การตรวจสอบความเหมาะสมของการจัดกลุ่มข้อมูลอนุกรมเวลา
The Consideration of Proper Time-series Clustering

ทิพยา ถินสูงเนิน* มาโนช ถินสูงเนิน กิตติศักดิ์ เกิดประสพ และนิตยา เกิดประสพ
สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครราชสีมา จ.นครราชสีมา 30000
*E-mail : tippayasot@hotmail.com

บทคัดย่อ

ปัญหาที่สำคัญสำหรับการจัดกลุ่มข้อมูลอนุกรมเวลาคือการไม่ทราบจำนวนกลุ่มที่เหมาะสมในการจัดกลุ่ม โดยเฉพาะอย่างยิ่งข้อมูลอนุกรมเวลาที่ไม่ทราบกลุ่มที่แท้จริง ข้อมูลอนุกรมเวลาเป็นข้อมูลที่มีคุณสมบัติที่น่าสนใจหลายอย่าง ทั้งขนาดข้อมูล รูปร่างของอนุกรม โครงสร้าง หรือโมเดล เป็นต้น คุณสมบัติเหล่านี้สามารถทำให้อนุกรมเวลาถูกจัดกลุ่มได้แตกต่างกัน สองเทคนิคพื้นฐานที่นิยมใช้ในการพิจารณาจำนวนกลุ่มที่เหมาะสมสำหรับการจัดกลุ่มโดยทั่วไปได้แก่ ซิลลูเอ็ต และค่าผลรวมความผิดพลาด ซึ่งไม่เพียงพอสำหรับการจัดกลุ่มข้อมูลอนุกรมเวลา ในงานวิจัยนี้ได้นำเสนอวิธีการสำหรับตรวจสอบความเหมาะสมของการจัดกลุ่มข้อมูลอนุกรมเวลา โดยอาศัย ซิลลูเอ็ต และค่าผลรวมความผิดพลาดเป็นพื้นฐาน ร่วมกับการแทนอนุกรมเวลาด้วยตัวแทนขององค์ประกอบ ในการทดลองใช้ทั้งข้อมูลสังเคราะห์และข้อมูลจริงจำนวนทั้งหมด 3 ชุดข้อมูล เปรียบเทียบผลการทดลองกับวิธีการวัดความคล้ายคลึง 6 วิธี และจัดกลุ่มด้วยเทคนิคแบบลำดับชั้น และแบบแบ่งแยก ผลการวิจัยพบว่าวิธีที่นำเสนอสำหรับใช้ตรวจสอบจำนวนกลุ่มที่เหมาะสมจะให้ผลสอดคล้องกันทั้งซิลลูเอ็ต และผลรวมความผิดพลาด รวมทั้งมีความคล้ายคลึงกันอย่างมีความหมาย

คำสำคัญ : การจัดกลุ่มข้อมูลอนุกรมเวลา จำนวนกลุ่มที่เหมาะสม ตัวแทนอนุกรมเวลา ซิลลูเอ็ต ผลรวมความผิดพลาด

Abstract

A key issue for the clustering of time-series dataset is not being aware of the appropriate number of clusters. In particular, the true cluster of most time-series is not known. There are many interesting properties of time-series data such as scale, shape, structure, model and others. These features can make the different results in clustering of time-series. Two basic techniques used to determine the appropriate number of clusters in general are the silhouette and the sum of squared error, which are not enough for considering clusters of time-series data. In this research, we present a method for evaluating the suitability of cluster by using the silhouette and the sum of squared error. Furthermore, we integrate evaluation method with the designated representation using agent of time-series components. The experimentation uses both synthetic and 2 sets of real time-series data to compare the results of partitioning clustering and hierarchical clustering with 6 similarity measures. The results of our proposed evaluation method showed that the number of clusters was in accordance with the silhouette and the sum of squared error. Moreover the similarity was meaningful.

Keywords : Time-series clustering; Appropriate number of clusters; Representation; Silhouette; Sum of Squared Error

1. บทนำ

ในยุคที่พัฒนาการด้านการประมวลผลข้อมูลและการจัดเก็บข้อมูลมีศักยภาพสูงขึ้น จึงมักจัดเก็บข้อมูลไว้เป็นช่วงระยะเวลายาวนานขึ้นพบได้จากการประยุกต์ใช้ในงานหลายแขนงที่เริ่มมีการจัดเก็บข้อมูลในรูปแบบอนุกรมเวลา (Time-series) ตัวอย่างเช่น ข้อมูลการขาย (Sale Data) ข้อมูลราคาหุ้น (Stock Prices) ข้อมูลสภาพอากาศ (Weather Data) การตรวจวัดทางชีวะการแพทย์ (Biomedical Measurements) (เช่น ความดันโลหิตและการวัดคลื่นไฟฟ้า) ข้อมูลชีวภาพ (Biometrics Data) (เช่น ข้อมูลรูปภาพเกี่ยวกับการจดจำใบหน้า) เป็นต้น [1] ลักษณะสำคัญของข้อมูลอนุกรมเวลา คือเป็นข้อมูลขนาดใหญ่ มิติสูงและซับซ้อน ซึ่งเป็นความท้าทายอย่างยิ่งสำหรับการขุดค้นหาความรู้ที่ซ่อนอยู่ในข้อมูลเหล่านี้ด้วยหลากหลายวิธีการ เช่น วิธี Subsequence Matching [2], [3] Anomaly Detection และ Motif Discovery [3] ลักษณะงานที่ศึกษา เช่น Clustering, Classification [4], Visualization [5], Segmentation [6], Identifying Patterns, Summarization [7] และ Forecasting เป็นต้น นอกจากนี้ยังมีงานวิจัยอีกมากที่พยายามปรับปรุงวิธีการเหล่านี้ให้ดีขึ้น [8], [9] อีกด้วย

การจัดกลุ่มข้อมูล (Data Clustering) เป็นเทคนิคการเรียนรู้ของเครื่องจักรแบบไม่มีผู้สอน (Unsupervised Learning) [10], [11] โดยมีหลักการบนพื้นฐานที่ว่า ข้อมูลที่ถูกจัดให้อยู่ในกลุ่มเดียวกันจะต้องมีลักษณะที่คล้ายคลึงกันมาก ในขณะที่ข้อมูลคนละกลุ่มย่อมคล้ายคลึงกันน้อยกว่า [12] ซึ่งอัลกอริทึมที่ใช้จะไม่รู้คลาสเป้าหมายมาก่อน เป็นผลให้งานจัดกลุ่มข้อมูลอนุกรมเวลามีประเด็นปัญหาที่น่าศึกษาหลายส่วนด้วยกัน ประการแรกคือการหามาตรวัดที่เหมาะสมสำหรับวัดคล้ายคลึง/ต่างกัน (Similarity / Dissimilarity Measure) ของแต่ละอนุกรม โดยอาจใช้วิธีการวัดระยะห่าง ตรวจสอบความคล้ายคลึงของรูปร่าง หรือตรวจจับลักษณะเด่นที่เกิดขึ้นบ่อยเป็นต้น ซึ่งมาตรวัดแต่ละแบบย่อมให้ผลแตกต่างกัน ประการที่สองคือการไม่ทราบจำนวนกลุ่มที่เหมาะสม (หรือค่า k) สำหรับใช้ในเทคนิคการจัดกลุ่ม โดยเฉพาะในข้อมูลที่ยากต่อการจำแนก เช่น ข้อมูลน้ำท่า (Runoff Data) ที่จัดเก็บอย่างต่อเนื่องในช่วงระยะเวลาหนึ่งทำให้ยากต่อการจัดกลุ่มข้อมูลที่รวบรวมไว้ด้วยวิธีการโดยทั่วไปได้ หรือข้อมูลการตรวจวัดทางชีวะการแพทย์ เช่น ข้อมูลคลื่นไฟฟ้าหัวใจ (ECGs) ที่ต้องวินิจฉัยด้วยผู้เชี่ยวชาญเฉพาะซึ่งอาศัยการตรวจจับสัญญาณที่ผิดปกติ เป็นต้น สำหรับเทคนิคพื้นฐานที่นิยมใช้ในการพิจารณาจำนวนกลุ่มที่เหมาะสมสำหรับการจัดกลุ่ม ได้แก่ ซิลลูเอ็ต (Silhouette) [13] และค่าผลรวมความผิดพลาด (Sum of Squared Errors: SSE) [12] มีงานวิจัยที่ศึกษาเทคนิคการตรวจสอบผลการจัดกลุ่มข้อมูล (Clustering Validity) เพื่อหาจำนวนกลุ่มที่เหมาะสม (Appropriate Number of Cluster) สำหรับการจัดกลุ่มข้อมูลด้วยอัลกอริทึมแบบ k-Means [14] โดยใช้ซิลลูเอ็ตและค่าผลรวมความผิดพลาด ซึ่งพบว่ากลุ่มที่เหมาะสมจะให้ผลสอดคล้องกันทั้งสองวิธี

ประการสุดท้ายคือการจัดการกับข้อมูลขนาดใหญ่ มิติสูงและมีความซับซ้อนที่มักส่งผลต่อประสิทธิภาพในการจัดกลุ่ม ซึ่งหัวใจของความสำเร็จอย่างหนึ่งสำหรับปัญหานี้คือ การแทนข้อมูลที่ดี [15] นั่นคือการเลือกใช้คุณลักษณะที่มีประโยชน์ หรือแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสม โดยหากตัวแทนข้อมูลที่ใช้ไม่ดีพออาจก่อปัญหาทำให้การจัดกลุ่มข้อมูลอนุกรมเวลามีคุณภาพต่ำได้โดยเฉพาะในการจัดกลุ่มแบบแบ่งแยก (Partitioning Approaches) อย่าง k-Means และ k-Medoids [16] ซึ่งถือเป็นเทคนิคการจัดกลุ่มข้อมูลที่นิยมใช้งานกันอย่างกว้างขวาง ส่วนการจัดกลุ่มแบบลำดับชั้นจะมีประโยชน์มากสำหรับจัดกลุ่มข้อมูลอนุกรมเวลาในเรื่องของการแสดงผลของการจัดกลุ่ม [17], [18] และลักษณะเด่นที่ไม่จำเป็นต้องกำหนดจำนวนกลุ่มไว้ก่อนล่วงหน้า ซึ่งถือเป็นจุดแข็งของการจัดกลุ่มแบบลำดับชั้น นอกจากนี้การจัดกลุ่มข้อมูลอนุกรมเวลาแบบลำดับชั้นยังเป็นวิธีที่ทำงานได้ดีกับอนุกรมเวลาที่มีขนาดไม่เท่ากันอีกด้วย [1] ซึ่งมีความเป็นไปได้ในการจัดกลุ่มข้อมูลอนุกรมเวลาที่มีขนาดไม่เท่ากันหากมีการเลือกใช้วิธีการวัดความคล้ายคลึง/ต่าง ที่เหมาะสมเช่นวิธี Dynamic Time Warping [19], [20] จะสามารถช่วยให้การจัดกลุ่มอนุกรมเวลามีคุณภาพดีได้

จากประเด็นปัญหาที่กล่าวมาข้างต้น งานวิจัยนี้จึงนำเสนอวิธีการตรวจสอบความเหมาะสมของการจัดกลุ่มข้อมูลอนุกรมเวลา โดยใช้ ซิลลูเอ็ตและค่าผลรวมความผิดพลาดเป็นฐาน ร่วมกับการแทนอนุกรมเวลาแบบใหม่ โดยใช้ตัวแทนองค์ประกอบของอนุกรมเวลา ซึ่งในการทดลองมีการใช้ทั้งข้อมูลสังเคราะห์และข้อมูลจริงจำนวนทั้งหมด 3 ชุดข้อมูล โดยเปรียบเทียบผลการจัดกลุ่มข้อมูลระหว่างข้อมูลดั้งเดิม (Raw Time-series) กับ ตัวแทนอนุกรมเวลา (วิธีที่นำเสนอ) ด้วยการใช้มาตรวัดความคล้ายคลึง 6 วิธี และใช้การจัดกลุ่มด้วยเทคนิคแบบลำดับชั้น และแบบแบ่งแยก

