MAPPER ALGORITHM AND IT'S APPLICATIONS

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

Abstract: In this paper we analyze and apply one of the main algorithms of TDA (Topological Data Analysis), Mapper, on some real data sets. We use Mapper for visualization of a data sets, and we tend to get some insights if some key characteristics of the data are captured by the visualization and how they are connected with human perception of the data. Also, we will discuss if the visualization can make progress in further work.

Keywords: MAPPER ALGORITHM, TOPOLOGICAL DATA ANALYSIS, ALGEBRAIC TOPOLOGY, DATA SCIENCE, COMPUTATIONAL TOPOLOGY

1. Introduction

Topological data analysis (**TDA**), is an approach for analyzing data using techniques from topology. Extraction of an information from the datasets which are high-dimensional, incomplete or noisy, is a wide field for researchers and scientists in last few years. TDA provides a general framework to analyze such data in a manner that is insensitive to a particular metric. Beyond this, it inherits functors , a fundamental concept of modern mathematics, which allows it to adapt to new mathematical tools.

Mapper algorithm was developed by Singh, M´emoli, and Carlsson in [1], and it gives a multi-resolution, low dimensional picture of point cloud. It's highly customizable, and has a track record of revealing structure that some other methods, like clustering and "projection pursuit" methods miss.

Mapper algorithm is one of the most important tools used in TDA for data visualization. For **input**, it use:

-point cloud;

- "filter function;"

-covering of a metric space;

-clustering algorithm;

- various other parameters.

Output is a Graph (or higher simplicial complex) which is tend to capture the main topological aspects of the point cloud.

2. Mathematical preliminaries

We will introduce some mathematical concepts, in order to construct a topological space from given dataset.

Let $n \geq 1$ be an integer, let $[n] = \{0, \ldots, n\}$.

An **n-simplex** σ is the convex hull of $n + 1$ affinely independent vertices $S = \{v^i\}, i \in [n]$ in R^d where $d \geq n$.

A simplex τ defined by $T \subseteq S$ is called a face.

A simplicial complex K is a finite set of simplices which meet along faces, every one of which is in K.

Let e^0 denote the origin in R^n and e^i the i-th standard basis vector for R^n.

The **standard n-simplex** $\Delta^n \subset R^n$ is the convex hull of $\{e^i\} i \in [n]$. Given any subset $J \subseteq [n]$, let Δ^J be the face of Δ^n spanned by $\{e^j\} j \in J$. The points of S are vertices of the simplex.

As basic examples, the low dimensional simplices (plural:simplices or simplexes) have special names:

- a 0-simplex is called a *vertex;*
- a 1-simplex is called *edge;*
- a 2-simplex is called *triangle*
- a 3-simplex is called *tetrahedron*,
- a 4-simplex is called a *5-cell*.

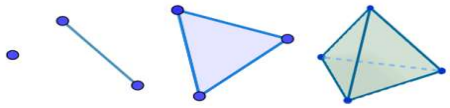

Figure 1. 0-simplex 1-simplex, 2-simplex, 3-simplex

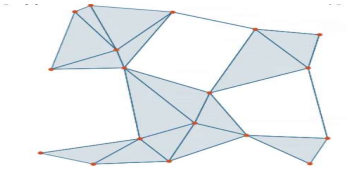

Figure 2. Example of simplicial complex

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 basic idea of Topological Data Analysis is to construct a simplicial complex from a dataset, i.e. in one hand, simplicial complexes are high dimensional analogues of graphs, and in other hand simplicial complexes are approximation of the topological space.

3.Mapper algorithm

The algorithm works very simple: put bin data into overlapping bins, cluster each bin, create a graph where vertices = clusters and two clusters are connected by an edge if they have points in common.

Mapper algorithm (implementation)

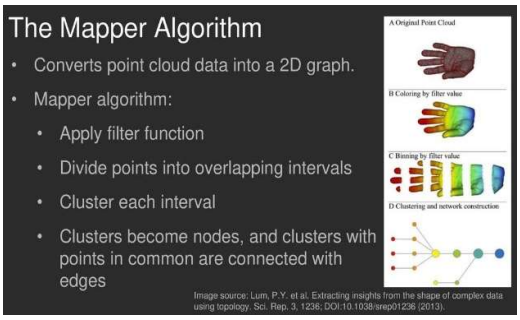

Figure 3 Mapper algorithm – steps

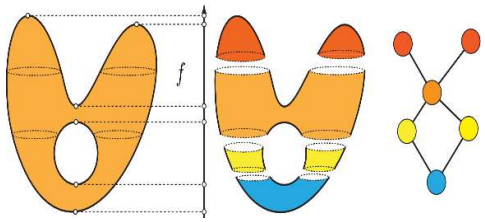

Figure 4. How works the Mapper algorithm – an illustration

Next, it is given a more precise description of the algorithm.

Given: X point cloud, |X| = N,

filter function f : X → R.

Assume we can always compute inter point distances.

