The Application of Topological Data Analysis in Practice and Its Effectiveness

Liang Cheng

Department of software and microelectronics, Harbin university of science and technology, Harbin, Heilongjiang, China

Abstract—Topological Data Analysis(TDA) is a new and fast growing field in data science. TDA provides an approach to analyze data sets and derive their relevant feature out of complex high-dimensional data, which greatly improves the working efficiency in many fields. In this paper, the author mainly discusses some mathematics concepts about topology, methods in TDA and the relation between these topological concepts and data sets (how to apply topological concepts on data). The problems of TDA, mathematical algorithm using in TDA and two application-examples are introduced in this paper. In addition, the advantages, limitations, and the direction of future development of TDA are discussed.

1 INTRODUCTION

Topology is the branch of pure mathematics that studies the notion of shape. The idea behind Topological Data Analysis (TDA) is to represent complex data sets as a network of nodes and edges, and create an intuitive map based on the similarity of the data points. The more similar the data points are, the closer they will be to each other on the map. The idea is to reduce high dimensional data sets to lower dimensions without sacrificing their most relevant topological properties. Topological methods provide a quick way to understand the structure of the data and obtain knowledge from data. Topology can be used to develop methods for recognizing shapes. In this paper, the author mainly discusses some mathematics concepts about topology, methods in TDA and the relation between these topological concepts and data sets (how to apply topological concepts on data). The problems of TDA, mathematical algorithm using in TDA and two application-examples are introduced in this paper. In addition, the advantages, limitations, and the direction of future development of TDA are discussed.

2 REVIEW

2.1 Simplicial Complexes

Simple complex is a concept in topology, which refers to a topological object that is bonded by simplex such as points, line segments, and triangles. It is also a tool for defining a class of topological spaces.

- Definition 1: A simplicial complex is a geometric figure composed of triangles. By correlating general graphics with these simpler graphics in a defined manner, topological (qualitative) studies of general graph ics can be simplified. These basic triangles are called two-dimensional simple complexes, abbreviated as two-dimensional simplex, and higher-dimensional simple complexes can also use triangular high-dimensional analogs (so-called N-dimensional simplex; eg: three-dimensional simplex is four-sided Body). These triangles must be in a certain way, either not intersecting, or the intersecting part is its common face.

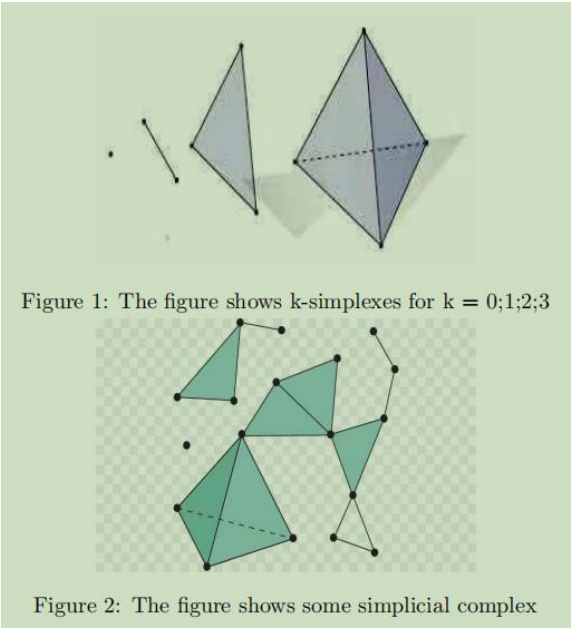

Figure 1: The figure shows k-simplexes for k = 0;1;2;3

Figure 2: The figure shows some simplicial complex

2.2 Homology

In mathematics, homology is a general way of associating a sequence of algebraic objects such as abelian groups or modules to other mathematical objects such as topological

angela@cas-harbour.org

spaces. Homology groups were originally defined in algebraic topology. There are different types of homology in topology such as simplicial homology, Group homology, Singular homology and so on.

- 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.

