

การเปรียบเทียบมาตรวัดระยะห่าง ในการวิเคราะห์การจัดกลุ่มสำหรับข้อมูลอนุกรมเวลา

Comparison of Distance Measures in Cluster Analysis for Time-Series Data

นัท กุลวานิช* และอักรินทร์ ไพบูลย์พานิช

ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย

แขวงวังใหม่ เขตปทุมวัน กรุงเทพมหานคร 10330

Nat Kulvanich* and Akarin Phaibulpanich

Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University,

Wangmai, Pathumwan, Bangkok 10330

บทคัดย่อ

การคำนวณหาระยะห่างระหว่างข้อมูลถือเป็นขั้นตอนสำคัญของการวิเคราะห์การจัดกลุ่ม (cluster analysis) ทั้งนี้การค้นหามาตรวัดที่เหมาะสมถือเป็นกระบวนการสำคัญเมื่อวิเคราะห์การจัดกลุ่มสำหรับข้อมูลอนุกรมเวลา เนื่องจากการเลือกใช้วิธีคำนวณระยะห่างระหว่างข้อมูลที่แตกต่างกันสามารถส่งผลให้ผลลัพธ์ของการจัดกลุ่มแตกต่างกัน งานวิจัยนี้เป็นการเปรียบเทียบประสิทธิภาพของวิธีการจัดกลุ่มข้อมูลด้วยวิธีวัดระยะห่างสำหรับข้อมูลอนุกรมเวลา 8 วิธี ประกอบด้วย (1) Euclidean (2) Minkowski (3) dynamic time warping (4) Chouakria-Douzal (5) Piccolo (6) Maharaj (7) discrete wavelet transform และ (8) cepstral-based distance โดยเปรียบเทียบทั้งในส่วนของการจำลองและข้อมูลจริง ในส่วนของข้อมูลจำลองที่ใช้ในการศึกษานั้นสร้างมาจากตัวแบบ ARIMA จำนวนทั้งสิ้น 14 ตัวแบบ และในส่วนข้อมูลจริงนั้นได้นำราคาปิดรายวันของ 68 หลักทรัพย์ ที่ผ่านการคัดเลือก 100 อันดับแรก ในตลาดหลักทรัพย์แห่งประเทศไทย (SET100) ในระหว่างเดือนมกราคมถึงเดือนเมษายน ปี พ.ศ. 2561 มาใช้ในการเปรียบเทียบ ผลการศึกษาพบว่ามาตรวัดระยะแบบ Dynamic Time Warping ให้ประสิทธิภาพในการจัดกลุ่มดีที่สุดทั้งในชุดข้อมูลจริงและชุดข้อมูลจำลอง

คำสำคัญ : มาตรวัดระยะห่าง; การวิเคราะห์การจัดกลุ่ม; ข้อมูลอนุกรมเวลา

Abstract

One important step in cluster analysis is to determine a distance or dissimilarity measure between data objects. Searching for suitable measure is an important process for time series

*ผู้รับผิดชอบบทความ : nat@cbs.chula.ac.th

clustering since a various measure of dissimilarity designed for time-series data could lead to different cluster solutions. This research is a comparative study of the effectiveness of 8 distance measures for time-series dataset: (1) Euclidean, (2) Minkowski, (3) dynamic time warping, (4) Chouakria-Douzal, (5) Piccolo, (6) Maharaj, (7) discrete wavelet transform, and (8) cepstral-based distance. This study extends prior research by using both real and simulated data for the comparison of clustering results. The simulated data is generated from 14 ARIMA processes. The real data is the daily stock close prices of 68 listed companies that are included in the list of stocks in stock exchange of Thailand 100 (SET100) index during January-April 2018. The results suggest that dynamic time warping is the most effective measure in both real and simulated data.

Keywords: distance measure; cluster analysis; time-series data

1. บทนำ

การวิเคราะห์การจัดกลุ่ม (cluster analysis) เป็นเทคนิคการจัดข้อมูลที่มีลักษณะเหมือนกันหรือคล้ายกันไว้ในกลุ่มเดียวกัน และข้อมูลที่มีลักษณะต่างกันก็จะจัดไว้ต่างกลุ่มกัน การวิเคราะห์การจัดกลุ่มข้อมูลนี้เป็นเทคนิคที่นิยมใช้กันอย่างแพร่หลายในด้านต่าง ๆ เช่น สังคมศาสตร์ การแพทย์ การตลาด การศึกษา เนื่องจากผู้วิเคราะห์ไม่จำเป็นต้องทราบจำนวนกลุ่มของข้อมูลมาก่อน หรือระบุจำนวนกลุ่มของข้อมูลไม่ได้ [1] สุชาติ [2] กล่าวถึงการใช้การวิเคราะห์การจัดกลุ่มว่าผู้วิเคราะห์จำเป็นต้องคำนวณหรือวัดระยะห่างระหว่างข้อมูล (distance) ก่อนใช้การวิเคราะห์การจัดกลุ่ม ซึ่งการเลือกใช้วิธีคำนวณระยะห่างระหว่างข้อมูลที่แตกต่างกันสามารถส่งผลให้ผลลัพธ์ของการจัดกลุ่มแตกต่างกันได้ [3,4]

มาตรวัดความห่างระหว่างข้อมูลสามารถทำการคำนวณได้หลายแบบ ได้แก่ ระยะห่างแบบ Euclidean (Euclidean distance) ระยะห่างแบบ Manhattan (Manhattan distance) ระยะห่างแบบ Chebychev (Chebychev distance) เป็นต้น อย่างไรก็ตาม วิธีการคำนวณระยะห่างที่กล่าวมานั้นมีข้อจำกัดในการนำไปใช้กับเวกเตอร์ของชุดข้อมูลอนุกรมเวลา

(time-series data) ซึ่งเป็นข้อมูลที่มีการเคลื่อนไหวของค่าข้อมูลอย่างต่อเนื่องและสามารถอธิบายการเคลื่อนไหวดังกล่าวด้วยค่าสหสัมพันธ์ในตัวเอง (autocorrelation) ดังนั้นหากนำวิธีการวัดระยะห่างสำหรับชุดข้อมูลภาคตัดขวาง (cross-sectional data) ไปใช้กับชุดข้อมูลอนุกรมเวลานั้น ก็อาจส่งผลให้การวิเคราะห์การจัดกลุ่มเกิดความผิดพลาดได้ เนื่องจากวิธีการวัดระยะห่างเหล่านั้นไม่ได้พิจารณาถึงโครงสร้างสหสัมพันธ์ (correlation structure) ในชุดข้อมูลประเภตอนุกรมเวลา [5,6]

