การหาแผ่นแบบเฉลี่ยสำหรับการจำแนกประเภทข้อมูลอนุกรมเวลาที่รวดเร็วและแม่นยำ

นายพงศกร เสถียรวิริยคุณ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2558

Fast and Accurate Template Averaging for Time Series Classification

Mr. Phongsakorn Sathianwiriyakhun

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 2015

หัวข้อวิทยานิพนธ์	การหาแผ่นแบบเฉลี่ยสำหรับการจำแนกประเภทข้อมูลอนุกรมเวลาที่รวดเร็วและแม่นยำ
โดย	นายพงศกร เสถียรวิริยคุณ
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ดร.ดวงดาว วิชาดากุล)

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

.....กรรมการ
(ดร.พีรพล เวทีกูล)

.....กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร.กฤษณะ ไวยมัย)

พงศกร เสถียรวิริยคุณ : การหาแผ่นแบบเฉลี่ยสำหรับการจำแนกประเภทข้อมูลอนุกรมเวลาที่รวดเร็วและแม่นยำ (Fast and Accurate Template Averaging for Time Series Classification) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร.โชติรัตน์ รัตนามหัทธนะ, 82 หน้า.

ข้อมูลอนุกรมเวลาเป็นข้อมูลที่น่าสนใจในการทำเหมืองข้อมูลอันเนื่องจากข้อมูลมีลำดับอย่างชัดเจนในตัวเอง การจำแนกประเภทข้อมูลอนุกรมเวลาเป็นวิธีการหนึ่งของการทำเหมืองข้อมูลอนุกรมเวลาที่ใช้อย่างแพร่หลายในหลากหลายสาขาวิชา เช่น สาขาการแพทย์ สาขาการเงิน หรือด้านอุตสาหกรรม งานวิจัยมากมายจึงได้เกิดขึ้นเพื่อพัฒนาการจำแนกประเภทข้อมูลอนุกรมเวลาให้มีความถูกต้องแม่นยำมากยิ่งขึ้น วิธีการหนึ่งที่ได้รับความนิยมและสามารถจำแนกประเภทได้ความแม่นยำสูง คือ การจำแนกประเภทข้อมูลแบบเพื่อนบ้านใกล้สุดลำดับที่ 1 โดยใช้การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง วิธีการนี้เป็นการเปรียบเทียบระยะทางข้อมูลเรียนรู้กับข้อมูลสอบถาม เพื่อกำหนดคลาสให้กับข้อมูลสอบถาม อย่างไรก็ตามการคำนวณระยะทางแบบไดนามิกไทม์วอร์ปปิงใช้เวลาสูงดังนั้นหากข้อมูลมีจำนวนมากจะทำให้เกิดข้อจำกัดในการทำงาน จากข้อจำกัดที่เกิดขึ้นทำให้มีงานวิจัยเกี่ยวกับการลดเวลาที่ใช้ในการคำนวณโดยลดจำนวนข้อมูลเรียนรู้ที่ใช้ในการวิเคราะห์ลง วิธีหนึ่งที่น่าสนใจคือการสร้างแผ่นแบบเฉลี่ยขึ้นเพื่อเป็นตัวแทนข้อมูลเรียนรู้ ส่งผลให้ลดข้อมูลเรียนรู้ที่ใช้ในการวิเคราะห์ลงได้มาก อย่างไรก็ตามแผ่นแบบเฉลี่ยเพียงแผ่นแบบเดียวต่อคลาสไม่เพียงพอต่อการได้ความแม่นยำสูง จึงมีงานวิจัยที่ใช้การจัดกลุ่มข้อมูลเข้ามาช่วยในการแยกข้อมูลในคลาสออกเป็นกลุ่มย่อยแล้วทำการสร้างแผ่นแบบหลายแผ่นแบบต่อหนึ่งคลาส แต่ทว่าการจัดกลุ่มข้อมูลเพื่อแยกข้อมูลออกเป็นกลุ่มย่อยนั้นมีความยากในการปรับค่าตัวแปรให้เหมาะสม และใช้เวลาในการจัดกลุ่มสูงเพื่อให้ได้ความแม่นยำที่สูง งานวิจัยนี้จึงมีแนวคิดที่จะนำเสนอวิธีการสร้างแผ่นแบบเฉลี่ยที่ปรับปรุงการจัดกลุ่มข้อมูลนั้นให้เหลือเพียงค่าตัวแปรเดียวที่ปรับได้ง่ายและใช้เวลาในการจัดกลุ่มน้อยลง ซึ่งจากผลการทดลองวิธีการที่นำเสนอสามารถลดเวลาที่ใช้ในการจัดกลุ่มข้อมูลก่อนสร้างแผ่นแบบลงได้มากเปรียบเทียบกับวิธีการล่าสุดในปัจจุบัน และยังคงความแม่นยำในการจำแนกประเภทข้อมูลไว้ได้

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ปีการศึกษา 2558

5770236321 : MAJOR COMPUTER ENGINEERING

KEYWORDS: TIME SERIES CLASSIFICATION / DYNAMIC TIME WARPING / TIME SERIES TEMPLATE AVERAGING

PHONGSAKORN SATHIANWIRIYAKHUN: Fast and Accurate Template Averaging for Time Series Classification. ADVISOR: ASST. PROF. CHOTIRAT RATANAMAHATANA, Ph.D., 82 pp.

Time series data is an interested data for data mining fields because the data have specific order within itself. Time series data classification is one of the data mining techniques that is used in many domains such as medical, financial, and industrial. Therefore, many researches have been focused on improving accuracy for time series data classification. One of the popular and accurate methods of time series data classification is one-nearest neighbor classification using Dynamic Time Warping (DTW) as a distance measure. This method calculates a distance between training data and testing data with an objective to assign class to unknown instances in the test data. However, Dynamic Time Warping distance measure requires large computation time, becoming a limitation for large datasets. Hence, many researches have attempted to reduce computation time by reducing the size of the training dataset. One of the methods is to build an average template to represent each class of the training data, so that it can reduce the number of training data for classification. Nevertheless, one template per class is insufficient to achieve high accuracy. As a remedy, some researches have attempted to use clustering techniques to split the data and build multiple templates per class. However, those algorithms still suffer from many predefined and hard-to-set parameters, while some require high computation time for high accuracy results. Therefore, this thesis work proposed a faster template averaging method that improves the data splitting process, and has only one easy-to-set parameter. From the experiments, the proposed method can reduce computation time in building templates compared to the state-of-the-art method and still have high accuracy.

Department: Computer Engineering Student's Signature .....

Field of Study: Computer Engineering Advisor's Signature .....

Academic Year: 2015

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงได้ด้วยความกรุณาจาก ผู้ช่วยศาสตราจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ อาจารย์ที่ปรึกษา ผู้คอยให้คำปรึกษา ให้แง่คิด ทั้งในด้านวิชาการและด้านอื่น ๆ และเป็นผู้ตรวจทานแก้ไขทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี ขอขอบพระคุณเป็นอย่างสูงมา ณ ที่นี้

ขอขอบพระคุณ ดร.พีรพล เวทีกูล ผู้ยินยอมให้เข้าร่วมฟังการประชุมงาน เพื่อศึกษาหาความรู้เพิ่มเติมในด้านของการทำเหมืองข้อมูล และให้เกียรติเป็นกรรมการสอบวิทยานิพนธ์ รวมถึงชี้แนะแนวทางในการปรับปรุงวิทยานิพนธ์ให้มีคุณภาพมากยิ่งขึ้น

ขอขอบพระคุณ ดร.ดวงดาว วิชาดากุล และรองศาสตราจารย์ ดร.กฤษณะ ไวยมัย ผู้ให้เกียรติเป็นประธานกรรมการสอบวิทยานิพนธ์และกรรมการสอบวิทยานิพนธ์ ที่ชี้แนะแนวทางในการปรับปรุงวิทยานิพนธ์ให้มีคุณภาพยิ่งขึ้น

ขอบคุณ โครงการทุนอัจฉริยะคืนรัง ของภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่สนับสนุนค่าใช้จ่ายในการทำงานวิจัย

สุดท้ายนี้ ขอขอบคุณครอบครัวทุกคน ที่คอยเป็นกำลังใจในการทำงานวิจัยนี้เสมอมา จนกระทั่งวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี

สารบัญ

หน้า

หน้า

หน้า

สารบัญตาราง

หน้า

สารบัญรูป

หน้า

หน้า

บทที่ 1 บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในการทำงานหรือกระทำสิ่งใดสิ่งหนึ่งให้สำเร็จนั้น "ข้อมูล (Data)" นับเป็นตัวแปรสำคัญที่ส่งผลต่อการตัดสินใจและทิศทางในการทำงานนั้น ๆ หากขาดแคลนข้อมูล การตัดสินใจต่าง ๆ ในการทำงานนั้น อาจเกิดการผิดพลาดหรือนำไปสู่ทิศทางที่ทำให้งานเกิดปัญหาขึ้น ส่งผลให้การที่จะทำงานนั้น ๆ ให้สำเร็จมีความยากอย่างยิ่ง เพราะฉะนั้น "ข้อมูล" นับเป็นสิ่งสำคัญที่ส่งผลในการทำงานอย่างยิ่งยวด อย่างไรก็ดีข้อมูลนั้นมีมากมายหลายประเภท และมีการใช้งานที่แตกต่างกันออกไป ตัวอย่างของประเภทข้อมูลได้แก่ ตัวเลข รูปภาพ เสียง หรือตัวอักษร เป็นต้น

ข้อมูลที่เป็นที่แพร่หลายในการทำงานต่าง ๆ ประเภทหนึ่งคือ ข้อมูลอนุกรมเวลา (Time Series Data) [1-7] ในข้อมูลอนุกรมเวลา 1 ตัวจะประกอบไปด้วยข้อมูลย่อย ๆ ที่แสดงถึงสิ่งใดสิ่งหนึ่ง โดยข้อมูลย่อย ๆ แต่ละตัวนั้นจะมีลำดับก่อนและหลังในกลุ่มของข้อมูลย่อย ๆ นั้น ซึ่งทำให้ข้อมูลมีการเกี่ยวข้องกับลำดับหรือเวลา สามารถนำมาสร้างกราฟที่แสดงความต่อเนื่องของข้อมูลได้ จึงส่งผลต่อลักษณะของข้อมูลที่จะมีลักษณะแตกต่างไปจากข้อมูลปกติทั่วไป ตัวอย่างข้อมูลประเภทนี้ได้แก่ ข้อมูลคลื่นไฟฟ้า ข้อมูลเสียง ข้อมูลหุ้น หรือข้อมูลยอดขายสินค้า เป็นต้น

แต่เพียงข้อมูลดิบนั้น ไม่สามารถนำมาใช้ในการทำงานได้ทันที จะต้องมีการวิเคราะห์และค้นหาความรู้ (Knowledge) ที่ถูกซ่อนอยู่ภายในข้อมูลนั้น ๆ ตัวอย่างเช่น หากต้องการทราบว่าข้อมูลเสียงที่มีอยู่เป็นเสียงของบุคคลใด จำเป็นต้องมีการวิเคราะห์ข้อมูลหาคุณลักษณะของเสียง แล้วจึงทำการเทียบเคียงกับบุคคลที่ต้องการทดสอบว่าเป็นบุคคลเดียวกันกับเจ้าของเสียงหรือไม่ ซึ่งโดยทั่วไปหากมีจำนวนข้อมูลไม่มาก หรือสามารถวิเคราะห์ได้โดยง่าย ผู้เชี่ยวชาญสามารถทำการวิเคราะห์เพื่อหาความรู้ที่ถูกซ่อนอยู่ในข้อมูลได้ แต่หากข้อมูลชุดนั้น ๆ มีจำนวนมากมายมหาศาลหรือยากแก่การวิเคราะห์ ผู้เชี่ยวชาญไม่สามารถวิเคราะห์ได้ถูกต้องแม่นยำหรือใช้เวลาสูงมากในการวิเคราะห์ การทำเหมืองข้อมูล (Data Mining) จึงเกิดขึ้น เพื่อใช้วิธีการต่าง ๆ ค้นหาและสกัดความรู้ออกมาจากข้อมูลจำนวนมากหรือข้อมูลที่ยากแก่การวิเคราะห์นั้น ให้มีความถูกต้องแม่นยำในการวิเคราะห์และใช้เวลาน้อยลง

ข้อมูลอนุกรมเวลาก็มีการทำเหมืองข้อมูลเช่นกัน ซึ่งการวิเคราะห์ข้อมูลจะแตกต่างจากการวิเคราะห์ข้อมูลปกติ เนื่องด้วยการมีลำดับของข้อมูลย่อย ๆ ในข้อมูลแต่ละตัว เมื่อนำไปเปรียบเทียบกับข้อมูลตัวอื่นที่ต้องการเปรียบเทียบหรือค้นหาความหมายบางอย่าง ลำดับของข้อมูลจึงมีความสำคัญมากในการวิเคราะห์ การทำเหมืองข้อมูลอนุกรมเวลานั้นมีวิธีการมากมาย ตัวอย่างเช่น

การจำแนกประเภทข้อมูล (Classification) การจัดกลุ่มข้อมูล (Clustering) หรือการหาสิ่งผิดปกติของข้อมูล (Anomaly Detection) เป็นต้น [3, 5, 8, 9]

การจำแนกประเภทข้อมูล (Classification) เป็นวิธีการหนึ่งที่นิยมใช้ในการทำเหมืองข้อมูล รวมถึงการทำเหมืองข้อมูลอนุกรมเวลาเช่นกัน โดยการจำแนกประเภทข้อมูลนั้นคือการกำหนดประเภทหรือ “คลาส” ของข้อมูลที่ไม่ทราบคลาส ให้อยู่ในคลาสใดคลาสหนึ่ง วิธีการที่มีการนำมาใช้นั้นมีอยู่หลากหลายวิธี รวมถึงวิธีการจำแนกประเภทข้อมูลใหม่ที่มีผู้คิดค้นขึ้นมาอยู่เสมอ แต่โดยส่วนใหญ่แล้ววิธีการจำแนกประเภทข้อมูลคือ “การเปรียบเทียบ” ข้อมูลที่ไม่ทราบคลาสเรียกว่า ข้อมูลทดสอบ (Testing Data) กับข้อมูลที่ทราบคลาสของข้อมูลอยู่แล้วเรียกว่า ข้อมูลเรียนรู้ (Training Data) ก่อนทำการเลือกคลาสจากข้อมูลที่ได้ วิธีการหนึ่งที่เป็นที่นิยมคือการเปรียบเทียบความคล้ายคลึง (Similarity) หรือระยะทางระหว่างข้อมูล (Distance) จากนั้นกำหนดคลาสของข้อมูลที่ต้องการทราบคลาสเป็นคลาสเดียวกับข้อมูลที่มีความคล้ายคลึงกันมากที่สุด เรียกวิธีการนี้ว่าการจำแนกประเภทแบบเพื่อนบ้านใกล้ที่สุดลำดับที่หนึ่ง (1-Nearest Neighbor Classification) [10-12]

การวัดความคล้ายคลึงหรือระยะทางระหว่างข้อมูลอนุกรมเวลา เพื่อนำมาใช้ในการเปรียบเทียบข้อมูลนั้นมีด้วยกันหลากหลายวิธีการ วิธีการที่เป็นที่นิยมประกอบด้วยตัววัดความคล้ายแบบยุคลิด (Euclidean Distance Metric) [13, 14] และการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง (Dynamic Time Warping Distance Measure) [9, 12, 15, 16] ซึ่งทั้งสองประเภทมีข้อดีแตกต่างกัน โดยตัววัดความคล้ายแบบยุคลิดมีการคำนวณโดยตรงระหว่างข้อมูลแต่ละจุดข้อมูลย่อย จึงคำนวณง่าย ไม่ซับซ้อน และสามารถบ่งบอกความแตกต่างของข้อมูลได้เด่นชัด แต่ไม่สามารถรองรับสิ่งรบกวน (Noise) และการแปรผันของเวลาในข้อมูล (Time shift) ได้ ทำให้การจำแนกประเภทข้อมูลที่มีสิ่งรบกวนหรือการแปรผันของเวลาเกิดความผิดพลาด ซึ่งการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงสามารถแก้ไขในเรื่องของสิ่งรบกวนและการแปรผันของเวลาในข้อมูลได้ในระดับหนึ่ง ทำให้ได้ความแม่นยำในการจำแนกประเภทข้อมูลในข้อมูลส่วนใหญ่สูงกว่าตัววัดความคล้ายแบบยุคลิด

อย่างไรก็ตามการคำนวณระยะทางแบบไดนามิกไทม์วอร์ปปิงมีความซับซ้อนมากกว่าตัววัดความคล้ายแบบยุคลิด ส่งผลให้เวลาที่ใช้ในการคำนวณระยะทางสูงกว่ามาก เมื่อนำมาหาระยะทางเพื่อใช้กับการจำแนกประเภทข้อมูลจึงส่งผลให้เวลาที่ใช้เพิ่มสูงขึ้น หากจำนวนข้อมูลเรียนรู้กับข้อมูลทดสอบมีจำนวนมากเวลาก็จะมากขึ้นเป็นทวีคูณ เพราะจำเป็นต้องคำนวณระยะทางระหว่างข้อมูลทดสอบทุกข้อมูลเทียบกับข้อมูลเรียนรู้ทุกข้อมูลเพื่อค้นหาข้อมูลที่เป็นข้อมูลเรียนรู้ที่ใกล้ที่สุด ทำให้เวลาที่ใช้ในการจำแนกประเภทข้อมูลทดสอบทั้งหมดด้วยการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงสูงมาก

อีกปัญหาที่อาจเกิดขึ้นกับวิธีการไดนามิกไทม์วอร์ปปิงได้คือ การที่ไดนามิกไทม์วอร์ปปิงมีการคำนวณเปรียบเทียบจุดเวลาที่ห่างกันมากเกินไป จนเกิดความไม่สมเหตุสมผลในการเปรียบเทียบ ส่งผลให้ข้อมูลบางประเภทมีความแม่นยำในการจำแนกประเภทลดลง หรือไม่เหมาะสมในการเปรียบเทียบกับลักษณะของข้อมูลบางชนิด เพื่อแก้ปัญหาที่เกิดขึ้น จึงได้มีผู้นำเสนอการกำหนดเงื่อนไขโดยรวม (Global Constraint) [16-18] เพื่อทำการจำกัดวิธีการเปรียบเทียบของไดนามิกไทม์วอร์ปปิงให้สามารถเปรียบเทียบระหว่างจุดข้อมูลที่ไม่ห่างกันมากจนเกินไป

นอกจากนี้ได้มีผู้พัฒนาฟังก์ชันขอบเขตล่างของไดนามิกไทม์วอร์ปปิงขึ้น (Lower Bounding Function of Dynamic Time Warping) [19-21] เพื่อทำการลดจำนวนของข้อมูลที่ต้องใช้ไดนามิกไทม์วอร์ปปิงในการเปรียบเทียบเมื่อใช้ในการจำแนกประเภทข้อมูลแบบเพื่อนบ้านใกล้ที่สุด โดยใช้ค่าฟังก์ชันขอบเขตล่างมาประมาณค่าของระยะทางไดนามิกไทม์วอร์ปปิงก่อนทำการคำนวณจริง ๆ เฉพาะบางข้อมูลเท่านั้น แต่อย่างไรก็ตามหากข้อมูลมีจำนวนมาก การคำนวณฟังก์ชันขอบเขตล่างก็มากเช่นกัน รวมถึงต้องคำนวณระยะทางแบบไดนามิกไทม์วอร์ปปิงกับข้อมูลที่คัดเลือกเช่นเดิม อาจทำให้ลดเวลาในการทำงานได้ไม่เพียงพอ

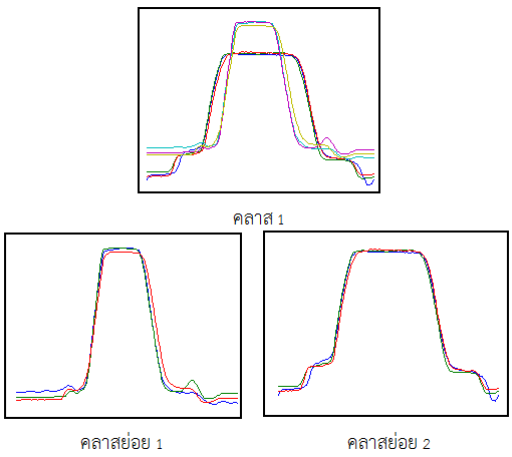
เพื่อแก้ปัญหาเวลาในการจำแนกประเภทด้วยการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงที่สูง แม้จะใช้ฟังก์ชันขอบเขตล่างแล้ว ได้มีผู้คิดค้นวิธีการต่าง ๆ มากมายที่จะสามารถลดระยะเวลาในการคำนวณลง โดยที่ยังคงความแม่นยำที่สูงดังเดิมหรือเทียบเคียงกับความแม่นยำเดิมไว้ได้ หนึ่งในนั้นคือวิธีการสร้างแผ่นแบบเฉลี่ย (Template Averaging) [22-26] ซึ่งเป็นวิธีหนึ่งสำหรับลดจำนวนข้อมูลเรียนรู้ที่มีจำนวนมากลง โดยสร้างข้อมูลขึ้นเป็นแผ่นแบบที่สามารถเป็นตัวแทนข้อมูลจำนวนมากในกลุ่มข้อมูลเรียนรู้ได้ แผ่นแบบที่สร้างขึ้นจะมีอย่างน้อยหนึ่งแผ่นแบบสำหรับข้อมูลในแต่ละคลาส

ในส่วนของการสร้างแผ่นแบบเฉลี่ยของข้อมูลอนุกรมเวลานั้น ไม่สามารถทำการเฉลี่ยค่าโดยทั่วไปแบบจุดต่อจุดได้ดังเช่นข้อมูลปกติ เนื่องด้วยข้อมูลอนุกรมเวลามีลำดับในข้อมูลย่อยแต่ละตัว อีกทั้งหากข้อมูลเกิดสิ่งรบกวนหรือการแปรผันของเวลาในข้อมูล จะส่งผลให้แผ่นแบบที่สร้างขึ้นมีความผิดเพี้ยนและส่งผลต่อความแม่นยำในการจำแนกประเภทข้อมูลอย่างมาก ดังนั้นจึงมีผู้เสนอการหาค่าเฉลี่ยข้อมูลอนุกรมเวลาด้วยการใช้การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง เพื่อทำการค้นหาเส้นทางการปรับแนว หรือวิถีของการวอร์ป (Warping paths) [22] ที่ใช้ในการเฉลี่ยข้อมูลในแต่ละจุดเวลา ทำให้แผ่นแบบเฉลี่ยที่สร้างขึ้นมีความแม่นยำในการจำแนกประเภทที่สูงกว่าแผ่นแบบเฉลี่ยที่สร้างขึ้นจากการเฉลี่ยแบบจุดต่อจุด

วิธีการสร้างแผ่นแบบเฉลี่ยได้มีการปรับปรุงเพื่อพัฒนาประสิทธิผลอย่างต่อเนื่อง ด้วยวิธีการต่าง ๆ เพื่อให้แผ่นแบบเฉลี่ยที่สร้างขึ้นแสดงลักษณะข้อมูลได้ครบถ้วนและสามารถนำไปใช้ในการจำแนกประเภทข้อมูลได้อย่างแม่นยำ โดยวิธีที่มีผู้เสนอมาประกอบไปด้วยวิธีการจับคู่ข้อมูลเพื่อทำการสร้างแผ่นแบบเฉลี่ย เพื่อปรับปรุงผลที่เกิดจากลำดับการสร้างแผ่นแบบเฉลี่ย และการปรับปรุง

วิธีการหาค่าเฉลี่ยของข้อมูลด้วยวิธีการทางการทำเหมืองข้อมูลต่าง ๆ แต่สุดท้ายได้มีผู้คิดค้นวิธีการสร้างแผ่นแบบเฉลี่ยด้วยการเฉลี่ยค่าของข้อมูลตามแกนเวลา เพื่อสามารถทำการหาค่าเฉลี่ยของข้อมูลทุกตัวพร้อมกันได้ ซึ่งช่วยเพิ่มความแม่นยำของแผ่นแบบในการจำแนกประเภทและความเร็วในการสร้างแผ่นแบบเฉลี่ยเป็นอย่างมาก [25]

อย่างไรก็ตามการสร้างแผ่นแบบเฉลี่ยโดยการเฉลี่ยข้อมูลอนุกรมเวลาทั้งหมดในคลาสเพื่อสร้างแผ่นแบบเฉลี่ยขึ้นนั้น ยังไม่สามารถได้ค่าความแม่นยำสูงเท่ากับการจำแนกประเภทโดยการใช้ข้อมูลเรียนรู้ทั้งหมด เนื่องด้วยลักษณะของข้อมูลที่ข้อมูลในคลาสเดียวกันอาจมีคุณลักษณะบางประการที่มีความแตกต่างจากกันอยู่ หรืออาจมีคลาสย่อย (Subclass) [26] ที่ทำให้ข้อมูลมีความแตกต่างกันมาก และอาจมีข้อมูลบางตัวที่เป็นข้อมูลที่มีสิ่งรบกวนมากจนรูปแบบของข้อมูลมีความผิดเพี้ยนไป เมื่อนำไปสร้างแผ่นแบบเฉลี่ยขึ้น คุณลักษณะบางประการได้ถูกลบหายไปหรือแสดงออกเพียงเล็กน้อย หรือมีการแสดงออกที่ผิดเพี้ยนไปจากที่ควรจะเป็น ส่งผลให้ความแม่นยำเมื่อทำการจำแนกประเภทกับข้อมูลทดสอบลดลง ตัวอย่างของข้อมูลที่ข้อมูลมีความแตกต่างกัน แต่จัดอยู่ในคลาสเดียวกัน แสดงได้ดังรูปที่ 1.1



รูปที่ 1.1 ตัวอย่างข้อมูลที่มีคลาสย่อยอยู่ภายในคลาสหลัก (ที่มา Y. Chen et al. [27])

เพื่อแก้ปัญหาที่การสร้างแผ่นแบบเฉลี่ยเพียงแผ่นแบบเดียวไม่สามารถแทนที่ข้อมูลทั้งหมดได้ จึงได้มีผู้คิดค้นการสร้างแผ่นแบบเฉลี่ยจำนวนมากกว่าหนึ่งแผ่นแบบเพื่อแทนที่ข้อมูลในหนึ่งคลาสขึ้นโดยใช้วิธีการจัดกลุ่มข้อมูล (Clustering) [8, 28] เพื่อทำการแบ่งแยกข้อมูลในคลาสเดียวกันออกเป็นกลุ่มย่อย ๆ ตั้งแต่สองกลุ่มขึ้นไป ก่อนที่จะทำการสร้างแผ่นแบบเฉลี่ยที่เป็นตัวแทนของแต่ละกลุ่มย่อย ๆ นี้ และรวมข้อมูลแผ่นแบบเฉลี่ยทั้งหมดเป็นแผ่นแบบเฉลี่ยของคลาสนั้น ๆ

ถึงแม้ว่าการจัดกลุ่มข้อมูลในแต่ละคลาสก่อนทำการสร้างแผ่นแบบเฉลี่ยจะสามารถปรับปรุงให้ค่าความแม่นยำของการจำแนกประเภทโดยใช้การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงสูงขึ้น เวลาในการสร้างแผ่นแบบเพื่อให้ได้แผ่นแบบที่สามารถให้ความแม่นยำสูงได้นั้นกลับเพิ่มขึ้นมาก อีกยัง

มีค่าตัวแปรบางค่าที่ส่งผลให้ข้อมูลแผ่นแบบเฉลี่ยไม่คงที่ จึงจำเป็นต้องทำการค้นหาค่าตัวแปรที่ให้ผลดี ส่งผลให้เวลาในการคำนวณของสร้างแผ่นแบบเฉลี่ยด้วยวิธีนี้สูงขึ้นอีก

ในงานวิจัยนี้จึงได้นำเสนอวิธีการสร้างแผ่นแบบเฉลี่ยวิธีใหม่ ซึ่งสามารถทำการหาคลาสย่อยของข้อมูล และทำการสร้างแผ่นแบบเฉลี่ยโดยไม่ต้องใช้ตัวแปรในการทำงาน โดยมีการพัฒนารูปแบบการสร้างแผ่นแบบเฉลี่ยจากวิธีหาค่าเฉลี่ยแบบไดนามิกไทม์วอร์ปปิงแบรีเซนเตอร์ (Dynamic Time Warping Barycenter Averaging หรือ DBA) [25] เป็นหลักในการทำงาน แต่มีการค้นหาคลาสย่อยของข้อมูลที่รวดเร็ว และสามารถใช้ค่าตัวแปรที่ง่ายในการปรับค่าในการทำงานในแต่ละขั้นตอน โดยที่ยังสามารถคงค่าความแม่นยำที่สูงเมื่อนำไปจำแนกประเภทโดยใช้การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงได้

1.2 วัตถุประสงค์ของงานวิจัย

เพื่อพัฒนาวิธีการสร้างแผ่นแบบของข้อมูลอนุกรมเวลาให้มีความเร็วในการทำงานมากขึ้น โดยสามารถคงความแม่นยำของการจำแนกประเภทข้อมูลไว้ได้

1.3 ขอบเขตของงานวิจัย

ข้อมูลอนุกรมเวลาที่นำมาใช้ทดสอบผลของวิธีการที่นำเสนอมาจาก UCR Time Series Classification Archive ในข้อมูลแต่ละชุดจะประกอบไปด้วยข้อมูลเรียนรู้ ข้อมูลทดสอบ คลาสของข้อมูลเรียนรู้แต่ละตัว คลาสของข้อมูลทดสอบแต่ละตัว จำนวนข้อมูลในแต่ละชุดข้อมูล ความยาวของข้อมูลอนุกรมเวลาแต่ละตัว และค่าความผิดพลาดในการคำนวณด้วยการใช้วิธีการจำแนกประเภทด้วยวิธีการจำแนกข้อมูลแบบเพื่อนบ้านใกล้สุดลำดับที่ 1 โดยใช้การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง ในกรณีที่ใช้ข้อมูลเรียนรู้ทุกข้อมูล ทดสอบกับข้อมูลทดสอบทุกข้อมูล

1.4 ประโยชน์ที่ได้รับจากงานวิจัย

ได้วิธีการสร้างแผ่นแบบเฉลี่ยที่สามารททำการสร้างแผ่นแบบได้อย่างรวดเร็ว โดยสามารถคงความแม่นยำในวิธีการจำแนกประเภทด้วยวิธีการจำแนกข้อมูลแบบเพื่อนบ้านใกล้สุดลำดับที่ 1 โดยใช้การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงได้

1.5 วิธีดำเนินงานวิจัย

1. ศึกษาเกี่ยวกับการทำเหมืองข้อมูลอนุกรมเวลา
2. ศึกษาเกี่ยวกับการจำแนกประเภทข้อมูลอนุกรมเวลาด้วยวิธีการจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่ 1 โดยใช้ระยะทางแบบไดนามิกไทม์วอร์ปปิง
3. ทดลองสร้างโปรแกรมเพื่อทำการจำแนกประเภทข้อมูลให้ได้ผลลัพธ์เท่ากับผลเฉลยจากข้อมูลที่มีอยู่

4. ศึกษาเกี่ยวกับวิธีการสร้างแผ่นแบบเฉลี่ย จากงานวิจัยที่เกี่ยวข้อง สรุปข้อดีข้อเสียของแต่ละงานวิจัย และกำหนดงานวิจัยที่จะใช้เป็นงานวิจัยหลักในการเปรียบเทียบกับวิธีการที่จะสร้างขึ้น โดยเลือกจากผลลัพธ์ที่ดีที่สุด

5. ศึกษาวิธีการขั้นตอนทางด้านการจำแนกประเภทและการจัดกลุ่มข้อมูล เพื่อค้นหา ดัดแปลง และเป็นแนวทางในการสร้างวิธีการใหม่ในการสร้างแผ่นแบบเฉลี่ย

6. ออกแบบและสร้างโปรแกรมสำหรับสร้างแผ่นแบบเฉลี่ย

7. ทดสอบโปรแกรมที่สร้างขึ้น โดยเก็บผลลัพธ์เป็นเวลาที่ใช้ในการสร้างแผ่นแบบและความแม่นยำจากการจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่ 1 โดยใช้ระยะทางแบบไดนามิกไทม์วอร์ปปิง

8. เปรียบเทียบผลลัพธ์จากวิธีการที่สร้างขึ้นกับผลลัพธ์จากงานวิจัยที่เกี่ยวข้องที่ดีที่สุด จากนั้นทำการปรับปรุงโปรแกรมด้วยวิธีการทางการเหมืองข้อมูลอนุกรมเวลาเพื่อให้ได้ผลลัพธ์ที่ดีที่สุด

9. วิเคราะห์และสรุปผลการทดลอง

10. สรุปผล เรียบเรียง และจัดทำวิทยานิพนธ์

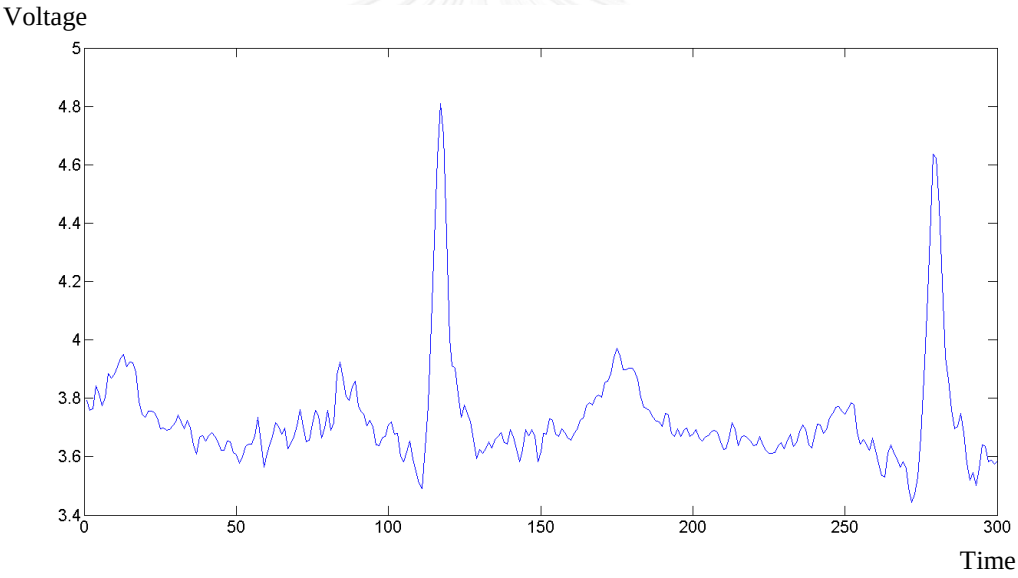
1.6 ผลงานวิจัยที่ได้ตีพิมพ์

- H. Sivaraks, P. Sathianwiriyakhun, T. Janyalikit, and C. A. Ratanamahatana, “Accurate Time series Classification Using Partial Dynamic Time Warping,” in Second International Conference On Advances in Applied Science and Environmental Technology, ASET 2015, pp.31-35, 2015.
- P. Sathianwiriyakhun, T. Janyalikit, and C. A. Ratanamahatana, “Fast and accurate template averaging for time series classification,” in 2016 8th International Conference on Knowledge and Smart Technology, KST 2016, pp. 49-54, 2016.
- T. Janyalikit, P. Sathianwiriyakhun, H. Sivaraks, and C. A. Ratanamahatana, “An enhanced support vector machine for faster time series classification,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9621, pp. 616-625, 2016.

