การลดมิติของข้อมูลอนุกรมเวลาโดยใช้การแทนข้อมูลบางส่วน

นายกึกก้อง ศิริสัมพันธ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2560

DIMENSIONALITY REDUCTION USING PARTIAL REPRESENTATION OF TIME SERIES DATA

Mr. Kukkong Sirisambhand

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2017

หัวข้อวิทยานิพนธ์	การลดมิติของข้อมูลอนุกรมเวลาโดยใช้การแทนข้อมูลบางส่วน
โดย	นายกึกก้อง ศิริสัมพันธ์
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์

(รองศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(ดร.ดวงดาว วิชาดากุล)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ)

.....กรรมการภายนอกมหาวิทยาลัย

(ดร.เหมวรรณ ศิวรักษ์)

กึกก้อง ศิริสัมพันธ์ : การลดมิติของข้อมูลอนุกรมเวลาโดยใช้การแทนข้อมูลบางส่วน (DIMENSIONALITY REDUCTION USING PARTIAL REPRESENTATION OF TIME SERIES DATA) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: รศ. ดร.โชติรัตน์ รัตนามหัทธนะ, หน้า.

ข้อมูลอนุกรมเวลานั้นกลายเป็นสิ่งหนึ่งที่ใช้กันอย่างแพร่หลายในงานเหมืองข้อมูล เนื่องจากข้อมูลอื่นที่ไม่ใช่ข้อมูลอนุกรมเวลาสามารถนำมาเปลี่ยนแปลงให้กลายเป็นข้อมูลอนุกรมเวลาได้ แต่เนื่องจากลักษณะโดยทั่วไปของอนุกรมเวลาที่มีจำนวนมิติที่มาก ส่งผลให้การทำงานที่เกี่ยวข้องกับข้อมูลอนุกรมเวลาจะใช้เวลามากเช่นเดียวกัน โดยมีหลายงานวิจัยที่พยายามลดจำนวนมิติข้อมูลลงเพื่อลดเวลาในการทำงานกับข้อมูลอนุกรมเวลา แต่เนื่องจากวิธีการหลายวิธีการยังคงต้องอาศัยการเรียนรู้พารามิเตอร์อยู่ทำให้เสียเวลาในการเรียนรู้พารามิเตอร์ที่ค่อนข้างมาก วิทยานิพนธ์นี้เสนอการลดมิติข้อมูลให้เหลือน้อยที่สุดและเท่ากันสำหรับทุกข้อมูล ซึ่งจะใช้ข้อมูลเพียงบางส่วนจากอนุกรมเวลาเดิมเพื่อลดเวลาในการทำงาน โดยผลการทดลองพบว่าวิธีการที่นำเสนอทำงานได้รวดเร็วกว่างานวิจัยอื่นที่นำมาเปรียบเทียบ โดยยังคงความแม่นยำในการจำแนกประเภทข้อมูลไว้ได้

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ปีการศึกษา 2560

5970112321 : MAJOR COMPUTER SCIENCE
KEYWORDS: TIME SERIES CLASSIFICATION / DIMENSIONALITY REDUCTION / TIME SERIES REPRESENTATION

KUKKONG SIRISAMBHAND: DIMENSIONALITY REDUCTION USING PARTIAL REPRESENTATION OF TIME SERIES DATA. ADVISOR: ASSOC. PROF. CHOTIRAT RATANAMAHATANA, Ph.D., pp.

Time series data is one of the data that mostly used in data mining because many kinds of data can be transformed into time series data. However, the nature of time series data is high dimensionality. Therefore, a lot of time has been wasted in many time series works. There are many research works that try to reduce the number of dimension. However, it still needs a lot of time for parameter training. In this thesis, a solution is proposed to minimize the number of dimensions which would be the same number for every kind of data. Only parts of time series are used to reduce processing time. From the experiments, the proposed method is much faster than other methods while maintaining high classification accuracy.

Department: Computer Engineering Student's Signature .....
Field of Study: Computer Science Advisor's Signature .....
Academic Year: 2017

กิตติกรรมประกาศ

วิทยานิพนธ์นี้จะสำเร็จลุล่วงไม่ได้หากปราศจากความกรุณาจาก รองศาสตราจารย์ ดร. โชติรัตน์ รัตนามหัทธนะ อาจารย์ที่ปรึกษา ผู้ให้คำแนะนำสำหรับแนวทางการทำงานวิจัยในด้านต่าง ๆ และเป็นผู้ให้คำแนะนำในด้านวิชาการและด้านอื่น ๆ รวมถึงเป้นผู้ตรวจทานแก้ไขวิทยานิพนธ์ฉบับนี้ให้สำเร็จลุล่วง ขอขอบพระคุณเป็นอย่างสูงมา ณ ที่นี้

ขอขอบพระคุณ ดร.ดวงดาว วิชาดากุล และ ดร.เหมวรรณ ศิวรักษ์ ผู้ให้เกียรติเป็นกรรมการสอบวิทยานิพนธ์ และชี้แนะแนวทางในการปรับปรุงวิทยานิพนธ์ให้ดียิ่งขึ้น

สุดท้ายนี้ขอขอบคุณสมาชิกครอบครัวทุกคน ที่คอยเป็นกำลังใจและสนับสนุนในทุกด้านจนทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี

สารบัญ

หน้า

หน้า

หน้า

หน้า

สารบัญรูปภาพ

สารบัญตาราง

บทที่ 1 บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ข้อมูล (Data) คือเรื่องราวที่เกี่ยวข้องกับสิ่งต่าง ๆ ซึ่งวิธีการในการได้มาซึ่งข้อมูลนั้นอาจมีหลากหลาย ทำให้ประเภทของข้อมูลมีความหลากหลายแตกต่างกันไปด้วย ในอดีตนั้นการวิเคราะห์หรือใช้ข้อมูลเพื่อนำไปสร้างประโยชน์ถือว่าอยู่ในวงแคบ แต่ในปัจจุบันมีการใช้ประโยชน์ของข้อมูลหลากหลายด้าน ทำให้ปัจจุบันข้อมูลเป็นสิ่งที่สำคัญและนับว่ามีค่าเป็นอย่างมาก

ข้อมูลที่ถูกเก็บรวบรวมมานั้นมีหลากหลายประเภท ข้อมูลประเภทหนึ่งที่น่าสนใจ คือ ข้อมูลอนุกรมเวลา (Time Series Data) [1] ซึ่งเป็นข้อมูลที่ถูกเก็บในช่วงระยะเวลาหนึ่งอย่างต่อเนื่อง ซึ่งสามารถพบเห็นข้อมูลลักษณะนี้ได้โดยทั่วไปตราบใดที่มีเวลาเป็นตัวกำหนดการเกิดข้อมูล ดังนั้นลักษณะการผันแปรเปลี่ยนไปของข้อมูลในแต่ละช่วงเวลานั้นย่อมมีความเกี่ยวข้องกัน ซึ่งการวิเคราะห์ข้อมูลอนุกรมเวลาย่อมทำให้เห็นเหตุในการเปลี่ยนไปของข้อมูลอย่างชัดเจน นอกจากนี้การเปลี่ยนแปลงของข้อมูลในแต่ละช่วงเวลายังสามารถนำไปพิจารณาแนวโน้มหรือทิศทางของข้อมูลที่จะเป็นไปในอนาคตได้ ซึ่งมักจะเห็นกันอย่างแพร่หลายในการทำนายในลักษณะต่าง ๆ เช่น ทำนายการขึ้นลงของหุ้น ทำนายสภาพอากาศ ทำนายสภาพเศรษฐกิจ เป็นต้น

จากประโยชน์ของข้อมูลอนุกรมเวลาทำให้มีการศึกษาวิจัยเกี่ยวกับข้อมูลอนุกรมเวลาอย่างกว้างขวาง สิ่งหนึ่งที่น่าสนใจ คือ การวัดความคล้ายคลึง (Similarity Measure) เป็นการจำแนกประเภทของข้อมูลอนุกรมเวลาในกลุ่มเดียวกัน โดยตัววัดความคล้ายคลึงที่มักถูกกล่าวถึงในหลากหลายงานวิจัย คือ การวัดระยะทางแบบยุคลิด (Euclidean distance) [2] ซึ่งเป็นวิธีที่รู้จักและใช้กันอย่างแพร่หลายในการจำแนกประเภทข้อมูล (Classification) [3] แต่เมื่อถูกนำมาใช้ในข้อมูลอนุกรมเวลายังไม่สามารถให้ความแม่นยำได้ดีมากนัก เนื่องจากการปรับแนว (Alignment) ของการวัดระยะทางแบบยุคลิดนั้นไม่เหมาะสมกับข้อมูลอนุกรมเวลาบางประเภท ต่อมาจึงมีการพัฒนาวิธีการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง (Dynamic Time Warping distance-DTW) [4] ซึ่งเป็นวิธีการที่ถูกพัฒนามาเพื่อแก้ไขการปรับแนวให้เหมาะสมกับลักษณะข้อมูลมากยิ่งขึ้น ส่งผลให้ความแม่นยำของผลลัพธ์เพิ่มมากขึ้นในหลายชุดข้อมูล แต่ระยะเวลาในการทำงานก็เพิ่มขึ้นมากเช่นกันเมื่อเทียบกับการวัดระยะทางแบบยุคลิด

จะเห็นว่าปัญหาทางด้านเวลาในข้อมูลอนุกรมเวลานั้นเกิดจากลักษณะโดยทั่วไปของข้อมูลอนุกรมเวลาที่มีมิติข้อมูลจำนวนมาก ดังนั้นการลดมิติข้อมูล (Dimensionality Reduction) จึงเป็นการแก้ไขปัญหาด้านเวลาได้มีประสิทธิภาพสูงสุด โดยเป็นการแทนข้อมูล (Data Representation) ใหม่จากชุดข้อมูลเดิม เพื่อให้ได้ข้อมูลชุดใหม่ที่สามารถแสดงถึงลักษณะต่าง ๆ ของข้อมูลเดิมด้วย

จำนวนมิติข้อมูลที่ลดลง งานวิจัยที่ถูกนำเสนอและรู้จักกันอย่างแพร่หลาย คือ การแทนข้อมูลแบบแซคซ์ (Symbolic Aggregate approXimation-SAX) [5] ซึ่งเป็นวิธีที่พัฒนาต่อจากการแทนข้อมูลแบบพีเอเอ (Piecewise Aggregate Approximation-PAA) [6] โดยการลดมิติข้อมูลของแซคซ์นั้นจำเป็นที่จะต้องเลือกขนาดมิติข้อมูลที่เหมาะสมก่อน ซึ่งในข้อมูลแต่ละประเภทก็มีจำนวนมิติของข้อมูลที่เหมาะสมต่างกันไป ดังนั้นการเลือกขนาดมิติข้อมูลของผู้ใช้อย่างไม่เหมาะสมย่อมส่งผลต่อความแม่นยำและระยะเวลาในการทำงานค่อนข้างมาก นอกจากนี้ยังมีวิธีการกระเป๋าคำ (Bag Of Words) [7] ซึ่งเป็นวิธีการที่ถูกพัฒนาต่อมาจากการแทนข้อมูลแบบแซกซ์ โดยการนับคำจากการตัดเพื่อนำไปแทนข้อมูลใหม่ ซึ่งเป็นการเพิ่มพารามิเตอร์ (Parameter) ให้มากยิ่งขึ้นจากเดิมทำให้ยากต่อการใช้งานสำหรับผู้ใช้ทั่วไป

ดังนั้นปัญหาในการเลือกขนาดมิติข้อมูลที่เหมาะสมจะหมดไป ถ้าหากสามารถลดขนาดมิติข้อมูลให้เหลือเท่ากันเสมอสำหรับทุกข้อมูล ซึ่งในการลดขนาดมิตินั้นถ้าหากสามารถลดให้เหลือขนาดของมิติจำนวนน้อยมาก ๆ ก็ยิ่งส่งผลให้เวลาในการทำงานน้อยลงมากเช่นเดียวกัน ดังนั้นผู้วิจัยจึงเล็งเห็นว่า ถ้าหากสามารถหาค่าจำนวนจริงใด ๆ ที่สามารถแสดงถึงลักษณะดั้งเดิมของอนุกรมเวลานั้น ๆ ได้ และสามารถแสดงได้ด้วยมิติข้อมูลที่น้อยที่สุดซึ่งเท่ากันในทุกข้อมูล ก็ย่อมจะทำให้เวลาในการทำงานเร็วกว่าจำนวนมิติข้อมูลที่มากกว่า และผู้ใช้ไม่ต้องกังวลต่อความยาวของมิติข้อมูลที่เหมาะสม

ในวิทยานิพนธ์ฉบับนี้ผู้วิจัยได้นำเสนอการลดมิติข้อมูลอนุกรมเวลาสำหรับทุกประเภทข้อมูลให้เหลือเพียง 2 มิติ และยังสามารถสะท้อนถึงลักษณะดั้งเดิมของอนุกรมเวลาได้ ส่งผลให้ยังคงคุณสมบัติในการจำแนกประเภทข้อมูลได้เป็นอย่างดี ซึ่งในการหาจำนวนจริงเพื่อแทนค่าข้อมูลดั้งเดิมนั้น ผู้วิจัยใช้ความชัน (slope) จากการลากเส้นขึ้นใหม่จากส่วนหนึ่งในอนุกรมเวลา ซึ่งค่าจำนวนจริงดังกล่าวจะยังคงทำให้ความแม่นยำและเวลามีประสิทธิภาพ เมื่อเปรียบเทียบกับการแทนข้อมูลรูปแบบอื่น ๆ ซึ่งในวิทยานิพนธ์ฉบับนี้ผู้วิจัยใช้วิธีการจำแนกประเภทข้อมูลแบบเพื่อนบ้านใกล้ที่สุดลำดับที่หนึ่ง (1-Nearest Neighbor) [8] ด้วยการทดสอบแบบการนำออกหนึ่ง (Leave-One-Out Cross Validation) ซึ่งการทดสอบด้วยวิธีการนี้จะสามารถวัดประสิทธิภาพได้อย่างชัดเจนและสามารถลดความเอนเอียงจากการแบ่งข้อมูลได้เป็นอย่างดี

1.2 วัตถุประสงค์ของงานวิจัย

1. ลดขนาดของมิติข้อมูลให้น้อยที่สุดเท่าที่จะเป็นไปได้ สำหรับข้อมูลอนุกรมเวลาหลากหลายประเภท โดยให้ยังคงมีประสิทธิภาพทั้งในด้านความแม่นยำและความเร็วในการทำงาน
2. นำเสนอวิธีการใหม่ในการลดมิติข้อมูลอนุกรมเวลาโดยการใช้ความชันจากส่วนใดส่วนหนึ่งของอนุกรมเวลา ในการแทนที่ข้อมูลเดิม
3. ผู้ใช้ไม่จำเป็นต้องกังวลกับหาขนาดมิติข้อมูลที่เหมาะสมของการแทนที่ข้อมูลที่นำเสนอ

1.3 ขอบเขตของการวิจัย

1. เปรียบเทียบประสิทธิภาพกับการลดมิติข้อมูลประเภทอื่น ๆ
2. ทดสอบความแม่นยำและเวลา ด้วยวิธีการจำแนกข้อมูลแบบเพื่อนบ้านใกล้ที่สุดลำดับที่หนึ่ง ด้วยการทดสอบแบบการนำออกหนึ่ง
3. ใช้ชุดข้อมูลในการทดลองจากฐานข้อมูล UCR Time Series Classification Archive เป็นหลัก

1.4 ประโยชน์ที่ได้รับ

การแทนที่ข้อมูลจากงานวิจัยนี้สามารถลดมิติข้อมูลให้เหลือขนาดเล็กที่สุดและมีขนาดเท่ากันสำหรับทุกประเภทข้อมูล โดยยังสามารถให้ผลลัพธ์ที่มีประสิทธิภาพทั้งในด้านของความแม่นยำและความเร็วในการทำงาน

1.5 วิธีการดำเนินงานวิจัย

1. ศึกษาเกี่ยวกับการทำเหมืองข้อมูลอนุกรมเวลา
2. ศึกษาเกี่ยวกับการจำแนกประเภทข้อมูลอนุกรมเวลาด้วยวิธีการจำแนกข้อมูลแบบเพื่อนบ้านใกล้ที่สุดลำดับที่หนึ่ง
3. ศึกษาการลดมิติของข้อมูลอนุกรมเวลาและตัววัดความคล้ายคลึงรูปแบบต่าง ๆ
4. ออกแบบและพัฒนาวิธีการลดมิติของข้อมูลโดยใช้ข้อมูลเพียงบางส่วนจากอนุกรมเวลา
5. ตีพิมพ์วิธีการลดมิติของข้อมูลส่วนหนึ่งที่ได้ออกแบบและพัฒนาไว้

6. พัฒนาวิธีการจากการตีพิมพ์ให้มีประสิทธิภาพมากยิ่งขึ้น

7. ทดสอบประสิทธิภาพของวิธีการแทนข้อมูลที่ออกแบบด้วยการประเมินด้วยการจำแนกข้อมูลแบบเพื่อนบ้านใกล้ที่สุดลำดับที่หนึ่ง และทดสอบด้วยการนำออกหนึ่ง

8. เปรียบเทียบผลการทดลองในด้านความแม่นยำและเวลากับการแทนข้อมูลแบบอื่น ๆ

9. วิเคราะห์และสรุปผลการทดลอง

10. สรุป เรียบเรียง และจัดทำวิทยานิพนธ์

1.6 ผลงานวิจัยที่ได้ตีพิมพ์

- K. Sirisambhand and C. A. Ratanamahatana, “A Dimensionality Reduction Technique for Time Series Classification Using Additive Representation,” in Third International Congress on Information and Communication Technology, ICICT 2018, at Brunel University London, London, United Kingdom, from 27 – 28 February 2018.

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ข้อมูลอนุกรมเวลา

เป็นค่าของข้อมูลที่ถูกเก็บควบคู่กับจุดเวลาที่เกิดขึ้น ดังนั้นทุกจุดข้อมูลในอนุกรมเวลาใด ๆ จะมีเวลากำกับเสมอ ซึ่งข้อมูลลักษณะนี้มักจะพบเห็นได้โดยทั่วไปทั้งใกล้และไกลตัว เช่นข้อมูลอนุกรมเวลาของตลาดหุ้น (stock market) [9] แสดงในรูปที่ 2.1 นอกจากนี้ยังมีข้อมูลคลื่นไฟฟ้าสมอง (Electroencephalography หรือ EEG) [10] ข้อมูลคลื่นไฟฟ้าหัวใจ (Electrocardiography หรือ ECG) [11] ข้อมูลสภาพอากาศในแต่ละวัน หรือแม้แต่ข้อมูลรายได้ของแม่ค้าที่ได้รับในแต่ละชั่วโมงก็นับเป็นข้อมูลอนุกรมเวลาเช่นเดียวกัน

รูปที่ 2.1 อนุกรมเวลาของราคาหุ้นรายวัน

(ที่มา : http://topicstock.pantip.com/sinthorn/topicstock/2011/04/I10481040/I10481040.html)

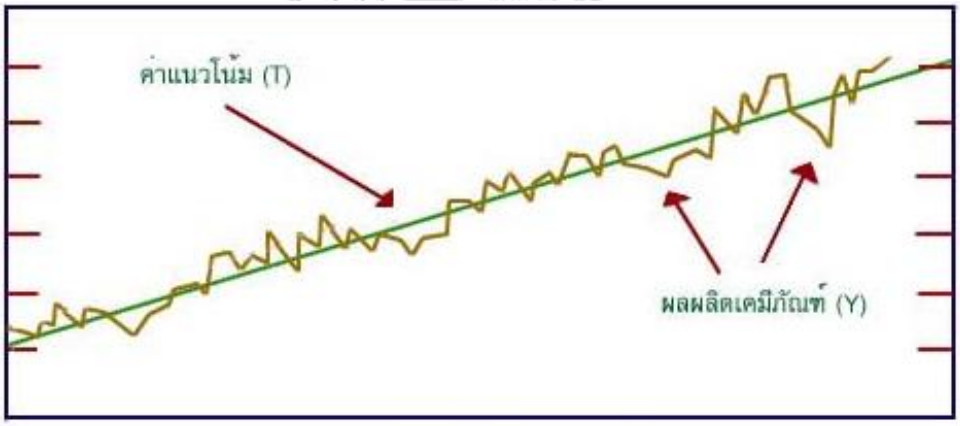
จะเห็นว่าข้อมูลอนุกรมเวลานั้นเป็นข้อมูลที่พบได้โดยทั่วไปในชีวิตประจำวัน ซึ่งนอกจากจะเป็นข้อมูลในลักษณะที่กล่าวไปข้างต้นแล้ว ข้อมูลอนุกรมเวลายังสามารถนำข้อมูลประเภทอื่นมาเปลี่ยนแปลงเป็นข้อมูลอนุกรมเวลาได้เช่นกัน เช่น ข้อมูลภาพใบหน้า (Faceall data) [12] ข้อมูลตัวอักษร (50Words data) [12] เป็นต้น ซึ่งการเปลี่ยนแปลงข้อมูลประเภทอื่นเป็นข้อมูลอนุกรมเวลาก็มีหลายรูปแบบขึ้นกับคุณลักษณะของข้อมูลที่ต้องการ ซึ่งประโยชน์ของการเปลี่ยนเป็นข้อมูลอนุกรมเวลานั้นคือการทำให้สามารถทำงานได้รวดเร็วยิ่งขึ้นในหลายชุดข้อมูล

2.2 การวิเคราะห์อนุกรมเวลา (Time Series Analysis) [13]

การวิเคราะห์อนุกรมเวลาเป็นการวิเคราะห์ลักษณะหรือรูปแบบของอนุกรมเวลา โดยสังเกตจากการเปลี่ยนแปลงของข้อมูลในแต่ละช่วงเวลาว่ามีการเปลี่ยนแปลงไปในลักษณะใด มีการเคลื่อนไหวของข้อมูลอย่างไร เนื่องจากอนุกรมเวลาแต่ละประเภทมีลักษณะที่แตกต่างกัน ดังนั้นจึงมีการกำหนดองค์ประกอบของอนุกรมเวลาเป็น 4 ส่วน ดังนี้

1. ส่วนแนวโน้ม (Trend Component)

   เป็นส่วนหลักที่สังเกตจากข้อมูลในระยะยาวโดยอาจมีการเติบโตหรือถดถอยในอนุกรมเวลา ซึ่งลักษณะของเส้นแนวโน้มนั้นขึ้นอยู่กับอนุกรมเวลา โดยอาจจะเป็นเส้นตรงหรือเส้นโค้งก็ได้ โดยเส้นแนวโน้มที่ถูกลากนั้นจะต้องเรียบไม่มีการเกิดมุมใด ๆ บนเส้น ดังตัวอย่างแสดงในรูปที่ 2.2 ซึ่งแสดงเส้นแนวโน้มที่น่าจะเป็นของข้อมูลอนุกรมเวลาผลผลิตเคมีภัณฑ์

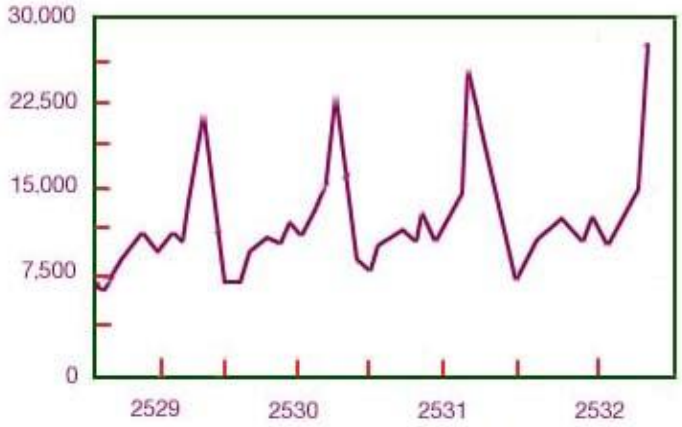


รูปที่ 2.2 เส้นแนวโน้มผลผลิตเคมีภัณฑ์ชนิดหนึ่ง

(ที่มา : http://www2.fpo.go.th/S-I/Source/ECO/ECO24.htm)

2. ส่วนฤดูกาล (Seasonal Component)

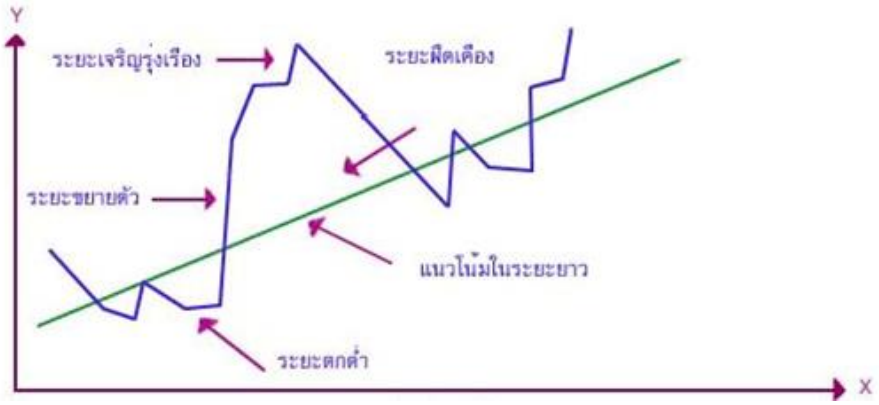
   เป็นการเปลี่ยนแปลงของอนุกรมเวลาในช่วงระยะเวลาหนึ่งในรูปแบบเดียวกันซ้ำ ๆ ซึ่งการเกิดของรูปแบบเดียวกันนี้จะเกิดในระยะเวลาสั้น ๆ ซึ่งในรูปที่ 2.3 จะพบลักษณะรูปแบบที่เกิดซ้ำกันในช่วงเวลาแต่ละปี



รูปที่ 2.3 แสดงยอดขายรายเดือนของห้างสรรพสินค้าแห่งหนึ่ง
(ที่มา : http://www2.fpo.go.th/S-I/Source/ECO/ECO24.htm)

3. ส่วนวัฏจักร (Cyclical Component)