ความท้าทายประการหนึ่งสำหรับการใช้การวิเคราะห์การจัดกลุ่มในชุดข้อมูลอนุกรมเวลา คือ มาตรวัดระยะห่างสำหรับข้อมูลชนิดนี้มีจำนวนมาก [7,8] Kianimajd และคณะ [9] แบ่งมาตรวัดระยะห่างสำหรับอนุกรมเวลาออกเป็น 2 กลุ่ม โดยกลุ่มแรก คือ มาตรวัดระยะห่างอนุกรมเวลาแบบขอบเขตเวลา (time domain distance measure) ซึ่งเป็นมาตรวัดที่เน้นศึกษารูปแบบการเคลื่อนไหวที่เหมาะสมที่สุดของชุดอนุกรมเวลาเมื่อเวลาเปลี่ยนไป ประกอบด้วยมาตรวัดระยะแบบ Minkowski มาตรวัดระยะแบบ dynamic time warping (DTW) [10] มาตรวัดระยะแบบ Chouakria-Douzal [11] มาตรวัดระยะแบบ

Piccolo [12] มาตรฐานวัดระยะแบบ Maharaj [13,14] เป็นต้น ส่วนกลุ่มที่สอง คือ มาตรฐานวัดระยะห่างฐานเปลี่ยนรูป (transform-based distance measure) ซึ่งเป็นมาตรฐานที่ใช้การแปลงฟูเรียร์ (Fourier transform) กับชุดข้อมูลอนุกรมเวลา เพื่อแปลงข้อมูลให้อยู่ในรูปแบบขอบเขตความถี่ (frequency domain) และจัดให้เป็นข้อมูลที่อยู่ในรูปแบบ sine หรือ cosine ในทางตรีโกณมิติ ประกอบด้วยมาตรฐานวัดระยะแบบการแปลงเวฟเลตแบบเต็มหน่วย (discrete wavelet transform, DWT) [15,16] มาตรฐานวัดระยะแบบ cepstral-based [17] เป็นต้น

Kianimajd และคณะ [14] เปรียบเทียบความถูกต้องของการใช้การวิเคราะห์การจัดกลุ่มจำนวน 20 กลุ่ม เมื่อใช้มาตรฐานวัดระยะห่างสำหรับอนุกรมเวลาที่ต่างกัน โดยใช้ข้อมูลระบบหัวใจและหลอดเลือดของคนไข้แต่ละคน ได้ข้อสรุปว่ามาตรฐานวัดระยะแบบ Chouakria-Douzal และมาตรฐานวัดระยะแบบ discrete wavelet transform ให้ผลความถูกต้องที่ดีกว่าการวัดระยะแบบอื่น แต่งานวิจัยนี้ก็ยังไม่มีการจำลองข้อมูลอนุกรมเวลาเพื่อศึกษาและเปรียบเทียบประสิทธิภาพของการใช้การวิเคราะห์การจัดกลุ่ม

Pertega และ Vilar [18] เปรียบเทียบประสิทธิภาพของการใช้การวิเคราะห์การจัดกลุ่มด้วยมาตรฐานวัดระยะห่าง 3 แบบ ประกอบด้วยมาตรฐานวัดระยะแบบ dynamic time warping มาตรฐานวัดระยะแบบ Piccolo และมาตรฐานวัดระยะห่างแบบ Euclidean ในชุดข้อมูลอนุกรมเวลาที่จำลองด้วยตัวแบบ autoregressive moving average (ARMA) ได้ข้อสรุปว่าการวัดระยะแบบ dynamic time warping ให้ผลความถูกต้องที่ดีกว่าการวัดระยะ 2 แบบที่เหลือ ส่วน Kalpakis และคณะ [17] เปรียบเทียบความถูกต้องของการใช้การวิเคราะห์การจัดกลุ่มด้วยการวัดระยะห่างในกลุ่มมาตรฐานเปลี่ยนรูป โดยใช้ชุดข้อมูลอนุกรม

เวลาที่จำลองด้วยตัวแบบ ARIMA ได้ข้อสรุปว่ามาตรฐานวัดระยะแบบ cepstral-based ให้ผลความถูกต้องที่ดีที่สุดในมาตรฐานวัดระยะห่างกลุ่มมาตรฐานเปลี่ยนรูป

การศึกษาครั้งนี้ ผู้วิจัยสนใจที่จะขยายขอบเขตงานวิจัยของ Pertega และ Vilar, Kianimajd และคณะ และ Kalpakis และคณะ เพื่อเปรียบเทียบความถูกต้องของการใช้การวิเคราะห์การจัดกลุ่มสำหรับชุดข้อมูลอนุกรมเวลา เมื่อใช้มาตรฐานวัดระยะ 8 แบบ ที่อยู่ในกลุ่มมาตรฐานวัดระยะห่างอนุกรมเวลาแบบขอบเขตเวลาและกลุ่มมาตรฐานวัดระยะห่างฐานเปลี่ยนรูป โดยใช้ชุดข้อมูลอนุกรมเวลาที่จำลองด้วยตัวแบบ ARIMA และใช้ชุดข้อมูลจริงของราคาปิดรายวันของ 68 หลักทรัพย์ ในดัชนี SET 100 ของตลาดหลักทรัพย์แห่งประเทศไทย ตั้งแต่วันที่ 3 มกราคม พ.ศ. 2561 ถึงวันที่ 30 เมษายน พ.ศ. 2561 ทั้งนี้ผู้วิเคราะห์ข้อมูลสามารถนำผลวิจัยที่ได้ไปใช้ในการวางแผนหรือใช้เป็นแนวทางในการเลือกมาตรฐานวัดระยะห่างสำหรับการวิเคราะห์การจัดกลุ่มของชุดข้อมูลอนุกรมเวลาในอนาคตได้

2. วิธีการวิจัย

2.1 มาตรฐานวัดระยะห่างและทฤษฎีที่เกี่ยวข้อง

2.1.1 มาตรฐานวัดระยะห่างในข้อมูลอนุกรมเวลา

ข้อมูลประเภทอนุกรมเวลามีความสำคัญมากในงานวิจัยหรืองานวิเคราะห์ที่มีลักษณะการเก็บข้อมูลแบบตามลำดับ (temporal ordering) ซึ่งมักจะเกิดขึ้นในด้านการแพทย์ที่เก็บข้อมูลจากคนไข้ตามระยะเวลาที่ทดลอง หรือในด้านการเงินที่เก็บข้อมูลงบการเงินของแต่ละบริษัทเป็นประจำ แม้กระทั่งในด้านวิทยาศาสตร์และสิ่งแวดล้อมที่มักจะมีข้อมูลเกิดขึ้นอย่างต่อเนื่อง การนำเทคนิคเหมืองข้อมูล (data

mining) โดยเฉพาะการวิเคราะห์การจัดกลุ่มไปใช้ในการวิเคราะห์ข้อมูลเหล่านี้จึงเกิดขึ้นอยู่เสมอ [3] สามารถแสดงตัวอย่างงานวิจัยที่เกี่ยวข้องดังตารางที่ 1 ตารางที่ 1 จะเห็นว่ามีวิธีการเลือกวิธีวัด

