Dataset Description

ASD

The Application Server Dataset (**ASD**) for multivariate time series anomaly detection and interpretation.

The ASD dataset is collected from a large Internet company. It characterizes the status of different servers (entities) using a group of metrics. A group of stable services are run on each server, thus the data does not experience service changes or concept drifts during the time period included in the dataset.

ASD contains 12 different entities (each for a server), each of which has 19 metrics characterizing the status of the server (including CPU-related metrics, memory-related metrics, network metrics, virtual machine metrics, *etc.*).

The data points in ASD are equally-spaced 5 minutes apart. The first 30-day data are training set (the last 30% data in training set are kept for validation), while the last 15-day data are testing set. Anomalies and their most anomalous dimensions in ASD testing set have been labeled by domain experts based on incident reports and domain knowledge.

For each entity, there are three data files under `processed/` folder. `omi-x_train.pkl` and `omi-x_test.pkl` are training data and testing data, respectively. `omi-x_test_label.pkl` is the ground-truth labels on test set. Under `interpretation_label/` folder, each entity has an corresponding `.txt` file, which shows the ground-truth interpretation labels. The format of interpretation labels is: $start - end : m_1, m_2, \cdots, m_k$, where "start" and "end" are the start and end indexes of the anomaly segment, $m_1, m_2, \cdots, m_k$ are the anomalous metrics in this segment (the metrics are listed in no particular order). `anomaly_type.txt` shows the anomaly type of each anomaly segment.

The overall anomaly ratio in ASD is 4.61%. Sensitive information in data (*e.g.*, IP address, metric names, concrete timestamps) has been removed to protect the confidentiality of data providers.

SMD

Server Machine Dataset (SMD) is a public dataset[1] for multivariate time series anomaly detection. Here we list part of SMD data that are used for evaluation in this paper, which has none or little concept drift during the data collection period.

The data points in SMD are equally-spaced 1 minute apart. The dataset contains 12 entities (each for a machine), each entity has 38 metrics. The first half of data is used as training set while the other half is testing set. The overall anomaly ratio in SMD is 5.84%.

For each entity, there are three data files under `processed/` folder. `machine-xxx_train.pkl` and `machine-xxx_test.pkl` are training data and testing data, respectively. `machine-xxx_test_label.pkl` is the ground-truth labels on test set. Under `interpretation_label/` folder, each entity has an corresponding `.txt` file, which shows the ground-truth interpretation labels. The format of interpretation labels is: $start - end : m_1, m_2, \cdots, m_k$, where "start" and "end" are the start and end indexes of the anomaly segment, $m_1, m_2, \cdots, m_k$ are the anomalous metrics in this segment (the metrics are listed in no particular order).

SWaT & WADI

SWaT and WADI are two MTS datasets about water treatment plants, which were firstly used for MTS anomaly detection in [2]. Both datasets are collected and released by iTrust, Centre for Research in Cyber Security, Singapore University of Technology and Design. You may follow the instructions on their website[3] to acquire the datasets. In this paper, we use SWaT.A2_Dec 2015, version 0 and WADI.A1[4].

References

[1] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. 2019. Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2828–2837.

[2] Dan Li, Dacheng Chen, Baihong Jin, Lei Shi, et al. 2019. Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In International Conference on Artificial Neural Networks. Springer, 703–716.

[3] https://itrust.sutd.edu.sg/itrust-labs_datasets/

[4] https://itrust.sutd.edu.sg/itrust-labs_datasets/dataset_info/