ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA

HUỲNH THỊ THU THỦY

PHÁT HIỆN NHỮNG ĐIỂM THAY ĐỔI VÀ CHUỖI CON BẤT THƯỜNG TRÊN DỮ LIỆU CHUỖI THỜI GIAN

Chuyên ngành: Khoa học máy tính
Mã số chuyên ngành: 62.48.01.01

TÓM TẮT LUẬN ÁN TIẾN SĨ

TP. HỒ CHÍ MINH - NĂM 2021

Công trình được hoàn thành tại **Trường Đại học Bách Khoa – ĐHQG-HCM**

Người hướng dẫn 1: DƯƠNG TUẤN ANH, Phó giáo sư, Tiến sĩ
Người hướng dẫn 2: VÕ THỊ NGỌC CHÂU, Tiến sĩ

Phản biện độc lập 1:
Phản biện độc lập 2:

Phản biện 1:
Phản biện 2:
Phản biện 3:

Luận án sẽ được bảo vệ trước Hội đồng đánh giá luận án họp tại
.....
.....
vào lúc giờ ngày tháng năm

Có thể tìm hiểu luận án tại thư viện:

- Thư viện Trường Đại học Bách Khoa – ĐHQG-HCM
- Thư viện Đại học Quốc gia Tp.HCM
- Thư viện Khoa học Tổng hợp Tp.HCM

Tạp chí quốc tế

1- [**CT01**] H. T. T. Thuy, D. T. Anh and V. T. N. Chau, "Anomaly repair-based approach to improve time series forecasting, "Intelligent Data Analysis, vol. 26, no. 2, pp. xxxx, 2022. *ISI Q4, SCIE, IF = 0.860* (*Unpublished Proceeding Paper*).

2- [**CT02**] H. T. T. Thuy, D. T. Anh and V. T. N. Chau, "Efficient segmentation-based methods in static and streaming time series under dynamic time warping," Journal of Intelligent Information Systems, vol. 56, no. 1, pp.121-146, 2021. *ISI Q2, SCIE, IF = 1.813*

Kỷ yếu hội nghị quốc tế

1- [**CT03**] H.T.T. Thuy, D.T. Anh, and V.T.N. Chau, "A new discord definition and an efficient time series discord detection method using GPUs," In 2021 3rd International Conference on Software Engineering and Development (ICSED), Xiamen, China, 19-21 November, pp. 63-70, 2021.

2- [**CT04**] H.T.T. Thuy, D.T. Anh, and V.T.N. Chau, "Segmentation-based methods for top-k discords detection in static and streaming time series under Euclidean distance," In International Conference on Context-Aware Systems and Applications (ICCASA), 28-29 October, pp. 147-163, 2021. Springer, Cham.

3- [**CT05**] H.T.T. Thuy, D.T. Anh, and V.T.N. Chau, "Incremental Clustering for Time Series Data Based on an Improved Leader Algorithm," In 2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF), Da Nang, Vietnam, 20 March, , pp. 1-6, 2019. IEEE.

4- [**CT06**] H.T.T. Thuy, D.T. Anh and V.T.N. Chau, "A Novel Method for Time Series Anomaly Detection based on Segmentation and Clustering," In 2018 10th International Conference on Knowledge and Systems Engineering (KSE), Ho Chi Minh, Vietnam, 1-3 November, pp. 276-281, 2018. IEEE.

5- [**CT07**] H. T. T. Thuy, D. T. Anh and V. T. N. Chau, "Comparing Three Time Series Segmentation Methods via Novel Evaluation Criteria," In 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, 1-3 November, pp. 171-176, 2017.

6- [**CT08**] H.T.T. Thuy, D.T. Anh and V.T.N. Chau, "An effective and efficient hash-based algorithm for time series discord discovery," In 2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS), Da Nang, Vietnam, 14-16 September, pp. 85-90, 2016. IEEE.

7- [**CT09**] H.T.T. Thuy, D.T. Anh and V.T.N. Chau, "Some Efficient Segmentation-Based Techniques to Improve Time Series Discord Discovery," In International Conference on Nature of Computation and Communication (ICTCC), Kien Giang, Vietnam, 17-18 March, pp. 179-188, 2016. Springer, Cham.

CHƯƠNG 1 GIỚI THIỆU

1.1 Động cơ nghiên cứu của đề tài

Giới thiệu ngữ cảnh

Bài toán được giải quyết trên ngữ cảnh dữ liệu chuỗi thời gian dạng tĩnh và ngữ cảnh dữ liệu chuỗi thời gian dạng luồng.

Giới thiệu bài toán

Bài toán cần nghiên cứu là bài toán *phát hiện chuỗi con bất thường nhất trên dữ liệu chuỗi thời gian* (time series data).

Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài là phát hiện bất thường trên *dữ liệu chuỗi thời gian dạng tĩnh* (static time series) và *chuỗi thời gian dạng luồng* (streaming time series). Trong đó, giá trị từng điểm dữ liệu của chuỗi thời gian là số thực như được trình bày trong Định nghĩa 2.1.

Phạm vi nghiên cứu của đề tài là giải bài toán phát hiện chuỗi con bất thường nhất trên dữ liệu chuỗi thời gian. Đề tài sẽ tập trung vào *chuỗi thời gian đơn biến* (univariate time series) và chuỗi con tìm được hướng về chuỗi con ngắn nhất.

Giới thiệu các công trình liên quan và hiện trạng giải quyết

Thách thức nổi trội đối với bài toán khám phá bất thường là tìm ra chiều dài chuỗi con bất thường phù hợp. Các công trình nổi tiếng như [6], [7], [37] cũng chưa thể vượt qua thách thức này. Kể cả các công trình mới gần đây của năm 2019 [38] và năm 2021 [39] cũng chưa vượt qua được thách thức này. Ngoài ra, có một thách thức khác xuất hiện gần đây hơn, đó là việc phát hiện chuỗi con bất thường trên dữ liệu chuỗi thời gian có kích thước lớn và dữ liệu chuỗi thời gian dạng luồng [40]. Năm 2018, Châu và các cộng sự đã đề xuất giải pháp HS-Squeezer-Stream sử dụng hướng tiếp cận dựa vào cửa sổ trượt để phát hiện bất thường trên dữ liệu chuỗi thời gian dạng luồng nhưng đề xuất này vẫn có chi phí thời gian cao [41].

Dựa vào công trình của Chandola và các cộng sự năm 2009 [22], công trình của Cheboli năm 2010 [23], và công trình của Braei và Wagner năm 2020 [4], luận án đúc kết lại có 4 hướng tiếp cận để giải bài toán phát hiện bất thường như sau:

1- *Hướng tiếp cận dựa vào cửa sổ trượt* (Window – based): Nhược điểm là cần phải xác định trước chiều dài của chuỗi con bất thường nhất cần tìm. Đây cũng chính là chiều dài của cửa sổ trượt.

2- *Hướng tiếp cận dựa vào dự báo* (Prediction – based): Khó khăn thứ nhất là cần xác định chiều dài lịch sử của dữ liệu để dự báo. Việc xác định chiều dài lịch sử dữ liệu không phù hợp sẽ ảnh hưởng đến kết quả dự báo. Khó khăn thứ hai là việc xác định giá trị ngưỡng để có thể kết luận dữ liệu là bất thường hay không. Nếu giá trị ngưỡng quá nhỏ thì số lượng *dương sai* (false positive, lỗi loại I) sẽ lớn. Nếu giá trị ngưỡng quá lớn thì số lượng *âm sai* (false negative, lỗi loại II) cũng sẽ lớn.

3- *Hướng tiếp cận dựa vào phân lớp* (Classification – based): Các chuỗi con trên chuỗi thời gian sẽ được chia thành hai lớp là bình thường hay bất thường. Đây là hướng tiếp cận học có giám sát. Nhược điểm của hướng tiếp cận dựa vào phân lớp là dữ liệu cần phải được gán nhãn trước là bình thường hay bất thường. Tuy nhiên, điều này không dễ có được hoặc nếu có thì chi phí quá đắt [5].

4- *Hướng tiếp cận dựa vào phân đoạn* (Segmentation - based): Trước hết chuỗi thời gian sẽ được chia thành các *phân đoạn* (segment). Sau đó, sử dụng một số kỹ thuật phát hiện bất thường phù hợp để xác định đoạn bất thường nhất trong các đoạn này. Điểm khó của hướng tiếp cận dựa vào phân đoạn là làm thế nào phân đoạn chuỗi thời gian cho hiệu quả. Sau đó, từ các phân đoạn được phân chia này, sử dụng các bước tiếp theo nào phù hợp để tìm ra phân đoạn bất thường nhất. Khó khăn này được khắc phục dễ dàng với các giải thuật phân đoạn chuỗi thời gian hiệu quả [25], [26], [27]. Với hiệu quả của các giải thuật phân đoạn chuỗi thời gian, việc phát hiện bất thường theo hướng phân đoạn sẽ hữu hiệu.