ระยะห่างระหว่างข้อมูลที่ต่างกันในแต่ละงานวิจัย ซึ่งใช้การวิเคราะห์การจัดกลุ่มกับข้อมูลอนุกรมเวลา เนื่องจากการเลือกวิธีวัดระยะห่างที่ต่างกันจะส่งผลต่อผลลัพธ์ของการวิเคราะห์การจัดกลุ่มที่ต่างกัน

ตารางที่ 1 การประยุกต์ใช้การวิเคราะห์การจัดกลุ่มในข้อมูลอนุกรมเวลาและวิธีการวัดระยะห่างที่เกี่ยวข้อง

งานวิจัย	ข้อมูล	มาตรวัดระยะห่างระหว่างข้อมูล
Khan และคณะ [19]	การเงิน	Euclidean
Kremer และคณะ [20]	สภาพอากาศ	dynamic time warping
Niennattrakul และ Ratanamahatana [21]	ระบบสื่อสาร	dynamic time warping
Jixue [22]	การเงิน	discrete wavelet transform
Verdoolaege และ Rosseel [23]	FMRI	discrete wavelet transform
Baragona [24]	สาธารณสุข	Chouakria-Douzal
Lhermitte [25]	สภาพแวดล้อม	Minkowski

2.1.2 การคำนวณค่าระยะห่างในข้อมูลอนุกรมเวลา

กำหนดให้ $X(t) = (X_1, X_2, \dots, X_N)$ และ $Y(t) = (Y_1, Y_2, \dots, Y_N)$ คือ เวกเตอร์ของชุดข้อมูลอนุกรมเวลา 2 ชุด ที่เก็บมาตามลำดับเป็นจำนวน N ระยะเวลา สามารถคำนวณระยะห่างข้อมูลอนุกรมเวลา $D(X(t), Y(t))$ ด้วยวิธีต่าง ๆ ได้ดังนี้

(1) ระยะห่างแบบ Euclidean (Euclidean distance) วิธีวัดระยะห่างแบบยุคลิดสำหรับอนุกรมเวลานั้น เป็นการคำนวณระยะห่างที่ง่ายที่สุดในวิธีการคำนวณทั้งหมด โดย Lhermitte และคณะ [26] เสนอสูตรในการคำนวณดังนี้

$$D(X(t), Y(t)) = \sum_{i=1}^N \sqrt{|X_i - Y_i|^2} \quad (1)$$

ซึ่งข้อเสียหลักของวิธีวัดระยะห่างแบบ Euclidean คือ ไม่สามารถวัดระยะห่างของข้อมูลอนุกรมเวลาที่มีจำนวนระยะเวลาในการเก็บไม่เท่ากัน

(2) ระยะห่างแบบ Minkowski (Minkowski distance) วิธีวัดระยะห่างแบบ Minkowski สำหรับอนุกรมเวลาสามารถคำนวณดังนี้

$$D(X(t), Y(t)) = \left(\sum_{i=1}^N |X_i - Y_i|^p \right)^{1/p} \quad (2)$$

โดยที่ p เป็นจำนวนเต็มบวก แต่ถ้า $p = 2$ วิธีวัดระยะห่างนี้จะเป็วิธีวัดระยะห่างแบบ Euclidean

(3) ระยะห่างแบบ dynamic time warping (dynamic time warping distance) โดย Sanko และ Kruskal [9] เสนอวิธีวัดระยะห่างแบบ dynamic time warping ซึ่งเป็นวิธีกำหนดพลวัต (dynamic programming) ที่มีความแม่นยำสูงสามารถใช้วัดความต่างระหว่างข้อมูลอนุกรมเวลา 2 ชุด ที่มีจำนวนระยะเวลาในการเก็บไม่เท่ากัน โดยกำหนดให้ข้อมูลอนุกรมเวลา 2 ชุด เป็นดังนี้

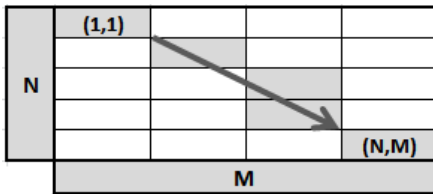
$X(t) = X_1, X_2, \dots, X_e, \dots, X_N$ ขนาด N ระยะเวลา
 $Y(t) = Y_1, Y_2, \dots, Y_f, \dots, Y_M$ ขนาด M ระยะเวลา

ซึ่งมีขั้นตอนในการคำนวณระยะทางดังนี้

(3.1) สร้างเมทริกซ์ขนาด $N \times M$ มิติเมื่อส่วนประกอบที่ (e^{th}, f^{th}) ของเมทริกซ์ คือ ระยะห่างกำลังสอง (squared distance) ของ X_e และ Y_f

$$d(X_e, Y_f) = (X_e - Y_f)^2 \quad (3)$$

(3.2) หาเส้นทางการวอร์ป (warping path) ทั้งหมดที่เป็นไปได้ โดยเส้นทางการวอร์ป คือ เส้นทางการเคลื่อนที่จาก $(1,1)$ ถึง (N, M) มีได้หลายรูปแบบ ดังแสดงในรูปที่ 1



รูปที่ 1 เส้นทางการวอร์ป (warping path)

(3.3) ดังนั้นเส้นทางที่ดีที่สุด คือ เส้นทางที่มีต้นทุนการวอร์ป (warping cost) ต่ำที่สุดเมื่อเปรียบเทียบกับเส้นทางการวอร์ปทั้งหมดที่เป็นไปได้ การวัดระยะทาง dynamic time warping ที่มีต้นทุนต่ำที่สุดระหว่าง $X(t)$ และ $Y(t)$ สามารถคำนวณดังนี้

$$D(X(t), Y(t)) = \min \sqrt{\sum_{g=1}^G w_g} \quad (4)$$

โดย w_g คือ องค์กรประกอบของเมทริกซ์เส้นทางการวอร์ป $(i, j)_g$ ในรูปที่ 1 และเป็นองค์กรประกอบที่ g ของเส้นทางการวอร์ป (warping path) จนถึง (N, M) และ G คือ จุดสุดท้ายของเส้นทาง

(4) ระยะห่างแบบ Chouakria-Douzal (Chouakria-Douzal distance) Chouakria-Douzal และ Nagabhushan [10] เสนอวิธีคำนวณระยะห่างโดยอาศัยแนวคิดของค่าสหสัมพันธ์ไขว้ (cross-correlation, CORRT) ของข้อมูลอนุกรมเวลา 2 ชุด

โดยเสนอสูตรการคำนวณค่าสหสัมพันธ์ไขว้ สำหรับข้อมูลอนุกรมเวลา 2 ชุด ที่มีจำนวนระยะเวลาในการเก็บเท่ากันดังนี้

$$CORRT(X(t), Y(t)) = \frac{\sum_{i=1}^{N-1} (X_{i+1} - X_i)(Y_{i+1} - Y_i)}{\sqrt{\sum_{i=1}^{N-1} (X_{i+1} - X_i)^2} \sqrt{\sum_{i=1}^{N-1} (Y_{i+1} - Y_i)^2}} \quad (5)$$