เป็นการเกิดของเหตุการณ์ที่คล้ายกับส่วนฤดูกาล คือการเกิดรูปแบบซ้ำ ๆ กันในอนุกรมเวลา แต่จะเกิดในช่วงระยะเวลาที่ยาวกว่า ซึ่งจำเป็นที่จะต้องสังเกตจากอนุกรมเวลาที่มีระยะเวลานานดังแสดงในรูปที่ 2.4



รูปที่ 2.4 แสดงวัฏจักรของธุรกิจ
(ที่มา : http://www2.fpo.go.th/S-I/Source/ECO/ECO24.htm)

4. ส่วนผิดปกติ (Irregular Component)

เป็นการเกิดขึ้นของเหตุการณ์หรือการเปลี่ยนแปลงไปของข้อมูลที่ไม่สามารถคาดการณ์ได้ โดยการเกิดขึ้นของเหตุการณ์นี้ไม่ได้อิงกับเหตุการณ์ใด ซึ่งเป็นการเปลี่ยนแปลงในเชิงสุ่ม (Random Variation)

จากองค์ประกอบของอนุกรมเวลาดังกล่าวทำให้สามารถสร้างแบบจำลองของข้อมูลอนุกรมเวลาได้ โดยมีแบบจำลองของอนุกรมเวลา ดังนี้

1. แบบจำลองผลบวก (Additive Model)

เป็นแบบจำลองที่องค์ประกอบทั้ง 4 ส่วนของอนุกรมเวลานั้น ๆ ไม่ขึ้นต่อกัน โดยเมื่อองค์ประกอบใด ๆ มีค่าเปลี่ยนแปลงไป จะไม่ส่งผลกระทบต่อการเปลี่ยนแปลงขององค์ประกอบอื่น ๆ

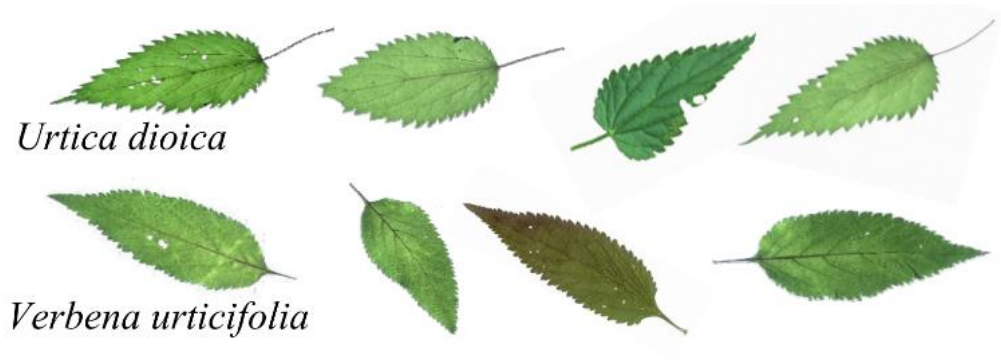
2. แบบจำลองผลคูณ (Multiplicative Model)

เป็นแบบจำลองที่องค์ประกอบทั้ง 4 ส่วนของอนุกรมเวลานั้น ๆ สัมพันธ์กัน โดยเมื่อองค์ประกอบใด ๆ มีค่าเปลี่ยนแปลงไป จะส่งผลให้องค์ประกอบอื่น ๆ มีการเปลี่ยนแปลงตามไปด้วย

2.3 เชปเล็ท (Shapelet) [14]

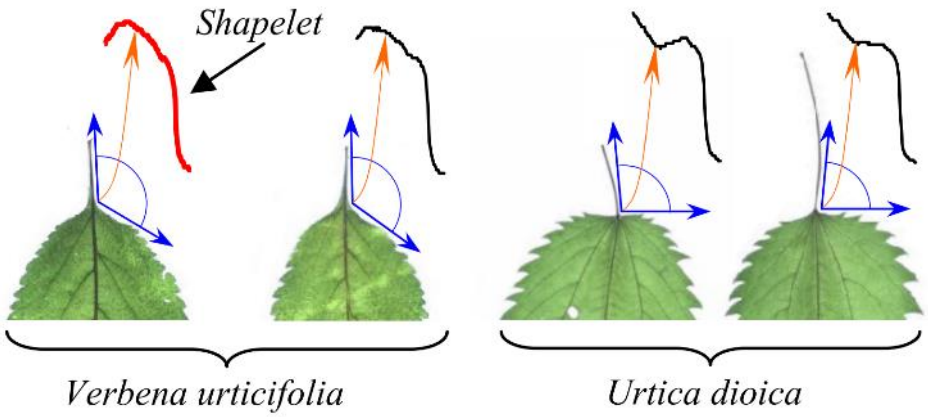
ในการทำเหมืองข้อมูลบนข้อมูลอนุกรมเวลาโดยทั่วไปนั้น ไม่ว่าจะเป็นการแบ่งกลุ่มหรือการจำแนกประเภทก็ตาม มักจะใช้ข้อมูลอนุกรมเวลาทั้งเส้น กล่าวคือมีการใช้ข้อมูลทุกมิติ ซึ่งในบางครั้งการใช้ข้อมูลทั้งเส้นของอนุกรมเวลาอาจทำให้เกิดปัญหาได้ เช่น สิ่งรบกวน (Noise) ที่อาจจะทำให้ความแตกต่างของอนุกรมเวลาที่อยู่คนละกลุ่มมีความคล้ายคลึงกัน ดังนั้นแทนที่จะใช้ข้อมูลทั้งเส้นของอนุกรมเวลา ก็เปลี่ยนเป็นใช้เพียงแค่ลำดับย่อยของอนุกรมเวลา (Subsequence) ซึ่งส่วนที่เลือกนี้จะเป็นส่วนที่สามารถแสดงถึงลักษณะของข้อมูลในกลุ่มนั้น ๆ ได้ โดยที่ไม่จำเป็นจะต้องพิจารณาข้อมูลอนุกรมเวลาทั้งหมด โดยส่วนที่เป็นตัวแทนของกลุ่มนี้เรียกว่าเชปเล็ท (Shapelet)

โดยขั้นตอนแรกในการหาเชปเล็ทเริ่มจากการหาลำดับย่อยทั้งหมดที่เป็นไปได้ของข้อมูลอนุกรมเวลา จากนั้นสังเกตว่าลำดับย่อยที่ทำการเลือกมานั้น สามารถแยกกลุ่มข้อมูลในกลุ่มเดียวกันและข้อมูลจากกลุ่มอื่นได้ดีมากน้อยเพียงใด ซึ่งลำดับย่อยที่สามารถแยกกลุ่มข้อมูลเดียวกันจากกลุ่มข้อมูลอื่นได้ดีที่สุด ก็จะเป็นเชปเล็ทที่ถูกเลือกเพื่อเป็นตัวแทนของกลุ่มข้อมูล นอกจากเชปเล็ทจะสามารถเป็นตัวแทนของกลุ่มข้อมูลได้แล้ว ยังสามารถแสดงให้เห็นถึงลักษณะของข้อมูลก่อนที่จะแปลงมาเป็นข้อมูลอนุกรมเวลาได้ด้วย เช่น ข้อมูลอนุกรมเวลาของใบไม้ Urtica dioica และ Verbena urticifolia ดังแสดงในรูปที่ 2.5



รูปที่ 2.5 ตัวอย่างใบไม้จากสองสปีชีส์ โดยใบไม้มีรอยการกัดแทะจากศัตรูพืช
(ที่มา : Ye และคณะ [14])

จากรูปที่ 2.5 จะเห็นว่าใบไม้มีรอยกัดแทะของศัตรูพืช หากใช้ข้อมูลอนุกรมเวลาทั้งเส้นมาแปลงเป็นเชปเล็ทอาจทำให้ได้ผลลัพธ์ไม่ดีเท่าที่ควรเนื่องจากสิ่งรบกวนดังกล่าว ดังนั้นจึงสามารถพิจารณาบางส่วนบนใบไม้ทั้งสองประเภทที่ต่างกันได้ ซึ่งจากรูปที่ 2.6 จะเห็นว่าองศาระหว่างก้านและใบของ Urtica dioica มีขนาดประมาณ 90 องศา ในขณะที่ก้านและใบของ Verbena urticifolia มีขนาดที่กว้างกว่า จะเห็นว่าเชปเล็ทนั้นสามารถแสดงถึงลักษณะของข้อมูลก่อนถูกแปลงมาเป็นข้อมูลอนุกรมเวลาได้ ด้วยคุณลักษณะของเชปเล็ทที่สามารถเป็นตัวแทนของกลุ่มข้อมูล ทำให้ไม่จำเป็นที่จะต้องพิจารณาข้อมูลอนุกรมเวลาทั้งหมด นอกจากจะแก้ปัญหาข้อมูลที่มีสิ่งรบกวนได้แล้วยังสามารถเพิ่มความเร็วในการทำงานได้อย่างมากอีกด้วย ทั้งนี้ขึ้นกับความยาวของเชปเล็ทว่ามีความยาวมากน้อยเพียงใด

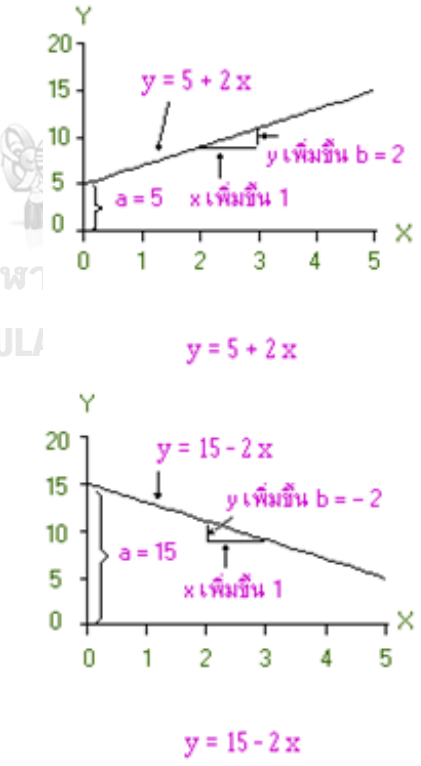


รูปที่ 2.6 เชปเล็ทแสดงถึงความแตกต่างของใบไม้ทั้งสองประเภท
(ที่มา : Ye และคณะ [14])

2.4 การถดถอยและการคะเนค่า

2.4.1 การวิเคราะห์ความถดถอยอย่างง่าย

การวิเคราะห์การถดถอยเป็นเทคนิคที่ใช้ในการประมาณค่าความสัมพันธ์ระหว่างตัวแปรสองประเภท ได้แก่ ตัวแปรอิสระ (independent variable-*x*) ซึ่งคือตัวแปรที่ทราบค่า และตัวแปรตาม (dependent variable-*y*) ซึ่งคือตัวแปรที่ต้องการประมาณค่า เนื่องจากตัวแปรอิสระเป็นตัวแปรที่ส่งผลต่อค่าของตัวแปรตาม แต่ในความเป็นจริงการรวบรวมตัวแปรอิสระให้ครบถ้วนเพื่อทำนายค่าตัวแปรตามให้ถูกต้องแน่นอนเป็นเรื่องยาก ดังนั้นจึงใช้การวิเคราะห์ความถดถอยอย่างง่าย กล่าวคือการที่ตัวแปรอิสระเพียงหนึ่งตัวที่ทำให้เกิดตัวแปรตาม ซึ่งก็จะลดปัญหาจากตัวแปรอิสระหลายตัวได้ ลักษณะที่พบได้บ่อยของความถดถอยอย่างง่ายนั้นคือลักษณะของสมการเส้นตรง โดยจะมีค่าความชันของเส้นตรงที่จะสามารถบอกลักษณะการแปรผันของตัวแปรอิสระและตัวแปรตามได้ ดังแสดงในรูปที่ 2.7 ซึ่งจะพบว่าเมื่อความชันเป็นบวกตัวแปรอิสระและตัวแปรตามจะผันตามกัน ในขณะที่ถ้าความชันเป็นลบตัวแปรอิสระและตัวแปรตามจะแปรผกผันกัน



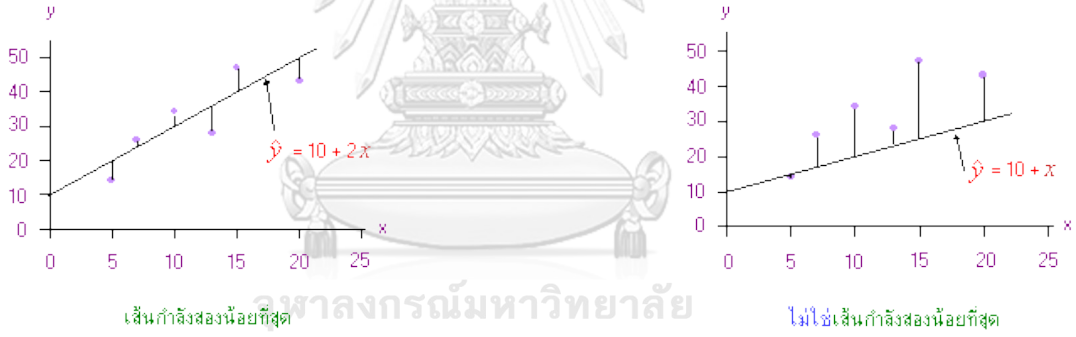
รูปที่ 2.7 แสดงการแปรผันของตัวแปรอิสระ x และตัวแปรตาม y

(ที่มา : https://web.ku.ac.th/schoolnet/snet2/knowledge_math/relation/relate3a.htm)

2.4.2 การสร้างสมการถดถอยอย่างง่าย

ในการที่จะสร้างเส้นถดถอยที่สามารถลากผ่านทุกจุดข้อมูลนั้นย่อมเป็นไปไม่ได้ กล่าวคือเส้นถดถอยไม่สามารถที่จะคาดคะเนค่าตัวแปรตามได้สำหรับทุกค่าตัวแปรอิสระ ดังนั้นจึงใช้เส้นตรงจากการถดถอยอย่างง่ายในการประมาณค่าตัวแปรตามให้ใกล้เคียงที่สุด

วิธีการที่เป็นที่นิยมในการประมาณค่าจุดตัดแกนตั้งและค่าความชันของเส้นตรง จากตัวแปรอิสระและตัวแปรตามใด ๆ คือ วิธีการกำลังสองน้อยที่สุด (Least Square Method) ซึ่งวิธีนี้จะให้ค่าคาดคะเนของค่าตัวแปรตามที่ใกล้เคียงกับค่าตัวแปรตามจริงมากที่สุดดังนั้น เมื่อเกิดการลากเส้นตรงถดถอยใหม่ เส้นตรงนี้จะลากผ่านกลางระหว่างจุดข้อมูลเดิมทำให้สามารถเห็นแนวโน้มของข้อมูลเดิมได้ ดังแสดงในรูปที่ 2.8 จะพบว่าเส้นตรงที่เกิดจากวิธีกำลังสองน้อยที่สุดจากลากผ่านกลางระหว่างจุดข้อมูล ซึ่งทำให้สามารถแสดงแนวโน้มของจุดข้อมูลได้ ในขณะที่เส้นตรงที่ไม่ได้เกิดจากวิธีกำลังสองน้อยที่สุด จะไม่สามารถแสดงแนวโน้มของจุดข้อมูลได้



รูปที่ 2.8 แสดงลักษณะของเส้นตรงที่เกิดจากวิธีการกำลังสองน้อยที่สุด

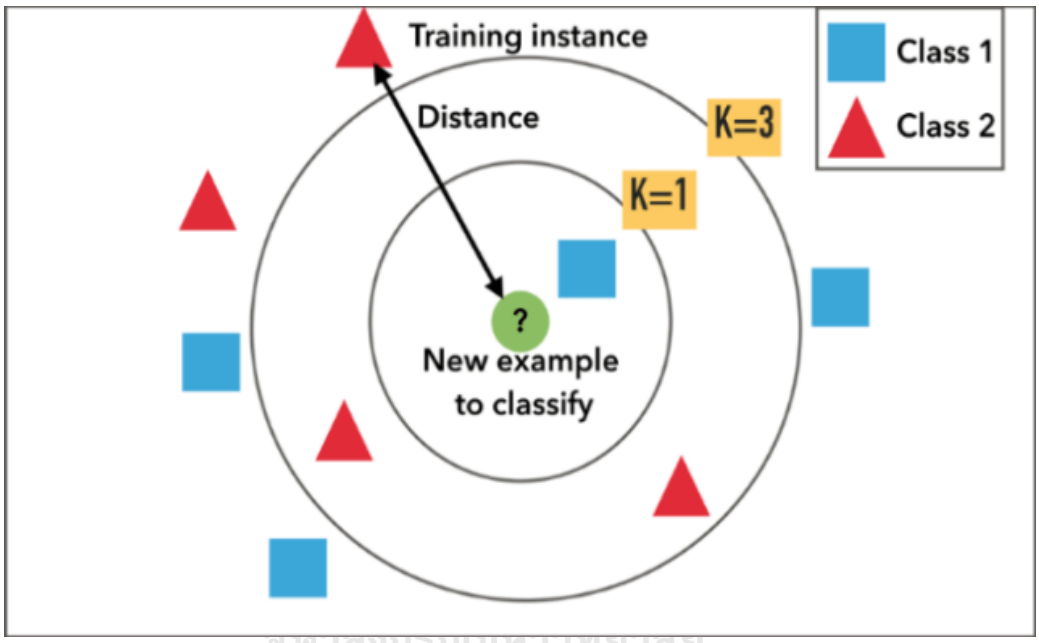
(ที่มา : https://web.ku.ac.th/schoolnet/snet2/knowledge_math/relation/relate3a.htm)

2.5 การจำแนกประเภทข้อมูลแบบเพื่อนบ้านใกล้ที่สุดลำดับที่หนึ่ง (1-Nearest Neighbor Classification) [8]

เป็นวิธีการจำแนกประเภทข้อมูลแบบหนึ่ง ซึ่งข้อมูลจำเป็นจะต้องถูกแบ่งเป็นข้อมูลทดสอบ (Test Data) และข้อมูลเรียนรู้ (Training Data) โดยการทำงานแบบเพื่อนบ้านใกล้ที่สุดลำดับหนึ่งนั้นจะเป็นการหาระยะทางระหว่างข้อมูลทดสอบและข้อมูลเรียนรู้ โดยมีขั้นตอนโดยสรุปดังนี้

1. สำหรับข้อมูลทดสอบหนึ่งตัว จะทำการหาระยะทางด้วยตัววัดระยะทางต่าง ๆ กับข้อมูลเรียนรู้ทุกข้อมูล

2. เลือกข้อมูลเรียนรู้ที่มีระยะทางใกล้กับข้อมูลทดสอบมากที่สุด เพื่อเป็นเพื่อนบ้านใกล้ที่สุดลำดับที่หนึ่ง
3. กำหนดประเภท (Class) ของข้อมูลทดสอบนั้น ๆ ว่าเป็นประเภทเดียวกันกับข้อมูลทดสอบที่ใกล้กันมากที่สุด
4. ทำตามข้อ 1 ถึง 3 สำหรับทุก ๆ ข้อมูลทดสอบ จะสามารถกำหนดประเภทให้กับข้อมูลทดสอบได้ครบทุกข้อมูล



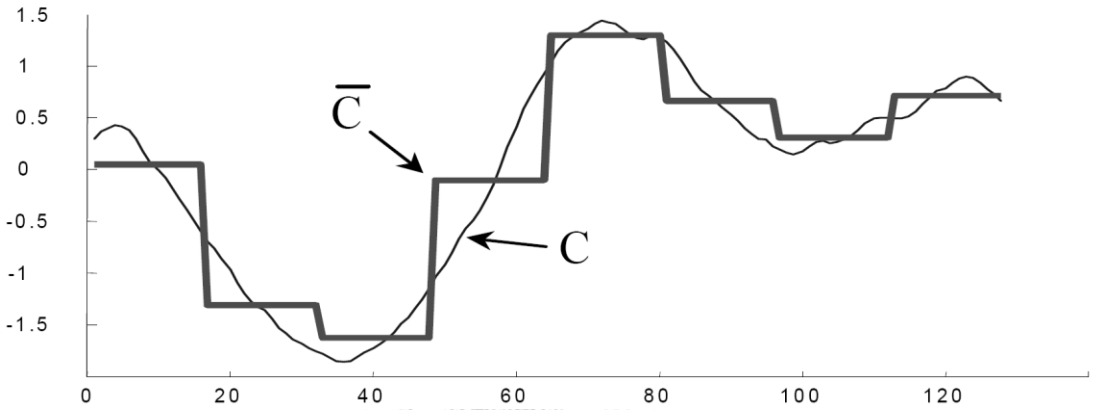
รูปที่ 2.9 การจำแนกประเภทข้อมูลแบบเพื่อนบ้านใกล้ที่สุดลำดับที่หนึ่งเมื่อ k=1
(ที่มา : https://arifuzzamanfaisal.com/k-nearest-neighbor-regression)

2.6 งานวิจัยที่เกี่ยวข้อง

ในการศึกษาวิจัยเกี่ยวกับข้อมูลอนุกรมเวลาในด้านของการแทนข้อมูลและการลดมิติของข้อมูล จะเน้นไปที่เรื่องของประสิทธิภาพด้านความเร็วในการทำงานเป็นหลัก ซึ่งมีงานวิจัยต่าง ๆ ที่เสนอวิธีการแทนข้อมูลและการลดมิติของข้อมูลดังต่อไปนี้

Yi และ Faloutsos (2000) [6] มีการเสนอการแทนข้อมูลแบบพีเอเอ โดยวิธีการแบบพีเอเอจะเป็นการแบ่งอนุกรมเวลาเป็นส่วน ๆ (Time Series Segmentation) โดยแต่ละส่วนมีความกว้างเท่ากัน จากนั้นหาค่าเฉลี่ยในช่วงนั้น ๆ เพื่อเป็นตัวแทนของข้อมูลแต่ละช่วง จากรูปที่ 2.10 จะเห็นว่า

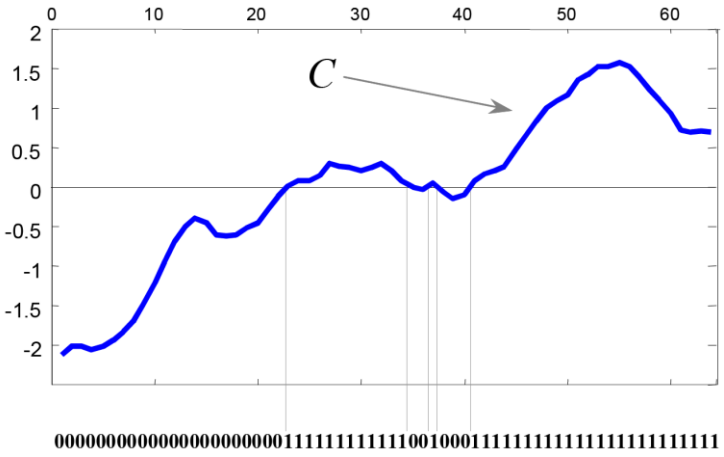
อนุกรมเวลาความยาว 128 จุด ถูกแบ่งเป็น 8 ช่วงโดยแต่ละช่วงมีความยาวของช่วงข้อมูล 16 จุด โดยในแต่ละช่วงจะนำค่าข้อมูลแต่ละจุดไปหาค่าเฉลี่ยเพื่อเป็นตัวแทนของช่วงข้อมูลนั้น ทำให้เหลืออนุกรมเวลาความยาว 8 จุด ซึ่งทำให้ช่วยในการทำงานได้เร็วขึ้นอย่างมาก แต่ในความเป็นจริงการเลือกความยาวหรือมิติข้อมูลที่เหมาะสมกับอนุกรมเวลาเป็นเรื่องที่มีความสำคัญ เนื่องจากถ้าหากเลือกมิติที่มากเกินไปก็จะทำให้ลดระยะเวลาไม่ได้มาก ในขณะเดียวกันถ้าเลือกมิติน้อยเกินไปก็อาจส่งผลต่อความแม่นยำของผลลัพธ์



รูปที่ 2.10 การแทนข้อมูลแบบพีเอเอ

(ที่มา : Yi และคณะ [6])

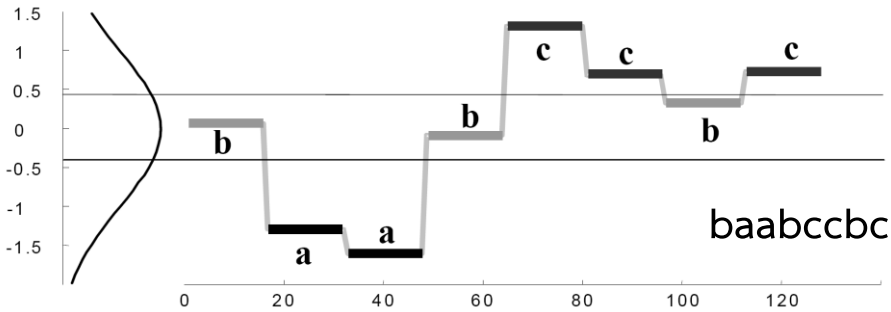
Bagnall และ Lornadi (2005) [16] ได้นำเสนอการแทนข้อมูลแบบคลิป (Clipped Representation) โดยเป็นการเปลี่ยนข้อมูลจากเลขจำนวนจริงให้เป็นเลขฐานสอง (Binary) โดยเกณฑ์ในการเปลี่ยนคือหาค่าเฉลี่ยจากข้อมูลอนุกรมเวลาจากนั้นถ้าจุดข้อมูลใดมากกว่าค่าเฉลี่ยจะมีค่าเป็น 1 และจุดข้อมูลใดมีค่าน้อยกว่าหรือเท่ากับค่าเฉลี่ยจะมีค่าเป็น 0 ดังตัวอย่างในรูปที่ 2.11 ซึ่งถึงแม้วิธีนี้จะเป็นการใช้เลขฐานสองเพื่อลดขนาดข้อมูล แต่การเปลี่ยนข้อมูลด้วยค่าเฉลี่ยนั้นอาจส่งผลให้ผลลัพธ์มีความผิดพลาดได้ง่ายในกรณีที่ข้อมูลมีโครงสร้างโดยรวมเหมือนกัน กล่าวคือถ้าข้อมูลมีการขึ้นลงผ่านค่าเฉลี่ยเท่า ๆ กันแต่ลักษณะของเส้นอนุกรมเวลาต่างกัน หลังจากผ่านการแทนข้อมูลแบบคลิปจะทำให้ได้รูปแบบข้อมูลที่เหมือนกัน ในขณะที่อนุกรมเวลานั้นมีความแตกต่างกัน



