Course Project: Multi-dataset Time Series Anomaly Detection

MSBD 5002 Data Mining, Fall 2021

Due: 11:59pm, December 3rd

1 Submission Guidelines

- Project should be submitted by CANVAS.
- You need to zip the following two files together:
  - proj_groupid_report.pdf/.docx: Please put your report in this file. (Attachments should be original .pdf or .docx, NOT compressed)
  - proj_groupid_code.zip: The zip file contains all your source codes for this project. Please do not submit data. You may assume that the local path is "../data-sets/KDD-Cup/data".
  - submission.csv:
- All attachments, including report and code, should be named in the format of: proj_groupid.zip. E.g., for a group with groupid as 4, the project can be named as: proj_group4.zip.
- You are allowed to form a group with up to 4 members. Please register your group through the link.
- Submissions not following the rules above are NOT accepted.
- 20 marks will be deducted for every 24 hours after the deadline.
- Your grade will be based on the correctness, efficiency and clarity.
- The email for Q&A: msbdt5002fall2021@gmail.com.
- **Plagiarism will lead to zero mark.**

2 Objective

The objective of this project is twofold:

- To acquire a better understanding of unsupervised data mining techiniques.
- Familiar with how to complete a given project by discussing with groupmates, surveying literatures and coding experiments.

3 Background

This project is based on the KDD Cup 2021, one of the most famous competitions in data mining area. The more details can be accessed through the link and the video.

Time series data is composed of a sequence of values over time, which exists in many applications, such as stock prices and websites. However, some values may deviate so much from other normal observations, and these unnormal values are called anomalies. Anomalous data can indicate critical incidents, such as technical incidents. Anomaly detection on time series data is to identify these unnormal data points, where machine learning technique is progressively being utilized to detect anomalies automatically.

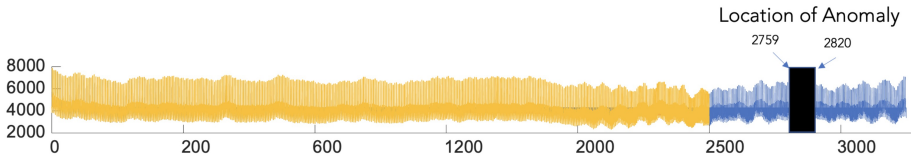

Figure 1: Example of anomaly of a single time series data.

4 Task Description

Here we introduce the task in detail. Given the time series data, your task is to discover the location of the anomaly. You can download data for this project via the link.

Specifically, KDD Cup 2021 provides many time series data. For each one time series data, the first section is training data and is completely free of anomalies, and the second section is test data set that contains exactly one anomaly. The files use a naming convention `id_name_split-number.txt` that provides a split between test and train. For example, there will be an anomaly from 35000 onwards in `001_UCR_Anomaly_35000.txt`. In Figure 1, the first section (yellow part) is training data, and the section (blue part) is test data. The test data contains an anomaly (black part). You need to find the location of this anomaly.

No.	Location of Anomaly
1	120
2	120
3	120
...	...
250	120

Table 1: The format of submission file.

4.1 Submission File

Some algorithms may report the location of the anomaly with different formats, such as the begin, the center, and the end location of the anomaly. In this task, we choose *center* location, and the submission file should contain two columns. As shown in Table 1, the first column is the id of the time series data, and the second column is the center location of each anomaly. Specifically, if you find the begin and end locations of an anomaly are 100 and 200, respectively, you should report 150 for this anomaly.

Note. If your model predicts multi anomalies, you should tweak them and only report one anomaly.

