Efficient Initial Data Selection and Labeling for Multi-Class Classification Using Topological Analysis

Lies Hadjadj[a], Emilie Devijver[a,*], Rémi Molinier[b] and Massih-Reza Amini[a]

[a]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[b]Univ. Grenoble Alpes, CNRS, Department of Mathematics, IF, 38000 Grenoble, France

Abstract. Machine learning techniques often require large labeled training sets to attain optimal performance. However, acquiring labeled data can pose challenges in practical scenarios. Pool-based active learning methods aim to select the most relevant data points for training from a pool of unlabeled data. Nonetheless, these methods heavily rely on the initial labeled dataset, often chosen randomly. In our study, we introduce a novel approach specifically tailored for multi-class classification tasks, utilizing Proper Topological Regions (PTR) derived from topological data analysis (TDA) to efficiently identify the initial set of points for labeling. Through experiments on various benchmark datasets, we demonstrate the efficacy of our method and its competitive performance compared to traditional approaches, as measured by average balanced classification accuracy.

1 Introduction

Machine learning has seen widespread applications across various domains in recent years, yet its dependence on labeled data remains a significant challenge. Despite the availability of large amounts of unlabeled data due to advances in computing and storage, labeling is often labor-intensive and costly. Semi-supervised learning methods aim to address this challenge by leveraging both labeled and unlabeled data, with active learning emerging as a promising approach to efficiently select data points for labeling. The fundamental assumption behind active learning is that machine learning algorithms can achieve higher performance levels by using fewer training labels, provided they have the ability to strategically select the training dataset [44]. Constructing the first training set is of particular interest in many applications, such as recommendation systems.

Pool-based methods [33], play a crucial role in selecting the most informative data points for labeling from a pool of unlabeled observations. Usually, a classifier is learnt on a first labelled set, to find the next points to be labeled. However, these methods face limitations, especially in low-budget scenarios, where the acquisition of labeled data is constrained and where a sufficient budget is needed to learn a weak model [40]. Addressing these limitations requires innovative approaches that can integrate regularization techniques [26, 37] usually found in other sub-domains, such as semi-supervised learning or self-learning [14], and tackle the cold-start problem associated with selecting initial seed points.

In this paper, we propose a novel meta-approach for pool-based active learning, grounded in concepts from topological data analysis (TDA), to mitigate the cold-start problem and enhance performance in low-budget regimes. TDA offers insights into the underlying structure of data [12] by examining its topological properties [22], with persistent homology being a key technique for extracting topological features. It has already shown impressive results in machine learning [31], especially for clustering. Many recent papers have benefited from these topological insights to understand the structure of the data: Singh et al. [45] use persistence homology to extract molecular topological fingerprints (MTFs) based on the persistence of molecular topological invariants, Lum et al. [36] use topological persistence to efficiently encode fMRI datasets, Carlsson and Gabrielsson [13] use persistence homology to automatically extract interpretable features from meta-organic datasets in order to predict methane and carbon dioxide adsorption levels for different materials, among others, and Li et al. [34] also make use of topological persistence in order to actively estimate the homology of the Bayes decision boundary, the resulted module is then used to do model selection from several families of classifiers. However, the use of TDA in NLP is very recent [41]. Leveraging TDA, we extend `ToMATo` [17], a persistent-based clustering method, to identify proper topological regions suitable for propagating labeling. Specifically, our approach introduces proper topological regions using the σ-Rips graph based on an adaptive threshold function, while extending the theoretical guarantees of Chazal et al. [17] to the σ-Rips graph. We then explore the use of proper topological regions in a zero-shot[1] learning framework, further enhancing the effectiveness of pool-based active learning.

To validate our approach, we conduct empirical studies comparing it with classical methods across various datasets. Our results demonstrate the efficacy of our approach in improving zero-shot learning when there is a topological structure in the dataset.

The remainder of this paper is structured as follows: Section 2 provides an overview of related work, while Section 3 introduces the theoretical framework. In Section 4, we detail our proposed method, followed by experimental validation in Section 5. Finally, Section 6 concludes with a summary and avenues for future research.

2 Related literature

Various efforts have been made to reduce the annotation burden of machine learning algorithms. Significant progress has been achieved in semi-supervised learning [3, 9], which utilizes a small set of labeled data along with many unlabeled examples. These methods typically incorporate consistency regularization into the supervised

* Corresponding Author. Email: Emilie.Devijver@univ-grenoble-alpes.fr.

[1] The term *zero-shot* is used to emphasize that active learning is applied to an unlabeled set.

loss function through data augmentation using unlabeled observations [14].

Among the most well-known pool-based strategies are uncertainty sampling [33], margin sampling, and entropy sampling strategies [44]. Some approaches adopt the query-by-committee approach [24, 32], which entails learning an ensemble of models at each iteration. Practical implementations include query by bagging and query by boosting, which employ bagging and boosting to construct the committees [1]. Extensive research has focused on deriving efficient disagreement measures and query strategies from a committee, including vote entropy, consensus entropy, and maximum disagreement [44], while Ali et al. [2] introduces model selection for a committee. Additionally, some strategies solve an optimization problem to select the most relevant queries, as seen in Roy and McCallum [43], where Monte Carlo estimation is used to estimate the expected error reduction on test examples. In contrast, others utilize Bayesian optimization on acquisition functions such as the probability of improvement or the expected improvement [25], while Auer et al. [5] cast the problem of selecting the most relevant active learning criterion as an instance of the multi-armed bandit problem. A recent survey introduce a benchmark for many active learning methods [49].

Recent advancements in active learning suggest enhancing pool-based methods by leveraging knowledge from the distribution of unlabeled examples [10]. For instance, Perez et al. [39] propose using clustering of unlabeled examples to enhance the performance of pool-based active learners, where experts annotate entire clusters at each iteration rather than individual examples. Such a strategy effectively reduces annotation effort, assuming that the cost of cluster annotation is comparable to single example labeling, as demonstrated in Citovsky et al. [20] for large-scale data. Recently, numerous studies have explored the use of clustering/segmentation for active sample selection in real applications [4].

This paper addresses the cold-start problem in pool-based active learning, focusing on selecting the first point to be labeled based on covariate knowledge. To optimize the initial active learning (AL) training set, clustering techniques are used to select the most representative examples, typically found at cluster centers. Some AL studies use k-Means or k-Medoids clustering methods. In text classification, where datasets are high-dimensional, these methods can introduce randomness that deteriorates results, so deterministic clusterings like FFT, AHC, and APC have been proposed to stabilize outcomes. In medical imaging, [19] used contrastive learning and pseudo-labeling. [48] introduced the Nearest Neighbor Criterion (NNC), which sequentially queries the most representative instance from unlabeled data to minimize the overall distance between queried and unlabeled data and ensure each class is observed.

3 σ-Rips graph and its use in ToMATo

In this section, we introduce the theoretical framework and how we extend ToMATo guarantees in our setting, where ToMATo operates as a clustering method that employs the hill climbing algorithm on a Rips graph $R_\delta(\mathcal{S}_\mathbf{x})$ (see Definition 1 below) and incorporates a merging rule based on the persistence of the Rips graph.

3.1 Framework and notations

We consider a multi-class classification problem with input space $\mathcal{X} \subset \mathbb{R}^m$, and with output space $\mathcal{Y} = \{1, \dots, c\}$ a set of unknown classes of size $c \in \mathbb{N}, c \geq 2$. Let d be a fixed distance on $\mathbb{R}^m$. In pool-based active learning, we observe a sample set $\mathcal{S}_\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^n$ drawn from an unknown marginal distribution $\mathbb{P}$, and we have access to an oracle $\mathcal{O} : \mathcal{X} \to \mathcal{Y}$ that can provide the true label y_i for each observation $\mathbf{x}_i$, for $1 \leq i \leq n$ at some (expensive) cost. We denote $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ as the labeled data sample of size n, which we do not have access to, generated by some unknown joint distribution over $\mathcal{X} \times \mathcal{Y}$. The goal of zero-shot learning is to construct the set $\mathcal{I} \subset \{1, \dots, n\}$ of points to be labeled, to be considered as the first labeled set.

In our method, as is generally the case in classification algorithms, we assume that close points (with respect to d) are associated with similar labels, also known as the *smoothness assumption*. In this setting, neighborhood graphs on the unlabeled sample $\mathcal{S}_\mathbf{x}$ can be considered. A graph is denoted as a couple (V, E) with V representing the set of vertices, and E the set of edges. For our purpose, we use a neighborhood graph induced by the metric d on $\mathcal{X}$.

Definition 1 (Rips graph)**.** *Given a finite point cloud $\mathcal{S}_\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^n$ from a metric space $(\mathcal{X}, d)$ and $\delta \geq 0$, the* Rips graph $R_\delta(\mathcal{S}_\mathbf{x})$ *is the graph with set of vertices $\mathcal{S}_\mathbf{x}$ and whose edges correspond to the pairs of points $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}_\mathbf{x}^2$ such that $d(\mathbf{x}_i, \mathbf{x}_j) \leq \delta$.*

Rips graphs, and more generally Rips complexes [18], are fundamental in topology and are commonly used in TDA, particularly with persistent homology. However, class similarity may vary across the metric space. For instance, the lower the density, the weaker the chance to detect a structure within points. Consequently, it becomes necessary to extend the definition of the Rips graph to accommodate such scenarios. This leads to the introduction of the σ-Rips graph $R_{\sigma(\cdot)}(\mathcal{S}_\mathbf{x})$, which utilizes an adaptive threshold function σ.

