Introduction to Topological Data Analysis

Ali Al-Barkawi, Montader Alasady

University of Utah

***Abstract*—Topological Data Analysis (TDA) is an emerging field of mathematics that combines topology with data analysis to understand complex, high-dimensional datasets. Unlike traditional statistical methods, TDA identifies patterns through clusters, holes, voids, and connections in data. This paper provides an introduction to TDA, its applications across various domains, key terminology, and interpretation of topological features in data analysis.**

***Index Terms*—topological data analysis, persistent homology, simplicial complex, high-dimensional data, pattern recognition**

I. Introduction

Topological Data Analysis (TDA) is a field of mathematics that combines topology—the study of properties of geometric objects—to graph and map complex datasets. Unlike conventional statistical methods, TDA uses pattern identification to reveal the underlying structure of data. These patterns are composed of clusters, holes, voids, and connections that provide insights into the data's inherent geometry.

TDA is particularly valuable when dealing with specific types of data. Two key characteristics indicate when TDA may be appropriate: (1) high-dimensional data, meaning data with more than three dimensions or features, and (2) noisy data, which contains errors, randomness, or information that might skew results in classical statistical analysis.

II. Applications of TDA

A. Biology and Medicine

TDA has demonstrated significant impact in biological and medical research. In cancer subtyping, TDA has been used to discover previously unknown subtypes of breast cancer by uncovering loops and clusters in gene expression data [1]. In neuroscience, brain connectome data, which is inherently high-dimensional, has been analyzed using TDA to help identify differences in neurological disorders [2].

B. Machine Learning

In the machine learning domain, TDA contributes to feature extraction through persistent homology, which produces topological features that improve classification accuracy when searching large databases [3]. Additionally, TDA provides tools for understanding neural networks by studying the structure of loss landscapes and analyzing hidden-layer activations.

C. Finance and Economics

Financial applications of TDA include market regime detection, where loops may correspond to repeating financial cycles. Anomaly detection is another important application, as holes or voids in data can signal abnormal behaviors in the market.

D. Physics and Materials Science

TDA has been applied to shape analysis of molecular structures, identifying voids or tunnels in materials that relate to properties such as strength and conductivity [4].

III. Key Terminology

A. Topology

Topology is a branch of mathematics concerned with properties of shapes that remain invariant under continuous deformations such as stretching or bending, but not tearing.

B. Point Cloud

A point cloud is a set of data points in a high-dimensional space, often serving as the raw input for TDA methods.

C. Simplicial Complex

A simplicial complex is a mathematical structure composed of points, line segments, triangles, and higher-dimensional analogs used to approximate the shape of data.

D. Filtration

Filtration is a process of building simplicial complexes at multiple scales by gradually increasing a distance threshold, allowing for multi-scale analysis of data.

E. Homology

Homology is a mathematical method for counting topological features such as connected components, loops, and voids within a space.

F. Persistent Homology

Persistent homology tracks how long topological features exist across different scales of the filtration. Features that persist across many scales are typically meaningful, while short-lived features often represent noise.

G. Barcode Diagram

A barcode diagram is a visual summary of persistent homology where each bar represents a topological feature, and its length indicates the persistence of that feature across scales.

H. Persistence Diagram

A persistence diagram plots feature birth versus death scale. Points positioned far from the diagonal represent strong, important features in the data.

I. Betti Numbers

Betti numbers are numerical invariants that count topological features:

- β_0: number of connected components
- β_1: number of loops (1-dimensional holes)
- β_2: number of voids (2-dimensional holes)

IV. Interpretation of Topological Features

A. Clusters (Connected Components – β_0)

Clusters represent groups of data points that are close together in the feature space. They indicate distinct categories or subpopulations within the dataset, such as different patient groups, customer segments, or behavioral modes.

B. Loops (1-Dimensional Holes – β_1)

Loops appear when data forms ring-like or cyclical structures. They often signal periodic or repeating behavior, such as circadian rhythms or financial cycles, or transition paths between states that return to the starting point. For example, a loop in wearable sensor data might indicate a person's daily activity cycle.

C. Voids or Cavities (2-Dimensional Holes – β_2)

Voids appear as hollow three-dimensional structures in higher-dimensional spaces. They can indicate missing data regions, multi-way cyclical interactions, or hollow structural features in molecules or materials. For instance, a void in molecular simulation data may correspond to a molecular pocket important for drug binding.

D. Higher-Dimensional Holes

Though difficult to visualize, persistent homology can detect holes in any dimension. These often represent complex relationships, high-dimensional cycles, or structural constraints. Higher-dimensional features are especially important in biology, physics, and network science.

V. Conclusion

Topological Data Analysis provides a powerful framework for understanding the structure of complex, high-dimensional datasets. By focusing on geometric properties—clusters, loops, and voids—TDA reveals patterns that traditional methods often overlook. Its applications span diverse fields including science, finance, medicine, and machine learning, making it an increasingly essential tool for modern data analysis. As computational methods continue to advance, TDA's role in extracting meaningful insights from complex data will likely continue to expand.

References

[1] M. Nicolau, A. J. Levine, and G. Carlsson, "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 108, no. 17, pp. 7265–7270, Apr. 2011, doi: 10.1073/pnas.1102826108.

[2] F. H. Xu, M. Gao, J. Chen, S. Garai, D. A. Duong-Tran, Y. Zhao, and L. Shen, "Topology-based clustering of functional brain networks in an Alzheimer's Disease cohort," *AMIA Jt. Summits Transl. Sci. Proc.*, pp. 449–458, May 2024.

[3] F. Hensel, M. Moor, and B. Rieck, "A survey of topological machine learning methods," *Front. Artif. Intell.*, vol. 4, 2021, doi: 10.3389/frai.2021.681108.

[4] S. Broderick, R. Dongol, T. Zhang *et al.*, "Classification of apatite structures via topological data analysis: A framework for a 'Materials Barcode' representation of structure maps," *Sci. Rep.*, vol. 11, art. 11599, 2021, doi: 10.1038/s41598-021-90070-4.

Acknowledgment — The author used ChatGPT (OpenAI) solely for grammar and wording suggestions. All technical content, analysis, and conclusions are the author's own.