2. วัสดุอุปกรณ์และวิธีการวิจัย

ทฤษฎีที่ใช้ในงานนี้ได้แก่ การแทนอนุกรมเวลา การวัดความคล้ายคลึง/ต่าง ของอนุกรมเวลา อัลกอริทึมสำหรับจัดกลุ่มข้อมูล ได้แก่ เค-มีนส์ เค-มีดอยส์ เอชคลัสต์ (Hclust) และพีดีซี และส่วนสุดท้ายคือวิธีประเมินผลการจัดกลุ่ม ได้แก่ ซิลลูเอ็ต และค่าผลรวมความผิดพลาด รายละเอียดอธิบายได้ดังนี้

2.1 การแทนข้อมูลอนุกรมเวลา

วิธีการแทนข้อมูล (Representation Method) อนุกรมเวลา หรือเรียกว่าการลดมิติ เป็นขั้นตอนปกติสำหรับงานจัดกลุ่มข้อมูลอนุกรมเวลาประเภท Whole Time-series Clustering ซึ่งมักดำเนินการด้วยหลายสาเหตุ เช่น ความต้องการลดการใช้หน่วยความจำในการประมวล เพื่อให้การคำนวณระยะในการจัดกลุ่มเร็วขึ้น [7], [21] หรือเพื่อป้องกันการวัดระยะบางวิธีที่อ่อนไหวมากกับรูปแบบที่ผิดปกติ [22], [23] จากการทบทวนงานวิจัยพบว่าการแทนอนุกรมเวลาแบ่งเป็น 4 ประเภท [7], [23]-[25] ได้แก่ 1) วิธีแบบปรับข้อมูล (Data Adaptive) 2) วิธีแบบไม่ปรับข้อมูล (Non-Data Adaptive) 3) วิธีแบบใช้โมเดล (Model-based) และ 4) วิธีแบบบังคับข้อมูล (Data Dictated) สำหรับประเภทที่ 1 และ 2 มีนักวิจัยได้นำเสนอเทคนิคการแทนอนุกรมเวลาไว้มาก ดัง (Figure 1) ที่แสดงผังเทคนิคย่อยสำหรับการแทนอนุกรมเวลา ด้วยวิธีแบบปรับข้อมูล และวิธีแบบไม่ปรับข้อมูล

วิธีแทนอนุกรมเวลาแบบไม่ปรับข้อมูลเป็นวิธีที่เหมาะกับอนุกรมเวลาที่มีความยาวเท่ากัน มีการวัดความคล้ายคลึงอย่างตรงไปตรงมา ซึ่งช่วยให้เกิดการจัดกลุ่มอย่างมีความหมายได้

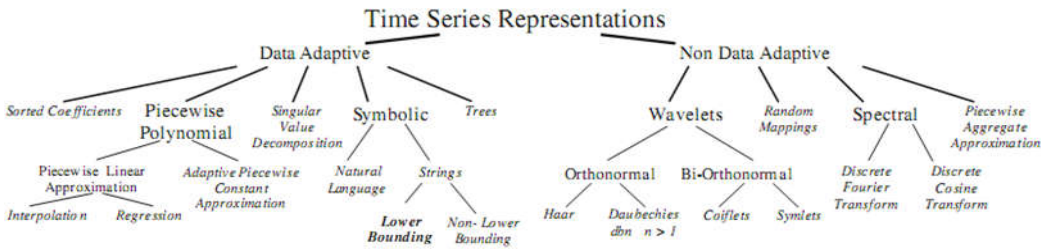


Figure 1. A Hierarchy of the various Time-series Representations in the literature [30].

เนื่องจากหากเลือกใช้ตัวแทนที่ไม่ดีเพียงบางส่วน (Subsequence Time-series) มาจัดกลุ่ม และยังวัดความคล้ายคลึงด้วยมาตรวัดที่ไม่เหมาะสมร่วมด้วย จะเป็นผลให้การจัดกลุ่มอนุกรมเวลาขาดความหมายได้ [2], [22] ตัวอย่างวิธีแทนอนุกรมเวลาแบบไม่ปรับข้อมูล ได้แก่ HAAR, DAUBECHIES, Symlets, Discrete Wavelet Transform (DWT), Spectral Chebyshev Polynomials [26], SpectralDFT [27], Random Mappings [28i], Piecewise Aggregate Approximation (PAA) [21] และ Indexable Piecewise Linear Approximation (IPLA) [29]

ในงานวิจัยนี้จึงได้นำเสนอวิธีแทนอนุกรมเวลาโดยกระทำกับข้อมูล Whole Time-series ด้วยวิธีการวิเคราะห์แยกองค์ประกอบอนุกรมเวลา เพื่อพัฒนา

วิธีการแทนอนุกรมเวลาที่สามารถช่วยให้จัดกลุ่มข้อมูลอนุกรมเวลาได้อย่างเหมาะสม และจัดกลุ่มอย่างมีความหมาย โดยอาศัยการตรวจสอบความเหมาะสมด้วยซิลลูเอ็ต ผลรวมค่าความผิดพลาด และแผนภาพการจัดกลุ่ม

2.2 การวัดความคล้ายคลึง/ต่าง ของอนุกรมเวลา

การจัดกลุ่มข้อมูลอนุกรมเวลาจะอาศัยมาตรวัดระยะระดับสูง ซึ่งมีหลายมาตรวัดที่ใช้สำหรับการแทนอนุกรมเวลา วิธีที่ง่ายที่สุดสำหรับวัดระยะระหว่างสองอนุกรมเวลาคือการใช้ข้อมูลอนุกรมเวลาเชิงเดี่ยว (Univariate) และวัดระยะระหว่างข้อมูลอนุกรมเวลาทุกจุดเวลาด้วยมาตรวัดที่เหมาะสม [1] สำหรับวิธีการของการวัดระยะในข้อมูลอนุกรมเวลานั้นมีนักวิจัยได้ศึกษาและนำเสนอวิธีการวัดระยะไว้มากมาย [1], [31] การเลือกใช้วิธีการวัดระยะที่เหมาะสมขึ้นอยู่กับคุณลักษณะ ขนาด การแทนข้อมูล และวัตถุประสงค์การจัดกลุ่ม สำหรับมาตรวัดที่นำมาใช้ในงานวิจัยนี้มี 5 กลุ่ม [31] กำหนดให้เป็น X_T และ Y_T เป็นอนุกรมเวลา 2 ชุดที่ต้องการวัดความคล้ายคลึง รายละเอียดมาตรวัดแต่ละกลุ่ม นิยามได้ต่อไปดังนี้ [31]

1) มาตรวัดตาม Raw Data ได้แก่

มาตรวัดระยะแบบ Euclidean Distance หรือ Minkowski Distance (diss.EUCL) หรือ L_q-norm Distance เมื่อ d_{L_q} คือระยะระหว่างอนุกรมเวลา X และ Y ที่จุดเวลา $t = 1$ ถึง T ด้วยมาตรวัด Minkowski Distance นิยามได้ดังสมการที่ 1

$$d_{L_q}(X_T, Y_T) = \left(\sum_{t=1}^{T}(X_t \quad Y_t)^q\right)^{1/q} \qquad (1)$$

โดยที่มี $q = 1$ คือ Manhattan Distance, $q = 2$ คือ Euclidean Distance และเมื่อ $q = \infty$ คือ Chebyshev Distance

มาตรวัด Dynamic Time Warping Distance หรือ DTW Distance (diss.DTW) คือการวัดความคล้ายของรูปร่าง $\boldsymbol{r}$ ในอนุกรมเวลา ด้วยการวัดระยะระหว่างสองลำดับย่อยของเวลา (X_{a_i}, Y_{b_i}) กำหนดให้ $r = ((X_{a1}, Y_{b1}), \dots, (X_{am}, Y_{bm}))$, M คือ เซตของลำดับที่เป็นไปได้ทั้งหมด m คู่ ดังนั้น d_{DTW} คือ DTW Distance ที่ใกล้ที่สุดนิยามได้ดังสมการที่ 2

$$d_{DTW}(X_T, Y_T) = \min_{r \in M}\left(\sum_{i=i,\dots m}|X_{ai} \quad Y_{bi}|\right) \qquad (2)$$

2) มาตรวัดตามคาบเวลา ได้แก่

มาตรวัดตามคาบเวลา จะมี Integrated Periodogram Based Dissimilarity (diss.IN.PER) เป็นมาตรวัดที่ใช้หาความคล้ายคลึงของอนุกรมเวลาที่ทำงานได้แบบบูรณาการทั้งกรณีปกติทั่วไป สนใจโครงสร้างความสัมพันธ์ และมีความแปรปรวน เมื่อ F_{X_T} และ F_{Y_T} คือ Integrated Periodogram ของ X_T และ Y_T ดังนั้น ความคล้ายคลึงของอนุกรมเวลาทั้งสองนิยามได้ดังนี้

$$d_{IP}(X_T, Y_T) = \int_{-\pi}^{\pi}\left|F_{X_T}(\lambda) \quad F_{Y_T}(\lambda)\right| d\lambda \qquad (3)$$

โดย

$F_{X_T}(\lambda_j) = C_{X_T}^{-1}\sum_{i=1}^{j} I_{X_T}(\lambda_i)$ และ $F_{Y_T}(\lambda_j) = C_{Y_T}^{-1}\sum_{i=1}^{j} I_{Y_T}(\lambda_i)$ ที่ใช้ $C_{X_T} = \sum_i I_X(\lambda_i)$, $C_{Y_T} = \sum_i I_{Y_T}(\lambda_i)$ กรณี Normalized และใช้ $C_{X_T} = C_{Y_T} = 1$ กรณี non-Normalized คาบเวลา และเมื่อ

$I_{X_T}(\lambda_k) = T^{-1}\left|\sum_{t=1}^{T} X_t e^{-i\lambda_k t}\right|^2$ และ

$I_{Y_T}(\lambda_k) = T^{-1}\left|\sum_{t=1}^{T} Y_t e^{-i\lambda_k t}\right|^2$ คือคาบเวลาของ X_T และ Y_T ตามลำดับ และมี $\lambda_k = 2\pi k/T$ ที่ $k = 1, \dots, n$ โดยที่ $n = [(T \quad 1)/2]$

3) มาตรวัดตาม Values และ Behavior ได้แก่

มาตรวัดที่วัดความคล้ายคลึงของ ค่าที่สังเกตหรือพฤติกรรม ซึ่งในงานนี้ใช้การวัดพฤติกรรม ด้วยการประเมินจากค่าเฉลี่ยของ Temporal Correlation Coefficient (diss.CORT) เมื่อกำหนดให้ $CORT(X_T, Y_T)$ คือความคล้ายคลึงของพฤติกรรมของอนุกรมเวลา X_T และ Y_T ซึ่งนิยามได้ดังสมการที่ 4

$$CORT(X_T, Y_T) = \frac{\sum_{t=1}^{T-1}(X_{t+1} \quad X_t)(Y_{t+1} \quad Y_t)}{\sqrt{\sum_{t=1}^{T-1}(X_{t+1} \quad X_t)^2}\sqrt{\sum_{t=1}^{T-1}(Y_{t+1} \quad Y_t)^2}} \qquad (4)$$

โดย $CORT(X_T, Y_T)$ จะมีค่าระหว่าง -1 ถึง 1 นั่นคือ หากมีค่าเป็น 1 แสดงว่าอนุกรมมีพฤติกรรมทิศทางเดียวกัน มีค่า -1 จะมีพฤติกรรมสวนทางกัน และมีค่า 0 แสดงว่ามีพฤติกรรมต่างกัน

4) มาตรวัดตาม Autocorrelations มีรูปแบบทั้ง Simple และ Partial ได้แก่ Autocorrelation-based Distance (diss.ACF) นิยามได้ดังสมการที่ 5

$$d_{ACF}(X_T, Y_T) = \sqrt{(\hat{\rho}_{X_T} \quad \hat{\rho}_{Y_T})^T \Omega (\hat{\rho}_{X_T} \quad \hat{\rho}_{Y_T})} \quad (5)$$

