กาจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา

นางสาววริศรา มีศรีกมลกุล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2554

SHAPE-BASED CLUSTERING FOR TIME SERIES DATA

Ms. Warissara Meesrikamolkul

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering Program in Computer Engineering

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2011

หัวข้อวิทยานิพนธ์	การจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา
โดย	นางสาววริศรา มีศรีกมลกุล
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ

คณะวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

…….. คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

……. ประธานกรรมการ
(ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

……... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ)

……. กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ)

……. กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร.กฤษณะ ไวยมัย)

วิศรา มีศรีกมลกุล : การจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา. (SHAPE-BASED CLUSTERING FOR TIME SERIES DATA) อ. ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ. ดร.โชติรัตน์ รัตนามหัทธนะ, 51 หน้า.

การจัดกลุ่มข้อมูลอนุกรมเวลา เป็นหนึ่งในการทำเหมืองข้อมูลของข้อมูลอนุกรมเวลาที่นักวิจัยส่วนใหญ่ให้ความสนใจ โดยอัลกอริทึมที่นิยมนำมาใช้ คือ การจัดกลุ่มแบบเคมีนส์ (*K*-means Clustering) ร่วมกับมาตรวัดระยะยุคลิด และหาตัวแทนกลุ่มด้วยวิธีการหาค่าเฉลี่ย หรือการเฉลี่ยแบบแอมพลิจูด ซึ่งเป็นวิธีที่ไม่เหมาะกับลักษณะของข้อมูลอนุกรมเวลา เพราะเป็นข้อมูลที่มีการเลื่อนในแนวแกนเวลา

งานวิจัยนี้จึงนำเสนอการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา โดยมีแนวคิดในการนำระยะไดนามิกไทม์วอร์ปปิง ซึ่งเป็นมาตรวัดที่เหมาะกับลักษณะของข้อมูลอนุกรมเวลามากกว่ามาใช้ร่วมกับการจัดกลุ่มแบบเคมีนส์แทนระยะยุคลิด และได้เสนอวิธี Ranking Shape-based Template Matching Framework (RSTMF) ซึ่งเป็นการหาตัวแทนกลุ่มโดยใช้ระยะไดนามิกไทม์วอร์ปปิง เพื่อนำมาใช้แทนการเฉลี่ยแบบแอมพลิจูด นอกจากนี้ยังได้ทำการวัดผลโดยการเปรียบเทียบความแม่นยำระหว่างการจัดกลุ่มแบบเคมีนส์แบบทั่วไปที่ใช้ระยะยุคลิดและการเฉลี่ยแบบแอมพลิจูด กับวิธีการจัดกลุ่มตามรูปร่าง ซึ่งเป็นการจัดกลุ่มแบบเคมีนส์ร่วมกับระยะไดนามิกไทม์วอร์ปปิงและการหาตัวแทนกลุ่มด้วยวิธี RSTMF ซึ่งให้ผลการจัดกลุ่มข้อมูลอนุกรมเวลาที่แม่นยำมากขึ้น เมื่อเทียบกับการจัดกลุ่มแบบเคมีนส์แบบทั่วไป

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา 2554

5370487121 : MAJOR COMPUTER ENGINEERING
KEYWORDS : TIME SERIES / CLUSTERING / SHAPE-BASED AVERAGING

WARISSARA MEESRIKAMOLKUL : SHAPE-BASED CLUSTERING FOR TIME SERIES DATA. ADVISOR : ASST.PROF. CHOTIRAT RATANAMAHATANA, Ph.D., 51 pp.

Time series data clustering is one of the most active tasks in time series mining. *K*-means clustering using Euclidean distance as a similarity measure is a popular clustering algorithm and a representative or a new cluster center is usually calculated using an amplitude averaging function. However, Euclidean distance metric and amplitude averaging are not suitable for time series data because time shifting can be occurred in time series data.

In this research, the Shape-based Clustering for Time Series Data (SCTS) which incorporates *k*-means clustering and DTW distance measure, together with our new averaging method, called Ranking Shape-based Template Matching Framework (RSTMF) as an averaging function, which can provide a new cluster center that preserves the overall characteristics of time series data. In the experiment, our proposed method outperforms the traditional *k*-means clustering technique in term of accuracy.

Department : Computer Engineering Student's Signature .....
Field of Study : Computer Engineering Advisor's Signature .....
Academic Year : 2011

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สามารถสำเร็จลุล่วงไปได้ด้วยดี เนื่องจากการสนับสนุนจากบุคคลหลายฝ่าย ที่มีส่วนช่วยให้ผู้จัดทำสามารถเรียนรู้และแก้ไขปัญหาต่าง ๆ ได้

ขอขอบพระคุณอาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ ที่คอยแนะนำและส่งเสริม พร้อมทั้งให้ข้อคิดเห็นอันเป็นแนวทางในการปรับปรุงข้อบกพร่องต่าง ๆ มากมาย

ขอขอบพระคุณคณะกรรมการสอบวิทยานิพนธ์ ซึ่งประกอบด้วย ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ และรองศาสตราจารย์ ดร.กฤษณะ ไวยมัย ที่ให้คำวิจารณ์และข้อเสนอแนะเพื่อนำไปพัฒนางานวิจัยต่อไป

ขอขอบพระคุณ ดร.วิชญ์ เนียรนาทตระกูล และสมาชิกในห้องปฏิบัติการทุกคน ที่เป็นแรงผลักดันสำคัญ ทำให้การทำวิจัยดำเนินไปได้อย่างมีชีวิตชีวา และทำให้ห้องปฏิบัติการเป็นมากกว่าสถานที่สำหรับการทำวิจัย

สารบัญ

หน้า

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

หน้า

หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

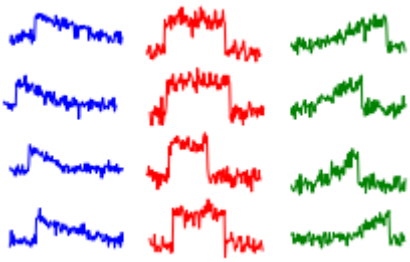
ปัจจุบันข้อมูลอนุกรมเวลาได้รับความนิยมมากขึ้นและมีการนำไปใช้ในงานต่าง ๆ อย่างแพร่หลาย เนื่องจากข้อมูลอนุกรมเวลานั้นสามารถประยุกต์ใช้กับข้อมูลได้หลายสาขา เช่น ข้อมูลทางชีวภาพ [1], มัลติมีเดีย [2][3], ข้อมูลทางการแพทย์ [4] และข้อมูลทางธรณีวิทยา [5] ด้วยเหตุนี้จึงมีงานวิจัยที่เกี่ยวข้องกับข้อมูลอนุกรมเวลาเกิดขึ้นมากมาย เช่น การจำแนก (Classification) [6], การจัดกลุ่ม (Clustering) [5][7], การค้นหาโมทีฟ (Motif discovery) [8][9], การทำดัชนี (Indexing) [10] และการตรวจสอบความผิดปกติ (Anomaly detection) [11]

การจัดกลุ่มข้อมูล (Clustering) ถือเป็นหนึ่งในการทำเหมืองข้อมูลที่นักวิจัยให้ความสนใจ ซึ่งอัลกอริทึมที่นิยมใช้คือ การจัดกลุ่มแบบเคมีนส์ (*K*-means clustering) [12] สำหรับข้อมูลทั่วไป การจัดกลุ่มแบบเคมีนส์จะนิยมใช้การวัดระยะยุคลิด และการหาตัวแทนของกลุ่มข้อมูลก็จะใช้วิธีการหาค่าเฉลี่ย (Arithmetic mean) อย่างไรก็ตามการวัดระยะยุคลิดซึ่งมีการจำกัดการวางแนวระหว่างอนุกรมเวลาให้เป็นแบบหนึ่งต่อหนึ่ง รวมไปถึงการหาตัวแทนกลุ่มข้อมูลโดยการหาค่าเฉลี่ยซึ่งเป็นการเฉลี่ยโดยใช้ระยะยุคลิดหรือการเฉลี่ยแอมพลิจูด นั้นไม่เหมาะกับลักษณะของข้อมูลอนุกรมเวลา จึงมีการนำระยะไดนามิกไทม์วอร์ปปิง [13] มาใช้เป็นมาตรวัดแทนระยะยุคลิดเพราะมีความยืดหยุ่นในการวางแนว (Alignment) ระหว่างอนุกรมเวลา ทำให้มีความแม่นยำสูง

ถึงแม้ว่าระยะไดนามิกไทม์วอร์ปปิงจะเป็นมาตรวัดที่มีความแม่นยำสูงเมื่อนำมาใช้ในงานด้านต่าง ๆ ที่เกี่ยวกับข้อมูลอนุกรมเวลา เช่น การจำแนกข้อมูล แต่การจัดกลุ่มแบบเคมีนส์โดยใช้ระยะไดนามิกไทม์วอร์ปปิงยังคงไม่สามารถใช้งานได้อย่างสมบูรณ์ [14] เนื่องจากวิธีการเฉลี่ยข้อมูลอนุกรมเวลาเพื่อหาตัวแทนกลุ่มในปัจจุบัน ยังไม่สามารถรักษารูปร่างของอนุกรมเวลาไว้ได้

ในงานวิทยานิพนธ์นี้จึงนำเสนอการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลาโดยใช้การจัดกลุ่มแบบเคมีนส์ร่วมกับระยะไดนามิกไทม์วอร์ปปิง นอกจากนี้ยังเสนอวิธีการเฉลี่ยเพื่อหาตัวแทนกลุ่มโดยทำการปรับปรุงจาก Shape-based Template Matching Framework (STMF) [15] ซึ่งเป็นการเฉลี่ยโดยใช้ระยะไดนามิกไทม์วอร์ปปิง เพื่อให้มีการทำงานที่รวดเร็วยิ่งขึ้น

แต่ยังคงสามารถนำไปประยุกต์ใช้กับการจัดกลุ่มแบบเคมีนส์ได้อย่างมีประสิทธิภาพ ดังแสดงในภาพที่ 1.1



ก)

ข)

ค)

ภาพที่ 1.1 ก) ตัวอย่างข้อมูลอนุกรมเวลาจากชุดข้อมูล CBF (Cylinder-Bell-Funnel) ซึ่งมี 3 คลาส ข) ตัวแทนของกลุ่มข้อมูล เมื่อทำการจัดกลุ่มแบบเคมีนส์ตามปกติ และ ค) เมื่อใช้วิธีการจัดกลุ่มตามรูปร่าง

1.2 วัตถุประสงค์ของการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อเพิ่มความแม่นยำในการจัดกลุ่มข้อมูลอนุกรมเวลา โดยใช้วิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา ซึ่งเป็นการนำการจัดกลุ่มแบบเคมีนส์มาประยุกต์ใช้ร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิงและการหาตัวแทนกลุ่มข้อมูลด้วยวิธีการเฉลี่ยแบบรูปร่าง

1.3 ขอบเขตของการวิจัย

1. พัฒนาวิธีการจัดกลุ่มข้อมูลสำหรับข้อมูลอนุกรมเวลาโดยนำการจัดกลุ่มแบบเคมีนส์มาประยุกต์ใช้ร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิงและการเฉลี่ยแบบรูปร่าง

2. พัฒนาวิธีการเฉลี่ยแบบรูปร่างสำหรับข้อมูลอนุกรมเวลาโดยการปรับปรุงเพิ่มเติมจากวิธีการ Shape-based Template Matching Framework (STMF) [15] ให้มีการทำงานที่รวดเร็วขึ้น

3. ทำการวัดผลวิธีการจัดกลุ่มโดยการทดลองกับข้อมูลอนุกรมเวลาโดยใช้ชุดข้อมูลสำหรับการจำแนกและการจัดกลุ่มจาก University of California, Riverside (UCR) [16]

4. ข้อมูลที่ใช้ในการทดลองนั้นได้มีการแบ่งเป็นคลาสไว้แล้ว จึงทำการวัดผลด้วยการเปรียบเทียบความแม่นยำกับการจัดกลุ่มข้อมูลแบบเคมีนส์ซึ่งใช้ระยะยุคลิดและการเฉลี่ยแบบแอมพลิจูด

1.4 ประโยชน์ที่ได้รับ

1. ได้อัลกอริทึมสำหรับการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา ที่มีความแม่นยำในการจัดกลุ่มข้อมูลเพิ่มมากขึ้น

2. ได้อัลกอริทึมสำหรับการหาตัวแทนกลุ่มของข้อมูลอนุกรมเวลาที่สามารถทำงานได้อย่างรวดเร็วมากยิ่งขึ้น นอกจากนี้ยังสามารถนำวิธีการหาตัวแทนกลุ่มดังกล่าวไปประยุกต์ใช้ในการหาแผ่นแบบ เพื่อใช้ในการจับคู่แผ่นแบบ (Template matching) ได้

1.5 วิธีดำเนินการวิจัย

1. ศึกษาทฤษฎี เอกสารและงานวิจัยที่เกี่ยวข้องกับข้อมูลอนุกรมเวลาและการทำเหมืองข้อมูลอนุกรมเวลาเพื่อทำการวิเคราะห์หาปัญหาสำหรับนำไปปรับปรุงและพัฒนา

2. ศึกษางานวิจัยที่เกี่ยวข้องกับการจัดกลุ่มข้อมูลอนุกรมเวลาเพื่อกำหนดขอบเขตของปัญหาในการทำวิจัย

3. ศึกษาการจัดกลุ่มแบบเคมีนส์ (*K*-means clustering) เพื่อหาแนวทางในการปรับปรุงและพัฒนา

4. ศึกษางานวิจัยที่เกี่ยวข้องกับการหาตัวแทนกลุ่มของข้อมูลอนุกรมเวลาเพื่อนำมาใช้ในการทำวิจัย

5. ออกแบบอัลกอริทึมและทำการทดลองเบื้องต้น

6. ปรับปรุงวิธีการหาตัวแทนกลุ่มให้มีความเหมาะสมกับลักษณะของข้อมูลอนุกรมเวลาเพื่อให้สามารถนำมาประยุกต์ใช้กับการจัดกลุ่มแบบเคมีนส์ได้อย่างมีประสิทธิภาพ

7. ทำการทดลองเพื่อทดสอบประสิทธิภาพและความแม่นยำของวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลาที่นำเสนอโดยการเปรียบเทียบกับวิธีการจัดกลุ่มแบบเคมีนส์ซึ่งใช้ระยะยุคลิดเป็นมาตรวัดความคล้ายและใช้การเฉลี่ยแบบแอมพลิจูดในการหาตัวแทนกลุ่มของข้อมูล

8. วิเคราะห์และสรุปผลการทดลอง

9. ตีพิมพ์ผลงานและจัดทำวิทยานิพนธ์

1.6 ผลงานตีพิมพ์จากการวิจัย

ส่วนหนึ่งของวิทยานิพนธ์นี้ ได้รับการตีพิมพ์เป็นบทความทางวิชาการ 2 เรื่อง ดังนี้

- Multiple Shape-based Template Matching for Time Series Data โดย วริศรา มีศรีกมลกุล วิชญ์ เนียรนาทตระกูล และ โชติรัตน์ รัตนามหัทธนะ ในงานประชุมวิชาการ “8th International Conference Organized by Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI) Association, Thailand” ซึ่งจัดขึ้น ณ จังหวัดขอนแก่น ประเทศไทย ระหว่างวันที่ 17 พฤษภาคม ถึง 19 พฤษภาคม 2554

- Shape-based Clustering for Time Series Data โดย วริศรา มีศรีกมลกุล วิชญ์ เนียรนาทตระกูล และ โชติรัตน์ รัตนามหัทธนะ ในงานประชุมวิชาการ “16th Pacific-Asia Conference on Knowledge Discovery and Data Mining

(PAKDD)” ซึ่งจัดขึ้น ณ เมืองกัวลาลัมเปอร์ ประเทศมาเลเซีย ระหว่างวันที่ 29 พฤษภาคม ถึง 1 มิถุนายน 2555

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

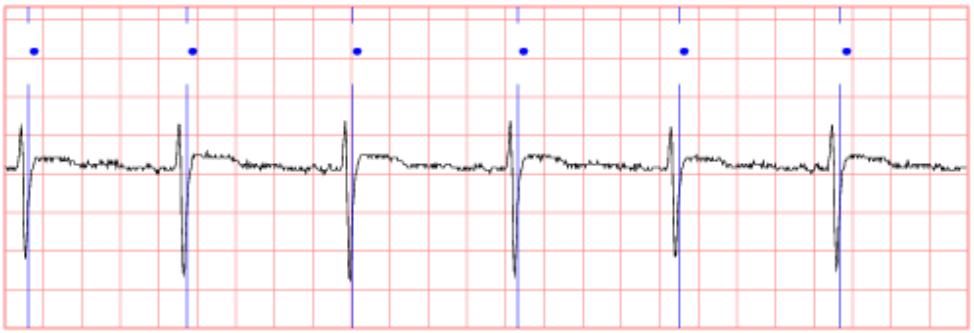
ในส่วนของเอกสารและงานวิจัยที่เกี่ยวข้องนั้น จะมีการกล่าวถึงแนวคิดและทฤษฎีที่ได้นำมาใช้ในงานวิทยานิพนธ์นี้ โดยเริ่มต้นจาก ข้อมูลอนุกรมเวลา (Time series data) เพื่อให้เกิดความเข้าใจในความหมายและลักษณะที่สำคัญของข้อมูลอนุกรมเวลา จากนั้นจึงกล่าวถึง การจัดกลุ่มแบบเคมีนส์ (*K*-means clustering) [12] โดยอธิบายถึงขั้นตอนการทำงานของการจัดกลุ่มแบบเคมีนส์ และลักษณะของการนำการจัดกลุ่มแบบเคมีนส์มาใช้กับข้อมูลอนุกรมเวลา นอกจากนี้ยังอธิบายเกี่ยวกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิง (DTW distance measure : Dynamic Time Warping distance measure) [13] ซึ่งเป็นมาตรวัดที่นิยมนำมาใช้กับข้อมูลอนุกรมเวลา เนื่องจากสามารถทำการปรับแนวได้ จึงทำให้มีความแม่นยำมากขึ้น และสุดท้ายได้อธิบายถึงการกำหนดเงื่อนไขบังคับโดยรวม (Global constraint) [17] ซึ่งเป็นวิธีการที่นำมาใช้ร่วมกับมาตรวัดระยะ ไดนามิกไทม์วอร์ปปิง เพื่อให้การปรับแนวเป็นไปอย่างเหมาะสม ซึ่งจะช่วยให้การวัดความคล้ายนั้นมีความแม่นยำมากขึ้น

นอกจากนี้ยังมีส่วนของงานวิจัยที่เกี่ยวข้องกับวิธีการจัดกลุ่มข้อมูลอนุกรมเวลา ซึ่งจะเป็นการนำเสนอถึงวิธีการจัดกลุ่มที่นิยมนำมาใช้กับข้อมูลอนุกรมเวลา และงานวิจัยเกี่ยวกับการเฉลี่ยแบบรูปร่างโดยใช้ระยะไดนามิกไทม์วอร์ปปิง ซึ่งเป็นวิธีการที่นำมาประยุกต์ใช้ในการหาตัวแทนกลุ่มกับข้อมูลอนุกรมเวลา