รูปที่ 2.11 การแทนข้อมูลแบบคลิป

(ที่มา : Bagnall และคณะ [16])

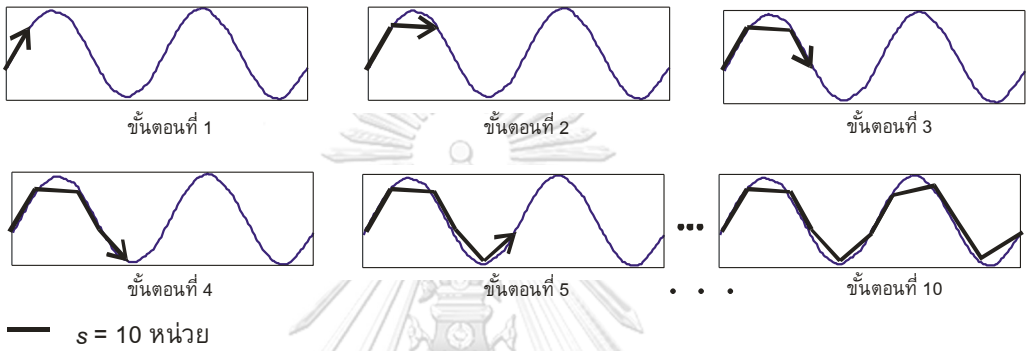
Lin และ คณะ (2007) [5] ได้เสนอการแทนข้อมูลอีกรูปแบบหนึ่งคือ การแทนข้อมูลแบบแซคซ์ โดยขั้นแรกใช้การแทนข้อมูลแบบพีเอเอเพื่อช่วยในการลดมิติของข้อมูล จากนั้นจะใช้ค่าเฉลี่ยที่ได้จากพีเอเอเพื่อแบ่งเป็นช่วงข้อมูลแล้วแทนด้วยสัญลักษณ์ ซึ่งการแบ่งช่วงข้อมูลนั้นถูกแบ่งจากการแจกแจงแบบเกาส์เซียน (Gaussian Distribution) และในการคำนวณในแต่ละครั้งจะต้องมีการทำซีนอร์มัลไลเซชัน (Z-Normalization) จากรูปที่ 2.12 จะเห็นว่าหลังจากการแทนที่ข้อมูลแบบพีเอเอในรูปที่ 2.10 แล้วจะถูกนำมาแบ่งช่วงของค่าเฉลี่ยด้วยการแจกแจงแบบเกาส์เซียน ซึ่งถึงแม้จะลดข้อมูลจากตัวเลขเป็นสัญลักษณ์ แต่ก็ยังคงต้องกำหนดจำนวนมิติที่เหมาะสมเช่นเดียวกับการแทนข้อมูลแบบพีเอเอ นอกจากนี้ยังต้องกำหนดพารามิเตอร์ของสัญลักษณ์อีกด้วยว่าเท่าใดจึงเหมาะสม ซึ่งการกำหนดพารามิเตอร์ของสัญลักษณ์ที่ไม่เหมาะสมก็ย่อมส่งผลต่อความแม่นยำไม่ต่างจากการกำหนดจำนวนมิติ



รูปที่ 2.12 การแทนที่ข้อมูลแบบแซคซ์

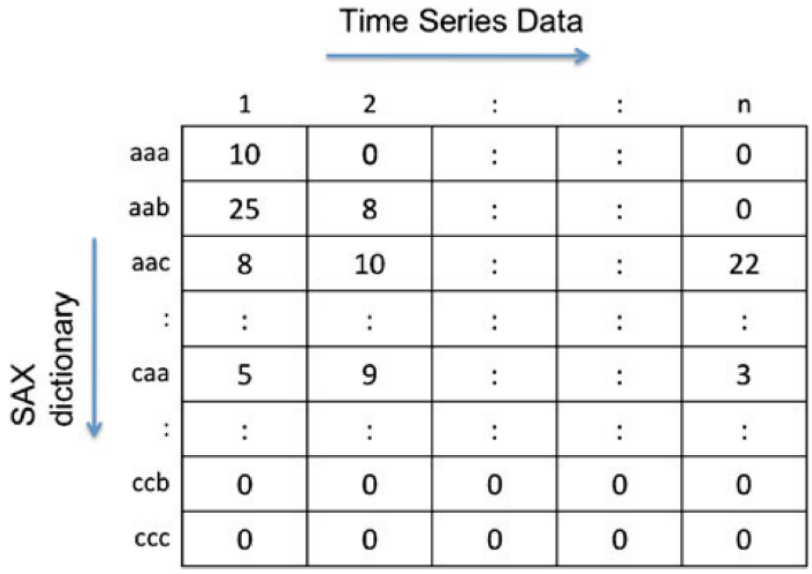
(ที่มา : Lin และคณะ [5])

Sajjipanon และ Ratanamahatana (2009) [17] นำเสนอการแทนข้อมูลแบบแฟรคทัล (Fractal Representation) โดยเป็นการลดจำนวนมิติให้เหลือเพียง 2 มิติ โดยใช้มิติเส้นขอบ (Compass Dimension) ในการแทนที่ข้อมูล ซึ่งถึงแม้จะมีการลดมิติข้อมูลแล้วก็ตาม แต่ผู้วิจัยแสดงให้เห็นว่าวิธีการนี้เหมาะสมกับข้อมูลที่มีความยาวมาก ๆ เท่านั้น และการทดสอบความแม่นยำเป็นการทดสอบจากข้อมูลที่เป็นคนละประเภทกัน แต่ไม่มีการทดลองกับข้อมูลประเภทเดียวกันแต่คนละกลุ่ม ซึ่งอาจจะไม่ได้ให้ผลลัพธ์ที่ดีเท่าเดิม



รูปที่ 2.13 แสดงการคำนวณค่ามิติเส้นขอบสำหรับการแทนข้อมูลแบบแฟรคทัล
(ที่มา : Sajjipanon และคณะ [17])

Lin และ คณะ (2012) [7] นำเสนอการแทนข้อมูลด้วยการใช้เทคนิคกระเป๋าคำ (Bag of Words) โดยจะทำการแทนข้อมูลแบบแซคซ์ก่อน จากนั้นเมื่อได้ข้อมูลอนุกรมเวลาในรูปแบบของสัญลักษณ์แล้ว จะทำการนับกลุ่มคำที่ซ้ำกันเพื่อนำไปเป็นข้อมูลของอนุกรมเวลานั้น ๆ ตัวอย่างจากรูปที่ 2.14 เป็นลักษณะการใช้เทคนิคกระเป๋าคำในการแทนข้อมูล โดยแนวตั้งคือข้อมูลอนุกรมเวลาและแนวนอนคือลักษณะของกลุ่มคำที่มีในอนุกรมเวลานั้น ๆ ซึ่งจะสังเกตเห็นว่ามีการนับกลุ่มคำว่ามีกลุ่มคำลักษณะใดเกิดขึ้นบ้างและเกิดขึ้นมากน้อยเพียงใด ซึ่งนอกจากจะทำให้เห็นความแตกต่างระหว่างอนุกรมเวลาได้ชัดเจนยังสามารถลดมิติข้อมูลได้มากขึ้นด้วย แต่เทคนิคกระเป๋าคำนี้จะทำให้เห็นความแตกต่างอย่างชัดเจนได้เมื่อข้อมูลอนุกรมเวลามีความยาวระดับหนึ่ง ถ้าหากข้อมูลอนุกรมเวลามีขนาดสั้นเกินไปจะทำให้กลุ่มคำที่เกิดไม่หลากหลาย ทำให้ผลลัพธ์ที่ได้ไม่มีประสิทธิภาพเท่าที่ควร

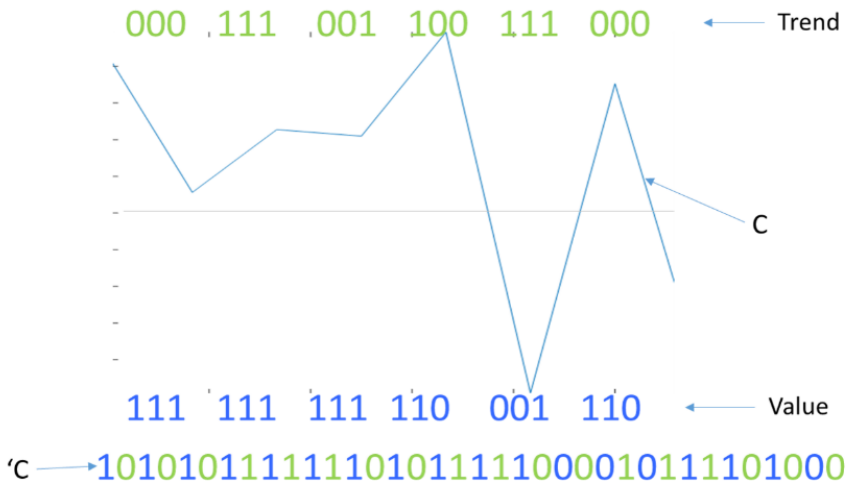


	1	2	:	:	n
aaa	10	0	:	:	0
aab	25	8	:	:	0
aac	8	10	:	:	22
:	:	:	:	:	:
caa	5	9	:	:	3
:	:	:	:	:	:
ccb	0	0	0	0	0
ccc	0	0	0	0	0

รูปที่ 2.14 ตัวอย่างข้อมูลของอนุกรมเวลาหลังจากใช้เทคนิคกระเป๋าคำ

(ที่มา : Lin และคณะ [7])

A. Kane (2017) นำเสนอการแทนข้อมูลรูปแบบใหม่ โดยนอกจากจะใช้ค่าข้อมูลของอนุกรมเวลาแล้วยังสังเกตจากลักษณะความเป็นไปของอนุกรมเวลา (Trend) โดยในการสังเกตความเป็นไปจะใช้ค่าข้อมูลจากการแทนข้อมูลแบบพีเอเอ แล้วสังเกตว่าจุดในอนุกรมเวลามีค่าเพิ่มขึ้นหรือลดลง ถ้าเพิ่มขึ้นจากจุดเวลาก่อนหน้าจะให้ค่าเป็น 1 ในขณะที่ถ้าหากค่าอนุกรมเวลาลดลงจากจุดเวลาก่อนหน้าจะให้ค่าเป็น 0 สำหรับค่าข้อมูลจะใช้ลักษณะของการแทนข้อมูลแบบคลิป คือเมื่อจุดข้อมูลใดมากกว่าค่าเฉลี่ยจะให้ค่าเป็น 1 และเมื่อจุดข้อมูลใดน้อยกว่าหรือเท่ากับค่าเฉลี่ยจะให้ค่าเป็น 0 ดังแสดงในรูปที่ 2.15 เพราะฉะนั้นอนุกรมเวลาใหม่จะเกิดจากการนำค่าข้อมูลหนึ่งตำแหน่งสลับกับค่าความเป็นไปหนึ่งตำแหน่ง ทำให้ข้อมูลใหม่มีความยาวเป็นสองเท่าจากความยาวเดิมของการแทนข้อมูลแบบพีเอเอ ซึ่งแม้ว่าวิธีการดังกล่าวจะมีความยาวเพิ่มเป็นสองเท่า แต่ลดขนาดของข้อมูลโดยตรงโดยใช้เลขฐานสอง



รูปที่ 2.15 การแทนข้อมูลโดยการดูความเป็นไปของอนุกรมเวลา
(ที่มา : A. Kane และคณะ [18])

ซึ่งจะเห็นว่างานวิจัยที่กล่าวไปนั้นถึงแม้จะมีการลดขนาดข้อมูลแต่ก็ยังคงมีขนาดมิติข้อมูลเท่าเดิมหรือเพิ่มขึ้น ในขณะที่งานวิจัยที่มีการลดมิติข้อมูลก็ยังไม่สามารถทำให้มิติข้อมูลมีขนาดเท่ากันได้สำหรับทุกชุดข้อมูล อีกทั้งหลายงานวิจัยยังมีความจำเป็นที่ต้องทำการหาพารามิเตอร์ที่เหมาะสมในแต่ละชุดข้อมูล ซึ่งทำให้มีความลำบากต่อการใช้งานจริงสำหรับผู้ใช้โดยทั่วไป

นอกจากนี้ความเร็วในการทำงานก็เป็นสิ่งสำคัญ ซึ่งการใช้ข้อมูลทั้งหมดจากอนุกรมเวลาเดิมจะทำให้เสียเวลามากกว่าเมื่อเปรียบเทียบกับการใช้ข้อมูลเพียงบางส่วน ดังนั้นผู้วิจัยจึงมีความพยายามที่จะพัฒนาการแทนข้อมูลที่มีประสิทธิภาพ โดยสามารถลดมิติข้อมูลให้เหลือน้อยที่สุดและมีขนาดเท่ากันสำหรับทุกชุดข้อมูล ซึ่งในการลดมิติข้อมูลนั้นจะใช้เพียงส่วนใดส่วนหนึ่งจากอนุกรมเวลา

บทที่ 3 การแทนข้อมูลบางส่วน (Partial Representation)

3.1 แนวคิดในการลดมิติด้วยการแทนข้อมูลบางส่วน

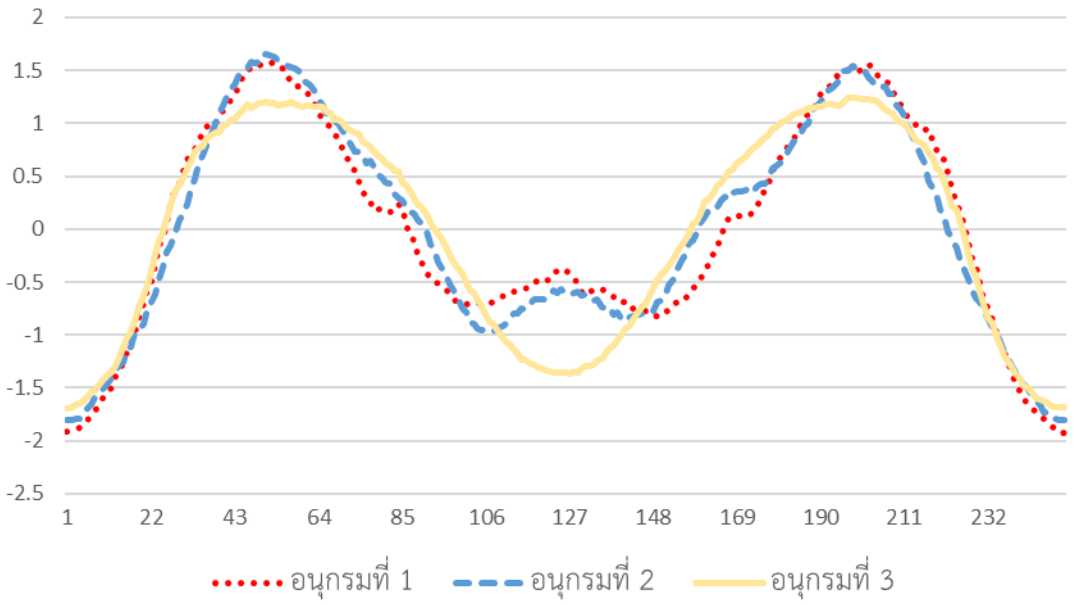
เทคนิคการแทนข้อมูลสำหรับข้อมูลอนุกรมเวลาส่วนใหญ่ มักจะใช้ข้อมูลอนุกรมเวลาทั้งหมดเพื่อสะท้อนลักษณะสำคัญของอนุกรมเวลานั้น ๆ ซึ่งในการที่จะต้องพิจารณาข้อมูลทั้งหมดจะทำให้เสียเวลาในการทำงานเพิ่มขึ้น ดังนั้นผู้วิจัยจึงมีแนวคิดที่จะใช้ข้อมูลเพียงบางส่วนจากอนุกรมเวลา

จากงานวิจัยหนึ่งที่มีการนำเสนอวิธีการที่ชื่อว่า เชปเล็ท ซึ่งเป็นการหาส่วนหนึ่งของอนุกรมเวลาเพื่อเป็นตัวแทนในการแสดงถึงข้อมูลกลุ่มใดกลุ่มหนึ่ง จากลักษณะของเชปเล็ทดังกล่าวจะเห็นว่าส่วนของเชปเล็ทควรจะเป็นส่วนข้อมูลที่สำคัญสำหรับกลุ่มข้อมูลนั้น ๆ ซึ่งสามารถที่จะสะท้อนคุณลักษณะของกลุ่มข้อมูลในแต่ละกลุ่มได้ ด้วยคุณสมบัติของเชปเล็ทที่สามารถแสดงถึงอนุกรมเวลาใดโดยใช้ข้อมูลเพียงส่วนใดส่วนหนึ่ง ทำให้ผู้วิจัยมีความคิดว่า หากมีเชปเล็ทที่สามารถแสดงถึงกลุ่มข้อมูลแต่ละกลุ่มได้อย่างมีประสิทธิภาพแล้ว ก็ไม่มีความจำเป็นที่จะต้องพิจารณาอนุกรมเวลาทั้งหมด

จากสมมติฐานดังกล่าวผู้วิจัยจึงลองสังเกตถึงลักษณะของข้อมูลอนุกรมเวลาหลากหลายประเภท จากรูปที่ 3.1 เป็นรูปภาพของหัวลูกศรที่ถูกแปลงไปเป็นข้อมูลอนุกรมเวลาชื่อว่า ArrowHead [12] ซึ่งจะเห็นว่าลักษณะของลูกศรทั้งสามแบบมีความแตกต่างกันค่อนข้างชัดเจน ทั้งในด้านความยาว ความกว้าง และลักษณะมุมของหัวลูกศร ด้วยความต่างเหล่านี้จะถูกสะท้อนไปยังข้อมูลอนุกรมเวลาที่แปลงมาจากภาพหัวลูกศร ซึ่งความต่างที่เกิดขึ้นนี้จะสามารถทำให้เราเลือกข้อมูลบางส่วนเพื่อสะท้อนข้อมูลอนุกรมเวลานี้ได้

รูปที่ 3.1 ภาพตัวอย่างหัวลูกศร 3 แบบ จากข้อมูล ArrowHead
(ที่มา : Chen และคณะ [12])

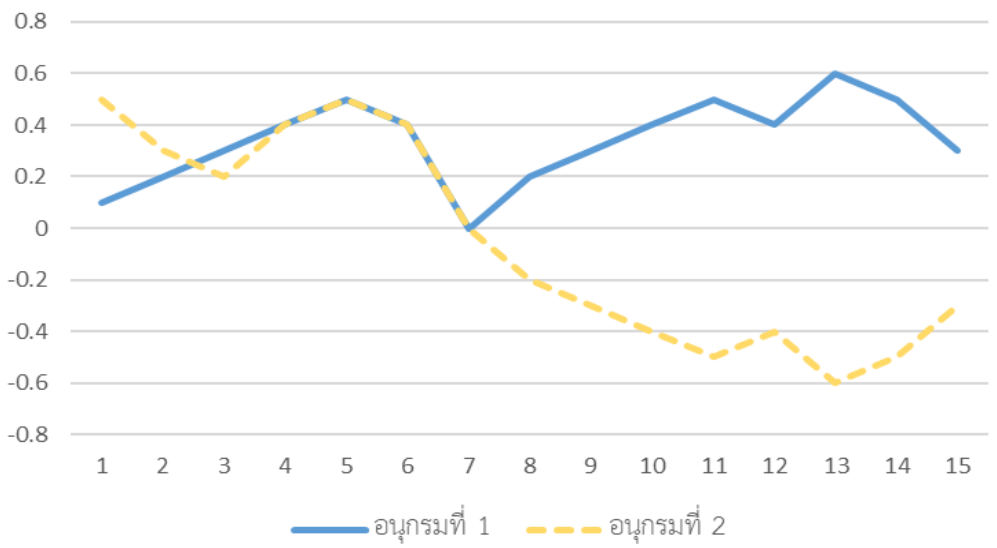
ในรูปที่ 3.2 เป็นภาพอนุกรมเวลาที่แปลงมาจากภาพหัวลูกศรในรูปที่ 3.1 ซึ่งจะสังเกตเห็นว่าแต่ละอนุกรมเวลาจะมีบางช่วงเวลาที่มีค่าแตกต่างกันอย่างชัดเจน โดยถ้าเราสามารถเลือกจุดหรือช่วงที่มีความแตกต่างนี้ได้ ก็ย่อมสามารถหาความแตกต่างระหว่างอนุกรมเวลาได้เช่นเดียวกัน เช่น ใช้การแทนข้อมูลโดยเลือกจุด 2 จุด จากอนุกรมเวลาเพื่อนำมาหาระยะทางด้วยค่าสัมบูรณ์ของผลต่างระหว่างจุดสองจุด โดยถ้าเลือกจุดที่ 127 และจุดที่ 190 จะได้ระยะทางคือ 1.6, 1.8 และ 2.6 ตามลำดับ จะเห็นว่าระยะทางดังกล่าวสามารถแยกความแตกต่างของอนุกรมเวลาทั้งสามได้ ซึ่งเป็นผลมาจากการเลือกจุดที่สามารถสะท้อนความแตกต่างของอนุกรมเวลา ขณะเดียวกันถ้าหาเลือกจุดที่ 64 และจุดที่ 190 จะได้ระยะทางคือ 0 สำหรับทุกอนุกรมเวลา ซึ่งไม่สามารถแยกความแตกต่างระหว่างอนุกรมเวลาทั้งสามได้ สังเกตได้ว่าการเลือกจุดหรือส่วนของอนุกรมเวลาที่ดี จะสามารถแยกอนุกรมเวลาทั้งสามออกจากกันได้โดยไม่จำเป็นต้องพิจารณาข้อมูลทั้งหมด



รูปที่ 3.2 ข้อมูลอนุกรมเวลาหัวลูกศร (ArrowHead) ที่ถูกแปลงมาจากหัวลูกศรทั้ง 3 แบบ

อย่างไรก็ตามในข้อมูลบางประเภทก็มีความจำเป็นที่ต้องเลือกคู่จุดข้อมูลหลายคู่ เพราะในหลาย ๆ ประเภทข้อมูลไม่สามารถหาจุดเพียงคู่เดียวที่สร้างความแตกต่างอย่างชัดเจนได้ จึงจำเป็นที่ต้องเลือกหลายคู่จุดข้อมูลเพื่อนำมาประกอบกันให้เกิดความแตกต่างได้อย่างชัดเจนยิ่งขึ้น และยังคงมีข้อมูลบางประเภทที่แต่ละกลุ่มข้อมูลมีลักษณะของอนุกรมเวลาที่ใกล้เคียงกันมาก ซึ่งไม่ว่าจะแบ่งคู่จุดข้อมูลอย่างไรก็ไม่สามารถหาค่าระยะทางที่แสดงความแตกต่างได้อย่างชัดเจนได้ ทำให้ค่าความแม่นยำของผลลัพธ์ไม่ดีเท่าที่ควร ดังนั้นการเลือกจุดหรือช่วงที่ดีจึงเป็นปัจจัยสำคัญในการสะท้อนความแตกต่างของข้อมูลอนุกรมเวลา

นอกจากปัญหาที่เกิดจากการเลือกจุดหรือช่วงดังที่กล่าวไปแล้ว ก็ยังคงมีปัญหาอันเนื่องมาจากวิธีการที่ใช้ในการคิดเพื่อหาค่าที่แสดง อย่างเช่นก่อนหน้านี้ที่ใช้การหาระยะทางระหว่างจุดสองจุด ในบางครั้งการหาค่าด้วยระยะทางก็ไม่สามารถสะท้อนถึงความแตกต่างได้ถึงแม้จะแบ่งได้ดีแล้วก็ตาม ดังตัวอย่างในรูปที่ 3.3 หากเลือกจุดที่ 7 และจุดที่ 13 จะสังเกตเห็นได้ว่าสำหรับอนุกรมที่ 1 และอนุกรมที่ 2 จะมีระยะทางเท่ากับ 0.6 เมื่อคำนวนระยะทางด้วยค่าสัมบูรณ์ของผลต่างระหว่างจุดสองจุด ซึ่งจะไม่สามารถแยกความแตกต่างของอนุกรมเวลาทั้งสองได้ แต่ถ้าหากเปลี่ยนจากการหาระยะทางดังกล่าวเป็นการหาผลต่างระหว่างจุดสองจุดจะได้ค่าเท่ากับ 0.6 และ -0.6 ตามลำดับ ซึ่งจะทำให้เกิดความแตกต่างระหว่างอนุกรมทั้งสองอย่างชัดเจน



รูปที่ 3.3 ตัวอย่างอนุกรมเวลาสองอนุกรมที่ทำให้การหาระยะทางไม่สามารถแยกความแตกต่างระหว่างสองอนุกรมนี้ได้ ถึงแม้จะมีการแบ่งที่ดีก็ตาม

จากเหตุผลด้านการเลือกจุดหรือช่วงและวิธีการในการหาค่าความต่างที่กล่าวไป ถ้าสามารถหาวิธีการเลือกจุดหรือช่วงที่มีประสิทธิผลและการหาค่าความต่างได้อย่างเหมาะสมแล้ว จะสามารถใช้ค่าที่ได้มานี้ในการสร้างการแทนข้อมูลโดยเป็นการใช้ข้อมูลเพียงบางส่วนจากอนุกรมเวลา ซึ่งจะสามารถเพิ่มความเร็วในการทำงานจากการที่ใช้ข้อมูลเพียงบางส่วนนี้ได้ นอกเหนือจากความเร็วที่เกิดจากการลดมิติข้อมูลหลังจากการแทนข้อมูลแล้ว