Definition 2 (σ-Rips graph)**.** *Given a finite point cloud $\mathcal{S}_\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^n$ from a metric space $(\mathcal{X}, d)$ and a real-valued function $\sigma \colon \mathcal{X}^2 \to \mathbb{R}_+^*$, the* σ-Rips graph $R_{\sigma(\cdot)}(\mathcal{S}_\mathbf{x})$ *is the graph with set of vertices $\mathcal{S}_\mathbf{x}$ and whose edges correspond to the pairs of points $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}_\mathbf{x}^2$ such that $d(\mathbf{x}_i, \mathbf{x}_j) \leq \sigma(\mathbf{x}_i, \mathbf{x}_j)$.*

These two notions of neighborhood graphs are illustrated in Figure 1 to elucidate their disparities. In situations of lower density, where there are fewer points, the σ-Rips graph tends to be more connected. This enhanced connectivity serves to emphasize the underlying structure and aids in its detection.

The σ-Rips graph serves as a generalization of the Rips graph considering a constant threshold function or as a usual Rips graph with parameter δ on the non-metric space $(\mathcal{X}, \hat{d})$, with

$$\begin{aligned} \hat{d} \colon \mathcal{X} \times \mathcal{X} &\longrightarrow \mathbb{R}^+ \\ (\mathbf{x}, \mathbf{x}') &\longrightarrow \frac{\delta d(\mathbf{x}, \mathbf{x}')}{\sigma(\mathbf{x}, \mathbf{x}')}. \end{aligned} \tag{1}$$

Most topological properties of Rips graphs on metric spaces hold true for Rips graphs on non-metric spaces, as discussed in Chazal et al. [18, Section 4.2.5].

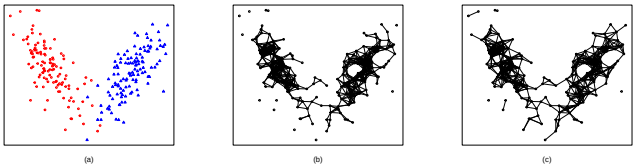


Figure 1: (a) A sample of 240 points from a mixture of two bivariate Gaussian distributions, with colors denoting true classes. (b) Associated Rips graph (Definition 1) with $\delta = 0.5$, using the Euclidean distance metric d. (c) Associated σ-Rips graph (Definition 2), using the parametric form from Equation (2), with the Euclidean distance metric d, $\delta = 0.5$, $r = 1.08$, and $t = 1/5$.

Here we adopt the following parametric threshold function:

$$\begin{aligned}\sigma(\cdot;\delta,r,t)\colon \mathcal{X}\times\mathcal{X} &\longrightarrow \mathbb{R}_+^* \\ (\mathbf{x},\mathbf{x}') &\longrightarrow \delta(r-\max\left(\mathbb{P}(\mathbf{x}),\mathbb{P}(\mathbf{x}')\right))^{\frac{1}{t}},\end{aligned} \tag{2}$$

with $t \in (0,1]$ and $(\delta, r) \in (\mathbb{R}_+^*)^2$, ensuring that $r > \max_{\mathbf{x}} \mathbb{P}(\mathbf{x})$. The temperature parameter t controls the curvature, while the max term ensures the symmetry of the function. δ and r represent the dilatation and translation parameters, respectively. This parametric form is illustrated in Figure 2. We demonstrate in Section 5.1 that the resulting curve from the optimal parameters of our function aligns with our intuition regarding class similarity as a density-aware measure.

ToMATo can be easily adapted to operate with a σ-Rips graph $R_{\sigma(\cdot)}(\mathcal{S}_{\mathbf{x}})$ by considering the non-metric space $(\mathcal{S}_{\mathbf{x}}, \hat{d})$ described above. In the next two sections, we explain that switching from Rips graph to σ-Rips graph in ToMATo yields about the same guarantees.

3.2 Persistence and upper-star filtrations

ToMATo relies on the notion of persistence, particularly persistent homology, which is a classical tool in Topological Data Analysis (TDA). We refrain from formally introducing these concepts and refer interested readers to [22, 12] for a more general and comprehensive treatment. Here we will only deal with nested families of graphs and their associated topological persistence diagrams.

Let G be a finite graph and $\mathcal{G} = (G_\alpha)_{\alpha\in\overline{\mathbb{R}}}$ be a nested family of subgraphs of G (here, if $\alpha_1 \leq \alpha_2$, then G_{α_2} is a subgraph of G_{α_1}). The idea of persistent homology is to keep track of the changes in the topology of G_α as α decreases (here we can think of α as the time but going backward). We are just interested in the connectedness of our graphs and we set, for G a graph, $H_0(G)$ to be the vector space generated by the connected components of G. Then, an inclusion $G_i \supseteq G_j$ of two graphs induces a linear map $H_0(G_j) \to H_0(G_i)$ which send a connected component $\mathcal{C}$ of G_j to the connected component of G_i which contains $\mathcal{C}$. The collection $\mathbf{G} = (H_0(G_\alpha))_{\alpha\in\overline{\mathbb{R}}}$ together with all these linear maps $H_0(G_{\alpha_1}) \to H_0(G_{\alpha_2})$ for all $\alpha_1 \geq \alpha_2$ is the *persistent module* of the nested family of graphs $(G_\alpha)_{t\in\overline{\mathbb{R}}}$. Then, one can keep track of the *persistence* of a connected component along the nested family $\mathcal{G}$ through this persistence module by looking at what time t it arises, its *birth* time, and at what time it dies (because glued to another one), its *death* time[2]. The *persistence diagram* $D\mathbf{G}$ of the nested family $\mathbf{G}$ is then the multi-subset[3] of $\overline{\mathbb{R}}^2$ of all given pair $(\text{death}, \text{birth})$ for each connected components (if two or more components have the same death and birth we get a point with multiplicity 2 or more) together with the diagonal $\Delta = \{(x,x) \mid x \in \overline{\mathbb{R}}\}$. When interpreting a persistence diagram, the distance of points from the diagonal, called their *prominence*, should be considered. Points with low prominence are considered as topological noise (they do not persist long), while points with high prominence signify relevant topological information.

We introduce now the families of graphs that we will consider: the upper-star filtration of $\mathbb{P}\colon \mathcal{X} \to \mathbb{R}$ restricted to a Rips graph.

Definition 3 (Upper-star Rips filtration)**.** *Given a finite point cloud $\mathcal{S}_{\mathbf{x}}$ from a metric space $(\mathcal{X}, d)$ with a probability function $\mathbb{P}$ and a real value $\delta \in \mathbb{R}^+$, the* upper-star Rips filtration *of $\mathbb{P}$, denoted $\mathcal{R}_\delta(\mathcal{S}_{\mathbf{x}}, \mathbb{P})$, is the nested family of subgraphs of the Rips graph $R_\delta(\mathcal{S}_{\mathbf{x}})$ defined as $\mathcal{R}_\delta(\mathcal{S}_{\mathbf{x}}, \mathbb{P}) = \left(R_\delta(\mathcal{S}_{\mathbf{x}} \cap \mathbb{P}^{-1}([\alpha, +\infty])\right)_{\alpha\in\overline{\mathbb{R}}}$. Such a nested family of graphs gives rise to a persistence module*

$$\mathbf{R}_\delta(\mathcal{S}_{\mathbf{x}}, \mathbb{P}) = \left(H_0\left(R_\delta(\mathcal{S}_{\mathbf{x}} \cap \mathbb{P}^{-1}([\alpha, +\infty])\right)\right)\right)_{\alpha\in\overline{\mathbb{R}}},$$

and to its associated persistence diagram $D\mathbf{R}_\delta(\mathcal{S}_{\mathbf{x}}, \mathbb{P})$.

This notion is illustrated in Figure 3.

Similarly, we define the *upper-star σ-Rips filtration* of $\mathbb{P}$, denoted $\mathcal{R}_{\sigma(\cdot)}(\mathcal{S}_{\mathbf{x}}, \mathbb{P})$, and the associated persistence module $\mathbf{R}_{\sigma(\cdot)}(\mathcal{S}_{\mathbf{x}}, \mathbb{P})$ and persistence diagram $D\mathbf{R}_{\sigma(\cdot)}(\mathcal{S}_{\mathbf{x}}, \mathbb{P})$.

3.3 Comparison of persistence diagrams for Rips graph and σ-Rips graph

In this section, we provide tools to control the difference between $D\mathbf{R}_\delta(\mathcal{S}_{\mathbf{x}}, \mathbb{P})$ and $D\mathbf{R}_{\sigma(\cdot)}(\mathcal{S}_{\mathbf{x}}, \mathbb{P})$.

The bottleneck distance serves as an effective and natural proximity measure to compare two persistence diagrams.

Definition 4 (Bottleneck distance)**.** *Given two multi-subsets A_1, A_2 of $\overline{\mathbb{R}}^2$, the* bottleneck distance $d_B^\infty(A_1, A_2)$ *between A_1 and A_2 is*

$$d_B^\infty(A_1, A_2) = \min_{\zeta : A_1 \to A_2} \max_{p\in A_1} \|p - \zeta(p)\|_\infty$$

where ζ runs over all possible multi-bijections between A_1 and A_2.

The bottleneck distance between two persistence diagrams induced by upper-star Rips (or σ-Rips) filtrations can be controlled by comparing the evolution of connected components along the filtration, which can be tracked with the appearance level.

[2] When several components gets attached at a given time, we can see that as the "death" of all of them except one. The one who survives is the "oldest" one (with an arbitrary choice if there are more than one). It is a bit more complicated than that in general but this suffices if for example the edges are added one at a time.

[3] Several connected components can have the same birth and death times which leads to points with multiplicity in the persistence diagram.

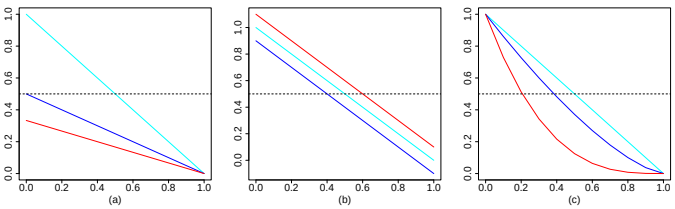


Figure 2: Representation of $s : u \mapsto \delta(r-u)^{1/t}$ as a proxy of the parametric form of σ given in (2), with varying parameters. By default, all parameters are fixed to 1. We vary in (a) $\delta \in \{1, 0.5, 1/3\}$, in (b) $r \in \{0.9, 1, 1.1\}$, in (c) $t \in \{1, 0.7, 1/3\}$. The dashed line represents a constant threshold function $\sigma = 0.5$.

