Pool-Based Active Learning with Proper Topological Regions

Lies Hadjadj, Emilie Devijver, Remi Molinier, Massih-Reza Amini

Pool-Based Active Learning with Proper Topological Regions

Lies Hadjadj[†], Emilie Devijver[†], Remi Molinier[‡], Massih-Reza Amini[†]
{Firstname.Lastname}@univ-grenoble-alpes.fr
[†] Computer Science Laboratory (LIG)
[‡] Department of Mathematics (IF)
University of Grenoble Alpes, France

Abstract

Machine learning methods usually rely on large sample size to have good performance, while it is difficult to provide labeled set in many applications. Pool-based active learning methods are there to detect, among a set of unlabeled data, the ones that are the most relevant for the training. We propose in this paper a meta-approach for pool-based active learning strategies in the context of multi-class classification tasks based on Proper Topological Regions. PTR, based on topological data analysis (TDA), are relevant regions used to sample cold-start points or within the active learning scheme. The proposed method is illustrated empirically on various benchmark datasets, being competitive to the classical methods from the literature.

1 Introduction

In recent years, machine learning has found gainful applications in diverse domains, but it still has a heavy dependence on expensive labeled data: Advances in cheap computing and storage have made it easier to store and process large amounts of unlabeled data, but the labeling often needs to be done by humans or using costly tools. Therefore, there is a need to develop general domain-independent methods to learn models effectively from a large amount of unlabeled data at the disposal, along with a minimal amount of labeled data: this is the framework of semi-supervised learning. Active learning aims explicitly to detect the observations to be labeled to optimize the learning process and efficiently reduce the labeling cost. The primary assumption behind active learning is that machine learning algorithms could reach a higher level of performance while using a smaller number of training labels if they were allowed to choose the training dataset (Settles, 2009). The most common active learning approaches are pool-based methods (Lewis and Catlett, 1994) based on a set of unlabeled observations. First, some points are labeled to train a classification model, and then, at each iteration, we choose unlabeled examples to query based on the predictions of the current model and a predefined priority score. These approaches show their limitations in low-budget regime scenarios because they need a sufficient budget to learn a weak model (Pourahmadi et al., 2021).

The literature has shown that for active learning to operate in a low-budget regime successfully, there is a need to introduce a form of regularization in training (Guyon et al., 2011) usually found in other sub-domains, such as semi-supervised learning or self-learning (Chapelle et al., 2006). Another line of work shows that the choice of the initial seed set in these approaches significantly impacts the end performance of their models (Hu et al., 2010; Chen et al., 2022), also known as the cold-start problem in active learning.

We close the gap in this paper by providing a theoretically founded meta-approach for pool-based active learning based on concepts from topological data analysis (TDA) to improve performance in a low-budget regime and avoid the cold-start problem. TDA aims to extract information on the structure of the data by examining its topological properties (Edelsbrunner and Harer, 2010), the insight being that nontrivial topologies should be exploited to improve data analysis (Carlsson, 2012). This structure can be detected by flexible tools based on algebraic topology, for example, using persistent homology based on Rips complexes (Hausmann, 1995): topological information is then encoded with persistence modules and diagrams (Edelsbrunner and Harer, 2010). It has already shown impressive results in machine learning (Rieck et al., 2020; Jiang et al., 2021; Krishnapriyan et al., 2021), especially for clustering. Many recent papers have benefited from these topological insights to understand the structure of the data: Singh et al. (2007) use persistence homology to extract molecular topological fingerprints (MTFs) based on the persistence of molecular topological invariants, Lum et al. (2013) use topological persistence to efficiently encode fMRI datasets, Carlsson and Gabrielsson (2020) 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. (2020) 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.

In this paper, we propose to extend `ToMATo` Chazal et al. (2013), a persistent-based clustering algorithm that respects the underlying topology, to detect proper topological regions where one can safely propagate labeling (assuming that the clustering is coherent with the metric). More precisely, our approach is based on the following:

- the introduction of proper topological regions using the σ-Rips graph based on an adaptive threshold function and the extension of `ToMATo`'s theoretical guarantees to the σ-Rips graph;
- the use of proper topological regions in a zero-shot learning method and a pool-based active learning scheme.

This is illustrated in an empirical study with several active learning strategies which shows that our approach for zero-shot learning and pool-based active learning improves over classical methods on several datasets.

The remainder of the paper is organized as follows. Section 2 describes the related work. The framework is introduced in Section 3. The method is developed in Section 4; and illustrated in Section 5. Finally, Section 6 concludes the paper.

2 Related literature

Different attempts have been made to reduce the annotation burden of machine learning algorithms. We can refer to the remarkable advances made in semi-supervised learning (Amini and Usunier, 2015; Berthelot et al., 2019). These methods take as input a small set of labeled training data together with a large number of unlabeled examples. They introduce a form of consistency regularization to the supervised loss function by applying data augmentation using unlabeled observations (Chapelle et al., 2006). The most commonly known pool-based strategies are uncertainty sampling (Lewis and Catlett, 1994; Zhu et al., 2008), margin sampling, and entropy sampling strategies (Settles, 2009). Some proposed strategies rely on the query-by-committee approach (Yan et al., 2011; Lakshminarayanan et al., 2017), which learns an ensemble of models at each round. Query by bagging and query by boosting are two practical implementations of this approach that use bagging and boosting to build the committees (Abe and Mamitsuka, 1998). There has been exhaustive research on how to derive efficient disagreement measures and query strategies from a committee, including vote entropy, consensus entropy, and maximum disagreement (Settles, 2009), whereas Ali et al. (2014) introduces model selection for a committee. Some research focuses on solving a derived optimization problem for optimal query selection, e.g. in Roy and McCallum (2001) they use Monte Carlo estimation of the expected error reduction on test examples. In contrast, other strategies employ Bayesian optimization on acquisition functions such as the probability of improvement or the expected improvement (Garnett, 2022), and in Auer et al. (2002), the authors propose to cast the problem of selecting the most relevant active learning criterion as an instance of the multi-armed bandit problem. Aside from the pool-based setting, in the stream-based setting (Lughofer, 2012; Baram et al., 2004), each unlabeled sample is given to the learner individually, and he queries its label if he finds it helpful.

Recent advances in active learning propose enhancing the pool-based methods by extracting knowledge from the distribution of unlabeled examples (Bonnin et al., 2011). Perez et al. (2018) propose to use clustering of unlabeled examples to boost the performance of pool-based active learners, with the expert annotating at each iteration cluster rather than single examples. Such a strategy effectively reduces the annotation effort, assuming that the cost of cluster annotation is comparable to single example labeling, as used in Citovsky et al. (2021) to operate on large-scale data. Similarly, Krempl et al. (2015) proposes to combine clustering with Bayesian optimization in the stream-based setting. Yu and Hansen (2017) propose a two-stage clustering constraint in the active learning algorithm, a first exploration phase to discover representative clusters of all classes, and a post-clustering reassignment phase where the learner is constrained on the initial clusters found at the first stage. Clustering methods also show promising results for addressing the cold-start problem in pool-based active learning strategies (Hu et al., 2010; Chen et al., 2022). In Urner et al. (2013) authors propose a procedure for binary domain feature sets to recover the labeling of a set of examples while minimizing the number of queries. They show that this routine reduces label complexity for training learners. Recently, many studies have explored the use of clustering/segmentation for active sample selection in real applications Andresini et al. (2023); Thoreau et al. (2022).

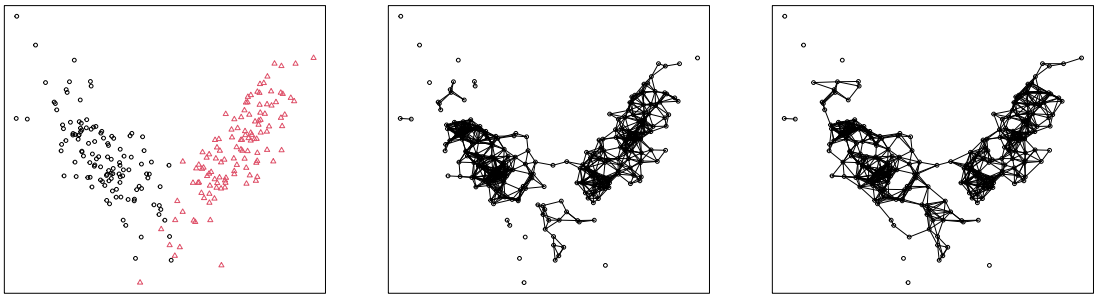


Figure 1: (a) A sample of 240 points, generated from a mixture of two bivariate Gaussian distributions. Colors represent the true classes. (b) Associated Rips graph as defined in Def. 1 with $\delta = 0.5$. (c) Associated σ-Rips graph as defined in Def. 2, using the parametric form given in Eq. (2) with $\delta = 0.5$, $r = 1.08$ and $t = 1/5$.

3 Framework and topological considerations

We introduce in this section the framework of active learning and the topological background needed to develop the proposed method. First, we introduce the framework and the main topological notions. Then we define the persistence and upper-star filtrations. Finally, we provide a comparison of persistence diagrams for Rips graph and σ-Rips graph, which allows us to extend `ToMATo` results to the σ-Rips graph.

3.1 Framework and notations

We consider a multi-class classification problem such that the input space is $\mathcal{X} \subset \mathbb{R}^m$ and the output space $\mathcal{Y} = \{1, \ldots, c\}$ is a set of unknown classes of cardinal $c \in \mathbb{N}, c \geq 2$. Let d be a fixed distance on $\mathbb{R}^m$. In pool-based active learning, we observe a sample set $\mathcal{S}_{\mathbf{x}} = \{\mathbf{x}_i\}_{i=1}^n$ coming 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 every observation $\mathbf{x}_i$, for $1 \leq i \leq n$ at some (expensive) cost. We denote $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ the labeled data sample of size n, which we do not have access to, generated by some unknown joint distribution over $\mathcal{X} \times \mathcal{Y}$.

In our method, as generally is the case in classification algorithms, we assume that close samples (with respect to d) are associated with similar labels, also known as the *smoothness assumption*. In that setting, one can consider neighborhood graphs on the unlabeled sample $\mathcal{S}_{\mathbf{x}}$. A graph is denoted as a couple (V, E) with V 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, or more generally Rips complexes (Chazal et al., 2014), are classical in topology and are classically used in TDA, in particular with persistent homology. However, class similarity might be different over the metric space. For example, lower

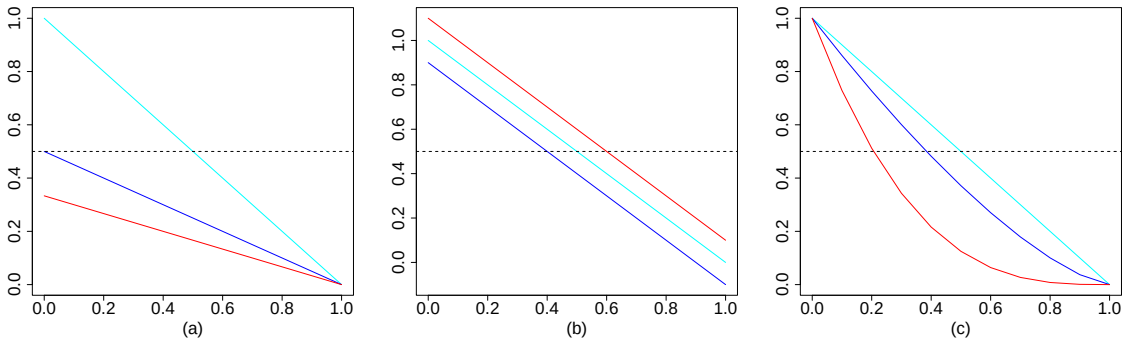


Figure 2: Representation of $s : u \mapsto \delta(r-u)^{1/t}$ as a proxy of the parametric form of σ given in Eq. (2), varying the parameters. By default, all the parameters are fixed to 1. We vary in (a) $\delta \in \{1, 0.5, 1/3\}$, in (b) $r \in \{0.9, 1, 1.1\}$, in (c) $t \in \{1, 0.7, 1/3\}$. Dashed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

is the density, weaker is the chance to detect a structure within points. Consequently, we need to generalize the definition of the Rips graph to take into account such cases, namely the σ-Rips graph $R_{\sigma(\cdot)}(\mathcal{S}_\mathbf{x})$ for 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 $d(\mathbf{x}_i, \mathbf{x}_j) \leq \sigma(\mathbf{x}_i, \mathbf{x}_j)$.*

Those two notions of neighborhood graph are illustrated in Figure 1 to understand the differences. When the density is lower (few points), the σ-Rips graph is more connected, to enforce the structure to appear.

