APPLICATION OF PERSISTENT HOMOLOGY ON BIO-MEDICAL DATA – A CASE STUDY

Eng. Sekuloski, P., Ass. Prof. Dr. Dimitrievska Ristovska, V.
Faculty of Computer Science and Engineering, ”Ss. Cyril and Methodius” University, Skopje, Macedonia
petar.sekuloski@finki.ukim.mk, vesna.dimitrievska.ristovska@finki.ukim.mk,

Abstract: In this paper we introduce, analyze and apply persistent homology, one of the main algorithms of TDA, on some real data sets from the bio-medical field. Topological data analysis (TDA) is a field which is a synergy between mathematics, data science and computer science. The main goal of TDA is studying the shape of data using topological techniques. TDA proposes new algorithms that deal with these problems based on tools or concepts from algebraic topology and pure mathematics. We analyze the results and give a topological characterization of the dataset and propose to use them in future work.

Keywords: PERSISTENT HOMOLOGY, TOPOLOGICAL DATA ANALYSIS, ALGEBRAIC TOPOLOGY, DATA SCIENCE, COMPUTATIONAL TOPOLOGY

1. Introduction

Topology is a mathematical field that studies properties of topological spaces, such as connectedness and compactness, invariant of continuous deformations. Algebraic topology studies topological spaces using techniques from algebra by associating algebraic objects such as groups with topological spaces. One of the main tools of algebraic topology is homology. Homology is a mathematical tool which associates sequences of algebraic objects with topological spaces. One way to study a topological space is to find and compute its homology groups. The motivation behind defining homology groups was that two shapes can be distinguished by examining their holes. For example, a disk is different from a circle, or a disk is not a circle, because the disk is solid while the circle has a hole through it. Homology groups are set of invariants of a topological space. These invariants characterize the topological space. The number of structures for some dimension k is the rank of the k-dimensional homology group of the topological space. The number of such structures is known as a Betti number (β_k) of dimension k.

The main idea of Topological Data Analysis is application of these mathematical concepts on real data. Persistent homology is an algorithm from TDA that use homology as main idea. The algorithm computes topological features of a space.

2. Mathematical Background

The starting point is to construct a topological space from a given dataset. We will define some necessary mathematical concepts.

Definition 1. A $\boldsymbol{k}$-simplex is a convex hull of $\boldsymbol{k+1}$ affinely independent points $\boldsymbol{S = \{x_0, x_1,, \dots, x_k\}} \subseteq \mathbb{R}^d$. The points of S are vertices of the simplex**.**

The low dimensional simplices (plural: simplices or simplexes) have special names:

- a 0-simplex is called a *vertex;*
- a 1-simplex is called *an edge;*
- a 2-simplex is called *a triangle:*

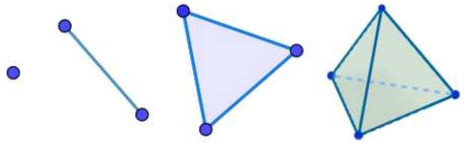

Figure 1. 0-simplex 1-simplex, 2-simplex, 3-simplex

Definition 2. Let σ be a k-simplex defined $\boldsymbol{on\ S} = \{\boldsymbol{x_0}, \boldsymbol{x_1},, \dots, \boldsymbol{x_k}\}$. A simplex $\boldsymbol{\tau}$ defined by $\boldsymbol{T} \subseteq \boldsymbol{S}$ is a face of $\boldsymbol{\sigma}$ and has $\boldsymbol{\sigma}$ as a coface. The relationship is denoted with $\boldsymbol{\sigma} \geq \boldsymbol{\tau}$ and $\boldsymbol{\tau} \leq \boldsymbol{\sigma}$.

