DTM-Filtrations Paper Summary

Annu Indraganti

December 2024

1 Introduction

The main issue this paper aims to solve is the sensitivity of the Cech and Rips filtrations to outliers. The paper presents the concept of DTM-filtations as a potential solution to the problem at hand.

To start, we are trying to understand the homological features of a dataset (X) when we create point clouds around each data point. Persistence diagrams are constructed in order to understand these features across different radii. The Rips and Cech filtrations are great choices for developing the simplicial complex of a dataset due to their theoretical advantage, being that they produce stable persistence diagrams with respect to the perturbations (small insignificant topological changes) of X in accordance of the Hausdorff metric.

Since the Hausdorff metric is sensitive to outliers, it becomes an issue to create inferences on X when it is very noisy. Therefore, DTM was developed to subvert this problem. Since its pure calculation is expensive, a k-nearest neighbors-esque variant is used which is much easier to compute and is accurate up to a fixed constant. Another way to try and approximate the expensive DTM is a weighted version of the Rips complex which is reviewed later.

2 Filtrations and Interleaving Distance

We define T as a set of positive real numbers: $T = R^+$
We define E as a set of points that are in d dimensional space.
A filtration is defined as a subset of E where we use T to index.
Another rule for filtrations is that given we take two values from T that we will call s and t such that $s \leq t$, the filtration must be non-decreasing with respect to inclusion, or formally:

$$V_s \subseteq V_t$$

Two filtrations V and W are said to be ϵ-interleaved if for every $t \in T$:

$$V_t \subseteq W_{t+\epsilon} \text{ and } W_t \subseteq V_{t+\epsilon}$$

We are essentially saying that are relatively close to each other in terms of inclusion where ϵ represents that leeway we give to the definition

Let's say we have two filtrations S_t^1 and S_t^2 of a simplicial complex. The interleaving pseudo-distance (represented as $d_i(S_t^1, S_t^2)$) is the smallest ϵ such that the property:

$$S_t^1 \subseteq S_{t+\epsilon}^2 \text{ and } S_t^2 \subseteq S_{t+\epsilon}^1$$

holds. This isn't an actual distance per-say because two filtrations can have an interleaving distance that is 0 but both filtrations can be different. This can happen, if both filtrations are similar structurally.

A persistence module over T is a structure used that consists of two components: family of vector spaces and family of linear maps. The linear maps have the following properties:

$$\text{For every } t \in T,\ V_t \to V_t \text{ is the identity map}$$

$$\text{For } r \leq s \leq t, \text{ the composition } v_{t,s} \cdot v_{s,r} = v_{t,r}$$

When $\epsilon \geq 0$, an ϵ-morphism of two persistence modules V and W is a family of linear maps

$$\varphi_t : V^t \to W^{t+\epsilon}$$

such that the diagram in Figure 1 is true for all $s \leq t \in T$.

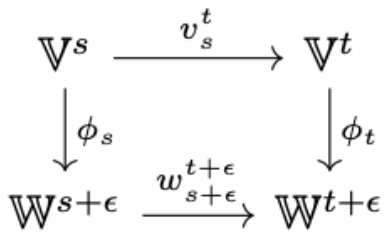

Figure 1: Persistence Diagram Mappings

If $\epsilon = 0$, then the family of φ_t is an isomorphism of the persistence modules. To expand the relationship, the ϵ-interleaving of V and W is a pair of ϵ-morphisms:

$$(\varphi_t : V_t \to W_{t+\epsilon})_{t \in T} \quad \text{and} \quad (\psi_t : W_t \to V_{t+\epsilon})_{t \in T}$$

The mapping that this provides is shown in the paper in the second section and is the second image in the section. Due to issues importing the image, it is not provided here.

The proximity between the persistence modules can be represented as a function as well. The mapping works the exact same as the image I mentioned above as long as the function holds that the function $\eta : T \to T$ is non-decreasing and

holds that $\eta(t) \geq t$.

Concepts of additive interleaving and multiplicative interleaving also exist are represented as such:

$$\eta : t \to t + c \text{ where } c \geq 0$$

$$\eta : t \to ct \text{ where } c \geq 1$$

For additive and multiplicative in that order.

The q-tameness of a persistence module also ensures that we can define a persistence diagram. This happens when for every $s, t \in T$ where st, the linear map $v_{t,s} : V_s \to V_t$ has a finite rank.

Given two q-tame persistence modules V and W along with their persistence diagrams, the isometry theorem states that the bottleneck distance between the two diagrams is equal to the interleaving pseudo-distance of the two persistence modules.

3 Weighted Filtrations

The section starts by generalizing the definition of a Cech complex. Here, we see that the radii of each point is transformed by some function f that is mapping the points from dataset X to a non-negative space R^+. Once we define that, the standard rules of a Cech complex are applied in constructing the filtration.

The paper gives an example of the weighted Cech filtration's capability to be resistant to noise under different scenarios. It then defines a proposition which states that if dataset $X \subseteq E$ and f is any function, then the persistence diagram using both of these is q-tame, which I defined above.

Another proposition the paper makes is that given another function that is similar in propoerties to f called g, if the max absolute distance between $f(x)$ and $g(x)$ is less than or equal to the defined leeway ϵ, then both persistence modules that were created based on f and g are considered ϵ-interleaved.

The last proposition that is made related to the weighted Cech complex is as follows:
Let's say we have a new dataset $Y \subseteq E$ and a function $f : X \cup Y \to R^+$ is c-Lipschitz where $c0$, we can say the two persistence diagrams created based on function f on both X and Y are k-interleaved where $k = \epsilon(1 + c^p)^{1/p}$. Though not mentioned above, p is a variable that is used to control the radius of each point cloud and is extremely important in doing so.

The paper states that under the presence of outliers, some of these propositions could fail, and that DTM, distance function that is robust to outliers, allows these propositions to hold much better under these circumstances.