The σ-Rips graph can be seen as a generalization of the Rips graph, which considers constant threshold function, or as a δ-Rips graph 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 of the topological properties of Rips graphs on metric spaces are true for Rips graphs on non-metric spaces, as mentioned in Chazal et al. (2014, Section 4.2.5).

In this work, we choose 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 $(\delta, r) \in (\mathbb{R}_+^*)^2$ such that $r > \max_\mathbf{x} \mathbb{P}(\mathbf{x})$. The temperature parameter t controls the curvature, the max term ensures that the function is symmetric. Then, δ and r are, respectively, dilatation and translation parameters. This parametric form is illustrated in Figure 2. We show in Section 5.1 that the curve resulting from the best parameters of our function confirms our intuition on the class similarity being a density-aware measure.

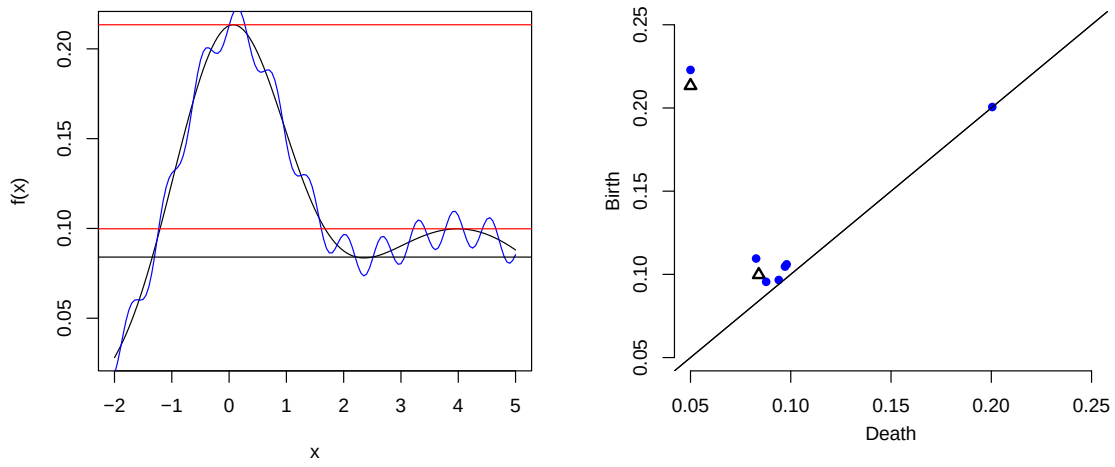


Figure 3: Persistence diagram $D\mathbb{P}$ (on the right) associated to the upper-star filtration of the functions whose graph is given on the left. The function in black is smooth, corresponding persistence diagram is given with black triangles. There are two elements in the persistence diagram, corresponding to the two peaks of the function. Peaks and valleys are highlighted on the left in red and black respectively, to read the values of birth and death in the persistent diagram. Notice that the high left point on the diagram is actually a point "at infinity", i.e. its death is $-\infty$. The function in blue is a noisy version with many peaks and valleys. Its persistence diagram corresponds to the blue points. There are two bluepoints a bit above the two black triangle, corresponding the the main topological features, and many points close to the diagonal, with low prominence, suggesting that this is topological noise.

3.2 Persistence and upper-star filtrations

In order to detect the underlying topology from a point cloud, our method is based on the notion of persistence, and more precisely, persistent homology, a classical tool in TDA (Singh et al., 2007; Edelsbrunner and Harer, 2010; Carlsson, 2012).

A persistence module is a sequence of vector spaces $\mathbf{X} = (X_\alpha)_{\alpha\in\overline{\mathbb{R}}}$ where $\overline{\mathbb{R}} = \mathbb{R}\cup\{-\infty,+\infty\}$ together with linear maps $\varphi_{\beta,\alpha}\colon X_\beta \to X_\alpha$ whenever $\alpha\leq\beta$ (setting $X_\alpha\to X_\alpha$ as the identity) and such that, if $\alpha\leq\beta\leq\gamma$, then $\varphi_{\gamma,\alpha} = \varphi_{\beta,\alpha}\circ\varphi_{\gamma,\beta}$. In such a framework, one can study the persistence of a vector. More precisely, given $\alpha\in\overline{\mathbb{R}}$ and $v\in X_\alpha$, we say that v is born at time[1] α if v is not in the image of $\varphi_{\beta,\alpha}$ for all $\beta>\alpha$, and we say that it dies at time $\gamma\leq\alpha$ if $\varphi_{\alpha,\gamma}(v)=0$ but $\varphi_{\alpha,\gamma'}(v)\neq 0$ for all γ' with $\gamma<\gamma'<\alpha$. Globally, we usually consider bases of the $X'_\alpha s$ (and related to the linear maps $(\varphi_{\alpha,\beta})_{\beta\leq\alpha}$) and summarize their persistence with a persistence diagram. More precisely, the persistence diagram $D\mathbf{X}$ of a persistence module $\mathbf{X}$ is the multi-

[1]Here, according to the way our spaces are connected, the time is flowing in the other direction: from $+\infty$ to $-\infty$.

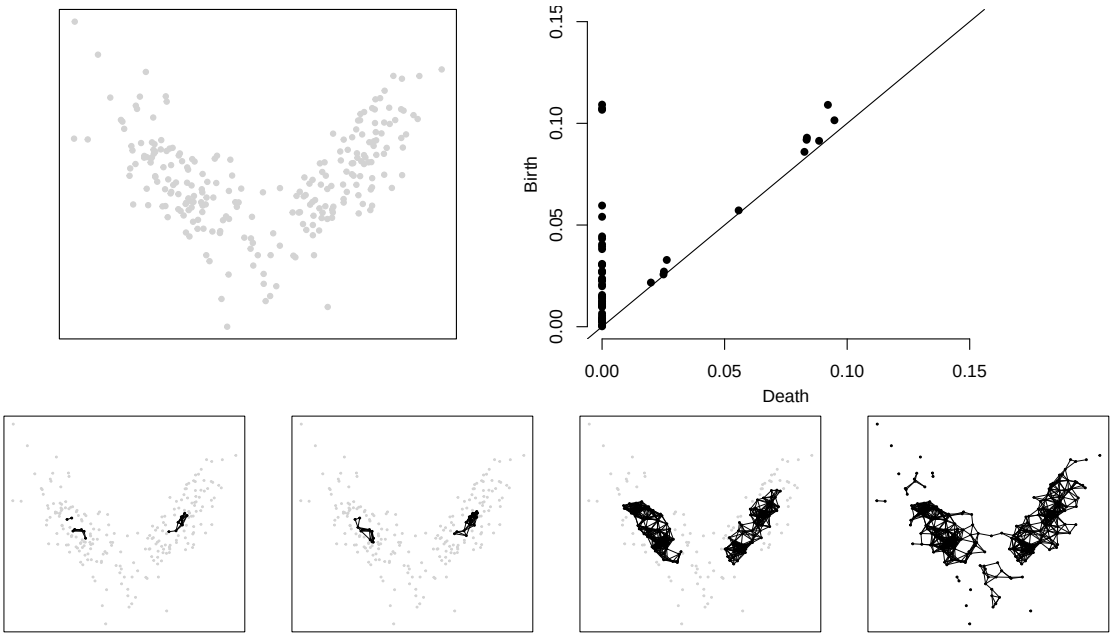


Figure 4: A point cloud generated from a mixture of two gaussians (top left), some realizations of its upper-star Rips filtration for $\delta = 0.5$ and $\alpha \in \{0.1, 0.095, 0.05, 0\}$ (bottom, from left to right) and the associated persistent diagram (top right).

set[2] of points in $\overline{\mathbb{R}}^2$ consisting in the diagonal[3] $\Delta = \{(x, x) \mid x \in \overline{\mathbb{R}}\}$ and points (i, j) for each basis element appearing at time i and dying at time $j < i$. When reading a persistence diagram, one should consider the distance of the points to the diagonal, i.e., their prominence. A point with low prominence should be considered as topological noise (they do not live long) whereas a point with high prominence as relevant topological information. Persistence modules and diagrams are often used with homology, and we refer the reader to Hatcher (2000) for more details. Here we only use the 0-dimensional homology, which detects connected components. More precisely, if T is a topological space or a graph, $H_0(T)$ is the vector space spanned by the (path) connected components of T. Moreover, a continuous map $T_1 \to T_2$, between spaces or a graph homomorphism between graphs, induces a natural linear application $H_0(T_1) \to H_0(T_2)$. With that in hand, a classical example of persistence module is induced by the upper-star filtration of a function $\mathbb{P}\colon \mathcal{X} \to \mathbb{R}^+$. If $\alpha \leq \beta$ are two reals, then there is an inclusion $\mathbb{P}^{-1}([\beta, +\infty]) \subseteq \mathbb{P}^{-1}([\alpha, +\infty])$, and this induces linear maps $H_0(\mathbb{P}^{-1}([\beta, +\infty])) \to H_0(\mathbb{P}^{-1}([\alpha, +\infty]))$ which defines a persistence module. We denote by $D\mathbb{P}$ the associated persistence diagram. This notion is illustrated in Figure 3 with two functions: the black one, very smooth with only two peaks, and the blue one, a noised version of the first one.

The persistence module that we consider here is 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*

[2]A multi-set A is a set with potential repetitions of elements, where we denote $\mu(p)$ the multiplicity of point $p \in \mathrm{Supp}(A)$. It can be denoted $A = \bigcup_{p \in |A|} \coprod_{i=1}^{\mu(p)} p$ with $\mathrm{Supp}(A)$ the support of A.

[3]The multiplicity of a point in the diagonal is $+\infty$.

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 $\mathcal{R}_\delta(\mathcal{S}_\mathbf{x}, \mathbb{P}) = (R_\delta(\mathcal{S}_\mathbf{x} \cap \mathbb{P}^{-1}([\alpha, +\infty]))_{\alpha \in \overline{\mathbb{R}}}$. Such a nested family of graphs gives rise to a persistence module

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

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

This notion is illustrated in Figure 4.

We define similarly the upper-star σ-Rips filtration of $\mathbb{P}$, $\mathcal{R}_{\sigma(\cdot)}(\mathcal{S}_\mathbf{x}, \mathbb{P})$ and the associated persistence module $\mathbf{R}_{\sigma(\cdot)}(\mathcal{S}_\mathbf{x}, \mathbb{P})$ and persistent diagram $D\mathbf{R}_{\sigma(\cdot)}(\mathcal{S}_\mathbf{x}, \mathbb{P})$. In the next section, we give some tools to control the difference between $D\mathbf{R}_\delta(\mathcal{S}_\mathbf{x}, \mathbb{P})$ and $D\mathbf{R}_{\sigma(\cdot)}(\mathcal{S}_\mathbf{x}, \mathbb{P})$.

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

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

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

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

One can control the bottleneck distance between two persistence diagrams from upper star Rips filtration by comparing the evolution of the connected components along the filtration which can be track with the appearance level.

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[4] in $R_\delta(\mathcal{S}_\mathbf{x})$ from the vertex $\mathbf{x}_i$ to the vertex $\mathbf{x}_j$.

For example, in Figure 4, between $\alpha = 0.1$ and $\alpha = 0.095$, we see two connected components that are finally connected (on the cluster on the left). The appearance level of the corresponding points is a value between 0.095 and 0.1.

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

Then, we are able to bound the bottleneck distance between the Rips graph and the σ-Rips graph.

[4] A path γ in a graph R is a sequence of vertices of R where two consecutive vertices of p are adjacent in R.

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,*

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

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 stands in Appendix A. It relies on Chazal et al. (2009) and their notion of ϵ-interleaving. The main idea is that one can control the birth and the death of connected components along the filtration by controlling the changes in appearance levels.

This theorem means that, when switching from the metric distance d to a closed (possibly) non-metric distance $\hat{d}$ defined in (1) (and then from the Rips graph to the σ-Rips graph), the dendrogram induced by the upper star Rips graph is mostly the same during the persistence process.