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ข้อมูลอนุกรมเวลา (Time Series Data)

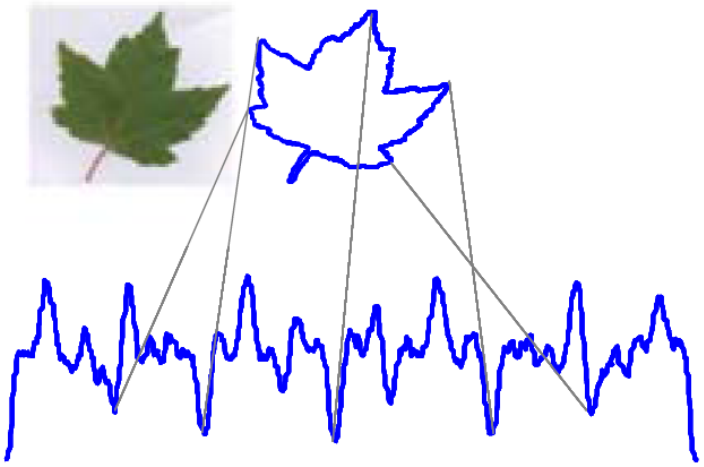
ข้อมูลอนุกรมเวลา คือข้อมูลของสิ่งใดสิ่งหนึ่งที่ทำการบันทึกค่าไว้ตามจุดต่าง ๆ ของเวลา กล่าวคือข้อมูลอนุกรมเวลาคือข้อมูลของสิ่งใดสิ่งหนึ่งที่มีการเปลี่ยนแปลงตามเวลา ข้อมูลย่อยแต่ละข้อมูลที่ทำการบันทึกไว้ก็จะเชื่อมกับเวลาใดเวลาหนึ่ง ทำให้ข้อมูลมีลำดับก่อนและหลัง ข้อมูลอนุกรมเวลาสามารถพบได้ทั่วไปในชีวิตประจำวัน เพราะข้อมูลของสิ่งต่าง ๆ ย่อมมีการเปลี่ยนแปลงไปตามกาลเวลาอยู่เสมอ ตัวอย่างเช่น ข้อมูลหุ้น (Stock market) [29] ข้อมูลคลื่นไฟฟ้าหัวใจของมนุษย์ (Electrocardiogram หรือ ECG) [3, 5] ข้อมูลระดับน้ำในเขื่อน [30] หรือข้อมูลการเปลี่ยนแปลงของอุณหภูมิ [31] เป็นต้น ดังแสดงในรูปที่ 2.1 ซึ่งเป็นกราฟข้อมูลคลื่นไฟฟ้าหัวใจของมนุษย์ สังเกตได้ว่าค่าของข้อมูลซึ่งคือแรงดันไฟฟ้า (Voltage) ที่ได้ทำการวัดมีการเปลี่ยนแปลงอยู่ตลอดเวลา



รูปที่ 2.1 ตัวอย่างข้อมูลคลื่นไฟฟ้าหัวใจ ที่เป็นข้อมูลอนุกรมเวลา (ที่มา Sivaraks และ Ratanamahatana [5])

ในส่วนของข้อมูลบางชนิดนั้นสามารถนำมาแปลงข้อมูลให้อยู่ในรูปแบบของข้อมูลอนุกรมเวลา เพื่อให้สามารถทำการวิเคราะห์ข้อมูลในรูปแบบของข้อมูลอนุกรมเวลาได้ ข้อมูลส่วนใหญ่ที่มีการแปลงเป็นข้อมูลอนุกรมเวลานั้นจะอยู่ในรูปแบบของภาพ ตัวอย่างเช่น ภาพใบไม้ [4] ภาพใบหน้าของมนุษย์ [2] หรือตัวอักษร [32] เป็นต้น ซึ่งการแปลงข้อมูลลักษณะนี้เป็นข้อมูลอนุกรมเวลานั้นจะใช้คุณลักษณะต่าง ๆ ของภาพสกัดเป็นค่าข้อมูลในข้อมูลอนุกรมเวลา ตัวอย่างดังรูปที่ 2.2 แสดงการแปลงภาพของใบไม้เป็นข้อมูลอนุกรมเวลา โดยพิจารณาจากรอยหยักรอบใบไม้ ตั้งแต่

จุดเริ่มต้นที่กำหนดไปจนถึงจุดสุดท้าย ซึ่งการแปลงข้อมูลลักษณะนี้สามารถตรวจสอบรอยหยักของใบไม้เพื่อทำการจำแนกประเภทได้รวดเร็วกว่าการตรวจสอบจากภาพสองมิติโดยตรง



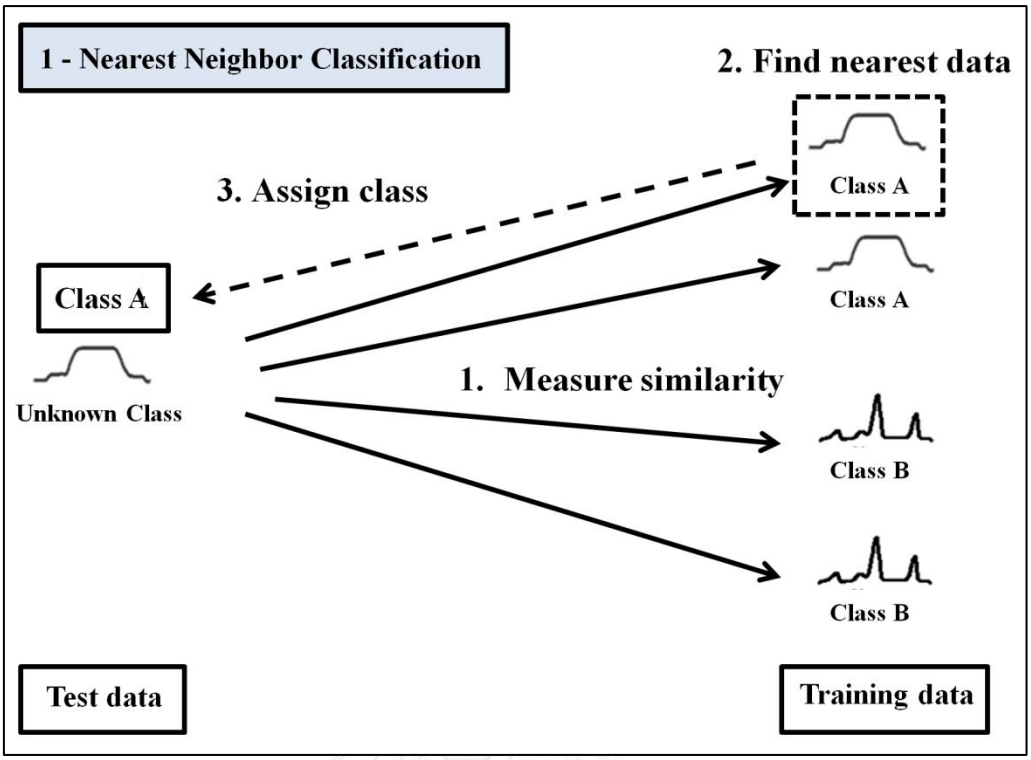
รูปที่ 2.2 การแปลงข้อมูลภาพของใบไม้ เป็นข้อมูลอนุกรมเวลา (ที่มา Ratanamahatana และ Keogh [2])

2.2 การจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่ 1 (1-Nearest Neighbor Classification)

การจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่ 1 [10-12] เป็นวิธีการหนึ่งที่เป็นที่นิยมในการจำแนกประเภทข้อมูล วิธีการจำแนกประเภทลักษณะนี้จำเป็นต้องมีข้อมูลเรียนรู้ที่เราทราบประเภทหรือคลาสของข้อมูลอยู่แล้ว เพื่อเป็นแบบอย่างในการเปรียบเทียบกับข้อมูลสอบถามที่เราต้องการทราบคลาส วิธีการจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่ 1 มีวิธีการดังนี้

- ทำการหาความคล้ายหรือระยะทางระหว่างข้อมูลสอบถามหนึ่งตัว กับข้อมูลเรียนรู้ทุกตัวที่มีอยู่ ในที่นี้จะใช้ระยะทางระหว่างข้อมูลเป็นตัววัดความคล้าย
- ทำการค้นหาข้อมูลเรียนรู้ตัวที่มีระยะทางน้อยที่สุดเมื่อเทียบกับข้อมูลสอบถามที่นำมาคำนวณ จากข้อมูลระยะทางระหว่างข้อมูลสอบถามข้อมูลนั้น ๆ กับข้อมูลเรียนรู้ทั้งหมด
- ทำนายประเภทหรือคลาสของข้อมูลสอบถามตัวนั้น ๆ ว่าเป็นคลาสเดียวกันกับข้อมูลเรียนรู้ตัวที่มีระยะทางน้อยที่สุด
- ทำวิธีการทำนายคลาสของข้อมูลสอบถามไปเรื่อย ๆ จนครบข้อมูลสอบถามทุกตัวที่มีอยู่

วิธีการจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่ 1 สามารถแสดงได้ดังรูปที่ 2.3

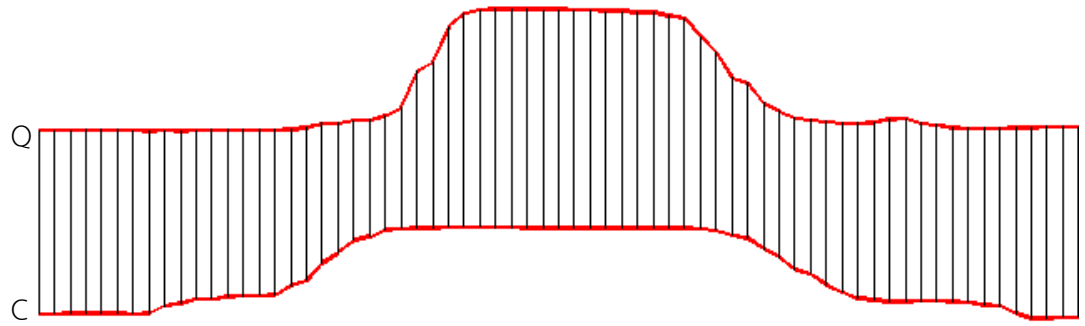


รูปที่ 2.3 วิธีการจำแนกประเภทแบบเพื่อนบ้านใกล้ที่สุดลำดับที่ 1

นอกจากนี้วิธีการจำแนกประเภทเพื่อนบ้านใกล้สุดอาจใช้ลำดับที่มากกว่า 1 ได้เช่นลำดับที่ 3 หรือลำดับที่ 5 ซึ่งจะเป็นการหาข้อมูลเรียนรู้ที่ใกล้ที่สุดลำดับต่อไปมาช่วยในการกำหนดคลาสของข้อมูลสอบถาม

2.3 ตัววัดความคล้ายแบบยุคลิด (Euclidean Distance Metric)

ตัววัดความคล้ายแบบยุคลิด [13, 14] เป็นวิธีการวัดความคล้ายคลึงกันของข้อมูลรูปแบบหนึ่งด้วยวิธีการวัดระยะทางระหว่างข้อมูล ซึ่งสามารถนำไปใช้กับการวัดระยะทางระหว่างข้อมูลอนุกรมเวลาสองข้อมูลได้เช่นกัน โดยเป็นวิธีที่เรียบง่ายที่สุดของการวัดระยะทางระหว่างข้อมูล ตัววัดความคล้ายแบบยุคลิดจะทำการวัดระยะทางระหว่างข้อมูลโดยตรง เมื่อนำมาปรับใช้กับข้อมูลอนุกรมเวลาจะทำการวัดเปรียบเทียบข้อมูลที่เกิดในจุดเวลาเดียวกันระหว่างข้อมูลอนุกรมเวลาสองข้อมูล ตัวอย่างเช่นข้อมูลในเวลาที่ 10 ของข้อมูลที่ 1 ก็จะทำการเปรียบเทียบกับข้อมูลในเวลาที่ 10 ของข้อมูลที่ 2 เช่นเดียวกัน โดยการหาระยะทางแบบยุคลิดของข้อมูลอนุกรมเวลาสามารถแสดงได้ดังรูปที่ 2.4 โดยเส้นระหว่างข้อมูลอนุกรมเวลาแสดงถึงจุดเวลาที่ทำการเปรียบเทียบค่า



รูปที่ 2.4 วิธีการวัดระยะทางด้วยตัววัดความคล้ายแบบยุคลิด เส้นระหว่างข้อมูลทั้งสองแสดงถึงการเปรียบเทียบกันระหว่างจุดเวลานั้น ๆ

สำหรับการนิยามวิธีการของยุคลิดนั้นสามารถอธิบายได้ดังนี้

- กำหนดข้อมูลอนุกรมเวลา Q เป็นข้อมูลสอบถาม (Query Sequence) ที่มีความยาวเท่ากับ n โดยที่ $Q = q_1, q_2, \dots, q_n$
- กำหนดข้อมูลอนุกรมเวลา C เป็นข้อมูลที่จะใช้ในการเปรียบเทียบ(Candidate Sequence) ที่มีความยาวเท่ากับ n โดยที่ $C = c_1, c_2, \dots, c_n$
- ระยะทางยุคลิดระหว่างข้อมูลอนุกรมเวลา Q และข้อมูลอนุกรมเวลา C สามารถหาได้ดังสมการ

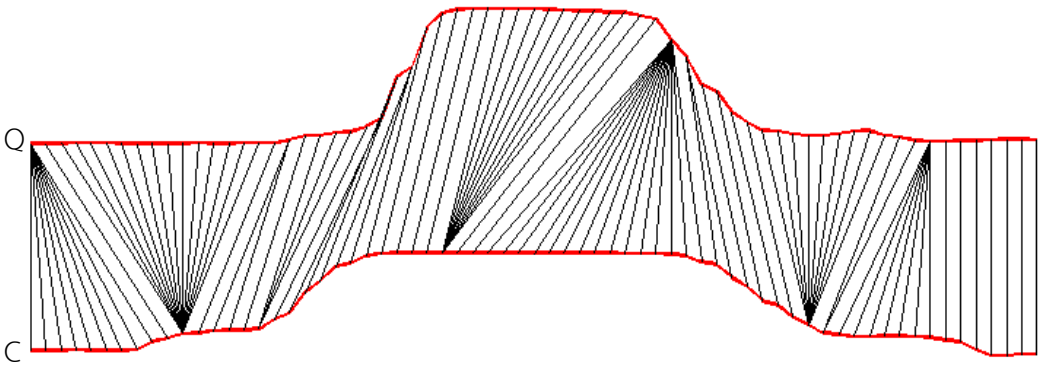
$$Euclidean(Q, C) = \sqrt{\sum_{i=1}^{n}(q_i - c_i)^2} \quad (1)$$

ข้อดีของตัววัดความคล้ายแบบยุคลิดนั้นคือความเร็วในการคำนวณ เนื่องจากเป็นวิธีการที่เรียบง่าย ทำให้ความเร็วในการคำนวณคือ $O(n)$ นอกจากนั้นการคำนวณเปรียบเทียบแบบจุดต่อจุดจะสามารถบอกความแตกต่างระหว่างข้อมูลสองข้อมูลได้ดี เมื่อต้องการแยกข้อมูลอนุกรมเวลาออกจากกัน

2.4 การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง (Dynamic Time Warping Distance Measure)

การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง [9, 12, 15, 16] เป็นวิธีการวัดความคล้ายคลึงกันระหว่างข้อมูลอนุกรมเวลาสองอนุกรม จุดเด่นที่สำคัญของการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงคือการใช้วิธีการกำหนดการพลวัต (Dynamic Programming) ในการ

คำนวณหาเส้นทางการปรับแนวหรือหรือวิถีของการวอร์ป (Warping paths) ก่อนที่จะทำการคำนวณหาระยะทางตามเส้นทางการปรับแนวที่ได้คำนวณไว้ ซึ่งทำให้การปรับแนวของการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงไม่เป็นแบบจุดต่อจุด ส่งผลให้สามารถทำการคำนวณระยะทางระหว่างข้อมูลอนุกรมเวลาที่มีขนาดไม่เท่ากันได้ และสามารถแก้ไขปัญหาของข้อมูลอนุกรมที่มีความแปรผันของเวลารวมถึงสิ่งรบกวนเล็กน้อยได้ ตัวอย่างเช่น ข้อมูลของเสียงที่ผู้พูดมีความเร็วในการพูดไม่เท่ากัน ถ้านำมาเปรียบเทียบระยะทางเพื่อค้นหาว่าผู้พูดเปล่งเสียงข้อความเดียวกันหรือไม่ ไดนามิกไทม์วอร์ปปิงจะคำนวณระยะทางระหว่างข้อมูลทั้งสองได้เหมาะสมกับการเปรียบเทียบมากกว่าการเทียบแบบจุดต่อจุดที่มีผลกระทบของการแปรผันของเวลาด้วย วิธีการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงสามารถแสดงได้ดังรูปที่ 2.5 โดยเส้นระหว่างข้อมูลอนุกรมเวลาแสดงถึงจุดเวลาที่ทำการเปรียบเทียบค่า



รูปที่ 2.5 วิธีการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง เส้นระหว่างข้อมูลทั้งสองแสดงถึงการเปรียบเทียบกันระหว่างจุดเวลานั้น ๆ

ระยะทางแบบไดนามิกไทม์วอร์ปปิงสามารถคำนวณได้จากค่าระยะทางสะสม (Cumulative Distance) ตามเส้นทางการปรับแนว ซึ่งจะทำให้ค่าระยะทางที่ได้มีค่าน้อยที่สุดจากทุกเส้นทางที่เป็นไปได้ โดยสามารถอธิบายวิธีการได้ดังนี้

- กำหนดข้อมูลอนุกรมเวลา Q เป็นข้อมูลสอบถาม (Query Sequence) ที่มีความยาวเท่ากับ m โดยที่ $Q = q_1, q_2, \dots, q_m$
- กำหนดข้อมูลอนุกรมเวลา C เป็นข้อมูลที่จะใช้ในการเปรียบเทียบ (Candidate Sequence) ที่มีความยาวเท่ากับ n โดยที่ $C = c_1, c_2, \dots, c_n$
- สร้างเมทริกซ์ D ขนาด $m\,x\,n$ ขึ้นเพื่อเก็บค่าระยะทางสะสมโดยค่าแต่ละตำแหน่งของเมทริกซ์สามารถคำนวณจากระยะทางของข้อมูลในตำแหน่งนั้นตามสมการ $d(q_i, c_j) =$

$(q_i - c_j)^2$ รวมกับค่าของระยะทางสะสมที่สั้นที่สุดจากตำแหน่งที่ติดกันก่อนหน้า 3 ตำแหน่ง ดังสมการ

$$\gamma(i,j) = d(q_i, c_j) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \quad (2)$$

โดย q_i และ c_j คือ ข้อมูลตำแหน่งที่ i และ j ในข้อมูลอนุกรมเวลา Q และ C ตามลำดับ

- เส้นทางการปรับแนวหรือวิถีการวอร์ป (Warping Path, W) สามารถคำนวณได้เส้นทางของระยะทางสะสมน้อยที่สุดระหว่างสองอนุกรม โดย $W = w_1, w_2, \dots, w_k, \dots, w_K$ และ W_k บ่งบอกถึงข้อมูลในตำแหน่งที่ i และตำแหน่งที่ j ที่ใช้ในการปรับแนวเข้าหากัน โดยค่า i จะอยู่ระหว่างค่า 1 ถึง m และค่า j จะอยู่ระหว่างค่า 1 ถึง n
- ค่าระยะทางแบบไดนามิกไทม์วอร์ปปิงคือค่าของระยะทางสะสมตำแหน่งสุดท้ายดังสมการ

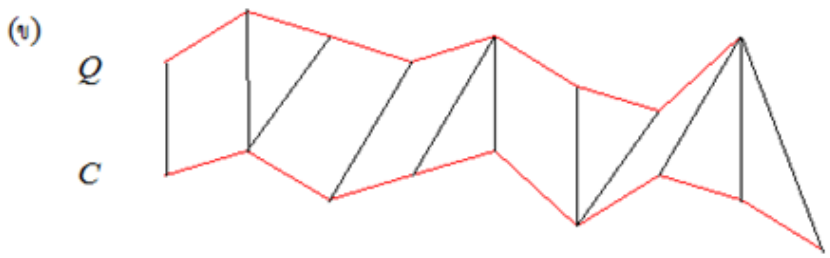
$$\mathrm{DTW}(Q, C) = \sqrt{\gamma(q_m, c_n)} \quad (3)$$

ตัวอย่างการคำนวณของการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงแสดงดังรูปที่ 2.6 โดยกำหนดให้ Q คือข้อมูลอนุกรมเวลาที่มีค่าเท่ากับ [3,1,2,3,2,4,5,2] และ C คือข้อมูลอนุกรมเวลาที่มีค่าเท่ากับ [2,1,3,2,1,4,2,3,5]

(ก)

C \ Q	3	1	2	3	2	4	5	2
2	1	2	2	3	3	7	16	16
1	5	1	2	6	4	12	23	17
3	5	5	2	2	3	4	8	9
2	6	6	2	3	2	6	13	8
1	10	6	3	6	3	11	22	9
4	11	15	7	4	7	3	4	8
2	12	12	7	5	4	7	12	4
3	12	16	8	5	5	5	9	5
5	16	28	17	9	14	6	5	14

(ข)

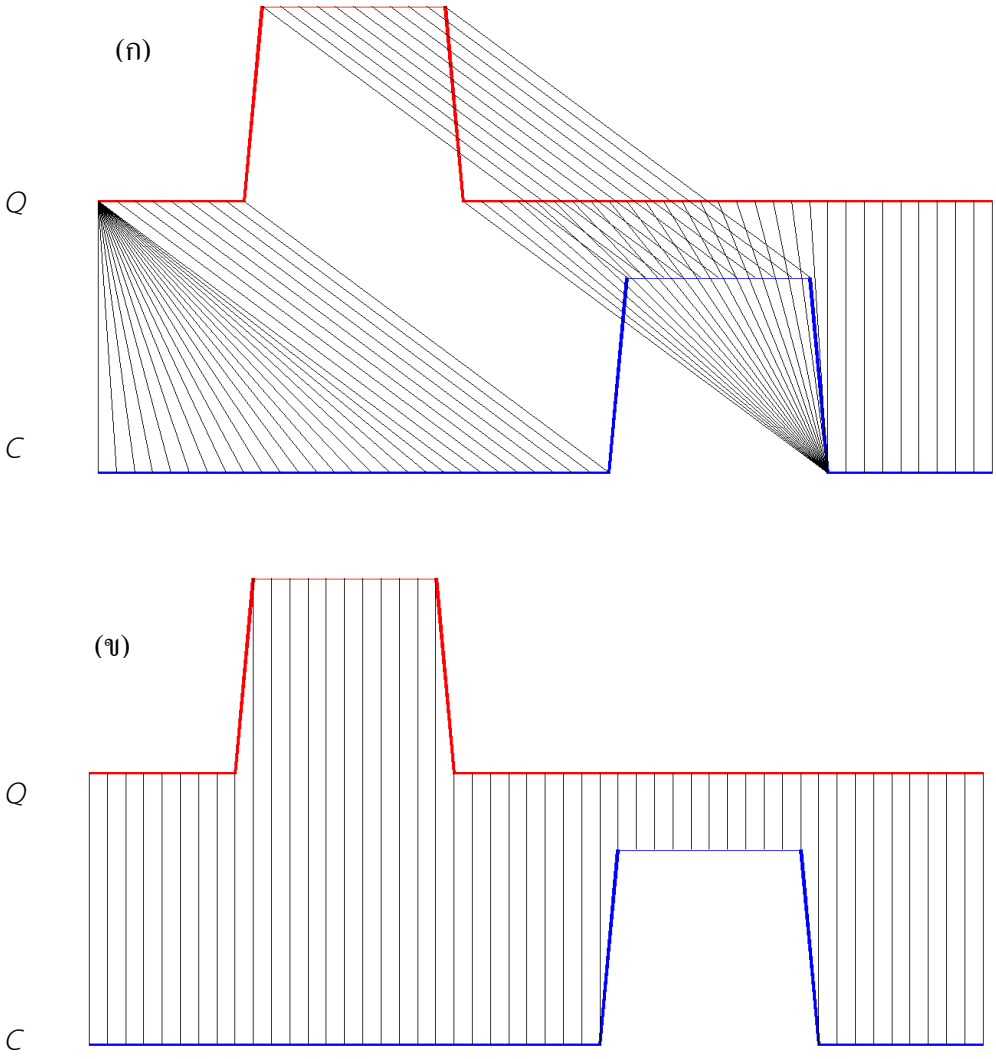


รูปที่ 2.6 (ก) เมทริกซ์สำหรับเก็บค่าระยะทางสะสม โดยช่องที่มีสีเข้มแสดงถึงเส้นทางการปรับแนวของข้อมูลทั้งสอง (ข) รูปภาพของข้อมูลอนุกรมเวลาทั้งสองและเส้นเชื่อมระหว่างข้อมูลที่แสดงถึงเส้นทางการปรับแนว

แม้ว่าการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงจะทำให้ระยะทางที่ได้ลดความผิดพลาดจากความแปรผันของเวลารวมถึงสิ่งรบกวนเล็กน้อยได้ จึงสามารถนำไปใช้กับการจำแนกประเภท การจัดกลุ่ม หรือวิธีการอื่น ๆ ได้อย่างมีประสิทธิภาพ แต่การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงนั้นมีข้อเสียที่ใช้เวลาในการคำนวณสูง โดยเวลาในการคำนวณนั้นอยู่ที่ $O(n^2)$ ทำให้เวลาในการคำนวณยิ่งมากขึ้นเมื่อข้อมูลมีจำนวนมาก นอกจากนี้การปรับแนวของไดนามิกไทม์วอร์ปปิงอาจไม่ถูกต้องเสมอไปในข้อมูลบางประเภทหรือบางกรณีเช่นกัน

2.5 การกำหนดเงื่อนไขบังคับโดยรวม (Global Constraint)

การกำหนดเงื่อนไขบังคับโดยรวม หรือ Global Constraint [16-18] คือวิธีการหนึ่งที่นำมาใช้ควบคู่ไปในวิธีการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง เนื่องจากวิธีการวัดระยะทางแบบไดนามิไทม์วอร์ปปิงนั้นจะทำการหาเส้นทางการปรับแนวระหว่างข้อมูลอนุกรมเวลาสองอนุกรม โดยเส้นทางการปรับแนวนั้นไม่จำเป็นที่จะต้องอยู่ใกล้กันหรืออยู่ติดกัน ส่งผลให้มีความเป็นไปได้ว่าเส้นทางการปรับแนวบางส่วนที่สร้างขึ้นนั้นทำการปรับแนวข้อมูลที่มีความต่างของเวลามากจนเกินไปจนกระทั่งระยะทางระหว่างข้อมูลอนุกรมเวลาสองข้อมูลน้อยและส่งผลไปถึงผลลัพธ์เมื่อนำไปทำการจำแนกประเภท จัดกลุ่มหรือวิธีการอื่น ๆ ที่ใช้การเปรียบเทียบระยะทางระหว่างข้อมูล การกำหนดเงื่อนไขบังคับโดยรวมให้เส้นทางการปรับแนวนั้นไม่ห่างมากเกินไปจากจุดเวลานั้นส่งผลให้ระยะทางที่ได้มีความเหมาะสมมากขึ้น รูปที่ 2.7 แสดงการเปรียบเทียบเส้นทางการปรับแนวระหว่างข้อมูลอนุกรมเวลาที่ไม่ได้ใช้การกำหนดเงื่อนไขบังคับโดยรวมและข้อมูลอนุกรมเวลาที่ใช้การกำหนดเงื่อนไขบังคับโดยรวม



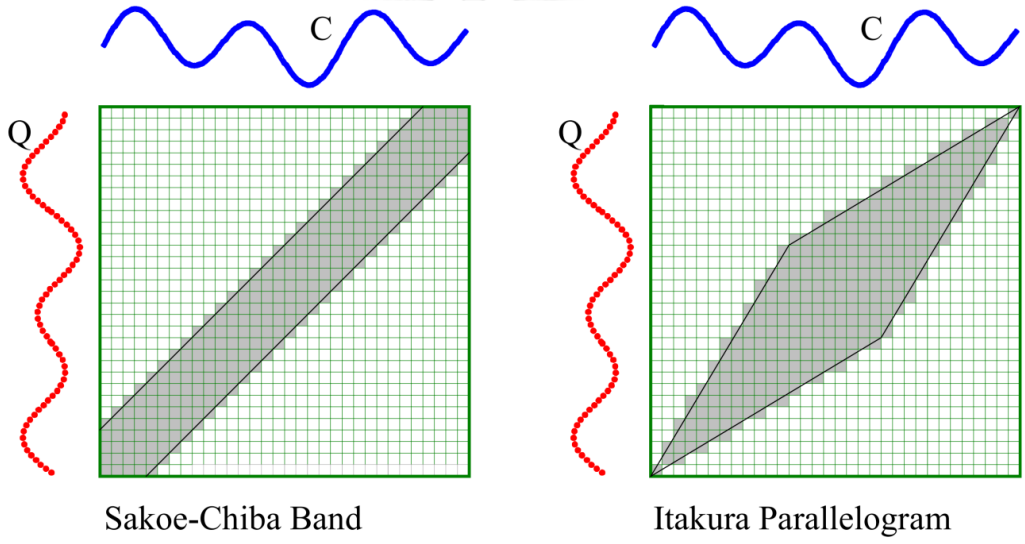
รูปที่ 2.7 การเปรียบเทียบระหว่าง (ก) ข้อมูลอนุกรมเวลาที่วัดระยะทางด้วยการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงโดยไม่ใช้การกำหนดเงื่อนไขบังคับโดยรวมและ (ข) ข้อมูลอนุกรมเวลาที่ด้วยการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงโดยใช้การกำหนดเงื่อนไขบังคับโดยรวม

การกำหนดค่าตัวแปรต่าง ๆ นั้นเป็นดังเช่นวิธีการที่อธิบายในหัวข้อที่ 2.4 การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง (Dynamic Time Warping Distance Measure) วิธีการกำหนดเงื่อนไขบังคับโดยรวมสามารถอธิบายได้ดังนี้

- กำหนดให้ $W = w_1, w_2, \dots, w_k, \dots, w_k$ และ $w_k = (i,j)_k$ โดย $\mathrm{j} - R_i \leq i \leq j + R_i$ เมื่อ R คือขอบเขตของเงื่อนไขบังคับโดยรวม

- กำหนดให้ $R_i=d$ เมื่อ $0 \leq d \leq m$ และ $1 \leq i \leq m$ เมื่อมองในตารางเมทริกซ์สำหรับเก็บค่าระยะทางสะสม (Cumulative Distance) ค่าของ Ri คือระยะทางด้านบนในแกน y และระยะทางด้านขวาในแกน x ของจุดเวลาเดียวกันของข้อมูลอนุกรมเวลาทั้งสองข้อมูล

วิธีการในการกำหนดเงื่อนไขบังคับโดยรวมนั้นมีอยู่หลากหลายวิธีการ โดยวิธีการที่เป็นที่นิยมที่สุดมีอยู่ทั้งหมดสองวิธีการได้แก่ การกำหนดเงื่อนไขโดยรวมแบบซาโก-ชิบะ [17] และการกำหนดเงื่อนไขบังคับโดยรวมแบบอิตาคุระ [18] ซึ่งมีรูปแบบของระยะทางในการกำหนดเงื่อนไขของจุดที่จะทำการปรับแนวได้แตกต่างกัน การกำหนดเงื่อนไขโดยรวมแบบซาโก-ชิบะจะกำหนดค่าของระยะทางเป็นค่าคงที่ การกำหนดเงื่อนไขบังคับโดยรวมแบบอิตาคุระจะกำหนดค่าระยะทางเป็นฟังก์ชันของค่า i รูปที่ 2.8 แสดงถึงขอบเขตที่สร้างขึ้นจากการกำหนดเงื่อนไขบังคับโดยรวม เส้นทางการปรับแนวจะอยู่ในกรอบของเงื่อนไขนี้ตั้งแต่จุดแรกเริ่มจนถึงจุดสุดท้าย



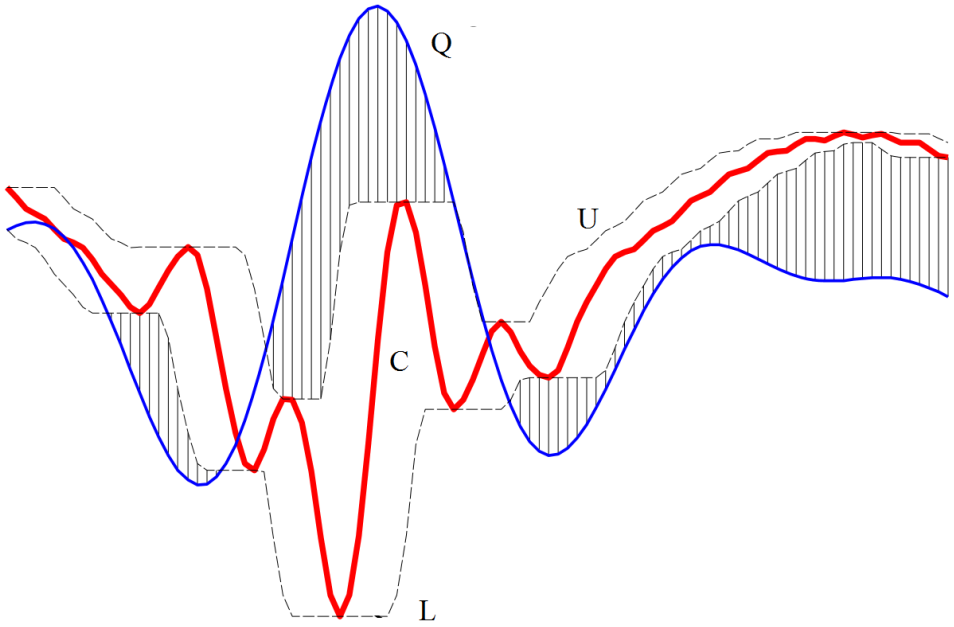
รูปที่ 2.8 แสดงขอบเขตของการกำหนดเงื่อนไขบังคับโดยรวมแบบซาโก-ชิบะและการกำหนดเงื่อนไขโดยรวมแบบอิตาคุระ (ที่มา Ratanamahatana และ Keogh [16])

การกำหนดเงื่อนไขบังคับโดยรวมในการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงนั้น ช่วยให้การนำระยะทางที่ได้ไปใช้งานมีประสิทธิภาพมากขึ้น รวมถึงสามารถลดเวลาในการคำนวณลงได้เล็กน้อยอันเนื่องจากเส้นทางการวอร์ปที่สั้นลง

2.6 ฟังก์ชันขอบเขตล่างของไดนามิกไทม์วอร์ปปิง (Lower Bounding Function of Dynamic Time Warping)

การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง รวมถึงการใช้การกำหนดเงื่อนไขบังคับโดยรวมนั้นมีประสิทธิภาพในการนำไปใช้งานมากขึ้น แต่ทว่าเวลาที่ใช้ในการคำนวณระยะทางระหว่างข้อมูลอนุกรมเวลาก็มากขึ้นเช่นกัน หากข้อมูลเหล่านั้นมีเป็นจำนวนมาก เวลาที่ใช้ในการคำนวณระยะทางให้ได้ระยะทางระหว่างข้อมูลอนุกรมเวลาทั้งหมดจะมากตามไปด้วย โดยปกติแล้วขีดจำกัดเชิงสัญกรณ์ในด้านเวลาของวิธีการไดนามิกไทม์วอร์ปปิงเท่ากับ $O(n^2)$ นั่นคือใช้เวลาในการคำนวณเป็นฟังก์ชันพหุนามกับความยาวของข้อมูลที่ใช้ในการเปรียบเทียบ ดังนั้นจึงได้มีผู้เสนอวิธีการที่ใช้ฟังก์ชันขอบเขตล่างของวิธีไดนามิกไทม์วอร์ปปิงขึ้น เพื่อใช้เป็นค่าประมาณค่าระยะทางระหว่างข้อมูลอนุกรมเวลาทั้งสองข้อมูล เพื่อลดจำนวนข้อมูลที่ต้องนำไปคำนวณด้วยวิธีการไดนามิกไทม์วอร์ปปิงจริง ๆ อีกครั้งหนึ่ง ส่งผลให้เวลาที่ใช้คำนวณรวดเร็วขึ้น

ที่ผ่านมาได้มีผู้นำเสนอฟังก์ชันขอบเขตล่างของไดนามิกไทม์วอร์ปปิงอยู่มากมาย **[19-21]** โดยวิธีการที่ดีที่สุดที่ได้นำมาใช้เพื่อวัดผลในงานวิจัยนี้คือวิธีการของ Eamonn Keogh [21] โดยมีการใช้การกำหนดเงื่อนไขบังคับโดยรวมมาใช้ในการประมาณค่าระยะทางของไดนามิกไทม์วอร์ปปิง โดยการสร้างข้อมูลอนุกรมเวลาเพื่อกำหนดขอบเขตช่วงบนของแต่ละจุดบนข้อมูลอนุกรมเวลา U และขอบเขตช่วงล่างของแต่ละจุดบนข้อมูลอนุกรมเวลา L จากค่าในการคำนวณกำหนดการพลวัตภายใต้เงื่อนไขบังคับโดยรวมสำหรับข้อมูลสอบถาม จากนั้นจะทำการหาค่าระยะทางโดยใช้ตัววัดความคล้ายแบบยุคลิด กับจุดข้อมูลที่มากกว่าอนุกรมเวลา U ที่เป็นขอบช่วงบน หรือน้อยกว่าอนุกรมเวลา L ซึ่งเป็นขอบเขตล่าง ซึ่งสามารถอธิบายได้ดังรูปที่ 2.9