Definition 3. Let K be a set. Simplicial complex S is a collection of subsets of $\boldsymbol{K}$ called simplices such that:

1. For all $\boldsymbol{x} \in \boldsymbol{K}, \{x\} \in \boldsymbol{S}$.

2. If τ ⊆ σ ∈ S, then τ ∈ S.

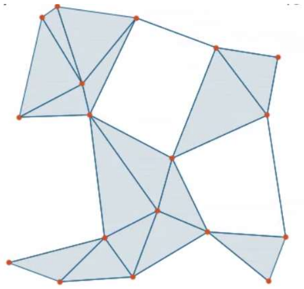

Figure 2. An example of a simplicial complex

We call the sets {x} the vertices of K. **Definition 3** gives a more abstract definition of simplicial complex that can be applied to a data where vertices will be the data points. Topological invariants of the space, such as holes and number of connected components, can be computed from a simplicial complex, see Figure 2. One of the key ideas of TDA is to construct a simplicial complex from a dataset. There are a few ways to construct such a simplicial complex [1]. In other words simplicial complexes are high dimensional analogues of graphs. We will explain the steps of the process.

1. Construction of a topological space from a given point cloud

The open (metric) ball of radius ε>0 centered at a point $m \in M$, usually denoted by $B(m; \varepsilon)$ is defined by

$$B(m; \varepsilon) = \{n \in M \mid d(m,n) \leq \varepsilon\}$$

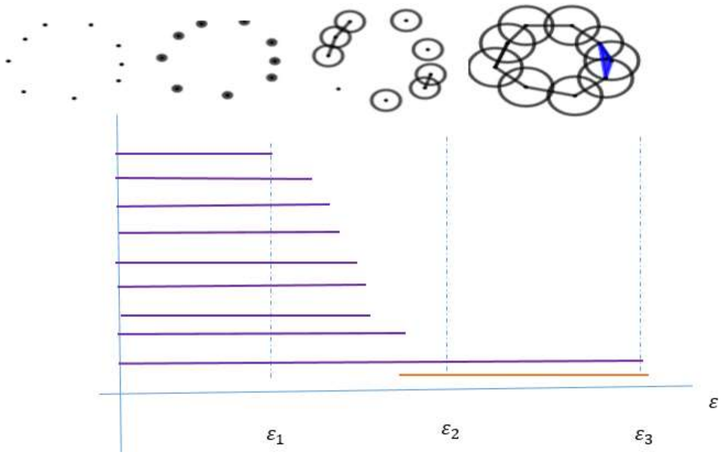

Figure 3. An example of Vietoris-Rips filtration of a space. There are different complexes for different values for ε. Violet horizontal lines shows barcodes in dimension 0 and orange line shows barcode for dimension 1.

Let $\boldsymbol{M}$ be a point cloud in $\mathbb{R}^{\boldsymbol{d}}$ and $\boldsymbol{\varepsilon} > \mathbf{0}$. The $\boldsymbol{\varepsilon}$ −neighborhood of the point cloud $\boldsymbol{M}$ is the set $\boldsymbol{S}(\boldsymbol{m}; \boldsymbol{\varepsilon})$, defined as

$$\boldsymbol{S}(\boldsymbol{m}; \boldsymbol{\varepsilon}) = \bigcup_{\boldsymbol{m}\boldsymbol{\varepsilon}\mathbf{M}} \boldsymbol{B}\,(\boldsymbol{m}, \boldsymbol{\varepsilon}), \qquad \boldsymbol{\varepsilon} \geq \mathbf{0}.$$

It is known that every $\boldsymbol{\varepsilon}$ −neighborhood is a topological space. PH gives a summary of a sequence of such topological spaces for different values for $\boldsymbol{\varepsilon}$. The key idea here is to see how topological characteristics are changing and which features are the same as $\boldsymbol{\varepsilon}$ increases.

