A TOPOLOGICAL APPROACH FOR SEMI-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine Learning and Deep Learning methods have become the state-of-the-art approach to solve data classification tasks. In order to use those methods, it is necessary to acquire, and label, a considerable amount of data; however, this is not straightforward in some fields since data annotation is time consuming and might require expert knowledge. This challenge can be tackled by means of semi-supervised learning methods that take advantage of both labelled and unlabelled data. In this work, we present a new semi-supervised learning method based on techniques from Topological Data Analysis. In particular, we have used a homological approach that consists in studying the persistence diagrams associated with data from binary classification tasks using the bottleneck and Wasserstein distances. In addition, we have carried out a thorough analysis of the developed method using 5 structured datasets. The results show that the semi-supervised method developed in this work outperforms both the results obtained with models trained with only manually labelled data, and those obtained with classical semi-supervised learning methods, improving the models up to a 16%.

1 INTRODUCTION

Machine Learning and Deep Learning techniques have become the state-of-the-art approach to solve classification problems in a wide variety of fields such as biology (Affonso et al., 2017), security (Akçay et al., 2016), or medicine (Araújo et al., 2017). One of the main problems of these techniques is the great amount of data that they need (Sun et al., 2017). This may not seem a problem due to the large amount of data that is generated in a daily basis; however, data acquisition is not easy in some fields due to, for example, a limited budget to obtain samples, the need to perform an invasive medical procedure or destructive processes. In addition, in supervised learning, one of the main paradigms in machine learning, data has to be manually annotated, and it is well-known that this might be a problem due to the time and experience that is required to conduct this task (Irvin et al., 2019). To tackle this challenge, a family of methods that has been successfully applied in the literature is semi-supervised learning (Berthelot et al., 2019; Laine & Aila, 2017).

Semi-supervised learning methods provide a mean of using unlabelled data to improve models' performance when we have access to a large corpus of data that is difficult to annotate. Traditional semi-supervised learning algorithms, such as Label Spreading (Zhou et al., 2004) or Label Propagation (Zhu & Ghahramani, 2002), focus on the distance among the data points to annotate unlabelled data points; that is, on the metric and density characteristics of the data in a dataset. However, topological characteristics of the data are not used, and this is the approach proposed in this paper.

Topological Data Analysis (from now on, TDA) has arisen as a field to extract topological and geometrical information from data, to reveal dynamical organisation of the brain (Saggar et al., 2018), to recognising atmospheric river patterns in large climate datasets (Muszynski et al., 2019), or to examine spreading processes on networks (Taylor et al., 2015). An important result of TDA is the Manifold Hypothesis (Fefferman et al., 2016), that states that high dimensional data tends to lie in low dimensional manifolds, and that has inspired our definition of a semi-supervised learning method for binary classification tasks. Intuitively, our method is based on the idea that given two sets of data points A and B, we can define two manifolds associated with each set, $\mathcal{M}_A$ and $\mathcal{M}_B$ respectively. Now, given an unlabelled data point x that belongs to either A or B; if x belongs to A,

analogously for B, then the manifold associated with $A \cup \{x\}$ and $\mathcal{M}_A$ will be more similar than the manifold associated with $B \cup \{x\}$ and $\mathcal{M}_B$.

The rest of the paper is devoted to introduce the aforementioned idea formally. Namely, we provide a complete description of our semi-supervised method based on TDA notions in Section 2. Subsequently, we conduct a thorough analysis for our method in 5 structured datasets, and compare its performance with classical semi-supervised learning methods, see Section 3. Finally, we end the paper with some conclusions and some ideas for further work.

All the code developed for this project and also the conducted experiments are available at the project webpage `https://anonymous.4open.science/r/TopologicalSSL-E12C/`.