เมื่อ $\hat{\rho}_{X_T} = (\hat{\rho}_{1,X_T}, \dots, \hat{\rho}_{L,X_T})^T$, $\hat{\rho}_{Y_T} = (\hat{\rho}_{1,Y_T}, \dots, \hat{\rho}_{L,Y_T})^T$ เป็นเวกเตอร์ค่าประมาณการ Autocorrelation ของ X_T และ Y_T และมี Ω เป็นเมตริกซ์ค่าน้ำหนัก อ่านเพิ่มเติมจาก [31]

5) มาตรวัดตาม Complexity ได้แก่

มาตรวัด Permutation Distribution Clustering (diss.PDC) เป็นมาตรวัดที่อาศัยของการกระจายของลำดับรูปแบบ (Permutation Distribution: PD) ที่มีขนาด m-embedding และ t-time delay ของข้อมูลดังเดิม X_T ซึ่งสร้างได้จาก $x'_m \equiv \{X'_m = (X_t, X_{t+1}, \dots, X_{t+m}), t = 1, \dots T \quad m\}$ จากนั้นวัดความคล้ายคลึงโดยอาศัยหลักการวัดแบบ Euclidean Distance

มาตรวัด Complexity-Invariant Dissimilarity (diss.CID) เป็นมาตรวัดที่อาศัยความคล้ายคลึง/ความต่างของอนุกรมเวลาในระดับที่สูงขึ้นโดยดูที่ความซับซ้อนของอนุกรมเวลา ซึ่ง Complexity-Invariant Dissimilarity นิยามได้ดังสมการที่ 6

$$d_{CID}(X_T, Y_T) = CF(X_T, Y_T) \;\; d(X_T, Y_T) \quad (6)$$

โดย $CE(X_T)$ คือ ตัวประมาณความซับซ้อนของ X_T

$CF(X_T, Y_T)$ คือ ปัจจัยการปรับความซับซ้อน ดังสมการที่ 7

$$CF(X_T, Y_T) = \frac{\max\{CE(X_T), CE(Y_T)\}}{\min\{CE(X_T), CE(Y_T)\}} \quad (7)$$

2.3 อัลกอริทึมแบบ เค-มีนส์

การจัดกลุ่มข้อมูลด้วยอัลกอริทึม k-Means (k-Means Algorithm) เป็นเทคนิคการจัดกลุ่มที่มีหลักการทำงานที่ง่าย ไม่ซับซ้อน ประมวลผลเร็ว [32], [33] สำหรับอัลกอริทึมในการจัดกลุ่มแบบ k-Means มีขั้นตอนการทำงาน 5 ขั้นตอน [33] ดัง (Table 1)

Table 1. The Five steps of k-Means Algorithm.

	Algorithm k-Means
1	Decide on a values for k.
2	Initialize k cluster centers (randomly, if necessary).
3	Decide the class memberships of the N objects by assigning them to the nearest cluster center.
4	Re-estimate the k cluster centers, by assuming the memberships found above are correct.
5	If none of the N object s changed membership in the last iteration, exit. Otherwise goto 3.

จากขั้นตอนในอัลกอริทึม k-Means ใน (Table 1) มีการทำงานหลัก 5 ขั้นตอนได้แก่ 1) การกำหนดค่า k เพื่อใช้สำหรับแบ่งกลุ่ม 2) การกำหนดจุดศูนย์กลางของกลุ่มจำนวน k จุด ซึ่งในรอบแรกอาจได้จากการสุ่มรอบถัดไปจะได้จากการคำนวณ 3) จัดข้อมูลให้สังกัดกลุ่มโดยพิจารณาให้สังกัดในกลุ่มที่ข้อมูลอยู่ใกล้จุดศูนย์กลางของกลุ่มนั้นมากที่สุด 4) หาจุดศูนย์กลางของแต่ละกลุ่มใหม่อีกครั้ง โดยใช้ค่าเฉลี่ยจากสมาชิกที่ถูกกำหนดให้สังกัดในกลุ่มนั้น ๆ และขั้นตอนที่ 5) หากสมาชิกในกลุ่มไม่มีการเปลี่ยนแปลง หรือจุดศูนย์กลางไม่เปลี่ยน ให้จบการทำงาน หรือไม่ให้เริ่มทำซ้ำขั้นตอนที่ 3

2.4 อัลกอริทึมแบบ เค-มีดอยส์

การจัดกลุ่มข้อมูลด้วยอัลกอริทึม k-Medoids เป็นวิธีการจัดกลุ่มที่อาศัยหลักการคล้ายคลึงกับ k-Means แตกต่างกันที่ k-Medoids จะกำหนดจุดศูนย์กลางของกลุ่มจากข้อมูลที่มีอยู่จริงในชุดข้อมูลนำเข้า และวิธีการวัดระยะหรือวัดความคล้ายคลึงโดย k-Medoids ใช้หลักการวัดระยะของ มินโคว์สกีที่มีค่า g=1 (r=1) ซึ่งก็คือหลักการของการวัดระยะทางแบบแมนฮัตตันนั่นเอง การจัดกลุ่มด้วยอัลกอริทึม k-Medoids ที่เหมาะสมคือ

การจัดกลุ่มที่มีค่าใช้จ่ายสุทธิ (Total Cost) ต่ำที่สุด โดยทั่วไป k-Medoids มักจะรู้จักกันดีในชื่อ PAM (Partitioning Around Medoids) Algorithm [32], [34], [35], [36] สำหรับการทำงานของอัลกอริทึม k-Medoids มี 6 ขั้นตอน [34] ดัง (Table 2)

Table 2. The Six steps of k-Medoids Algorithm.

	Algorithm k-Medoids
1	Decide on a values for *k*.
2	Select *k* points as the initial representative objects.
3	Assign each point to the cluster with the nearest representative object.
4	Randomly select a non-representative, object $\boldsymbol{x_i}$.
5	Compute the total cost referencing S of swapping the representative object $\boldsymbol{C}$ with $\boldsymbol{x_i}$.
6	If $S < 0$, then swap $\boldsymbol{C}$ with $\boldsymbol{x_i}$ to form the new set of *k* representative objects and goto 3. Otherwise exit.

จากขั้นตอนในอัลกอริทึม k-Medoids ใน (Table 2) มีการทำงานหลัก 6 ขั้นตอนได้แก่ 1) กำหนดค่า k เพื่อใช้สำหรับแบ่งกลุ่ม 2) การกำหนดให้จุด k จุดเป็นตัวแทนกลุ่ม 3) จัดข้อมูลให้สังกัดกลุ่มโดยพิจารณาให้สังกัดในกลุ่มที่ข้อมูลอยู่ใกล้จุดศูนย์กลางของกลุ่มนั้นมากที่สุด 4) สุ่มหาข้อมูลใหม่ที่ไม่ใช่จุดศูนย์กลาง เพื่อใช้เป็นจุดศูนย์กลางใหม่ 5) คำนวณ Total Cost และผลต่าง Total Cost ระหว่างจุดศูนย์กลางก่อนหน้า และจุดที่สุ่มใหม่ 6) หาผลต่าง Total Cost $S < 0$ ให้เปลี่ยนจุดศูนย์กลางเป็นจุดสุ่มใหม่ แล้วไปขั้นตอนที่ 3 ต่อ หรือไม่ก็จบการทำงาน

2.5 อัลกอริทึมการจัดกลุ่มแบบลำดับชั้น

เป็นเทคนิคการจัดกลุ่มที่มีจุดเด่นในการแสดงภาพผลการจัดกลุ่ม [37] ในรูปแบบของแผนภาพต้นไม้ โดยเทคนิควิธีนี้สามารถแบ่งได้เป็น 2 แบบได้แก่ Agglomerative Hierarchical Clustering (Bottom-Up) และแบบ Divisive Hierarchical Clustering (Top-Down) [34] วิธีที่นิยมคือ Agglomerative เป็นเทคนิคการจัดกลุ่มซึ่งจัดเป็นวิธีที่แบ่งกลุ่มได้ดีมีความแม่นยำสูง [38] แต่มีข้อจำกัดคือเหมาะกับชุดข้อมูลขนาดไม่ใหญ่ ขั้นตอนการทำงานของอัลกอริทึมการจัดกลุ่มแบบลำดับขั้น [33] แสดงดัง (Table 3)

Table 3. The Five steps of Hierarchical Clustering Algorithm.

	Algorithm Hierarchical Clustering
1	Calculate the distance between all object. Store the results in a distance matrix.
2	Search through the distance matrix and find the two most similar clusters/objects.
3	Join the two clusters/objects to produce a cluster that now has at least 2 objects.
4	Update the matrix by calculating the distances between this new cluster and all other clusters.
5	Repeat step 2 until all cased are in one cluster

จากขั้นตอนใน (Table 3) มีการทำงานหลัก 5 ขั้นตอน ได้แก่ 1) คำนวณระยะระหว่างวัตถุแต่ละตัวและเก็บไว้เป็นเมตริกซ์ 2) ตรวจสอบระยะห่างระหว่างกลุ่ม เพื่อหากลุ่มที่ใกล้ชิดกันมากที่สุดทีละคู่ 3) จับคู่กลุ่มเพื่อจัดให้เป็นกลุ่มเดียวกันทีละคู่ 4) ปรับปรุงเมตริกซ์ระยะห่างระหว่างกลุ่มใหม่ และกลุ่มอื่นทุกกลุ่ม 5) ทำซ้ำตั้งแต่ขั้นตอนที่ 2 จนกระทั่งกลุ่มทุกกลุ่มรวมกันเป็นกลุ่มเดียว จึงจบการทำงาน

2.6 การจัดกลุ่มด้วยวิธี พีดีซี

การจัดกลุ่มการกระจายของการเปลี่ยนแปลง (Permutation Distribution Clustering: PDC) เสนอโดย[39], [40] เป็นวิธีการจัดกลุ่มข้อมูลอนุกรมเวลาประเภทอาศัยความซับซ้อนเป็นฐาน (Complexity-based Approach) ขั้นตอนการทำงานใช้หลักการของอัลกอริทึมการจัดกลุ่มแบบ Agglomerative Hierarchical Clustering เป็นพื้นฐาน [39] โดยนำเสนอหลักการสำคัญสำหรับใช้ในการจัดกลุ่มข้อมูลอนุกรมเวลา ดังนี้

1) การแทนอนุกรมเวลาด้วย รูปแบบของการกระจายการเปลี่ยนแปลง (Permutation Distribution: PD) ซึ่ง PD ของ X’ นิยามได้จาการนับความถี่ของช่วงของรูปแบบ (Patterns) ที่ไม่ซ้ำกันของสมาชิกลำดับย่อยใน x’ ดังสมการที่ 8

$$p_\pi = \frac{\#\{x' \in X' | \Pi(x') = \pi\}}{T'} \quad (8)$$

เมื่อ X' คือ ชุดของข้อมูลอนุกรมเวลาที่ถูกจัดรูปแบบความน่าจะเป็นของรูปแบบการกระจายการเปลี่ยนแปลง x' แล้ว

2) วัดความคล้ายระหว่าง PD ด้วย Squared Hellinger Distance เมื่อ P และ Q แทน PD สองชุด ดังนั้น Squared Hellinger Distance นิยามได้ดังสมการ 9

$$D(P,Q) = \frac{1}{\sqrt{2}} \left\| \sqrt{P} \quad \sqrt{Q} \right\|_2^2 \quad (9)$$

2.7 ซิลลูเอ็ต (Silhouette)

การหาคำตอบว่าควรแบ่งข้อมูลเป็นกี่กลุ่ม (ค่า k) ซึ่งเป็นองค์ประกอบพื้นฐานหลักของวิธีการจัดกลุ่ม [41], [42] ถึงแม้ว่าในบางเทคนิคก็ไม่จำเป็น และมีการพัฒนาอัลกอริทึมที่ไม่ต้องกำหนดค่า k ก่อนแล้วก็ตาม แต่ในที่สุดผู้ใช้ย่อมต้องการระบุว่าจะแบ่งเป็นกี่กลุ่มจึงจะเหมาะสมสำหรับบางข้อมูล ซึ่ง ซิลลูเอ็ต เป็นมาตรวัดที่อาศัยทั้งการยึดเหนี่ยวภายใน และการแยกกันระหว่างกลุ่ม โดยใช้ค่าเฉลี่ยของระยะห่างระหว่างจุดกับกลุ่มที่อยู่ใกล้ชิด เทียบกับจุดที่อยู่ภายในกลุ่มเดียวกัน [43] ดัง (Figure 2) แสดงตัวอย่างการวัดระยะสำหรับหาค่า Silhouette

สำหรับแต่ละวัตถุ x_i เราสามารถคำนวณหา Silhouette Coefficient (s_i) ได้ดังสมการที่ 10 ดังนี้

$$s_i = \frac{\mu_{out}^{min}(x_i) \quad \mu_{in}(x_i)}{\max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}} \quad (10)$$

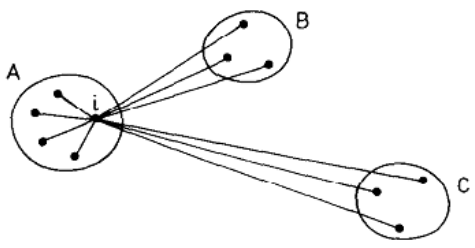


Figure 2. The Computation a Silhouette of *i* [13].

