Advances in Social Science, Education and Humanities Research, volume 412
International Conference on Modern Educational Technology and Innovation and Entrepreneurship (ICMETIE 2020)

Topological Methods for the Analysis of Applications

Yumiao Lei
Department of Mathematics, Faculty of Information And Computing Science, Taiyuan University of Technology, Taiyuan, Shanxi, 030024, China
Corresponding E-mail：angela@cas-harbour.org

Keywords: TDA, Persistent Homology, Hausdorff Distance, text classification, face detection

Abstract. Topological Data Analysis(TDA) is a rapidly developing data analysis field in recent years. It provides topological and geometric methods to obtain the relevant features of high-dimensional data. This paper introduces the related mathematical principles of Persistent Homology, Mapper, Hausdorff Distance in topology and enumerates two applications of TDA. One is about text classification of natural language. It uses persistent homology to analyze poetry data and mapper algorithm to analyze and visualize data sets. The other application is based on the principle of Robust Hausdorff Distance, and proposes a fast and accurate shape comparison method for face detection. The result shows that the TDA method is not only accurate, but also can realize data visualization.

1. Introduction

Data is everywhere, there are many connections hidden in the complex data. Generally, four ”V” namely, Volume, Variety, Value, Velocity, are used to summarize the characteristics of data. However, massive and complex data sets that cannot be extracted, stored, searched, shared, analyzed and processed with the current software tools. At the core of the time of big data, forecasting analysis has been widely used in business and society. Because of the huge volume of data, various types of data and low value density, different requirements for data processing have been put forward in various fields. Deal with high dimensional data and transform it into data with less dimensionality in order to make it easier for analyzing. How to purify data and get valuable information is a big problem. The use of topology, in particular, algebraic topology has been used to address a wide variety of problems[4]. People use topological methods to reduce the dimensionality of high-dimensional data, analyze the topological structure or shape of data, and finally cluster complex data [3].

This paper selects the application of TDA in text classification and face detection in order to illustrate the advantages of topology in data analysis. These two applications use three main methods about TDA , Persistent Homology, Mapper and Hausdorff Distance. In Persistent homology, a filtration of combinatorial objects, simplicial complexes, is constructed and then main topological structures of data is derived. The mapper is used to analyze the result as a simplificial complex which is interactive and can be quantified in several ways using statistics [3]. Simplicial complexes can be seen as higher dimensional generalization of graphs. They are mathematical objects that are both topological and combinatorial, a property making them particularly useful for TDA [5]. Then these two methods are applied to analyze authorship attribution (data set of poems) and obtain high accuracy results [3]. Another application about topological methods is that robust face detection based on enhanced Hausdorff Distance (HD) which provides higher efficiency and more reliability. In terms of algorithm complexity , HD is more faster.

2. Preliminaries

Definition 1 (Convex combination) If A_1, A_2...A_p are points in R^d. A convex combination is a point of the form $\lambda_1 A_+ \ldots + \lambda_p A_p$ with $\lambda_1 + \ldots + \lambda_p = 1$ and $\forall$ i $0 \leq \lambda_i \leq 1$.

The set of all convex combination of $A_1,\ldots,A_p$ is called the convex hall of $A_1,\ldots,A_p$. [1]

Definition 2 (Simplicail complex) A simplicial complex is a collection K of finite nonempty sets such that if A is an element of K, then so is every nonempty subset of A [2].

Definition 3 (Simplicial Homology). Given $n \in Z+$, the n-th homology group of a simplicial complex K, is denoted by Hn(K,F) and is defined as (1) [2] :

$$H_n(k,\mathbb{F}) := \frac{\mathbb{Z}_n(K,\mathbb{F})}{\beta_\bullet(K,\mathbb{F})}. \tag{1}$$

Definition 4 Hausdorff distance between A and B is defined by any of the two following equalities (2) [2]:

$$d_H(A,B)=max\{\sup_{b\in B} d(b,A),\sup_{a\in A} d(a,B)\}=\sup_{x\in M} | d(x,A)-d(x,B) | = \| d(.,A)-d(.,B) \|_\infty \tag{2}$$

3. TDA of Applications

3.1 New Text classification for Natural Language Processing

3.1.1 Difficulty in Text classification

Text classification is a hot research topic at present, and one of its difficulty is the high dimension of feature space. In high-dimensional feature space, the features may be redundant or unrelated, resulting in the inconvenience of high dimensional spatial processing, prone to over-learning, time and space overhead, without affecting the classification accuracy, it is necessary to carry out feature dimension reduction [6].