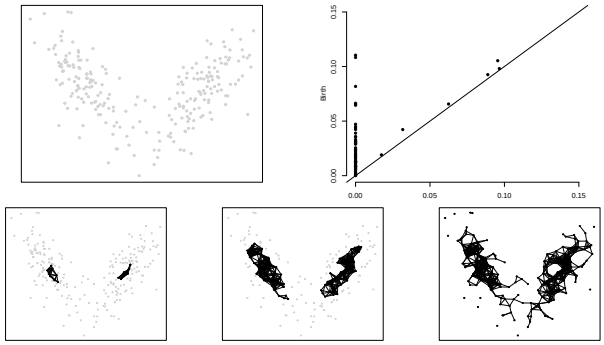


Figure 3: A point cloud generated from a mixture of two Gaussians (top left), some realizations of its upper-star Rips filtration for $\delta = 0.5$ and $\alpha \in \{0.1, 0.05, 0\}$ (bottom, from left to right) and the associated persistence diagram (top right).

Definition 5 (Appearance level)**.** *Given a finite point cloud $\mathcal{S}_{\mathbf{x}} = \{\mathbf{x}_i\}_{i=1}^n$ from a metric space $(\mathcal{X}, d)$ with a probability function $\mathbb{P}$ and δ such that $R_\delta(\mathcal{S}_{\mathbf{x}})$ is connected. For two distinct points $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}_{\mathbf{x}}^2$, we define the* appearance level $\alpha_\delta(\mathbf{x}_i, \mathbf{x}_j)$ *as the highest level of the upper-star Rips filtration $\mathcal{R}_\delta(\mathcal{S}_{\mathbf{x}}, \mathbb{P})$ at which $\mathbf{x}_i$ and $\mathbf{x}_j$ are in the same connected component:*

$$\alpha_\delta(\mathbf{x}_i, \mathbf{x}_j) = \max_{\gamma \in \mathcal{P}(\mathbf{x}_i, \mathbf{x}_j)} \min_{\mathbf{x} \in \gamma} \mathbb{P}(\mathbf{x})$$

where $\mathcal{P}(\mathbf{x}_i, \mathbf{x}_j)$ is the set of all paths in $R_\delta(\mathcal{S}_{\mathbf{x}})$ from the vertex $\mathbf{x}_i$ to the vertex $\mathbf{x}_j$, where a path γ in a graph R is a sequence of vertices of R where two consecutive vertices are adjacent in R.

For example, in Figure 3, between $\alpha = 0.05$ and $\alpha = 0$, we observe two connected components that are eventually connected, on the cluster on the left. The appearance level of the corresponding points lies between 0.05 and 0.

Similarly, we define $\alpha_{\sigma(\cdot)}$ the appearance level for an upper-star σ-Rips filtration $\mathcal{R}_{\sigma(\cdot)}(\mathcal{S}_{\mathbf{x}}, \mathbb{P})$.

This enables us to bound the bottleneck distance between the Rips graph and the σ-Rips graph.

Theorem 1. *Given a finite point cloud $\mathcal{S}_{\mathbf{x}} = \{\mathbf{x}_i\}_{i=1}^n$ from a metric space $(\mathcal{X}, d)$ with probability function $\mathbb{P}$. Let $R_\delta(\mathcal{S}_{\mathbf{x}})$ be the Rips graph with parameter δ, $R_{\sigma(\cdot)}(\mathcal{S}_{\mathbf{x}})$ the σ-Rips graph with threshold function σ and assume that they share the same connected components. Then,*

$$\begin{aligned} &d_B^\infty \left(D\mathbf{R}_\delta(\mathcal{S}_{\mathbf{x}}, \mathbb{P}), D\mathbf{R}_{\sigma(\cdot)}(\mathcal{S}_{\mathbf{x}}, \mathbb{P}) \right) \\ &\leq \max_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}_{\mathbf{x}}^2} |\alpha_\delta(\mathbf{x}_i, \mathbf{x}_j) - \alpha_{\sigma(\cdot)}(\mathbf{x}_i, \mathbf{x}_j)|, \end{aligned}$$

setting $\alpha_\delta(\mathbf{x}_i, \mathbf{x}_j) = \alpha_{\sigma(\cdot)}(\mathbf{x}_i, \mathbf{x}_j) = 0$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ are not in the same connected component.

The proof is provided in the extended version (Appendix A)[4], leveraging the notion of ε-interleaving from [15]. The core concept is to regulate the birth and death of connected components along the filtration by controlling the changes in appearance levels.

This theorem implies that when shifting from the metric distance d to the neighboring (possibly) non-metric distance $\hat{d}$ defined in (1) (and subsequently from the Rips graph to the σ-Rips graph) results in a dendrogram induced by the upper-star Rips graph that remains largely consistent throughout the persistence process.

Since all theoretical guarantees of ToMATo [17] are derived from the persistence diagram of the Rips graph, Theorem 1 also tells us that, when shifting to σ-Rips graphs in ToMATo, we get about the same guaranties: with reasonable assumption on the point cloud we can recover the numbers of clusters induced by the density $\mathbb{P}$ as well as the basins of attractions of the prominent peaks of $\mathbb{P}$ (see [17] for more details).

4 Proper topological regions and zero-shot learning

In this section, we first introduce the proper topological regions and then explain how to use them in zero-shot learning.

4.1 Proper topological regions

In our method, we leverage the concept of topological regions, which is based on the algorithm ToMATo where the granularity of the clustering is controlled by a merging hyperparameter $\tau \geq 0$, which retains only clusters with a prominence higher than τ. These regions are defined as the clusters obtained by ToMATo with a σ-Rips graph.

[4] https://aptikal.imag.fr/~amini/Publis/TDA_ECAI24.pdf

Definition 6 (Topological regions)**.** *The* topological regions *of a sample set $\mathcal{S}_{\mathbf{x}}$ coming from an unknown marginal distribution $\mathbb{P}$ and with parameters (δ, r, t, τ) are the clusters given by the clustering algorithm ToMATo on the σ-Rips graph $R_{\sigma(\cdot;\delta,r,t)}(\mathcal{S}_{\mathbf{x}})$:*

$$\mathrm{TR}_{\delta,r,t,\tau}^{\mathbb{P}}(\mathcal{S}_{\mathbf{x}}) = \texttt{ToMATo}_\tau \left(R_{\sigma(\cdot;\delta,r,t)}(\mathcal{S}_{\mathbf{x}}), \mathbb{P} \right).$$

When the underlying density $\mathbb{P}$, and the parameters are understood, we will simply denote $\mathrm{TR}(\mathcal{S}_{\mathbf{x}})$ the clustering, seeing the mapping function as $\mathrm{TR}\colon \mathcal{S}_{\mathbf{x}} \to \{1, \ldots, k\}$ for a clustering into k topological regions.

For a clustering $\mathrm{TR}(\mathcal{S}_{\mathbf{x}})$ of the sample $\mathcal{S}_{\mathbf{x}}$, we define $\mathcal{L}_{\mathrm{TR}}^{\mathbb{P}}$ as the labeling function that assigns, within a given topological region, the label of the sample with the highest density according to $\mathbb{P}$:

$$\begin{aligned} &\mathcal{L}_{\mathrm{TR}}^{\mathbb{P}}\colon \mathcal{S}_{\mathbf{x}} \to \mathcal{Y} \\ &\qquad \mathbf{x}_i \mapsto \mathcal{O}\left(\underset{\mathbf{x}_j : \mathrm{TR}(\mathbf{x}_j) = \mathrm{TR}(\mathbf{x}_i)}{\arg\max} \mathbb{P}(\mathbf{x}_j) \right). \end{aligned} \tag{3}$$

If access to the labeled data $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is available, we define the *Purity Size function* PS as the objective function that considers the labeling error when applying $\mathcal{L}_{\mathrm{TR}}^{\mathbb{P}}$ within the topological regions, penalized by the number of topological regions k in TR:

$$\mathrm{PS}\left(\mathcal{S}, \mathbb{P}, \mathrm{TR}\right) = \left[\frac{k}{n} + \frac{1}{n} \sum_{i=1}^n 1_{\mathcal{L}_{\mathrm{TR}}^{\mathbb{P}}(\mathbf{x}_i) \neq y_i} \right] \in [0, 1].$$

Then, we introduce the concept of proper topological regions, which will be a fundamental element in our method.

Definition 7. *The* proper topological regions *of a sample set $\mathcal{S}_{\mathbf{x}}$ coming from an unknown marginal distribution $\mathbb{P}$ are the topological regions of $\mathrm{TR}_{\delta^*,r^*,t^*,\tau^*}^{\mathcal{S}_{\mathbf{x}},\mathbb{P}}$ where*

$$(\delta^*, r^*, t^*, \tau^*) = \underset{(\delta,r,t,\tau)}{\arg\min} \left\{ \mathrm{PS}\left(\mathcal{S}, \mathbb{P}, \mathrm{TR}_{\delta,r,t,\tau}^{\mathcal{S}_{\mathbf{x}},\mathbb{P}} \right) \right\}. \tag{4}$$

The main problem in this notion is the use of labels in the purity size function, which are not available in our active learning context, where labeled data is scarce. We then need to use an unsupervised objective function. We consider a trade-off between the Silhouette score (other potential unsupervised criteria typically used to assess the clustering quality are discussed in Appendix B[4]) and the coverage compactness of a clustering TR of $\mathcal{S}_{\mathbf{x}}$ into k topological regions $\{R_1, \ldots R_k\}$.

