Amelioration of Clustering Algorithms Using Topological Methods

Xijia Tao

August 2019

Abstract

Topological data analysis(TDA) is a burgeoning field that concerns with the measurement and representation of shapes in the context of large, complex and high-dimensional data sets. While machine learning algorithms alone are capable of uncovering critical patterns and relationships within a data set, TDA can provide an extremely simple and efficient method of partitioning data to comprehend its underlying properties and reduce the possibility of missing critical insights. This paper will introduce a few selected theoretical aspects of topology, including one of the best methods in TDA, Persistent Homology, and discuss their applications in enhancing the performance and efficiency of clustering algorithm using Betti numbers.

Keywords: Topological Data Analysis, Persistent Homology, Unsupervised Machine Learning, Clustering, Data Analysis

1 INTRODUCTION

With the advances in information technology, every and each one of us is now bombarded with the ever-growing amount of data in our daily life. A tremendous amount of information will have to be elaborately analyzed from multiple facets and various approaches so as to, for example, manage a company's interactions with current and future customers[1] or obtain a better comprehension of force fields in granular media[2]. Apparently, applying conventional techniques of data analysis would be insufficient and inefficient when it comes to dealing with complex and high dimensional data and transforming it into a simpler version. Here, is where topological methods are introduced.

Topology is the branch of pure mathematics that studies the notion of shape. It permits one to discover shape related properties within the data, such as the presence of loops, and it provides methods for creating compressed representations of data sets that retain features, and which reflect the relationships among points in the data set. The representation is often in the form of a combination of simplicial complexes, which are very simple and intuitive objects to work with using Vietoris-Rips construction algorithm. Due to the unique methodology of TDA, it can solve some difficult and intriguing problems in data science as a clustering method that is robust to perturbations. One of the vitally important techniques based on topology is Persistent Homology, which is a filtration of combinatorial objects, simplicial complexes, After constructing a persistent homology, major topological features of a data set are derived and can be represented by tools like "Betti Numbers" and "Persistent Landscape".

In this paper first we explain the math of simplicial complexes and its homology, and thus introduce the concept of Betti numbers. Next we elaborate the essence of persistent homology and how it works in TDA. The methodology of summarizing extracted topological features will be introduced. The last section describes the CBN(Clustering using Betti Numbers) algorithm in comparison with other conventional methods of clustering in data analysis.

2 Preliminaries

In this section, we define and explain the pre-requisites needed to develop the theory of persistent homology.

2.1 Metric Spaces

As topological and geometric features are often associated to continuous spaces, we are not able to observe any topological information per se directly from finite data sets. A natural approach to highlight some topological structures out of data is to "connect" data points that are intuitively close to each other so as to uncover a global continuous shape underlying the data. In TDA, convenient methods of quantifying the notion of closeness between data points typically involve the definition of a distance between each pair of elements within a set by considering data sets as discrete *metric spaces*.

Definition 1(Metrics)**.** *A metric on a set E is a map d: $E \times E \to \mathbb{R}$*

Definition 2(Metric Spaces)**.** *A metric space (E, d) is a set E with a metric d, such that for any $x, y, z \in E$:*

i) $d(x, y) \geq 0$

ii) $d(x, y) = d(y, x)$

iii) $(d(x, y) = 0) \iff (x = y)$

iv) $d(x, z) \leq d(x, y) + d(y, z)$

Examples of Common Metrics with $E = \mathbb{R}^n$:

Taxicab/Manhattan Distance:

$$d_1(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Euclidean Distance:

$$d_2(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Chebyshev/Supremum Distance:

$$d_\infty(x, y) = \max_{1 \leq i \leq n} |x_i - y_i|$$

It is an easy exercise to prove that the three distances mentioned above are legitimate metrics by showing that they all satisfy the four criteria defining a metric space.

