Zero-shot Active Learning with Topological Clustering for Multiclass Classification

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present a novel approach for zero-shot active learning for multi-class classification based on a clustering technique, called ToMATo, which is guided by topological persistence. Our objective is to identify effective regions in the feature space for label querying. The labeling of examples in these regions will allow the training of efficient multi-class classification prediction functions. We have adapted ToMATo with a density aware δ-Rips graph in order to obtain homogeneous simplicial trees. From these trees, informative simplices are identified with respect to the annotation effort, or the budget. Representative examples from each of them are labeled by an oracle and these labels are then propagated through the trees. We adapt ToMATo by computing our *persistence diagram* (PD) from a δ-Rips graph that is estimated using a k-nearest neighbor distance matrix This allows the application of the method to large scale scenarios. From this perspective we also propose a local density estimator from the same distance matrix. Comparisons on different benchmarks show that the proposed approach greatly improves performance with respect to a random querying strategy for label assignment that has been found outperforming state-of-the art approaches in previous works.

1 Introduction

In many real-life applications, the labeling of training observations for learning is costly and sometimes not even realistic. For example, in web oriented applications, huge amount of observations are collected sequentially. However, there is not enough time to label these data for different purpose while unlabeled data are abundant. Different attempts have been made to reduce the annotation burden. For example, we can refer to the so many successful *semi-supervised* and *active learning* approaches that have been proposed until now [Baram et al., 2004, Settles, 2012, Chapelle et al., 2006, Amini and Usunier, 2015].

All of these approaches suppose that there exists a small set of labeled training data together with a large set of unlabeled examples. They also tend to identify additional informative unlabeled observations to be (pseudo-)labeled for learning. Besides, some of the proposed strategies are based on the approximation of the risk of selection, *e.g.* with iterative methods [Zhao et al., 2006, Zhu et al., 2008], or the selection of the most uncertain observations regarding some confidence measures, *e.g.* with model-driven models [Lakshminarayanan et al., 2017, Yan et al., 2011].

In this work, we suppose that there is no initial labeled training data and propose a new zero-shot active learning strategy based on topological persistence in order to find informative observations to be labeled for learning. More precisely, our approach is based on topological data analysis which has recently brought exciting new ideas to the machine learning community, especially in *unsupervised learning* with *topological clustering* [Bonis and Oudot, 2018, Cabanes et al., 2013]. Among these

Submitted to the Topological Data Analysis and Beyond Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

studies, ToMATo [Chazal et al., 2013] is a mode-seeking clustering algorithm with a cluster merging phase guided by *topological persistence*. It relies on the concept of *prominence* by computing the PD which reflects the modes of the density, and a prominence threshold is estimated from the PD to merge clusters and discard noise. This method is adapted and used in this work in order to detect informative observations to be labeled.

The contributions of this work are the following:

- we propose a generic data driven approach for zero-shot learning based on the ToMATo clustering technique;
- we investigate different ways to scale up computations in use in this approach and we derive a density aware formulation of δ-Rips graph.

We validate our approach by comparing it to a random querying strategy for label assignment on different benchmarks.

2 TCAL: Topological Clustering for Active Learning

Let $(\mathcal{X}, d)$ be a metric space, where $d\colon \mathcal{X} \times \mathcal{X} \to [0, \infty)$ is the distance metric. We assume that we have a set $X := \{x_i\}_{i=1}^n \subset \mathcal{X}$ of i.i.d unlabeled examples drawn from an arbitrary unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} := [c]$ is a set of $c > 1$ unknown classes for which each example x_i has a unique, yet unknown, label y_i. We also have at our disposal a perfect oracle (labeling from the expert) $\mathcal{O}\colon \mathcal{X} \to \mathcal{Y}$ over X where $\mathbb{P}_{x_i \sim X}(\mathcal{O}(x_i) = y_i) = 1$ and a budget $\mathcal{B}$ which corresponds to the maximum number of examples the expert can label. We describe the main steps of the proposed method denoted by TCAL and which is summarized in Algorithm 1.

Algorithm 1: TCAL

Input: $X := \{x_i\}_{i=1}^n$, oracle $\mathcal{O}$, budget $\mathcal{B}$ and a distance metric $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$.

