Machine learning meets false discovery rate: application to graph anomaly detection

Ariane Marandon

Sorbonne Université, Laboratoire de Probabilités, Statistiques and Modélisation (LPSM), ariane.marandon-carlhian@sorbonne-universite.fr

February 25, 2022

Novelty/outlier detection is the problem of identifying observations that do not conform to a well defined notion of normal behavior. More formally, the problem we consider is as follows: we have at hand a sample of observations $Y_1, ..., Y_n$ each with a common unknown distribution P_0, which we call nominal observations, and a sample of observations $X_1, ..., X_m$ that may contain both nominals, i.e. observations marginally distributed according to P_0, and novelties if otherwise. Specifically, the aim is to control the false discovery rate (FDR) defined as the proportion of detections that are false (true nominals declared as novelties) at some fixed error margin, say 10% for instance, while maximizing the number of detections under this constraint.

Under this setting, Conformal Anomaly Detection (CAD) (Bates et al., 2021; Mary and Roquain, 2022; Yang et al., 2021) is a breakthrough novelty detection technique that provides the guarantee to have finite-sample control of the FDR. In this work, we propose a powerful extension of CAD, called AdaDetect (Marandon et al., 2022), that leverages the unlabeled data and classification methods. We prove that AdaDetect comes with finite sample guarantees: it controls the FDR strongly and approximates the oracle in terms of the power, with explicit remainder terms that are small under mild conditions. In practice, AdaDetect can be used in combination with any classifier, which allows the user to choose the most relevant classification approach. We illustrate the versatility of our method on three semi-supervised learning problems related to graph data: graph-level anomaly detection, node-level anomaly detection, and link prediction.

Keywords: multiple testing, novelty detection, false discovery rate, conformal p-values, graphs

References

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. (2021). Testing for outliers with conformal p-values.

Marandon, A., Lei, L., Mary, D., and Roquain, E. (2022). Machine learning meets false discovery rate.

Mary, D. and Roquain, E. (2022). Semi-supervised multiple testing. Electronic Journal of Statistics, 16(2):4926 – 4981.

Yang, C.-Y., Lei, L., Ho, N., and Fithian, W. (2021). Bonus: Multiple multivariate testing with a data-adaptive test statistic.