Anomaly detection at scale and the role of Bayesian ensembles for detector selection

Dusan Randjelovic[1,2]

[1]Senior Data Scientist, SmartCat, www.smartcat.io, dusan.randjelovic@smartcat.io
[2]University Centre for Applied Statistics, Novi Sad, dusan.randjelovic@uns.ac.rs

Anomaly detection services could prove very useful in securing SLA (service-level agreement) requirements, like latency or high-availability, for applications deployed in the cloud. However, a good anomaly detection service (ADS) is itself a non-trivial requirement. Difficulties come from the fact that hundreds of collected infrastructure metrics represent unlabeled, unbalanced, multivariate time series that are usually autocorrelated or otherwise non-stationary in nature and can exhibit complex contextual or collective anomalies and high false positive rate of point anomalies. Furthermore, ADS itself is usually required to serve several other functions besides obvious diagnostics and real-time detection of anomalous events, for example: predictive maintenance, root-cause analysis, alerting, accountability, reporting. First goal of this research project is to scope and properly define data science, data engineering and business-related requirements for ADS solution as cloud-based analytics platform, with synthetic data, that can serve as a blueprint for real-life implementation.

From data science perspective, anomaly detection is predominantly done in unsupervised fashion. There are many approaches to this problem: machine learning detectors like one-class SVM or robust PCA, forecasting methods like ARIMAX or Holt-Winters or deep learning methods for anomaly detection with GANs [1], LSTMs or robust autoencoders [2]. There is a general consensus that combination of multiple detectors into ensembles could be beneficial to overall accuracy of detection, although ensembles for unsupervised anomaly detection are more recent and emerging area of research [3]. Second goal of this research project is to investigate various bayesian ensemble learning models with emphasis on usage of bayesian approach for detector selection. Work presented is a continuation of previous efforts [4], with implementation in pyMC3 python library [5], on Yahoo Webscope S5 dataset [6].

References

[1] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth and G. Langs, Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, CoRR (2017), abs/1703.05921.

[2] R. Chalapathy, A. K. Menon and S. Chawla, Robust, deep and inductive anomaly detection, CoRR (2017), abs/1704.06743.

[3] A. Zimek, R. J. G. B. Campello and J. Sander, Ensembles for unsupervised outlier detection: Challenges and research questions a position paper, SIGKDD Explor. Newsl. **5**(1) (2014), 11–22.

[4] E. Yu and P. Parekh. A Bayesian Ensemble for Unsupervised Anomaly Detection, ArXiv e-prints, (2016).

[5] J. Salvatier, T. V. Wiecki and C. Fonnesbeck, Probabilistic programming in python using pymc3, PeerJ Computer Science (2016), 2:e55.

[6] N. Laptev and S. Amizadeh, Yahoo anomaly detection dataset s5 (2015), available from `http://webscope.sandbox.yahoo.com/catalog.php?datatype=s/\&did=70`