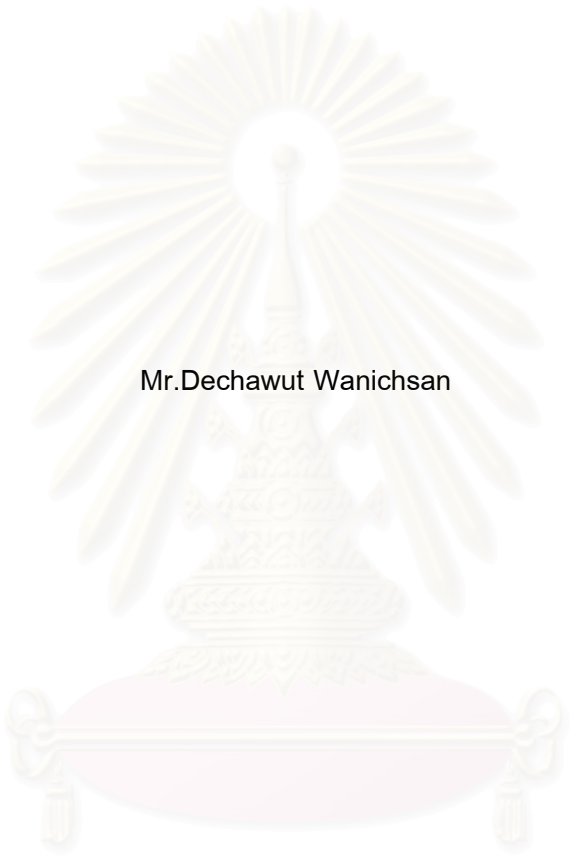
การเพิ่มความแม่นยำให้กับการเลือกเกณฑ์หยุดสำหรับการจำแนกคลาสข้อมูลอนุกรมเวลาแบบกึ่งมีผู้สอน

นายเดชาวุฒิ วานิชสรรพ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2550

ACCURACY IMPROVEMENT OF A STOPPING CRITERION SELECTION FOR SEMI-SUPERVISED TIME SERIES CLASSIFICATION

Mr.Dechawut Wanichsan



A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2007

หัวข้อวิทยานิพนธ์	การเพิ่มความแม่นยำให้กับการเลือกเกณฑ์หยุดสำหรับการจำแนกคลาสข้อมูลอนุกรมเวลาแบบกึ่งมีผู้สอน
โดย	นายเดชาวุฒิ วานิชสรรพ์
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษา	อาจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

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

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

..... อาจารย์ที่ปรึกษา
(อาจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ชาญยศ ปลื้มปิติวิริยะเวช)

เดชาวุฒิ วานิชสรรพ์ : การเพิ่มความแม่นยำให้กับการเลือกเกณฑ์หยุดสำหรับการจำแนกคลาสข้อมูลอนุกรมเวลาแบบกึ่งมีผู้สอน. (ACCURACY IMPROVEMENT OF A STOPPING CRITERION SELECTION FOR SEMI-SUPERVISED TIME SERIES CLASSIFICATION) อาจารย์ที่ปรึกษา : อาจารย์ ดร. โชติรัตน์ รัตนามหัทธนะ, 90 หน้า.

การสร้างตัวจำแนกคลาสสำหรับข้อมูลอนุกรมเวลาให้สามารถจำแนกคลาสได้อย่างมีประสิทธิภาพจะต้องอาศัยข้อมูลที่ทราบคลาสเป็นจำนวนมาก แต่จำนวนข้อมูลประเภทนี้มีอยู่อย่างจำกัด ในขณะที่ข้อมูลที่ไม่ทราบคลาสนั้นมีอยู่ทั่วไป จึงได้มีงานวิจัยอื่นที่นำเสนอการเรียนรู้แบบกึ่งมีผู้สอนด้วยวิธีการฝึกสอนด้วยตนเองที่สามารถสร้างตัวจำแนกคลาสที่ดีแม้ว่าจะใช้ข้อมูลที่ทราบคลาสจำนวนไม่มาก อย่างไรก็ตามการเรียนรู้ประเภทนี้มีข้อจำกัดเกี่ยวกับการหาเกณฑ์หยุด ทำให้ได้ผลการจำแนกคลาสที่ไม่ดีเท่าที่ควร งานวิจัยนี้ได้เสนอการหาเกณฑ์หยุดโดยใช้ค่าระยะทางที่เปลี่ยนแปลงสำหรับการจำแนกคลาสข้อมูลอนุกรมเวลาด้วยการเรียนรู้แบบกึ่งมีผู้สอนและใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงเพื่อช่วยเพิ่มความแม่นยำในการเลือกข้อมูลอนุกรมเวลาขณะทำการฝึกสอน จากการทดลองกับข้อมูลอนุกรมเวลา จำนวน 10 ชุดข้อมูลที่มีความหลากหลาย แสดงให้เห็นว่าตัวจำแนกคลาสที่สร้างจากเกณฑ์หยุดด้วยวิธีที่นำเสนอนั้นสามารถจำแนกคลาสได้ด้วยความถูกต้องแม่นยำมากกว่าการใช้เกณฑ์หยุดแบบเดิม นอกจากนี้งานวิจัยชิ้นนี้ยังได้พัฒนาวิธีการสร้างตัวจำแนกคลาสแบบหลายคลาสที่ให้ผลการจำแนกคลาสที่น่าพอใจอีกด้วย

ภาควิชา...วิทยาศาสตร์คอมพิวเตอร์... ลายมือชื่อนิสิต...เดชาวุฒิ วานิชสรรพ์...
สาขาวิชา...วิศวกรรมคอมพิวเตอร์... ลายมือชื่ออาจารย์ที่ปรึกษา...
ปีการศึกษา...2550...

4970324821 : MAJOR COMPUTER SCIENCE

KEY WORD: SEMI-SUPERVISED LEARNING / SELF TRAINING METHOD / TIME SERIES / CLASSIFICATION / DYNAMIC TIME WARPING

DECHAWUT WANICHSAN : ACCURACY IMPROVEMENT OF A STOPPING CRITERION SELECTION FOR SEMI-SUPERVISED TIME SERIES CLASSIFICATION, THESIS ADVISOR : CHOTIRAT RATANAMAHATANA, Ph.D., 90 pp.

Building a good Time Series classifier necessarily requires a large amount of labeled data. However labeled training data are difficult to obtain, while unlabeled data are largely available. Many typically researchers have proposed Semi-Supervised learning with Self-Training methods, which can build satisfactory classifiers by using only a small amount of labeled data. However, the main limitation of the previous method is the way to determine an optimal stopping criterion. This work proposes a novel stopping criterion for Semi-Supervised Time Series classification, and Dynamic Time Warping technique is used to improve selection data performance during Self Training. The experimental results on 10 different datasets show that this approach can build a better classifier that achieves higher classification accuracy than the previous approach. In addition, the extended proposed work is also shown to have satisfactory result for multi-class semi-supervised time series classifier.

Department..Computer Engineering... Student's signature..DECHAWUT WANICHSAN.....

Field of study..Computer Science.... Advisor's signature.....

Academic year ..2007.....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี เนื่องด้วยความช่วยเหลือ และการให้คำปรึกษาอย่างดีในทุกช่วงเวลาที่ผู้วิจัยต้องการคำปรึกษาของอาจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่กรุณาให้คำปรึกษา เสนอแนะแนวทางการทำวิทยานิพนธ์ ตลอดจนให้การดูแล ให้คำแนะนำ และข้อคิดเห็น ด้วยความเมตตาอย่างที่สุดแก่ผู้วิจัยตลอดการดำเนินการวิจัย ผู้วิจัยขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้ กราบขอบคุณความกรุณาจากศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล ประธานการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.โปรดปราน บุณยพุกกณะ และผู้ช่วยศาสตราจารย์ ดร.ชาญยศ ปลื้มปิติวิริยะเวช ที่กรุณาให้ข้อคิดเห็น และข้อเสนอแนะสำหรับวิทยานิพนธ์ฉบับนี้ และขอบคุณคณาจารย์ภาควิชาวิศวกรรมคอมพิวเตอร์ทุกท่าน ที่ช่วยประสิทธิ์ประสาทความรู้แก่ผู้วิจัยเป็นอย่างดี

ขอบคุณเพื่อน ๆ และน้อง ๆ ทุกคนในห้องปฏิบัติการภาควิชาวิศวกรรมคอมพิวเตอร์ ชั้น 18 ที่ให้คำแนะนำ ให้ข้อคิดเห็น ให้กำลังใจ สร้างเสียงหัวเราะและบรรยากาศในการทำวิจัยที่มีค่ายิ่งแก่ผู้วิจัย ทำให้งานวิจัยชิ้นนี้สำเร็จลุล่วงไปได้ด้วยดี

ขอกราบขอบคุณพ่อและแม่ที่ให้การสนับสนุน และดูและเอาใจใส่ผู้วิจัยด้วยความรัก ความเมตตา และเป็นกำลังใจให้ผู้วิจัย ทำให้การวิจัยสำเร็จลุล่วงไปด้วยดี

สารบัญ

หน้า

หน้า

สารบัญภาพ

หน้า

หน้า

สารบัญตาราง

หน้า

หน้า

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ข้อมูลอนุกรมเวลา (Time Series Data) สามารถพบได้ทั่วไปในชีวิตประจำวัน และเป็นที่สนใจในหลาย ๆ วงการ เช่น วงการแพทย์ (ความดันโลหิต คลื่นหัวใจ) วงการธุรกิจ (ข้อมูลดัชนีหุ้นในตลาดหลักทรัพย์ ผลกำไรจากการค้ารายเดือน) วงการอุตุนิยมวิทยา (ปริมาณน้ำฝน อุณหภูมิในแต่ละวัน) และวงการสังคม (ความถี่ของการเกิดอาชญากรรม ปริมาณการจ้างงาน) นอกจากนี้ยังมีการนำสื่อประสม (Multimedia) เช่น รูปภาพ และวีดีโอ มาแปลงเป็นข้อมูลอนุกรมเวลาอีกด้วย [1] และปัญหาหนึ่งที่น่าสนใจสำหรับการทำเหมืองข้อมูลอนุกรมเวลา (Time Series Data Mining) คือ ปัญหาการจำแนกคลาสข้อมูล (Classification) ซึ่งมีจุดประสงค์เพื่อจำแนกข้อมูลที่ยังไม่ทราบคลาสมาก่อน ด้วยตัวจำแนกคลาส (Classifier) ซึ่งใช้ข้อมูลที่ทราบคลาส (Labeled Data) จำนวนหนึ่งเพื่อนำมาฝึกสอน (Training) ด้วยเกณฑ์การเรียนรู้บางอย่าง ตัวอย่างของการจำแนกคลาสข้อมูลที่พบในชีวิตประจำวัน [2] คือ การจำแนกคลาสข้อมูลคลื่นหัวใจ โดยจำแนกเป็นคลาสของคลื่นหัวใจที่เต้นเป็นปกติและผิดปกติ

การจำแนกคลาสข้อมูลโดยทั่วไปเป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) ซึ่งใช้เฉพาะข้อมูลที่ทราบคลาสเพื่อสร้างตัวจำแนก แต่บางครั้งข้อมูลที่ทราบคลาสนั้นมีจำนวนน้อยเกินไปและข้อมูลส่วนใหญ่ที่มีเป็นข้อมูลที่ไม่ทราบคลาส (Unlabeled Data) ทำให้ได้ผลการจำแนกไม่ดีพอ จึงได้มีการพัฒนาการเรียนรู้แบบกึ่งมีผู้สอน (Semi-Supervised Learning) [3, 4] เพื่อแก้ปัญหากรณีข้อมูลที่ทราบคลาสมีจำนวนไม่มาก

การเรียนรู้แบบกึ่งมีผู้สอน [3, 4] สร้างตัวจำแนกคลาสโดยใช้ข้อมูลที่ทราบและไม่ทราบคลาสมาทำการฝึกสอนร่วมกัน ในกรณีที่ข้อมูลที่ทราบคลาสมีจำนวนไม่มาก การเรียนรู้ประเภทนี้จะสร้างตัวจำแนกที่จำแนกคลาสได้ถูกต้องแม่นยำกว่าการเรียนรู้แบบมีผู้สอน [5] ในปัจจุบันนี้มีเทคนิคการเรียนรู้แบบกึ่งมีผู้สอนหลายวิธี แต่สำหรับข้อมูลอนุกรมเวลาแล้ว วิธีที่นำมาใช้คือ วิธีการฝึกสอนด้วยตนเอง (Self Training) วิธีนี้ทำการฝึกสอนตนเองเพื่อเพิ่มขนาดของเซตข้อมูลในคลาสที่สนใจ ด้วยการกำหนดคลาสให้ข้อมูลที่ไม่ทราบคลาสที่มีความคล้ายกับเซตข้อมูลในคลาสที่สนใจมากที่สุด แต่ถ้ามีการเลือกข้อมูลที่ผิดคลาสขณะทำการฝึกสอน จะส่งผลให้ได้ตัวจำแนกที่ไม่ดีนัก [6] Wei และ Keogh [7] ได้เสนอวิธีการฝึกสอนด้วยตนเองกับข้อมูลอนุกรมเวลา โดยวัดความคล้ายคลึงของข้อมูลด้วยวิธีวัดระยะทางแบบยุคลิด (Euclidean Distance) และหาเกณฑ์หยุด (Stopping Criterion) เพื่อเลือกข้อมูลที่จะนำมาสร้างตัวจำแนกคลาสข้อมูลที่เหมาะสม แต่ในบางกรณี เกณฑ์หยุดที่หาได้นั้นไม่อยู่ในตำแหน่งที่เหมาะสม และ

วิธีวัดระยะทางแบบยุคลิดอาจจำแนกคลาสผิดพลาดขณะทำการฝึกสอนได้ นอกจากนี้แล้ว โดยเฉพาะอย่างยิ่งในกรณีข้อมูลที่อยู่ในคลาสที่ไม่สนใจมีอยู่หลายกลุ่ม เกณฑ์หยุดแบบเดิมจะพบในตำแหน่งของการเปลี่ยนกลุ่มของข้อมูลในคลาสที่เราไม่สนใจ ทำให้เกณฑ์หยุดแบบเดิมไม่สามารถสร้างตัวจำแนกคลาสข้อมูลอนุกรมเวลาที่ดีในกรณีนี้ได้

จากการที่เกณฑ์หยุดไม่สามารถแบ่งข้อมูลสำหรับการสร้างตัวจำแนกที่ดีได้ งานวิจัยนี้จึงได้เสนอวิธีการหาเกณฑ์หยุดแบบใหม่ที่เหมาะสมสำหรับการสร้างตัวจำแนก วิธีที่เสนอจะช่วยแก้ปัญหาในกรณีที่พบเกณฑ์หยุดหลายแห่ง และนอกจากนี้ ขณะทำการฝึกสอนจะวัดความคล้ายคลึงระหว่างข้อมูลอนุกรมเวลาด้วยวิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง (Dynamic Time Warping Distance) วิธีนี้จะให้ผลการจำแนกคลาสที่ดีกว่าวิธีวัดระยะทางแบบยุคลิด [8, 9] ซึ่งจะช่วยแก้ปัญหาการเลือกข้อมูลที่ผิดคลาสขณะทำการฝึกสอนได้

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อพัฒนาตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบกึ่งมีผู้สอนที่มีผลการจำแนกคลาสดีขึ้น
2. เพื่อแก้ปัญหาการเลือกเกณฑ์หยุดให้อยู่ในบริเวณที่เหมาะสมมากยิ่งขึ้น

1.3 ขอบเขตของการวิจัย

1. สร้างตัวจำแนกคลาสอนุกรมเวลาแบบกึ่งมีผู้สอนด้วยวิธีฝึกสอนด้วยตนเอง
2. เสนอวิธีการเลือกเกณฑ์หยุดเพื่อแก้ปัญหาให้อยู่ในบริเวณที่เหมาะสมมากยิ่งขึ้น
3. ทดสอบประสิทธิภาพของตัวจำแนกคลาสที่สร้างจากเกณฑ์หยุดที่คำนวณด้วยวิธีที่งานวิจัยชิ้นนี้นำเสนอกับ ตัวจำแนกคลาสที่สร้างจากเกณฑ์หยุดด้วยวิธีของ Wei และ Keogh [7]
4. ทำการทดลองกับชุดข้อมูลอนุกรมเวลาที่มีความหลากหลายจำนวน 10 ชุดข้อมูล

1.4 ขั้นตอนของการวิจัย

1. ศึกษาลักษณะของข้อมูลอนุกรมเวลา และการจำแนกคลาสข้อมูลอนุกรมเวลา
2. ศึกษาการเรียนรู้แบบกึ่งมีผู้สอน
3. ศึกษาการฝึกสอนด้วยตนเอง
4. ศึกษาการสร้างตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบกึ่งมีผู้สอน
5. ออกแบบวิธีการเลือกเกณฑ์หยุดเพื่อสร้างตัวจำแนกคลาส

6. ทดลองผลการจำแนกคลาสของตัวจำแนกที่สร้างได้
7. ปรับปรุงวิธีเลือกเกณฑ์หยุดที่นำเสนอ
8. สรุปผลและเรียบเรียงวิทยานิพนธ์

1.5 ประโยชน์ที่ได้รับ

1. ได้ตัวจำแนกที่ให้ผลการจำแนกคลาสที่ดีได้แม้ว่าข้อมูลที่ทราบคลาสจะมีจำนวนไม่มากนัก
2. ได้วิธีการเลือกเกณฑ์หยุดที่เหมาะสมสำหรับวิธีการฝึกสอนด้วยตนเองซึ่งช่วยให้ตัวจำแนกที่ได้มีผลการจำแนกคลาสที่แม่นยำมากยิ่งขึ้น
3. ได้วิธีการสร้างตัวจำแนกแบบกึ่งมีผู้สอนที่มีประสิทธิภาพ
4. ได้วิธีการสร้างตัวแนกแบบหลายคลาสที่ช่วยลดจำนวนข้อมูลแปลกแยก

1.6 โครงสร้างของวิทยานิพนธ์

เนื้อหาของวิทยานิพนธ์ฉบับนี้แบ่งออกเป็น 5 บท ดังนี้คือ บทที่ 1 เป็นบทนำ บทที่ 2 กล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องกับการทำงานวิจัยชิ้นนี้ บทที่ 3 กล่าวถึงการดำเนินงานวิจัย โดยอธิบายเป็นขั้นตอนต่าง ๆ อย่างละเอียด ส่วนในบทที่ 4 เป็นการทดลองและผลที่ได้จากการทดลองตามชุดข้อมูลต่าง ๆ และบทที่ 5 เป็นการสรุปผลการทดลองและข้อเสนอแนะของงานวิจัย ซึ่งอาจจะเป็นประโยชน์ต่องานวิจัยอื่น ๆ ในอนาคต

1.7 ผลงานตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของงานวิทยานิพนธ์นี้ได้รับการตีพิมพ์เป็นบทความวิชาการในหัวเรื่อง “การหาเกณฑ์หยุดสำหรับตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบกึ่งมีผู้สอน” โดย เดชาวุฒิ วานิชสรรพ์ และ โชติรัตน์ รัตนามหัทธนะ ในงานประชุมวิชาการ “11th National Computer Science and Engineering Conference (NCSEC 2007)” ซึ่งจัดขึ้น ณ โรงแรมมิราเคิลแกรนด์ กรุงเทพมหานคร ประเทศไทย ระหว่างวันที่ 19-21 พฤศจิกายน 2550 ดังภาคผนวก ข หน้า 75 และนอกจากนี้ยังมีงานวิจัยอื่นที่ได้รับการตีพิมพ์ขณะที่กำลังศึกษาอยู่ ซึ่งมีหัวข้อเรื่อง “Hand Geometry Verification using Time Series Representation” โดย วิชญ์ เนียรนาทตระกูล เดชาวุฒิ วานิชสรรพ์ และ โชติรัตน์ รัตนามหัทธนะ ในงานประชุมวิชาการ “11th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (KES 2007)” ซึ่งจัดขึ้นที่เมือง Vietri sul Mare ประเทศอิตาลี ระหว่างวันที่ 12-14 กันยายน 2550 ดังภาคผนวก ข หน้า 83

บทที่ 2

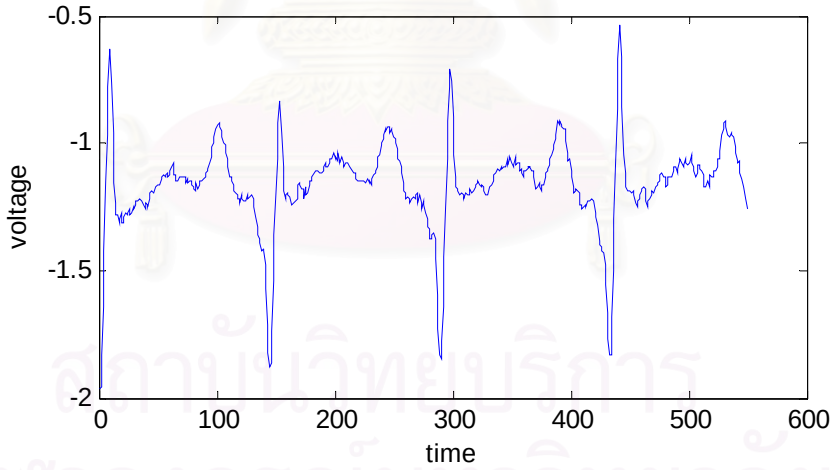
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการทำการวิจัยเกี่ยวกับการคำนวณหาเกณฑ์หยุดเพื่อสร้างตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบกึ่งมีผู้สอนนั้น มีความจำเป็นต้องศึกษาความรู้เพื่อเป็นพื้นฐานต่อการทำการวิจัย ซึ่งแสดงในส่วนของทฤษฎีและงานวิจัยที่เกี่ยวข้องดังต่อไปนี้

2.1 ข้อมูลอนุกรมเวลา

ข้อมูลอนุกรมเวลา หมายถึง ลำดับข้อมูลที่มีค่าเป็นจำนวนจริงที่เกิดขึ้นตามเวลาที่เปลี่ยนแปลงไป เช่น ข้อมูลคลื่นหัวใจ ดังรูปที่ 2.1 รายรับรายจ่ายต่อเดือน ข้อมูลเสียง และข้อมูลดัชนีหุ้นในตลาดหลักทรัพย์ นอกเหนือจากข้อมูลเหล่านี้แล้ว ยังมีการนำสื่อประสมมาแปลงเป็นข้อมูลอนุกรมเวลา เช่น ลายมือ รูปภาพ และวีดีโอ อีกด้วย [1]

กำหนดให้ข้อมูลอนุกรมเวลา $T = t_1, t_2, \ldots, t_n$ โดยที่ t_i เป็นข้อมูลที่มีค่าเป็นจำนวนจริงใด ๆ และ i มีค่าตั้งแต่ 1 ถึง n

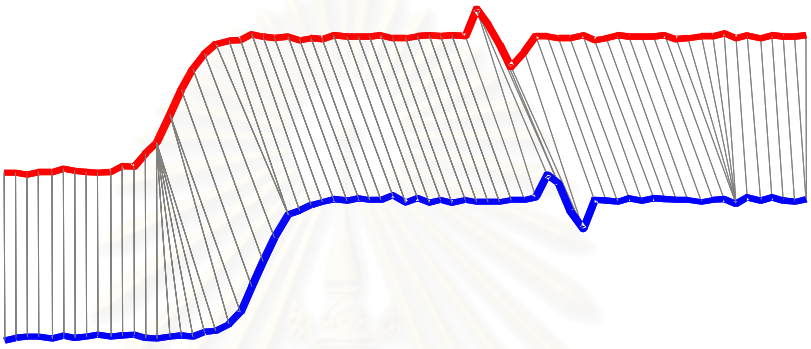


รูปที่ 2.1 ข้อมูลคลื่นหัวใจ

จากข้อมูลคลื่นหัวใจในรูปที่ 2.1 ค่าในแกน x แสดงด้วยเวลา (1/100วินาที) และแกน y แสดงค่าความต่างศักย์ไฟฟ้า (โวลต์)

2.2 วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง

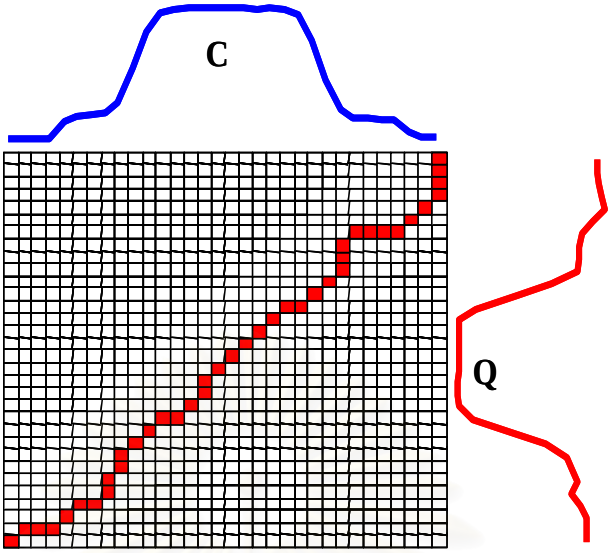
การวัดความคล้ายคลึงระหว่างข้อมูลอนุกรมเวลานั้น สามารถทำได้ด้วยการใช้วิธีวัดระยะทาง เช่น วิธีวัดระยะทางแบบยุคลิด ในงานวิจัยนี้จะใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง [8, 9] ซึ่งมีการคำนวณค่าระยะทางแบบไม่เป็นเชิงเส้น (Non-Linear Alignment) ดังรูปที่ 2.2



รูปที่ 2.2 วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงซึ่งมีการจับคู่จุดข้อมูลในตำแหน่งที่คล้ายกัน (ที่มา : Ratanamahatana และ Keogh [10])

จากรูปที่ 2.2 เห็นได้ว่าหนึ่งจุดของข้อมูลอนุกรมเวลาด้านบน สามารถจับคู่คำนวณค่าระยะทางกับหลายจุดบนข้อมูลอนุกรมเวลาด้านล่าง โดยการคำนวณค่าระยะทางที่มีลักษณะแบบไม่เป็นเชิงเส้นนี้ จะทำให้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงสามารถวัดความคล้ายคลึงระหว่างข้อมูลอนุกรมเวลาได้ดีกว่าวิธีวัดระยะทางแบบยุคลิด

การคำนวณค่าระยะทางของวิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงนั้นใช้หลักการของกำหนดการพลวัต (Dynamic Programming) ซึ่งจะคำนวณค่าระยะทางในทุกเส้นทางที่เป็นไปได้ เพื่อให้ได้ค่าระยะทางสะสมที่มีค่าน้อยที่สุดภายในเมทริกซ์ระยะทาง (Distance Matrix) ดังรูปที่ 2.3



รูปที่ 2.3 การคำนวณค่าระยะทางภายในเมทริกซ์ระยะทาง
(ที่มา : Ratanamahatana และ Keogh [10])

บริเวณที่แรเงาในรูปที่ 2.3 คือเส้นทางการวอร์ปภายในเมทริกซ์ระยะทาง โดยที่จุดสุดท้ายของเส้นทางการวอร์ปคือ ค่าระยะทางที่คำนวณได้จากวิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง โดยวิธีการคำนวณค่าระยะทางเป็นดังรายละเอียดต่อไปนี้

กำหนดให้ข้อมูลอนุกรมเวลา $Q = q_1, q_2, \ldots, q_m$ และ $C = c_1, c_2, \ldots, c_n$ ซึ่งมีความยาวของข้อมูล m และ n ตามลำดับ การคำนวณค่าระยะทางแบบไดนามิกไทม์วอร์ปปิง จะใช้เมทริกซ์ระยะทางที่มีขนาด m x n สำหรับเก็บค่าระยะทางระหว่างจุดบนข้อมูลทั้งสอง ซึ่งคำนวณจากผลรวมระหว่างเซลล์ (i,j) และเซลล์ข้างเคียงที่มีค่าน้อยที่สุด ดังสมการที่ 2.1