3.1.2 Process

In an experiment [3] that is in order to classify Persian poems which has been composed by two of the best Iranian poets namely *Ferdowsi* and *Hafez*. In this experiment, the author used two R packages, TDA and TDA status, for text classification. Those two R packages were implemented persistent homology. The textual data (poems) of two Iranian poets (Hafez and Ferdowsi) was used and the data set was gathered from *Shahnameh* and *Ghazaliat-e-Hafez* [3], which included about 8000 hemistich from each book. After preprocessing the data was fed to TF-IDF algorithm in order to make document term matrix, next the document term matrix was fed to the persistent homology algorithm.

First, it sketches persistent diagram, barcode and persistent landscapes for a sample of Ferdowsi poems including 1000 hemistich [3]. It is also divided hemistich of *hafez* into some parts, then it computed persistent diagram and first landscape of each part. Finally it sketched the mean landscape of these parts. Then it did same work for hemistich of *Ferdowsi*. At last step it computed Wasserstein distances between persistent Diagrams of correspondence parts of *hafez's* and *Ferdowsi's* poems.

The topological method about Mapper can be explained as follows: Suppose we have a point cloud data that represents a shape, such as a knot [3]. Firstly, we projected the whole data onto a coordinate system with a smaller dimension, so as to reduce the complexity by dimensionality reduction.Put the data into the overlapping receipt divided by parameter space, classify points by clustering algorithm, and finally create an interactive model.

The experiment examined two accuracy tests shape graph. Firstly, it partitions the whole graph into 3 clusters: Hafez, Ferdowsi and Both. In *Hafez* cluster it have the nodes which include the high percent of Hafezian poems, similarly in *Ferdowsi* cluster it has the nodes which include the high percent of poems of Ferdowsi and in the "Both" cluster it have about the same amount of both poems [3]. To do this, it simply divides the number of Hafezian poems in each node in the *Hafez* cluster by the number of all poems in each node in the same cluster, and it does the same test to other clusters as well.

3.1.3 Evaluation

The key of text classification is to reduce the dimension of unstructured data sets. Generally, feature selection and feature extraction are used to reduce the dimension. However, according to the existing experimental results made by other people, the degree of dimensionality reduction of data

space is not the same, which requires different methods to improve the accuracy and get better classification effect, so compared with TDA, it is more complex.

But the topological methods provide innovative data mining methods that can improve the efficiency of machine learning techniques. Some visualization tools about persistent homology, such as Persistent Diagram, Barcode and Persistent Landscape are invented to indicate the main topological features of data.

3.2 Robust Face Detection Using the Hausdorff Distance

3.2.1 Proposed Hausdorff Distance

Face detection is one of major research areas in AI.As one of the human identification features, facial features have the advantages of easy acquisition of sample images compared with finger prints and iris features.At present, the research of face detection is mainly aimed at static face detection, and the research object is often static face image without depth rotation [8].

In order to adapt to the closure of some sports fields, effectively use the continuous motion image sequence to improve the recognition efficiency, and minimize the decline of the recognition effect caused by the motion fuzzy image, it is meaningful to propose that the recognition suitable for dynamic situations is meaningful. A similarity measure using Hausdorff Distance(HD) can tolerate to perturbations in the point locations better than others [10]. This is because it measures the proximity rather than the exactness of superimposition. Previous research of applications of HD emphasized locating an object under translation and scaling [9]. In addition, many researchers have improved the performance of the conventional HD measure in terms of speed and accuracy.

3.2.2 Process

In the preprocessing step of dynamic face detection, Hausdorff Distance is used to locate the face image, which optimizes the next step to a certain extent.This section introduces an efficient implementation method for face location, works on grayscale still images, which is suitable for real-time applications.

This method presents a shape comparison approach to achieve fast, accurate face detection that is robust to changes in illumination and background. A two-step-process that allows both coarse detection and exact localization of faces is presented [7].