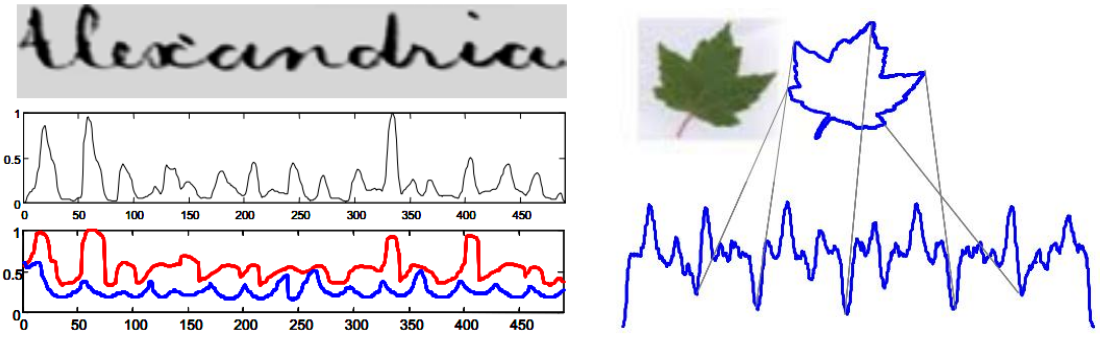
2.1 แนวคิดและทฤษฎี

2.1.1 ข้อมูลอนุกรมเวลา (Time series data)

ข้อมูลอนุกรมเวลา คือ ข้อมูลที่มีการเปลี่ยนแปลงไปตามเวลา ซึ่งพบได้ทั่วไปในชีวิตประจำวัน เช่น ข้อมูลอัตราการเต้นของหัวใจ [17] (ดังภาพที่ 2.1) นอกจากนี้ข้อมูลประเภทต่าง ๆ ยังสามารถนำมาแปลงให้เป็นข้อมูลอนุกรมเวลาได้อีกด้วย เช่น ลายมือ รูปภาพ และภาพเคลื่อนไหว ดังแสดงในภาพที่ 2.2

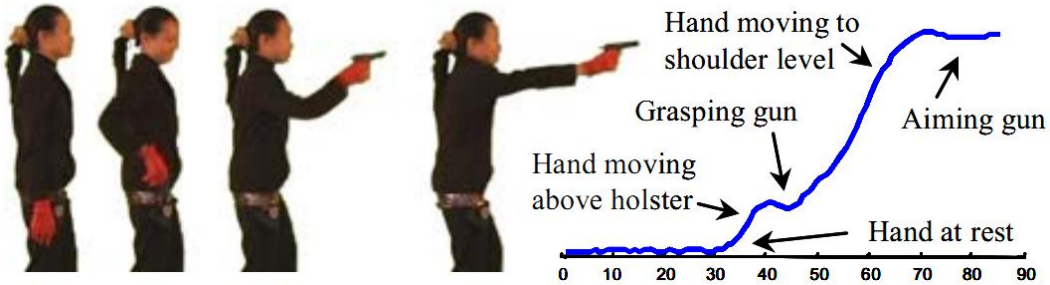


ภาพที่ 2.1 อัตราการเต้นของหัวใจ (ที่มา : Goldberger และคนอื่นๆ [17])



ก)

ข)



ค)

ภาพที่ 2.2 การแปลงข้อมูลที่อยู่ในรูปแบบอื่น ๆ ให้เป็นข้อมูลอนุกรมเวลา ก) ลายมือ ข) รูปภาพ ค) ภาพเคลื่อนไหว (ที่มา : Ratanamahatana และ Keogh [2])

2.1.2 การจัดกลุ่มแบบเคมีนส์ (*K*-means clustering)

การจัดกลุ่มแบบเคมีนส์ [12] ถือเป็นการจัดกลุ่มข้อมูลแบบแบ่งส่วนรูปแบบหนึ่ง (Partitional clustering) ซึ่งวิธีการจัดกลุ่มแบบนี้จะทำการจัดกลุ่มข้อมูลที่มีลักษณะคล้ายกันให้อยู่ในกลุ่มเดียวกันด้วยการหาค่าเหมาะสุด (Optimization) ของฟังก์ชันที่ใช้เป็นเกณฑ์ในการจัดกลุ่ม โดยมีขั้นตอนดังนี้

1. สุ่มข้อมูลเริ่มต้นขึ้นมา *K* ตัวเท่ากับจำนวนกลุ่มที่ต้องการแบ่งเพื่อใช้เป็นตัวแทนของกลุ่ม
2. ทำการจัดกลุ่มข้อมูลโดยการวัดค่าความคล้าย (Similarity measure) ระหว่างข้อมูลที่ต้องการจัดกลุ่มกับข้อมูลตัวแทน และจัดกลุ่มข้อมูลให้อยู่ในกลุ่มเดียวกับข้อมูลตัวแทนที่คล้ายกันมากที่สุด
3. ทำการเฉลี่ยข้อมูลแต่ละกลุ่มเพื่อหาตัวแทนของกลุ่มตัวใหม่

จากนั้นวนกลับไปทำซ้ำในขั้นตอนที่ 2 และ 3 ไปเรื่อย ๆ จนกระทั่งข้อมูลที่เป็นสมาชิกในแต่ละกลุ่มไม่มีการเปลี่ยนแปลง ซึ่งการสุ่มข้อมูลเริ่มต้นและผลลัพธ์ที่ได้จากการจัดกลุ่มแบบเคมีนส์แสดงในภาพที่ 2.3 ก) และ ข) ตามลำดับ

ก) ข)

ภาพที่ 2.3 ก) ลักษณะของข้อมูลการสุ่มข้อมูลเริ่มต้นก่อนทำการจัดกลุ่ม
ข) ผลลัพธ์เมื่อทำการจัดกลุ่มแบบเคมีนส์ (ที่มา : Tan [19])

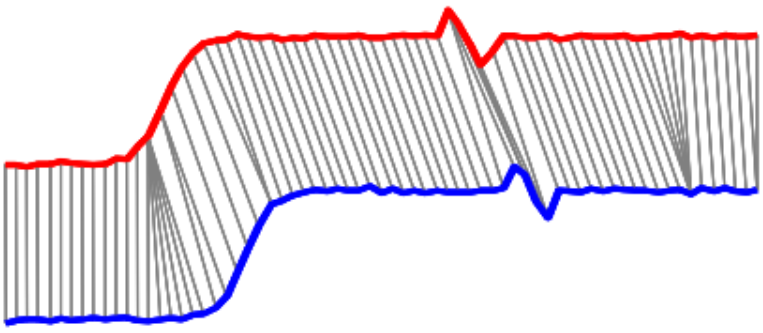
จะเห็นได้ว่า การจัดกลุ่มแบบเคมีนส์ประกอบไปด้วยส่วนสำคัญ 2 ส่วน โดยส่วนแรก คือ มาตรวัดที่ใช้ในการวัดความคล้ายของข้อมูลเพื่อวัดความคล้ายระหว่างข้อมูลแต่ละตัวกับ

ข้อมูลที่เป็นตัวแทนกลุ่ม เพื่อทำการจัดข้อมูลให้อยู่ในกลุ่ม ส่วนที่สอง คือ วิธีการหาตัวแทนข้อมูลเพื่อนำมาใช้เป็นตัวแทนกลุ่มสำหรับการจัดกลุ่มในรอบถัดไป

การจัดกลุ่มข้อมูลอนุกรมเวลาด้วยวิธีการจัดกลุ่มแบบเคมีนส์ โดยทั่วไปจะใช้มาตรวัดระยะยุคลิดเป็นมาตรวัดความคล้ายระหว่างข้อมูลและใช้การเฉลี่ยแบบแอมพลิจูดในการหาตัวแทนกลุ่ม อย่างไรก็ตาม เนื่องจากลักษณะการวางแนวที่เป็นแบบหนึ่งต่อหนึ่งของมาตรวัดระยะยุคลิดนั้นไม่เหมาะกับลักษณะของข้อมูลอนุกรมเวลาที่มักจะมีการเลื่อนของข้อมูลในแนวแกนเวลา รวมถึงวิธีการเฉลี่ยข้อมูลแบบแอมพลิจูด ซึ่งจะทำให้อนุกรมเวลาที่ได้จากการเฉลี่ยนั้นมีลักษณะที่แตกต่างจากอนุกรมเวลาที่นำมาเฉลี่ย จึงทำให้ผลที่ได้จากการจัดกลุ่มข้อมูลอนุกรมเวลาด้วยวิธีการดังกล่าวยังไม่มีความแม่นยำ ดังนั้นในงานวิจัยจึงจะเสนอวิธีการเพื่อปรับปรุงการจัดกลุ่มข้อมูลอนุกรมเวลาให้มีความแม่นยำมากยิ่งขึ้น

2.1.3 **มาตรวัดระยะไดนามิกไทม์วอร์ปปิง** (DTW distance measure : Dynamic Time Warping distance measure)

มาตรวัดระยะไดนามิกไทม์วอร์ปปิง [13] เป็นมาตรวัดที่ใช้ในการวัดความคล้าย (Similarity measure) โดยสามารถนำมาวัดความคล้ายระหว่างข้อมูลอนุกรมเวลาสองอนุกรม โดยนิยมนำมาใช้กับข้อมูลอนุกรมเวลาเนื่องจากมีความยืดหยุ่น จึงสามารถปรับการวางแนวระหว่างสองอนุกรมเวลาได้เหมาะสมที่สุด (ดังแสดงในภาพที่ 2.4) ทำให้การวัดความคล้ายของอนุกรมเวลามีความแม่นยำมากยิ่งขึ้น



ภาพที่ 2.4 การวัดระยะไดนามิกไทม์วอร์ปปิง (ที่มา : Keogh และ Ratanamahatana [17])

สมมติให้ P และ Q เป็นข้อมูลอนุกรมเวลามีขนาดยาว n และ m ตามลำดับ ที่ต้องการทำการวัดระยะไดนามิกไทม์วอร์ปปิง มี p_i และ q_j เป็นข้อมูลแต่ละจุด โดยที่ i และ j แทนค่าตำแหน่งในแนวแกนเวลาของข้อมูลอนุกรมเวลา P และ Q ตามลำดับ จากนั้นสร้างเมทริกซ์

ขนาด n x m และคำนวณระยะระหว่างแต่ละจุดข้อมูลของอนุกรมตามสมการที่ (2.1) จากนั้นทำการคำนวณแบบพลวัต (Dynamic programming) โดยใช้สมการที่ (2.2) ในการสะสมค่าระยะระหว่างแต่ละจุดข้อมูลของอนุกรมเวลา ซึ่งค่าระยะไดนามิกไทม์วอร์ปปิงจะเป็นค่าสุดท้ายและสามารถหาได้ตามสมการที่ (2.3)

$$D(p_i, q_j) = (q_i - c_j)^2 \tag{2.1}$$

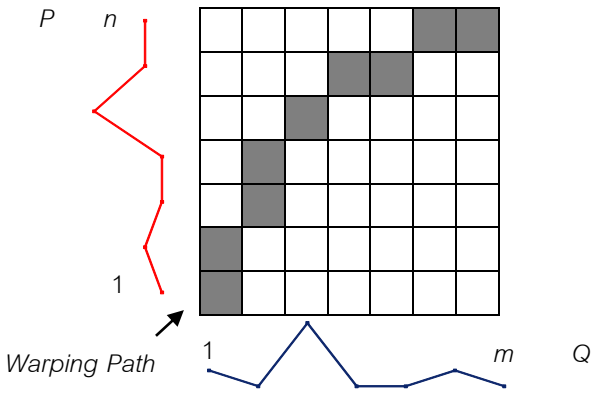
$$dist(p_i, q_j) = D(p_i, q_j) + \min\begin{cases} dist(p_{i-1}, q_j) \\ dist(p_i, q_{j-1}) \\ dist(p_{i-1}, q_{j-1}) \end{cases} \tag{2.2}$$

$$DTW(P, Q) = \sqrt{dist(p_i, q_j)} \tag{2.3}$$

นอกจากนี้ยังสามารถหาวิถี (Path) ของการวอร์ประหว่างคู่อนุกรมเวลาที่ทำให้ได้ค่าความคล้ายที่น้อยที่สุดได้ โดยวิถี

$$W = \langle w_1, w_2, \ldots, w_k, \ldots, w_K \rangle \qquad \text{เมื่อ } \max(n, m) \le K \le n + m - 1$$

สามารถหาได้หลังจากทำการคำนวณระยะไดนามิกไทม์วอร์ปปิงแล้ว ดังแสดงในภาพที่ 2.5

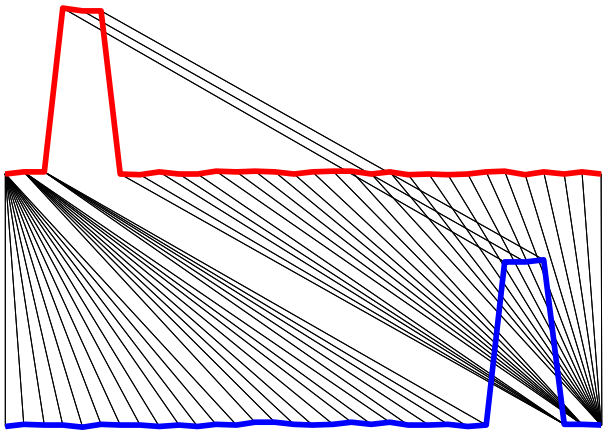


ภาพที่ 2.5 วิถี (Path) ที่ได้ภายหลังจากการคำนวณระยะไดนามิกไทม์วอร์ปปิง

ในงานวิจัยนี้ได้นำมาตรวัดระยะไดนามิกไทม์วอร์ปปิงมาใช้เป็นมาตรวัดความคล้ายของข้อมูลอนุกรมเวลาแทนมาตรวัดระยะยุคลิดในการกำหนดกลุ่มของข้อมูลอนุกรมเวลาแต่ละข้อมูล เพื่อให้การกำหนดกลุ่มให้ข้อมูลอนุกรมเวลานั้นมีความแม่นยำมากขึ้น

2.1.4 **การกำหนดเงื่อนไขบังคับโดยรวม** (Global constraint)

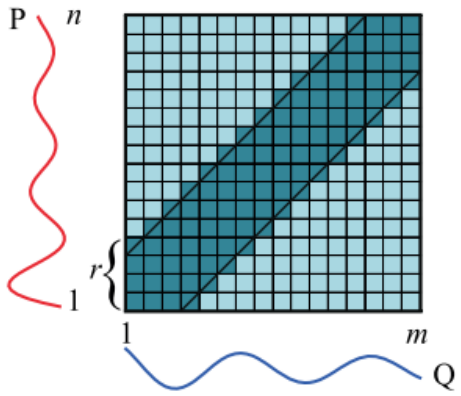
การกำหนดเงื่อนไขบังคับโดยรวมจะนำมาใช้เมื่อต้องการจำกัดขอบเขตของการคำนวณระยะไดนามิกไทม์วอร์ปปิง เนื่องจากวิธีการคำนวณระยะไดนามิกไทม์วอร์ปปิงอาจมีการปรับแนวที่ไม่เหมาะสมเกิดขึ้นได้ [17] (ดังภาพที่ 2.6) เพราะข้อมูลอนุกรมเวลาบางข้อมูลจะเป็นข้อมูลในคลาสเดียวกันเมื่อมีการเลื่อนในแนวแกนเวลาเกิดขึ้นเพียงเล็กน้อยเท่านั้น ดังนั้นการกำหนดเงื่อนไขบังคับโดยรวม จะช่วยให้ข้อมูลอนุกรมเวลาที่มีรูปร่างคล้ายกันแต่มีการเลื่อนในแนวแกนเวลามาก ซึ่งอาจเป็นข้อมูลคนละคลาสกัน ไม่เกิดการปรับแนวที่ไม่เหมาะสมซึ่งจะทำให้ข้อมูลดังกล่าวถูกพิจารณาว่าอยู่ในคลาสเดียวกัน



ภาพที่ 2.6 การคำนวณระยะไดนามิกไทม์วอร์ปปิงของข้อมูลอนุกรมเวลาซึ่งอาจอยู่คนละคลาสกัน แต่สามารถเกิดการปรับแนวที่นำไปสู่การกำหนดคลาสของข้อมูลที่ผิดพลาด
(ที่มา : Keogh และ Ratanamahatana [20])

การกำหนดเงื่อนไขบังคับโดยรวมนี้มีการนำไปใช้ในงานด้านต่าง ๆ เช่น การรู้จำเสียงพูด [21] เพื่อช่วยให้การจำแนกข้อมูลอนุกรมเวลานั้นมีความแม่นยำมากยิ่งขึ้น สำหรับในงานวิจัยนี้ได้ใช้การกำหนดเงื่อนไขบังคับโดยรวมแบบซาโก-ชิบะ [21] ซึ่งได้มีการนำเสนอขึ้นเพื่อใช้ในงานทางด้านเสียงพูด และมีการนำไปประยุกต์ใช้ในงานด้านอื่น ๆ อีกมากมาย [17]

ขอบเขตในการคำนวณระยะไดนามิกไทม์วอร์ปปิงของการกำหนดเงื่อนไขบังคับโดยรวมแบบซาโก-ชิบะ จะถูกกำหนดโดยตัวแปร *r* (ดังภาพที่ 2.7) ซึ่งเป็นความกว้างของบริเวณที่อนุญาตให้มีการปรับแนวเพื่อทำการคำนวณระยะไดนามิกไทม์วอร์ปปิงของข้อมูลอนุกรมเวลา และจะมีค่าเท่ากันทั้งตามแนวตั้งและแนวขวางจากเส้นทแยงมุม



ภาพที่ 2.7 ขอบเขตในการคำนวณระยะไดนามิกไทม์วอร์ปปิงของ
ข้อมูลอนุกรมเวลา *P* และ *Q* ซึ่งถูกกำหนดโดยตัวแปร *r*

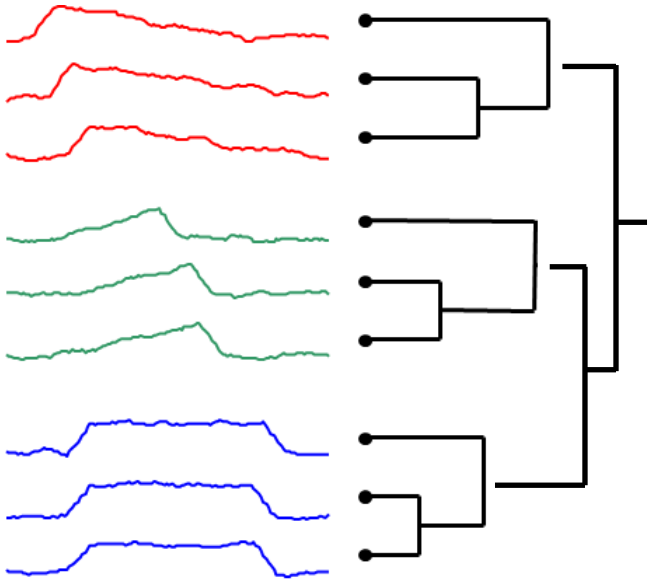
เมื่อนำการกำหนดเงื่อนไขบังคับโดยรวมมาประยุกต์ใช้ร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิง ซึ่งใช้ในการวัดค่าความคล้ายของข้อมูลและใช้เป็นส่วนหนึ่งในการเฉลี่ยแบบรูปร่างเพื่อหาตัวแทนกลุ่มของข้อมูล จะช่วยให้การจัดกลุ่มข้อมูลอนุกรมเวลานั้นมีความแม่นยำมากยิ่งขึ้น เนื่องจากการวัดความคล้ายระหว่างข้อมูลอนุกรมเวลาแต่ละตัวกับข้อมูลอนุกรมเวลาที่เป็นตัวแทนของกลุ่มนั้นมีความถูกต้องมากขึ้น