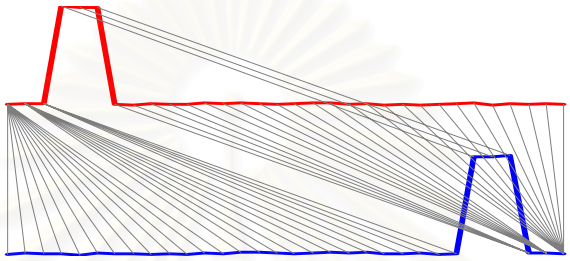
$$\gamma(i,j) = d(q_i, c_j) + \min\{\gamma(i-1,j), \gamma(i,j-1), \gamma(i-1,j-1)\} \tag{2.1}$$

โดยที่ $\gamma(i,j)$ คือผลรวมของค่าระยะทางสะสม ณ เซลล์ (i, j) ส่วน $d(q_i, c_j)$ คือค่าระยะทางซึ่งคำนวณจาก $(q_i - c_j)^2$ ค่าระยะทางแบบไดนามิกไทม์วอร์ปปิงคำนวณจากผลรวมของค่าระยะทางบนเส้นทางที่ให้ค่าระยะทางทางน้อยที่สุด ดังสมการที่ 2.2

$$DTW(Q,C) = \min_{\forall w \in P} \sqrt{\sum_{k=1}^{K} d_{w_k}} \tag{2.2}$$

โดยที่ w คือ เส้นทางการวอร์ปเส้นทางหนึ่งที่มีความยาว K ส่วน P คือเซตของเส้นทางที่เป็นไปได้ทั้งหมด และ d_{w_k} คือ ค่าระยะทางของเส้นทางการวอร์ป w ในลำดับที่ k

โดยปกติแล้ว วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงคำนวณค่าระยะทางโดยเลือกเส้นทางการวอร์ปจากทุกเส้นทางที่เป็นไปได้เพื่อให้ได้ผลรวมค่าระยะทางที่น้อยที่สุด แต่ในความเป็นจริงแล้ว ผลรวมนั้นอาจรวมเส้นทางการวอร์ปที่ไม่เหมาะสมอยู่ด้วย ดังรูปที่ 2.4

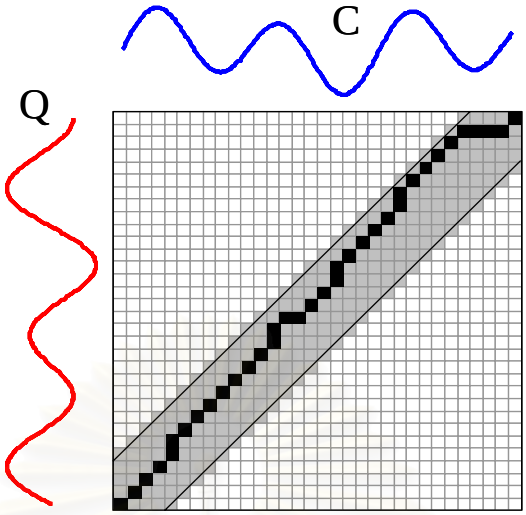


รูปที่ 2.4 การวอร์ปที่ไม่เหมาะสมระหว่างข้อมูลอนุกรมเวลา
(ที่มา : Ratanamahatana และ Keogh [10])

รูปที่ 2.4 แสดงข้อมูลอนุกรมเวลาทั้งสองที่ถูกกำหนดให้มีคลาสที่ต่างกัน แต่วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงจะพยายามหาเส้นทางที่ให้ค่าระยะทางน้อยที่สุด ซึ่งในกรณีนี้มีการวอร์ปเส้นทางการคำนวณที่มากเกินไป ทำให้ได้ผลการจำแนกคลาสที่ไม่ถูกต้อง ดังนั้นจึงควรบังคับให้มีการคำนวณค่าระยะทางไม่ให้เกินจากขอบเขตที่กำหนดไว้ โดยขอบเขตนั้นเรียกว่าเงื่อนไขบังคับโดยรวม (Global Constraint) [9]

เงื่อนไขบังคับโดยรวมเป็นวิธีปรับปรุงประสิทธิภาพของวิธีไดนามิกไทม์วอร์ปปิง โดยวิธีนี้จะบังคับให้การจับคู่คำนวณค่าระยะทางของวิธีไดนามิกไทม์วอร์ปปิงทำงานอยู่ภายในขอบเขตที่กำหนดไว้ ทั้งนี้เพื่อแก้ปัญหาการจับคู่การคำนวณที่ห่างกันมากเกินไปมาพิจารณาในการหาระยะทาง ซึ่งอาจส่งผลโดยตรงต่อการหาค่าระยะทางของกระบวนการนี้

ในงานวิจัยนี้จะใช้เงื่อนไขบังคับโดยรวมแบบซาโก-ชิบะ (Sakoe-Chiba) [11] ซึ่งคือขอบเขตบังคับที่มีลักษณะเป็นเส้นขนานจากเส้นทแยงมุมเป็นดังรูปที่ 2.5



รูปที่ 2.5 เงื่อนไขบังคับโดยรวมแบบซาโก-ชิบะ
(ที่มา : Ratanamahatana และ Keogh [9])

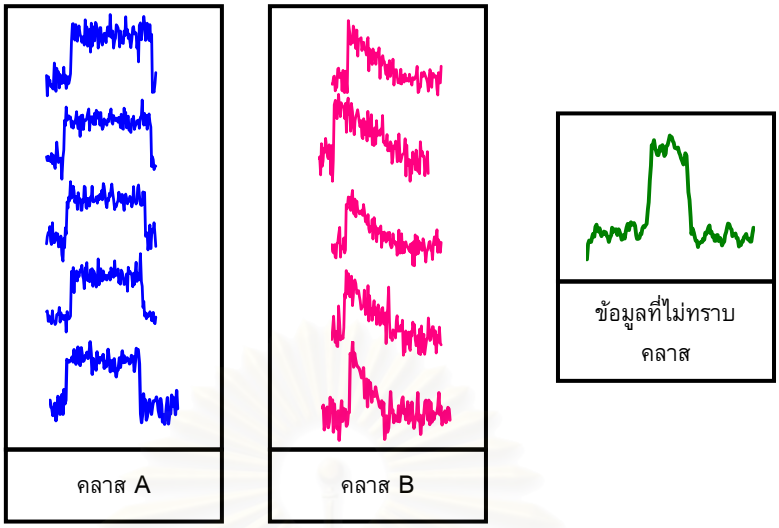
รูปที่ 2.5 แสดงเส้นทางการวอร์ปของไดนามิกไทม์วอร์ปปิงที่จำกัดอยู่ภายในขอบเขตที่กำหนด การจำกัดขอบเขตการวอร์ปด้วยเงื่อนไขบังคับโดยรวมที่มีความเหมาะสมกับข้อมูล จะช่วยลดปัญหาการวอร์ปที่ไม่เหมาะสม ทำให้การจำแนกข้อมูลมีความถูกต้องยิ่งขึ้น

วิธีการวัดระยะทางระหว่างข้อมูลนั้นได้รับการนำไปใช้ในการทำเหมืองข้อมูลอนุกรมเวลาหลายด้าน เช่น การจำแนกคลาสข้อมูล และการจัดกลุ่มข้อมูล เป็นต้น แต่ในงานวิจัยนี้ให้ความสนใจในส่วนของการจำแนกคลาสข้อมูล

2.3 การจำแนกคลาสข้อมูลอนุกรมเวลา

การจำแนกคลาสข้อมูลมีจุดประสงค์เพื่อจำแนกประเภทให้ข้อมูลที่ไม่ทราบคลาสอย่างถูกต้อง ด้วยการใช้ตัวจำแนกคลาส ซึ่งสร้างจากข้อมูลที่ทราบคลาสจำนวนหนึ่ง อย่างไรก็ตาม การสร้างตัวจำแนกด้วยข้อมูลที่ทราบคลาสเพียงอย่างเดียว จะเรียกว่าการเรียนรู้แบบมีผู้สอน (Supervised Learning)

การจำแนกคลาสข้อมูลอนุกรมเวลา (Time Series Classification) [8, 9] สามารถทำได้ด้วยการใช้การรู้จำแบบ (Pattern Recognition) โดยข้อมูลที่มีรูปแบบคล้ายกันจะมีโอกาสมากที่จะอยู่ในคลาสเดียวกัน ดังรูปที่ 2.6 จะเห็นได้ว่าข้อมูลที่ไม่ทราบคลาสมีลักษณะคล้ายกับข้อมูลในคลาส A ข้อมูลนั้นจึงมีโอกาสอยู่ในคลาส A มากกว่าคลาส B



รูปที่ 2.6 ข้อมูลอนุกรมเวลาที่ทราบคลาสสองคลาส และข้อมูลที่ไม่ทราบคลาส
(ที่มา : Ratanamahatana และ Keogh [9])

การวัดความคล้ายคลึงระหว่างข้อมูลอนุกรมเวลาสามารถทำได้ด้วยการใช้วิธีวัดระยะทาง เช่น วิธีวัดระยะทางแบบยุคลิด และวิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง หากค่าระยะทางที่คำนวณได้มีค่าน้อย แสดงว่าข้อมูลมีความคล้ายคลึงกันมาก

การสร้างตัวจำแนกด้วยการเรียนรู้แบบมีผู้สอนจำเป็นต้องใช้ข้อมูลที่ทราบคลาสจำนวนหนึ่ง แต่บางครั้ง ข้อมูลที่ทราบคลาสมีจำนวนน้อยเกินไป ทำให้ได้ผลการจำแนกที่ไม่ดีพอ จึงได้มีการพัฒนาการเรียนรู้แบบกึ่งมีผู้สอน [3-5] ขึ้นสำหรับการแก้ปัญหาในกรณีข้อมูลที่ทราบคลาสมีจำนวนไม่มาก

2.4 การเรียนรู้แบบกึ่งมีผู้สอน

การเรียนรู้แบบกึ่งมีผู้สอน (Semi-Supervised Learning) [3-5] เป็นเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ที่ใช้ทั้งข้อมูลที่ทราบคลาส (Labeled Data) และข้อมูลที่ไม่ทราบคลาส (Unlabeled Data) มาทำการฝึกสอนร่วมกัน โดยนิยมใช้ในกรณีที่ข้อมูลในคลาสที่สนใจมีจำนวนไม่มาก และข้อมูลที่ไม่ทราบคลาสมีอยู่เป็นจำนวนมาก การเรียนรู้แบบกึ่งมีผู้สอนมีอยู่หลายวิธี โดยประสิทธิภาพ จะขึ้นอยู่กับชนิดของข้อมูลที่นำมาใช้ โดยเทคนิคการเรียนรู้แบบกึ่งมีผู้สอนสามารถแบ่งได้เป็น 5 ประเภทหลัก [3-5] ดังนี้

1. แบบจำลองเพิ่มพูน (Generative Models) [12] เป็นการเรียนรู้แบบกึ่งมีผู้สอนที่ใช้เป็นประเภทแรก [4] วิธีนี้เชื่อว่าข้อมูลมีการแจกแจงแบบผสมกัน (Mixture Distribution) โดยข้อมูลที่ไม่ทราบคลาสมีจำนวนมาก สามารถช่วยให้เห็นการแจกแจงของข้อมูลทั้งหมด วิธีการนี้จะใช้ได้ผลที่ดีเมื่อข้อมูลแต่ละคลาสที่เราทำการเรียนรู้มีความแตกต่างระหว่างคลาสพอสมควร อย่างไรก็ตามวิธีนี้ให้ผลการจำแนกคลาสที่ไม่ดีนัก ในกรณีที่ข้อมูลมีจำนวนหลายคลาส

2. วิธีเชิงกราฟ (Graph-Based Methods) วิธีนี้จะสร้างกราฟขึ้นโดยกำหนดโหนด (Node) บนกราฟด้วยข้อมูลทั้งสองประเภท และกำหนดค่าบนเส้นเชื่อม (Edge) ด้วยค่าน้ำหนักความคล้ายคลึงระหว่างโหนดในขั้นแรก จากนั้นจะใช้วิธีการต่าง ๆ เช่น Mincuts [13] สำหรับการแบ่งคลาสข้อมูล โดยประสิทธิภาพของวิธีเชิงกราฟจะขึ้นอยู่กับความสมบูรณ์ของกราฟที่สร้างได้ [4]

3. วิธีแบ่งบริเวณที่มีความหนาแน่นต่ำ (Low Density Based Approaches) [14] วิธีนี้สร้างเส้นแบ่งข้อมูลบริเวณที่มีความหนาแน่นต่ำของข้อมูล เพื่อแบ่งแยกข้อมูลแต่ละคลาสออกจากกัน อย่างไรก็ตาม ข้อมูลอนุกรมเวลามีขนาดมิติของข้อมูลมาก ทำให้ข้อมูลที่อยู่ในคลาสเดียวกันอาจไม่อยู่ในบริเวณที่ใกล้เคียงกัน ดังนั้นการหาบริเวณที่มีความหนาแน่นต่ำของข้อมูลดังกล่าวจึงเป็นเรื่องยาก

4. วิธีฝึกสอนร่วมกัน (Co-Training Methods) [15] วิธีนี้จะแบ่งข้อมูลทั้งหมดออกเป็นสองกลุ่ม แล้วนำข้อมูลแต่ละกลุ่มมาสร้างตัวจำแนก โดยตัวจำแนกที่ได้ในแต่ละกลุ่มจะทำการจำแนกคลาสด้วยการใช้คุณลักษณะ (Feature) ที่ต่างกัน ผลการจำแนกคลาสของแต่ละกลุ่ม จะทำให้จำนวนข้อมูลของอีกกลุ่มเพิ่มมากขึ้น โดยมีความเชื่อว่าแต่ละคุณลักษณะของข้อมูลไม่ขึ้นต่อกัน (Independent) และสามารถแบ่งแยกออกจากกันได้ แต่การใช้เพียงค่าใดค่าหนึ่งบนข้อมูลอนุกรมเวลา ไม่สามารถใช้จำแนกคลาสข้อมูลได้ วิธีการนี้จึงไม่เหมาะกับข้อมูลอนุกรมเวลา

5. วิธีฝึกสอนด้วยตนเอง (Self Training Methods) [4, 6] จะเพิ่มจำนวนข้อมูลที่ทราบคลาส ด้วยข้อมูลที่ไม่ทราบคลาสที่มีความเหมาะสมที่สุดในแต่ละรอบของการฝึกสอน ในงานวิจัยนี้จะสนใจการเรียนรู้แบบกึ่งมีผู้สอนด้วยวิธีฝึกสอนด้วยตนเอง เพราะวิธีนี้มีความเหมาะสมกับลักษณะของข้อมูลอนุกรมเวลาที่ใช้วิธีวัดระยะทางในการวัดความคล้ายคลึงระหว่างข้อมูล (Distance Based) โดยรายละเอียดของวิธีนี้จะกล่าวในหัวข้อถัดไป

2.5 การเรียนรู้แบบกึ่งมีผู้สอนด้วยวิธีฝึกสอนด้วยตนเอง

วิธีฝึกสอนด้วยตนเอง [4, 6] เป็นวิธีการหนึ่งของการเรียนรู้แบบกึ่งมีผู้สอน โดยขั้นตอนวิธีทั่วไปของการฝึกสอนด้วยตนเองแสดงในรูปที่ 2.7

ขั้นตอนวิธี : การฝึกสอนด้วยตนเอง

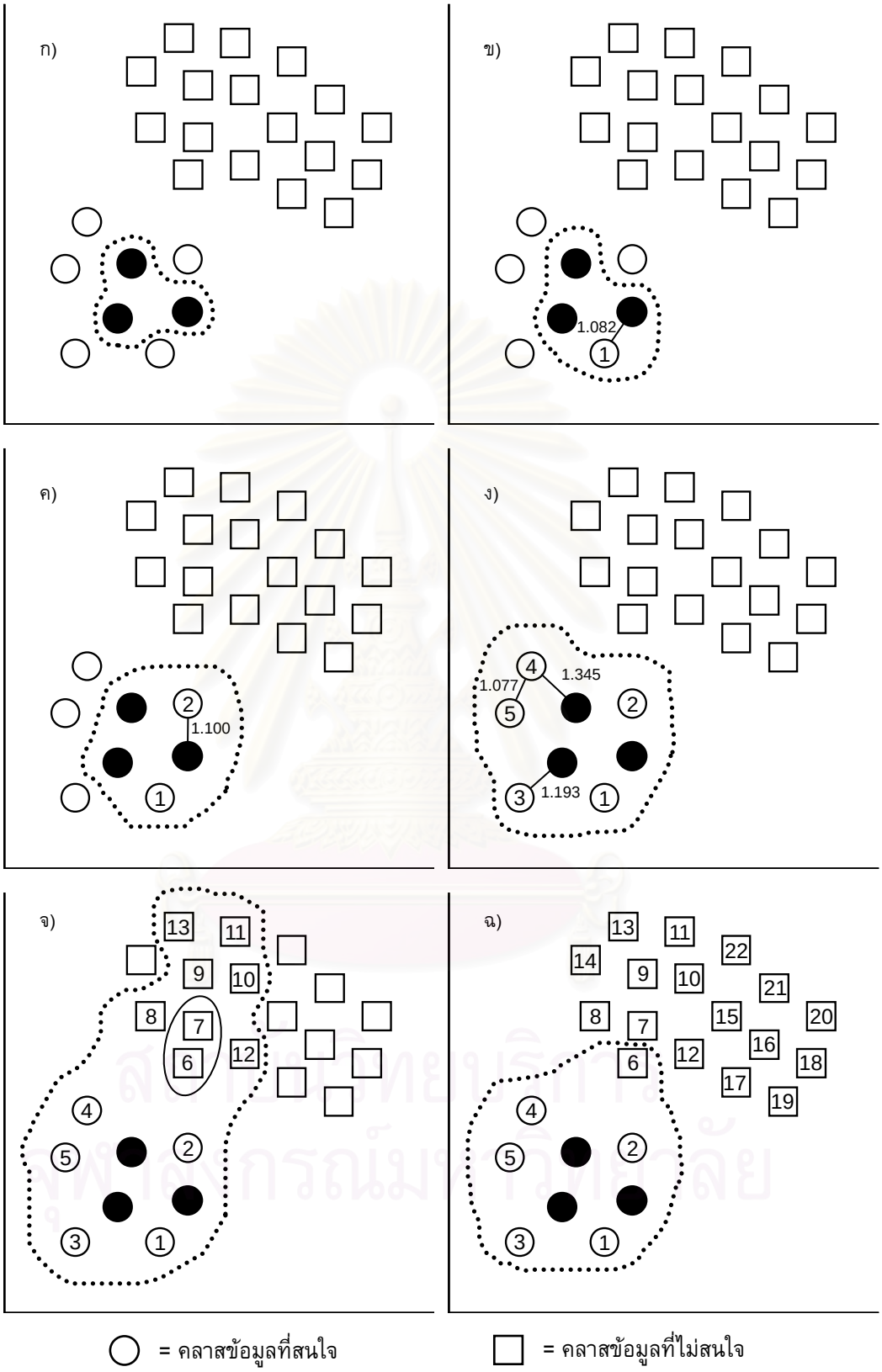
1: สร้างตัวจำแนกเริ่มต้น ด้วยข้อมูลในคลาสที่สนใจ
2: ใช้ตัวจำแนกที่ได้ จำแนกข้อมูลที่ไม่ทราบคลาส
3: เลือกข้อมูลที่ได้รับการจำแนก ที่มีความเหมาะสมที่สุด
4: ย้ายข้อมูลนั้นไปยังคลาสข้อมูลที่สนใจ และสร้างตัวจำแนกคลาสอีกครั้ง
5: ทำขั้นตอนที่ 2 ถึง 4 ซ้ำ จนกระทั่งพบเกณฑ์หยุดที่เหมาะสม

รูปที่ 2.7 ขั้นตอนวิธีการฝึกสอนด้วยตนเอง

การฝึกสอนของวิธีนี้ เริ่มที่การสร้างตัวจำแนกด้วยข้อมูลที่ทราบคลาสจำนวนหนึ่งซึ่งเป็นคลาสที่สนใจ จากนั้นนำตัวจำแนกไปใช้กับข้อมูลที่ไม่ทราบคลาส และเลือกข้อมูลที่ได้รับการจำแนกที่มีความเหมาะสมที่สุด เพื่อเพิ่มเข้าในกลุ่มข้อมูลที่ทราบคลาส กระบวนการฝึกสอนนี้จะทำซ้ำเพื่อเพิ่มจำนวนข้อมูลที่ทราบคลาส จนกระทั่งพบเกณฑ์หยุด ซึ่งเป็นเกณฑ์ที่ใช้เพื่อบอกว่าควรหยุดทำการฝึกสอนเมื่อไร อย่างไรก็ตาม วิธีฝึกสอนด้วยตนเองมีข้อควรระวังสองข้อ คือ การเลือกข้อมูลที่ผิดคลาสขณะที่ทำการฝึกสอนในแต่ละรอบ ทำให้การฝึกสอนในรอบต่อไปอาจเลือกข้อมูลที่ผิดคลาสเพิ่มเข้ามาอีก [16] และตัวจำแนกที่ได้จากการฝึกสอนจะมีลักษณะที่ไม่เหมาะสม [17] ข้อควรระวังอีกข้อหนึ่งคือ เกณฑ์หยุดที่ไม่เหมาะสมทำให้สร้างตัวจำแนกให้ผลการจำแนกคลาสที่ไม่ดีนัก สำหรับข้อมูลอนุกรมเวลา Wei และ Keogh [7] ได้เสนอการเรียนรู้แบบกึ่งมีผู้สอนสำหรับสร้างตัวจำแนกคลาสข้อมูลอนุกรมเวลาด้วยวิธีฝึกสอนด้วยตนเอง ซึ่งจะกล่าวต่อไปในหัวข้อที่ 2.6

2.6 การจำแนกคลาสข้อมูลอนุกรมเวลาด้วยการเรียนรู้แบบกึ่งมีผู้สอน

สำหรับข้อมูลอนุกรมเวลา Wei และ Keogh [7] เสนอวิธีการสร้างตัวจำแนกคลาสข้อมูลอนุกรมเวลาด้วยวิธีการฝึกสอนด้วยตนเอง โดยใช้วิธีวัดระยะทางแบบยุคลิดเพื่อวัดความคล้ายคลึงระหว่างข้อมูลขณะทำการฝึกสอนด้วยตนเอง งานวิจัยนี้เสนอวิธีการหาเกณฑ์หยุดโดยใช้ค่าระยะทางที่มีค่าน้อยที่สุดที่เกิดขึ้นขณะทำการฝึกสอน รูปที่ 2.8 แสดงการฝึกสอนด้วยตนเองในแต่ละรอบการทำงาน



รูปที่ 2.8 การหาเกณฑ์หยุดด้วยค่าระยะทางที่น้อยที่สุด
โดยแต่ละรูป แสดงการฝึกสอนด้วยตนเองในแต่ละรอบการทำงาน

จากรูปที่ 2.8 สัญลักษณ์วงกลมคือ คลาสข้อมูลที่สนใจ สัญลักษณ์สี่เหลี่ยมคือ คลาสข้อมูลที่ไม่สนใจ สัญลักษณ์วงกลมทึบคือ ข้อมูลที่ทราบคลาสขณะเริ่มทำการฝึกสอน กรอบเส้นไข่ปลาคือ กลุ่มข้อมูลที่เป็นตัวจำแนกคลาส ตัวเลขที่อยู่ภายในสัญลักษณ์วงกลมและสี่เหลี่ยมคือ ลำดับข้อมูลที่เพิ่มในกลุ่มตัวจำแนกในแต่รอบขณะทำการฝึกสอน แต่ข้อมูลที่นำมาสร้างตัวจำแนกนั้น

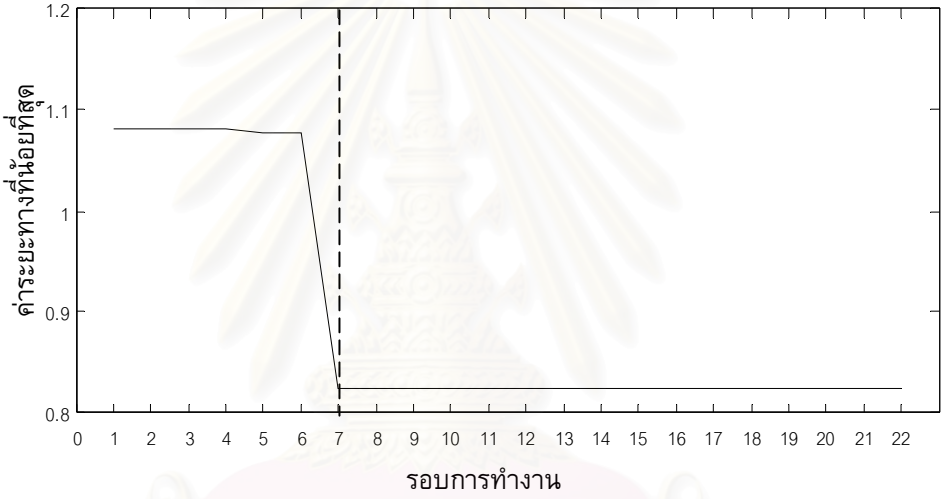
ในที่นี้ข้อมูลที่ทราบคลาส ก่อนเริ่มทำการฝึกสอนมีอยู่ 3 ตัว (สัญลักษณ์วงกลมทึบ) ดังรูปที่ 2.8 ก) การฝึกสอนจะใช้ข้อมูลกลุ่มนี้เลือกข้อมูลที่ไม่ทราบคลาสที่มีความเหมาะสมที่สุด (มีค่าระยะทางน้อยที่สุด) เพื่อเพิ่มเป็นข้อมูลในกลุ่มที่ทราบคลาส โดยจะบันทึกค่าระยะทางและค่าระยะทางที่น้อยที่สุดที่เคยเกิดขึ้น ขณะทำการฝึกสอนแต่ละรอบเอาไว้ด้วย การฝึกสอนรอบแรกทำให้ข้อมูลที่ทราบคลาสมีจำนวนเพิ่มขึ้นเป็น 4 ตัว ดังรูปที่ 2.8 ข) โดยค่าระยะทางระหว่างข้อมูลที่คำนวณได้คือ 1.082 จะได้รับการบันทึกไว้ การฝึกสอนในรอบที่ 2 ดังรูปที่ 2.8 ค) พบว่าค่าระยะทางมีค่า 1.100 ซึ่งมากกว่าค่าระยะทางของการฝึกสอนในรอบแรกคือ 1.082 ค่าระยะทางในรอบนี้จึงยังคงค่าเดิมคือ 1.082 ซึ่งเป็นค่าระยะทางที่มีค่าน้อยที่สุด โดยค่าระยะทางของการฝึกสอนในแต่ละรอบได้รับการบันทึกไว้ ดังตารางที่ 2.1

ตารางที่ 2.1 ค่าระยะทางในแต่ละรอบของการฝึกสอนด้วยตนเอง

รอบการทำงาน	ค่าระยะทาง	ค่าระยะทางที่น้อยที่สุด
1	1.082	1.082
2	1.100	1.082
3	1.193	1.082
4	1.345	1.082
5	1.077	1.077
6	1.800	1.077
7	0.8246	0.8246
8	1.020	0.8246
9	1.100	0.8246
10	1.005	0.8246
11	1.020	0.8246
12	1.166	0.8246
13	1.086	0.8246
14-22	...	0.8246