โดยที่ $\mu_{in}(x_i)$ คือค่าเฉลี่ยของระยะระหว่างจุด x_i กับจุดอื่นภายในกลุ่ม $\hat{y}_i$ ซึ่งเป็นกลุ่มของตัวมันเองซึ่งหาได้ดังนี้

$$\mu_{in}(x_i) = \frac{\sum_{x_j \in C_{\hat{y}_i}, j \neq i} \delta(x_i, x_j)}{n_{\hat{y}_i} \quad 1} \quad (11)$$

และ $\mu_{out}^{min}(x_i)$ คือค่าเฉลี่ยของระยะระหว่างจุด x_i กับจุดอื่นในกลุ่มที่อยู่ใกล้ชิด หาได้จากสมการที่ 12 ดังนี้

$$\mu_{out}^{min}(x_i) = \min_{j \neq \hat{y}_i} \left\{ \frac{\sum_{y \in C_j} \delta(x_i, y)}{n_j} \right\} \quad (12)$$

ดังนั้นค่า Silhouette ของแต่ละกลุ่มนิยามได้จากค่าเฉลี่ยของ s_i จากทุกข้อมูลภายในกลุ่ม ดังสมการที่ 13 ดังนี้

$$SC = \frac{1}{n} \sum_{i=1}^{n} s_i \quad (13)$$

โดยที่ n คือ จำนวนข้อมูลทั้งหมด ซึ่งจะได้ว่าหากค่า SC มีค่ามาก ๆ เข้าใกล้ +1 แสดงว่าเป็นการจัดกลุ่มที่ดี

2.8 ผลรวมความผิดพลาด

ผลรวมความผิดพลาด คือ เครื่องมือตรวจความเหมาะสมในการจัดกลุ่ม โดยการพิจารณาจากค่าผลรวมความผิดพลาด (Sum of Squared Error: SSE) ในการแบ่งวัตถุให้อยู่ในกลุ่ม โดยเฉพาะในกรณี Unsupervised Learning จำเป็นต้องมีค่าคะแนนบางอย่างเป็นตัวชี้วัดหรือประเมินว่ากลุ่มที่จัดแบ่งให้กับวัตถุแต่ละตัวนั้นเหมาะสมดีแล้วหรือไม่ ซึ่งในที่นี้จะใช้ผลรวมของค่าความผิดพลาด (Sum of Squared Errors: SSE) นิยามได้ดังสมการที่ 14 [43]

$$SSE(C) = \sum_{i=1}^{k} \sum_{x_j \epsilon C_i} \left\| x_j \quad \mu_i \right\|^2 \quad (14)$$

เมื่อ x_j คือ ข้อมูลใด ๆ ตัวที่ j

μ_i คือ ค่าเฉลี่ยของกลุ่ม i

SSE เป็นเครื่องมือหนึ่งที่ใช้ในการวัดคุณภาพของการจัดกลุ่มที่ให้ผลดีไม่แพ้กันและยังเป็นที่รู้จักอย่างกว้างขวาง ในการพิจารณาค่า SSE ให้สังเกตกราฟที่สร้างกราฟจากความสัมพันธ์ระหว่าง SSE กับค่า k ที่กำหนด ณ จุดเปลี่ยนความชัน (Significant Local Change) หรือจุดที่มีลักษณะหัวเข่า “knee” (Significant “knee”) ซึ่งเป็นตำแหน่งที่สามารถบ่งชี้จำนวนกลุ่มที่เหมาะสม ในการจัดกลุ่มได้จากการศึกษาของ Tippaya และคณะ [14] เกี่ยวกับการใช้ SSE เพื่อตรวจสอบหาจำนวนกลุ่มที่เหมาะสมสำหรับการจัดกลุ่มข้อมูลสรุปไว้ว่าควรเลือก k ที่มีมีอัตราการเปลี่ยนแปลงสูงสุด โดยอัตราการเปลี่ยนแปลง (%Change) นิยามได้ดังสมาการที่ 15 [14]

$$\%C\ ange = \frac{(SSEofK_{i-1} \quad SSEofK_i) \quad 100}{SSEofK_{i-1}} \qquad (15)$$

ตัวอย่างการแสดงกราฟความสัมพันธ์ระหว่างค่าk และค่า SSE ณ จุดเปลี่ยนความชัน แสดงดัง (Figure 3)

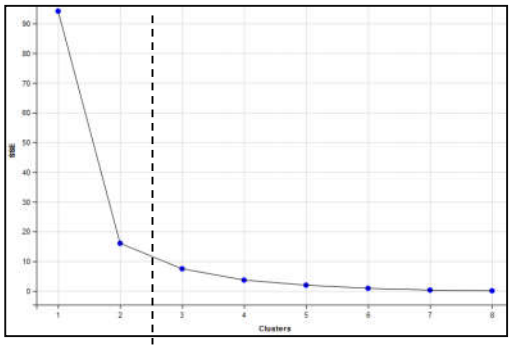


Figure 3. A Relationship Graph for k value and SSE.

จาก (Figure 3) แสดงการตรวจสอบ k ที่เหมาะสมสำหรับการจัดกลุ่มโดยทดลองจัดกลุ่มโดยกำหนดค่า k ให้มีค่าตั้งแต่ 2 ถึง 8 แล้วนำมาสร้างเป็นกราฟแสดงความสัมพันธ์ระหว่างค่า k และค่า SSEสรุปได้ว่าการจัดกลุ่มนี้ควรกำหนดค่า k =2 เนื่องจาก ณ k =2 เป็นตำแหน่งที่เป็นจุดเปลี่ยนความชัน (มี %Change มีค่าสูงที่สุด)

2.9 ชุดข้อมูลสำหรับการทดลอง

ข้อมูลที่ใช้ในการวิจัยนี้มี 3 ชุดข้อมูลได้แก่ ข้อมูล ECG ข้อมูลน้ำท่า และข้อมูลอนุกรมเวลาสังเคราะห์ รายละเอียดดังนี้

1) ข้อมูล ECG (Electrocardiogram) หรือคลื่นไฟฟ้าหัวใจที่ได้จาก [44], [45] เป็นข้อมูลจากการวินิจฉัยคลื่นไฟฟ้าหัวใจ ซึ่งแต่ละอนุกรมมาจากแต่ละขั้วไฟฟ้าในช่วงหนึ่งการเต้นของหัวใจผ่านการวินิจฉัยจากผู้เชี่ยวชาญ ข้อมูลใช้ในการวิจัยครั้งนี้มีขนาด 96 จุดเวลา มี 2 คลาสคือ ปกติ (Normal) จำนวน 133 อนุกรม และไม่ปกติ (Abnormal) 67 อนุกรม รวมทั้งหมด 200 อนุกรม

2) ข้อมูลน้ำท่าจากกรมชลประทาน [46] (Runoff) เป็นข้อมูลรายเดือนที่จัดเก็บตลอดทั้งปีจาก 9 สถานีตรวจวัดในเขต 3 จังหวัด ได้แก่ นครราชสีมา อุบลราชธานี และศรีษะเกษ ระหว่างปี ค.ศ. 2005-2014 รายละเอียดดัง (Table 4)

Table 4. The Details of Runoff Dataset.

Station	Catchment	District	Province
M2A	แม่น้ำมูล	เฉลิมพระเกียรติ	นครราชสีมา
M43A	ลำตะคอง	สีคิ้ว	นครราชสีมา
M145	ลำพระเพลิง	ปากช่อง	นครราชสีมา
M69	ลำเชบก	กันทรลักษ์	ศรีสะเกษ
M5	แม่น้ำมูล	ราศีไศล	ศรีสะเกษ
M9	ห้วยสำราญ	เมือง	ศรีสะเกษ
M7	แม่น้ำมูล	เมือง	อุบลราชธานี
M127	ห้วยตาเทียว	ตระการพืชผล	อุบลราชธานี

3) ข้อมูลสังเคราะห์ (Synthetic.tseries) หรือข้อมูลอนุกรมเวลาสังเคราะห์ ได้จาก [31] โดยข้อมูลสังเคราะห์จาก 6 โมเดลที่แตกต่างกันโมเดลละ 3 ชุด ที่มีทั้งแบบเชิงเส้นและไม่เชิงเส้น รายละเอียดแต่ละโมเดล แสดงดัง (Table 5)

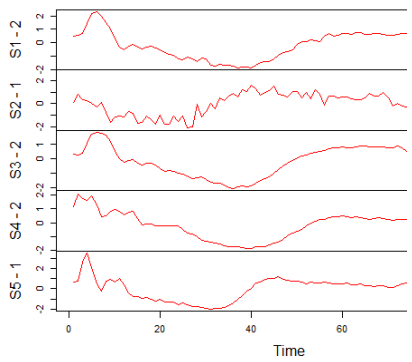
Table 5. The Details of the models of Synthetic Dataset. [31]

Name	Models
AR	$X_t = 0.6X_{t-1} + \varepsilon_t$
Bilinear	$X_t = (0.3 \quad 0.2\varepsilon_{t-1})X_{t-1} + 1.0 + \varepsilon_t$
EXPAR	$X_t = (0.9\exp(\ X_{t-1}^2) \quad 0.6)X_{t-1} + 1.0 + \varepsilon_t$
SETAR	$X_t = (0.3X_{t-1} + 1)I(X_{t-1} \geq 0.2) \quad (0.3X_{t-1} + 1)I(X_{t-1} < 0.2) + \varepsilon_t$
NLAR	$X_t = 0.7\lvert X_{t-1}\rvert(2 + \ X_{t-1}\ )^{-1} + \varepsilon_t$
STAR	$X_t = 0.8X_{t-1} \quad 0.8X_{t-1}\big(1 + \exp(\ 10X_{t-1})\big)^{-1} + \varepsilon_t$

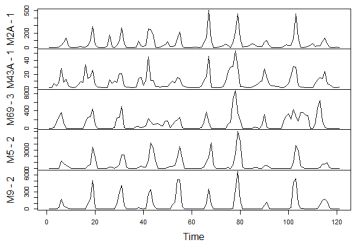
สำหรับรายละเอียดของข้อมูลทั้ง 3 ชุดข้อมูล เกี่ยวกับขนาดเวลา จำนวนอนุกรม และจำนวนคลาส แสดงดัง (Table 6) และแสดงตัวอย่างรูปร่างอนุกรมเวลาของข้อมูล ECG ข้อมูล Runoff และข้อมูลสังเคราะห์ดัง (Figure 4) ที่ (a), (b) และ (c) ตามลำดับ

Table 6. The Dataset details for the Experimental.

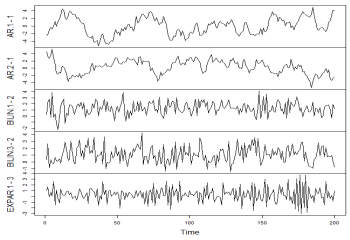
Dataset	Length	Samples	Class
ECG200	96	200	2
Runoff	120	9	-
Synthetic	200	18	6



(a) ECG Dataset



(b) Runoff Dataset



(c) Synthetic Time-series

Figure 4. An Example for Shape of Time-series.