2.2 เอกสารและงานวิจัยที่เกี่ยวข้อง

ในส่วนของเอกสารและงานวิจัยที่เกี่ยวข้องนี้จะแบ่งออกเป็น 2 ส่วน คือ ส่วนของงานวิจัยที่เกี่ยวกับการจัดกลุ่มข้อมูลอนุกรมเวลา ซึ่งจะกล่าวถึงวิธีการจัดกลุ่มข้อมูลอนุกรมเวลาแบบต่าง ๆ ที่มีการนำไปใช้อย่างแพร่หลาย รวมไปถึงข้อจำกัดของแต่ละวิธีการจัดกลุ่มข้อมูลที่ส่งผลต่อการนำมาใช้ในการจัดกลุ่มข้อมูลอนุกรมเวลา และอีกส่วนจะกล่าวถึงงานวิจัยที่เกี่ยวข้องกับการเฉลี่ยแบบรูปร่างของข้อมูลอนุกรมเวลาซึ่งมีการนำระยะไดนามิกไทม์วอร์ปปิงมาประยุกต์ใช้ เพื่อให้ข้อมูลอนุกรมเวลาที่ได้จากการเฉลี่ย สามารถรักษาลักษณะรูปร่างของข้อมูลอนุกรมเวลาที่นำมาเฉลี่ยไว้ได้

2.2.1 งานวิจัยที่เกี่ยวข้องกับการจัดกลุ่มข้อมูลอนุกรมเวลา

ในช่วงหลายปีที่ผ่านมา ได้มีการนำเสนอวิธีการสำหรับการจัดกลุ่มข้อมูลอนุกรมเวลาไว้เป็นจำนวนมาก [22] วิธีหนึ่งที่มีการนำมาใช้ทั่วไป คือ วิธีการจัดกลุ่มตามลำดับขั้น (Hierarchical clustering) [5] ซึ่งเป็นวิธีที่จะค่อย ๆ รวมข้อมูลอนุกรมเวลาคู่ที่มีความคล้ายกันมากที่สุดทีละคู่ไปเรื่อย ๆ จนกระทั่งข้อมูลอนุกรมเวลาทุกตัวได้รับการจัดกลุ่ม (ดังภาพที่ 2.8)



ภาพที่ 2.8 ตัวอย่างการจัดกลุ่มข้อมูลตามลำดับขั้น

อย่างไรก็ตาม วิธีการจัดกลุ่มแบบขั้นนี้จะต้องมีการคำนวณค่าความคล้ายระหว่างข้อมูลทุกคู่เพื่อหาข้อมูลคู่ที่มีความคล้ายกันมากที่สุดในแต่ละรอบของการรวมกลุ่ม จึงทำให้ต้องใช้เวลาในการประมวลผลมาก โดยเฉพาะเมื่อข้อมูลที่ถูกจัดกลุ่มมีขนาดใหญ่ นอกจากนี้ยังอาจได้ผลการจัดกลุ่มที่คลาดเคลื่อน หากชุดข้อมูลที่ถูกจัดกลุ่มมีข้อมูลที่เป็นตัวแปลกแยก (Outlier)

มาตรวัดความคล้ายที่เหมาะสำหรับข้อมูลอนุกรมเวลาคือ มาตรวัดระยะไดนามิกไทม์วอร์ปปิง แต่เนื่องจากการคำนวณระยะไดนามิกไทม์วอร์ปปิงจะต้องใช้เวลาในการคำนวณสูง ดังนั้นเมื่อนำมาตรวัดระยะไดนามิกไทม์วอร์ปปิงมาใช้ร่วมกับการจัดกลุ่มแบบขั้นซึ่งจะต้องมีการคำนวณความคล้ายของข้อมูลทุกคู่ในทุกรอบของการจัดกลุ่ม ก็จะทำให้สิ้นเปลืองเวลาในการประมวลผลมากขึ้น

วิธีการจัดกลุ่มข้อมูลที่เป็นที่นิยมอีกวิธีหนึ่ง คือ การจัดกลุ่มแบบแบ่งส่วน (Partitional clustering) ซึ่งมีความแตกต่างกับการจัดกลุ่มแบบลำดับขั้น โดยจะต้องมีการกำหนดจำนวนกลุ่มเริ่มต้นให้กับข้อมูลที่ต้องการจัดกลุ่ม การจัดกลุ่มแบบแบ่งส่วนที่เป็นที่รู้จัก ได้แก่ วิธีการจัดกลุ่มแบบเคมีดอยส์ (*K*-medoids clustering) และวิธีการจัดกลุ่มแบบเคมีนส์ (*K*-means Clustering) [12] โดยความแตกต่างระหว่างการจัดกลุ่มสองแบบนี้ คือ วิธีการในการหาตัวแทนของแต่ละกลุ่มข้อมูลตัวใหม่ [22]

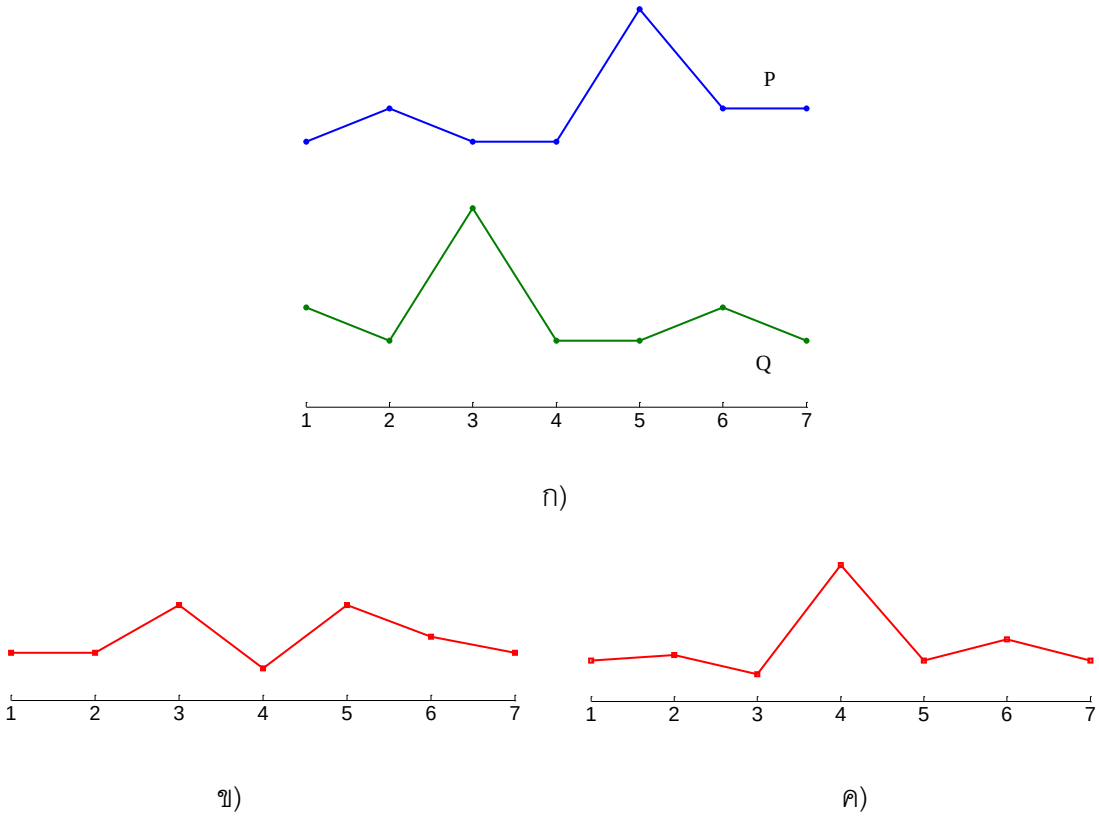
ในบางงานวิจัย [7] ได้มีการประยุกต์ใช้การจัดกลุ่มข้อมูลอนุกรมเวลาแบบเคมีดอยส์ โดยนำมาตรวัดระยะไดนามิกไทม์วอร์ปปิงมาใช้เป็นมาตรวัดความคล้ายระหว่างข้อมูล และใช้ข้อมูลที่มีผลรวมของค่าความคล้ายระหว่างข้อมูลนั้นไปยังข้อมูลที่เหลือในกลุ่มน้อยที่สุดเป็นตัวแทนของกลุ่มข้อมูลตัวใหม่

อย่างไรก็ตาม การหาตัวแทนกลุ่มข้อมูลด้วยวิธีนี้ จะทำให้ได้ตัวแทนที่อาจไม่อยู่ในตำแหน่งเซนทรอยด์ (Centroid) ของกลุ่มข้อมูล และทำให้เกิดความคลาดเคลื่อนในการจัดกลุ่มข้อมูลได้

ในขณะที่การจัดกลุ่มแบบเคมีนส์โดยทั่วไปจะนิยมใช้มาตรวัดระยะยุคลิด ร่วมกับการหาตัวแทนกลุ่มด้วยวิธีการเฉลี่ยข้อมูล หรือการเฉลี่ยแบบแอมพลิจูด [23] แต่เนื่องจากลักษณะมาตรวัดระยะยุคลิด รวมถึงวิธีการเฉลี่ยแบบแอมพลิจูดนั้นไม่เหมาะสมกับลักษณะของข้อมูลอนุกรมเวลา ซึ่งส่งผลให้อนุกรมเวลาที่ได้จากการเฉลี่ยด้วยวิธีดังกล่าว จะมีรูปร่างที่เปลี่ยนไปจากอนุกรมเวลาที่นำมาเฉลี่ย ทำให้ได้ตัวแทนกลุ่มข้อมูลตัวใหม่ที่มีลักษณะแตกต่างจากเดิม และส่งผลให้การจัดกลุ่มข้อมูลนั้นเกิดความผิดพลาด

2.2.2 งานวิจัยที่เกี่ยวข้องกับการเฉลี่ยข้อมูลอนุกรมเวลา

การเฉลี่ยข้อมูลโดยทั่วไปสามารถทำได้โดยใช้วิธีการเฉลี่ยทางคณิตศาสตร์ (Arithmetic mean) ซึ่งค่าที่ได้จากการเฉลี่ยจะถือเป็นค่ากลางของข้อมูลที่นำมาเฉลี่ย อย่างไรก็ตาม วิธีการเฉลี่ยทางคณิตศาสตร์หรือการเฉลี่ยแบบแอมพลิจูดนั้น ไม่เหมาะสมกับลักษณะของข้อมูลอนุกรมเวลาที่มักมีการเลื่อนของข้อมูลในแนวแกนเวลา เนื่องจากข้อมูลที่ได้จากการเฉลี่ยจะมีรูปร่างที่เปลี่ยนไป (ดังภาพที่ 2.9 ข) ด้วยเหตุนี้ จึงมีงานวิจัยที่นำเสนอวิธีการเฉลี่ยแบบรูปร่าง (ดังภาพที่ 2.9 ค) สำหรับนำมาใช้ในการเฉลี่ยข้อมูลอนุกรมเวลา เพื่อให้อนุกรมเวลาที่ได้จากการเฉลี่ยสามารถรักษารูปร่างของข้อมูลเอาไว้ได้



ภาพที่ 2.9 ก) ข้อมูลอนุกรมเวลา *P* และ *Q* ข) เมื่อทำการเฉลี่ยแบบแอมพลิจูด

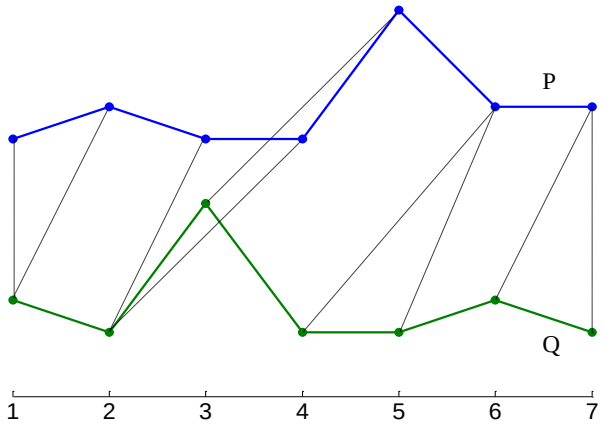
ค) เมื่อทำการเฉลี่ยแบบรูปร่าง

ในช่วงแรก ได้มีผู้นำเสนอวิธีการเฉลี่ยเรียกว่า Non Linear Alignment and Averaging Filter (NLAAF) [24] ซึ่งเป็นเฉลี่ยข้อมูลอนุกรมเวลาที่ใช้ระยะไดนามิกไทม์วอร์ปปิงมาช่วยในการเฉลี่ย โดยจะทำการเฉลี่ยข้อมูลทีละ 2 อนุกรม วิธีการเฉลี่ยจะเริ่มจากการคำนวณระยะไดนามิกไทม์วอร์ปปิงเพื่อหาวิถีระหว่างข้อมูลทั้ง 2 อนุกรมนั้น หลังจากนั้นจึงทำการเฉลี่ยจุดของข้อมูลแต่ละคู่อันดับในวิถีนั้น โดยใช้สมการที่ (2.4)

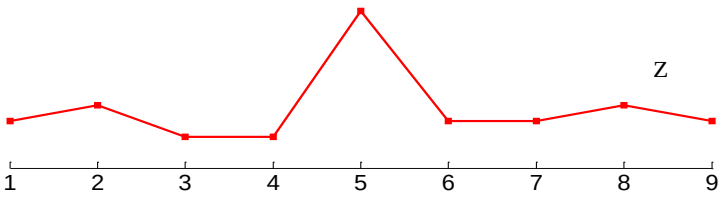
$$z_k = \frac{\omega_P p_i + \omega_Q q_j}{\omega_P + \omega_Q} \tag{2.4}$$

สมมติให้ *P* และ *Q* เป็นอนุกรมเวลาที่ต้องการเฉลี่ย โดยมี p_i และ q_j เป็นข้อมูลแต่ละจุดและมีค่า ω_P และ ω_Q เป็นค่าน้ำหนักของอนุกรมเวลา *P* และ *Q* ตามลำดับ จะสามารถทำการเฉลี่ยอนุกรมเวลาทั้งสองอนุกรมโดยใช้วิธี NLAAF และได้ผลการเฉลี่ยเป็นอนุกรมเวลา *Z*

ซึ่งประกอบด้วย z_k เป็นข้อมูลแต่ละจุด (ดังภาพที่ 2.10) โดยเส้นประระหว่างอนุกรมเวลา *P* และ *Q* แทนวิถีที่ได้ภายหลังจากการคำนวณระยะไดนามิกไทม์วอร์ปปิง



ก)



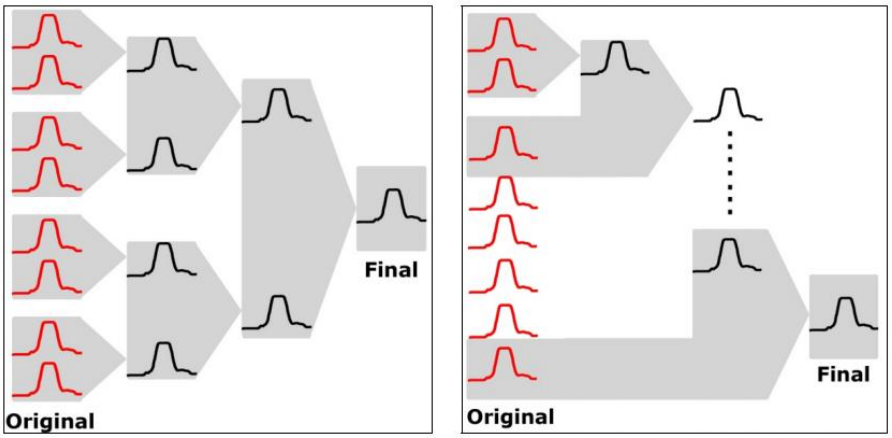
ข)

ภาพที่ 2.10 ก) ข้อมูลอนุกกรมเวลา *P* และ *Q* ที่นำมาเฉลี่ยด้วยวิธี NLAAF
ข) อนุกรมเวลา *Z* เป็นผลลัพธ์ที่ได้จากการเฉลี่ย

ถึงแม้ว่าระยะไดนามิกไทม์วอร์ปปิงจะเป็นมาตรวัดที่เหมาะสมกับข้อมูลอนุกรมเวลามากกว่ามาตรวัดระยะยุคลิด แต่การเฉลี่ยข้อมูลอนุกรมเวลาด้วยระยะไดนามิกไทม์วอร์ปปิงยังไม่สามารถให้ผลลัพธ์ที่รักษาลักษณะของข้อมูลไว้ได้ จากภาพที่ 2.10 ข) จะเห็นว่าอนุกรมเวลา *Z* ซึ่งเป็นผลลัพธ์จากการเฉลี่ยนั้นมีความยาวมากกว่าอนุกรมเวลา *P* และ *Q* ที่นำมาเฉลี่ย ซึ่งหากมีการเฉลี่ยหลายครั้ง ก็จะยิ่งทำให้ผลลัพธ์นั้นมีความแตกต่างจากอนุกรมเวลาที่นำมาเฉลี่ยเพิ่มมากขึ้น

นอกจากนี้ลำดับในการเฉลี่ยข้อมูลของวิธี NLAAF จะเป็นแบบสุ่ม โดยขึ้นอยู่กับจำนวนของข้อมูลที่ต้องการเฉลี่ย ดังแสดงในภาพที่ 2.11 โดยภาพที่ 2.11 ก) เป็นลำดับในการเฉลี่ยข้อมูลด้วยวิธี NLAFF1 ซึ่งจะใช้เมื่อมีจำนวนข้อมูลเป็น 2^n ส่วนภาพที่ 2.11 ข) จะใช้ในกรณี

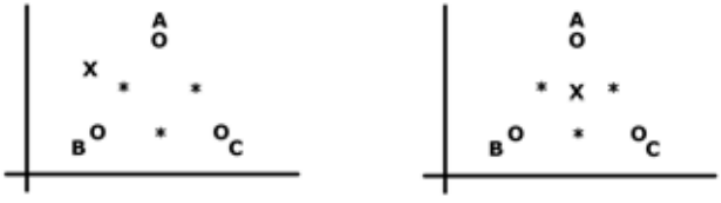
ที่จำนวนข้อมูลไม่เท่ากับ 2^n ซึ่งหากเกิดการเฉลี่ยข้อมูลที่มีลักษณะแตกต่างกันมากก่อน ก็จะส่งผลต่อลักษณะรูปร่างของผลลัพธ์ได้ [14]



ก) ข)

ภาพที่ 2.11 ลำดับในการเฉลี่ยข้อมูลด้วยวิธี ก) NLAAF1 ข) NLAAF2
(ที่มา : Niennattrakul และ Ratanamahatana [14])

อย่างไรก็ตาม วิธีการเฉลี่ยแบบรูปร่างโดยการนำระยะไดนามิกไทม์วอร์ปปิงมาประยุกต์ใช้เพื่อให้เหมาะสมกับลักษณะของข้อมูลอนุกรมเวลา ก็ยังคงทำให้เกิดความคลาดเคลื่อนเมื่อนำไปประยุกต์ใช้ร่วมกับการจัดกลุ่มแบบเคมีนส์ [14] ดังภาพที่ 2.12 ซึ่งแสดงให้เห็นว่าเมื่อทำการเฉลี่ยข้อมูลแล้ว ผลที่ได้จากการเฉลี่ยอาจเกิดความคลาดเคลื่อน ทำให้เมื่อคำนวณระยะ ไดนามิกไทม์วอร์ปปิงระหว่างผลลัพธ์กับข้อมูลที่นำมาเฉลี่ยแต่ละตัวแล้ว พบว่าผลลัพธ์นั้นไม่อยู่ในตำแหน่งที่ควรจะเป็น (ดังภาพที่ 2.12 ก) เมื่อเทียบกับตำแหน่งของข้อมูลเฉลี่ยที่ควรจะอยู่ตรงกลางของกลุ่ม (ดังภาพที่ 2.12 ข)