Kết quả nghiên cứu trên đã dẫn đến định hướng nghiên cứu của luận án là áp dụng hướng tiếp cận dựa vào phân đoạn nhằm khắc phục các thách thức của bài toán phát hiện chuỗi con bất thường nhất trên dữ liệu chuỗi thời gian.

1.2 Ý nghĩa khoa học và ý nghĩa thực tiễn của đề tài

Ý nghĩa khoa học của đề tài nghiên cứu

Đóng góp của luận án là tìm giải pháp hiệu quả cho bài toán phát hiện chuỗi con bất thường trên dữ liệu chuỗi thời gian và đặc biệt là dữ liệu chuỗi thời gian có kích thước lớn cũng như dữ liệu chuỗi thời gian dạng luồng. Giải pháp được chọn theo hướng tiếp cận mới từ các kết quả đạt được của *bài toán tìm các điểm thay đổi* và *không yêu cầu người dùng xác định chiều dài chuỗi con bất thường*. Một đóng góp quan trọng nữa của luận án là hỗ trợ cho bài toán khai phá dữ liệu chuỗi

thời gian khác như: bài toán *dự báo* (forecasting), bài toán *làm sạch dữ liệu* (data cleaning).

Ý nghĩa thực tiễn của đề tài nghiên cứu

Bài toán phát hiện bất thường có các ứng dụng như: phát hiện nhịp tim bất thường [48], [49]; tìm kiếm các hình dạng không bình thường trong cơ sở dữ liệu hình ảnh lớn [50]; sử dụng trong hệ thống giám sát mực nước của một đập thủy điện [51], sử dụng trong hệ thống giám sát lượng dữ liệu lưu thông trên *mạng dữ liệu* (data network) [52].

Tóm lại, phát hiện chuỗi con bất thường trên dữ liệu chuỗi thời gian được ứng dụng phổ biến trong nhiều lĩnh vực: tài chính, kinh tế [13], [14], giải trí, nghệ thuật, khoa học và kỹ thuật [50], [53], [54], y khoa [48], [49], thời tiết [9], [55], [11], [46], khí tượng thủy văn [51], [10], [12], [47], môi trường [56], giám sát mạng dữ liệu [52].

1.3 Mục tiêu và nhiệm vụ nghiên cứu

Luận án đề ra 2 mục tiêu chính:

- *Đề xuất giải pháp mới để phát hiện hiệu quả chuỗi con bất thường nhất trên dữ liệu chuỗi thời gian **dạng tĩnh**.*
- *Đề xuất giải pháp mới để phát hiện hiệu quả chuỗi con bất thường nhất trên dữ liệu chuỗi thời gian **dạng luồng** (còn được gọi là xử lý online).*

1.4 Những đóng góp của luận án

- **Chương 3**: Đề xuất mới độ đo *PALS* (Percentage of Average Length Segments): Đánh giá các phương pháp phát hiện các điểm thay đổi (phương pháp phân đoạn). [CT07]

- **Chương 4**: Trình bày 03 đề xuất cải tiến các phương pháp phát hiện bất thường dựa vào hướng tiếp cận cửa sổ trượt gồm:

i- Đề xuất cải tiến giải thuật *I-HOTSAX* (Improved - HOT SAX): Giảm độ khó cho việc thiết lập tham số và tăng tốc giải thuật HOT SAX phát hiện chuỗi con bất thường trên chuỗi thời gian. [CT09]

ii- Đề xuất cải tiến giải thuật *Hash_DD* (Hash-based algorithm for Time series Discord Discovery): Sử dụng bảng băm nhằm cải thiện chi phí bộ nhớ và tăng tốc giải thuật HOT SAX. [CT08]

iii- Đề xuất cải tiến giải thuật *KBF_GPU*: Sử dụng kỹ thuật lập trình song song nhằm tăng tốc giải thuật KBF - một cải biên của Brute-Force nhằm phát hiện chuỗi con bất thường khi có sự xuất hiện của *bất thường đôi* (twin freak). [CT03]

- **Chương 5**: Trình bày 01 đề xuất cải tiến giải thuật gom cụm hỗ trợ cho bài toán phát hiện bất thường và 03 đề xuất mới các phương pháp phát hiện bất thường dựa vào phân đoạn trên chuỗi thời gian dạng tĩnh và dạng luồng với độ đo Euclid gồm:

i- Đề xuất giải thuật *I-Leader*: Một cải tiến từ giải thuật gom cụm Leader cho bài toán gom cụm các chuỗi con. [CT05]

ii- Đề xuất mới giải thuật *EP-ILeader*: Phát hiện chuỗi con bất thường trên chuỗi thời gian dạng tĩnh theo hướng tiếp cận dựa vào phân đoạn. [CT06]

iii- Đề xuất mới giải thuật *TopK-EP-ALeader*: Phát hiện *k* chuỗi con bất thường nhất trên chuỗi thời gian dạng tĩnh và theo hướng phân đoạn. [CT04]

iv- Đề xuất mới giải thuật *TopK-EP-ALeader-S*: phát hiện *k* chuỗi con bất thường nhất trên chuỗi thời gian dạng luồng và theo hướng phân đoạn. [CT04]

- **Chương 6**: Trình bày 02 đề xuất mới phát hiện bất thường dựa vào phân đoạn trên chuỗi thời gian dạng tĩnh và dạng luồng với độ đo DTW gồm:

i- Đề xuất mới giải thuật *EP-Leader-DTW*: Phát hiện chuỗi con bất thường trên chuỗi thời gian dạng tĩnh với độ đo DTW và theo hướng phân đoạn. [CT02]

ii- Đề xuất mới giải thuật *SEP-Leader-DTW*: Phát hiện chuỗi con bất thường trên chuỗi thời gian dạng luồng với độ đo DTW và theo hướng phân đoạn. [CT02]

- **Chương 7**: Trình bày đề xuất mới hướng tiếp cận *EPL_S_X:* Cải thiện chất lượng dự báo cho các phương pháp dự báo dữ liệu chuỗi thời gian dựa vào phát hiện bất thường và khử bất thường. [CT01]

CHƯƠNG 2 CƠ SỞ LÝ THUYẾT VÀ CÁC CÔNG TRÌNH LIÊN QUAN

2.1 Định nghĩa

2.1.1 Định nghĩa 2.1. Dữ liệu chuỗi thời gian

Một chuỗi T có thứ tự của m biến giá trị thực [1] và được ghi nhận tại các thời điểm đều nhau theo thời gian được gọi là *dữ liệu chuỗi thời gian* (time series data) [2].

$$T = \{t(i) \mid i = 1 \ldots m\},\ t_i \in R$$

2.1.2 Định nghĩa 2.2. Dữ liệu chuỗi thời gian dạng luồng

Một chuỗi các quan sát $T = \{t(i) \mid i = 1 \ldots \infty\}$ được ghi nhận ở nhiều thời điểm khác nhau, cách đều nhau và đến liên tục theo thứ tự thời gian được gọi là *chuỗi thời gian dạng luồng* [21].

2.1.3 Định nghĩa 2.3. Chuỗi con

Cho một *chuỗi thời gian T* có chiều dài *m, một chuỗi con* (Subsequence) *S* của *T* là một mẫu gồm *n* vị trí liên tục được lấy từ *T* với $n < m$.

Khi đó $S = t_p, \ldots, t_{p+n-1}$ với $1 \leq p \leq m - n + 1$.

2.1.4 Định nghĩa 2.4. Trùng khớp không tầm thường

Cho chuỗi thời gian *T* chứa chuỗi con *C* bắt đầu ở vị trí *p* với chiều dài *n* và một chuỗi con trùng khớp *M* bắt đầu ở vị trí *q*, ta nói *M* là một *trùng khớp không tầm thường* (non-self match) của *C* nếu như $|p - q| \geq n$.

2.1.5 Định nghĩa 2.5. Chuỗi con bất thường nhất

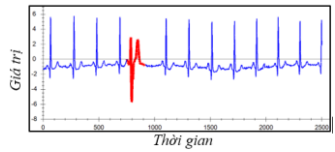
Cho chuỗi thời gian *T*, chuỗi con *C* của *T* được xem là *chuỗi con bất thường nhất* (còn được gọi là *1-discord* hoặc *top-1 discord*) trong *T* nếu như *C* có khoảng cách xa nhất đến chuỗi con trùng khớp không tầm thường của nó (Hình 2.1).

2.1.6 Định nghĩa 2.6. Chuỗi con bất thường thứ k

Cho chuỗi thời gian *T*, một chuỗi con *D* có chiều dài *n* bắt đầu ở vị trí *p* là *chuỗi con bất thường thứ k* (k^{th} *- discord*) trong *T* nếu *D* có khoảng cách lớn thứ *k* đến lân cận trùng khớp không tầm thường của nó và không có sự chồng lên nhau đến chuỗi con bất thường thứ i^{th} bắt đầu ở vị trí thứ p_i, với $1 \leq i \leq k$. Nghĩa là $|p - p_i| \geq n$.