จาก (Figure 4) ที่ (a) เป็นตัวอย่างรูปร่างอนุกรมข้อมูล ECG ซึ่งเป็นข้อมูลที่ทราบกลุ่มที่แท้จริง ซึ่งมี 2 กลุ่มได้แก่กลุ่ม 1 คืออนุกรม S2 และ S5 กลุ่ม 2 คือ S1, S3 และ S4 ซึ่งพบว่าข้อมูล ECG ค่อนข้างเป็นข้อมูลที่มีรูปร่างที่แตกต่างกันชัดเจนระหว่างสองกลุ่ม ใน (b) เป็นตัวอย่างรูปร่างอนุกรมข้อมูล Runoff ซึ่งเป็นข้อมูลที่ไม่ทราบกลุ่มที่แท้จริง แต่หากจะพิจารณาตามเขตจังหวัดก็ยากต่อการจัดกลุ่มตัวอย่างเช่น สถานี M43A และ M2A เป็นข้อมูลจากสถานีในจังหวัดเดียวกัน แต่มีรูปร่างแตกต่าง และยังไปคล้ายคลึงกับสถานี M69 ซึ่งเป็นสถานีในจังหวัดหมายเลข 3 ส่วน (c) เป็นตัวอย่างอนุกรมเวลาจากข้อมูลสังเคราะห์ เป็นอนุกรมที่มีความซับซ้อนของรูปร่าง แต่สังเคราะห์จากโมเดลที่แตกต่าง จึงค่อนข้างจำแนกได้ง่ายกว่าเมื่อเทียบที่ความซับซ้อน

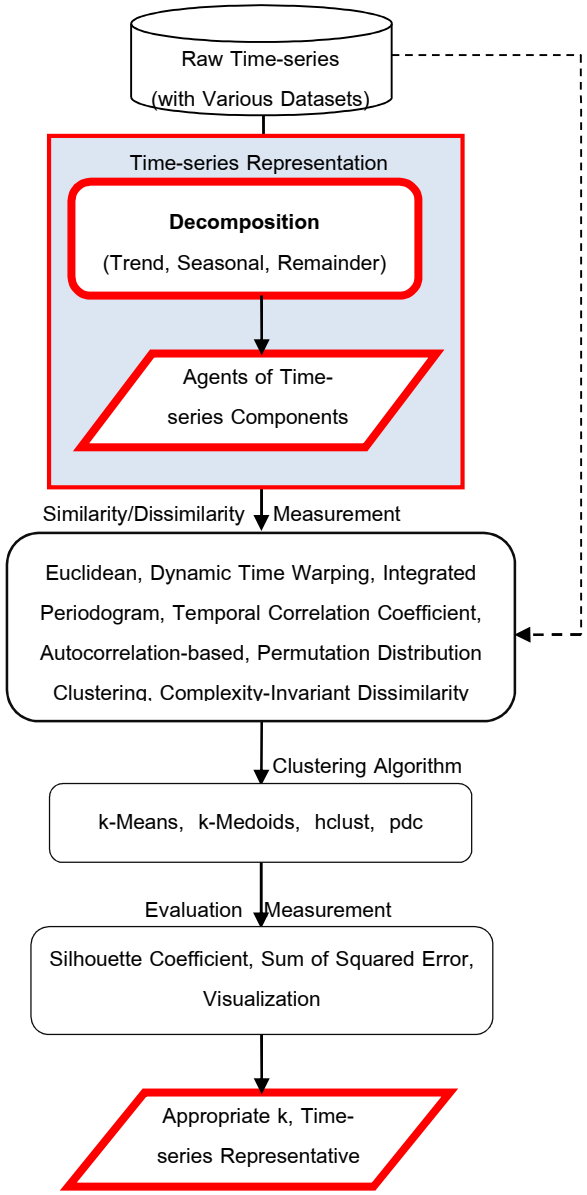


Figure 5 The Proposed of the Research Framework.

2.10 วิธีดำเนินการวิจัย

เพื่อเป็นการตรวจสอบแนวคิดที่นำเสนองานวิจัยนี้ได้ออกแบบกรอบในการดำเนินงานวิจัย โดยมีขั้นตอนการทำงาน (Figure 5) ซึ่งสรุปขั้นตอนได้ดังนี้

1) นำเข้าข้อมูลอนุกรมเวลาดั้งเดิม (Raw Time-series)

2) ขั้นตอนการแทนอนุกรมเวลา โดยอาศัยแนวคิดการแยกองค์ประกอบและวิเคราะห์หาตัวแทนองค์ประกอบที่เหมาะสม (งานวิจัยนี้นำเสนอ)

3) เมื่อได้ตัวแทนองค์ประกอบแล้ว กำหนดให้ตัวแทนองค์ประกอบเป็นตัวแทนอนุกรมเวลาสำหรับการจัดกลุ่ม

4) วัดความคล้ายคลึง/ความต่าง ระหว่างตัวแทนอนุกรมเวลาที่ได้จากแนวคิดที่นำเสนอ ด้วย 6 มาตรวัดได้แก่ Euclidean Distance, Dynamic Time Warping, Integrated Periodogram, Temporal Correlation Coefficient, Autocorrelation-based, Permutation Distribution Clustering, Complexity-Invariant Dissimilarity

5) จัดกลุ่มข้อมูลอนุกรมเวลาด้วย 4 เทคนิค ได้แก่ k-Means, k-Medoids, Hierarchical Clustering และ PDC

6) ประเมินผลการจัดกลุ่ม ด้วยค่าซิลลูเอ็ต ค่าผลรวมความผิดพลาด และตรวจสอบความคล้ายคลึงของอนุกรมจากแผนภาพผลลัพธ์ของการจัดกลุ่ม

7) ผลลัพธ์ท้ายสุดของการจัดกลุ่มสำหรับงานวิจัยนี้ คือ ตัวแทนข้อมูลอนุกรมเวลาที่ให้ค่า k ที่เหมาะสมสำหรับการจัดกลุ่ม

2.11 การกำหนดตัวแทนอนุกรมเวลา

ในงานวิจัยนี้ได้นำเสนอวิธีการแทนอนุกรมเวลาโดยวิเคราะห์หาตัวแทนองค์ประกอบอนุกรมเวลา เพื่อช่วยลดรูปแบบที่ไม่สำคัญ หรือเพิ่มความโดดเด่นชัดเจนให้กับรูปร่างของ Raw Data และวิเคราะห์หาตัวแทนขององค์ประกอบที่สามารถกำหนดเป็นตัวแทนอนุกรมเวลาได้ โดยแนวคิดของวิธีการที่นำเสนอแสดงได้ดัง (Figure 5) (ภายในสี่เหลี่ยมทึบ) มีขั้นตอนการทำงานดังต่อไปนี้

1) แยกองค์ประกอบของอนุกรมเวลา กำหนดให้ X_T คืออนุกรมเวลาที่มีขนาดเวลา เป็น T จะได้องค์ประกอบของอนุกรมเวลาดังต่อไปนี้

Trend(X_T) = องค์ประกอบที่เป็นค่าแนวโน้ม ซึ่งเบื้องต้นได้จากการคำนวณด้วย Moving Average

Seasonal(X_T) = องค์ประกอบที่เป็นฤดูกาลได้จาก

$= X_T - Trend(X_T)$

Random(X_T) = องค์ประกอบส่วนที่เหลือจาก X_T แยก Trend(X_T) และ Seasonal(X_T) ออกไป

$= (X_T - Trend(X_T)) - Seasonal(X_T)$

จากตัวอย่างองค์ประกอบของอนุกรมเวลา ข้อมูล Runoff ใน (Figure 6) ซึ่งแสดงองค์ประกอบของข้อมูล Runoff โดยแสดง Trend, Seasonal, Random (Remainder Part) และ Agent ใน(a), (b), (c) และ (d) ตามลำดับ

2) วิเคราะห์หาตัวแทนองค์ประกอบอนุกรมเวลาที่เหมาะสม เพื่อใช้เป็นตัวแทนข้อมูลอนุกรมเวลาทั้งชุด โดยการทดสอบความสามารถในการแบ่งแยกขององค์ประกอบแต่ละตัว จากนั้น ออกแบบตัวแทนองค์ประกอบ ดังนี้

$Agent1 = Trend(X_T)$

Agent2 = Seasonal(X_T)

$Agent3 = SeasonalAdjust\ (Seasonal(X_T))$

$Agent4 = TrendAdjust(Trend(X_T))$

3) กำหนดตัวแทนองค์ประกอบเพื่อเป็นตัวแทนอนุกรมเวลา ซึ่งทดลองจัดกลุ่มข้อมูลโดยใช้ตัวแทนองค์ประกอบจากขั้นตอนที่ 2 ซึ่งเมื่อกำหนดตัวแทนองค์ประกอบจะได้ผลลัพธ์อนุกรมดัง (Figure 7)

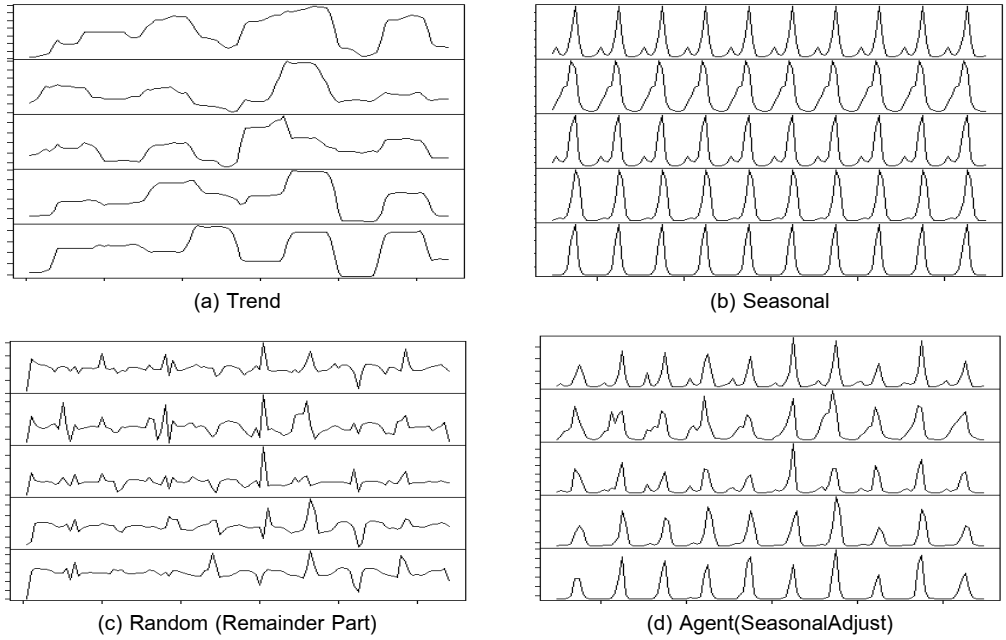


Figure 6. An Example of the Results for Runoff Representative.