Definition 8. *For $1 \leq q \leq k$, let π_q be the size of the topological region $R_q = \{\mathbf{x} \in \mathcal{S}_{\mathbf{x}} : \mathrm{TR}(\mathbf{x}) = q\}$. For $\lambda \in \mathbb{R}^+$, we define the* penalized silhouette score *by:*

$$\begin{aligned} \mathit{SilSize}_\lambda(\mathcal{S}_{\mathbf{x}}, \mathrm{TR}) &= \left[\frac{1}{k} \sum_{q=1}^k \frac{1}{\pi_q} \sum_{\mathbf{x} \in R_q} s_{il}(\mathbf{x}) \right] - \lambda \frac{k}{n} \\ &\in \left[-1 - \lambda, 1 - \frac{\lambda}{n} \right], \\ \text{with} \qquad s_{il}(\mathbf{x}) &= \frac{\nu^c(\mathbf{x}) - \nu(\mathbf{x})}{\max(\nu(\mathbf{x}), \nu^c(\mathbf{x}))} \end{aligned} \tag{5}$$

where, for all q and all $\mathbf{x} \in R_q$, $\nu(\mathbf{x})$ represents the average distance of sample $\mathbf{x}$ within its cluster R_q and $\nu^c(\mathbf{x})$ stands for the average distance of sample $\mathbf{x}$ to his nearest neighbor cluster:

$$\nu(\mathbf{x}) = \frac{1}{\pi_q - 1} \sum_{\mathbf{x}' \in R_q} d(\mathbf{x}, \mathbf{x}'), \quad \nu^c(\mathbf{x}) = \min_{q' \neq q} \frac{1}{|C_{q'}|} \sum_{\mathbf{x}' \in C_{q'}} d(\mathbf{x}, \mathbf{x}').$$

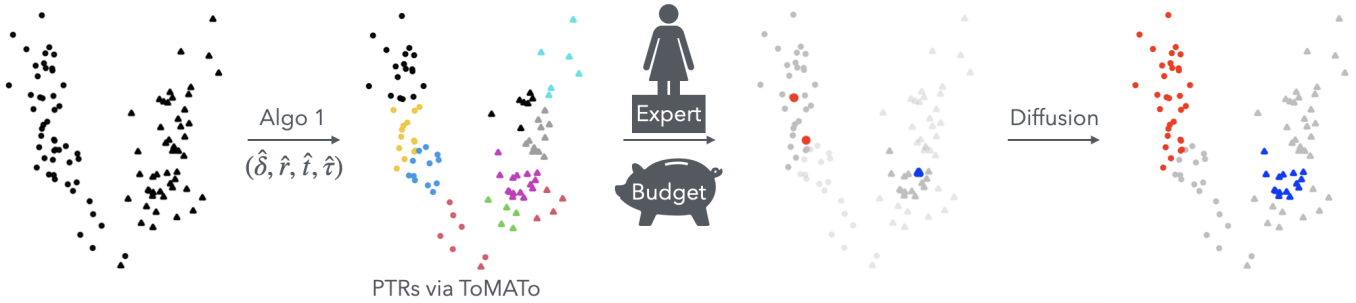


Figure 4: Flowchart of the method. We observe the data, and we run Algorithm 1 to construct the proper topological regions, seen as clusters through colors. Then, an expert labels one point in each of the $\mathcal{B}$ largest clusters, for budget $\mathcal{B}$ (here, $\mathcal{B} = 3$), and we consider the pseudo-labels by diffusing the labeling to the proper topological regions.

The trade-off parameter λ in (5) plays a crucial role in identifying the proper topological regions within the sample set $\mathcal{S}_{\mathbf{x}}$. Higher values of λ penalize coverage compactness, leading to partitions with a higher degree of agglomeration, meaning fewer topological regions with larger sizes. However, an additional way to control the labeling propagation error term of the Purity Size objective in an unsupervised setting is to control the size distribution of groups in the resulting partition. Conversely, lower λ values tend to produce highly fragmented partitions with numerous small-sized groups. In such scenarios, the Silhouette score tends to converge to graphs with a single non-singleton connected component and many singletons.

Thus, we approximate the optimization problem (4) by

$$\underset{(\delta,r,t,\tau)}{\operatorname{argmax}} \left\{ \mathit{SilSize}_\lambda \left(\mathcal{S}_{\mathbf{x}}, \mathrm{TR}^{\mathcal{S}_{\mathbf{x}},\mathbb{P}}_{\delta,r,t,\tau} \right) \right\}. \tag{6}$$

Unfortunately, optimizing this objective function directly is computationally expensive because it requires running `ToMATo` multiple times for different parameter settings. Instead, we propose a proxy optimization problem that involves running `ToMATo` only once:

$$(\delta^\sharp, r^\sharp, t^\sharp) = \underset{(\delta,r,t)}{\operatorname{argmax}} \left\{ \mathit{SilSize}_\lambda(\mathcal{S}_{\mathbf{x}}, R_{\sigma(\cdot;\delta,r,t)}(\mathcal{S}_{\mathbf{x}})) \right\} \tag{7}$$

$$\tau^\sharp = \underset{\tau}{\operatorname{argmax}} \left\{ \mathit{SilSize}_\lambda \left(\mathcal{S}_{\mathbf{x}}, \mathrm{TR}^{\mathcal{S}_{\mathbf{x}},\mathbb{P}}_{\delta^\sharp,r^\sharp,t^\sharp,\tau} \right) \right\} \tag{8}$$

with a slight abuse of notations in (7) between the Rips graph $R_{\sigma(\cdot;\delta,r,t)}(\mathcal{S}_{\mathbf{x}})$ and its connected components seen as a clustering. The best hyperparameters $a^\sharp, r^\sharp, t^\sharp$ for the silhouette of the σ-Rips graph are then used to find the best hyperparameter $\tau^\sharp$ for `ToMATo`.

Since the underlying density is usually unknown, we need to estimate it from the data. For that purpose, we use the distance to a measure [17], which computes the root-mean-squared distance to the ℓ nearest neighbors of the considered query point: for all $i \in \{1, \ldots, n\}$,

$$\hat{\mathbb{P}}(\mathbf{x}_i) = \left(\frac{1}{\ell} \sum_{j=1}^{n} d(\mathbf{x}_i, \mathbf{x}_j)^2 1_{\mathbf{x}_j \text{ is a } \ell\text{-nearest neighbors of } \mathbf{x}_i} \right)^{-1/2} \tag{9}$$

The entire procedure used to approximate the proper topological regions is data-driven using the unlabeled set $\mathcal{S}_{\mathbf{x}}$. We denote the corresponding estimated proper topological regions with parameters $(\hat{\delta}, \hat{r}, \hat{t}, \hat{\tau})$ as $\widehat{\mathrm{PTR}}$. Algorithm 1 describes a two-stage black-box optimization scheme to estimate the σ-Rips graph parameters (δ^*, r^*, t^*) by $(\hat{\delta}, \hat{r}, \hat{t})$, and the merging parameter τ^* by $\hat{\tau}$, which solves our optimization problem given in (4) for the proper topological regions of $\mathcal{S}$.

[5] A graph is degenerate if the sizes of the connected components are imbalanced (we do not allow very small connected components).

4.2 *Cold-start learning*

In this section, we present the application of proper topological regions to the zero-shot learning problem.

Given the unlabeled set $\mathcal{S}_{\mathbf{x}}$, along with the estimated proper topological regions $\widehat{\mathrm{PTR}}$ obtained from Algorithm 1, and access to an oracle for providing a limited number of labels (denoted as a budget $\mathcal{B}$), as well as the density estimation $\hat{\mathbb{P}}$, our strategy is as follows:

1. Selection of Largest Regions: We focus on the $\mathcal{B}$ largest proper topological regions $R_1, R_2, \ldots, R_{\mathcal{B}}$.
2. Querying the Oracle: For each R_q, where $q \in \{1, \ldots, \mathcal{B}\}$, we query the oracle for the $\mathcal{B}$ points with highest density within R_q.
3. Label Propagation: We propagate the labels using the labelling function $\mathcal{L}^{\hat{\mathbb{P}}}_{\widehat{\mathrm{PTR}}}$ introduced in (3), resulting in labeling $\sum_{q=1}^{\mathcal{B}} |R_q|$ points by propagating the true labels in each topological region.
4. Initial Labeled Set: We denote by $\hat{\mathcal{S}}^0$ this first set of labeled and pseudo-labeled points, which includes true labels obtained directly from the oracle, and estimated labels while diffusing the true labels to the topological regions.

The label propagation scheme is employed on proper topological regions to augment the sample size for training in a small budget scenario with a fixed number of calls to the oracle, as proposed in [20]. This procedure is summarized in Figure 4.

The use of proper topological regions confers distinct advantages over conventional clustering methodologies, manifesting in several key facets. Unlike generic clustering methods such as k-means, which impose assumptions of specific structural configurations such

Algorithm 1 Optimization procedure for PTR

Require: $\mathcal{S}_{\mathbf{x}} := \{\mathbf{x}_i\}_{i=1}^n$, $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$, s the step size for the linear search, and l the number of trials for the optimization strategy.

Initialize $\lambda = s$.
Compute the density estimator $\hat{\mathbb{P}}$ with (9) based on d and $\mathcal{S}_{\mathbf{x}}$.
Optimize the problem (7) for l trials, and return $(\hat{\delta}, \hat{r}, \hat{t})$.
Build the σ-Rips graph $R_{\sigma(\cdot;\hat{\delta},\hat{r},\hat{t})}(\mathcal{S}_{\mathbf{x}})$.
while $R_{\sigma(\cdot;\hat{\delta},\hat{r},\hat{t})}(\mathcal{S}_{\mathbf{x}})$ is not a *degenerate graph*[5] **do**
    Update $\lambda \longleftarrow \lambda + s$.
    Optimize the problem (7) for l trials, updating $\hat{\delta}, \hat{r}, \hat{t}$.
    Build the σ-Rips graph $R_{\sigma(\cdot;\hat{\delta},\hat{r},\hat{t})}(\mathcal{S}_{\mathbf{x}})$.
end while
Update $\lambda \longleftarrow \lambda - s$.
Optimize problem (8) for l trials
Output: parameters $\hat{\delta}, \hat{r}, \hat{t}, \hat{\tau}$ and the corresponding $\widehat{\mathrm{PTR}}$.