2. Construction of a simplicial complex from topological space

In our experiments we will use Vietoris-Rips complexes. For a given point cloud $\boldsymbol{M}$ and $\boldsymbol{\varepsilon} \geq \mathbf{0}$ we construct Vietoris-Rips complex denoted as $\boldsymbol{VR}(\boldsymbol{M}; \boldsymbol{\varepsilon})$. $\boldsymbol{VR}(\boldsymbol{M}; \boldsymbol{\varepsilon})$ is defined as:

$$\boldsymbol{VR}(\boldsymbol{M}; \boldsymbol{\varepsilon}) = \bigcup_{\boldsymbol{n}\geq\mathbf{0}} \boldsymbol{VR}(\boldsymbol{M}; \boldsymbol{\varepsilon})_{\boldsymbol{n}}$$

$$\boldsymbol{VR}(\boldsymbol{M}; \boldsymbol{\varepsilon})_{\boldsymbol{n}} = \left\{(\boldsymbol{m}_{\mathbf{0}}, \ldots, \boldsymbol{m}_{\boldsymbol{n}}) \middle| \boldsymbol{d}(\boldsymbol{m}_{\boldsymbol{i}}\boldsymbol{m}_{\boldsymbol{j}}) \leq \boldsymbol{\varepsilon}, \textbf{for all i, j } \in \{\mathbf{1}, \mathbf{2}, \ldots, \mathbf{n}\}\right\}$$

Note that $\boldsymbol{VR}(\boldsymbol{M}; \boldsymbol{\varepsilon})_{\boldsymbol{n}}$ is the set of all n-simplexes of the simplicial complex. The simplicial complex constructed from the topological space is the approximation of the topological space. Hence, every simplicial complex is a topological space which is why we can analyze its topological features.

3. Computing and representing homology groups

Linear algebra is used for computing homology groups of a given simplicial complex. The $\boldsymbol{k}^{\boldsymbol{th}}$ homology group $\boldsymbol{H}_{\boldsymbol{k}}$(S) of a simplicial complex $\boldsymbol{S}$ is defined as abelian quotient group. The rank of the $\boldsymbol{H}_{\boldsymbol{k}}$, $\boldsymbol{rank}(\boldsymbol{H}_{\boldsymbol{k}}(\boldsymbol{S}))$, is called $\boldsymbol{k}^{\boldsymbol{th}}$ Betti number of $\boldsymbol{S}$. It gives a measure of the number of k-dimensional holes in S. The homology groups are computed for every simplicial complex derived from the topological space for each $\boldsymbol{\varepsilon}$. Thus, by increasing $\boldsymbol{\varepsilon}$ we can trail elements of homology groups of the corresponding complex $\boldsymbol{VR}(\boldsymbol{M}; \boldsymbol{\varepsilon})_.$ We can visualize the existence of homology groups as $\boldsymbol{\varepsilon}$ increases using a persistent barcode. Persistent barcode is a topological summary of a topological space. When an element shows at some $\boldsymbol{\varepsilon}$, we say that an element is born and denote that $\boldsymbol{\varepsilon}$ as $\boldsymbol{\varepsilon}_{\boldsymbol{birth}}$. When the element disappears at some $\boldsymbol{\varepsilon}$ (it is mapped to 0), we say that the element has died and we denote that $\boldsymbol{\varepsilon}$ as $\boldsymbol{\varepsilon}_{\boldsymbol{death}}$. Every element is represented with a “bar” (a line in the persistent barcode) on the interval $[\,\boldsymbol{\varepsilon}_{\boldsymbol{birth}}.\ \boldsymbol{\varepsilon}_{\boldsymbol{death}})$. For example, in $\boldsymbol{H}_{\mathbf{0}}$, this will correspond to the formation of a connected component in the simplicial complex at $\boldsymbol{\varepsilon}_{\boldsymbol{birth}}$ and connecting that component with others in a way that they will form a circle in $\boldsymbol{\varepsilon}_{\boldsymbol{death}}$, see Figure 3. If we observe the Figure 3, we can see that the orange line is a bar which corresponds to an element of a homology group of dimension 1, which appears near $\boldsymbol{\varepsilon}_{\mathbf{2}}$. It clearly be seen that there is one circle at the last simplex. Also, we can see that near $\boldsymbol{\varepsilon}_{\mathbf{2}}$ there is one violet line which means that we have one connected component which corresponds with the given simplex.