3. ผลการวิจัย

สำหรับผลการวิจัย โดยใช้ตัวแทนอนุกรมเวลา และตรวจสอบการจัดกลุ่มตามแนวคิดที่นำเสนอ ได้ผลดังต่อไปนี้

3.1 ผลการกำหนดตัวแทนอนุกรมเวลา

จากการทดลองจัดกลุ่มข้อมูลกับตัวแทนอนุกรมเวลาทั้ง 6 Agents ตัวแทนอนุกรมเวลาที่เหมาะสมที่ได้จากงานวิจัยนี้ คือตัวแทนอนุกรมเวลาที่ให้ผลการทดลองสอดคล้องกันทั้ง 3 ชุดข้อมูล คือตัวแทนองค์ประกอบ Agent3 (ปรับ Seasonal: SeasonalAdjust) โดยตัวแทนอนุกรมเวลาข้อมูล ECG ข้อมูล Runoff และข้อมูลสังเคราะห์ ที่ได้แสดงดัง(Figure 7) ที่ (b), (d) และ (f) ตามลำดับ

จาก (Figure 7) เป็นการแสดงเปรียบเทียบผลลัพธ์ตัวแทนอนุกรมเวลากับข้อมูลดังเดิม (Raw Data) สำหรับข้อมูลแต่ละชุด ซึ่งพบว่าลักษณะของอนุกรมที่มีความชัดเจนขึ้นคือข้อมูล Runoff ดังตัวอย่างตัวแทนอนุกรมของข้อมูล Runoff ที่สถานี M69 และ M34A เป็นต้น ส่วนตัวแทนของ ECG และข้อมูลสังเคราะห์มีการเปลี่ยนแปลงของรูปร่างเพียงเล็กน้อย

3.2 ผลการจัดกลุ่มข้อมูล

เมื่อนำตัวแทนอนุกรมที่ได้จากการวิเคราะห์ไปจัดกลุ่มด้วยเทคนิค 4 อัลกอริทึม ได้แก่ k-Means, Hierarchical Clustering (Hclust), Permutation Distribution Clustering (PDC) และ k-Medoids ซึ่งรายละเอียดของผลการจัดกลุ่มแบ่งออกเป็น 2 ส่วน ดังต่อไปนี้

1) ประเมินค่า k ที่เหมาะสมสำหรับการจัดกลุ่มอนุกรมเวลาจากผลการทดลองเมื่อประเมินจากการพิจารณาค่า k ที่สอดคล้องกันทุกชุดข้อมูล ทั้งจากค่าซิลลูเอ็ด และค่าผลรวมความผิด คือตัวแทนองค์ประกอบ Agent3 (SeasonalAdjust) นั่นคือตัวแทนองค์ประกอบที่ได้จากการปรับแต่ง Seasonal ซึ่งใช้กับวิธีวัดความคล้ายคลึงแบบ PDC ผลการวิจัยแสดงได้ดัง (Table 7-9) เมื่อพิจารณาแต่ละชุดข้อมูลมีรายละเอียดดังนี้

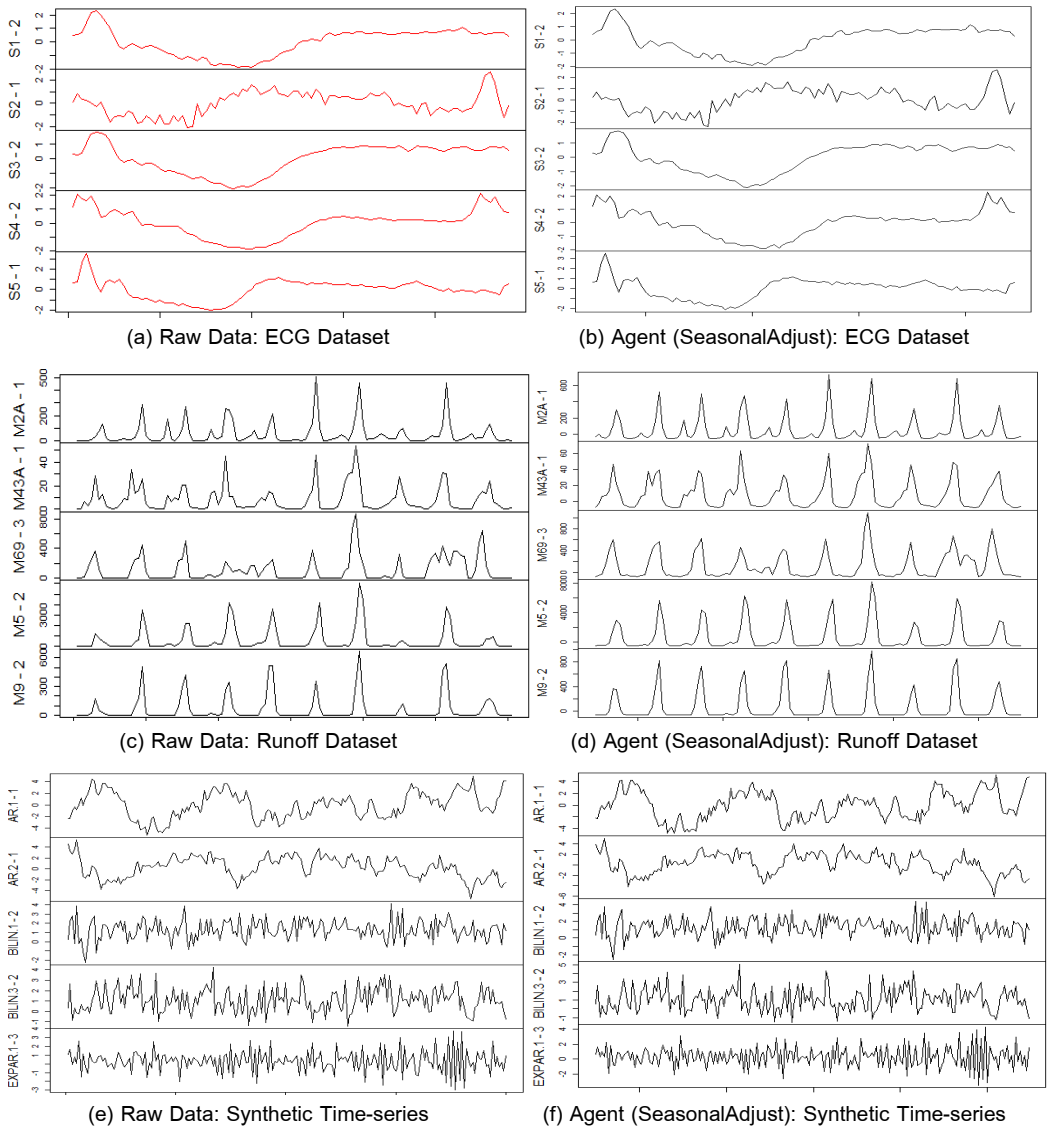


Figure 7. The Comparison of Raw Data and Time-series Representative.

จาก (Table 7) เป็นผลการตรวจสอบความเหมาะสมด้วย ซิลลูเอ็ต และผลรวมความผิดพลาดสำหรับข้อมูล ECG ซึ่งพบว่า Raw Data ให้ค่า k ที่สอดคล้องกันในทุกอัลกอริทึมการจัดกลุ่ม เมื่อวัดความคล้ายคลึงด้วยวิธีแบบ Periodograms (IN.PER), Autocorrelation-based (ACF) และประเภท Complexity-based (ได้แก่ วิธี Complexity-Invariant Dissimilarity: CID และ PDC)

จาก (Table 9) เป็นผลการตรวจสอบสำหรับข้อมูลสังเคราะห์ ซึ่งพบว่า Raw Data ให้ค่า k ที่สอดคล้องกันในทุกอัลกอริทึมการจัดกลุ่มเมื่อวัดความคล้ายคลึงด้วย Raw Data เป็นฐาน และประเภท Temporal Correlation Coefficient (CORT) เมื่อพิจารณาข้อมูล Runoff จาก (Table 8) จะแตกต่างข้อมูลทั้งสองชุดที่กล่าวไปแล้ว คือ Raw Data ของข้อมูล Runoff ให้ผลสอดคล้องกันทั้งค่าซิลลูเอ็ด และ

ค่าผลรวมความผิดพลาดแค่บางอัลกอริทึม และบางวิธีการวัดความคล้ายคลึงเท่านั้น

2) แผนภาพแสดงการจัดกลุ่มอนุกรมเวลา

เมื่อนำตัวแทนองค์ประกอบ Agent3 จากข้อมูลทั้ง 3 ชุดไปจัดกลุ่ม โดยวัดความคล้ายคลึง และจัดกลุ่มข้อมูลอนุกรมเวลาวิธีแบบ PDC โดยเปรียบเทียบกับผลลัพธ์กับการจัดกลุ่ม Raw Data แสดงผลการวิจัยของข้อมูล ECG ได้ดัง (Figure 8) ผลการวิจัยของข้อมูล Runoff ดัง (Figure 9) และผลการวิจัยสำหรับข้อมูลสังเคราะห์ ดัง (Figure 10) ตามลำดับ

3.3 การอภิปรายผล

จากผลการวิจัยสามารถอภิปรายผลได้ดังต่อไปนี้

1) จากผลการวิจัยสำหรับข้อมูล ECG ใช้อนุกรมเวลา 10 อนุกรมในการจัดกลุ่ม ดัง (Figure 8) พบว่าเมื่อนำข้อมูล Raw Time-series ไปจัดกลุ่มโดยกำหนดค่า k=2 ได้ผลลัพธ์ดังรูป (a) และเมื่อตรวจสอบกลุ่มจริงจากรูป (b) จะพบว่ามีการจัดผิดกลุ่ม 2 อนุกรมคือ S5 และ S10 และหากพิจารณาที่อนุกรม S2, S6 และ S9 ที่ถูกจัดให้อยู่ในกลุ่มเดียวกัน จะพบว่ารูปร่างของอนุกรมทั้งสามแตกต่างอย่างเห็นได้ชัด โดยเฉพาะอนุกรม S2

Table 7. The Comparison of k values on Silhouette (sc) and SSE: ECG Dataset.

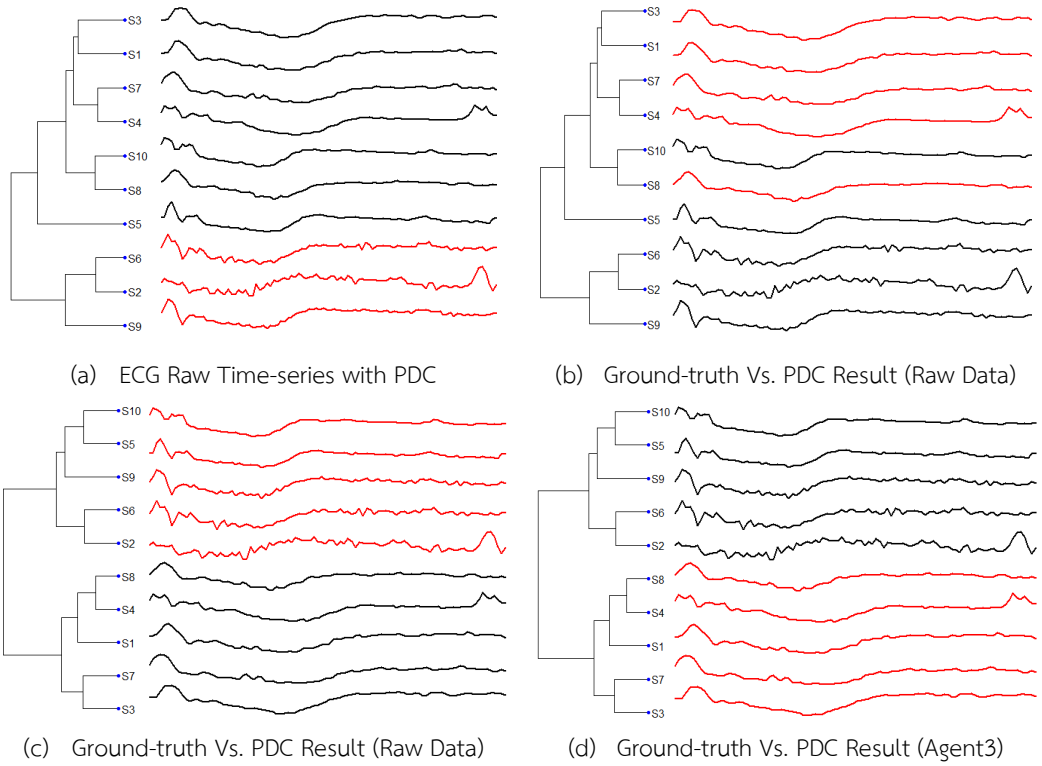
Representation	Distance Measures	EUCL		DTW		IN.PER		ACF		CORT		CID		PDC	
	Clustering Algorithms	sc	sse	sc	sse	sc	sse	sc	sse	sc	sse	sc	sse	sc	sse
Raw Data	k-Means	3	2	3	2	2	2	2	2	3	2	2	2	2	2
	Hclust	5	2	2	2	2	2	2	2	5	3	2	2	2	2
	PDC	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	k-Medoids	3	2	3	2	2	2	2	2	3	2	2	2	2	2
SeasonalAdjust	k-Means	4	2	2	2	2	2	2	2	3	3	3	3	2	2
	Hclust	5	2	2	2	2	3	2	2	5	3	2	2	2	2
	PDC	5	2	2	2	2	2	2	2	2	2	2	2	2	2
	k-Medoids	3	2	3	2	2	2	2	2	3	2	2	2	2	2

Table 8. The Comparison of k values on Silhouette (sc) and SSE: Runoff Dataset.