2.1.1 Điểm thay đổi

Định nghĩa 2.7. *Điểm thay đổi* (change point) là điểm mà tại đó tính chất của dữ liệu thay đổi một cách đột ngột.



Hình 2.1: Chuỗi con bất thường nhất (màu đỏ) trên chuỗi thời gian điện tâm đồ - ECG

Định nghĩa 2.8. *Điểm thay đổi* là điểm kết nối giữa hai phân đoạn kế cận.

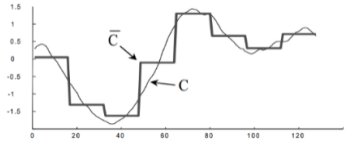
2.2 Thu giảm số chiều

Phương pháp *thu giảm số chiều* (Dimensionality Reduction) là cách thức biểu diễn lại chuỗi thời gian $X = \{x_1, x_2, \ldots, x_m\}$ thành chuỗi dữ liệu $Y = \{y_1, y_2, \ldots, y_k\}$, với k là hệ số biến đổi và $k < m$.

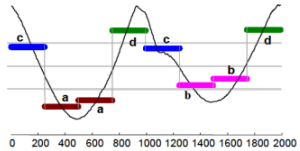
Sau đây là phương pháp thu giảm số chiều được sử dụng trong luận án.

2.2.1 Phương pháp xấp xỉ gộp từng đoạn

Phương pháp xấp xỉ gộp từng đoạn (Piecewise Aggregate Approximation - PAA) tuần tự thực hiện xấp xỉ k điểm giá trị liền kề nhau thành cùng một giá trị trung bình cộng của k điểm đó. Quá trình được thực hiện từ trái sang phải và kết quả cuối cùng ta được một đường dạng bậc thang như Hình 2.2.



Hình 2.2: Phép biến đổi PAA



Hình 2.3: Phương pháp SAX

2.3 Rời rạc hóa dữ liệu

Dữ liệu chuỗi thời gian thường là những dữ liệu liên tục và có chiều dài lớn nên ta cần chia dữ liệu thành những đoạn rời rạc nhỏ hơn và ký hiệu hóa các đoạn này dựa vào đặc trưng của chúng, từ đó giúp cho việc phân tích và tính toán dễ dàng hơn, quá trình này gọi là quá trình *rời rạc hóa* (discretization). Sau đây là phương pháp rời rạc hóa được sử dụng trong luận án.

2.3.1 Phương pháp xấp xỉ gộp ký hiệu hóa

Phương pháp biểu diễn chuỗi thời gian bằng *chuỗi bit* (bit string) chỉ dùng 2 ký tự 0 và 1 để biểu diễn nên thường không thể hiện được hết đặc tính của dữ liệu. Do đó, J. Lin và các cộng sự đã đề xuất phương pháp *xấp xỉ gộp ký hiệu hóa* (Symbolic Aggregate approXimation - SAX) vào năm 2003 để thực hiện rời rạc hóa dữ liệu chuỗi thời gian. Chuỗi thời gian ban đầu được chia thành từng đoạn bằng phương pháp PAA. Sau đó, dựa trên giá trị trung bình cộng của từng đoạn, ta sẽ biểu diễn đặc trưng của đoạn thành các ký tự. Khi đó, chuỗi thời gian ban đầu sẽ được mã hóa rời rạc thành một chuỗi các ký tự như Hình 2.3. Phương pháp này biểu diễn dữ liệu chuỗi thời gian thành dạng chuỗi ký tự nên từ đó có thể áp dụng được các kỹ thuật xử lý trên dữ liệu tràng ký tự để thực hiện xử lý, phân tích dữ liệu chuỗi thời gian.

2.4 Giới thiệu các tập dữ liệu thực nghiệm

Các tập dữ liệu thực nghiệm có được từ các trang web có uy tín:

http://www.cs.ucr.edu/~eamonn/discords/
www.cs.ucr.edu/~eamonn/time_series_data
https://www.physionet.org/

Các tập dữ liệu đại diện phổ biến như: ECG: dữ liệu điện tâm đồ, POWER: dữ liệu tiêu thụ điện năng, Stock: dữ liệu chứng khoán, Chromosome: dữ liệu nhiễm sắc thể, nprs43: dữ liệu nhịp thở, Tek16: dữ liệu cảm biến. Các tập dữ liệu này thuộc các lĩnh vực y học, công nghiệp, tài chính, sinh học.

CHƯƠNG 3 PHÁT HIỆN NHỮNG ĐIỂM THAY ĐỔI TRÊN CHUỖI THỜI GIAN VÀ CÁC PHƯƠNG PHÁP PHÂN ĐOẠN

3.1 Các phương pháp phát hiện các điểm thay đổi

Phương pháp phát hiện các điểm thay đổi trên dữ liệu chuỗi thời gian còn được gọi là phương pháp phân đoạn dữ liệu chuỗi thời gian. Có 3 phương pháp (PP) phát hiện các điểm thay đổi cần được nghiên cứu để lựa chọn hỗ trợ cho bài toán phát hiện chuỗi con bất thường bao gồm:

- PP *điểm cực trị quan trọng* (Important Extrema) của Fink và Gandhi.
- PP *điểm quan trọng cảm nhận được* (Perceptually Important Points-PIP) của Fu và Chung.

- PP *xấp xỉ bình phương tối thiểu đa thức* (Polynomial Least-Squares Approximations - PLSA) của Fuchs và cộng sự.

3.2 Đề xuất tiêu chí đánh giá các phương pháp phân đoạn

3.2.1 Đề xuất độ đo PALS đánh giá chất lượng phương pháp phân đoạn

Định nghĩa: Gọi T là chuỗi thời gian có chiều dài m và S_i là phân đoạn thứ i có chiều dài n_i trong n phân đoạn được trích từ T:

$$PALS = \frac{\left|\{S_i \mid \overline{n_s} - \delta \le n_i \le \overline{n_s} + \delta, \forall i = 1..n\}\right|}{n} \quad (3.1)$$

trong đó:

$\overline{n_s} = \frac{\sum_{i=1..n} n_i}{n}$ và δ là ngưỡng xấp xỉ cho chiều dài trung bình của các phân đoạn.

Độ đo này có giả định là độ dài của các phân đoạn được rút trích tuân theo phân phối Gauss và kỳ vọng rằng một phương pháp phân đoạn tốt hơn sẽ đạt được số phân đoạn có độ dài xấp xỉ độ dài trung bình của tất cả các phân đoạn cao hơn. Ưu điểm của độ đo PALS so với các độ đo đã có là: *không cần dữ liệu phải có đánh dấu của chuyên gia*.

3.2.2 Đánh giá độ đo PALS

Độ đo PALS có thể được áp dụng để so sánh chất lượng của các phương pháp phân đoạn.

CHƯƠNG 4 CẢI TIẾN CÁC PHƯƠNG PHÁP PHÁT HIỆN CHUỖI CON BẤT THƯỜNG NHẤT DỰA VÀO CỬA SỔ TRƯỢT TRUYỀN THỐNG TRÊN DỮ LIỆU CHUỖI THỜI GIAN DẠNG TĨNH

4.1 Giải thuật cải tiến I-HOTSAX

Giải thuật I-HOTSAX sử dụng 3 kỹ thuật ước lượng tham số và một đóng góp mới như sau:

- *Ước lượng kích thước khung PAA* dựa vào kỹ thuật phân đoạn PLA.

- *Ước lượng chiều dài chuỗi con bất thường và chiều dài từ SAX* dựa vào các điểm cực trị quan trọng.

- *Cách trượt cửa sổ trượt mới* là trượt cửa sổ qua từng đoạn PAA thay vì cách trượt cửa sổ truyền thống là trượt cửa sổ qua từng điểm dữ liệu.

- *Đóng góp mới khác: Cách tính khoảng cách giữa các chuỗi con trong I-HOTSAX:* Trong giải thuật HOT SAX, việc chuyển đổi các chuỗi con thành các từ SAX ngụ ý cho việc sử dụng khoảng cách MINDIST giữa hai từ SAX, được đưa ra trong giải thuật [76]. Bên cạnh cách tính khoảng cách giữa hai từ SAX, còn cách tính khác là tham chiếu trở lại hai chuỗi con trên chuỗi thời gian ban đầu tương ứng với hai từ SAX và tính khoảng cách Euclid giữa hai chuỗi con này. Nhờ cách tính khoảng cách Euclid giữa hai chuỗi con, I-HOTSAX đạt được độ chính xác tốt hơn so với HOT SAX dùng cách tính thứ nhất.