4 Proper topological regions and their use in active learning

In this section, we start by defining the proper topological regions. We then use the proper topological regions for zero-shot learning and for pool-based active learning.

4.1 Proper topological regions

The main tool used in our method is the notion of topological regions that are based on the algorithm `ToMATo` (Chazal et al., 2013). `ToMATo` is a clustering method that uses the hill climbing algorithm on a Rips graph $R_\delta(\mathcal{S}_{\mathbf{x}})$ along with a merging rule on the Rips graph's persistence. It depends on a merging hyperparameter $\tau \geq 0$ which drives the granularity: it keeps only clusters with prominence higher than τ. It can be easily adapted to work with a σ-Rips graph $R_{\sigma(\cdot)}(\mathcal{S}_{\mathbf{x}})$ by considering the non-metric space $(\mathcal{S}_{\mathbf{x}}, \hat{d})$ with $\hat{d}$ introduced in Eq. (1). The topological regions correspond to the clusters given by `ToMATo` for a σ-Rips graph, defined formally as follows.

Definition 6. *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*

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

When the set of covariates $\mathcal{S}_{\mathbf{x}}$*, the underlying density* $\mathbb{P}$*, and the parameters are understood, we will simply denote* TR*.*

For $\mathrm{TR}\colon \mathcal{S}_{\mathbf{x}} \to \{1, \ldots, k\}$ a clustering into k topological regions of $\mathcal{S}_{\mathbf{x}}$, we denote $\mathcal{L}_{\mathrm{TR}}^{\mathbb{P}}$ the labeling function that propagates, in a given topological region, the label of

the sample with the highest density with respect to $\mathbb{P}$:

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

If one could have access to the labeled data, we define the Purity Size function PS as the objective function that considers the labeling error when propagating the labels inside the topological regions with $\mathcal{L}^{\mathbb{P}}_{\mathrm{TR}}$, penalized by the number of topological regions k in TR:

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

Then, we introduce the notion of proper topological regions, that will be the key 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 $\mathrm{TR}^{\mathcal{S}_{\mathbf{x}},\mathbb{P}}_{\delta^*,r^*,t^*,\tau^*}$ where*

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

However, in our active learning context, we need to use an unsupervised objective function. We consider a trade-off between the Silhouette score[5] and the coverage compactness of a clustering TR of $\mathcal{S}_{\mathbf{x}}$ into k topological regions $\{R_1, \dots R_k\}$: for $1 \leq q \leq k$, let π_q be the cardinal 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

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

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

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

Note that the trade-off parameter λ in (5) is key in uncovering the proper topological regions of the sample set $\mathcal{S}_{\mathbf{x}}$. High values of λ penalize the coverage compactness, resulting in partitions with a high degree of agglomeration, i.e. fewer topological regions with large cardinals. 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 result in

[5]Other potential unsupervised criteria typically used to assess the clustering quality are discussed in Appendix C, but we have observed on an empirical study the benefit of the Silhouette.

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*[6] **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 $\widehat{\mathrm{PTR}}$.

highly fragmented partitions with many groups with small cardinals, and the Silhouette score tends to converge to graphs with a single non-singleton connected component and many singletons.

Thus, the optimization problem (4) is approximated by the following:

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

Unfortunately, this optimization problem is too costly, because the set of parameters leads to running the `ToMATo` function many times. So instead, we propose to solve the following proxy, which is 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 the `ToMATo` algorithm.

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 Chazal et al. (2013), which computes the root-mean-squared distance to the ℓ nearest neighbors of the considered query point: for all $i \in \{1, \ldots, n\}$,

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

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

Algorithm 2 Zero-shot learning based on proper topological regions

Require: $\mathcal{S}_{\mathbf{x}} := \{\mathbf{x}_i\}_{i=1}^n$, oracle $\mathcal{O}$, budget $\mathcal{B}$, $\hat{\mathbb{P}}$, and proper topological regions $\widehat{\text{PTR}}$.

1: Detect the $\mathcal{B}$ largest proper topological regions $R_1, \ldots, R_{\mathcal{B}}$ of $\widehat{\text{PTR}}$.
2: Set $\mathcal{S}_{\mathbf{x}}^0 = \cup_{q=1}^{\mathcal{B}} R_q$
3: **for** $\mathbf{x}_i \in \mathcal{S}_{\mathbf{x}}^0$ **do**
4: Label the corresponding points using the oracle $\mathcal{O}$: $\hat{y}_i = \mathcal{L}^{\hat{\mathbb{P}}}_{\widehat{\text{PTR}}}(\mathbf{x}_i)$.
5: **end for**
6: **Output:** $\hat{\mathcal{S}}^0 = (\mathbf{x}_i, \hat{y}_i)_{\mathbf{x}_i \in \mathcal{S}_{\mathbf{x}}^0}$

The whole procedure used to approximate the proper topological regions is data-driven using the unlabeled set $\mathcal{S}_{\mathbf{x}}$, and we will denote in the following $\widehat{\text{PTR}}$ the corresponding estimated proper topological regions with parameters $(\hat{\delta}, \hat{r}, \hat{t}, \hat{\tau})$. We describe in Algorithm 1 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 $\hat{\tau}$, solution to our optimization problem given in Eq. (4), for the proper topological regions of $\mathcal{S}$.

As we are extending `ToMATo` to σ-Rips graph, proper topological regions enjoy the same theoretical guarantees as the topological region given by `ToMATo` applied to the usual Rips graph. More precisely, under some topological assumptions on the persistence diagrams, there is a range of values of τ such that the number of topological regions output by $\texttt{ToMATo}_\tau(R_\delta(\mathcal{S}_{\mathbf{x}}), \mathbb{P})$ is equal to the number of peaks (i.e., local maximum) of $\mathbb{P}$ with prominence at least τ in $D\mathbb{P}$ with high probability with respect to n. Moreover, each of these topological regions contain a neighborhood of the basins of attraction[7] of the corresponding peak. In this context, Theorem 1 tells us that, under reasonable conditions on the threshold function σ, we get about the same persistence diagram when considering the σ-Rips graph, and thus we derive the same kind of theoretical guarantees when applying `ToMATo` with a σ-Rips graph. Those results are summarized in Appendix B.

4.2 Proposed meta-learning strategy

In this section, we introduce two strategies based on the proper topological regions found by Algorithm 1. First, they are used in a zero-shot learning algorithm and second, in a pool-based strategy in a meta fashion, independently from the estimator and the strategy. To do so, we propose to use the label propagation scheme on proper topological regions in order to increase the sample size for training in a small budget scenario with a fixed number of calls to the oracle, as proposed in Perez et al. (2018); Citovsky et al. (2021).

Zero-shot learning We observe the unlabelled set $\mathcal{S}_{\mathbf{x}}$, and we have access to the proper topological regions $\widehat{\text{PTR}}$ estimated by Algorithm 1; to an oracle to give few labels (a budget $\mathcal{B}$ is considered); and the density estimation $\hat{\mathbb{P}}$. The strategy is the following: we label the $\mathcal{B}$ largest proper topological regions $R_1, R_2, \ldots, R_{\mathcal{B}}$ using the

[7] The basin of attraction of $\mathbb{P}: \mathcal{X} \to \mathbb{R}$ of a peak p of $\mathbb{P}$ corresponds to all the points of $\mathcal{X}$ flowing into p along the flow defined by the gradient vector field of $\mathbb{P}$.

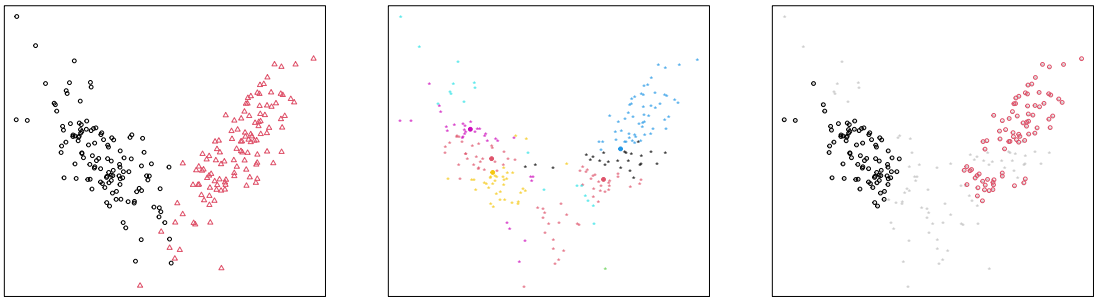


Figure 5: Illustration of Algorithm 2 (a) data with the oracle for the labels. (b) clustering given by tomato, describing the proper topological regions with the estimated parameters given by Algorithm 1 (c) output of Algorithm 2, where budget = 5, and with propagation. Not all points are labelled, and not all labels are sure.

labelling function $\mathcal{L}^{\hat{\mathbb{P}}}_{\widehat{\text{PTR}}}$ defined in Eq. (3): we ask to the oracle $\mathcal{B}$ points (the one with highest density in each R_q), and we then label $\sum_{q=1}^{\mathcal{B}} |R_q|$ points by propagating in each topological region. We denote by $\hat{\mathcal{S}}^0$ this first set of labeled points, which includes true labels obtained directly from the oracle, and estimated labels while diffusing the true labels to the topological regions. This procedure is summarized in Algorithm 2 and illustrated in Figure 5. The benefit to use proper topological regions instead of any clustering method is in details. No structure is assumed, as in k-means for example where clusters have a spherical shape. Here, only the topology is important, thus the algorithm can retrieve connected components even with an ambiguous shape. Moreover, fine hyperparameter tuning in `ToMATo` allows to merge or distinguish between regions in a precise way (see Results 1 and 2 in Appendix B).

Meta-approach for training pool-based active learning The idea is again to diffuse the labels asked to the oracle to the proper topological regions to get more (pseudo)-labeled points. The unlabeled sample set $\mathcal{S}_{\mathbf{x}}$, the oracle $\mathcal{O}$, the budget $\mathcal{B}$, the density estimation $\hat{\mathbb{P}}$ and the proper topological regions $\widehat{\text{PTR}} = \{R_1, \ldots, R_k\}$ estimated by Algorithm 1 are common inputs for active learning techniques. Additionally, a pool-based active learning technique $h_{st}(\mathcal{S}_{\mathbf{x}}, \mathcal{B})$ is also provided as input. The algorithm then performs r rounds of active learning. Within each round, the active learner agent is asked to detect $\mathcal{B}$ points, which are defining at most $\mathcal{B}$ topological regions. If two points detected by the active learner agent belong to the same topological region, we use the extra budget to label the largest topological regions without any detected points. Then, the oracle is asked to label $\mathcal{B}$ unlabeled examples that correspond to the points of high density in each considered topological region. The labeled set $\hat{\mathcal{S}}^r$ returned by the algorithm has $r \times \mathcal{B}$ labels and many pseudo-labels, given by the label propagation with $\mathcal{L}^{\hat{\mathbb{P}}}_{\widehat{\text{PTR}}}$ to increase the size of the training set during each round of active learning.

Remark that the choice for the extra budget to label the largest unlabelled proper topological regions is not driven by an active learner agent. Good results have been observed in Section 5, but one can think of different ways to use the extra budget, such as, for example, asking the learner for more examples to label. Algorithm 3 describes

Algorithm 3 Pool-based active learning on proper topological regions (PAL_{PTR})

Require: $\mathcal{S}_{\mathbf{x}} := \{\mathbf{x}_i\}_{i=1}^{n}$, oracle $\mathcal{O}$, budget $\mathcal{B}$, $\hat{\mathbb{P}}$, proper topological regions $\widehat{\text{PTR}} = \{R_1, \ldots, R_k\}$, active learner agent $h_{st}(\mathcal{S}_{\mathbf{x}}, \mathcal{B})$ with an underlying pool-based strategy st, and r the active training rounds.