ก) ข)

ภาพที่ 2.12 ก) *X* เป็นข้อมูลที่ได้จากการเฉลี่ยอนุกรมเวลา *A B* และ *C* มีตำแหน่งที่เลื่อนออกไปอยู่นอกกลุ่ม ข) *X* เป็นข้อมูลที่ได้จากการเฉลี่ยอนุกรมเวลา *A B* และ *C* มีตำแหน่งอยู่ภายในกลุ่ม
(ที่มา : Niennattrakul และ Ratanamahatana [14])

ต่อมาได้มีการนำเสนอวิธีการหาแผ่นแบบ (Template) สำหรับข้อมูลอนุกรมเวลา เรียกว่า Shape-based Template Matching Framework (STMF) [15] ซึ่งเป็นงานวิจัยที่เสนอวิธีการจับคู่แผ่นแบบ (Template matching) โดยการหาตัวแทนกลุ่มของข้อมูลอนุกรมเวลา เพื่อนำมาใช้เป็นแผ่นแบบสำหรับการจำแนก (Classification) ข้อมูลอนุกรมเวลาตัวใหม่ ซึ่งวิธีดังกล่าวจะช่วยลดระยะเวลาในการคำนวณระยะไดนามิกไทม์วอร์ปปิงระหว่างข้อมูลลงได้ และช่วยให้การจำแนกข้อมูลทำได้รวดเร็วมากขึ้น

การหาแผ่นแบบของข้อมูลอนุกรมเวลาด้วยวิธี STMF นั้นเป็นการปรับปรุงวิธีการเฉลี่ยข้อมูลอนุกรมเวลาที่ใช้ระยะไดนามิกไทม์วอร์ปปิง โดยมีขั้นตอนในการทำงานดังแสดงในตารางที่ 2.1 ซึ่งจะเฉลี่ยข้อมูลทีละสองอนุกรม โดยเฉลี่ยคู่อนุกรมเวลาที่คล้ายกันมากที่สุดเมื่อใช้มาตรวัดระยะไดนามิกไทม์วอร์ปปิง

ตารางที่ 2.1 รหัสเทียมแสดงขั้นตอนการทำงานของ STMF

Algorithm STMF(D)	
1.	D is the set of time series data to be averaged
2.	initialize weight $\omega = 1$ for every sequences in D
3.	while(size(D) > 1)
4.	$\{P, Q\}$ = the most similar pair of sequences in D
5.	Z = CDTW(P, Q, ω_P, ω_Q)
6.	$\omega_Z = \omega_P + \omega_Q$
7.	add Z to D
8.	remove P, Q from D
9.	end while
10.	return Z

จากตารางที่ 2.1 เมื่อได้อนุกรมเวลา P และ Q ซึ่งเป็นคู่ที่มีความคล้ายกันมากที่สุด จากนั้นจึงทำการเฉลี่ย P และ Q ด้วยวิธี Cubic-Spline Dynamic Time Warping (CDTW) โดยเริ่มจากการหาวิถี W ซึ่ง

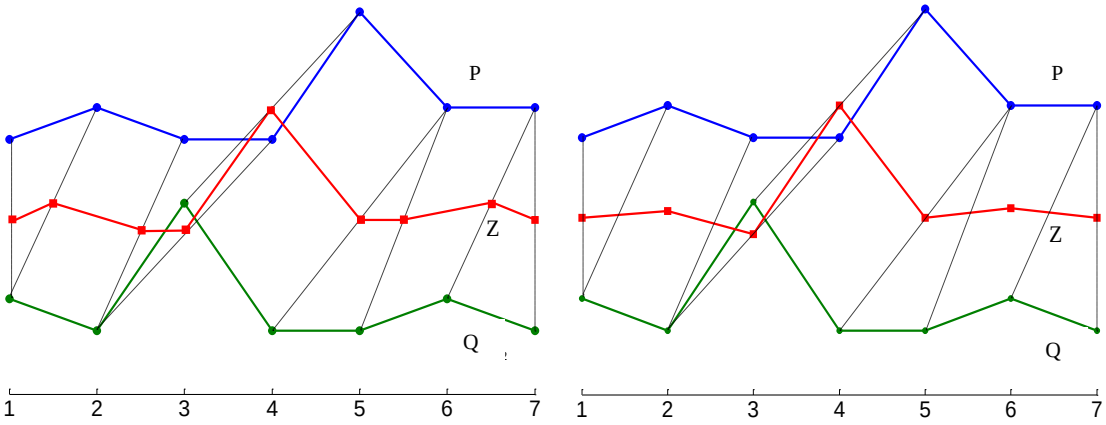
$$W = \langle w_1, w_2, \ldots, w_k, \ldots, w_K \rangle$$

โดยแต่ละ w_k จะประกอบด้วยคู่อันดับ $(i_k,\ j_k)$ ซึ่งเก็บค่าตำแหน่งของข้อมูลที่มีการวอร์ปกันในอนุกรมเวลา P และ Q ที่ต้องการนำมาเฉลี่ยตามลำดับ จากนั้นทำการเฉลี่ยอนุกรมเวลาทีละจุดตามคู่อันดับในวิถี โดยเฉลี่ยทั้งในแนวแกน x ซึ่งเป็นตำแหน่งของข้อมูล และแนวแกน y ซึ่งเป็นค่าของข้อมูล โดยใช้สมการที่ (2.5) และ (2.6) ตามลำดับ

$$z_k(x) = \frac{\omega_P i_k + \omega_Q j_k}{\omega_P + \omega_Q} \tag{2.5}$$

$$z_k(y) = \frac{\omega_P p_{i_k} + \omega_Q q_{j_k}}{\omega_P + \omega_Q} \tag{2.6}$$

ค่า ω_P และ ω_Q เป็นค่าน้ำหนักของอนุกรมเวลา P และ Q ตามลำดับ เมื่อทำการเฉลี่ยเสร็จแล้ว อนุกรมเวลาที่ได้จะมีจำนวนข้อมูล 9 จุด (ดังภาพที่ 2.13 ก) ซึ่งมากกว่าอนุกรมที่นำมาเฉลี่ยซึ่งมี 7 จุด จึงต้องทำการปรับขนาดของอนุกรมเวลาแบบคิวบิคสไปลน์ (Cubic-Spline interpolation) ให้จำนวนจุดของข้อมูลลดลงเหลือ 7 จุด (ดังภาพที่ 2.13 ข)



ก) ข)

ภาพที่ 2.13 อนุกรมเวลาที่ได้จากการเฉลี่ย ก) ก่อนการปรับความยาว และ ข) หลังการปรับความยาวแบบคิวบิคสไปลน์ (Cubic-Spline interpolation)

อย่างไรก็ตาม การคำนวณเพื่อหาแผ่นแบบของข้อมูลอนุกรมเวลาด้วยวิธี STMF นั้นจะใช้เวลาในการคำนวณมาก เนื่องจากต้องหาระยะไดนามิกไทม์วอร์ปปิงระหว่างข้อมูลอนุกรมเวลาทุกคู่ เพื่อทำการเฉลี่ยข้อมูลอนุกรมคู่ที่มีความคล้ายกันมากที่สุดทีละคู่ไปเรื่อย ๆ จนกระทั่งเหลือตัวแทนของข้อมูลอนุกรมเวลาเพียงหนึ่งอนุกรม ซึ่งเมื่อนำวิธีการหาแผ่นแบบดังกล่าวมา

ประยุกต์ใช้เป็นการหาตัวแทนกลุ่มของข้อมูล (Cluster center) ร่วมกับการจัดกลุ่มแบบเคมีนส์ และมาตรวัดยะระไดนามิกไทม์วอร์ปปิง ก็จะทำให้เวลาที่ต้องใช้ในการคำนวณที่เพิ่มสูงมากขึ้นตามไปด้วย

เพราะฉะนั้น งานวิจัยนี้จึงเสนอวิธีการจัดกลุ่มข้อมูลอนุกรมเวลาให้มีความแม่นยำมากขึ้น โดยใช้การจัดกลุ่มแบบเคมีนส์ และนำมาตรวัดระยะไดนามิกไทม์วอร์ปปิงมาใช้แทน มาตรวัดระยะยุคลิด และใช้วิธีการเฉลี่ยเพื่อหาตัวแทนกลุ่มของข้อมูลที่สามารถรักษารูปร่างของอนุกรมเวลาได้ เพื่อให้การกำหนดกลุ่มของข้อมูลอนุกรมเวลามีความถูกต้องมากขึ้น

บทที่ 3
การจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา

งานวิจัยนี้ ได้นำเสนอวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา หรือ Shape-based Clustering for Time Series Data (SCTS) ซึ่งเป็นการประยุกต์ใช้การจัดกลุ่มแบบเคมีนส์ ร่วมกับการใช้ระยะไดนามิกไทม์วอร์ปปิงเป็นมาตรวัดความคล้าย นอกจากนี้ยังได้นำเสนอวิธีการหาตัวแทนกลุ่ม (Cluster Center) ซึ่งมีประสิทธิภาพในการทำงานที่รวดเร็วขึ้น เรียกว่า Ranked Shape-based Template Matching Framework (RSTMF) โดยทำการปรับปรุงเพิ่มเติมจากวิธี Shape-based Template Matching Framework (STMF) ซึ่งเป็นการหาแผ่นแบบของข้อมูลอนุกรมเวลา อย่างไรก็ตาม เนื่องจากการหาแผ่นแบบโดยใช้วิธี STMF นั้น จะต้องทำการคำนวณระยะไดนามิกไทม์วอร์ปปิงระหว่างข้อมูลทุกคู่ เพื่อหาข้อมูลคู่ที่มีความคล้ายกันมากที่สุดก่อนจะนำมาเฉลี่ย จึงส่งผลให้ใช้เวลาในการประมวลผลสูง เพราะฉะนั้นวิธีการ RSTMF จึงจะทำการประมาณลำดับของข้อมูลก่อนจะทำการเฉลี่ยข้อมูลอนุกรมเวลาเพื่อหาตัวแทนกลุ่ม เพื่อลดเวลาในการคำนวณระยะไดนามิกไทม์วอร์ปปิง โดยขณะเดียวกันก็จะยังคงรักษารูปร่างของข้อมูลเอาไว้ได้

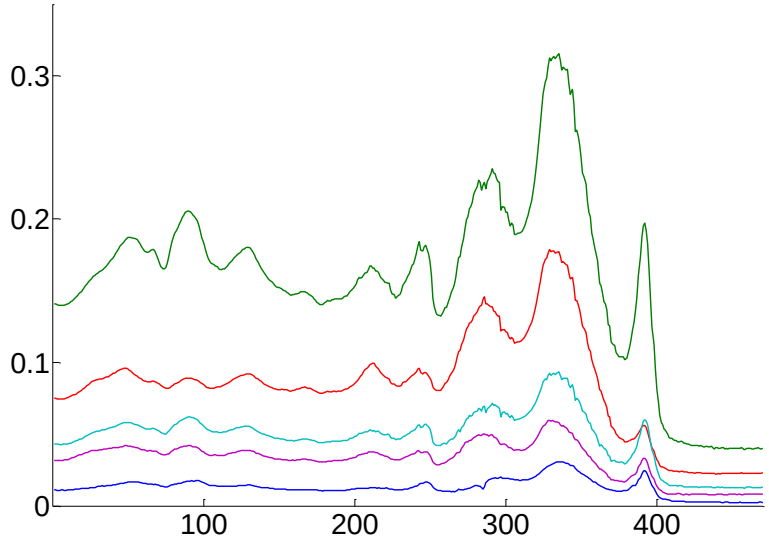
ในส่วนของบทที่ 3 จะเป็นการนำเสนอขั้นตอนวิธีในการจัดกลุ่มข้อมูลอนุกรมเวลา โดยเริ่มจากการแปลงข้อมูลอนุกรมเวลาให้เป็นบรรทัดฐาน (Normalization) โดยจะเป็นการปรับข้อมูลอนุกรมเวลาให้มีมาตราส่วนที่เท่ากัน จากนั้นจะเป็นการอธิบายถึงแนวคิดของวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา หรือ Shape-based Clustering for Time Series Data (SCTS) และขั้นตอนของการหาตัวแทนของกลุ่มของข้อมูลอนุกรมเวลาด้วยวิธี Ranked Shape-based Template Matching Framework (RSTMF) โดยจะอธิบายถึงแนวคิดของวิธีการประมาณลำดับของข้อมูลอนุกรมเวลาที่จะนำมาเฉลี่ย และการปรับค่าประมาณความคล้ายของข้อมูลอนุกรมเวลาที่ได้จากการเฉลี่ย ซึ่งเป็นขั้นตอนสำคัญที่จะช่วยทำให้สามารถลดเวลาในการหาตัวแทนกลุ่มลงได้

นอกจากนี้ ยังมีการอธิบายถึงแนวคิดในการนำการกำหนดเงื่อนไขบังคับโดยรวมมาประยุกต์ใช้ร่วมกับวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลาเพื่อช่วยเพิ่มความแม่นยำให้กับการจัดกลุ่มข้อมูลอนุกรมเวลาอีกด้วย

3.1 การแปลงข้อมูลอนุกรมเวลาให้เป็นบรรทัดฐาน (Normalization)

สำหรับข้อมูลอนุกรมเวลาซึ่งเป็นข้อมูลที่มีการเก็บบันทึกในช่วงเวลาต่าง ๆ กัน จึงส่งผลทำให้ข้อมูลที่มีลักษณะรูปร่างคล้ายคลึงกันแต่มีการเก็บบันทึกในช่วงเวลาที่ไม่ตรงกันสามารถเกิดการเลื่อนของข้อมูลในแนวแกนเวลาขึ้นได้ โดยลักษณะเช่นนี้ของข้อมูลอนุกรมเวลาจะส่งผลต่อการใช้มาตรวัดระยะยุคลิดในการวัดค่าความคล้ายของข้อมูล ด้วยเหตุนี้ จึงมีการนำมาตรวัดระยะ ไดนามิกไทม์วอร์ปปิงซึ่งเป็นมาตรวัดที่สามารถปรับการวางแนวให้เหมาะสม จึงสามารถรองรับลักษณะของข้อมูลอนุกรมเวลาที่มีการเลื่อนในแนวแกนเวลาได้ดีกว่ามาตรวัดระยะยุคลิด

อย่างไรก็ตาม ถึงแม้ว่ามาตรวัดระยะไดนามิกไทม์วอร์ปปิงจะมีความเหมาะสมกับลักษณะของข้อมูลอนุกรมเวลา แต่สำหรับในบางกรณี ข้อมูลอนุกรมเวลาที่มีรูปร่างคล้ายกันอาจมีมาตราส่วนที่แตกต่างกันดังแสดงในภาพที่ 3.1 โดยลักษณะเช่นนี้ย่อมส่งผลให้เกิดความคลาดเคลื่อนขึ้น เมื่อทำการวัดความคล้ายของข้อมูลที่มีลักษณะดังกล่าวโดยใช้มาตรวัดระยะไดนามิก ไทม์วอร์ปปิง



ภาพที่ 3.1 ลักษณะของข้อมูลอนุกรมเวลาซึ่งมีมาตราส่วนที่แตกต่างกัน

จากภาพที่ 3.1 ข้อมูลอนุกรมเวลาที่มีรูปร่างคล้ายกันแต่มีมาตราส่วนที่แตกต่างกันอาจมีผลต่อค่าความคล้ายที่ได้จากมาตรวัดระยะไดนามิกไทม์วอร์ปปิง ดังนั้นการเปรียบเทียบข้อมูลจึงควรมีการปรับลักษณะของข้อมูลให้เป็นบรรทัดฐานเดียวกัน ซึ่งโดยทั่วไปการปรับข้อมูลให้เป็นบรรทัดฐานเดียวกันนั้น สามารถทำได้โดยใช้วิธีการให้คะแนน Z (Z-score normalization)

สมมติ P เป็นอนุกรมเวลาที่ต้องการนำมาปรับให้เป็นบรรทัดฐานโดยใช้วิธีการให้คะแนน Z โดยมี p_i เป็นข้อมูลแต่ละจุดของ P วิธีการให้คะแนน Z เพื่อทำการปรับอนุกรมเวลา P สามารถทำได้โดยใช้สมการ (3.1)

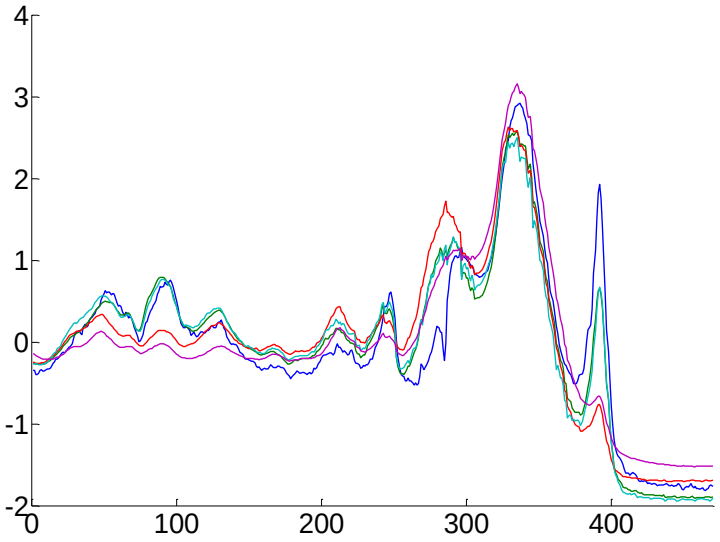
$$Pz_i = \frac{p_i - \bar{p}}{SD} \tag{3.1}$$

$$\bar{p} = \frac{\sum_{i=1}^{n} p_i}{n} \tag{3.2}$$

$$SD = \sum_{i=1}^{n} \sqrt{\frac{(p_i - \bar{p})^2}{n}} \tag{3.3}$$

จากสมการที่ (3.1) $\bar{p}$ คือค่าเฉลี่ย (Mean) ของข้อมูลทุกจุดในอนุกรมเวลา P และ SD คือส่วนเบี่ยงเบนมาตรฐาน (Standard deviation) ของข้อมูลทุกจุดใน P ซึ่งสามารถคำนวณได้โดยใช้สมการที่ (3.2) และ (3.3) ตามลำดับ

หลังจากที่ทำการปรับข้อมูลอนุกรมเวลาให้เป็นบรรทัดฐานเดียวกันแล้ว จะได้ลักษณะของข้อมูลอนุกรมเวลาเป็นดังภาพที่ 3.2



ภาพที่ 3.2 ข้อมูลอนุกรมเวลาหลังจากการได้รับทำให้เป็นบรรทัดฐานเดียวกัน

3.2 การจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา (Shape-based Clustering for Time Series Data (SCTS))

สำหรับข้อมูลอนุกรมเวลาซึ่งมีการบันทึกข้อมูลในแต่ละช่วงเวลานั้น ข้อมูลที่มีรูปร่างคล้ายคลึงกัน อาจถูกเก็บบันทึกในช่วงเวลาที่แตกต่างกันได้ เพราะฉะนั้นวิธีการจัดกลุ่มข้อมูลอนุกรมเวลา จึงควรต้องคำนึงถึงลักษณะดังกล่าวของข้อมูลอนุกรมเวลาด้วย