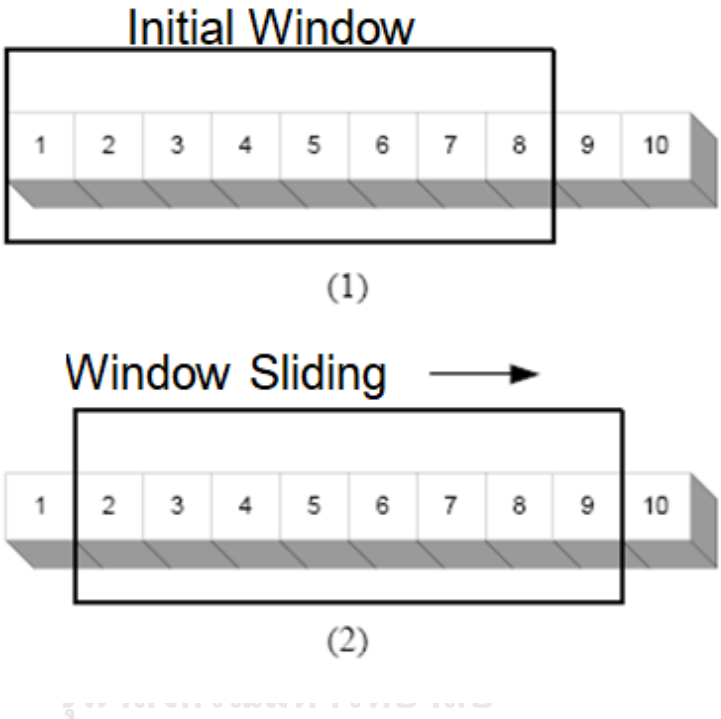
3.2 การแทนข้อมูลบางส่วน (Partial Representation)

จากแนวคิดที่กล่าวไปข้างต้นแสดงให้เห็นว่าการใช้ข้อมูลอนุกรมเวลาเพียงบางส่วน น่าจะสามารถสร้างข้อมูลใหม่ที่แทนข้อมูลเดิมได้ ดังนั้นการใช้เชปเล็ทจึงเป็นแนวทางที่ดีที่จะนำมาใช้ในการแทนข้อมูลเนื่องจากเชปเล็ทเป็นส่วนใดส่วนหนึ่งของอนุกรมเวลา ที่สามารถแสดงถึงลักษณะดั้งเดิมของอนุกรมเวลานั้น ๆ ได้ แต่การใช้เพียงแค่เชปเล็ทนั้นก็ไม่เพียงพอในการลดมิติข้อมูลให้มีขนาดน้อยที่สุด ดังนั้นผู้วิจัยจึงมีแนวคิดที่จะใช้ข้อมูลจากเชปเล็ทเพื่อสร้างข้อมูลใหม่ โดยถ้าหากทำการลากเส้นตรงให้อยู่ตรงกลางระหว่างทุกจุดข้อมูลของเชปเล็ท จะทำให้สามารถสร้างเส้นตรงใหม่ที่แสดงถึงแนวโน้มของเชปเล็ทนั้นได้ ซึ่งค่าความชันที่ได้จากเส้นตรงดังกล่าวคือสิ่งที่สามารถบอกถึงลักษณะและทิศทางของเส้นตรง ทำให้น่าจะสามารถใช้ค่านี้ในการแทนข้อมูลอนุกรมเวลาเดิมได้ แต่เนื่องจากการลากเส้นตรงเพื่อหาความชันนั้น ถ้าหากนำมาใช้กับเชปเล็ทอาจจะไม่ได้ให้ผลลัพธ์ที่ดีเหมือนตัวเชปเล็ทเอง และยังทำให้เสียเวลาในการหาเชปเล็ทที่ค่อนข้างมากอีกด้วย ดังนั้นผู้วิจัยจึงเล็งเห็นว่าถ้าหากใช้ความชันจากเส้นตรงที่ลากผ่านกลางระหว่างลำดับย่อยแทนเชปเล็ท ก็ย่อมที่จะสามารถลดขนาดมิติข้อมูลและลดระยะเวลาในการทำงานได้เป็นอย่างมาก โดยวิธีการแทนข้อมูลบางส่วนอธิบายได้ดังนี้

3.2.1 การหาเชปเล็ท

การหาเชปเล็ทจากอนุกรมเวลาใด ๆ นั้นจะเริ่มจากการแบ่งอนุกรมเวลาออกเป็นหลาย ๆ ส่วน จากนั้นนำแต่ละส่วนไปทดสอบความสามารถในการแบ่งแยกข้อมูลด้วยฟังก์ชันทดสอบ ลำดับย่อยใดที่สามารถแบ่งแยกข้อมูลสองกลุ่มออกจากกันได้มากที่สุดจะเป็นส่วนที่ถูกเลือก ซึ่งกลายเป็นเชปเล็ทที่สามารถแสดงถึงอนุกรมเวลาเดิมได้ โดยก่อนที่จะทำการแบ่งอนุกรมเวลานั้นจำเป็นที่จะต้องรู้ถึงขนาดของเชปเล็ทที่เหมาะสมสำหรับแต่ละข้อมูลก่อน ดังนั้นในการหาความยาวที่เหมาะสม จึงมีความจำเป็นที่จะต้องทำการทดสอบในหลาย ๆ ความยาว ซึ่งจะทำให้เสียเวลาในการหาเป็นอย่างมาก นอกจากนี้ในส่วนของการแบ่งอนุกรมเวลาหลังจากทราบความยาวที่เหมาะสมแล้ว เพื่อให้ได้เชปเล็ทที่เหมาะสมและครบถ้วน

จึงจำเป็นที่จะต้องแบ่งเซปเล็ทด้วยวิธีการหน้าต่างบานเลื่อน (Sliding Window) ซึ่งขนาดของหน้าต่างคือความยาวของเซปเล็ทที่ได้จากการทดสอบ โดยในการแบ่งด้วยวิธีการหน้าต่างบานเลื่อนนั้นจะทำการสร้างหน้าต่างตามขนาดที่กำหนด จากนั้นนำไปครอบโดยเริ่มตั้งแต่จุดแรกของอนุกรมเวลา จะทำให้ได้ลำดับย่อยของอนุกรมเวลาออกมาจากนั้นจะทำการเลื่อนหน้าต่างไปหนึ่งตำแหน่ง ซึ่งจะทำให้ได้ลำดับย่อยของอนุกรมเวลาแบบใหม่ดังรูปที่ 3.4



รูปที่ 3.4 วิธีการหน้าต่างบานเลื่อน

(ที่มา : http://wiki.treck.com/Introduction_to_TCP/IP)

จะเห็นว่านอกจากการหาความยาวที่เหมาะสมจะทำให้เสียเวลาแล้ว การหาลำดับย่อยด้วยวิธีการหน้าต่างบานเลื่อนก็ทำให้เสียเวลามากเช่นเดียวกัน นอกจากนี้ยังมีความจำเป็นที่จะต้องนำลำดับย่อยที่ทำการแบ่งไว้แล้ว ไปตรวจสอบความสามารถในการแบ่งกลุ่มข้อมูลเพื่อหาเซปเล็ท ซึ่งจะทำให้เสียเวลามากขึ้นเป็นอย่างมาก

จากปัญหาดังกล่าวผู้วิจัยจึงจะทำการหาลำดับย่อยด้วยการแบ่งอนุกรมเป็นส่วน ๆ โดยไม่ใช้วิธีการหน้าต่างบานเลื่อน โดยผู้วิจัยจะทำการหาความยาวที่เหมาะสมสำหรับข้อมูลนั้น ๆ จากนั้นจะนำความยาวดังกล่าวมาใช้เป็นหน้าต่างแต่จะไม่ทำการเลื่อน แต่จะทำการนำหน้าต่างใหม่ที่มีขนาดเท่าเดิมมาวางต่อกันไปเรื่อย ๆ จนสุดอนุกรมเวลา ตัวอย่างเช่น ถ้า

หากมีอนุกรมเวลายาว 500 จุด และขนาดหน้าต่างที่เหมาะสมคือ 5 เมื่อทำการวางหน้าต่างใหม่เรื่อย ๆ จะสามารถได้ลำดับย่อยของอนุกรมเวลาออกมาทั้งหมด 100 ส่วน ซึ่งจะลดเวลาลงมากเมื่อเทียบกับวิธีการหน้าต่างบานเลื่อน และเนื่องจากผู้วิจัยใช้การลากเส้นตรงเพื่อหาความชัน ดังนั้นจึงไม่มีความจำเป็นที่ผู้วิจัยจะต้องหาเซปเล็ทจากลำดับย่อยที่ได้มา เนื่องจากค่าความชันที่ได้มาจากเซปเล็ทนั้นอาจจะไม่ใช่ค่าที่สามารถแบ่งข้อมูลได้ดีที่สุดเหมือนเซปเล็ท ดังนั้นจึงไม่มีความจำเป็นที่จะต้องหาเซปเล็ท โดยผู้วิจัยจะทำการลากเส้นตรงจากลำดับย่อยที่ได้มาทันที ซึ่งจะลดเวลาในการหาเซปเล็ทออกไปได้เป็นอย่างมาก

3.2.2 การลากเส้นตรง

หลังจากทำการแบ่งอนุกรมเวลาเป็นลำดับย่อย ๆ สำหรับทุกลำดับนั้นจะทำการลากเส้นตรงผ่านกลางระหว่างจุดข้อมูล ซึ่งวิธีที่จะใช้ในการลากเส้นตรงนี้คือ วิธีกำลังสองน้อยที่สุด จากสมการที่ (3.1) จะเป็นสมการเส้นตรงสำหรับวิธีกำลังสองน้อยที่สุดโดยกำหนดให้ $\hat{y}$ คือค่าคาดคะเนของตัวแปรตาม a คือค่าจุดตัดแกนตั้ง b คือค่าความชัน และ x คือค่าตัวแปรอิสระ

$$\hat{y} = a + bx \tag{3.1}$$

ในการที่จะลากเส้นตรงใหม่จากข้อมูลที่มีนั้นจำเป็นที่จะต้องทราบถึงค่าจุดตัดแกนตั้งและค่าความชันก่อน โดยการหาค่าความชันสำหรับสมการเส้นตรงนั้นหาได้จากสมการที่ (3.2) โดย x_i คือตำแหน่งของจุดข้อมูล i บนเส้นอนุกรมเวลา y_i คือค่าข้อมูลของจุดข้อมูล i บนเส้นอนุกรมเวลา $\bar{x}$ คือค่าเฉลี่ยของตำแหน่งสำหรับทุกจุดข้อมูล i โดยมี $\bar{y}$ คือค่าเฉลี่ยข้อมูลสำหรับทุกจุดข้อมูล i และ n คือจำนวนมิติหรือความยาวของอนุกรมเวลา

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.2}$$

เนื่องจากในการแทนข้อมูลแบบใหม่จะใช้เพียงแค่ค่าความชัน ดังนั้นจึงไม่มีความจำเป็นที่จะต้องคำนวณค่าจุดตัดแกนตั้ง เพราะฉะนั้นเมื่อมีอนุกรมเวลาความยาว n เข้ามาจะถูกทำการแบ่งเป็นลำดับย่อยด้วยความยาว w ซึ่งจะได้จำนวนลำดับย่อยทั้งหมด n / w ส่วน จากนั้นสำหรับทุก ๆ ลำดับย่อยจะถูกนำมาลากเส้นตรงใหม่ด้วยสมการกำลังสองโดยจะใช้เพียงแค่ค่าความชัน b ซึ่งหมายความว่าจะได้ค่าความชันทั้งหมด n / w ส่วน

3.2.3 การเลือกจำนวนมิติและค่าความชัน

ในการพิจารณาจำนวนมิตินั้น เนื่องจากผู้วิจัยต้องการลดมิติข้อมูลให้มีขนาดน้อยที่สุด ดังนั้นจากการใช้ค่าความชันผู้วิจัยสามารถลดค่ามิติให้เหลือ 1 มิติได้โดยการเลือกความชันค่าความชันต่ำสุดหรือค่าความชันสูงสุดออกมาจากค่าความชันทั้งหมด ซึ่งในการเลือกค่าความชันค่าใดค่าหนึ่งออกมานั้นจำเป็นที่จะต้องตรวจสอบค่าความชันทั้งหมดก่อน ดังนั้นถ้าหากทำการเลือกค่าความชันออกมาทั้งสองค่าคือค่าความชันสูงสุดและค่าความชันต่ำสุด ซึ่งถึงแม้จะมีมิติที่เพิ่มขึ้นเป็น 2 มิติ แต่จะใช้เวลาในการตรวจหาค่าความชันเท่ากันหรือใกล้เคียงกับการลดค่ามิติให้เหลือ 1 มิติ และนอกจากนี้การเลือกค่าสูงสุดและต่ำสุดของความชันนั้นย่อมจะสามารถบอกถึงลักษณะของอนุกรมเวลาได้ดีกว่าการบอกด้วยค่าความชันสูงสุดหรือต่ำสุดเพียงค่าเดียว

แต่จะเห็นว่าถึงแม้จะพิจารณาอนุกรมเวลาเป็นลำดับย่อยแล้วก็ตามแต่ก็ต้องพิจารณาทุกจุดเวลาของลำดับย่อย หมายความว่าจะต้องพิจารณาทั้งอนุกรมเวลาเดิมที่ความยาว n ซึ่งจะไม่ได้ลดเวลาการทำงานลงไปเมื่อเทียบกับการวัดระยะทางแบบยุคลิด ดังนั้นผู้วิจัยจะใช้เพียงแค่จุดเริ่มต้นและจุดจบของลำดับย่อย ซึ่งจะทำให้ลดเวลาในการทำงานของการลากเส้นตรงลงไปเป็นอย่างมาก

3.3 รหัสเทียมการแทนข้อมูลบางส่วน

ขั้นตอนในการแทนข้อมูลบางส่วนสามารถอธิบายได้ตามรหัสเทียมในรูปที่ 3.5 ในบรรทัดที่ 1 จะวนสำหรับทุก ๆ อนุกรมเวลา ในบรรทัดที่ 2 และ 3 จะเป็นการระบุความยาวของลำดับย่อย `Length` โดยใช้ความยาวของอนุกรม `Length_TS` เวลาหารด้วยจำนวนลำดับย่อย `Subseq_Num` ซึ่งจำนวนลำดับย่อยนั้นในกรณีที่ต้องการให้ได้ความแม่นยำสูงสุดจำเป็นที่จะต้องทำการเรียนรู้กับข้อมูลเรียนรู้ แต่ผลลัพธ์จากการเรียนรู้นั้นไม่ได้แตกต่างกับการแบ่งด้วยจำนวนลำดับย่อยค่าเดียว ดังนั้นผู้วิจัยจึงระบุให้ใช้ลำดับย่อยทั้งหมด 4 ส่วนสำหรับทุกอนุกรมเวลา เนื่องจากจะสามารถแบ่ง

อนุกรมเวลาออกเป็นลำดับย่อยที่มีความยาว 25 เปอร์เซ็นต์ได้อย่างลงตัวและมีจำนวนลำดับย่อยน้อยที่สุด ในบรรทัดที่ 4 ถึง 8 จะเป็นการระบุตำแหน่งจุดเริ่มต้นของลำดับย่อย `Start_Pos` และจุดสุดท้ายของลำดับย่อย `Stop_Pos` โดยมี `Count_Loop` เป็นเงื่อนไขที่ใช้ในการหยุดสำหรับการวนแต่ละลำดับย่อย ส่วนค่า `Max_Slope` และ `Min_Slope` เป็นตัวแปรที่ไว้ใช้เก็บค่าความชันสูงสุดและต่ำสุดของลำดับย่อยที่ถูกแบ่งไว้

สำหรับบรรทัดที่ 9 ถึง 24 จะเป็นลูปที่ใช้ในการวนสำหรับทุก ๆ ลำดับย่อย โดยในทุก ๆ ลำดับย่อยจะทำการค้นหาค่าความชันจาก `FindSlope` ซึ่งหาได้จากสมการที่ (3.2) โดยใช้ค่าเฉลี่ยตำแหน่งข้อมูล `X_bar` ของตำแหน่งจุดข้อมูลเริ่มต้น `Start_Pos` และตำแหน่งจุดข้อมูลสุดท้าย `Stop_Pos` และค่าเฉลี่ยข้อมูล `Y_bar` ของค่าข้อมูลเริ่มต้น `Start_Value` และค่าข้อมูลสุดท้าย `Stop_Value` ในส่วนของบรรทัดที่ 13 ถึง 17 จะเป็นการเก็บค่าความชันที่มีค่าสูงสุด `Max_Slope` และต่ำสุด `Min_Slope` เพื่อนำไปเป็นค่าใหม่ที่ใช้แทนอนุกรมเวลาเดิม ในบรรทัดที่ 18 ถึง 22 จะเป็นการขยับจุดเริ่มต้นของลำดับย่อยและจุดสุดท้ายของลำดับย่อยให้เป็นลำดับย่อยใหม่ ซึ่งในบางกรณีที่ความยาวของอนุกรมเวลาไม่สามารถหารลงตัวกับจำนวนส่วนย่อยที่กำหนดได้ จำเป็นที่จะต้องทำให้ลำดับย่อยอันสุดท้ายมีความยาวแตกต่างจากเดิมเพื่อให้สามารถแบ่งลำดับย่อยได้ครอบคลุมทั้งอนุกรมเวลา โดยจะให้จุดสุดท้ายของลำดับย่อยสุดท้ายเป็นจุดสุดท้ายของอนุกรมเวลา ซึ่งสุดท้ายในบรรทัดที่ 25 จะได้ข้อมูลใหม่ `NEW_TS` สำหรับอนุกรมเวลานั้นใด ๆ ที่มีขนาด 2 มิติ ซึ่งประกอบไปด้วยค่าความชันสูงสุดและค่าความชันสูงสุดของลำดับย่อยที่ถูกแบ่งจากอนุกรมเวลา

ซึ่งจะเห็นว่าการลดมิติข้อมูลด้วยการแทนข้อมูลบางส่วนนี้จะทำให้มิติข้อมูลสำหรับข้อมูลทุกประเภทมีขนาดเท่ากันคือ 2 มิติ และเนื่องจากผลต่างของจำนวนลำดับย่อยไม่ได้ส่งผลต่อความแม่นยำมากนักเมื่อเทียบกับเวลาที่ต้องเสียไปในการเรียนรู้เพื่อหาจำนวนลำดับย่อยที่เหมาะสมสำหรับข้อมูล ทำให้สามารถตั้งค่าจำนวนลำดับย่อยให้คงที่ไว้ได้ ทำให้ผู้ใช้งานไม่จำเป็นต้องกังวลกับการตั้งค่าพารามิเตอร์ใด ๆ ทำให้สะดวกสบายต่อการใช้งานสำหรับผู้ใช้งานทุกประเภท

Algorithm : Partial Representation (TS)

```
 1:  Foreach Time_Series do
 2:    Subseq_Num ← 4
 3:    Length ← Length_TS / Subseq_Num
 4:    Start_Pos ← 1
 5:    Stop_Pos ← Start_Pos + Length – 1
 6:    Count_Loop ← 1
 7:    Max_Slope = MINIMUM
 8:    Min_Slope = MAXIMUM
 9:    While Count_Loop ≤ 4 do
10:      X_bar ← (Stop_Pos + Start_Pos) / 2
11:      Y_bar ← (Stop_Value + Start_Value) / 2
12:      Slope ← FindSlope(X_bar, Y_Bar, Stop_P, Start_P)
13:      If Slope > Max_Slope
14:        Max_Slope ← Slope
15:      ElseIf Slope < Min_Slope
16:        Min_Slope ← Slope
17:      EndIf
18:      Start_Pos = Stop_Pos + 1
19:      If Count_Loop = 3
20:        Stop_Pos = Length_TS
21:      Else
22:        Stop_Pos = Start_Pos + Length - 1
23:      EndIf
24:    EndWhile
25:    NEW_TS ← [Max_Slope, Min_Slope]
26:  EndFor
27:  Return NEW_TS
```


รูปที่ 3.5 รหัสเทียมสำหรับการแทนข้อมูลบางส่วน

บทที่ 4 การทดลองและวิเคราะห์ผล

ในบทนี้กล่าวถึงประเภทชุดข้อมูลและการทดสอบประสิทธิภาพของการแทนข้อมูล ซึ่งจะแบ่งการทดลองออกได้เป็น 2 ส่วน คือ การทดสอบความเร็วและความแม่นยำสำหรับการเรียนรู้พารามิเตอร์กับพารามิเตอร์คงที่ และการทดสอบความเร็วและความแม่นยำกับการแทนข้อมูลประเภทอื่น ๆ ด้วยข้อมูลรูปแบบต่าง ๆ ซึ่งในการทดสอบนั้นผู้วิจัยใช้วิธีการจำแนกประเภทข้อมูลแบบเพื่อนบ้านใกล้ที่สุดอันดับที่หนึ่ง และทดสอบด้วยวิธีการนำออกหนึ่ง ซึ่งเป็นวิธีที่สามารถทดสอบประสิทธิภาพในด้านความเร็วและความแม่นยำได้อย่างชัดเจน และลดความเอนเอียงในการแบ่งชุดข้อมูลเรียนรู้และข้อมูลทดสอบ

4.1 ประเภทของข้อมูล

สำหรับข้อมูลที่นำมาใช้ในการทดสอบนั้นเป็นข้อมูลจากฐานข้อมูลอนุกรมเวลามหาวิทยาลัยแคลิฟอร์เนีย (UCR Time Series Archive) [12] ซึ่งเป็นฐานข้อมูลขนาดใหญ่ที่มีจำนวนข้อมูล จำนวนประเภท และความยาว ของอนุกรมเวลาที่หลากหลาย ซึ่งในการสร้างชุดข้อมูลจำเป็นต้องสร้างอนุกรมเวลาที่มีขนาดมิติเท่ากันในทุกข้อมูล จึงจำเป็นต้องตัดข้อมูลหรือการนำข้อมูลมาต่อให้มีขนาดเท่ากันจากนั้นนำไปทำซีนอร์มัลไลเซชัน (Z-Normalization) โดยข้อมูลที่ถูกทดสอบสามารถแบ่งได้ดังนี้

4.1.1 ชุดข้อมูลที่หนึ่ง

เป็นชุดข้อมูลที่ประกอบด้วยข้อมูลหลายขนาด หลายความยาว และหลายประเภท ซึ่งเป็นชุดข้อมูลดั้งเดิมที่ไม่ได้มีการตัดต่อเพิ่มเติมใด ๆ โดยชุดข้อมูลนี้ประกอบไปด้วยชุดข้อมูล 20 ชุดย่อย ซึ่งจะเป็นการทดสอบการจำแนกข้อมูลภายในประเภทเดียวกัน ซึ่งมีความยากกว่าการจำแนกข้อมูลประเภทต่างกัน โดยรายละเอียดของข้อมูลแสดงในตารางที่ 4.1

ตารางที่ 4.1 ชุดข้อมูลที่หนึ่งมีจำนวนข้อมูล 20 ประเภท

Data	Classes	Data size	Length
50Words	50	450	270
Adiac	37	390	176
Beef	5	30	470
CBF	3	30	128

Coffee	2	28	286
ECG200	2	100	96
FaceAll	14	560	131
FaceFour	4	24	350
Fish	7	175	463
Gun-Point	2	50	150
Lightning2	2	60	637
Lightning7	7	70	319
OliveOil	4	30	570
OSULeaf	6	200	427
SwedishLeaf	15	500	625
synthetic control	6	300	60
Trace	4	100	275
Two Patterns	4	1000	128
wafer	2	1000	152
yoga	2	300	426

4.1.2 ชุดข้อมูลที่สอง

เป็นชุดข้อมูลที่สร้างขึ้นเพื่อวิเคราะห์ประสิทธิภาพในการจำแนกประเภทข้อมูลในกรณีที่มีประเภทข้อมูลจำนวนมากแต่มีข้อมูลแต่ละประเภทน้อย โดยข้อมูลชุดนี้มีจำนวนข้อมูล 30 อนุกรม และมีความยาว 300 จุด ประกอบไปด้วยข้อมูล 15 ประเภท ประเภทละ 2 อนุกรม โดยรายละเอียดของข้อมูลแสดงในตารางที่ 4.2

ตารางที่ 4.2 ชุดข้อมูลที่สองมีจำนวนข้อมูล 30 อนุกรม 15 ประเภท

Name	Data size
Gun-Point	2
Wafer	2
ECG	2
Yoga	2

Coffee	2
ItalyPowerDemand	2
MoteStrain	2
SonyAIBORobot Surface	2
BirdChicken	2
Ham	2
Herring	2
ToeSegmentation1	2
ShapeletSim	2
Wine	2
Computers	2

4.1.3 ชุดข้อมูลที่สาม

เป็นชุดข้อมูลที่สร้างขึ้นเพื่อวิเคราะห์ประสิทธิภาพในการจำแนกประเภทข้อมูลในกรณีที่มีประเภทข้อมูลจำนวนมากและมีจำนวนข้อมูลเท่ากัน โดยข้อมูลชุดนี้มีจำนวนข้อมูล 500 อนุกรม และมีความยาว 100 จุดประกอบไปด้วยข้อมูล 10 ประเภท โดยรายละเอียดของข้อมูลแสดงในตารางที่ 4.3

ตารางที่ 4.3 ชุดข้อมูลที่สามมีจำนวนข้อมูล 500 อนุกรม 10 ประเภท

Name	Data Size
CBF	50
FaceFour	50
Lightning7	50
Plane	50
Car	50
Beef	50
OliveOil	50
ChlorineConcentration	50
DiatomSizeReduction	50
Symbols	50

4.1.4 ชุดข้อมูลที่สี่

เป็นชุดข้อมูลที่สร้างขึ้นเพื่อวิเคราะห์ประสิทธิภาพในการจำแนกประเภทข้อมูลในกรณีที่มีประเภทข้อมูลจำนวนมากแต่มีข้อมูลแต่ละประเภทไม่เท่ากัน เพื่อทดสอบประสิทธิภาพการจัดการความเอนเอียงต่อข้อมูลใดข้อมูลหนึ่ง โดยมีประเภทข้อมูลเหมือนกับชุดข้อมูลที่สามแต่จำนวนประเภทไม่เท่ากัน สำหรับข้อมูลชุดนี้มีจำนวนข้อมูล 559 อนุกรม และความยาว 100 จุด ประกอบด้วยข้อมูล 10 ประเภท ซึ่งแสดงในตารางที่ 4.4

ตารางที่ 4.4 ชุดข้อมูลที่สี่มีจำนวนข้อมูล 559 อนุกรม 10 ประเภท

Name	Data size
CBF	24
FaceFour	24
Lightning7	70
Plane	105
Car	60
Beef	30
OliveOil	30
ChlorineConcentration	175
DiatomSizeReduction	16
Symbols	25

