Article

Topological Data Analysis: Merging Algebraic Topology with Machine Learning

Topic: Computational Topology

Written on: 13 May 2025

Keywords: Topological Data Analysis, Machine Learning, Algebraic Topology, Persistent Homology, Feature Extraction

Shamit Fatin, Kahlert School of Computing, University of Utah
Email, Homepage, Orcid

Topological Data Analysis (TDA) offers a novel framework for understanding the intrinsic geometric and topological structures within complex datasets. By leveraging concepts from algebraic topology, TDA provides tools such as persistent homology to capture multi-scale features that are often elusive to traditional statistical methods. This paper explores the integration of TDA into machine learning pipelines, highlighting how topological insights can enhance model robustness, interpretability, and performance, especially in high-dimensional and noisy data environments. We discuss foundational concepts, practical implementations, and potential applications, aiming to bridge the gap between abstract mathematical theory and practical machine learning solutions.

The exponential growth of data in various domains necessitates advanced analytical tools capable of uncovering hidden structures and patterns. Traditional machine learning techniques, while powerful, often struggle with high-dimensional, noisy, or incomplete data. Topological Data Analysis (TDA) emerges as a promising approach, utilizing principles from algebraic topology to study the ”shape” of data. By focusing on the connectivity and continuity of data points, TDA provides a lens to capture global features that are invariant under continuous transformations, offering robustness against noise and perturbations. This paper delves into the synergy between TDA and machine learning, exploring how topological insights can be harnessed to improve data analysis and model performance

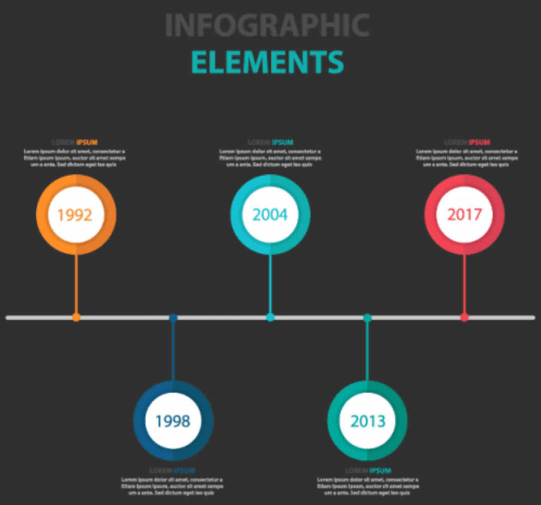

Figure 1: Example figure caption.

Foundations of Algebraic Topology in Data Analysis

Topological Data Analysis (TDA) draws on core principles of algebraic topology to extract structural and shape-based features from data. This section outlines the mathematical foundations necessary for understanding how topology is applied to data, including the construction of simplicial complexes, the computation of homology, and the use of persistent homology for multi-scale analysis [3, 6, 4].

Simplicial Complexes and Filtrations

A **simplicial complex** is a combinatorial object that generalizes the notion of a graph to higher dimensions. Formally, a simplicial complex K is a collection of finite sets closed under the subset operation: if $\sigma \in K$ and $\tau \subseteq \sigma$, then $\tau \in K$. Each k-dimensional simplex corresponds to a set of $k+1$ vertices:

$$\sigma_k = [v_0, v_1, \ldots, v_k], \quad \text{with } v_i \in V$$

For data given as a point cloud $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, two common constructions are:

- **Vietoris-Rips Complex** $VR_\epsilon(X)$: includes a k-simplex for every set of $k+1$ points that are pairwise within distance ϵ.
- **Čech Complex** $\check{C}_\epsilon(X)$: includes a k-simplex if the intersection of all balls of radius ϵ centered at the $k+1$ points is non-empty.

To analyze data at multiple scales, a filtration $\{K_\epsilon\}_{\epsilon \geq 0}$ is constructed:

$$K_{\epsilon_0} \subseteq K_{\epsilon_1} \subseteq \cdots \subseteq K_{\epsilon_n}$$

This filtration allows us to observe the birth and death of topological features as ϵ increases [5].