รูปที่ 2.9 แสดงการคำนวณค่าฟังก์ชันขอบเขตล่างของไดนามิกไทม์วอร์ปปิง โดนเส้นที่โยงระหว่างข้อมูลอนุกรมเวลาคือส่วนที่ทำการคำนวณโดยใช้ตัววัดความคล้ายแบบยุคลิด (ที่มา Keogh และ Ratanamahatana [21])

การกำหนดค่าตัวแปรต่าง ๆ นั้นเป็นดังเช่นวิธีการที่อธิบายในหัวข้อที่ 2.4 การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง (Dynamic Time Warping Distance Measure) กำหนดให้ $U = u_1, u_2, \dots, u_m$ เป็นอนุกรมเวลาที่เป็นขอบเขตบนของอนุกรมเวลา Q และกำหนดให้ $L = l_1, l_2, \dots, l_m$ เป็นอนุกรมเวลาที่เป็นขอบเขตล่างของอนุกรมเวลา Q สามารถหาค่าของ U และ L ได้ดังสมการที่ 4 และ 5

$$U = \max(q_{\min(1,i-r)}, \dots q_{\max(i+r,m)}) \quad (4)$$

$$L = \min(q_{\min(1,i-r)}, \dots q_{\max(i+r,m)}) \quad (5)$$

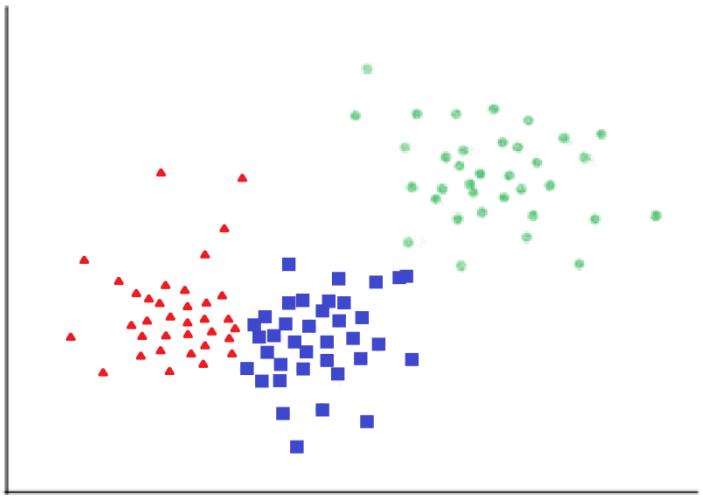
กำหนดให้ $LB\,(Q,C)$ เป็นค่าของฟังก์ชันขอบเขตล่างของอนุกรมเวลา Q และ C โดยสามารถคำนวณค่าได้ดังสมการที่ 6

$$LB(Q,C) = \sqrt{\sum_{i=1}^{m} \begin{cases} (U_i - c_i)^2 \; if \; c_i > U_i \\ (L_i - c_i)^2 \; if \; c_i < L_i \\ 0 \qquad otherwise \end{cases}} \qquad (6)$$

2.7 การจัดกลุ่มข้อมูล (Clustering)

การจัดกลุ่มข้อมูล [8, 28] คือการพิจารณาข้อมูลจำนวนหนึ่ง และทำการแบ่งข้อมูลนั้นออกเป็นกลุ่ม ๆ ตามความคล้ายคลึงของข้อมูล วิธีการจัดกลุ่มมีมากมายหลากหลายวิธี ซึ่งสามารถจัดเป็นสามรูปแบบหลักได้แก่

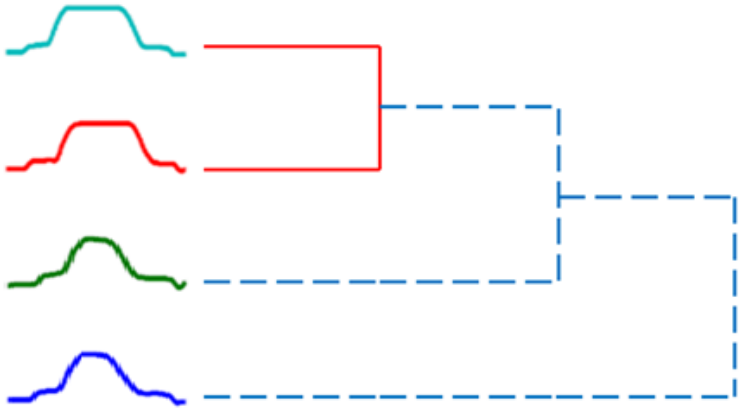
- การจัดกลุ่มด้วยการแบ่งส่วน (Partition Method) เป็นการจัดกลุ่มแบบแบ่งส่วนของข้อมูลออกเป็นส่วน ๆ แยกออกจากกันแล้วจึงจับข้อมูลรวมเป็นกลุ่ม ๆ จำนวน K กลุ่ม ซึ่งต้องมีการกำหนดจำนวนกลุ่มของข้อมูลก่อนทำการจัดกลุ่ม ตัวอย่างวิธีการในรูปแบบนี้เช่น การจัดกลุ่มแบบเคมีนส์ (K-means Clustering) [33] การจัดกลุ่มแบบเคเมดอยด์ (K-medoid Clustering) [34] CLARA (Clustering LARge Application) [35] เป็นต้น วิธีการจัดกลุ่มแบบแบ่งส่วนแสดงดังรูปที่ 2.10



รูปที่ 2.10 วิธีการจัดกลุ่มข้อมูลด้วยการแบ่งส่วน รูปร่างของจุดแสดงถึงกลุ่มที่แตกต่างกัน

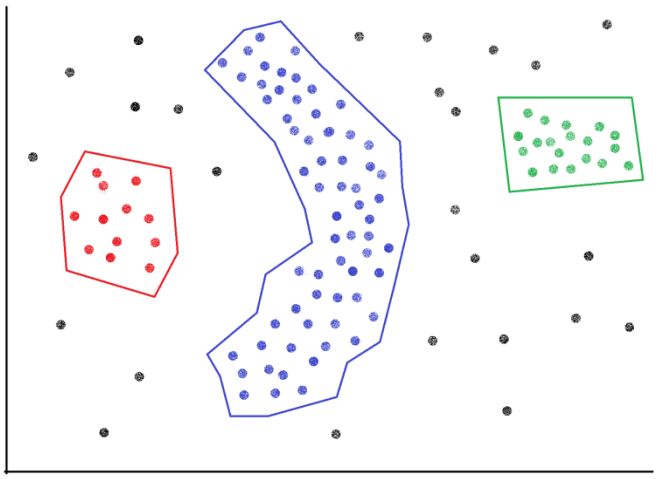
- การจัดกลุ่มตามลำดับขั้น (Hierarchical Method) คือการจัดกลุ่มโดยการจัดลำดับขั้นของข้อมูลคล้ายกับต้นไม้ แล้วจึงทำการจัดกลุ่ม โดยจะมีการแบ่งการจัดลำดับขึ้นเป็นสองรูปแบบคือ แบบล่างขึ้นบน (Agglomerative) และแบบบนลงล่าง (Divisive) ตัวอย่างวิธีการในรูปแบบนี้เช่น AGNES (Agglomerative NESting) [35] DIANA

(Divisive ANAlysis) [35] BIRCH (Balanced Iterative Reducing and clustering using hierarchies) [36] เป็นต้น วิธีการจัดกลุ่มตามลำดับขั้นแสดงดังรูปที่ 2.11



รูปที่ 2.11 วิธีการจัดกลุ่มข้อมูลอนุกรมเวลาแบบลำดับขั้น รูปแบบของเส้นเชื่อมข้อมูลแสดงถึงกลุ่มข้อมูลที่แตกต่างกัน

- การจัดกลุ่มตามความหนาแน่นของข้อมูล (Density-based method) เป็นการจัดกลุ่มข้อมูลตามความหนาแน่นของข้อมูล ซึ่งแตกต่างจากการจัดกลุ่มด้วยการแบ่งส่วนที่ไม่ต้องการค่าของจำนวนกลุ่มที่ต้องการจะแบ่ง แต่จะต้องการข้อกำหนดหรือตัวแปรบางประการเพื่อทำการแบ่งข้อมูลออกจากกัน ตัวอย่างวิธีการในรูปแบบนี้เช่น Density-based spatial clustering of applications with noise (DBSCAN) [37] หรือ Affinity propagation (AP) [38] วิธีการจัดกลุ่มตามความหนาแน่นของข้อมูลแสดงดังรูปที่ 2.12



รูปที่ 2.12 วิธีการจัดกลุ่มข้อมูลตามความหนาแน่นของข้อมูล

ในข้อมูลอนุกรมเวลานั้นก็มีการจัดกลุ่มข้อมูลโดยการใช้ระยะทางระหว่างข้อมูลเช่นกัน ซึ่งมักใช้ร่วมกับการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงซึ่งมีเวลาในการคำนวณที่สูง เมื่อผนวกเข้ากับเวลาคำนวณของวิธีการจัดกลุ่มที่ใช้เวลานานเช่นกัน ส่งผลให้การแบ่งกลุ่มข้อมูลอนุกรมเวลานั้นใช้เวลาสูง

2.8 ค่าเฉลี่ยแบบไดนามิกไทม์วอร์ปปิงแบรีเซนเตอร์ (Dynamic Time Warping Barycenter Averaging หรือ DBA)

ค่าเฉลี่ยแบบไดนามิกไทม์วอร์ปปิงแบรีเซนเตอร์ หรือ DBA [25, 26] เป็นวิธีการหนึ่งในการหาค่าเฉลี่ยของข้อมูลอนุกรมเวลา จุดเด่นของวิธี DBA นั้นคือการที่สามารถหาค่าเฉลี่ยจากข้อมูลอนุกรมเวลามากกว่าสองข้อมูลได้พร้อมกัน ซึ่งแตกต่างจากวิธีการอื่น ๆ ที่ทำการหาค่าเฉลี่ยได้ทีละคู่ โดยใช้การหาค่าเฉลี่ยของข้อมูลอนุกรมเวลาตามแนวแกนของเวลาแทนที่จะทำการหาค่าเฉลี่ยตามจุดข้อมูล ซึ่งส่งผลให้ข้อมูลที่สร้างขึ้นจากการหาค่าเฉลี่ยนั้น มีจุดเวลาที่ไม่เพิ่มขึ้นและกำจัดผลกระทบจากลำดับของการจับคู่ข้อมูลอนุกรมเวลาเมื่อต้องการหาค่าเฉลี่ยมากกว่าสองข้อมูล วิธีการ ของ DBA สามารถอธิบายได้ดังนี้

- ค้นหาข้อมูลที่เป็นตัวหลักในการหาค่าเฉลี่ยเรียกว่า Pivot ซึ่งในงานวิจัยที่ผ่านมาได้มีการใช้วิธีการหลัก ๆ อยู่สองวิธีการได้แก่ การสุ่มและการใช้ข้อมูลอนุกรมเวลาที่เป็นข้อมูลเมดอยด์ของข้อมูลทั้งหมด โดยข้อมูลเมดอยด์คือข้อมูลที่มีความคล้ายมากที่สุดกับข้อมูลตัวอื่น ๆ ทั้งหมด

- คำนวณการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงระหว่างข้อมูลตัวหลักกับข้อมูลตัวอื่นที่เหลือทุกตัว แต่สิ่งที่ต้องการในขั้นตอนนี้ไม่ใช่ค่าของระยะทาง แต่เป็นเส้นทางการปรับแนวหรือ วิถีการวอร์ปของข้อมูลตัวอื่น ๆ ทุกตัวเทียบกับข้อมูลตัวหลัก
- หลังจากนั้นจะทำการเริ่มคำนวณจากจุดเวลาที่ 1 ไปจนถึงเวลาสุดท้ายโดยจะทำการเฉลี่ยค่าของข้อมูลทุกจุดเวลาของข้อมูลตัวอื่นทุกตัวที่มีการปรับแนวเข้ากับจุดเวลานั้นของข้อมูลตัวหลัก จนกระทั่งได้ข้อมูลค่าเฉลี่ยครบทุกจุดเวลา แล้วทำการกำหนดข้อมูลค่าเฉลี่ยที่ได้เป็นข้อมูลตัวหลัก ทำการคำนวณระยะทาง และหาค่าเฉลี่ยอีกครั้งหนึ่งหรือมากกว่า ตามแต่ผู้ใช้จะกำหนด

ค่าเฉลี่ยแบบไดนามิกไทม์วอร์ปปิงแบรีเซนเตอร์นั้นสามารถใช้สร้างแผ่นแบบเฉลี่ยที่ให้ความแม่นยำมากขึ้นเมื่อนำไปทำการจำแนกประเภทด้วยวิธีการจำแนกข้อมูลแบบเพื่อนบ้านใกล้สุดอันดับที่หนึ่ง โดยใช้การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง รวมถึงความเร็วในการหาแผ่นแบบเฉลี่ยที่รวดเร็วขึ้น เพราะมีการคำนวณหาระยะทางระหว่างข้อมูลตัวอื่น ๆ กับข้อมูลตัวหลักเท่านั้น อย่างไรก็ตามการเลือกข้อมูลตัวหลักก็ส่งผลต่อรูปร่างของแผ่นแบบเฉลี่ยเช่นกัน วิธีการหาค่าเฉลี่ยแบบไดนามิกไทม์วอร์ปปิงแบรีเซนเตอร์สามารถอธิบายได้ดังรหัสเทียมหรือ Pseudocode ดังตารางที่ 2.1

ตารางที่ 2.1 รหัสเทียมหรือ Pseudocode ของวิธีการ DBA (ที่มา Petitjean, Ketterlin, และ Gançarski [25])

Algorithm 1. DBA (*D*, *i*)
require *D* : The set of sequences with length *L* require *i* : number of iterations
T = medoid (*D*) Do *i* times *T* = DBA_update (*D*,*T*) return *T*
Algorithm 2. DBA_update(*D*,*P*)
require *P* : the average sequence to refine (with length *L*) require *D* : the set of sequences with length *L*

alignment = [Ø,Ø,...,Ø] //array of empty set with length *L*
for each *S* in *D*

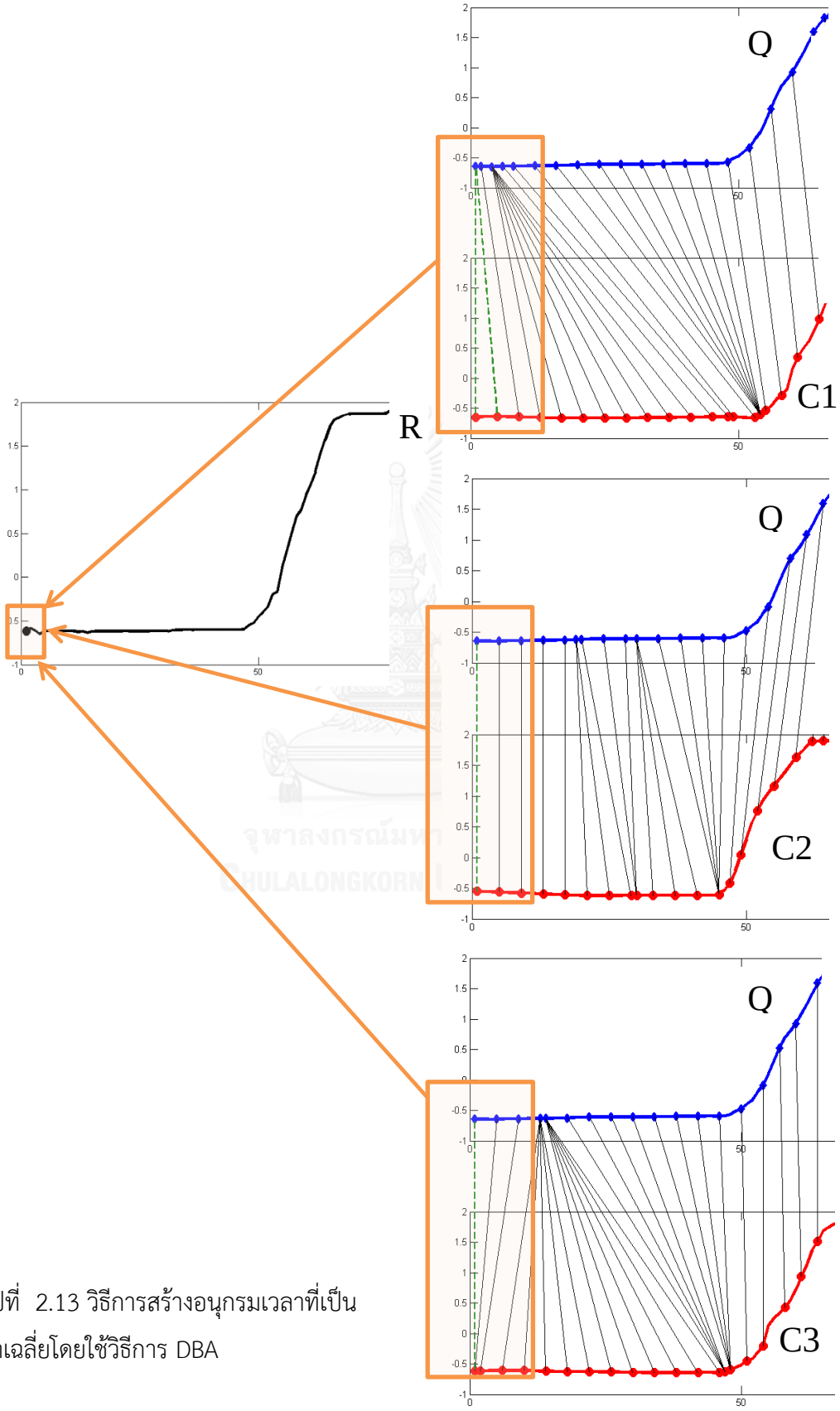
```
        alignment_for_S = DTW_multiple_alignment (P,S)
        for i = 1 to L
                alignment[i] = alignment [i] ∪ alignment_for_S[i]
        end for
end for
T = sequences with length L
for i = 1 to L
        T(i) = mean(alignment[i])
end for
return T
```

Algorithm 3. DTW_multiple_alignment (*R,S*)

```
require R : the main sequence for which the alignment is computed
require S : the sequence to align to R using DTW
cost = DTW(R,S) //compute the cumulative matrix of DTW between R and S
L = length(R)
alignment = [Ø,Ø,...,Ø] //array of empty set with length L
[i , j] = size (cost) //iterates over the element of R and S
while (i>1) && (j>1)
        alignment[i] = alignment[i] ∪ S(j)
        if (i == 1) j = j-1
        else if (j == 1) i = i-1
        else
                score = min (cost[i-1][j-1] , cost[i,j-1] , cost[i-1,j])
                if (score == cost[i-1][j-1] )
                        i = i -1
                        j = j - 1
                else if ( score == cost[i-1][j]) then i = i - 1
                else j = j -1
                end if
        end if
```

```
end while
return alignment
```

จากรูปที่ 2.13 กำหนดให้ข้อมูล Pivot คือข้อมูล Q และข้อมูลอื่นที่นำมาทำการหาแผ่นแบบเฉลี่ยคือ C1,C2 และ C3 ข้อมูล R คือข้อมูลอนุกรมเวลาที่เป็นแผ่นแบบเฉลี่ย เส้นระหว่างกราฟแสดงถึงเส้นทางการปรับแนวที่ได้จากวิธีการหาระยะทางแบบไดนามิกไทม์วอร์ปปิง ได้แสดงถึงการเฉลี่ยข้อมูลอนุกรมเวลาที่จุดเวลาที่ 1 โดยค่าในจุดที่ 1 ของอนุกรมเวลา R จะเกิดจากการเฉลี่ยของค่าของจุด 1 ของอนุกรมเวลา C2 และ C3 และจุดเวลาที่ 1 กับ 2 ของอนุกรมเวลา C1 จากนั้นทำวิธีการลักษณะนี้กับจุดเวลาถัดไปจนกระทั่งครบทุกจุดเวลาของข้อมูล Pivot Q ก็จะได้ข้อมูลอนุกรมเวลาใหม่ที่เป็นแผ่นแบบขึ้น โดยในรูปได้ทำการตัดเฉพาะส่วนเวลาช่วงต้นของข้อมูลอนุกรมเวลาเพื่อแสดงให้ดูเท่านั้น



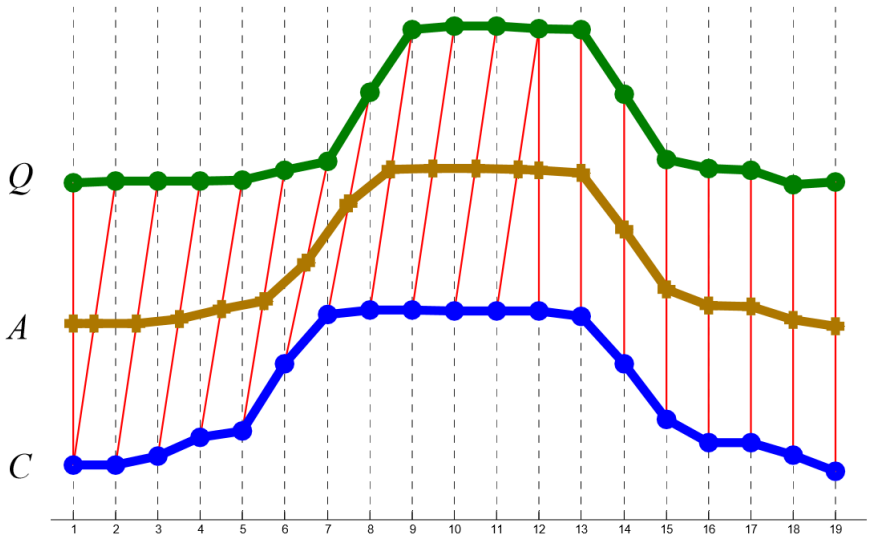
รูปที่ 2.13 วิธีการสร้างอนุกรมเวลาที่เป็นค่าเฉลี่ยโดยใช้วิธีการ DBA

2.9 งานวิจัยที่เกี่ยวข้อง

งานวิจัยค้นคว้าเกี่ยวกับการสร้างแผ่นแบบเฉลี่ยของข้อมูลอนุกรมเวลาเริ่มต้นมานาน ในช่วงแรกมักจะเป็นงานวิจัยในด้านการประมวลผลสัญญาณ (Signal Processing) โดยนำแผ่นแบบเฉลี่ยที่สร้างขึ้นไปใช้ทั้งในการจัดกลุ่มข้อมูลและการจำแนกประเภทข้อมูล [39]

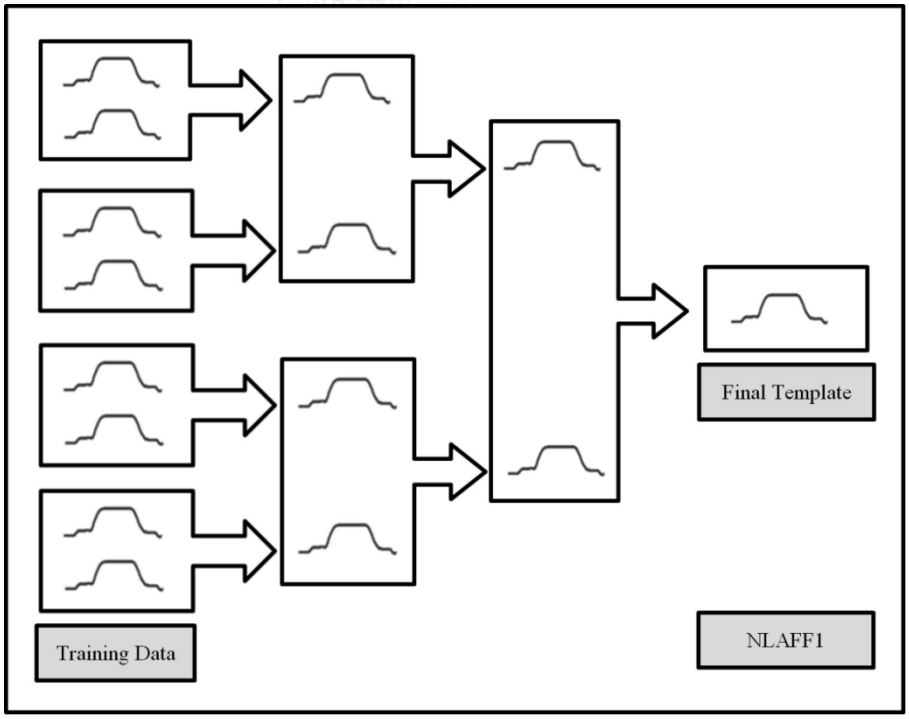
ในปี 1996 Gupta และคณะได้นำเสนอวิธีการสร้างแผ่นแบบเฉลี่ยของข้อมูลอนุกรมเวลาขึ้น โดยทำการสร้างแผ่นแบบเฉลี่ยจากข้อมูลอนุกรมเวลาสองข้อมูลด้วยวิธีการปรับแนวแบบไม่เชิงเส้น (Non-linear alignment) ชื่อว่า Non-linear Alignment and Averaging Filters (NLAFF) [22] ซึ่งการสร้างแผ่นแบบเฉลี่ยด้วยวิธี NLAFF ต่างจากการหาค่าเฉลี่ยทั่วไปที่ทำการเฉลี่ยค่าข้อมูลอนุกรมเวลาสองข้อมูลโดยเฉลี่ยข้อมูลตามจุดเวลาเดียวกันเข้าด้วยกัน ซึ่งส่งผลให้แผ่นแบบเฉลี่ยที่หาด้วยวิธีทั่วไปไม่สามารถรองรับสิ่งรบกวนหรือการแปรผันของเวลาได้ แผ่นแบบเฉลี่ยที่สร้างขึ้นจึงได้คุณภาพที่ไม่ดีเมื่อนำไปทำการจัดกลุ่มข้อมูลหรือจำแนกประเภทข้อมูล

วิธีการ NLAFF จึงได้ใช้การปรับแนวแบบไม่เชิงเส้น โดยใช้เส้นทางการปรับแนว (Warping Path) จากวิธีการหาระยะทางแบบไดนามิกไทม์วอร์ปปิง ในการหาค่าเฉลี่ยของข้อมูลอนุกรมเวลาสองข้อมูล ซึ่งในแต่ละข้อมูลย่อยของข้อมูลอนุกรมเวลาที่เป็นแผ่นแบบเฉลี่ยใหม่นั้น เกิดจากการเฉลี่ยของคู่ข้อมูลในแต่ละเส้นทางการปรับแนว ส่งผลให้แผ่นแบบเฉลี่ยนั้นอาจจะมีความยาวของข้อมูลอนุกรมเวลาที่มากขึ้นกว่าข้อมูลตั้งต้นทั้งสองข้อมูล วิธีการปรับแนวแบบไม่เชิงเส้นสามารถแสดงได้ดัง รูปที่ 2.14

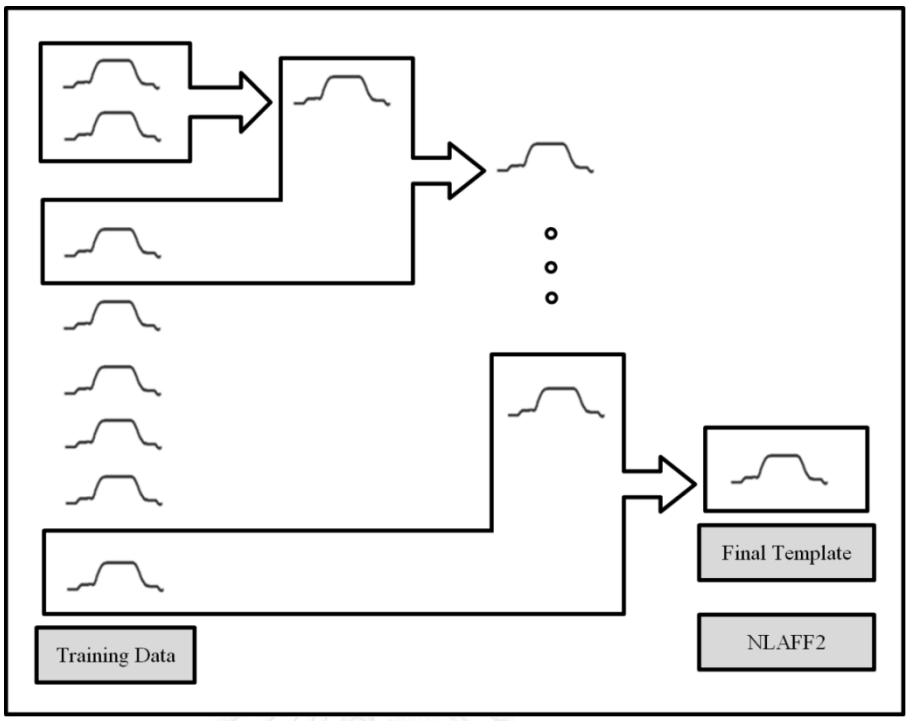


รูปที่ 2.14 วิธีการปรับแนวแบบไม่เชิงเส้น (ที่มา Niennattrakul, Srisai และ Ratanamahatana [24])

วิธีการ NLAFF สามารถทำการหาแผ่นแบบเฉลี่ยได้จากข้อมูลอนุกรมเวลาทีละคู่เท่านั้น จึงได้มีการใช้วิธีการเลือกข้อมูลอนุกรมเวลามาทีละคู่เพื่อทำการหาแผ่นแบบเฉลี่ย 2 วิธีการ โดยแต่ละวิธีการนั้นจะเหมือนกันคือใช้ข้อมูลตามลำดับที่ข้อมูลเรียนรู้มีอยู่ทำการเลือกข้อมูลทีละคู่เพื่อทำการหาค่าเฉลี่ย วิธีการแรกนั้นจะทำการหาค่าเฉลี่ยข้อมูลโดยการจับคู่ข้อมูลตามลำดับเป็นคู่ ๆ ตัวอย่างเช่น ข้อมูลลำดับที่ 1 จับคู่กับข้อมูลลำดับที่ 2 และข้อมูลลำดับที่ 3 จับกับข้อมูลลำดับที่ 4 ไปเรื่อย ๆ จนครบทุกข้อมูล จากนั้นทำการหาแผ่นแบบเฉลี่ยของแต่ละคู่ แล้วใช้แผ่นแบบเฉลี่ยที่ได้มาทำการจับคู่ในแบบเดียวกันอีก กล่าวคือแผ่นแบบที่ 1 จะคู่กับแผ่นแบบที่ 2 แผ่นแบบที่ 3 จะจับคู่กับแผ่นแบบที่ 4 ไปเรื่อย ๆ และทำการหาแผ่นแบบเฉลี่ยซ้ำไปเรื่อย ๆ จนกระทั่งเหลือเพียงแผ่นแบบเดียว ซึ่งวิธีการนี้จะมีข้อกำกัดคือ หากจะทำให้สุดท้ายเหลือแผ่นแบบเดียวได้นั้น ข้อมูลเรียนรู้ทั้งหมดต้องจำนวนเท่ากับค่ากำลังของ 2 ใดๆ เช่น 2 4 8 16 32 64 128 เท่านั้นทำให้การใช้งานมีข้อจำกัดมาก เรียกวิธีนี้ว่า NLAFF1 จึงได้มีอีกหนึ่งวิธีการจับคู่ข้อมูลคือ ทำการหาแผ่นแบบเฉลี่ยของข้อมูลลำดับที่ 1 กับข้อมูลลำดับที่ 2 ให้ได้แผ่นแบบเฉลี่ยมาก่อนหนึ่งแผ่นแบบ จากนั้นทำการหาแผ่นแบบเฉลี่ยใหม่จากแผ่นแบบเฉลี่ยที่ได้มากับข้อมูลลำดับถัดไปคือข้อมูลลำดับที่ 3 ทำวนซ้ำไปเรื่อย ๆ จนกระทั่งครบทุกข้อมูล แล้วได้แผ่นแบบเฉลี่ยสุดท้ายเพียงแผ่นแบบเดียว เรียกวิธีการนี้ว่า NLAFF2 วิธีการจับคู่ข้อมูลทั้งสองแบบนั้นสามารถแสดงได้ดังรูปที่ 2.15 และรูปที่ 2.16



รูปที่ 2.15 วิธีการจับคู่หาแผ่นแบบเฉลี่ยแบบ NLAFF1



รูปที่ 2.16 วิธีการจับคู่หาแผ่นแบบเฉลี่ยแบบ NLAFF2

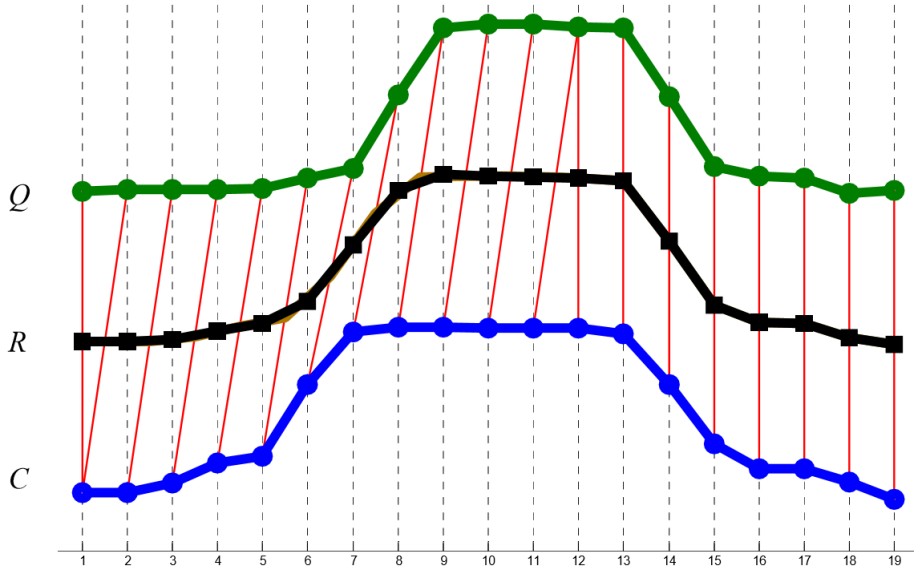
อย่างไรก็ตามการหาแผ่นแบบเฉลี่ยด้วยวิธี NLAFF นั้นยังมีข้อเสียสำคัญคือแผ่นแบบเฉลี่ยที่ได้จะมีความยาวของข้อมูลมากขึ้นเรื่อย ๆ ทำให้จุดที่ทำการหาค่าเฉลี่ยมีมากกว่าจุดข้อมูลเดิมทุกครั้ง ยกเว้นกรณีที่ทุกจุดหาค่าเฉลี่ยกันโดยตรง แม้ว่าการใช้การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง จะทำให้สามารถวัดระยะทางระหว่างข้อมูลที่มีความยาวของแกนเวลาไม่เท่ากันได้ก็ตาม แต่เวลาที่ต้องใช้ในการคำนวณก็จะเพิ่มขึ้นมากเช่นกัน นอกจากนี้การจับคู่ข้อมูลเพื่อทำการหาแผ่นแบบเฉลี่ยของ NLAFF ก็มีข้อเสียสำคัญเช่นกันคือ การจับคู่ข้อมูลแบบแรกการจับคู่ทุกคู่พร้อมกันมีข้อเสียสำคัญคือข้อจำกัดของจำนวนข้อมูล และอีกวิธีหนึ่งจะมีข้อจำกัดในลำดับของข้อมูลซึ่งมีผลมากต่อค่าของแผ่นแบบเฉลี่ยสุดท้าย โดยแผ่นแบบเฉลี่ยสุดท้ายที่สร้างขึ้นจะมีค่าหรือรูปร่างของกราฟข้อมูลเหมือนกับข้อมูลลำดับสุดท้ายมากกว่าลำดับแรก และค่าหรือรูปร่างของข้อมูลลำดับแรกจะส่งผลน้อยมากในแผ่นแบบเฉลี่ยสุดท้ายที่ได้ ทำให้แผ่นแบบที่สร้างขึ้นยังได้ผลลัพธ์ที่ไม่ดีนักในการจำแนกประเภทข้อมูล

