Massih-Reza Amini
Émilie Devijver

Research project

Multi-class semi-supervised learning through pseudo-labelling

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 $\mathcal{S} = (\mathbf{x}_i, y_i)_{1 \leq i \leq \ell} \sim \mathcal{D}^\ell$ and try to take advantage of the huge amount of unlabeled observations $\mathcal{Z}_{\mathcal{U}} = (\mathbf{x}_i)_{\ell+1 \leq i \leq \ell+u} \sim (\mathcal{D}_X)^u$, with $u >> \ell$, to find a more efficient prediction function that the one that can be found using only the labeled training set.

Many approaches have been proposed in the literature to learn a classifier in this context, mainly based on the cluster assumption stipulating that the decision boundary should not pass through data clusters found over $\mathcal{Z}_{\mathcal{U}}$.

In this project, we consider the self-learning approach which consists in first learning a classifier using the labeled training set $\mathcal{S}$ and then use the predictions as scores of confidence by assigning iteratively pseudo-labels to unlabeled data having prediction scores above a given threshold and then training a new classifier.

The main question here is the choice of the threshold which remains an open question in the case of multi-class classification problems. In the binary case, a solution has been proposed under the PAC-Bayes framework which consists in bounding the error of the Bayes classifier estimated over the unlabeled data having an absolute prediction score higher than a threshold; and then choosing the threshold which gives the tightest bound [1].

In this project we propose first, to extent the previous theoretical result to the multiclass case, using for example the confusion matrix as introduced in [2]. And to use the result for automatically estimating different thresholds for pseudo-labeling unlabebed examples into different classes. The resulting algorithm will be then tested on real-world applications using genetics data or data extracted from different information retrieval problems.

For this position, we are looking for inquisitive minds who are interested in both theory and applications of machine learning and who have a good back-ground on statistics and good programming skills.

References

[1] Massih-Reza Amini, François Laviolette, and Nicolas Usunier. A transductive bound for the voted classifier with an application to semi-supervised learning. In *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pages 65–72, 2008.

[2] Emilie Morvant, Sokol Koço, and Liva Ralaivola. Pac-bayesian generalization bound on confusion matrix for multi-class classification. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.