4.2 Giải thuật cải tiến Hash_DD

Giải thuật phát hiện chuỗi con bất thường Hash_DD (Hash-based algorithm for Time series Discord Discovery) được cải tiến so với HOT SAX ở các điểm sau:

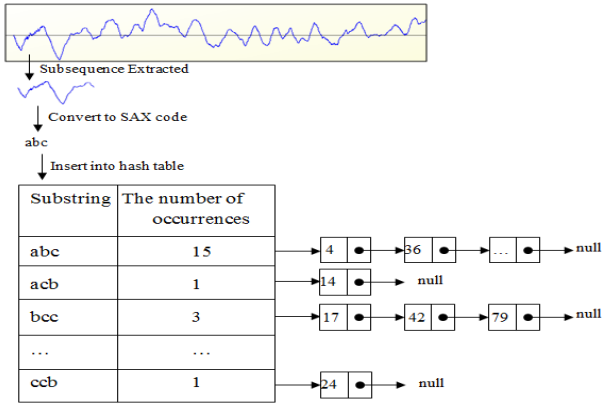
- *Kế thừa các đặc điểm từ I-HOTSAX*: tự động ước lượng chiều dài các chuỗi con và chiều dài từ SAX, trượt cửa sổ qua từng đoạn PAA thay cho cách trượt cửa sổ truyền thông là qua từng điểm, sử dụng cách tính khoảng cách Euclid giữa hai chuỗi con thay cho khoảng cách MINDIST.

- Sử dụng cấu trúc bảng băm (hình 4.1) thay cho *cây gia tố* (augment trie).

Với giải thuật Hash_DD, các chuỗi con dạng tràng kí tự SAX được băm vào bảng băm. Những chuỗi con giống nhau được băm vào cùng một thùng trong bảng băm. Mỗi thùng trong bảng băm sẽ chứa chuỗi con dạng *tràng kí tự* (còn gọi là các từ *SAX*) và số lần xuất hiện của chuỗi con trong thùng băm. Hai vòng lặp có heuristic của Hash_DD hoạt động như sau:

- *Vòng lặp ngoài*: Sau khi bảng băm đã được xây dựng, các chuỗi con có số lần xuất hiện *nhỏ nhất* sẽ được xem xét ở vòng lặp ngoài và các chuỗi con có số lần xuất hiện lớn hơn sẽ không được xem xét ở vòng lặp ngoài.
- *Vòng lặp trong*: Thứ tự của các chuỗi con ở vòng lặp trong chỉ là thứ tự của các chuỗi con mà chúng được tìm thấy khi duyệt qua các thùng trong bảng băm theo thứ tự tăng dần của số lần xuất hiện chuỗi con. Bên cạnh đó, ở vòng lặp trong, thực sự không cần phải tìm lân cận thực sự gần nhất với chuỗi con

ứng viên hiện tại. Ngay khi tìm thấy bất kỳ chuỗi con nào gần giống với chuỗi con ứng viên hiện tại hơn so với giá trị của *best_so_far_dist* (nghĩa là chuỗi con này không có cơ hội trở thành chuỗi con bất thường cần tìm), vòng lặp trong có thể được kết thúc sớm, điều này là an toàn vì chuỗi con ứng viên hiện tại không thể là một chuỗi con bất thường.



Hình 4.1: Cấu trúc bảng băm hỗ trợ vòng lặp trong và vòng lặp ngoài của Hash_DD

4.3 Giải thuật cải tiến KBF_GPU

Định nghĩa 4.1: *Khoảng cách K* (K-distance): Cho một số dương *K*, khoảng cách *K* của chuỗi con S_p, kí hiệu là *K-dist(Sp)* được định nghĩa là tổng các khoảng cách từ chuỗi con S_p đến *K* lân cận trùng khớp không tầm thường của nó.

Định nghĩa 4.2: *Chuỗi con bất thường theo khoảng cách K* (K-distance discord): Cho một chuỗi thời gian *T*, chuỗi con S_d có chiều dài *n* bắt đầu ở vị trí *d* được gọi là chuỗi con bất thường nhất của *T* nếu S_d có khoảng cách *K* lớn nhất trong số các khoảng cách *K* của tất cả các chuỗi con của *T*. Nghĩa là, bất kỳ chuỗi con S_c có chiều dài *n* bắt đầu ở vị trí *c* trong *T*, $|d - c| \geq n$, $K\text{-}dist(S_d) \geq K\text{-}dist(S_c)$.

Tìm kiếm chuỗi con bất thường theo khoảng cách K:

Theo tinh thần của giải thuật Brute-Force [7], để tìm chuỗi con bất thường theo khoảng cách *K*, giải thuật Brute-Force được hiệu chỉnh thành giải thuật mới với tên gọi là KBF (Brute-Force for K-distance discord).

Tăng tốc giải thuật KBF với GPU

Phiên bản song song đề xuất cho KBF được đặt tên là KBF_GPU (Brute-Force for K-distance discord using GPU) gồm 3 bước:

- *Bước 1*: CPU chép toàn bộ chuỗi thời gian vào bộ nhớ GPU.

- *Bước 2*: Ứng với mỗi chuỗi con ứng viên C_p tại vị trí *p* của *vòng lặp ngoài* (outer loop), CPU sẽ gọi thực hiện hàm *kernel* trong GPU. Mỗi tiến trình *kernel* sẽ thực hiện cho một chuỗi con ứng viên để tính tất cả các khoảng cách từ chuỗi con ứng viên C_p đến những chuỗi trùng khớp không tầm thường của nó. Tiến trình *kernel* cũng sẽ lưu tất cả các khoảng cách tính được vào một mảng có tên là *List-Dist*. Kết thúc mỗi tiến trình *kernel*, mảng *List-Dist* sẽ được chép trở lại vào CPU.

- *Bước 3*: CPU sẽ xác định *K* khoảng cách từ chuỗi con đang xét đến *K* lân cận trùng khớp không tầm thường của nó và lưu vào mảng Array-K. CPU cũng tính khoảng cách *K* của mỗi chuỗi con dựa vào mảng Array-K và xác định chuỗi con bất thường nhất là chuỗi con có khoảng cách *K* lớn nhất.

Mỗi tiến trình *kernel* được dùng cho một chuỗi con ứng viên C_p tại vị trí *p* trong vòng lặp ngoài. Khi đó sẽ cần $(|T| - n + 1)$ tiến trình *kernel* để thực hiện công việc tính toán chính cho giải thuật KBF_GPU. Vậy, độ phức tạp của giải thuật KBF_GPU là $\boldsymbol{O}(|T|)$.

Ngoài ra, KBF_GPU không cần người sử dụng xác định trước chiều dài của chuỗi con bất thường. Thay vào đó, KBF_GPU sẽ tự động xác định giá trị chiều dài phù hợp cho chuỗi con bất thường dựa vào giải thuật phát hiện điểm cực trị quan trọng. Điều này làm cho KBF_GPU dễ sử dụng hơn so với các phương pháp dựa trên cửa sổ đã được đánh giá trước đây để phát hiện chuỗi con bất thường trên chuỗi thời gian.

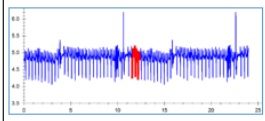
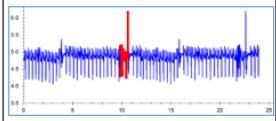
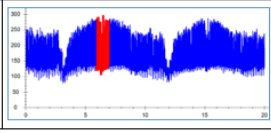
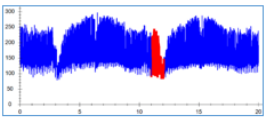
4.4 Đánh giá các giải thuật cải tiến

- Giải thuật I-HOTSAX và Hash_DD phát hiện chuỗi con bất thường chính xác. Giải thuật I-HOTSAX thực thi nhanh gấp **2,8** lần so với HOT SAX và Hash_DD nhanh gấp **8,24** lần so với HOT SAX.

- Giải thuật KBF_GPU:

- *Tính chính xác*: Khi chuỗi thời gian có *bất thường đôi* (twin freak) như trong hình 4.2, KBF_GPU phát hiện chính xác một trong số hai chuỗi con bất thường nhất giống nhau. Trong khi Brute-Force [7] không tìm được chuỗi

nào trong số hai chuỗi con bất thường nhất này. Trong trường hợp chuỗi con bất thường nhất chỉ xuất hiện một lần, chuỗi con bất thường nhất do KBF_GPU và Brute-Force tìm được đều như nhau trên các tập dữ liệu thực nghiệm.

