Time Series Analytics

CS5344 Individual Project

Project Statement

Please complete the following statement to define your project scope:

I am working on [TASK] for [NATURE] Time Series by [APPROACH].

Please select:

- **Task:** Classification / Clustering / Anomaly Detection
- **Nature:** Univariate / Multivariate
- **Approach:** Designing a new algorithm / Improving an existing algorithm

Project Overview

In this individual project, you will explore time series data analytics. You have the option to **choose one** of the following tasks for your project:

1. **Classification**
2. **Clustering**
3. **Anomaly Detection**

Your objective is to design a data mining pipeline to solve the chosen problem. You can either:

- Design your own algorithms, or
- Improve existing algorithms by adding new insights.

Note: You are restricted to using data mining techniques only. Deep learning or neural networks are not allowed.

Background on Time Series (TS) Data

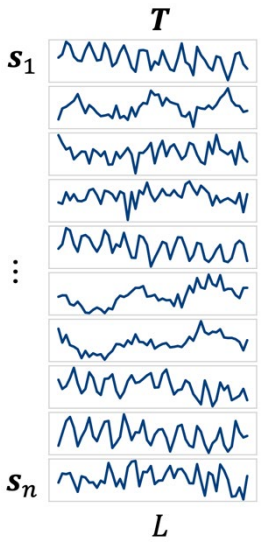
Time series data consists of sequences of data points collected or recorded at specific time intervals. This type of data is prevalent in various domains such as manufacturing, finance, healthcare, environmental studies, and more. Analyzing time series data involves understanding patterns, trends, and other temporal structures within the data, which can be crucial for making informed decisions.

Multivariate vs. Univariate Time Series Data

- **Univariate Time Series:** This type of data consists of a single variable recorded over time. Examples include daily stock prices, temperature recordings, or heart rate measurements.
- **Multivariate Time Series:** This type of data includes multiple variables recorded over time, often simultaneously. Examples include multiple sensors recording different environmental conditions (temperature, humidity, and pressure) or financial data that tracks various economic indicators (interest rates, stock prices, and exchange rates).

Definitions:

- **A Multivariate Time Series (MTS)** $\boldsymbol{T}$ with n series is often represented as a matrix, i.e., $\boldsymbol{T} = (\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n)^{\mathrm{T}}$. Let $|\boldsymbol{T}|$ be the length of the time series of $\boldsymbol{T}$. Then, each series $\boldsymbol{s}_i$ $(1 \leq i \leq n)$ can be denoted as a $|\boldsymbol{T}|$-dimensional vector, i.e., $\boldsymbol{s}_i = (x_{i,1}, \cdots, x_{i,|\boldsymbol{T}|})$ where $x_{i,j}$ $(1 \leq j \leq |\boldsymbol{T}|)$ is a series value of $\boldsymbol{s}_i$ from a single time point t_j. For simplicity, we assume that the time interval between any two consecutive time points t_j and t_{j+1} is the same. L is the length for each series.
- **A Univariate Time Series (UTS)** $\boldsymbol{s}$ can be defined as a vector,. i.e., $\boldsymbol{s} = (x_1, \cdots, x_{|\boldsymbol{s}|})$ where x_j $(1 \leq j \leq |\boldsymbol{s}|)$ is a series value of $\boldsymbol{s}$ from a single time point t_j. L is the length for the series.



Key Differences:

- **Complexity:** Multivariate time series data is generally more complex due to the interaction between multiple variables, which can provide a richer context for analysis but also requires more sophisticated techniques.
- **Correlation:** In multivariate data, correlations between different time series can be leveraged to improve analysis and predictions.
- **Dimensionality:** Multivariate data involves higher dimensionality, which could lead to challenges such as increased computational requirements and the need for dimensionality reduction techniques.

Project Tasks

Task 1: Classification

Objective: Design a pipeline to classify time series data into predefined categories.

Real-world Applications:

- **Activity Recognition:** Identifying different types of physical activities (e.g., walking, running, cycling) using wearable sensor data.
- **Finance:** Categorizing financial market conditions (e.g., bull, bear) based on stock price movements.

Task 2: Clustering

Objective: Create a pipeline to group similar time series data points together without predefined labels.

Real-world Applications:

- **Climate Analysis:** Grouping weather patterns to identify similar climate zones or detect climate change indicators.
- **Customer Segmentation:** Grouping customers based on transaction behaviors over time to tailor marketing strategies.
- **Healthcare:** Clustering patient health metrics to identify common health conditions or responses to treatments.

Task 3: Anomaly Detection

Objective: Develop a pipeline to detect anomalies in time series data.

Real-world Applications:

- **Fraud Detection:** Identifying unusual transaction patterns in financial data that may indicate fraudulent activity.
- **Industrial Monitoring:** Detecting abnormal behavior in sensor data from machinery to prevent equipment failures.
- **Healthcare Monitoring:** Identifying irregularities in patient vital signs that could indicate medical emergencies.

Project Requirements

1. **Algorithm Development**
   - Design and implement a new algorithm for the chosen task, or
   - Improve an existing algorithm by integrating new insights or modifications.
2. **Performance Evaluation**
   - Compare your method against the current state-of-the-art data mining method if you design a new method, or
   - Compare your improved method against the original version if you enhance an existing algorithm.

Step-by-Step Guide

1. **Select a Task**
   - Choose one task from Classification, Clustering, or Anomaly Detection.
2. **Decide on Working with Multivariate or Univariate Data**
   - Determine if you will work with multiple variables (MTS) or a single variable (UTS).
3. **Select a Problem**
   - Decide whether you will design a new algorithm or improve an existing algorithm.
4. **Identify Available Datasets**
   - Collect at least 10 datasets relevant to your task.
5. **Survey on the Evaluation**
   - Identify evaluation metrics for your task such as:
     - **Effectiveness:** Accuracy, precision, recall, F1-score.
     - **Efficiency:** Time complexity, computational cost.
     - **Scalability:** Performance with increasing data size.
6. **Work on Your Own Pipeline and Methods**
   - Design and implement your data mining pipeline and algorithm/improvement.
7. **Benchmark on the Selected Datasets**
   - Run your pipeline on the selected datasets.
   - Collect and compare the performance results.
8. **Results Analysis**
   - Analyze the results to understand the strengths and weaknesses of your method.
   - Discuss the implications of your findings and suggest possible improvements.

Useful Resources

Example Datasets: https://www.aeon-toolkit.org/en/latest/examples/datasets/load_data_from_web.html

Classification and Clustering: https://www.timeseriesclassification.com/dataset.php

Anomaly Detection: https://timeeval.github.io/evaluation-paper/notebooks/Datasets.html, https://tods-doc.github.io/ https://pyod.readthedocs.io/en/latest/