- Compute distance matrix M and density estimator $(\widetilde{f}(i))_{1 \le i \le n}$ with (1).
- Build graph G with the threshold distance δ.
- Run ToMATo$(G, \widetilde{f}(i), \tau)$ to obtain the set of clusters $(\mathcal{C}_k)_{1 \le k \le \mathcal{K}}$.
- Ask an oracle to label the $\mathcal{B}$ peaks from $(\mathcal{C}_k)_{1 \le k \le \mathcal{K}}$.
- Propagate the labels and subsample majority classes as in (2).

Output: Set of labeled training examples S

For each $x_i \in X$ we compute its distance to its l-nearest neighbors[1] and we obtain a sparse distance matrix $M = (m_{i,j}) \in \mathbb{R}^{n \times n}$ (with only l non zero values in each row) where

$$m_{i,j} = \begin{cases} d(x_i, x_j) & \text{if } x_j \text{ is one of the } k\text{-nearest neighbors of } x_i, \\ 0 & \text{elsewhere.} \end{cases}$$

The use of l-nearest neighbors allows us to consider large data sets, and we observe in practice no loss in accuracy.

To estimate the density from M, we propose to use $\widetilde{f}$, where, for all $1 \le i \le n$,

$$\widetilde{f}(i) = \left(\frac{1}{l} \sum_{j=1}^{n} m_{i,j}^2 \right)^{-1/2}. \tag{1}$$

A δ-Rips graph $G = (V, E)$ is constructed from $\widetilde{f}$ where $V = [n]$ and

$$E = \{(i, j) \mid i \in V, j \in [k] \text{ and } 0 < m_{i,j} \le \delta(i)\}.$$

Here, for all $1 \le i \le n$, the threshold distance $\delta(i)$ is defined with respect to the density estimation at x_i by $\delta(i) = \delta_0 \left(\alpha - \widetilde{f}(i)\right)^{1/\beta}$, where δ_0 corresponds to the initial threshold, α allows to take into account the shift in density distribution, and β can be interpreted as the evolution of similarity in different density levels.

[1] In practice, we use $l = \lfloor n/10 \rfloor$.

A set of $\mathcal{K}$ clusters $(\mathcal{C}_k)_{1<k\leq\mathcal{K}}$ is computed by running ToMATo on the graph G with density $\widetilde{f}$. The prominence threshold τ to merge clusters has to be set and its selection is discussed in Section 3. Those clusters are sorted by decreasing size, to focus first on the largest cluster. Then, the oracle is providing labels associated to the $\mathcal{B}$ peaks from $(\mathcal{C}_k)_{1\leq k\leq\mathcal{K}}$ starting from the largest, and these labels are propagated in their respective clusters, the result being an imbalanced set of labeled data:

$$S = \bigcup_{k=1}^{\min(\mathcal{K},\mathcal{B})} S_k, \qquad \text{where } S_k = \left\{ \left(x_l, \mathcal{O}\left(\arg\max_{i\in\mathcal{C}_k} \widetilde{f}(i) \right) \right) \right\}_{l\in\mathcal{C}_k}. \tag{2}$$

If the hyper-parameters are carefully tuned, the set S have labels from most, if not all, of the c classes, but with an unequal distribution. At this last step, the large classes are randomly subsampled to obtain a balanced training set.

3 Parameter selection

Graph parameters Finding the right graph representation that explains class similarity across different data collections is a universal problem for graph based methods. The threshold distance function δ for the graph G has three parameters δ_0, β, α, describing that the similarity between data points is not uniform overall density levels, especially for multi-class data sets. A radius that gives good representatives of dense classes will capture less information in low density classes. Inversely, larger radius will capture representatives in low density levels but also diffuse noise in high density regions. In practice, we consider hyperparameters $(\delta_0, \beta, \alpha)$ such that the threshold function has a decreasing behavior in the region $[\min \widetilde{f}, \max \widetilde{f}] \times [\min M_{avg}, \max M_{avg}]$ where $M_{avg} = \left\{ \sum_{j=1}^{n} m_{i,j}/l \mid 1 \leq i \leq n \right\}$. We notice that on image data sets (Coil20, MNIST and Statlog), the same hyperparameters are providing good performance.