Representation	Distance Measures	EUCL		DTW		IN.PER		ACF		CORT		CID		PDC	
	Clustering Algorithms	sc	sse	sc	sse	sc	sse	sc	sse	sc	sse	sc	sse	sc	sse
Raw Data	k-Means	2	3	2	3	2	7	2	8	3	3	3	3	8	7
	Hclust	2	3	2	3	2	2	2	8	2	2	2	2	2	2
	PDC	8	6	8	6	2	8	2	7	8	6	8	3	2	2
	k-Medoids	2	3	2	3	2	2	2	8	2	2	2	2	2	2
SeasonalAdjust	k-Means	2	3	2	8	2	7	2	7	3	3	3	3	7	7
	Hclust	2	3	2	3	2	2	2	8	2	2	2	2	2	2
	PDC	8	8	8	8	2	8	8	8	8	8	8	8	2	2
	k-Medoids	2	3	2	3	2	2	2	8	2	2	2	2	2	2

Table 9. The Comparison of k values on Silhouette (sc) and SSE: Synthetic Dataset.

Representation	Distance Measures	EUCL		DTW		IN.PER		ACF		CORT		CID		PDC	
	Clustering Algorithms	sc	sse	sc	sse	sc	sse	sc	sse	sc	sse	sc	sse	sc	sse
Raw Data	k-Means	2	2	2	2	2	2	3	8	2	2	5	2	7	2
	Hclust	2	2	2	2	3	3	2	2	2	2	2	2	2	2
	PDC	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	k-Medoids	2	2	2	2	4	2	2	2	2	2	2	2	2	2
SeasonalAdjust	k-Means	2	2	2	2	4	4	7	3	2	2	2	2	2	2
	Hclust	2	2	2	2	4	2	2	2	5	2	2	2	2	2
	PDC	2	2	2	2	2	2	2	6	2	2	2	2	2	2
	k-Medoids	2	2	2	2	4	2	2	2	5	2	3	2	2	2



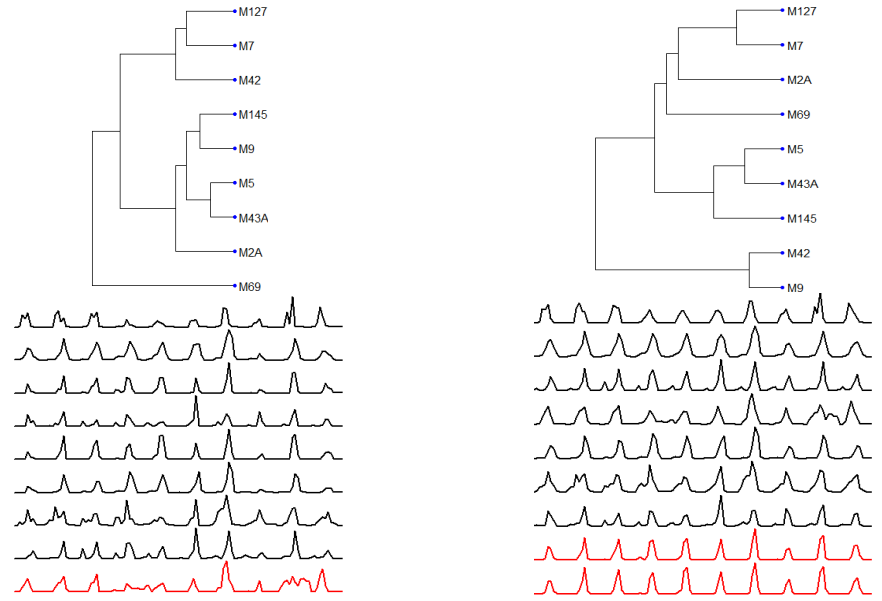
(a) ECG Raw Time-series with PDC

(b) Ground-truth Vs. PDC Result (Raw Data)

(c) Ground-truth Vs. PDC Result (Raw Data)

(d) Ground-truth Vs. PDC Result (Agent3)

Figure 8. The Comparison of the Clustering Results for ECG Dataset and Representative.



(a) Runoff Raw Time-series with PDC (b) Runoff Agent3 with PDC

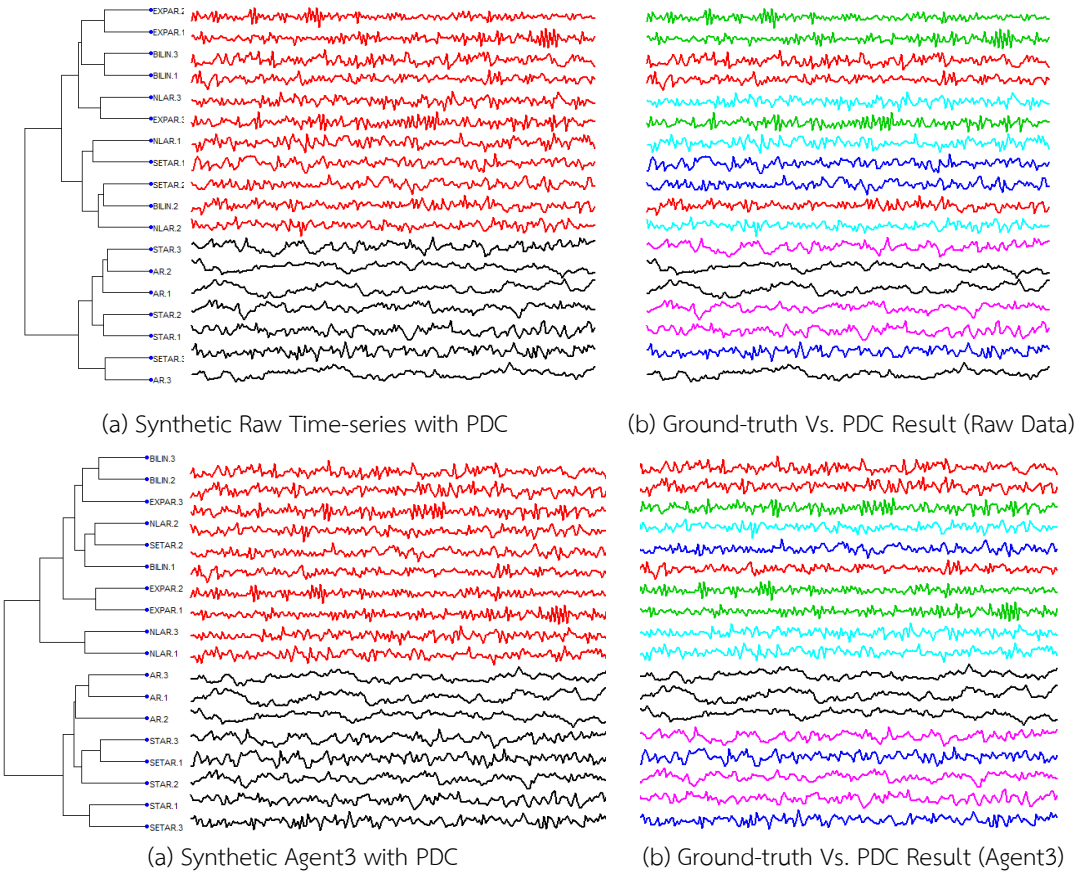
Figure 9. The Comparison of the Clustering Results for Runoff Dataset and Representative.

เมื่อนำตัวแทนอนุกรม Agent3 ไปจัดกลุ่มโดยกำหนดค่า k=2 ได้ผลลัพธ์ดังรูป (c) และเมื่อตรวจสอบกลุ่มจริงจากรูป (d) จะพบว่าจัดกลุ่มได้ถูกต้องตามกลุ่มจริงทุกอนุกรม และหากพิจารณาที่รูปร่างของอนุกรมจะเห็นว่าอนุกรมที่อยู่ในกลุ่มเดียวกันมีความคล้ายคลึงกันอย่างมาก ซึ่งจัดว่าเป็นความคล้ายคลึงกันอย่างมีความหมาย

2) จากผลการวิจัยสำหรับข้อมูล Runoff ในการจัดกลุ่ม ดัง (Figure 9) พบว่าเมื่อจัดกลุ่ม Raw Data ที่ค่า k=2 ได้ผลลัพธ์ดังรูป (a) จะพบว่าอนุกรมที่ถูกจัดกลุ่มแยกออกคือ M69 ซึ่งมีรูปร่างคล้ายคลึงเล็กน้อยกับบางอนุกรม เช่น M7 หรือ M42 แต่พิจารณากลุ่มใหญ่จะพบว่ามีบางอนุกรมที่รูปร่างแตกต่างจากอนุกรมอื่นอย่างมาก เช่น M43A เป็นต้น จึงวิเคราะห์ได้ว่าการจัดกลุ่มนี้อาจยังไม่เหมาะสมเท่าที่ควร

เมื่อนำตัวแทนอนุกรม Agent3 ไปจัดกลุ่มโดยกำหนดค่า k=2 ได้ผลลัพธ์ดัง (b) จะพบว่าอนุกรม M42 และ M9 ถูกจัดกลุ่มให้อยู่ในกลุ่มเดียวกันนั้น มีความคล้ายคลึงกันมากทีเดียว ซึ่งจัดว่าเป็นความคล้ายคลึงกันอย่างมีความหมาย ในขณะที่อนุกรมที่ถูกจัดให้อยู่ในอีกกลุ่มส่วนใหญ่จะมีรูปร่างใกล้เคียงกัน มีบางอนุกรมที่คล้ายคลึงกับกลุ่มแรก เช่น M5 เป็นต้น

3) สำหรับผลการวิจัยในการจัดกลุ่มข้อมูลสังเคราะห์ ดัง (Figure 10) พบว่าเมื่อจัดกลุ่มกับ Raw Data และตัวแทนอนุกรม Agent3 ที่ค่า k=2 ได้ผลลัพธ์ดังรูป (a) และ (c) พบว่าให้ผลลัพธ์ใกล้เคียงกันมาก ต่างกันเพียงอนุกรมเดียวคืออนุกรม SETAR.1 ซึ่งจัดกลุ่มโดยการแทนด้วย Agent3 จะจัด SETAR.1 คล้ายคลึงกับ STAR.3 ในขณะที่จัดกลุ่มกับ Raw Data จะมองว่า SETAR.1 คล้ายคลึงกับ NLAR.1 มากกว่าเมื่อพิจารณาที่อนุกรม AR.1, AR.2 และ AR.3 ในรูป (c) เปรียบเทียบกับกลุ่มจริงในรูป (d) ซึ่งเป็นอนุกรมที่มีกลุ่มจริงอยู่กลุ่มเดียวกัน พบว่าเมื่อใช้ตัวแทนเป็น Agent3 จะให้ผลลัพธ์การจัดกลุ่มได้เหมาะสมกว่าใช้เมื่อ Raw Data ในการจัดกลุ่ม และในการวัดความคล้ายคลึงสำหรับตัวแทนอนุกรมยังถูกมองว่ามีความคล้ายคลึงกันมากที่สุดอีกด้วย



(a) Synthetic Raw Time-series with PDC (b) Ground-truth Vs. PDC Result (Raw Data)

(a) Synthetic Agent3 with PDC (b) Ground-truth Vs. PDC Result (Agent3)

Figure 10. The Comparison of the Clustering Results for Synthetic Dataset and Representative.