Table 1: Dataset statistics: n_{train} is the size of the training set, n_{test} is the size of the test set, m is the number of features, c is the number of classes, and imbalance corresponds to the class imbalance ratio.

Dataset	n_{train}	n_{test}	m	c	imbalance
protein [27]	756	324	77	8	0.70
banknote [42]	943	405	4	2	0.83
coil-20 [47]	1008	432	1024	20	1.00
isolet [23]	4366	1872	617	26	0.99
pendigits [42]	7694	3298	16	10	0.92
nursery [42]	9070	3888	8	4	0.09

as spherical clusters, our approach prioritizes the inherent topology of the data. This inherent flexibility enables our algorithm to discern and delineate connected components even amidst ambiguity in shape, thus circumventing the limitations imposed by rigid structural assumptions. Furthermore, the efficacy of our pseudo-labeling strategy hinges upon the purity of clusters, a criterion facilitated by the judicious selection of clustering methods and a sufficiently rich assortment of clusters. When these conditions are met, pseudo-labeling emerges as a potent tool, yielding commendable results that underscore its utility as a pivotal component within our methodology. This nuanced interplay between proper topological regions and pseudo-labeling not only enhances the discriminative power of our approach but also imbues it with resilience against the intricacies and uncertainties inherent in real-world datasets. In essence, our methodological framework represents a paradigm shift in the realm of multi-class classification, offering a versatile and robust alternative to traditional clustering methodologies by prioritizing topological integrity and harnessing the power of pseudo-labeling.

5 Empirical results

We conduct experiments to evaluate how the proposed approach identifies valuable examples for the initial training set. We use six datasets commonly employed in active learning, known for their topological structure, with statistics presented in Table 1.

We use Euclidean distance d, but note that the parametrization in (2) allows for more general geometries. For optimizing Algorithm 1, we use the Tree-structured Parzen Estimator (TPE) [8] with $l = 500$ trials and a step size s of 0.01 for the line search procedure. To estimate $\mathbb{P}$, we use (9) with the distance to measure based on the ℓ nearest neighbors with ℓ being the sample size, if smaller than 2000, and 2000 otherwise. In all experiments, we use the random forest classifier [28] as the base estimator with default parameters. We consider several budgets $\mathcal{B} \in \{3, 10, 20\}$, and conduct 20 stratified random splits, allocating 70% of the data to training and 30% to testing. We evaluate performance using balanced classification accuracy [11]. For data preprocessing, we remove duplicate samples and those with null values, then apply standard min-max normalization.

5.1 *Rips graph vs σ-Rips graph*

To validate our hypothesis regarding the density-aware threshold defined by Equation (2) for class similarity and to justify our extension of the Rips graph to accommodate this concept, we conduct a comparative study shown in Figure 5. This comparison is carried out between the Rips graph and the σ-Rips graph using the protein dataset. The results for the remaining datasets are provided in Figure 6 in C[4]. The plot illustrates the threshold optimization process for both the Rips graph and the σ-Rips graph, with the aim of minimizing the Purity Size cost function. In the plot, the threshold for the Rips graph (depicted in blue) remains constant and is represented as a horizontal line, whereas the threshold for the σ-Rips graph (depicted in orange) varies as a non-constant function. The respective parameters are optimized to minimize the purity size function, resulting in a noticeable reduction in the objective function from 0.2111 in the constant case to 0.1722 in the non-constant case.

In Figure 5, we also include two additional side plots: the distribution of the dataset's density estimation $\hat{\mathbb{P}}$ along the x-axis and the distribution of Euclidean distances in the distance matrix D along the y-axis. It is important to note that according to the definitions of the Rips graph (Def. 1) and the σ-Rips graph (Def. 2), threshold values larger than the maximum distance result in a fully connected graph. From this figure, it is evident that the values of the optimal threshold rule found in the hypothesis class of the σ-Rips graph, using our proposed threshold function $\sigma(\cdot; (\delta, r, t, \tau))$ given in (2), are negatively correlated with the density estimation $\hat{\mathbb{P}}$. These observations hold for other datasets reported in the appendix[4], except for coil-20 and nursery collections, where both achieve the same performance. These findings provide empirical evidence supporting our hypothesis that class similarity is a density-aware measure. They also validate our choice of $\sigma(\cdot; (\delta, r, t, \tau))$ given in (2) as an appropriate threshold function to generalize the Rips graph.

5.2 *Cold-start results*

For the cold-start experiments, we compare our approach with the following unsupervised methods:

- **K-Means clustering** (KM) has been used for active learning in Zhu et al. [50], to generate the initial training set by labeling the closest sample to each centroid.
- **K-Means clustering with model examples (KM+ME).** A variant of KM proposed in Kang et al. [30], Hu et al. [29] adds artificial samples from the centroids, named *model examples*, to the initial training set. This approach leads to an initial training set twice as large as the one created using K-Means.
- **K-Medoids clustering** (Km) is very similar to K-Means except that it uses the actual samples for centers, namely the medoids, as the center of each cluster. These medoids are then used to form the initial training set in active learning [38].
- **Agglomerative Hierarchical Clustering** (AHC) is a bottom-up clustering approach that builds a hierarchy of clusters. Initially, each sample is a singleton cluster. Then, the algorithm recursively

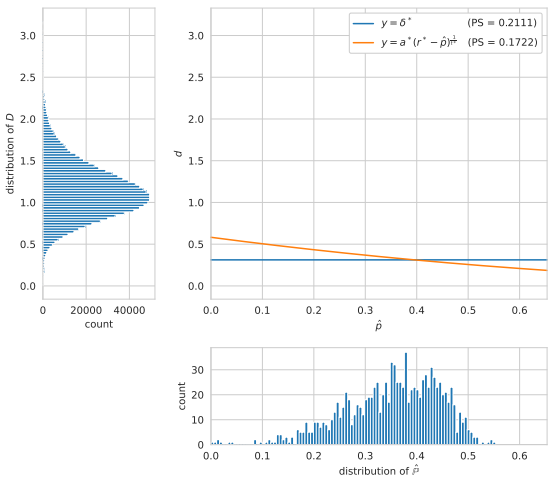


Figure 5: Comparison study between the Rips graph and the σ-Rips graph on the protein dataset: the Purity Size score is reported for each minimizer on top right.

Table 2: Average balanced classification accuracy (in %) and standard deviation of random forest classifier with the initial training set obtained from different methods over 20 stratified random splits for different budgets $\mathcal{B}$. $^{\uparrow}/^{\downarrow}$ indicate statistically significantly better/worse performance than Random Selection RS, according to a Wilcoxon rank sum test ($p < 0.05$) [46].