Tập dữ liệu	Brute-Force	KBF_GPU
ECG 24K		
AEM		

Hình 4.2: Chuỗi con bất thường (đoạn màu đỏ) tìm được bởi Brute-Force và KBF_GPU

- *Tính hữu hiệu về thời gian thực thi:* Giải thuật KBF_GPU nhanh gấp 10.216 lần so với giải thuật KBF và KBF_GPU nhanh gấp 28 lần so với giải thuật HOT SAX. Với tập dữ liệu dưới 3.000 điểm, giải thuật KBF_GPU có thể thực thi trong thời gian tính bằng miligiây. Từ đó cho thấy KBF_GPU có tiềm năng áp dụng được cho dữ liệu dạng luồng.

CHƯƠNG 5 ĐỀ XUẤT CÁC PHƯƠNG PHÁP PHÁT HIỆN CHUỖI CON BẤT THƯỜNG NHẤT DỰA VÀO PHÂN ĐOẠN VỚI ĐỘ ĐO EUCLID

5.1 Giải thuật đề xuất cải tiến I-Leader cho bài toán gom cụm các chuỗi con

Giải thuật I-Leader được sử dụng cho bài toán gom cụm gia tăng. Ở đây, ngụ ý của việc gia tăng là các điểm dữ liệu mới đến liên tục và các điểm cũ bị xóa đi. Các điểm dữ liệu được xem xét là các điểm nằm trong vùng đệm xoay vòng - nơi chứa các điểm dữ liệu mới đến thay cho các điểm cũ. Giải thuật gom cụm gia tăng I-Leader cho chuỗi thời gian gồm các ý tưởng mới như sau:

i/. I-Leader sử dụng "*centroid*" thay vì "*leader*" để làm phần tử đại diện cụm.

ii/. I-Leader tính *tâm cụm* (centroid) theo cách tính gia tăng.

iii/. Chất lượng gom cụm được duy trì tốt ở mỗi lần cập nhật cụm bằng cách kiểm tra loại cụm.

Đánh giá giải thuật gom cụm cải tiến I-Leader

- Chất lượng gom cụm: Tốt hơn Leader và *k*-Means.

- Tính hữu hiệu của I-Leader: I-Leader thực thi nhanh hơn Leader và *k*-Means.

5.2 Giải thuật đề xuất mới EP-ILeader cho bài toán phát hiện chuỗi con bất thường trên dữ liệu chuỗi thời gian tĩnh

Ý tưởng chính: Hai giải thuật phổ biến cho khám phá bất thường trên chuỗi thời gian tĩnh là Brute-Force và HOT SAX [7] đều dựa vào cửa sổ trượt. Vì vậy, hai giải thuật này có độ phức tạp thời gian cao. Một cách khác biệt, EP-ILeader (Extreme Points and Improved Leader) thì rất hiệu quả cho bài toán phát hiện bất thường trên chuỗi thời gian dạng tĩnh. EP-ILeader sử dụng phương pháp các điểm cực trị quan trọng và giải thuật gom cụm gia tăng I-Leader. Nghĩa là EP-ILeader làm việc theo hướng tiếp cận phân đoạn và gom cụm. Hướng tiếp cận phân đoạn và gom cụm này không cần tham số chiều dài chuỗi con bất thường.
Ý tưởng chính của hướng tiếp cận này như sau: Trước tiên, chuỗi thời gian sẽ được phân đoạn thành nhiều chuỗi con dựa vào các điểm cực trị quan trọng. Sau đó, sử dụng giải thuật gom cụm để gom các chuỗi con vào các cụm. Tiếp đến, mỗi chuỗi con sẽ được tính hệ số bất thường và cuối cùng chuỗi con nào có hệ số bất thường lớn nhất sẽ là chuỗi con bất thường cần tìm.

Đánh giá giải thuật mới EP-ILeader phát hiện chuỗi con bất thường

-Tính chính xác: Chuỗi con bất thường do EP-ILeader tìm được trùng khớp với chuỗi con bất thường do giải thuật cơ sở Brute-Force tìm được.

-Tính hữu hiệu về thời gian thực thi: EP-ILeader có thể thực thi nhanh gấp **2794** lần so với giải thuật HOT SAX. Hơn nữa, EP-ILeader có thể phát hiện bất thường trên tập dữ liệu hàng trăm ngàn điểm với tốc độ tính bằng mili giây.

5.3 Đề xuất mới giải thuật TopK-EP-ALeader phát hiện *k* chuỗi con bất thường nhất trên chuỗi thời gian dạng tĩnh

5.3.1 Các kỹ thuật hỗ trợ cho giải thuật TopK-EP-ALeader:

- Tăng tốc tính khoảng cách Euclid cho giải thuật gom cụm I-Leader

Sử dụng hai kỹ thuật tăng tốc được lấy cảm hứng từ bộ kỹ thuật tăng tốc UCR-ED được giới thiệu bởi Rakthanmanon và các cộng sự năm 2012 [33]:

- Sử dụng Khoảng cách Bình phương. ED sử dụng phép tính căn bậc hai. Tuy nhiên, nếu bỏ qua bước này, thứ hạng tương đối của các chuỗi con so sánh sẽ không thay đổi, vì hàm ED là đơn điệu và lõm [97]. Hơn nữa, sự vắng mặt của hàm căn bậc hai làm cho việc tính toán ED nhanh hơn.
- Từ bỏ sớm cho ED. Trong quá trình tính toán ED, nếu tổng hiện tại của sự khác biệt bình phương giữa mỗi cặp điểm dữ liệu tương ứng (x_i - y_i) (i = 1..kz, kz < n) vượt quá giá trị ngưỡng ε trong giải thuật gom cụm Leader, thì ngừng việc tính toán. Hình 5.1 minh họa ý tưởng *từ bỏ sớm* (early abandoning) cho độ đo khoảng cách Euclid.

Hình 5.1: Minh họa độ đo euclid có sử dụng kỹ thuật từ bỏ sớm

- Giải thuật gom cụm A-Leader

A-Leader là phiên bản được cải thiện từ I-Leader cho bài toán gom cụm. Giải thuật A-Leader chỉ khác giải thuật gom cụm I-Leader ở giai đoạn tính khoảng cách từ chuỗi con S_i đến các cụm C_j để quyết định chọn cụm C_j nào phù hợp cho chuỗi con S_i. Giải thuật A-Leader có sử dụng thêm kỹ thuật từ bỏ sớm để tăng tốc. Giá trị ngưỡng cho việc tăng tốc tính khoảng cách ED là ngưỡng ε của giải thuật gom cụm I-Leader.

- Giải thuật phát hiện chuỗi con bất thường EP-ALeader

EP-ALeader là phiên bản cải tiến từ giải thuật EP-ILeader. Nhìn chung, giải thuật EP-ALeader chỉ khác giải thuật EP-ILeader ở bước sử dụng giải thuật gom cụm A-Leader thay cho giải thuật gom cụm I-Leader trong EP-ILeader.

5.3.2 Đề xuất mới giải thuật TopK-EP-ALeader phát hiện k chuỗi con bất thường nhất trên chuỗi thời gian dạng tĩnh

TopK-EP-ALeader là phiên bản mở rộng của giải thuật EP-ALeader. Giải thuật TopK-EP-ALeader gồm bốn bước chính giống như trong giải thuật EP-ALeader.

Điểm khác biệt duy nhất giữa giải thuật TopK-EP-ALeader và giải thuật EP-ALeader là tại bước 4, giải thuật TopK-EP-ALeader cho kết quả là *k* chuỗi con bất thường nhất chính là những chuỗi con có hệ số bất thường lớn đến thứ *k* trong khi giải thuật EP-ALeader trả về một chuỗi con bất thường nhất chính là chuỗi con có hệ số bất thường lớn nhất. Nhờ sự hiện diện của các hệ số bất thường, việc TopK-EP-ALeader trả về *k* chuỗi con bất thường nhất không gây thêm chi phí tính toán nào cả.

Đánh giá các giải thuật

- A-Leader thực thi nhanh hơn I-Leader bình quân là **1,37** lần.

- EP-ALeader thực thi nhanh hơn EP-ILeader bình quân là **16,7** lần.

- TopK-EP-ALeader nhanh hơn TopK-EP-ILeader bình quân là **1,9** lần. Trong đó, TopK-EP-ILeader là giải thuật tìm *k* chuỗi con bất thường nhất dựa vào giải thuật EP-ILeader, TopK-EP-ALeader là giải thuật tìm *k* chuỗi con bất thường nhất dựa vào giải thuật EP-ALeader. Thực nghiệm cho thấy giải thuật TopK-EP-ALeader cho kết quả phát hiện bất thường chính xác.

5.4 Đề xuất mới giải thuật TopK-EP-ALeader-S phát hiện k chuỗi con bất thường nhất trên chuỗi thời gian dạng luồng

TopK-EP-ALeader-S là phiên bản mở rộng của TopK-EP-ALeader nhằm áp dụng cho chuỗi thời gian dạng luồng. Các tính năng mở rộng của TopK-EP-ALeader-S là nhằm vượt qua các thách thức trong việc tìm *k* chuỗi con bất thường nhất trên dữ liệu chuỗi thời gian dạng luồng. Chi tiết của giải thuật được trình bày như sau.