ต่อมาในปี 2009 ได้มีผู้พัฒนาวิธีการหาแผ่นแบบเฉลี่ยของข้อมูลอนุกรมเวลาใหม่ขึ้น ชื่อว่า Prioritized Shape Averaging (PSA) [23] ซึ่งเป็นการปรับปรุงจากวิธีการ NLAFF โดยทำการปรับปรุงส่วนของการจับคู่ข้อมูลเพื่อทำการเฉลี่ยค่าระหว่างข้อมูลอนุกรมเวลา 2 ข้อมูล วิธีการนี้จะไม่ใช้การจับคู่ข้อมูลเรียงตามลำดับตามแต่เดิม แต่ใช้ระยะทางระหว่างข้อมูลแต่ละคู่เป็นตัวตัดสินว่าข้อมูลคู่ใดสมควรที่จะทำการหาแผ่นแบบเฉลี่ยก่อน โดยเลือกจากข้อมูลที่มีระยะทางระหว่างกันน้อยที่สุดนำมาทำการหาแผ่นแบบเฉลี่ยก่อน และนำแผ่นแบบเฉลี่ยที่ได้ไปหาระยะทางเทียบกับข้อมูลที่

เหลืออยู่ทั้งหมด แล้วทำการเลือกคู่ที่มีระยะทางระหว่างกันน้อยที่สุดใหม่อีกครั้ง ทำวนซ้ำไปเรื่อย ๆ จนกระทั่งเหลือเพียงแผ่นแบบเดียวต่อหนึ่งคลาส

วิธีการหาระยะทางที่ใช้ใน PSA นั้นใช้การหาระยะทางแบบไดนามิกไทม์วอร์ปปิง เพื่อให้ได้แผ่นแบบเฉลี่ยที่มีคุณภาพมากขึ้น และลำดับของข้อมูลส่งผลน้อยลง อย่างไรก็ตามการหาระยะทางระหว่างข้อมูลทุกตัวด้วยการหาระยะทางแบบไดนามิกไทม์วอร์ปปิงนั้นจะใช้เวลาสูงเมื่อจำนวนข้อมูลเรียนรู้มีจำนวนมาก ส่งผลใช้เวลาในการสร้างแผ่นแบบสูง แม้จะลดเวลาในการจำแนกประเภทได้มาก แต่เวลาในการสร้างแผ่นแบบที่สูงส่งผลให้เวลารวมที่ใช้จะลดลงไม่มาก อีกทั้งแม้ลำดับของข้อมูลจะส่งผลน้อยลงจะมีการคัดเลือกคู่ข้อมูลในการสร้างแผ่นแบบเฉลี่ยแล้วก็ตาม แต่ข้อมูลลำดับหลัง ๆ ก็ยังส่งผลกับแผ่นแบบเฉลี่ยสุดท้ายอยู่เช่นเดิม

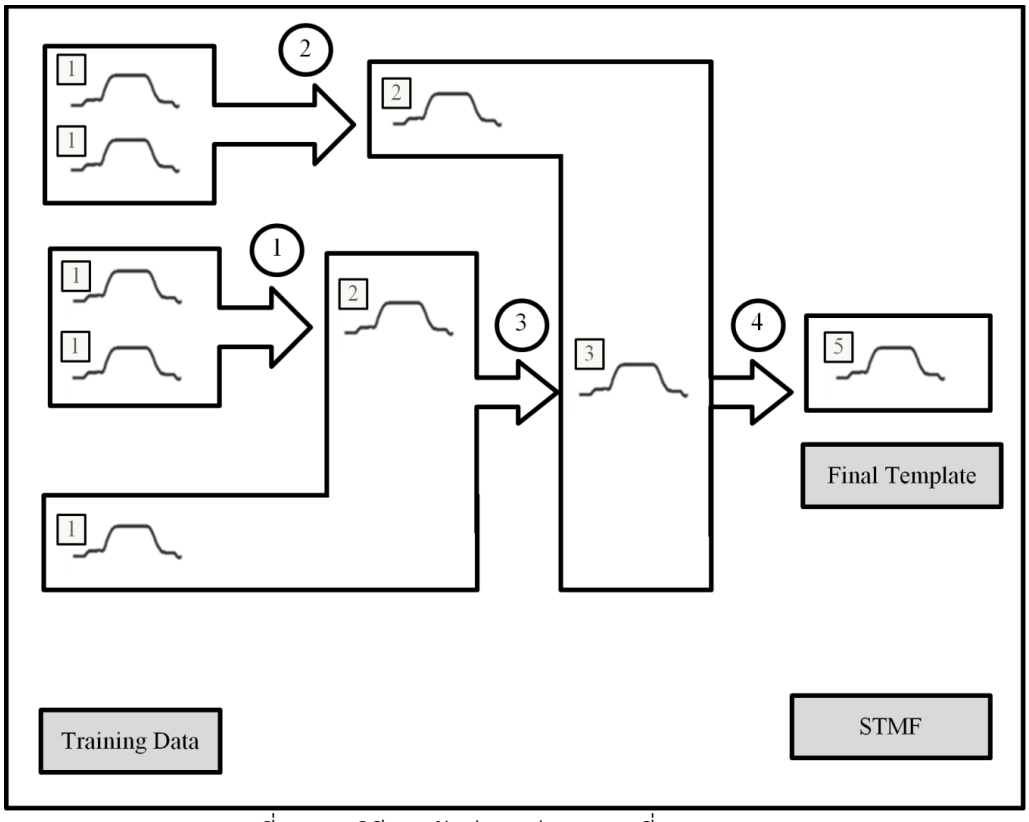
ในปี 2012 ผู้พัฒนา PSA จึงได้ปรับปรุงวิธีการ PSA ขึ้นใหม่โดยตั้งชื่อว่า Shape-based Template Matching Framework (STMF) [24] วิธีการนี้ได้ทำการปรับปรุงวิธีการเฉลี่ยข้อมูลที่ยังคงมีจำนวนจุดข้อมูลมากกว่าคู่ข้อมูลเดิมที่นำมาทำการหาค่าแผ่นแบบเฉลี่ย โดยใช้วิธีการประมาณค่าด้วยวิธีกระดูกงูกำลังสาม (Cubic Spline Approximation) [40] ในการลดจำนวนจุดที่เพิ่มขึ้นในการสร้างแผ่นแบบแต่ละครั้ง วิธีการลดจุดข้อมูลด้วยการใช้การประมาณค่าด้วยวิธีกระดูกงูกำลังสามนั้นสามารถแสดงได้ดังรูปที่ 2.17



รูปที่ 2.17 วิธีการลดจำนวนจุดที่เพิ่มขึ้นในการสร้างแผ่นแบบโดยใช้วิธีกระดูกงูกำลังสาม (ที่มา Niennattrakul, Srisai และ Ratanamahatana [24])

นอกจากปรับปรุงในส่วนของการเฉลี่ยข้อมูลให้มีจำนวนจุดที่ลดลงแล้ว วิธีการ STMF ยังได้ปรับปรุงขั้นตอนการคัดเลือกคู่ข้อมูลอนุกรมเวลาที่นำมาสร้างแผ่นแบบด้วย โดยทำการเพิ่มน้ำหนัก (Weight) เข้าไปในข้อมูลอนุกรมเวลาทุก ๆ ข้อมูล เริ่มต้นจากข้อมูลเรียนรู้ที่รับมาให้ทุกข้อมูลมี

น้ำหนักเท่ากับ 1 เมื่อทำการสร้างแผ่นแบบเฉลี่ยใหม่เพื่อนำมาใช้แทนแผ่นแบบคู่ใด ๆ จะนำน้ำหนักของคู่แผ่นแบบนั้นเป็นเกณฑ์ว่า แผ่นแบบเฉลี่ยที่สร้างขึ้นสมควรจะมีรูปร่างใกล้เคียงกับข้อมูลอนุกรมเวลาข้อมูลใดมากกว่ากัน จากนั้นแผ่นแบบใหม่จะมีค่าน้ำหนักเท่ากับค่าน้ำหนักของข้อมูลอนุกรมเวลาทั้งสองที่นำมาหาค่าเฉลี่ยรวมกัน วิธีการ STMF สามารถแสดงได้ดังรูปที่ 2.18 โดยตัวเลขในวงกลมคือลำดับการทำงานของการจับคู่ข้อมูลเพื่อสร้างแผ่นแบบเฉลี่ย และตัวเลขในกล่องสี่เหลี่ยมแสดงถึงค่าของน้ำหนักของข้อมูลอนุกรมเวลานั้น ๆ



รูปที่ 2.18 วิธีการจับคู่หาแผ่นแบบเฉลี่ยแบบ STMF

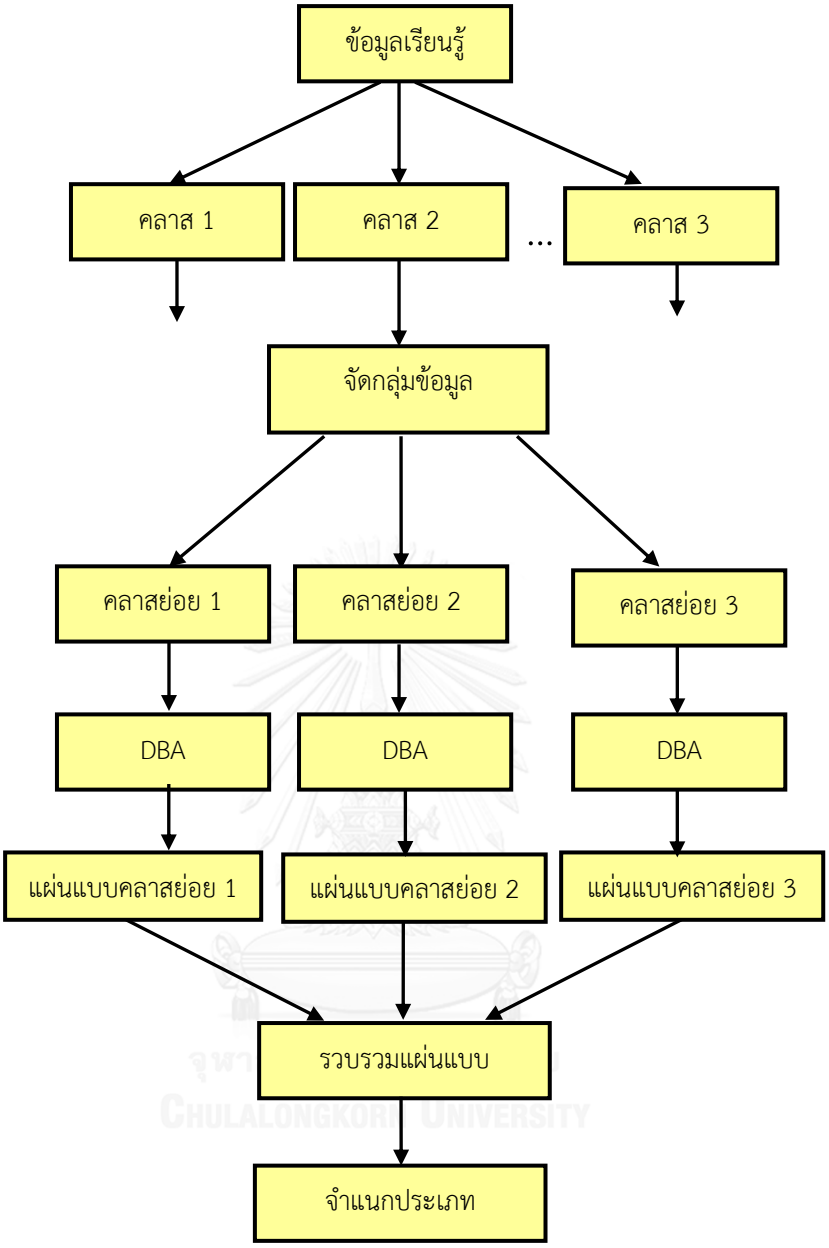
ถึงแม้ว่าวิธีการ STMF จะสามารถจัดการข้อจำกัดของวิธีการ NLAFF ไปได้มาก หากแต่เวลาที่ใช้ในการสร้างแผ่นแบบก็สูงขึ้นมากเช่นเดียวกัน เนื่องด้วยต้องทำการคำนวณระยะทางแบบไดนามิกไทม์วอร์ปปิงระหว่างข้อมูลเรียนรู้ทุกข้อมูลเทียบกัน อีกทั้งเมื่อได้แผ่นแบบใหม่ก็ต้องทำการคำนวณระยะทางแบบไดนามิกไทม์วอร์ปปิงกับข้อมูลที่เหลือใหม่ทั้งหมดอีกครั้ง วิธีนี้จึงไม่เหมาะกับข้อมูลที่มีข้อมูลเรียนรู้จำนวนมากกว่าข้อมูลสอบถามมาก ๆ เพราะจะทำให้เวลาที่ลดลงจากการสร้างแผ่นแบบน้อย หรืออาจจะเพิ่มขึ้นกว่าเดิมในบางกรณี

ในปี 2011 Petitjean และคณะได้นำเสนอวิธีการสร้างแผ่นแบบใหม่ขึ้นชื่อว่าวิธีการหาแผ่นแบบเฉลี่ยแบบไดนามิกไทม์วอร์ปปิงแบรีเซนเตอร์ (Dynamic Time Warping Barycenter Averaging หรือ DBA) [25] โดยได้ทำการเปลี่ยนแปลงวิธีการปรับแนวรวมถึงวิธีการคัดเลือกคู่ลำดับ

ของข้อมูลอนุกรมเวลาที่นำมาสร้างแผ่นแบบเฉลี่ยใหม่ทั้งหมด วิธีการ DBA นี้จะเลือกข้อมูลอนุกรมเวลาเพียงหนึ่งตัวเป็นตัวหลักในการสร้างแผ่นแบบเฉลี่ยเรียกว่า Pivot จากนั้นทำการหาค่าเฉลี่ยข้อมูลอนุกรมเวลาทั้งหมดพร้อมกันในแกนของเวลาแทนที่จะเฉลี่ยข้อมูลทีละคู่ โดยค้นหาเส้นทางการปรับแนวของข้อมูลที่เหลือเทียบกับข้อมูล Pivot จากนั้นทำการหาค่าเฉลี่ยของเส้นทางที่เชื่อมกับข้อมูลย่อยที่จุดเวลานั้น ๆ ของข้อมูล Pivot ทุกจุดเพื่อเป็นค่าในจุดเวลานั้น ๆ ของแผ่นแบบเฉลี่ย ทำการไล่จุดเวลาของข้อมูล Pivot ตั้งแต่จุดแรกจนถึงจุดสุดท้ายก็จะได้ข้อมูลอนุกรมเวลาที่เป็นแผ่นแบบเฉลี่ยขึ้น อนึ่งวิธีการเลือกข้อมูล Pivot ของวิธีการที่ DBA นำเสนอนั้นจะใช้การสุ่มตัวใดตัวหนึ่งมาเป็นข้อมูล Pivot วิธีการ DBA โดยวิธีการอย่างละเอียดได้แสดงไว้ในหัวข้อ 2.8

สุดท้ายในปี 2014 Petitjean และคณะได้ทำการพัฒนาวิธีการ DBA อีกครั้งหนึ่ง เพื่อพัฒนาความแม่นยำในการจำแนกประเภทของวิธีการ DBA ชื่อว่า Nearest Centroid Classifier (NCC) [26] โดยแก้ปัญหาการแกว่งของความแม่นยำในการจำแนกประเภทอันเนื่องจากการสุ่มของจุด Pivot โดยเปลี่ยนเป็นการใช้ข้อมูลที่เป็นเมดอยด์ (Medoid) ของกลุ่มข้อมูลในคลาสนั้น ๆ เป็นจุด Pivot ที่ใช้ในการสร้างแผ่นแบบเฉลี่ย

นอกจากนั้นวิธีการ NCC ได้พัฒนาวิธีการให้สามารถสร้างแผ่นแบบที่มากกว่าหนึ่งแผ่นแบบต่อหนึ่งคลาสได้ โดยการแบ่งกลุ่มข้อมูลในคลาสเดียวกันออกเป็นกลุ่มย่อย ๆ ก่อนที่จะนำแต่ละกลุ่มย่อยไปทำการสร้างแผ่นแบบเฉลี่ย แม้แผ่นแบบจะมากขึ้นแต่ทำให้ความแม่นยำในการจำแนกประเภทข้อมูลสูงขึ้นเช่นกัน โดยวิธีการแบ่งกลุ่มที่ NCC ใช้นั้นมีหลากหลายวิธีการ แต่วิธีการที่สรุปว่าให้ผลดีนั้นคือการแบ่งกลุ่มข้อมูลแบบเคมีนส์ (*K*-means clustering) โดยใช้การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงในการวัดระยะทางของวิธีการแบ่งข้อมูล อย่างไรก็ตามวิธีการแบ่งกลุ่มข้อมูลแบบเคมีนส์นั้นยังต้องการจำนวนกลุ่มที่เหมาะสมในการแบ่งกลุ่มคลาสย่อยอยู่ ซึ่งในส่วนนี้ยังไม่มีวิธีการหาค่าที่ให้ผลลัพธ์ที่ดี ได้เพียงทำการใช้ค่าทดสอบไปตามลำดับ เพราะฉะนั้นจึงส่งผลให้วิธี NCC ใช้เวลามากในการสร้างแผ่นแบบเฉลี่ย โดยขั้นตอนของวิธีการ NCC สามารถอธิบายได้ดังรูปที่ 2.19



รูปที่ 2.19 แผนภูมิแสดงขั้นตอนวิธี NCC ที่พัฒนาเพิ่มเติมในส่วนของการจัดกลุ่มข้อมูลเพิ่มเติมก่อนทำการหาแผ่นแบบเฉลี่ยด้วยวิธีการ DBA

บทที่ 3 การสร้างแผ่นแบบเป็นตัวแทนข้อมูลเรียนรู้เพื่อใช้ในการจำแนกประเภทข้อมูล

แนวคิดที่งานวิจัยนี้จะนำเสนอเพื่อปรับปรุงวิธีการสร้างแผ่นแบบเฉลี่ยนั้น คือการปรับปรุงขั้นตอนการแบ่งข้อมูลในคลาสเดียวกันออกเป็นกลุ่มย่อย ๆ ก่อนที่จะนำไปสร้างแผ่นแบบเฉลี่ย โดยจากเดิมที่มีการใช้วิธีการต่าง ๆ ของการจัดกลุ่มข้อมูลที่มีแต่เดิมเช่น การจัดกลุ่มข้อมูลแบบเคมีนส์ (*K*-means clustering) การจัดกลุ่มข้อมูลแบบเคเมดอยด์ (*K*-medoid clustering) หรือการจัดกลุ่มตามลำดับขั้นแบบล่างขึ้นบน (Agglomerative Hierarchical clustering) เป็นต้น [26] ซึ่งวิธีเหล่านี้แม้ว่าจะได้ผลลัพธ์ที่ดีในแง่ของความแม่นยำในการจำแนกประเภท แต่มีความยากในการค้นหาค่าตัวแปรที่เหมาะสม ส่งผลให้ใช้เวลานานในการคำนวณและการสร้างแผ่นแบบเฉลี่ย งานวิจัยนี้จึงได้นำเสนอวิธีการใหม่ในการจัดกลุ่มข้อมูลอนุกรมเวลาที่อยู่ในคลาสเดียวกันออกเป็นกลุ่มย่อย ๆ ก่อนที่จะนำไปสร้างแผ่นแบบเฉลี่ย โดยสามารถลดเวลาในการสร้างแผ่นแบบเฉลี่ยลงในขณะที่คงไว้ซึ่งความแม่นยำในการจำแนกประเภท

ในบทที่ 3 นี้จะนำเสนอขั้นตอนวิธีการในการสร้างแผ่นแบบเป็นตัวแทนข้อมูลเรียนรู้ โดยประกอบด้วย การวิเคราะห์ที่มาของข้อมูลอนุกรมเวลา การจัดเตรียมข้อมูล ขั้นตอนวิธีการสร้างแผ่นแบบเฉลี่ย และสุดท้ายคือ การวัดผลแผ่นแบบที่สร้างขึ้น

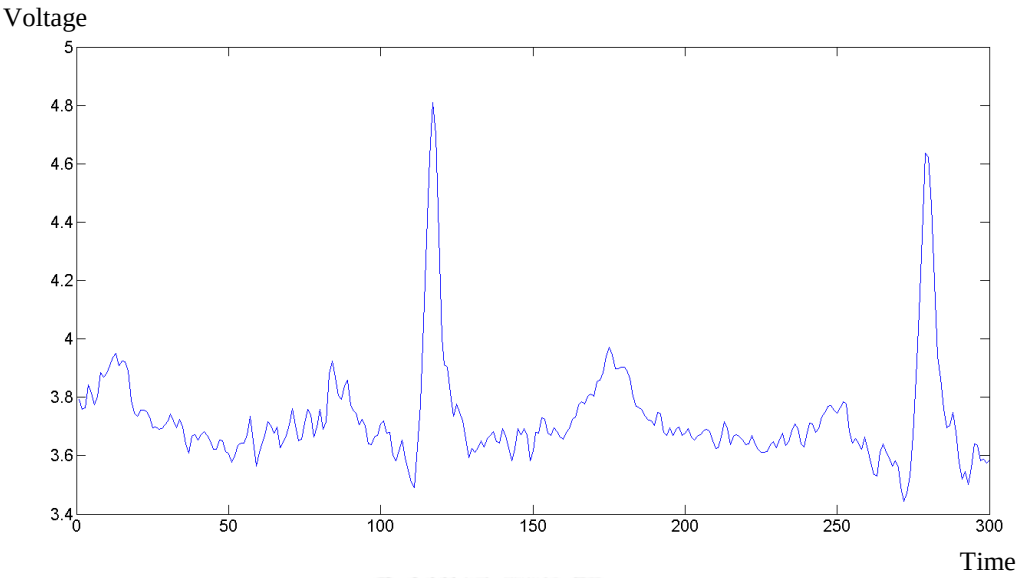
3.1 การวิเคราะห์ที่มาของข้อมูลอนุกรมเวลา

ข้อมูลอนุกรมเวลาที่งานวิจัยนี้ได้นำมาใช้ประกอบด้วยข้อมูลอนุกรมเวลาทั้งหมด 20 ชุดข้อมูล จากแหล่งข้อมูล UCR Time Series Classification Archive [27] โดยข้อมูลอนุกรมเวลาทั้ง 20 ชุดข้อมูลนั้นมีลักษณะแตกต่างกันไปทั้งขนาด ความยาวของเวลา หรือจำนวนคลาสของข้อมูล การวิเคราะห์ผลลัพธ์ที่ได้จากวิธีการที่สร้างขึ้นต้องอาศัยความรู้เกี่ยวกับลักษณะของข้อมูลที่ได้นำมาทำการทดลอง เนื่องด้วยข้อมูลอนุกรมเวลานั้นไม่จำเป็นที่จะต้องเกิดจากการเก็บข้อมูลไปเรื่อย ๆ ตามเวลาที่ผ่านไปเสมอ ยังมีการปรับใช้การสกัดลักษณะของข้อมูลในรูปแบบอื่น ๆ มาเป็นข้อมูลอนุกรมเวลาเพื่อให้เกิดความสะดวกในการการทำเหมืองข้อมูลอีกด้วย ซึ่งที่มาของข้อมูลอนุกรมเวลาทั้ง 20 ชุดข้อมูลที่ได้นำมาใช้นั้นประกอบด้วย 4 รูปแบบหลักดังนี้

3.1.1 ข้อมูลอนุกรมเวลาที่เกิดจากการเก็บข้อมูลตามเวลาโดยตรง

ข้อมูลอนุกรมเวลาลักษณะนี้เป็นการเก็บข้อมูลโดยใช้เครื่องมือในการวัดค่า การสังเกตจากผู้เชี่ยวชาญ หรือข้อมูลที่เกิดจากการจดบันทึกทางสถิติ ทำการบันทึกค่าของข้อมูลลงไปโดยแต่ละข้อมูลจะเชื่อมกับจุดเวลาที่แตกต่างกัน ตัวอย่างของข้อมูลในลักษณะนี้เช่น ข้อมูลคลื่นไฟฟ้าหัวใจ

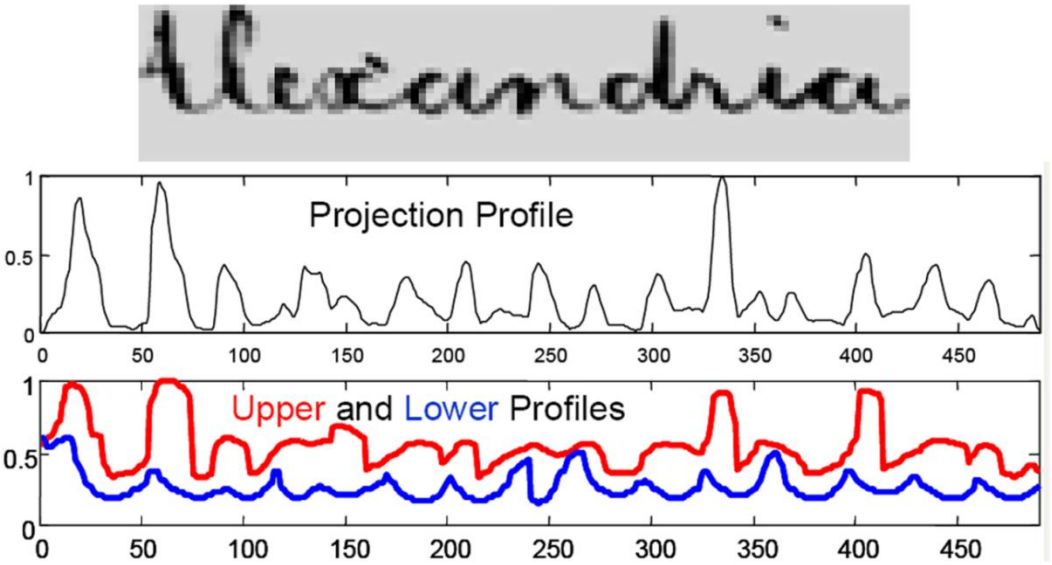
(Electrocardiogram) หรือข้อมูลค่าความต่างศักย์ของอุปกรณ์ไฟฟ้า รูปที่ 3.1 แสดงตัวอย่างลักษณะของข้อมูลคลื่นไฟฟ้าหัวใจโดยแกนตั้งเป็นค่าที่ทำการวัดและแกนนอนคือค่าของเวลา



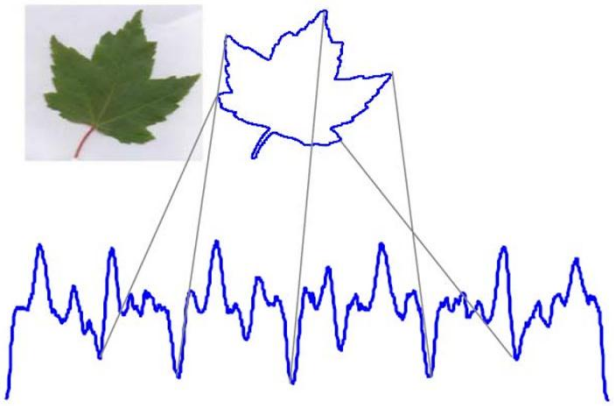
รูปที่ 3.1 ตัวอย่างข้อมูลคลื่นไฟฟ้าหัวใจ ที่เป็นข้อมูลอนุกรมเวลา (ที่มา Sivaraks และ Ratanamahatana [5])

3.1.2 ข้อมูลอนุกรมเวลาที่เกิดจากการแปลงข้อมูลรูปภาพ

ข้อมูลรูปภาพนับว่าเป็นข้อมูลที่มีความซับซ้อนสูง โดยทั่วไปจะเป็นการเก็บข้อมูลในลักษณะของตารางเมทริกซ์สองมิติ ในแต่ละช่องข้อมูลจะกล่าวถึงค่าของสีในแต่ละจุด เมื่อนำไปทำการเปรียบเทียบกันแบบจุดต่อจุดเพื่อวัดความคล้ายคลึงกัน อาจเป็นไปได้ที่ภาพทั้งสองมีลักษณะของสีที่ใกล้เคียงกัน ส่งผลให้ภาพมีความคล้ายมาก ในขณะที่ลักษณะของภาพอาจมีความแตกต่างกันมาก ในทำนองเดียวกัน ภาพที่มีลักษณะคล้ายคลึงกันแต่มีสีที่แตกต่างกันมาก ก็ส่งผลให้เมื่อวัดค่าความคล้ายคลึงอาจได้ผลลัพธ์ว่าแตกต่างกันมาก ส่งผลให้การเปรียบเทียบความคล้ายคลึงกันของข้อมูลรูปภาพมักจะมีการแปลงข้อมูลโดย “การสกัดลักษณะสำคัญของข้อมูล” (Feature Extraction) ออกมาเพื่อเน้นไปในส่วนของข้อมูลภาพที่สนใจทำการเปรียบเทียบ ในส่วนของแกนเวลาของข้อมูลอนุกรมที่จะได้นั้น เป็นไปได้ในหลายแนวทางเช่น การใช้แกนตั้งเป็นค่าของข้อมูลที่สนใจแล้วใช้แกนนอนเป็นแกนเวลา ดังแสดงในรูปที่ 3.2 ซึ่งเป็นการสกัดข้อมูลจากกข้อมูลภาพถ่ายลายมือ [32] หรือการสกัดข้อมูลรูปภาพโดยเริ่มจากค่าของจุดใดจุดหนึ่งของส่วนภาพที่สนใจแล้วเมื่อทำการเก็บข้อมูลจุดต่อไปก็ทำการเลื่อนแกนของเวลาตามไปด้วย ดังตัวอย่างที่แสดงในรูปที่ 3.3 ซึ่งเป็นการสกัดข้อมูลของรูปภาพใบไม้เป็นข้อมูลอนุกรมเวลา [2]



รูปที่ 3.2 การสกัดข้อมูลภาพลายมือเป็นข้อมูลอนุกรมเวลาโดยใช้ Projection Profile และ Upper and Lower Profiles (ที่มา Rath และ Manmatha [32])



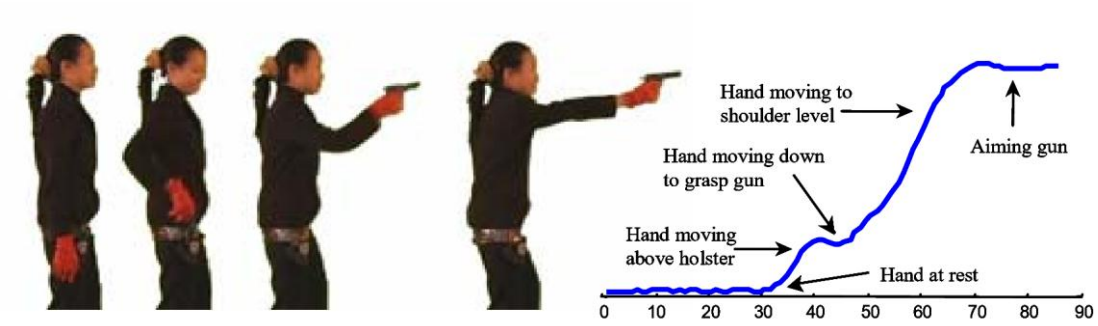
รูปที่ 3.3 การสกัดข้อมูลรูปภาพใบไม้เป็นข้อมูลอนุกรมเวลา (ที่มา Ratanamahatana และ Keogh [2])

ตัวอย่างของข้อมูลลักษณะนี้ได้แก่ ข้อมูลลักษณะของใบไม้ชนิดต่าง ๆ ข้อมูลที่สกัดจากตัวอักษร ข้อมูลที่สกัดจากภาพใบหน้ามนุษย์ในมุมมองต่าง ๆ เป็นต้น

3.1.3 ข้อมูลอนุกรมเวลาที่เกิดจากการแปลงจากข้อมูลการเคลื่อนไหว

ข้อมูลลักษณะนี้จะมีความคล้ายคลึงกับข้อมูลที่ทำการแปลงจากรูปภาพ แต่จะทำการเก็บข้อมูลในส่วนที่สนใจในรูปภาพจากการใช้เซนเซอร์ และเก็บลักษณะของการเคลื่อนไหวที่เกิดขึ้นเป็นค่าที่เปลี่ยนแปลงไปตามเวลา ตัวอย่างเช่น ข้อมูลการเคลื่อนไหวของมือเมื่อผู้ทดลองทำการ

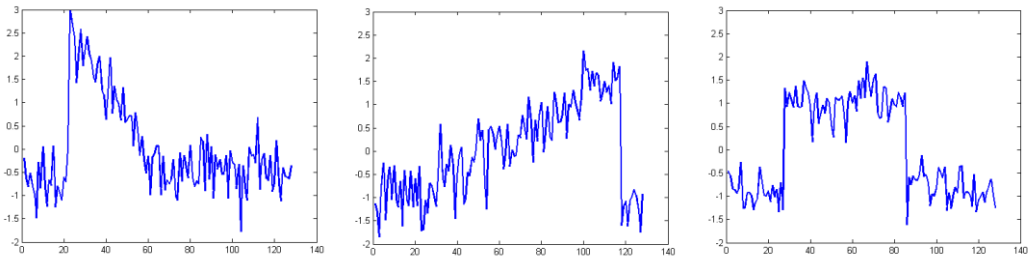
เคลื่อนไหวที่แตกต่างกันไป รูปที่ 3.4 แสดงลักษณะของการเก็บข้อมูลการเคลื่อนไหวของมือที่ทำการหยิบปืน



รูปที่ 3.4 ลักษณะการเก็บข้อมูลการเคลื่อนไหวของมือขณะทำการหยิบปืน (ที่มา Ratanamahatana และ Keogh [2])

3.1.4 ข้อมูลอนุกรมเวลาที่เกิดจากการสังเคราะห์ขึ้น

ข้อมูลในลักษณะนี้จะไม่เป็นข้อมูลจริงที่เกิดจากการเก็บค่า แต่เป็นข้อมูลที่เกิดจากการสังเคราะห์ขึ้นเพื่อทำการทดลองในลักษณะต่าง ๆ ซึ่งมีลักษณะของข้อมูลแตกต่างกันออกไปตามแต่ผู้สังเคราะห์ต้องการ โดยมีการสังเคราะห์ให้ข้อมูลมีความแตกต่างกันตั้งแต่ข้อมูลที่มีความแตกต่างกันน้อย ไปจนถึงข้อมูลที่มีการแทรกของสิ่งรบกวน (Noise) ลงไปในข้อมูลบางส่วน เพื่อทดสอบวิธีการต่าง ๆ ในการวิเคราะห์ข้อมูลอนุกรมเวลาในกรณีที่มีการเกิดสิ่งรบกวน ตัวอย่างข้อมูลในลักษณะนี้เช่น ข้อมูล Cylinder Bell Funnel (CBF) หรือข้อมูลลักษณะแพทเทิร์นขึ้นลงที่มีการเกิดการเคลื่อนของเวลา เป็นต้น รูปที่ 3.5 แสดงตัวอย่างของข้อมูล CBF ทั้งสามคลาส



รูปที่ 3.5 ข้อมูลทั้งสามคลาสของข้อมูล CBF ซึ่งถูกสังเคราะห์ขึ้น

3.2 การจัดเตรียมข้อมูล

ในขั้นตอนการจัดเตรียมข้อมูลนั้นเป็นการเตรียมพร้อมของข้อมูลในแต่ละชุดข้อมูลเพื่อให้พร้อมที่จะทำการสร้างแผ่นแบบเฉลี่ยในขั้นตอนถัดไป ประกอบด้วยการแปลงข้อมูลอนุกรมเวลาให้เป็นบรรทัดฐาน และการแบ่งข้อมูลแต่ละคลาสออกจากกัน

3.2.1 การแปลงข้อมูลอนุกรมเวลาให้เป็นบรรทัดฐาน

การเปรียบเทียบข้อมูลอนุกรมเวลาด้วยการใช้ตัววัดระยะทางต่าง ๆ นั้น สามารถใช้ข้อมูลของระยะทางในการวัดผลด้วยวิธีการต่าง ๆ ได้ แต่อย่างไรก็ตามข้อมูลระยะทางที่ได้นั้นอาจเกิดความผิดพลาดได้หากข้อมูลอนุกรมเวลาที่ใช้เปรียบเทียบระยะทางนั้นมีมาตราส่วน (Scale) ที่แตกต่างกัน ดังนั้นหากไม่ทราบแน่ชัดว่าข้อมูลอนุกรมเวลาที่นำมาใช้มีการทำให้ข้อมูลอนุกรมเวลาอยู่ในบรรทัดฐานเดียวกันหรือไม่ การแปลงข้อมูลอนุกรมเวลาให้เป็นบรรทัดฐานจึงเป็นสิ่งจำเป็นก่อนทำการวัดระยะทางระหว่างข้อมูลอนุกรมเวลา วิธีการที่ทำให้ข้อมูลอนุกรมเวลาอยู่ในบรรทัดฐานเดียวกันนั้นคือการปรับมาตราส่วนและแอมพลิจูด (Amplitude) ของข้อมูลให้อยู่ในระดับเดียวกัน วิธีการที่นำมาใช้ในการแปลงข้อมูลอนุกรมเวลาให้เป็นบรรทัดฐานที่ใช้ในงานวิจัยนี้คือการใช้คะแนน Z (Z Normalization) [41]