4.2 การเปรียบเทียบกับงานวิจัยอื่น

ในส่วนนี้จะกล่าวถึงวิธีการทำงานของงานวิจัยอื่นที่ถูกนำมาเปรียบเทียบกับการแทนข้อมูลบางส่วน ซึ่งในการจำแนกประเภทข้อมูลแบบเพื่อนบ้านใกล้ที่สุดลำดับที่ 1 โดยทดสอบด้วยการนำออกหนึ่งนั้น พารามิเตอร์ที่ต้องการในแต่ละงานวิจัยก็จะต่างกัน สำหรับการแทนข้อมูลบางส่วนจะใช้ระยะทางแบบยุคลิดแต่ไม่ถอดรากที่สอง

สำหรับงานวิจัยที่มีการลดมิติข้อมูลที่จะนำมาทดสอบ ได้แก่ การแทนข้อมูลแบบแซคซ์ การแทนข้อมูลแบบแฟรคทัล การแทนข้อมูลด้วยความเป็นไปของอนุกรมเวลา และรวมไปถึงตัววัดระยะทางแบบยุคลิด ซึ่งจะมีรายละเอียดของแต่ละวิธีดังต่อไปนี้

4.2.1 การวัดระยะทางแบบยุคลิด

เป็นการวัดระยะทางของคู่จัดข้อมูลยกกำลังสอง หลังจากนั้นนำผลรวมของระยะทางทุกคู่จุดข้อมูลมาถอดรากที่สอง โดยจะเปรียบเทียบกับทุก ๆ ข้อมูลอนุกรมเวลา โดยถ้าคู่อนุกรมเวลาใดมีระยะทางระหว่างกันน้อยที่สุด ก็จะถูกจัดเป็นอนุกรมเวลาประเภทเดียวกัน

4.2.2 การแทนข้อมูลแบบแซคซ์

เป็นการลดขนาดของข้อมูลโดยอาศัยผลลัพธ์จากการแทนข้อมูลแบบพีเอเอ ซึ่งในการแทนข้อมูลแบบพีเอเอจำเป็นต้องกำหนดพารามิเตอร์ที่เหมาะสมก่อน หลังจากนั้นจะนำผลลัพธ์มาผ่านตารางเกาส์เซียนเพื่อแบ่งช่วงข้อมูลจากการกำหนดจำนวนตัวอักษรที่เหมาะสม ในการกำหนดความยาวจะทำการเรียนรู้ตั้งแต่ 2 ถึงครึ่งหนึ่งของความยาวอนุกรมเวลา โดยจะทำการคูณสองค่าความยาวในการเรียนรู้ทุก ๆ รอบ สำหรับจำนวนตัวอักษรนั้นจะใช้ค่าเริ่มต้นของแซคซ์ คือ 3

4.2.3 การแทนข้อมูลแบบแฟรคทัล

เป็นการแทนข้อมูลโดยใช้มิติเส้นขอบเพื่อลดขนาดมิติข้อมูลให้เหลือเพียง 2 มิติ ซึ่งในมิติแรกเป็นการวัดระยะทางด้วยความยาวที่เท่ากัน และมิติที่สองเป็นการวัดระยะทางด้วยความกว้างที่เท่ากัน โดยการกำหนดพารามิเตอร์ในการแบ่งข้อมูลนั้นจะใช้การกำหนดเช่นเดียวกับที่งานวิจัยนี้ได้กล่าวไว้ คือจะหยุดเมื่อการแบ่งข้อมูลได้ขนาดเท่าเดิม หรือค่าความแม่นยำจากการเรียนรู้ที่ได้นั้นเท่ากับการแบ่งข้อมูลครั้งก่อนหน้า

4.2.4 การแทนข้อมูลจากแนวโน้มอนุกรมเวลา

เป็นการแทนข้อมูลที่อาศัยการสังเกตุแนวโน้มของอนุกรมเวลาและข้อมูลบนจุดของอนุกรมเวลา ซึ่งจะทำการแปลงข้อมูลจากผลลัพธ์การแทนข้อมูลแบบพีเอเอ โดยจะทำการกำหนดความยาวในการเรียนรู้เช่นเดียวกับแซคซ์ คือเรียนรู้ตั้งแต่ 2 ถึงครึ่งหนึ่งของความยาวอนุกรมเวลา โดยจะทำการคูณสองค่าความยาวในการเรียนรู้ทุก ๆ รอบ เมื่อได้ผลลัพธ์มาแล้วจะแบ่งเป็นสองส่วนในการคิด ส่วนแรกคือส่วนที่ดูความเป็นไปของข้อมูล เมื่อข้อมูลของจุดปัจจุบันเพิ่มขึ้นจากเดิมจะแทนค่าข้อมูลใหม่ด้วย 1 เมื่อข้อมูลของจุดปัจจุบันลดลงจากเดิมจะแทนค่าข้อมูลใหม่ด้วย 0 ส่วนที่สองคือดูค่าของข้อมูลซึ่งเป็นการใช้การแทนข้อมูลแบบคลิปโดยหาค่าเฉลี่ยของข้อมูลทั้งหมดก่อน เมื่อข้อมูลใดมีค่ามากกว่าค่าเฉลี่ยจะแทนค่าข้อมูลใหม่ด้วย 1 และข้อมูลใดทีมีค่าน้อยกว่าหรือเท่ากับค่าเฉลี่ยจะแทนค่าข้อมูลใหม่ด้วย 0 ซึ่ง

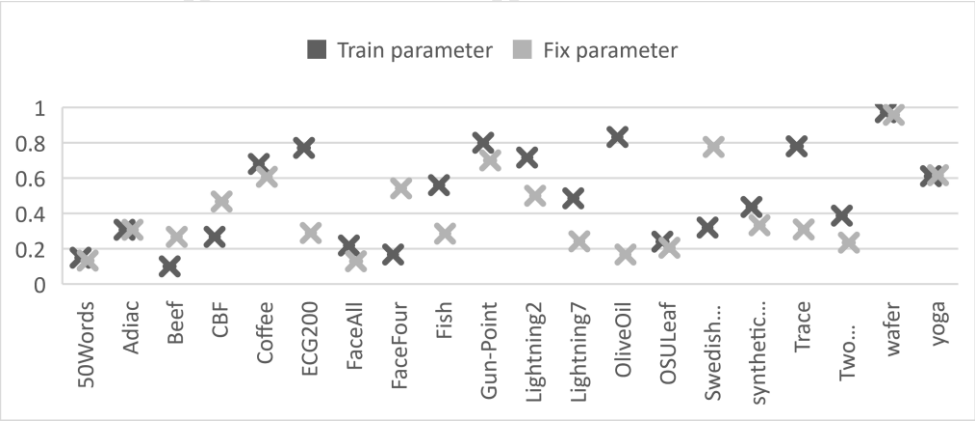
ท้ายที่สุดจะได้ข้อมูล 2 เส้น คือเส้นของแนวโน้มข้อมูลกับค่าของข้อมูล โดยจะนำเส้นทั้งสองเส้นนี้มาสลับกันระหว่างค่าของข้อมูลและค่าของแนวโน้ม

4.3 การทดลองเพื่อวิเคราะห์ประสิทธิภาพของการแทนข้อมูลบางส่วน

ในส่วนต่อไปนี้จะกล่าวถึงการทดลองรูปแบบต่าง ๆ สำหรับการแทนข้อมูลบางส่วน เพื่อทดสอบประสิทธิภาพในการทำงาน และมองถึงความเป็นไปได้ที่จะนำการแทนข้อมูลบางส่วนมาใช้งานจริง สำหรับการการทดสอบจะมุ่งเน้นไปที่การลดเวลาในการทำงานของการแทนข้อมูลบางส่วน โดยจะพยายามคงความแม่นยำไว้ให้มากที่สุด

4.3.1 การทดลองเพื่อวิเคราะห์ความคุ้มค่าในการเรียนรู้พารามิเตอร์และค่าพารามิเตอร์คงที่

ในส่วนนี้ผู้วิจัยได้ทำการทดสอบความแม่นยำระหว่างการเรียนรู้พารามิเตอร์และค่าพารามิเตอร์คงที่ของการแทนข้อมูลบางส่วน โดยในการเรียนรู้พารามิเตอร์จะเรียนรู้ตั้งแต่จำนวนลำดับย่อย 3 ถึง 10 หรือก็คือความยาวของลำดับย่อยประมาณ 30 เปอร์เซ็นต์ถึง 10 เปอร์เซ็นต์ โดยในส่วนของค่าพารามิเตอร์คงที่จะใช้ค่าเท่ากับ 4 โดยในการทดสอบจะใช้ชุดข้อมูลชุดที่หนึ่ง ที่ประกอบด้วยข้อมูล 20 ชุดย่อย เมื่อทดลองเปรียบเทียบประสิทธิภาพของความเร็วและความแม่นยำระหว่างการเรียนรู้พารามิเตอร์และค่าพารามิเตอร์คงที่ ได้ผลความแม่นยำเปรียบเทียบสำหรับทุกชุดข้อมูลแสดงในรูปที่ 4.1 และผลของเวลาสำหรับทุกชุดข้อมูลแสดงในตารางที่ 4.5



รูปที่ 4.1 ความแม่นยำของข้อมูล 20 ประเภท ระหว่างการเรียนรู้พารามิเตอร์กับพารามิเตอร์คงที่

ซึ่งจากรูปที่ 4.1 จะเห็นว่าความแม่นยำของพารามิเตอร์คงที่มีความใกล้เคียงกับความแม่นยำจากการเรียนรู้พารามิเตอร์ในหลายข้อมูล ซึ่งจะมีเพียงสามข้อมูลเท่านั้นที่มีค่าความแม่นยำน้อยกว่ามาก ๆ คือข้อมูล ECG200 OliveOil และข้อมูล Trace ซึ่ง ในขณะที่เวลาที่แสดงในตารางที่ 4.5 แสดงให้เห็นว่าความเร็วที่ใช้ต่างกันเป็นอย่างมาก จากผลลัพธ์นี้แสดงให้เห็นว่าการใช้ค่าพารามิเตอร์คงที่ก็เพียงพอที่จะใช้ในการแทนข้อมูลบางส่วน ซึ่งจะช่วยให้การทำงานเร็วขึ้นเป็นอย่างมากเมื่อเทียบกับการเรียนรู้พารามิเตอร์

ตารางที่ 4.5 เวลาในการทำงานระหว่างการเรียนรู้พารามิเตอร์กับพารามิเตอร์คงที่

Data	Time (Trained Parameters) (ns)	Time (Fixed Parameter) (ns)
50Words	$5.79 \times 10^{\wedge 11}$	$3.15 \times 10^{\wedge 7}$
Adiac	$2.62 \times 10^{\wedge 11}$	$1.64 \times 10^{\wedge 7}$
Beef	$2.34 \times 10^{\wedge 8}$	$1.18 \times 10^{\wedge 5}$
CBF	$8.35 \times 10^{\wedge 7}$	$9.47 \times 10^{\wedge 4}$
Coffee	$1.24 \times 10^{\wedge 8}$	$8.57 \times 10^{\wedge 4}$
ECG200	$2.71 \times 10^{\wedge 9}$	$1.09 \times 10^{\wedge 6}$
FaceAll	$9.74 \times 10^{\wedge 7}$	$6.12 \times 10^{\wedge 4}$
FaceFour	$5.72 \times 10^{\wedge 10}$	$3.04 \times 10^{\wedge 6}$
Fish	$4.61 \times 10^{\wedge 8}$	$2.52 \times 10^{\wedge 5}$
Gun-Point	$3.08 \times 10^{\wedge 9}$	$3.66 \times 10^{\wedge 5}$
Lightning2	$2.54 \times 10^{\wedge 9}$	$4.78 \times 10^{\wedge 5}$
Lightning7	$3.07 \times 10^{\wedge 8}$	$9.20 \times 10^{\wedge 4}$
OliveOil	$5.79 \times 10^{\wedge 11}$	$5.79 \times 10^{\wedge 11}$
OSULeaf	$8.10 \times 10^{\wedge 10}$	$4.30 \times 10^{\wedge 6}$
SwedishLeaf	$5.29 \times 10^{\wedge 11}$	$1.92 \times 10^{\wedge 7}$
synthetic control	$7.10 \times 10^{\wedge 10}$	$7.50 \times 10^{\wedge 6}$
Trace	$7.56 \times 10^{\wedge 9}$	$8.59 \times 10^{\wedge 5}$
Two Patterns	$5.16 \times 10^{\wedge 12}$	$9.15 \times 10^{\wedge 7}$
wafer	$5.51 \times 10^{\wedge 12}$	$9.01 \times 10^{\wedge 17}$
yoga	$3.30 \times 10^{\wedge 11}$	$9.23 \times 10^{\wedge 6}$

4.3.2 การทดลองและเปรียบเทียบการแทนข้อมูลบางส่วนกับการแทนข้อมูลที่มีการลดมิติ

ในส่วนนี้จะเป็นการทดสอบความแม่นยำและเวลาในการทำงาน เมื่อเปรียบเทียบกับการแทนข้อมูลที่มีการลดมิติจากงานวิจัยต่าง ๆ ได้แก่ การแทนข้อมูลแบบแซคซ์ การแทนข้อมูลแบบแฟรคทัล การแทนข้อมูลจากแนวโน้มอนุกรมเวลา และตัววัดระยะทางแบบยุคลิด

4.3.2.1 การทดสอบด้วยชุดข้อมูลที่หนึ่ง

การทดสอบด้วยข้อมูลชุดที่หนึ่งจะเป็นการทดสอบกับข้อมูล 20 ชุดข้อมูล โดยจะเป็นการทดสอบการจำแนกประเภทข้อมูลของแต่ละชุดข้อมูลย่อย ซึ่งจะมีความยากในการจำแนกประเภทข้อมูลมากกว่าข้อมูลที่เกิดจากชุดข้อมูลย่อยคนละประเภทกัน โดยข้อมูลชุดที่หนึ่งนี้ประกอบไปด้วยชุดข้อมูลย่อยที่มีขนาด ความยาว และจำนวนประเภทที่หลากหลาย โดยผลการทดลองสำหรับการแทนข้อมูลบางส่วนเมื่อเปรียบเทียบกับวิธีอื่นด้านความแม่นยำแสดงในตารางที่ 4.6 และประสิทธิภาพด้านเวลาแสดงในตารางที่ 4.7

จากตารางที่ 4.6 พบว่าความแม่นยำของการแทนข้อมูลบางส่วนมีความใกล้เคียงกับการแทนข้อมูลแบบแซคซ์และการแทนข้อมูลแบบแฟรคทัล ในขณะที่เมื่อเปรียบเทียบกับการวัดระยะทางแบบยุคลิดหรือการแทนข้อมูลจากแนวโน้มข้อมูลอนุกรมเวลาแล้วนั้น จะเห็นความแตกต่างของความแม่นยำที่ค่อนข้างชัดเจนเป็นอย่างมาก จุดนี้แสดงให้เห็นว่าการแทนข้อมูลบางส่วนยังไม่สามารถที่จะแยกข้อมูลประเภทเดียวกันแต่คนละกลุ่มออกจากกันได้อย่างมีประสิทธิภาพมากนัก อย่างไรก็ตามความแม่นยำของการแทนข้อมูลบางส่วนสามารถให้ค่าที่ใกล้เคียงกับการแทนข้อมูลแบบแซคซ์ที่ใช้กันแพร่หลายในปัจจุบัน

จากตารางที่ 4.7 จะพบว่าเวลาในการทำงานของการแทนข้อมูลบางส่วนมีความเร็วกว่าการแทนข้อมูลแบบแซคซ์ การแทนข้อมูลแบบแฟรคทัล และการแทนข้อมูลจากแนวโน้มอนุกรมเวลา เป็นอย่างมากโดยยิ่งอนุกรมเวลามีขนาดยาว หรืออนุกรมเวลามีจำนวนมาก จะยิ่งทำให้เวลาในการทำงานของการแทนข้อมูลบางส่วนกับการแทนข้อมูลทั้งสามต่างกันมากยิ่งขึ้น สำหรับเวลาของการแทนข้อมูลบางส่วนนั้นมีความเร็วมากกว่าการวัดระยะทางแบบยุคลิดค่อนข้างมาก โดยมี 6 ชุดข้อมูลที่มีความเร็วเฉลี่ยเพิ่มขึ้นถึง 184 เท่า และอีก 11 ชุดข้อมูลที่มีความเร็วเฉลี่ยเพิ่มขึ้น 40 เท่า โดยจากผลการทดลองจะเห็นว่า การแทนข้อมูลบางส่วนจะทำงานเร็วมากยิ่งขึ้นเมื่อเทียบกับการวัดระยะทางแบบยุคลิดในกรณีที่ความยาวของข้อมูลเพิ่มขึ้น

ตารางที่ 4.6 ความแม่นยำในการทำงานของข้อมูลชุดที่หนึ่ง 20 ประเภท

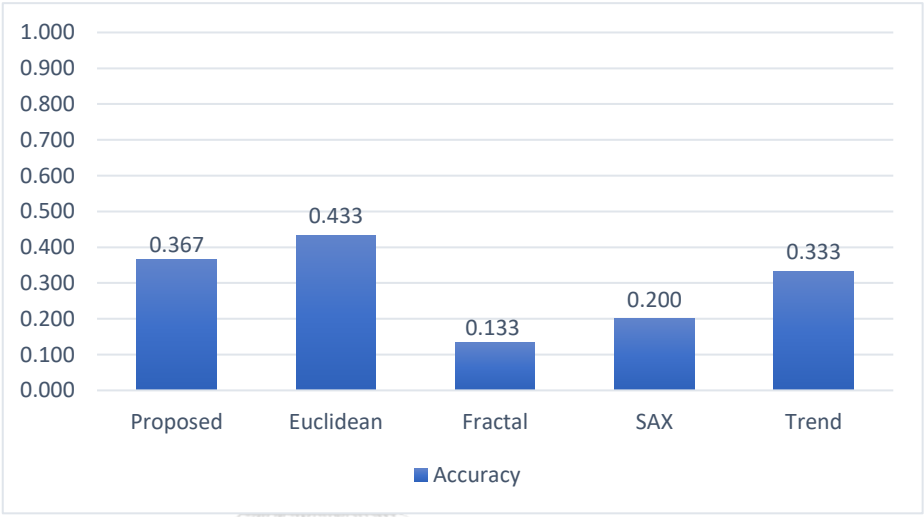
Data	Proposed	Euclidean	Fractal	SAX	Trend
50Words	0.133	0.636	0.180	0.158	0.593
Adiac	0.308	0.605	0.228	0.095	0.159
Beef	0.267	0.500	0.433	0.333	0.433
CBF	0.467	0.833	0.733	0.433	0.800
Coffee	0.607	1.000	0.250	0.536	0.964
ECG200	0.710	0.860	0.700	0.720	0.860
FaceAll	0.129	0.887	0.216	0.116	0.736
FaceFour	0.458	0.667	0.250	0.333	0.958
Fish	0.286	0.760	0.331	0.286	0.709
Gun-Point	0.700	0.960	0.620	0.580	0.860
Lightning2	0.500	0.717	0.050	0.483	0.517
Lightning7	0.243	0.643	0.100	0.214	0.100
OliveOil	0.833	0.900	0.400	0.533	0.633
OSULeaf	0.205	0.620	0.450	0.260	0.615
SwedishLeaf	0.224	0.750	0.392	0.174	0.472
synthetic control	0.333	0.133	0.507	0.673	0.670
Trace	0.690	0.840	0.990	0.490	0.620
Two Patterns	0.235	0.914	0.287	0.423	0.353
wafer	0.957	0.903	0.574	0.959	0.999
yoga	0.617	0.540	0.653	0.543	0.763

ตารางที่ 4.7 ความเร็วในการทำงานของข้อมูลชุดที่หนึ่งในหน่วยนาโนวินาที

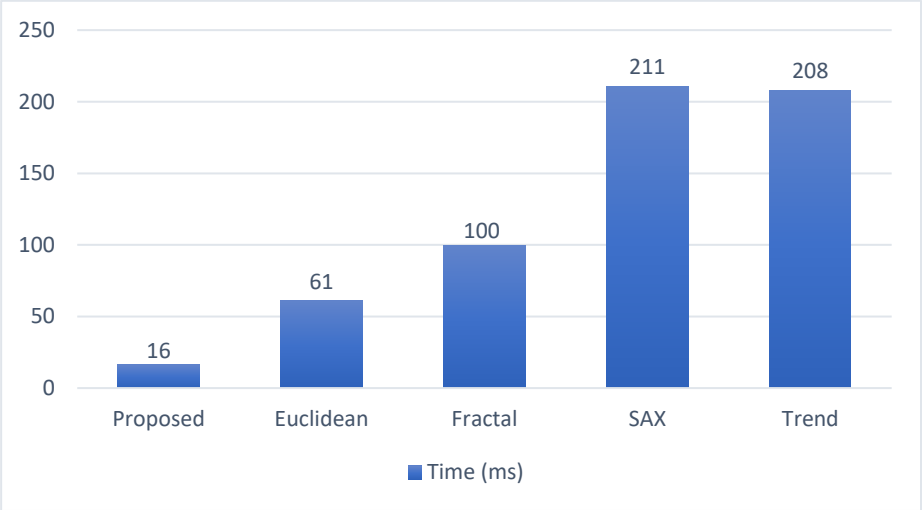
Data	Proposed	Euclidean	Fractal	SAX	Trend
50Words	3.15 x 10^7	9.42 x 10^8	1.19 x 10^12	1.67 x 10^12	1.36 x 10^12
Adiac	1.64 x 10^7	4.73 x 10^8	4.07 x 10^11	5.19 x 10^11	4.24 x 10^11
Beef	1.18 x 10^5	1.10 x 10^7	4.32 x 10^8	5.12 x 10^8	4.40 x 10^8
CBF	9.47 x 10^4	2.49 x 10^6	1.07 x 10^8	1.34 x 10^8	9.93 x 10^7
Coffee	8.57 x 10^4	8.82 x 10^6	2.24 x 10^8	3.25 x 10^8	2.43 x 10^8
ECG200	1.09 x 10^6	1.94 x 10^7	3.56 x 10^9	4.26 x 10^9	3.33 x 10^9
FaceAll	3.15 x 10^7	3.62 x 10^8	1.09 x 10^12	1.45 x 10^12	1.23 x 10^12
FaceFour	6.12 x 10^4	9.00 x 10^6	1.37 x 10^8	2.26 x 10^8	1.80 x 10^8
Fish	3.04 x 10^6	1.68 x 10^8	8.28 x 10^10	1.15 x 10^11	9.18 x 10^10
Gun-Point	2.52 x 10^5	1.24 x 10^7	4.93 x 10^8	9.49 x 10^8	6.91 x 10^8
Lightning2	3.66 x 10^5	1.15 x 10^8	1.72 x 10^9	7.53 x 10^9	6.54 x 10^9
Lightning7	4.78 x 10^5	7.63 x 10^7	1.49 x 10^9	5.58 x 10^9	4.78 x 10^9
OliveOil	9.20 x 10^4	2.19 x 10^7	4.48 x 10^8	8.24 x 10^8	6.21 x 10^8
OSULeaf	4.30 x 10^6	1.95 x 10^8	1.41 x 10^11	1.67 x 10^11	1.42 x 10^11
SwedishLeaf	1.92 x 10^7	5.66 x 10^8	7.39 x 10^11	6.82 x 10^11	5.13 x 10^11
synthetic control	7.50 x 10^6	6.49 x 10^7	1.29 x 10^11	7.12 x 10^10	5.77 x 10^10
Trace	8.59 x 10^5	1.23 x 10^8	1.41 x 10^10	1.53 x 10^10	1.72 x 10^10
Two Patterns	9.15 x 10^7	4.28 x 10^8	6.55 x 10^12	6.78 x 10^12	5.63 x 10^12
wafer	9.01 x 10^7	6.35 x 10^8	6.11 x 10^12	9.12 x 10^12	8.12 x 10^12
yoga	9.23 x 10^6	4.58 x 10^8	5.13 x 10^11	5.64 x 10^11	4.33 x 10^11

4.3.2.2 การทดสอบด้วยชุดข้อมูลที่สอง

การทดสอบด้วยชุดข้อมูลชุดที่สองจะเป็นการทดสอบประสิทธิภาพในการจำแนกประเภทข้อมูลหลายประเภทโดยที่มีจำนวนอนุกรมเวลาน้อยมาก ๆ คือเพียง 2 อนุกรมต่อประเภท ซึ่งจะทำให้มีความยากมากยิ่งขึ้นในการจำแนกประเภทข้อมูล โดยผลการทดลองสำหรับการแทนข้อมูลบางส่วนเมื่อเปรียบเทียบกับวิธีอื่นด้านความแม่นยำแสดงในรูปที่ 4.2 และประสิทธิภาพด้านเวลาแสดงในรูปที่ 4.3



