Topological Data Analysis

Anthony Tan

McMaster University

tana4@mcmaster.ca

April 2019

Overview

1. Introduction
2. Techniques and Applications
3. Comments
4. References

What is Topological Data Analysis?

- Topological Data Analysis (TDA) is the analysis of data using geometric and topological methodology.
- Topology is the study of shapes and spaces that undergo continuous deformations without tearing it.

Figure: A coffee mug and donut are “topologically equivalent”.

- A popular technique is to construct topological spaces out of data sets and examine its structure.

A Pinch of Topology I

Unless otherwise stated, a space X will always refer to a topological space.

- Topology generalizes metric spaces through set theoretic constructs to define a notion of proximity.
- Classification of shapes or spaces are conducted by tracking properties that remain unchanged under transformations, or invariants. Some invariants include connectedness, path connectedness, compactness, and more.
- Among these invariants, we will focus on connectedness.

A Pinch of Topology II

Let X be a space.

Definition (Connected)

A **separation** of X is a pair of non-empty open subsets U and V of X such that $U \cap V = \emptyset$ and $X = U \cup V$. We say that X is **connected** if there does not exist a separation. Otherwise, we say it is **disconnected**.

Definition (Path Connected)

We say X is **path connected** if, for all $x, y \in X$, there exists $\gamma \colon [0,1] \to X$ such that $\gamma(0) = x$ and $\gamma(1) = y$.

Proposition

If X is path connected, then it is connected.

Preview of the Main Attractions

- The goal of statistical learning is to understand given data.
- We will describe two techniques of TDA that share similarities with clustering.
- The goal of clustering is to partition a data set into subgroups, or clusters, where the members are related to each other.
- Connectivity information of a space is similar to clustering in the sense that we are examining how the space is organized.
- A space may be written as a union of its “clusters”, or more formally, (connected) components.

Persistent Homology I

Let $X \subset \mathbb{R}^n$ be a finite set. We regard X as our data set, or point cloud.

Sketch

1. For each $x \in X$, enclose it by an (open) ball of centered at x with radius ϵ. We denote this by $B_\epsilon(x) := \{x' \in X \mid d(x', x) \leq \epsilon\}$. Note that $d(-,-)$ is the usual Euclidean metric.
2. Let $\epsilon \to \infty$. For $x, y \in X$, join the two points by an (undirected) edge whenever $B_\epsilon(x) \cap B_\epsilon(y) \neq \emptyset$. This creates a skeleton-like object known as a **simplicial complex**, which is a topological space!
3. Apply **homology** to the resulting space to **algebraically** determine its connectivity information.
4. Record the connectivity information in the form of **barcodes** to analyze the **persistence**.

Persistent Homology II

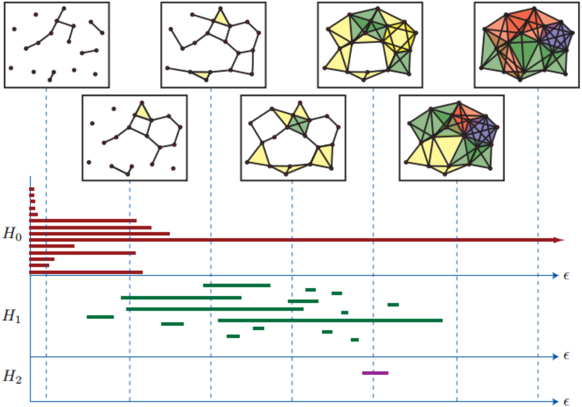

Figure: Construction of the Vietoris-Rips complex and its corresponding barcodes as $\epsilon \to \infty$. (Retrieved from Ghrist 2009.)

Persistent Homology III

We mentioned in the previous slide the words **homology** and **algebra**. I owe you all an explanation...

- Homology is an algebraic method used to detect i-dimensional holes. We call i the degree.
- Thanks to linear algebra, the number of holes in dimension i can be computed. These are called **Betti numbers**, denoted by β_i.

For example, the following degrees count

- $i = 0$: path-connected components,
- $i = 1$: loops,
- $i = 2$: voids.

The Mapper Algorithm

Let $X \subset \mathbb{R}^n$ be a point cloud.