3. Diabetes datasets

For this case study we picked two diabetes datasets. First dataset is the Miller-Reaven dataset. Reaven and Miller (1979) examined the relationship among blood chemistry measures of glucose tolerance and insulin in 145 non-obese adults [10]. They used the PRIM9 system to visualize the data in 3D, and discovered a peculiar pattern that looked like a large blob with two wings in different directions. In this dataset, the data is split up in three categories. Data from non-diabetic patients, data from patients with diabetes classified as overt and data from patients with diabetes classified as chemical diabetes. Overt diabetes is the most advanced stage, characterized by elevated fasting blood glucose concentration and classical symptoms. Preceding overt diabetes is the latent or chemical diabetic stage, with no symptoms of diabetes but demonstrable abnormality of oral or intravenous glucose tolerance. There are 145 observations on the following 6 variables:

relwt

relative weight, expressed as the ratio of actual weight to expected weight, given the person's height, a numeric vector

glufast

fasting plasma glucose level, a numeric vector

glutest

test plasma glucose level, a measure of glucose intolerance, a numeric vector

instest

plasma insulin during test, a measure of insulin response to oral glucose, a numeric vector

sspg

steady state plasma glucose, a measure of insulin resistance, a numeric vector

group

diagnostic group, a factor with levels Normal, Chemical_Diabetic, Overt_Diabetic.

4. Preliminary results and discussion

First, we apply persistent homology for each diabetic group of data. For the Chemical_Diabetic group the results are given in Figure 4 and for Overt_Diabetic group the results are given in Figure 5.

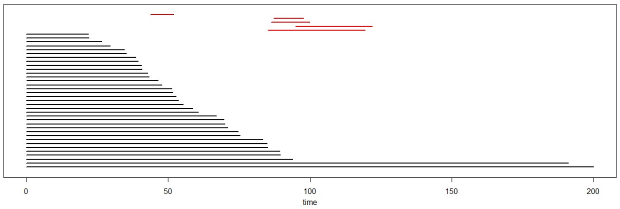

Figure 4. Persistent barcode for the Chemical_Diabetic group

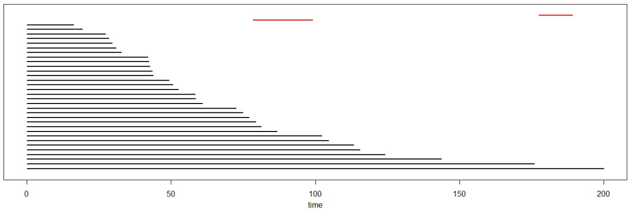

Figure 5. Persistent barcode for the Overt_Diabetic group

We can see that the persistent barcodes are different. In Figure 4, the persistent barcode has more red bars, which means that there are more circles in the simplex constructed from the data for the Chemical_Diabetic group. In this case, there is significant topological difference in the simplexes which means the shape of the data of the two groups is different. A question that arises here is which physical or real factor makes the difference? These factors may be crucial for better understanding the different types of diabetes.

Next, we apply persistent homology on both the diabetic group and the non-diabetic group. The results are given in Figure 6 and Figure 7.

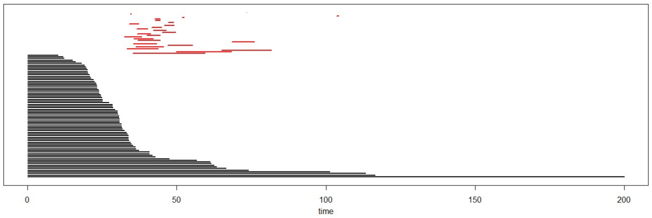

Figure 6. Persistent for non-diabetic group

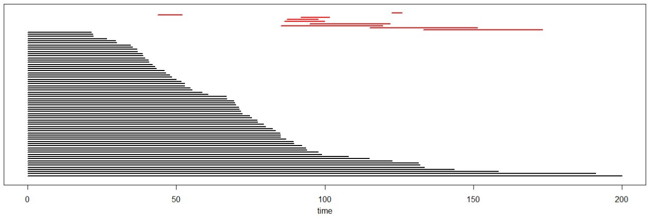

Figure 7. Persistent for diabetic groups

According to the barcodes in Figure 6 and Figure 7, we can conclude that topological characteristics in the data of diabetic and non-diabetic groups are obvious. In the second persistent barcode, there are circles which are present most of the time.