The specific step is to refine the facial parameters in the second stage after roughly detecting the facial region. On two large test-sets, a relative error is given to measure the performance of the system by comparing the estimated eye position with the manual eye position. Relative error measure that is independent of both the dimension of the input images and the scale of the faces [7]. The better location results show that the system has strong robustness under different backgrounds and illumination conditions. The run time behaviour allows the use in real time video applications [9].

3.2.3 Evaluation

Different from the traditional HD, Robust Hausdorff Distance(RHD) not only makes use of the position information of edge points, but also considers other types of information, such as the total number of edge points satisfying a directed distance and some pseudo-edges composed of very few edge points [7]. RHD takes occlusion and pseudo edge into account, which is not easily affected by blur in dynamic image recognition.

4. Conclusion

In this paper, we cite two implementations of TDA applications. One is about text classification of natural language, which uses persistent homology algorithm to analyze poetry data-sets. Applying a new method called Mapper to author attribution. The results are analyzed as a complex system, and these statistics can be quantified in many ways. The other application is an efficient algorithm for an automatic face detection system has been proposed. The Hausdorff distance is used as a similarity measure between a general face model and possible instances of the object within the image. The method performs robust and accurate face detection and its efficiency makes it suitable for real-time applications. The face detection algorithm is simple and less computation complexity

than traditional methods. The experimental results have shown that the algorithm is the most efficient approach in terms of speed, accuracy and reliability compared to others [10].

In conclusion, TDA can be used in many different fields widely. For example, persistent homology is a tool to study data sets and has been previously used in pulse crystal structures, analyzing 3D images, image analysis, analyzing breast cancer. Persistent homology now is used to understanding biological systems, as an algebraic tool for measuring high dimensional data to represent the topological features of point clouds. Researchers extend applications of computational homology to the analysis of genetic data from breast cancer patients [11]. These topological data analysis methods will be very useful, just like seemingly unrelated discrete points, we can mine out their topology and display the data vividly.

Acknowledgment

First and foremost, I would like to show my deepest gratitude to my teachers and professors in my university, who have provided me with valuable guidance in every stage of the writing of this thesis. Further, I would like to thank all my friends and roommates for their encouragement and support. Without all their enlightening instruction and impressive kindness, I could not have completed my thesis.

References

[1] F. Chazal, B. Michel, An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists, *math. ST,* vol.43, pp: 3-6, 2017.

[2] F. Memoli, K. Singhal, A Primer on Persistent Homology of Finite Metric Spaces, *math. AT,* vol.38, pp:7-13, 2019.

[3] N. Elyasi, An Introduction to a New Text Classifification and Visualization for Natural Language Processing Using Topological Data Analysis, 2019. https://arxiv.xilesou.top/abs/1906.01726 .

[4] R. Rivera-Castro, P. Pilyugina, P. Pletnev, I. Maksimov, W. Wyz and E. Burnaev, Topological Data Analysis of Time Series Data for B2B Customer Relationship Management, *cs. LG,* 2019. https://arxiv.xilesou.top/abs/1906.03956

[5] P. Bubenik, Statistical Topological Data Analysis using Persistence Landscapes, *Journal of Machine Learning Research,* vol. 25, pp: 77-102, 2015.

[6] T. Chen, Y. Xie, Literature Review of Feature Dimension Reduction in Text Categorization, *Information and learning newspaper,* vol. 24(6): pp: 690-694, 2005.

[7] O. Jesorsky, K. J. Kirchberg and R. V. Frischholz, Robust Face Detection Using the Hausdorffff Distance, *Lecture Notes in Computer Science,* pp. 90-95, 2001.

[8] S. Srisuk and W. Kurutach, New Robust Hausdorff Distance Based Face Detection, pp:1022-1025, 2001.

[9] Y. Wang, Image Matching Based on Robust Hausdorff Distance, *Journal of computer aided design and graphics,* vol.14(3), pp: 238-241, 2002.

[10] Y. Liu and L. Shen, Face Image Location Using Hausdorff Distance, *The research and development of the counter-computing machine,* vol.38(14), pp: 475-481, 2011.

[11] D. DeWoskin, J. Climent, I. Cruz-White, M. Vazquez, C. Park and J. Arsuaga, Applications of computational homology to the analysis of treatment response in breast cancer patients, *Topology and its Applications,* vol.157, pp: 157–164, 2010.