ในวิทยานิพนธ์นี้ได้เสนอวิธีการจัดกลุ่มสำหรับข้อมูลอนุกรมเวลา เรียกว่า การจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา (Shape-based Clustering for Time Series Data (SCTS)) โดยนำวิธีการจัดกลุ่มแบบเคมีนส์มาประยุกต์ใช้ ซึ่งสามารถแบ่งขั้นตอนการจัดกลุ่มออกเป็น 4 ขั้นตอนหลัก ๆ ได้ดังนี้

1. ทำการสุ่มข้อมูลขึ้นมาจำนวน *K* ตัว เพื่อใช้เป็นตัวแทนกลุ่ม
2. กำหนดกลุ่มให้กับข้อมูลทุกตัว ด้วยการวัดความคล้ายระหว่างข้อมูลแต่ละตัวกับข้อมูลตัวแทนกลุ่มโดยใช้มาตรวัดระยะไดนามิกไทม์วอร์ปปิง
3. หาตัวแทนของกลุ่มข้อมูลตัวใหม่โดยวิธี Ranked Shape-based Template Matching Framework (RSTMF)

เมื่อได้ตัวแทนกลุ่มตัวใหม่ครบทุกกลุ่มแล้ว ก็จะทำการจัดกลุ่มให้กับอนุกรมเวลาทั้งหมดอีกครั้ง และจะทำไปเรื่อย ๆ จนกระทั่งสมาชิกในกลุ่มทุกกลุ่มไม่มีการเปลี่ยนแปลง โดยมีรหัสเทียมแสดงการทำงานของการจัดกลุ่มดังตารางที่ 3.1

ตารางที่ 3.1 รหัสเทียมแสดงขั้นตอนการทำงานของอัลกอริทึม SCTS

Algorithm SCTS(D, K)	
1.	D is the set of time series data
2.	K is the number of cluster in C
3.	C is the set of cluster centers
4.	M is the set of data in each cluster
5.	*Dist* is the matrix of the distance between data sequences and all cluster centers
6.	Initialize C as cluster centers of K clusters
7.	do
8.	for i = 1:size(D)
9.	for k = 1:K
10.	$Dist_{D_i, C_k}$ = DTW(D_i, C_k)
11.	end for
12.	If($Dist_{D_i, C_k}$ is minimal)
13.	assign D_i into M_k
14.	end if
15.	end for
16.	for k = 1:K
17.	C_k = RSTMF(M_k, *Dist*)
18.	end for
19.	while(the cluster membership changes)
20.	return the cluster members and the cluster centers

3.3 การหาตัวแทนกลุ่มของข้อมูลอนุกรมเวลาด้วยวิธี Ranked Shape-based Template Matching Framework (RSTMF)

งานวิจัยนี้ ได้นำเสนอวิธีการหาตัวแทนกลุ่มของข้อมูลอนุกรมเวลาที่เรียกว่า Ranked Shape-based Template Matching Framework (RSTMF) ซึ่งทำการปรับปรุงเพิ่มเติมจาก Shape-based Template Matching Framework (STMF) [15] ให้มีการทำงานที่รวดเร็วขึ้น เนื่องจาก STMF ต้องคำนวณระยะไดนามิกไทม์วอร์ปปิงระหว่างอนุกรมเวลาที่ต้องการเฉลี่ยทุกคู่

เพื่อหาข้อมูลอนุกรมเวลาคู่ที่มีความคล้ายกันมากที่สุดมาเฉลี่ย เมื่อนำวิธีการ STMF มาใช้ร่วมกับการจัดกลุ่มแบบเคมีนส์ จะเพิ่มเวลาในการประมวลผลมากยิ่งขึ้น

เพราะฉะนั้น การหาตัวแทนกลุ่มของข้อมูลอนุกรมเวลาโดยใช้วิธี RSTMF จึงมีแนวคิดในการประมาณลำดับการเฉลี่ยข้อมูลอนุกรมเวลา เพื่อลดขั้นตอนในการคำนวณระยะไดนามิกไทม์วอร์ปปิง โดยได้แสดงการทำงานไว้ดังตารางที่ 3.2

ตารางที่ 3.2 รหัสเทียมแสดงการทำงานของอัลกอริทึม RSTMF

Algorithm RSTMF(*M*, *Dist*)	
1.	*M* is the set of data in each cluster
2.	*Dist* is the matrix of the distance between data sequences and all cluster centers
3.	*S* is the matrix of the distance between data sequence in *M*
4.	Initialize weight $\omega = 1$ for every sequences in *M*
5.	for i = 1:size(*M*)
6.	for $j = i + 1$:size(*M*)
7.	$S_{M_i, C_j} = S_{M_j, C_i} = dist_{approx}(Dist_{M_i,...}, Dist_{M_j,...})$
8.	end for
9.	end for
10.	while(size(*M*) > 1) $Dist_{D_i, C_k} = \text{DTW}(D_i, C_k)$
11.	S_{M_i, C_j} = minimum value in *S*
12.	$M_z = \text{CDTW}(M_i, M_j,)$
13.	$\omega_{M_z} = \omega_{M_i} + \omega_{M_j}$
14.	add M_z to *M*
15.	UPDATE(*S*, *i*, *j*, *z*)
16.	remove M_j, M_i from *M*
17.	end while
18.	return M_z

จากตารางที่ 3.2 สามารถสรุปขั้นตอนการทำงานของวิธีการหาตัวแทนกลุ่ม RSTMF ได้ดังนี้

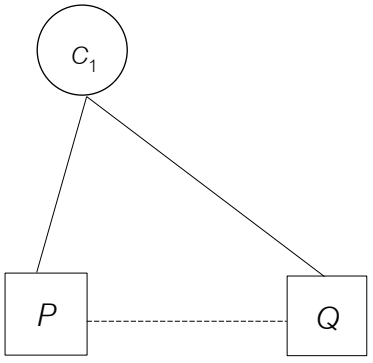
1. คำนวณค่าประมาณความคล้ายของอนุกรมเวลาแต่ละคู่เพื่อใช้ในการประมาณลำดับของคู่อนุกรมเวลาที่ต้องการเฉลี่ย จากนั้นจึงเฉลี่ยอนุกรมคู่ที่มีค่าประมาณความคล้ายน้อยที่สุดทีละคู่โดยใช้วิธี Cubic-spline Dynamic Time Warping (CDTW) จนกระทั่งเหลืออนุกรมสุดท้ายที่เป็นตัวแทนของกลุ่ม

2. ทำการปรับค่าประมาณความคล้ายระหว่างอนุกรมเวลาที่ได้จากการเฉลี่ยกับอนุกรมเวลาที่เหลือ

3.3.1 การประมาณลำดับของข้อมูลอนุกรมเวลาก่อนนำมาเฉลี่ย

จากขั้นตอนการทำงานของวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลาที่ได้นำเสนอไปนั้น ในขั้นตอนของการกำหนดกลุ่มให้กับอนุกรมเวลาแต่ละตัว จะต้องทำการคำนวณระยะไดนามิกไทม์วอร์ปปิงระหว่างอนุกรมเวลาตัวนั้นกับข้อมูลอนุกรมเวลาที่เป็นตัวแทนของกลุ่ม โดยจะสามารถนำค่าความคล้ายดังกล่าวมาช่วยในการประมาณลำดับของข้อมูลอนุกรมเวลาที่ต้องการเฉลี่ย ซึ่งจะช่วยลดการคำนวณระยะไดนามิกไทม์วอร์ปปิงของทุกคู่อนุกรมเวลา และช่วยให้การประมวลผลรวดเร็วขึ้น

ค่าระยะไดนามิกไทม์วอร์ปปิงที่คำนวณได้จะถูกเก็บเอาไว้ในตัวแปร *Dist* ซึ่งเป็นเมทริกซ์ขนาด $K \times n$ โดยที่ K เป็นจำนวนกลุ่มของข้อมูล และ n เป็นจำนวนข้อมูลอนุกรมเวลาทั้งหมด โดย *Dist* คือค่าระยะไดนามิกไทม์วอร์ปปิงที่คำนวณได้จากอนุกรมเวลาแต่ละตัวกับตัวแทนของข้อมูลกลุ่มที่ k ซึ่งทำการคำนวณไว้ในขั้นตอนของการจัดอนุกรมแต่ละตัวให้เข้ากลุ่ม ซึ่งสามารถนำค่าดังกล่าวมาใช้ในการประมาณค่าความคล้ายของอนุกรมเวลา P และ Q ได้ดังแสดงในภาพที่ 3.3



ภาพที่ 3.3 การประมาณค่าความคล้ายของอนุกรมเวลา P และ Q

จากภาพที่ 3.3 เมื่อ P และ Q เป็นอนุกรมเวลาที่เราต้องการคำนวณค่าประมาณความคล้าย หรือ $dist_{approx}$ ในขั้นตอนของการจัดกลุ่มจะมีการคำนวณค่าความคล้ายด้วยมาตรวัดระยะไดนามิกไทม์วอร์ปปิงระหว่าง P และ Q กับข้อมูลตัวแทนกลุ่ม (Cluster center) C_k ซึ่งเราจะทำการเก็บค่าไว้ จากนั้นจึงนำค่าดังกล่าวมาใช้ประมาณความคล้ายหรือ ระหว่าง P และ Q โดยใช้ฟังก์ชัน $dist_{approx}$ ในการประมาณค่าระยะไดนามิกไทม์วอร์ปปิงของแต่ละคู่อนุกรมเวลาดังแสดงในสมการที่ (3.4)

$$dist_{approx}(Dist_{M_P,...}, Dist_{M_Q,...}) = \max_{1 \le k \le K} \left| Dist_{M_P, C_k} - Dist_{M_Q, C_k} \right| \quad (3.4)$$

จากสมการที่ (3.4) สมมติ P และ Q เป็นอนุกรมเวลาของกลุ่ม M ที่ต้องการหาค่า $dist_{approx}$ ซึ่งสามารถคำนวณได้จากค่า $Dist$ เมื่อ

$$Dist_{M_P,...} = \langle Dist_{M_P, C_1}, \dots, Dist_{M_P, C_k}, \dots, Dist_{M_P, C_K} \rangle \text{ และ}$$

$$Dist_{M_Q,...} = \langle Dist_{M_Q, C_1}, \dots, Dist_{M_Q, C_k}, \dots, Dist_{M_Q, C_K} \rangle$$

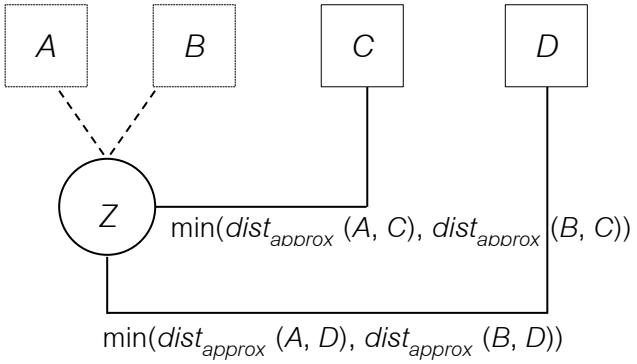
โดยที่ $Dist_{M_P,...}$ และ $Dist_{M_Q,...}$ ประกอบด้วยค่าระยะไดนามิกไทม์วอร์ปปิงระหว่าง P และ Q กับข้อมูลตัวแทนกลุ่ม K ตัว

หลังจากที่คำนวณค่า $dist_{approx}$ ระหว่างอนุกรมเวลาทุกคู่จนครบ จึงทำการเฉลี่ยอนุกรมเวลาแต่ละคู่โดยเลือกคู่ที่มีค่า $dist_{approx}$ น้อยที่สุดมาเฉลี่ยด้วยวิธี CDTW และเมื่อเฉลี่ยเสร็จ จึงทำการปรับค่า $dist_{approx}$ ของแต่ละคู่อนุกรมเวลาใหม่

3.3.2 การปรับค่าประมาณความคล้ายของอนุกรมเวลาที่ได้จากการเฉลี่ยกับอนุกรมเวลาตัวที่เหลือ

เมื่อทำการเฉลี่ยอนุกรมเวลาด้วยวิธี Cubic-spline Dynamic Time Warping (CDTW) เสร็จแล้ว อนุกรมเวลาที่ถูกเฉลี่ยไปแล้วจะถูกลบออกไป และอนุกรมเวลาที่ได้จากการเฉลี่ยจะถูกเพิ่มเข้าไปใหม่ จากนั้นจึงทำการเลือกอนุกรมเวลาคู่ถัดไปที่มีค่าประมาณความคล้ายน้อยที่สุดเพื่อมาเฉลี่ยด้วยวิธี CDTW อีกครั้ง และทำเช่นนี้ไปเรื่อย ๆ จนกระทั่งเหลืออนุกรมเวลาตัวสุดท้ายที่เป็นตัวแทนของกลุ่มตัวใหม่

เพราะฉะนั้นจึงต้องมีการปรับค่าประมาณความคล้ายระหว่างอนุกรมเวลาที่ได้จากการเฉลี่ยที่ทำการเพิ่มเข้าไปใหม่กับอนุกรมเวลาที่เหลืออยู่ซึ่งยังไม่ได้ทำการเฉลี่ย เพื่อนำมาใช้ในการเลือกคู่ของอนุกรมเวลาที่มีค่าประมาณความคล้ายน้อยที่สุดคู่ถัดไปสำหรับนำมาเฉลี่ย ดังแสดงในภาพที่ 3.4



ภาพที่ 3.4 การปรับค่าประมาณความคล้ายของอนุกรมเวลา Z ที่ได้จากการเฉลี่ย

จากภาพที่ 3.4 A B C และ D เป็นอนุกรมเวลาในกลุ่มที่ต้องการหาตัวแทนกลุ่มตัวใหม่ สมมติ A และ B เป็นอนุกรมเวลาที่มี $dist_{approx}$ น้อยที่สุด หลังจากทำการเฉลี่ย A และ B ด้วยวิธี CDTW แล้ว จะได้อนุกรมเวลา Z เป็นผลลัพธ์เพิ่มเข้ามาในกลุ่ม และ A กับ B จะถูกลบออกไปจากกลุ่ม จากนั้นจึงทำการปรับค่า $dist_{approx}$ ระหว่าง Z และอนุกรมเวลาที่เหลือในกลุ่มคือ C และ D โดยจะเลือกค่า $dist_{approx}$ ที่น้อยที่สุดระหว่างอนุกรมเวลา A และ B กับอนุกรมเวลา C และ D เพื่อเป็นค่า $dist_{approx}$ ของ Z และอนุกรมเวลาที่เหลือ โดยสามารถแสดงขั้นตอนวิธีการปรับค่า $dist_{approx}$ ดังแสดงในตารางที่ 3.3

ตารางที่ 3.3 รหัสเทียมแสดงการทำงานของอัลกอริทึม UPDATE

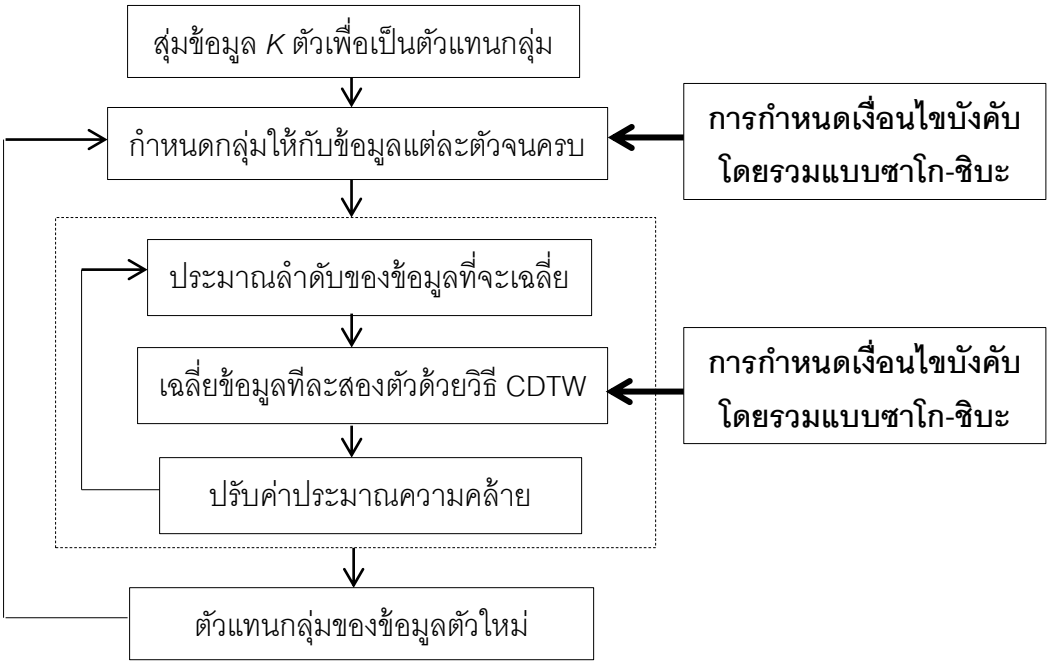
Algorithm RSTMF(S, a, b, z)	
1.	S is the matrix of the distance between data sequences in M
2.	for i = 1:size(S)
3.	$S_{M_i, M_z} = S_{M_z, M_i} = \min(S_{M_a, M_i}, S_{M_b, M_i})$
4.	end for
5.	remove $S_{M_a, \ldots}$, $S_{\ldots, M_a}$, $S_{M_b, \ldots}$, $S_{\ldots, M_b}$ from S

ตารางที่ 3.3 แสดงขั้นตอนการปรับค่า $dist_{approx}$ ของข้อมูลอนุกรมเวลาแต่ละคู่ ซึ่งถูกเก็บไว้ในเมตริกซ์ *S* โดย *a* และ *b* เป็นดัชนีของคู่ของอนุกรมเวลาที่ทำการเฉลี่ย และ *z* เป็นดัชนีของอนุกรมเวลาที่ได้จากการเฉลี่ย หลังจากที่ทำการปรับค่า (บรรทัดที่ 3) ครบทุกค่าแล้ว จึงลบค่า $dist_{approx}$ ระหว่างอนุกรมเวลาที่เฉลี่ยไปแล้วกับอนุกรมเวลาที่เหลืออยู่ออกไป

ด้วยขั้นตอนในการประมาณลำดับของอนุกรมเวลาเพื่อนำมาเฉลี่ย โดยใช้ค่าระยะไดนามิกไทม์วอร์ปปิงที่ได้จากการคำนวณในการกำหนดกลุ่ม ร่วมกับขั้นตอนของการปรับค่าประมาณความคล้ายระหว่างข้อมูลอนุกรมเวลาตัวใหม่ที่ได้จากการเฉลี่ยกับข้อมูลอนุกรมเวลาที่เหลือซึ่งยังไม่ได้ทำการเฉลี่ย สามารถลดเวลาในการหาตัวแทนกลุ่มของข้อมูลอนุกรมเวลา และช่วยให้การจัดกลุ่มข้อมูลอนุกรมเวลามีความรวดเร็วมากขึ้น

3.4 การประยุกต์ใช้การกำหนดเงื่อนไขบังคับโดยรวม (Global constraint) ร่วมกับการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา

การกำหนดเงื่อนไขบังคับโดยรวมแบบซาโก-ชิบะ สามารถนำมาประยุกต์ใช้ร่วมกับวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา เพื่อเพิ่มความแม่นยำในการจัดกลุ่ม โดยนำไปใช้ในขั้นตอนต่าง ๆ ของการจัดกลุ่มดังแสดงในภาพที่ 3.5



ภาพที่ 3.5 การประยุกต์ใช้การกำหนดเงื่อนไขบังคับโดยรวมในขั้นตอนต่าง ๆ ของวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา

จากภาพที่ 3.5 ได้มีการนำการกำหนดเงื่อนไขบังคับโดยรวมมาประยุกต์ใช้ในขั้นตอนของการกำหนดกลุ่มข้อมูลอนุกรมเวลา ซึ่งนำมาตรวัดระยะไดนามิกไทม์วอร์ปปิงมาใช้ในการวัดความคล้ายระหว่างข้อมูล เพื่อเพิ่มความแม่นยำในการกำหนดกลุ่มของข้อมูล นอกจากนี้ยังได้นำมาใช้ร่วมกับการเฉลี่ยข้อมูลอนุกรมเวลาด้วยวิธี Cubic-spline Dynamic Time Warping (CDTW) เพื่อให้เกิดการปรับแนวที่เหมาะสม และได้วิถี (Path) สำหรับนำไปใช้ในการเฉลี่ยเพื่อหาข้อมูลอนุกรมเวลาตัวใหม่

บทที่ 4

การทดลองและวิเคราะห์ผลการทดลอง

สำหรับการทดลองและวิเคราะห์ผลการทดลองนั้น จะประเมินคุณภาพของแนวคิดที่ได้นำเสนอโดยทำการทดลองกับข้อมูลอนุกรมเวลา โดยเริ่มจากการอธิบายลักษณะของชุดข้อมูลอนุกรมเวลาต่าง ๆ ที่มีการนำมาใช้ในการทดลองเพื่อเปรียบเทียบประสิทธิภาพของวิธีการจัดกลุ่มข้อมูลอนุกรมเวลาที่ได้นำเสนอ

การทดลองในส่วนแรก จะทำการเปรียบเทียบความแม่นยำในการจัดกลุ่มข้อมูลอนุกรมเวลา รวมถึงมีการนำเกณฑ์สำหรับกรณีที่ทราบกลุ่มของข้อมูล (Criteria based on known ground truth) [22] มาใช้ในการประเมินผลการจัดกลุ่มข้อมูล ถัดมาได้ทำการประเมินคุณภาพของกลุ่มข้อมูลที่ได้จากการจัดกลุ่มด้วยวิธีต่าง ๆ โดยใช้ดัชนีเงา (Silhouette index) [25] ซึ่งเป็นดัชนีที่มีการนำไปใช้ในการประเมินคุณภาพของกลุ่มข้อมูล นอกจากนี้ยังใช้ค่าความคล้ายรวมของข้อมูลภายในกลุ่ม (Intracluster distance) เพื่อเปรียบเทียบความคล้ายคลึงของตัวแทนกลุ่มและข้อมูลสมาชิกในกลุ่มภายหลังทำการจัดกลุ่ม

การทดลองในส่วนถัดมา จะเปรียบเทียบระหว่างการจัดกลุ่มตามรูปร่างซึ่งใช้วิธีการเฉลี่ยข้อมูลแบบ Shape-based Template Matching Framework (STMF) และวิธีการจัดกลุ่มตามรูปร่างร่วมกับการเฉลี่ยข้อมูลแบบ Ranked Shape-based Template Matching Framework (RSTMF) ซึ่งได้นำเสนอไว้ในบทก่อนหน้า โดยจะทำการเปรียบเทียบเวลาที่ใช้ในการจัดกลุ่มข้อมูลอนุกรมเวลา และผลลัพธ์ที่ได้หลังจากการจัดกลุ่ม

และในการทดลองส่วนสุดท้าย จะแสดงผลที่ได้จากวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลาเมื่อนำการกำหนดเงื่อนไขบังคับโดยรวม (Global constraint) มาประยุกต์ใช้

4.1 ชุดข้อมูลที่ใช้ในการทดลอง

สำหรับในงานวิจัยนี้ ได้ใช้ชุดข้อมูลอนุกรมเวลาสำหรับการจำแนก และการจัดกลุ่มจาก University of California, Riverside (UCR) [16] โดยชุดข้อมูลอนุกรมเวลาเหล่านี้เป็นข้อมูลที่มีการเก็บรวบรวมเพื่อนำมาใช้ในการทดลองด้านต่าง ๆ ที่เกี่ยวข้องกับข้อมูลอนุกรมเวลา

เนื่องจากเป็นข้อมูลจริงที่ได้ทำการเก็บรวบรวมไว้ เพื่อใช้สำหรับงานวิจัยด้านการทำเหมืองข้อมูลอนุกรมเวลา

ข้อมูลแต่ละชุดจะถูกแบ่งออก 2 ส่วน คือ ส่วนที่ใช้สำหรับเป็นข้อมูลฝึก และส่วนที่ใช้เป็นข้อมูลทดสอบ ซึ่งข้อมูลทุกตัวจะมีการกำหนดคลาสของข้อมูลเอาไว้แล้ว โดยจำนวนคลาสนั้นจะแตกต่างกันไปตามชุดของข้อมูล นอกจากนี้ ข้อมูลแต่ละชุดยังมีความยาวที่ไม่เท่ากัน อย่างไรก็ตาม ข้อมูลภายในชุดเดียวกันนั้นจะมีความยาวที่เท่ากัน โดยรายละเอียดต่าง ๆ ของชุดข้อมูลทั้งหมดที่นำมาใช้ในงานวิจัยนี้ได้แสดงไว้ในตารางที่ 4.1

ตารางที่ 4.1 รายละเอียดของชุดข้อมูลที่นำมาใช้ในการทดลอง

ชุดข้อมูล	จำนวนคลาส	ความยาว	จำนวนข้อมูลฝึก	จำนวนข้อมูลทดสอบ
▲ Trace	4	275	100	100
◆ CBF	3	128	30	900
+ Synthetic Control	6	60	300	300
▬ Face Four	4	350	24	88
▪ Fish	7	463	175	175
● ECG	2	96	100	100
◆ Gunpoint	2	150	50	150
✳ Lightning-7	7	319	70	73
■ Lightning-2	2	637	60	61
× Olive Oil	4	570	30	30

จากตารางที่ 4.1 แสดงรายระเอียดของข้อมูลแต่ละชุดที่ใช้ในการทดลองเพื่อประเมินคุณภาพของวิธีการจัดกลุ่มข้อมูลอนุกรมเวลาที่นำเสนอ โดยแต่ละชุดข้อมูลจะใช้สัญลักษณ์กำกับเพื่อความสะดวกในการแสดงผลการเปรียบเทียบ จะเห็นได้ว่าชุดข้อมูลนั้นมีความแตกต่างทั้งด้านความยาว จำนวนคลาสของข้อมูล และจำนวนข้อมูลทั้งหมดในชุดข้อมูลนั้น

นอกจากนี้ลักษณะร่วมของข้อมูลที่อยู่ในคลาสเดียวกันนั้นยังมีความแตกต่างกันในแต่ละชุดข้อมูลอีกด้วย

ความหลากหลายของชุดข้อมูลที่นำมาใช้ในงานวิจัยนี้ จะช่วยแสดงให้เห็นว่าวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลาที่ได้นำเสนอนั้นสามารถนำไปประยุกต์ใช้กับข้อมูลอนุกรมเวลาประเภทต่าง ๆ ได้

4.2 การทดลองเพื่อประเมินผลลัพธ์ที่ได้จากวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา (Shape-based Clustering for Time Series Data (SCTS))

ในการทดลองส่วนที่เป็นการจัดกลุ่มแบบแบ่งส่วน (Partitional clustering) ซึ่งประกอบด้วย การจัดกลุ่มแบบเคมีนส์ การจัดกลุ่มแบบเคมีดอยส์ รวมถึงวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลาที่ได้นำเสนอนั้น จะต้องมีการกำหนดค่าเริ่มต้นของจำนวนกลุ่มที่ต้องการทำการจัดกลุ่มข้อมูล หรือค่า K ซึ่งในการทดลองได้กำหนดค่า K เป็นจำนวนเท่ากับจำนวนคลาสของแต่ละชุดข้อมูล และนำข้อมูลฝึกรวมกับข้อมูลทดสอบเพื่อใช้เป็นข้อมูลสำหรับการจัดกลุ่มด้วยวิธีต่าง ๆ

นอกจากนี้ ผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลแบบแบ่งส่วนในแต่ละรอบจะมีความแตกต่างกัน เนื่องจากข้อมูลตัวแทนกลุ่มเริ่มต้นในแต่ละครั้งของการจัดกลุ่มนั้นได้จากการสุ่ม เพราะฉะนั้น จึงได้ทำการทดลองจัดกลุ่มทั้งหมด 40 ครั้ง และนำค่าที่ทำการวัดในแต่ละครั้งทั้งหมดมาหาค่าเฉลี่ย

สำหรับวิธีการจัดกลุ่มแบบลำดับขั้น (Hierarchical clustering) นั้น ผลลัพธ์ที่ได้จากการจัดกลุ่มจะคงเดิมไม่มีการเปลี่ยนแปลง และยังสามารถทำการกำหนดจำนวนกลุ่มให้กับข้อมูลที่ต้องการจัดกลุ่มได้ภายหลังจากการจัดกลุ่มเสร็จสิ้น ซึ่งสำหรับในการทดลองนี้ จำนวนกลุ่มของข้อมูลก็จะถูกกำหนดให้เท่ากับจำนวนคลาสของแต่ละชุดข้อมูลเช่นเดียวกับการจัดกลุ่มแบบแบ่งส่วน

4.2.1 การเปรียบเทียบโดยการวัดความแม่นยำในการจัดกลุ่มข้อมูลอนุกรมเวลา

สำหรับการจัดกลุ่มข้อมูลโดยทั่วไป ข้อมูลที่นำมาใช้ในการจัดกลุ่มจะยังไม่มีการกำหนดคลาสของข้อมูลไว้ล่วงหน้า ซึ่งทำให้วิธีการวัดผลโดยทั่วไปจะสามารถทำได้โดยใช้การเปรียบเทียบลักษณะของข้อมูลภายในกลุ่มเดียวกัน และข้อมูลต่างกลุ่มกันภายหลังจากที่ทำการ

จัดกลุ่มข้อมูลเสร็จสิ้น อย่างไรก็ตาม เนื่องจากชุดข้อมูลอนุกรมเวลาที่นำมาใช้ในการทดลองนี้ได้มีการกำหนดคลาสของข้อมูลแต่ละตัวเอาไว้แล้ว จึงทำให้ทราบว่าข้อมูลใดที่ควรจะถูกจัดให้อยู่ในกลุ่มเดียวกัน

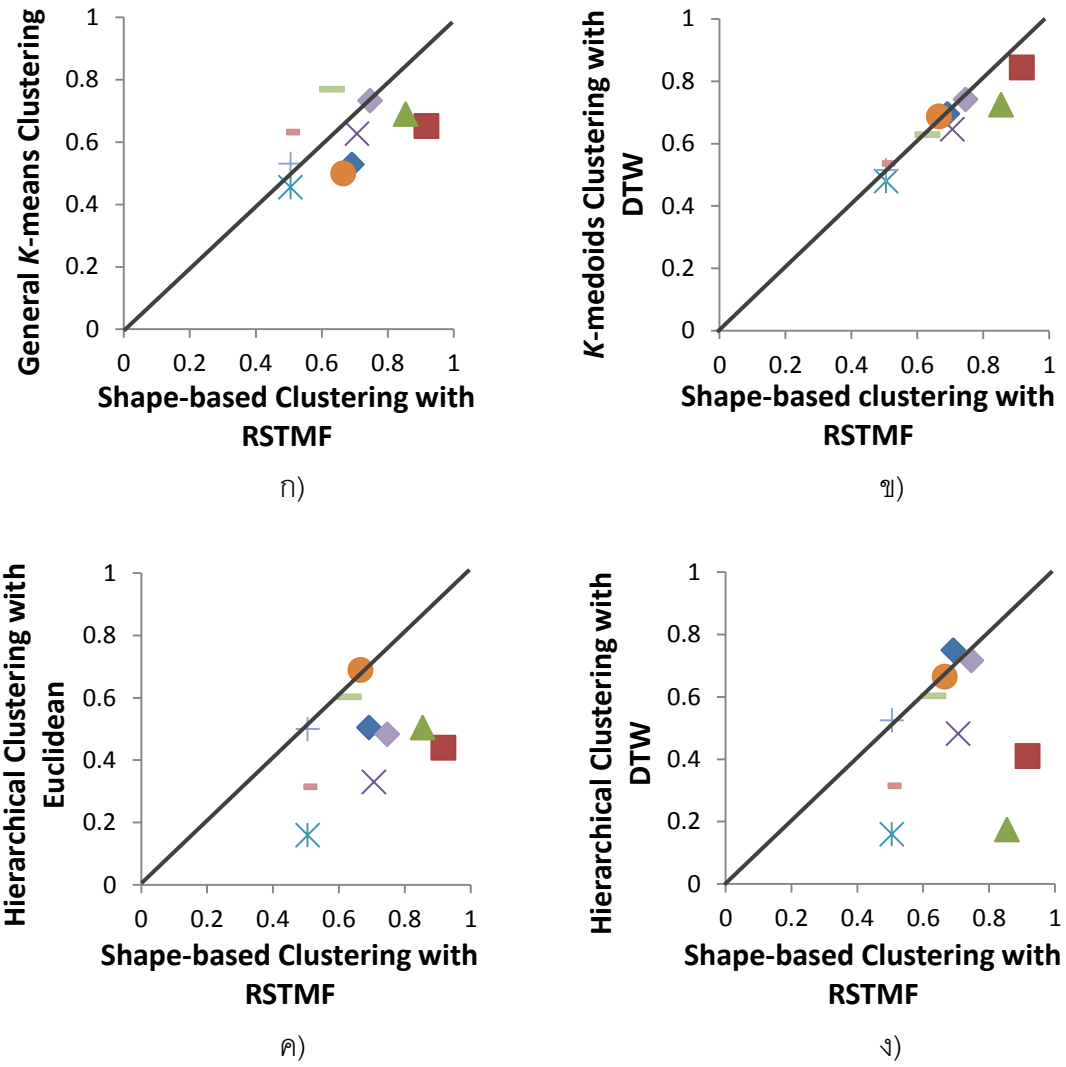
ในส่วนของความแม่นยำของผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลนั้น สามารถคำนวณได้โดยใช้สมการที่ (4.1)

$$\text{accuracy} = \frac{\sum_{k=1}^{K} T_k}{N} \tag{4.1}$$

จากสมการที่ (4.1) T_k คือ จำนวนของข้อมูลในกลุ่ม k ที่ถูกจัดกลุ่มได้อย่างถูกต้อง และ N คือ จำนวนข้อมูลทั้งหมดที่นำมาจัดกลุ่ม เมื่อทำการรวมจำนวนข้อมูลที่จัดกลุ่มได้ถูกต้องจนครบ K กลุ่ม และหารด้วยจำนวนข้อมูลทั้งหมดที่ถูกจัดกลุ่ม ก็จะได้ค่าความแม่นยำ (Accuracy) ของผลลัพธ์ที่ได้จากการจัดกลุ่ม

ซึ่งการทดลองนี้ จะทำการวัดความแม่นยำในการจัดกลุ่มข้อมูลอนุกรมเวลา โดยแสดงผลการทดลอง (ดังภาพที่ 4.1) ซึ่งเป็นการเปรียบเทียบระหว่างวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา กับวิธีการจัดกลุ่มข้อมูลแบบต่าง ๆ ที่นิยมนำมาใช้จัดกลุ่มข้อมูลอนุกรมเวลา ดังนี้

1. การจัดกลุ่มแบบเคมีนส์ร่วมกับมาตรวัดระยะยุคลิด และการเฉลี่ยแบบแอมพลิจูด ดังแสดงในภาพที่ 4.1 ก)
2. การจัดกลุ่มแบบเคมีดอยส์ร่วมกับร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิง ดังแสดงในภาพที่ 4.1 ข)
3. การจัดกลุ่มแบบลำดับขั้นร่วมกับมาตรวัดระยะยุคลิด ดังแสดงในภาพที่ 4.1 ค)
4. การจัดกลุ่มแบบลำดับขั้นร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิง ดังแสดงในภาพที่ 4.1 ง)



ภาพที่ 4.1 เปรียบเทียบความแม่นยำของผลลัพธ์ที่ได้จากวิธีจัดกลุ่มที่นำเสนอกับ

ก) การจัดกลุ่มแบบเคมีนส์ร่วมกับมาตรวัดระยะยุคลิด และการเฉลี่ยแบบแอมพลิจูด

ข) การจัดกลุ่มแบบเคมีดอยส์ร่วมกับร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิง

ค) การจัดกลุ่มแบบลำดับขั้นร่วมกับมาตรวัดระยะยุคลิด และ

ง) การจัดกลุ่มแบบลำดับขั้นร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิง

จากภาพที่ 4.1 เป็นความแม่นยำของผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลอนุกรมเวลาด้วยวิธีการจัดกลุ่มตามรูปร่าง ร่วมกับการหาตัวแทนกลุ่มข้อมูลตัวใหม่ด้วยวิธี Ranked Shape-based Template Matching Framework (RSTMF) โดยเปรียบเทียบกับวิธีการจัดกลุ่มแบบต่าง ๆ

จะเห็นได้ว่าการจัดกลุ่มข้อมูลอนุกรมเวลาด้วยวิธีการจัดกลุ่มตามรูปร่างนั้น ให้ผลลัพธ์ของการจัดกลุ่มที่มีความแม่นยำมากกว่าเมื่อเปรียบเทียบกับการจัดกลุ่มแบบเคมีนส์โดยทั่วไปที่ใช้มาตรวัดระยะยุคลิด และการเฉลี่ยแบบแอมพลิจูด (ดังภาพที่ 4.1 ก) เนื่องจากทั้งมาตรวัดระยะยุคลิดและการเฉลี่ยแบบแอมพลิจูดนั้นไม่เหมาะกับลักษณะของข้อมูลอนุกรมเวลาที่มีการเลื่อนในแนวแกนเวลาเกิดขึ้นบ่อยครั้ง

ในขณะที่เมื่อเปรียบเทียบกับวิธีการจัดกลุ่มแบบลำดับขั้น ผลลัพธ์ที่ได้จากวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลานั้นมีความแม่นยำกว่าอย่างมาก ไม่ว่าจะเป็นกรณีที่ใช้มาตรวัดระยะยุคลิด (ดังภาพที่ 4.1 ค) หรือมาตรวัดระยะไดนามิกไทม์วอร์ปปิง (ดังภาพที่ 4.1 ง) ทั้งนี้เนื่องจากผลลัพธ์ที่ได้จากการจัดกลุ่มแบบลำดับขั้นนั้น หากมีข้อมูลซึ่งมีลักษณะเป็นตัวแปลกแยก (Outlier) จะทำให้ข้อมูลดังกล่าวถูกจัดกลุ่มแยกออกไป ซึ่งทำให้ผลการจัดกลุ่มโดยรวมนั้นเกิดความผิดพลาด

สำหรับวิธีการจัดกลุ่มแบบเคมีดอยส์นั้น (ดังภาพที่ 4.1 ข) เนื่องจากเป็นการจัดกลุ่มแบบแบ่งส่วน และมีการนำมาตรวัดระยะไดนามิกไทม์วอร์ปปิงมาใช้ในการวัดความคล้าย จึงทำให้ผลลัพธ์ที่ได้จากการจัดกลุ่มมีความแม่นยำที่ใกล้เคียงกันในหลายชุดข้อมูล แต่อย่างไรก็ตามวิธีการจัดกลุ่มตามรูปร่างก็ยังคงมีความแม่นยำมากกว่า เนื่องจากวิธีการหาตัวแทนกลุ่มของการจัดกลุ่มแบบ เคมีดอยส์อาจทำให้ได้ตัวแทนกลุ่มที่ไม่ได้อยู่ในตำแหน่งเซนทรอยด์ (Centroid) ซึ่งส่งผลต่อความแม่นยำในการจัดกลุ่ม