เมื่อทำการฝึกสอนไปแล้ว 5 รอบดังรูปที่ 2.8 ง) ค่าระยะทางที่น้อยที่สุด คือ 1.077 การฝึกสอนในรอบนี้ทำให้ข้อมูลในคลาสที่สนใจทั้งหมดอยู่ในกลุ่มของตัวจำแนก และเมื่อทำการฝึกสอนไปถึงรอบการฝึกสอนที่ 13 ดังรูป 2.8 จ) ค่าระยะทางที่น้อยที่สุดในตารางที่ 2.1 ได้เปลี่ยนค่าเป็น 0.8246 ในรอบที่ 7 ของการฝึกสอน และการฝึกสอนจะดำเนินต่อไปจนกระทั่งข้อมูลที่ไม่ทราบคลาสทั้งหมดได้รับการกำหนดคลาส และข้อมูลที่นำไปใช้เป็นตัวจำแนกคลาสคือข้อมูลที่ได้รับการฝึกสอนในรอบที่ 1 ถึง 6 แสดงบริเวณกรอบเส้นประ ดังรูปที่ 2.8 ฉ)

เมื่อนำค่าระยะทางที่น้อยที่สุดที่บันทึกไว้ในทุก ๆ รอบของการฝึกสอนไปวาดกราฟ จะเป็นดังรูปที่ 2.9



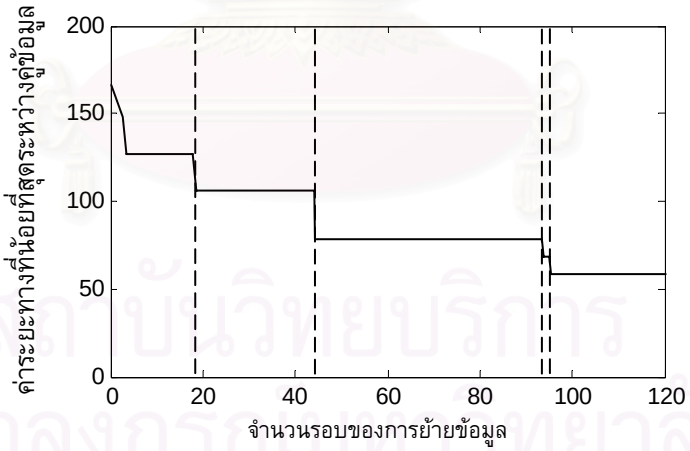
รูปที่ 2.9 ค่าระยะทางที่น้อยที่สุดในแต่ละรอบของการฝึกสอน

ในรูปที่ 2.9 บริเวณเส้นประคือ บริเวณที่กราฟตกลงอย่างเห็นได้ชัดในรอบที่ 7 ของการฝึกสอน ซึ่งเป็นบริเวณที่เป็นเกณฑ์หยุด ดังนั้นข้อมูลที่จะนำไปใช้เป็นตัวจำแนกคลาสคือข้อมูลที่ได้รับการฝึกสอนในรอบก่อนที่พบเกณฑ์หยุด

จากตารางที่ 2.1 จะเห็นว่า ณ รอบที่ 5 ของการฝึกสอน ค่าระยะทางที่น้อยที่สุดมีค่าเปลี่ยนแปลงเช่นเดียวกับรอบที่ 7 ของการฝึกสอน และสาเหตุที่ไม่พิจารณารอบการทำงานนี้เป็นเกณฑ์หยุด เพราะบริเวณที่กราฟระยะทางมีค่าลดลงในรอบแรก ๆ และไม่ใช่ตำแหน่งที่กราฟระยะทางมีค่าลดลงมากที่สุด โดยตำแหน่งที่กราฟระยะทางมีค่าลดลงในรอบแรก ๆ นั้นเป็นค่าระยะทางที่เกิดจากการย้ายข้อมูลในคลาสที่สนใจเข้ามายังตัวจำแนก จึงไม่พิจารณารอบการทำงานนี้เป็นเกณฑ์หยุด

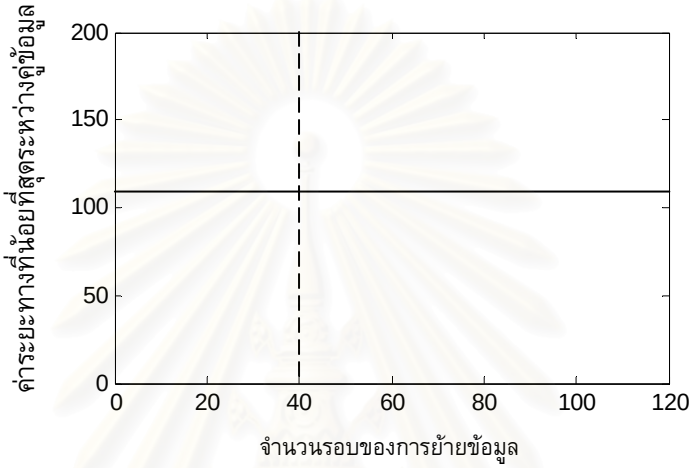
การหาเกณฑ์หยุดด้วยค่าระยะทางที่น้อยที่สุดนี้ มีแนวคิดที่ว่า ข้อมูลที่ทราบคลาสและอยู่ในคลาสที่สนใจมีอยู่เป็นจำนวนน้อย และข้อมูลที่อยู่ในคลาสที่ไม่สนใจมีอยู่เป็นจำนวนมาก ทำให้ข้อมูลที่อยู่ในคลาสนี้น่าจะมีระยะห่างระหว่างคู่ข้อมูลที่ไม่มาก ดังนั้นเมื่อพบค่าระยะทางที่มีค่าน้อยแล้วจึงทราบว่า ข้อมูลในคลาสที่ไม่สนใจ เพิ่มเข้ามายังเซตข้อมูลที่ทราบคลาสแล้ว

อย่างไรก็ตาม การหาเกณฑ์หยุดด้วยการวิธีนี้มีข้อจำกัดหลายข้อคือ การฝึกสอนในบางครั้งจะพบว่า กราฟระยะทางมีค่าระยะทางลดลงหลายครั้งขณะทำการฝึกสอน บริเวณที่จะนำมาพิจารณาเป็นเกณฑ์หยุดจึงพบหลายแห่งซึ่งมักเกิดในกรณีที่ข้อมูลในคลาสที่ไม่สนใจมีอยู่หลายกลุ่มข้อมูล ดังรูปที่ 2.10 ทำให้ไม่สามารถทราบได้ว่าบริเวณใดเป็นบริเวณที่เป็นเกณฑ์หยุดที่เหมาะสมสำหรับตัวจำแนกคลาสที่ดี เส้นประในรูปที่ 2.10 คือบริเวณที่อาจเป็นเกณฑ์หยุดที่เหมาะสม แต่หากสังเกตรูปที่ 2.10 จะเห็นว่าในรอบการฝึกสอนที่ 4 จะไม่นำมาพิจารณาเป็นเกณฑ์หยุดด้วย สาเหตุที่ไม่พิจารณารอบนี้เพราะบริเวณที่กราฟระยะทางมีค่าลดลงในรอบแรก ๆ ซึ่งในรอบนี้เป็นค่าระยะทางที่เกิดจากการย้ายข้อมูลในคลาสที่เราสนใจเข้ามายังตัวจำแนก จึงไม่พิจารณารอบการทำงานนี้เป็นเกณฑ์หยุด โดยเกณฑ์ในการเลือกเกณฑ์หยุดคือ ตำแหน่งที่กราฟระยะทางมีค่าลดลงมากที่สุด แม้ว่าตำแหน่งก่อนพบเกณฑ์หยุดจะพบว่ากราฟระยะทางลดลง แต่ตำแหน่งนั้นจะไม่นำมาพิจารณาเป็นเกณฑ์หยุด



รูปที่ 2.10 กราฟค่าระยะทางที่พบบริเวณที่น่าจะเป็นเกณฑ์หยุดหลายแห่ง

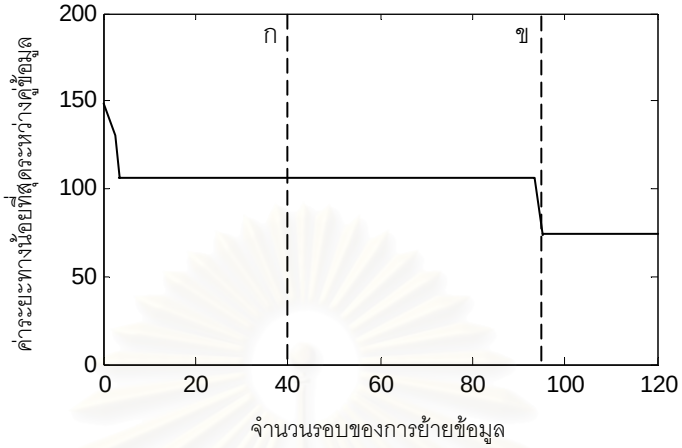
ข้อจำกัดอีกข้อหนึ่งของการหาเกณฑ์หยุดด้วยวิธีนี้คือ กรณีที่ข้อมูลที่ไม่ทราบคลาสมีการกระจายตัวสูง ทำให้การฝึกสอนไม่พบค่าระยะทางที่มีค่าลดลงในบริเวณที่ควรจะเป็นเกณฑ์หยุดที่เหมาะสม ดังรูปที่ 2.11 ที่กราฟระยะทางมีรูปร่างคงที่ และไม่สามารถเห็นบริเวณที่กราฟมีระยะทางมีค่าลดลง ทำให้ไม่สามารถหาเกณฑ์หยุดสำหรับการสร้างตัวจำแนกได้ได้ โดยเส้นประแสดงบริเวณที่เป็นเกณฑ์หยุดเหมาะสม



รูปที่ 2.11 กราฟค่าระยะทางมีรูปร่างคงที่ทำให้ไม่สามารถหาเกณฑ์หยุดได้

จากรูปที่ 2.11 แสดงรูปกราฟระยะทางที่มีรูปร่างคงที่ ทำให้การฝึกสอนด้วยตนเองไม่สามารถเลือกข้อมูลตัวใด ๆ เพิ่มเข้ามายังเซตข้อมูลที่ทราบคลาสได้เลย ทำให้การฝึกสอนด้วยตนเองให้ตัวจำแนกที่มีความสามารถในการจำแนกคลาสได้ไม่แตกต่างจากตัวจำแนกที่ไม่ได้ผ่านการเรียนรู้ แต่สำหรับวิธีการเลือกเกณฑ์หยุดที่นำเสนอนั้นสามารถแก้ปัญหาในกรณีที่นี้ได้

นอกจากนี้ การฝึกสอนด้วยตนเองในกรณีข้อมูลที่ไม่ทราบคลาสมีการกระจายตัวสูง อาจทำให้พบค่าระยะทางที่มีค่าลดลงในรอบหลัง ๆ ของการฝึกสอน ซึ่งเป็นผลทำให้เกณฑ์หยุดที่พบอยู่ในบริเวณที่ไม่เหมาะสม ดังรูปที่ 2.12 ซึ่งพบเกณฑ์หยุดในรอบการทำงานที่ 93 ของการฝึกสอน (บริเวณเส้นประ ข) แต่บริเวณที่เป็นเกณฑ์หยุดที่เหมาะสมอยู่ในรอบที่ 40 ของการฝึกสอน (บริเวณเส้นประ ก) จากปัญหานี้แสดงให้เห็นว่าวิธีการเลือกเกณฑ์หยุดที่ไม่แม่นยำส่งผลให้เลือกข้อมูลที่ไม่ถูกต้องเข้ามาเป็นตัวจำแนกเป็นจำนวนมาก



รูปที่ 2.12 กราฟค่าระยะทางที่พบเกณฑ์หยุดในบริเวณที่ไม่เหมาะสม

จากรูปที่ 2.12 บริเวณเส้นประ ก) แสดงตำแหน่งที่เป็นเกณฑ์หยุดที่เหมาะสม แต่วิธีการหาเกณฑ์หยุดด้วยการใช้ค่าระยะทางที่น้อยที่สุดไม่สามารถเลือกได้ และบริเวณเส้นประ ข) แสดงตำแหน่งที่ไม่ถูกต้องที่วิธีนี้เลือกได้ ซึ่งแสดงให้เห็นว่าวิธีการนี้อาจเลือกเกณฑ์หยุดในตำแหน่งที่ไม่แม่นยำในกรณีที่ข้อมูลในคลาสที่ไม่สนใจมีการกระจายตัวสูง

นอกจากปัญหาการเลือกเกณฑ์หยุดที่อาจส่งผลต่อประสิทธิภาพของตัวจำแนกที่สร้างได้ดังที่กล่าวมาแล้ว การใช้วิธีวัดระยะทางแบบยุคลิดในการวัดค่าระยะทางระหว่างข้อมูลอนุกรมเวลาขณะทำการฝึกสอนด้วยตนเอง อาจทำให้เกิดปัญหาเกี่ยวกับการเลือกข้อมูลที่ผิดคลาสขณะทำการฝึกสอน ซึ่งจะส่งผลต่อการเลือกข้อมูลตัวที่อยู่ในคลาสที่ไม่สนใจในรอบถัดไปของการฝึกสอนอีกด้วย

จากปัญหาของการเลือกเกณฑ์หยุดที่กล่าวมาทั้งหมดนี้ ทำให้ไม่สามารถแบ่งข้อมูลสำหรับการสร้างตัวจำแนกที่ดีได้ ในงานวิจัยชิ้นนี้จึงได้เสนอวิธีการหาเกณฑ์หยุดแบบใหม่ที่เหมาะสมสำหรับการสร้างตัวจำแนกที่ดี นอกจากนี้ยังนำวิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง มาใช้วัดความคล้ายคลึงระหว่างข้อมูลอนุกรมเวลาขณะทำการฝึกสอนด้วยตนเอง โดยวิธีดำเนินงานวิจัยจะกล่าวในบทที่ 3 ต่อไป

บทที่ 3

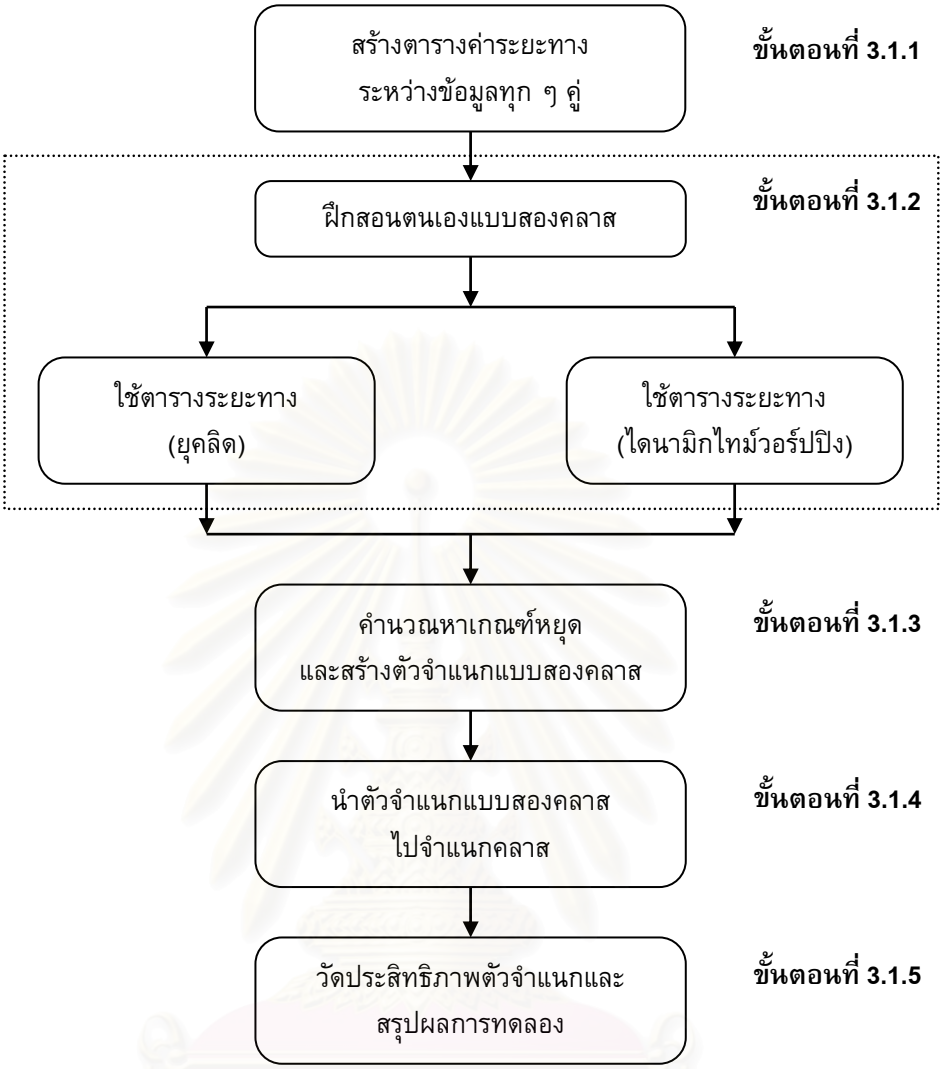
วิธีดำเนินงานวิจัย

จากปัญหาของการเลือกเกณฑ์หยุดด้วยวิธีของ Wei และ Keogh [7] ซึ่งยังมีข้อจำกัดอยู่ งานวิจัยนี้จึงมีจุดมุ่งหมายเพื่อปรับปรุงวิธีการเลือกเกณฑ์หยุดสำหรับการสร้างตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบกึ่งมีผู้สอนให้มีผลการจำแนกคลาสที่ดีกว่าการสร้างตัวจำแนกคลาสด้วยวิธีเลือกเกณฑ์หยุดแบบเดิม นอกจากนี้ในงานวิจัยชิ้นนี้ยังได้ดัดแปลงกระบวนวิธีการฝึกสอนด้วยตนเองให้สามารถฝึกสอนตัวจำแนกได้หลายคลาส ต่างจากเดิมที่สามารถจำแนกคลาสสองคลาส โดยวิธีดำเนินงานวิจัยทั้งหมดมีดังนี้

3.1 ตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบสองคลาส

ในงานวิจัยชิ้นนี้นำเสนอวิธีการสร้างตัวจำแนกคลาสแบบสองคลาสที่มีประสิทธิภาพ โดยกรณีที่นำวิธีนี้มาใช้เป็นกรณีที่เราสนใจข้อมูลในคลาสใดคลาสหนึ่งมากเป็นพิเศษ และข้อมูลในคลาสที่สนใจนั้นมีจำนวนไม่มากพอที่จะสร้างตัวจำแนกคลาสที่ดีได้ ตัวจำแนกคลาสประเภทนี้จะจัดแบ่งข้อมูลขณะเริ่มทำการฝึกสอนออกเป็นสองคลาส คลาสแรกคือข้อมูลที่ทราบคลาสและเราให้ความสนใจที่จะสร้างตัวจำแนกจากข้อมูลเหล่านี้ ส่วนข้อมูลอีกคลาสหนึ่งคือ ข้อมูลอื่นที่ไม่อยู่ในคลาสแรก

ขั้นตอนในการสร้างตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบสองคลาสสามารถแบ่งได้เป็น 5 ขั้นตอน ดังรูปที่ 3.1 โดยในขั้นตอนแรกคือการสร้างตารางค่าระยะทางระหว่างข้อมูลทุก ๆ คู่ข้อมูลเพื่อลดจำนวนครั้งของการคำนวณค่าระยะทาง ต่อมาในขั้นตอนที่ 3.1.2 คือการฝึกสอนด้วยตนเอง ซึ่งทำการฝึกสอนด้วยการใช้วิธีวัดระยะทางที่แตกต่างกันคือวิธีวัดระยะทางแบบยุคลิดและวิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง ส่วนในขั้นตอนที่ 3.1.3 คือการคำนวณหาเกณฑ์หยุดที่เหมาะสม และทำการสร้างตัวจำแนกคลาส จากนั้นขั้นตอนที่ 3.1.4 คือการนำตัวจำแนกคลาสที่สร้างในขั้นตอนที่แล้ว และนำตัวจำแนกที่สร้างได้ไปใช้จำแนกคลาสเพื่อบอกว่าข้อมูลตัวที่ได้รับการจำแนกเป็นข้อมูลที่มีคลาสเดียวกับกลุ่มข้อมูลที่นำมาสร้างเป็นตัวจำแนกหรือไม่ จากนั้นจะเข้าสู่ขั้นตอนสุดท้ายคือการวัดประสิทธิภาพของตัวจำแนกและสรุปผลการทดลอง โดยในส่วนของขั้นตอนที่ 3.1.1-3.1.5 มีรายละเอียดดังต่อไปนี้



รูปที่ 3.1 ขั้นตอนการสร้างตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบสองคลาส

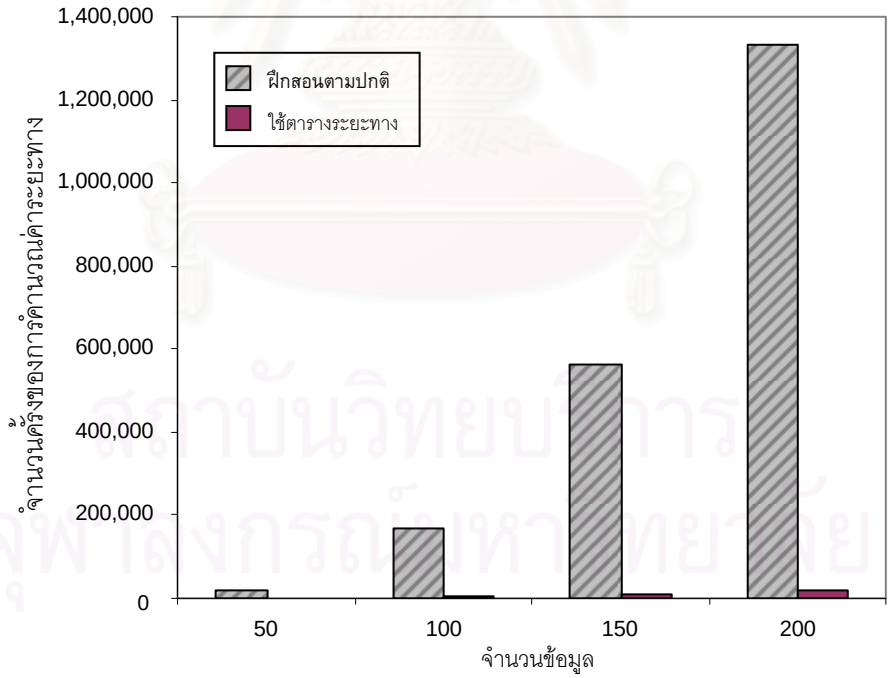
3.1.1 การสร้างตารางค่าระยะทางระหว่างทุกคู่ข้อมูล

เนื่องจากขณะฝึกสอนด้วยตนเองนั้น มีการคำนวณค่าระยะทางระหว่างข้อมูลคู่เดิมซ้ำกันหลายครั้ง งานวิจัยนี้จึงได้นำเทคนิคบางอย่างเพื่อความเร็วในการคำนวณค่าระยะทางด้วยการเรียกใช้ค่าระยะทางที่เคยคำนวณได้รับการคำนวณมาแล้ว (Memoization) มาใช้เพื่อลดจำนวนครั้งของการคำนวณค่าระยะทางลง โดยสามารถดำเนินการได้โดยทำการคำนวณค่าระยะทางระหว่างข้อมูลทุก ๆ คู่ แล้วบันทึกค่าลงในตารางระยะทาง ก่อนที่จะทำการฝึกสอนเพื่อช่วยลดเวลาของการฝึกสอนด้วยตนเอง โดยตารางที่ 3.1 แสดงจำนวนครั้งของการคำนวณระยะทางระหว่างคู่ข้อมูลของการฝึกสอนตามปกติ และการฝึกสอนด้วยการใช้ตารางระยะทาง

ตารางที่ 3.1 การเปรียบเทียบจำนวนครั้งที่คำนวณค่าระยะทางระหว่างการฝึกสอนตามปกติและการใช้ตารางระยะทาง

จำนวนข้อมูล			จำนวนครั้งที่คำนวณค่าระยะทาง	
ทั้งหมด	ทราบคลาส	ไม่ทราบคลาส	การฝึกสอนตามปกติ	การใช้ตารางระยะทาง
50	1	49	20,825	1,225
100	1	99	166,650	4,950
150	1	149	562,475	11,175
200	1	199	1,333,300	19,900

จากตารางที่ 3.1 จะเห็นว่าการคำนวณค่าระยะทางด้วยวิธีการฝึกสอนแบบใช้ค่าระยะทางจากตาราง จะช่วยลดจำนวนครั้งของการคำนวณค่าระยะทางเป็นอย่างมาก เมื่อเทียบกับการฝึกสอนด้วยตนเองแบบปกติ โดยเฉพาะเมื่อข้อมูลมีจำนวนมากขึ้น ดังรูปที่ 3.2 ซึ่งนำข้อมูลจากตารางที่ 3.1 มาวาดกราฟ



รูปที่ 3.2 กราฟเปรียบเทียบจำนวนครั้งของการคำนวณระหว่างการฝึกสอนตามปกติและการใช้ตารางระยะทาง ซึ่งวาดด้วยข้อมูลจากตารางที่ 3.1

3.1.2 การฝึกสอนด้วยตนเองของตัวจำแนกแบบสองคลาส

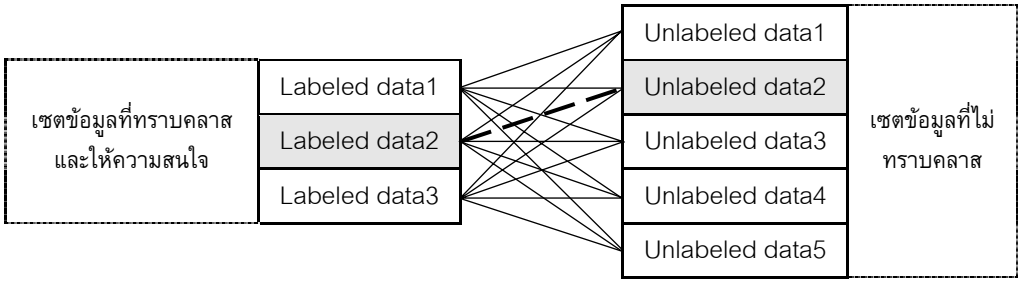
งานวิจัยนี้จะทำการทดลองด้วยการใช้วิธีวัดระยะทางที่แตกต่างกันขณะทำการฝึกสอนด้วยตนเอง คือการนำค่าระยะทางแต่ละคู่ข้อมูลจากตารางระยะทางที่คำนวณจากขั้นตอนที่ 3.1.1 ด้วยวิธีวัดระยะทางแบบยุคลิดและวิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงที่กำหนดค่าเงื่อนไขบังคับโดยรวม (Global Constraint) มาใช้ขณะทำการฝึกสอนด้วยตนเอง โดยขั้นตอนวิธีการฝึกสอนด้วยตนเองเพื่อสร้างตัวจำแนกคลาสอนุกรมเวลาแบบสองคลาส แสดงในรูปที่ 3.3

ขั้นตอนวิธี : การฝึกสอนด้วยตนเองเพื่อสร้างตัวจำแนกคลาสแบบสองคลาส

1: แบ่งข้อมูลทั้งหมดออกเป็นเซตข้อมูลที่ทราบคลาสและให้ความสนใจ กับเซตข้อมูลที่ไม่ทราบคลาส
2: เลือกข้อมูลที่ไม่ทราบที่มีค่าระยะทางที่มีค่าน้อยที่สุดระหว่างทั้ง 2 เซตข้อมูลที่แบ่งในขั้นตอนที่ 1
3: ย้ายข้อมูลที่เลือกได้จากขั้นตอนที่ 2 จากเซตข้อมูลที่ไม่ทราบคลาส ไปยังเซตข้อมูลที่ทราบคลาส พร้อมทั้งบันทึกข้อมูลที่จำเป็นต่อการเลือกเกณฑ์หยุด
4: ตรวจสอบจำนวนข้อมูลในเซตข้อมูลที่ไม่ทราบคลาส
   $\Longrightarrow$ ถ้ายังมีข้อมูลเหลืออยู่ ให้ทำขั้นตอนที่ 2-4 อีกครั้ง
   $\Longrightarrow$ ถ้าไม่มีข้อมูลเหลืออยู่ ให้ทำขั้นตอนที่ 5
5: คำนวณหาเกณฑ์หยุดที่เหมาะสม
6: สร้างตัวจำแนกคลาสแบบสองคลาสจากเกณฑ์หยุดที่คำนวณได้ในขั้นตอนที่ 5