Let I denote the "range" of f: explicitly I = [m, M] ⊂ R where m = $\min_{x \in X}$ {f (x)}, M = $\max_{x \in X}$ {f (x) }

Divide I into a set S of smaller intervals (of uniform length) which overlap. Obtain two resolution controlling parameters: *l* the length of the intervals, and *p* the percentage overlap between successive intervals.

- For each interval $I_j \in S$, let $X_j := \{x : f(x) \in I_j\}$. Then the collection of all such X_j is a covering of X.
- (2) For each X_j , perform a clustering algorithm to obtain clusters $\{X_{jk}\}$.
- Each cluster defines a vertex of our simplicial complex: draw an edge between vertices whenever $X_{jk} \cap X_{lm} \neq \emptyset$.

4. Application of the Mapper algorithm on the Tori (two rings) dataset

In this section, we choose 3D object form of two rings (tori), see Figure 5. It's synthetic dataset, consisted of 2048 points. We apply Mapper algorithm on that dataset.

In these experiments, made in mathematical software *R*, we use different values of the parameters:

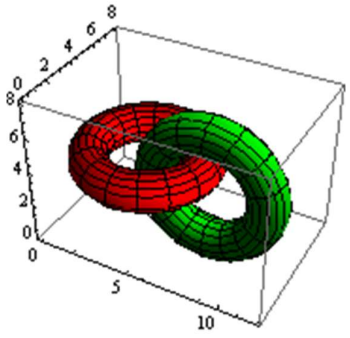

Figure 5. Tori

- *n*=number of intervals, varying between 6 and 16,
- *p*=percent of overlapping, between 20 and 80,
- *b*=number of overlapping bins when clustering, between 5 and 15.

The results from Mapper algorithm for Dvatorusi datset are given in Figure 6, Figure 7, Figure 8 and Figure 9. Every figure corresponds to a Mapper algorithm results for different parametars.

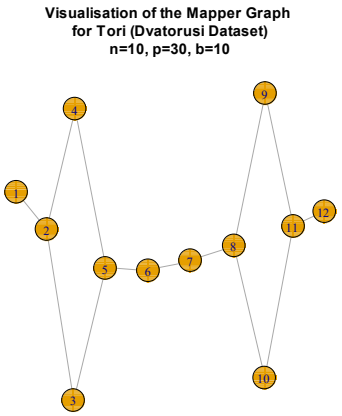

Figure 6. Mapper algorithm on Tori- 2 obtained cycles

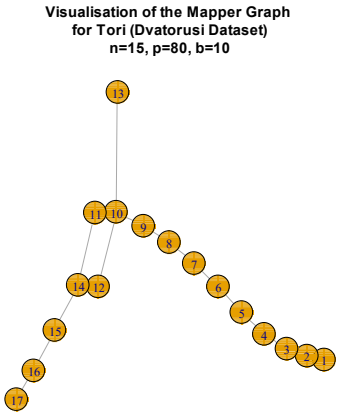

Figure 7. Mapper algorithm on Tori - 1 obtained cycle

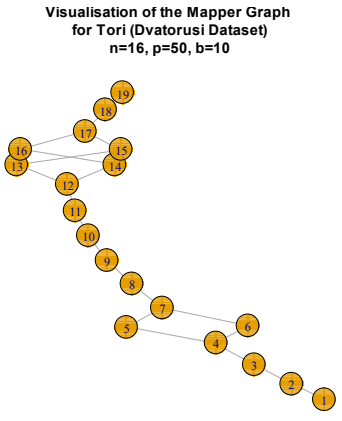

Figure 8. Mapper algorithm on Tori - 3 obtained cycles

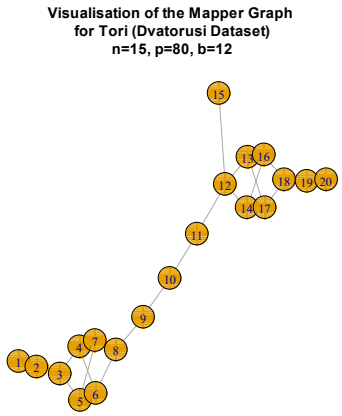

Figure 9. Mapper algorithm on Tori - 4 obtained cycles

We can conclude that there are different graphs obtained for different values of parameters. There is no one way of choosing parameters of Mapper algorithm. It depends on the subject of the research.

5. Application of the Mapper algorithm on the Torus dataset

In this section, we choose Torus- 3D object see Figure 10 and apply Mapper algorithm over the database.

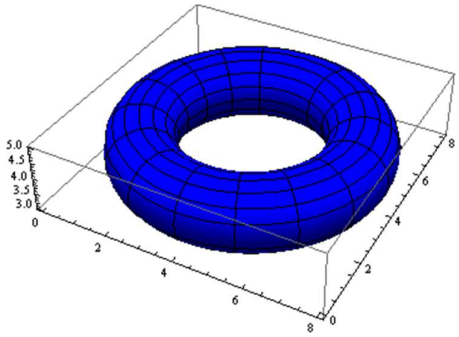

Figure 10. Torus

In these experiments, it is interesting that Mapper graphs with one dimensional filter, for all values of the different parameters, are of the same form, showed in Figure 11.