รูปที่ 4.2 ความแม่นยำของชุดข้อมูลที่สอง

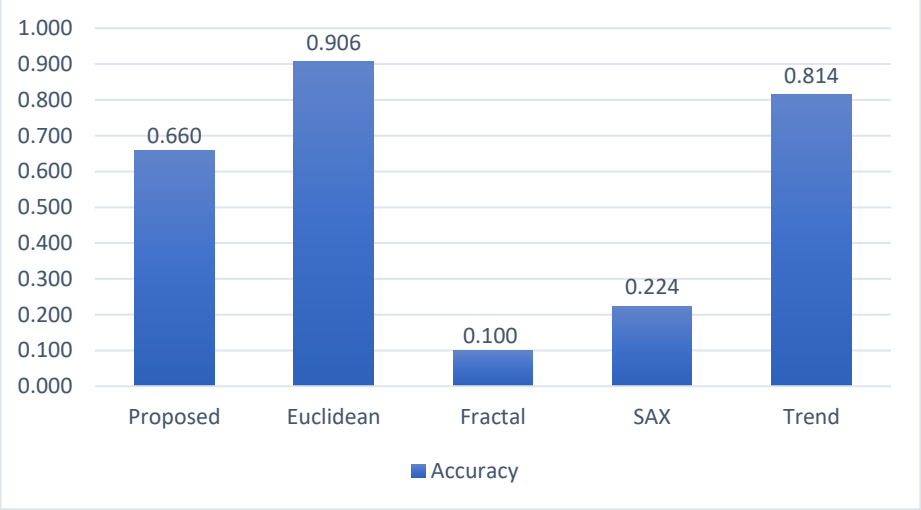


รูปที่ 4.3 เวลาในการทำงานของข้อมูลชุดที่สอง

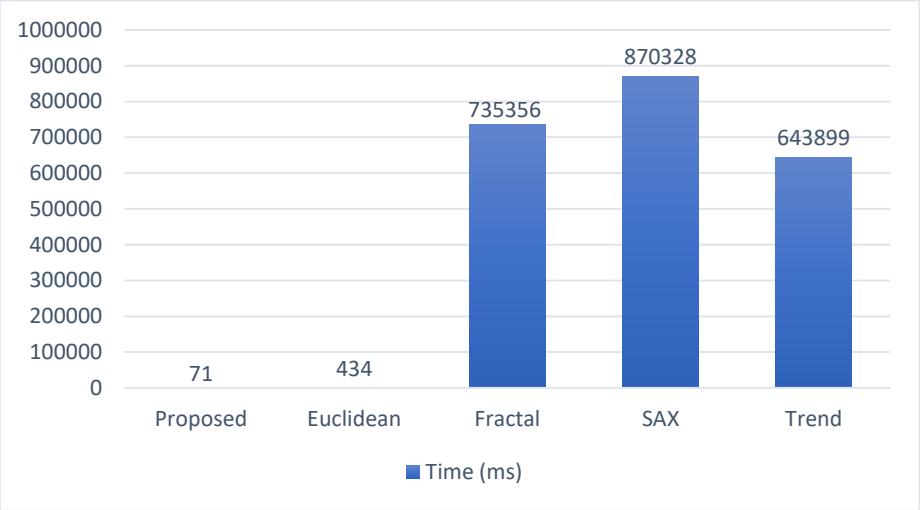
สำหรับผลลัพธ์ส่วนของความแม่นยำ ในรูปที่ 4.2 จะเห็นว่าการแทนข้อมูลบางส่วนสามารถเอาชนะการแทนแบบแฟรคทัลและการแทนข้อมูลแบบแซกซ์ไป 23 เปอร์เซ็นต์ และ 16 เปอร์เซ็นต์ โดยมีความแม่นยำที่ใกล้เคียงกับการแทนข้อมูลแบบแนวโน้มของอนุกรมเวลาและการวัดระยะทางแบบยุคลิด อย่างไรก็ตามเวลาในการทำงานจากรูปที่ 4.3 การแทนข้อมูลบางส่วนมีความเร็วมากกว่าการวัดระยะทางแบบประมาณ 4 เท่า และเมื่อเทียบระยะเวลากับการแทนข้อมูลแบบแนวโน้มของอนุกรมเวลาจะมีความเร็วมากกว่าถึง 12 เท่า ซึ่งจะเห็นว่าความต่างของเวลาจะไม่ต่างกันมากนัก เนื่องจากจำนวนอนุกรมเวลามีจำนวนน้อยและขนาดอนุกรมเวลามีขนาดสั้น

4.3.2.3 การทดสอบด้วยชุดข้อมูลที่สาม

ในชุดข้อมูลนี้จะทำการทดสอบการจำแนกประเภทข้อมูลที่มีหลายประเภทเช่นเดียวกับข้อมูลชุดที่สอง แต่จำนวนข้อมูลจะมากกว่าโดยที่มีจำนวนข้อมูลเท่ากัน ซึ่งจะเป็นกรณีการทดสอบที่ไม่มีการเอนเอียงไปยังข้อมูลใดข้อมูลหนึ่ง โดยผลการทดลองแสดงในรูปที่ 4.4 และรูปที่ 4.5 ซึ่งเป็นผลลัพธ์ในด้านความแม่นยำและเวลาตามลำดับ



รูปที่ 4.4 ความแม่นยำของข้อมูลชุดที่สาม

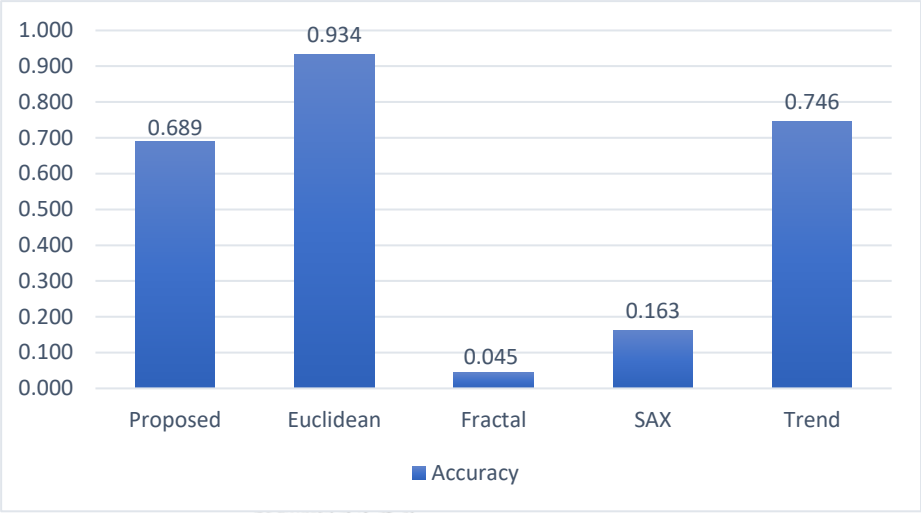


รูปที่ 4.5 เวลาในการทำงานของข้อมูลชุดที่สาม

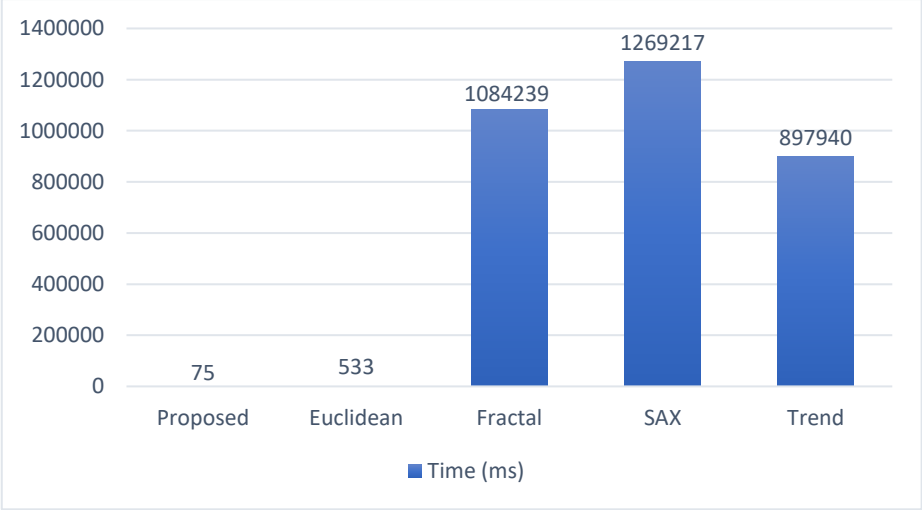
จากรูปที่ 4.4 การแทนข้อมูลบางส่วนสามารถเอาชนะการแทนข้อมูลแบบแฟรคทัลไปถึง 56 เปอร์เซ็นต์ และเอาชนะการแทนข้อมูลแบบแซคซ์ไป 44 เปอร์เซ็นต์ ซึ่งในชุดข้อมูลนี้ความแม่นยำของการแทนข้อมูลบางส่วนไม่ได้ใกล้เคียงกับการวัดระยะทางแบบยุคลิดเท่ากับข้อมูลชุดที่สอง แต่เมื่อพิจารณารูปที่ 4.5 จะเห็นว่าการแทนข้อมูลบางส่วนทำงานได้เร็วกว่าการวัดระยะทางแบบยุคลิดประมาณ 6 เท่า โดยที่เมื่อเทียบกับการแทนข้อมูลอีกสามแบบจะใช้เวลาต่างกันมากกว่า 9000 เท่าเนื่องจากไม่มีการเรียนรู้พารามิเตอร์ใด ๆ

4.3.2.4 การทดสอบด้วยชุดข้อมูลที่สี่

ในชุดข้อมูลนี้จะทำการทดสอบการจำแนกประเภทข้อมูลที่มีหลายประเภท โดยมีประเภทชนิดเดียวกันกับข้อมูลชุดที่สาม แต่จำนวนข้อมูลจะมีขนาดไม่เท่ากัน ซึ่งจะเป็นกรณีการทดสอบที่มีการเอนเอียงไปยังข้อมูลใดข้อมูลหนึ่ง โดยผลการทดลองแสดงในรูปที่ 4.6 และรูปที่ 4.7 ซึ่งเป็นผลลัพธ์ในด้านความแม่นยำและเวลาตามลำดับ



รูปที่ 4.6 ความแม่นยำของข้อมูลชุดที่สี่



รูปที่ 4.7 เวลาในการทำงานของข้อมูลชุดที่สี่

จากผลลัพธ์ในรูปที่ 4.6 และ 4.7 จะมีลักษณะของความแม่นยำที่เหมือนกับชุดข้อมูลที่สามคือ การแทนข้อมูลบางส่วนแพ้การวัดระยะทางแบบยุคลิดและการวัดระยะทางแบบแนวโน้มอนุกรมเวลา แต่ความแตกต่างกับการวัดระยะทางแบบแนวโน้มอนุกรมเวลามีความใกล้เคียงกันมากขึ้น ในขณะที่สามารถเอาชนะการแทนข้อมูลแบบแฟรคทัลและการแทนข้อมูลแบบแซคซ์ไปได้ด้วยความแม่นยำที่ค่อนข้างมาก ซึ่งในข้อมูลที่สี่ที่มีความเอนเอียงของข้อมูลจะเห็นว่าการแทนข้อมูล

บางส่วนและการวัดระยะทางแบบยุคลิดยังสามารถที่จะคงประสิทธิภาพการจำแนกประเภทไว้ได้ จากความแม่นยำที่ไม่ลดลงจากชุดข้อมูลที่สาม ในขณะที่การแทนข้อมูลทั้งสามแบบมีความแม่นยำลดลงจากเดิม สำหรับเวลาในการทำงานนั้นการแทนข้อมูลบางส่วนและการวัดระยะทางยุคลิดไม่ได้เปลี่ยนแปลงไปมากนักตามจำนวนอนุกรมเวลาที่เพิ่มขึ้น ในขณะที่การแทนข้อมูลทั้งสามแบบที่ถูกนำมาเปรียบเทียบมีเวลาที่เพิ่มขึ้นอย่างชัดเจนเมื่อจำนวนอนุกรมเวลาเพิ่มขึ้น

บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ

วิทยานิพนธ์นี้ได้พัฒนาการแทนข้อมูลที่ลดมิติข้อมูลให้เหลือขนาดน้อยที่สุดโดยมีขนาดเท่ากันสำหรับทุกข้อมูล โดยการแทนข้อมูลนี้มีชื่อว่าการแทนข้อมูลบางส่วน ซึ่งสามารถลดขนาดมิติข้อมูลให้เหลือเพียง 2 มิติ โดยใช้ค่าความชันจากการลากเส้นตรงของจุดเริ่มต้นและจุดสิ้นสุดของส่วนหนึ่งในอนุกรมเวลา จากผลการทดลองที่แสดงไว้ในบทที่ 4 สามารถสรุปภาพรวมของงานวิจัยได้ดังนี้

5.1 สรุปผลการวิจัย

สำหรับการทดลองเพื่อทดสอบประสิทธิภาพสำหรับการแทนข้อมูลบางส่วนนั้น พบว่าการแทนข้อมูลบางส่วนสามารถทำงานได้ดีในข้อมูลคนละประเภทกัน นั่นคือข้อมูลที่มีรูปร่างแตกต่างกัน สำหรับข้อมูลประเภทเดียวที่มีรูปร่างใกล้เคียงกันจะได้ความแม่นยำที่ไม่ดีเท่ากับข้อมูลคนละประเภท เนื่องจากการลดมิติข้อมูลจะทำให้สูญเสียลักษณะข้อมูลไป ทำให้การแยกข้อมูลที่มีความคล้ายกันนั้นไม่มีประสิทธิภาพมากเท่าที่ควร

ในส่วนของการแทนข้อมูลอื่นที่ถูกนำมาเปรียบเทียบนั้น ในชุดข้อมูลแรกที่เป็นข้อมูลคนละประเภทกัน พบว่าประสิทธิภาพในการจำแนกประเภทข้อมูลที่มีจำนวนน้อยและมีประเภทมาก สามารถทำได้ดีใกล้เคียงกับการวัดระยะทางแบบยุคลิด และสามารถเอาชนะการแทนข้อมูลทั้งสามที่นำมาเปรียบเทียบได้ ในส่วนของการทดสอบเพื่อวัดประสิทธิภาพในการรองรับการเอนเอียงไปยังประเภทใดประเภทหนึ่งของข้อมูล พบว่าการแทนข้อมูลบางส่วนสามารถรองรับการเอนเอียงนี้ได้เช่นเดียวกับการวัดระยะทางแบบยุคลิด ในขณะที่การแทนข้อมูลทั้งสามแบบนั้นมีค่าความแม่นยำที่ลดลง ซึ่งสำหรับเวลาในการทดลองนั้นพบว่ามีความเร็วในการทำงานมากกว่าการวัดระยะทางแบบยุคลิดประมาณ 5 เท่า แต่เมื่อเทียบกับการแทนข้อมูลที่นำมาเปรียบเทียบนั้นพบว่ามีความเร็วมากกว่าถึง 9,000 เท่า เนื่องจากการแทนข้อมูลบางส่วนจะไม่เสียเวลาในการเรียนรู้พารามิเตอร์

สำหรับการเปรียบเทียบกับชุดข้อมูลที่สองที่เป็นข้อมูลประเภทเดียวกันพบว่า การแทนข้อมูลบางส่วนมีความแม่นยำที่ใกล้เคียงกับการแทนข้อมูลแบบแซคซ์และการแทนข้อมูลแบบแฟรคทัล ในขณะที่เมื่อเปรียบเทียบความแม่นยำกับการแทนข้อมูลจากแนวโน้มอนุกรมเวลาและการวัดระยะทางแบบยุคลิด จะเห็นว่าไม่สามารถให้ความแม่นยำที่ใกล้เคียงได้ เนื่องจากการสูญเสียคุณลักษณะของข้อมูลในการลดมิติข้อมูลไป แต่เมื่อดูผลการทดลองด้านเวลาจะพบว่านอกจากที่เวลาในการทำงานจะมากกว่าการแทนข้อมูลทั้งสามแล้ว ยังมีความเร็วมากกว่าการวัดระยะทางแบบยุคลิดอีกด้วย โดยมี 6 ข้อมูลที่มีความเร็วเฉลี่ยมากกว่าการวัดระยะทางแบบยุคลิดถึง 184 เท่า และอีก 11 ข้อมูลที่มีความเร็วเฉลี่ยมากกว่า 40 เท่า ซึ่งจะพบว่าใน 6 ข้อมูลที่มีความเร็วต่างกันมากนั้นเป็น

อนุกรมเวลาที่มีความยาวมากกว่าอนุกรมอื่น ซึ่งจะเห็นได้ว่ายิ่งถ้าข้อมูลอนุกรมเวลามีความยาวมาก ความต่างของเวลาจะยิ่งเพิ่มมากขึ้น

จากผลการทดลองจึงสรุปได้ว่าการแทนข้อมูลบางส่วนสามารถจำแนกประเภทข้อมูลได้ดีสำหรับข้อมูลคนละประเภท แต่จะทำงานได้ไม่ดีมากนักในข้อมูลประเภทเดียวกันที่มีความคล้ายกัน เนื่องจากมีการสูญเสียข้อมูลบางส่วนไป แต่เวลาในการทำงานนั้นมีความเร็วมากกว่าวิธีการอื่นที่ถูกนำมาเปรียบเทียบเป็นอย่างมากรวมถึงการวัดระยะทางแบบยุคลิด นอกจากนี้การแทนข้อมูลบางส่วนยังสามารถรองรับการเอนเอียงของข้อมูลต่อประเภทใดประเภทหนึ่งได้เป็นอย่างดี

5.2 ข้อเสนอแนะ

จากผลการทดลองจะเห็นว่าประเด็นหลักที่ต้องพัฒนาคือความสามารถในการจำแนกข้อมูลประเภทเดียวกัน ซึ่งเป็นเหตุเนื่องมาจากการสูญเสียข้อมูลไประหว่างการลดมิติข้อมูล ในส่วนนี้สามารถที่จะพัฒนาให้ดีได้ยิ่งขึ้นโดยการลากเส้นตรงผ่านทุกจุดบนลำดับย่อยอนุกรมเวลา ซึ่งจะทำให้การสูญเสียข้อมูลนั้นลดลง แต่ถ้าหากใช้ข้อมูลทุกจุดจากลำดับย่อยโดยปกติ จะทำให้เวลาในการทำงานใกล้เคียงกับการวัดระยะทางแบบยุคลิดในกรณีที่ไม่มีการเรียนรู้พารามิเตอร์จำนวนลำดับย่อย ซึ่งถ้าหากมีการเรียนรู้พารามิเตอร์ก็จะทำให้ใช้เวลามากกว่าการวัดระยะทางแบบยุคลิด ซึ่งจะไม่ใช่เป้าหมายในการลดมิติข้อมูลที่ต้องการให้เวลาในการทำงานลดลง

เพราะฉะนั้นแนวทางในการพัฒนานั้นน่าจะพุ่งเป้าไปยังการใช้จุดทุกจุดบนลำดับย่อย เนื่องจากการใช้เพียงจุดเริ่มต้นและจุดจบของลำดับย่อยจะทำให้สูญเสียข้อมูลไปค่อนข้างมาก โดยมีหลายส่วนที่สามารถพัฒนาได้ คือการเลือกส่วนลำดับย่อยที่สามารถให้ค่าความชันซึ่งแยกข้อมูลประเภทเดียวกันได้ดีขึ้น โดยที่ในการเลือกลำดับย่อยจะต้องใช้เวลาในการเลือกที่น้อย เพราะต้องใช้ข้อมูลทุกจุดจากลำดับย่อย ถ้าหากใช้เวลามากในการเลือกจะทำให้เวลาในการทำงานมากขึ้น นอกจากนี้ยังสามารถพัฒนาได้ในส่วนของค่าข้อมูลใหม่ที่ใช้แทนอนุกรมเวลา ซึ่งอาจจะมีวิธีการที่ดีกว่าการลากเส้นตรงเพื่อหาความชันจากลำดับย่อย โดยค่าข้อมูลจากวิธีนี้จะให้ค่าข้อมูลต่างกันระหว่างประเภทข้อมูลที่ต่างกันมากขึ้น ทำให้สามารถจำแนกประเภทข้อมูลได้มีประสิทธิภาพมากยิ่งขึ้น

รายการอ้างอิง

[1] Hamilton, James Douglas. *Time series analysis*. Vol. 2. Princeton: Princeton university press, 1994.

[2] Danielsson, Per-Erik. "Euclidean distance mapping." *Computer Graphics and image processing* 14.3 (1980): 227-248.

[3] Piatetsky-Shapiro, Gregory. *Advances in knowledge discovery and data mining*. Eds. Usama M. Fayyad, Padhraic Smyth, and Ramasamy Uthurusamy. Vol. 21. Menlo Park: AAAI press, 1996.

[4] Berndt, Donald J., and James Clifford. "Using dynamic time warping to find patterns in time series." *KDD workshop*. Vol. 10. No. 16. 1994.

[5] Lin, J., Keogh, E., Wei, L., & Lonardi, S. (2007). Experiencing SAX: a novel symbolic representation of time series. Data Mining and knowledge discovery, 15(2), 107-144.

[6] Yi, B. K., & Faloutsos, C. (2000). Fast time sequence indexing for arbitrary Lp norms. VLDB.

[7] Lin, J., Khade, R., & Li, Y. (2012). Rotation-invariant similarity in time series using bag-of-patterns representation. Journal of Intelligent Information Systems, 39(2), 287-315.

[8] Cover, Thomas, and Peter Hart. "Nearest neighbor pattern classification." *IEEE transactions on information theory* 13.1 (1967): 21-27.

[9] Fama, Eugene F. "The behavior of stock-market prices." *The journal of Business* 38.1 (1965): 34-105.

[10] Periodic, E. E. G. "Clinical significance of periodic EEG patterns." *Arch Neurol* 37 (1980): 15-20.

[11] Blackburn, Henry, et al. "The electrocardiogram in population studies: a classification system." *Circulation* 21.6 (1960): 1160-1175.

[12] Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall A., Mueen A., & Batista, G. (2015). The UCR Time Series Classification Archive [Online]. Available from: www.cs.ucr.edu/~eamonn/time_series_data/ [Accessed June 20, 2018].

[13] Fiscal Policy Office (2005). Time Series [Online]. Available from: www.fpo.go.th/S-I/Source/ECO/ECO24.htm [Accessed June 20, 2018].

[14] Ye, L., & Keogh, E. (2009). Time series shapelets: a new primitive for data mining. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 947-956). ACM.

[15] Neter, John, et al. *Applied linear statistical models*. Vol. 4. Chicago: Irwin, 1996.

[16] Ratanamahatana, C., Keogh, E., Bagnall, A. J., & Lonardi, S. (2005). A novel bit level time series representation with implication of similarity search and clustering. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 771-777). Springer, Berlin, Heidelberg.

[17] Sajjipanon, Poat, and Chotirat Ann Ratanamahatana. "A Novel Fractal Representation for Dimensionality Reduction of Large Time Series Data." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Berlin, Heidelberg, 2009.

[18] Kane, Aminata. "Trend and value based time series representation for similarity search." *Multimedia Big Data (BigMM), 2017 IEEE Third International Conference on*. IEEE, 2017.

ภาคผนวก

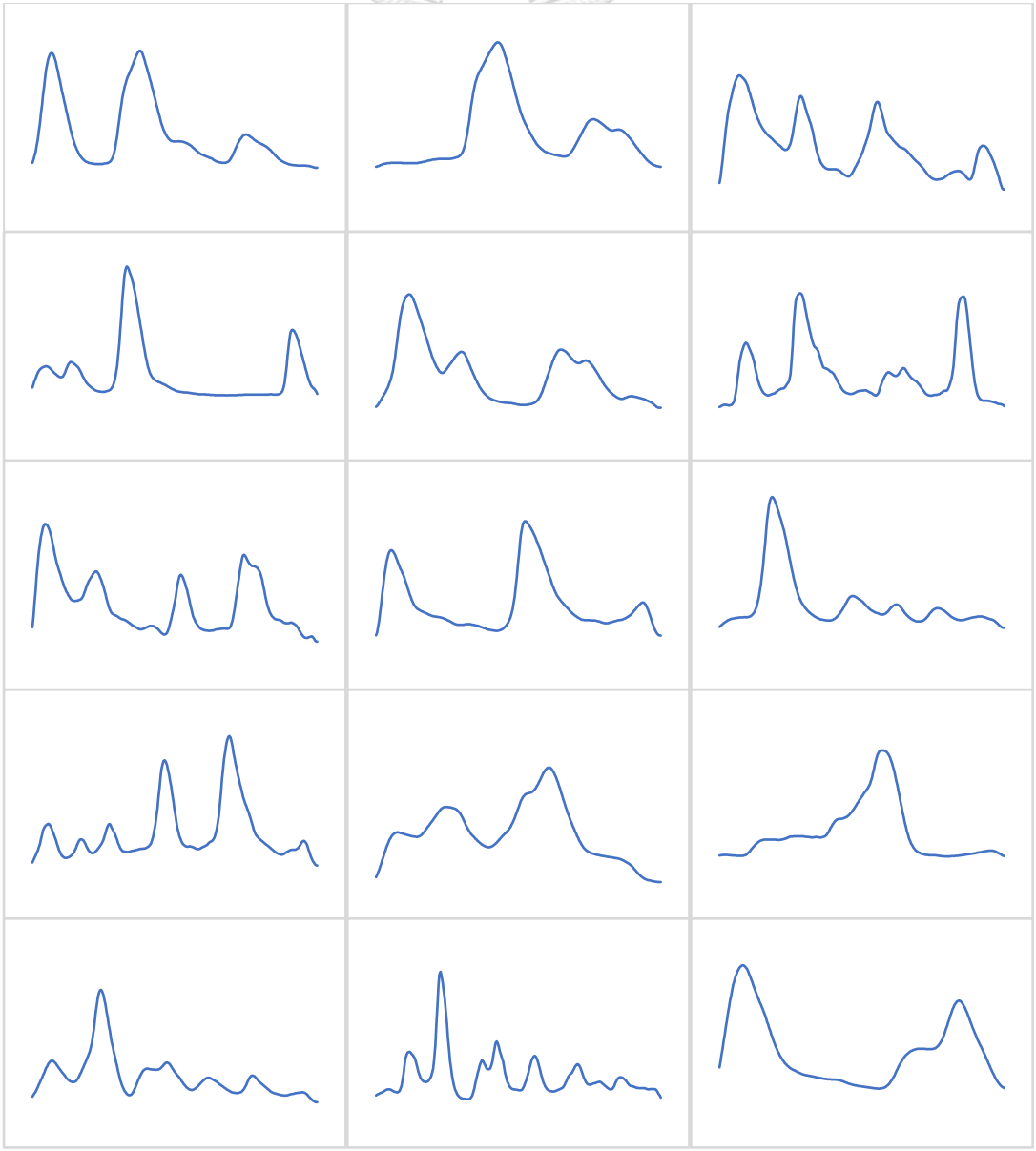
ภาคผนวก ก

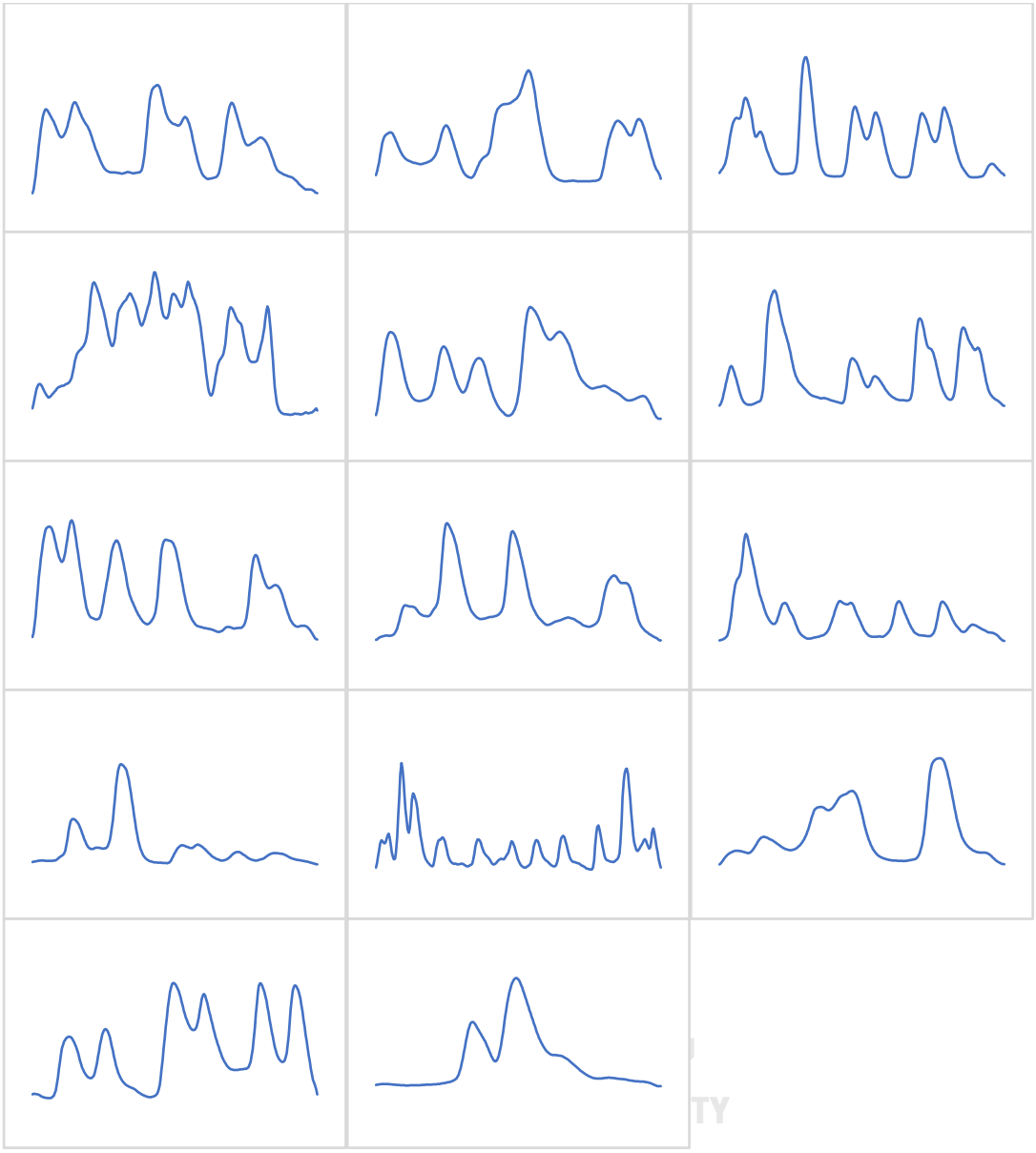
ในส่วนนี้จะแสดงตัวอย่างข้อมูลที่นำมาใช้สำหรับการทดลอง โดยข้อมูลนั้นนำมาจากฐานข้อมูล UCR Time Series Classification Archive ซึ่งเป็นฐานข้อมูลของมหาวิทยาลัยแคลิฟอร์เนีย โดยมีทั้งหมด 4 ชุดข้อมูล ดังนี้