Dataset	Feature Representation	Topological Dimension
MNIST	Persistence Image	β_1
CIFAR-10	Betti Curve	β_0, β_1
fMRI Brain Data	Persistence Diagram	β_0, β_2
Protein Graphs	Persistence Landscape	β_1
Synthetic Torus	Persistence Barcode	β_1, β_2
Financial Time Series	Betti Curve	β_0
Materials Microstructure	Persistence Image	β_2
Gene Expression	Persistence Diagram	β_0, β_1

Table 1: Topological feature extraction methods applied across various datasets and their associated Betti dimensions.

Homology and Betti Numbers

Homology provides an algebraic summary of topological spaces by quantifying features such as connected components, loops, and voids. Given a simplicial complex K, we define:

- The k-th chain group $C_k(K; \mathbb{F})$ as the vector space over a field $\mathbb{F}$ spanned by the k-simplices of K.
- The boundary operator $\partial_k : C_k \to C_{k-1}$ satisfying $\partial_k \circ \partial_{k+1} = 0$.

The k-th homology group is then:

$$H_k(K; \mathbb{F}) = \frac{\ker(\partial_k)}{\mathrm{im}(\partial_{k+1})}$$

The dimension of H_k is the k-th **Betti number** β_k, which intuitively counts:

$$\begin{aligned} \beta_0 &= \text{number of connected components} \\ \beta_1 &= \text{number of independent loops} \\ \beta_2 &= \text{number of voids or cavities} \end{aligned}$$

These numbers provide coarse yet powerful summaries of topological shape [6].

Persistent Homology

Persistent homology tracks the evolution of homology classes across a filtration. For each k-dimensional feature, we record its birth ϵ_b and death ϵ_d values:

$$\text{Persistence Pair: } (\epsilon_b, \epsilon_d)$$

These pairs are often visualized via:

- **Persistence Diagrams** $\mathcal{D}_k$: a multiset of points $(\epsilon_b, \epsilon_d) \in \mathbb{R}^2$ above the diagonal.
- **Barcodes**: horizontal lines spanning from ϵ_b to ϵ_d on a 1D axis.

The p-Wasserstein distance between two persistence diagrams $\mathcal{D}_1$ and $\mathcal{D}_2$ gives a notion of similarity between topological signatures:

$$W_p(\mathcal{D}_1, \mathcal{D}_2) = \left(\inf_{\gamma: \mathcal{D}_1 \to \mathcal{D}_2} \sum_{x \in \mathcal{D}_1} \|x - \gamma(x)\|_\infty^p \right)^{1/p}$$

where γ is a bijection between diagrams (possibly allowing diagonal projections).

Persistent homology distinguishes noise from signal by measuring the longevity of features: long intervals in barcodes often correspond to significant topological structure, while short ones indicate noise [5, 4].

Example: Circle and Torus

Consider data sampled from a noisy circle S^1. Its ideal homology is:

$$\beta_0 = 1, \quad \beta_1 = 1, \quad \beta_k = 0 \text{ for } k > 1$$

For a torus T^2, we expect:

$$\beta_0 = 1, \quad \beta_1 = 2, \quad \beta_2 = 1$$

TDA can recover these signatures even from scattered point clouds using persistent homology, providing strong evidence for underlying topological structure [3, 8].

Integrating TDA into Machine Learning

While persistent homology provides powerful topological summaries, its integration into machine learning workflows requires careful encoding, interpretation, and computational adaptation. This section outlines how topological features are translated into vectorized representations, used to improve model robustness, and applied across various domains [7, 8].

Feature Extraction and Representation

Persistence diagrams $\mathcal{D}_k$ are not naturally suited for direct consumption by standard machine learning algorithms due to their variable size and set-based nature. To bridge this gap, several techniques have been developed to convert diagrams into fixed-dimensional vector spaces:

- **Persistence Landscapes** [2] represent $\mathcal{D}_k$ as a sequence of piecewise-linear functions:

  $$\lambda_k(t) = \sup_{(b,d) \in \mathcal{D}_k} \left[\min\left(t - b, d - t\right)_+ \right]$$

  where $(\cdot)_+ = \max(\cdot, 0)$, giving a functional summary that supports statistical analysis in Banach spaces.

