Journal of the Egyptian Mathematical Society 000 (2017) 1–8

Contents lists available at ScienceDirect

Journal of the Egyptian Mathematical Society

journal homepage: www.elsevier.com/locate/joems

Original Article

On the Topological Data Analysis extensions and comparisons

H.N. Alaa*, S.A. Mohamed

Department of Mathematics, Faculty of Science Aswan University, Egypt

ARTICLE INFO

Article history:
Received 1 April 2017
Revised 16 June 2017
Accepted 8 July 2017
Available online xxx

Keywords:
55N35
55U05
62H99
Topology – homology-persistence landscape
Topological Data Analysis
Testing hypotheses

ABSTRACT

Topological Data Analysis is an emerging field at the intersection of algebraic topology and statistical inference aimed at describing the shapes objects represented as point cloud data in the multidimensional space. Since the range of applications of shape analysis is enormous, new tests have given birth to the field of TDA. In this habilitation study three TDA-oriented tests are discussed. A new test based on metric functions is proposed. A small simulation study among the preceding tests has been employed via Monte Carlo simulation. All the mentioned tests in the vignette are activated by real world data within educational field.

© 2017 Egyptian Mathematical Society. Production and hosting by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license.
(http://creativecommons.org/licenses/by-nc-nd/4.0/)

1. Introduction

In a wide variety of disciplines, it is of great practical importance to measure, sketch and compare the shapes between different objects. Dryden and Mardia [1] defined the shapes of certain objects as all the geometric information that remains when location, scale and rotational effects are filtered out. If the size information is also of interest, then the scale will be omitted from the definition. Here the size of the information will be taken into consideration. In other words, we can claim that two objects have the same shape if by the translation, shifting or rotation operations the two objects will coincided, see [2]. The fundamental field concerning with studying the geometric properties of the objects is topology. Indeed, topology has been present in mathematics for quite a long time without anticipating applications to real-world applications until the beginning of this century. As, Carlsson in [3] proposed his survey article which produced another new area of research known as computational topology that enables the researchers to extract the quantitative and qualitative information that describe the point cloud data's shapes.

Computational topology is a set of algorithmic methods developed to understand topological invariants such as loops and holes in high-dimensional data sets. The specialized approach that employs the statistical tools to compute and analyze the topological features is called TDA. Generally speaking, TDA refers to a collection of methods and tools that enable the researchers for finding and studying the topological invariants structure in data. The input of these procedures typically takes the form of a point cloud data which is usually represented as a large finite dataset sampled from a geometrical object in a n-dimensional metric space, possibly with some noise. The output is a collection of data summaries and diagrams that are used to estimate the statistical features of the data. Lesnick [4] divided TDA tools into two parts: the first one is the descriptors TDA which are the procedures that aim at describing, summarizing, discovering, and visualizing point cloud data. However, the second is TDA inference which uses the probability theory to investigate or test the statistical features of the sample data (e.g. mean, variance…etc.).

In the last few years, community topology has witnessed important progress in supporting complex data analysis. In consequence, TDA plays a crucial role in a variety of different fields range from industry [5] shape classification Chazal et al. in [6, 7], clustering and histology images for breast cancer analysis [8]. In addition, TDA has received recently much attention by statisticians which gives a birth to a competitor approach in the data mining. For instance, Singh et al in [9] proposed a new classification tool based on simplicial complexes figures called Mapper, Kent et al. in [10] introduced k-tree level sets which can be utilized in the classification and comparison purposes, Turner [11] defined the means and medians for the persistent homology diagrams, from [12] derived confidence band for the persistence diagram that allows us to separate topological signal from topological noise, Chazal in [13] proposed sub-sampling methods for analyzing the shape of sets and functions from point cloud data in the case of the sample is too large.

* Corresponding author.
E-mail addresses: ala2222000@yahoo.com (H.N. Alaa), statisticsMS.2010@gmail.com (S.A. Mohamed).

http://dx.doi.org/10.1016/j.joems.2017.07.001
1110-256X/© 2017 Egyptian Mathematical Society. Production and hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license.
(http://creativecommons.org/licenses/by-nc-nd/4.0/)

Please cite this article as: H.N. Alaa, S.A. Mohamed, On the Topological Data Analysis extensions and comparisons, Journal of the Egyptian Mathematical Society (2017), http://dx.doi.org/10.1016/j.joems.2017.07.001

The major motivation beyond the present study is to provide a review for the three tests based on TDA using for testing the similarities between the objects. Further, propose a new test based on metric functions can be employed for the same purpose. In addition, conducting a power comparison study between the tests based on TDA and the proposed tests a benchmarking test. This article is structured as follows: the next section will give a snapshot of TDA tools. The third section includes all the tests that can be employed for testing the closeness between the objects. The followed section is devoted for the Monte Carlo results. The final section presents the results concerned to the real life applications.

2. Topological Data Analysis

The general framework of TDA for computing topological features from point cloud data usually contains two necessary steps: constructing simplicial complexes and applying TDA techniques on the simplicial complexes frequently are the persistent homology, barcodes and the persistent landscape. The main textbook for this section is Edelsbrunner and Harer [14].