A Simple Illustration: With $E = \mathbb{R}^2$, consider the point $(-10, 3)$ and the point $(15, -8)$. By definition, d_1 is the sum of differences between corresponding coordinates of two points, i.e. $d_1\left((-10, 3), (15, -8)\right) = |-10 - 15| + |3 - (-8)| = 36$. d_2 is the straight-line distance between two points and can be calculated using Pythagorean theorem, i.e. $d_2\left((-10, 3), (15, -8)\right) = \sqrt{(-10 - 15)^2 + (3 - (-8)^2} = \sqrt{746} = 27.3$ (to 3 s.f.). Lastly, d_∞ is the maximum difference between corresponding coordinates of two points in a plane, i.e. $d_\infty\left((-10, 3), (15, -8)\right) = \max\limits_{1 \leq i \leq 2} |x_i - y_i| = \max\{|-10 - 15|, |3 - (-8)|\} = \max\{25, 11\} = 25$

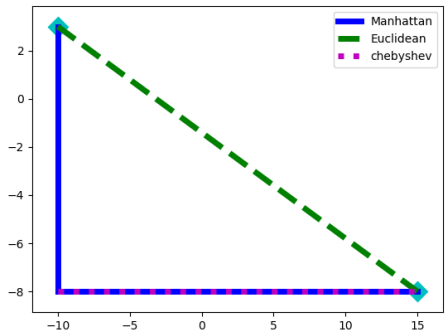

Figure 1: A Simple Illustration: It is straightforward to calculate d_1, d_2 and d_∞ between the two points.

Taking the distance between two data points in a metric space into consideration, one is able to construct a neighbouring graph through connecting pairs by edges. Consequently, the connectivity of the data can be analyzed from the graph, e.g. using some clustering or dimensionality

reduction algorithms. To explore other valuable topological properties other than connectivity, a central idea in TDA is to build higher dimensional equivalent of neighboring graphs in a topological sense by not only connecting pairs but also (k + 1)-uple of nearby data points.[3] The resulting objects are called simplicial complexes, which we will introduce in the next section.

2.2 Simplicial Complexes

Since it is generally difficult to extract desired information from a set of discrete data points, we can equip the point cloud $\mathbb{X}$ with an *simplicial complex* so as to approximate the topological structure or shape underneath the data set and therefore discover its topological and geometric features that we concern.A simplicial complex is a topological object that can be considered as a union of vertexes, edges, triangles, tetrahedrons and their higher dimensional counterparts.

Definition 1(Simplicial Complex)**.** *A simplicial complex is a collection K of finite non-empty sets such that if A is an element of K, then so is every non-empty subset of A.*

Definition 2(Simplex)**.** *Every element of a simplicial complex K is a simplex of K.*

Definition 3(Dimension of Simplicial Complex)**.** *The dimension of a simplicial complex K is the largest dimension of its simplices, where the dimension of a simplex X is one less than cardinality of X .*

For example, the collection

$$S = \{\{a, b, c\}, \{a, b\}, \{b, c\}, \{a, c\}, \{a\}, \{b\}, \{c\}, \{d\}\}$$

is a simplicial complex. Some of its simplices are {a, b, c}, {a, c}, {b} and {d}. The dimension of S is 2 as its element with the largest dimension is the simplex {a, b, c}, which has a cardinality of 3 and thus its dimension is 2.

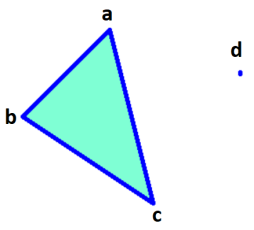

Figure 2: Visualization of the simplicial complex M

There are many different types of simplicial complexes that data analysts are interested in. When it comes to constructing simplicial complexes from finite metric spaces, an appropriate choice depends on several factors such as nature of the data, time and space complexity. While some simplicial complexes have desirable theoretical properties, they may, at the same time, suffer from being computationally inefficient, or vice versa. The Vietoris-Rips complex and Čech complex are two of the widely applied types of simplicial complexes in data analysis for their fast computational implementation and easy construction.[6]

Definition 4(Vietoris-Rips Complex)**.** *For a set of data points $\mathbb{X}$ in a metric space (M, $d_{\mathbb{X}}$) and a real number $\delta \geq 0$, the Vietoris-Rips complex $Rips_\delta(\mathbb{X})$ is the set of simplices $[x_0, ..., x_k]$ such that $d_{\mathbb{X}}(x_i, x_j) \leq \delta$ for all (i, j).*

If $\mathbb{X} \subseteq \mathbb{R}^d$, a 0-simplex can be visualized as an node, a 1-simplex can be visualized as an edge, a 2-simplex can be visualized as a triangle, and a 3-simplex can be visualized as a tetrahedron. Notice that one cannot identify a simplex with dimension greater than three. Also note that a simplicial complex can be of dimension higher than d.