Dataset	$\mathcal{B}$	RS	KM	KM+ME	Km	AHC	FFT	APC	PTR
protein [27]	3	16.9(4.0)	21.2$^{\uparrow}$(1.8)	**23.9**$^{\uparrow}$(2.2)	21.2$^{\uparrow}$(4.4)	22.7$^{\uparrow}$(2.5)	17.4(3.3)	16.7(3.3)	22.1$^{\uparrow}$(6.2)
	10	28.2(3.2)	30.6$^{\uparrow}$(4.6)	31.4$^{\uparrow}$(4.5)	29.3$^{\uparrow}$(4.4)	31.6$^{\uparrow}$(3.7)	21.8(3.8)	28.8(3.4)	**40.5**$^{\uparrow}$(3.9)
	20	36.4(3.8)	42.1$^{\uparrow}$(3.9)	45.5$^{\uparrow}$(2.5)	39.2(4.4)	43.4$^{\uparrow}$(3.4)	26.1$^{\downarrow}$(3.4)	39.2(3.7)	**54.0**$^{\uparrow}$(3.4)
banknote [42]	3	55.5(7.2)	74.0$^{\uparrow}$(4.6)	**84.3**$^{\uparrow}$(5.6)	62.5$^{\uparrow}$(3.3)	63.7$^{\uparrow}$(4.5)	58.2$^{\uparrow}$(7.3)	58.7(8.0)	70.2$^{\uparrow}$(14.7)
	10	79.9(9.9)	85.2$^{\uparrow}$(5.7)	86.8$^{\uparrow}$(4.8)	87.6$^{\uparrow}$(3.3)	85.6$^{\uparrow}$(5.0)	70.6$^{\downarrow}$(5.3)	82.4(6.9)	**88.7**$^{\uparrow}$(4.4)
	20	87.6(2.9)	90.7$^{\uparrow}$(2.4)	92.4$^{\uparrow}$(2.0)	92.3$^{\uparrow}$(2.4)	92.6$^{\uparrow}$(2.9)	71.9$^{\downarrow}$(7.2)	90.9$^{\uparrow}$(3.2)	**93.9**$^{\uparrow}$(3.4)
coil-20 [47]	3	12.6(2.6)	**15.0**$^{\uparrow}$(0.0)	**15.0**$^{\uparrow}$(0.0)	**15.0**$^{\uparrow}$(0.0)	**15.0**$^{\uparrow}$(0.0)	10.8$^{\downarrow}$(2.0)	11.7(2.3)	13.6(1.7)
	10	29.0(5.7)	36.7$^{\uparrow}$(4.2)	38.2$^{\uparrow}$(2.7)	32.9$^{\uparrow}$(5.1)	36.0$^{\uparrow}$(3.7)	18.6$^{\downarrow}$(3.4)	27.2(4.8)	**44.2**$^{\uparrow}$(2.4)
	20	42.0(5.8)	56.7$^{\uparrow}$(3.7)	63.0$^{\uparrow}$(2.8)	42.3(3.5)	58.1$^{\uparrow}$(4.1)	25.6$^{\downarrow}$(2.5)	41.4(4.7)	**71.1**$^{\uparrow}$(3.8)
isolet [23]	3	07.6(1.5)	08.7$^{\uparrow}$(0.9)	09.7$^{\uparrow}$(0.6)	07.8(1.6)	09.1$^{\uparrow}$(1.9)	09.2$^{\uparrow}$(1.0)	07.5(1.8)	**10.8**$^{\uparrow}$(1.1)
	10	13.8(2.3)	22.3$^{\uparrow}$(1.6)	**27.6**$^{\uparrow}$(1.6)	07.1$^{\downarrow}$(1.9)	23.3$^{\uparrow}$(1.8)	16.5$^{\uparrow}$(1.7)	15.4(3.2)	27.5$^{\uparrow}$(2.8)
	20	19.2(2.7)	27.9$^{\uparrow}$(2.5)	**40.4**$^{\uparrow}$(3.2)	10.7$^{\downarrow}$(2.0)	28.2$^{\uparrow}$(2.1)	18.8(2.4)	21.1$^{\uparrow}$(3.1)	38.6$^{\uparrow}$(3.2)
pendigits [42]	3	21.5(3.5)	21.3(1.9)	22.5(2.1)	26.6$^{\uparrow}$(2.6)	19.4$^{\downarrow}$(1.8)	17.3$^{\downarrow}$(3.7)	17.8$^{\downarrow}$(4.9)	**29.9**$^{\uparrow}$(0.0)
	10	37.4(7.2)	62.5$^{\uparrow}$(3.5)	65.6$^{\uparrow}$(2.3)	53.9$^{\uparrow}$(5.2)	61.4$^{\uparrow}$(1.9)	27.2$^{\downarrow}$(4.9)	38.3(8.2)	**80.1**$^{\uparrow}$(2.6)
	20	54.3(5.9)	72.3$^{\uparrow}$(2.7)	75.8$^{\uparrow}$(2.3)	64.0$^{\uparrow}$(3.6)	72.3$^{\uparrow}$(2.5)	34.8$^{\downarrow}$(4.5)	52.2(5.9)	**87.7**$^{\uparrow}$(4.1)
nursery [42]	3	30.7(4.0)	29.2(5.2)	30.2(6.5)	25.0$^{\downarrow}$(0.2)	28.3$^{\downarrow}$(3.9)	30.0(3.2)	30.0(3.7)	**35.1**$^{\uparrow}$(5.6)
	10	42.7(7.2)	44.5(5.7)	**49.3**$^{\uparrow}$(4.0)	28.4$^{\downarrow}$(1.3)	44.9(7.2)	39.1(3.5)	45.1(6.7)	46.5(6.0)
	20	**55.3**(2.8)	52.8$^{\downarrow}$(3.3)	54.4(3.0)	32.9$^{\downarrow}$(1.1)	53.8(2.7)	39.8$^{\downarrow}$(1.1)	52.5$^{\downarrow}$(4.9)	54.1(4.5)

merges the closest clusters using a *linkage function* (Ward linkage here) until one cluster is left. This process is depicted in a dendogram, with each level representing a merge. AHC has been used for active learning by pruning the dendogram at a certain level to form clusters and then selecting the samples closest to the cluster centroids for the initial training set [21].

- **Furthest-First-Traversal** (FFT) selects a sequence of examples where the first example is chosen arbitrarily, and each subsequent example in the chain is placed as far away from the set of previously chosen examples as possible. The resulting sequence is then used as the initial training set for active learning [6].
- **Affinity Propagation Clustering** (APC) find *exemplars* of the sample set which are representative of clusters. It simultaneously considers all the sample set as possible *exemplars* and uses the message-passing procedure to converge to a relevant set of *exemplars* that are used as an initial training set for active learning [29].

Our meta-approach for zero-shot learning, PTR, uses the σ-Rips graph from Algorithm 1. Table 2 presents the average balanced classification accuracy for the random forest classifier on Random Selection (RS), competitors, and PTR across all datasets. Propagation within clusters detected by `ToMATo` is applied only to PTR, while competitors use $\mathcal{B}$ clusters. Our findings show PTR's efficacy and superiority over random selection and other methods. PTR consistently outperforms random selection, highlighting its effectiveness for initiating pool-based active learning. The success of PTR underscores the importance of Algorithm 1 in identifying the largest proper topological regions for the initial training set, providing accurate pseudo-labels that enhance model performance. For instance, in the protein dataset, labeling 3 points allowed for 94 pseudo labels with an accuracy mean (and variance) of 0.9 (0.1) over 20 splits.

PTR competes strongly against baseline strategies for the cold-start problem in active learning, outperforming others on 4 out of 6 datasets. While methods like APC align with RS, and FFT and Km sometimes lag behind RS, KM, KM+ME, and AHL show improvements, but PTR remains superior. The Nursery dataset's class imbalance and our label propagation step degrade training data quality. The estimator training lacks mechanisms to handle this imbalance, leading to suboptimal performance. On the COIL-20 dataset, PTR excels with budgets of 10 and 20 but performs unusually at a budget of 3, where all top approaches show zero variance, indicating potential issues with the balanced accuracy metric in low-class scenarios.

KM+ME improves classification accuracy by creating artificial examples from K-means centroids, mitigating class imbalance, reducing overfitting, and ensuring better data distribution coverage. This enhances classifier robustness, making KM+ME a strong benchmark despite differing assumptions from our routine. Overall, PTR demonstrates robustness and versatility in active learning, especially for topologically structured data, offering enhanced model performance and accelerated learning in multi-class classification.

6 Conclusion

In this study, we introduce a data-driven meta-approach for pool-based active learning strategies, specifically addressing multi-class classification problems. Central to our methodology is the concept of proper topological regions within a sample set. We explain the theoretical basis of this concept and formulate a black-box optimization problem to identify these regions. We also demonstrate the application of proper topological regions in zero-shot learning, showing their effectiveness in guiding the selection of initial data points for labeling. Our extensive empirical evaluation across diverse benchmark datasets under resource-constrained conditions highlights the robustness and efficacy of our approach. However, further research is needed. We aim to conduct a rigorous theoretical analysis, including the derivation of generalization bounds to ensure optimal performance. We also advocate exploring semi-supervised methodologies that integrate a regularization term from the proper topological regions. Additionally, the issue of class imbalance, crucial in real-world scenarios, requires careful investigation. Inspired by works like [7], we plan to explore imbalanced learning settings and develop tailored solutions for asymmetrical class distributions. Finally, while this method focuses on H_0 features, similar to `ToMATo`, it would be interesting to include all TDA features (Betti numbers from zero to infinity) and select the relevant ones. This is beyond the scope of this study but represents an intriguing area for future research.

References

[1] N. Abe and H. Mamitsuka. Query learning strategies using boosting and bagging. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, page 1–9, 1998.

[2] A. Ali, R. Caruana, and A. Kapoor. Active learning with model selection. In *AAAI*, 2014.

[3] M.-R. Amini and N. Usunier. *Learning with Partially Labeled and Interdependent Data*. Springer, New York, USA, 2015.

[4] G. Andresini, A. Appice, D. Ienco, and D. Malerba. Seneca: Change detection in optical imagery using siamese networks with active-transfer learning. *Expert Systems with Applications*, 214:119123, 2023.

[5] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

[6] Y. Baram, R. El-Yaniv, and K. Luz. Online choice of active learning algorithms. *J. Mach. Learn. Res.*, 5:255–291, 2004.

[7] R. Barata, M. Leite, R. Pacheco, M. O. P. Sampaio, J. a. T. Ascensão, and P. Bizarro. Active learning for imbalanced data under cold start. In *Proceedings of the Second ACM International Conference on AI in Finance*, ICAIF '21, 2022.

[8] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

[9] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2019.

[10] A. Bonnin, R. Borràs, and J. Vitrià. A cluster-based strategy for active learning of rgb-d object detectors. In *ICCV Workshops*, pages 1215–1220, New York, USA, 2011. IEEE.

[11] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pages 3121–3124, 2010.

[12] G. Carlsson. *Foundations of Computational Mathematics, Budapest 2011*, chapter The Shape of Data, page 16–44. London Mathematical Society Lecture Note Series. Cambridge University Press, 2012.

[13] G. Carlsson and R. B. Gabrielsson. Topological approaches to deep learning. In N. A. Baas, G. E. Carlsson, G. Quick, M. Szymik, and M. Thaule, editors, *Topological Data Analysis*, pages 119–146, 2020.

[14] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, Massachusetts, USA, 2006.

[15] F. Chazal, D. Cohen-Steiner, M. Glisse, L. J. Guibas, and S. Y. Oudot. Proximity of persistence modules and their diagrams. In *Proceedings of the 25^{th} Annual Symposium on Computational Geometry*, page 237–246, 2009.

[16] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba. Scalar field analysis over point cloud data. *Discrete & Computational Geometry*, 46(4):743, 2011.

[17] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba. Persistence-based clustering in riemannian manifolds. *J. ACM*, 60(6), 2013.

[18] F. Chazal, de Vin Silva, and S. Oudot. Persistence stability for geometric complexes. *Geometriae Dedicata*, 173(1):193–214, 2014.

[19] L. Chen, Y. Bai, S. Huang, Y. Lu, B. Wen, A. Yuille, and Z. Zhou. Making your first choice: To address cold start problem in medical active learning. In *Medical Imaging with Deep Learning*, 2023.

[20] G. Citovsky, G. DeSalvo, C. Gentile, L. Karydas, A. Rajagopalan, A. Rostamizadeh, and S. Kumar. Batch active learning at scale. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[21] S. Dasgupta and D. Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 208–215, 2008.