และนำค่าสหสัมพันธ์ไขว้ใน (5) ไปคำนวณค่าระยะห่างระหว่างอนุกรมเวลาได้ดังนี้

$$D(X(t), Y(t)) = h(CORRT(X(t), Y(t))) \delta(X(t), Y(t)) \quad (6)$$

โดยที่ $h(\cdot)$ เป็นฟังก์ชันปรับค่าดัดแปลง (adaptive tuning function) และ $\delta(X(t), Y(t))$ เป็นค่าระยะห่างแบบอื่นตามแต่ผู้วิเคราะห์จะเลือกใช้ โดยสามารถใช้ค่าระยะห่างแบบ Euclidean หรือ Minkowski ก็ได้

(5) ระยะห่างแบบ Piccolo (Piccolo distance) โดย Piccolo [11] เสนอวิธีคำนวณระยะห่างโดยอาศัยสมบัติผกผัน (invertible) ของแบบจำลอง ARMA ที่สามารถเขียนสมการของตัวแบบในรูป $AR(\infty)$ โดยกำหนดให้ $\Pi_X = (\pi_{1,X}, \pi_{2,X}, \dots, \pi_{k_1,X})^T$ เป็นค่าสัมประสิทธิ์หน้า X_t ของอนุกรมเวลา $X(t)$ ในรูป $AR(k_1)$ และ $\Pi_Y = (\pi_{1,Y}, \pi_{2,Y}, \dots, \pi_{k_2,Y})^T$ เป็นค่าสัมประสิทธิ์หน้า Y_t ของอนุกรมเวลา $Y(t)$ ในรูป $AR(k_2)$ ซึ่งสามารถคำนวณระยะห่างแบบ Piccolo ดังนี้

$$D(X(t), Y(t)) = \sqrt{\sum_{i=1}^k (\pi_{i,X} - \pi_{i,Y})^2} \quad (7)$$

เมื่อ $k = \max(k_1, k_2)$ โดยที่ $\pi_{i,X} = 0$ เมื่อ $i > k_1$ และ $\pi_{i,Y} = 0$ เมื่อ $i > k_2$

(6) ระยะห่างแบบ Maharaj (Maharaj distance) โดย Maharaj [12] ปรับเปลี่ยนการวัดระยะห่างในสมการ (5) โดยใช้ข้อมูลจากค่าความคลาดเคลื่อนของตัวแบบ ARMA มาปรับค่าคำนวณ ได้สูตรวัดระยะห่างดังต่อไปนี้

$$D(X(t), Y(t)) = \sqrt{N} (\Pi_X - \Pi_Y)' V^{-1} (\Pi_X - \Pi_Y) \quad (8)$$

เมื่อ $V = \sigma_X^2 R_X^{-1}(k) + \sigma_Y^2 R_Y^{-1}(k)$ โดยที่ σ_X^2 และ σ_Y^2 เป็นความแปรปรวนของค่าความคลาดเคลื่อนของอนุกรมเวลา $X(t)$ และ $Y(t)$ ตามลำดับ เช่นเดียวกับ $R_X(k)$ และ $R_Y(k)$ ซึ่งเป็นเมทริกซ์ความแปรปรวนร่วม (covariance matrix) ของค่าความคลาดเคลื่อน

(7) ระยะห่างแบบ discrete wavelet transform (discrete wavelet transform distance) โดย Shasha [15] เสนอให้แปลงข้อมูลอนุกรมเวลาที่ต้องการหาระยะห่างให้อยู่ในรูปแบบขอบเขตความถี่ (frequency domain) เพื่อใช้ในการหาค่าของการเกิดด้วย periodogram โดยกำหนดให้ $I_X(\lambda_k) = N^{-1} |\sum_{t=1}^N X_t \exp(-i\lambda_k t)|^2$ เป็น periodogram ของ $X(t)$ และ $I_Y(\lambda_k) = N^{-1} |\sum_{t=1}^N Y_t \exp(-i\lambda_k t)|^2$ เป็น periodogram ของ $Y(t)$ ทั้งนี้ $\lambda_k = 2\pi k/N$ และสามารถคำนวณระยะห่างดังนี้

$$D(X(t), Y(t)) = \left(\frac{2}{N-1} \right) \sqrt{\sum_{k=1}^{(N-1)/2} (I_X(\lambda_k) - I_Y(\lambda_k))^2} \quad (9)$$

(8) ระยะห่างแบบ cepstral-based (cepstral-based distance) โดย Kalpakis [17] ใช้เทคนิคการแปลงกลับฟูเรียร์ (inverse Fourier transform) มาใช้คำนวณค่าสัมประสิทธิ์ cepstral เพื่อใช้ในการหาระยะทางสำหรับอนุกรมเวลาที่สามารถเขียนให้อยู่ในรูป $AR(p)$ โดยสามารถคำนวณค่าระยะห่างระหว่างอนุกรมเวลาดังนี้

$$D(X(t), Y(t)) = \sum_{i=1}^N \sqrt{|\psi_{i,X} - \psi_{i,Y}|^2} \quad (10)$$

ซึ่งค่าสัมประสิทธิ์ cepstral สามารถคำนวณดังนี้

$$\psi_i = \begin{cases} \phi_i; i = 1 \\ \phi_i + \sum_{m=1}^{i-1} (\phi_m - \psi_{i-m}); 1 < i \leq p \\ \sum_{m=1}^p \left(1 - \frac{m}{i}\right) \phi_m \psi_{i-m}; p < i \end{cases}$$

เมื่อ ϕ_i คือ ค่าสัมประสิทธิ์ในตัวแบบอนุกรม $AR(p)$

และ $\psi_{i,X}$ คือ ค่าสัมประสิทธิ์ cepstral ของอนุกรมเวลา $X(t)$ ส่วน $\psi_{i,Y}$ คือ ค่าสัมประสิทธิ์ cepstral ของอนุกรมเวลา $Y(t)$

2.1.3 การวิเคราะห์การจัดกลุ่มข้อมูลด้วย

algorithm K-medoids

algorithm K-medoids พัฒนาขึ้นจาก algorithm K-means เพื่อแก้ปัญหาค่าความผิดพลาดที่จะเกิดขึ้น เมื่อมีบางข้อมูลที่มีค่าแตกต่างจากข้อมูลอื่นมากหรือเรียกว่า outlier ทั้งนี้ algorithm K-medoids สามารถจัดกลุ่มหน่วยข้อมูลหรือวัตถุที่ชุดข้อมูลมีทั้งตัวแปรเชิงคุณภาพและตัวแปรเชิงปริมาณผสมกันอยู่ในชุดข้อมูลเดียวกัน โดยหลักการของ K-medoids คือ การจัดกลุ่มข้อมูลโดยพิจารณาความคล้ายคลึงกันของข้อมูลกับ median vector (medoids) ซึ่งเป็นตัวแทนในแต่ละกลุ่ม