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

2.3 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 $\mu_p^{i,j}$ 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 $\mu_p^{i,j}$, 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:

$$W\infty\ (D_1, D_2) = \inf \sup_{\varphi \in B} \sup_{x \in D_1} \|x - \varphi(x)\|\infty \quad (3)$$

- 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:

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

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 $\in t_1$ in x-axis and if die in $\in t_2$, we stop drawing the line and end of the line would be at $\in t_2$. 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.

2.4 The Relation Between Topology and Data Sets

In topology [4], researchers learn space by assigning algebraic objects called invariant which may be as simple as integers, but usually are more complex algebraic structures. For TDA, the chosen invariant is persistent homology. The collected data mostly are ordered set of N-tuples that contain features such as coordinates and dimensions. Researchers can think of them as definition vectors for European space when these features are numbers. Researchers use a filter function which can be a linear projection of the data matrix(like PCA) or the density estimate or the centrality index of the distance matrix to calculate a filtered value for each data point. The data points are divided into different filter value intervals from small to large according to their filtered values. In most cases, adjacent filter value intervals are set with a certain overlap area. That means the points of the overlap area belong to two intervals at the same time. And then, the data are clustered in each interval separately. If there are identical raw data points between the two classes, an edge need to be added between them. A layer of mechanical layout is applied to the above-mentioned circle and edge graphics to achieve equilibrium, and the final data pattern is obtained.

3 THE ANALYSIS OF THE APPLICATION CASES

The author studies two papers Topological Data Analysis of Time Series Data for B2B Customer Relationship Management [1] and An Introduction to a New Text Classification and Visualization for Natural Language Processing Using Topological Data Analysis [2]. TDA is applied to both these two articles to solve some practical problems in different areas.

3.1 Problems and Solutions of customer relationship

In order to improve the demand forecast of its services, one of the world's largest cloud computing provider need to use the previous customer relationship management(CRM) data to get customer demand prediction. However, considering the data available is limited at customer level, it is hard to use the traditional way to make predictions. Furthermore, the Recency, Frequency, Monetary(RFM) framework is a popular method but it also can be misleading.

So based on the actual situation, researchers collect four different data sets from internet and use two other methods Time Series Clustering and TDA combined with RFM to analyze these data. They examine the accuracy of the prediction at the end of the experiment and see whether these two new methods really help to improve the effectiveness.

In this case, TDA combined with RFM are applied on analyzing data and make predictions. Unlike to time series clustering, TDA is not likely to be affected by noise, so this method is expected to be the most effective way to make predictions on the same data compared to the above two methods. The idea of TDA is to lower the dimension of complex and multi-dimensional data sets and extract the topological structure from data sets which is more easily to analyze and compare. The first step is similar to the first step in *Time Series RFM*. Then they start construct point clouds. Once three point clouds are obtained, an algorithm in TDA called Rip filtration is used to obtain death and birth complexes. Next, barcode diagrams are generated for both 0- and 1- dimensional homologies which help to visualize the birth-death filtered complexes. After that, using K-means based on featured extracted from the barcodes to clusters. In the end, gradient tree boosting is used for doing prediction again.

3.2 Problems and Solutions of Text Classification

Nowadays, internet is full of based information that we need to classify them in order to better analyze and understand these data. In the Text Classification case, researchers classify text (Persian poems) which has been composed by two of the best Iranian poets namely *Ferdowsi* and *Hafez* by TDA.

TDA is one of the newest and fast growing branches of data science which is effective in analyzing data by studying its shape and transform it into data with less dimensionality that makes it easier for analyzing. At this time, researchers use two popular algorithms in TDA, Persistent Homology and Mapper to achieve the classification.

Persistent Homology Algorithm: The idea of this algorithm [5] is that we take P as point cloud data. Firstly, constructing the the Vietoris-Rips complex of P as follows: consider the increasing sequence of positive real numbers $a1 \leq a2 \leq a3 \leq ...$ Then a cover of circles with point is in P and diameter a1 is constructed, so we have many circles as the number of data points in the point cloud data. And then we draw the edges between the centers of the two circles with any intersections. As a result, we have a simplicial complex VR (a1), and we do the same for all i = 1, 2, 3... After that we we have a filtration of complexes VR (ai). Account that it is hard for us to analyze the information about homological groups and holes, we can use a visualization method called barcode. The idea of this method is if a hole appears in $\alpha t1$, we start to draw a line, the starting point of the line is on the x-axis of $\alpha t1$, if it dies in $\alpha t2$, we stop drawing the line and the end of the line will be $\alpha t2$.