- Để sử dụng được TopK-EP-ALeader-S, một *cửa sổ di chuyển* (moving window) được định nghĩa để chứa chuỗi thời gian theo ngữ cảnh luồng. Trong cửa sổ này, một đoạn con của chuỗi thời gian dạng luồng được lưu trữ theo thời gian. Chỉ những điểm dữ liệu mới đến mới được chứa trong cửa sổ này. Cửa sổ di chuyển thường được hiện thực dưới dạng *vùng đệm xoay vòng* (circular buffer).

- Ngoài ra, TopK-EP-ALeader-S làm việc theo chiến lược *cập nhật trễ* (delayed update) thay cho chiến lược cập nhật tức thì để tăng tính hữu hiệu quả. Nhờ vào chiến lược trễ này, mỗi khi có điểm cực trị mới đến thì TopK-EP-ALeader-S mới

bắt đầu thực hiện tìm *k* chuỗi con bất thường mới nhất thay cho việc cứ mỗi điểm dữ liệu mới đến là phải đi tìm *k* chuỗi con bất thường mới nhất.

Đánh giá giải thuật TopK-EP-ALeader-S

Trong phần thực nghiệm với dữ liệu chuỗi thời gian dạng luồng, các tập dữ liệu chứa chuỗi dữ liệu thời gian được mô phỏng thành dạng luồng. Đối với giải thuật TopK-EP-ALeader-S, giá trị cho tham số chiều dài vùng đệm được thiết lập dựa vào chu kỳ của chuỗi dữ liệu thời gian. Nếu dữ liệu không có chu kỳ thì kích thước vùng đệm được ước lượng thông qua thực nghiệm.

Về tính chính xác: Các chuỗi con bất thường do TopK-EP-ALeader-S tìm được khớp với các chuỗi con bất thường do chuyên gia đánh dấu.

Tính đáp ứng tức thời: Cần tìm câu trả lời cho câu hỏi: "*Phương pháp phát hiện bất thường trực tuyến TopK-EP-ALeader-S có đáp ứng yêu cầu truyền dữ liệu thực tế không*?"

Đối với bộ dữ liệu điện năng POWER, tần suất ghi nhận dữ liệu là một giờ ghi nhận một lần. Vì vậy, thời gian đáp ứng cần cho một điểm dữ liệu mới đến của bộ dữ liệu POPWER là 1 giờ. Trong khi đó, TopK-EP-ALeader-S có thời gian đáp ứng cho mỗi điểm dữ liệu mới đến là 7 mili giây đối với dữ liệu điện năng. Xét trường hợp nhanh nhất, mỗi điểm dữ liệu mới đến cũng chính là điểm cực trị thì tốc độ của TopK-EP-ALeader-S nhanh gấp 514.286 lần so với tốc độ truyền của dữ liệu POWER.

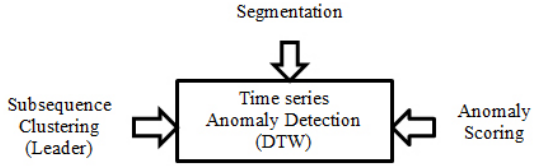
Đối với bộ dữ liệu điện tâm đồ ECG, chu kỳ của một nhịp tim là 1 giây. Mỗi điểm cực trị là tương ứng với nửa chu kỳ nhịp tim (nửa chuỗi con). TopK-EP-ALeader-S có thể phát hiện *k* chuỗi con bất thường nhất trong khoảng thời gian 6 mili giây. Như vậy, tốc độ phát hiện bất thường của TopK-EP-ALeader-S nhanh gấp **83** lần so với tốc độ truyền của dữ liệu ECG.

Kết luận: *Những phân tích trên cho thấy giải thuật TopK-EP-ALeader-S phát hiện k chuỗi con bất thường nhất trên chuỗi thời gian dạng luồng có thể đáp ứng yêu cầu truyền thực tế đối với bộ dữ liệu điện năng và điện tâm đồ. Ngoài ra, thực nghiệm cũng cho thấy giải thuật TopK-EP-ALeader-S cho kết quả phát hiện bất thường chính xác.*

CHƯƠNG 6 ĐỀ XUẤT PHÁT HIỆN BẤT THƯỜNG DỰA VÀO PHÂN ĐOẠN TRÊN DỮ LIỆU CHUỖI THỜI GIAN DẠNG TĨNH VÀ DỮ LIỆU CHUỖI THỜI GIAN DẠNG LUỒNG VỚI ĐỘ ĐO DTW

6.1 Đề xuất mới giải thuật EP-Leader-DTW phát hiện chuỗi con bất thường trên chuỗi thời gian dạng tĩnh với độ đo DTW

Ý tưởng chính của giải thuật đề xuất mới EP-Leader-DTW được trình bày trong hình 6.1.



Hình 6.1: Ý tưởng phát hiện bất thường trên dữ liệu chuỗi thời gian với khoảng cách DTW

Giải thuật EP-Leader-DTW sử dụng độ đo DTW gồm các bước:

- *Bước 1*: Sử dụng phương pháp điểm cực trị quan trọng để phân chia dữ liệu chuỗi thời gian thành các chuỗi con.
- *Bước 2*: Sử dụng phép *biến hình vị tự* (homothetic transform) để chuyển đổi các chuỗi con có chiều dài khác nhau về cùng một chiều dài, với chiều dài được chọn để biến hình vị tự là chiều dài trung bình của các chuỗi con.
- *Bước 3*: Giải thuật Leader sẽ gom cụm các chuỗi con đã được biến hình vị tự.
- *Bước 4*: Sử dụng các chuỗi con trong các cụm đã được gom cụm ở bước 3 để tính hệ số bất thường cho từng chuỗi con. Chuỗi con bất thường tìm được chính là chuỗi con có hệ số bất thường lớn nhất.

Đánh giá kết quả thực nghiệm

- Tính chính xác: Thực nghiệm cho thấy chuỗi con bất thường được phát hiện bởi EP-Leader-DTW gần giống với chuỗi con bất thường được phát hiện bởi giải thuật BF-DTW. Giải thuật BF_DTW là giải thuật Brute-Force trong đó độ đo khoảng cách được dùng là độ đo DTW. Giải thuật EP-Leader-DTW không phát hiện sai chuỗi con bất thường nào. Ngoài ra, đối với tập dữ liệu Tek16, chuỗi con bất thường do EP-Leader-DTW tìm được là hoàn toàn giống với chuỗi con bất thường được đánh dấu bởi các chuyên gia.

- Tính hữu hiệu: Trong thực nghiệm, thời gian thực thi và số lần gọi hàm tính khoảng cách DTW được sử dụng làm độ đo tính hữu hiệu của EP-Leader-DTW và QR-AF. Phương pháp QR-AF (Quadratic Regression and Anomaly Factors) do Leng và các cộng sự đề xuất về việc phát hiện chuỗi con bất thường trên dữ liệu chuỗi thời gian dưới độ đo DTW và dựa vào việc phân đoạn và hệ số bất thường. Để tăng tốc, độ đo DTW được sử dụng trong thực nghiệm là độ đo DTW có sử dụng kỹ thuật cận dưới LB_Keogh. Để đảm bảo tính công bằng của sự so sánh này, độ đo DTW có sử dụng kỹ thuật cận dưới LB_Keogh được dùng cho cả hai giải thuật EP-Leader-DTW và QR-AF. Thực nghiệm cho thấy, số lần gọi hàm tính khoảng cách DTW của EP-Leader-DTW ít hơn nhiều so với của QR-AF. Về thời gian thực thi, trung bình EP-Leader-DTW thực thi nhanh gấp **8,8** lần so với giải thuật QR-AF.

