IJETIR

ISSN (Online) 2583-0554

International Journal of Emerging Technologies and Innovative Research (IJETIR)

Volume 4, Issue 12, December 2024

Impact Factor: 5.731

Evaluating the Practical Effectiveness of Topological Data Analysis in High-Dimensional Data Interpretation

Sagar Chandrabhan Pawar[1] and Dr. Brijpal Singh[2]

Research Scholar, Department of Mathematics[1]

Professor, Department of Mathematics[2]

Sunrise University, Alwar, Rajasthan, India

Abstract: *A relatively new and rapidly expanding area of data science is topological data analysis. TDA offers a method for analyzing data sets and extracting pertinent features from intricate, high-dimensional data, significantly increasing productivity in a variety of industries. The author primarily covers topological mathematics topics, TDA techniques, and the connections between topological concepts and data sets in this study. This article introduces the TDA issues, the mathematical technique used in TDA, and two application examples. Furthermore, the benefits, drawbacks, and potential growth path of TDA are examined..*

Keywords: Data Clustering, Dimensionality Reduction, Persistent Homology

I. INTRODUCTION

The area of pure mathematics that examines the concept of shape is called topology. By representing complicated data sets as a network of nodes and edges, topological data analysis aims to provide an understandable map based on the data points' similarities. The data points will seem closer to one another on the map the more similar they are. Reducing high dimensional data sets to lesser dimensions while maintaining their most essential topological characteristics is the goal.

Topological approaches provide a rapid means of deriving information from data and comprehending its structure. It is possible to create techniques for form recognition using topology. The author primarily covers topological mathematics topics, TDA techniques, and the connections between topological concepts and data sets in this study. This article introduces the TDA issues, the mathematical technique used in TDA, and two application examples. Furthermore, the benefits, drawbacks, and potential growth path of TDA are examined.

II. REVIEW

1. Simplicial Complexes

In topology, the term "simple complex" describes a topological object, such as triangles, points, or line segments, that is joined by a simplex. Additionally, it may be used to define a class of topological spaces.

Definition 1: A geometric shape made up of triangles is called a simplicial complex. Topological investigations of general graphics may be made easier by establishing a specified correlation between general graphics and these simpler graphics. These fundamental triangles are known as two-dimensional simple complexes, or simply two-dimensional simplex. Triangular high-dimensional analogs may also be used for higher-dimensional simple complexes. These triangles have to be in a certain orientation, either not intersecting or having a shared face.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 12, December 2024
Impact Factor: 5.731

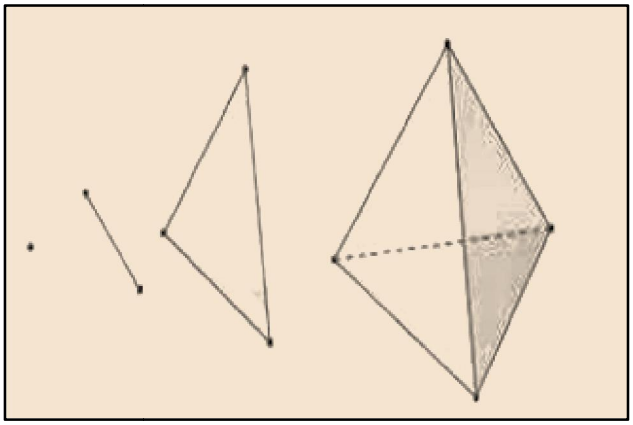

Figure 1: The figure shows k-simplexes for k = 0;1;2;3

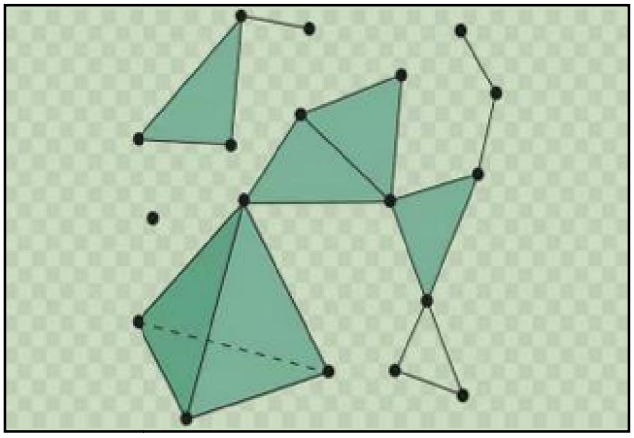

Figure 2: The figure shows some simplicial complex

2. Homology

A broad method of connecting a series of algebraic objects, such abelian groups or modules, to other mathematical objects, like topological spaces, is called homology. Algebraic topology was the first to define homology groups. In topology, there are several kinds of homology, including singular, group, and simplicial homology, among others.