We apply persistent homology on the second dataset which contains data from diabetic and non-diabetic patients. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients in this dataset are females at least 21 years old and of Pima Indian heritage. The results are given in Figure 8 and Figure 9.

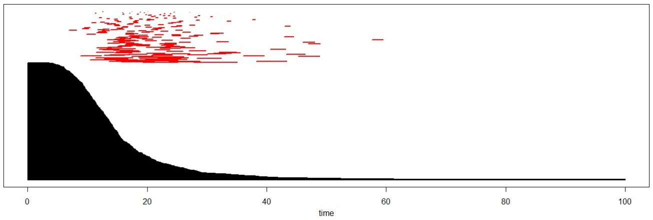

Figure 8. Persistent barcode for non-diabetic data

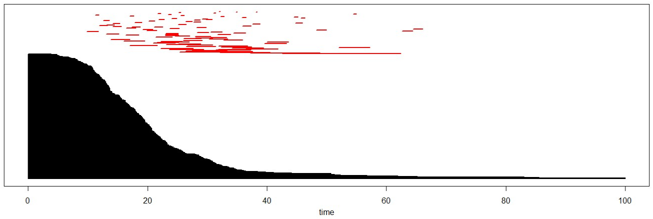

Figure 9. Persistent barcode for diabetic data

5. Further work and application in bio-medical field

The main goal is to link the differences of the topological characterizations of the two types of diabetes to real factors. Persistent homology, and in general, TDA, can be applied in the bio-medical field in many areas. The application of statistics allowed significant progress in understanding diseases. Knowing that, and the fact that TDA gives a new way of analyzing the data, specifically, analyzing the shape of the data, we think that TDA will be useful for medicine. It can be used to see how one factor changes the topological characteristics of the topological space underneath the given data, and how it is related to a disease. If we work in three dimensional Euclidean space, we may find some structural deformations of a system in the body. For example, to observe the deformations of the vasculature of some organ or tissue. In the future, we will investigate how persistent homology can be applied to characterize retinal and liver vasculature networks. TDA can also be applied on big data from the healthcare field.

6. Acknowledgement

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius University", Skopje.

References

[1] H. Edelsbrunner, “Persistent homology: theory and practice”, 2014.

[2] Gunnar Carlsson, “Topology and data”. Bulletin of the American Mathematical Society. 46 (2), 2009, pp. 255–308.

[3] J. R. Munkres, Topology. vol. 2. Upper Saddle River: Prentice Hall, 2000

[4] llen Hatcher, Algebraic topology. Cambridge University Press, 2002

[5] G. Carlsson, A. Zomorodian, A. Collins, L. Guibas, J. (2005-12-01). "Persistence barcodes for shapes". International Journal of Shape Modeling.

[6] Brittany Terese Fasy, Jisu Kim, Fabrizio Lecci, and Clément Maria. “Introduction to the r package tda. “, arXiv preprint arXiv:1411.1830, 2014.

[7] https://s3.amazonaws.com/cdn.ayasdi.com/wp-content/uploads/2018/11/12131418/TDA-Based-Approaches-to-Deep-Learning.pdf

[8] https://en.wikipedia.org/wiki/Fluoroscopy

[9] Ulrich Bauer and Michael Lesnick.” Induced matchings of barcodes and the algebraic stability of persistence. In Proceedings of the thirtieth annual symposium on Computational geometry”, p. 355, 2014.

[10] Reaven, G. M. and Miller, R. G. (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis. Diabetologia, 16, 17-24.

[11] A. J. Zomorodian, Topology for Computing, Cambridge, 2005

[12] J. Nicponski and J.-H. Jung, Topological data analysis of vascular disease: A theoretical framework, BioRxiv, (2019), p. 637090.

[13] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, Stability of persistence diagrams, Discrete & Computational Geometry, (2007), pp. 103–120

[14] A. Zomorodian and G. Carlsson, Computing persistent homology, Discrete & Computational Geometry, 33 (2005), pp. 249–274