รูปที่ 3.3 ขั้นตอนวิธีการฝึกสอนตนเองเพื่อสร้างตัวจำแนกคลาสแบบสองคลาส

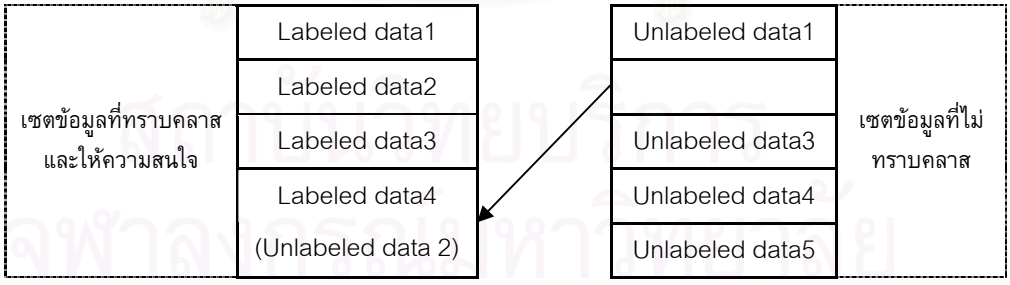
จากรูปที่ 3.3 ขั้นตอนวิธีการฝึกสอนด้วยตนเองจะทำการฝึกสอนเพื่อเพิ่มจำนวนข้อมูลในเซตข้อมูลที่ทราบคลาสและให้ความสนใจ โดยในขั้นตอนที่ 1 ในรูปที่ 3.3 นั้น หากมีข้อมูลที่ทราบคลาสแต่ข้อมูลตัวนั้นเป็นข้อมูลที่ไม่สนใจ ข้อมูลตัวนั้นจะได้รับการแบ่งอยู่ในเซตข้อมูลที่ไม่ทราบคลาสด้วย ขั้นตอนที่สองจะเป็นการหาค่าระยะทางที่มีค่าน้อยที่สุดระหว่างทั้งสองเซตข้อมูลที่ทราบคลาส และเซตข้อมูลที่ไม่ทราบคลาส โดยดึงค่าระยะทางจากตารางระยะทางที่สร้างจากหัวข้อ 3.1.1 (ค่าระยะทางคำนวณจากวิธีวัดระยะทางแบบยุคลิด และวิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง) การที่จะหาว่าค่าระยะทางที่น้อยที่สุดนั้นเกิดขึ้นระหว่างข้อมูลคู่ใดนั้น จำเป็นต้องคำนวณค่าระยะทางในทุกความเป็นไปได้ระหว่างทั้งสองเซตข้อมูลดังรูปที่ 3.4



รูปที่ 3.4 การหาค่าระยะทางที่น้อยที่สุดระหว่างทั้ง 2 เซตข้อมูล

รูปที่ 3.4 แสดงการจับคู่ข้อมูลเพื่อคำนวณค่าระยะทางระหว่างทั้งสองเซตข้อมูล กล่องทางด้านซ้ายคือเซตข้อมูลที่ทราบคลาสและให้ความสนใจ ซึ่งมีสมาชิกสามตัวคือ Labeled data 1-3 ส่วนในกล่องทางขวาคือเซตข้อมูลที่ไม่ทราบคลาสมีสมาชิกห้าตัว คือ Unlabeled data 1-5 ในที่นี้สมมติให้ค่าระยะทางที่น้อยที่สุดเกิดขึ้นระหว่างข้อมูลที่ทราบคลาสตัวที่สอง และข้อมูลที่ไม่ทราบคลาสตัวที่สอง ซึ่งคือข้อมูลที่ได้รับการแรเงา และเส้นเชื่อมระหว่างข้อมูลสองตัวนี้แสดงเป็นเส้นประ

เมื่อได้ข้อมูลที่ทำให้เกิดค่าระยะทางที่น้อยที่สุดแล้ว การดำเนินการในขั้นตอนวิธีที่ 3 ในรูปที่ 3.3 จะทำการย้ายข้อมูลตัวนั้นจากเซตข้อมูลที่ไม่ทราบคลาส (ในที่นี้คือข้อมูลที่ไม่ทราบคลาสตัวที่สองในรูปที่ 3.4) ไปยังเซตข้อมูลที่ทราบคลาสและให้ความสนใจ ซึ่งจะทำให้เซตข้อมูลที่ไม่ทราบคลาสมีจำนวนข้อมูลลดลง และเซตข้อมูลที่ทราบคลาสและให้ความสนใจมีจำนวนเพิ่มมากขึ้น โดยผลลัพธ์ของการย้ายคลาสข้อมูลแสดงในรูปที่ 3.5



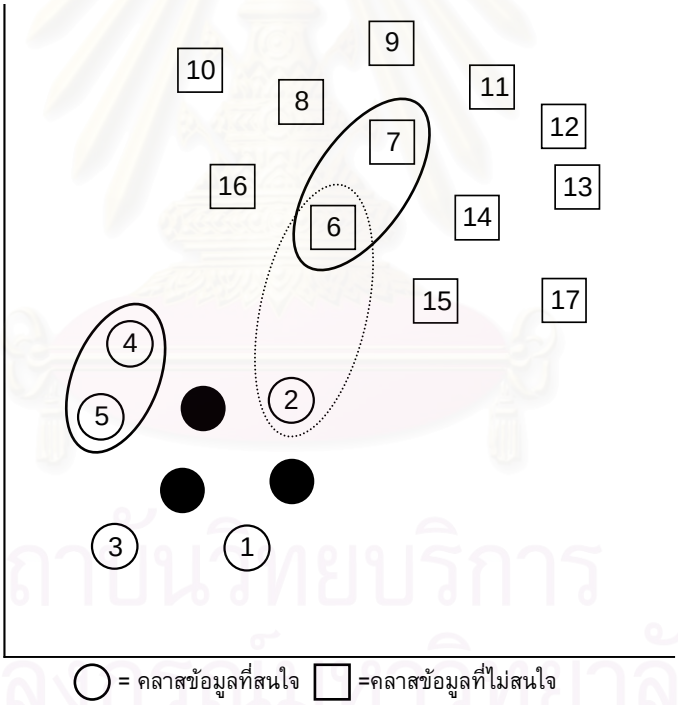
รูปที่ 3.5 ข้อมูลในแต่ละเซตหลังผ่านการย้ายข้อมูล

การฝึกสอนด้วยตนเองจะดำเนินการซ้ำไปเรื่อย ๆ จนกระทั่งไม่มีข้อมูลในเซตข้อมูลที่ไม่ทราบคลาสเหลืออยู่ ในส่วนของการคำนวณหาเกณฑ์หยุด และการสร้างตัวจำแนกจะอธิบายอย่างละเอียดในหัวข้อ 3.1.3

3.1.3 การคำนวณหาเกณฑ์หยุดของตัวจำแนกแบบสองคลาส

การฝึกสอนด้วยตนเองนั้น จะทำการเพิ่มจำนวนข้อมูลในเซตข้อมูลที่ทราบคลาส แต่หากทำการฝึกสอนด้วยจำนวนรอบที่มากเกินไป อาจรวมข้อมูลที่ไม่เหมาะสมเข้าไปในเซตข้อมูลที่ทราบคลาสด้วย จึงจำเป็นต้องมีการคำนวณหาเกณฑ์หยุดสำหรับการเลือกข้อมูลเพื่อสร้างตัวจำแนกคลาสให้อยู่ในตำแหน่งที่เหมาะสม

งานวิจัยนี้จะเสนอวิธีการเลือกเกณฑ์หยุดแบบใหม่ โดยมีแนวคิดที่ว่าข้อมูลที่อยู่ในคลาสเดียวกันควรมีค่าระยะทางน้อย และข้อมูลที่อยู่ต่างคลาสกันควรมีค่าระยะทางมาก และเมื่อพิจารณาถึงลำดับของการย้ายข้อมูลขณะทำการฝึกสอนแล้ว ช่วงที่ผลต่างค่าระยะทางของข้อมูลที่มีลำดับการย้ายข้อมูลที่ติดกันขณะทำการฝึกสอนมีค่ามาก แสดงว่าข้อมูลที่ไม่อยู่ในคลาสที่เราสนใจเพิ่มเข้ามาในเซตข้อมูลที่สนใจแล้ว ดังนั้น บริเวณที่มีผลต่างค่าระยะทางสูงจึงควรได้รับการพิจารณาเป็นเกณฑ์หยุด



รูปที่ 3.6 ระยะทางระหว่างข้อมูลภายในคลาสเดียวกันและต่างคลาสกัน

จากรูปที่ 3.6 สัญลักษณ์วงกลมคือ ข้อมูลในคลาสที่สนใจ สัญลักษณ์สี่เหลี่ยมคือ ข้อมูลในคลาสที่ไม่สนใจ สัญลักษณ์วงกลมทึบคือ ข้อมูลที่ทราบคลาสขณะเริ่มทำการฝึกสอน ข้อมูลที่อยู่ภายในวงรีเส้นทึบคือ ตัวอย่างข้อมูลที่มีคลาสเหมือนกัน ซึ่งมีค่าระยะทางน้อยระหว่างข้อมูล ส่วนข้อมูลที่อยู่ภายในวงรีจุดไข่ปลาคือ ตัวอย่างข้อมูลที่อยู่ต่างคลาสกัน ซึ่ง

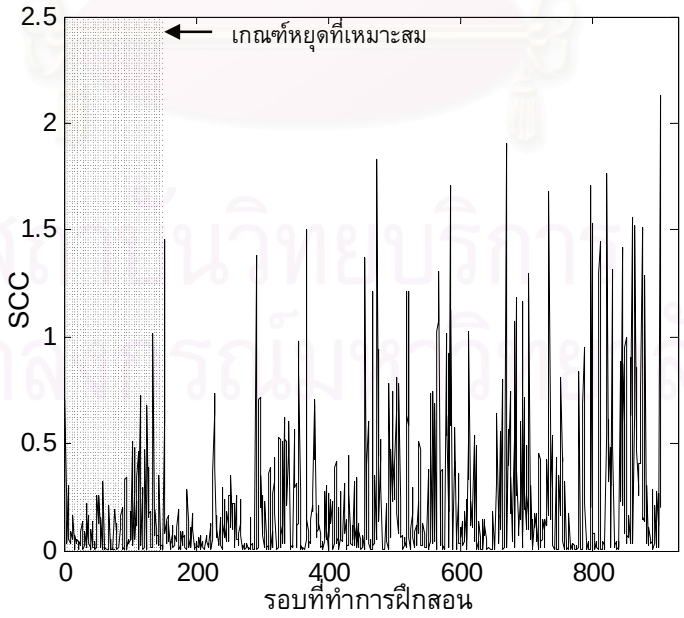
มีค่าระยะทางมากระหว่างข้อมูล ตัวเลขที่อยู่ภายในสัญลักษณ์วงกลมและสี่เหลี่ยมคือ ลำดับข้อมูลที่ย้ายเซตในแต่ละรอบของการฝึกสอน โดยข้อมูลในรูปที่ 3.6 มีการกระจายตัวของข้อมูลที่ใกล้เคียงกัน ระหว่างข้อมูลในคลาสที่สนใจและไม่สนใจ

จากแนวคิดเบื้องต้นของการหาเกณฑ์หยุดจากผลต่างค่าระยะทางของข้อมูลที่มีลำดับการย้ายข้อมูลที่ติดกัน ค่าเกณฑ์หยุด (Stopping Criterion Confidence) หรือ *SCC* จะคำนวณค่าคะแนนสำหรับการเลือกเกณฑ์หยุดขณะทำการฝึกสอนด้วยตนเองจากสมการที่ (3.1)

$$SCC(i) = \left| MovingDist(i) - MovingDist(i-1) \right| \tag{3.1}$$

โดยที่ *i* คือรอบของการฝึกสอน ค่า *SCC*(*i*) คือค่าคะแนนที่บอกว่ารอบการฝึกสอนที่ *i* มีความเหมาะสมที่จะใช้เป็นเกณฑ์หยุดหรือไม่ ค่า *MovingDist*(*i*) คือค่าระยะทางที่มีค่าน้อยที่สุดระหว่างคลาสที่สนใจและ คลาสที่ไม่สนใจ ในรอบการฝึกสอนที่ *i*

เหตุผลที่ต้องใส่ค่าสัมบูรณ์กับผลต่างของค่า *MovingDist* ในสมการที่ (3.1) เป็นเพราะว่าวิธีการหาเกณฑ์หยุดของงานวิจัยชิ้นนี้จะไม่สนใจว่าค่าผลต่างที่มีค่าสูง (ที่คำนวณจากค่าสัมบูรณ์) นั้นเกิดจากค่าระยะทางที่มีการเปลี่ยนแปลงมากขึ้น หรือเกิดจากค่าระยะทางที่มีการเปลี่ยนแปลงน้อยลง และการนำค่า *SCC* ที่คำนวณได้จากทุก ๆ รอบของการฝึกสอนที่บันทึกได้จากการคำนวณด้วยสมการที่ (3.1) ไปวาดกราฟ จะเป็นดังรูปที่ 3.7



รูปที่ 3.7 ค่า *SCC* ของทุกรอบของการฝึกสอนด้วยตนเอง

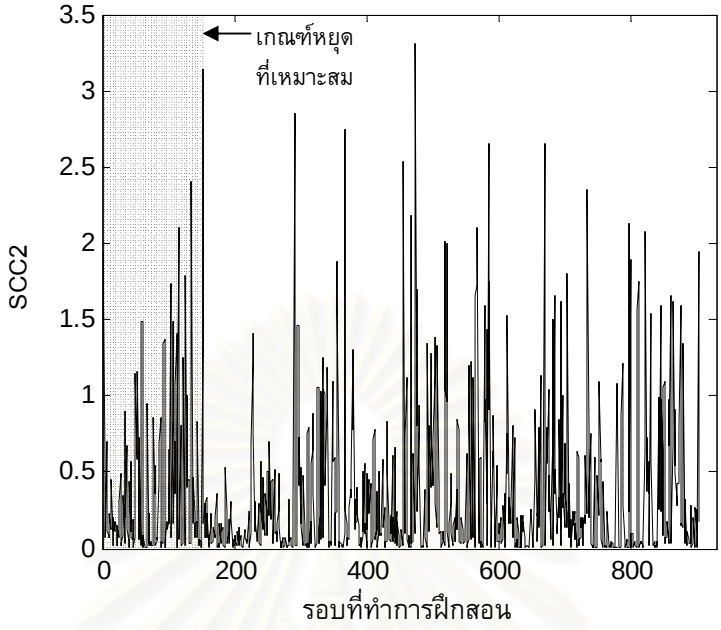
กราฟในรูปที่ 3.7 เป็นกราฟของค่า *SCC* ที่บันทึกไว้จากการฝึกสอนข้อมูลอนุกรมเวลาของลายมือ (รายละเอียดของข้อมูลอยู่ในภาคผนวก ก) โดยตำแหน่งที่ค่า *SCC* มีค่ามากในรูปที่ 3.7 เป็นตำแหน่งที่ควรพิจารณาเป็นเกณฑ์หยุด ซึ่งมีอยู่หลายตำแหน่ง แต่ตำแหน่งที่เป็นเกณฑ์หยุดที่เหมาะสมที่สุดอยู่ในรอบที่ 151 ของการฝึกสอน ข้อมูลรอบที่ทำการฝึกสอนก่อนพบเกณฑ์หยุดแสดงบริเวณเส้นแบ่งของพื้นหลังสีเทาและสีขาวในรูปที่ 3.7 แต่การที่ตำแหน่งของรอบการฝึกสอนอื่น ๆ มีค่า *SCC* สูงกว่ารอบที่ 151 แสดงให้เห็นว่าการนำค่าสัมบูรณ์ของผลต่างค่าระยะทางของข้อมูลที่มีลำดับการย้ายข้อมูลที่ติดกันมาใช้สำหรับการคำนวณหาเกณฑ์หยุดเพียงอย่างเดียวยังไม่เพียงพอต่อการหาบริเวณที่เป็นเกณฑ์หยุดที่เหมาะสมได้ สมการที่ (3.1) จึงได้รับการคำนวณเพิ่มเติมด้วยการนำค่าการกระจายตัวของข้อมูลในรอบของการฝึกสอนตั้งแต่ตำแหน่งแรกจนถึงตำแหน่งปัจจุบันของการฝึกสอน โดยค่าการกระจายตัวของข้อมูลนี้ดัดแปลงมาจากสมการการคำนวณค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) ซึ่งคำนวณจากสมการที่ (3.2) เป็นดังสมการที่ (3.3)

$$Std(x_1..x_n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.2}$$

โดยที่ค่า i คือตำแหน่งของข้อมูล ค่า x_i ข้อมูลใด ๆ ที่อยู่ในตำแหน่งที่ i ค่า n คือจำนวนข้อมูลทั้งหมด และค่า $\bar{x}$ คือค่าเฉลี่ยเลขคณิต โดยฟังก์ชันนี้จะคำนวณการกระจายตัวของค่ะระยะทาง โดยมีข้อมูลเข้าคือ ค่าระยะทางที่ได้จากการฝึกสอนตั้งแต่รอบแรกจนถึงรอบปัจจุบันของการฝึกสอน

$$SCC2(i) = \frac{|MovingDist(i) - MovingDist(i-1)|}{Std(MovingDist(1)..Movingdist(i))} \tag{3.3}$$

การนำผลต่างค่าระยะทางของข้อมูลที่มีลำดับการย้ายข้อมูลที่ติดกันไปหารด้วยค่าเบี่ยงเบนมาตรฐาน จะทำให้ค่าคะแนนของ *SCC2* ตำแหน่งที่มีค่ามากมีความแม่นยำในการใช้เป็นเกณฑ์หยุดมากขึ้น และเมื่อนำ *SCC2* ทุก ๆ รอบของการฝึกสอนที่บันทึกได้จากการคำนวณด้วยสมการที่ (3.3) ไปวาดกราฟ รูปร่างของกราฟที่ได้จะเป็นดังรูปที่ 3.8



รูปที่ 3.8 ค่า *SCC2* ของทุกรอบของการฝึกสอนด้วยตนเอง

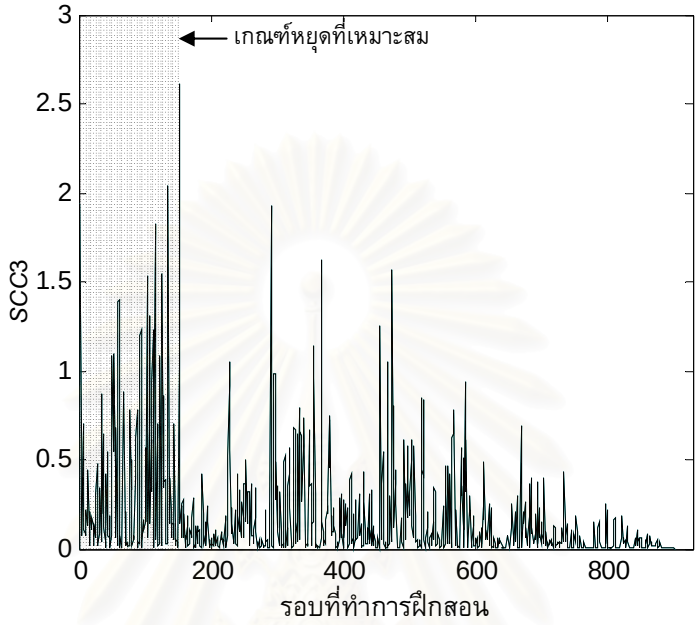
จากรูปที่ 3.8 แสดงให้เห็นว่าตำแหน่งที่ค่า *SCC2* ที่มีค่ามากมีความแม่นยำในการใช้เป็นเกณฑ์หยุดมากขึ้น แต่ในรอบที่ 452 ของการฝึกสอนยังคงเป็นตำแหน่งที่มีค่า *SCC2* มากที่สุด ซึ่งไม่ตรงกับตำแหน่งที่เหมาะสมจะเป็นเกณฑ์หยุดมากที่สุด คือตำแหน่งเส้นแบ่งของพื้นหลังสีเทาและสีขาวในรูปที่ 3.8

สมการที่ (3.3) จึงได้รับการคำนวณเพิ่มเติมด้วยการนำไปเพิ่มค่าน้ำหนักให้กับค่า *SCC2* โดยรอบการฝึกสอนเริ่มต้นจะมีค่าน้ำหนักมากที่สุด และจะมีค่าน้ำหนักลดลงเรื่อย ๆ เมื่อรอบของการฝึกสอนมากขึ้น แสดงดังสมการที่ (3.4)

$$SCC3(i) = \frac{\left|MovingDist(i) - MovingDist(i-1)\right|}{Std(MovingDist(1)..Movingdist(i))} \times \frac{InitialUnknown - (i-1)}{InitialUnknown} \qquad (3.4)$$

โดยที่ค่า *InitialUnknown* คือจำนวนข้อมูลของเซตข้อมูลที่ไม่ทราบคลาสเมื่อเริ่มทำการฝึกสอนด้วยตนเอง โดยค่านี้จะคงที่ตั้งแต่เริ่มทำการฝึกสอนจนกระทั่งการฝึกสอนสิ้นสุดลง

เมื่อนำ *SCC*3 จากทุก ๆ รอบของการฝึกสอนที่บันทึกได้จากการคำนวณด้วยสมการที่ (3.4) ไปวาดกราฟ รูปร่างของกราฟที่ได้จะเป็นดังรูปที่ 3.9



รูปที่ 3.9 ค่า *SCC*3 ของทุกรอบของการฝึกสอนด้วยตนเอง

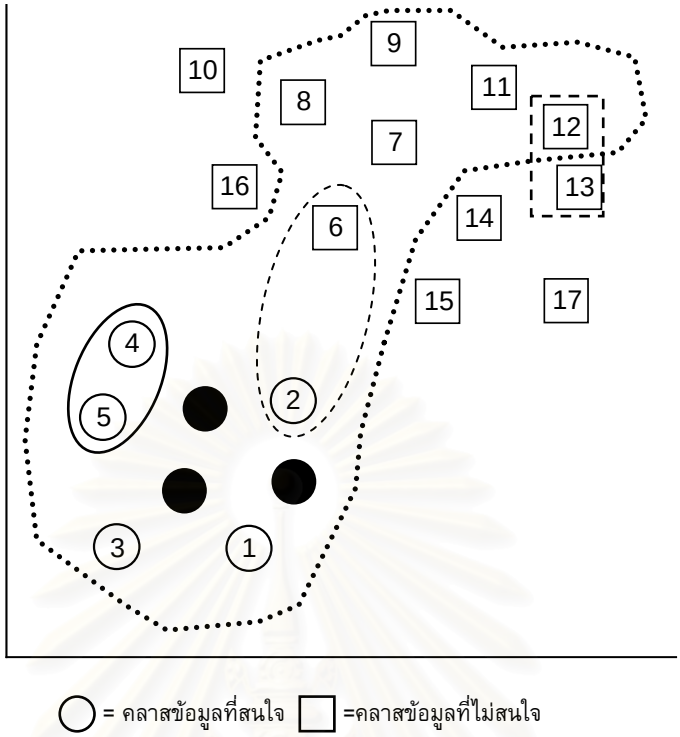
จากรูปที่ 3.9 แสดงให้เห็นว่าตำแหน่งที่ค่า *SCC3* ที่มีค่ามากที่สุดในรอบที่ 151 ของการฝึกสอนตรงกับบริเวณที่เป็นเกณฑ์หยุดที่เหมาะสม คือตำแหน่งเส้นแบ่งของพื้นหลังสีเทาและสีขาวในรูปที่ 3.9 ดังนั้นวิธีการหาเกณฑ์หยุดสำหรับตัวจำแนกคลาสอนุกรมเวลาแบบสองคลาสจึงคำนวณได้จากสมการที่ (3.4)

หากคำนวณหาเกณฑ์หยุดด้วยวิธีที่นำเสนอจากสมการที่ (3.4) จะได้ค่าจากการคำนวณดังตารางที่ 3.2 ซึ่งจะเห็นได้ว่าการหาเกณฑ์หยุดด้วยวิธีที่เสนอจะพบเกณฑ์หยุดในรอบการสอนที่ 6 ที่ค่า *SCC*3 มีค่ามากที่สุด ซึ่งเป็นการเลือกเกณฑ์หยุดในตำแหน่งที่มีความเหมาะสมเมื่อดูจากรูปที่ 3.10 แต่การหาเกณฑ์หยุดด้วยวิธีของ Wei และ Keogh ซึ่งใช้ความเปลี่ยนแปลงที่มากที่สุดของค่าระยะทางที่ลดน้อยลงในการหาเกณฑ์หยุด ซึ่งจะพบเกณฑ์หยุดในรอบการสอนที่ 13 แสดงในสี่เหลี่ยมประของรูปที่ 3.10 โดยตัวจำแนกที่ได้ประกอบด้วยข้อมูลจากคลาสที่ไม่สนใจเป็นจำนวนมาก

ตารางที่ 3.2 การคำนวณหาค่า *Stopping Criterion Confidence* (*SCC*3)

รอบการฝึกสอน	ค่าระยะทาง	ผลต่างค่าระยะทางที่มีรอบการฝึกสอนที่ติดกัน	ค่าระยะทางที่น้อยที่สุด	*SCC*3
1	1.0820	N/A	1.082	N/A
2	1.100	0.0180	1.082	1.3310
3	1.1930	0.1110	1.082	1.3774
4	1.3450	0.1520	1.082	1.0408
5	1.0770	0.2680	1.077	1.7994
6	2.4738	1.3968	1.077	1.8052
7	1.4422	1.0316	1.077	1.3372
8	1.2950	0.1472	1.077	0.1869
9	1.3120	0.0170	1.077	0.0207
10	1.3270	0.0150	1.077	0.0172
11	1.3450	0.0180	1.077	0.0191
12	1.1280	0.2170	1.077	0.2035
13	0.7430	0.3850	0.7430	0.2852
14	1.3520	0.6090	0.7430	0.3754
15	1.3115	0.0405	0.7430	0.0194
16	1.3625	0.0510	0.7430	0.0169
17	1.4239	0.0614	0.7430	0.0105

จากตารางที่ 3.2 ค่าระยะทาง คือค่าระยะทางที่มีค่าน้อยที่สุดระหว่างคู่ข้อมูลในคลาสข้อมูลที่สนใจและไม่สนใจ ผลต่างค่าระยะทาง คือผลต่างค่าระยะทางที่มีรอบการฝึกสอนที่ติดกันนั้นคำนวณจากสมการที่ (3.1) ค่าระยะทางที่น้อยที่สุดคือ ค่าระยะทางที่มีค่าน้อยที่สุดที่เคยเกิดขึ้นตั้งแต่มีการฝึกสอนมาทุกรอบ *SCC3* คือค่าคะแนนสำหรับการเลือกเกณฑ์หยุดขณะทำการฝึกสอนด้วยตนเอง ซึ่งคำนวณได้จากสมการที่ (3.4)