Compute $\hat{\mathcal{S}}^0$ using Algorithm 2
for $u = 0, \ldots, r-1$ **do**
    Train the active learner agent $h_{st}(\hat{\mathcal{S}}^u, \mathcal{B})$.
    Ask a set $S_{\mathbf{x}}$ from h_{st} of size $\mathcal{B}$.
    $\mathcal{S}_{\mathbf{x}}^{u+1} = \cup_{R_q \cap S_{\mathbf{x}} \neq \emptyset} R_q$
    if $\tilde{\mathcal{B}} = |\{q \mid R_q \cap S_{\mathbf{x}} \neq \emptyset\}| < \mathcal{B}$ **then**
        Detect the $\mathcal{B} - \tilde{\mathcal{B}}$ unlabeled largest topological regions without any points which do not intersect $S_{\mathbf{x}}$ and add them to $\mathcal{S}_{\mathbf{x}}^{u+1}$
    end if
    for $\mathbf{x}_i \in \mathcal{S}_{\mathbf{x}}^{u+1}$ **do**
        Label the corresponding points using the oracle $\mathcal{O}$: $\hat{y}_i = \mathcal{L}^{\hat{\mathbb{P}}}_{\widehat{\text{PTR}}}(\mathbf{x}_i)$.
    end for
    $\hat{\mathcal{S}}^{u+1} = \hat{\mathcal{S}}^u \cup \{(\mathbf{x}_i, \hat{y}_i) \mid \mathbf{x}_i \in \mathcal{S}_{\mathbf{x}}^{u+1}\}$
end for
Output: the labeled set $\hat{\mathcal{S}}^r$

this meta-approach.

5 Empirical results

We conduct a number of experiments aimed at evaluating how the proposed approach can identify valuable examples to be labeled for learning. To this end, we consider two scenarios for the identification of an initial training set to be labeled from an unlabeled set, and the increase of the training sample size during the rounds with active learning while operating under a low-budget regime.

We carry out experiments on data collections that are frequently used in active learning. Table 1 presents statistics of these datasets.

For the metric function, d, we consider the Euclidean distance, and we choose the Tree-structured Parzen Estimator (TPE) (Bergstra et al., 2011) for the optimization procedure of Algorithm 1, with a number of trials $l = 500$, and a step size s of 0.01 for the line search procedure. We estimate $\mathbb{P}$ using Eq. (9) with the distance to measure based on the ℓ nearest neighbors with ℓ the sample size, if smaller than 2000, and 2000 elsewhere. In all our experiments, we use the random forest classifier (Ho, 1995) as the base estimator for the different strategies with default parameters, we also consider several budgets $\mathcal{B} \in \{3, 10, 20\}$, and 20 stratified random splits, with 70% of the data in the training set and 30% in the test set. We report the balanced classification accuracy (Brodersen et al., 2010) over all experiments. Regarding the data preprocessing, we drop sample duplicates and samples with null values. Then we apply a standard min-max normalization to the filtered datasets.

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 (Higuera et al., 2015)	756	324	77	8	0.70
banknote (Romano et al., 2021)	943	405	4	2	0.83
coil-20 (Yang et al., 2011)	1008	432	1024	20	1.00
isolet (Fanty and Cole, 1991)	4366	1872	617	26	0.99
pendigits (Romano et al., 2021)	7694	3298	16	10	0.92
nursery (Romano et al., 2021)	9070	3888	8	4	0.09

5.1 Rips graph vs σ-Rips graph

To validate our hypothesis of a density-aware threshold given by Eq. (2) for class similarity and to motivate our generalization of the Rips graph to express this notion, we present a comparison study in Figure 6 between the Rips and the σ-Rips graphs on the protein dataset. The results for the other collections are shown in Figure 8 in Appendix D.1.

The plot represents the threshold of the best Rips graph and σ-Rips graph in minimizing the Purity Size cost function. The Rips graph's threshold (in blue) is a constant presented as a horizontal line in the plot.

In this figure, we also include two more side plots which are the distribution of the dataset's density estimation $\hat{\mathbb{P}}$ under the x-axis and the distance matrix D's distribution of Euclidean distances on the left of the y-axis. Note that from the definitions of the Rips graph (Def. 1), and the σ-Rips graph (Def. 2), threshold values larger than the maximum distance lead to a full graph.

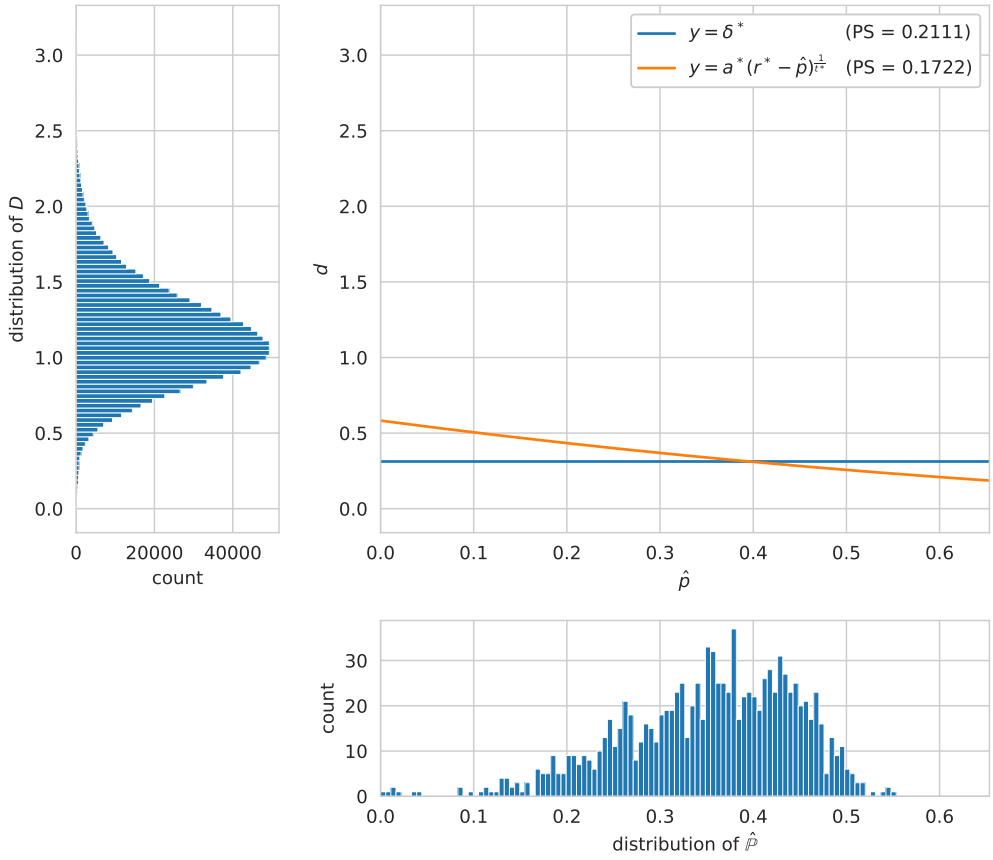
From this figure, it comes out that the optimal threshold rule's values found in the hypothesis class of the σ-Rips graph with our proposed threshold function $\sigma(\cdot;(\delta,r,t,\tau))$ given in Eq. (2) are negatively correlated to the estimation density $\hat{\mathbb{P}}$. We also observe that the best σ-Rips graph achieves better Purity Size than the best Rips graph. These observations are consistent with other datasets reported in the appendix, except for coil-20 and nursery collection, where they have the same performance. These findings provide empirical evidence for our hypothesis that class similarity is a density-aware measure. It also supports our choice of $\sigma(\cdot;(\delta,r,t,\tau))$ given in Eq. (2) as an appropriate threshold function to generalize the Rips graph.

5.2 Cold-start results

For the cold-start experiments, we consider the following unsupervised approaches to compare with our approach:

- **K-Means clustering (KM).** The K-Means algorithm (Lloyd, 1982) partitions a collection of examples into K clusters by minimizing the sum of squared distances to the cluster centers. It has been used for active learning in Zhu et al. (2008), to generate the initial training set by labeling the closest sample to each centroid.

Figure 6: Comparison study between the Rips graph and the σ-Rips graph on the protein dataset: the Purity Size score is reported for each minimizer.



- **K-Means clustering with model examples (KM+ME).** A variant of KM proposed in Kang et al. (2004) 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).** The K-Medoids algorithm (Kaufman and Rousseeuw, 1990) 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.

- **Agglomerative Hierarchical Clustering (AHC).** Agglomerative hierarchical clustering (Voorhees, 1985) is a bottom-up clustering approach that builds a hierarchy of clusters. Initially, each sample represents a singleton cluster. Then, the algorithm recursively merges the closest clusters using a *linkage function* (here the Ward linkage is used) until one cluster is left. This process is usually presented in a dendogram, where each level refers to a merge in the algorithm. AHC has been used for active learning in Dasgupta and Hsu (2008) by pruning the dendogram at a certain level to obtain clusters, then similar to the strategy used with K-Means, selecting the closest samples to the clusters centroids to generate the initial training set.

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 a budget $\mathcal{B} = 10$. $^{\uparrow}/^{\downarrow}$ indicate statistically significantly better/worse performance than Random Selection RS, according to a Wilcoxon rank sum test with $p < 0.05$ (Wolfe, 2012).

Data	RS	KM	KM+ME	Km	AHC	FFT	APC	PTR
protein	28.2	30.6$^{\uparrow}$	31.4$^{\uparrow}$	29.3$^{\uparrow}$	31.6$^{\uparrow}$	21.8	28.8	**40.5**$^{\uparrow}$
	(3.2)	(4.6)	(4.5)	(4.4)	(3.7)	(3.8)	(3.4)	(3.9)
banknote	79.9	85.2$^{\uparrow}$	86.8$^{\uparrow}$	87.6$^{\uparrow}$	85.6$^{\uparrow}$	70.6$^{\downarrow}$	82.4	**88.7**$^{\uparrow}$
	(9.9)	(5.7)	(4.8)	(3.3)	(5.0)	(5.3)	(6.9)	(4.4)
coil-20	29.0	36.7$^{\uparrow}$	38.2$^{\uparrow}$	32.9$^{\uparrow}$	36.0$^{\uparrow}$	18.6$^{\downarrow}$	27.2	**44.2**$^{\uparrow}$
	(5.7)	(4.2)	(2.7)	(5.1)	(3.7)	(3.4)	(4.8)	(2.4)
isolet	13.8	22.3$^{\uparrow}$	**27.6**$^{\uparrow}$	07.1$^{\downarrow}$	23.3$^{\uparrow}$	16.5$^{\uparrow}$	15.4	27.5$^{\uparrow}$
	(2.3)	(1.6)	(1.6)	(1.9)	(1.8)	(1.7)	(3.2)	(2.8)
pendigits	37.4	62.5$^{\uparrow}$	65.6$^{\uparrow}$	53.9$^{\uparrow}$	61.4$^{\uparrow}$	27.2$^{\downarrow}$	38.3	**80.1**$^{\uparrow}$
	(7.2)	(3.5)	(2.3)	(5.2)	(1.9)	(4.9)	(8.2)	(2.6)
nursery	42.7	44.5	**49.3**$^{\uparrow}$	28.4$^{\downarrow}$	44.9	39.1	45.1	46.5
	(7.2)	(5.7)	(4.0)	(1.3)	(7.2)	(3.5)	(6.7)	(6.0)

- **Furthest-First-Traversal (FFT).** The furthest-first traversal of a sample set is a sequence of a selected 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 as in Baram et al. (2004).

- **Affinity Propagation Clustering (APC).** Affinity propagation is a clustering algorithm designed to 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*. The *exemplars* found are then used as an initial training set for active learning (Hu et al., 2010).

Our meta-approach for zero-shot learning is called PTR, where the σ-Rips graph is obtained by Algorithm 1. The results for Random Selection (RS) strategy, competitors, and our method PTR are shown in Table 2 over all collections for a budget $\mathcal{B} = 10$. Note that we do propagation within clusters detected by `ToMATo` for our approach, but not for competitors, for which we consider $\mathcal{B}$ clusters.

Except for the unbalanced dataset nursery, PTR consistently outperforms the random selection method, which has been demonstrated to be very difficult to surpass in different studies. This demonstrates that a preferable starting point for pool-based active learning procedures than random selection is to use the biggest proper topological regions discovered by Algorithm 1 as an initial training set (Line 2 of Algorithm 2). Furthermore, when compared to baseline approaches that are exclusively created to address the cold-start problem in active learning, our meta-approach exhibits very competitive results on different datasets. APC is equivalent to RS, while FFT and Km may have worst performance than RS in some settings. KM, KM+ME and AHL are better than

RS, but we get the best results on four datasets out of the 6. We present further results for budgets $\mathcal{B}$ equal to 3 and 20 in Table 3 in Appendix D.2, with similar conclusions.