วิธีการแปลงข้อมูลอนุกรมเวลาให้เป็นบรรทัดฐานโดยการใช้คะแนน Z เป็นการแทนที่จุดข้อมูลเดิมของข้อมูลอนุกรมเวลาด้วยค่าใหม่ซึ่งเป็นค่าคะแนน Z ของข้อมูลแต่ละข้อมูล สามารถอธิบายวิธีการอย่างละเอียดได้ดังนี้

- กำหนดข้อมูลอนุกรมเวลา Q เป็นข้อมูลที่จะทำการแปลงเป็นบรรทัดฐาน ที่มีความยาวเท่ากับ n โดยที่ $Q = q_1, q_2, \dots, q_n$
- กำหนดข้อมูลอนุกรมเวลา Q_z เป็นข้อมูลที่ได้ทำการแปลงเป็นบรรทัดฐานจากข้อมูล Q โดย $Q_z = qz_1, qz_2, \dots, qz_n$ และค่าของแต่ละจุดข้อมูลสามารถคำนวณได้ดังสมการที่ 7

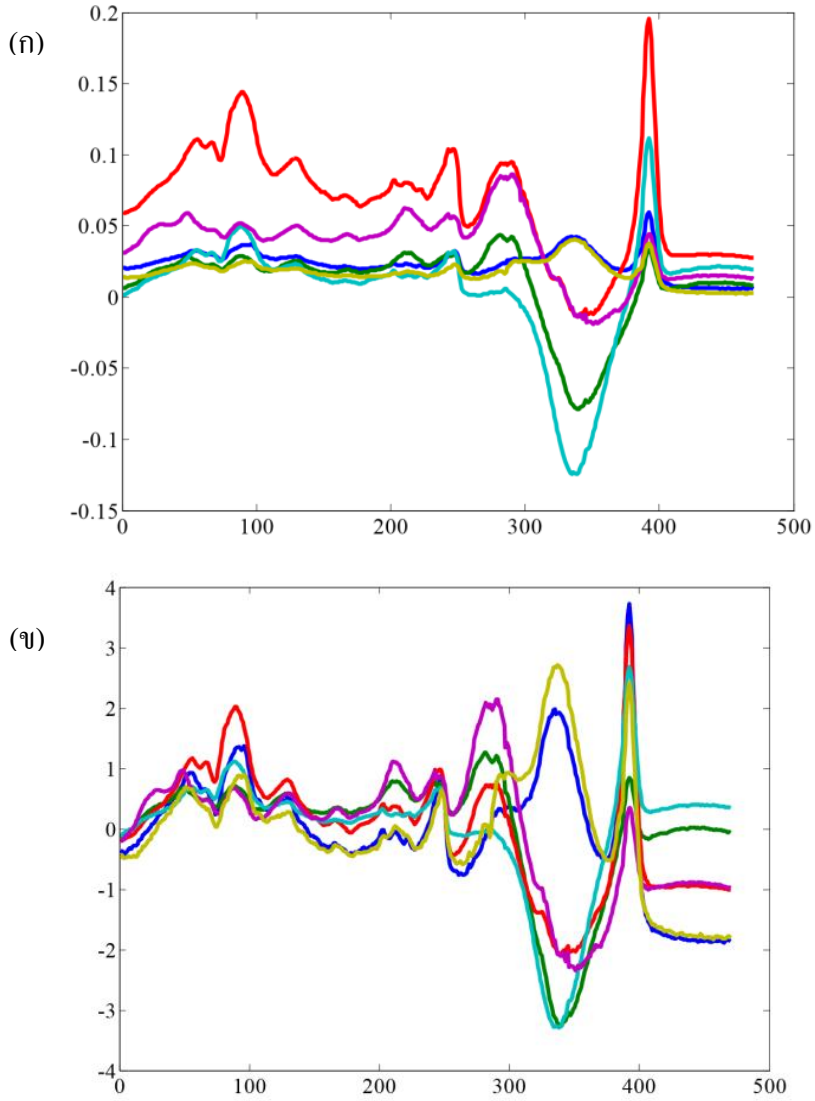
$$qz_i = \frac{q_i - \bar{q}}{SD} \tag{7}$$

โดยที่ค่า $\bar{q}$ และค่า SD คือค่าเฉลี่ยเลขคณิตของทุกจุดของข้อมูลอนุกรมเวลา Q และส่วนเบี่ยงเบนมาตรฐานของทุกจุดข้อมูลอนุกรมเวลา Q ตามลำดับ ซึ่งสามารถคำนวณได้จากสมการที่ 8 และ 9

$$\bar{q} = \frac{\sum_{i=1}^{n} q_i}{n} \tag{8}$$

$$SD = \sum_{i=1}^{n} \sqrt{\frac{(q_i - \bar{q})^2}{n}} \quad (9)$$

รูปที่ 3.6 แสดงตัวอย่างของข้อมูลอนุกรมเวลาก่อนการแปลงเป็นบรรทัดฐานและข้อมูลอนุกรมเวลาหลังจากทำการแปลงข้อมูลเป็นบรรทัดฐานแล้ว



รูปที่ 3.6 ข้อมูลอนุกรมเวลาก่อน (ก) และหลัง (ข) จากการทำการแปลงข้อมูลให้เป็นบรรทัดฐาน

3.2.2 การแบ่งข้อมูลแต่ละคลาสออกจากกัน

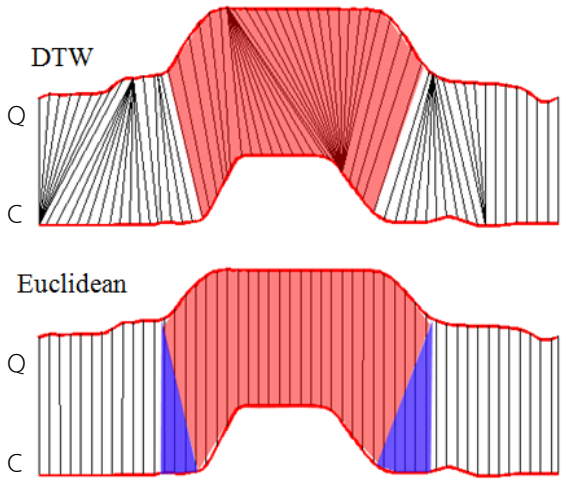
เนื่องด้วยการสร้างแผ่นแบบเฉลี่ยจากข้อมูลอนุกรมเวลานั้น ต้องการได้แผ่นแบบที่เป็นตัวแทนของข้อมูลเรียนรู้ ข้อมูลที่นำมาเฉลี่ยค่านั้นจึงเป็นข้อมูลในคลาสเดียวกันเท่านั้น หากนำข้อมูล

ที่แตกต่างคลาสมาทำการเฉลี่ยกัน ข้อมูลที่ได้จะไม่สามารถนำไปใช้งานได้ ในขั้นตอนนี้จึงได้ทำการแบ่งข้อมูลอนุกรมเวลาในแต่ละชุดข้อมูลออกเป็นกลุ่มย่อย ๆ โดยในแต่ละกลุ่มจะมีข้อมูลเพียงคลาสเดียวเท่านั้น ซึ่งการแบ่งข้อมูลแต่ละคลาสนั้นจะคงลำดับการเข้ามาของข้อมูลไว้ไม่มีการสลับที่

3.3 ขั้นตอนวิธีการสร้างแผ่นแบบเฉลี่ย (Template Averaging)

แนวคิดในการปรับปรุงวิธีการสร้างแผ่นแบบเฉลี่ย เพื่อให้สามารถสร้างแผ่นแบบเฉลี่ยโดยใช้เวลาที่น้อยลง ในขณะที่สามารถคงความแม่นยำในการจำแนกประเภทไว้ได้ ประกอบด้วยแนวคิดสำคัญที่ใช้ในการปรับปรุงวิธีการสามประการได้แก่

- การปรับใช้ตัววัดระยะทางแบบยุคลิดแทนที่การใช้การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงที่ใช้ระยะเวลามากในการคำนวณ โดยความเร็วของการวัดระยะทางแบบยุคลิดนั้นคือ $O(n)$ ในขณะที่การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงคือ $O(n^2)$ นอกจากนั้นการวัดระยะทางแบบยุคลิดยังส่งผลให้ระยะทางที่แสดงถึงความแตกต่างของข้อมูลอนุกรมเวลาที่ได้เด่นชัดกว่า รูปที่ 3.7แสดงการเปรียบเทียบระหว่างการวัดระยะทั้งสองแบบ กับข้อมูลอนุกรมเวลาคู่หนึ่งและส่วนของระยะทางที่เกิดจากความแตกต่างกันของการวัดระยะทางทั้งสองรูปแบบ



รูปที่ 3.7 การเปรียบเทียบการปรับแนว และส่วนของระยะทางที่จะเกิดความแตกต่างกัน

- การใช้วิธีการจัดกลุ่มแบบใหม่ที่สร้างขึ้น เพื่อทำการแบ่งข้อมูลอนุกรมเวลาในคลาสเดียวกันออกเป็นกลุ่มย่อย ๆ ก่อนที่จะทำการสร้างแผ่นแบบเฉลี่ย แนวคิดของวิธีการที่สร้างขึ้นคือการทำการจัดเรียงข้อมูลอนุกรมเวลาโดยเลือกข้อมูลอนุกรมเวลาตัวหนึ่งออกมาเป็นข้อมูลตัวหลัก (Pivot) ในส่วนของวิธีการนี้แล้วทำการเปรียบเทียบข้อมูลตัวอื่นก่อนที่จะทำการเรียงลำดับข้อมูล จากนั้นเมื่อข้อมูลทั้งหมดเรียงลำดับกันแล้ว จะใช้ค่าขีดแบ่ง (Threshold) ที่คำนวณจากข้อมูลที่มีเป็นหลักในการแบ่งแยกย่อยกลุ่มข้อมูล

- การใช้วิธีการอื่นในการหาข้อมูลตัวหลักของวิธีการ DBA ให้มีความรวดเร็วมากขึ้นกว่าการค้นหาเมดอยด์ด้วยระยะทางแบบไดนามิกไทม์วอร์ปปิงที่ไม่ใช้การสุ่มข้อมูล เนื่องด้วยผลลัพธ์ของแผ่นแบบเฉลี่ยที่ได้จะแตกต่างกันออกไปตามข้อมูลตัวหลัก ในงานวิจัยนี้ใช้วิธีการให้คะแนนกับข้อมูลแต่ละตัวด้วยการค้นหาข้อมูลเพื่อนบ้านใกล้ที่สุดแล้วให้คะแนน [12] จากนั้นทำการเลือกข้อมูลที่ได้คะแนนดีที่สุดมาเป็นข้อมูลตัวหลักในวิธีการ DBA

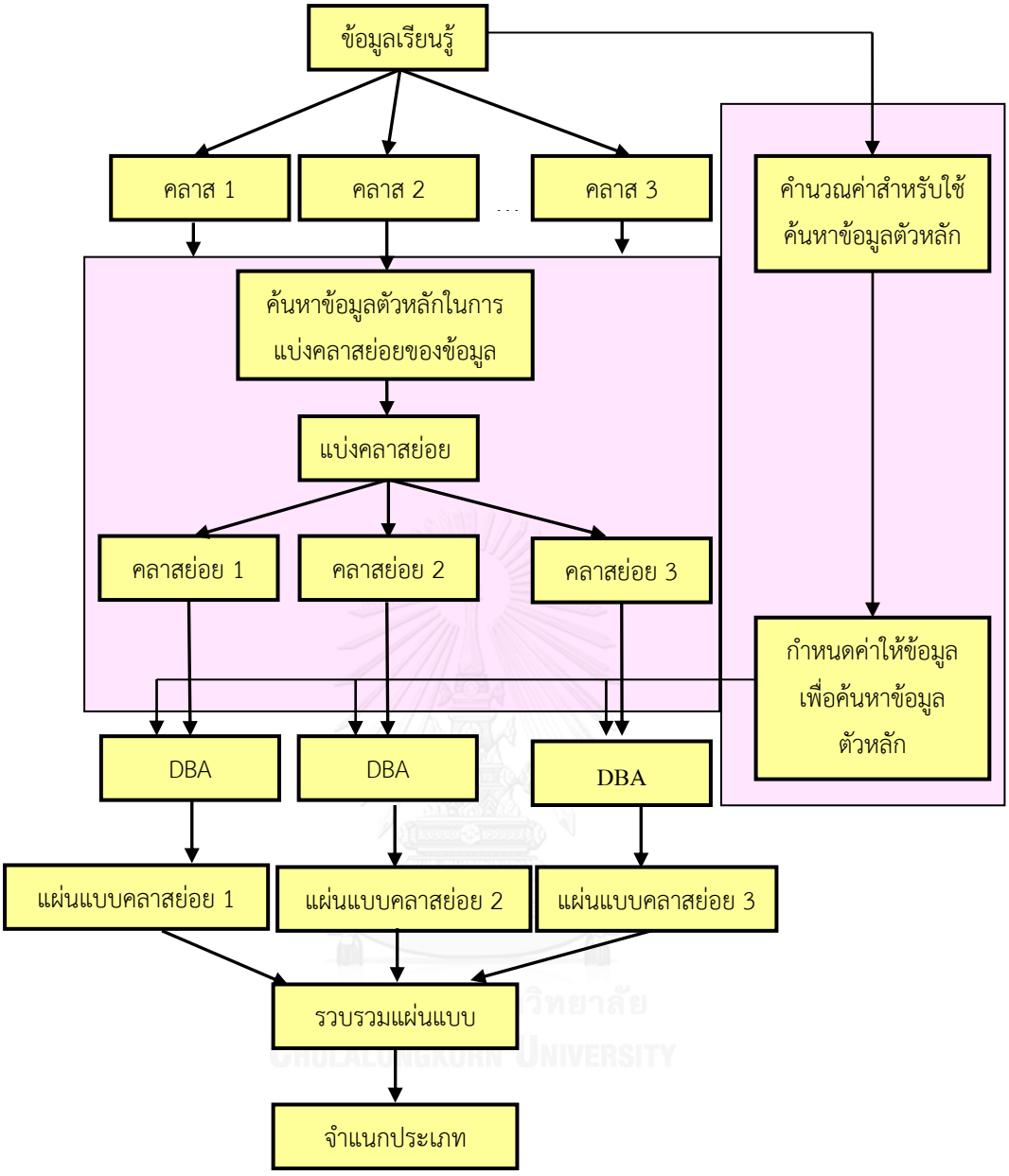
วิธีการสร้างแผ่นแบบที่ได้นำเสนอโดยปรับปรุงตามแนวคิดสามประการที่ได้กล่าวไว้นั้น สามารถแบ่งเป็นขั้นตอนได้ทั้งหมด 4 ขั้นตอนได้แก่

- ขั้นตอนที่ 1 การค้นหาข้อมูลตัวหลักสำหรับใช้ในการจัดลำดับข้อมูลอนุกรมเวลา
- ขั้นตอนที่ 2 การจัดลำดับข้อมูลอนุกรมเวลาโดยเปรียบเทียบกับข้อมูลตัวหลัก
- ขั้นตอนที่ 3 การคำนวณหาค่าขีดแบ่ง แล้วทำการแบ่งข้อมูลออกเป็นกลุ่ม ๆ
- ขั้นตอนที่ 4 การให้คะแนนข้อมูล เพื่อทำการหาข้อมูลตัวหลักของวิธีการ DBA

รูปที่ 3.8 แสดงขั้นตอนทั้งหมดของวิธีการสร้างแผ่นแบบเฉลี่ย โดยจุดที่ทำการปรับปรุงและทำการทดลอง จะอยู่ในส่วนที่มีการแรเงาพื้นหลัง

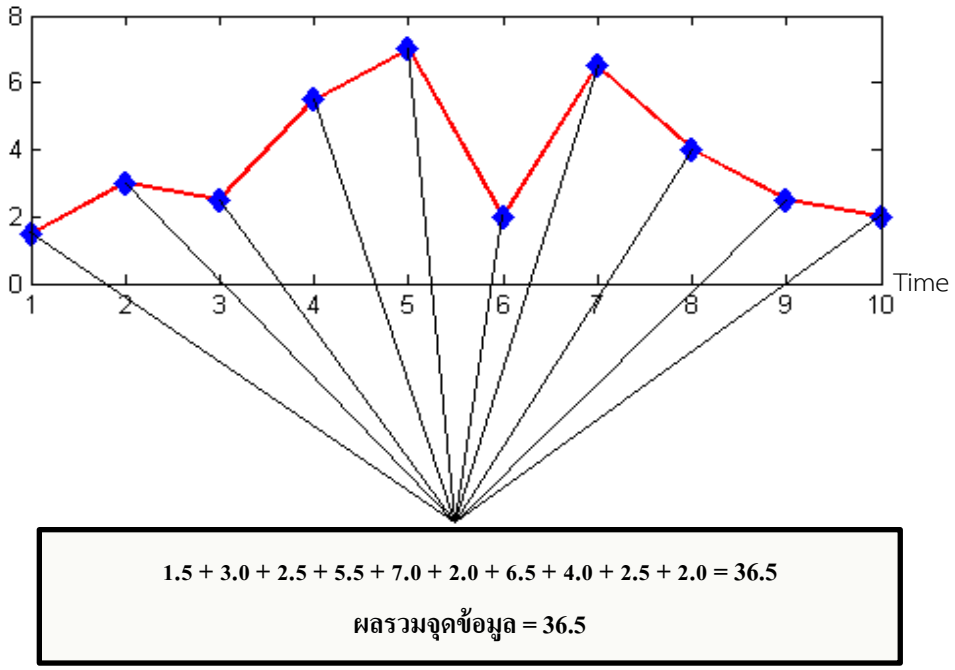
3.3.1 การค้นหาข้อมูลตัวหลักสำหรับใช้ในการจัดลำดับข้อมูลอนุกรมเวลา

เนื่องด้วยข้อมูลอนุกรมเวลาแต่ละข้อมูล ไม่ได้ประกอบด้วยข้อมูลเพียงค่าเดียว แต่ประกอบไปด้วยข้อมูลที่มีค่าหลากหลายแตกต่างกันที่เชื่อมต่อกับจุดเวลา และมีลำดับในตัวเอง ฉะนั้นข้อมูลอนุกรมเวลาจึงไม่สามารถที่จะจัดเรียงลำดับของข้อมูลก่อนหลังได้ในทันทีดังเช่นข้อมูลที่เป็นตัวเลขที่มีความมากน้อยที่ชัดเจน หรือข้อมูลตัวอักษรที่มีลำดับก่อนหลังที่ชัดเจน ในที่นี้จึงได้มีแนวคิดที่จะทำการเปรียบเทียบข้อมูลอนุกรมเวลา โดยใช้ข้อมูลตัวใดตัวหนึ่งในกลุ่มข้อมูลเป็นข้อมูลตัวหลักแล้วทำการวัดระยะทางจากข้อมูลตัวหลักนั้น ๆ ก่อนที่จะนำข้อมูลระยะทางนั้น ๆ ไปใช้ในการจัดลำดับข้อมูล



รูปที่ 3.8 แผนภูมิแสดงขั้นตอนวิธีการทั้งหมดในการสร้างแผ่นแบบเฉลี่ย

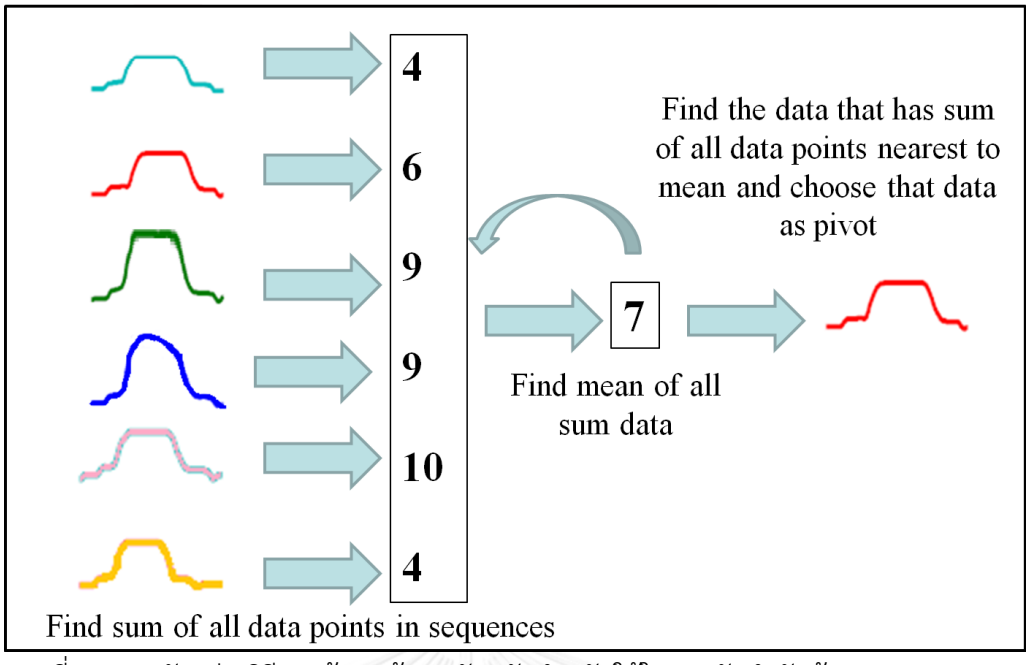
ในส่วนของขั้นตอนนี้จะทำการค้นหาข้อมูลตัวหลักที่จะใช้ในขั้นตอนจัดลำดับข้อมูลอนุกรมเวลา โดยข้อมูลที่ต้องการเลือกนั้นเป็นข้อมูลที่เป็นข้อมูลกึ่งกลางของข้อมูลทั้งหมดในคลาส เริ่มต้นจากการคำนวณ “ผลรวมจุดข้อมูล” โดยทำการรวมค่าของจุดข้อมูลทุกจุดในแต่ละจุดเวลาออกเป็นค่า ๆ เดียว ค่าของจุดเวลาในที่นี้จะเป็นค่าที่เปรียบเทียบกับค่า 0 เพราะฉะนั้นหากค่าจุดข้อมูลในจุดเวลาใดมีค่าติดลบ ค่านั้นจะผ่านการแปลงเป็นค่าบวกก่อนที่จะนำมารวมกับค่าผลรวมจุดข้อมูล วิธีการหาผลรวมจุดข้อมูลสามารถอธิบายได้ดังรูปที่ 3.9



รูปที่ 3.9 แสดงวิธีการหาค่าผลรวมจุดข้อมูล

หลังจากที่ได้ค่าผลรวมจุดข้อมูลของข้อมูลอนุกรมเวลาครบทุกตัวแล้ว จะทำการคำนวณหาค่าเฉลี่ยของค่าผลรวมจุดข้อมูลทั้งหมดที่ได้มา แล้วทำการค้นหาข้อมูลอนุกรมเวลาที่มีค่าผลรวมจุดข้อมูลใกล้เคียงกับค่าเฉลี่ยของค่าผลรวมจุดข้อมูลมากที่สุด เพื่อใช้เป็นข้อมูลอนุกรมเวลาข้อมูลนั้นเป็นข้อมูลตัวหลักในการจัดอันดับข้อมูล หากมีข้อมูลอนุกรมเวลาที่มีผลรวมจุดข้อมูลใกล้เคียงกับค่าเฉลี่ยมากที่สุดมากกว่า 1 ข้อมูลจะทำการเลือกข้อมูลตัวตามลำดับของข้อมูลที่เข้ามา

ตัวอย่างของวิธีการค้นหาข้อมูลตัวหลักสำหรับใช้ในการจัดลำดับข้อมูลอนุกรมเวลาสามารถอธิบายได้ดังรูปที่ 3.10 สมมติให้มีข้อมูลอนุกรมเวลาทั้งหมด 6 ข้อมูล ขั้นตอนแรกทำการหาผลรวมค่าจุดข้อมูลของแต่ละข้อมูล โดยสมมติให้มีค่าเท่ากับ 4 6 9 9 10 4 ดังรูป จะสังเกตเห็นได้ว่ามีความเป็นไปได้ที่ข้อมูลอนุกรมเวลาสองข้อมูลที่มีรูปร่างแตกต่างกันจะมีค่าผลรวมของจุดข้อมูลเท่ากัน จากนั้นทำการคำนวณค่าเฉลี่ยของค่าผลรวมจุดข้อมูลนั้นได้เท่ากับ (4+6+9+9+10+4)/6 = 7 สุดท้ายจึงค้นหาว่าข้อมูลใดมีผลรวมจุดข้อมูลใกล้เคียงกับค่า 7 มากที่สุด ในตัวอย่างคือข้อมูลที่ 2 ที่มีค่าเท่ากับ 6 ที่มีความใกล้เคียงกับค่าผลเฉลี่ยมากที่สุด จึงได้เลือกข้อมูลที่ 2 เป็นข้อมูลตัวหลักในการจัดอันดับข้อมูล อนึ่งกรณีที่มีข้อมูลใกล้กับค่าเฉลี่ยของค่าผลรวมจุดข้อมูลมากกว่าหนึ่งข้อมูล จะทำการเลือกข้อมูลแรกตามลำดับการคำนวณของข้อมูล

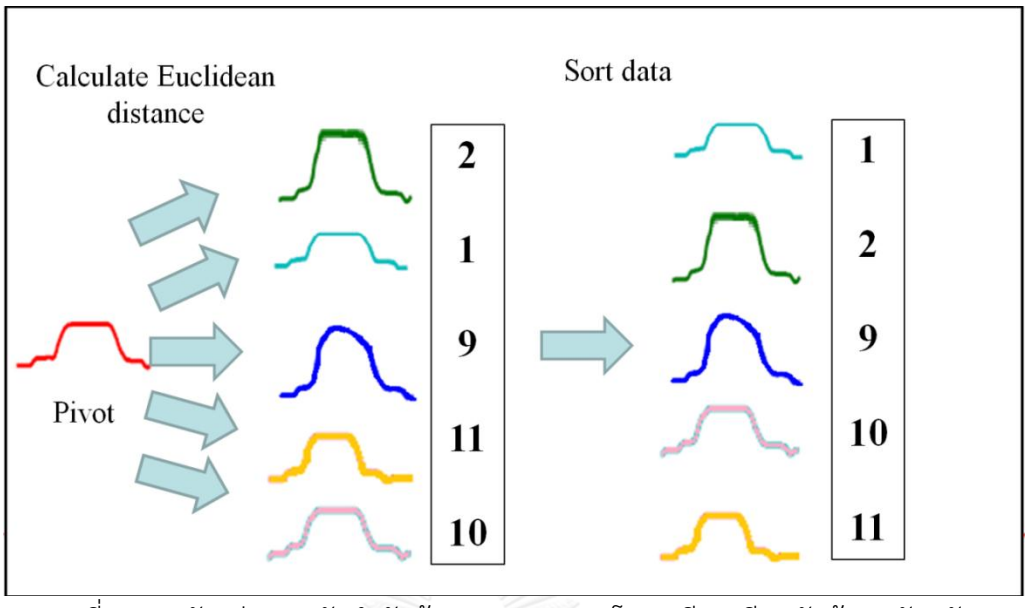


รูปที่ 3.10 ตัวอย่างวิธีการค้นหาข้อมูลตัวหลักสำหรับใช้ในการจัดลำดับข้อมูลอนุกรมเวลา

3.3.2 การจัดลำดับข้อมูลอนุกรมเวลาโดยเปรียบเทียบกับข้อมูลตัวหลัก

หลังจากได้ข้อมูลตัวหลักที่จะใช้ในการเปรียบเทียบมาแล้ว ให้ทำการคำนวณหาระยะทางระหว่างข้อมูลตัวอื่น ๆ ที่เหลืออยู่กับข้อมูลตัวหลักโดยใช้ตัววัดความคล้ายแบบยุคลิด ซึ่งในการคำนวณระยะทางในขั้นตอนนี้จะทำการคำนวณระยะทางแบบเทียบทั้งข้อมูลอนุกรมเวลา ดังวิธีที่ได้นำเสนอในบทที่ 2 หัวข้อที่ 2.3 หลังจากที่ได้ระยะทางแบบยุคลิดของข้อมูลตัวอื่น ๆ เทียบกับตัวหลักแล้ว จะทำการจัดลำดับข้อมูลโดยเรียงจากระยะทางที่น้อยที่สุดไปหาระยะทางที่มากที่สุด

ตัวอย่างของการจัดลำดับข้อมูลอนุกรมเวลาโดยเปรียบเทียบกับข้อมูลตัวหลักแสดงได้ดังรูปที่ 3.11 โดยจะสังเกตว่าค่าระยะทางแบบยุคลิดที่ได้ในขั้นตอนนี้ไม่จำเป็นต้องเรียงลำดับตามความใกล้เคียงของค่าผลรวมจุดข้อมูลที่หาในขั้นตอนก่อนหน้า หลังจากที่ได้ค่าระยะทางครบทุกข้อมูลแล้วก็ทำการเรียงลำดับข้อมูลจากน้อยไปหามากตามค่าระยะทางของแต่ละข้อมูล



รูปที่ 3.11 ตัวอย่างการจัดลำดับข้อมูลอนุกรมเวลาโดยเปรียบเทียบกับข้อมูลตัวหลัก

3.3.3 การคำนวณหาค่าขีดแบ่ง แล้วทำการแบ่งข้อมูลออกเป็นกลุ่ม ๆ

เมื่อข้อมูลทั้งหมดมีการเรียงลำดับตามระยะทางเทียบกับข้อมูลตัวหลักแล้ว จะทำการคำนวณหาค่าขีดแบ่งเพื่อนำมาใช้ในการแบ่งข้อมูลที่มีการเรียงลำดับแล้ว โดยค่าขีดแบ่งที่จะนำมาใช้ในวิธีการที่ได้นำเสนอคือค่าเบี่ยงเบนมาตรฐาน ซึ่งเป็นค่าที่ใช้ในการวัดการกระจายตัวของข้อมูล แต่ค่าเบี่ยงเบนมาตรฐานนี้ไม่สามารถใช้ค่าที่คำนวณได้โดยตรงจากค่าระยะทางของข้อมูลเทียบกับตัวหลักในทันที เนื่องจากค่าเบี่ยงเบนมาตรฐานที่คำนวณจากวิธีนี้จะบ่งบอกถึงเพียงแค่ระยะทางจากจุดกึ่งกลางของข้อมูลเท่านั้นทำให้ไม่สามารถแบ่งกลุ่มข้อมูลเป็นจำนวนมากได้

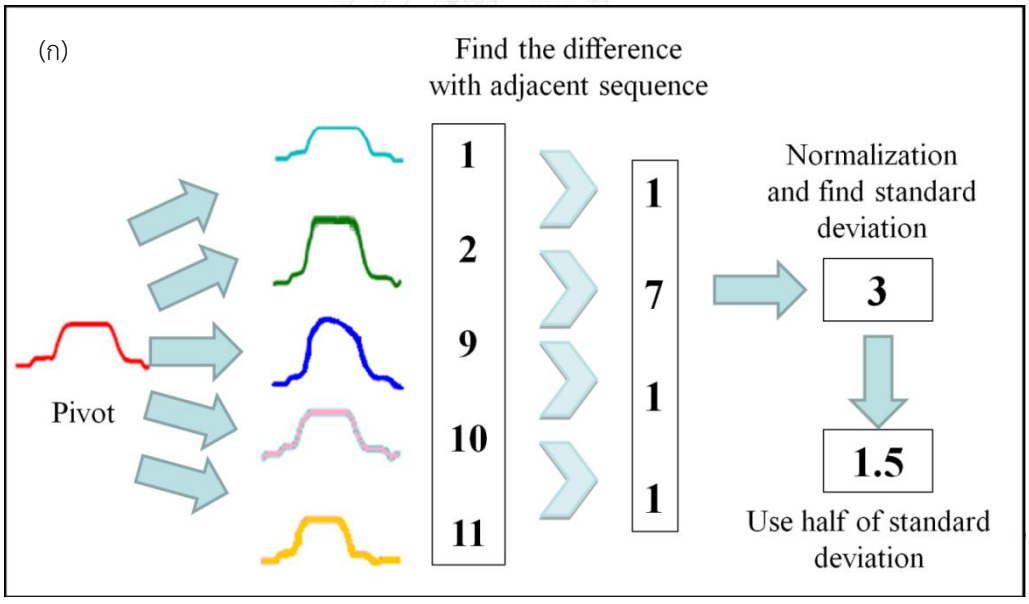
ในที่นี้จึงได้ทำการคำนวณค่าความต่างของระยะทางที่เกิดจากการคำนวณเมื่อเทียบกับข้อมูลตัวหลักระหว่างข้อมูลที่ติดกัน ซึ่งในที่นี้จะเรียกว่า “ส่วนต่างระหว่างข้อมูลที่ติดกัน” (Difference of adjacent data) โดยทำการคำนวณจากค่าระยะทางของข้อมูลอนุกรมเวลาแต่ละข้อมูลเมื่อทำการคำนวณระยะทางเปรียบเทียบกับข้อมูลตัวหลัก ซึ่งค่าส่วนต่างระหว่างข้อมูลที่ติดกันนี้จะบ่งบอกว่าข้อมูลตัวนั้น ๆ มีความแตกต่างจากข้อมูลก่อนหน้ามากเพียงใด และเนื่องจากนำข้อมูลการเปรียบเทียบของหลายคู่ข้อมูลมาใช้ จึงได้ทำการแปลงข้อมูลให้เป็นบรรทัดฐานก่อน เพื่อให้ข้อมูลทุกค่าอยู่ในบรรทัดฐานเดียวกัน จากนั้นจึงทำการคำนวณค่าส่วนเบี่ยงเบนมาตรฐานจากข้อมูลของส่วนต่างระหว่างข้อมูลที่ติดกันทั้งหมดที่ได้มา

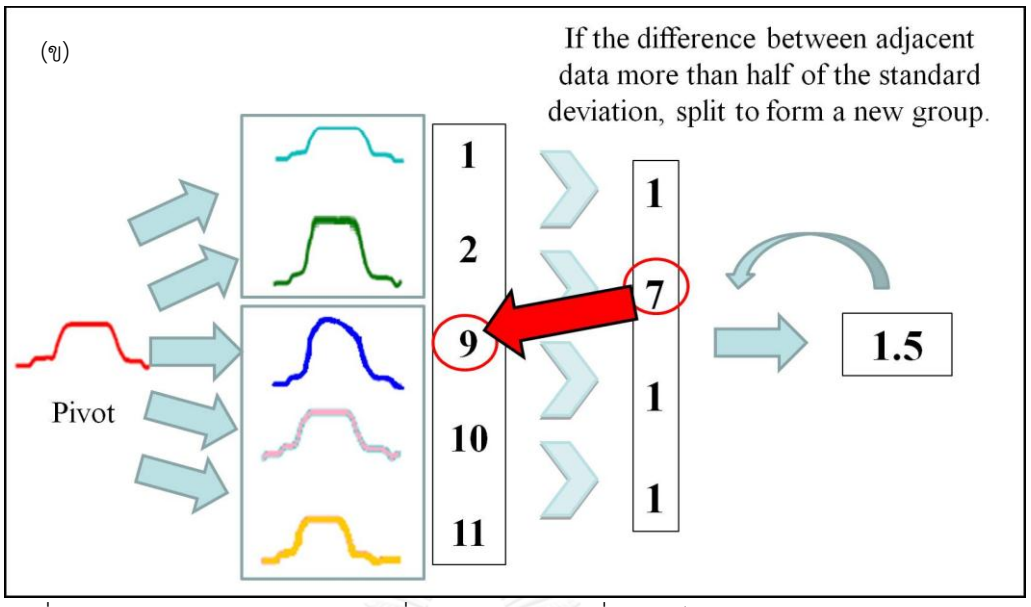
เมื่อได้ค่าส่วนเบี่ยงเบนมาตรฐานแล้ว จะใช้ค่าที่ได้ไปเปรียบเทียบข้อมูลค่าส่วนต่างระหว่างข้อมูลแต่ละตัว หากค่าส่วนต่างระหว่างข้อมูลตัวใดมีค่าสูงเกินกว่าครึ่งหนึ่งของค่าส่วนเบี่ยงเบน

