

With the tremendous growth of data available electronically, the constitution of labeled training sets for learning, which often requires a skilled human agent or a physical experiment, becomes unrealistic. One alternative is to gather a small set of labeled training examples $\mathcal{S} = (\mathbf{x}_i, y_i)_{1 \leq i \leq l} \sim \mathcal{D}^l$ and try to take advantage of the huge amount of unlabeled observations $\mathcal{Z}_l = (\mathbf{x}_i)_{l+1 \leq i \leq l+u} \sim \mathcal{D}_X^u$, with $u \gg l$, to find a prediction function more efficient than the one found based on the labeled training set.

However, in some cases, the set of labeled training examples is not well constructed, in the sense that it does not reveal the complexity of the data set. For example, it would be true if labeled observations were originally drawn randomly from a huge set of unlabeled data, without revealing the intrinsic nature of the whole set.

In this project, we propose to consider topological data analysis (TDA) to improve the selection of informative labeled examples in conjunction with semi-supervised algorithms [1, 4]. TDA is a growing mathematical field which is starting to be widely applied in many different contexts (see [3] for a survey on the subject). It consists in using algebraic tools, which can be computed, to detect some topological features of the data we are considering. TDA has already been used for machine learning algorithm as introduced in [2, 5] among others.

The aim here is to detect interesting observations (regarding the topology of the data) in order to be labeled, to ask for their labels (active learning step) and then to use semi-supervised learning algorithms to improve the performance of the prediction function. The resulting algorithm will be tested on real-world applications using data extracted from different information retrieval problems. We have in mind a specific data set coming from biology application (ongoing collaboration with I2BC).

For this position, we are looking for inquisitive minds who are interested in both theory and applications of machine learning, who have a good background in mathematics and good programming skills, and who are curious about topology and want to learn more from the maths point of view.

References

- [1] Massih-Reza Amini, Nicolas Usunier, and François Laviolette. A transductive bound for the voted classifier with an application to semi-supervised learning. In *Advances in Neural Information Processing Systems 21*, pages 65–72. 2009.
- [2] Frédéric Chazal, Leonidas J. Guibas, Steve Y. Oudot, and Primoz Skraba. Persistence-based clustering in riemannian manifolds. *J. ACM*, 60(6):41:1–41:38, November 2013.
- [3] Frédéric Chazal and Bertrand Michel. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. *ArXiv e-prints*, October 2017.
- [4] Yury Maximov, Massih-Reza Amini, and Zaid Harchaoui. Rademacher complexity bounds for a penalized multi-class semi-supervised algorithm. *Journal of Artificial Intelligence Research*, 61(1):761–786, 2018.
- [5] Larry Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5(1):501–532, 2018.