ATD: Anomalous Topic Discovery in High Dimensional Discrete Data

We propose an algorithm for detecting patterns exhibited by anomalous clusters in high dimensional discrete data. Unlike most anomaly detection (AD) methods, which detect individual anomalies, our proposed method detects groups (clusters) of anomalies; i.e. sets of points which collectively exhibit abnormal patterns. In many applications this can lead to better understanding of the nature of the atypical behavior and to identifying the sources of the anomalies. Moreover, we consider the case where the atypical patterns exhibit on only a small (salient) subset of the very high dimensional feature space. Individual AD techniques and techniques that detect anomalies using all the features typically fail to detect such anomalies, but our method can detect such instances collectively, discover the shared anomalous patterns exhibited by them, and identify the subsets of salient features. In this paper, we focus on detecting anomalous topics in a batch of text documents, developing our algorithm based on topic models. Results of our experiments show that our method can accurately detect anomalous topics and salient features (words) under each such topic in a synthetic data set and two real-world text corpora and achieves better performance compared to both standard group AD and individual AD techniques. All required code to reproduce our experiments is available from https://github.com/hsoleimani/ATD.