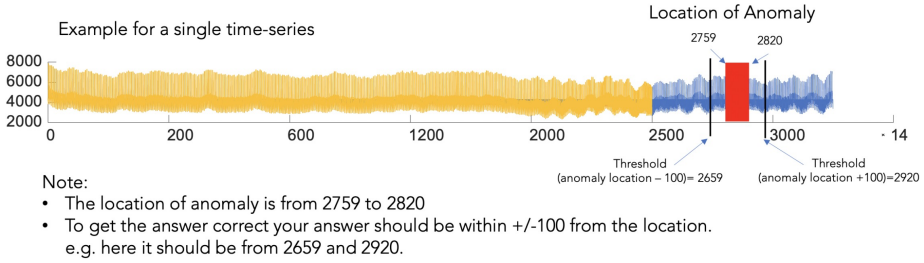

Figure 2: The illustration to evaluation measurement.

4.2 Evaluation Metric

In this project, we evaluate the result as follows. Given n time series data, for each time series data i, we denote the begin and the end location of the anomaly in i as b_i and e_i, respectively. We will count your result is correct if your report result $r_i \in [b_i - 100, e_i + 100]$. The total accuracy is defined as follows.

$$Acc = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_i(r_i \in [b_i - 100, e_i + 100]) \tag{1}$$

where $\mathbb{I}_i(r_i \in [b_i - 100, e_i + 100]) = 1$ if $r_i \in [b_i - 100, e_i + 100]$. More details can be referred to Figure. 2

timestamp	value	label	KPI ID
1476460800	0.0126036806477	0	da10a69f-d836-3baa-ad40-3e548ecf1fbd
1476460860	0.017785768231900003	0	da10a69f-d836-3baa-ad40-3e548ecf1fbd
1476460920	0.012013594431100002	0	da10a69f-d836-3baa-ad40-3e548ecf1fbd
...	...	...	...
1497697020	24.71	0	43115f2a-baeb-3b01-96f7-4ea14188343c
...	...	...	...

Table 2: The format of KPI training data.

5 Description of Another Data Set

The task presented in Section 4 is unsupervised learning. There is another supervised data set, KPI, which is provided by AIOps Challenge. The data set is collected from many real-world Internet companies, such as Tencent, eBay, Alibaba. Currently, KPI has been a benchmark data set for comparing the performance of academic papers.

As shown in Table 2, `timestamp` is the time when value was recorded. You need to classify whether `value` is an anomaly, and `label` indicates its groundtruth (`1` represents anomaly). `KPI ID` identifies the data belongs to a time-series. Note that please report `classification_report` of your model in the report.pdf/.docx if your group conducts the empirical study on this data set.

6 Project Guidelines

The objective of this project is to prompt you to learn how you complete a given project. In other words, what we expect from you is to learn how to solve a problem instead of a well-performed models.

Given any task or project, taking a good survey on literatures is one of the most important steps. You can learn from previous works and implement them to solve your own task. Please pay attention to some acedemic conferences will make it easy for you to solve problems in career. Some conferences related to data mining are listed as:

- **Data Mining:** SIGKDD, Webconf (WWW)
- **Database:** SIGMOD, VLDB, ICDE
- **Artificial Intelligence:** ICML, NeurIPS, ICLR, IJCAI, AAAI

Most of data mining algorithms you learned from MSBD 5002 are published on SIGKDD, SIGMOD and VLDB.

7 Grading Schema

The project grading schema has been listed as:

- **Coding:** (50 marks) Essential comment will be helpful for your grading.
  - Loading data. (10 marks)
  - Data preprocessing. (10 marks)
  - Prediction algoritm. (20 marks)
  - TAs will evaluate your model performance based on Acc presented in Equation (1). (10 marks)
- **Project Report:** (50 marks)
  - Introducing your group members (Name, Student No., ITSC email).
  - Introducing your data engineering in details. (15 marks)
  - Introducing your model in details. (20 marks)
  - Introducing your model's performance. (5 marks)
  - Well-organized and clearly written. (10 marks)
- **Bonus:** (10 marks)
  - Conduct the supervised anomaly detection on KPI data set. (10 marks)

Your program and report should be based on your group independent effort. In case you seek help from any person or reference source, you should state it clearly in your report. Failure to do so is considered plagiarism which will lead to appropriate disciplinary actions.