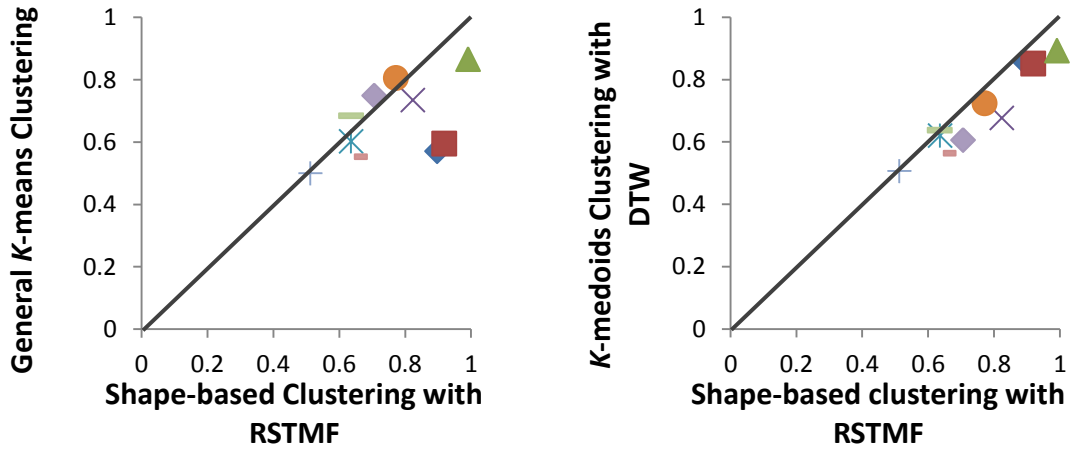
4.2.2 **การเปรียบเทียบโดยใช้เกณฑ์สำหรับกรณีที่ทราบกลุ่มของข้อมูล** (Criteria based on known ground truth)

บางกรณี ในแต่ละคลาสของข้อมูลที่ถูกนำมาจัดกลุ่มนั้นอาจจะมีจำนวนของข้อมูลที่แตกต่างกันเป็นจำนวนมาก เพราะฉะนั้น จึงได้ทำการวัดผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลเพิ่มเติม โดยการใช้เกณฑ์สำหรับกรณีที่ทราบกลุ่มของข้อมูล (Criteria based on known ground truth) [22] ซึ่งสามารถคำนวณได้โดยใช้สมการที่ (4.2) และ (4.3)

$$Sim(G, C)= \frac{1}{K}\sum_{i=1}^{K} \max_{1\le j\le K} sim(G_i, C_j) \tag{4.2}$$

$$sim(G_i, C_j)= \frac{2\left|G_i \cap C_j\right|}{\left|G_i\right|+\left|C_j\right|} \tag{4.3}$$

จากสมการที่ (4.2) และ (4.3) G และ C เป็นเซตของข้อมูลที่ทราบกลุ่มจำนวน K กลุ่ม และผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลออกเป็น K กลุ่มด้วยวิธีต่าง ๆ ตามลำดับ โดยผลการทดลองได้แสดงไว้ดังภาพที่ 4.2



ก) ข)

ภาพที่ 4.2 เปรียบเทียบค่าที่ได้จากเกณฑ์สำหรับกรณีที่ทราบกลุ่มของข้อมูล (Criteria based on known ground truth) ของวิธีจัดกลุ่มที่นำเสนอกับ
ก) การจัดกลุ่มแบบเคมีนส์ร่วมกับมาตรวัดระยะยุคลิด และการเฉลี่ยแบบแอมพลิจูด
ข) การจัดกลุ่มแบบเคมีดอยส์ร่วมกับร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิง

จากภาพที่ 4.2 เป็นการเปรียบเทียบค่าซึ่งทำการคำนวณโดยใช้เกณฑ์สำหรับกรณีที่ทราบกลุ่มของข้อมูล (Criteria based on known ground truth) โดยในการทดลองส่วนนี้จะทำการเปรียบเทียบวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา กับการจัดกลุ่มแบบเคมีนส์ร่วมกับมาตรวัดระยะยุคลิด และการเฉลี่ยแบบแอมพลิจูด (ดังภาพที่ 4.2 ก) และการจัดกลุ่มแบบเคมีดอยส์ร่วมกับร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิง (ดังภาพที่ 4.2 ข) เนื่องจากผลลัพธ์ที่ได้จากการจัดกลุ่มทั้งสองแบบนี้จะมีการเปลี่ยนแปลงเพราะการสุ่มข้อมูลตัวแทนกลุ่มเริ่มต้น

จะเห็นได้ว่าวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลาที่ได้นำเสนอยังคงให้ผลลัพธ์ของการจัดกลุ่มที่ดีกว่าเมื่อทำการวัดด้วยเกณฑ์สำหรับกรณีที่ทราบกลุ่มของข้อมูล ทั้งนี้เนื่องจากการใช้วิธีการหาตัวแทนกลุ่มที่สามารถรักษารูปร่างลักษณะของอนุกรมเวลาเอาไว้ได้ รวมถึงการใช้มาตรวัดความคล้ายที่เหมาะสมกับลักษณะของข้อมูลอนุกรมเวลา

4.2.3 การเปรียบเทียบโดยการวัดค่าดัชนีเงา (Silhouette index) ของผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูล

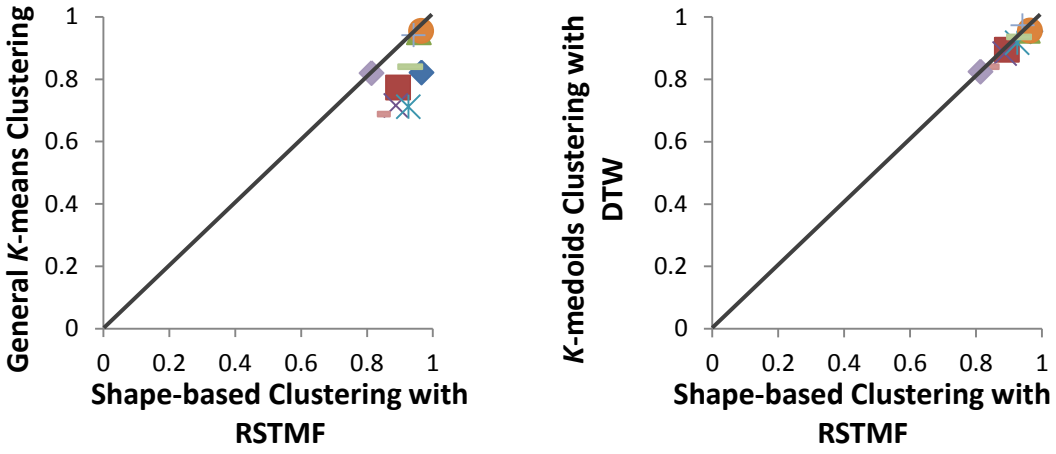
โดยทั่วไปแล้ว การจัดกลุ่มข้อมูลนั้นจะกระทำบนข้อมูลที่ไม่มีการกำหนดคลาสไว้ล่วงหน้า ซึ่งการจัดกลุ่มจะเป็นลักษณะของการจัดข้อมูลที่มีความคล้ายกันเอาไว้ในกลุ่มเดียวกัน โดยวิธีการวัดคุณภาพของกลุ่มข้อมูลที่ได้จากการจัดกลุ่มนั้น สามารถทำได้โดยใช้ค่าดัชนีเงา (Silhouette index) [25] ซึ่งเป็นเทคนิคที่มีการนำไปใช้ในการวัดคุณภาพของกลุ่มข้อมูลที่ได้ภายหลังการจัดกลุ่ม [6]

ค่าดัชนีเงา (Silhouette index) จะเป็นการวัดคุณภาพของกลุ่มข้อมูลโดยการวัดความคล้ายระหว่างข้อมูลภายในกลุ่มเดียวกัน และข้อมูลที่อยู่ต่างกลุ่มกัน ซึ่งการจัดกลุ่มที่มีคุณภาพ ข้อมูลภายในกลุ่มเดียวกันควรจะมีความคล้ายกันมาก ๆ ในขณะที่ข้อมูลที่อยู่ต่างกลุ่มควรจะมีความแตกต่างกันมาก ๆ ซึ่งการคำนวณค่าดัชนีเงานั้น สามารถทำได้โดยใช้สมการที่ (4.4) และ (4.5)

$$s(i) = \frac{b(i) - a(i)}{max\{b(i), a(i)\}} \tag{4.4}$$

$$S = \frac{1}{n} \sum_{i=1}^{n} s(i) \tag{4.5}$$

จากสมการที่ (4.4) และ (4.5) S คือ ค่าดัชนีเงาของข้อมูลทั้งหมดภายหลังจากการจัดกลุ่มเสร็จสิ้นแล้ว ซึ่งสามารถคำนวณได้จากค่า $s(i)$ ซึ่งเป็นค่าดัชนีเงาของแต่ละข้อมูลอนุกรมเวลา i โดย $a(i)$ คือ ค่าความคล้ายเฉลี่ยระหว่างข้อมูลอนุกรมเวลาตัวที่ i กับอนุกรมเวลาตัวอื่น ๆ ที่อยู่ภายในกลุ่มเดียวกัน ส่วน $b(i)$ คือ ค่าที่น้อยที่สุดของค่าความคล้ายเฉลี่ยระหว่างข้อมูลอนุกรมเวลาตัวที่ i กับอนุกรมเวลาตัวอื่น ๆ ที่อยู่ต่างกลุ่ม โดยผลการเปรียบเทียบค่าดัชนีเงาของวิธีจัดกลุ่มที่นำเสนอได้แสดงไว้ดังภาพที่ 4.3



ก) ข)

ภาพที่ 4.3 เปรียบเทียบค่าดัชนีเงา (Silhouette index) ของวิธีจัดกลุ่มที่นำเสนอกับ
ก) การจัดกลุ่มแบบเคมีนส์ร่วมกับมาตรวัดระยะยุคลิด และการเฉลี่ยแบบแอมพลิจูด
ข) การจัดกลุ่มแบบเคมีดอยส์ร่วมกับร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิง

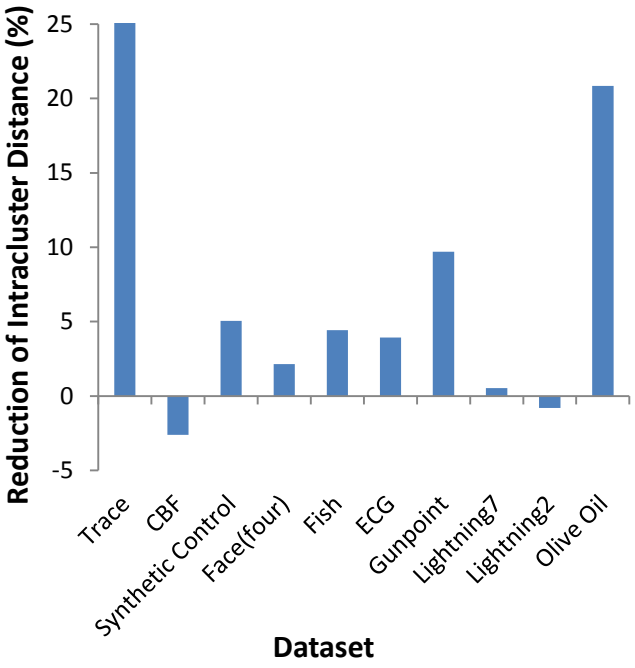
ภาพที่ 4.3 แสดงผลการเปรียบเทียบค่าดัชนีเงาของวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลาที่นำเสนอ กับวิธีการจัดกลุ่มแบบเคมีนส์ร่วมกับมาตรวัดระยะยุคลิดและการเฉลี่ยแบบแอมพลิจูด (ดังภาพที่ 4.3 ก) โดยจะเห็นว่า วิธีการจัดกลุ่มตามรูปร่างให้ผลลัพธ์ของกลุ่มที่มีค่าดัชนีเงาสูงกว่า ซึ่งแสดงว่าวิธีดังกล่าวสามารถจัดกลุ่มข้อมูลอนุกรมเวลาที่มีลักษณะคล้ายกันให้อยู่ในกลุ่มเดียวกันได้ดีกว่า ทั้งนี้เนื่องจากมาตรวัดระยะไดนามิกไทม์วอร์ปปิงและวิธีการหาตัวแทนกลุ่มแบบใหม่ที่ใช้การเฉลี่ยแบบรูปร่างนั้นเหมาะสมกับลักษณะของข้อมูลอนุกรมเวลามากกว่า และเมื่อเทียบค่าดัชนีเงากับวิธีการจัดกลุ่มแบบเคมีดอยส์ร่วมกับร่วมกับมาตรวัดระยะ ไดนามิกไทม์วอร์ปปิง (ดังภาพที่ 4.3 ข) จะพบว่ามีค่าใกล้เคียงกัน ดังนั้นในการทดลองถัดไปจึงได้ทำการเปรียบเทียบวิธีการจัดกลุ่มตามรูปร่างกับวิธีการจัดกลุ่มแบบเคมีดอยส์เพิ่มเติม

4.2.4 การเปรียบเทียบโดยใช้ค่าความคล้ายรวมของข้อมูลในกลุ่ม (Intracluster distance)

การวัดค่าความคล้ายรวมของข้อมูลในกลุ่ม (Intracluster distance) เป็นวิธีหนึ่งที่ใช้ในการวัดคุณภาพของข้อมูลตัวแทนกลุ่มที่ได้ [15] เนื่องจากตัวแทนของกลุ่มข้อมูลและข้อมูลที่เป็นสมาชิกภายในกลุ่มควรจะมีความคล้ายคลึงกัน ซึ่งการทดลองในส่วนนี้ได้ทำการวัดความคล้ายของตัวแทนกลุ่มและสมาชิกภายในกลุ่มที่ได้ โดยคำนวณจากสมการที่ (4.6)

$$\text{Intracluster Distance} = \sum_{k=1}^{K} \sum_{i=1}^{n} \text{DTW}(C_k, M_{k_i}) \quad (4.6)$$

จากสมการที่ (4.6) C_k และ M_{k_i} คือ ข้อมูลตัวแทนกลุ่มที่ k และสมาชิกในกลุ่มที่ k ตามลำดับ ส่วนค่าความคล้ายระหว่างข้อมูลหาได้โดยใช้มาตรวัดระยะไดนามิกไทม์วอร์ปปิง ซึ่งค่าความคล้ายรวม คำนวณได้จากผลรวมของความคล้ายระหว่างข้อมูลตัวแทนกลุ่มและสมาชิกภายในกลุ่มทุกกลุ่ม โดยได้แสดงการเปรียบเทียบค่าความคล้ายรวมของข้อมูลภายในกลุ่มไว้ดังภาพที่ 4.4



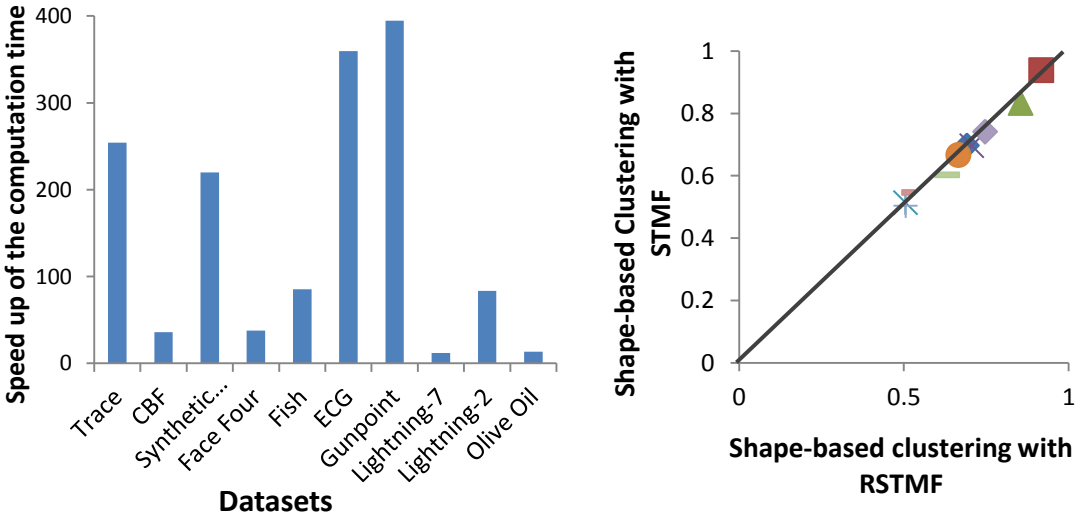
ภาพที่ 4.4 ค่าความคล้ายรวมของข้อมูลภายในกลุ่มที่ลดลง เมื่อใช้วิธีการจัดกลุ่มตามรูปร่างเทียบกับวิธีการจัดกลุ่มแบบเคมีดอยส์ร่วมกับร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิง

ภาพที่ 4.4 แสดงเป็นร้อยละของค่าความคล้ายรวมของข้อมูลภายในกลุ่ม เมื่อใช้วิธีการจัดกลุ่มตามรูปร่างที่ลดลงเทียบกับค่าค่าความคล้ายรวมของข้อมูลภายในกลุ่ม เมื่อใช้วิธีการจัดกลุ่มแบบเคมีดอยส์ร่วมกับร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิง ซึ่งจะพบว่าวิธีการจัดกลุ่มตามรูปร่าง สามารถลดค่าความคล้ายรวมของข้อมูลภายในกลุ่มลงได้ แสดงว่าข้อมูลตัวแทนกลุ่มและสมาชิกภายในกลุ่มที่ได้จากการจัดกลุ่มตามรูปร่างนั้นมีความคล้ายคลึงกันมากกว่าการจัดกลุ่มแบบเคมีดอยส์ร่วมกับร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิง เนื่องจากตัวแทนกลุ่มที่ได้จากการเฉลี่ยด้วยวิธี RSTMF นั้น สามารถรักษารูปร่างลักษณะของข้อมูลอนุกรมเวลาภายในกลุ่มที่ถูกนำมาเฉลี่ยเอาไว้ได้

4.3 การทดลองเพื่อประเมินประสิทธิภาพของวิธีการหาตัวแทนกลุ่ม Ranked Shape-based Template Matching Framework (RSTMF)

จากวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา ซึ่งเป็นการประยุกต์ใช้การจัดกลุ่มแบบเคมีนส์ร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิงแทนการใช้มาตรวัดระยะยุคลิด นอกจากนี้ยังได้มีการนำเสนอวิธีการหาตัวแทนกลุ่มที่เหมาะกับลักษณะของข้อมูลอนุกรมเวลาเพื่อใช้แทนวิธีการเฉลี่ยแบบแอมพลิจูดอีกด้วย

ซึ่งวิธีการหาตัวแทนกลุ่มที่นำเสนอนั้น คือ Ranked Shape-based Template Matching Framework (RSTMF) ซึ่งเป็นวิธีที่ปรับปรุงเพิ่มเติมจาก Shape-based Template Matching Framework (STMF) ให้สามารถหาตัวแทนข้อมูลได้รวดเร็วยิ่งขึ้น เนื่องจาก STMF จะต้องทำการคำนวณระยะไดนามิกไทม์วอร์ปปิงระหว่างข้อมูลทุกคู่เพื่อหาข้อมูลคู่ที่มีความคล้ายกันมากที่สุดมาเฉลี่ย ซึ่งระยะไดนามิกไทม์วอร์ปปิงนั้นต้องใช้เวลาในการคำนวณสูง เพราะฉะนั้น RSTMF จึงทำการประมาณลำดับของข้อมูลอนุกรมเวลาที่ต้องการเฉลี่ย เพื่อลดเวลาในการคำนวณระยะไดนามิกไทม์วอร์ปปิง โดยแสดงผลการเปรียบเทียบดังภาพที่ 4.5