2.1.4 ตัวแบบ autoregressive integrated

moving average (ARIMA)

ตัวแบบอนุกรมเวลาด้วยวิธีของ Box-Jenkins เป็นวิธีวิเคราะห์อนุกรมเวลาทางสถิติที่เกิดขึ้นใน ปี ค.ศ. 1970 ซึ่งได้รับความนิยมจนถึงปัจจุบัน [27] โดยเลือกรูปแบบในการพยากรณ์อย่างเป็นระบบ ใช้ค่าสัมประสิทธิ์สหสัมพันธ์ในตัวเองและค่าสัมประสิทธิ์สหสัมพันธ์ในตัวเองบางส่วน ทดสอบทางสถิติและการหาช่วงความเชื่อมั่นของค่าสังเกตในอนาคต ซึ่งก่อให้เกิดตัวแบบ ARIMA โดยข้อมูลอนุกรมเวลาที่นำมาวิเคราะห์ต้องเป็นแบบคงที่ (stationary) และไม่มีแนวโน้ม

ตัวแบบ ARIMA(p,d,q) คือ ตัวแบบ ARMA(p,q) ที่หาผลต่าง (differencing) อันดับ d เพื่อแปลงข้อมูลอนุกรมเวลาแบบไม่คงที่ (non-stationary time-series) ให้เป็นข้อมูลอนุกรมเวลาแบบคงที่ ซึ่งสามารถเขียนในรูปสมการดังนี้

$$\phi(B)(1-B)^d x_t = \theta(B)w_t \quad (11)$$

หรือสามารถเขียนสมการที่ (11) ได้เป็น

$$(1-\phi_1B-\phi_2B^2-\dots-\phi_pB^p)(1-B)^d x_t = (1-\theta_1B-\theta_2B^2-\dots-\theta_qB^q)w_t \quad (12)$$

เมื่อ $\phi_1, \phi_2, \dots, \phi_p$ คือ ค่าสัมประสิทธิ์แสดงอันดับของกระบวนการ autoregressive โดยที่ $\phi_p \neq 0$ และ $\theta_1, \theta_2, \dots, \theta_q$ คือ ค่าสัมประสิทธิ์แสดงอันดับของกระบวนการ moving average โดยที่ $\theta_q \neq 0$ และ w_t คือ ค่าความเคลื่อนที่มีการแจกแจงแบบเกาส์เซียน (Gaussian) โดยมีค่าเฉลี่ยเป็น 0 และความแปรปรวน σ^2

2.1.5 การวัดประสิทธิภาพของการวิเคราะห์การจัดกลุ่ม

เมื่อผู้วิเคราะห์ได้นำการวิเคราะห์การจัดกลุ่มไปใช้กับชุดข้อมูลอนุกรมเวลา และหากทราบกลุ่มที่แท้จริง (true cluster solution) ของแต่ละหน่วยในชุดข้อมูล ก็จะสามารถวัดประสิทธิภาพของกลุ่มที่จัดด้วยมาตรวัดต่าง ๆ 2 แบบ ดังนี้

(1) ค่าอัตราการจำแนกผิด (misclassification rate) เป็นค่าที่บอกถึงความผิดพลาดในการจัดกลุ่ม โดยในงานวิจัยนี้ได้จำลองชุดข้อมูลอนุกรมเวลาขึ้นมา ซึ่งถือว่ากลุ่มที่จำลองขึ้นมานั้นเป็นกลุ่มที่แท้จริง และเมื่อใช้การวิเคราะห์การจัดกลุ่มกับชุดข้อมูลที่จำลองเสร็จสิ้น กลุ่มที่จัดมาโดยการวิเคราะห์การจัดกลุ่มนั้นก็จะเป็นมาตรวจสอบกับกลุ่มที่แท้จริงที่ได้จำลองไว้ตั้งแต่เริ่มต้น ดังสูตรต่อไปนี้

$$\text{misclassification rate} = \frac{\text{จำนวนข้อมูลที่ทำนายผิด}}{\text{จำนวนข้อมูลทั้งหมด}} \quad (13)$$

ทั้งนี้ หากค่าอัตราการจำแนกผิด (misclassification rate) มีค่าน้อย จะแสดงถึงประสิทธิภาพในการจัดกลุ่มที่ดี

(2) ค่าการประเมินการจัดกลุ่มของ Kalpakis โดย Kalpakis [17] เสนอการคำนวณค่าการประเมินการจัดกลุ่มของ Kalpakis หรือ $Sim(G, C)$

สำหรับใช้วัดประสิทธิภาพในการจัดกลุ่มเมื่อทราบกลุ่มที่แท้จริงดังนี้

$$Sim(G, C) = 1 / k \sum_{i=1}^k \max_{1 \leq j \leq k} Sim(G_i, C_j) \quad (14)$$

โดยที่ G เป็นเวกเตอร์แสดงกลุ่มที่แท้จริงของข้อมูลแต่ละหน่วยในชุดข้อมูล และ C เป็นเวกเตอร์แสดงกลุ่มที่ถูกจัดโดยการวิเคราะห์การจัดกลุ่มของแต่ละหน่วยในชุดข้อมูล

$$\text{ซึ่ง } Sim(G_i, C_j) = \frac{2|G_i \cap C_j|}{|G_i| + |C_j|} \text{ และ}$$

$| \cdot |$ คือ การนับจำนวนสมาชิกในแต่ละเซต และ k คือ จำนวนกลุ่ม ทั้งนี้หากค่า $Sim(G, C)$ มีค่ามาก จะแสดงถึงประสิทธิภาพในการจัดกลุ่มที่ดี

2.2 การเปรียบเทียบในชุดข้อมูลจำลอง

เปรียบเทียบในชุดข้อมูลจำลอง โดยมีขั้นตอนดังต่อไปนี้

ขั้นตอนที่ 1 : จำลองชุดข้อมูลอนุกรมเวลา โดยใช้โปรแกรม R ทั้งหมด 14 กลุ่ม โดยในแต่ละกลุ่มจำลองด้วยตัวแบบ ARIMA ประกอบด้วยตัวแบบดังนี้ ARIMA(1,0,0) , ARIMA(0,0,1) , ARIMA(1,0,1) , ARIMA(2,0,0) , ARIMA(0,0,2) , ARIMA(2,0,1) , ARIMA(1,0,2) , ARIMA(2,1,1) , ARIMA(1,1,2) , ARIMA(1,1,0) , ARIMA(0,1,1) , ARIMA(1,1,1) , ARIMA(2,1,0) และ ARIMA(0,1,2) โดยกำหนดค่าพารามิเตอร์ในแต่ละตัวแบบดังแสดงในตารางที่ 2 ซึ่งขนาดของข้อมูลหรือระยะเวลาในแต่ละกลุ่ม (N) เท่ากับ 50, 100 และ 300 ตามลำดับ