ก.1 ชุดข้อมูลที่หนึ่ง

ข้อมูลชุดที่หนึ่งมีประกอบไปด้วยชุดข้อมูล 20 ชุดย่อย ดังต่อไปนี้

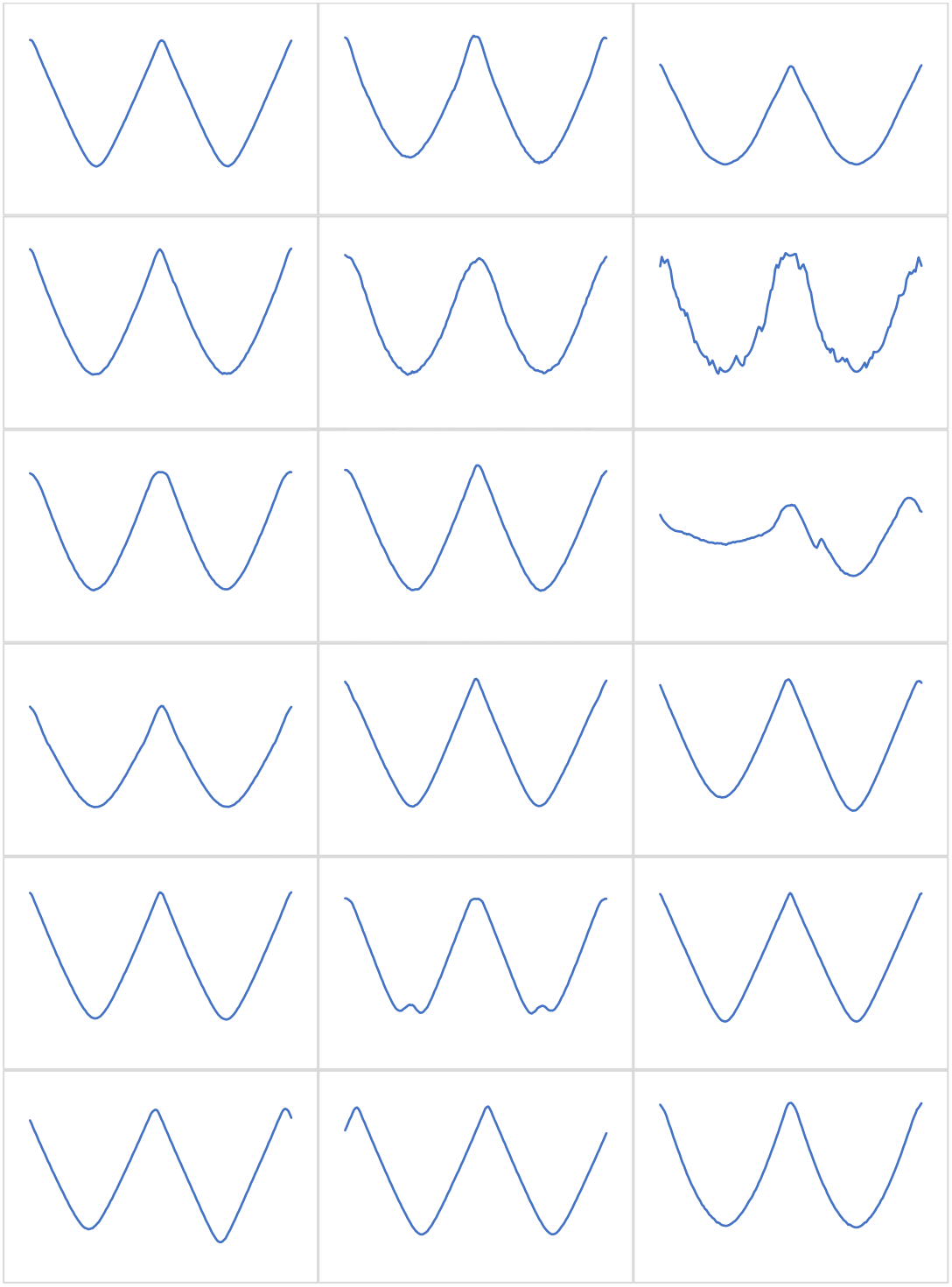
ก.1.1 ตัวอย่างข้อมูล 50Words มี 50 ประเภท แสดงดังรูป ก.1