2 METHOD

In this section, we describe the semi-supervised learning algorithm that we have designed to tackle binary classification tasks. We start with a set X_1 of points from class 1, a set X_2 of points from class 2, and a set X of unlabelled points. The objective of our algorithms is to annotate the elements of X by using topological properties of X_1 and X_2. We assume some familiarity with notions employed in TDA such as Vietoris-Rips filtration (we denote by V_X to the Vietoris-Rips filtration associated with a set X), persistence diagrams (we denote by $P(F)$ to the persistence diagram associated with a filtration F), and the bottleneck and Wasserstein distances (denoted by d_B and d_W respectively). For a detailed introduction to these topics see (Zomorodian, 2012).

Our semi-supervised learning algorithm takes as input the sets X_1 and X_2, a point $x \in X$, a threshold value t, and a flag that indicates whether the bottleneck or the Wasserstein distance should be used, we denote the chosen distance as d. The output produced by our algorithm is whether the point x belongs to X_1, X_2 or none of them. In order to decide the output of the algorithm, our hypothesis is that if a point belongs to X_1, analogously for X_2, the topological variation that X_1 will suffer when adding the point will be minimal; whereas if the point does not belong to X_1, the variation will be greater. In particular, we proceed as follows:

1. Construct the Vietoris-Rips filtrations V_{X_1}, V_{X_2}, $V_{X_1 \cup \{x\}}$ and $V_{X_2 \cup \{x\}}$;
2. Construct the persistence diagrams $P(V_{X_1})$, $P(V_{X_2})$, $P(V_{X_1 \cup \{x\}})$ and $P(V_{X_2 \cup \{x\}})$;
3. Compute the distances $d(P(V_{X_1}), P(V_{X_1 \cup \{x\}}))$ and $d(P(V_{X_2}), P(V_{X_2 \cup \{x\}}))$, from now on d_1 and d_2 respectively;
4. If both d_1 and d_2 are greater than the threshold t, return none; otherwise, return the set associated with the minimum of the distances d_1 and d_2.

The above algorithm is diagrammatically described in Figure 1, and it is applied for all the points of the set of unlabelled points X. Note that if we use a threshold value of 0, the algorithm will annotate all the points of X; however, this might introduce some noise as we will see in the experiments of the next section.

3 EVALUATION

In this work, we have used 5 different datasets taken from the UCI Machine Learning Repository (Dua & Graff, 2017) — the datasets are banknote, breast cancer, ionosphere, pima indian diabetes, and sonar; a summary of the features of these datasets is provided in the appendix. For our study, we have split each of the datasets of the benchmark into two different sets: a training set with the 80% of the data, and a testing set with the 20% of the data. In addition, for each training dataset, we have selected 25 samples per class using them as labelled data, and removing the annotation of the rest of the training data to test the semi-supervised learning methods.

To check the correct performance of our methods we have trained two machine learning algorithms that are SVM (Cortes & Vapnik, 1995) and Random Forest (Ho, 1995) using the scikit-learn library (Pedregosa et al., 2011). In particular, we have trained these models with the initial annotated data obtaining a base result. Subsequently we have used the developed methods, and three classical semi-supervised learning techniques (namely, Label Propagation (Zhu & Ghahramani, 2002), Label