[22] H. Edelsbrunner and J. Harer. *Computational Topology - an Introduction*. American Mathematical Society, Boston, USA, 2010.

[23] M. Fanty and R. Cole. Spoken letter recognition. In R. P. Lippmann, J. E. Moody, and D. S. Touretzky, editors, *Advances in Neural Information Processing Systems 3*, pages 220–226. Morgan-Kaufmann, Massachusetts, USA, 1991.

[24] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Mach. Learn.*, 28(2-3):133–168, 1997.

[25] R. Garnett. *Bayesian Optimization*. Cambridge University Press, Cambridge, United Kingdoms, 2022.

[26] I. Guyon, G. C. Cawley, G. Dror, and V. Lemaire. Results of the active learning challenge. In I. Guyon, G. Cawley, G. Dror, V. Lemaire, and A. Statnikov, editors, *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, volume 16 of *Proceedings of Machine Learning Research*, pages 19–45, Sardinia, Italy, 2011. PMLR.

[27] C. Higuera, K. J. Gardiner, and K. J. Cios. Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome. *PLOS ONE*, 10(6):1–28, 2015.

[28] T. K. Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[29] R. Hu, B. M. Namee, and S. J. Delany. Off to a good start: Using clustering to select the initial training set in active learning. In *FLAIRS*, 2010.

[30] J. Kang, K. R. Ryu, and H.-C. Kwon. Using cluster-based sampling to select initial training set for active learning in text classification. In H. Dai, R. Srikant, and C. Zhang, editors, *Advances in Knowledge Discovery and Data Mining*, pages 384–388, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[31] A. S. Krishnapriyan, J. Montoya, M. Haranczyk, J. Hummelshøj, and D. Morozov. Machine learning with persistent homology and chemical word embeddings improves prediction accuracy and interpretability in metal-organic frameworks. *Scientific Reports*, 11(1):8888, 2021.

[32] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30*, pages 6402–6413, 2017.

[33] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine Learning Proceedings 1994*, pages 148–156. 1994.

[34] W. Li, G. Dasarathy, K. Natesan Ramamurthy, and V. Berisha. Finding the homology of decision boundaries with active learning. In *Advances in Neural Information Processing Systems*, pages 8355–8365, 2020.

[35] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[36] P. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, and G. Carlsson. Extracting insights from the shape of complex data using topology. *Scientific reports*, 3:1236, 2013.

[37] Y. Maximov, M.-R. Amini, and Z. Harchaoui. Rademacher complexity bounds for a penalized multi-class semi-supervised algorithm. *Journal of Artificial Intelligence Research*, 61:761–786, 2018.

[38] H. T. Nguyen and A. Smeulders. Active learning using pre-clustering. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 79, New York, NY, USA, 2004. Association for Computing Machinery.

[39] F. Perez, R. Lebret, and K. Aberer. Cluster-based active learning. *CoRR*, page abs/1812.11780, 2018.

[40] K. Pourahmadi, P. Nooralinejad, and H. Pirsiavash. A simple baseline for low-budget active learning, 2021.

[41] A. Rathore, Y. Zhou, V. Srikumar, and B. Wang. Topobert: Exploring the topology of fine-tuned word representations. *Information Visualization*, 22(3):186–208, 2023.

[42] J. D. Romano, T. T. Le, W. La Cava, J. T. Gregg, D. J. Goldberg, P. Chakraborty, N. L. Ray, D. Himmelstein, W. Fu, and J. H. Moore. PMLB v1.0: an open-source dataset collection for benchmarking machine learning methods. *Bioinformatics*, 38(3):878–880, 2021.

[43] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In C. E. Brodley and A. P. Danyluk, editors, *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 441–448, 2001.

[44] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[45] G. Singh, F. Mémoli, and G. Carlsson. Topological methods for the analysis of high dimensional data sets and 3d object recognition. pages 91–100, 2007.

[46] D. A. Wolfe. *Nonparametrics: Statistical Methods Based on Ranks and Its Impact on the Field of Nonparametric Statistics*, pages 1101–1110. Springer US, Boston, MA, 2012.

[47] J. Yang, Z. Chen, W.-S. Chen, and Y. Chen. Robust affine invariant descriptors. *Mathematical Problems in Engineering*, 2011.

[48] Y. Yang and M. Loog. To actively initialize active learning. *Pattern Recognition*, 131:108836, 2022.

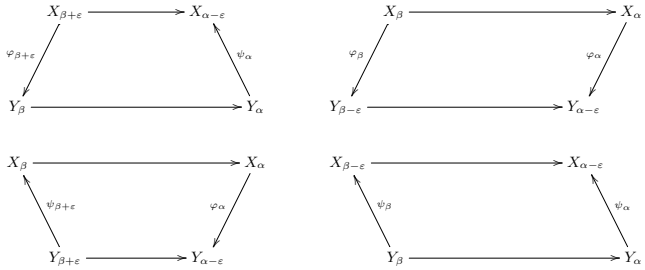
[49] X. Zhan, H. Liu, Q. Li, and A. B. Chan. A comparative survey: Benchmarking for pool-based active learning. In Z.-H. Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4679–4686. International Joint Conferences on Artificial Intelligence Organization, 2021. Survey Track.

[50] J. Zhu, H. Wang, T. Yao, and B. K. Tsou. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1137–1144, Manchester, UK, 2008.

A Proof of Theorem 1

To show that two persistence diagrams are close to one another with respect to the bottleneck distance, one can use the following notion of proximity at the level of persistence modules introduced in [15].

Definition 9 (ε-interleaved)**.** *Let* $\mathbf{X} = (X_\alpha)_{\alpha\in\overline{\mathbb{R}}}$ *and* $\mathbf{Y} = (Y_\alpha)_{\alpha\in\overline{\mathbb{R}}}$ *be two persistence modules. We say that* $\mathbf{X}$ *and* $\mathbf{Y}$ *are* strongly ε-interleaved *if there exist two families of linear application* $\{\varphi_\alpha \colon X_\alpha \to Y_{\alpha-\varepsilon}\}_{\alpha\in\mathbb{R}}$ *and* $\{\psi_\alpha \colon Y_\alpha \to X_{\alpha-\varepsilon}\}_{\alpha\in\mathbb{R}}$*, such that for all* $\alpha, \beta \in \mathbb{R}$*, if* $\alpha \leq \beta$*, then the following diagrams, whenever they make sense, are commutative:*



The idea is that every component born (resp. died) in $\mathbf{X}$ at some time α must appear (resp. die) in $\mathbf{Y}$ within $[\alpha-\varepsilon, \alpha+\varepsilon]$, and vice-versa. The following lemma highlights how important this notion is.

Lemma 2. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be two persistence modules such that* $D\mathbf{X}$ *and* $D\mathbf{Y}$ *have only finitely many points away from the diagonal, and let* $\varepsilon > 0$*. If* $\mathbf{X}$ *and* $\mathbf{Y}$ *are strongly* ε*-interleaved, then* $D\mathbf{X}$ *and* $D\mathbf{Y}$ *are at a distance at most* ε *with respect to the bottleneck distance.*

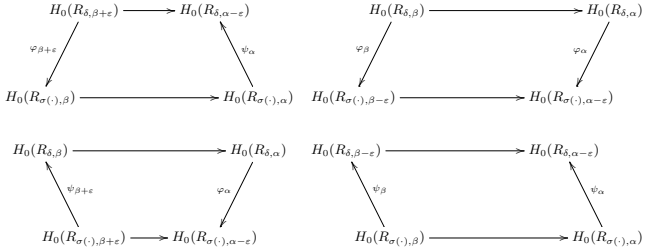
This lemma is a direct consequence of Chazal et al. [15, Theorem 4.4] where the result is proven for every homological dimension.

For example, in Chazal et al. [16, Theorem 5], it is proven that given the density function $\mathbb{P}$ on a point cloud $\mathcal{S}_\mathbf{x}$ with sufficient sampling density, the persistence diagram $D\mathbb{R}_\delta(\mathcal{S}_\mathbf{x}, \mathbb{P})$ built upon the Rips graph $R_\delta(\mathcal{S}_\mathbf{x})$ with an appropriate δ is a good approximation of the persistence diagram of $\mathbb{P}$. Consequently, $D\mathbb{R}_\delta(\mathcal{S}_\mathbf{x}, \mathbb{P})$ highlights the *0th homology groups* of the underlying space of $\mathcal{S}_\mathbf{x}$, this is a crucial ingredient in the proof of the theoretical guarantees of ToMATo.

Proof of Theorem 1. To simplify the notations let $\varepsilon = \max_{(\mathbf{x}_i,\mathbf{x}_j)\in\mathcal{S}_\mathbf{x}^2} |\alpha_\delta(\mathbf{x}_i,\mathbf{x}_j) - \alpha_{\sigma(\cdot)}(\mathbf{x}_i,\mathbf{x}_j)|$, $\mathbf{R}_\delta = \mathbf{R}_\delta(\mathcal{S}_\mathbf{x}, \mathbb{P})$ and $\mathbf{R}_{\sigma(\cdot)} = \mathbf{R}_{\sigma(\cdot)}(\mathcal{S}_\mathbf{x}, \mathbb{P})$, and, for $\alpha \in \mathbb{R}$,

$$R_{\delta,\alpha} = R_\delta\left(\mathcal{S}_\mathbf{x} \cap \mathbb{P}^{-1}([\alpha,+\infty])\right), \text{ and}$$
$$R_{\sigma(\cdot),\alpha} = R_{\sigma(\cdot)}\left(\mathcal{S}_\mathbf{x} \cap \mathbb{P}^{-1}([\alpha,+\infty])\right).$$

For $\alpha \in \mathbb{R}$, let $\mathcal{C}_1, \ldots, \mathcal{C}_k$ be the connected components of $R_{\delta,\alpha}$. For every $q \in \{1,\ldots,k\}$, and every $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}_q$, we have that $\alpha_\delta(\mathbf{x}_i,\mathbf{x}_j) \geq \alpha$ and thus, by definition of ε, $\alpha_{\sigma(\cdot)}(\mathbf{x}_i,\mathbf{x}_j) \geq \alpha - \varepsilon$. Hence $\mathcal{C}_q$ is contained in a connected component of $R_{\sigma(\cdot),\alpha-\varepsilon}$. This gives a linear map $\varphi_\alpha \colon H_0(R_{\delta,\alpha}) \to H_0(R_{\sigma(\cdot),\alpha-\varepsilon})$. By a similar argument, we get a linear map $\psi_\alpha \colon H_0(R_{\sigma(\cdot),\alpha}) \to H_0(R_{\delta,\alpha-\varepsilon})$ and, by construction, the following diagrams are commutative.