6.2 Đề xuất giải thuật SEP-Leader-DTW để phát hiện chuỗi con bất thường trên chuỗi thời gian dạng luồng với độ đo DTW

Giải thuật SEP-Leader-DTW bao gồm các ý chính như sau:

- *Giải thuật SEP-Leader-DTW sử dụng vùng đệm xoay vòng* để chứa các điểm dữ liệu đang được xem xét của chuỗi thời gian dạng luồng.
- *Giải thuật EP-Leader-DTW được sử dụng lại trong SEP-Leader-DTW:* Giải thuật EP-Leader-DTW được gọi để phát hiện chuỗi con bất thường trên phần chuỗi thời gian đang được lưu trữ trong vùng đệm.
- *Trình kích hoạt dựa vào chuỗi con được sử dụng để gọi EP-Leader-DTW một cách hiệu quả:* EP-Leader-DTW không được gọi mỗi khi có một điểm dữ liệu mới đến. Thay vào đó, quá trình sẽ bị trì hoãn cho đến khi điểm dữ liệu mới đến thực sự là một điểm cực trị quan trọng. Khi đó, một chuỗi con mới được hình thành và việc hình thành chuỗi con mới này giúp kích hoạt quá trình phát hiện chuỗi con bất thường trong phần chuỗi thời gian đang được lưu trữ trong vùng đệm.
- *Cập nhật gia tăng được áp dụng cho quá trình gom cụm các chuỗi con theo thời gian:* Các điểm dữ liệu cũ nhất thuộc về chuỗi con cũ nhất ở vùng đệm sẽ được xóa khỏi vùng đệm. Vì vậy, khi những điểm dữ liệu cũ nhất bị xóa đi,

cũng là lúc cần xóa chuỗi con cũ nhất ra khỏi cụm đang chứa nó (*chuỗi con cũ nhất là chuỗi con chứa các điểm dữ liệu cũ nhất*). Chiến lược xóa chuỗi con cũ nhất được thực hiện như sau: Khi EP-Leader-DTW được gọi lần đầu tiên, tất cả các chuỗi con trong vùng đệm được gom vào các cụm. Chuỗi con cũ nhất bị xóa khỏi vùng đệm sẽ rơi vào 1 trong 3 trường hợp sau:

- *Trường hợp 1*: Nếu chuỗi con cũ nhất không phải là phần tử đại diện cụm, chuỗi con này sẽ bị xóa khỏi cụm.
- *Trường hợp 2*: Nếu chuỗi con cũ nhất là phần tử đại diện cụm và cụm này chỉ có duy nhất một phần tử đại diện cụm thì cụm này sẽ bị xóa.
- *Trường hợp 3*: Nếu chuỗi con cũ nhất là phần tử đại diện cụm và cụm này có nhiều hơn một phần tử thì chuỗi con cũ nhất sẽ bị xóa khỏi cụm và chuỗi con thứ hai đứng ngay sau phần tử đại diện cụm sẽ được chuyển lên làm phần tử đại diện cho cụm này.

Thảo luận về giải thuật SEP-Leader-DTW

Trước hết, SEP-Leader-DTW có thể vượt qua các thách thức của bài toán phát hiện bất thường trên chuỗi thời gian dạng luồng, đó là việc xác định chiều dài của chuỗi con bất thường và tính đáp ứng tức thời cho dữ liệu dạng luồng. Đối với thách thức thứ nhất, chiều dài chuỗi con bất thường có thể được xác định một cách tự động dựa vào kết quả của quá trình phân đoạn. Với thách thức thứ hai, phương pháp phân đoạn và gom cụm gia tăng giúp chuyển đổi EP-Leader-DTW từ việc áp dụng cho dữ liệu tĩnh sang áp dụng cho dữ liệu luồng một cách phù hợp và hiệu quả. Thêm nữa, SEP-Leader-DTW được kế thừa tính hiệu quả của EP-Leader-DTW để đưa ra phản hồi tức thời nhằm phát hiện chuỗi con bất thường trên chuỗi thời gian dạng luồng mỗi khi có điểm cực trị mới xuất hiện.

Với tất cả những điều biện luận trên, SEP-Leader-DTW có thể được xem là phương pháp mới đầu tiên để phát hiện chuỗi con bất thường trên chuỗi thời gian dạng luồng với độ đo DTW so với các phương pháp hiện có.

Đánh giá kết quả thực nghiệm

Tính chính xác: Tính chính xác của phương pháp phát hiện bất thường trên dữ liệu chuỗi thời gian dạng luồng được kiểm tra nhờ vào hai tập dữ liệu chuỗi thời

gian có đánh dấu chuỗi con bất thường bởi các chuyên gia (*ground truth*), đó là: tập Tek16 và tập ECG. Ngoài ra, các kết quả chuỗi con bất thường tìm được bởi SEP-Leader-DTW cũng được so sánh với các chuỗi con bất thường tìm được bởi HOT SAX, một giải thuật phát hiện chuỗi con bất thường trên chuỗi thời gian dạng tĩnh được báo cáo bởi Keogh và cộng sự.

Thực nghiệm cho thấy chuỗi con bất thường trên tập dữ liệu điện tâm đồ ECG được đánh dấu bởi chuyên gia giống với kết quả chuỗi con bất thường tìm được bởi SEP-Leader-DTW. Hơn nữa, chuỗi con bất thường do SEP-Leader-DTW tìm được hoàn toàn trùng khớp với chuỗi con bất thường do phương pháp HOT SAX tìm được trên dữ liệu chuỗi thời gian dạng tĩnh Tek16.

Tính hữu hiệu: Tính hữu hiệu của phương pháp phát hiện bất thường trên dữ liệu chuỗi thời gian dạng luồng được đánh giá thông qua việc kiểm tra xem SEP-Leader-DTW có đáp ứng được yêu cầu trực tuyến của các ứng dụng hay không.

Thời gian thực thi trung bình của SEP-Leader-DTW là 0,043 giây trên toàn bộ tập dữ liệu ECG với chiều dài 20.000 điểm dữ liệu. Thời gian thực thi trung bình của SEP-Leader-DTW là 0,006 giây trên vùng đệm xoay vòng mỗi khi có điểm cực trị quan trọng mới của tập dữ liệu ECG được đưa vào vùng đệm. Mỗi điểm cực trị quan trọng mới đánh dấu một nửa nhịp tim. Trong khi đó, trên thực tế, thời gian đến của một nhịp tim là khoảng 1 giây. Vì vậy, tốc độ phát hiện bất thường của SEP-Leader-DTW nhanh gấp 83 lần so với tốc độ truyền của dữ liệu điện tâm đồ. Đối với tập dữ liệu điện năng, tốc độ phát hiện chuỗi con bất thường của SEP-Leader-DTW nhanh hơn khoảng 18.000.000 lần so với tốc độ ghi dữ liệu tiêu thụ điện năng.

Những phân tích trên cho thấy giải thuật SEP-Leader-DTW phát hiện chuỗi con bất thường trên chuỗi thời gian dạng luồng có thể đáp ứng yêu cầu truyền thực tế đối với chuỗi thời gian dạng luồng điện năng POWER và điện tâm đồ ECG.

CHƯƠNG 7 ỨNG DỤNG PHÁT HIỆN BẤT THƯỜNG VÀO VIỆC CẢI THIỆN CHẤT LƯỢNG DỰ BÁO DỮ LIỆU CHUỖI THỜI GIAN

7.1 Đề xuất hướng tiếp cận mới EPL_S_*X* cho dự báo dữ liệu chuỗi thời gian

EPL_S_*X* là hướng tiếp cận được đề xuất để dự báo dữ liệu chuỗi thời gian có sự xuất hiện của yếu tố bất thường. Trong đó, *X* có thể là bất kỳ một phương pháp dự báo chuỗi thời gian hiện có nào đó. Thông thường, bất thường được xử lý trong giai đoạn tiền xử lý dữ liệu của quá trình khai phá dữ liệu. Từ đó, luận án xác định hiệu chỉnh bất thường là giai đoạn tiền xử lý dữ liệu trước khi dự báo chuỗi thời gian bằng bất kỳ một phương pháp dự báo hiện có nào. Hướng tiếp cận EPL_S_*X* bao gồm ba bước chính:

- **Bước 1**: *Phát hiện bất thường*. Ở bước này, chuỗi thời gian ban đầu được xử lý nhằm phát hiện bất thường (nếu có). Những bất thường này chính là những chuỗi con bất thường có trong chuỗi thời gian ban đầu.
- **Bước 2**: *Hiệu chỉnh bất thường*. Các chuỗi con bất thường ở bước 1 được hiệu chỉnh thành chuỗi con bình thường. Điều đó có nghĩa là những chuỗi con bất thường sẽ được làm mịn. Sau bước 2 này, chuỗi thời gian được xem như là *sạch* (clean), không có chuỗi con bất thường.
- **Bước 3**: *Dự báo*. Chuỗi dữ liệu thời gian sạch ở bước 2 được sử dụng để dự báo. Bất kỳ một phương pháp dự báo chuỗi thời gian nào cũng đều có thể được áp dụng trên chuỗi thời gian sạch này. Kết quả đầu ra của bước này là các điểm dữ liệu được dự báo. Đây cũng chính là kết quả đầu ra của hướng tiếp cận dự báo dữ liệu chuỗi thời gian EPL_S_*X*.

Đối với hướng tiếp cận dự báo EPL_S_*X*, không có bất kỳ ràng buộc nào về tính chất của dữ liệu. Ngoài ra, ở bước 2, hướng tiếp cận EPL_S_*X* chỉ làm sạch những chuỗi con bất thường và giữ nguyên tất cả những dữ liệu còn lại trên chuỗi thời gian.

7.2 Đánh giá kết quả thực nghiệm

Hướng tiếp cận EPL_S_*X* được thực nghiệm tương ứng với hai câu hỏi nghiên cứu và một số thiết lập thực nghiệm như sau:

- *Câu hỏi 1*: Hướng tiếp cận đề xuất EPL_S_*X* có vượt trội hơn các phương pháp RHW và RHW' trong [111] khi dự báo dữ liệu chuỗi thời gian hay không?