Definition 5(Cech Complex)**.** *For a set of data points $\mathbb{X}$ in a metric space (M, $d_{\mathbb{X}}$) and a real number $\delta \geq 0$, the Cech complex $Cech_\delta(\mathbb{X})$ the set of simplices $[x_0, ..., x_k]$ such that the k + 1 closed balls $B(x_i, \alpha) = \{x \in M | d_{\mathbb{X}}(x_i, x) \leq \delta\}$ have a non-empty intersection.*

The figure below shows the Cech complex $Cech_\delta(\mathbb{X})$ and the and Vietoris-Rips $Rips_2\delta(\mathbb{X})$ of a finite point cloud in the plane $\mathbb{R}^2$. The bottom part of $Cech_\delta(\mathbb{X})$ is the union of two adjacent triangles, while the bottom part of $Rips_2\delta(\mathbb{X})$ is the tetrahedron spanned by the four vertices

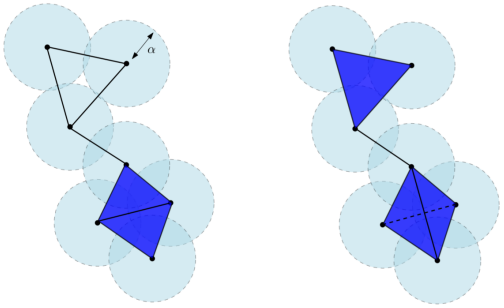

Figure 3: [3]$Cech_\delta(\mathbb{X})$(left) $Rips_2\delta(\mathbb{X})$(right)

and all its faces. The dimension of the Cech complex is 2, where as the dimension of the Vietoris-Rips complex is 3, which is thus not embedded in $\mathbb{R}^2$. Notice that in $\mathbb{R}^2$, the parameter δ in Cech complex defines the radius of the ball, whereas δ in the Vietoris-Rips complex defines the diameter. Additionally, it is not difficult to see that

$$Rips_\delta(\mathbb{X}) \subseteq Cech_\delta(\mathbb{X}) \subseteq Rips_2\delta(\mathbb{X})$$

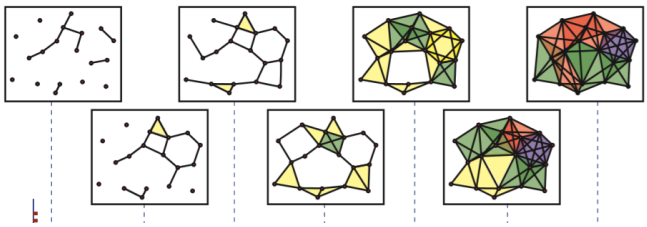

Figure 4: [4]Note that a simplicial complex with smaller parameter is the inclusion of a complex built on the same data set with larger one.

Having grasped the basics of simplicial complexes, one is able to construct different simplicial complexes from a finite data set and obtain some insight into its topological structure by varying the parameter δ and generating presumably different simplicial complexes.

2.3 Simplicial Homology

In this section, we will introduce the homology of simplicial complexes. *Simplicial homology* formalizes the idea of the number of holes of a given dimension(i.e. Betti number), which is a generalization of the number of connected components at 0 dimension.