มาตรฐาน จะทำการแบ่งข้อมูลตัวถัดไปที่ทำการคำนวณค่าส่วนต่างด้วยเป็นข้อมูลกลุ่มใหม่ โดยตีความว่ามีความแตกต่างจากข้อมูลก่อนหน้ามากจนมีความเหมาะสมที่จะสร้างแผ่นแบบเฉลี่ยใหม่เพิ่มขึ้นสำหรับข้อมูลตัวถัดไป จากนั้นทำการเปรียบเทียบแล้วแบ่งกลุ่มข้อมูลจนกระทั่งครบทุกข้อมูลที่มีอยู่ อนึ่งข้อมูลตัวหลักที่ได้นั้นจะถูกจัดเป็นแผ่นแบบทันทีเนื่องจากไม่มีข้อมูลตัวอื่นมาเป็นหลักในการเปรียบเทียบ

สาเหตุที่ใช้ค่าครึ่งหนึ่งของส่วนเบี่ยงเบนมาตรฐานนั้น มาจากเหตุผลที่ว่าเมื่อใช้ค่าส่วนเบี่ยงเบนมาตรฐานเต็มจำนวน จะครอบคลุมการกระจายระหว่างข้อมูลที่มาก [42] อาจส่งผลให้ไม่สามารถแบ่งกลุ่มย่อยบางกลุ่มได้ ซึ่งจากการทดลองกับข้อมูลทุกชุดข้อมูล ได้ผลที่ว่าค่าครึ่งหนึ่งของส่วนเบี่ยงเบนมาตรฐานนั้นมีความเหมาะสมในการแบ่งกลุ่มย่อยให้กับข้อมูลในคลาสเดียวกัน อีกทั้งในการใช้งานในข้อมูลบางกลุ่มมีการใช้ค่าครึ่งหนึ่งของส่วนเบี่ยงเบนมาตรฐานเช่นกัน [43]

รูปที่ 3.12 แสดงวิธีการการคำนวณหาค่าขีดแบ่ง แล้วทำการแบ่งข้อมูลออกเป็นกลุ่ม โดยในที่นี้ได้ใช้ข้อมูลที่มีระยะทางระหว่างข้อมูลตัวหลักที่ผ่านการเรียงลำดับมาแล้วจากรูปที่ 3.10 จะสังเกตเห็นได้ว่าค่าส่วนต่างระหว่างข้อมูลที่ติดกันของข้อมูลลำดับที่ 2 กับ 3 จะมีความแตกต่างกันมาก เมื่อทำการใช้ค่าส่วนเบี่ยงเบนมาตรฐานมาทำการตัดแบ่งจะเป็นจุดที่ข้อมูลถูกแบ่งออกจากกัน





รูปที่ 3.12 (ก) การคำนวณค่าส่วนเบี่ยงเบนมาตรฐานที่จะใช้เป็นค่าขีดแบ่ง (ข) การใช้ค่าส่วนเบี่ยงเบนมาตรฐานในการแบ่งกลุ่มข้อมูลออกจากกัน

หลังจากแบ่งกลุ่มข้อมูลเสร็จแล้วจะทำการสร้างแผ่นแบบเฉลี่ยของแต่ละกลุ่มข้อมูลออกมา โดยที่แผ่นแบบเฉลี่ยของทุกกลุ่มข้อมูลนั้นจะมีคลาสเดียวกันกับคลาสตั้งต้นเดิม

3.3.4 การให้คะแนนข้อมูล เพื่อทำการหาข้อมูลตัวหลักของวิธีการ DBA

ในวิธีการหาค่าเฉลี่ยแบบไดนามิกไทม์วอร์ปปิงแบรีเซนเตอร์ หรือ DBA ที่ได้กล่าวไว้ในหัวข้อที่ 2.8 นั้นต้องการข้อมูลตัวหลักหรือ Pivot ในการคำนวณเพื่อสร้างแผ่นแบบเฉลี่ย โดยเริ่มต้นนั้นวิธีการ DBA ได้ทำการสุ่มข้อมูลเริ่มต้นจากข้อมูลอนุกรมเวลาที่มีอยู่ [25] แต่เนื่องจากค่าผลลัพธ์ของแผ่นแบบที่ได้มีความแตกต่างกันขึ้นอยู่กับข้อมูลตัวหลักที่เลือก การใช้ค่าสุ่มจึงเกิดการแกว่งของผลลัพธ์ ในวิธีการ Nearest Centroid Classifier (NCC) ที่ได้พัฒนาต่อมาจึงได้เปลี่ยนการเลือกตัวหลักของวิธีการ DBA เป็นการใช้ข้อมูลที่เป็นเมดอยด์ของข้อมูล

เมดอยด์ของข้อมูลนั้นคือข้อมูลที่มีความใกล้เคียงกับข้อมูลตัวอื่น ๆ ที่เหลือมากที่สุด วิธีการค้นหาข้อมูลเมดอยด์จะเริ่มจากการค้นหาระยะทางระหว่างข้อมูลทุกตัวเช่น ข้อมูลที่ 1 ค้นหาระยะทางเทียบกับข้อมูลที่ 2 3 และ 4 จากนั้นข้อมูลที่ 2 ค้นหาระยะทางเทียบกับข้อมูลที่ 3 และ 4 แล้วใช้ข้อมูลระยะทางเทียบกับข้อมูลที่ 1 ที่ได้หามาแล้ว ทำซ้ำจนกระทั่งครบทุกคู่ข้อมูล อย่างไรก็ตามในวิธีการ NCC นั้นใช้การคำนวณระยะทางแบบไดนามิกไทม์วอร์ปปิง ส่งผลให้การคำนวณหาเม

ดอยด์เพื่อนำมาใช้ในวิธีการ DBA ใช้เวลาสูง ในงานวิจัยนี้จึงได้นำเสนอวิธีการใหม่เพื่อนำมาใช้ค้นหาข้อมูลตัวหลักในวิธีการ DBA

วิธีการที่นำเสนอนั้นอยู่ในแนวคิดของการให้คะแนนข้อมูลซึ่งอ้างอิงมาจากวิธีการ Simple Rank Method (SR) [12] ซึ่งเป็นวิธีการในการให้คะแนนจากการคำนวณหาระยะทางระหว่างข้อมูลเรียนรู้ทั้งหมด ซึ่งในที่นี้จะใช้ตัววัดความคล้ายแบบยุคลิดในการคำนวณระยะทาง จากนั้นค้นหาข้อมูลเพื่อนบ้านใกล้ที่สุดของข้อมูลทุกตัว คะแนนของข้อมูลแต่ละตัวจะอยู่ในเงื่อนไขที่ว่า “ข้อมูลที่เป็นเพื่อนบ้านใกล้ที่สุดให้ข้อมูลใด ๆ จะมีคะแนนเพิ่มขึ้น 1 ต่อข้อมูล 1 ตัวที่เป็นเพื่อนบ้านใกล้ที่สุดด้วยหากข้อมูลนั้น ๆกับข้อมูลที่เป็นเพื่อนบ้านใกล้ที่สุดมีคลาสเดียวกัน” ในที่นี้จะเรียกคะแนนนี้ว่า Nearest Neighbor Point หรือ NN-point ซึ่งสามารถอธิบายได้ดังสมการที่ 3.1 โดย x คือข้อมูลใด ๆ และ x_i คือข้อมูลที่ x เป็นเพื่อนบ้านใกล้สุดด้วย

$$NN-point(x) = \sum_i \begin{Bmatrix} 1 \; if \; class(x) = class(x_i) \\ 0 \qquad\qquad otherwise \end{Bmatrix} \tag{10}$$

ตัวอย่างการคำนวณค่า NN-point สามารถอธิบายได้ดังรูปที่ 3.13 ยกตัวอย่างข้อมูลอนุกรมเวลาทั้งหมด 5 ข้อมูล โดยข้อมูลที่ 1,2 และ 3 มีคลาส 1 และข้อมูลที่ 4 และ 5 มีคลาสที่ 2 ตาราง E แสดงถึงระยะทางระหว่างข้อมูลแต่ละตัวที่คำนวณโดยการใช้ตัววัดความคล้ายแบบยุคลิด ช่องที่ถูกแรเงาแสดงถึงข้อมูลที่เป็นเพื่อนบ้านใกล้ที่สุดของข้อมูลอนุกรมเวลาที่ทำการเปรียบเทียบ ตัวอย่างเช่น ช่องที่เกิดการวงกลมไว้จะหมายถึงข้อมูลที่สองเป็นข้อมูลเพื่อนบ้านใกล้สุดลำดับที่ 1 ของข้อมูลที่ 1 จากนั้นทำการตรวจสอบว่าข้อมูลทั้งสองข้อมูลเป็นข้อมูลคลาสเดียวกันหรือไม่แล้วทำการให้คะแนนเท่ากับ 1 คะแนนหากเป็นคลาสเดียวกัน สุดท้ายทำการรวมคะแนนทั้งหมดให้แก่ข้อมูลอนุกรมเวลาทุกตัว โดยคำนวณคะแนนจากการการที่ข้อมูลนั้น ๆ เป็นข้อมูลเพื่อนบ้านใกล้ที่สุดลำดับที่ 1 กับข้อมูลอื่นแล้วเป็นคลาสเดียวกันมากน้อยเพียงใด ตัวอย่างเช่นข้อมูลอนุกรมเวลาที่ 2 เป็นข้อมูลเพื่อนบ้านใกล้ที่สุดลำดับที่ 1 กับข้อมูลที่ 1 และข้อมูลที่ 4 แต่ข้อมูลที่ 4 เป็นข้อมูลคนละคลาสกับข้อมูลที่ 2 ข้อมูลที่ 2 จึงมีคะแนน NN-point เท่ากับ 1 คะแนน

หลังจากนั้นใช้คะแนนของข้อมูลนั้น ๆ ในการใช้คัดเลือกข้อมูลที่จะเป็นข้อมูลตัวหลักของวิธีการ DBA โดยใช้ข้อมูลที่มีคะแนนมากที่สุดในแต่ละคลาสเป็นข้อมูลตัวหลัก กล่าวคือในตัวอย่างนี้จะใช้ข้อมูลที่ 1 เป็นข้อมูลตัวหลักของคลาสที่ 1 และใช้ข้อมูลที่ 4 เป็นข้อมูลตัวหลักของคลาสที่ 2 อนึ่งหากมีข้อมูลที่มีค่า NN-point เท่ากันจะเลือกข้อมูลแรกที่เข้ามาตามลำดับของข้อมูล

E	Seq 1	Seq 2	Seq 3	Seq 4	Seq 5
Seq 1 (1)	-	2	3	8	7
Seq 2 (1)	2	-	5	3	5
Seq 3 (1)	3	5	-	6	6
Seq 4 (2)	8	3	6	-	4
Seq 5 (2)	7	5	6	4	-

NN	Seq 1	Seq 2	Seq 3	Seq 4	Seq 5
Seq 1 (1)	-	1	0	0	0
Seq 2 (1)	1	-	0	0	0
Seq 3 (1)	1	0	-	0	0
Seq 4 (2)	0	0	0	-	0
Seq 5 (2)	0	0	0	1	-
Sum of NN-point	**2**	**1**	**0**	**1**	**0**

รูปที่ 3.13 ตัวอย่างการคำนวณค่าคะแนนที่จะใช้คัดเลือกตัวหลักของ DBA ชื่อว่า NN-point

3.4 การวัดผลแผ่นแบบเฉลี่ยที่สร้างขึ้น

ในส่วนของการวัดผลแผ่นแบบเฉลี่ยที่สร้างขึ้นนั้น จะใช้การเปรียบเทียบประสิทธิภาพการทำงานของวิธีการที่นำเสนอและคุณภาพของแผ่นแบบสองประการได้แก่ การเปรียบเทียบเวลาที่ใช้ในการสร้างแผ่นแบบและการเปรียบเทียบความแม่นยำในการจำแนกประเภทโดยวิธีการจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่หนึ่งโดยใช้การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง แผ่นแบบเฉลี่ยที่สร้างขึ้นจะถูกนำไปใช้แทนที่ข้อมูลอนุกรมเวลาที่เป็นข้อมูลเรียนรู้ทั้งหมด ซึ่งจะมีการอธิบายอย่างละเอียดและผลลัพธ์ของการวัดผลในบทถัดไป

บทที่ 4 การทดลองและวิเคราะห์ผล

ในบทนี้จะกล่าวถึงการประเมินคุณภาพและประสิทธิภาพของสิ่งที่งานวิจัยได้ทำการนำเสนอขึ้น วิธีการอื่นที่จะนำมาทำการเปรียบเทียบกับวิธีการที่ได้นำเสนอขึ้นประกอบด้วยสองวิธีการได้แก่ วิธีการ Nearest Centroid Classifier (NCC) และ Shape-based Template Matching Framework (STMF) โดยวิธีการ NCC นั้นเป็นวิธีการค้นหาแผ่นแบบที่มีความแม่นยำสูงจากการสร้างแผ่นแบบเฉลี่ยจากข้อมูลทั้งหมดพร้อมกันและสามารถสร้างแผ่นแบบที่มากกว่าหนึ่งแผ่นแบบต่อหนึ่งคลาสได้ ในส่วนของวิธีการ STMF นั้นเป็นวิธีการที่ใช้วิธีการจับคู่แผ่นแบบแล้วทำการสร้างแผ่นแบบเฉลี่ยที่ให้ผลลัพธ์ที่ดีที่สุดเพียงแผ่นแบบเดียว ในส่วนการทดสอบนั้นจะประกอบไปด้วยทั้งหมดสองส่วนหลัก ๆ คือ การทดสอบความแม่นยำของแผ่นแบบเมื่อนำไปทำการจำแนกประเภทข้อมูล และความเร็วในการสร้างแผ่นแบบ

4.1 ชุดข้อมูลที่ใช้ในงานวิจัย

ข้อมูลอนุกรมเวลาทั้งหมดที่ได้นำมาทำการทดลองในงานวิจัยนี้ประกอบด้วยข้อมูลอนุกรมเวลาทั้งหมด 20 ชุด จาก UCR Time Series Classification Archive [27] ซึ่งเป็นข้อมูลอนุกรมเวลาที่มีการเปิดเผยเพื่อใช้สำหรับงานวิจัยที่เกี่ยวข้องกับข้อมูลอนุกรมเวลา ข้อมูลแต่ละชุดมีลักษณะของข้อมูลที่แตกต่างกัน ทั้งในด้านของความยาวของอนุกรมเวลา จำนวนของคลาส จำนวนของข้อมูลเรียนรู้ และจำนวนของข้อมูลทดสอบ ตารางที่ 4.1 แสดงลักษณะของข้อมูลทั้ง 20 ชุดข้อมูลที่ได้นำมาทดลองในงานวิจัยนี้ประกอบไปด้วยชื่อของชุดข้อมูล จำนวนคลาส ความยาวของข้อมูลอนุกรมเวลา จำนวนข้อมูลเรียนรู้ และจำนวนข้อมูลทดสอบ

ตารางที่ 4.1 ลักษณะของข้อมูลที่ใช้ในการทดลอง

ชุดข้อมูล	จำนวนคลาส	ความยาวของข้อมูล	จำนวนข้อมูลเรียนรู้	จำนวนข้อมูลทดสอบ
50words	50	270	450	455
Adiac	37	175	390	391
Beef	5	470	30	30
CBF	3	128	30	900
Coffee	2	286	28	28
ECG200	2	96	100	100
FaceAll	14	131	560	1690

FaceFour	4	350	24	88
FISH	7	463	175	175
Gun_Point	2	150	50	150
Lightning2	2	637	60	61
Lightning7	7	319	70	73
OliveOil	4	570	30	30
OSUleaf	6	427	200	242
SwedishLeaf	15	128	500	625
Synthetic_control	6	60	300	300
Trace	4	275	100	100
Two_Patterns	4	128	1,000	4,000
Wafer	2	152	1,000	6,174
Yoga	2	426	300	3,000

4.2 การทดสอบความแม่นยำของแผ่นแบบเมื่อนำไปทำการจำแนกประเภทข้อมูล

ในส่วนของการทดสอบด้านความแม่นยำของแผ่นแบบที่สร้างขึ้นนั้นจะทำการเปรียบเทียบแผ่นแบบที่สร้างขึ้นจากวิธีการที่นำเสนอกับแผ่นแบบที่สร้างขึ้นด้วยวิธีการ STMF วิธีการ NCC และความแม่นยำเมื่อทำการจำแนกประเภทโดยการใช้ข้อมูลเรียนรู้ครบทุกข้อมูล ข้อมูลเรียนรู้ที่ใช้ในการจำแนกประเภทนั้นจะถูกแทนที่ด้วยแผ่นแบบที่สร้างขึ้นของแต่ละชุดข้อมูล และข้อมูลทดสอบจะใช้ข้อมูลคงเดิมที่มาจาก UCR Time Series Classification Archive เพื่อวัดความแม่นยำของแผ่นแบบ

วิธีการจำแนกประเภทที่ใช้ในการวัดความแม่นยำของข้อมูลนั้นประกอบด้วยการจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่ 1 ดังที่ได้อธิบายไว้ในบทที่ 2 หัวข้อ 2.2 ซึ่งเป็นการวัดผลด้วยการใช้การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงควบคู่ไปกับการจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่หนึ่งแบบปกติ อีกวิธีหนึ่งที่ใช้วัดผลความแม่นยำในการทดลองนี้คือ การใช้การกำหนดเงื่อนไขบังคับโดยรวมควบคู่กับการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงในการวัดระยะทางแล้วทำการจำแนกประเภทโดยการใช้การจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่ 1 เช่นเดิม รวมถึงการใช้ฟังก์ชันขอบเขตล่างของไดนามิกไทม์วอร์ปปิงของ Eamonn Keogh หรือเรียกว่า LB_Keogh ซึ่งเป็นฟังก์ชันขอบเขตล่างที่มีความเร็วในการทำงานสูงและสามารถใช้การกำหนดเงื่อนไขโดยรวมเข้าไปในการวัดระยะทางด้วยได้ ดังอธิบายไว้ในบทที่ 2.6

ตารางที่ 4.2 แสดงความแม่นยำของการจำแนกประเภทโดยการใช้การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงควบคู่ไปกับการจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่หนึ่งแบบปกติ และตารางที่ 4.3 แสดงจำนวนแผ่นแบบที่สร้างขึ้นจากวิธีการต่างๆ

ตารางที่ 4.2 ความแม่นยำของการจำแนกประเภทด้วยการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงควบคู่ไปกับการจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่หนึ่งแบบปกติ ของแผ่นแบบที่สร้างจากวิธีการที่นำเสนอ NCC STMF และการใช้ข้อมูลเรียนรู้ครบทุกข้อมูล

ชุดข้อมูล	**ความแม่นยำ (%)**			
	วิธีการที่นำเสนอ	***NCC***	***STMF***	***ข้อมูลเรียนรู้***
50words	70.77	69.01	60	69
Adiac	57.03	59.82	49	61.4
Beef	60	63.33	47	63.3
CBF	99.89	100	96	99.7
Coffee	92.86	98.89	96	100
ECG200	77	83	70	77
FaceAll	80.77	79.59	83	80.8
FaceFour	84.09	84.09	83	83
FISH	74.29	82.86	58	82.3
Gun_Point	85.33	90.67	64	90.7
Lightning2	86.89	88.25	56	86.9
Lightning7	78.08	78.08	66	72.6
OliveOil	86.67	86.67	77	83.3
OSUleaf	55.79	57.85	41	59.1
SwedishLeaf	78.88	78.88	69	79.2
Synthetic_control	99	98.67	97	99.3
Trace	100	100	98	100
Two_Patterns	100	100	97	100
Wafer	95.81	97.27	64	98
Yoga	73.67	83.07	48	83.6

ตารางที่ 4.3 จำนวนแผ่นแบบที่สร้างขึ้นด้วยวิธีการที่นำเสนอ NCC STMF และการใช้ข้อมูลเรียนรู้ทั้งหมด

ชุดข้อมูล	**จำนวนแผ่นแบบทั้งหมด**			
	วิธีการที่นำเสนอ	***NCC***	***STMF***	***ข้อมูลเรียนรู้***
50words	256	282	50	450
Adiac	218	246	37	390
Beef	18	15	5	30
CBF	21	9	3	30
Coffee	19	7	2	28
ECG200	43	18	2	100
FaceAll	267	28	14	560
FaceFour	16	8	4	24
FISH	66	158	7	175
Gun_Point	31	22	2	50
Lightning2	39	12	2	60
Lightning7	32	14	7	70
OliveOil	13	12	4	30
OSUleaf	117	60	6	200
SwedishLeaf	205	476	15	500
Synthetic_control	132	30	6	300
Trace	55	4	4	100
Two_Patterns	175	12	4	1,000
Wafer	124	130	2	1,000
Yoga	112	281	2	300

จากตารางที่ 4.2 จะเห็นได้ว่าวิธีการที่นำเสนอนั้นมีความแม่นยำของการจำแนกประเภทข้อมูลอนุกรมเวลาที่สามารถเทียบเคียงกับวิธีการที่ดีที่สุดในขณะนี้ของการสร้างแผ่นแบบเฉลี่ยคือวิธีการ NCC ได้และสามารถได้ผลลัพธ์ที่ดีกว่าในบางชุดข้อมูล ในขณะที่มีความแม่นยำสูงกว่าวิธีการ

STMF เกือบทุกชุดข้อมูล อีกทั้งเมื่อเทียบกับผลลัพธ์ที่ใช้ข้อมูลเรียนรู้ครบทุกตัวแล้ว วิธีการที่นำเสนอมีความแม่นยำใกล้เคียงการใช้ข้อมูลเรียนรู้ครบทุกตัวในการจำแนกประเภท และมีความแม่นยำสูงกว่าในหลายชุดข้อมูล

ชุดข้อมูลที่วิธีการที่นำเสนอได้ผลลัพธ์น้อยกว่าวิธีการ NCC มากประกอบไปด้วยชุดข้อมูล Coffee ชุดข้อมูล FISH และชุดข้อมูล yoga โดยลักษณะของข้อมูล Coffee นั้นข้อมูลทั้งสองคลาสจะมีลักษณะใกล้เคียงกันและแตกต่างกันเล็กน้อยในบางจุด แต่มีความผันแปรในส่วนของความมากน้อยของค่าที่ใช้วัดหรือความสูงต่ำของกราฟมากในคลาสเดียวกัน อีกทั้งข้อมูลเรียนรู้ของ Coffee นั้นมีจำนวนน้อย ทำให้แผ่นแบบที่สร้างขึ้นมีความใกล้เคียงกันมากระหว่างสองคลาสและเกิดความผิดพลาดในการจำแนกประเภทขึ้นในขณะที่วิธีการ NCC มีการค้นหาค่าของจำนวนกลุ่มที่เหมาะสมกับวิธีการ STMF ทำการสร้างแผ่นแบบจากข้อมูลทั้งหมด ทำให้ความแตกต่างระหว่างคลาสของข้อมูลชัดเจนกว่าวิธีการที่นำเสนอ

ในขณะที่ข้อมูลของ FISH นั้นมีจำนวนคลาส 7 คลาสและแต่ละคลาสข้อมูลมีความใกล้เคียงกันสูง และข้อมูล yoga นั้นมีความแปรผันของข้อมูลในคลาสเดียวกันสูงมาก ทำให้ผลลัพธ์ความแม่นยำในวิธีการที่นำเสนอนั้นต่ำกว่าการใช้ข้อมูลเรียนรู้ครบทุกตัวมาก แต่ในขณะเดียวกันวิธีการ NCC สามารถได้ผลลัพธ์ที่ดีจากการใช้จำนวนแผ่นแบบจำนวนมากจนมีความใกล้เคียงกับการใช้ข้อมูลเรียนรู้ครบทุกตัว และสามารถลดเวลาที่ใช้ในการจำแนกประเภทได้น้อยกว่า และสุดท้ายวิธีการ STMF นั้นได้ผลลัพธ์ที่ต่ำกว่าการใช้ข้อมูลเรียนรู้ทุกข้อมูลมากเช่นกัน โดยลักษณะของตัวอย่างข้อมูลจะทั้งหมดจะแสดงในภาคผนวก ก

ในส่วนถัดไปจะกล่าวถึงการใช้ฟังก์ชันขอบเขตล่างของไดนามิกไทม์วอร์ปปิงของ Eamonn Keogh หรือ LB_Keogh ร่วมกับการกำหนดเงื่อนไขโดยรวมเพื่อทำการจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่ 1 ในส่วนของการกำหนดเงื่อนไขโดยรวมนั้นจำเป็นที่จะต้องทำการค้นหาขอบเขตของการกำหนดเงื่อนไขโดยรวมที่เหมาะสมกับแผ่นแบบที่สร้างขึ้นก่อนที่จะนำไปทำการใช้ฟังก์ชันขอบเขตล่างของไดนามิกไทม์วอร์ปปิงในการหาระยะทางและจำแนกประเภท

วิธีการที่จะค้นหาขอบเขตของเงื่อนไขโดยรวมที่เหมาะสมในที่นั้นใช้วิธีการจำแนกประเภทแบบเพื่อนบ้านลำดับที่ 1 โดยใช้การกำหนดเงื่อนไขโดยรวมระหว่างแผ่นแบบกับแผ่นแบบด้วยกันเองหรือเรียกว่า Leave-one-out Cross Validation [44] เพื่อหาค่าของขอบเขตการกำหนดเงื่อนไขโดยรวมที่เหมาะสมที่สุด ในข้อมูลแต่ละตัวของแผ่นแบบนั้นจะทำการหาระยะทางแบบไดนามิกไทม์วอร์ปปิงควบคู่กับการกำหนดเงื่อนไขโดยรวมกับแผ่นแบบทั้งหมดที่เหลืออยู่ โดยใช้ฟังก์ชันขอบเขตล่างของไดนามิกไทม์วอร์ปปิงควบคู่ไปด้วย หลังจากหาระยะทางของข้อมูลทุกตัวได้จะทำการจำแนกประเภทแบบเพื่อนบ้านลำดับที่ 1 เพื่อค้นหาค่าของขอบเขตการกำหนดเงื่อนไขโดยรวมที่ทำให้ความแม่นยำในการจำแนกประเภทของแผ่นแบบด้วยกันเองสูงที่สุด

ในการทดลองนี้ได้กำหนดค่าของขอบเขตการกำหนดเงื่อนไขโดยรวมที่จะทำการค้นหาเป็นร้อยละของความยาวของแผ่นแบบโดยเริ่มจากร้อยละ 0 จนกระทั่งถึงร้อยละ 100 ตัวอย่างเช่นความยาวของแผ่นแบบเป็น 150 ค่าขอบเขตการกำหนดเงื่อนไขโดยรวมที่ร้อยละ 10 จะเท่ากับ 150 x (10/100) = 15 เป็นต้น อนึ่งหากจำนวนของแผ่นแบบที่เหมาะสมที่สุดเป็นจำนวน 1 แผ่นแบบต่อคลาสดังเช่นวิธีการ STMF หรือบางกรณีที่อาจเกิดขึ้นได้ในวิธีอื่นจะไม่สามารถใช้วิธีที่กล่าวไปในการคำนวณได้ เนื่องจากความแม่นยำในการจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่ 1 ระหว่างแผ่นแบบด้วยกันเองจะเป็น 0 เสมอ จากการที่มีไม่มีข้อมูลในคลาสเดียวกันเหลืออยู่ให้ทำการกำหนดคลาสที่ถูกต้องได้แล้ว ในกรณีดังกล่าวนั้นจะใช้ค่าขอบเขตการกำหนดเงื่อนไขเท่ากับ 10% และในที่นี้จะไม่ทำการเปรียบเทียบความแม่นยำร่วมกับวิธีการ STMF ที่มีเพียงแผ่นแบบเดียวเสมอ

ตารางที่ 4.4 แสดงความแม่นยำของการจำแนกประเภทโดยการใช้การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงควบคู่ไปกับการจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่หนึ่งและใช้ฟังก์ชันขอบเขตล่างของไดนามิกไทม์วอร์ปปิงแบบ LB_Keogh

ตารางที่ 4.4 ความแม่นยำของการจำแนกประเภทด้วยการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงโดยใช้การกำหนดเงื่อนไขโดยรวม ควบคู่ไปกับการจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่หนึ่ง ของแผ่นแบบที่สร้างจากวิธีการที่นำเสนอ NCC และการใช้ข้อมูลเรียนรู้ครบทั้งหมด

ชุดข้อมูล	**ความแม่นยำ (%)**		
	วิธีการที่นำเสนอ	***NCC***	***ข้อมูลเรียนรู้***
50words	**76.48**	76.03	75.8
Adiac	**57.54**	57.30	60.9
Beef	**66.67**	60	66.7
CBF	99.78	100	99.6
Coffee	96.43	98.89	100
ECG200	**90.00**	84	88
FaceAll	**82.49**	79.29	80.8
FaceFour	**88.64**	77.27	88.6
FISH	72.00	82.29	84.6
Gun_Point	93.33	93.33	91.3
Lightning2	85.25	85.25	86.9
Lightning7	**80.82**	61.64	71.2

OliveOil	86.67	90	86.7
OSUleaf	53.72	56.2	61.2
SwedishLeaf	78.40	84.32	84.6
synthetic_control	98.00	98.67	98.3
Trace	**97.00**	57	99
Two_Patterns	95.10	99.38	99.8
Wafer	97.86	98.36	99.5
Yoga	64.23	78.57	84.5

จากตารางที่ 4.4 จะเห็นได้ว่าเมื่อทำการวัดผลด้วยการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงควบคู่ไปกับการใช้ฟังก์ชันขอบเขตล่างของไดนามิกไทม์วอร์ปปิงแบบ LB_Keogh ซึ่งมีการกำหนดเงื่อนไขบังคับโดยรวมในการวัดผลรวมอยู่ด้วย ทำให้ความแม่นยำของวิธีการที่นำเสนอส่วนใหญ่เพิ่มสูงขึ้นในขณะที่ความแม่นยำของข้อมูลบางส่วนที่ตกลงนั้นตกลงเพียงเล็กน้อยเท่านั้น ในขณะที่วิธีการ NCC นั้นเมื่อใช้การกำหนดเงื่อนไขโดยรวมในการวัดผลด้วย ความแม่นยำของข้อมูลหลายชุดข้อมูลตกลง รวมถึงบางข้อมูลที่มีจำนวนแผ่นแบบน้อยจนเกินไปเช่นข้อมูล Trace จะตกลงอย่างมาก และส่งผลให้วิธีการที่นำเสนอมีความแม่นยำสูงกว่าหลายชุดข้อมูล ทำให้สรุปได้ว่าวิธีการที่นำเสนอมีความเหมาะสมที่จะใช้วัดผลด้วยวิธีการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงเมื่อใช้การกำหนดเงื่อนไขบังคับโดยรวมควบคู่ไปด้วยมากกว่า

อย่างไรก็ตามแม้วิธีการ NCC จะให้ความแม่นยำที่สูงกว่าวิธีการที่นำเสนอ แต่เวลาที่ใช้ในการหาตัวแปรที่เหมาะสมเพื่อให้สามารถสร้างแผ่นแบบที่เหมาะสมในการจำแนกประเภทของแต่ละชุดข้อมูลนั้นสูงมาก ดังนำเสนอในหัวข้อถัดไป

4.3 ความเร็วในการสร้างแผ่นแบบ

ในส่วนถัดไปจะนำเสนอเปรียบเทียบความเร็วของการสร้างแผ่นแบบระหว่างวิธีการที่นำเสนอกับวิธีการ NCC สาเหตุที่ไม่ได้ทำการเปรียบเทียบกับวิธีการ STMF นั้นเนื่องด้วยวิธีการที่นำเสนอและวิธีการ NCC ต่างใช้การสร้างแผ่นแบบเฉลี่ยด้วยวิธีการ DBA ซึ่งเป็นการคำนวณคนละวิธีการกับการจับคู่แผ่นแบบด้วยวิธีการ STMF อีกทั้งวิธีการ STMF มีผลลัพธ์ของความแม่นยำส่วนใหญ่น้อยกว่าวิธีการ NCC และวิธีการที่นำเสนอ จึงได้ทำการเปรียบเทียบเวลาที่ใช้ในการสร้างแผ่นแบบเฉพาะวิธีการที่นำเสนอและวิธีการ STMF เท่านั้น โดยเวลาที่ใช้ในการสร้างแผ่นแบบแสดงดังตารางที่ 4.5

ตารางที่ 4.5 เวลาที่ใช้ในการสร้างแผ่นแบบเปรียบเทียบระหว่างวิธีการที่นำเสนอและวิธีการ NCC

ชุดข้อมูล	**เวลาที่ใช้การสร้างแผ่นแบบ (วินาที)**	
	วิธีการที่นำเสนอ	***NCC***
50words	27.39	3153.30
Adiac	11.49	1340.02
Beef	4.8	178.55
CBF	0.98	47.82
Coffee	1.64	335.37
ECG200	1.2	47.06
FaceAll	13.48	23.99
FaceFour	2.56	17.60
FISH	38.96	16550.78
Gun_Point	0.896	726.40
Lightning2	9.42	372.07
Lightning7	7.13	44.98
OliveOil	8.62	86.91
OSUleaf	28.27	5250.30
SwedishLeaf	9.424	6811.18
synthetic_control	1.87	2458.07
Trace	5.6	17.11
Two_Patterns	27.99	1027.00
Wafer	44.92	898599.86
Yoga	53.99	2078992.62

จากตารางที่ 4.5 จะเห็นได้ว่าเวลาในการสร้างแผ่นแบบของวิธีการที่นำเสนอนั้นเร็วกว่าวิธีการ NCC มาก เมื่อเทียบกับความแม่นยำที่ลดลงเพียงเล็กน้อยของวิธีการที่นำเสนอแล้วนั้น เป็นการแลกเปลี่ยนที่คุ้มค่าเมื่อนำไปใช้งาน อนึ่งในข้อมูลบางกรณีเช่นข้อมูล FaceAll หรือ FaceFour นั้นเนื่องจากวิธีการ NCC ได้กำหนดไว้ว่าหากวิธีการสามารถชนะความแม่นยำของการใช้ข้อมูลเรียนรู้ครบทุกตัวได้แล้วจะหยุดการทำงาน ส่งผลให้ข้อมูลบางข้อมูลที่มีความแม่นยำสูงเมื่อนำไปสร้างแผ่นแบบตั้งแต่การใช้แผ่นแบบจำนวนน้อย ๆ จะหยุดการทำงานเร็ว และใช้เวลาทำงานไม่นาน แต่อย่างไร

ก็ตามวิธีการที่นำเสนอก็มีการทำงานที่เร็วกว่าอยู่เช่นกัน เมื่อมองถึงข้อมูลที่มีจำนวนข้อมูลมากและมีการทำงานของวิธีการ NCC ครบทุกรูปแบบที่เป็นไปได้เนื่องจากความแม่นยำของการจำแนกประเภทด้วยแผ่นแบบในแต่ละรูปแบบไม่สามารถชนะการใช้ข้อมูลเรียนรู้ครบทุกข้อมูลได้ พบว่าเวลาที่ใช้สร้างแผ่นแบบของวิธีการ NCC จะสูงมาก ตัวอย่างเช่นชุดข้อมูล Yoga หรือ Wafer ซึ่งวิธีการที่นำเสนอสามารถสร้างแผ่นแบบได้รวดเร็วกว่ามาก

บทที่ 5 สรุปผลงานวิจัยและข้อเสนอแนะ

5.1 สรุปผลงานวิจัย

การจำแนกประเภทข้อมูลอนุกรมเวลาแบบเพื่อนบ้านใกล้สุดลำดับที่ 1 นั้นนับว่าเป็นการจำแนกประเภทข้อมูลอนุกรมเวลาที่มีประสิทธิภาพสูง โดยเฉพาะอย่างยิ่งเมื่อนำมาใช้ร่วมกับการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงซึ่งเป็นการวัดระยะทางที่มีการปรับแนวแบบไม่เป็นจุดต่อจุด ซึ่งช่วยให้ระยะทางที่ได้เหมาะสมกับการจำแนกประเภทข้อมูลมากขึ้น แต่อย่างไรก็ตามเวลาที่ใช้คำนวณระยะทางแบบไดนามิกไทม์วอร์ปปิงนั้นสูง โดยมีขีดจำกัดเชิงสัญกรณ์ในการคำนวณระยะทาง $O(n^2)$ ซึ่งเมื่อนำไปใช้ควบคู่กับการจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่หนึ่งที่ต้องทำการคำนวณระยะทางเทียบกับข้อมูลเรียนรู้ทุกข้อมูลแล้วนั้น เวลาที่ใช้ในการทำงานจะสูงมาก

