

4 April, 9:40 - 10:00, Room 13.1

Randomly perturbed random forests

Laura Anderlucci¹, Angela Montanari¹

¹ University of Bologna, Italy, laura.anderlucci@unibo.it, angela.montanari@unibo.it

In supervised classification, a change in the distribution of a single feature, a combination of features, or the class boundaries, may be observed between the training and the test set. This situation is known as dataset shift. As a result, in real data applications, the common assumption that the training and testing data follow the same distribution is often violated. In order to address dataset shift we propose to randomly introduce more variability in the training set by sketching the input data matrix resorting to random projections of units. We then modify the random forests algorithm to involve sketched, rather than bootstrapped, versions of the original data. Results on real data show that perturbing the training data via matrix sketching improves the prediction accuracy of test units that have a different distribution in terms of variance structure.

Keywords: classification, dataset shift, data perturbation

Dataset shift occurs when the testing (unseen) data experience a phenomenon that leads to a change in the distribution of a single feature, a combination of features, or the class boundaries ([2]), i.e., when the training and test joint distributions are different:

$$P_{tr}(\mathbf{X}|\mathbf{y}) \neq P_{te}(\mathbf{X}|\mathbf{y}),$$

where $P_{tr}(\mathbf{X}|\mathbf{y})$ indicate the features' distribution according to $\mathbf{y}$ in the training set, while $P_{te}(\mathbf{X}|\mathbf{y})$ in the test set.

An example of dataset shift can be found in the classification of metastatic cancer. Specifically, when cancer spreads, the secondary cancer cells usually look like abnormal versions of the primary cancer cells (in the tissue where the cancer began). For example, if breast cancer spreads to the lungs, the metastatic tumor in the lung is made up of cancerous breast cells (not lung cells) and is then described as metastatic breast cancer (not lung cancer).

A tool that can be employed to identify the tissue of origin of the metastasis is the modelling of Micro-RNA (mi-RNA) expression profiles. Recently, [3] proposed a miRNA-based tissue classifier to identify the tissue origin of metastatic tumors and they tested on a dataset including 205 primary tumors and 131 metastatic tumors, representing 22 different tumor origins. Exploratory data analysis carried on these data highlighted three main aspects:

- The cancer classes are highly imbalanced;

- The primary and the metastatic cells have similar expression mean values;
- The primary and the metastatic cells have different variability.

Such features suggest that the problem can indeed be cast within the dataset shift framework.

In this work we propose a novel method that accurately classifies the secondary from the primary cancer cells, allowing to identify the most likely tissue of origin of metastatic samples. In order to deal with dataset shift, data perturbation via Random Projections (RP) is proposed.

In [1], it was shown that using random projections to generate new ‘compressed’ points would allow not only to fix the imbalance problem, but also to explore information outside the convex hull of the original data, so as to avoid the risk of overfitting. The perturbation induced by matrix sketching modified the variability of the training set while preserving the mean values.

Based on such result, we propose a novel classifier that modifies the random forests algorithm to involve sketched, rather than bootstrapped, versions of the original data. Results on real data show that perturbing the training data via matrix sketching improves the prediction accuracy of test units that have a different distribution in terms of variance structure.

References

- [1] R. Falcone, L. Anderlucci, and A. Montanari. Matrix sketching for supervised classification with imbalanced classes. *Data Mining and Knowledge Discovery*, 36(1):174–208, 2022.
- [2] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.
- [3] N. Rosenfeld, R. Aharonov, E. Meiri, S. Rosenwald, Y. Spector, M. Zepeniuk, H. Benjamin, N. Shabes, S.t Tabak, A. Levy, et al. Micrornas accurately identify cancer tissue origin. *Nature Biotechnology*, 26(4):462–469, 2008.