รูปที่ 3.10 ค่าระยะทางระหว่างข้อมูลภายในคลาสเดียวกันและต่างคลาสกัน

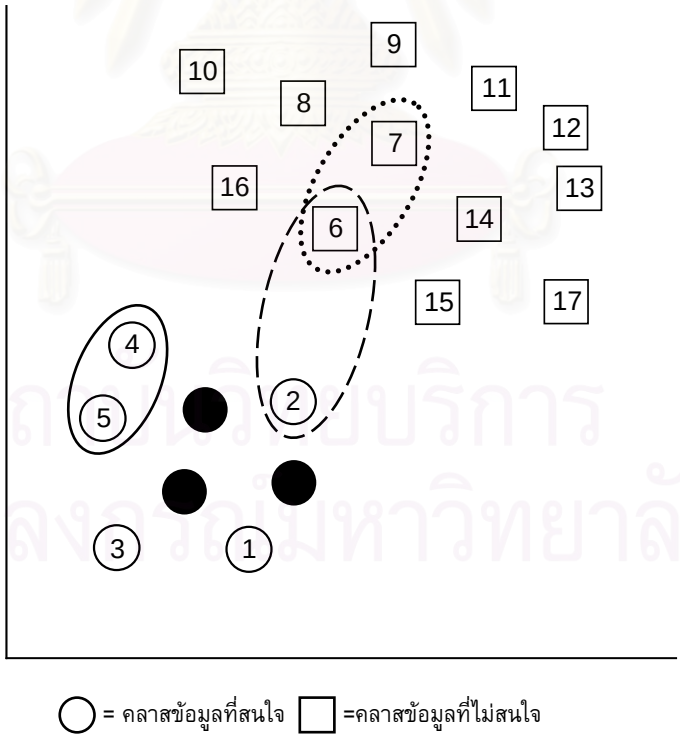
จากรูปที่ 3.10 ข้อมูลในกรอบวงรีเส้นทึบแสดงคู่ข้อมูลที่ให้ค่าระยะทางที่มีค่าน้อยที่สุดในรอบการฝึกสอนที่ 5 ข้อมูลในกรอบวงรีเส้นประแสดงคู่ข้อมูลที่ให้ค่าระยะทางที่มีค่าน้อยที่สุดในรอบการฝึกสอนที่ 6 ข้อมูลในกรอบเส้นประแสดงคู่ข้อมูลที่ทำให้เกิดค่าระยะทางที่มีค่าน้อยที่สุดตั้งแต่เกิดการฝึกสอน ซึ่งเกิดในรอบการฝึกสอนที่ 13 โดยข้อมูลที่นำมาสร้างเป็นตัวจำแนกคือข้อมูลในรอบการฝึกสอนที่ 1-12 แสดงในกรอบไข่ปลาในรูปที่ 3.10 ซึ่งตัวจำแนกประกอบด้วยข้อมูลที่อยู่คลาสที่ไม่สนใจจำนวนหลายตัว

เมื่อพิจารณาถึงจำนวนกลุ่มของข้อมูล (Cluster) ภายในคลาสข้อมูลที่ทราบคลาสและสนใจ และภายในคลาสข้อมูลที่ไม่สนใจแล้ว พบว่าการเลือกเกณฑ์หยุดด้วยวิธีที่เสนอมีข้อกำหนดเริ่มต้นก่อนการฝึกสอน โดยข้อมูลที่เริ่มต้นทำการฝึกสอนต้องเป็นกลุ่มที่มาจากคลาสข้อมูลที่ทราบคลาสและสนใจ และต้องไม่มาจากกลุ่มคลาสข้อมูลที่ไม่สนใจ ในกรณีที่ข้อมูลที่ทราบคลาสและสนใจมีจำนวนหลายกลุ่ม ถ้าขณะเริ่มต้นฝึกสอน กลุ่มข้อมูลใด ๆ ที่ไม่มีสมาชิกที่เป็นข้อมูลที่ทราบคลาสแล้ว จะมีผลทำให้กลุ่มข้อมูลนั้นไม่ได้รับการเรียนรู้ ทำให้ตัวจำแนกที่ทำการเรียนรู้ได้ไม่สามารถจำแนกคลาสให้ข้อมูลภายในกลุ่มนั้นได้ เพราะลักษณะของวิธีการฝึกสอนด้วยตนเองนั้น จะเลือกข้อมูลตัวที่มีความคล้ายคลึงที่สุดกับคลาสข้อมูลที่ทราบคลาสและสนใจขณะทำการฝึกสอน

สำหรับวิธีการสร้างตัวจำแนกของงานวิจัยนี้ ถ้าเราสามารถเลือกเกณฑ์หยุดได้อย่างเหมาะสมแล้ว จะสามารถทำการสร้างตัวจำแนกคลาสจากข้อมูลที่มีการย้ายเซตก่อนลำดับที่พบเกณฑ์หยุดตั้งแต่ตำแหน่งแรกจนถึงตำแหน่งก่อนพบเกณฑ์หยุดสองตำแหน่งมาใช้เป็นตัวจำแนก

อย่างไรก็ตาม วิธีการสร้างตัวจำแนกคลาสของ Wei และ Keogh นั้นสร้างตัวจำแนกคลาสจากข้อมูลที่มีการย้ายเซตก่อนลำดับที่พบเกณฑ์หยุดตั้งแต่ตำแหน่งแรกจนถึงตำแหน่งก่อนพบเกณฑ์หยุดเพียงตำแหน่งเดียวมาใช้เป็นตัวจำแนก

โดยเหตุผลที่เลือกข้อมูลที่มีการย้ายเซตก่อนลำดับที่พบเกณฑ์หยุดตั้งแต่ตำแหน่งแรกจนถึงตำแหน่งก่อนพบเกณฑ์หยุดสองตำแหน่งมาเป็นตัวจำแนกเป็นเพราะลักษณะการหาเกณฑ์หยุดด้วยวิธีที่นำเสนอ จากสมการที่ (3.4) ในหน้า 27 โดยในส่วนที่คำนวณค่าผลต่างของระยะทางในรอบใด ๆ คือส่วนของ $\left|MovingDist(i) - MovingDist(i-1)\right|$ ซึ่งส่วนนี้จะไม่สนใจว่าสาเหตุที่ค่าผลต่างที่มีค่ามาก (มีค่าเป็นบวกเท่านั้น เพราะการคำนวณจากการใส่ค่าสัมบูรณ์) อาจเกิดจากกรณีที่ค่าระยะทางมีการเปลี่ยนแปลงมากขึ้น หรือกรณีที่ค่าระยะทางมีการเปลี่ยนแปลงน้อยลง



รูปที่ 3.11 ตำแหน่งที่พบค่าระยะทางระหว่างข้อมูลในรอบการฝึกสอนที่ 5-7

จากรูปที่ 3.11 ข้อมูลในกรอบวงรีเส้นทึบแสดงคู่ข้อมูลที่ให้ค่าระยะทางที่มีค่าน้อยที่สุดในรอบการฝึกสอนที่ 5 ข้อมูลในกรอบวงรีเส้นประแสดงคู่ข้อมูลที่ให้ค่าระยะทางที่มีค่าน้อยที่สุดในรอบการฝึกสอนที่ 6 ข้อมูลในกรอบวงรีจุดไข่ปลาแสดงคู่ข้อมูลที่ให้ค่าระยะทางที่มีค่าน้อยที่สุดในรอบการฝึกสอนที่ 7

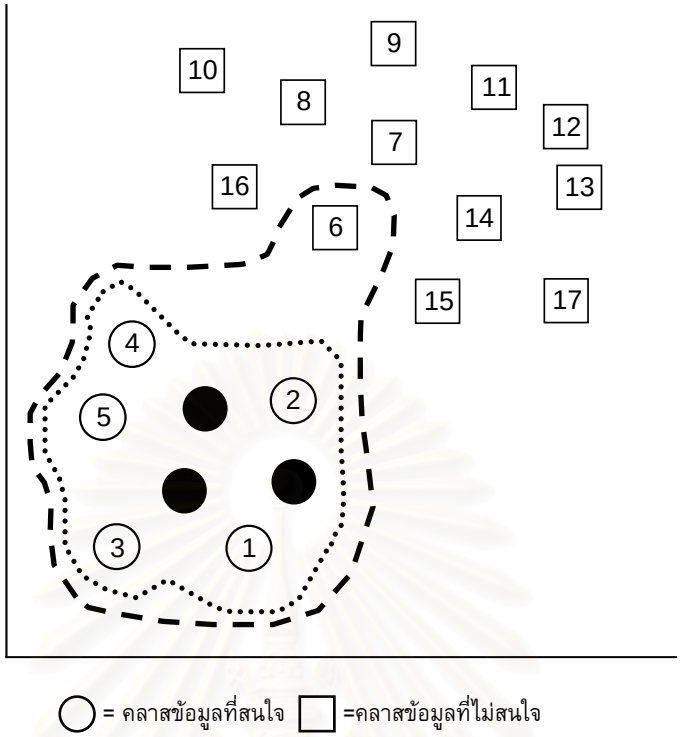
จากสมการที่ (3.4) ส่วนหนึ่งของสมการที่คำนวณค่าผลต่างของค่าระยะทางในรอบใด ๆ คือ $|MovingDist(i) - MovingDist(i-1)|$ ซึ่งคำนวณจากค่าระยะทางของข้อมูลที่ย้ายคลาสในรอบปัจจุบันลบกับค่าระยะทางของข้อมูลที่ย้ายคลาสในรอบก่อนหน้า ตัวอย่างเช่น ค่าผลต่างของค่าระยะทางในรอบที่ 7 จะคำนวณจากค่าระยะทางของข้อมูลที่มีลำดับการย้ายข้อมูลในรอบการฝึกสอนที่ 7 ลบกับค่าระยะทางของข้อมูลที่มีลำดับการย้ายในรอบก่อนหน้าคือรอบการฝึกสอนที่ 6

เมื่อสังเกตจากลักษณะการคำนวณผลต่างของค่าระยะทาง จากรูปที่ 3.11 เมื่อพิจารณาจากการคำนวณที่ต้องใส่ค่าสัมบูรณ์

- ค่าผลต่างที่เกิดจากกรณีที่ค่าระยะทางมีการเปลี่ยนแปลงมากขึ้นที่มีโอกาสได้รับเลือกเกณฑ์หยุดมากที่สุด คือรอบการฝึกสอนที่ 5 และรอบการฝึกสอนที่ 6 ในรอบการฝึกสอนที่ 5 ที่ค่าระยะห่างของคู่ข้อมูลในบริเวณวงรีเส้นทึบซึ่งมีค่าระยะทางต่ำ เมื่อเทียบกับรอบการฝึกสอนที่ 6 ที่ค่าระยะห่างของคู่ข้อมูลในบริเวณวงรีจุดไข่ปลาซึ่งมีค่าระยะทางสูง
- ค่าผลต่างที่เกิดจากกรณีที่ค่าระยะทางมีการเปลี่ยนแปลงน้อยลงที่มีโอกาสได้รับเลือกเกณฑ์หยุดมากที่สุด คือรอบการฝึกสอนที่ 6 ที่ค่าระยะห่างของคู่ข้อมูลในบริเวณวงรีจุดไข่ปลาซึ่งมีค่าระยะทางสูง เมื่อเทียบกับรอบการฝึกสอนที่ 7 ที่ค่าระยะห่างของคู่ข้อมูลในบริเวณวงรีเส้นประ ซึ่งมีค่าระยะทางต่ำ

สมมติให้ค่า *SCC* ของรอบการฝึกสอนที่ 6 มีค่าเท่ากับรอบการฝึกสอนที่ 7 และมีโอกาสเลือกเป็นเกณฑ์หยุดที่เหมาะสมได้เท่ากัน เพราะโอกาสเกิดค่า *SCC* ที่มีค่าสูงที่สุดอยู่ระหว่างสองตำแหน่งนี้

ถ้าเลือกเกณฑ์หยุดในรอบการฝึกสอนที่ 6 และรอบการฝึกสอนที่ 7 และข้อมูลที่จะนำมาสร้างเป็นตัวจำแนก ใช้ข้อมูลตั้งแต่ตำแหน่งแรกจนถึงตำแหน่งก่อนพบเกณฑ์หยุดเพียงตำแหน่งเดียว กลุ่มข้อมูลที่จะนำไปสร้างเป็นตัวจำแนก แสดงในกรอบไข่ปลาและกรอบเส้นประในรูปที่ 3.12

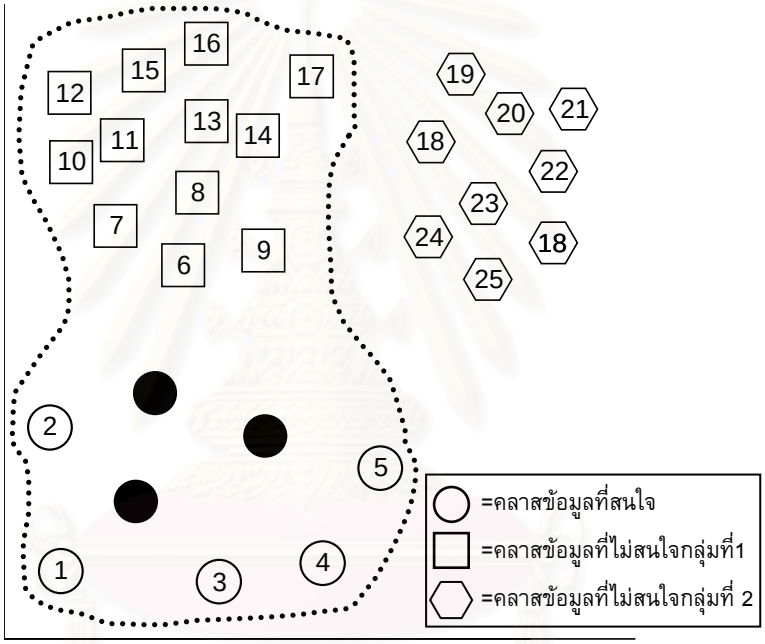


รูปที่ 3.12 ข้อมูลที่นำมาสร้างเป็นตัวจำแนกเมื่อเลือกเกณฑ์หยุดในรอบการฝึกสอนที่ 6 และ 7

รูปที่ 3.12 บริเวณกรอบไข่ปลาแสดงข้อมูลที่นำมาสร้างตัวจำแนกเมื่อเลือกเกณฑ์หยุดในรอบการฝึกสอนที่ 6 คือข้อมูลที่มีลำดับการย้ายคลาสลำดับที่ 1-5 และบริเวณกรอบเส้นประแสดงข้อมูลที่นำมาสร้างตัวจำแนกเมื่อเลือกเกณฑ์หยุดในรอบการฝึกสอนที่ 7 คือข้อมูลที่มีลำดับการย้ายคลาสลำดับที่ 1-6

จากรูปที่ 3.12 การเลือกเกณฑ์หยุดในรอบการฝึกสอนที่ 6 ซึ่งค่าผลต่างที่เกิดจากการที่ค่าระยะทางมีการเปลี่ยนแปลงมาก จะเห็นว่าข้อมูลที่นำไปใช้สร้างตัวจำแนกทั้ง 5 ตัว แสดงภายในตำแหน่งกรอบวงรีไข่ปลาในรูปที่ 3.12 เป็นข้อมูลที่อยู่ในคลาสที่สนใจทั้งหมด ซึ่งเป็นเรื่องที่สมเหตุสมผล แต่ถ้าเลือกเกณฑ์หยุดในรอบการฝึกสอนที่ 7 ซึ่งค่าผลต่างที่เกิดจากการที่ค่าระยะทางที่เปลี่ยนแปลงน้อย จะเห็นว่าข้อมูลที่นำไปใช้สร้างตัวจำแนกทั้ง 6 ตัว แสดงภายในตำแหน่งกรอบวงรีในรูปที่ 3.12 มีเพียงข้อมูลที่มีลำดับการย้ายคลาสในรอบที่ 6 ตัวเดียวเท่านั้นที่ไม่อยู่ในคลาสที่สนใจ แต่ข้อมูลที่ตัวนี้เพียงตัวเดียวอาจส่งผลต่อความแม่นยำในการจำแนกของตัวจำแนกคลาสในภายหลังได้

จากตัวอย่างที่ผ่านมา แม้ว่ากรณีที่ค่าผลต่างที่เกิดจากการที่ค่าระยะทางเปลี่ยนแปลงมากขึ้น และการใช้ข้อมูลที่มีการย้ายเซตก่อนลำดับที่พบเกณฑ์หยุดตั้งแต่ตำแหน่งแรกสำหรับการสร้างตัวจำแนกนั้น ไม่ได้มีข้อมูลในคลาสที่ไม่สนใจปะปนเข้ามาเลย แต่การคิดค่าผลต่างเพียงเฉพาะกรณีที่ค่าผลต่างอันเกิดจากการที่ค่าระยะทางเปลี่ยนแปลงมากขึ้นเพียงกรณีเดียว โดยไม่คิดในกรณีที่ผลต่างระยะทางที่เกิดจากค่าที่เปลี่ยนแปลงน้อยลงด้วย อาจทำให้เกิดปัญหาในการสร้างตัวจำแนก เช่น ในรูปที่ 3.13 ที่ข้อมูลในคลาสที่สนใจมีจำนวน 1 กลุ่ม แทนด้วยสัญลักษณ์วงกลม ข้อมูลที่อยู่ในคลาสที่ไม่สนใจมีจำนวน 2 กลุ่ม โดยกลุ่มแรกแทนด้วยสัญลักษณ์สี่เหลี่ยม กลุ่มที่สองแทนด้วยสัญลักษณ์หกเหลี่ยม



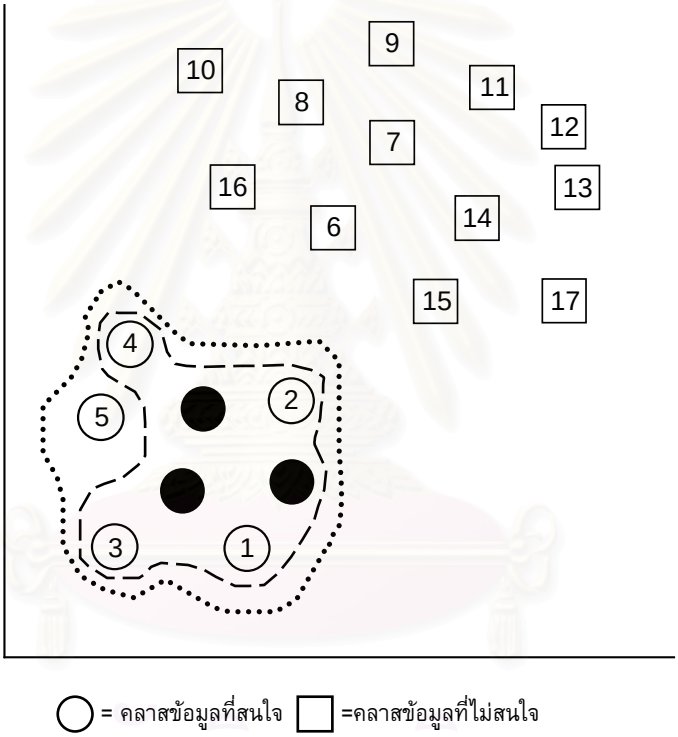
รูปที่ 3.13 การพบเกณฑ์หยุดในรอบการฝึกสอนที่18

รูปที่ 3.13 บริเวณกรอบจุดไข่ปลาแสดงข้อมูลที่นำมาสร้างตัวจำแนกเมื่อเลือกเกณฑ์หยุดในรอบการฝึกสอนที่ 18 คือข้อมูลที่มีลำดับการย้ายคลาสตั้งแต่ลำดับแรกจนถึงลำดับที่ 17

การคำนวณหาเกณฑ์หยุดในกรณีที่ค่าผลต่างที่เกิดจากการที่ค่าระยะทางเปลี่ยนแปลงมากขึ้นเพียงกรณีเดียวนั้น ไม่สามารถเลือกเกณฑ์ให้อยู่ในตำแหน่งที่เหมาะสมได้ ดังกรณีในรูปที่ 3.13 เพราะค่าผลต่างของระยะทางที่เปลี่ยนแปลงมากขึ้นนั้น เกิดจากค่าระยะทางในลำดับการฝึกสอนรอบปัจจุบันที่มีค่าระยะทางมาก ลบกับค่าระยะทางในลำดับการฝึกสอนรอบก่อนหน้าที่มีค่าระยะทางน้อย และจากรูปที่ 3.13 ที่พบเกณฑ์หยุดในรอบการ

ฝึกสอนที่ 18 ซึ่งข้อมูลที่จะนำไปสร้างตัวจำแนก (ข้อมูลที่มีลำดับการย้ายคลาสที่ 1-17) นั้นรวมข้อมูลที่ผิดคลาสเข้ามามากมาย ดังนั้นการคำนวณหาเกณฑ์หยุดด้วยการค่าหาผลต่างที่เกิดจากการที่ค่าระยะทางเปลี่ยนแปลงมากขึ้น จึงไม่เพียงพอต่อการได้มาซึ่งตัวจำแนกที่ดีในบางกรณี

จากเหตุผลที่กล่าวมาทั้งหมด แสดงให้เห็นการเลือกเกณฑ์หยุดควรพิจารณาลำดับการย้ายข้อมูลลำดับที่ให้ค่า *SCC*3 สูงที่สุด ไม่ว่าค่าผลต่างนั้นจะเกิดจากการที่ค่าระยะทางเปลี่ยนแปลงน้อยลง หรือเปลี่ยนแปลงมากขึ้นก็ตาม โดยจากรูปที่ 3.14 ข้อมูลที่จะนำไปสร้างตัวจำแนกจะเป็นข้อมูลที่มีการย้ายคลาสก่อนลำดับที่พบเกณฑ์หยุดตั้งแต่ตำแหน่งแรกจนถึงตำแหน่งก่อนพบเกณฑ์หยุดสองตำแหน่ง



รูปที่ 3.14 ข้อมูลที่นำมาสร้างเป็นตัวจำแนกเมื่อเลือกเกณฑ์หยุดในรอบการฝึกสอนที่ 6 และ 7

จากรูปที่ 3.14 กรอบจุดไข่ปลาแสดงข้อมูลที่นำไปสร้างตัวจำแนกกรณีที่พบเกณฑ์หยุดในรอบการฝึกสอนที่ 7 ซึ่งมีข้อมูลลำดับที่ 1-5 เป็นส่วนหนึ่งของตัวจำแนกที่ได้ กรอบเส้นประแสดงข้อมูลที่นำไปสร้างเป็นตัวจำแนกกรณีที่พบเกณฑ์หยุดในรอบการฝึกสอนที่ 6 ซึ่งมีข้อมูลลำดับที่ 1-4 เป็นตำแหน่งของตัวจำแนกที่สร้างได้

การที่พบเกณฑ์หยุดในรอบการฝึกสอนที่ 6 นั้นทำให้ข้อมูลที่มีลำดับการย้ายคลาสลำดับที่ 1-4 ได้รับการเลือกสำหรับการสร้างตัวจำแนกแสดงในกรอบเส้นประในรูปที่ 3.14 ซึ่งในกรณีนี้แม้ว่าตัวจำแนกจะขาดข้อมูลในรอบการฝึกสอนที่ 5 ไป 1 ตัว แต่หากเปรียบเทียบการขาดข้อมูลในคลาสที่สนใจไปหนึ่งตัว กับการเพิ่มข้อมูลในคลาสที่ไม่สนใจ 1 ตัวเป็นสมาชิกในตัวจำแนกแล้ว ข้อมูลที่ผิดจากการเพิ่ม 1 ตัวอาจส่งผลต่อการจำแนกข้อมูลที่อยู่ในคลาสที่เราไม่สนใจเข้ามาได้อีกมากมาย ซึ่งจะเป็นผลเสียต่อตัวจำแนกที่สร้างได้มากกว่า ดังนั้นในงานวิจัยชิ้นนี้จึงเลือกข้อมูลสำหรับการสร้างตัวจำแนกคลาส จากข้อมูลที่มีการย้ายเซตก่อนลำดับที่พบเกณฑ์หยุดตั้งแต่ตำแหน่งแรกจนถึงตำแหน่งก่อนพบเกณฑ์หยุดสองตำแหน่ง

3.1.4 การนำตัวจำแนกคลาสที่สร้างได้ไปใช้สำหรับการจำแนกคลาสข้อมูล

การจำแนกคลาสข้อมูลจะใช้การจำแนกคลาสแบบเลือกข้อมูลข้างเคียงที่ใกล้ที่สุดหนึ่งตัว (One Nearest Neighbor Classification) โดยวิธีนี้จะหาค่าระยะทางที่มีค่าน้อยที่สุดระหว่างตัวจำแนก (สร้างจากขั้นตอนที่ 3.3) กับข้อมูลที่นำมาทดสอบ และใช้ค่าขีดแบ่ง (*Threshold*) เพื่อตัดสินว่าข้อมูลที่ไม่ทราบคลาส มีโอกาสมากที่จะเป็นคลาสเดียวกับตัวจำแนกหรือไม่ ขั้นตอนวิธีจำแนกคลาสของตัวจำแนกแบบสองคลาสเป็นดังรูปที่ 3.15

ขั้นตอนวิธี : การจำแนกคลาสข้อมูลแบบสองคลาส

1: แบ่งข้อมูลออกเป็น 2 กลุ่ม คือกลุ่มของตัวจำแนกคลาส และกลุ่มข้อมูลทดสอบ
2: สำหรับสมาชิกในกลุ่มข้อมูลทดสอบแต่ละตัว ให้เลือกข้อมูลในตัวจำแนกที่ใกล้กับตัวเองมากที่สุด (มีค่าระยะทางที่น้อยที่สุด)
3: นำค่าระยะทางระหว่างข้อมูลในขั้นตอนที่ 2 ไปเปรียบเทียบกับค่าขีดแบ่ง
   - ถ้าค่าระยะทางน้อยกว่าหรือเท่ากับค่าขีดแบ่ง ให้คลาสของข้อมูลที่เลือกได้ในขั้นตอนที่สองเป็นคลาสเดียวกับตัวจำแนก
   - ถ้าค่าระยะทางมากกว่าค่าขีดแบ่ง ข้อมูลตัวนั้นจะได้รับการจำแนกเป็นคลาสคนละคลาสกับตัวจำแนก
4: ทำซ้ำในขั้นตอนที่ 2-3 จนกระทั่งจำแนกคลาสให้ข้อมูลในกลุ่มทดสอบหมดทุกตัว

รูปที่ 3.15 ขั้นตอนวิธีการการจำแนกคลาสข้อมูลแบบสองคลาส

ค่าขีดแบ่งสามารถคำนวณได้จาก ผลรวมของค่าเฉลี่ยของค่าระยะทางที่มีค่าน้อยที่สุดของแต่ละข้อมูลแต่ละตัวที่ได้รับการจำแนกคลาส และค่าเบี่ยงเบนมาตรฐานของค่าระยะทางของข้อมูลทุกตัวที่ได้รับการจำแนกคลาส ซึ่งสามารถคำนวณได้ดังสมการที่ (3.5)

$$Threshold = (\frac{1}{C}\sum_{i=1}^{C} Cdistance(i)) + Std(Cdistance(1)..Cdistance(C)) \quad (3.5)$$

จากสมการที่ (3.5) ค่า *Threshold* คือค่าขีดแบ่ง ค่า *C* คือจำนวนข้อมูลในกลุ่มของตัวจำแนกคลาส ค่า Cdistance(i) คือค่าระยะทางที่มีค่าน้อยที่สุดระหว่างข้อมูลตัวที่ i และข้อมูลตัวอื่น ๆ ตัวที่ใกล้กับตัวเองมากที่สุด และค่า *Std*(*Cdistance*(1)..*Cdistance*(*n*)) คือฟังก์ชันคำนวณการกระจายตัวของค่าระยะทาง โดยฟังก์ชันนี้จะมีข้อมูลเข้าคือค่า *Cdistance* ของข้อมูลตัวแรกจนถึงตัวสุดท้าย

