Semisupervised learning

Suppose to be given a labeled set of objects (samples) (X_i, Y_i) for $i = 1, \ldots, n$ and an unlabelled set X_j, $j = n{+}1, \ldots, n{+}m$. The aim is to develop a method which utilizes the unlabeled set (X_j) to improve the quality of classification. The idea of label propagation suggests to perform a clustering procedure at the first step of the algorithm using the whole set $X_1, \ldots, X_{n+m}$. Then each cluster is labeled due to majority of labeled data within this cluster. This procedure AWSL (semisupervised learning) is implemented in terms of the weight matrix $W_{ij}^{(k)}$ obtained from the clustering procedure and the vector $\widetilde{\boldsymbol{\theta}}^{(k)} = (\widetilde{\theta}_i^{(k)})$ which estimates the success probabilities $\theta_i = I\!P(Y_i = 1)$ for all i.

Workpackages:

1. Efficient scalable implementation of the AWSL for training set of large size
2. Retraining with the testing dataset
3. Exploring the theoretical properties of the AWSL including propagation, separation, and consistency
4. Application to tracking problem: given a collection of screenshots $\boldsymbol{X}(t), \boldsymbol{Y}(t)$, track a possibly moving object descibed by the labels $\boldsymbol{Y}(t)$
5. Application to social, media, bio, medicine, financial data

Literature: Efimov, Adamyan, Spokoiny (2017) Adaptive nonparametric clustering. arxiv 1709.09102.

Contact: Maxim Panov, Igor Silin, Kirill Efimov, VS