- **Persistence Images** [1] map each diagram to a fixed-size grid with Gaussian kernels centered at each point (b, d):

  $$\rho(x, y) = \sum_{(b,d) \in \mathcal{D}_k} \exp\left(-\frac{(x - b)^2 + (y - d)^2}{2\sigma^2} \right)$$

  This transforms topological features into image-like tensors that can be fed directly into neural networks or classifiers.

- **Betti Curves** plot Betti numbers over the filtration parameter ϵ:

  $$\beta_k(\epsilon) = \mathrm{rank}(H_k(K_\epsilon))$$

  giving a coarse yet intuitive descriptor for the evolution of topological complexity across scales.

These representations retain essential topological structure while conforming to the expectations of downstream models.

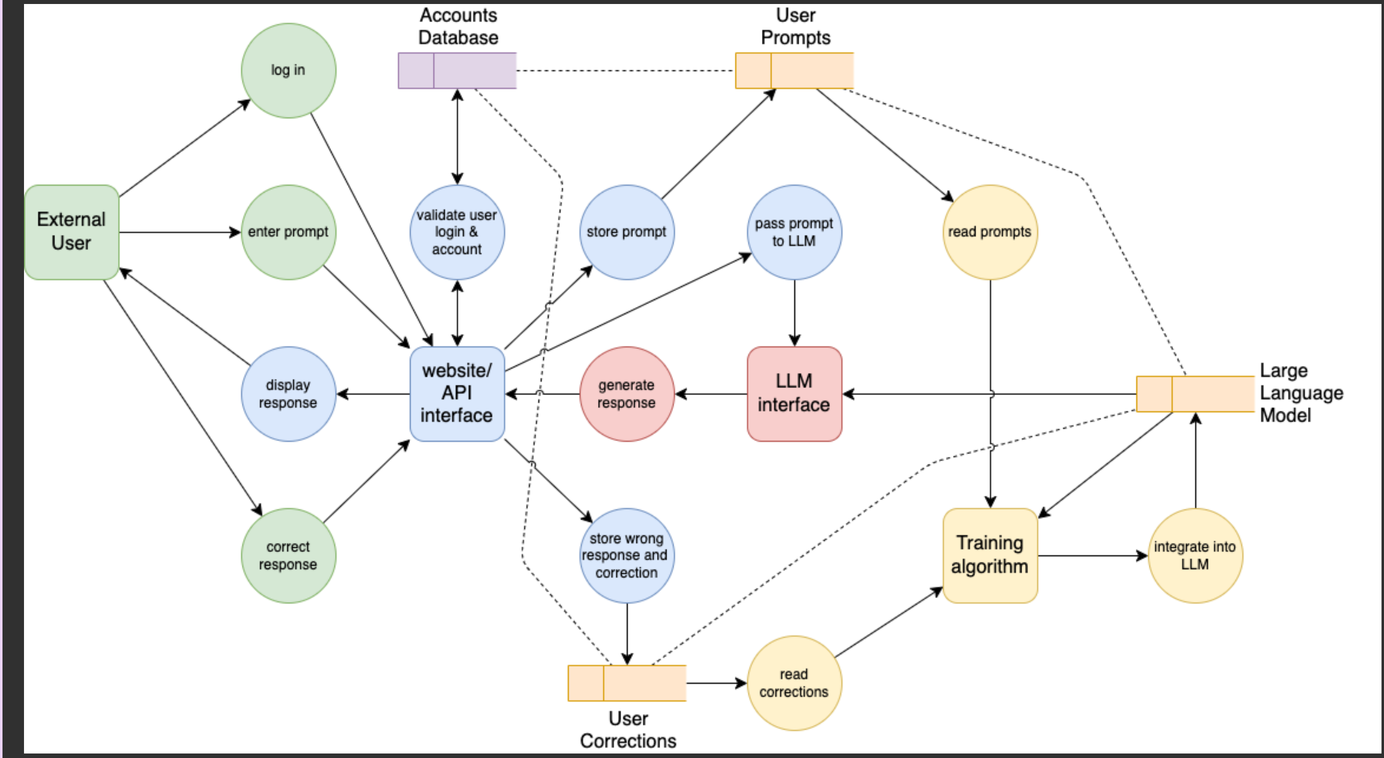

Figure 2: Example of a figure spanning two columns.