3.1.5 การวัดประสิทธิภาพของตัวจำแนกคลาสแบบสองคลาส

งานวิจัยนี้ได้นำเสนอตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบสองคลาส และเพื่อแสดงให้เห็นประสิทธิภาพของตัวจำแนก จึงจำเป็นต้องทดสอบและวัดประสิทธิภาพในการทำงาน โดยวิธีการใช้ในการวัดประสิทธิภาพมีดังต่อไปนี้

(1) ค่าความเที่ยง (*Precision*) [18] คือการหาอัตราส่วนของการจำแนกคลาสเป็นคลาสที่สนใจได้อย่างถูกต้องต่อจำนวนครั้งที่จำแนกคลาสเป็นคลาสที่สนใจ ดังสมการที่ (3.6)

$$Precision = \frac{C}{A_c} \quad (3.6)$$

โดยที่ *Precision* คือ ค่าความเที่ยง *C* คือ จำนวนครั้งที่จำแนกคลาสเป็นคลาสที่สนใจได้อย่างถูกต้อง A_c คือ จำนวนครั้งที่จำแนกคลาสเป็นคลาสที่สนใจ

ตัวอย่างการหาค่าความเที่ยงของการจำแนกคลาสแบบสองคลาส เช่น จำนวนข้อมูลทดสอบได้รับการจำแนกคลาสเป็นคลาสที่สนใจทั้งหมด 10 ครั้ง และในจำนวนนั้นสามารถจำแนกคลาสข้อมูลได้ถูกต้อง 8 ตัว ดังนั้นค่าความเที่ยงมีค่าเท่ากับ $\frac{8}{10}$ หรือเท่ากับ 0.8

(2) ค่าความระลึก (*Recall*) [18] คือ การหาอัตราส่วนของการจำแนกคลาสเป็นคลาสที่สนใจได้อย่างถูกต้อง ต่อจำนวนข้อมูลในคลาสที่สนใจทั้งหมด ดังสมการที่ (3.7)

$$Recall = \frac{C}{A_i} \tag{3.7}$$

โดยที่ *Recall* คือ ค่าความระลึก *C* คือ จำนวนครั้งที่จำแนกคลาสเป็นคลาสที่สนใจได้อย่างถูกต้อง A_i คือ จำนวนข้อมูลในคลาสที่สนใจทั้งหมด ซึ่งรวมข้อมูลขณะเริ่มต้นทำการฝึกสอนด้วย

ตัวอย่างการหาค่าความระลึกของการจำแนกคลาสแบบสองคลาส เช่น จำนวนข้อมูลในคลาสที่สนใจที่นำมาทดสอบทั้งหมดมี 8 ตัว และข้อมูลทดสอบได้รับการจำแนกเป็นคลาสข้อมูลที่สนใจทั้งหมด 10 ตัว แต่จำแนกได้ถูกต้อง 8 ตัว และไม่ถูกต้อง 2 ตัว ดังนั้นค่าความระลึกมีค่าเท่ากับ $\frac{8}{8}$ หรือเท่ากับ 1

(3) ค่ามาตรวัด F (*F-Measure*) [18] เป็นค่ามาตรวัดที่แสดงความสัมพันธ์ระหว่างค่าความเที่ยง (*Precision*) และ ค่าความระลึก (*Recall*) ดังสมการที่ (3.8)

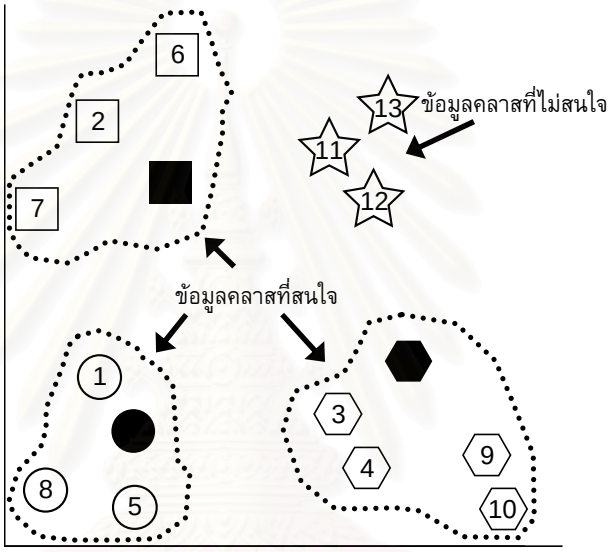
$$F\text{-}Measure = \frac{(2 \times Precision \times Recall)}{Precision + Recall} \tag{3.8}$$

เนื่องจากการวัดประสิทธิภาพของการตัวจำแนกคลาสที่จะต้องดูทั้งค่าความเที่ยงและค่าความระลึกซึ่งเป็นการไม่สะดวก โดยเฉพาะอย่างยิ่งในกรณีที่ค่าความเที่ยงและค่าความระลึกที่ได้จากตัวจำแนกทั้งสองตัวมีค่าไม่เท่ากัน ทำให้การเปรียบเทียบประสิทธิภาพของตัวจำแนกเป็นเรื่องยาก

ดังนั้นงานวิจัยนี้จึงใช้ค่ามาตรวัด F ในการวัดประสิทธิภาพของตัวจำแนกคลาส ตัวอย่างการหาค่ามาตรวัด F ของตัวจำแนกคลาสแบบสองคลาส เช่น จำนวนข้อมูลในคลาสที่สนใจของข้อมูลที่นำมาทดสอบทั้งหมดมี 8 ตัว และข้อมูลที่ได้รับการจำแนกคลาสเป็นคลาสที่สนใจมีทั้งหมด 10 ตัว โดยสามารถจำแนกคลาสได้ถูกต้อง 8 ตัว ดังนั้นค่ามาตรวัด F มีค่าเท่ากับ $\frac{(2 \times 0.8 \times 1)}{0.8 + 1}$ หรือเท่ากับ 0.88

3.2 วิธีดำเนินงานวิจัยเพิ่มเติม

นอกเหนือจากวิธีดำเนินงานวิจัยในส่วนหลักของวิทยานิพนธ์ฉบับนี้แล้ว งานวิจัยนี้ยังได้ทำการทดลองเพิ่มเติมนอกเหนือจากส่วนหลักที่กล่าวไปในหัวข้อที่ 3.1 โดยจะทำการดัดแปลงกระบวนวิธีสำหรับการสร้างตัวจำแนกคลาสจากเดิม ที่จำแนกคลาสได้สองคลาส ให้สามารถจำแนกคลาสได้หลายคลาส ตัวจำแนกนี้จะมีข้อมูลที่สนใจขณะเริ่มต้นทำการฝึกสอน เป็นข้อมูลจากทุก ๆ คลาส และเมื่อทำการเรียนรู้สำเร็จจะได้ตัวจำแนกที่สามารถจำแนกคลาสได้หลายคลาส ลักษณะของกลุ่มข้อมูลเป็นดังรูปที่ 3.16



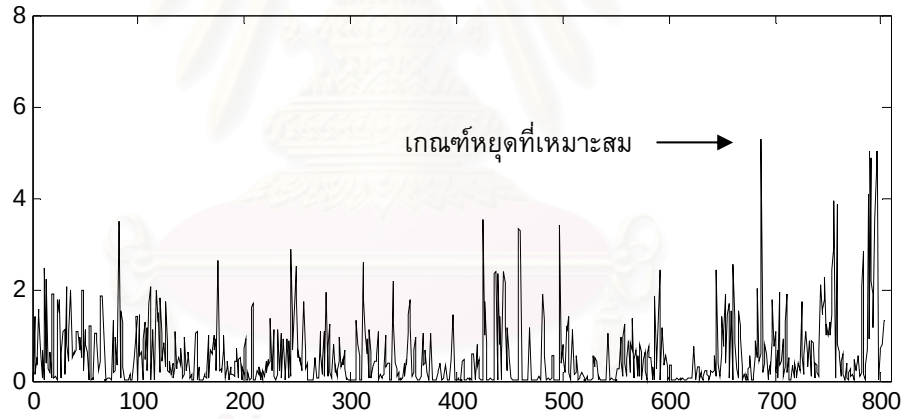
รูปที่ 3.16 ข้อมูลในคลาสที่สนใจที่ประกอบด้วยข้อมูลสามคลาส และข้อมูลคลาสที่ไม่สนใจ

จากรูปที่ 3.16 สัญลักษณ์ที่มีพื้นหลังเป็นสีดำคือ ข้อมูลที่ทราบคลาสเมื่อเริ่มทำการฝึกสอน สัญลักษณ์วงกลมคือข้อมูลที่อยู่ในคลาสแรก สัญลักษณ์หกเหลี่ยมคือข้อมูลที่อยู่ในคลาสที่สอง สัญลักษณ์สี่เหลี่ยมคือข้อมูลที่อยู่ในคลาสที่สาม ซึ่งทั้งสามคลาสนี้เป็นข้อมูลที่สนใจ และในส่วนของสัญลักษณ์ดาวคือข้อมูลคลาสที่ไม่สนใจ ตัวเลขในสัญลักษณ์ต่าง ๆ คือลำดับข้อมูลที่ได้รับการย้ายคลาสในแต่ละรอบของการฝึกสอน

การฝึกสอนตัวจำแนกแบบหลายคลาสจะเพิ่มจำนวนข้อมูลในแต่ละคลาสให้มากขึ้น และจะพบเกณฑ์หยุดในรอบการฝึกสอนที่ผลต่างของค่าระยะทางของข้อมูลที่ย้ายคลาสที่มีลำดับที่ติดกันมีค่ามาก ซึ่งตรงกับรอบการฝึกสอนที่ 11 โดยข้อมูลที่ได้รับการย้ายคลาสในรอบการฝึกสอนที่ 11 เป็นข้อมูลคลาสที่ไม่สนใจตัวแรกที่ได้รับการย้ายคลาสพอดี โดยข้อมูลที่จะนำไปสร้างตัวจำแนกคลาสคือข้อมูลที่มีลำดับการย้ายคลาสก่อนตำแหน่งที่พบเกณฑ์หยุด 1-10 ไปสร้างเป็นตัวจำแนกในแต่ละคลาส ดังข้อมูลในกรอบไข่ปลาแต่ละกรอบ

ขั้นตอนในการสร้างตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบหลายคลาสมีความคล้ายกับการสร้างตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบสองคลาสแต่ในบางขั้นตอนจะมีรายละเอียดปลีกย่อยที่แตกต่างกัน โดยในส่วนที่ปลีกย่อยมีคือในส่วนของการหาเกณฑ์หยุดการจำแนกคลาส และการวัดประสิทธิภาพของตัวจำแนกคลาส

การหาเกณฑ์หยุดสำหรับสร้างตัวจำแนกแบบหลายคลาสมีจุดประสงค์เพื่อพัฒนาตัวจำแนกคลาสให้สามารถจำแนกคลาสได้อย่างแม่นยำมากขึ้นเมื่อจำนวนข้อมูลที่ทราบคลาสตอนเริ่มต้นทำการฝึกสอนนั้นมีอยู่จำนวนไม่มาก โดยการฝึกสอนตัวจำแนกจะเลือกข้อมูลที่มีความใกล้เคียงกับข้อมูลที่ทราบคลาสที่อยู่ในแต่ละคลาสเข้ามาเป็นตัวจำแนก แต่วิธีการหาเกณฑ์หยุดจะใช้ค่า *SCC2* ที่คำนวณได้จากสมการที่ 3.3 ในหน้าที่ 25 ตำแหน่งที่พบเกณฑ์หยุดจะเป็นตำแหน่งที่ข้อมูลตัวใดตัวหนึ่งซึ่งไม่มีความใกล้เคียงกับข้อมูลในกลุ่มที่ทราบคลาสและสนใจ เริ่มย้ายคลาสเข้ามายังกลุ่มข้อมูลที่เราสนใจ และเมื่อทำการทดลองกับข้อมูลคลื่นหัวใจ และนำ *SCC2* ทุก ๆ รอบของการฝึกสอนที่บันทึกได้จากการคำนวณด้วยสมการที่ (3.3) ไปวาดกราฟ รูปร่างของกราฟที่ได้จะเป็นดังรูปที่ 3.17



รูปที่ 3.17 ค่า *SCC2* ของทุกรอบของการฝึกสอนด้วยตนเองที่คำนวณจากสมการที่ 3.3

รูปที่ 3.17 เป็นกราฟที่นำค่า *SCC2* มาใช้สำหรับการคำนวณหาเกณฑ์หยุดของการสร้างตัวจำแนกแบบหลายคลาส และจากกราฟพบว่าบริเวณที่พบเกณฑ์หยุดอยู่ในรอบการฝึกสอนที่ 678

เมื่อได้เกณฑ์หยุดที่เหมาะสมแล้ว เราจะนำข้อมูลที่มีลำดับการย้ายเซตตั้งแต่ตำแหน่งแรกจนถึงตำแหน่งก่อนพบเกณฑ์หยุดสองตำแหน่ง (อธิบายในหัวข้อที่ 3.1.3) มาใช้เป็นตัวจำแนก โดยจะจำแนกคลาสข้อมูลด้วยวิธีตัวการจำแนกคลาสแบบเลือกสมาชิกข้างเคียงที่ใกล้ที่สุด 1 ตัว ข้อมูลที่ไม่ทราบคลาสที่มีค่าระยะทางใกล้กับข้อมูลที่ทราบคลาสมากที่สุด จะมีโอกาสที่เป็นข้อมูลที่อยู่ในคลาสเดียวกัน ขั้นตอนวิธีจำแนกคลาสของตัวจำแนกแบบหลายคลาสเป็นดังรูปที่ 3.18

ขั้นตอนวิธี : การจำแนกคลาสข้อมูลแบบหลายคลาส

1: สำหรับสมาชิกในกลุ่มข้อมูลทดสอบแต่ละตัว ให้เลือกข้อมูลในตัวจำแนกที่ใกล้กับตัวเองมากที่สุด (มีค่าระยะทางที่น้อยที่สุด)
2: กำหนดคลาสให้ข้อมูลทดสอบ โดยให้มีคลาสเดียวกับข้อมูลในตัวจำแนกตัวที่มีค่าระยะทางน้อยที่สุด
3: ทำซ้ำขั้นตอนที่ 1-2 จนกระทั่งจำแนกคลาสให้ข้อมูลในกลุ่มทดสอบทุกตัว

รูปที่ 3.18 ขั้นตอนวิธีการการจำแนกคลาสข้อมูลแบบหลายคลาส

สำหรับตัวจำแนกแบบหลายคลาส ถ้าขณะเริ่มต้นฝึกสอน ข้อมูลที่สนใจไม่ได้ประกอบด้วยสมาชิกจากทุก ๆ คลาส ข้อมูลจากคลาสนั้นจะไม่ได้รับการเรียนรู้ ทำให้ตัวจำแนกจะจำแนกข้อมูลคลาสนั้นเป็นคลาสอื่น ซึ่งเป็นการจำแนกคลาสที่ผิด ดังนั้นข้อจำกัดของตัวจำแนกแบบหลายคลาสคือ ข้อมูลที่เริ่มต้นทำการฝึกสอนควรมีสมาชิกข้อมูลมาจากทุก ๆ คลาส

ในส่วนของการวัดประสิทธิภาพของตัวจำแนกแบบหลายคลาส ซึ่งจะทำการทดสอบและวัดประสิทธิภาพในการทำงาน โดยวิธีการที่นำมาใช้ในการหาค่าความแม่นยำ (*Accuracy*) ดังสมการที่ (3.9)

$$Accuracy = \frac{CorrectlyClassified}{Total\ \#\ of\ Instances\ Classified} \tag{3.9}$$

โดยที่ *Accuracy* คือ ค่าความแม่นยำของการจำแนกคลาส *CorrectlyClassified* คือ ข้อมูลที่ได้รับการจำแนกคลาสได้ถูกต้อง *Total # of Instances Classified* คือ จำนวนข้อมูลทดสอบ

ตัวอย่างการหาค่าความแม่นยำของการจำแนกคลาสแบบหลายคลาส เช่น จำนวนข้อมูลทดสอบที่ได้รับการจำแนกคลาสทั้งหมด 10 ตัว และจำแนกคลาสข้อมูลได้ถูกต้อง 8 ตัว ดังนั้นค่าความแม่นยำมีค่าเท่ากับ $\frac{8}{10}$ หรือเท่ากับ 0.8

บทที่ 4

การทดลองและผลการทดลอง

ในบทนี้จะกล่าวถึงวิธีการทดลองและผลการทดลองของงานวิจัยเรื่องการเพิ่มความแม่นยำให้กับการเลือกเกณฑ์หยุดสำหรับการสร้างตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบกึ่งมีผู้สอน ซึ่งเป็นการทดลองวัดค่าความแม่นยำของการจำแนกคลาสของตัวจำแนกที่สร้างจากวิธีการเลือกเกณฑ์หยุดที่งานวิจัยชิ้นนี้นำเสนอ ข้อมูลอนุกรมเวลาที่ใช้ในการทดลองมีจำนวน 10 ชุดข้อมูล ซึ่งมาจาก 2 แหล่ง คือ จาก University of California, Riverside's archive [19] และจากงานวิจัยชื่อ Semi-Supervised Time Series Classification ของ Wei และ Keogh [20]

จำนวนข้อมูลที่นำมาใช้ในงานวิจัยนี้มีทั้งหมด 10 ชุดข้อมูล แต่ละชุดข้อมูลมีจำนวนข้อมูล ความยาว และจำนวนคลาสภายในชุดข้อมูลที่แตกต่างกันไป โดยรายละเอียดของแต่ละชุดข้อมูล แสดงในตารางที่ 4.1 และในส่วนของรายละเอียดของแต่ละคลาสในแต่ละชุดข้อมูลอย่างละเอียดแสดงในภาคผนวก ก

ตารางที่ 4.1 ข้อมูลที่นำมาทำการทดลอง

ข้อมูล	จำนวนข้อมูล (ตัว)	ความยาว (มิติ)	หมายเลขคลาสและรายละเอียดของข้อมูลในคลาส	รูปแบบกลุ่ม (Cluster) ข้อมูลภายในคลาสที่ทำการนิยามไว้
คลื่นหัวใจ (ECG)	2,026	87	1.คลื่นหัวใจที่เต้นปกติ	หลายกลุ่ม
			2.คลื่นหัวใจที่เต้นผิดปกติ	หลายกลุ่ม
ลายมือ (Word spotting)	1,710	272	1.เขียนคำว่า the	หลายกลุ่ม
			2.เขียนคำอื่น ๆ	หลายกลุ่ม
โยคะ (Yoga)	612	428	1.ผู้เล่นเพศชาย	หลายกลุ่ม
			2.ผู้เล่นเพศหญิง	หลายกลุ่ม
ปืน (Gun)	247	152	1.บุคคลแรกทำท่าทางถือปืน	1 กลุ่ม
			2.บุคคลแรกไม่ได้ถือปืน	1 กลุ่ม
			3.บุคคลที่สองทำท่าทางถือปืน	1 กลุ่ม
			4.บุคคลที่สองไม่ได้ถือปืน	1 กลุ่ม
ถ้วยกาแฟ (Coffee cup)	56	286	1.ถ้วยเล็ก	1 กลุ่ม
			2.ถ้วยใหญ่	1 กลุ่ม

ตารางที่ 4.1 (ต่อ) ข้อมูลที่นำมาทำการทดลอง

ข้อมูล	จำนวนข้อมูล (ตัว)	ความยาว (มิติ)	หมายเลขคลาสและรายละเอียดของข้อมูลในคลาส	รูปแบบกลุ่ม (Cluster) ข้อมูลภายในคลาสที่ทำการนิยามไว้
น้ำมันมะกอก (Olive oil)	60	570	1.แบบที่ 1	1 กลุ่ม
			2.แบบที่ 2	1 กลุ่ม
			3.แบบที่ 3	1 กลุ่ม
			4.แบบที่ 4	1 กลุ่ม
ซีบีเอฟ (CBF)	930	128	1.กระบอก	หลายกลุ่ม
			2.ระฆัง	หลายกลุ่ม
			3.กรวย	หลายกลุ่ม
สองรูปแบบ (2 patterns)	5,000	128	1.รูปแบบลง-ลง	1 กลุ่ม
			2.รูปแบบขึ้น-ลง	1 กลุ่ม
			3.รูปแบบลง-ขึ้น	1 กลุ่ม
			4.รูปแบบขึ้น-ขึ้น	1 กลุ่ม
นิวเคลียร์เทรซ (Nuclear trace)	200	275	1.แบบที่ 1	1 กลุ่ม
			2.แบบที่ 2	1 กลุ่ม
			3.แบบที่ 3	1 กลุ่ม
			4.แบบที่ 4	1 กลุ่ม
สังเคราะห์ (Synthetic Control)	600	60	1.รูปแบบปกติ	หลายกลุ่ม
			2.รูปแบบวงกลม	1 กลุ่ม
			3.มีแนวโน้มเพิ่มขึ้น	1 กลุ่ม
			4.มีแนวโน้มลดลง	1 กลุ่ม
			5.เพิ่มระดับขึ้น	1 กลุ่ม
			6.ลดระดับลง	1 กลุ่ม

การนิยามจำนวนกลุ่มข้อมูลว่ามีหลายกลุ่มนั้นเป็นการนิยามจากงานวิจัยชิ้นอื่น โดย Wei และ Keogh [7] นิยามข้อมูลคลื่นหัวใจ ลายมือ และโยคะ ว่ามีกลุ่มข้อมูลภายในคลาสเดียวกันหลายกลุ่ม และสำหรับและ Keogh และคณะ [21] นิยามข้อมูลซีบีเอฟว่ามีกลุ่มข้อมูลภายในคลาสหลายกลุ่ม

ในหัวข้อต่อไปจะกล่าวถึงวิธีทำการทดลองในงานวิจัยชิ้นนี้ โดยจะแบ่งการทดลองเป็นสามส่วนหลัก ๆ คือ 1. การทดลองเพื่อวัดความสามารถในการจำแนกคลาสด้วยตัวจำแนกที่สร้างด้วยวิธีที่นำเสนอ โดยเปรียบเทียบผลการทดลองกับวิธีเลือกเกณฑ์หยุดของ Wei และ Keogh 2. การทดลองเพื่อวัดผลว่าขนาดของค่าเงื่อนไขบังคับโดยรวมที่แตกต่างกันนั้น จะส่งผลต่อความสามารถในการจำแนกคลาสของตัวจำแนกหรือไม่ และ 3. การทดลองเพื่อวัดผลว่าจำนวนข้อมูลขณะเริ่มต้นทำการฝึกสอนนั้น จะส่งผลต่อคุณภาพของตัวจำแนกหรือไม่

4.1 การทดลองเพื่อวัดความสามารถของตัวจำแนกที่สร้างด้วยวิธีเลือกเกณฑ์หยุดที่นำเสนอ

เพื่อที่จะประเมินความสามารถในการจำแนกคลาสของตัวจำแนกที่สร้างจากวิธีที่นำเสนอและตัวจำแนกที่สร้างจากวิธีของ Wei และ Keogh ว่าวิธีการใดจะให้ผลการจำแนกคลาสที่ดีกว่ากัน จึงเตรียมข้อมูลสำหรับการทดลอง โดยมีรายละเอียดของข้อมูลที่นำมาทำการทดลอง คลาสที่นำมาใช้ จำนวนข้อมูลสำหรับการฝึกสอน และจำนวนข้อมูลสำหรับการทดสอบ ดังตารางที่ 4.2 โดยการทดลองในหัวข้อนี้จะแบ่งการทดลองเป็น 3 การทดลองย่อย ๆ

ตารางที่ 4.2 จำนวนข้อมูลที่ใช้สำหรับการสร้างตัวจำแนกคลาสแบบสองคลาส

ข้อมูล	หมายเลขคลาสที่นำมาใช้	จำนวนข้อมูลที่ทราบคลาสเมื่อเริ่มทำการฝึกสอน	จำนวนข้อมูลที่ไม่ทราบคลาสที่ใช้ทำการฝึกสอน		จำนวนข้อมูลที่ใช้ทำการทดสอบ
			คลาสที่สนใจ	คลาสที่ไม่สนใจ	
คลื่นหัวใจ	1	10	208	602	1,216
ลายมือ	1	10	109	696	905
โยคะ	1	10	156	156	306
ปืน	3	3	27	98	125
ถ้วยกาแฟ	1	3	14	14	28
น้ำมันมะกอก	4	3	13	17	30
ซีบีเอฟ	1	10	155	310	465
สองรูปแบบ	1	3	271	729	4,000
นิวเคลียร์เทรซ	1	3	26	74	100
สังเคราะห์	2	3	50	250	300

จากตารางที่ 4.2 คลาสข้อมูลที่นำมาใช้คือ คลาสข้อมูลที่สนใจที่จะนำไปทำการสร้างตัวจำแนกในการทดลอง ข้อมูลที่ทราบคลาสเมื่อเริ่มทำการฝึกสอนคือข้อมูลจำนวนหนึ่งที่ได้รับการสุ่มจากข้อมูลในคลาสที่สนใจ โดยข้อมูลกลุ่มนี้จะนำมาใช้ทำการฝึกสอน จำนวนข้อมูลที่ทำการสุ่มนั้นพิจารณาจากจำนวนกลุ่มข้อมูลภายในคลาสข้อมูลที่สนใจ (ดูรายละเอียดของกลุ่มข้อมูลภายในตารางที่ 4.1) หากคลาสข้อมูลที่สนใจมีหลายกลุ่ม จะเริ่มต้นทำการสุ่มข้อมูลจำนวน 10 ตัวสำหรับการเริ่มทำการฝึกสอน และหากคลาสข้อมูลที่สนใจมีกลุ่มเดียว จะเริ่มต้นทำการสุ่มข้อมูลจำนวน 3 ตัวสำหรับการเริ่มทำการฝึกสอน

การทดลองแรกจะวัดค่าความแม่นยำของตัวจำแนกที่สร้างด้วยวิธีที่เสนอที่ใช้วิธีเลือกเกณฑ์หยุดที่เสนอร่วมกับการใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง ซึ่งกำหนดเงื่อนไขบังคับโดยรวมขนาด 5% สำหรับการฝึกสอนด้วยตนเอง ในส่วนของการเปรียบเทียบผลการทดลองนั้น จะเทียบผลกับวิธีการสร้างตัวจำแนกในงานวิจัยของ Wei และ Keogh รายละเอียดของข้อมูลที่นำมาทำการฝึกสอนและทำการทดสอบ แสดงในตารางที่ 4.2

ในส่วนของข้อมูลที่นำมาทำการฝึกสอนจะแบ่งเป็นคลาสที่สนใจ และคลาสที่ไม่สนใจ (คลาสที่จะนำมาสร้างเป็นตัวจำแนก) ข้อมูลในคลาสที่ไม่สนใจคือข้อมูลในคลาสอื่นๆ ที่ไม่เป็นสมาชิกในคลาสที่สนใจ การทดลองนี้ทำการทดลองชุดข้อมูลละ 30 รอบ และทำการบันทึกค่าความเที่ยง (Precision) และค่าความระลึก (Recall) เพื่อนำไปหาค่ามาตรวัด F (F-measure) ผลการทดลองที่ได้แสดงในตารางที่ 4.3