- *Câu hỏi 2*: Kỹ thuật *hiệu chỉnh bất thường* (anomaly – repair) được đề xuất có cải thiện được độ chính xác của kết quả dự báo của các phương pháp dự báo khác hay không?

Với câu hỏi 1, hiệu suất dự báo của phương pháp EPL_S_*k*NN được so sánh với hiệu suất dự báo của các phương pháp do Gelper và các cộng sự đề xuất.

Kết quả thực nghiệm cho thấy, giá trị lỗi bình phương trung bình - *MSE* (mean squared error) của các phương pháp RHW và RHW' lớn hơn giá trị lỗi MSE của phương pháp EPL_S_*k*NN trong cả hai trường hợp thực nghiệm. Bất kể điểm dữ liệu bất thường xuất hiện ở đâu, việc làm mịn bất thường theo hướng tiếp cận EPL_S_*X* cũng giúp cải thiện chất lượng dự báo.

Kết luận đánh giá cho câu hỏi 1: EPL_S_*k*NN hiệu quả hơn cả hai phương pháp RHW'và RHW do Gelper và các cộng sự đề xuất.

Với câu hỏi 2, thực nghiệm được tiến hành trên cả 8 phương pháp trong đó có 4 phương pháp không áp dụng hướng tiếp cận EPL_S_*X* gồm: LR, *k*NN, ANN, Hybrid và 4 phương pháp áp dụng hướng tiếp cận EPL_S_*X* bao gồm: EPL_S_LR, EPL_S_*k*NN, EPLS_ANN, EPL_S_Hybrid. Kết quả thực nghiệm của 8 phương pháp được trình bày như sau:

- Đối với phương pháp dự báo đơn giản là *hồi qui tuyến tính* (linear regression - LR), tỉ lệ cải thiện MSE của EPL_S_LR đối với LR là từ 1,05 là 568,63 lần. Tỉ lệ cải thiện MAE của EPL_S_LR hơn MAE của LR từ 1,08 đến 23,21 lần và tỉ lệ cải thiện MAPE của EPL_S_LR hơn MAPE của LR từ 1,08 đến 23,147 lần.

- Đối với phương pháp dự báo *k- lân cận gần nhất* (k-nearest neighbors- *k*NN), tỉ lệ cải thiện MSE của EPL_S_*k*NN so với *k*-NN là từ 1,01 đến 7,9 lần. Tỉ lệ cải thiện MAE của EPL_S_*k*NN so với *k*-NN là 1,03 đến 2,73 lần và tỉ lệ cải thiện MAPE là từ 1,04 đến 2,74 lần.

- Đối với phương pháp dự báo *mạng nơ ron nhân tạo* (artificial neural network - ANN), tỉ lệ cải thiện MSE, MAE và MAPE của EPL_S_ANN so với ANN lần lượt là từ 1,68 đến 1.382,92, từ 1,37 đến 33,59 và từ 1,3 đến 33,52 lần.

- Đối với phương pháp Hybrid kết hợp phương pháp dự báo làm mịn theo hàm mũ Holt-Winters và phương pháp dự báo mạng nơ ron nhân tạo do Bao và cộng sự đề xuất, kết quả thực nghiệm cho thấy tỉ lệ cải thiện MSE, MAE, và MAPE

của EPL_S_Hybrid so với Hybrid tương ứng là từ 1,11 đến 32,37, từ 1,05 đến 4,88, và từ 1,06 đến 3,95 lần.

Kết luận đánh giá cho câu hỏi 2: Kết quả dự báo dữ liệu chuỗi thời gian của các phương pháp dự báo được cải thiện khi có áp dụng hướng tiếp cận EPL_S_*X*.

Tóm lại, đóng góp chính của EPL_S_*X* là nhằm bổ trợ cho các phương pháp dự báo đã có dựa vào việc giảm nhiễu và tăng độ chính xác của kết quả dự báo. Ngoài ra, EPL_S_*X* cũng góp phần phát hiện phương pháp dự báo nào nhạy cảm với nhiễu hơn dựa vào quan sát tỉ lệ cải thiện chất lượng dự báo sau khi giảm nhiễu.

CHƯƠNG 8 KẾT LUẬN

8.1 Các đóng góp chính

- Cải tiến một số giải pháp (dựa vào cửa sổ trượt) đã có cho bài toán phát hiện chuỗi con bất thường trên dữ liệu chuỗi thời gian.

- Phát triển được các giải pháp mới và hiệu quả cho bài toán phát hiện chuỗi con bất thường (chuỗi con bất thường nhất, *k* chuỗi con bất thường nhất) trên dữ liệu chuỗi thời gian, đặc biệt là dữ liệu chuỗi thời gian dạng luồng với độ đo Euclid và độ đo DTW. Giải quyết được thách thức chung của bài toán: dữ liệu lớn, thay đổi liên tục, yêu cầu xử lý nhanh và đáp ứng tức thời, tổng quát cho nhiều miền ứng dụng khác nhau. Giải quyết được hạn chế chung của các phương pháp hiện có cho bài toán phát hiện bất thường: không yêu cầu tham số kích thước của chuỗi con bất thường.

- Phát triển được khung thức mới để dự báo dữ liệu chuỗi thời gian dựa vào phát hiện bất thường và hiệu chỉnh bất thường giúp nâng cao hiệu quả của các phương pháp dự báo hiện có khi chuỗi thời gian có chứa đựng bất thường.

8.2 Các kết quả đạt được

- Cải thiện đáng kể hiệu suất của các giải thuật phát hiện chuỗi con bất thường trên dữ liệu chuỗi thời gian theo hướng tiếp cận dựa vào cửa sổ trượt. Các giải thuật đề xuất cải tiến bao gồm: *I-HOTSAX, Hash_DD,* và *KBF_GPU*. Riêng giải thuật KBF_GPU dựa vào công nghệ GPU có tốc độ thực thi cao, giúp thích ứng với quy mô dữ liệu lớn.

- Đề xuất mới các giải thuật hiệu quả cho bài toán phát hiện chuỗi con bất thường trên dữ liệu chuỗi thời gian dạng tĩnh và dữ liệu chuỗi thời gian dạng luồng theo hướng tiếp cận dựa vào phân đoạn trên độ đo Euclid và độ đo DTW. Các giải thuật đề xuất mới bao gồm: *EP-ILeader, EP-ALeader, EP-Leader-DTW, SEP-Leader-DTW.*

- Đề xuất mới các giải thuật phát hiện *k* chuỗi con bất thường nhất trên dữ liệu chuỗi thời gian dạng tĩnh và dạng luồng theo hướng tiếp cận dựa vào phân đoạn. Các giải thuật đề xuất mới bao gồm: *TopK-EP-ILeader, TopK-EP-ALeader* và *TopK-EP-ALeader-S.*

- Đề xuất hướng tiếp cận mới ứng dụng kết quả phát hiện chuỗi con bất thường vào bài toán dự báo dữ liệu chuỗi thời gian. Hướng tiếp cận đề xuất có tên là *EPL_S_X*.

- Đề xuất cải tiến giải thuật gom cụm *I-Leader* và *A-Leader* hiệu quả cho bài toán gom cụm.

- Đề xuất mới độ đo *PALS* đánh giá chất lượng của các phương pháp phân đoạn chuỗi thời gian.

8.3 Hướng phát triển

- Mở rộng KBF_GPU để có thể làm việc được với độ đo DTW và có thể khai phá motif (chuỗi được lặp lại nhiều nhất) trên dữ liệu chuỗi thời gian.

- Nghiên cứu áp dụng hướng lập trình phân bố dựa vào Spark hay Map Reduce cho giải thuật KBF.

- Mở rộng hướng tiếp cận EPL_S_*X* để dự báo *dữ liệu chuỗi thời gian dạng luồng* thuộc nhiều miền ứng dụng cần dự báo theo thời gian thực.

- Áp dụng hướng tiếp cận EPL_S_*X* cho phương pháp dự báo dựa trên *mạng nơ ron học sâu* (deep neural networks).

- Mở rộng phương pháp EP-Leader-DTW để phát hiện các ảnh bất thường trong cơ sở dữ liệu hình ảnh trong đó hình ảnh đã được chuyển thành dữ liệu chuỗi thời gian. Ứng dụng này dự kiến sẽ sử dụng khoảng cách DTW bất biến xoay vòng được đề xuất trong [50]. Bên cạnh đó, giải thuật EP-Leader-DTW cũng sẽ được cải thiện bằng cách sử dụng bộ kỹ thuật cho DTW có tên là UCR-DTW được đề xuất trong [150] để tăng tốc độ tính toán khoảng cách DTW hơn nữa.