Consequently, $\mathbf{R}_\delta$ and $\mathbf{R}^\sigma$ are strongly ε-interleaved, then the bottleneck distance is bounded. □

B How to approximate the Purity Size objective function

For a given graph $R(\mathcal{S}_\mathbf{x})$, let $\mathcal{C}_1, \ldots, \mathcal{C}_k$ be the connected components of this graph, we define the mean-sample per connected component $\mathcal{C}_q$, and the mean-sample of $\mathcal{S}_\mathbf{x}$ as follow:

$$\mu_q = \frac{1}{|\mathcal{C}_q|}\sum_{\mathbf{x}\in\mathcal{C}_q} \mathbf{x}, \forall q \in \{1,\ldots,k\}$$
$$\mu = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i.$$

Then, apart from the adapted Silhouette score (5) that we use here the following scores can be used to approximate the purity size objective function:

- Calinski-Harabasz score

$$S_{ch}(R(\mathcal{S}_\mathbf{x})) = \left[\frac{(n-k)B}{(k-1)\sum_{q=1}^{k} W_q}\right] \in [0,+\infty),$$

  with $B = \sum_{q=1}^{k} |\mathcal{C}_q| \|\mu_q - \mu\|^2$ is the inter-group variance, and $W_q = \sum_{\mathbf{x}\in\mathcal{C}_q} \|\mathbf{x} - \mu_q\|^2$ is the intra-group variance, for all $q \in \{1,\ldots,k\}$. It translates that good partitioning should maximize the average inter-group variance and minimize the average intra-group variance; some well-known clustering algorithms, such as K-means [35], maximize this criterion by construction.
- Davies-Bouldin score

$$S_{db}(R(\mathcal{S}_\mathbf{x})) = \left[\frac{1}{k}\sum_{q=1}^{k} \max_{j\neq q}\left(\frac{\bar{\delta}_q + \bar{\delta}_j}{d(\mu_q,\mu_j)}\right)\right] \in (+\infty, 0],$$

  with $\bar{\delta}_q = \frac{1}{|\mathcal{C}_q|}\sum_{\mathbf{x}\in\mathcal{C}_q} d(\mathbf{x},\mu_q)$ is the average distance of all samples in the group to their mean-sample group, for all $q \in \{1,\ldots,k\}$.
- Dunn score

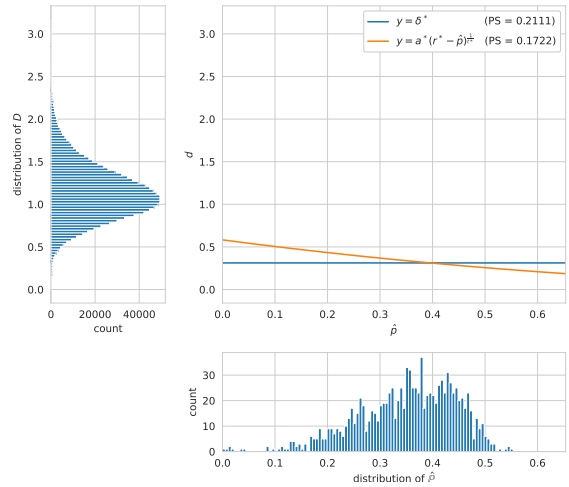
$$S_d(R(\mathcal{S}_\mathbf{x})) = \left[\frac{\min_{q,j} d(\mu_q,\mu_j)}{\max_q \Delta_q}\right] \in [0,+\infty),$$

  with $\Delta_q = \max_{\mathbf{x},\mathbf{x}'\in\mathcal{C}_q} d(\mathbf{x},\mathbf{x}')$ being the diameter of the group $\mathcal{C}_q$, similar to the Calinski-Harbasz score, we aim to maximize the minimum distance between the mean-sample groups and minimize the maximum group diameter.

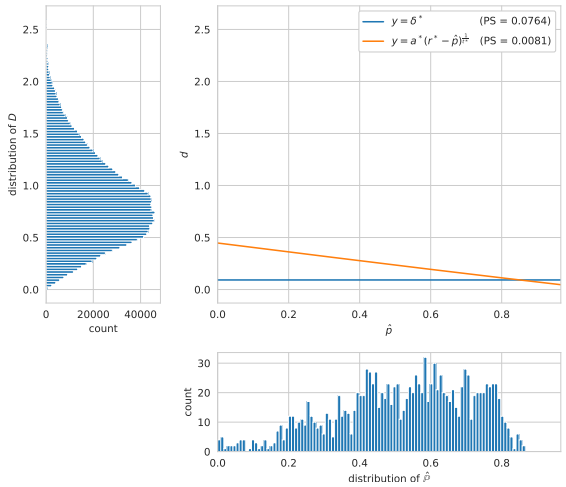
C More empirical results: Rips graph vs σ-Rips graph

In this section we show an empirical comparison in Figure 6 between the Rips graph and the σ-Rips graph using the considered datasets. Overall, the σ-Rips graph achieves a better *PuritySize* (PS) score than the Rips graph, except for coil-20 datasets where the scores are comparable. Note that the retrieved decision curves of the σ-Rips graphs are anti-correlated to the density estimation, which validates our intuition, and motivates the σ-Rips graph formulation.

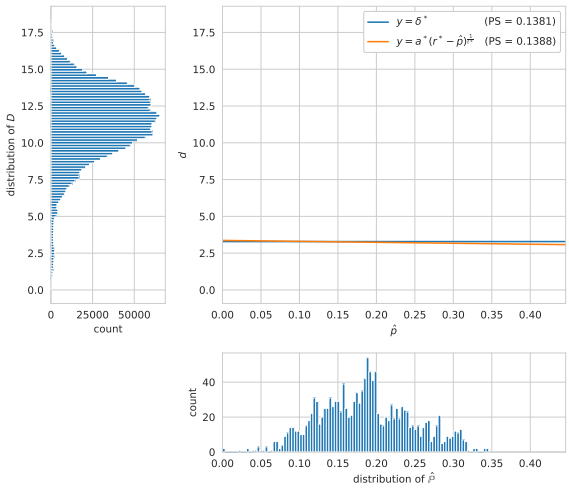
Figure 6: Comparison study between the Rips graph and the σ-Rips graph over all datasets, the Purity Size score is reported for each minimizer.



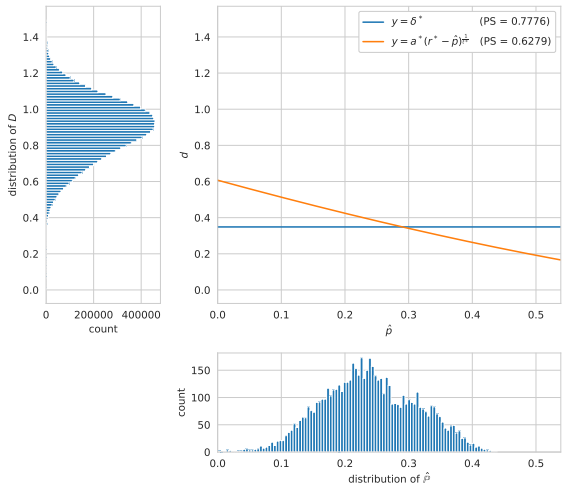
(a) protein



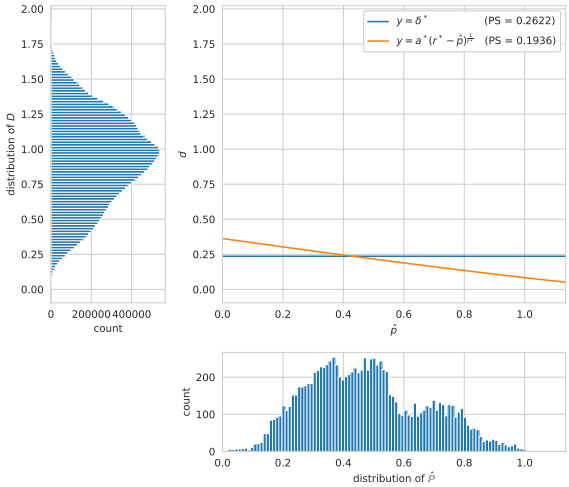
(b) banknote



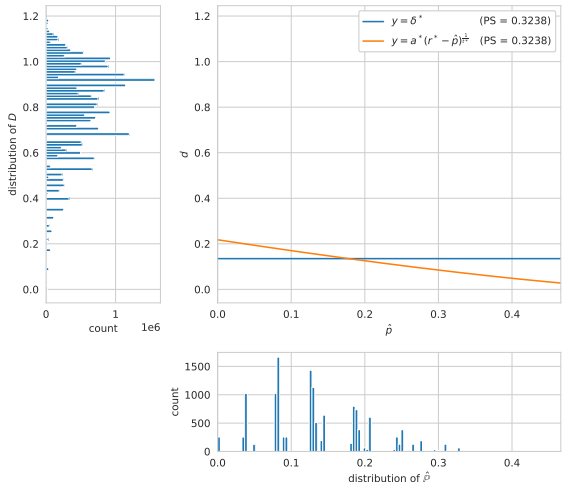
(c) coil-20



(d) isolet



(e) pendigits



(f) nursery