Definition 1(Chain Complex)**.**[5] *Let K be a simplicial complex and K_n denote the collection of all simplices of K with dimension exactly n. For every positive integer n, we define*

$$C_n = \Big\{ \sum c_i \sigma_i | c_i \subseteq \mathbb{R}, \sigma_i \subseteq K_n \Big\}$$

By definition, we know that C_n is a free vector space.
The boundary map $\delta_n : C_n \to C_{n-1}$ is defined as follow:
, for $\sigma = [x_0, ..., x_n] \in K_n$ and $0 \leq i \leq n$, note that the element $[x_0, x_1, ..., x_n] \backslash x_i \in K_{n-1}$. Then we define

$$\delta_n(\sigma) = \sum_{i=0}^{n} (-1)^i [x_0, x_1, ..., x_n] \backslash x_i$$

Since C_n is a free vector space over K_n,it suffices to define the boundary maps on elements of K_n. The chain complex associated to K, denoted by $\mathcal{C}^*(K, \mathbb{R})$, is defined to be the sequence of vector spaces $\{C_n\}_{n \in \mathbb{Z}_+}$, along with the boundary maps $\delta_n : C_n \to C_{n-1}$. Precisely, we have

$$\mathcal{C}^*(K, \mathbb{R}) = \ ... \xrightarrow{\delta_{n+1}} C_n \xrightarrow{\delta_n} C_{n-1} \xrightarrow{\delta_{n-1}} ... \xrightarrow{\delta_2} C_1 \xrightarrow{\delta_1} C_0 \xrightarrow{\delta_0} 0$$

As an illustration, consider the simplicial complex $S = \{\{a,b,c\},\{a,b\},\{b,c\},\{a,c\},\{a\},\{b\},\{c\},\{d\}\}$. It can be visualized as a combination of simple topological objects in figure 2. We know that $K_2 = \{\{a,b,c\}\}, K_1 = \{\{a,b\},\{b,c\},\{a,c\}\}$ and $K_0 = \{\{a\},\{b\},\{c\},\{d\}\}$. To explore the definition of the chain complex, we proceed by recognizing that C_2 can be written in the form $\lambda\{a,b,c\}$, where λ is an arbitrary real number. We can take, for example, $\{a,b,c\}$ as an element of the free vector space C_2. By applying δ_2 to $\{a,b,c\}$, we have

$$\delta_2(\{a,b,c\}) = \{b,c\} - \{a,c\} + \{a,b\}$$

, which is an element of C_1. By applying δ_1 again to the resulting combination of 1-simplice, we now have

$$\delta_1(\{b,c\}-\{a,c\}+\{a,b\}) = \delta_1(\{b,c\})-\delta_1(\{a,c\})+\delta_1(\{a,b\}) = \{c\}-\{b\}-(\{c\}-\{a\})+\{b\}-\{a\} = 0$$

, which is a element of C_0 with the coefficients of the 0-simplice all equal to 0.

With further mathematical deductions, one can discover with ease that for all $n \in \mathbb{Z}_+$, $\delta_n \circ \delta_{n+1} = 0$. Based on this fact, we also know that *the image of δ_{n+1} is contained in the kernel(the set of elements that map to the zero vector) of δ_n.*

Definition 2(Simplicial Homology)**.** We denote the kernel of the map δ_n by $\mathcal{Z}_n(K,\mathbb{R})$ and denote the image of δ_{n+1} by $\mathcal{B}_n(K,\mathbb{R})$. As a result, we have $\mathcal{B}_n(K,\mathbb{R}) \subseteq \mathcal{Z}_n(K,\mathbb{R})$.

For $n \in \mathbb{Z}_+$, the n-th homology group of a simplicial complex K, denoted by $H_n(K,\mathbb{R})$, is defined as

$$H_n(K,\mathbb{R}) = \frac{\mathcal{Z}_n(K,\mathbb{R})}{\mathcal{B}_n(K,\mathbb{R})}$$

Definition 2(Betti Number)**.** *We define the n-th Betti number $\beta_n(K)$ of a simplicial complex K as the dimension of $H_n(K,\mathbb{R})$, where $dim(H_n(K,\mathbb{R})) = dim(\mathcal{Z}_n(K,\mathbb{R})) - dim(\mathcal{B}_n(K,\mathbb{R}))$.*

Based on the definition of Betti number, we can show that for a given simpicial complex, β_0 is the number of connected components, β_1 is the number of one-dimensional holes(i.e. tunnels) and β_2 is the number of two-dimensional holes(i.e. cavities). More about Betti numbers and other ways of exploring the topological and geometric features of data sets will be introduced and explained in detail in the next section.