Prominence threshold The parameter τ in ToMATo is used to filter out topological noise and to distinguish between relevant peaks from subsidiary peaks, coming from parents in $\widetilde{f}$. It makes a trade-off between purity and cluster size. More precisely, increasing τ will merge more clusters (thus reduce the number $\mathcal{K}$ of clusters) and diffuse more noise, so that $(\mathcal{C}_k)_{1<k\leq\mathcal{K}}$ contains larger clusters with less purity with respect to their true labels. We follow a similar but more conservative procedure to Chazal et al. [2013]: we minimize the diffusion of the noise and keep reasonable cluster sizes. To do so, we sort the prominent peaks given by $\widetilde{f}$ by decreasing order in V, and we fix τ to be the more stable value after a significant gap in the distribution. In practice, we compute the variance on a sliding window of size $\lfloor |V|/10 \rfloor$, we select the window that has the lowest variance after the window of maximum variance, and we fix τ to its median value.

4 Experiments

Data sets We conduct experiments on benchmark data sets for classification problems also often used in active learning: MNIST [LeCun et al., 1998], COIL-20 [Yang et al., 2011], Isolet [Fanty and Cole, 1991] and sensorless drive diagnosis SDD [Paschke et al., 2013] as well as two imbalanced data sets including Protein [Higuera et al., 2015] and Statlog [King et al., 2000]. Table 1 presents statistics of the data sets on the four first columns.

Baseline Following results from Siméoni et al. [2019], we only use random labeling strategy, as it outperforms many recent strategies in active learning with small budget scenarios.

A simple linear support vector machine with stochastic gradient descent is used for the classifier for both methods, with default parameters and a single epoch, since our objective is to show the gap in learning performance of any hypothesis class with respect to the training set in an online setting. For the metric distance for TCAL, we use the euclidean distance overall data sets.

We run the two procedures for several budgets $\mathcal{B}$ (less than 0.5% of the sample size) and compare the evolution. As the expert labels only one observation per cluster, in our experiments the budget is upper bounded by the number of clusters detected by ToMATo. 20 random splits are considered, with 70% of the data in the training set and 30% in the test set.

Table 1: Average classification accuracy (in %) and standard deviation over 20 random splits for different budgets $\mathcal{B}$. The third column corresponds to the dimension of the feature space $\mathbb{R}^p$.

Dataset	n	p	c	Budget $\mathcal{B}$	Random	TCAL
Protein	1080	77	8	5	18.70 ± 5.54	$\mathbf{32.29 \pm 4.47}$
				10	20.35 ± 8.38	$\mathbf{38.56 \pm 4.60}$
COIL-20	1440	1024	20	10	23.00 ± 3.24	$\mathbf{41.57 \pm 3.48}$
				50	57.28 ± 5.90	$\mathbf{82.74 \pm 3.43}$
				100	75.44 ± 4.55	$\mathbf{95.08 \pm 1.59}$
Isolet	6238	617	26	10	13.22 ± 2.53	$\mathbf{33.29 \pm 1.88}$
				50	26.28 ± 2.81	$\mathbf{55.09 \pm 2.81}$
				100	43.69 ± 4.18	$\mathbf{63.13 \pm 4.05}$
Statlog	6435	36	6	10	31.32 ± 12.83	$\mathbf{65.60 \pm 2.64}$
				20	32.16 ± 14.61	$\mathbf{66.28 \pm 1.97}$
SDD	58.5k	48	11	100	33.84 ± 6.55	$\mathbf{46.92 \pm 3.01}$
				250	38.57 ± 6.50	$\mathbf{47.54 \pm 4.05}$
				500	43.13 ± 5.44	$\mathbf{49.37 \pm 3.15}$
MNIST	70k	784	10	100	68.22 ± 2.96	$\mathbf{81.00 \pm 1.01}$
				700	82.91 ± 1.05	$\mathbf{87.29 \pm 0.50}$
				1000	83.45 ± 1.24	$\mathbf{88.12 \pm 0.54}$
				1400	84.03 ± 0.71	$\mathbf{88.75 \pm 0.46}$

Results Table 1 presents the results of our approach and of the random labeling strategy over all data sets. In all cases, TCAL provides significantly better results than the random strategy. For imbalanced data sets (Protein and Statlog), the random strategy is affected by the imbalance in class distribution and performs badly, whereas our method has benefit from the clustering step and performs the best. We remark that even with very few labels, if the labels are given for observations particularly discriminant, the performance are very high, *e.g.* for COIL-20 with 100 labels over the 20 classes, where the accuracy of TCAL is 95%, while the performance for the random strategy is 75%.