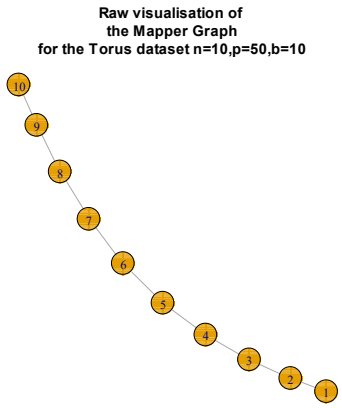

Figure 11. Mapper algorithm on Torus dataset

But, if the filter is bi-dimensional [11], the obtained Mapper graph is of the form, showed in Figure 12.

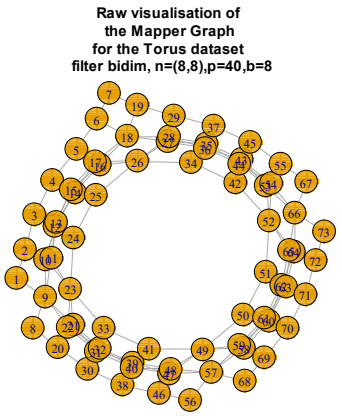

Figure 12. Mapper algorithm on Torus- bi-dimensional, filter

6. Application of the Mapper algorithm on the Diabetes dataset

In the following case, we apply Mapper algorithm on the Diabetes dataset, consists of 145 lines, with 6 attributes in each line- 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 visualized 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.

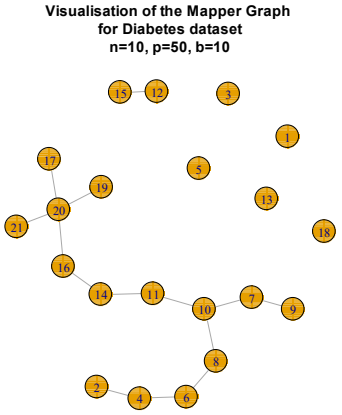

Figure 13. Mapper algorithm on Diabetes dataset (n=10)

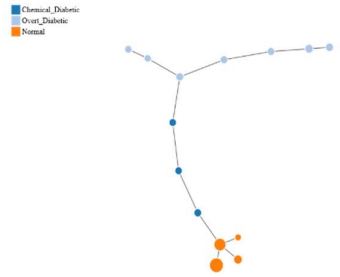

Figure 14. Colored Mapper Graph over Diabetes dataset (from Fig.13)

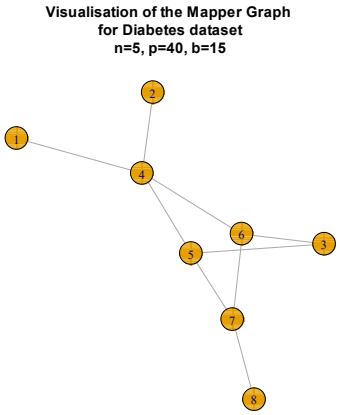

Figure 15. Mapper Graph on Diabetes data set (n=5)

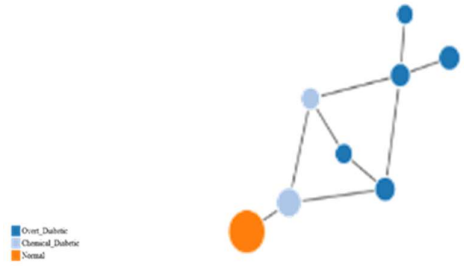

Figure 16. Colored Mapper Graph on Diabetes dataset (from Fig.15)

The peculiar pattern visualized in [10], can be seen on Figure 14 and Figure 16. The two types of diabetes are distinguished on the obtained graphs.

7. Discusion

Mapper algorithm is useful tool for visualization of datasets. There are many open problems in the process of choosing parameters, as it can be seen on the visualizations in this work. It is open research area. In the future, we like to optimize that process. Also, we plan to apply Mapper algorithm on bio-medical data and used it for categorize or group observations of some diseases.

8. Acknowledgement

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

Literature

[1] Gurjeet Singh, Facundo Mémoli and Gunnar Carlsson, "Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition", Eurographics Symposium on Point-Based Graphics, 2007

[2] Andrew W. Herring, "The Mapper Algorithm", Western TDA Learning Seminar, Department of Mathematics, Western University, 2018

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

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

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

[6] http://archive.ics.uci.edu/ml/datasets/statlog+(heart)

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

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

[9] https://en.wikipedia.org/wiki/Topological_data_analysis

[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] http://bertrand.michel.perso.math.cnrs.fr/Enseignements/TDA/ Mapper .html

[12] Mariam Pirashvili, Lee Steinberg, Francisco Belchi Guillamon, Mahesan Niranjan, Jeremy G. Frey, Jacek Brodzki, "Improved understanding of aqueous solubility modeling through topological data analysis", Journal of Cheminformatics, 2018