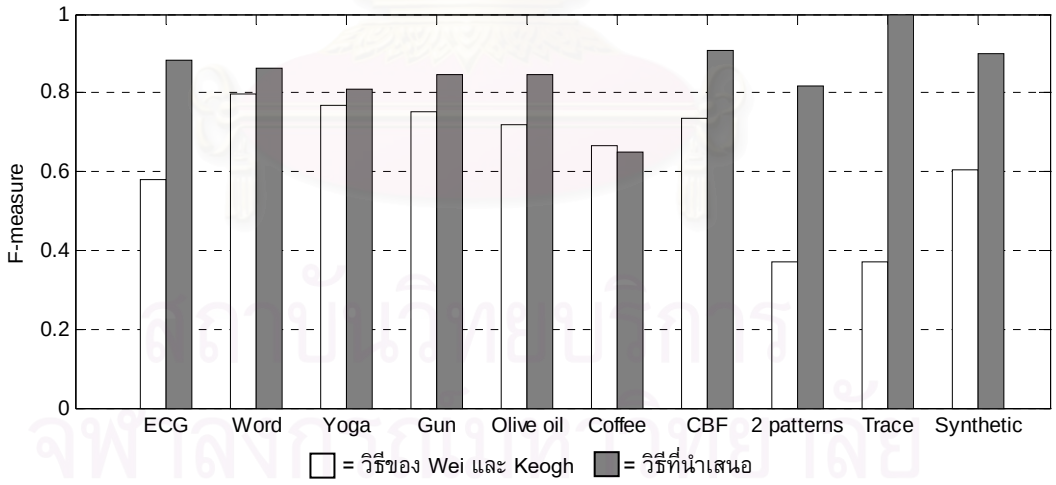
ตารางที่ 4.3 ผลการทดลองการวัดประสิทธิภาพของตัวจำแนกคลาสแบบสองคลาสที่ใช้เกณฑ์หยุดของ Wei และ Keogh และตัวจำแนกที่สร้างจากการใช้เกณฑ์หยุดของงานวิจัยชิ้นนี้

ข้อมูล	วิธีสร้างตัวจำแนกแบบกึ่งมีผู้สอนของ Wei และ Keogh			วิธีสร้างตัวจำแนกแบบกึ่งมีผู้สอนที่งานวิจัยชิ้นนี้นำเสนอ		
	ค่าความเที่ยง	ค่าความระลึก	ค่ามาตรวัด F	ค่าความเที่ยง	ค่าความระลึก	ค่ามาตรวัด F
คลื่นหัวใจ	0.4616	0.7736	0.5782	0.9668	0.8106	**0.8818**
ลายมือ	0.8480	0.7486	0.7952	0.9837	0.7713	**0.8646**
โยคะ	0.6346	0.9744	0.7686	0.7326	0.9081	**0.8110**
ปืน	0.9892	0.6089	0.7538	1.0000	0.7333	**0.8462**
น้ำมันมะกอก	0.6444	0.8167	0.7204	0.9451	0.7667	**0.8466**

ตารางที่ 4.3 (ต่อ) ผลการทดลองการวัดประสิทธิภาพของตัวจำแนกคลาสแบบสองคลาสที่ใช้เกณฑ์หยุดของ Wei และ Keogh และตัวจำแนกที่สร้างจากการใช้เกณฑ์หยุดของงานวิจัยชิ้นนี้

ข้อมูล	วิธีสร้างตัวจำแนกแบบกึ่งมีผู้สอนของ Wei และ Keogh			วิธีสร้างตัวจำแนกแบบกึ่งมีผู้สอนที่งานวิจัยชิ้นนี้นำเสนอ		
	ค่าความเที่ยง	ค่าความระลึก	ค่ามาตรวัด F	ค่าความเที่ยง	ค่าความระลึก	ค่ามาตรวัด F
ถ้วยกาแฟ	0.6220	0.7128	**0.6643**	0.6671	0.6308	0.6484
ซีบีเอฟ	0.7966	0.6843	0.7362	0.9848	0.8426	**0.9082**
สองรูปแบบ	0.7391	0.2493	0.3729	1.0000	0.6907	**0.8170**
นิวเคลียร์เทรซ	0.5027	0.2917	0.3691	1.0000	1.0000	**1.0000**
สังเคราะห์	0.9382	0.4467	0.6052	1.0000	0.8200	**0.9011**

ตารางที่ 4.3 แสดงรายละเอียดของการทดลองที่แสดงค่าความเที่ยง ค่าความระลึก และค่ามาตรวัด F ของตัวจำแนกแบบสองคลาสที่ทำการฝึกสอนตนเองด้วยวิธีที่แตกต่างกันทั้ง 10 ชุดข้อมูล ซึ่งเมื่อนำค่าจากตารางที่ 4.3 ไปวาดกราฟจะได้ดังรูปที่ 4.1



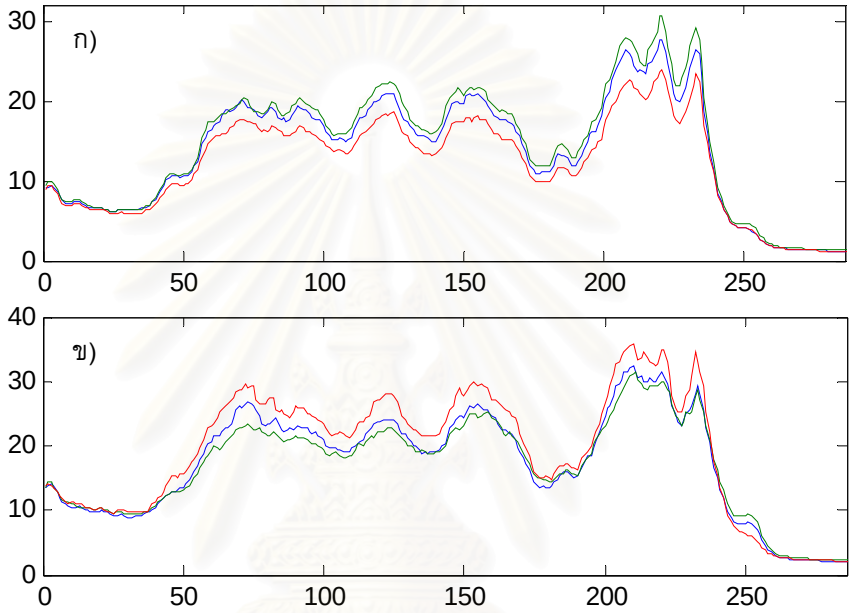
รูปที่ 4.1 ค่ามาตรวัด F ของตัวจำแนกแบบสองคลาสที่สร้างจากเกณฑ์หยุดที่นำเสนอและเกณฑ์หยุดแบบเดิม

ค่าเฉลี่ยของค่าความเที่ยงและค่าความระลึกคำนวณด้วยการใช้ค่าเฉลี่ยเลขคณิต (*mean*) ตัวอักษรหนาในตารางที่ 4.3 แสดงค่ามาตรวัด F ที่มีค่ามากกว่าอีกวิธีหนึ่ง ซึ่งจากการทดลองนี้จะเห็นว่าวิธีการที่นำเสนอนั้นสามารถทำการสร้างตัวจำแนกที่ให้ผลการจำแนกคลาสที่ดีกว่าวิธีการเดิมทุกชุดข้อมูล ยกเว้นชุดข้อมูลถ้วยกาแฟที่ค่ามาตรวัด F ของวิธีที่เสนอมีค่าน้อยกว่าวิธีการเดิมเพียงเล็กน้อยเท่านั้น ส่วนค่ามาตรวัด F ของชุดการทดลองอื่น ๆ ที่ได้จากตัวจำแนกที่สร้างจากวิธีที่นำเสนอนั้น มีค่ามากกว่า 0.8000 ทุกชุดข้อมูล โดยเฉพาะชุดข้อมูลนิวเคลียร์เทรซที่ได้ค่ามาตรวัด F 1.0000 ซึ่งมีค่ามากขึ้นกว่าการใช้วิธีเลือกเกณฑ์หยุดแบบเดิมเกือบสามเท่า แสดงให้เห็นว่าวิธีการที่เสนอสำหรับการสร้างตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบกึ่งมีผู้สอนนั้นมีประสิทธิภาพมากกว่าการใช้วิธีของ Wei และ Keogh โดยเฉพาะเมื่อดูจากค่าความเที่ยงเพียงอย่างเดียว ค่าความเที่ยงของวิธีการที่เสนอจากตารางที่ 4.3 ให้ค่ามากกว่าวิธีการของ Wei และ Keogh ในทุกกรณี

เมื่อวิเคราะห์ที่ค่าความเที่ยง และค่าความระลึกของชุดข้อมูลต่าง ๆ แล้วจากผลการทดลองที่ได้ จะได้ข้อสังเกตอยู่หลายข้อดังนี้

- หากผลการทดลองที่ได้มีค่าความระลึกที่มีค่าสูง และค่าความเที่ยงที่มีค่าต่ำ แสดงว่าวิธีการเลือกเกณฑ์หยุดนั้น ไม่สามารถหาเกณฑ์หยุดได้ในตำแหน่งที่ดีและพบเกณฑ์หยุดในรอบการฝึกสอนหลัง ๆ ซึ่งเป็นผลทำให้ตัวจำแนกที่สร้างได้มีขนาดใหญ่และรวมข้อมูลในคลาสที่ไม่สนใจปะปนเข้าไปเป็นจำนวนมาก ในกรณีนี้ไม่สามารถสรุปได้แน่ชัดว่าวิธีวัดระยะทางที่ใช้สำหรับการฝึกสอนมีความเหมาะสมสำหรับทำการฝึกสอนหรือไม่
- หากผลการทดลองที่ได้มีค่าความเที่ยงที่มีค่าสูง และค่าความระลึกที่มีค่าต่ำ แสดงว่าวิธีวัดระยะทางที่ใช้สำหรับการฝึกสอนมีความเหมาะสม (เห็นได้จากการเลือกข้อมูลที่ถูกต้องจำนวนมากเข้าสู่ตัวจำแนก) และในส่วนของวิธีการเลือกเกณฑ์หยุดนั้น พบเกณฑ์หยุดในรอบการฝึกสอนแรก ๆ ซึ่งเป็นผลทำให้ตัวจำแนกที่สร้างได้มีขนาดเล็ก ทำให้ไม่สามารถใช้แทนถึงข้อมูลที่ไม่เคยเห็นได้ทั้งหมด
- หากผลการทดลองที่ได้มีค่าความเที่ยงที่มีค่าสูง และค่าความระลึกที่มีค่าสูงเช่นเดียวกัน แสดงว่าวิธีการเลือกเกณฑ์หยุดและวิธีวัดระยะทางที่ใช้มีความเหมาะสมดีแล้ว จึงทำให้ตัวจำแนกที่สร้างได้มีความแม่นยำในการจำแนกคลาสข้อมูลได้อย่างถูกต้อง

และเมื่อพิจารณาจากข้อสังเกตที่สรุปได้กับชุดข้อมูลถ้วยกาแฟแล้ว สามารถตีความได้ว่าตัวจำแนกที่ได้จากวิธีที่เสนอนั้นสร้างตัวจำแนกที่มีขนาดเล็กกว่าตัวจำแนกที่สร้างจากวิธีของ Wei และ Keogh ซึ่งมีขนาดใหญ่กว่าและประกอบด้วยข้อมูลในคลาสที่ไม่สนใจปะปนอยู่จำนวนมากกว่า และเมื่อนำข้อมูลถ้วยกาแฟในแต่ละคลาสไปวาดกราฟจะมีรูปร่างดังรูปที่ 4.2



รูปที่ 4.2 ข้อมูลถ้วยกาแฟในแต่ละคลาส

รูปที่ 4.2 เป็นชุดข้อมูลถ้วยกาแฟในคลาสที่ 1 และคลาสที่ 2 และรูปร่างของข้อมูลอนุกรมเวลาทั้งสองคลาสมีความใกล้เคียงกันมาก ค่าต่าง ๆ บนข้อมูลอนุกรมเวลารูป 4.2 ก) และรูป 4.2 ข) ให้ค่าที่ไม่แตกต่างกันมากนักเมื่อพิจารณาจากตำแหน่งของแกน x ที่มีตำแหน่งตรงกันดังนั้นการใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงกับข้อมูลชนิดนี้จึงไม่ส่งผลมากนักต่อความสามารถของตัวจำแนกคลาส เพราะวิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงจะช่วยแก้ปัญหาในกรณีที่ข้อมูลมีการเลื่อนในแกน x

อย่างไรก็ตามค่ามาตรวัด F ของทั้งสองวิธียังให้ผลที่ไม่ดีนัก คือได้ค่ามาตรวัด F ที่ 0.6643 และ 0.6484 ตามลำดับ จึงอาจสรุปได้ว่าชุดข้อมูลนี้อาจมีการแปลงข้อมูลจากข้อมูลต้นแบบมาเป็นข้อมูลอนุกรมเวลาด้วยวิธีที่ไม่เหมาะสม หรือหากเลือกเพียงแค่ค่าคุณสมบัติบางช่วงของข้อมูลอนุกรมเวลามาใช้อาจทำให้ได้ผลที่ดีมากยิ่งขึ้น

การทดลองที่ผ่านมาจะเห็นได้ว่าการสร้างตัวจำแนกด้วยวิธีที่นำเสนอนั้นให้ค่ามาตรวัด F สูงกว่าการสร้างตัวจำแนกด้วยวิธีการเดิม โดยการสร้างตัวจำแนกด้วยวิธีที่นำเสนอนั้นมีการปรับปรุงค่าที่ต่างจากวิธีการเดิมสองค่า คือวิธีเลือกเกณฑ์หยุด และวิธีวัดระยะทาง จึงอาจทำให้เห็นภาพไม่ชัดเจนว่าการที่ค่ามาตรวัด F มีค่าสูงขึ้นนั้น เกิดขึ้นจากการปรับพารามิเตอร์ตัวใดตัวหนึ่งเพียงตัวเดียวหรือไม่

ในการทดลองที่สอง จะพิจารณาในการปรับปรุงค่าวิธีการเลือกเกณฑ์หยุดเพียงอย่างเดียว โดยจะใช้วิธีวัดระยะทางที่เหมือนกันในการทดลอง คือใช้วิธีวัดระยะทางแบบยุคลิดในการฝึกสอนด้วยตนเอง การทดลองนี้จะเปรียบเทียบค่ามาตรวัด F ที่ได้จากวิธีเลือกเกณฑ์หยุดที่นำเสนอ กับวิธีเลือกเกณฑ์หยุดของ Wei และ Keogh รายละเอียดของข้อมูลที่นำมาทำการฝึกสอนและทำการทดสอบ แสดงในตารางที่ 4.2 การทดลองนี้ทำการทดลองชุดข้อมูลละ 30 รอบ และทำการบันทึกค่าความเที่ยง (Precision) และค่าความระลึก (Recall) เพื่อนำไปหาค่ามาตรวัด F (F-measure) ผลการทดลองที่ได้แสดงในตารางที่ 4.4

ตารางที่ 4.4 ผลการทดลองการวัดประสิทธิภาพของตัวจำแนกคลาสแบบสองคลาสที่ใช้เกณฑ์หยุดของ Wei และ Keogh และตัวจำแนกที่สร้างจากการใช้เกณฑ์หยุดของงานวิจัยชิ้นนี้ โดยฝึกสอนตนเองด้วยการใช้วิธีวัดระยะทางแบบยุคลิด

ข้อมูล	เกณฑ์หยุดของ Wei และ Keogh (ยุคลิด)			เกณฑ์หยุดด้วยวิธีที่เสนอ (ยุคลิด)		
	ค่าความเที่ยง	ค่าความระลึก	ค่ามาตรวัด F	ค่าความเที่ยง	ค่าความระลึก	ค่ามาตรวัด F
คลื่นหัวใจ	0.4616	0.7736	0.5782	0.7316	0.7296	**0.7306**
ลายมือ	0.8480	0.7486	0.7952	0.9002	0.7517	**0.8193**
โยคะ	0.6346	0.9744	0.7686	0.7823	0.8944	**0.8346**
ปืน	0.9892	0.6089	0.7538	0.9975	0.7044	**0.8258**
น้ำมันมะกอก	0.6444	0.8167	0.7204	0.8963	0.7500	**0.8167**

ตารางที่ 4.4 (ต่อ) ผลการทดลองการวัดประสิทธิภาพของตัวจำแนกคลาสแบบสองคลาสที่ใช้เกณฑ์หยุดของ Wei และ Keogh และตัวจำแนกที่สร้างจากการใช้เกณฑ์หยุดของงานวิจัยชิ้นนี้ โดยฝึกสอนตนเองด้วยการใช้วิธีวัดระยะทางแบบยุคลิด

ข้อมูล	เกณฑ์หยุดของ Wei และ Keogh (ยุคลิด)			เกณฑ์หยุดด้วยวิธีที่เสนอ (ยุคลิด)		
	ค่าความเที่ยง	ค่าความระลึก	ค่ามาตรวัด F	ค่าความเที่ยง	ค่าความระลึก	ค่ามาตรวัด F
ถ้วยกาแฟ	0.6220	0.7128	**0.6643**	0.5908	0.7077	0.6440
ซีบีเอฟ	0.7966	0.6843	**0.7362**	0.9894	0.5222	0.6836
สองรูปแบบ	0.7391	0.2493	**0.3729**	0.5708	0.1883	0.2831
นิวเคลียร์เทรซ	0.5027	0.2917	0.3691	0.4807	0.5333	**0.5056**
สังเคราะห์	0.9382	0.4467	0.6052	1.0000	0.8467	**0.9170**

ตารางที่ 4.4 แสดงรายละเอียดของการทดลองที่แสดงค่าความเที่ยง ค่าความระลึก และค่ามาตรวัด F ของตัวจำแนกแบบสองคลาสที่ทำการฝึกสอนตนเองด้วยวิธีการเลือกเกณฑ์หยุดที่แตกต่างกัน และใช้วิธีวัดระยะทางแบบยุคลิดในการฝึกสอนด้วยตนเองของทั้ง 10 ชุดข้อมูล

ค่าเฉลี่ยของค่าความเที่ยงและค่าความระลึกคำนวณด้วยการใช้ค่าเฉลี่ยเลขคณิต (*mean*) ตัวอักษรหนาในตารางที่ 4.4 แสดงค่ามาตรวัด F ที่มีค่ามากกว่าอีกวิธีหนึ่ง ซึ่งจากการทดลองนี้จะเห็นว่าวิธีการที่นำเสนอนั้นสามารถทำการสร้างตัวจำแนกที่ให้ผลการจำแนกคลาสที่ดีกว่าวิธีการเดิมเป็นส่วนใหญ่ อย่างไรก็ตามการเปรียบเทียบผลของการทดลองนี้อาจเปรียบเทียบผลแบบละเอียดไม่ได้มากเพราะวิธีเลือกเกณฑ์หยุดที่เสนอนั้นจะให้ผลดีที่สุดเมื่อใช้กับวิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง

ในส่วนของการทดลองสุดท้ายคือ การทดลองที่ใช้วิธีการเลือกเกณฑ์หยุดที่แตกต่างกัน แต่ใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงที่กำหนดเงื่อนไขบังคับโดยรวมขนาด 5% ขณะทำการฝึกสอนด้วยตนเอง การทดลองนี้จะเปรียบเทียบระหว่างวิธีการเลือกเกณฑ์หยุดที่นำเสนอ กับวิธีการเลือกเกณฑ์หยุดของ Wei และ Keogh รายละเอียดของข้อมูลที่นำมาทำการฝึกสอนและทำการทดสอบ แสดงในตารางที่ 4.2 การทดลองนี้ทำการทดลองชุดข้อมูลละ 30 รอบ และทำการบันทึกค่าความเที่ยง (Precision) และค่าความระลึก (Recall) เพื่อนำไปหาค่ามาตรวัด F (F-measure) ผลการทดลองที่ได้แสดงในตารางที่ 4.5

ตารางที่ 4.5 ผลการทดลองการวัดประสิทธิภาพของการจำแนกคลาสแบบสองคลาสที่สร้างจากการใช้เกณฑ์หยุดของ Wei และ Keogh และตัวจำแนกที่สร้างจากการใช้เกณฑ์หยุดด้วยวิธีที่เสนอ โดยใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงที่กำหนดเงื่อนไขบังคับโดยรวมขนาด 5% ขณะทำการฝึกสอนด้วยตนเอง

ข้อมูล	เกณฑ์หยุดของ Wei และ Keogh (ไดนามิกไทม์วอร์ปปิง)			เกณฑ์หยุดด้วยวิธีที่เสนอ (ไดนามิกไทม์วอร์ปปิง)		
	ค่าความเที่ยง	ค่าความระลึก	ค่ามาตรวัด F	ค่าความเที่ยง	ค่าความระลึก	ค่ามาตรวัด F
คลื่นหัวใจ	0.6510	0.8218	0.7265	0.9668	0.8106	**0.8818**
ลายมือ	0.5754	0.9498	0.7167	0.9837	0.7713	**0.8646**
โยคะ	0.6794	0.9739	0.8004	0.7326	0.9081	**0.8110**
ปืน	0.9783	0.7489	**0.8484**	1.0000	0.7333	0.8462
น้ำมันมะกอก	0.4231	0.9167	0.5790	0.9451	0.7667	**0.8466**
ถ้วยกาแฟ	0.5193	0.7949	0.6282	0.6671	0.6308	**0.6484**
ซีบีเอฟ	0.9396	0.8753	0.9063	0.9848	0.8426	**0.9082**
สองรูปแบบ	0.4332	0.6389	0.5163	1.0000	0.6907	**0.817**
นิวเคลียร์เทรซ	0.3692	1.0000	0.5393	1.0000	1.0000	**1.0000**
สังเคราะห์	0.3472	1.0000	0.5155	1.0000	0.82	**0.9011**

ผลการทดลองในตารางที่ 4.5 แสดงให้เห็นว่าการวิธีการเลือกเกณฑ์หยุดที่งานวิจัยชิ้นนี้นำเสนอนั้น ให้ผลของค่ามาตรวัด F ที่สูงกว่าเกณฑ์หยุดที่เสนอโดย Wei และ Keogh เป็นส่วนใหญ่เหมือนกับสองการทดลองที่ผ่านมาในตารางที่ 4.3 และ 4.4

ค่ามาตรวัด F ที่ได้จากการทดลองในตารางที่ 4.5 ที่ได้จากวิธีเลือกเกณฑ์หยุดที่เสนอนั้นมีค่าสูงกว่าวิธีเลือกเกณฑ์หยุดของ Wei และ Keogh อย่างมากในบางชุดข้อมูล เช่น ชุดข้อมูลนิวเคลียร์เทรซ และข้อมูลสังเคราะห์ที่ได้ค่ามาตรวัด F ที่สูงกว่าเกือบ 2 เท่า

เมื่อเปรียบเทียบค่ามาตรวัด F ของวิธีเลือกเกณฑ์หยุดที่เสนอ ในตารางที่ 4.4 และตารางที่ 4.5 จะเห็นว่าวิธีเลือกเกณฑ์หยุดที่เสนอจะให้ผลการสร้างตัวจำแนกคลาสที่ดีเมื่อใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง

โดยสรุปแล้ว จากผลการทดลองทั้งหมดในหัวข้อนี้ จะเห็นได้ว่า การใช้วิธีการเลือกเกณฑ์หยุดที่ดีเพียงอย่างเดียวนั้น สามารถสร้างตัวจำแนกที่ดีได้ แต่วิธีวัดระยะทางก็สำคัญและส่งผลต่อการฝึกสอนด้วยตนเองเช่นเดียวกัน ดังนั้นการที่จะได้ค่าจากการจำแนกคลาสที่ดีที่สุดนั้นควรใช้ทั้งวิธีการเลือกเกณฑ์หยุดที่ดี และใช้วิธีวัดระยะทางที่มีความเหมาะสมประกอบกัน ซึ่งเห็นได้จากการความแม่นยำของการจำแนกคลาสด้วยวิธีที่นำเสนอที่ใช้วิธีเลือกเกณฑ์หยุดที่เสนอร่วมกับวิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง สามารถสร้างตัวจำแนกที่ดีกว่าการใช้วิธีเลือกเกณฑ์หยุดของ Wei และ Keogh ร่วมกับวิธีวัดระยะทางแบบยุคลิด

4.2 การทดลองเพื่อวัดความสามารถของตัวจำแนกที่ใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงที่ปรับเปลี่ยนขนาดของเงื่อนไขบังคับโดยรวม

จากการทดลองในหัวข้อที่ผ่านมาแสดงให้เห็นว่าการใช้วิธีเลือกเกณฑ์หยุดที่ดีและใช้วิธีวัดระยะทางที่เหมาะสมนั้นช่วยให้การฝึกสอนด้วยตนเองสามารถสร้างตัวจำแนกคลาสแบบกึ่งมีผู้สอนที่ดีได้ในการกำหนดเงื่อนไขบังคับโดยรวมขนาด 5% ในหัวข้อนี้จึงมีแนวคิดที่จะทำการทดลองเพื่อให้ทราบผลว่าการปรับขนาดของเงื่อนไขบังคับโดยรวมของวิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงนั้นส่งผลต่อความสามารถของตัวจำแนกที่สร้างได้จากการฝึกสอนด้วยตนเองหรือไม่

การทดลองชุดนี้จะทำการปรับค่าเงื่อนไขบังคับโดยรวมให้มีค่า 5% 10% และ 100% โดยจำนวนข้อมูลในคลาสที่สนใจที่สุ่มเมื่อเริ่มทำการฝึกสอนนั้น เป็นดังตารางที่ 4.2 การทดลองนี้จะทำการทดลองชุดข้อมูลละ 30 รอบ และนำค่าความเที่ยงและค่าความระลึกไปหาค่าเฉลี่ยเลขคณิต เพื่อนำไปคำนวณค่ามาตรวัด F ต่อไป ผลการทดลองที่ได้จากการทดลองชุดนี้แสดงในตารางที่ 4.6

ตารางที่ 4.6 แสดงให้เห็นว่าค่ามาตรวัด F ที่มีค่ามากที่สุดนั้นไม่อยู่ในการปรับค่าเงื่อนไขบังคับโดยรวมค่าใดค่าหนึ่งเสมอไป เพราะการได้มาซึ่งค่ามาตรวัด F ที่มีค่ามากนั้นจะขึ้นอยู่กับความเหมาะสมของขนาดเงื่อนไขบังคับโดยรวมกับชุดข้อมูลที่นำมาฝึกสอน หากนำขนาดของเงื่อนไขบังคับโดยรวมเป็นค่าที่ไม่เหมาะสมกับชุดข้อมูลนั้น ๆ อาจเกิดผลเสียขณะทำการฝึกสอนได้ ซึ่งทำให้ได้ผลการจำแนกคลาสที่ไม่ดี ผลการทดลองในตารางที่ 4.6 แสดงให้เห็นว่าขนาดของเงื่อนไขบังคับโดยรวมที่แตกต่างกันนั้นส่งผลต่อการเลือกข้อมูลขณะทำการฝึกสอน

ตารางที่ 4.6 ผลการทดลองการวัดประสิทธิภาพของการจำแนกคลาสแบบสองคลาสที่ใช้เกณฑ์หยุดด้วยวิธีที่เสนอ โดยใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงที่ปรับขนาดเงื่อนไขบังคับโดยรวม 5% 10% 100% ขณะทำการฝึกสอนด้วยตนเอง

ข้อมูล	เงื่อนไขบังคับโดยรวม 5%			เงื่อนไขบังคับโดยรวม 10%			เงื่อนไขบังคับโดยรวม 100%		
	ค่าความเที่ยง	ค่าความระลึก	ค่ามาตรวัด F	ค่าความเที่ยง	ค่าความระลึก	ค่ามาตรวัด F	ค่าความเที่ยง	ค่าความระลึก	ค่ามาตรวัด F
คลื่นหัวใจ	0.9668	0.8106	**0.8818**	0.9480	0.8183	0.8783	0.9687	0.7505	0.8457
ลายมือ	0.9837	0.7713	**0.8646**	0.7940	0.9040	0.8454	0.8017	0.8826	0.8402
โยคะ	0.7326	0.9081	0.811	0.7976	0.9132	**0.8515**	0.7134	0.8759	0.7864
ปืน	1.0000	0.7333	0.8462	1.0000	0.8000	0.8889	1.0000	0.9667	**0.9831**
น้ำมันมะกอก	0.9451	0.7667	**0.8466**	0.9615	0.7167	0.8212	1.0000	0.7167	0.8350
ถ้วยกาแฟ	0.6671	0.6308	0.6484	0.5757	0.7538	0.6529	0.5770	0.7538	**0.6537**
ซีบีเอฟ	0.9848	0.8426	0.9082	0.9996	0.8671	**0.9286**	0.9991	0.8568	0.9225
สองรูปแบบ	1.0000	0.6907	0.817	1.0000	0.8221	**0.9024**	1.0000	0.8560	0.9224
นิวเคลียร์เทรซ	1.0000	1.0000	**1.0000**	1.0000	1.0000	**1.0000**	1.0000	1.0000	**1.0000**
สังเคราะห์	1.0000	0.8200	**0.9011**	1.0000	0.8187	0.9003	1.0000	0.8200	**0.9011**