5.3 Active learning results

In this section, we present the results of our meta-approach for pool-based active learning strategies (described in Algorithm 3 and denoted PAL_{PTR}).

For the active learning experiments and following results from Siméoni et al. (2019), we only consider the comparison against the Random Selection strategy (RS) as a cold-start strategy, as it outperforms many recent strategies in active learning in small-budget scenarios. We compare our meta-approach and RS strategy for different pool-based active learning strategies, namely the uncertainty sampling query, the entropy sampling query, and the margin sampling query strategies (Danka and Horvath, 2018).

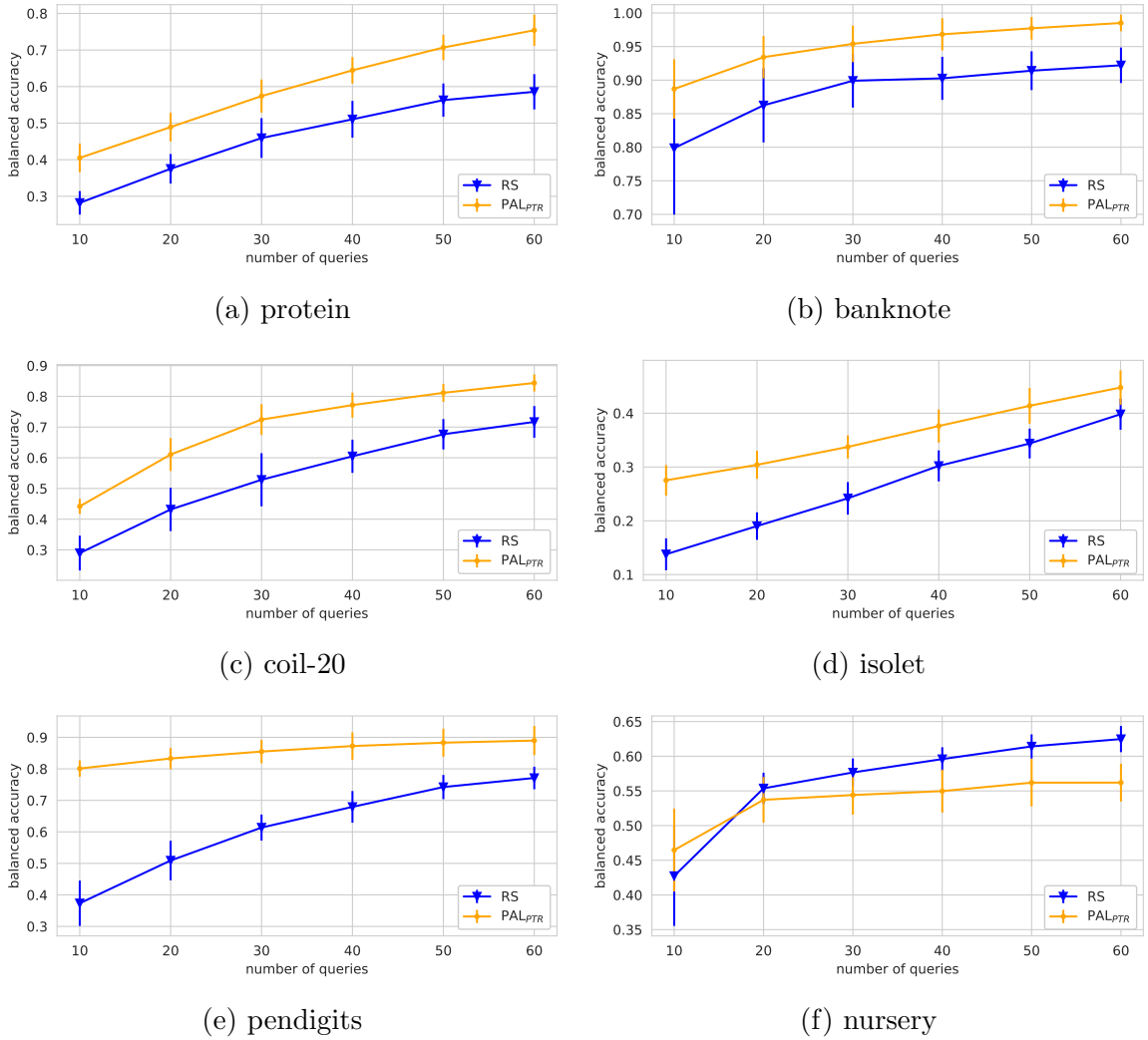
Figure 7 depicts the results corresponding to the uncertainty sampling query strategy with budget $\mathcal{B} = 10$ and a subfigure for each dataset made up of error bar plots that show the average balanced accuracy and standard deviation across the splits for all active learning rounds. Results for other strategies and budgets are shown in Figures 9, 10 and 11 in Appendix D.3, the plots are organized so that each row and column represents a certain budget and active learning technique, but similar conclusions can be drawn.

The results show that, as compared to the use of the random selection technique, all of the pool-based active learning strategies that were taken into consideration gain significantly from our method. Only in the nursery dataset do we not observe a gain. The primary cause is the nursery's significant class imbalance. In Algorithm 3, we decide to put the increase in training sample size ahead of class discovery or class ratio, which may help us understand the current class imbalance. These findings indicate that, when training with highly class-imbalanced datasets, various sample criteria of PTR in Algorithm 3 should be taken into account in addition to only picking the largest ones.

6 Conclusion

We propose a data driven meta-approach for pool-based active learning strategies for multi-class classification problems. Our approach is based on the introduced notion of proper topological regions of a given sample set. We showed the theoretical foundations of this notion and derived a black-box optimization problem to uncover the proper topological regions. Then, we describe how to use those proper topological regions to select the first points to label in a zero-shot learning task, and we derive a meta-approach for pool-based active learning strategies. Our empirical study validates our meta-approach on different benchmarks, in low-budget scenarios, and for various pool-based active learning strategies. Challenging open questions are left: a theoretical analysis that guarantees good performance in active learning, such as generalization bounds, and the use of semi-supervised approaches to conclude the analysis with a model-dependent approach by having a regularization term derived from the PTR.

Figure 7: Average balanced classification accuracy and standard deviation of pool-based active learning with the uncertainty strategy and budget $\mathcal{B} = 10$ on protein dataset, using random forest estimator over 20 stratified random splits.



(a) protein

(b) banknote

(c) coil-20

(d) isolet

(e) pendigits

(f) nursery

References

Abe, N. and Mamitsuka, H. (1998). Query learning strategies using boosting and bagging. In Proceedings of the Fifteenth International Conference on Machine Learning, ICML '98, page 1–9, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ali, A., Caruana, R., and Kapoor, A. (2014). Active learning with model selection. In AAAI.

Amini, M.-R. and Usunier, N. (2015). Learning with Partially Labeled and Interdependent Data. Springer, New York, USA.

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

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

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

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, Advances in Neural Information Processing Systems, volume 24, Red Hook, New York, USA. Curran Associates, Inc.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., and Garnett, R., editors, Advances in Neural Information Processing Systems, volume 32, Red Hook, New York, USA. Curran Associates, Inc.

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

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

Carlsson, G. (2012). The Shape of Data, page 16–44. London Mathematical Society Lecture Note Series. Cambridge University Press, Cambridge, United Kingdom.

Carlsson, G. and Gabrielsson, R. B. (2020). Topological approaches to deep learning. In Baas, N. A., Carlsson, G. E., Quick, G., Szymik, M., and Thaule, M., editors, Topological Data Analysis, pages 119–146, Cham. Springer International Publishing.

Chapelle, O., Schölkopf, B., and Zien, A. (2006). Semi-Supervised Learning. MIT Press, Cambridge, Massachusetts, USA.

Chazal, F., Cohen-Steiner, D., Glisse, M., Guibas, L. J., and Oudot, S. Y. (2009). Proximity of persistence modules and their diagrams. In Proceedings of the Twenty-Fifth Annual Symposium on Computational Geometry, SCG '09, page 237–246, New York, NY, USA. Association for Computing Machinery.

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

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

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

Chen, L., Bai, Y., Huang, S., Lu, Y., Wen, B., Yuille, A. L., and Zhou, Z. (2022). Making your first choice: To address cold start problem in vision active learning. ArXiv, abs/2210.02442.

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

Danka, T. and Horvath, P. (2018). modAL: A modular active learning framework for Python. available on arXiv at https://arxiv.org/abs/1805.00979.

Dasgupta, S. and Hsu, D. (2008). Hierarchical sampling for active learning. In Proceedings of the 25th International Conference on Machine Learning, ICML '08, page 208–215, New York, NY, USA. Association for Computing Machinery.

Edelsbrunner, H. and Harer, J. (2010). Computational Topology - an Introduction. American Mathematical Society, Boston, USA.

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

Garnett, R. (2022). Bayesian Optimization. Cambridge University Press, Cambridge, United Kingdoms.

Guyon, I., Cawley, G. C., Dror, G., and Lemaire, V. (2011). Results of the active learning challenge. In Guyon, I., Cawley, G., Dror, G., Lemaire, V., and Statnikov, A., 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. PMLR.

Hatcher, A. (2000). Algebraic topology. Cambridge Univ. Press, Cambridge.

Hausmann, J.-C. (1995). On the Vietoris-Rips complexes and a cohomology theory for metric spaces, pages 175–188. Prospects in topology : proceedings of a conference in honor of William Browder. Princeton University Press, Princeton, N.J. ID: unige:12821.

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

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

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

Jiang, Y., Chen, D., Chen, X., Li, T., Wei, G.-W., and Pan, F. (2021). Topological representations of crystalline compounds for the machine-learning prediction of materials properties. npj Computational Materials, 7(1):28.

Kang, J., Ryu, K. R., and chul Kwon, H. (2004). Using cluster-based sampling to select initial training set for active learning in text classification. In PAKDD.

Kaufman, L. and Rousseeuw, P. (1990). Finding Groups in Data: An Introduction To Cluster Analysis.

Krempl, G., Ha, T. C., and Spiliopoulou, M. (2015). Clustering-based optimised probabilistic active learning (copal). In Japkowicz, N. and Matwin, S., editors, Discovery Science, pages 101–115, Cham. Springer International Publishing.

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

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

Lewis, D. D. and Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning. In Cohen, W. W. and Hirsh, H., editors, Machine Learning Proceedings 1994, pages 148–156. Morgan Kaufmann, San Francisco (CA).

Li, W., Dasarathy, G., Natesan Ramamurthy, K., and Berisha, V. (2020). Finding the homology of decision boundaries with active learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, Advances in Neural Information Processing Systems, volume 33, pages 8355–8365. Curran Associates, Inc.

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

Lughofer, E. (2012). Single-pass active learning with conflict and ignorance. Evolving Systems, 3(4):251–271.

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

Perez, F., Lebret, R., and Aberer, K. (2018). Cluster-based active learning. CoRR, page abs/1812.11780.

Pourahmadi, K., Nooralinejad, P., and Pirsiavash, H. (2021). A simple baseline for low-budget active learning. arXiv preprint arXiv:2110.12033.

Rieck, B., Yates, T., Bock, C., Borgwardt, K., Wolf, G., Turk-Browne, N., and Krishnaswamy, S. (2020). Uncovering the topology of time-varying fmri data using cubical persistence. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, Advances in Neural Information Processing Systems, volume 33, pages 6900–6912, Red Hook, New York, USA. Curran Associates, Inc.

Romano, J. D., Le, T. T., La Cava, W., Gregg, J. T., Goldberg, D. J., Chakraborty, P., Ray, N. L., Himmelstein, D., Fu, W., and Moore, J. H. (2021). Pmlb v1.0: an open source dataset collection for benchmarking machine learning methods. arXiv preprint arXiv:2012.00058v2.

Roy, N. and McCallum, A. (2001). Toward optimal active learning through sampling estimation of error reduction. In Brodley, C. E. and Danyluk, A. P., editors, Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001, pages 441–448, Massachusetts, USA. Morgan Kaufmann.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Siméoni, O., Budnik, M., Avrithis, Y., and Gravier, G. (2019). Rethinking deep active learning: Using unlabeled data at model training. ICPR.