ก)

ข)

ภาพที่ 4.5 เปรียบเทียบการจัดกลุ่มตามรูปร่างร่วมกับวิธี STMF และ RSTMF

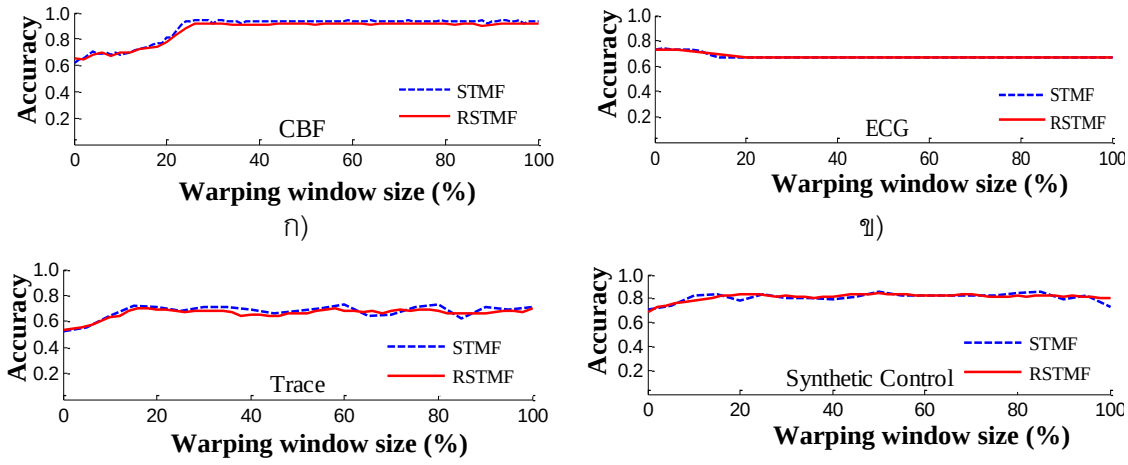
ก) ความเร็วในการคำนวณที่เพิ่มขึ้นเมื่อใช้วิธี RSTMF

ข) ความแม่นยำของผลลัพธ์ที่ได้จากการจัดกลุ่ม

ภาพที่ 4.5 แสดงผลการเปรียบเทียบวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลาเมื่อหาตัวแทนกลุ่มด้วยวิธี STMF และ RSTMF ซึ่งวิธีหาตัวแทนกลุ่ม RSTMF สามารถลดเวลาในการคำนวณระยะไดนามิกไทม์วอร์ปปิงเพื่อหาคู่ที่คล้ายกันมากที่สุดมาเฉลี่ย โดยใช้วิธีการประมาณลำดับของข้อมูลอนุกรมเวลาที่ต้องการเฉลี่ยซึ่งได้กล่าวไว้ในบทก่อนหน้า จึงทำให้เพิ่มความเร็วในการจัดกลุ่มได้หลายเท่าตัว (ดังภาพที่ 4.5 ก) ในขณะที่ความแม่นยำที่ได้จากการจัดกลุ่มนั้นยังคงใกล้เคียงกัน (ดังภาพที่ 4.5 ข)

4.4 การทดลองเพื่อแสดงผลการจัดกลุ่มตามรูปร่างโดยนำการกำหนดเงื่อนไขบังคับโดยรวม (Global constraint) มาประยุกต์ใช้

ในบางกรณี การใช้มาตรวัดระยะไดนามิกไทม์วอร์ปปิงในการวัดความคล้ายของข้อมูลอนุกรมเวลานั้น อาจเกิดการปรับแนวที่ไม่เหมาะสมดังที่เคยกล่าวไว้ในบทที่ 2 ซึ่งลักษณะดังกล่าวจะทำให้เกิดความผิดพลาดในการจัดกลุ่มข้อมูลได้ เพราะฉะนั้น การนำการกำหนดเงื่อนไขบังคับโดยรวม (Global constraint) มาประยุกต์ใช้ร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิงนั้น ช่วยลดการปรับแนวที่ไม่เหมาะสมลงได้ เนื่องจากมีการจำกัดขอบเขตของการปรับแนว โดยผลการทดลองการจัดกลุ่มตามรูปร่างที่นำการกำหนดเงื่อนไขบังคับโดยรวมมาประยุกต์ใช้ได้แสดงไว้ดังภาพที่ 4.6



ภาพที่ 4.6 เปรียบเทียบความแม่นยำของผลลัพธ์จากชุดข้อมูล ก) CBF ข) ECG ค) Trace และ ง) Synthetic Control ซึ่งได้จากวิธีการจัดกลุ่มตามรูปร่างร่วมกับการหาตัวแทนกลุ่มแบบ STMF และ RSTMF ที่การกำหนดเงื่อนไขบังคับโดยรวมค่าต่าง ๆ

ภาพที่ 4.6 แสดงความแม่นยำของผลลัพธ์ที่ได้จากวิธีการจัดกลุ่มตามรูปร่างร่วมกับวิธีการหาตัวแทนกลุ่มแบบ STMF และ RSTMF ซึ่งมีการประยุกต์ใช้การกำหนดเงื่อนไขโดยรวม จะเห็นว่าเมื่อทำการจำกัดขอบเขตการปรับแนวเป็นค่าต่าง ๆ การจัดกลุ่มตามรูปร่างร่วมกับการหาตัวแทนกลุ่มทั้งสองวิธียังคงให้ผลลัพธ์ของกลุ่มที่มีความแม่นยำใกล้เคียงกัน นอกจากนี้ผลการทดลองยังแสดงให้เห็นว่า การกำหนดขอบเขตของการปรับแนวให้มีขนาดใหญ่ไม่จำเป็นต้องให้ผลลัพธ์ของการจัดกลุ่มที่ดีขึ้นเสมอไป เนื่องจากข้อมูลอนุกรมเวลาอาจเกิดการปรับแนวที่ไม่เหมาะสมขึ้นได้ ซึ่งจากการทดลองจะเห็นได้ว่า ขอบเขตการปรับแนวที่เหมาะสมจะอยู่ที่ประมาณ 20 เปอร์เซนต์ของความยาวของข้อมูลอนุกรมเวลา

บทที่ 5

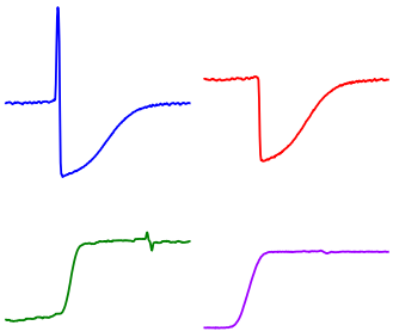
สรุปผลการวิจัย และข้อเสนอแนะ

วิทยานิพนธ์นี้ได้นำเสนอวิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา หรือ Shape-based Clustering for Time Series Data (SCTS) ซึ่งเป็นการประยุกต์ใช้วิธีการจัดกลุ่มแบบเคมีนส์ร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิง นอกจากนี้ยังได้นำเสนอวิธีการหาตัวแทนกลุ่มของข้อมูลอนุกรมเวลาโดยใช้วิธีการเฉลี่ยข้อมูลอนุกรมเวลาที่เรียกว่า Ranked Shape-based Template Matching Framework (RSTMF) เพื่อนำมาประยุกต์ใช้กับวิธีการจัดกลุ่มตามรูปร่างในการหาตัวแทนกลุ่มตัวใหม่ ซึ่งจะช่วยให้ได้ตัวแทนกลุ่มของข้อมูลที่สามารถรักษารูปร่างลักษณะของข้อมูลอนุกรมเวลาเอาไว้ได้ และทำให้ผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลนั้นมีความแม่นยำมากยิ่งขึ้น ดังรายละเอียดที่กล่าวไว้ในบทที่ 4 ซึ่งผลที่ได้จากการวิจัยที่ได้นำเสนอไปนั้น สามารถสรุปได้ดังนี้

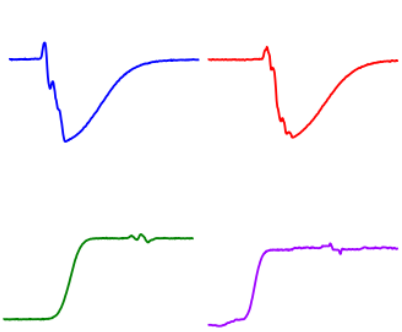
5.1 สรุปผลการวิจัย

การจัดกลุ่มแบบเคมีนส์ซึ่งนิยมใช้ระยะยุคลิดเป็นมาตรวัดความคล้ายระหว่างข้อมูล และใช้การเฉลี่ยแบบแอมพลิจูดในการหาตัวแทนของข้อมูลตัวใหม่นั้น เมื่อนำวิธีดังกล่าวมาใช้ในการจัดกลุ่มข้อมูลอนุกรมเวลา จะทำให้ผลลัพธ์ที่ได้จากการจัดกลุ่มนั้นเกิดความผิดพลาด เนื่องจากคุณสมบัติของมาตรวัดระยะยุคลิด และวิธีการเฉลี่ยแบบแอมพลิจูดนั้นไม่เหมาะสมกับลักษณะของข้อมูลอนุกรมเวลา

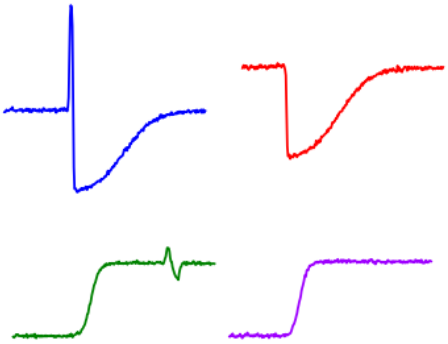
จากแนวคิดทั้งหมดที่ได้นำเสนอมาในบทก่อนหน้า จะเห็นได้ว่า วิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา ซึ่งเป็นการนำมาตรวัดระยะไดนามิกไทม์วอร์ปปิงและวิธีการเฉลี่ยแบบรูปร่างมาใช้ร่วมกับการจัดกลุ่มแบบเคมีนส์นั้น สามารถเพิ่มความแม่นยำของผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลได้ เนื่องจากมาตรวัดระยะไดนามิกไทม์วอร์ปปิงและวิธีการเฉลี่ยแบบรูปร่างนั้นมีความเหมาะสมกับลักษณะของข้อมูลอนุกรมเวลาที่มักจะเกิดการเลื่อนของข้อมูลในแนวแกนเวลาขึ้นได้ โดยการเปรียบเทียบลักษณะตัวแทนกลุ่มของข้อมูลอนุกรมเวลาภายหลังจากการจัดกลุ่มแบบเคมีนส์ตามปกติ และการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลา เสร็จสิ้นได้แสดงไว้ดังภาพที่ 5.1



ก)



ข)



ค)

ภาพที่ 5.1 ก) ตัวอย่างข้อมูลอนุกรมเวลาจากชุดข้อมูล Trace ซึ่งมี 4 คลาส
ข) ตัวแทนของกลุ่มข้อมูล เมื่อทำการจัดกลุ่มแบบเคมีนส์ตามปกติ และ
ค) เมื่อใช้วิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลาที่นำเสนอ

เมื่อทำการเปรียบเทียบกับวิธีการจัดกลุ่มที่มีการนำมาใช้กับข้อมูลอนุกรมเวลาวิธีอื่น ๆ ไม่ว่าจะเป็น การจัดกลุ่มแบบลำดับขั้นทั้งที่ใช้มาตรวัดระยะยุคลิดและมาตรวัดระยะไดนามิกไทม์วอร์ปปิง หรือการจัดกลุ่มแบบเคมีดอยส์ร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิง วิธีการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลาที่ได้นำเสนอนั้น ก็ยังคงให้ผลลัพธ์ของการจัดกลุ่มที่แม่นยำกว่า

นอกจากนี้วิธีการหาตัวแทนกลุ่มของข้อมูลอนุกรมเวลาโดยใช้วิธีการเฉลี่ยแบบ Ranked Shape-based Template Matching Framework (RSTMF) ซึ่งได้ทำการปรับปรุงเพิ่มเติมจากวิธี Shape-based Template Matching Framework (STMF) เพื่อให้มีการทำงานที่รวดเร็วขึ้นนั้น ก็จะเห็นว่า วิธีการหาตัวแทนกลุ่มด้วยการเฉลี่ยแบบ RSTMF สามารถลดเวลาการคำนวณลงได้

หลายเท่าตัว โดยใช้การประมาณลำดับของข้อมูลอนุกรมเวลาที่ต้องการเฉลี่ย แทนการคำนวณระยะไดนามิกไทม์วอร์ปปิงที่ต้องใช้เวลาในการคำนวณสูง

5.2 ข้อเสนอแนะ

จากแนวคิดของการจัดกลุ่มตามรูปร่างสำหรับข้อมูลอนุกรมเวลาที่นำเสนอในงานวิทยานิพนธ์นี้ ซึ่งเป็นการนำการจัดกลุ่มแบบเคมีนส์มาประยุกต์ใช้ร่วมกับมาตรวัดระยะไดนามิกไทม์วอร์ปปิง และการหาตัวแทนกลุ่มด้วยวิธีการเฉลี่ยแบบ Ranked Shape-based Template Matching Framework (RSTMF) นั้น จะเห็นว่าการจัดกลุ่มตามรูปร่างให้ผลลัพธ์ของการจัดกลุ่มที่ดีกว่าเนื่องจากวิธีการหาตัวแทนกลุ่มที่สามารถรักษารูปร่างของข้อมูลอนุกรมเวลาเอาไว้ได้ อย่างไรก็ตาม การหาตัวแทนกลุ่มของข้อมูลอนุกรมเวลาด้วยการเฉลี่ยแบบรูปร่างนั้น ผลลัพธ์ที่ได้ยังคงไม่ใช่ค่าเฉลี่ย (Mean) ของข้อมูลอนุกรมเวลา ซึ่งหากสามารถปรับปรุงวิธีการเฉลี่ยแบบรูปร่างให้มีผลลัพธ์ที่ดีขึ้น ก็จะสามารถนำไปใช้ประโยชน์ได้มากขึ้น

รายการอ้างอิง

[1] Niennattrakul, V., Wanichsan, D., and Ratanamahatana, C. A. Hand Geometry Verification Using Time Series Representation. In Proceedings of the 11th International Conference on Knowledge-Based Intelligent Information and Engineering Systems 2007.

[2] Ratanamahatana, C. A., and Keogh, E. Multimedia Retrieval Using Time Series Representation and Relevance Feedback. Digital Libraries: Implementing Strategies and Sharing Experiences 2005: 400-405.

[3] Niennattrakul, V., Ratanamahatana, C. A. On clustering multimedia time series data using k-means and dynamic time warping. In Proceedings of the International Conference on Multimedia and Ubiquitous Engineering. 2007: 733-738.

[4] Wismuller, A., Lange, O., Dersch, D.R., Leinsinger, G.L., Hahn, K., Pütz, B., and Auer, D. Cluster analysis of biomedical image time-series. International Journal of Computer Vision 46 (2002): 103-128.

[5] Shumway, R.H. Time-frequency clustering and discriminant analysis. Statistics and Probability Letters 63. 2003: 307-314.

[6] Meesrikamolkul, W., Niennattrakul, V., and Ratanamahatana, C.A. Multiple shape-based template matching for time series data. In Proceedings of the 8th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2011) Khonkan, Thailand. 2011: 464 -467.

[7] Liao, T.W., et al. Understanding and projecting battle states. In Proceedings of 23rd Army Science Conference 2002.

[8] Mueen, A., Keogh, E., Zhu, Q., Cash, S., and Westover, M.B. Exact discovery of time series motifs. In Proceedings of the SIAM Data Mining Conference (SDM 2009) 2009: 473-484.

[9] Nunthanid, P., Niennattrakul, V., Ratanamahatana, C. A. Discovery of variable-length time series motif. In Proceedings of the 8th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications, and Information Technology (ECTI-CON 2011) 2011: 472-475.

[10] Niennattrakul, V., Ruengronghirunya, P., Ratanamahatana, C. A. Exact indexing for massive time series databases under time warping distance. Data Mining and Knowledge Discovery 21 (2010): 509-541.

[11] Wei, L., Kumar, N., Lolla, V., Keogh, E. J., Lonardi, S., Ratanamahatana, C. A. Assumption-free anomaly detection in time series. In Proceedings of the 17th international conference on Scientic and statistical database management 2005: 237-240.

[12] Bradley, P. S., and Fayyad, U. M. Refining Initial Points for K-Means Clustering. In Proceedings of the 15th International Conference on Machine Learning 1998.

[13] Berndt, D. J., and Clifford, J. Using Dynamic Time Warping to Find Patterns in Time Series. In Proceedings of AAAI Workshop on Knowledge Discovery in Databases 1994: 359-370.

[14] Niennattrakul, V., and Ratanamahatana, C. A. Inaccuracies of shape averaging method using dynamic time warping for time series data. In Proceedings of the 7th international conference on Computational Science 2007 : 513-520.

[15] Niennattrakul, V., Srisai, D, and Ratanamahatana, C. A. Shape-based Template Matching for Time Series Data. Knowledge-Based Systems 26 (2011): 1-8.

[16] Keogh, E., Xi, X., Wei, L., and Ratanamahatana, C. A. The UCR Time Series Classification/Clustering Homepage. [Online]. 2008. Available from: www.cs.ucr.edu/~eamonn/time_series_data [2010/6]

[17] Ratanamahatana, C. A., and Keogh, E. Making time-series classification more accurate using learned constraints. In Proceedings of SIAM International Conference on Data Mining 2004: 11-22.

[18] Goldberger, A., et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. [Online]. 2000. Available from: http://www.physionet.org [2011/6]

[19] Tan, PN., Introduction to data mining. USA: Addison-Wesley, 2006.

[20] Ratanamahatana, C.A., and Keogh, E. Indexing and Mining Large Time Series Databases. Tutorial at 12th International Conference on Database Systems for Advanced Applications (DASFAA 2007) Bangkok, Thailand. 2007.

[21] Sakoe, H., and Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing 26 (1978): 43-49.

[22] Liao, T.W. Clustering of time series data-a survey. Pattern Recognition 38 (2005): 1857-1874.

[23] Vlachos, M., Lin, J., Keogh, E., and Gunopulos, D. A wavelet-based anytime algorithm for k-means clustering of time series. In Proceedings of Workshop on Clustering High Dimensionality Data and Its Applications 2003: 23-30.

[24] Gupta, L., Molfese, D. L., and Simos, P. G. Nonlinear alignment and averaging for estimating the evoked potential. IEEE Transactions on Biomedical Engineering 43 (1996): 348-356.

[25] Rousseeuw, P. J. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics 20 (1987): 53-65.

ประวัติผู้เขียนวิทยานิพนธ์

นางสาววริศรา มีศรีกมลกุล จบการศึกษาในระดับมัธยมศึกษาตอนปลายจากโรงเรียนนวมินทราชินูทิศ บดินทรเดชา จากนั้นเข้าศึกษาต่อในปีการศึกษา 2549 ที่ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย และสำเร็จการศึกษาในหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต ปีการศึกษา 2552 และได้เข้าศึกษาต่อในหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต ที่ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2553