AIN429 Data Mining Lab Project Proposal

Student Number : 2220765010
Name Surname : Alperen Demirci

Project Title : Pattern Mining in Seismic Time Series Data Using the Stanford Earthquake Dataset (STEAD)

1.Problem

Earthquakes are complex natural phenomena that occur as a result of the sudden release of energy in the Earth's crust. Türkiye is located at the intersection of major tectonic plates, including the Anatolian, Eurasian, and Arabian plates, which makes it one of the most seismically active countries in the world. The devastating Kahramanmaraş earthquakes on February 6, 2023, deeply impacted millions of people [1], including myself, and demonstrated once again how critical it is to understand seismic behavior and identify early indicators of destructive earthquakes.

Traditional earthquake prediction remains extremely challenging because seismic signals are highly nonlinear, noisy, and influenced by various geological conditions. However, with the increasing availability of large-scale seismic datasets and modern data mining techniques, it is now possible to extract meaningful temporal patterns and relationships that were previously hidden.

This project aims to apply **pattern discovery and time series mining methods** on the Stanford Earthquake Dataset (STEAD) to uncover repeating waveform structures and trends related to earthquake magnitudes, depths, and wave phases. The main objective is not to predict earthquakes directly but to identify recurring signal patterns that could enhance our understanding of how earthquakes behave in specific geological contexts. The findings may contribute to building better monitoring or early-warning frameworks in the future.

2. Data

The dataset that will be used in this project is the **Stanford Earthquake Dataset (STEAD)** developed by [2]. It is one of the largest and most comprehensive global seismic datasets, containing over **1.2 million three-component seismograms** recorded between 1984 and 2018. The dataset includes records from more than 2,600 seismic stations distributed around the world. Each sample consists of 60 seconds of waveform data recorded at 100 Hz sampling rate, along with a rich set of metadata such as event magnitude, depth, distance from the station, latitude, longitude, and phase arrival times.

STEAD provides both **earthquake** and **non-earthquake (noise)** signals, which makes it suitable for both supervised and unsupervised analysis. The metadata allows filtering based on magnitude range, geographic region, or station type. In this project, I will focus on a subset of the data that represents **earthquake events in or near Türkiye** or regions with similar tectonic characteristics. By doing so, the analysis will be more contextually relevant to local seismic behavior.

The dataset's quality and diversity make it ideal for pattern discovery tasks. Since it contains waveform time series of various magnitudes and distances, it can reveal how signal structures change under different conditions. Additionally, the dataset has been preprocessed for noise reduction and labeled with verified metadata, which makes it easier to integrate into the KDD process.

3. Methodology

This project will follow the **Knowledge Discovery in Databases (KDD)** process, which includes data selection, cleaning, transformation, mining, and interpretation.

1. **Data Selection:**
   The first step involves selecting a manageable subset of STEAD that contains relevant seismic signals. I will filter the data by geographic region and event type, focusing on shallow and moderate-to-large magnitude earthquakes ($M \geq 4$). This ensures that the selected signals have well-defined seismic phases that can reveal meaningful temporal patterns.

2. **Data Cleaning:**
   Seismic recordings often contain background noise, missing values, or inconsistent metadata. Cleaning will include removing corrupted or incomplete waveform samples, handling missing magnitude or distance fields, and ensuring that all signals have a consistent sampling rate and duration.

3. **Data Transformation:**
   After cleaning, the waveform data will be transformed into a format suitable for pattern mining. Time series can be converted into symbolic representations using **Symbolic Aggregate Approximation (SAX)** or **Piecewise Aggregate Approximation (PAA)** or there exists other methods like **Matrix Profiling** for similarity matching. These methods simplify continuous time series into discrete symbolic sequences, making it easier to identify frequent subsequences or motifs. In addition, statistical and frequency-based features such as energy, mean amplitude, zero-crossing rate, and spectral entropy will be extracted for clustering and visualization.

4. **Data Mining:**
   The transformed data will be analyzed using **sequential pattern mining** and **clustering** techniques. Algorithms like **STUMP[3]** or **SPADE[4]** will be used to discover frequent subsequences that appear across multiple earthquake events. These patterns may correspond to repeating waveform shapes or seismic phase transitions. To complement this, **Dynamic Time Warping (DTW)[5]**-based K-means clustering will be applied to group signals with similar temporal shapes. This combination of symbolic mining and similarity-based clustering will help reveal structural patterns in seismic signals.
   Additionally, **Apache Kafka** can be integrated into the system to simulate the real-time streaming of seismic data. Kafka producers can publish waveform data to a topic, while consumers can process incoming signals and check for known patterns as they appear. This demonstrates how pattern mining results could be operationalized for early warning systems.

5. **Interpretation and Evaluation:**
The final stage will focus on interpreting the discovered patterns and evaluating their significance. Visualization techniques such as waveform overlays, cluster heatmaps, and frequency plots will be used to illustrate recurring structures. I will analyze whether certain motifs correspond to specific magnitude ranges, depths, or seismic phases (P-wave, S-wave). The project will conclude with a discussion on how these insights can improve understanding of seismic behavior and how they might contribute to future prediction or detection frameworks.

4. Plan

Week	Task
Week 1	Literature review and understanding STEAD dataset structure
Week 2	Data selection and preprocessing (cleaning and transformation)
Week 3	Feature extraction(must) and SAX transformation(optional)
Week 4	Apply sequential pattern mining algorithms
Week 5	Perform clustering and visualize waveform groups
Week 6	Integrate Kafka-based streaming simulation (optional)
Week 7	Analyze results, prepare report and presentation

5. References

1. Zhe Qu, Feijian Wang, Xiangzhao Chen, Xiaoting Wang, Zhiguang Zhou, Rapid report of seismic damage to hospitals in the 2023 Turkey earthquake sequences, Earthquake Research Advances, Volume 3, Issue 4, 2023, 100234, ISSN 2772-4670, *https://doi.org/10.1016/j.eqrea.2023.100234*

2. Mousavi, S. M., Sheng, Y., Zhu, W., & Beroza, G. C. (2019). *STanford EArthquake Dataset (STEAD): A global data set of seismic signals for AI*. IEEE Access, 7, 179464–179475. *https://doi.org/10.1109/ACCESS.2019.2947848*

3. S.M. Law, (2019). *STUMPY: A Powerful and Scalable Python Library for Time Series Data Mining*. Journal of Open Source Software, 4(39), 1504.

4. Zaki, M.J. SPADE: An Efficient Algorithm for Mining Frequent Sequences. Machine Learning 42, 31–60 (2001). *https://doi.org/10.1023/A:1007652502315*

5. Dynamic Time Warping. In: Information Retrieval for Music and Motion. Springer, Berlin, Heidelberg. (2007) https://doi.org/10.1007/978-3-540-74048-3_4