Definition 1(simplicial homology): Given n ∈ Z+, the n-th homology group of a simplicial complex K, is denoted by Hn (K, F), and is defined as formula (1). That is, Hn (K, F) is a quotient vector space and the elements of Hn (K, F) are equivalence classes of n-cycles of C∗(K, F).

$$Hn(K,F) = \frac{Zn(K,F)}{Bn(K,F)} \quad (1)$$

Definition 2 (Betti numbers): Given n ∈ Z+, the n-th Betti number of a simplicial complex K is denoted by βn(K), and is defined as βn(K) := dim(Hn(K, F)).

Lemma 1 (fundamental lemma of homology): For every (p+1)-chain d we have θkθk+1d = 0.

Lemma 2: For every simplicial complex K, β0(K) is equal to the number of connected components of K.

Definition 3 (contiguous simplicial maps): Given simplicial complexes K and L, simplicial maps f, g: K -¿ L are said to be contiguous if for every simplex σ ∈ K, f(σ) ∪ g(σ) is a simplex in L.

Definition 4: A sequence of nested simplicial subcomplexes is called a filtration.

$$\phi \subseteq K_0 \subseteq K_1 \subseteq K_2 \subseteq \ldots \quad (2)$$

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 12, December 2024
Impact Factor: 5.731

3. Persistent Homology

Persistent Homlogy is an efficient tool which is widely used on studying and analyzing data sets. It tries to find and track homological groups and holes with the help of filtrations. A visual representation of persistent homology is persistent diagram.

Definition 1(filtered simplicial complex): A filtered simplicial complex is a sequence of simplicial complexes {Kβ} (β ∈ R) such that for all β ≤ β0, Kβ ⊆ Kβ0. We now see some examples of filtered simplicial complexes that can be constructed from a finite metric space (X, dx).

Definition 2: The p-persistent diagram D of a filtration is defined as formula (2). Let μ be the number of independent p- dimensional classes that are Born in Ki and die entering Kj then D is obtained by drawing a set of points (i, j) with multiplicity μ, where the diagonal is added with infinite multiplicity. For comparing two persistent diagrams some metrics are defined that two of most important of them are bottleneck and Wasserstein distances.

Definition 3: Let D1, D2 be two persistent diagrams and B be the set of all bijective functions ϕ: D1∞ > D2. If ||.||∞ be the supremum norm, then the bottleneck distance between two persistent diagrams D1, D2 denoted by W∞(D1, D2) is defined as follows:

$$W\infty\ (D_1, D_2) = \inf \sup \varphi \in B_x \in D_1 \left\| x - \varphi(x) \right\| \infty \quad (3)$$

Definition 4: Let D1, D2 two persistent diagrams and B be the set of all bijective functions ϕ: D1∞ > D2, then the Wasserstein distance between two persistent diagrams D1, D2 denoted by WP (D1, D2) is defined as follows:

$$W_p(D_1, D_2) = \left[i\inf_{\varphi \in B} \sum_{x \in D_1} \left\| x - \varphi(x) \right\|_\infty^p \right]^{1/p} \quad (4)$$

Since it is very hard to analyze the information about homological groups and holes we can use a visualization method called barcode, the idea is as follows: if a hole appears in $t_1 \in$ in x-axis and if die in $t_2 \in$, we stop drawing the line and end of the line would be at $t_2 \in$. Persistent landscape is another method introduced by Bebunik [2] to visualize persistent homology.

Definition 5: The persistence landscape is a function λ: N × RR > R where R denotes the extended real numbers [−∞, ∞]. Alternatively, it may be thought of a sequence of functions λk: RR > R, where λk(t) = λ (k, t). Def ineλk(t) = sup {m ≥ 0|β (ttm), (t+m) ≥ k} where β (i, j) is the dimension of group Hi/Hj.