Enhancing Model Robustness and Interpretability

Topological features provide complementary information to geometric and statistical features. In particular:

- **Robustness**: Persistent homology is stable under perturbations of the data [4], meaning features with long lifespans persist even under noise. This makes topological descriptors ideal for noisy or incomplete datasets.
- **Global Structure Awareness**: Unlike local features (e.g., gradients or kernels), topological features capture global connectivity, loops, and voids—critical for problems where such structure encodes meaning, such as protein folding or brain connectivity.
- **Interpretability**: Betti numbers and persistence diagrams are human-interpretable. For instance, a significant β_1 feature may correspond to a loop in the data manifold—something that geometric embeddings like PCA would fail to express.

Recent work has shown that neural networks augmented with topological features, or trained using topologically-informed regularizers, exhibit improved generalization and adversarial resilience [7, 8].

Applications Across Domains

Topological data analysis has found practical utility in a wide range of scientific and industrial applications:

- **Biology**: In gene expression analysis, TDA captures clustering structures and loops in expression profiles that may correspond to cyclic biological processes or subpopulations [3].
- **Neuroscience**: Persistent homology has been applied to functional MRI data to identify topological patterns in brain activity, such as connectedness and modularity of neural networks [4].
- **Materials Science**: Topological fingerprints of porous structures in materials help in quantifying permeability and mechanical properties.
- **Finance**: TDA can detect regime changes and anomalous market behavior by capturing shape-based anomalies in high-dimensional stock trajectories [8].

These examples showcase TDA's versatility as a domain-agnostic tool for extracting meaningful structure in complex datasets.

Conclusion

Topological Data Analysis offers a powerful and mathematically principled approach to understanding the shape and structure of complex datasets. By leveraging concepts from algebraic topology—such as simplicial complexes, homology groups, and persistent homology—TDA provides a robust, noise-tolerant framework for uncovering multiscale geometric features that are often invisible to conventional statistical techniques [3, 4].

The integration of TDA into machine learning pipelines has opened new avenues for both theoretical exploration and practical application. Through persistence-based feature extraction methods like persistence landscapes, images, and Betti curves, topological summaries can be transformed into vectorized representations suitable for modern learning algorithms [2, 1]. These features enhance interpretability, improve robustness to noise and adversarial attacks, and often yield performance gains in settings where data is high-dimensional, sparse, or structured in nontrivial ways.

Despite its potential, TDA faces several challenges that must be addressed for widespread adoption. Computational complexity, particularly in high dimensions, remains a bottleneck. Additionally, the interpretability of topological features across domains, as well as their seamless integration with deep learning architectures, are areas of active research [7, 8].

Nonetheless, as the field matures and software tools continue to evolve, the application of topological methods in machine learning is expected to grow significantly. The synergy between rigorous mathematical abstraction and practical data analysis embodied by TDA is emblematic of a broader trend—where geometry, topology, and learning theory converge to create next-generation analytical tools.

Font Acknowledgment

This document uses the 'Harding' typeface, designed by Paul Barnes and Christian Schwartz. The font is distributed by Commercial Type and is available at `https://commercialtype.com/`. *It is included here strictly for educational and non-commercial purposes. No part of this document will be used in professional, commercial, or revenue-generating contexts. The author does not claim any rights over the font and fully acknowledges the intellectual property of its creators.*

References

[1] H. Adams, G. Carlsson, and D. Sheehy. Persistence images: A stable vector representation of persistent homology. *Discrete & Computational Geometry*, 57(3):547–587, 2017.

[2] P. Bubenik and P. Kim. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16(1):77–102, 2015.

[3] G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[4] F. Chazal and B. Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4:667963, 2021.

[5] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, 2002.

[6] A. Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.

[7] C. Hofer, R. Kwitt, M. Niethammer, and A. Uhl. Deep learning with topological signatures. In *Advances in neural information processing systems*, volume 30, 2017.

[8] B. Rieck, C. Bock, M. Moor, M. Horn, L. Roethlisberger, and K. Borgwardt. Topological machine learning with persistence indicator functions. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8275–8287, 2020.