ขั้นตอนที่ 2 : จัดกลุ่มข้อมูลด้วยการวิเคราะห์การจัดกลุ่มข้อมูลด้วย algorithm K-medoids โดยเลือกจำนวนกลุ่มของการจัดกลุ่มเป็น 14 กลุ่ม ซึ่งเท่ากับจำนวนกลุ่มที่แท้จริงจากขั้นตอนที่ 1 และเลือกใช้มาตรวัดระยะห่างสำหรับข้อมูลอนุกรมเวลา 8 แบบ ประกอบด้วย Euclidean, Minkowski,

dynamic time warping, Chouakria-Douzal, Piccolo, Maharaj, discrete wavelet transform และ cepstral-based

ตารางที่ 2 ตัวแบบ ARIMA และค่าพารามิเตอร์ที่กำหนด

ตัวแบบ	ค่าพารามิเตอร์
ARIMA(1,0,0)	$\phi_1 = 0.5$
ARIMA(0,0,1)	$\theta_1 = -0.5$
ARIMA(1,0,1)	$\phi_1 = 0.5, \theta_1 = -0.25$
ARIMA(2,0,0)	$\phi_1 = 0.5, \phi_2 = -0.5$
ARIMA(0,0,2)	$\theta_1 = 0.5, \theta_2 = -0.25$
ARIMA(2,0,1)	$\phi_1 = 0.5, \phi_2 = -0.5, \theta_1 = -0.25$
ARIMA(1,0,2)	$\phi_1 = 0.5, \theta_1 = 0.5, \theta_2 = -0.25$
ARIMA(2,1,1)	$\phi_1 = 0.5, \phi_2 = -0.5, \theta_1 = -0.25$
ARIMA(1,1,2)	$\phi_1 = 0.5, \theta_1 = 0.5, \theta_2 = -0.25$
ARIMA(1,1,0)	$\phi_1 = 0.5$
ARIMA(0,1,1)	$\theta_1 = -0.5$
ARIMA(1,1,1)	$\phi_1 = 0.5, \theta_1 = -0.25$
ARIMA(2,1,0)	$\phi_1 = 0.5, \phi_2 = -0.5$
ARIMA(0,1,2)	$\theta_1 = 0.5, \theta_2 = -0.25$

ขั้นตอนที่ 3 : คำนวณค่าอัตราการจำแนกผิด (misclassification rate) และค่าการประเมินการจัดกลุ่มของ Kalpakis ของแต่ละมาตรวัดระยะห่างทั้ง 8 แบบ เพื่อใช้ในการเปรียบเทียบประสิทธิภาพการจัดกลุ่มของแต่ละมาตรวัดระยะห่างต่อไป

ขั้นตอนที่ 4 : ทำซ้ำ 1,000 ครั้ง และคำนวณหาค่าเฉลี่ยของค่าอัตราการจำแนกผิด (misclassification rate) และค่าการประเมินการจัดกลุ่มของ Kalpakis ของแต่ละมาตรวัดระยะห่างเมื่อทำการจำลองครบ 1,000 ครั้ง

ขั้นตอนที่ 5 : เปรียบเทียบประสิทธิภาพการจัดกลุ่มของมาตรวัดระยะห่างทั้ง 8 แบบ โดยมาตรวัดใดมีค่าเฉลี่ยของอัตราการจำแนกผิด (misclassification rate) ต่ำที่สุด หรือมีค่าเฉลี่ยการประเมินการจัดกลุ่มของ Kalpakis สูงที่สุด จะเป็นมาตรวัดระยะห่างที่มีประสิทธิภาพมากที่สุดสำหรับข้อมูลประเภทอนุกรมเวลา

2.3 การเปรียบเทียบในชุดข้อมูลจริง

เปรียบเทียบในชุดข้อมูลจำลอง โดยมีขั้นตอนดังต่อไปนี้

ขั้นตอนที่ 1 : เก็บรวบรวมข้อมูลราคาปิดรายวันของ 68 หลักทรัพย์ ในดัชนี SET100 ประจำปี พ.ศ. 2561 ระหว่างวันที่ 3 มกราคม พ.ศ. 2561 ถึงวันที่ 30 เมษายน พ.ศ. 2561 จำนวนทั้งสิ้น 80 วัน

ขั้นตอนที่ 2 : วิเคราะห์ตัวแบบอนุกรมเวลา ARIMA สำหรับข้อมูลราคาปิดของหลักทรัพย์ในขั้นตอนที่ 1 ทั้งนี้ผู้วิจัยสามารถแบ่งกลุ่มหลักทรัพย์ 6 กลุ่ม ตามลักษณะตัวแบบอนุกรมเวลาดังแสดงในตารางที่ 3

ขั้นตอนที่ 3 : จัดกลุ่มข้อมูลด้วยการวิเคราะห์ การจัดกลุ่มข้อมูลด้วย algorithm K-medoids โดยเลือกจำนวนกลุ่มของการจัดกลุ่มเป็น 6 กลุ่ม ซึ่งเท่ากับจำนวนกลุ่มที่แท้จริงจากขั้นตอนที่ 2 และเลือกใช้มาตรวัดระยะห่างสำหรับข้อมูลอนุกรม 8 แบบ ประกอบด้วย Euclidean, Minkowski, dynamic time warping, Chouakria- Douzal, Piccolo, Maharaj, discrete wavelet transform และ cepstral-based

ขั้นตอนที่ 4 : คำนวณค่าอัตราการจำแนกผิด (misclassification rate) และค่าการประเมินการจัดกลุ่มของ Kalpakis ของแต่ละมาตรวัดระยะห่างทั้ง 8 วิธี เพื่อใช้ในการเปรียบเทียบประสิทธิภาพการจัดกลุ่มของแต่ละมาตรวัดระยะห่างต่อไป

ขั้นตอนที่ 5 : เปรียบเทียบประสิทธิภาพ เปรียบเทียบกับผลที่ได้จากข้อมูลจำลอง
การจัดกลุ่มของมาตรวัดระยะห่างทั้ง 8 แบบ และนำไป

ตารางที่ 3 หลักทรัพย์ในแต่ละกลุ่มและลักษณะตัวแบบอนุกรมเวลาที่เกี่ยวข้อง

กลุ่ม	หลักทรัพย์	จำนวนหลักทรัพย์ในกลุ่ม	รูปแบบอนุกรมเวลา
1	PSH BEC MAJOR ESSO TASCO BANPU UV HANA MINT CENTEL	10	ARIMA(1,1,0)
2	TTA STA GFPT TRUE BH CK BPP BCPG BIG KTC KTB PTG KKP	13	ARIMA(0,1,1)
3	SGP THAI BA TKN HMPRO SCC	6	ARIMA(2,1,1)
4	CPN BTS SPRC JMART KBANK LPN JWD AAV BEAUTY GLOBAL TVO	11	ARIMA(0,1,2)
5	PTT PTTGC BCH MEGA INTUCH BBL AOT EGCO CKP CPALL BEM	11	ARIMA(2,0,0)
6	SIRI CPF GUNKUL STEC GGC WHA LH WORK TMB BJC QH MC SAWAD ITD SCB GPSC IVL	17	ARIMA(1,1,1)