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

Thoreau, R., Achard, V., Risser, L., Berthelot, B., and Briottet, X. (2022). Active learning on large hyperspectral datasets: A preprocessing method. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B3-2022:435–442.

Urner, R., Wulff, S., and Ben-David, S. (2013). Plal: Cluster-based active learning. In Shalev-Shwartz, S. and Steinwart, I., editors, Proceedings of the 26th Annual Conference on Learning Theory, volume 30 of Proceedings of Machine Learning Research, pages 376–397, Princeton, NJ, USA. PMLR.

Voorhees, E. M. (1985). The Effectiveness & Efficiency of Agglomerative Hierarchic Clustering in Document Retrieval. PhD thesis, Cornell University, USA.

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

Yan, Y., Rosales, R., Fung, G., and Dy, J. G. (2011). Active learning from crowds. In Proceedings of the 28th International Conference on International Conference on Machine Learning, page 1161–1168.

Yang, J., Chen, Z., Chen, W.-S., and Chen, Y. (2011). Robust affine invariant descriptors. Mathematical Problems in Engineering.

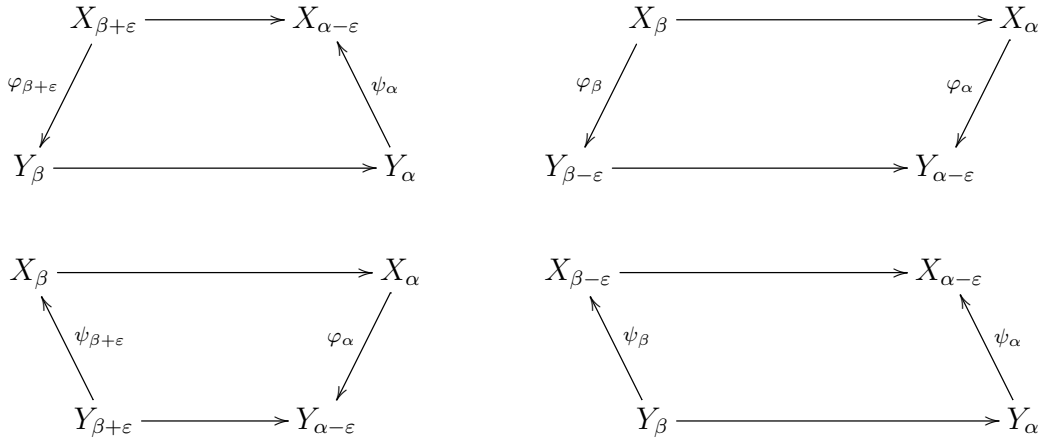
Yu, C. and Hansen, J. H. L. (2017). Active learning based constrained clustering for speaker diarization. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(11):2188–2198.

Zhu, J., Wang, H., Yao, T., and Tsou, B. K. (2008). 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.

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 introduced in Chazal et al. (2009).

Definition 8 (ε-interleaved)**.** *Let $\mathbf{X} = (X_\alpha)_{\alpha\in\mathbb{R}}$ and $\mathbf{Y} = (Y_\alpha)_{\alpha\in\mathbb{R}}$ be two persistence modules and let $D\mathbf{X}$. 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 behind these diagrams is that every component appearing (resp. dying) 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 1. *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. (2009, Theorem 4.4) where the result is proven for every homological dimension.

For example, in Chazal et al. (2011, 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 $D\mathbb{P}$ the persistence diagram of $\mathbb{P}$. Consequently, $D\mathbb{R}_\delta(\mathcal{S}_\mathbf{x}, \mathbb{P})$ encodes 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. Let denote by $\mathbf{R}_\delta = \mathbf{R}_\delta(\mathcal{S}_\mathbf{x}, \mathbb{P})$ and $\mathbf{R}_{\sigma(\cdot)} = \mathbf{R}_{\sigma(\cdot)}(\mathcal{S}_\mathbf{x}, \mathbb{P})$. $R_\delta = R_\delta(\mathcal{S}_\mathbf{x})$, and $R_{\sigma(\cdot)} = R_{\sigma(\cdot)}(\mathcal{S}_\mathbf{x})$ and, for $\alpha \in \mathbb{R}$, we set

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

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 each vertices $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{C}_q$, 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:

$$\psi_\alpha \colon H_0(R_{\sigma(\cdot),\alpha}) \to H_0(R_{\delta,\alpha-\varepsilon}).$$

By construction, the following diagrams are commutative (the linear maps involved are induced by inclusions on connected components).

$$\begin{array}{ccc} H_0(R_{\delta,\beta+\varepsilon}) & \longrightarrow & H_0(R_{\delta,\alpha-\varepsilon}) \\ {\scriptstyle\varphi_{\beta+\varepsilon}}\searrow & & \nearrow{\scriptstyle\psi_\alpha} \\ H_0(R_{\sigma(\cdot),\beta}) & \longrightarrow & H_0(R_{\sigma(\cdot),\alpha}) \end{array} \qquad \begin{array}{ccc} H_0(R_{\delta,\beta}) & \longrightarrow & H_0(R_{\delta,\alpha}) \\ {\scriptstyle\varphi_\beta}\searrow & & \searrow{\scriptstyle\varphi_\alpha} \\ H_0(R_{\sigma(\cdot),\beta-\varepsilon}) & \longrightarrow & H_0(R_{\sigma(\cdot),\alpha-\varepsilon}) \end{array}$$

$$\begin{array}{ccc} H_0(R_{\delta,\beta}) & \longrightarrow & H_0(R_{\delta,\alpha}) \\ {\scriptstyle\psi_{\beta+\varepsilon}}\nwarrow & & \searrow{\scriptstyle\varphi_\alpha} \\ H_0(R_{\sigma(\cdot),\beta+\varepsilon}) & \longrightarrow & H_0(R_{\sigma(\cdot),\alpha-\varepsilon}) \end{array} \qquad \begin{array}{ccc} H_0(R_{\delta,\beta-\varepsilon}) & \longrightarrow & H_0(R_{\delta,\alpha-\varepsilon}) \\ {\scriptstyle\psi_\beta}\nwarrow & & \nearrow{\scriptstyle\psi_\alpha} \\ H_0(R_{\sigma(\cdot),\beta}) & \longrightarrow & H_0(R_{\sigma(\cdot),\alpha}) \end{array}$$

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

B More details on the theoretical guarantees of ToMATo

Let us assume that $\mathcal{X}$ is a Riemannian m-manifold and the density function $\mathbb{P} \colon \mathcal{X} \to \mathbb{R}$ is a κ-Lipschitz probability density function with respect to the m-dimensional Hausdorff measure and $\kappa > 0$. In order to draw the theoretical guarantees of ToMATo we need some assumption on the persistence diagram of $\mathbb{P}$, more precisely, on the spatial distribution of the points in this diagram.

Definition 9. *Let $d_1, d_2 \in \mathcal{R}$ be two non-negative real numbers such that $d_1 < d_2$. The persistent diagram $D\mathbb{P}$ is called (d_1, d_2)-separated if every point of $D\mathbb{P}$ lies either in the region D_1 above the diagonal line $y = x - d_1$ or in the region D_2 below the diagonal line $y = x - d_2$ and to the right of the vertical line $x = d_2$.*

The points in the region D_2 will be considered as the prominence peaks and the points in the region D_1 as "topological noise".

We highlight the theoretical guarantees of ToMATo. We refer the reader interested in more details and the proof of the statements to Chazal et al. (2013).

The first guarantee ensures that with reasonable assumption on the point cloud and the density $\mathbb{P}$ ToMATo can recover the numbers of clusters induced by $\mathbb{P}$.

Result 1 ((Chazal et al., 2013, Theorem 9.2)). *Assume that $\mathcal{S}_\mathbf{x}$ is i.i.d with respect to $\mathbb{P}$. If $D\mathbb{P}$ is (d_1, d_2)-separated and if the parameter δ is smaller than a fraction of*

$d_2 - d_1$ and of the convexity radius of $\mathcal{X}$, then there is a range $(d_1 + 2\kappa\delta, d_2 - 3\kappa\delta)$ of values of τ such that the number of topological regions output by $\texttt{ToMATo}_\tau(R_\delta(S), \mathbb{P})$ is equal to the number of peaks, of $\mathbb{P}$ with prominence at least τ with probability at least $1 - e^{\Omega(n)}$ where n is the number of data points.

$\Omega(n)$ hides a factor increasing monotonically with c and δ and depending on certain geometric quantities of the manifold $\mathcal{X}$.

The following result tells us that, under the same hypotheses, we can recover the basins of attractions of the prominent peaks of $\mathbb{P}$.

Result 2 ((Chazal et al., 2013, Theorem 10.1)). *Under the same hypotheses as in Result 1 and with the same probability, we have that, for every point $p \in D_2$, $\texttt{ToMATo}_\tau(R_\delta(S), \mathbb{P})$ outputs a topological region R such that $R \cap \mathbb{P}^{-1}([\alpha, +\infty]) = B_\tau(m_p) \cap \mathcal{S} \cap \mathbb{P}^{-1}([\alpha, +\infty])$ for all $\alpha \in (\alpha_\tau(p) + d_1 + \frac{5}{2}\kappa\delta, p_x]$, where m_p is the peaks of $\mathbb{P}$ corresponding to p, $B_\tau(m_p)$ is the basin of attraction of m_p in the underlying manifold $\mathcal{X}$, and $\alpha_\tau(m_p)$ is the first value of α at which $B_\tau(m_p)$ gets connected to the basin of attraction of other peaks of $\mathbb{P}$ of prominence at least τ in the superlevel-set $\mathbb{P}^{-1}([\alpha, +\infty])$.*

In other words, R is the trace of the basin of attraction $B_\tau(m_p)$ over the point of $\mathcal{S}$, until the value $\alpha_\tau(m_p)$ at which the basin of attraction meets another τ-prominent peak.

Finally, it is worth mentioning that in (Chazal et al., 2013, Section 11), they also study the robustness of the approach when considering an estimation of the density function.

Theorem 1 tells us that one way to ensure the bottleneck constraint in (4) is to apply the same post-processing phase used in the `ToMATo` algorithm on the σ-Rips graph. It consists of applying a merging rule along the hill-climbing method on the graph with $\mathbb{P}$. This merging rule compares the *topological persistence* of connected components to an additional merging parameter $\tau \in [0, \max_{\mathbf{x} \in \mathcal{S}_\mathbf{x}} \mathbb{P}(\mathbf{x})]$ (Chazal et al., 2013).

C 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 (Lloyd, 1982), 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 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.

D More empirical results

D.1 Rips graph vs σ-Rips graph

In this section we show an empirical comparison in Figure 8 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.

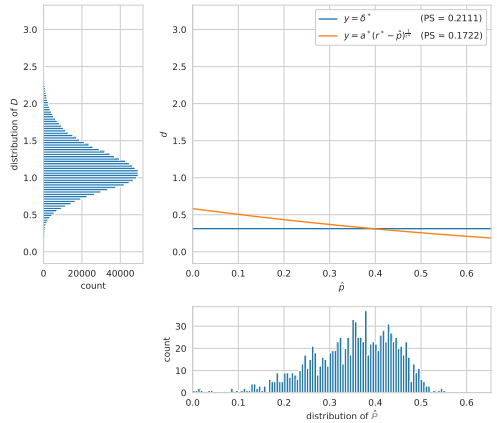
D.2 Cold-start results

This section presents the numerical results obtained in our empirical investigation of the cold-start problem in Table 3 for several budgets. It shows that the proposed PTR approach achieves competitive performance compared to the baseline methods, we also notice that most of the methods except for the k-means-based methods and PTR suffer from degraded performance compared to the random selection strategy.

