# THESIS INFORMATION

Title: **Detecting change points and the most unusual subsequence in time series**


Major Computer Science

Major Code: 62.48.01.01

PhD student: Huỳnh Thị Thu Thủy

Advisors: Assoc. Prof. Dương Tuấn Anh

Dr. Võ Thị Ngọc Châu

University: Ho Chi Minh City University of Technology, VNU-HCMC


## 1. ABSTRACT


Time series data is present in a large number of various practical domains such as medicine, finance, e-commerce, hydrometeorology, etc. Data mining and knowledge discovery from time series data are increasingly interested; where anomaly detection in time series is a challenging problem that needs to be considered. The challenges of this problem are listed as follows: (1). pre-specifying the anomaly subsequence's length, (2). the accuracy and time efficiency of finding the most unusual subsequences, and especially, (3). dealing with large time series and instant response requirements. Currently, the research works on the problem of detecting the most unusual subsequences in time series have not yet solved the above challenges effectively and efficiently in a practical context, and especially in the context of streaming time series.

Therefore, the aim of this thesis is to develop effective and efficient solving methods to the problem of finding the most unusual subsequences in time series which can handle nicely the above challenges. The approach of our solving methods is based on change points to segment time series and detect the most unusual subsequence based on the extracted segments.

Experiments show that the segmentation-based approach helps to accurately identify the most unusual subsequence with high computational efficiency and without pre-defining the length of the most unusual subsequence. In addition, the segmentation-based approach

can effectively and efficiently detect the most unusual subsequence even on streaming time series. The thesis brings out 09 published articles, among which 02 articles are in the list of prestigious international journals (1 article in SCIE-Q2 and 1 article in SCIE-Q4), and 07 articles are in the proceedings of international scientific conferences.

## 2. MAIN CONTRIBUTIONS

- **Proposing three improved versions (I-HOTSAX, Hash_DD, and KBF_GPU)** that significantly improve the performance of the two well-known time series anomaly detection algorithms (Brute-Force and HOT SAX) which are based on sliding window. Particularly, the KBF_GPU algorithm is based on GPU technology, brings out high execution efficiency, and helps to adapt to very large time series data.
- **Proposing four new algorithms (EP-ILeader, EP-ALeader, EP-Leader-DTW, and SEP-Leader-DTW)** that help to effectively detect the most unusual subsequence in static and streaming time series using segmentation-based approach under Euclidean measure and Dynamic Time Warping distance.
- **Proposing three new algorithms (TopK-EP-ILeader, TopK-EP-ALeader, and TopK-EP-Aleader-S)** which can detect top-*k* discords in static and streaming time series using segmentation-based approach.
- **Proposing new approach named EPL_S_X** which can apply the result of anomaly detection to improve the performance of forecasting time series.
- **Proposing two improvement clustering algorithms named I-Leader and A-Leader** that can efficiently cluster time series subsequences incrementaly.
- **Proposing new measure named PALS** for evaluating the quality of time series segmentation methods.

## 3. ISSUES FOR FUTURE RESEARCH

The research results in the thesis have solved all original research objectives. However, the thesis still has some research directions for further development:

- Extend the KBF_GPU algorithm so that it can work with DTW distance and find motif (most repeated subsequence) in time series.
- Research and apply distributed programming technique based on Apache Spark or Map Reduce to KBF algorithm.
- Extend EPL_S_*X* approach for forecasting streaming time series in many application domains that require real-time forecasting.
- Apply EPL_S_*X* approach to time series forecasting which uses deep neural networks.
- Extend EP-Leader-DTW algorithm to detect anomalous images in the image database where the images have been converted to time series. This application is expected to use rotation-invariant DTW distance. Besides, the EP-Leader-DTW algorithm will also be improved by using the suite of techniques for DTW distance named UCR-DTW to further speed up the DTW distance calculation.

| ADVISOR 1 | ADVISOR 2 | Ph.D. STUDENT |
|---|---|---|
| Assoc. Prof. Dương Tuấn Anh | Dr. Võ Thị Ngọc Châu | Huỳnh Thị Thu Thủy |