วิธีการต่าง ๆ จึงได้พัฒนาขึ้นเพื่อทำการลดเวลาที่ใช้ในการจำแนกประเภท วิธีการหนึ่งที่ได้ผลลัพธ์ที่ดีในการลดเวลาที่ใช้ในการจำแนกประเภทคือ การสร้างแผ่นแบบเฉลี่ย (Template Averaging) ซึ่งสามารถลดเวลาในการจำแนกประเภทลงได้จากการที่ลดจำนวนข้อมูลเรียนรู้ที่ต้องการคำนวณลง ซึ่งวิธีการที่ดีที่สุดในปัจจุบันคือวิธีการ Nearest Centroid Classifier (NCC) ซึ่งสามารถสร้างแผ่นแบบที่มีจำนวนน้อยกว่าข้อมูลเรียนรู้และลดเวลาในการจำแนกประเภทลงได้ ในขณะที่สามารถมีความแม่นยำที่สูงเทียบเท่ากับการจำแนกประเภทโดยการใช้ข้อมูลเรียนรู้ครบทุกข้อมูล แต่อย่างไรก็ตามเวลาที่ใช้ในการสร้างแผ่นแบบเฉลี่ยของวิธีการ NCC นั้นสูงมาก เนื่องจากการปรับแต่งตัวแปรที่เหมาะสมเพื่อให้ได้ความแม่นยำที่สูงนั้นยาก ทำให้ใช้เวลาในการคำนวณทั้งหมดนาน

งานวิจัยนี้จึงได้นำเสนอวิธีการสร้างแผ่นแบบเฉลี่ยของข้อมูลอนุกรมเวลา โดยปรับปรุงจากวิธีการ NCC ซึ่งใช้เวลาในการสร้างแผ่นแบบสูง มาใช้วิธีการใหม่ที่สามารถลดเวลาในการสร้างแผ่นแบบให้น้อยลง และคงไว้ซึ่งความแม่นยำของการจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่ 1 เมื่อใช้ข้อมูลแผ่นแบบที่สร้างขึ้นแทนการใช้ข้อมูลเรียนรู้ครบทุกข้อมูลได้ โดยมีแนวคิดของงานวิจัยที่จะทำการปรับปรุงการแบ่งกลุ่มข้อมูลในวิธีการ NCC และวิธีการหาข้อมูลตัวหลัก (Pivot) ในการสร้างแผ่นแบบเฉลี่ยที่ใช้เวลาในการคำนวณนานหรือปรับแต่งค่าตัวแปรยาก ให้สามารถปรับแต่งค่าตัวแปรได้ง่ายและใช้เวลาในการคำนวณน้อยลง ผลลัพธ์ของเวลาที่ใช้ในการสร้างแผ่นแบบได้ทำการนำเสนอไว้ในบทที่ 4 ซึ่งจะเห็นได้ว่าวิธีการที่นำเสนอนั้นสามารถสร้างแผ่นแบบเฉลี่ยขึ้นได้ใช้เวลาน้อยลงกว่าวิธีการ NCC มากในทุกชุดข้อมูลที่ทำการทดสอบ

ในส่วนของความแม่นยำในการจำแนกประเภทนั้น จากข้อมูลในบทที่ 4 จะเห็นได้ว่าวิธีการที่นำเสนอสามารถคงไว้ซึ่งความแม่นยำเมื่อเปรียบเทียบกับวิธีการ NCC และเปรียบเทียบกับการใช้ข้อมูลเรียนรู้ครบทุกข้อมูลได้ แม้ว่าความแม่นยำส่วนใหญ่จะไม่สามารถเทียบเท่าได้ แต่ความแม่นยำ

นั้นตกลงไม่มาก เมื่อแลกเปลี่ยนกับเวลาในการสร้างแผ่นแบบที่รวดเร็วขึ้นอย่างมากแล้วสามารถนับเป็นการแลกเปลี่ยนที่คุ้มค่าได้ แม้ว่าบางชุดข้อมูลจะมีความแม่นยำที่แตกต่างกันมากกับวิธีการ NCC แต่ชุดข้อมูลเหล่านั้นล้วนแต่มีจำนวนข้อมูลที่น้อยมากหรือมีการใช้แผ่นแบบใกล้เคียงกับจำนวนของข้อมูลเรียนรู้ซึ่งส่งผลให้แผ่นแบบที่ได้ลดเวลาในการจำแนกประเภทได้น้อย อย่างไรก็ตามเมื่อทำการวัดผลโดยการใช้การกำหนดเงื่อนไขโดยรวม (Global Constraint) แล้วผลลัพธ์ของวิธีการที่นำเสนอสามารถเทียบกับกับวิธีการ NCC ได้ดียิ่งขึ้นอีก

เมื่อมองโดยภาพรวมวิธีการที่นำเสนอสามารถลดเวลาที่ใช้ในการสร้างแผ่นแบบลงได้มาก ส่งผลให้เวลาในการคำนวณทั้งหมดเพื่อทำการจำแนกประเภทข้อมูลอนุกรมเวลาลดลงอย่างมาก อีกทั้งสามารถคงความแม่นยำในการจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่ 1 ไว้ได้ทั้งในการใช้การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงปกติและการใช้การกำหนดเงื่อนไขโดยรวมร่วมด้วย

5.2 ข้อเสนอแนะ

ข้อเสนอแนะต่อไปนี้เป็นแนวทางในการวิจัยที่จะนำเสนอเพื่อพัฒนาวิธีการสร้างแผ่นแบบเฉลี่ยของข้อมูลอนุกรมเวลาให้ดียิ่งขึ้น ประเด็นหลักที่ต้องมีการพัฒนาต่อจากงานวิจัยนี้คือความแม่นยำของการจำแนกประเภทด้วยการจำแนกประเภทแบบเพื่อนบ้านใกล้สุดลำดับที่ 1 ถึงแม้ว่าวิธีการที่นำเสนอนั้นจะสามารถจำแนกประเภทข้อมูลได้แม่นยำเทียบเคียงกับการใช้ข้อมูลเรียนรู้ครบทุกข้อมูล แต่วิธีการ NCC นั้นมีข้อมูลบางชุดข้อมูลที่มีความแม่นยำที่สูงกว่าในขณะที่แผ่นแบบมีจำนวนข้อมูลน้อยกว่าจำนวนข้อมูลเรียนรู้มาก ซึ่งแสดงให้เห็นว่าจำนวนแผ่นแบบเฉลี่ยที่เหมาะสมมีประสิทธิภาพสูงกว่าข้อมูลเรียนรู้ครบทุกข้อมูล โดยแนวทางที่ยังสามารถปรับปรุงต่อไปได้ในการสร้างแผ่นแบบเฉลี่ยของวิธีการที่นำเสนอนั้นประกอบด้วยการจัดเรียงข้อมูลเพื่อทำการแบ่งข้อมูลเป็นกลุ่มย่อย การเลือกข้อมูลตัวหลักในวิธีการสร้างแผ่นแบบเฉลี่ยด้วยวิธีการ DBA และการแยกข้อมูลที่เป็นแตกต่างจากข้อมูลอื่น ๆ มากออกจากกลุ่มข้อมูล (Outlier)

การจัดเรียงข้อมูลเพื่อทำการแบ่งข้อมูลเป็นกลุ่มย่อยนั้น ดังที่ได้เสนอไปในบทที่ 3 มีการจัดเรียงข้อมูลโดยทำการเลือกข้อมูลตัวหลักมาทำการเปรียบเทียบ เนื่องด้วยยังไม่มีวิธีการจัดเรียงข้อมูลอนุกรมเวลาที่ยังมีประสิทธิผลพอที่จะได้บอกได้ว่าข้อมูลอนุกรมเวลาข้อมูลใดสมควรมาก่อนหรือมาหลัง การเลือกข้อมูลตัวหลักเพื่อเป็นหลักในการเปรียบเทียบข้อมูลตัวอื่น ๆ นั้นแม้จะสามารถเรียงข้อมูลและแบ่งแยกข้อมูลที่แตกต่างกันออกได้ในระดับหนึ่ง แต่หากข้อมูลมีกลุ่มข้อมูลจำนวนมาก ตัวอย่างเช่นข้อมูล yoga การแบ่งกลุ่มข้อมูลนั้นแบ่งกลุ่มได้คุณภาพน้อยลงเนื่องจากมีข้อมูลบางกลุ่มย่อยที่เมื่อเทียบกับข้อมูลตัวหลักแล้วนั้น มีความใกล้เคียงกัน ถ้าสามารถใช้วิธีการที่สามารถจัดเรียงข้อมูลได้อย่างมีคุณภาพมากขึ้นก็จะสามารถแบ่งกลุ่มข้อมูลเป็นกลุ่มย่อยได้ดีและสร้างแผ่นแบบเฉลี่ยได้อย่างมีคุณภาพมากขึ้นทั้งความรวดเร็วและความแม่นยำในการจำแนกประเภท

อีกหนึ่งส่วนที่สามารถปรับปรุงเพิ่มเติมได้คือการเลือกข้อมูลตัวหลักในวิธีการสร้างแผ่นแบบเฉลี่ยด้วยวิธีการ DBA ดังที่ได้นำเสนอไปในบทที่ 3 ในงานวิจัยนี้ได้เลือกวิธีการที่ปรับปรุงจากวิธีการ Simple Rank (SR) วิธีการ NCC ใช้การค้นหาข้อมูลเมดอยด์ และวิธีการ DBA ดั้งเดิมใช้การสุ่มข้อมูล ซึ่งทั้งสามวิธีการนั้นยังไม่สามารถเลือกข้อมูลที่เหมาะสมที่สุดในการนำไปสร้างแผ่นแบบเฉลี่ยด้วยวิธีการ DBA ได้ ดังนั้นแล้วหากมีวิธีที่สามารถเลือกข้อมูลที่เหมาะสมมาใช้ควบคู่กับวิธีการ DBA หรือมีการปรับปรุงวิธีการ DBA ให้สามารถสร้างแผ่นแบบเฉลี่ยโดยไม่ต้องการข้อมูลตัวหลัก ก็จะสามารถพัฒนาคุณภาพของแผ่นแบบได้เช่นกัน

ข้อมูลทุกชุดข้อมูลนั้นจะมีข้อมูลบางส่วนที่อาจเกิดจากการเก็บข้อมูลที่ผิดพลาดหรือมีสิ่งรบกวนในการเก็บข้อมูล ส่งผลให้ข้อมูลนั้น ๆ หรือข้อมูลจำนวนหนึ่งมีรูปร่างที่ผิดแปลกไปจากข้อมูลตัวอื่น ๆ มาก หากข้อมูลนั้นเป็นข้อมูลในกลุ่มข้อมูลเรียนรู้ที่ใช้เปรียบเทียบก็จะมีมีการกำหนดคลาสของข้อมูลเป็นคลาสใดคลาสหนึ่ง ซึ่งข้อมูลที่แตกต่างจากข้อมูลอื่น ๆ มากนี้จะส่งผลอย่างมากในการแบ่งกลุ่มข้อมูลเป็นกลุ่มย่อยทั้งในวิธีการ NCC และวิธีการที่นำเสนอ แม้ว่าวิธีการที่นำเสนอจะสามารถตัดข้อมูลที่มีความแตกต่างอย่างเด่นชัดออกไปได้ในระดับหนึ่งโดยการสร้างกลุ่มย่อยใหม่ที่มีข้อมูลนั้น ๆ ข้อมูลเดียวได้ แต่หากข้อมูลที่อยู่ในกลุ่มย่อยเดียวกันมีความกระจายตัวมาก ข้อมูลที่แตกต่างจากข้อมูลอื่น ๆ มากนี้ก็อาจจะถูกนับรวมเป็นกลุ่มเดียวกันและถูกนำไปสร้างแผ่นแบบเฉลี่ยที่ผิดพลาดได้เช่นกัน ทั้งนี้หากมีวิธีการที่สามารถแบ่งแยกข้อมูลที่แตกต่างจากข้อมูลอื่น ๆ มากออกไปก่อนที่จะทำการแบ่งกลุ่มย่อยก็สามารถทำให้แผ่นแบบที่สร้างขึ้นมีคุณภาพสูงขึ้นเช่นกัน

รายการอ้างอิง

[1] F. Iglesias and W. Kastner, "Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns," *Energies,* vol. 6, pp. 579, 2013.

[2] C. A. Ratanamahatana and E. Keogh, "Everything you know about dynamic time warping is wrong," in *3rd Workshop on Mining Temporal and Sequential Data, in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)*, 2004.

[3] G. P. Shorten and M. J. Burke, "Use of dynamic time warping for accurate ECG signal timing characterization," *J Med Eng Technol,* vol. 38, pp. 188-201, May, 2014.

[4] C. Caballero and M. C. Aranda, "Plant species identification using leaf image retrieval," presented at the *Proceedings of the ACM International Conference on Image and Video Retrieval,* Xi'an, China, 2010.

[5] H. Sivaraks and C. A. Ratanamahatana, "Robust and Accurate Anomaly Detection in ECG Artifacts Using Time Series Motif Discovery," *Computational and Mathematical Methods in Medicine,* vol. 2015, pp. 20, 2015.

[6] K. Chen, T.-Y. Wu, and H.-J. Zhang, "On the use of nearest feature line for speaker identification," *Pattern Recogn. Lett.,* vol. 23, pp. 1735-1746, 2002.

[7] J. D. Hamilton, "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica,* vol. 57, pp. 357-384, 1989.

[8] S. R. a. G. Sikka, "Recent Techniques of Clustering of Time Series Data: A Survey," *International Journal of Computer Applications,* vol. 52, pp. 1-9, August, 2012.

[9] E. J. Keogh and M. J. Pazzani, "Scaling up dynamic time warping for datamining applications," presented at the Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Boston, Massachusetts, USA, 2000.

[10] K. Fukunaga and P. M. Narendra, "A Branch and Bound Algorithm for Computing k-Nearest Neighbors," *IEEE Transactions on Computers,* vol. C-24, pp. 750-753, 1975.

[11] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy K-nearest neighbor algorithm," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. SMC-15, pp. 580-585, 1985.

[12] K. Ueno, X. Xi, E. Keogh, and D.-J. Lee, "Anytime Classification Using the Nearest Neighbor Algorithm with Applications to Stream Mining," presented at *the Proceedings of the Sixth International Conference on Data Mining*, 2006.

[13] R. A. Lotufo and F. A. Zampirolli, "Fast multidimensional parallel Euclidean distance transform based on mathematical morphology," in *Computer Graphics and Image Processing, 2001 Proceedings of XIV Brazilian Symposium on*, pp. 100-105, 2001.

[14] A. Meijster, J. B. T. M. Roerdink, and W. H. Hesselink, "A General Algorithm for Computing Distance Transforms in Linear Time," in *Mathematical Morphology and its Applications to Image and Signal Processing*. vol. 18, J. Goutsias, L. Vincent, and D. Bloomberg, Eds., ed: Springer US, pp. 331-340, 2000.

[15] X. Xi, E. Keogh, C. Shelton, L. Wei, and C. A. Ratanamahatana, "Fast time series classification using numerosity reduction," presented at the *Proceedings of the 23rd international conference on Machine learning,* Pittsburgh, Pennsylvania, USA, 2006.

[16] C. Ratanamahatana and E. Keogh, "Making Time-Series Classification More Accurate Using Learned Constraints," in *SDM*, 2004.

[17] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," in *Readings in speech recognition*, W. Alex and L. Kai-Fu, Eds., ed: Morgan Kaufmann Publishers Inc., pp. 159-165, 1990.

[18] F. Itakura, "Minimum prediction residual principle applied to speech recognition," in *Readings in speech recognition*, W. Alex and L. Kai-Fu, Eds., ed: Morgan Kaufmann Publishers Inc., pp. 154-158, 1990.

[19] S.-W. Kim, S. Park, and W. W. Chu, "An Index-Based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases," presented at

the *Proceedings of the 17th International Conference on Data Engineering*, 2001.

[20] B.-K. Yi, H. V. Jagadish, and C. Faloutsos, "Efficient Retrieval of Similar Time Sequences Under Time Warping," presented at the *Proceedings of the Fourteenth International Conference on Data Engineering*, 1998.

[21] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowl. Inf. Syst.,* vol. 7, pp. 358-386, 2005.

[22] L. Gupta, D. L. Molfese, R. Tammana, and P. G. Simos, "Nonlinear alignment and averaging for estimating the evoked potential," *IEEE Trans Biomed Eng,* vol. 43, pp. 348-56, Apr, 1996.

[23] V. Niennattrakul and C. A. Ratanamahatana, "Shape averaging under Time Warping," in *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2009. ECTI-CON 2009. 6th International Conference on*, pp. 626-629, 2009.

[24] V. Niennattrakul, D. Srisai, and C. A. Ratanamahatana, "Shape-based template matching for time series data," *Knowledge-Based Systems,* vol. 26, pp. 1-8, Febuary, 2012.

[25] F. Petitjean, A. Ketterlin, and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognition,* vol. 44, pp. 678-693, March, 2011.

[26] F. Petitjean, G. Forestier, G. I. Webb, A. E. Nicholson, C. Yanping, and E. Keogh, "Dynamic Time Warping Averaging of Time Series Allows Faster and More Accurate Classification," in *Data Mining (ICDM), 2014 IEEE International Conference on*, pp. 470-479, 2014.

[27] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, and A. M. a. G. Batista, "The UCR Time Series Classification Archive," 2015.

[28] S. Rodpongpun, V. Niennattrakul, and C. A. Ratanamahatana, "Selective Subsequence Time Series clustering," *Knowledge-Based Systems,* vol. 35, pp. 361-368, November, 2012.

[29] T. Bollerslev and H. Ole Mikkelsen, "Modeling and pricing long memory in stock market volatility," *Journal of Econometrics,* vol. 73, pp. 151-184, July, 1996.

[30] W.-C. Wang, K.-W. Chau, C.-T. Cheng, and L. Qiu, "A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series," *Journal of Hydrology,* vol. 374, pp. 294-306, August, 2009.

[31] M. Ghil and R. Vautard, "Interdecadal oscillations and the warming trend in global temperature time series," *Nature,* vol. 350, pp. 324-327, March, 1991.

[32] T. Rath and R. Manmatha,. "Word Image Matching Using Dynamic Time Warping," *CVPR,* vol. 2, pp. 521-527, 2003.

[33] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Applied Statistics,* vol. 28, pp. 100-108, 1979.

[34] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert Systems with Applications,* vol. 36, pp. 3336-3341, March, 2009.

[35] L. K. a. P. J. Rousseeuw, *Finding Groups in Data: an Introduction to Cluster Analysis.*, 1990.

[36] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," *SIGMOD Rec.,* vol. 25, pp. 103-114, 1996.

[37] M. Ester, H.-p. Kriegel, S. Jörg, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," pp. 226-231, 1996.

[38] B. J. Frey and D. Dueck, "Clustering by Passing Messages Between Data Points," *Science,* vol. 315, pp. 972-976, 2007.

[39] L. R. Rabiner, B. Gold, and C. K. Yuen, "Theory and Application of Digital Signal Processing," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 8, pp. 146-146, 1978.

[40] P. Costantini and R. Morandi, "Monotone and convex cubic spline interpolation," *CALCOLO,* vol. 21, pp. 281-294, 1984.

[41] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition,* vol. 38, pp. 2270-2285, December, 2005.

[42] J. M. Bland and D. G. Altman, "Statistics Notes: Measurement error," *BMJ,* vol. 313, pp. 744, September, 1996.

[43] G. R. Norman, J. A. Sloan, and K. W. Wyrwich, "Interpretation of Changes in Health-Related Quality of Life: The Remarkable Universality of Half a Standard Deviation," *Medical Care,* vol. 41, pp. 582-592, 2003.

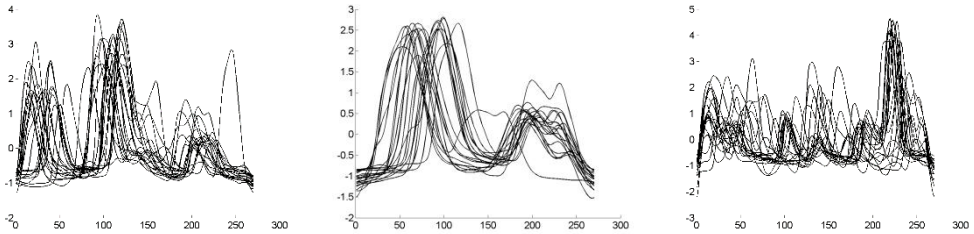
[44] J. Shao, "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association,* vol. 88, pp. 486-494, 1993.

ภาคผนวก

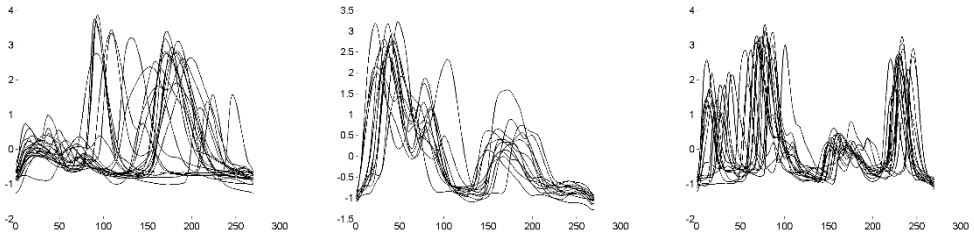
ภาคผนวก ก

ข้อมูลทั้งหมด 20 ชุดที่ใช้ในการทดลอง ซึ่งได้กล่าวถึงลักษณะต่าง ๆ ของข้อมูลไว้ในบทที่ 4 จะถูกแสดงในภาคผนวก ก โดยจะมีการแจกแจงในรูปแบบของกราฟ แต่ละกราฟประกอบด้วยตัวอย่างข้อมูลหนึ่งคลาสของชุดข้อมูล ในที่นี้จะทำการแจกแจงข้อมูลเฉพาะข้อมูลเรียนรู้ที่ได้นำมาใช้ในการทดลองสร้างแผ่นแบบเท่านั้น

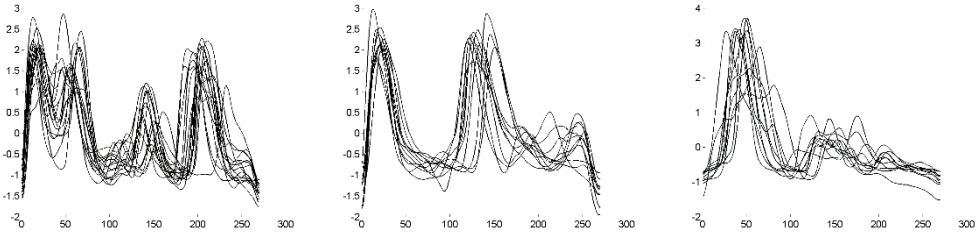
1. ชุดข้อมูล 50 words มีจำนวนข้อมูลทั้งหมด 50 คลาส แสดงดังรูป ก.1



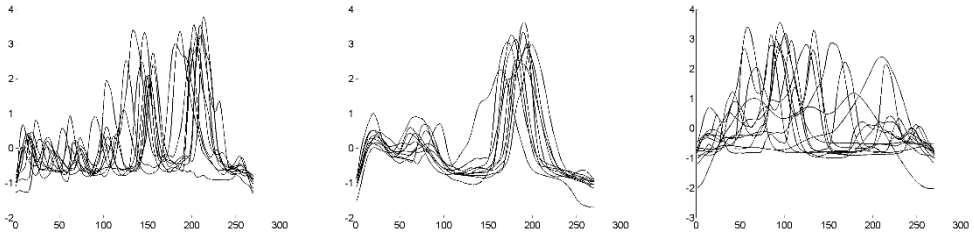
คลาส 1 คลาส 2 คลาส 3



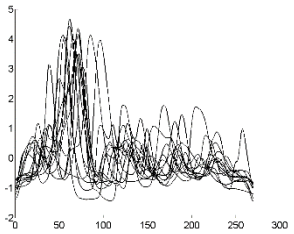
คลาส 4 คลาส 5 คลาส 6



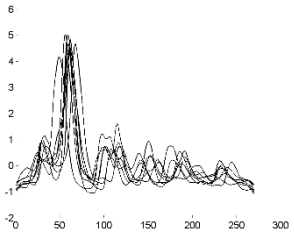
คลาส 7 คลาส 8 คลาส 9



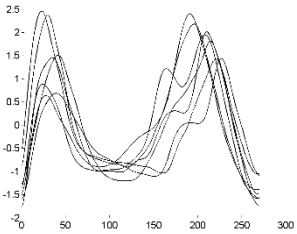
คลาส 10 คลาส 11 คลาส 12



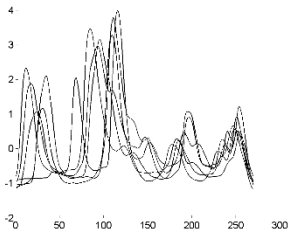
คลาส 13



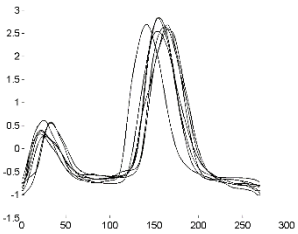
คลาส 14



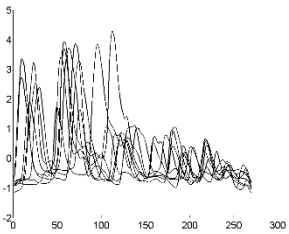
คลาส 15



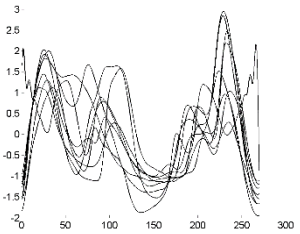
คลาส 16



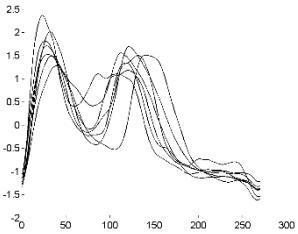
คลาส 17



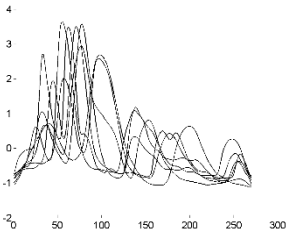
คลาส 18



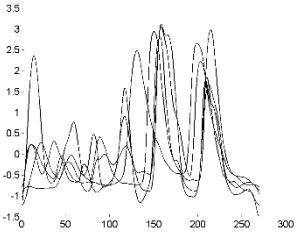
คลาส 19



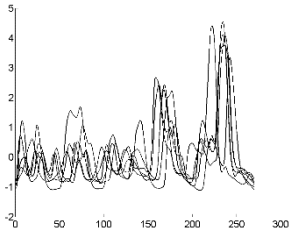
คลาส 20



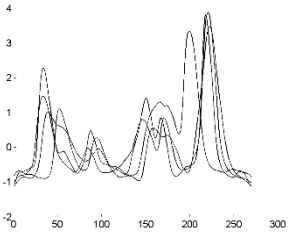
คลาส 21



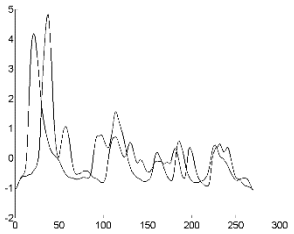
คลาส 22



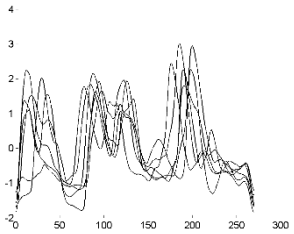
คลาส 23



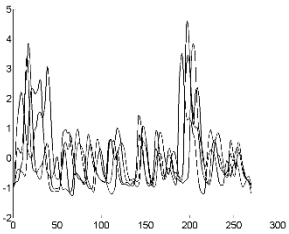
คลาส 24



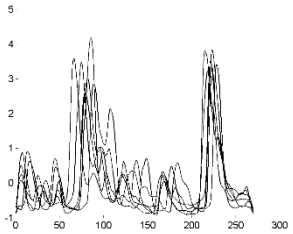
คลาส 25



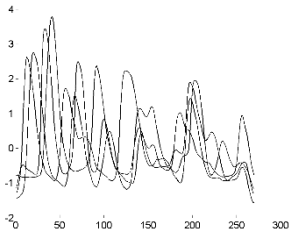
คลาส 26



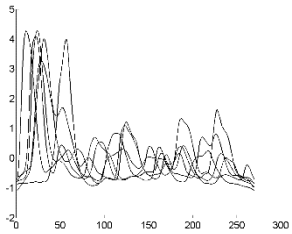
คลาส 27



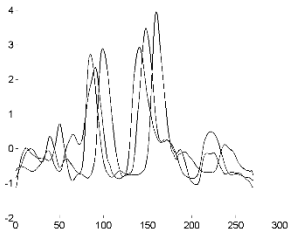
คลาส 28



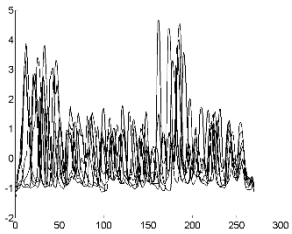
คลาส 29



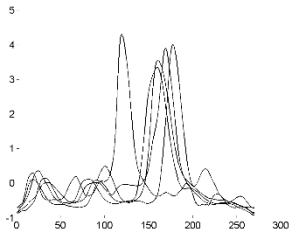
คลาส 30



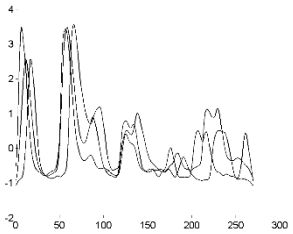
คลาส 31



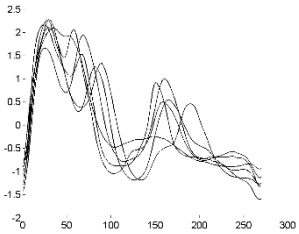
คลาส 32



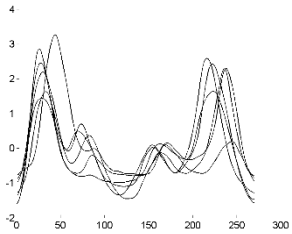
คลาส 33



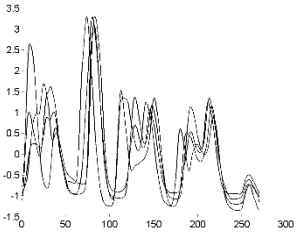
คลาส 34



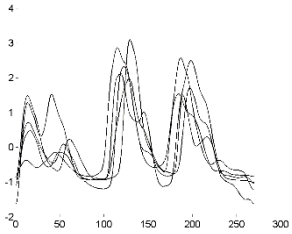
คลาส 35



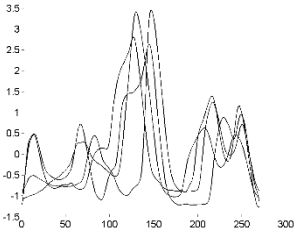
คลาส 36



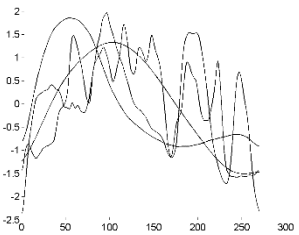
คลาส 37



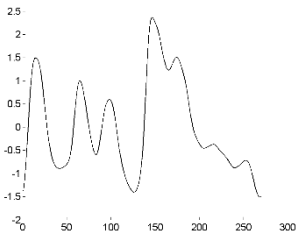
คลาส 38



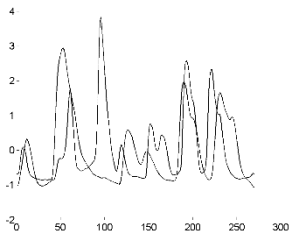
คลาส 39



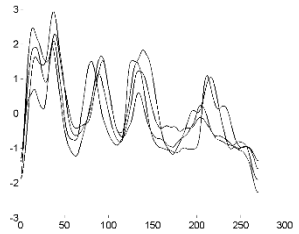
คลาส 40



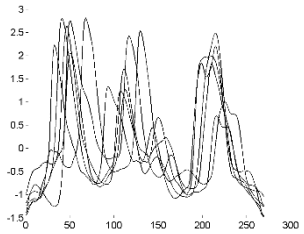
คลาส 41



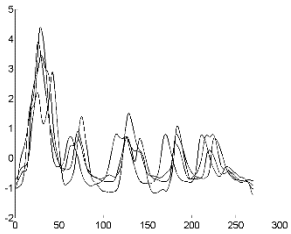
คลาส 42



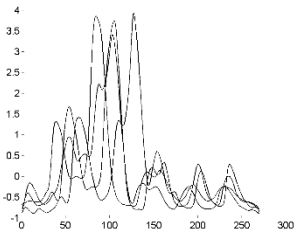
คลาส 43



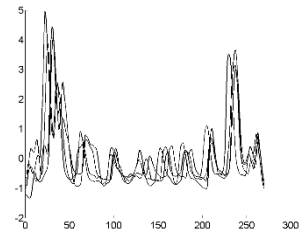
คลาส 44



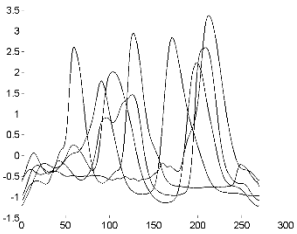
คลาส 45



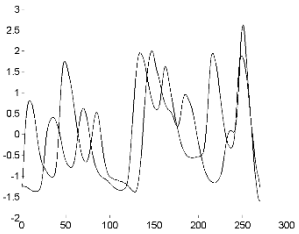
คลาส 46



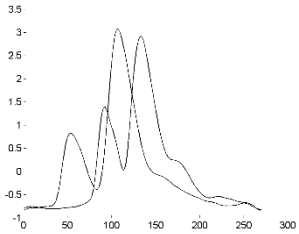
คลาส 47



คลาส 48



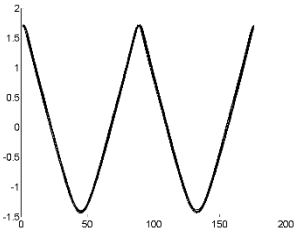
คลาส 49



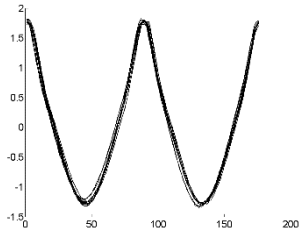
คลาส 50

รูปที่ ก.1 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล 50 words

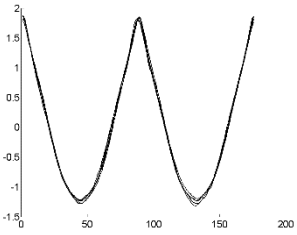
2. ชุดข้อมูล Adiac มีจำนวนข้อมูลทั้งหมด 37 คลาส แสดงดังรูป ก.2