Sketch

1. Choose (carefully) a continuous function (called a **filter**) $f : X \to Z$, where Z is a parameter space. For simplicity, we often have that $Z = \mathbb{R}$. We may also take on $\mathbb{R}^m$ or $S^1 := \{x \in \mathbb{R}^2 \mid ||x|| = 1\}$.
2. Let $\mathcal{U}$ be any finite **cover** of Z. Upon taking the **pre-image** of the elements of $\mathcal{U}$, we obtain a finite cover of X. This is denoted by $\mathcal{V}$.
3. The elements of $\mathcal{V}$ are "partial clusters". Organize these clusters using **any** clustering algorithm.
4. If these partial clusters "overlap", join them by an edge. This creates easy-to-visualize simplicial complex.

Let's look at a couple examples.

Visualing the Shape of the Normal Distribution

Input

- Let $f : [-M, M] \to [0, \infty)$ where $(x, y) \mapsto y$.
- Take $\mathcal{U} = \{[0,5), (4,10), (9,15), (14,\infty)\}$ be a covering of $[0, \infty)$.

Output

- $f^{-1}([0,5))$ consists of one partial cluster.
- Otherwise, every other pre-image consists of two partial clusters.
- Obtain a heat-map-like simplicial complex as seen in the figure below.

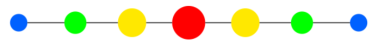

Figure: Retrieved from Singh et. al. 2007.

Redefining the Positions of Basketball (1/3)

- There are 5 positions in basketball: point guard (PG), shooting guard (SG), small forward (SF), power forward (PF), and center (C).
- Each position has a traditional skills to learn.
- Versatility is key to the most elite players. That is, they have mastered the skills of at least 2 positions.
- Elite teams may have specific play styles.

Observation

Basketball team compositions, including NBA teams, do not revolve around these “positions” anymore.

Question

How do we determine if a player is a good fit for a team?

Redefining the Positions of Basketball (2/3)

- Muthu Alagappan, a former student at Stanford University, conducted a study on NBA player data during his internship at Ayasdi, a data science company utilizing TDA.
- The sample size was 452 NBA players, each with 7 associated normalized statistics.

Results

- A graph of 13 "clusters" was constructed. In particular, these are the redefined positions.
- Player strengths are easier to identify.
- Team compositions are easier to identify.

Redefining the Positions of Basketball (3/3)

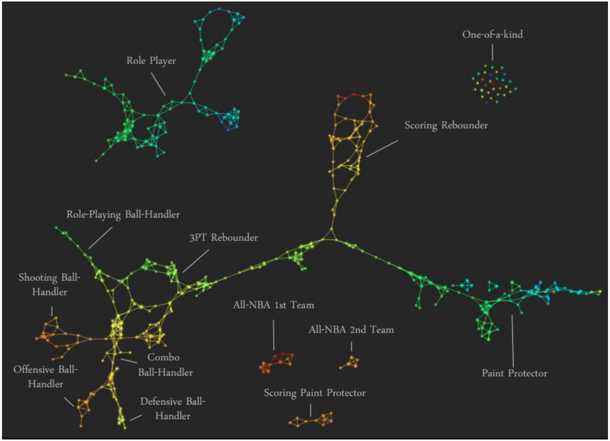

Figure: 2010 - 2011 NBA Player Data (Retrieved from Alagappan 2012).

Some Comments

In the words of Robert Ghrist:

> “Homological methods are, almost by definition, robust, relying on neither precise coordinates nor careful estimates for efficacy. As such, they are most useful in settings where geometric precision fails. With great robustness comes both great flexibility and great weakness.”

- It is still unclear whether or not there is a “go-to” application for the described TDA methods.
- There are underlying problems with the methodology. For instance, density is an issue for both methods.

References

M. Alagappan.
From 5 to 13: Redefining the Positions in Basketball.
MIT Sloan Sports Analytics Conference, 2012. Date Accessed: April 2019.

G. Carlsson.
Topology and Data.
Bulletin of the American Mathematical Society 46, 255-308, 2009.

R. Ghrist.
Homological Algebra and Data.
The Mathematics of Data, IAS/Park City Mathematics Volume 25, 273-325, 2017.

G. James, D. Witten, T. Hastie, and R. Tibshirani.
An Introduction to Statistical Learning.
Springer New York, 2014.

References

R. Ghrist.
Barcodes: The Persistent Topology of Data.
Bull. Amer. Math. Soc., 45(1) 61-75, 2007.

A. Hatcher.
Algebraic Topology.
Cambridge University Press, 2001.

J. Munkres.
Topology.
Prentice-Hall, Incorporated, Second Edition, 2000.

G. Singh, F. Mémoli, and G. Carlsson.
Topological Methods for the Analysis of High Dimensional Data Sets and 3D Recognition.
Eurographics Symposium on Point-Based Graphics, 2007.

Questions?

Thank you for listening!