The graph of landscape indicates persistent and non persistent betti numbers, for example the support of persistent landscape denotes non-persistent Betti numbers and the maximum of landscape graph indicate the most persistent Betti number.

4. Relation Between Topology and Data Sets

By assigning algebraic objects known as invariants which may be as basic as integers but are often more complicated algebraic structures topologists [4] can learn space. Persistent homology is the invariant selected for TDA. The majority of the gathered data is an ordered collection of N tuples with attributes like dimensions and locations. When these qualities are numerical, researchers might consider them as defining vectors for European space. To determine a filtered value for every data point, researchers use a filter function, which might be a linear projection of the data matrix, the density estimate, or the centrality index of the distance matrix.

Based on their filtered values, the data points are separated into several filter value intervals ranging from tiny to big. Adjacent filter value intervals are often set with a certain overlap region. This indicates that the overlap area's points simultaneously belong to two intervals. Following that, each interval's data are clustered independently. An edge must be placed between the two classes if their raw data points are identical. The final data pattern is created by applying a layer of mechanical layout to the circle and edge graphics stated previously in order to attain equilibrium.

III. ANALYSIS OF THE APPLICATION CASES

The author examines two articles: An Introduction to a New Text Classification and Visualization for Natural Language Processing Using Topological Data Analysis [2] and Topological Data Analysis of Time Series Data for B2B

Customer Relationship Management [1]. Both of these papers use TDA to address certain real-world issues in various fields.

1. Problems and Solutions of customer relationship

One of the biggest cloud computing providers in the world needs to leverage historical customer relationship management data to estimate client demand in order to enhance its service demand projection. However, it is challenging to generate forecasts using the conventional method because of the little customer-level data available. Additionally, while the Recency, Frequency, Monetary framework is a widely used approach, it may also be deceptive.

In light of this, researchers gather four distinct data sets from the internet and analyze them using two additional techniques: Time Series Clustering and TDA in conjunction with RFM. At the conclusion of the trial, they assess the prediction's accuracy to see if these two novel approaches really contribute to increased efficacy.

In this instance, TDA and RFM are used in conjunction to analyze data and provide predictions. In contrast to time series clustering, TDA is unlikely to be impacted by noise, making it the most efficient approach for generating predictions on the same data when compared to the other two approaches.

In order to make complicated and multi-dimensional data sets easier to study and compare, TDA aims to reduce their dimension and extract their topological structure. The first step is comparable to Time Series RFM's first step. After that, they begin building point clouds. After obtaining three point clouds, death and birth complexes are created using a TDA method known as rip filtering. The birth-death filtered complexes may then be seen thanks to the creation of barcode diagrams for both 0- and 1-dimensional homologies. Following that, clusters were created using K-means based on features that were taken from the barcodes. Gradient tree boosting is ultimately used to repeat the prediction.

2. Problems and Solutions of Text Classification

In order to properly evaluate and comprehend the vast amount of data available on the internet nowadays, we must categorize it. In the Text Classification instance, researchers use TDA to categorize texts written by Ferdowsi and Hafez, two of the greatest Iranian poets.

One of the newest and fastest-growing areas of data science is TDA, which effectively analyzes data by examining its structure and converting it into less dimensional data that is simpler to analyze. Currently, Persistent Homology and Mapper, two well-liked TDA algorithms, are used by academics to accomplish the categorization.

The algorithm for persistent homology: This algorithm's concept [5] is to use P as point cloud data. The Vietoris-Rips complex of P is first constructed as follows: take the ascending series of positive real numbers $a1 \leq a_2 \leq a_3 \leq..$ A cover of circles with a diameter of a1 and a point in P is then created, giving us a large number of circles equal to the number of data points in the point cloud data. The boundaries between the two circles' centers are then drawn, along with any intersections. We get a simplicial complex VR (a_1) as a consequence, and we repeat the process for all i = 1, 2, 3,... Next, we have a VR (ai) filtering of complexes. Given that we find it challenging to evaluate the data on homological groups and holes, we may use a visualization technique known as barcode. This approach works by drawing a line if a hole opens in $\alpha t1$, with the line's beginning point on the x-axis of $\alpha t1$. If the hole dies in αt_2, the line is stopped and ends at αt_2.

The Mapper Algorithm's basic concept is that by taking a high-dimensional data point cloud as input, we may get a network that represents the point cloud's topological information. The point cloud is divided into many parts using the filter function, and each region is clustered independently. The output network has nodes for each class.