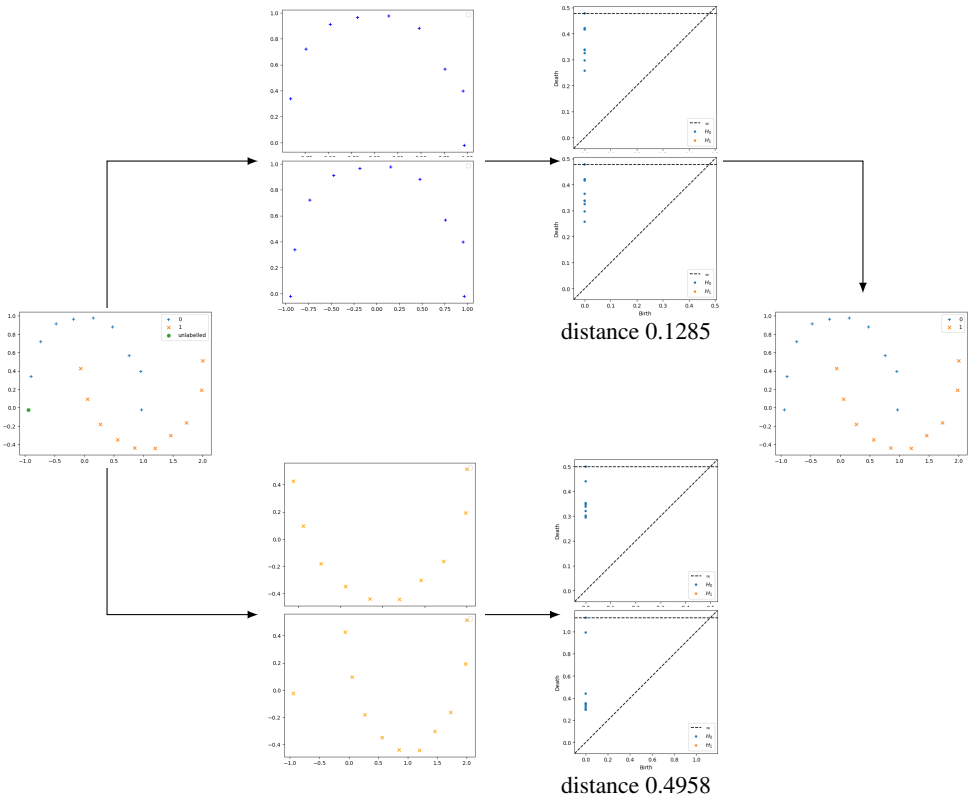

Figure 1: Example of the application of our method using the bottleneck distance, and using 0.6 as threshold value.

Spreading (Zhou et al., 2004), and self-training (Yarowsky, 1995)) to annotate the unlabelled data. Finally, we have retrained the two ML models with all the annotated data, to see the variation in the models' performance. Such a performance has been evaluated using the accuracy metric. In addition, in order to evaluate the behaviour of the annotation methods we have taken into account the percentage of the data points correctly labelled and the percentage of data labelled with respect to the total available data. For testing our methods, we have used 10 variations of our semi-supervised learning method by considering the bottleneck and the Wasserstein distance and using 5 different threshold values (0.8, 0.6, 0.4, 0.2 and 0.0). Our method has been implemented in the Python programming language by using the functionality provided by the scikit-tda library (Saul & Tralie, 2019).

The performance of the aforementioned methods on the studied datasets is included in Table 1. From these results we can withdraw several conclusions. In general, our method offers good results, improving the base results in 8 out of the 10 models. Moreover, our method obtains better results than the classical semi-supervised learning techniques in 8 out of the 10 models, the only drawback is that different threshold and distances must be tested. Regarding the choice of distance, there are not considerable differences between the bottleneck and the Wasserstein distance. On the contrary, the value of the threshold is relevant for the performance of the models. When using the bottleneck distance, we have observed that the best results are obtained when setting the threshold value to 0.8. This is also the case for most of the experiments conducted with the Wasserstein distance, but for that distance there is not an optimum threshold — although, threshold values of 0.8 and 0.6 produce the best results for 9 out of 10 experiments.

Finally, we analyse the percentage of points that are labelled by each version of our algorithm, see Table 2. As expected, when the threshold value decreases, the percentage of labelled points increases. However, when the threshold value decreases, the percentage of correctly labelled points also decreases; hence, using such an annotation might introduce some noise. As we have previously

Table 1: Accuracy results for the SVM and RF classifiers trained with data annotated by each of the annotation methods (classical, homological and connectivity) together with the results obtained with the initial data (base) in the 5 structured datasets. Best result for each dataset is highlighted in bold face.