3 Persistent Homology

By now, we are able to equip a given point cloud with either Vietoris-Rips complex or Cech complex based on the knowledge of simplicial complexes mentioned in the previous section. And this paves the way for studying desired topological and geometric properties through the lens of *persistent homology*[6]. There are essentially 3 steps in the persistent homology pipeline[3]:

1. We first build a filtration of simplicial complexes corresponding to an increasing sequence of the parameter δ on the metric space (X, d_x). The filtration here refers to a nested family of simplicial complexes, where the simplicial complex with larger parameter δ contains the complex with smaller one.

2. Second we compute relevant topological summaries such as the number of connected components, tunnels and cavities, i.e. n-th Betti number β_n.

3. Then we track the behavior of these summaries as δ increases. And we may represent this variation of Betti numbers in the form of *persistent barcode*.

Definition 1(Persistent Barcode)**.** Persistent barcode is a collection of stacked horizontal bars whose ends correspond to the appearance(or birth) and disappearance(or death) of the respective features as the threshold increases.

We can see from Figure 5 that Betti numbers are closely related to persistent barcode and can be easily extracted from it.

Summaries of topological and geometric features like persistent barcode could be easily visualized and provide deeper insight into the data set we interest in. Intuitively speaking, a feature(e.g. n-th Betti number) is more likely to be a true and common property if it persists over a significant range of parameter δ, while one that has a short life span could be a by-product of perturbations. The idea of taking a whole range of parameter values rather than finding an optimal one is the essence of persistent homology.

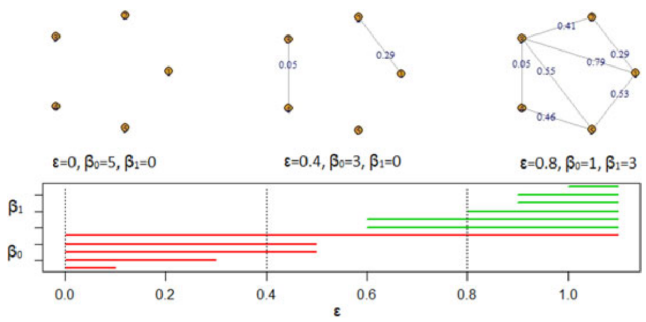

Figure 5: The figure illustrates the persistent barcode of one-dimensional Vietoris-Rips filtration built over 5 points. The ϵ appeared in the graph specifies the parameter δ mentioned before. The interpoint distances less than or equal to ϵ are depicted in blue. The ϵ values corresponding to the two ends of the horizontal bars mark the birth and death of topological features. To find the Betti numbers, we count the number of times the respective horizontal bars intersect the vertical line through ϵ. For example, for $\epsilon = 0.8$, $\beta_0 = 1$ and $\beta_1 = 3$.

4 Clustering Using Betti Numbers(CBN) Algorithm

In TDA, our method of data analysis is to primarily consider data sets as metric spaces. A point cloud $\mathbb{X}$ is a set of data points that belong to some metric space in $\mathbb{R}^n$. These data points are often defined by components representing spatial coordinates, quantitative features or repeated measurements. We define d to be the metric between two arbitrary points in the metric space. Our objective in data analysis is to partition $\mathbb{X}$ into non-overlapping clusters[6]. One example of algorithms that are capable of doing this is DBSCAN(density-based spatial clustering of applications with noise). These algorithms are dedicated to identify clusters such that any point in the cluster can be reached from any other one through the path consisting of points belonging to the cluster. Moreover, the consecutive points are close enough that their neighbors are similar in shape. Recall that we can build a filtration of simplicial complexes around each data point, and thus obtain the topological summaries(i.e. Betti numbers) concerning their neighbors. Two consecutive points are likely to be in a cluster if all of their n-th Betti numbers are similar. Alternatively, one could use persistent barcodes to compare the similarity between the neighbors of two points. The CBN applies both the metrics between points and local geometric information around points[6]. Below we outline the main steps used in CBN:

1. For each point i in $\mathbb{X}$, let $N(i)$ represent the collection of the neighborhood of point i consisting k of its nearest points with respect to the metric d. Define $D(i)$ to be the set of distances between each pair of points in $N(i)$.