4. สรุปและเสนอแนะ

ในงานวิจัยนี้ได้นำเสนอวิธีการสำหรับตรวจสอบความเหมาะสมของการจัดกลุ่มข้อมูลอนุกรมเวลาตามรูปโดยอาศัย ซิลลูเอ็ต และค่าผลรวมความผิดพลาดเป็นพื้นฐาน ร่วมกับการนำเสนอวิธีการแทนอนุกรมเวลาด้วยตัวแทนขององค์ประกอบ โดยทดลองกับข้อมูลสังเคราะห์และข้อมูลจริงจำนวนทั้งหมด 3 ชุดข้อมูล เปรียบเทียบผลการทดลองกับวิธีการวัดความคล้ายคลึง 6 วิธี และจัดกลุ่มด้วยเทคนิคแบบลำดับชั้น ได้แก่การจัดกลุ่มแบบ Agglomerative Hierarchical Clustering และ Permutation Distribution Clustering และแบบแบ่งแยก ได้แก่ k-Means Algorithm และ k-Medoids Algorithm ผลการวิจัยพบว่าวิธีที่นำเสนอสามารถแสดงจำนวนกลุ่มที่เหมาะสมสอดคล้องกันทั้งซิลลูเอ็ต และค่าผลรวมความผิดพลาด เมื่อใช้งานร่วมกับการวัดความคล้ายคลึง/ความต่าง ประเภท Complexity-based โดยทุกชุดข้อมูลจะให้ผลสอดคล้องไปในทิศทางเดียวกัน นอกจากนี้เมื่อตรวจสอบแผนภาพการจัดกลุ่มจะพบว่าอนุกรมเวลาที่ถูกจัดให้อยู่ในกลุ่มเดียวกันจะมีความคล้ายคลึงกันตามรูปมากกว่า ซึ่งสะท้อนให้เห็นถึงการจัดกลุ่มที่มีความคล้ายคลึงอย่างมีความหมาย

แนวทางวิจัยในอนาคต เพื่อช่วยให้สามารถลดมิติของข้อมูล หรือปริมาณการคำนวณความคล้ายคลึงโดยเฉพาะข้อมูลที่มีขนาดใหญ่อย่างเช่น EEG สามารถวิจัยเพื่อค้นหาตัวแทนอนุกรมเวลาที่ลดมิติได้มากขึ้น และวิจัยเพื่อค้นหาวิธีการวัดความคล้ายคลึง ที่เหมาะสมกับตัวแทนอนุกรมเหล่านั้น

กิตติกรรมประกาศ

ขอขอบพระคุณ มหาวิทยาลัยราชภัฏนครราชสีมาที่ได้ให้การสนับสนุนทุนการศึกษา ขอขอบพระคุณโปรแกรมวิชาวิทยาการสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยีที่ให้การสนับสนุนการทำวิจัยให้โอกาสในการศึกษาค้นคว้าข้อมูล รวมถึงขอขอบพระคุณมหาวิทยาลัยเทคโนโลยีสุรนารีที่ให้การสนับสนุนทุนอุดหนุนการวิจัยในครั้งนี้

เอกสารอ้างอิง

[1] Aghabozorgi, S., Shirkorshidi, S. A. and Wan, Y. T. (2015). Time-series clustering-A decade review. **Information Systems**. Vol. 53, pp.16-38.

[2] Niennattrakul, V. (2010). **Meaningful Subsequence Clustering for Time Series Data Stream**. A Dissertation Submitted, Philosophy Program in Computer Engineering, Department of Computer Engineering. Chulalongkorn University.

[3] Lin, J., Keogh, E., Lonardi, S., Lankford, J., Nystrom, D. (2004). Visually mining and monitoring massive time series, in: **Proceedings of 2004 ACMSIGKDD International Conference on Knowledge Discovery and data Mining** – KDD'04, pp. 460.

[4] Keogh, E. and Kasetty, S. (2003). On the need for time series data mining benchmarks: a survey and empirical demonstration, **Data Min. Knowl. Discov**. Vol. 7, no. 4, pp. 349–371.

[5] Haigh, K., Foslien, W. and Guralnik, V. (2004). Visual query language: finding patterns in and relationships among time series data, **Seventh Workshop on Mining Scientific And Engineering Datasets**, pp. 324–332.

[6] Keogh, E., Chu, S. and Hart, D. (2004). Segmenting time series: a survey and novel approach, **Data Min. Time Ser. Databases 57**. Vol. 1, pp. 1–21.

[7] Lin, J., Keogh, E., Lonardi, S. and Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms, in: **Proceedings of 8th ACM SIGMOD Workshop on Research Issues Data Mining and Knowledge Discovery** – DMKD '03, pp. 2.

[8] Zakaria, J., Rotschafer, S., Mueen, A., Razak, K., Keogh, E. (2012). Mining massive archives of mice sounds with symbolized representations, in: **SIGKDD**, pp. 1–10.

[9] Rakthanmanon, T., Campana, A.B., Batista, G., Zakaria, J., Keogh, E. (2012). Searching and mining trillions of time series subsequences under dynamic time warping, in: **proceedings of the Conference on Knowledge Discovery and Data Mining**, pp. 262–270.

[10] Roiger, R. J., Geatz, M. W. (2003). **Data Mining A Tutorial – Based Primer**. Pearson Education, Inc. Addison Wesley. pp. 11-12.

[11] Shahbaba, M., Beheshti, S. (2014). MACE-means clustering. **Singnal Processing**. Vol.105, pp.216-225.

[12] Kwedlo, W. (2011). A clustering method combining differential evolution with the K-means algorithm. **Pattern Recognition Letters**. Vol.32, pp.1613–1621.

[13] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**. Vol.20, pp.53-65.

[14] Tippaya, T., Nuttawut, K., Pongsakorn, D., Kittisan, K., Nittaya, K. (2015). The Clustering Validity with Silhouette and Sum of Squared Errors. **Proceedings of the 3rd International Conference on Industrial Application Engineering 2015**. pp.44-51.

[15] Längkvist, M., Karlsson, L. and Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. **Pattern Recognition Letters**, Vol. 42, pp. 11-24.

[16] Hautamaki, V., Nykanen, P. and Franti, P. (2008). Time-series clustering by approximate prototypes, in: **Proceedings of 19th International Conference on Pattern Recognition**, 2008, ICPR 2008, pp. 1–4.

[17] Keogh, E., Pazzani, M. (1998). An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback, in: **Proceedings of the 4th International Conference of Knowledge Discovery and Data Mining**, pp. 239–241.

[18] Van Wijk, J.J. and Van Selow, E.R. (1999). Cluster and calendar based visualization of time series data, in: **Proceedings of 1999 IEEE Symposium on Information Vision**, pp. 4–9.

[19] Sakoe, H. and Chiba, S. (1971). A dynamic programming approach to continuous speech recognition, **Proceedings of the Seventh International Congress on Acoustics**. Vol. 3, pp. 65-69.

[20] Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition, **IEEE Trans. Acoust. Speech Signal Process**. Vol. 26, no. 1, pp. 43-49.

[21] Keogh, E., Pazzani, M., Chakrabarti, K. and Mehrotra, S. (2000). A simple dimensionality reduction technique for fast similarity search in large time series databases, **Knowl. Inf. Syst**. Vol. 1805. No, 1, pp. 122-133.

[22] Ratanamahatana, C. (2005). Multimedia retrieval using time series representation and relevance feedback, in: **Proceedings of 8th International Conference on Asian Digital Libraries (ICADL 2005)**, pp. 400-405.

[23] Ratanamahatana, C., Keogh, E., Bagnall, A.J. and Lonardi, S. (2005). A novel bit level time series representation with implications for similarity search and clustering, in: **Proceedings of 9th Pacific-Asian International Conference on Knowledge Discovery and Data Mining (PAKDD'05),** pp. 771–777.

[24] Bagnall, A.A.J., Ratanamahatana, C. "Ann", Keogh, E., Lonardi, S. and Janacek, G. (2006). A bit level representation for time series data mining with shape based similarity, **Data Min. Knowl. Discov**. Vol. 13, no. 1, pp. 11–40.

[25] Shieh, J. and Keogh, E. (2009). iSAX: disk-aware mining and indexing of massive time series datasets, **Data Min. Knowl. Discov**. Vol. 19, no. 1, pp. 24–57.

[26] Cai, Y. and Ng, R. (2004). Indexing spatio-temporal trajectories with Chebyshev polynomials, in: **Procedings of 2004 ACM SIGMOD International**, pp. 599.

[27] Faloutsos, C., Ranganathan, M. and Manolopoulos, Y. (1994). Fast subsequence matching in time-series databases, **ACM SIGMOD Rec**. Vol. 23, pp. 419–429.

[28] Bingham, E. (2001). Random projection in dimensionality reduction: applications to image and text data, in: **Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, pp. 245–250.

[29] Chen, Q., Chen, L., Lian, X. and Liu,Y. (2007). Indexable PLA for efficient similarity search, in: **Proceedings of the 33rd International Conference on Very large Data Bases**, pp. 435–446.

[30] Lin, J., Keogh, E., Wei, L., Lonardi, S. (2007). Experiencing SAX: a novel symbolic representation of time series, **Data Min. Knowl. Discov**. Vol. 15, Issue 2, pp. 107–144.

[31] Montero, P. and Vilar, A.J. (2014). TSclust: An R Package for Time Series Clustering. **Journal of Statistical Software**. November 2014, Vol. 62, Issue 1, pp. 1-43.

[32] Theodoridis, S., Pikrakis, A., Koutroumbas, K., Cavouras, D. (2010). An Introduction to Pattern Reccognition : A MATLAB Approach. **Academic Press, USA**.

[33] Keogh, E., Lin, J. (2005). Clustering of Time Series Subsequences is meaningless: implications for previous and future research, **Knowledge and information systems**. Vol.8, No. 2, pp. 154–177.

[34] Aggarwal, C. C. and Reddy, C. K. (2014). Data Clustering Algorithms and Applications, CRC Press, **Taylor & Fancis Group. A Chapman & Hall Book**.

[35] Zhao, Y. (2013). R and Data Mining Examples and Case Studies. **Academic Press**, USA, UK, The Netherlands.

[36] Maimon, O. and Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook: **Second Edition. Springer, New York Dordrecht Heidelberg London**, pp. 270.

[37] Keogh E, Kasetty S (2002) On the need for time series data mining benchmarks: a survey and empirical demonstration. In: **Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining**. Edmonton, Alberta, Canada, 23–26 July, pp 102–111.

[38] **Hierarchical clustering**. (2016). Wikipedia the free encyclopedia, April 25, 2016. [Online]. Avialable:https://en.wikipedia.org/wiki/Hierarchical_clustering.

[39] Brandmaier A (2012). Permutation Distribution Clustering and Structural Equation Model Trees. **Dissertation, Saarland University, Saarbrücken**.

[40] Brandmaier, A. (2015). pdc: **Permutation Distribution Clustering. R package**

version 1.0.3, URL http://CRAN.R-project.org/package =pdc.

[41] Aggarwal, C. C., Reddy, C. K. (2013). Data Clustering: Algorithms and Applications, Vol.31, CRC Press, Hoboken, New Jersey, p.648. (Chapman & Hall/CRC **Data Mining and Knowledge Discovery Series**, ISBN: 1466558210).

[42] Jain, A. K. (2010). Data clustering: 50 years beyond k-means, **Pattern Recogn. Lett**. Vol.31(8), pp. 651–666, http://dx.doi.org/10.1016/j.patrec.2009.09.011.

[43] Mohammed J. Zaki and Wagner Meira Jr. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. **Cambridge University Press**, USA.

[44] Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen and Gustavo Batista (2015). **The UCR Time Series Classification Archive**. URL http://www.cs.ucr.edu/~eamonn/time_series_data/

[45] Olszewski, R. T. (2001). Generalized feature extraction for structural pattern recognition in time-series data (No. CMU-CS-01-108). **CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE**.

[46] Royal Irrigation Department. (2016). “Runoff Data Files”, HYDROLOGY IRRIGATION Lower North Eastern Region Hydrological Irrigation Center (NAKHON RATCHASIMA, THAILAND), Ministry of Agriculture and Cooperatives. [Online]. **Available:** http://hydro-4.com/.