is a numeric value for how many intervals (for example, 7 intervals, 20 intervals) should be used. The overlap is the degree of overlap between intervals (for example, 20%).
4. Compute Cartesians to determine points in common. Perform Cartesian products of the range intervals and assign the original data points to the resulting two-dimensional regions according to their filter values.
5. Perform clustering. Each cluster is a node, and if two nodes share a point, then there is an edge between them. It doesn't really matter which kind of clustering used. It should be based on the type of data being analyzed.
6. Build the TDA network by determining a color scheme to capture localized behavior and derive hidden insights from the data (Singh, 1991, p. 1-2).

Few recommendations or best practices for choosing specific covers and lenses for the data currently exist. Obviously, these choices will be based on the specific data and the quantity of the data and is a trial-and-error process. Liu, Xie, and Yi developed a faster algorithm to overcome this; however, the algorithm is limited by storage in memory and computing time.

2 TDA Applications

The following section will provide an overview of some examples of TDA used successfully in several different disciplines.

2.1 Breast cancer

TDA identified subgroups in breast cancer survival, relapse, and fatality rates that point to future avenues of research. Lum et. al used TDA to group patients in more detail than standard clustering methods. They also identified small groups that may be "important for targeted therapy" (Lum, 2013, p. 3). The researchers used two older data sets, NKI and GSE2034, and applied two filter functions, L-infinity centrality and survival.

2.2 Sports

Lum et al. also successfully used TDA to show 13 scoring positions in basketball versus the normal 5 positions of point guard, shooting guard, small forward, power forward, and center. They considered factors such as minutes played, minutes played, number of rebounds, assists, turnovers, steals, blocked shots, personal fouls, and points scored.

2.3 Voting trends in the House of Representatives

Lum et al. explored voting trends in the US House of Representatives from 1990 to 2011. The study showed that most members voted along party lines. However, the data from 2009 to 2011 showed more diversity in voting behavior. Principal component analysis did not show these trends.

2.4 Zebrafish

TDA has analyzed and predicted skin patterns on zebrafish. This enables quantitative predictions in patterns that occur in wild-type and mutant zebrafish. This work can be expanded for large-scale analysis of biological data. McGuirl, Volkening, and Sandstede uses TDA methods and "interpretable machine learning for quantifying both agent-level features and global pattern attributes on a large scale" for zebrafish (McGuirl, 2020, p. 5113). Before the use of TDA, a researcher might use manual inspection or smoothing algorithms to find patterns in a data set. These methods take time and could accidentally discard relevant data. The

researchers used persistent homology so that they would not have to use labeled training data. As a result, the researchers were able to create

> an automated, interpretable framework for counting stripes and spots, detecting broken stripes, measuring stripe widths, quantifying stripe straightness, calculating spot size and roundness, measuring spot placement, and estimating the onset of stripe formation from pattern data. (McGuirl, 2020, p. 5114)

2.5 Time-series and Internet of things

TDA has shown better results than single and multi-attribute techniques for time series data. Diaz et. al showed TDA performed better on classifying incomplete and noisy Internet-of-things (IOT) data. These data consisted of audio speakers, video cameras, doorbells, fitbits, game controllers, IOT hubs, lights, 3D printers, Roombas, routers, environmental sensors, switches, tablets, televisions, television dongles, weather stations, and other devices with a total of 183 devices studied over 9 months. This team used government-owned software, Time Series Analysis Tool (TSAT), to perform some of the analysis (Diaz, 2019, p. 1543, 1548).

2.6 Darknet

Coudriau used the Mapper algorithm to visualize network monitoring for cybersecurity for darknet data. Security analysts quickly analyzed large numbers of IP packets and showed patterns that were missed by Suricata, a popular intrusion detection system (Coudriau, 2016, p. 1-6).

3 Current Technologies

Many open-source technologies demonstrate TDA principals. Below are a few investigated during the course of this project. Most of them focus on using the Mapper algorithm.

3.1 Python Mapper

Python Mapper is an open-source software whose main focus is the Mapper algorithm. It provides a graphical user interface (GUI), which contains filter functions and visualization of the network, and is released under the GNU's not Unix! (GNU) General Public License version 3 (GPLv3) license. Python Mapper requires Python 2.6 or higher. The GUI needs Python 2, which reached end of life January 2020. Python Mapper has not been updated since April 19, 2017.

3.2 Topology ToolKit

The Topology ToolKit (TTK) is an open-source toolkit that aides in TDA and provides visualization using ParaView as a front end. The developers offer a virtual box consisting of TTK and ParaView preloaded with examples from their website. The interface shows the topological graph along with persistent graphs. The website was updated with a tutorial in May 2020, and the code was updated March 2020. TTK is licensed under the Berkeley Software Distribution (BSD).