2. Transform the distances in $D(i)$ with range $[0, 1]$ using the empirical cumulative distribution function of the distances.

3. Build a filtration of Vietoris-Rips complexes for the point i with a sequence of increasing parameter δ, i.e. $\delta_1 < \delta_2 < ... < \delta_n$. We have $VR_1(i) \subset VR_2(i) \subset ... \subset VR_n(i)$.

4. Compute the sequence of Betti-0 numbers and Betti-1 numbers on the corresponding filtration of simplicial complexes.

5. Compare the sequence of Betti numbers of each point in $N(i)$ with that of i. If the sequences show great similarity, that point will remain in $N(i)$. Otherwise it will be discarded. Consequently, we can see that Betti numbers help to refine $N(i)$, retaining only those nearest points whose local neighborhoods are relatively similar to that of i.

6. Form an adjacency matrix which keeps a record of the points in the updated $N(i)$. Compute the strongly connected components(i.e. clusters) of the directed graph defined by the matrix.

The algorithm is capable of clustering accurately without knowing the exact numbers of clusters, which is of great importance when it comes to data analysis. Note that if two point

clouds are similar in a topological/geometrical way, their corresponding persistent diagrams of Cech/Vietoris-Rips filtration are also close. Also note that as the parameter δ increases uniformly, the corresponding VR complexes contain more and more 1-simplices(or edges), resulting in a quick decrease in the number of connected components(i.e. δ_0) and a gradual increase of one-dimensional holes(i.e. δ_1).

Unlike other TDA-based clustering algorithms such as Mapper[7] and ToMATo(Topological Mode Analysis Tool)[8], which typically depend on a filter function or another clustering algorithm, CBN only rely on the construction of Vietoris-Rips simplicial complexes and computation of Betti numbers, thereby differing in its approach to clustering.

We will now test the CBN algorithm on a synthetic data set[9] consisting of 3,800 points in $\mathbb{R}$ and 13 clusters in distinct shapes in comparison with other different classes of clustering methods.