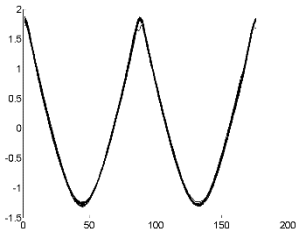
คลาส 1



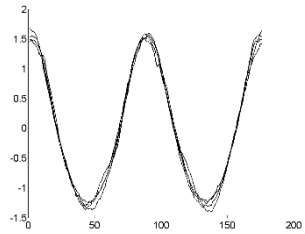
คลาส 2



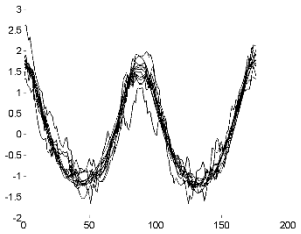
คลาส 3



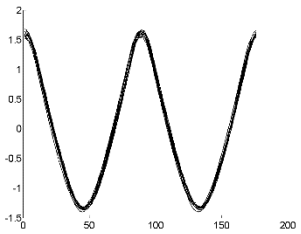
คลาส 4



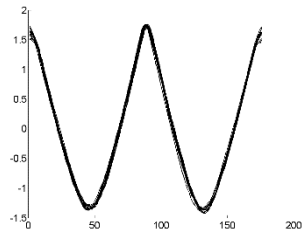
คลาส 5



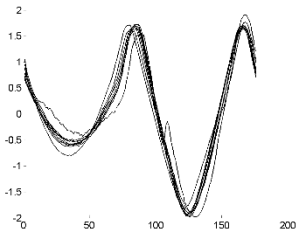
คลาส 6



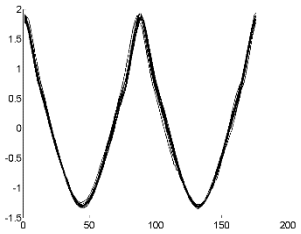
คลาส 7



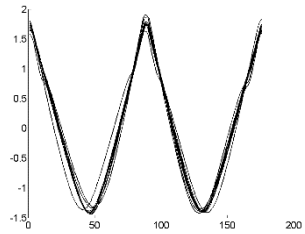
คลาส 8



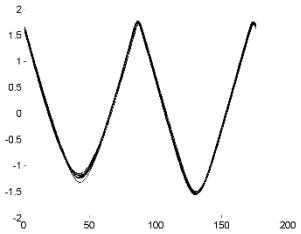
คลาส 9



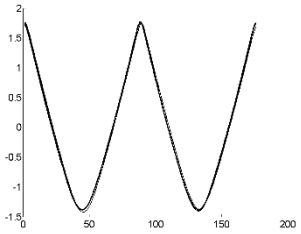
คลาส 10



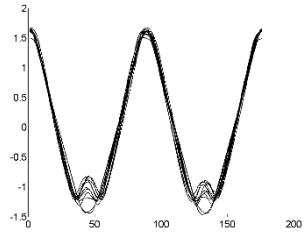
คลาส 11



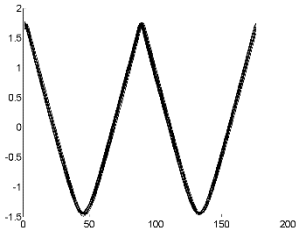
คลาส 12



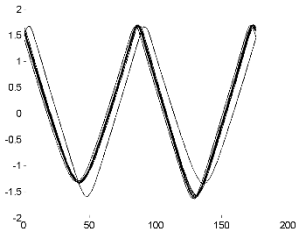
คลาส 13



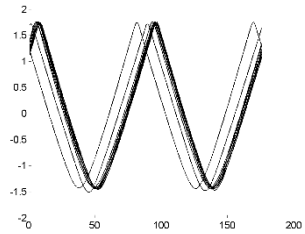
คลาส 14



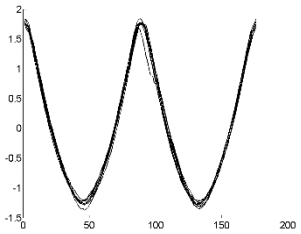
คลาส 15



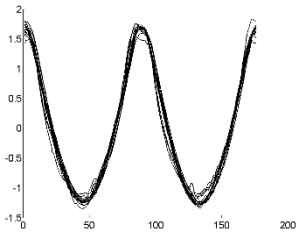
คลาส 16



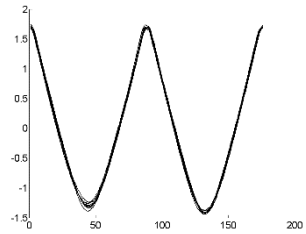
คลาส 17



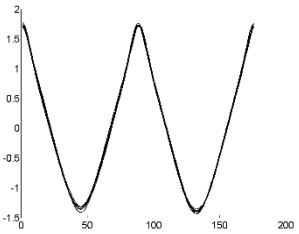
คลาส 18



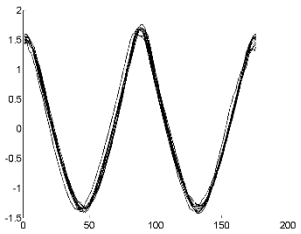
คลาส 19



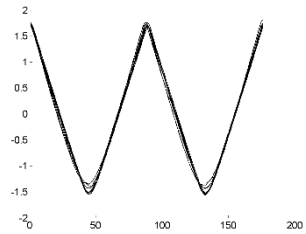
คลาส 20



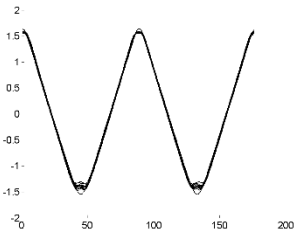
คลาส 21



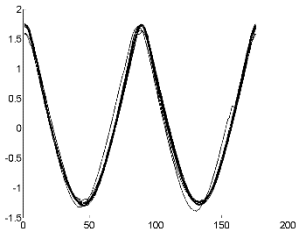
คลาส 22



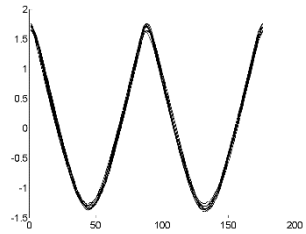
คลาส 23



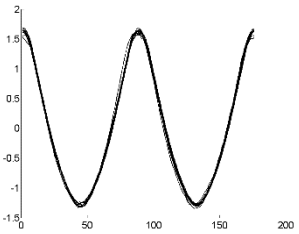
คลาส 24



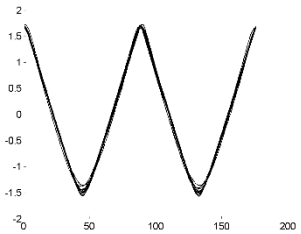
คลาส 25



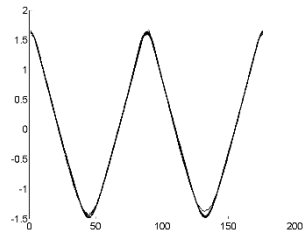
คลาส 26



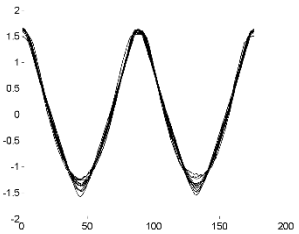
คลาส 27



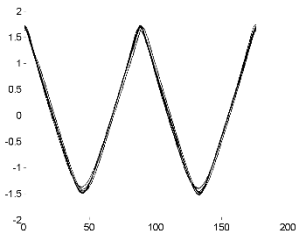
คลาส 28



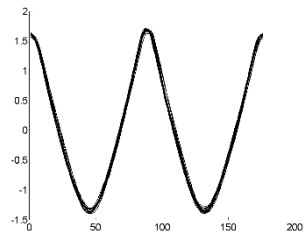
คลาส 29



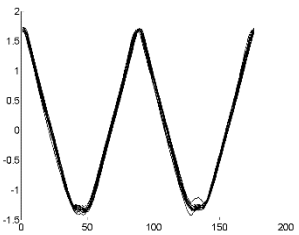
คลาส 30



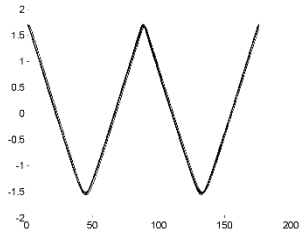
คลาส 31



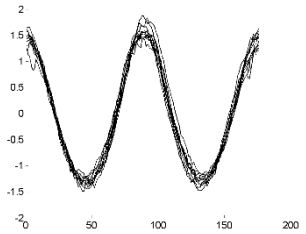
คลาส 32



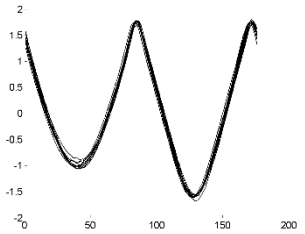
คลาส 33



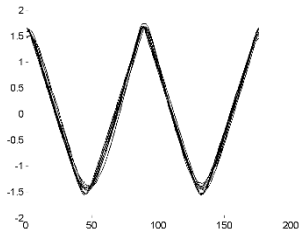
คลาส 34



คลาส 35



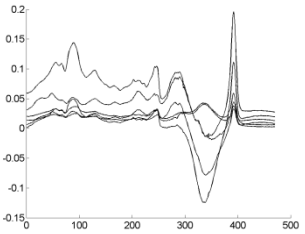
คลาส 36



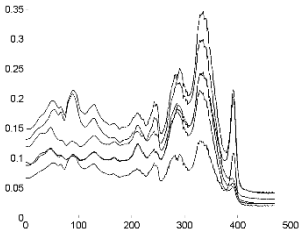
คลาส 37

รูปที่ ก.2 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล Adiac

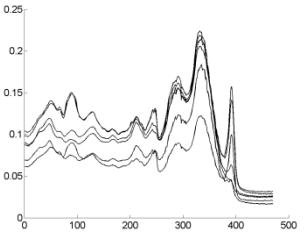
3. ชุดข้อมูล Beef มีจำนวนข้อมูลทั้งหมด 5 คลาส แสดงดังรูป ก.3



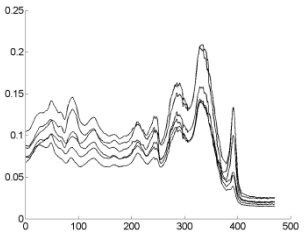
คลาส 1



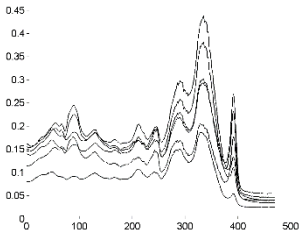
คลาส 2



คลาส 3



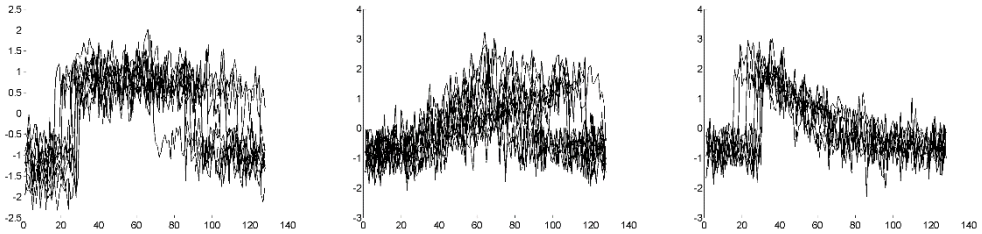
คลาส 4



คลาส 5

รูปที่ ก.3 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล Beef

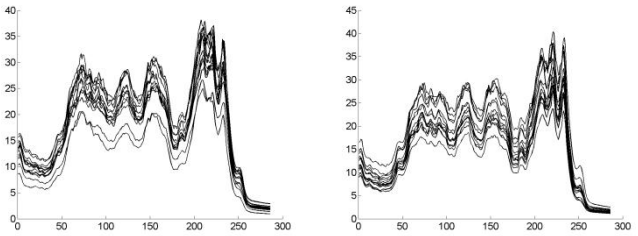
4. ชุดข้อมูล CBF มีจำนวนข้อมูลทั้งหมด 3 คลาส แสดงดังรูป ก.4



คลาส 1 คลาส 2 คลาส 3

รูปที่ ก.4 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล CBF

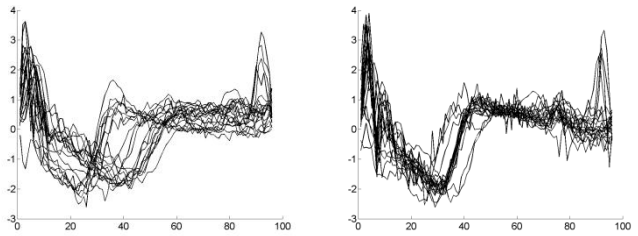
5. ชุดข้อมูล Coffee มีจำนวนข้อมูลทั้งหมด 2 คลาส แสดงดังรูป ก.5



คลาส 1 คลาส 2

รูปที่ ก.5 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล Coffee

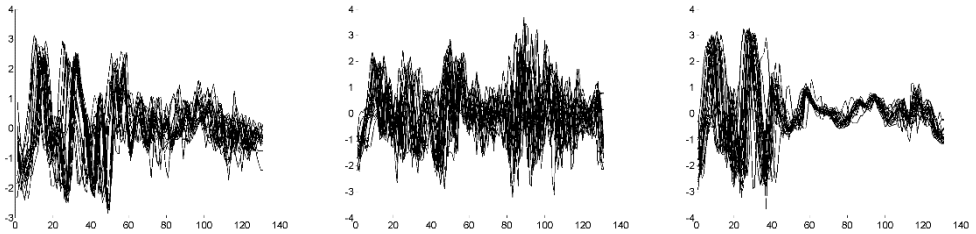
6. ชุดข้อมูล ECG200 มีจำนวนข้อมูลทั้งหมด 2 คลาส แสดงดังรูป ก.6



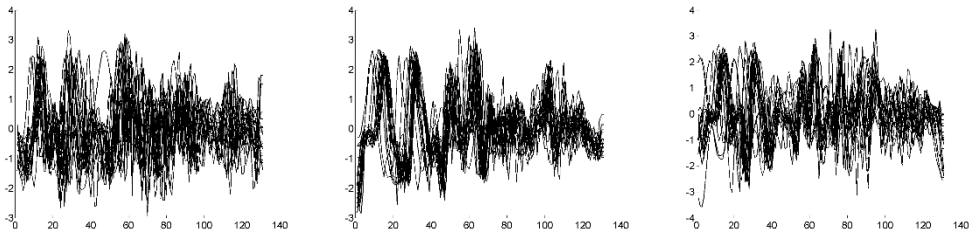
คลาส 1 คลาส 2

รูปที่ ก.6 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล ECG200

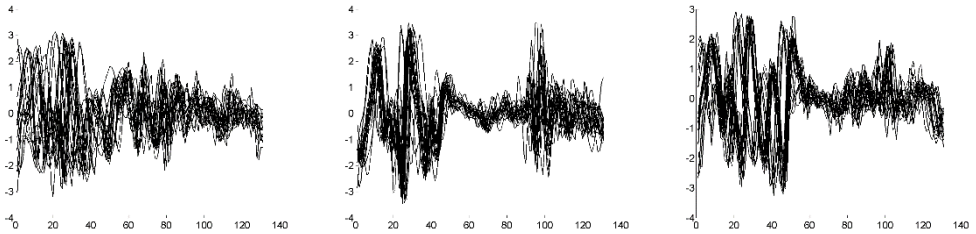
7. ชุดข้อมูล FaceAll มีจำนวนข้อมูลทั้งหมด 14 คลาส แสดงดังรูป ก.7



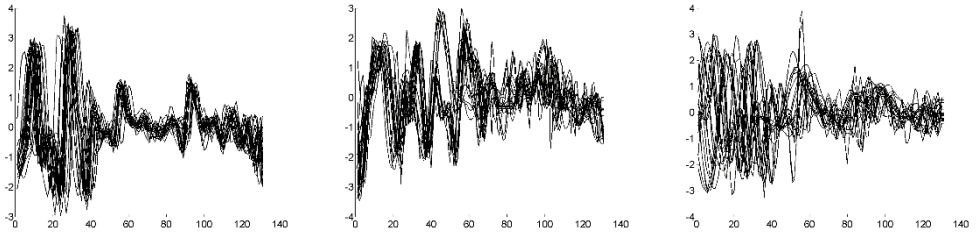
คลาส 1 คลาส 2 คลาส 3



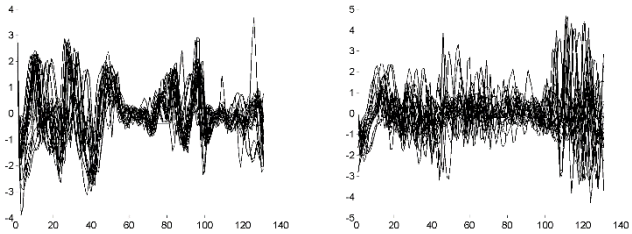
คลาส 4 คลาส 5 คลาส 6



คลาส 7 คลาส 8 คลาส 9



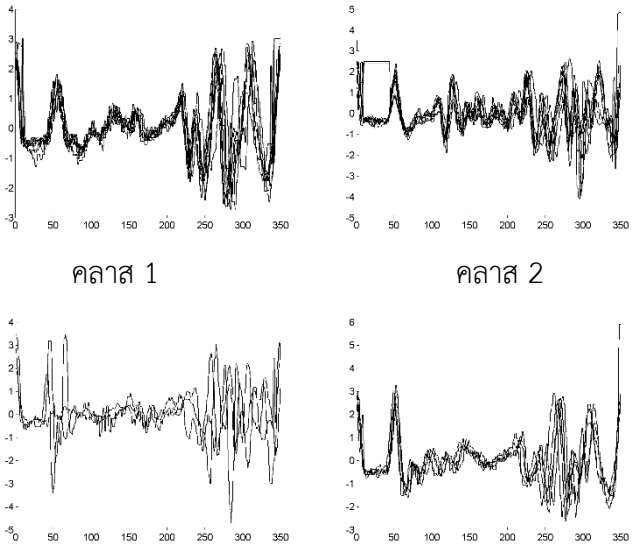
คลาส 10 คลาส 11 คลาส 12



คลาส 13 คลาส 14

รูปที่ ก.7 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล FaceAll

8. ชุดข้อมูล FaceFour มีจำนวนข้อมูลทั้งหมด 4 คลาส แสดงดังรูป ก.8

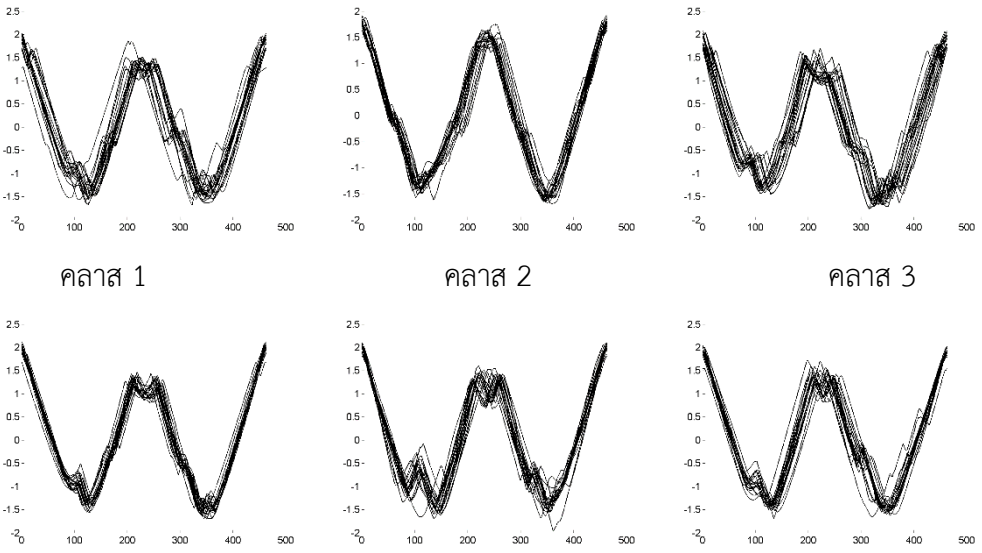


คลาส 1 คลาส 2

คลาส 3 คลาส 4

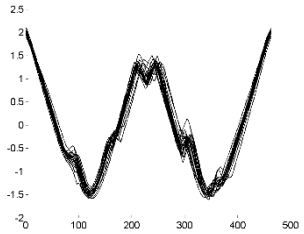
รูปที่ ก.8 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล FaceFour

9. ชุดข้อมูล FISH มีจำนวนข้อมูลทั้งหมด 7 คลาส แสดงดังรูป ก.9



คลาส 1 คลาส 2 คลาส 3

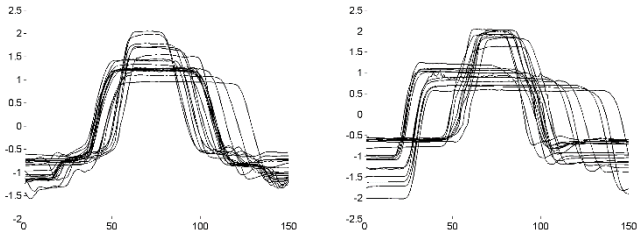
คลาส 4 คลาส 5 คลาส 6



คลาส 7

รูปที่ ก.9 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล FISH

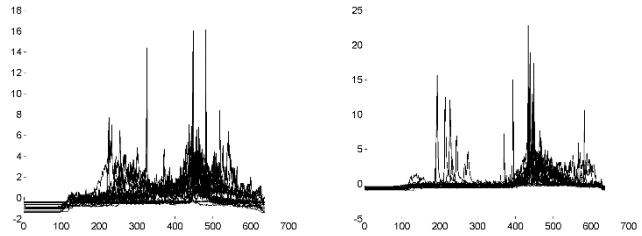
10. ชุดข้อมูล Gun_Point มีจำนวนข้อมูลทั้งหมด 2 คลาส แสดงดังรูป ก.10



คลาส 1 คลาส 2

รูปที่ ก.10 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล Gun_Point

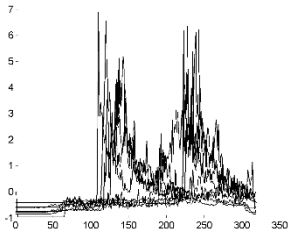
11. ชุดข้อมูล Lightning2 มีจำนวนข้อมูลทั้งหมด 2 คลาส แสดงดังรูป ก.11



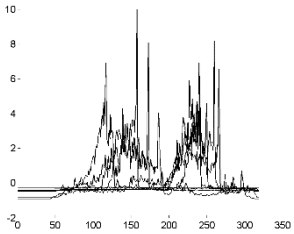
คลาส 1 คลาส 2

รูปที่ ก.11 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล Lightning2

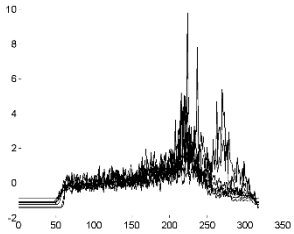
12. ชุดข้อมูล Lightning7 มีจำนวนข้อมูลทั้งหมด 7 คลาส แสดงดังรูป ก.12



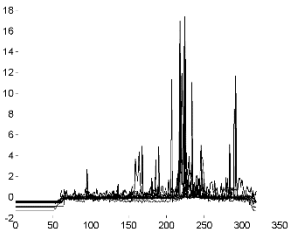
คลาส 1



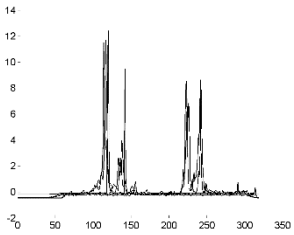
คลาส 2



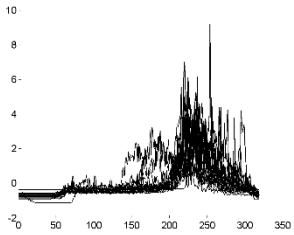
คลาส 3



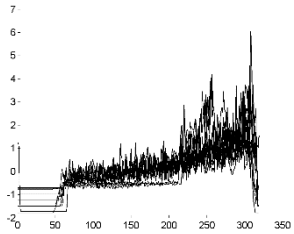
คลาส 4



คลาส 5



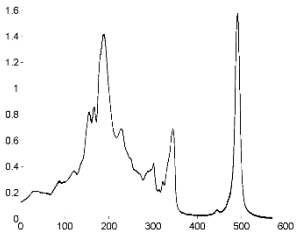
คลาส 6



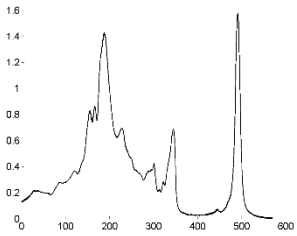
คลาส 7

รูปที่ ก.12 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล Lightning7

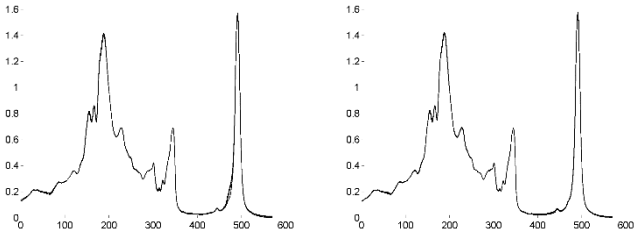
13. ชุดข้อมูล OliveOil มีจำนวนข้อมูลทั้งหมด 4 คลาส แสดงดังรูป ก.13 จากรูปจะเห็นว่าข้อมูลแต่ละตัวมีความใกล้เคียงกันมาก ส่งผลให้รูปของกราฟมีความใกล้เคียงกับข้อมูลอนุกรมเวลาเพียงหนึ่งข้อมูลเท่านั้น



คลาส 1



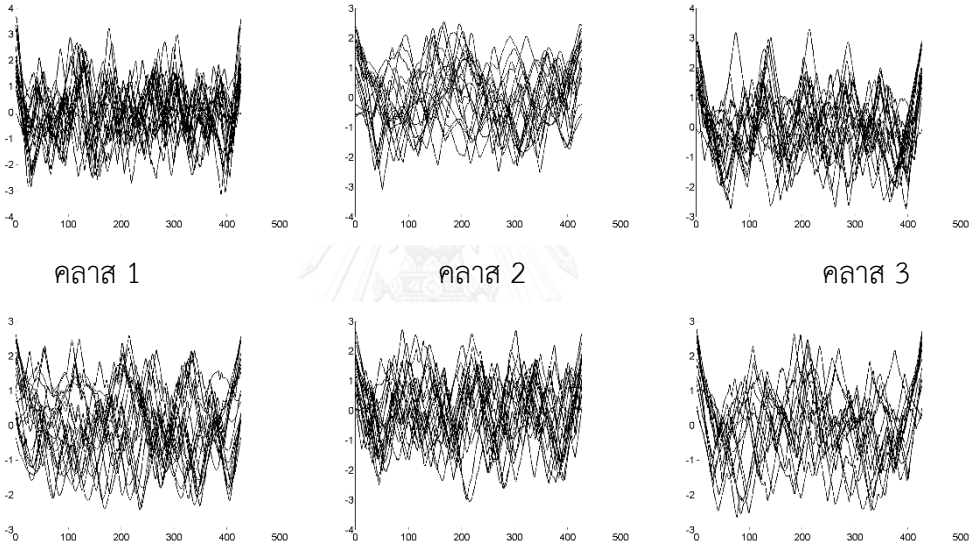
คลาส 2



คลาส 3 คลาส 4

รูปที่ ก.13 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล OliveOil

14. ชุดข้อมูล OSULeaf มีจำนวนข้อมูลทั้งหมด 6 คลาส แสดงดังรูป ก.14

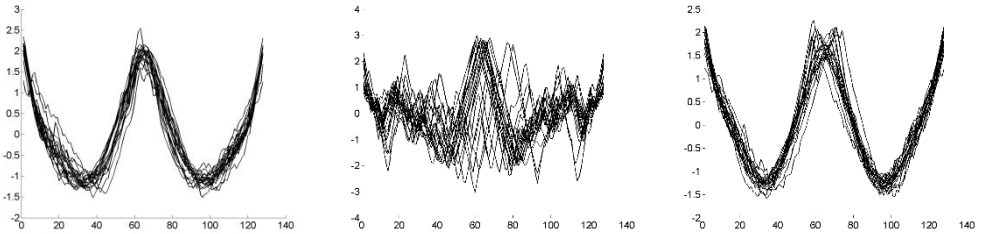


คลาส 1 คลาส 2 คลาส 3

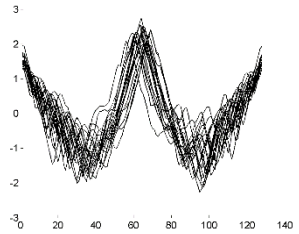
คลาส 4 คลาส 5 คลาส 6

รูปที่ ก.14 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล OSULeaf

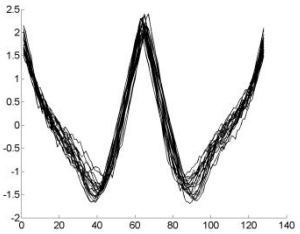
15. ชุดข้อมูล SwedishLeaf มีจำนวนข้อมูลทั้งหมด 15 คลาส แสดงดังรูป ก.15



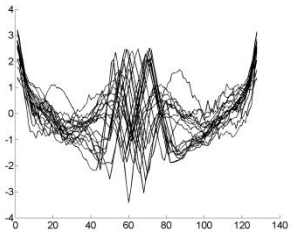
คลาส 1 คลาส 2 คลาส 3



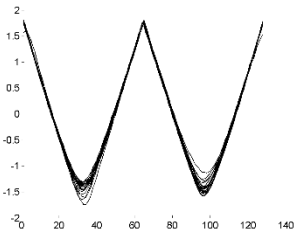
คลาส 4



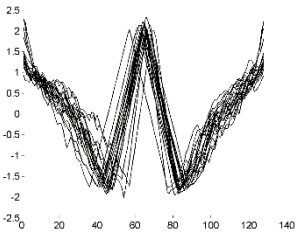
คลาส 5



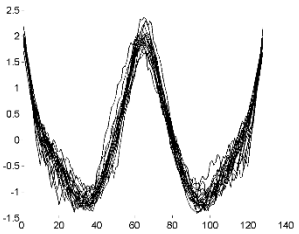
คลาส 6



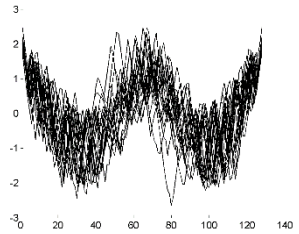
คลาส 7



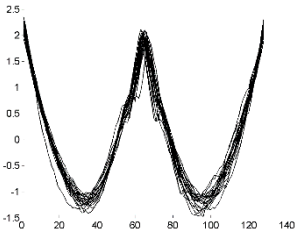
คลาส 8



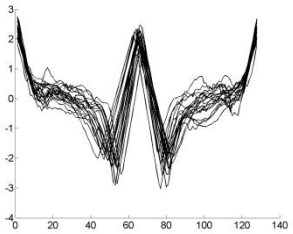
คลาส 9



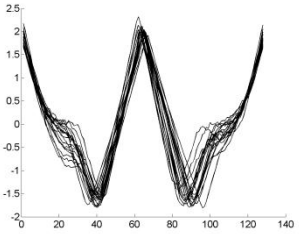
คลาส 10



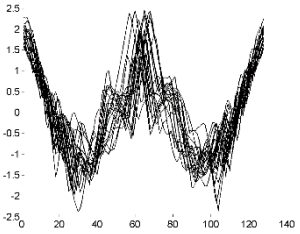
คลาส 11



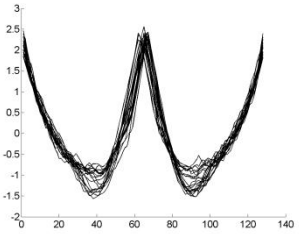
คลาส 12



คลาส 13



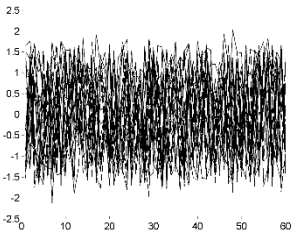
คลาส 14



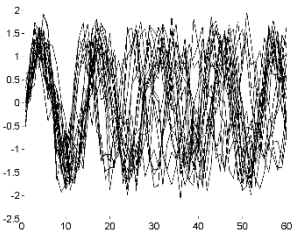
คลาส 15

รูปที่ ก.15 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล SwedishLeaf

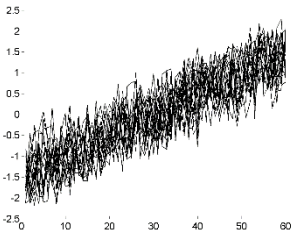
16. ชุดข้อมูล Synthetic_control มีจำนวนข้อมูลทั้งหมด 6 คลาส แสดงดังรูป ก.16



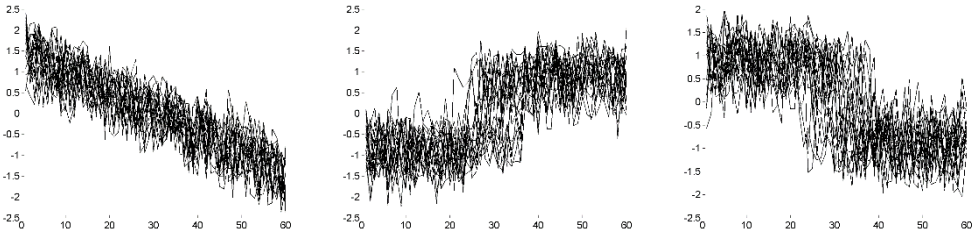
คลาส 1



คลาส 2



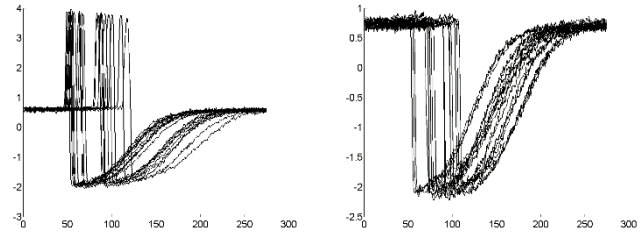
คลาส 3



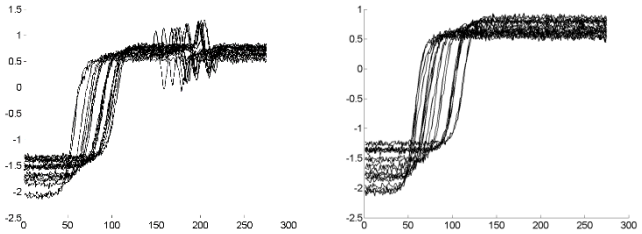
คลาส 4 คลาส 5 คลาส 6

รูปที่ ก.16 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล Synthetic_control

17. ชุดข้อมูล Trace มีจำนวนข้อมูลทั้งหมด 4 คลาส แสดงดังรูป ก.17



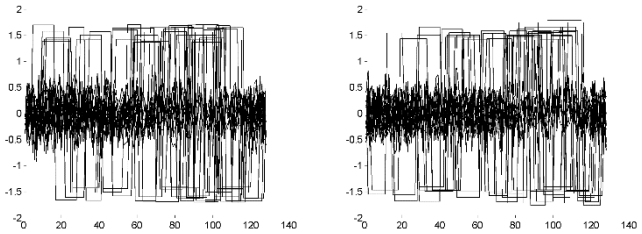
คลาส 1 คลาส 2



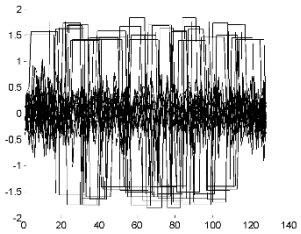
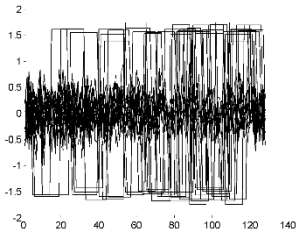
คลาส 3 คลาส 4

รูปที่ ก.17 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล Trace

18. ชุดข้อมูล Two_Patterns มีจำนวนข้อมูลทั้งหมด 4 คลาส แสดงดังรูป ก.18



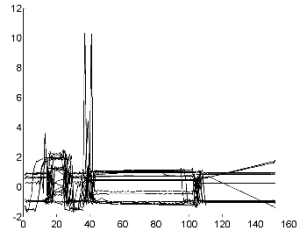
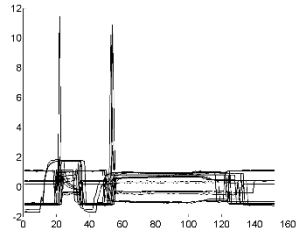
คลาส 1 คลาส 2

คลาส 3 คลาส 4

รูปที่ ก.18 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล Two_Patterns

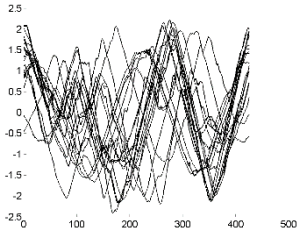
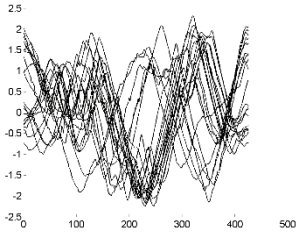
19. ชุดข้อมูล Wafer มีจำนวนข้อมูลทั้งหมด 2 คลาส แสดงดังรูป ก.19

คลาส 1 คลาส 2

รูปที่ ก.19 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล Wafer

20. ชุดข้อมูล Yoga มีจำนวนข้อมูลทั้งหมด 2 คลาส แสดงดังรูป ก.20

คลาส 1 คลาส 2

รูปที่ ก.20 กราฟของข้อมูลอนุกรมเวลาแต่ละคลาสของชุดข้อมูล Yoga

ประวัติผู้เขียนวิทยานิพนธ์

นายพงศกร เสถียรวิริยคุณ เกิดวันที่ 1 พฤศจิกายน พ.ศ. 2535 สำเร็จการศึกษาระดับมัธยมศึกษาจากโรงเรียนสวนกุหลาบวิทยาลัย จากนั้นทำการศึกษาต่อที่คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2553 และสำเร็จการศึกษาปริญญาวิศวกรรมศาสตรบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ ในปีการศึกษา 2556 และเข้าศึกษาในหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ ที่ภาควิชาคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2557