Mapper Algorithm: The intuitive idea behind Mapper is that we use a high-dimensional data point cloud (distance matrix) as an input and we can get a network representing the topology information of the point cloud as a result. We use filter function to divide the point cloud into several regions, and each region is separately clustered. Each class serves as a node in the output network. If a point in the coincidence portion of the point cloud belongs to two or multiple clusters, then we connect them. About the way to divide the point cloud, a point cloud of size m is processed into m n-dimensional vectors (or points) by n filter functions. First, decide how many sub-areas are to be divided, and how many coincidences of each sub-area and other sub-areas are, and then start the Division.

4 DISCUSSION

4.1 Problems and Solutions of customer relationship

The input to the TDA can be a distance matrix representing the distance between any two data points. TDA studies shapes that are independent of coordinates and is completely unconstrained by coordinates. This also means that the construction of topological shapes depends on the definition of the distance function, or the definition of the concept of similarity. Coordinate-independent features allow TDA to integrate data from different platforms. Although the structure of the data is not the same, we only need to give a reasonable distance function. Moreover, the data shape of TDA research can tolerate small deformation and distortion of data. In addition, if we want to sketch a lake outline roughly, the simplest one is to use a polygon. Topological processing is an abstract shape. The most typical example is to use a hexagon to represent a circle. This requires only 6 points and 6 edges. TDA uses this form to compress data and use a limited number of points and edges to represent large amounts of data.

4.2 Practical Assistance

Both two cases mentioned above that study customer relationship and text classification, applying TDA algorithms to analyze data sets and obtaining ideal results. Thus, it shows that TDA has a high performance in economical area. In addition to research text data, TDA can be applied to any other kinds of data, such as image data, sensor data, and even audio data, owing to TDA mainly focus on studying the structure and shape of data. Therefore, TDA has been successfully applied to many fields, such as tumors, nerves, image processing and biophysics [6].

4.3 Limitations

The main limitation in topology data analysis is that the computing resources it required. For some TDA algorithms, such as persistent homology [7], the method used in Text Classification, they are not capable to handle considerable amount of big data. In this situation, they will stall or crash, partly owing to the use of KNN maps in the algorithm. Besides this, there are almost nothing but some purely technical problems in applying TDA algorithms. For instance, how to efficiently compute multi-D persistence in practice, nearly all about studying the metric space under persistent homology.

4.4 Development Direction

Up till now, although topology is a field with a wide range, TDA only has scratched the surface of one or two tiny pieces of research. In the future, researchers need to better combine TDA with existing techniques, especially with machine learning. Through this way, researchers are able to better extract and utilize high dimensional data to overcome existing drawbacks and make it more efficient when dealing with big data.

5 CONCLUSION

Due to its merits, TDA has begun to be widely used. TDA is a very powerful tool in machine learning, and it can be used with machine learning methods to get better results than using a single technology. More importantly, it has changed the way we analyze data to a large extent. Combining topology and the discipline of pure mathematics with data analysis is a very cutting-edge and bold technology. Considering the limitation of TDA that it can not have an ideal performance when dealing with a large amount of data. Researchers ought to better join the machine learning into TDA to improve the efficiency. Because it not only can cover the drawbacks of TDA as much as possible, but also can apply the most obvious merit of TDA, versatility on machine learning to gain a deeper insight into data. In the future, TDA-based algorithms will be constantly proposed, and more used with other improving technique.

ACKNOWLEDGMENT

First and foremost, I would like to show my deepest gratitude to my teachers and professors in my university, from whom I have gotten the best guidance. Furthermore, I would like to thank my classmates and roommates for their encouragement and support. Without their help, I can not finish my paper.

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. A. 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.