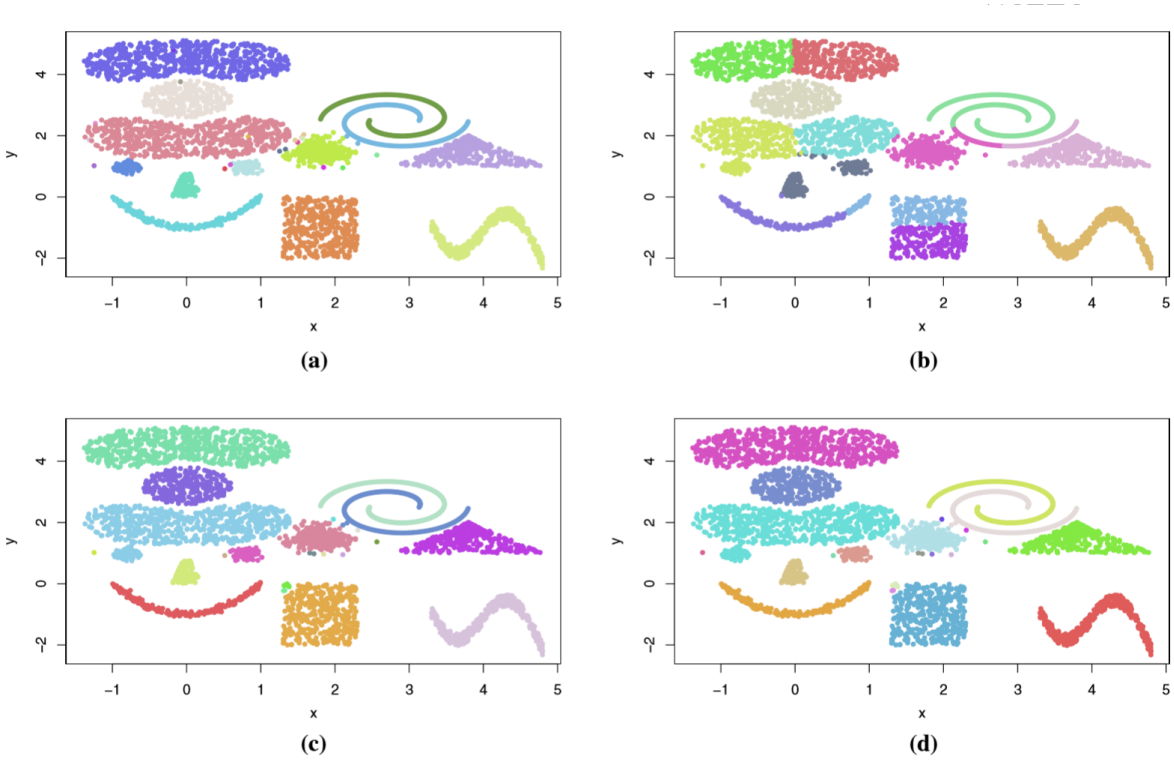

Figure 6: [6]Performance of (a)CBN (b)K-means[10] (c)Hierarchical Clustering Algorithm[11] (c)DBSCAN(density-based spatial clustering of applications with noise)[12]

It is straightforward to see from Figure 5 that CBN outperforms the rest – it successfully identify all the 13 clusters with few mismatches, whereas K-means indicate the poorest performance. The remaining algorithms fail to identify the "left eye" of the "smiley face" in the bottom left of the graph.

5 CONCLUSIONS

In this paper, we have introduced the topological concepts that the CBN algorithm based on and also the way CBN works. One of the significant properties that distinguish it from other clustering methods is that by integrating the conventional clustering algorithm with Betti numbers, it can systematically provide insights into data shapes and geometry without requiring the knowledge of the number of clusters. Nevertheless, there are still many problems (e.g. if there are too many "bridges" between clusters, it may be rather difficult for CBN to seperate them.) that need to be resolved in the future studies.

References

[1] Rodrigo Rivera-Castro, Polina Pilyugina, Alexander Pletnev, Ivan Maksimov, Wanyi Wyz, and Evgeny Burnaev. *Topological Data Analysis of Time Series Data for B2B Customer Relationship Management.* Industrial Marketing & Purchasing Group Conference (IMP19), 2019.

[2] M. Kramar, A. Goullet, L. Kondic, and K. Mischaikow. *Persistence of force networks in compressed granular media.* PHYSICAL REVIEW E 87, 042207 (2013).

[3] Chazal, Frédéric and Michel, Bertrand *An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists.*

[4] Robert Ghrist *Barcodes: The Persistent Topology of Data* Bulletin of the American Mathematical Society (2008)

[5] Facundo Mémoli and Kritika Singhal *A Primer on Persistent Homology of Finite Metric Spaces* Bulletin of Mathematical Biology, (2019), 1-43

[6] Islambekov, Umar and R. Gel, Yulia *Unsupervised space-time clustering using persistent homology: Clustering using Persistent Homology.* Environmetrics (2018)

[7] Singh, G., Mémoli, F. and Carlsson, G. *Topological methods for the analysis of high dimensional data sets and 3D object recognition* Eurographics Symposiumon Point-Based Graphics(SPBG), Prague, Czech Republic, 91–100 (2007)

[8] Chazal, F., Guibas, L., Oudot, S. and Skraba, P. *Persistence-based clustering in riemannian manifolds* Journal of the ACM, 60(6), 1–38 (2013)

[9] Jeong, M. H., Cai, Y., Sullivan, C. J. and Wang, S. *Data depth based clustering analysis* Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Burlingame, CA (2016)

[10] MacQueen, J. *Some methods for classification and analysis of multivariate observations* Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability(Vol.1, pp.281–297). Los Angeles, CA: University of California Press (1967)

[11] Johnson, S. *Hierarchical clustering schemes* Psychometrika, 32, 241–254 (1967)

[12] Ester, M., Kriegrel, H., Sander, J., and Xu, X. *A density-based algorithm for discovering clusters in large spatial databases with noise* Proceedings of the Second ACM SIGKDD International Conference on Knowledge Discovery and DataMining, Portland, Oregon, 226–231(1996)