Method	Banknote		Breast Cancer		Ionosphere		Prima Indian		Sonar		Mean (std)	
	SVM	RF	SVM	RF	SVM	RF	SVM	RF	SVM	RF	SVM	RF
Base	97.0	88.6	89.3	**96.1**	80.0	93.3	65.7	60.8	61.3	64.5	78.7(15.2)	80.7(16.7)
Label Propagation	97.4	93.2	90.3	89.3	86.7	86.7	64.3	68.5	58.1	54.8	79.3(17.1)	78.5(16.3)
Label Spreading	97.4	93.2	90.3	89.3	86.7	86.7	64.3	68.5	58.1	54.8	79.3(17.1)	78.5(16.3)
Self Training classifier	95.1	93.6	35.9	35.9	85.0	86.7	66.4	66.4	58.1	67.7	68.1(23.2)	70.1(22.4)
Bottleneck threshold 0.8	**99.2**	92.4	93.2	91.3	78.3	**95.0**	63.6	64.3	61.3	64.5	79.1(17.0)	**81.5(15.6)**
Bottleneck threshold 0.6	**99.2**	91.3	89.3	90.3	75.0	88.3	59.4	63.6	48.4	45.2	74.3(20.9)	75.7(20.6)
Bottleneck threshold 0.4	97.4	90.5	87.4	85.4	78.3	86.7	63.6	62.9	45.2	45.2	74.4(20.5)	74.1(19.5)
Bottleneck threshold 0.2	97.4	90.5	87.4	85.4	78.3	86.7	63.6	62.9	45.2	45.2	74.4(20.5)	74.1(19.5)
Bottleneck threshold 0.0	97.4	90.5	87.4	85.4	78.3	86.7	63.6	62.9	45.2	45.2	77.1(22.6)	74.1(19.5)
Wasserstein threshold 0.8	97.4	89.8	92.2	88.4	80.0	**95.0**	**68.5**	67.8	61.3	64.5	79.9(15.3)	81.1(13.9)
Wasserstein threshold 0.6	**99.2**	93.6	89.3	87.4	70.0	91.7	61.5	61.5	**74.2**	61.3	78.9(15.2)	79.1(16.3)
Wasserstein threshold 0.4	97.0	96.2	87.4	87.4	76.7	81.7	60.8	62.9	71.0	71.0	78.6(14.1)	79.8(13.2)
Wasserstein threshold 0.2	97.0	96.2	87.4	87.4	76.7	81.7	60.8	62.9	71.0	71.0	78.6(14.1)	79.8(13.2)
Wasserstein threshold 0.0	97.0	96.2	87.4	87.4	76.7	81.7	60.8	62.9	71.0	71.0	78.6(14.1)	79.8(13.2)

seen in Table 1, models trained with annotations produced by our algorithm using threshold values of 0.8 and 0.6 usually produce better results than using smaller threshold values. In particular, when we fix a threshold value that is lower than 0.4, all points are labelled; but, the percentage of correctly labelled points is usually lower than the percentages of classical semi-supervised learning methods. On the contrary, the amount of points that are correctly annotated when using a threshold value of 0.8 or 0.6 is greater than in the classical techniques; and, as we have previously seen this produces better models. Therefore, we can conclude that is more important the quality than the quantity of data.

Table 2: Percentage of labelled data points and percentage of correctly labelled data points by each semi-supervised method