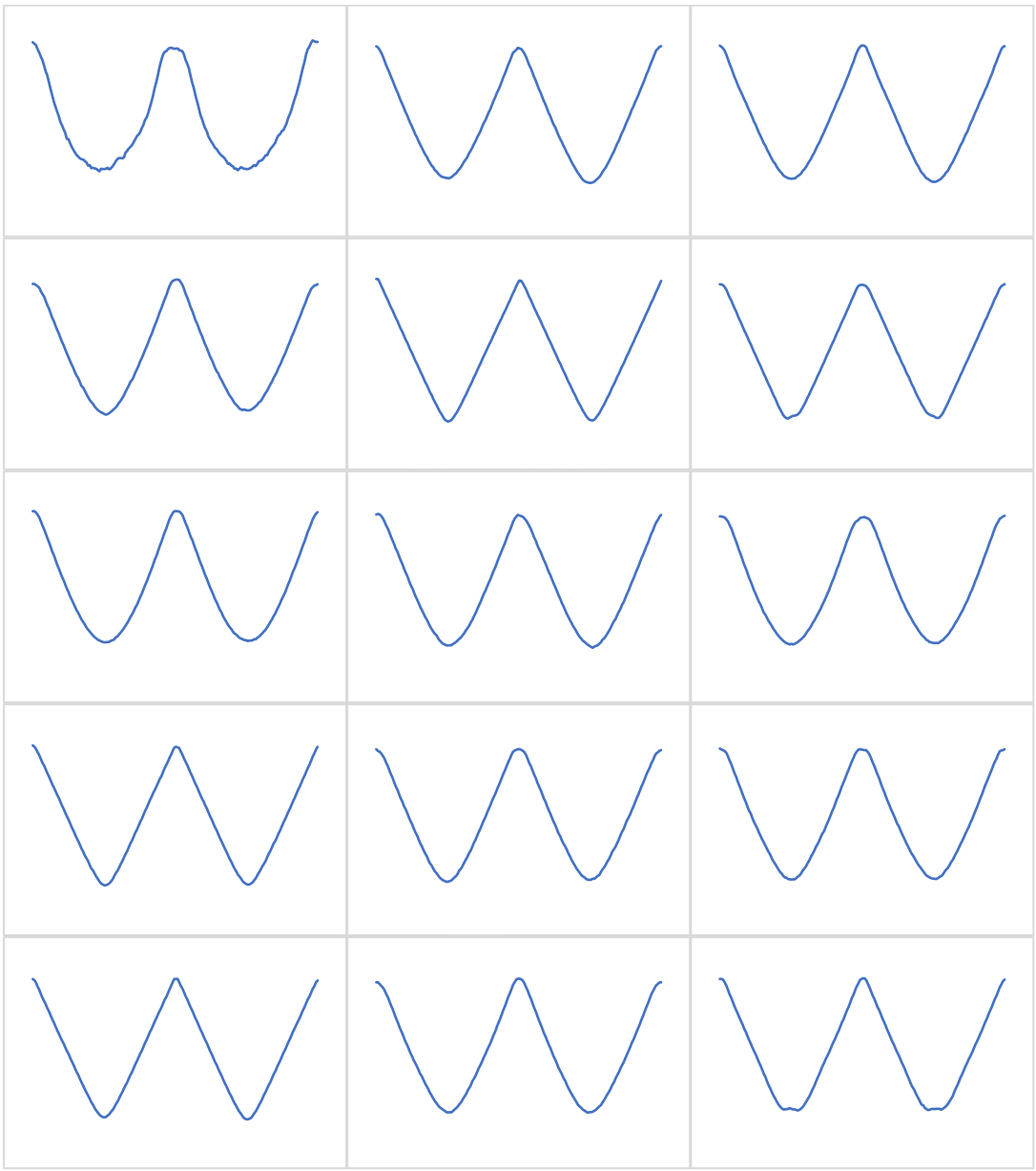


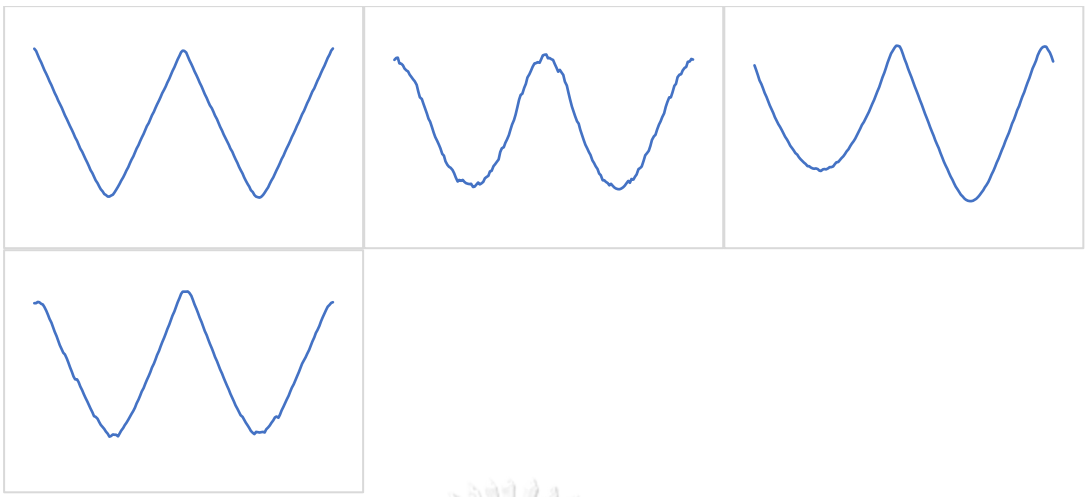


รูปที่ ก.1 ชุดข้อมูล 50Words

ก.1.2 ตัวอย่างข้อมูล Adiac มี 37 ประเภท แสดงดังรูป ก.2

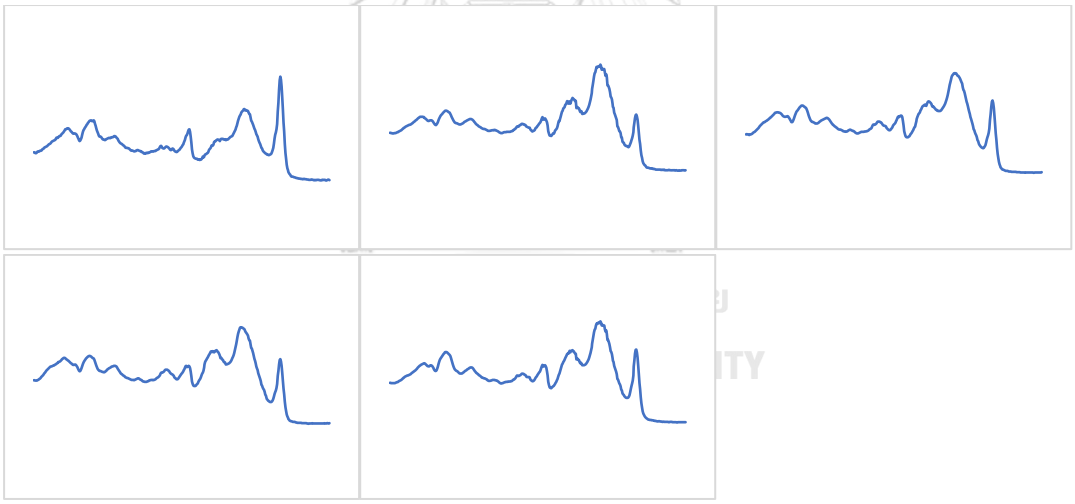






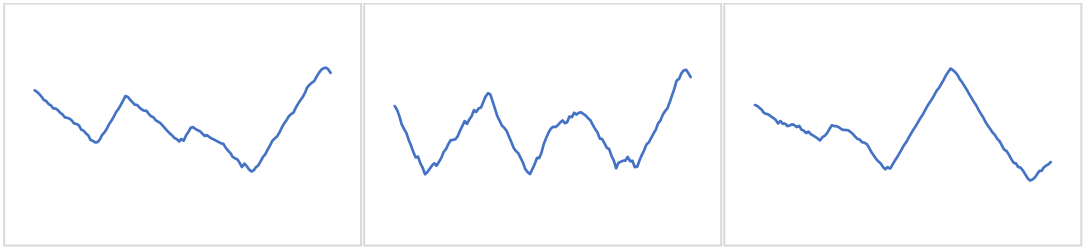
รูปที่ ก.2 ชุดข้อมูล Adiac

ก.1.3 ตัวอย่างข้อมูล Beef มี 5 ประเภท แสดงดังรูป ก.3



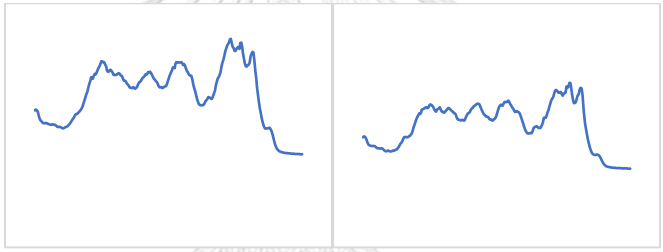
รูปที่ ก.3 ชุดข้อมูล Beef

ก.1.4 ตัวอย่างข้อมูล CBF มี 3 ประเภท แสดงดังรูป ก.4



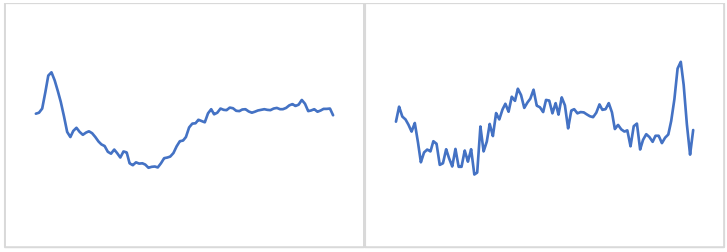
รูปที่ ก.4 ชุดข้อมูล CBF

ก.1.5 ตัวอย่างข้อมูล Coffee มี 2 ประเภท แสดงดังรูป ก.5



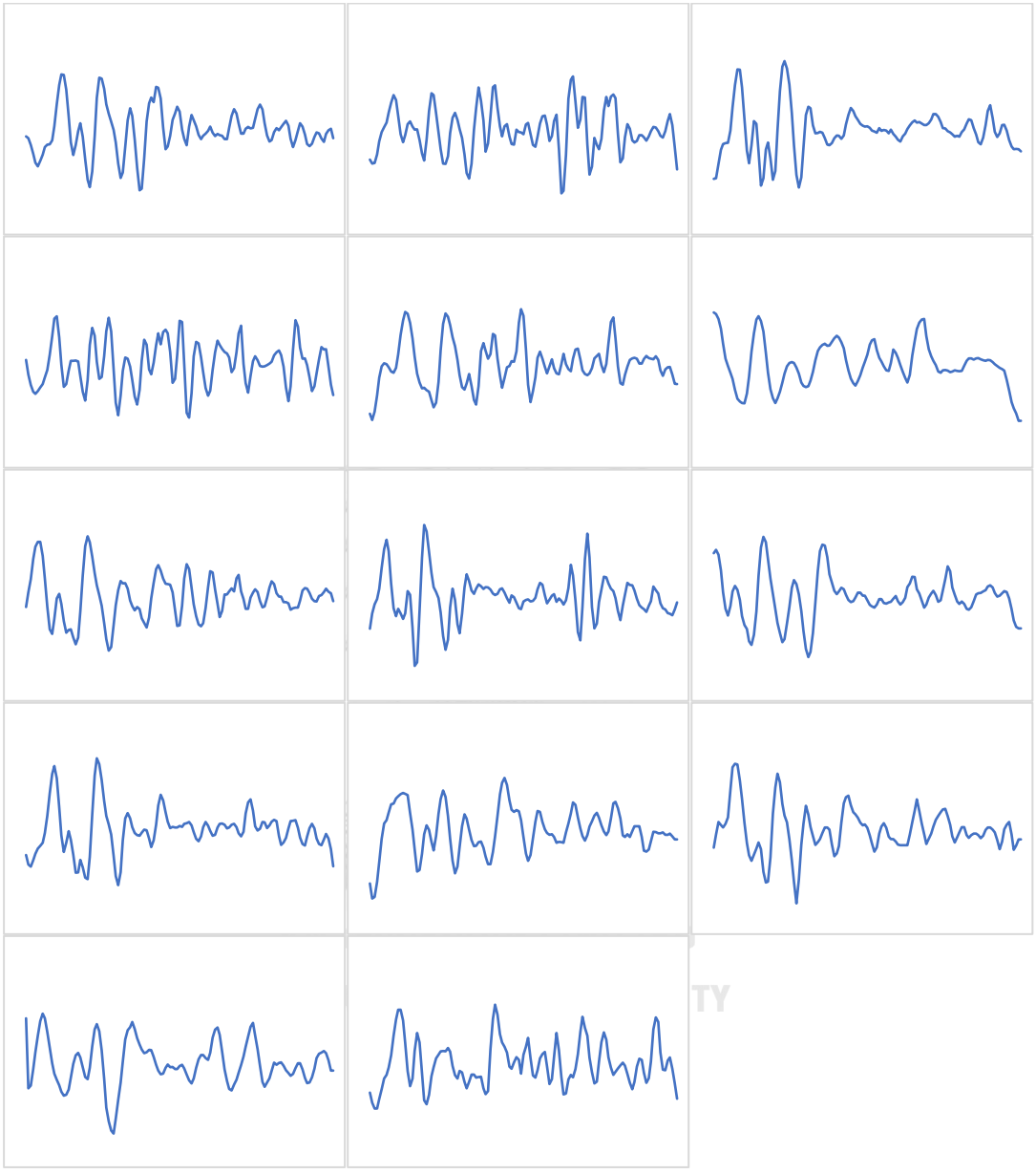
รูปที่ ก.5 ชุดข้อมูล Coffee

ก.1.6 ตัวอย่างข้อมูล ECG200 มี 2 ประเภท แสดงดังรูป ก.6



รูปที่ ก.6 ชุดข้อมูล ECG200

ก.1.7 ตัวอย่างข้อมูล FaceAll มี 14 ประเภท แสดงดังรูป ก.7

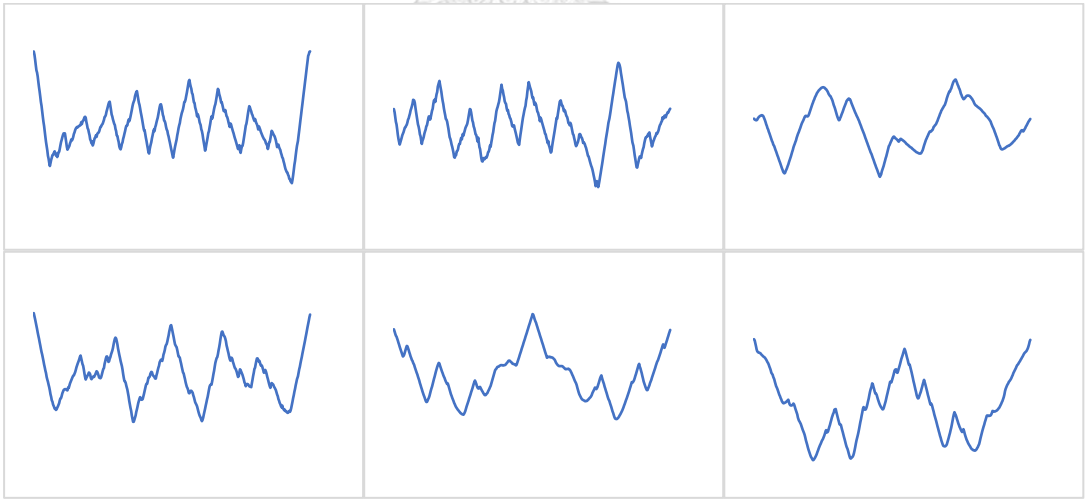


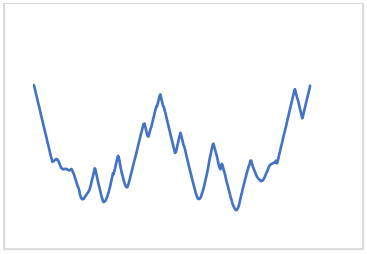
รูปที่ ก.7 ชุดข้อมูล FaceAll

ก.1.8 ตัวอย่างข้อมูล FaceFour มี 4 ประเภท แสดงดังรูป ก.8

รูปที่ ก.8 ชุดข้อมูล FaceFour

ก.1.9 ตัวอย่างข้อมูล FISH มี 7 ประเภท แสดงดังรูป ก.9



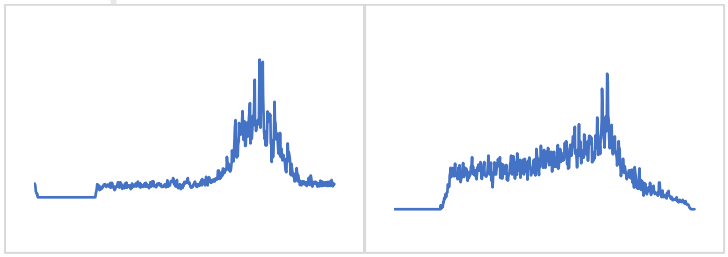


รูปที่ ก.9 ชุดข้อมูล FISH

ก.1.10 ตัวอย่างข้อมูล Gun-Point มี 2 ประเภท แสดงดังรูป ก.10

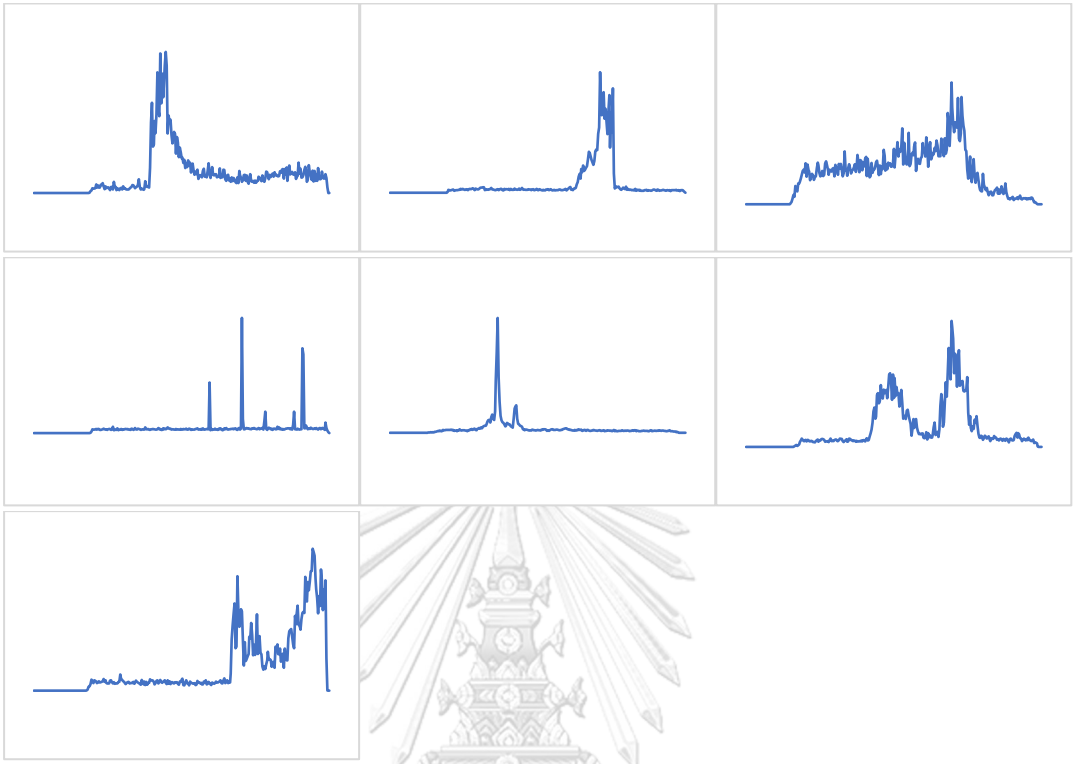
รูปที่ ก.10 ชุดข้อมูล Gun-Point

ก.1.11 ตัวอย่างข้อมูล Lightning2 มี 2 ประเภท แสดงดังรูป ก.11



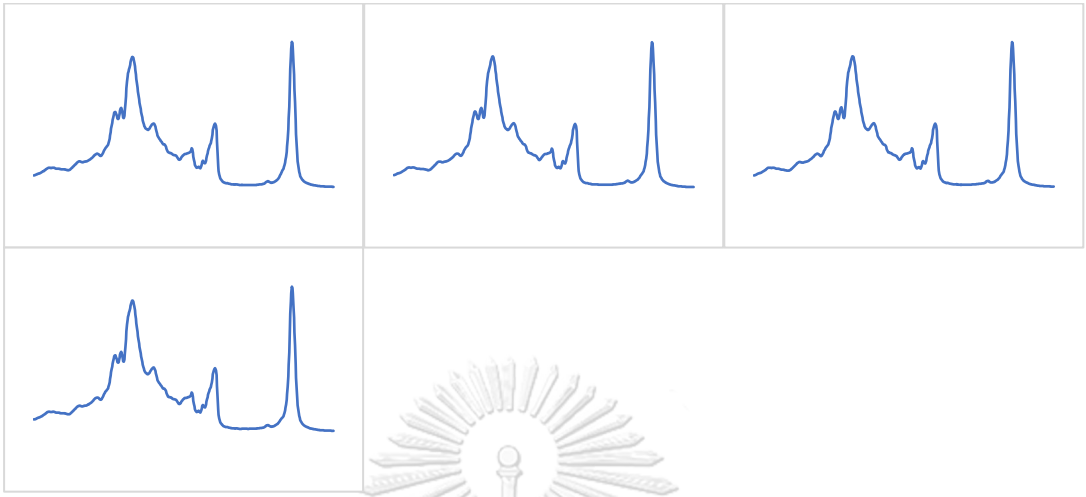
รูปที่ ก.11 ชุดข้อมูล Lightning2

ก.1.12 ตัวอย่างข้อมูล Lightning7 มี 7 ประเภท แสดงดังรูป ก.12



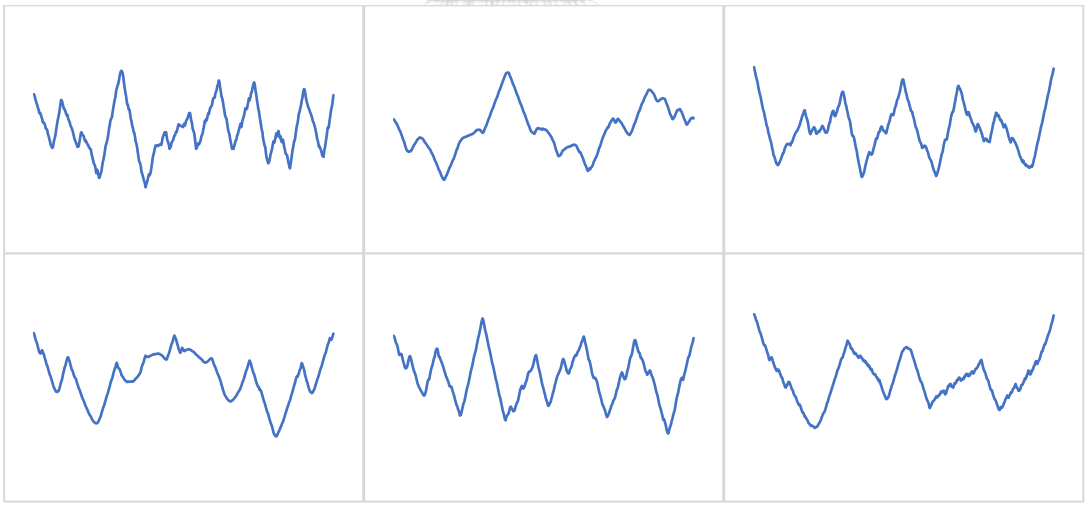
รูปที่ ก.12 ชุดข้อมูล Lightning7

ก.1.13 ตัวอย่างข้อมูล OliveOil มี 4 ประเภท แสดงดังรูป ก.13



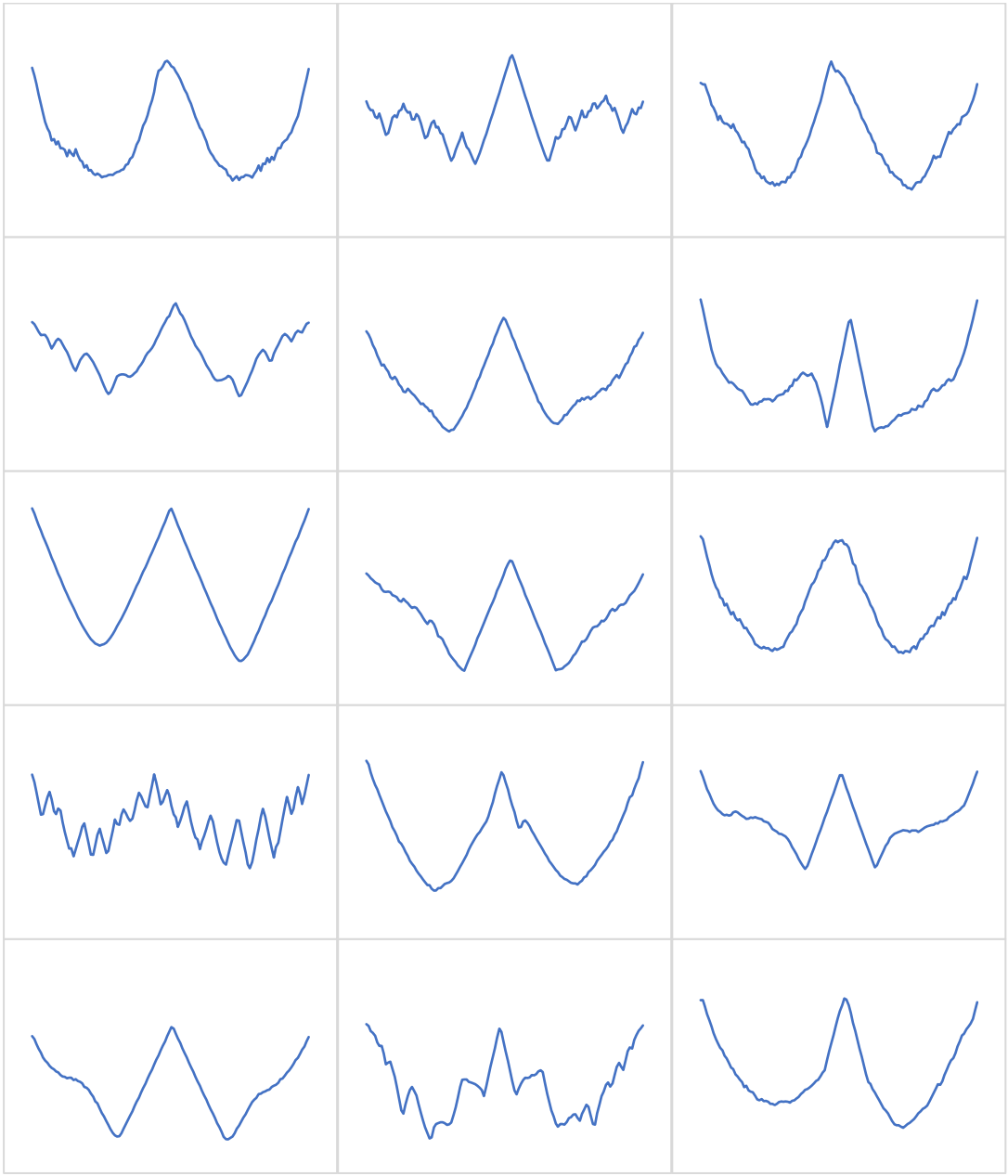
รูปที่ ก.13 ชุดข้อมูล OliveOil

ก.1.14 ตัวอย่างข้อมูล OSULeaf มี 6 ประเภท แสดงดังรูป ก.14



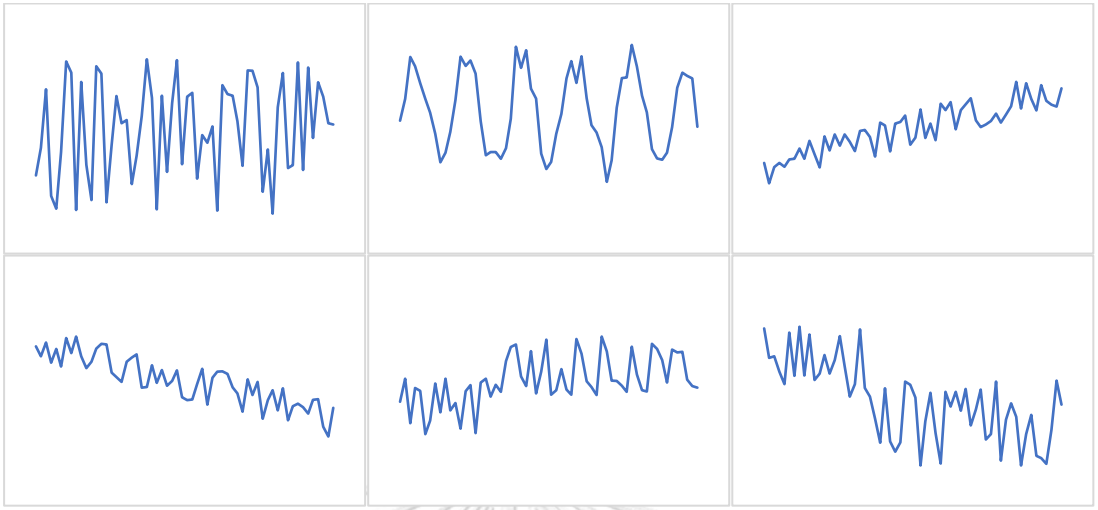
รูปที่ ก.14 ชุดข้อมูล OSULeaf

ก.1.15 ตัวอย่างข้อมูล SwedishLeaf มี 15 ประเภท แสดงดังรูป ก.15



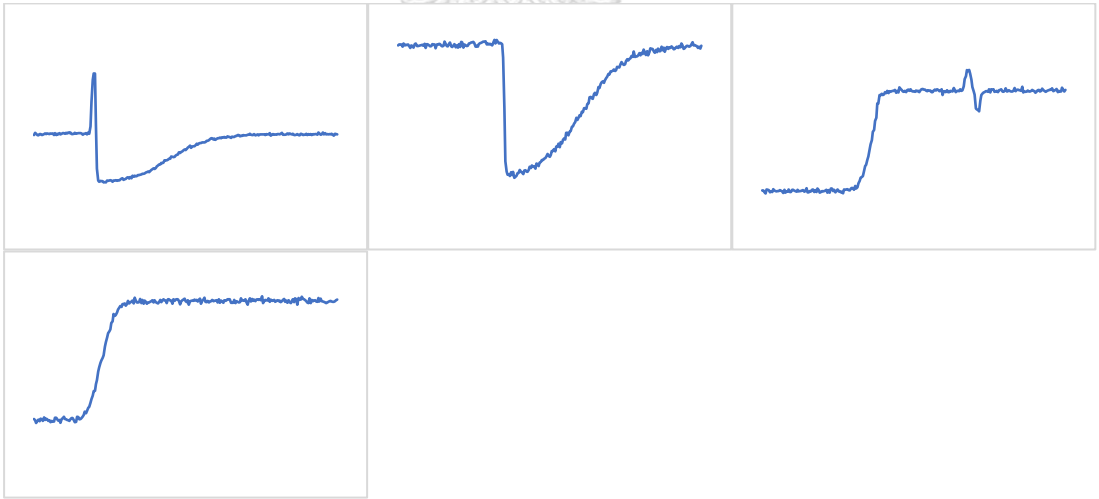
รูปที่ ก.15 ชุดข้อมูล SwedishLeaf

ก.1.16 ตัวอย่างข้อมูล synthetic control มี 6 ประเภท แสดงดังรูป ก.16



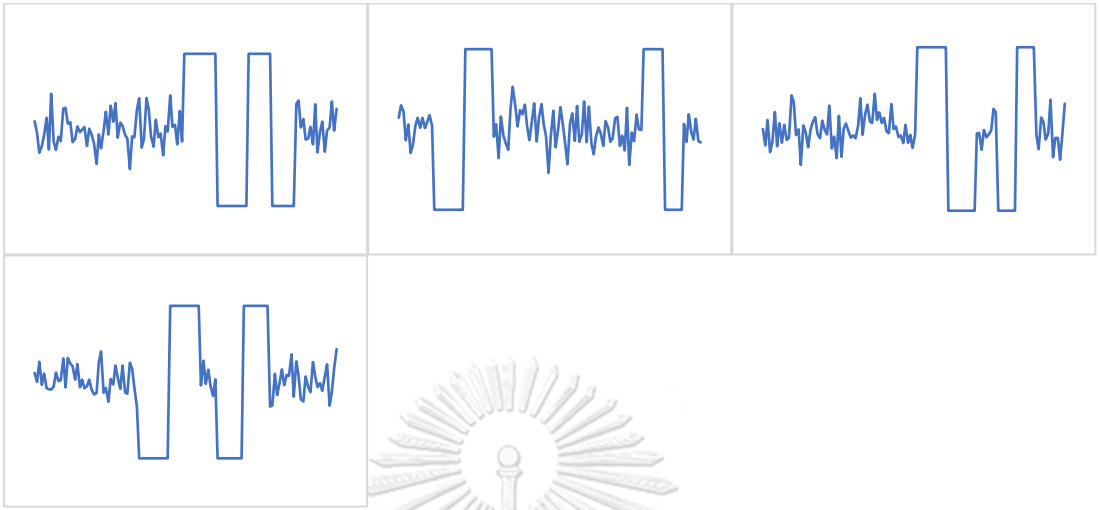
รูปที่ ก.16 ชุดข้อมูล synthetic control

ก.1.17 ตัวอย่างข้อมูล Trace มี 4 ประเภท แสดงดังรูป ก.17



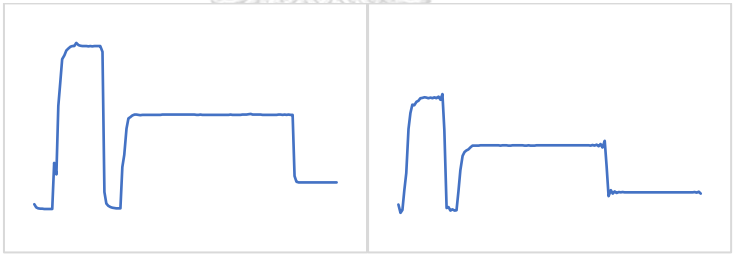
รูปที่ ก.17 ชุดข้อมูล Trace

ก.1.18 ตัวอย่างข้อมูล Two Patterns มี 4 ประเภท แสดงดังรูป ก.18



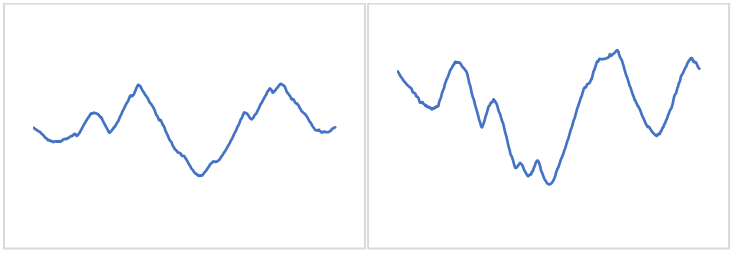
รูปที่ ก.18 ชุดข้อมูล Two Patterns

ก.1.19 ตัวอย่างข้อมูล wafer มี 2 ประเภท แสดงดังรูป ก.19



รูปที่ ก.19 ชุดข้อมูล wafer

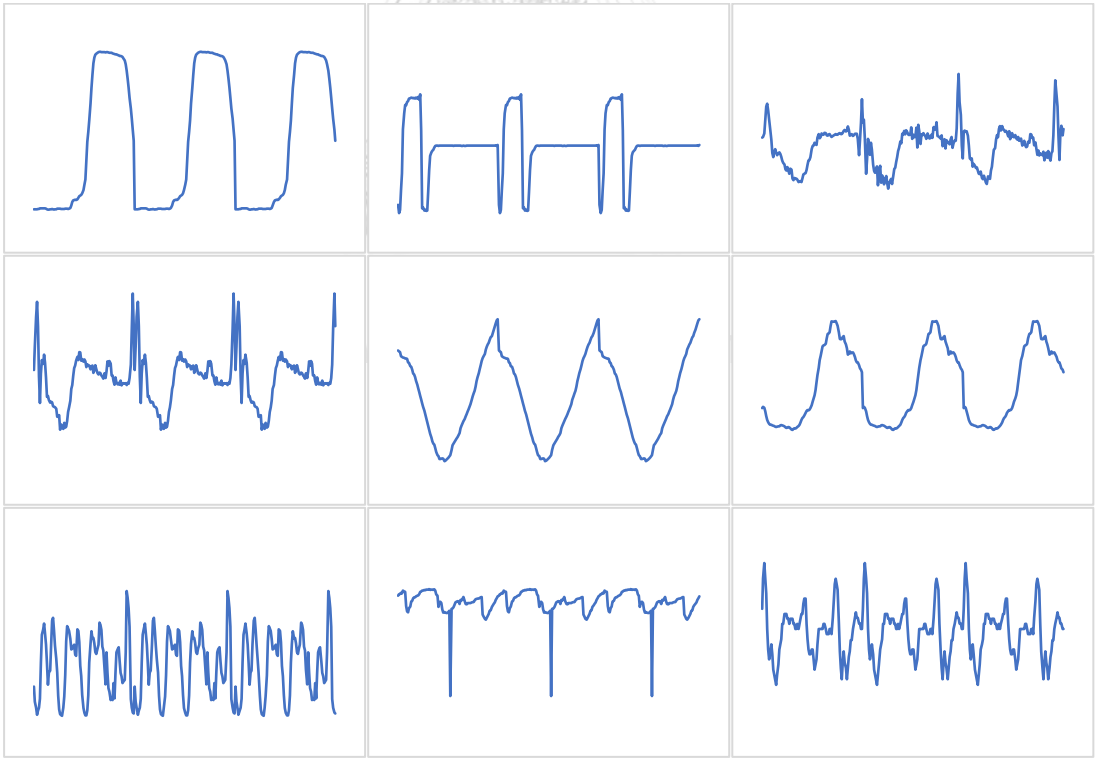
ก.1.20 ตัวอย่างข้อมูล yoga มี 2 ประเภท แสดงดังรูป ก.20

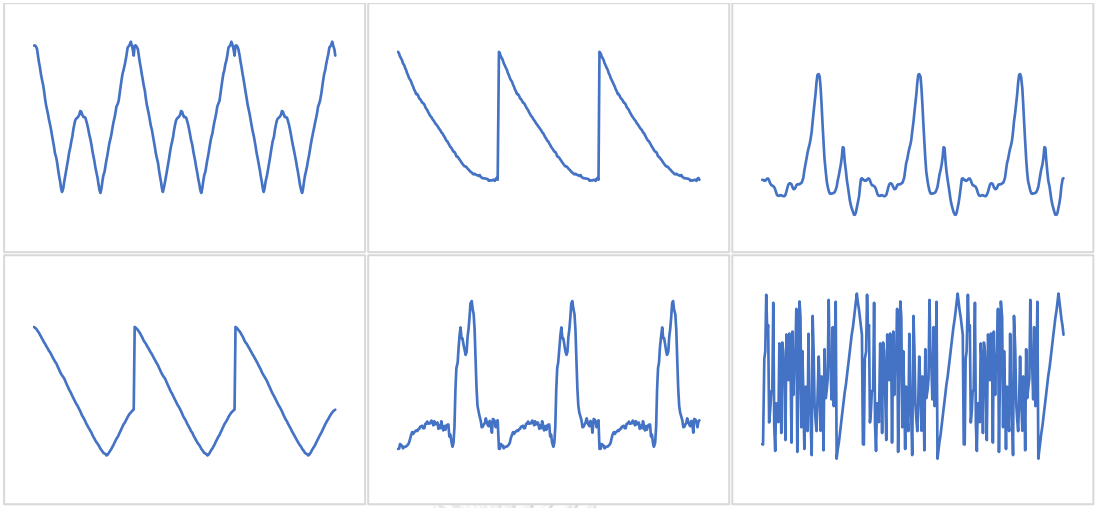


รูปที่ ก.20 ชุดข้อมูล yoga

ก.2 ชุดข้อมูลที่สอง

ข้อมูลชุดที่สองมีจำนวนข้อมูล 30 อนุกรม และมีความยาว 300 จุด ประกอบไปด้วยข้อมูล 15 ประเภท แสดงดังรูป ก.21

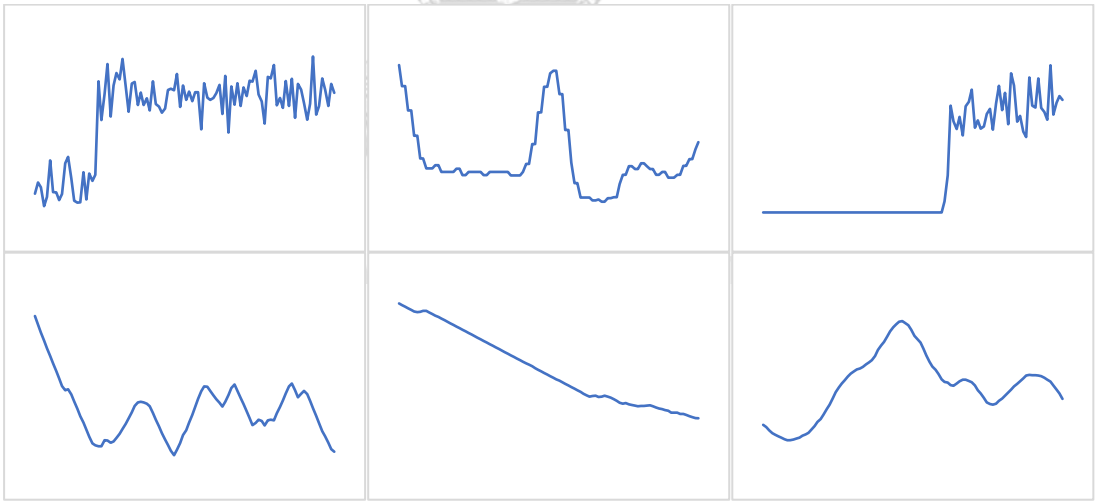


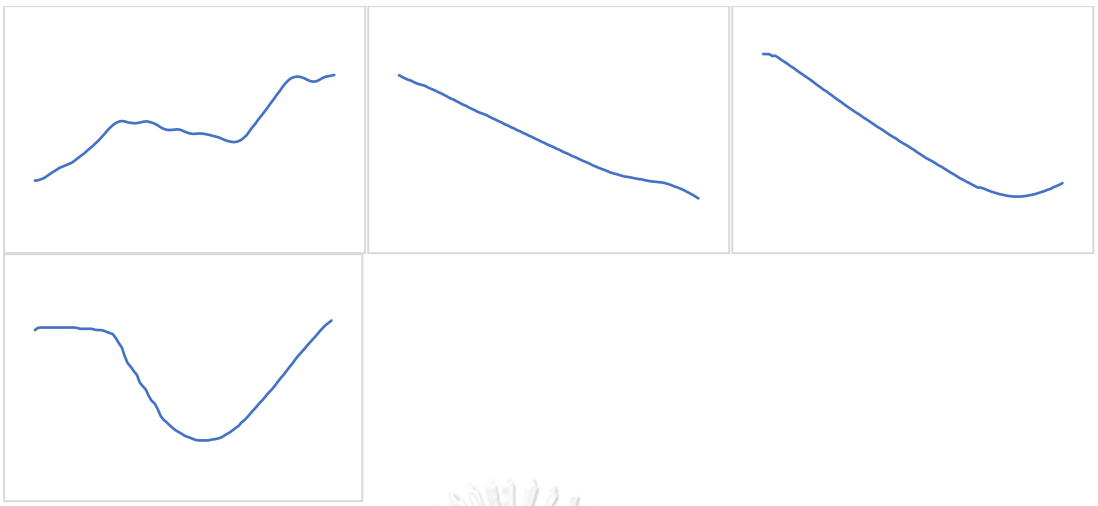


รูปที่ ก.21 ชุดข้อมูลที่สอง

ก.3 ชุดข้อมูลที่สามและสี่

เป็นชุดข้อมูลที่มีความยาว 100 จุด ประกอบไปด้วยข้อมูล 10 ประเภท แสดงดังรูป ก.22





รูปที่ ก.22 ชุดข้อมูลที่สามและสี่

ประวัติผู้เขียนวิทยานิพนธ์

นายกีกก้อง ศิริสัมพันธ์ เกิดเมื่อวันที่ 4 มีนาคม พศ.2537 สำเร็จการศึกษาระดับมัธยมศึกษาจากโรงเรียนฤทธิยะวรรณาลัย จากนั้นทำการศึกษาต่อที่คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2555 และสำเร็จการศึกษาปริญญาวิทยาศาสตร์บัณฑิต สาขาวิทยาการคอมพิวเตอร์ ในปีการศึกษา 2559 และเข้าศึกษาในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2559