Highlight: the effect of the prominence threshold We investigate the effect of the prominence threshold τ on the classifier performance. Figure 1 shows the average accuracy curve of 20 random splits on MNIST data set for 100 points of τ uniformly in $[0, 10^{-2}]$ with a budget of 100, where $\hat{\tau}$ is estimated following the procedure in Section 3. First, we remark that whatever the value for τ (small enough), we get good performance compared with the random labeling strategy. In addition, the strategy of largest gap for τ estimation in ToMATo fails in this case, where it gives a value around 0.02. Finally, note that our estimation is very close to the best one.

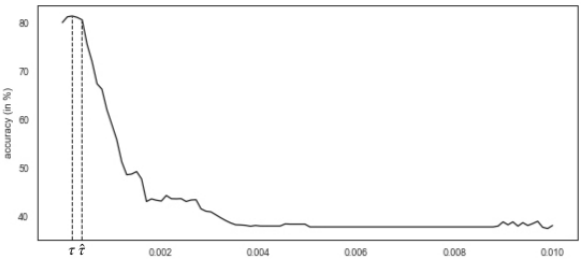

Figure 1: Average classification accuracy (in %) of 20 random splits on MNIST data set with different values for the prominence threshold. Optimal value τ is shown with dashed line whereas $\hat{\tau}$ is computed as explained in Section 3.

5 Conclusion

We propose a data driven method for zero-shot active learning in multiclass classification problems with topological clustering. Our empirical study validates this method on different benchmark data sets. This work is, to our knowledge, the first significant step to use topological data analysis to detect relevant observations for zero-shot active learning. Challenging open questions are left, as the use of semi-supervised model to conclude the analysis (instead of a supervised classifier) and theoretical results that guarantee good performance in active learning.

References

Yoram Baram, Ran El-Yaniv, and Kobi Luz. Online choice of active learning algorithms. *J. Mach. Learn. Res.*, 5:255–291, 2004.

Burr Settles. *Active Learning*. Morgan & Claypool Publishers, 2012.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2006.

Massih-Reza Amini and Nicolas Usunier. *Learning with Partially Labeled and Interdependent Data*. Springer, 2015.

Yue Zhao, Ciwen Xu, and Yongcun Cao. Research on query-by-committee method of active learning and application. In Xue Li, Osmar R. Zaïane, and Zhanhuai Li, editors, *Advanced Data Mining and Applications*, pages 985–991, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin 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.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30*, pages 6402–6413, 2017.

Yan Yan, Romer Rosales, Glenn Fung, and Jennifer G. Dy. Active learning from crowds. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, page 1161–1168, 2011.

Thomas Bonis and Steve Oudot. A fuzzy clustering algorithm for the mode-seeking framework. *Pattern Recognition Letters*, 102:37 – 43, 2018.

Guénaël Cabanes, Younès Bennani, Renaud Destenay, and André Hardy. A new topological clustering algorithm for interval data. *Pattern Recognition*, 46(11):3030 – 3039, 2013.

Frédéric Chazal, Leonidas J. Guibas, Steve Y. Oudot, and Primoz Skraba. Persistence-based clustering in riemannian manifolds. *J. ACM*, 60(6), 2013.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jianwei Yang, Zirun Chen, Wen-Sheng Chen, and Yunjie Chen. Robust affine invariant descriptors. *Mathematical Problems in Engineering*, 2011.

Mark Fanty and Ronald 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, 1991.

Fabian Paschke, Christian Bayer, Martyna Bator, Uwe Mönks, Alexander Dicks, Olaf Enge-Rosenblatt, and Volker Lohweg. Sensorlose Zustandsüberwachung an Synchronmotoren. In F. Hoffmann, editor, *23. Workshop Computational Intelligence 2013. Proceedings*, volume 46, pages 211–225. KIT Scientific Publishing, 2013.

Clara Higuera, Katheleen J. Gardiner, and Krzysztof 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.

Ross D. King, Chen-Chieh Feng, and Alistair A.P. Sutherland. Statlog: Comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence*, 9, 2000.

Oriane Siméoni, Mateusz Budnik, Yannis Avrithis, and Guillaume Gravier. Rethinking deep active learning: Using unlabeled data at model training. 2019.