A simplicial complex S is a set consisting of a finite collection of p-simplices (simple pieces), where a 0-simplex is a vertex, a 1-simplex is an edge, a 2-simplex is a triangle, a 3-simplex is a tetrahedron, and so on. In more precise way, the simplicial complex divided the space into smaller and topologically simpler pieces, which when assembled back together carry the same aggregate topological information as the original space. These simplices should satisfy two conditions. First, for every set σ in S, every nonempty subset $\tau \subset \sigma$ also belongs in S. For instance, if tetrahedron abcd is in S, then the triangles abc, abd, acd, bcd, the edges ab, ac, ab and the vertices a, b, c, d are also in S. Second, two p-simplices are either empty or they intersect in a lower dimensional simplex. In order to obtain simplicial complex sets, Vietoris–Rips filter is advocated in this study.

Homology is a tool from algebraic topology that measures the features of a topological space such as an annulus, sphere, torus, or more complicated surface. In particular, homology can distinguish these spaces from one another by quantifying their connected components, loops, voids, and so forth. One interesting feather associated with the homology group is the Betti numbers, as they provide meaningful information about the complex. Roughly speaking, the pth Betti number β_p is the number of pth dimensional independent holes in the homology groups, so that β_0 is the number of connectedcomponents, β_1 is the number of loops, β_2 is the number of enclosed voids and so on. Persistent homology is the primary algebraic topology tool was developed by Edelsbrunner et al. [29] used in the TDA methods in order to track long persist features. It provides a way to measure the lifespan of a topological feature, which is the persistence of the feature, whereas short-lived features may be ignored as noise.

A convenient way to visualize persistent homology is through a graphical representation called a barcode which can summarize the information encoded in the persistence diagram in a different vision. There is a distinct barcode for each homology space from which we infer the Betti number. In other words, the length of every line in the Barcodes diagrams refers to the distance between the time of death j and the time of born i, the number of the lines associated to dimension zero equals to β_0, while the number of the lines associated to dimension one equals to β_1 and so on.

Another graphical way that can summarize the information contained in the persistent homology diagram is the persistent Landscape proposed by Bubenik [15]. Persistent Landscape can be considered as a rotated version of barcode plot. The main advantage of the Persistent Landscapes is it allows us to calculate and summarize the data with the standard statistics indicators e.g. means, median, variance…etc, as opposite to either persistence diagram or barcode plot. To define the landscape, construct a triangle whose base corresponds to a persistence intervals and the top vertex by tenting each persistence point using the following function:

$$\Lambda_s(\varepsilon) = \begin{cases} \varepsilon - i & \varepsilon \in \left[i, \frac{i+j}{2}\right] \\ j - \varepsilon & \varepsilon \in \left(\frac{i+j}{2}, j\right] \\ 0 & otherwise \end{cases}$$

where ε is the filtered simplicial complex time and s takes 1 to n, n is the number of the points in the persistent diagram. It should be noted that $\Lambda_s(\varepsilon)$ obtained separately to each p-dimension. Formally, $\lambda_s(\varepsilon)$ is the sth largest value of $\Lambda_s(\varepsilon)$ taken into consideration the homology dimension. When $s = 1$, of course, $\lambda_s(\varepsilon)$ can be interpreted as the maximal possible distance of an interval centered about ε. Fig. 1 applied all the TDA's tools, mentioned above, to a sample drawn from tours.

3. Statistical shape analysis

Shape analysis is an active subject of academic research in the both of mathematical and applied sciences. It has extensive applications in many fields as it is great practical importance to carry out hypotheses tests that distinguish between objects under uncertainty. A plenty of tests have been suggested in the literature (see [2]). However, three different tests will be focused in this context. Assume that you have K-objects and that we would like to test the null hypothesis that all the objects are similar and have the same shape versus the alternative hypothesis that states that at least one object differs than the others. This can be achieved by the following tests which are so called k-sample tests.

3.1. Statistical inference using persistent homology

Gamble in [2] produced a new test which can be dependable for testing the similarity between two persistent homology diagrams using Wasserstein distance. Robinson and Turner in [16] generalized the test of Gamble in the multivariate case; as if it is required to test between two sets of persistent homology. In the present paper, it will generalize from [2], test into K samples. The test statistic that can be utilized to test between K persistent homology diagrams Pin the light of Gamble and Heo may be expressed as:

$$T_R = \frac{1}{\binom{k}{2}} \sum_{i=2}^{k} \sum_{j=1}^{i-1} W\left(P_i, P_j\right)$$

where $W(P_i, P_j)$ is the Wasserstein distance between P_i and P_j. Obviously, T_R can be considered as the average of all pair wise Wasserstein distances. Robinson in [2] recommended using the Hungarian algorithm to compute the Wasserstein distance.

Given $p_{1,1}, p_{2,1} \dots p_{n_1,1}$ and $p_{1,2}, p_{2,2} \dots p_{n_2,2}$ are the points corresponding to P_1 and P_2 respectively.The Hungarian algorithm required, first, that the two persistent homology have to be the same size, this is done via adding n_2 points to the first sample and n_1 points to the second sample, which yields we have $n_1 + n_2$ points for the both persistent homology. The added points are copy of a diagonal that are the perpendicular distances. Then, constructing the cost matrix where its entries are the squared Euclidean distances. Next, match every row with the optimum column.[1] Finally, the Wasserstein distance is the sum up for the optimum distances,

[1] The optimum column means that the column that has least distance.

Please cite this article as: H.N. Alaa, S.A. Mohamed, On the Topological Data Analysis extensions and comparisons, Journal of the Egyptian Mathematical Society (2017), http://dx.doi.org/10.1016/j.joems.2017.07.001

Download English Version:

https://daneshyari.com/en/article/6898940

Download Persian Version:

https://daneshyari.com/article/6898940

Daneshyari.com