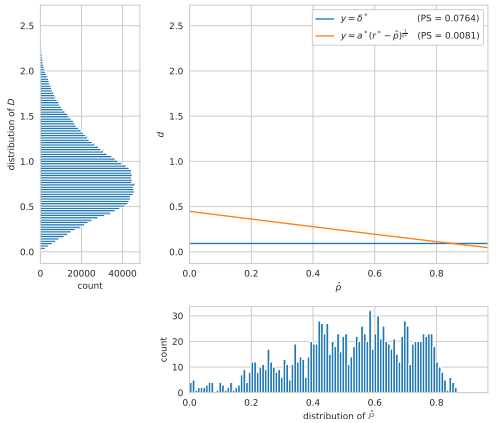
D.3 Active learning results

This last section illustrates the empirical results of pool-based active learning for the rest of the considered datasets using our proposed approach, compared against the random sampling strategy. Figures 9, 10 and 11(a) show a clear and significant gain of the PTR methodology for pool-based active learning strategies against random selection. The

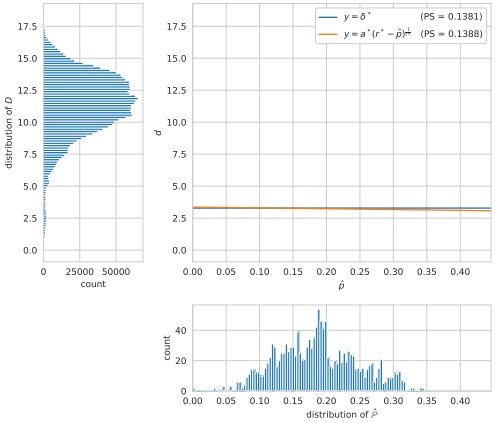
Figure 8: Comparison study between the Rips graph and the σ-Rips graph over all datasets, the Purity Size score is reported for each minimizer.



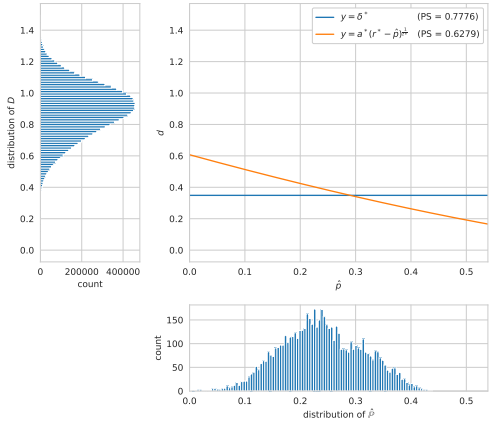
(a) protein



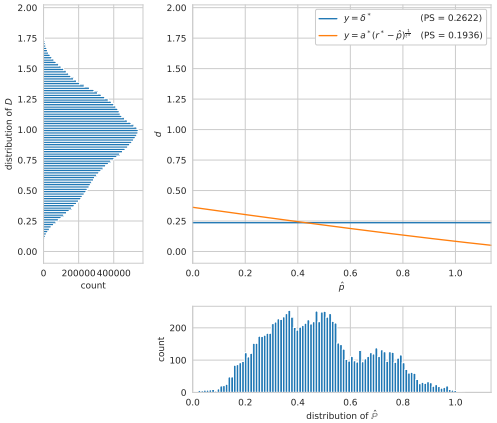
(b) banknote



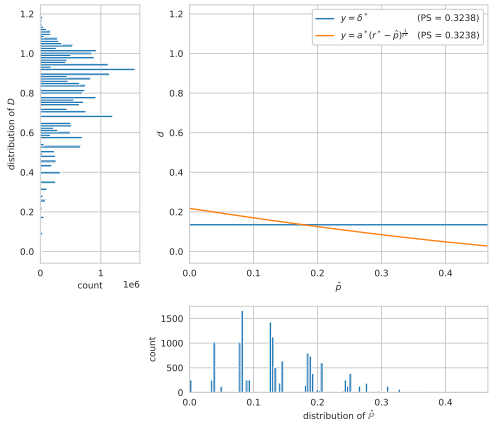
(c) coil-20



(d) isolet



(e) pendigits



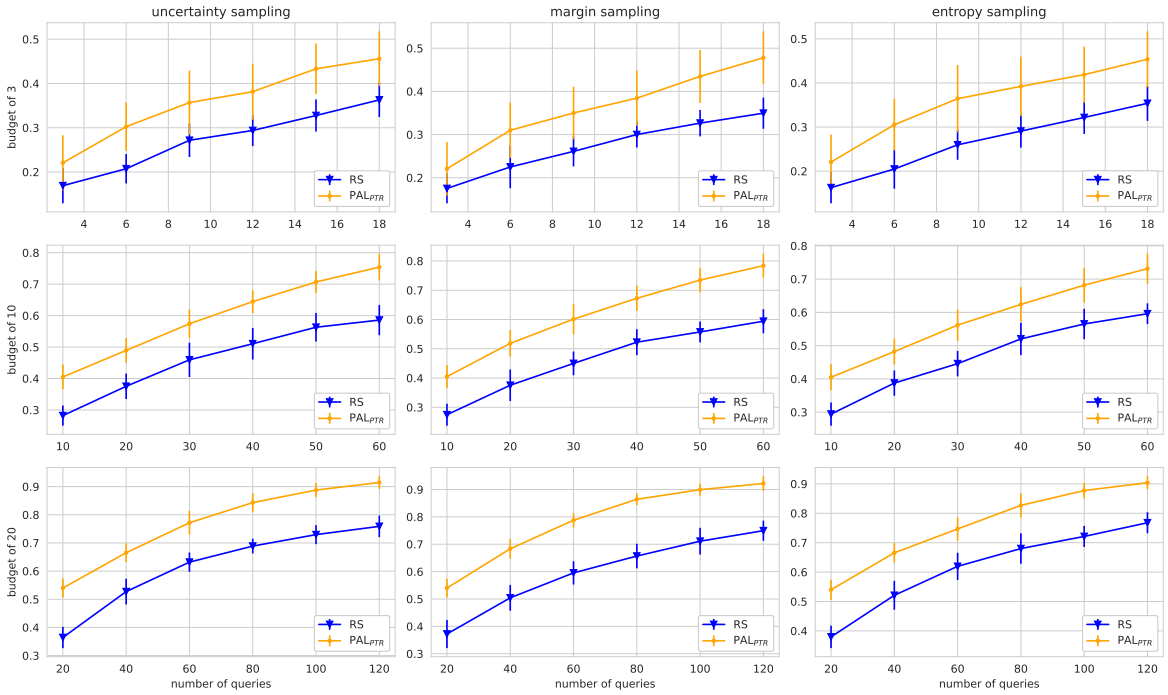
(f) nursery

results in Figure 11(b) indicate that the PTR approach is not robust against the high label imbalance datasets. We observed that the propagation step over the PTR tends to amplify the class imbalance in the resulting training set.

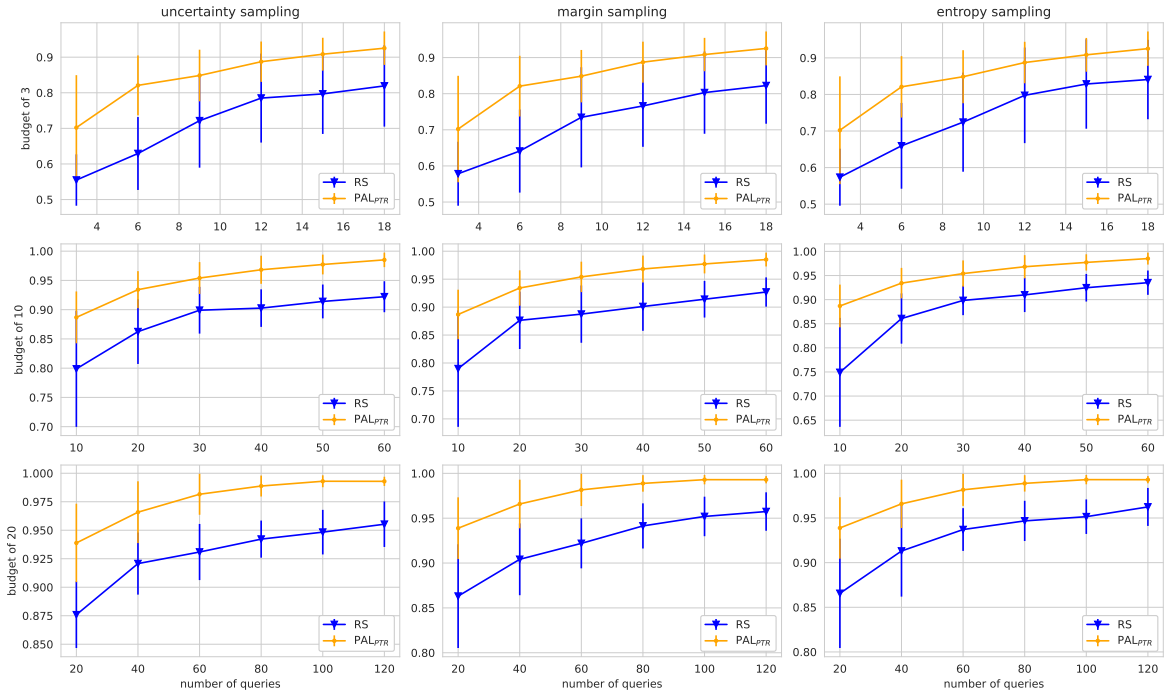
Table 3: Average balanced classification accuracy (in %) and standard deviation of random forest classifier with the initial training set obtained from different methods over 20 stratified random splits for different budgets $\mathcal{B}$. $^{\uparrow}/^{\downarrow}$ indicate statistically significantly better/worse performance than Random Selection RS, according to a Wilcoxon rank sum test ($p < 0.05$) (Wolfe, 2012).

Dataset	$\mathcal{B}$	RS	KM	KM+ME	Km	AHC	FFT	APC	PTR
protein	3	16.9	21.2$^{\uparrow}$	**23.9**$^{\uparrow}$	21.2$^{\uparrow}$	22.7$^{\uparrow}$	17.4	16.7	22.1$^{\uparrow}$
		(4.0)	(1.8)	(2.2)	(4.4)	(2.5)	(3.3)	(3.3)	(6.2)
	20	36.4	42.1$^{\uparrow}$	45.5$^{\uparrow}$	39.2	43.4$^{\uparrow}$	26.1$^{\downarrow}$	39.2	**54.0**$^{\uparrow}$
		(3.8)	(3.9)	(2.5)	(4.4)	(3.4)	(3.4)	(3.7)	(3.4)
banknote	3	55.5	74.0$^{\uparrow}$	**84.3**$^{\uparrow}$	62.5$^{\uparrow}$	63.7$^{\uparrow}$	58.2$^{\uparrow}$	58.7	70.2$^{\uparrow}$
		(7.2)	(4.6)	(5.6)	(3.3)	(4.5)	(7.3)	(8.0)	(14.7)
	20	87.6	90.7$^{\uparrow}$	92.4$^{\uparrow}$	92.3$^{\uparrow}$	92.6$^{\uparrow}$	71.9$^{\downarrow}$	90.9$^{\uparrow}$	**93.9**$^{\uparrow}$
		(2.9)	(2.4)	(2.0)	(2.4)	(2.9)	(7.2)	(3.2)	(3.4)
coil-20	3	12.6	**15.0**$^{\uparrow}$	**15.0**$^{\uparrow}$	**15.0**$^{\uparrow}$	**15.0**$^{\uparrow}$	10.8$^{\downarrow}$	11.7	13.6
		(2.6)	(0.0)	(0.0)	(0.0)	(0.0)	(2.0)	(2.3)	(1.7)
	20	42.0	56.7$^{\uparrow}$	63.0$^{\uparrow}$	42.3	58.1$^{\uparrow}$	25.6$^{\downarrow}$	41.4	**71.1**$^{\uparrow}$
		(5.8)	(3.7)	(2.8)	(3.5)	(4.1)	(2.5)	(4.7)	(3.8)
isolet	3	07.6	08.7$^{\uparrow}$	09.7$^{\uparrow}$	07.8	09.1$^{\uparrow}$	09.2$^{\uparrow}$	07.5	**10.8**$^{\uparrow}$
		(1.5)	(0.9)	(0.6)	(1.6)	(1.9)	(1.0)	(1.8)	(1.1)
	20	19.2	27.9$^{\uparrow}$	**40.4**$^{\uparrow}$	10.7$^{\downarrow}$	28.2$^{\uparrow}$	18.8	21.1$^{\uparrow}$	38.6$^{\uparrow}$
		(2.7)	(2.5)	(3.2)	(2.0)	(2.1)	(2.4)	(3.1)	(3.2)
pendigits	3	21.5	21.3	22.5	26.6$^{\uparrow}$	19.4$^{\downarrow}$	17.3$^{\downarrow}$	17.8$^{\downarrow}$	**29.9**$^{\uparrow}$
		(3.5)	(1.9)	(2.1)	(2.6)	(1.8)	(3.7)	(4.9)	(0.0)
	20	54.3	72.3$^{\uparrow}$	75.8$^{\uparrow}$	64.0$^{\uparrow}$	72.3$^{\uparrow}$	34.8$^{\downarrow}$	52.2	**87.7**$^{\uparrow}$
		(5.9)	(2.7)	(2.3)	(3.6)	(2.5)	(4.5)	(5.9)	(4.1)
nursery	3	30.7	29.2	30.2	25.0$^{\downarrow}$	28.3$^{\downarrow}$	30.0	30.0	**35.1**$^{\uparrow}$
		(4.0)	(5.2)	(6.5)	(0.2)	(3.9)	(3.2)	(3.7)	(5.6)
	20	**55.3**	52.8$^{\downarrow}$	54.4	32.9$^{\downarrow}$	53.8	39.8$^{\downarrow}$	52.5$^{\downarrow}$	54.1
		(2.8)	(3.3)	(3.0)	(1.1)	(2.7)	(1.1)	(4.9)	(4.5)

Figure 9: Average balanced classification accuracy and standard deviation of different pool-based active learning strategies and budgets on protein and banknote datasets, using random forest estimator over 20 stratified random splits.