ตารางที่ 4 ผลการเปรียบเทียบประสิทธิภาพของมาตรวัดระยะห่างในการวิเคราะห์การจัดกลุ่มเมื่อขนาดข้อมูล
จำลองในแต่ละกลุ่มเป็น 50, 100 และ 300

มาตรวัดระยะห่างใน การวิเคราะห์การจัดกลุ่ม	ค่าอัตราการจำแนกผิด			ค่าการประเมินการจัดกลุ่มของ Kalpakis		
	$N = 50$	$N = 100$	$N = 300$	$N = 50$	$N = 100$	$N = 300$
Time domain distance measure						
Euclidean	0.477	0.475	0.469	0.502	0.505	0.601
Minkowski	0.346	0.339	0.337	0.582	0.591	0.595
Dynamic time warping	0.201	0.195	0.188	0.820	0.834	0.835
Chouakria-Douzal	0.261	0.259	0.254	0.765	0.766	0.771
Piccolo	0.279	0.279	0.279	0.721	0.721	0.721
Maharaj	0.266	0.266	0.266	0.733	0.733	0.733
Transform-based distance measure						
Discrete wavelet transform	0.374	0.370	0.368	0.601	0.604	0.607
Cepstral-based	0.318	0.305	0.299	0.628	0.630	0.631

3. ผลการวิจัย

งานวิจัยนี้นำเสนอผลการเปรียบเทียบโดยแบ่งออกเป็น 2 ส่วน ในส่วนที่ 1 เปรียบเทียบประสิทธิภาพของมาตรวัดระยะห่างในการวิเคราะห์การจัดกลุ่ม เมื่อมีการใช้มาตรวัดระยะห่างที่แตกต่างกัน 8 แบบ โดยใช้ชุดข้อมูลจำลอง และส่วนที่ 2 เปรียบเทียบประสิทธิภาพ

ภาพของการใช้มาตรวัดระยะห่างในการวิเคราะห์การจัดกลุ่ม เมื่อมีการใช้มาตรวัดชุดเดียวกันกับในส่วนที่ 1 โดยใช้ชุดข้อมูลจริง ซึ่งเป็นราคาปิดรายวันของ 68 หลักทรัพย์ ในดัชนี SET100 จำนวนทั้งสิ้น 80 วัน โดยผลการเปรียบเทียบสำหรับทั้งสองส่วนนั้น แสดงในตารางที่ 4 และ 5 ตามลำดับ

ตารางที่ 5 ผลการเปรียบเทียบประสิทธิภาพของมาตรวัดระยะห่างในการวิเคราะห์การจัดกลุ่มในชุดข้อมูลจริง

มาตรวัดระยะห่างในการวิเคราะห์การจัดกลุ่ม	ค่าอัตราการจำแนกผิด	ค่าการประเมินการจัดกลุ่มของ Kalpakis
Time domain distance measure		
Euclidean	0.521	0.531
Minkowski	0.487	0.590
Dynamic time warping	0.248	0.785
Chouakria-Douzal	0.314	0.692
Piccolo	0.388	0.645
Maharaj	0.375	0.638
Transform-based distance measure		
Discrete wavelet transform	0.466	0.576
Cepstral-based	0.459	0.588

ตารางที่ 5 พบว่าสำหรับชุดข้อมูลจริงที่เป็นข้อมูลราคาปิดรายวันของ 68 หลักทรัพย์ ในดัชนี SET100 จำนวนทั้งสิ้น 80 วัน มาตรวัดระยะห่างแบบ dynamic time warping ให้ประสิทธิภาพในการจัดกลุ่มสูงที่สุด เนื่องจากมีค่าอัตราการจำแนกผิดต่ำที่สุด และมีค่าการประเมินการจัดกลุ่มของ Kalpakis สูงที่สุด หลังจากจัดกลุ่มด้วย algorithm K-medoids ส่วนมาตรวัดระยะห่างแบบ Chouakria-Douzal ให้ประสิทธิภาพในการจัดกลุ่มรองลงมา และมาตรวัดที่มีประสิทธิภาพต่ำสุด คือ มาตรวัดระยะห่างแบบ Euclidean เนื่องจากมีค่าอัตราการจำแนกผิดสูงที่สุด และมีค่าการประเมินการจัดกลุ่มของ Kalpakis ต่ำ

ที่สุด ซึ่งผลการเปรียบเทียบที่ได้จากข้อมูลจริงสอดคล้องกับผลการเปรียบเทียบในชุดข้อมูลจำลอง

4. สรุปผลการวิจัย

งานวิจัยนี้นำเสนอผลการเปรียบเทียบประสิทธิภาพของการใช้การวิเคราะห์การจัดกลุ่มในชุดข้อมูลอนุกรมเวลา เมื่อมีการใช้มาตรวัดระยะห่าง 8 แบบ โดยแบ่งการวิเคราะห์เป็น 2 ส่วน ส่วนที่ 1 เปรียบเทียบประสิทธิภาพของการใช้การวิเคราะห์การจัดกลุ่มด้วยชุดข้อมูลจำลองที่แต่ละกลุ่มถูกจำลองด้วยตัวแบบ ARIMA และส่วนที่ 2 เปรียบเทียบประสิทธิภาพของการใช้การวิเคราะห์การจัดกลุ่มด้วยชุดข้อมูลจริง โดย

ใช้ราคาปิดรายวันของ 68 หลักทรัพย์ ในดัชนี SET100 จำนวนทั้งสิ้น 80 วัน ตั้งแต่วันที่ 3 มกราคม พ.ศ. 2561 ถึงวันที่ 30 เมษายน พ.ศ. 2561

เมื่อใช้ค่าอัตราการจัดจำแนกผิด (misclassification rate) และค่าการประเมินการจัดกลุ่มของ Kalpakis เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพการจัดกลุ่ม พบว่ามาตรวัดระยะห่างแบบ dynamic time warping ให้ประสิทธิภาพในการจัดกลุ่มสูงสุด ซึ่งสอดคล้องกับผลวิจัยของ Pertega และ Vilar [18] รองลงมา คือ มาตรวัดระยะห่างแบบ Chouakria-Douzal ส่วนมาตรวัดที่มีประสิทธิภาพต่ำสุด คือ มาตรวัดระยะห่างแบบ Euclidean ซึ่งให้ผลสอดคล้องกันทั้งในชุดข้อมูลจริงและชุดข้อมูลจำลอง และหากพิจารณาถึงขนาดระยะเวลาในแต่ละกลุ่ม พบว่าเมื่อชุดข้อมูลอนุกรมเวลามีขนาดใหญ่ขึ้น มาตรวัดข้อมูลส่วนใหญ่จะมีประสิทธิภาพสูงขึ้น ยกเว้นมาตรวัดแบบ Piccolo และ Maharaj

