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 $s \leq t \in T$.

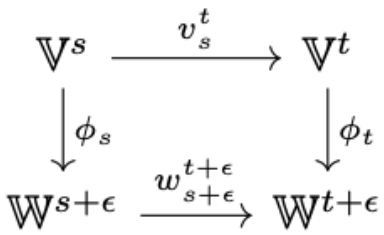

Figure 1: Persistence Diagram Mappings

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 $s \leq t$, 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 paper first begins by creating a general definition of the Cech filtration. We have Euclidean space $E = R^d$, index set $T = R^+$, and real number $p \geq 1$. Given X is some subset or equal to E and f maps X to R^+, for every $x \in X$ and $t \in T$, we define:

$$r_x(t) = \begin{cases} -\infty & \text{if } t < f(x) \\ (t^p - f(x)^p)^{1/p} & \text{otherwise} \end{cases}$$

which is a function that defines the radius of every point in X.

Given a persistence module $V^t[X, f]$, based on the Cech filtration, the family of filtrations where $t \geq 0$ is a weighted Cech filtration of parameters $X, f,$ and p.

We have cover $V^t[X, f]$ and $N(V^t[X, f])$ is the nerve of the cover, which is the simplicial ccomplex over vertex set X. The familiy of nerves where $t \geq 0$ is a filtered simplicial complex. Due to the nerve theorum, $V[X, f]$ and $V_N[X, f]$ are isomorphic persistence modules.

When $f = 0$, $N(V^t[X, f])$ is the usual Cech complex. When $p = 2$, the filtration value (called y) is the power distance of y associated with the weighted set (X, f). As such, $V[X, f]$ is the weighted Cech filtration.

Based on Proposition 2, if X is a bounded subset of E and f is any function, then $V[X, f]$ is q-tame.

The stability guarantees are as follows:

Prop 3 says given $g : X \to R^+$ which is a function such that $\sup_{x \in X} |f(x) - g(x)| \leq \epsilon$, then the modules $V[X, f]$ and $X[X, g]$ are -interleaved.

Prop 4 considers new subset of E called Y. Suppose $f : X \cup Y \to R^+$ (maps all unique points between X and Y to some positive dimensional space), is c-Lipschitz, where $c \geq 0$. We also assume that X and Y are compact and Hausdorff distance $d_H(X, Y) \leq \epsilon$. If so, $V[X, f]$ and $V[Y, f]$ are k-interleaved where $k = \epsilon(1 + c^p)^{1/p}$.

Proposition 3 and 4 become delicate in the wake of outliers. To mitigate this, DTM is introduced next, which is used as f.

Since Cech complexes are difficult to compute, most would wonder about the Rips complex and these concepts shifting over to them.

G is a graph of vertex set X. X's clique complex (subset of vertices and edges that form a complete graph) is the simplicial complex over X consisting of the cliques of G.

If S is a simplicial complex, its flag complex (generalization of a complete graph) is the clique complex of its 1-skeleton (bounded to 1-dimensional homological features).

We have $N(V^t[X, f])$ where its flag complex is $\text{Rips}(V^t[X, f])$. This Rips complex is called the weighted Rips complex with the same parameters as the weighted Cech complex. Both of these complexes are 2-interleaved, creating a generalization.

In Prop 6, we see that:

$$N(V^t[X, f]) \subseteq \text{Rips}(V^t[X, f]) \subseteq N(V^{2t}[X, f])$$

We can improve the second part of this inequality to:

$$\text{Rips}(V^t[X, f]) \subseteq N(V^{ct}[X, f]), \quad \text{where} \quad c = \sqrt{\frac{2d}{d+1}}$$

and d is the space $E = R^d$.
To compute the Rips complex, it enough to know the values of its 0 and 1-simplicies. Below describes the values:
Let $p < +\infty$, the filtration value of a vertex x be defined as:

$$t_X(x, y) = \begin{cases} \max(f(x), f(y)) & \text{if } \|x - y\| \leq |f(x)^p - f(y)^p|^{1/p}, \\ t & \text{otherwise} \end{cases}$$ where t is the only positive root of:

$$\|x - y\| = (t^p - f(x)^p)^{1/p} + (t^p - f(y)^p)^{1/p}$$

There are times this equation does not have a closed-form solution and the paper talks about those cases. For the sake of not delving too deep into theory, I will skip over this aspect.
The paper also says that non-trivial points on a persistence diagram are those that do not lie on the line $y = x$, which is something already reviewed through the introduction paper. We also have the following as being true:

$$D_0(N(V[X, f, p])) = D_0(Rips(V[X, f, p]))$$

since they both share the same 1-skeleton.

Lastly, it is stated that the number of non-trivial points in both filtrations is non-increasing under $p \in [1, +\infty)$.

4 DTM-Filtrations

The section begins again with what exactly DTM is. It goes over the true definition of DTM and how it is not practical to compute this practically. Therefore, we are introduced to the "nearest-neighbors" variant of the formula, which has already been reviewed in the introduction paper.

The weighted Cech filtration where it uses the DTM measure is defined as $W[X]$. The nerve of this along with its family of nerves where $t \geq 0$ are a filtered simplicial complex. The persistence modules of the weighted Cech filtration and its nerve are isomorphic.

Based on what we read about the weighted Rips complex above, we also see that the stated also makes sense:

$$\text{Rips}(W_t[X]) \text{ denotes the flag complex of } N(W_t[X]), \text{ and Rips}(W[X])$$

The paper makes a distance guarantee between two weighted Cech complexes using DTM as f. When two probability measure μ and v are on E with compact supports X and Y, we obtain:

$$d_i(W[X], W[Y]) \leq m - \frac{1}{2}W_2(X, Y) + \frac{2}{1}pd_H(X, Y)$$

This produces worse stability than the normal Cech filtration. The paper says that when the Hausdorff distance is small between X and Y, it is better to simply use standard Cech filtrations. But when it is larger, it may be better to use DTM-filtrations.

Stability is tested when $p = 1$. $m \in (0, 1)$ and μ is a probability measure on E with support $\text{supp}(\mu)$. It defines $c(\mu, m) = \sup_{\text{supp}(\mu)} (d_\mu, m)$.

Prop 14 states that given X such that $T \subseteq E$, both of their weighted Cech filtrations using DTM as f are $c(\mu, m)$-interleaved. If there is $Y \subseteq E$ is another set such that $T \subseteq Y$, then their DTM-filtrations are $c(\mu, m)$-interleaved. Also, if T is a finite subset and μ_T is its empirical measure, then T's DTM-filtration and X's DTM-filtration are $c(T, m)$-interleaved.