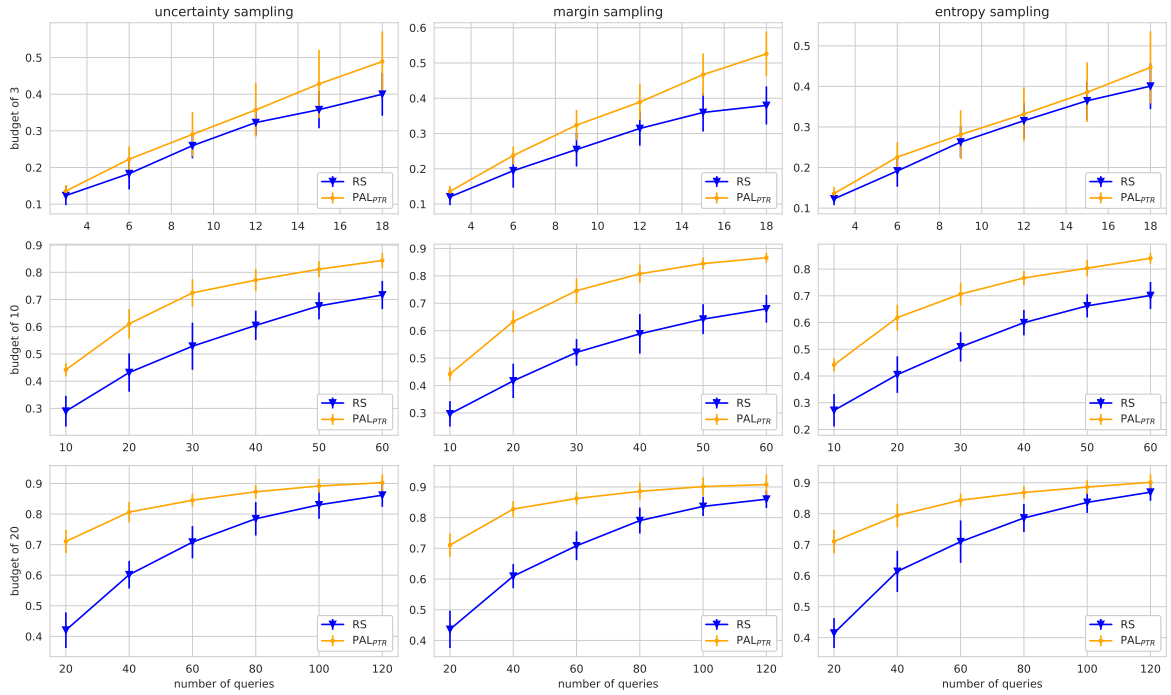


(a) protein

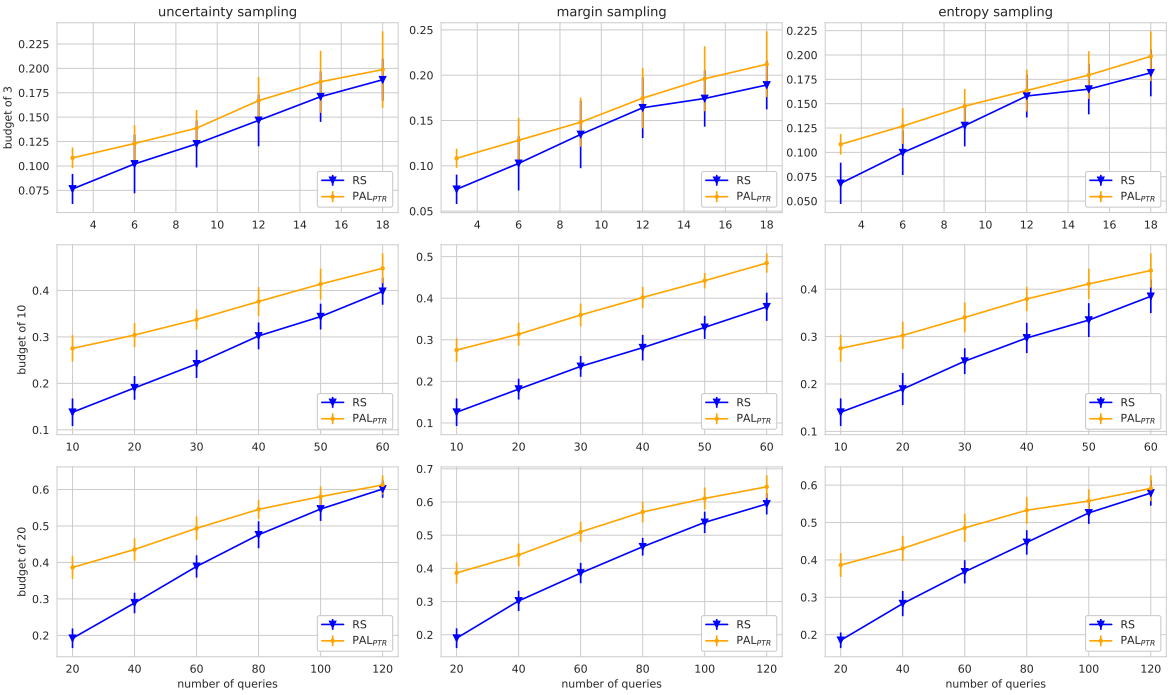


(b) banknote

Figure 10: Average balanced classification accuracy and standard deviation of different pool-based active learning strategies and budgets on coil-20 and isolet datasets, using random forest estimator over 20 stratified random splits.

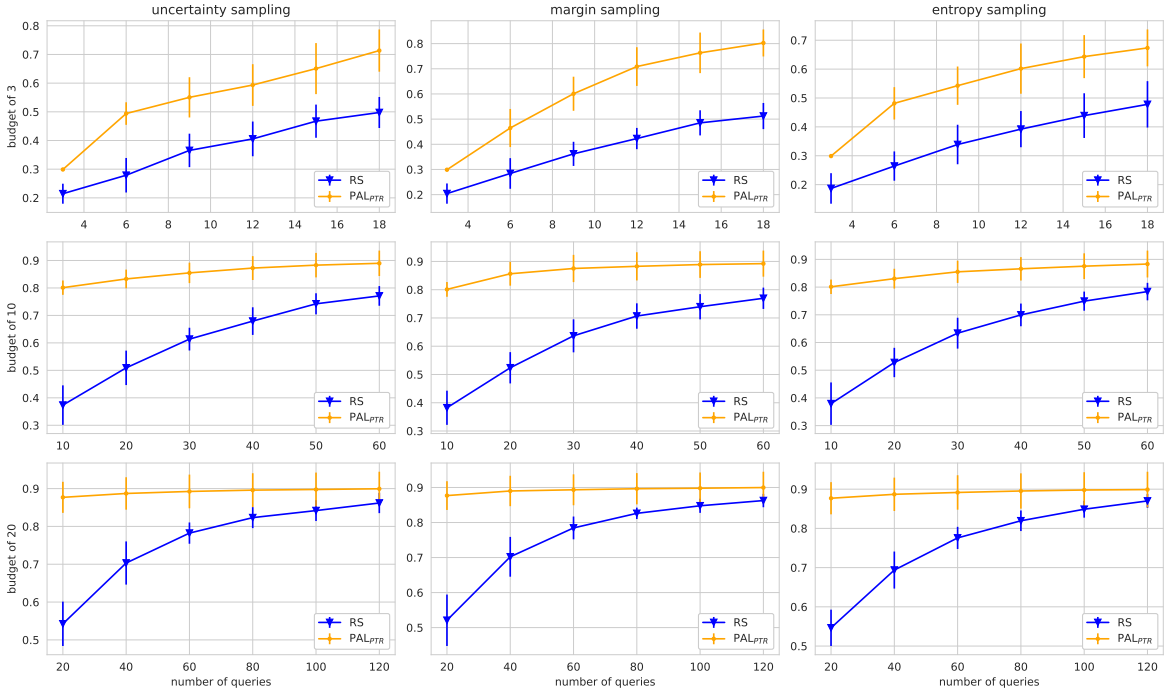


(a) coil20

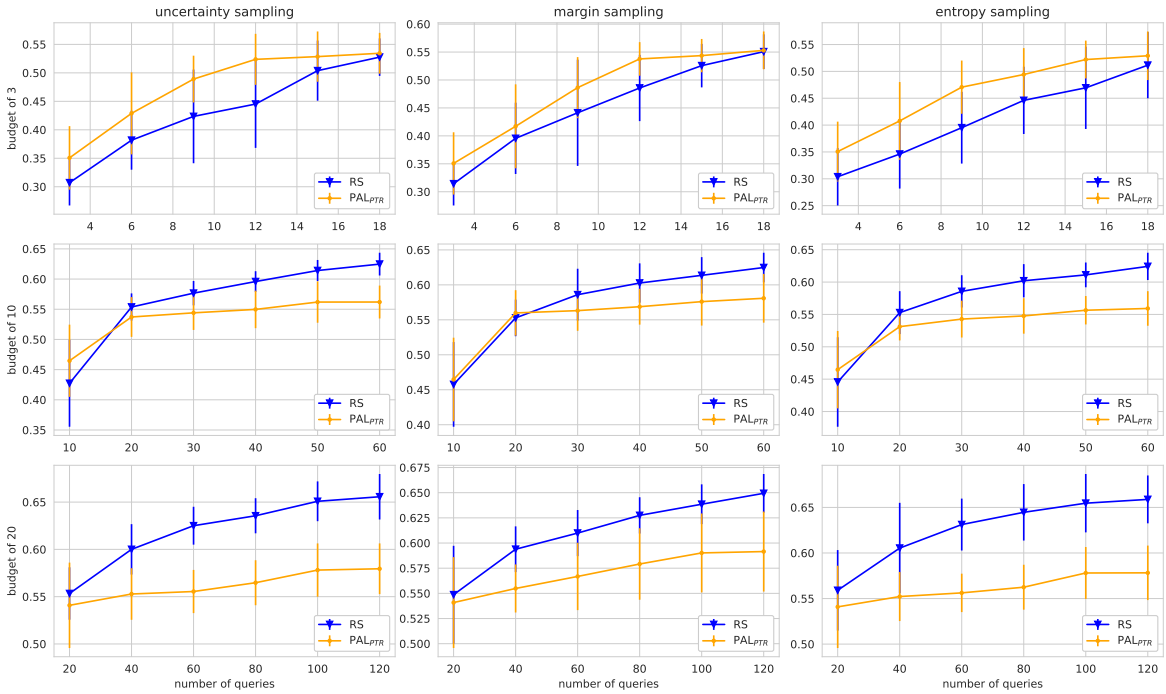


(b) isolet

Figure 11: Average balanced classification accuracy and standard deviation of different pool-based active learning strategies and budgets on pendigits and nursery datasets, using random forest estimator over 20 stratified random splits.



(a) pendigits



(b) nursery