ทั้งนี้ถ้าเปรียบเทียบประสิทธิภาพของการจัดกลุ่มในกลุ่มมาตรวัดระยะห่างฐานเปลี่ยนรูปนั้น (transform-based distance measure) พบว่ามาตรวัดระยะห่างแบบ cepstral-based มีประสิทธิภาพสูงกว่ามาตรวัดระยะห่างแบบ discrete wavelet transform ซึ่งสอดคล้องกับผลวิจัยของ Kalpakis [17] แต่ก็ยังไม่สูงเท่ามาตรวัดระยะห่างแบบ dynamic time warping

5. ข้อเสนอแนะ

ผลที่ได้จากงานวิจัยนี้พบว่ามาตรวัดที่มีประสิทธิภาพสูงที่สุดในการใช้การวิเคราะห์การจัดกลุ่มสำหรับข้อมูลอนุกรมเวลา คือ มาตรวัดระยะห่างแบบ dynamic time warping ทั้งนี้ปัญหาหนึ่งของมาตรวัดดังกล่าว คือ การใช้เวลาในการคำนวณที่นานกว่ามาตรวัดแบบอื่น (computationally expensive) ดังนั้นผู้

สนใจสามารถเปรียบเทียบมาตรวัดระยะห่างที่ดัดแปลงจากมาตรวัดระยะห่างแบบ dynamic time warping เช่น มาตรวัดระยะห่างแบบ global alignment Kernel distance [27] หรือมาตรวัดระยะห่างแบบ soft-DTW [28] ในอนาคต

6. รายการอ้างอิง

- [1] กัลยา วานิชย์บัญชา, 2551, การวิเคราะห์ข้อมูลหลายตัวแปร, พิมพ์ครั้งที่ 3, ธรรมสาร, กรุงเทพฯ.
- [2] สุชาติ ประสิทธิ์รัฐสินธุ์, 2540, เทคนิคการวิเคราะห์ตัวแปรหลายตัวสำหรับการวิจัยทางสังคมศาสตร์, พิมพ์ครั้งที่ 4, บริษัท เพื่องฟ้าพรินต์ติ้ง จำกัด, กรุงเทพฯ.
- [3] Patidar, A. K., Agrawal, J. and Mishra, N., 2012, Analysis of different similarity measure functions and their impacts on shared nearest neighbor clustering approach, Int. J. Comp. Appl. 40(16): 1-5.
- [4] Tong, H. and Dabas, P., 1990, Cluster of time series models: An example, J. Appl. Stat. 17: 187-198.
- [5] Rani, S. and Sikka, G., 2012, Recent techniques of clustering of time series data: A survey, Int. J. Comp. Appl. 52(15): 1-9.
- [6] Galeano, P. and Peña, D. P., 2000, Multivariate analysis in vector time series, Resenhas do Instituto de Matemática e Estatística da Universidade de São Paulo 4: 383-403.
- [7] Keogh, E. and Kasetty, S., 2003, On the need for time series data mining benchmarks: A survey and empirical

- demonstration, Data Min. Knowl. Disc. 7: 349-371.
- [8] Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P. and Keogh, E., 2013, Experimental comparison of representation methods and distance measures for time series data, Data Min. Knowl. Disc. 26: 275-309.
- [9] Kianimajd, A., Ruano, M.G., Carvalho, P., Henriques, J., Rocha, T., Paredes, S. and Ruano, A.E., 2017, Comparison of different methods of measuring similarity in physiologic time series, IFAC- Papers On Line 50(1): 11005-11010.
- [10] Sankoff, D., 1983, Time Warps, String Edits and Macromolecules, The Theory and Practice of Sequence Comparison, Reading.
- [11] Chouakria, A.D. and Nagabhushan, P.N., 2007, Adaptive dissimilarity index for measuring time series proximity, Adv. Data Anal. Classif. 1: 5-21.
- [12] Piccolo, D., 1990, A distance measure for classifying ARIMA models, J. Time Ser. Anal. 11: 153-164.
- [13] Maharaj, E.A., 1996, A significance test for classifying ARMA models, J. Stat. Comp. Simul. 54: 305-331.
- [14] Maharaj, E. A. , 2000, Cluster of time series, J. Classif. 17: 297-314.
- [15] Shasha, D. and Zhu, Y. , 2004, High Performance Discovery in Time Series: Techniques and Case Studies, Ph. D. Thesis, Available Source: https://cs.nyu.edu/media/publications/zhu_yunyue.pdf.
- [16] Wei, W. S. , 2006, Time Series Analysis: Univariate and Multivariate Methods, Pearson Addison Wesley, 605 p.
- [17] Kalpakis, K., Gada, D. and Puttagunta, V., 2001, Distance measures for effective clustering of ARIMA time-series, pp. 273-280, In Data Mining, ICDM 2001, IEEE International Conference.
- [18] Diaz, S.P. and Vilar, J.A., 2010, Comparing several parametric and nonparametric approaches to time series clustering: A simulation study, J. Classif. 27: 333-362.
- [19] Khan, A., Khan, K. and Baharudin, B.B., 2009, Frequent patterns mining of stock data using hybrid clustering association algorithm, pp. 667-671, In Information Management and Engineering, ICIME 2009, International Conference.
- [20] Kremer, H., Gunnemann, S. and Seidl, T., 2010, Detecting climate change in multivariate time series data by novel clustering and cluster tracing techniques, pp. 96-97, In Data Mining Workshops (ICDMW) , 2010 IEEE International Conference.
- [21] Niennattrakul, V. and Ratanamahatana, C.A., 2007, On clustering multimedia time series data using K-means and dynamic time warping, pp. 733-738, IEEE.
- [22] Jixue, D., 2009, Data mining of time series based on wave cluster, pp. 697-699, In

- Information Technology and Applications, IFITA 2009, International Forum, IEEE.
- [23] Verdoolaege, G. and Rosseel, Y., 2010, Activation detection in event-related fMRI through clustering of wavelet distributions, pp. 4393-4396, In Image Processing (ICIP) , 2010 17th IEEE International Conference.
- [24] Baragona, R., 2001, A simulation study on clustering time series with metaheuristic methods, Quaderni di Statistica 3: 1-26.
- [25] Lhermitte, S., Verbesselt, J., Verstraeten, W.W. and Coppin, P., 2011, A comparison of time series similarity measures for classification and change detection of ecosystem dynamics, Remote Sens. Environ. 115: 3129-3152.
- [26] Shumway, R.H. and Stoffer, D.S., 2010, Time Series Analysis and Its Applications with R Examples, 3rd Ed., Springer, 576 p.
- [27] Cuturi, M., 2011, Fast global alignment Kernels) , pp. 929- 936, In 28th International Conference on Machine Learning (ICML-11).
- [28] Cuturi, M. and Blondel, M., 2017, Soft-DTW: A differentiable loss function for time- series, pp. 894-903, In 34th International Conference on Machine Learning.