3.3 Gudhi Library

Gudhi is an open-source library for computational topology and TDA with the following capabilities: various types of simplicial complexes, topological descriptors computation, manifold reconstruction, and topological descriptors tools such as persistence diagrams and barcode. Gudhi is licensed under Massachusetts Institute of Technology (MIT), but many of the modules have dependencies licensed under GPLv3 and Lesser General Public License (LGPL). This code was last updated September 2020.

3.4 KeplerMapper

KeplerMapper is an open-source library that shows the topological network built using the mapper algorithm whose main Mapper workflow is to "to project the data, group the image, apply clustering to the preimage

of the groups, and then build a simplicial complex" (https://kepler-mapper.scikit-tda.org/started.html). KeplerMapper is licensed under MIT and was last updated September 2020.

3.5 Symphony Ayasdi

Symphony Ayasdi is a private company founded by Gunnar Carlsson who was the lead investigator on the Defense Advanced Research Projects Agency (DARPA) "Topological Data Analysis" project from 2005 to 2010. They have developed their own software and have used TDA in several different areas from financial crime to healthcare. They have partnered with several universities around the world including The University of California, University of North Carolina Chapel Hill, Karolinksa Institutet, and many others. (www.ayasdi.com)

3.6 IBM

International Business Machines (IBM) is conducting research in the field of TDA, with a focus on persistent homology, and the Computation Genomics Group is pursuing TDA in their genomics research. This group has received a National Science Foundation award and written several papers. They have not made clear what software they are using to calculate the topology in the data (https://researcher.watson.ibm.com/researcher/view_group.php?id=6585).

4 Conclusions

TDA is a capability and technology that ERDC should continue to research as part of ERDC's data analytics capabilities. TDA allows examining large data sets and is not constrained by choice of metrics. Topological networks compress information in high-dimensional data. Thus, no information is lost, which provides a way to make sure critical insights are not missed due to researchers discarding data. I would like to thank ERS RDA for the funding to perform this TDA research.

References

Carlsson, G. 2009. "Topology and Data". *Bull. Amer. Math. Soc.* 46: 255-308.

Coudriau, M., A. Lahmadi, and J. Francois. 2016. "Topological Analysis and Visualisation of Network Monitoring Data: Darknet case study," 2016 IEEE International Workshop on Information Forensics and Security (WIFS).

Diaz, C., M. S. Postol, R. Simon, and D. Wicke. 2019. "Time-series Data Analysis for Classification of Noisy and Incomplete Internet-of-Things Data sets," 2019 18th IEEE International Conference of Machine Learning and Applications.

KeplerMapper 1.2.0 documentation. "Getting Started, " https://kepler-mapper.scikit-tda.org/started.html, accessed July 12, 2020.

Liu, XU, Z. Xie, and D. Yi. 2012. "A Fast Algorithm for Constructing Topological Structure in Large Data." *Homology, Homotopy and Applications*, 14(1).

Lum, P. Y., G. Singh, A. Lehman, T. Ishkanov, Vejdemo-Johansson, M. Algappan, J. Carlsson, and G. Carlsson. 2013. "Extracting Insights from the Shape of Complex Data Using Topology." Scientific Reports 3, Article number:1236.

McGuirl, M. R., A. Volkening, and B. Sandstede. 2020. "Topological Data Analysis of Zebrafish Patterns." In *Proceedings of the National Academy of Sciences*, 117(10): 5113-5124. (https://www.pnas.org/content/117/10/5113)

Singh, G., F. Memoli, and G. Carlsson. 1991. "Mapper: a topological mapping tool for point cloud data". *Eurographics symposium on point-based graphics*.

Singh, G., F. Memoli, and G. Carlsson. 2007. "Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition." *Eurographics Symposium on Pont-Based Graphics*.

Tierny, J. 2006. "Introduction to Topological Data Analysis," *Sorbonne Universities*, UPMC Univ Paris 06.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
June 2021	Final	

4. TITLE AND SUBTITLE

Topological Data Analysis: An Overview

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
0603833D8Z

6. AUTHOR(S)

Amy E. W. Bednar

5d. PROJECT NUMBER
403

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Information Technology Laboratory
US Army Engineer Research and Development Center
3909 Halls Ferry Road
Vicksburg, MS 39180-6199

8. PERFORMING ORGANIZATION REPORT NUMBER

ERDC/ITL SR-21-5

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of the Secretary of Defense

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Project Data Analytics, "Topological Data Analysis"

14. ABSTRACT

A growing area of mathematics topological data analysis (TDA) uses fundamental concepts of topology to analyze complex, high-dimensional data. A topological network represents the data, and the TDA uses the network to analyze the shape of the data and identify features in the network that correspond to patterns in the data. These patterns extract knowledge from the data. TDA provides a framework to advance machine learning's ability to understand and analyze large, complex data. This paper provides background information about TDA, TDA applications for large data sets, and details related to the investigation and implementation of existing tools and environments.

15. SUBJECT TERMS

Topology – Electronic data processing
Big data
Machine learning
High-performance computing

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
Unclassified	Unclassified	Unclassified	SAR	17	