	Banknote		Breast Cancer		Ionosphere		Prima Indian		Sonar	
	% labelled	% correct	% labelled	% correct	% labelled	% correct	% labelled	% correct	% labelled	% correct
Label Propagation	100	93.67	100	90.87	100	81.74	100	70.43	100	63.78
Label Spreading	100	93.67	100	90.87	100	81.74	100	70.43	100	63.78
Self Training Classifier	100	95.65	100	39.18	100	82.99	100	66.26	100	70.87
Bottleneck threshold 0.8	15.12	100	50.72	97.16	30.71	100	5.91	88.24	0	0
Bottleneck threshold 0.6	84.59	92.73	91.82	90.31	90.04	83.41	32.87	68.78	87.4	53.15
Bottleneck threshold 0.4	100	90.26	100	85.58	100	80.08	100	57.22	100	48.03
Bottleneck threshold 0.2	100	90.26	100	85.58	100	80.08	100	57.22	100	48.03
Bottleneck threshold 0.0	100	90.26	100	85.58	100	80.08	100	57.22	100	48.03
Wasserstein threshold 0.8	17.49	100	38.7	97.52	24.9	100	9.57	92.73	0	0
Wasserstein threshold 0.6	73.53	99.1	86.3	91.92	57.68	89.21	55.65	73.44	32.28	97.56
Wasserstein threshold 0.4	100	94.99	100	88.22	100	78.42	100	65.57	100	74.02
Wasserstein threshold 0.2	100	94.99	100	88.22	100	78.42	100	65.57	100	74.02
Wasserstein threshold 0.0	100	94.99	100	88.22	100	78.42	100	65.57	100	74.02

4 Conclusions and further work

In this work, we have studied the application of Topological Data Analysis techniques to the semi-supervised learning setting to tackle binary classification problems with a limited amount of labelled data. The results show that our method can create classification models that achieve better results than those obtained when using classical semi-supervised learning methods.

We plan to extend our work in different ways. First of all, the proposed method can be expanded to multi-class classification tasks, and, an iterative version of the algorithm can be easily developed. In addition, we plan to design new semi-supervised learning algorithms based on other notions from TDA, such as the connectivity of the data.

References

C. Affonso, A. L. Debiasso Rossi, F. H. A. Vieira, et al. Deep learning for biological image classification. *Expert Systems with Applications*, 85(1):114–122, 2017.

S. Akçay, M. E. Kundegorski, M. Devereux, et al. Transfer learning using convolutional neural networks for object classification within x-ray baggage security imagery. In *2016 IEEE International Conference on Image Processing*, ICIP'16, pp. 1057–1061, 2016.

T. Araújo, G. Aresta, E. Castro, et al. Classification of breast cancer histology images using convolutional neural networks. *PLoS ONE*, 12(6), 2017.

D. Berthelot et al. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems 32*, pp. 5049–5059. Curran Associates, Inc., 2019.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pp. 278–282. IEEE, 1995.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

S. Laine and T. Aila. Temporal Ensembling for Semi-Supervised Learning. In *5th International Conference on Learning Representations*, ICLR'17, pp. 1–13, 2017.

Grzegorz Muszynski, Karthik Kashinath, Vitaliy Kurlin, Michael Wehner, et al. Topological data analysis and machine learning for recognizing atmospheric river patterns in large climate datasets. *Geoscientific Model Development*, 12(2):613–628, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Manish Saggar, Olaf Sporns, Javier Gonzalez-Castillo, Peter A Bandettini, Gunnar Carlsson, Gary Glover, and Allan L Reiss. Towards a new approach to reveal dynamical organization of the brain using topological data analysis. *Nature communications*, 9(1):1–14, 2018.

Nathaniel Saul and Chris Tralie. Scikit-tda: Topological data analysis for python, 2019. URL `https://doi.org/10.5281/zenodo.2533369`.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

Dane Taylor, Florian Klimm, Heather A Harrington, Miroslav Kramár, Konstantin Mischaikow, Mason A Porter, and Peter J Mucha. Topological data analysis of contagion maps for examining spreading processes on networks. *Nature communications*, 6(1):1–11, 2015.

David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pp. 189–196, 1995.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, pp. 321–328. MIT Press, 2004.

Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, 2002.

Afra Zomorodian. Topological data analysis. *Advances in applied and computational topology*, 70: 1–39, 2012.

A FEATURES OF THE STUDIED DATASETS

Dataset	# Examples	# Unlabelled examples	# Features
Banknote	1372	1322	4
Breast Cancer	569	519	30
Ionosphere	351	301	34
Pima Indian Diabetes	768	718	8
Sonar	208	158	60

Table 3: Description of the datasets employed in our experiments.