We link points in the coincidence section of the point cloud that belong to two or more clusters. Regarding how to split a point cloud, n filter functions convert a point cloud of size m into m n-dimensional vectors. Determine the number of sub-areas to be split, as well as the number of coincidences between each sub-area and other sub-areas, before beginning the division.

IV. DISCUSSION

1. Problems and Solutions of customer relationship

A distance matrix that shows the separation between any two data points may be used as the TDA's input. TDA is entirely unrestricted by coordinates and investigates forms that are not reliant on them. This implies that the definition

of the distance function or the notion of similarity is necessary for the building of topological forms. TDA can integrate data from several platforms thanks to coordinate-independent characteristics.

We simply need to provide a decent distance function, even if the data format is different. Furthermore, TDA research's data shape can withstand little data distortion and deformation. Additionally, the easiest method for drawing a basic contour of a lake is to utilize a polygon. The shape of topological processing is abstract. The most common example is when a circle is represented by a hexagon. Just six points and six edges are needed for this. This shape is used by TDA to compress data and represent vast quantities of data using a small number of points and edges.

2. Practical Assistance

Both of the aforementioned scenarios, which examine customer relationships and text categorization, use TDA algorithms to evaluate data sets and provide optimal outcomes. Consequently, it demonstrates that TDA performs well in the economic domain. Because TDA primarily focuses on examining the structure and organization of data, it may be used to any kind of data, including picture, sensor, and even audio data, in addition to research text data. Consequently, TDA has been effectively used in a variety of domains, including biophysics, image processing, cancers, and nerves [6].

3. Limitations

The primary constraint in the study of topological data is the computational resources needed. Some TDA methods, like the Text Classification method's persistent homology [7], cannot manage a significant volume of huge data. They will slow or fail in this scenario, in part because the method uses KNN maps. Apart from this, the only real issues with using TDA algorithms are a few strictly technical ones. For example, the study of the metric space under persistent homology is almost entirely necessary to determine how to effectively calculate multi-D persistence in reality.

4. Development Direction

To now, TDA has only touched on a few small areas of study, despite the fact that topology is a broad topic. Researchers must better integrate TDA with current methods in the future, particularly with machine learning. By doing this, researchers may more effectively extract and use high-dimensional data to get around current problems and increase efficiency when working with large data.

V. CONCLUSION

TDA has gained popularity due to its advantages. TDA is a very effective machine learning tool that can be used with other machine learning techniques to provide greater outcomes than if it were used alone. More significantly, it has significantly altered the way we evaluate data. It is a highly innovative and daring technique to combine data analysis with topology and the field of pure mathematics. taking into account TDA's drawback, which is that it cannot function at its best when handling a lot of data. To increase efficiency, researchers should better integrate machine learning into TDA. Because it can employ the most evident benefit of TDA its adaptability in machine learning to get a better understanding of data, in addition to covering its shortcomings as much as feasible. TDA-based algorithms will continue to be suggested and utilized more often in the future along with other advancements in technology.

REFERENCES

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

[2]. N. Elyasi, M. Hosseini.Moghadam, An Introduction to a New Text Classification and Visualiza tion for Natural Language Processsing Using Topological Data Analysis. *IEEE.* 2019.

[3]. R. Rivera-Castrol, P. Pilyugina, A. Pletnev, I. Maksimov, W. Wyz, E. Burnaev, Topological Data Analysis of Time Series Data for B2B Customer Relationship Management. *IEEE.* 2019.

[4]. Hatcher, Algebraic topology. *Cambridge Univ. Press* 2000.

[5]. U. Bauer; M. Kerber; and J. Reininghaus, PHAT (Persistent Homology Algorithm Toolbox), 2012, https://bitbucket.org/phat-code/phat

[6]. P. G. Camara, D. I.S. Rosenbloom, K. J. Emmett, A. J. Levine, R. Rabadan, Topological data analysis generates high-resolution, genome-wide maps of human recombination. *Cell Systems,* 2016.

[7]. S. Bhattacharya, R. Ghrist, V. Kumar, Persistent Homology for Path Planning in Uncertain Environ ments. *IEEE TRANSACTIONS ON ROBOTICS*, vol. 31, no. 3 Jun. 2015