สำหรับการเรียนรู้แบบมีผู้สอนแล้ว การนำข้อมูลที่ทราบคลาสที่มีอยู่ไปทำการเรียนรู้ค่าเงื่อนไขบังคับโดยรวมที่เหมาะสมนั้นไม่ใช่เรื่องยาก เพราะข้อมูลที่ทราบคลาสนั้นมีจำนวนมากพอที่จะทำให้ทราบได้ว่าควรใช้ขนาดของเงื่อนไขบังคับโดยรวมเท่าไรจึงจะมีความเหมาะสม แต่สำหรับการเรียนรู้แบบกึ่งมีผู้สอนแล้ว ด้วยข้อมูลที่ทราบคลาสที่มีจำนวนจำกัดที่มีอยู่นั้นไม่เพียงพอต่อการนำไปคำนวณหาขนาดของเงื่อนไขบังคับโดยรวมที่เหมาะสมได้

จากผลการทดลองที่ 4.6 เมื่อทำการหาค่าเฉลี่ยของค่ามาตรวัด F ที่ได้จากขนาดเงื่อนไขบังคับโดยรวมที่มีค่าที่แตกต่างกันแล้วพบว่าได้ค่าเฉลี่ยของการใช้เงื่อนไขบังคับโดยรวมขนาด 5% 10% และ 100% คือ 0.8525 0.8670 0.8690 จะเห็นว่าเงื่อนไขบังคับโดยรวมขนาด 100% ให้ค่ามาตรวัด F ที่มีค่ามากที่สุด ซึ่งทำให้ตัดสินใจได้ว่าควรใช้ขนาดของเงื่อนไขบังคับโดยรวม 100% และเหตุผลอีกประการที่ควรเลือกใช้ขนาดของเงื่อนไขบังคับ

โดยรวมเป็น 100% คือการเริ่มฝึกสอนด้วยตนเองด้วยข้อมูลจำนวนน้อย ๆ นั้นควรกำหนดค่าเงื่อนไขบังคับโดยรวมให้มีค่ามาก ๆ ไว้ก่อน เพื่อการเลือกข้อมูลที่เกี่ยวข้องเข้ามายังตัวจำแนกให้ได้จำนวนมาก เพราะเมื่อได้ข้อมูลที่ทราบคลาสเพิ่มจากการฝึกสอนแล้ว จึงค่อยสามารถนำข้อมูลเหล่านั้นไปหาขนาดเงื่อนไขบังคับโดยรวมค่าอื่น ๆ ที่มีความเหมาะสมกว่านี้ในขั้นตอนต่อไปได้

4.3 การทดลองเพื่อวัดความสามารถของตัวจำแนกที่ใช้จำนวนข้อมูลขณะเริ่มทำการฝึกสอนที่แตกต่างกัน

นอกเหนือจากการเลือกใช้วิธีวัดระยะทาง วิธีเลือกเกณฑ์หยุด ขนาดของค่าเงื่อนไขบังคับโดยรวมของวิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงแล้ว ยังมีปัจจัยอื่นที่ส่งผลต่อการสร้างตัวจำแนกแบบกึ่งมีผู้สอนอีกอย่างหนึ่ง คือจำนวนข้อมูลที่ทราบคลาสที่ใช้ขณะเริ่มต้นทำการฝึกสอน

การทดลองชุดนี้ทำการทดลองเพื่อให้ทราบว่าจำนวนข้อมูลที่ทราบคลาสขณะเริ่มทำการฝึกสอน ส่งผลต่อตัวจำแนกที่ได้จากการฝึกสอนด้วยตนเองหรือไม่ การทดลองนี้จะใช้ข้อมูลเริ่มต้นสำหรับการฝึกสอนจำนวนไม่เท่ากันโดยจะลดจำนวนข้อมูลที่ใช้ในการฝึกสอนลงและเพิ่มข้อมูลที่ใช้ในการฝึกสอนมากขึ้น สำหรับข้อมูลที่นิยามว่ามีจำนวนกลุ่มข้อมูล (Cluster) 1 กลุ่มภายในคลาสที่สนใจ จะทำการทดลองด้วยการใช้จำนวนข้อมูล 1 3 และ 5 ตัว นอกจากนี้ยังได้ทำการทดลองด้วยการใช้จำนวนข้อมูลทั้งหมดสำหรับการฝึกสอน เพื่อเป็นค่าพื้นฐานว่าการเรียนรู้แบบกึ่งมีผู้สอนได้ผลการทดลองเป็นอย่างไรเมื่อเทียบกับการเรียนรู้แบบมีผู้สอน การทดลองนี้จะทำการฝึกสอนด้วยตนเองด้วยการใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงที่กำหนดเงื่อนไขบังคับโดยรวม 5% โดยจำนวนข้อมูลสำหรับการทดสอบแสดงในตารางที่ 4.2 ในส่วนของผลการทดลองที่ได้แสดงในตารางที่ 4.7

ผลการทดลองในตารางที่ 4.7 แสดงให้เห็นว่าเมื่อใช้จำนวนข้อมูลเมื่อเริ่มฝึกสอนที่มีจำนวนจำนวนมากขึ้นจะทำให้ค่าความระลึกมีค่าคงที่หรือมีค่าเพิ่มขึ้นในทุกชุดข้อมูล ซึ่งในหากพิจารณาจากค่ามาตรวัด F แล้ว การใช้ข้อมูลเริ่มต้นที่มีจำนวนลดลงจาก 3 ตัวเป็น 1 ตัว ทำให้ได้ค่ามาตรวัด F ที่มีค่าน้อยกว่าหรือมีค่าเท่าเดิมในทุกกรณี และการใช้จำนวนข้อมูลเมื่อเริ่มฝึกสอนที่มีจำนวนเพิ่มขึ้นจาก 3 ตัวเป็น 5 ตัว ทำให้ได้ค่ามาตรวัด F ที่มีค่ามากกว่าหรือเท่าเดิมในทุกกรณี โดยเมื่อเปรียบเทียบผลที่ได้ของการเรียนรู้แบบกึ่งมีผู้สอนกับการเรียนรู้แบบมีผู้สอนแล้วจะพบว่าได้ผลลัพธ์ที่น่าพอใจในหลายชุดข้อมูลที่ให้มาตรวัด F มีค่าเท่ากัน เช่นชุดข้อมูลปืน และชุดข้อมูลนิวเคลียร์เทรซ

ตารางที่ 4.7 ผลการจำแนกคลาสของตัวจำแนกที่ใช้ชุดข้อมูลที่นิยามว่ามีจำนวนกลุ่มข้อมูล 1 กลุ่ม โดยใช้จำนวนข้อมูลเมื่อเริ่มทำการฝึกสอน 1 3 5 และใช้ข้อมูลทุกตัวในการฝึกสอน

จำนวนข้อมูลที่ทราบคลาสเมื่อเริ่มทำการฝึกสอน	ค่าวัดที่วัดได้จากการทดลอง	ชุดข้อมูลที่นำมาทำการทดลอง					
		ปืน	น้ำมันมะกอก	ถ้วยกาแฟ	สองรูปแบบ	นิวเคลียร์เทรซ	สังเคราะห์
1 ตัว	ค่าความเที่ยง	1.0000	0.8667	0.6566	1.0000	1.0000	1.0000
	ค่าความระลึก	0.7333	0.5056	0.2238	0.6283	1.0000	0.8200
	ค่ามาตรวัด F	0.8462	0.6386	0.3338	0.7717	1.0000	0.9011
3 ตัว	ค่าความเที่ยง	1.0000	0.9451	0.6671	1.0000	1.0000	1.0000
	ค่าความระลึก	0.7333	0.7667	0.6308	0.6907	1.0000	0.8200
	ค่ามาตรวัด F	0.8462	0.8466	0.6484	0.8170	1.0000	0.9011
5 ตัว	ค่าความเที่ยง	1.0000	0.9408	0.6092	1.0000	1.0000	1.0000
	ค่าความระลึก	0.7333	0.8222	0.7333	0.7539	1.0000	0.8200
	ค่ามาตรวัด F	0.8462	0.8775	0.6655	0.8597	1.0000	0.9011
ทุกตัว	ค่าความเที่ยง	1.0000	0.9408	0.6092	1.0000	1.0000	1.0000
	ค่าความระลึก	0.7333	0.9167	0.8462	0.8920	1.0000	0.8600
	ค่ามาตรวัด F	0.8462	0.9286	0.7156	0.9429	1.0000	0.9247

สำหรับข้อมูลที่นิยามว่ามีจำนวนกลุ่มข้อมูล (Cluster) หลายกลุ่มภายในคลาสที่สนใจ จะทำการทดลองด้วยการใช้จำนวนข้อมูล 5 10 และ 15 ตัวสำหรับทำการฝึกสอน นอกจากนี้ยังได้ทดลองด้วยการใช้ข้อมูลเริ่มต้นทำการฝึกสอนจำนวน 1 ตัวและใช้ข้อมูลทุกตัวในการทดลองอีกด้วย การทดลองนี้จะทำการฝึกสอนด้วยตนเองด้วยการใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงที่กำหนดเงื่อนไขบังคับโดยรวม 5% โดยจำนวนข้อมูลสำหรับการทดสอบแสดงในตารางที่ 4.2 ในส่วนของผลการทดลองที่ได้แสดงในตารางที่ 4.8

ผลการทดลองในตารางที่ 4.8 แสดงให้เห็นว่าจำนวนข้อมูลเมื่อเริ่มฝึกสอนที่มีจำนวนลดลงจาก 10 ตัวเป็น 5 ตัว ทำให้ได้ค่ามาตรวัด F ที่มีค่าน้อยกว่าเดิมในทุกกรณี การใช้จำนวนข้อมูลเมื่อเริ่มฝึกสอนที่มีจำนวนเพิ่มขึ้นจาก 10 ตัวเป็น 15 ตัว ทำให้ได้ค่ามาตรวัด F ที่มีค่ามากกว่าเดิมในทุกกรณี การใช้จำนวนข้อมูลเริ่มต้นทำการฝึกสอนจำนวน 1 ตัวนั้นให้ค่ามาตรวัด F ที่มีค่าน้อยที่สุดเมื่อเทียบกับการใช้จำนวนข้อมูลเมื่อเริ่มฝึกสอนที่มากกว่า 1 ตัว และสำหรับการใช้ข้อมูลทุกตัวการฝึกสอนนั้นให้ค่ามาตรวัด F ที่มีค่ามากที่สุดเมื่อเทียบกับการใช้จำนวนข้อมูลเมื่อเริ่มฝึกสอนอื่น ๆ

ตารางที่ 4.8 ผลการจำแนกคลาสของตัวจำแนกที่ใช้ชุดข้อมูลที่นิยามว่าจะมีจำนวนกลุ่มข้อมูลหลายกลุ่ม โดยใช้จำนวนข้อมูลเมื่อเริ่มทำการฝึกสอน 1 5 10 15 และใช้ข้อมูลทุกตัวในการฝึกสอน

จำนวนข้อมูลที่ทราบคลาสเมื่อเริ่มทำการฝึกสอน	ค่าวัดที่วัดได้จากการทดลอง	ชุดข้อมูลที่นำมาทำการทดลอง			
		คลื่นหัวใจ	ลายมือ	โยคะ	ซีบีเอฟ
1 ตัว	ค่าความเที่ยง	0.9690	0.8847	0.8983	1.0000
	ค่าความระลึก	0.7406	0.4972	0.4808	0.1824
	ค่ามาตรวัด F	0.8395	0.6367	0.6263	0.3085
5 ตัว	ค่าความเที่ยง	0.9667	0.9841	0.6919	0.8402
	ค่าความระลึก	0.7747	0.5278	0.9107	0.6684
	ค่ามาตรวัด F	0.8601	0.6871	0.7863	0.7445
10 ตัว	ค่าความเที่ยง	0.9668	0.9837	0.7326	0.9848
	ค่าความระลึก	0.8106	0.7713	0.9081	0.8426
	ค่ามาตรวัด F	0.8818	0.8646	0.8110	0.9082
15 ตัว	ค่าความเที่ยง	0.9675	0.9762	0.7394	0.9954
	ค่าความระลึก	0.8387	0.8287	0.9692	0.8465
	ค่ามาตรวัด F	0.8985	0.8964	0.8388	0.9149
ทุกตัว	ค่าความเที่ยง	0.9711	0.9561	0.9123	1.0000
	ค่าความระลึก	0.8878	1.0000	1.0000	0.9226
	ค่ามาตรวัด F	0.9276	0.9776	0.9541	0.9597

จากการทดลองในตารางที่ 4.7 และตารางที่ 4.8 ทำให้สรุปได้ว่าจำนวนข้อมูลเมื่อเริ่มฝึกสอนหากมีจำนวนมากแล้ว จะทำให้การฝึกสอนตนเองสามารถสร้างตัวจำแนกที่จำแนกคลาสได้ดีขึ้น และเมื่อวิเคราะห์จากค่ามาตรวัด F ที่เปลี่ยนแปลงไปอย่างละเอียดทำให้ได้ข้อสรุปบางอย่างเกี่ยวกับจำนวนกลุ่มข้อมูลในคลาสที่สนใจได้ดังนี้

- กรณีที่ชุดข้อมูลที่จำนวนข้อมูลเปลี่ยนแปลงเมื่อเริ่มต้นทำการฝึกสอนไม่ส่งผลต่อค่ามาตรวัด F เช่น ข้อมูลปืน นิวเคลียร์เทรซ และข้อมูลสังเคราะห์ แสดงให้เห็นว่าชุดข้อมูลนั้นเป็นข้อมูลที่มีกลุ่มข้อมูลภายในคลาสที่สนใจเพียงกลุ่มเดียว และมีการกระจายตัวของข้อมูลที่ไม่มาก นั่นคือข้อมูลในคลาสที่สนใจมีการเกาะกลุ่มกันมาก

- กรณีที่ชุดข้อมูลที่จำนวนข้อมูลเปลี่ยนแปลงเมื่อเริ่มต้นทำการฝึกสอนส่งผลต่อค่ามาตรวัด F แสดงให้เห็นว่าชุดข้อมูลนั้นอาจมีกลุ่มข้อมูลภายในคลาสที่สนใจหลายกลุ่ม หรือชุดข้อมูลนั้นอาจมีกลุ่มข้อมูลภายใจคลาสที่สนใจกลุ่มเดียวแต่มีการกระจายตัวของข้อมูลในกลุ่มนั้น ๆ มาก

4.4 วิธีการทดลองการดัดแปลงขั้นตอนวิธีให้ตัวจำแนกสามารถได้รับการฝึกสอนและจำแนกคลาสได้ทีละหลายคลาส

ในหัวข้อนี้เป็นการทดลองเพิ่มเติมนอกเหนือจากในส่วนของงานวิจัยหลักโดยมีจุดประสงค์เพื่อสร้างตัวจำแนกคลาสแบบหลายคลาส การทดลองนี้จะวัดว่าตัวจำแนกแบบหลายคลาสที่ออกแบบมานั้น เมื่อเทียบผลการจำแนกคลาสกับการเรียนรู้แบบมีผู้สอนแล้ว จะมีผลเป็นอย่างไรโดยการทดลองชุดนี้จะใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงที่ไม่กำหนดขนาดเงื่อนไขบังคับโดยรวม โดยข้อมูลที่ใช้เริ่มทำการฝึกสอนนั้นแสดงในตารางที่ 4.9

ตารางที่ 4.9 จำนวนข้อมูลที่เตรียมสำหรับการสร้างตัวจำแนกคลาสแบบหลายคลาส

ข้อมูล	จำนวนข้อมูลที่ทราบคลาสเมื่อเริ่มทำการฝึกสอน (แต่ละคลาส)	จำนวนข้อมูลที่ใช้ทำการฝึกสอน	จำนวนข้อมูลที่ใช้ทำการทดสอบ
คลื่นหัวใจ	(5+5) = 10	810	1,216
ลายมือ	(5+1) = 6	805	905
โยคะ	(5+5) = 10	306	306
ปืน	(1+1+1+1) = 4	122	125
น้ำมันมะกอก	(1+1+1+1) = 4	30	30
ถ้วยกาแฟ	(1+1) = 2	28	28
ซีบีเอฟ	(5+5+5) = 3	465	465
สองรูปแบบ	(1+1+1+1) = 4	1,000	4,000
นิวเคลียร์เทรซ	(1+1+1+1) = 4	100	100
สังเคราะห์	(5+1+1+1+1+1) = 10	300	300

ตารางที่ 4.9 แสดงจำนวนข้อมูลที่ทำการสุ่มในแต่ละคลาสของแต่ละชุดข้อมูลที่แตกต่างกัน โดยข้อมูลในวงเล็บคือข้อมูลในแต่ละคลาส เช่นข้อมูลสังเคราะห์ที่มีจำนวนข้อมูล (5+1+1+1+1+1) แสดงว่าจะทำการสุ่มข้อมูลจากคลาสแรกของข้อมูลชุดนี้ 5 ตัว คลาสสองถึงคลาสที่หก จะทำการสุ่มคลาสละ 1 ตัวตามลำดับ โดยในคลาสที่ทำการสุ่มข้อมูล 5 ตัวนั้นเป็นคลาสที่นิยามว่ามีจำนวนกลุ่มข้อมูลภายในคลาสหลายกลุ่ม และในคลาสที่ทำการสุ่มข้อมูล 1 ตัวนั้นเป็นคลาสที่นิยามว่ามีจำนวนกลุ่มข้อมูลภายในคลาสนั้น ๆ เพียงกลุ่มเดียว ในส่วนของคอลัมน์อื่น ๆ ของตารางที่ 4.9 แสดงจำนวนข้อมูลสำหรับการฝึกสอน และจำนวนข้อมูลสำหรับการทดสอบ

เมื่อสุ่มข้อมูลในแต่ละคลาสได้จำนวนข้อมูลตามที่ต้องการแล้ว และเมื่อข้อมูลกลุ่มนี้ได้รับการฝึกสอนและนำไปใช้สำหรับการจำแนกคลาส การทดลองนี้จะทำการทดลองทั้งหมด 20 รอบ เพื่อการคำนวณหาค่าความแม่นยำเฉลี่ย และบันทึกค่าที่มากที่สุดไว้ ผลลัพธ์ที่ได้จากการทดลองแสดงในตารางที่ 4.10

ตารางที่ 4.10 ผลการทดลองของตัวจำแนกแบบหลายคลาส ซึ่งทำการฝึกสอนด้วยตนเองด้วยการใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงที่ไม่กำหนดค่าเงื่อนไขบังคับโดยรวม

ชุดข้อมูล	ค่าความแม่นยำ (%)		
	การเรียนรู้แบบกึ่งมีผู้สอน		การเรียนรู้แบบมีผู้สอน
	ค่าเฉลี่ย	ค่าที่มากที่สุด	
คลื่นหัวใจ	79.37	79.76	**99.01**
ลายมือ	80.66	83.24	**100.00**
โยคะ	80.68	82.60	**100.00**
ปืน	86.08	**90.40**	88.80
น้ำมันมะกอก	82.33	**90.00**	86.77
ถ้วยกาแฟ	63.21	67.85	**82.01**
ซีบีเอฟ	63.65	66.02	**99.07**
สองรูปแบบ	91.79	99.70	**100.00**
นิวเคลียร์เทรซ	84.40	87.00	**100.00**
สังเคราะห์	87.46	91.00	**99.03**

ผลการทดลองจากตารางที่ 4.10 ตัวเลขที่เป็นตัวหนาแสดงความแม่นยำของวิธีที่ให้ค่ามากกว่าอีกวิธีหนึ่ง และจากค่าในตารางที่ 4.10 แสดงให้เห็นว่าวิธีการสร้างตัวจำแนกคลาสแบบหลายคลาสสามารถนำไปใช้งานได้ดีในระดับหนึ่ง เพราะมีบางชุดข้อมูลเมื่อผ่านการฝึกสอนด้วยตนเองแล้ว ทำให้ค่าความแม่นยำของตัวจำแนกคลาสในกรณีที่ได้ค่ามากที่สุดมีค่ามากกว่าการใช้ข้อมูลในเซตฝึกสอนทั้งหมดมาทำการจำแนกคลาส เช่น ข้อมูลปืน และข้อมูลน้ำมันมะกอก โดยเหตุการณ์นี้อธิบายได้ด้วยเหตุผลที่ว่าข้อมูลที่ได้รับการสุ่มตอนเริ่มทำการฝึกสอนนั้นอาจอยู่ในตำแหน่งที่เป็นศูนย์กลางของกลุ่มข้อมูลนั้น ๆ ทำให้การฝึกสอนด้วยตนเองสามารถทำได้อย่างมีประสิทธิภาพ และการใช้เกณฑ์หยุดวิธีที่นำเสนอนี้สามารถเลือกตำแหน่งที่ควรจะเป็นเกณฑ์หยุดได้ดี ซึ่งตำแหน่งนั้นคือตำแหน่งที่เมื่อกลุ่มที่ไม่มีลักษณะใกล้เคียงกับข้อมูลที่สนใจ เริ่มได้รับการฝึกสอนนั่นเอง โดยผลการทดลองที่ได้เป็นผลที่ดีเกินความคาดหมาย โดยเฉพาะเมื่อพิจารณาถึงจำนวนข้อมูลจำนวนไม่มากที่ใช้สำหรับการเริ่มฝึกสอน แต่สามารถให้ผลการจำแนกคลาสที่ไม่แตกต่างจากการเรียนรู้แบบมีผู้สอนมาก

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

งานวิจัยนี้เป็นการเลือกเกณฑ์หยุดที่เหมาะสมเพื่อช่วยให้ตัวจำแนกคลาสแบบกึ่งมีผู้สอนที่สร้างได้ สามารถจำแนกคลาสได้ด้วยความแม่นยำที่เพิ่มมากขึ้น และจากการทดลองในบทที่ 4 สามารถสรุปผลการทดลองได้ดังนี้

5.1 สรุปผลการวิจัย

งานวิจัยนี้ศึกษาเกี่ยวกับการสร้างตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบกึ่งมีผู้สอนด้วยเกณฑ์หยุดแบบใหม่ที่คำนวณจากผลต่างของค่าระยะทางให้สามารถจำแนกคลาสด้วยความแม่นยำมากกว่าการใช้เกณฑ์หยุดแบบเดิมที่คำนวณจากค่าระยะทางที่มีค่าน้อยที่สุดที่เคยเกิดขึ้น ในการทดลองจะทำการเรียนรู้แบบกึ่งมีผู้สอนด้วยการใช้วิธีการฝึกสอนด้วยตนเอง และเปรียบเทียบผลการจำแนกคลาสระหว่างตัวจำแนกที่สร้างจากเกณฑ์หยุดที่งานวิจัยชิ้นนี้นำเสนอ และตัวจำแนกที่สร้างจากเกณฑ์หยุดของ Wei and Keogh [7] โดยชุดข้อมูลอนุกรมเวลาที่นำมาใช้มีจำนวน 10 ชุดข้อมูลที่มีความหลากหลายในการทดลองครั้งนี้

จากผลการทดลองของการจำแนกคลาสของตัวจำแนกแบบสองคลาสที่สร้างจากวิธีการเลือกเกณฑ์หยุดที่งานวิจัยชิ้นนี้นำเสนอรวมกับการใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงให้ค่ามาตรวัด F (F-measure) เฉลี่ยของทั้ง 10 ชุดข้อมูลคือ 85.25% ซึ่งมากกว่าวิธีการเลือกเกณฑ์หยุดที่ Wei และ Keogh กับการใช้วิธีวัดระยะทางแบบยุคลิด ซึ่งให้ค่ามาตรวัดคือ 63.60% อยู่ 21.61% โดยเมื่อแยกพิจารณาเป็นกรณีที่ใช้วิธีการเลือกเกณฑ์หยุดที่แตกต่างกันแต่ ใช้วิธีวัดระยะทางที่เหมือนกัน ผลการทดลองที่ได้จากการใช้วิธีวัดระยะทางแบบยุคลิดกับการเลือกเกณฑ์หยุดด้วยวิธีที่งานวิจัยชิ้นนี้นำเสนอให้ค่ามาตรวัด F เฉลี่ยของทั้ง 10 ชุดข้อมูลคือ 70.60% ซึ่งมากกว่าวิธีที่ Wei และ Keogh นำเสนอคือ 63.60% อยู่ 7.00% และผลการทดลองที่ได้จากการใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงกับการเลือกเกณฑ์หยุดด้วยวิธีที่งานวิจัยชิ้นนี้นำเสนอให้ค่ามาตรวัด F เฉลี่ยของทั้ง 10 ชุดข้อมูลคือ 85.25% ซึ่งมากกว่าวิธีที่ Wei และ Keogh นำเสนอคือ 67.77% อยู่ 17.48% จึงสามารถสรุปได้ว่าการหาเกณฑ์หยุดด้วยวิธีที่งานวิจัยชิ้นนี้นำเสนอช่วยทำให้ตัวจำแนกคลาสมีผลการจำแนกคลาสแบบสองคลาสที่ดีกว่าการหาเกณฑ์หยุดของ Wei และ Keogh และหากวิเคราะห์จากค่าความเที่ยง (Precision) และค่าความระลึก (Recall) แล้วจะได้ข้อสรุปว่าวิธีการเลือกเกณฑ์หยุดที่เหมาะสมส่งผลต่อการสร้างตัวจำแนกคลาสที่ดี นอกจากนี้วิธีวัดระยะทางที่เหมาะสมจะช่วยให้ค่าความเที่ยงของตัวจำแนกแบบมีค่าสูง และทำให้ประสิทธิภาพโดยรวมของการจำแนกคลาสดีขึ้นอีกด้วย

แม้ว่าการใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงจะได้ผลการจำแนกคลาสที่ดี แต่จากการทดลองอื่นเพิ่มเติม พบว่าปัจจัยที่ส่งผลต่อตัวจำแนกแบบกึ่งมีผู้สอนอีกอย่างหนึ่งคือขนาดของเงื่อนไขบังคับโดยรวม ซึ่งถ้ามีขนาดที่เหมาะสมจะช่วยให้ตัวจำแนกสามารถจำแนกคลาสได้อย่างแม่นยำ แต่ถ้ามีขนาดที่ไม่เหมาะสมอาจทำให้ได้ผลการจำแนกที่ไม่ดีได้ ปัจจัยที่มีผลอีกประการหนึ่งคือ จำนวนข้อมูลที่ใช้ขณะเริ่มทำการฝึกสอน ถ้าจำนวนข้อมูลตอนเริ่มทำการฝึกสอนมีอยู่เป็นจำนวนมาก ผลของการจำแนกคลาสที่ได้จะมีแนวโน้มที่มีค่าความถูกต้องที่ดีขึ้นด้วย

ในส่วนผลการทดลองของตัวจำแนกคลาสแบบหลายคลาสนั้นจะเห็นว่าได้ผลการทดลองที่น่าพึงพอใจ ไม่แตกต่างจากการเรียนรู้แบบมีผู้สอนมากนักในบางชุดข้อมูล และสามารถใช้งานได้ดีในกรณีที่ข้อมูลที่สนใจจะทำการจำแนกมีอยู่หลายคลาส และข้อมูลที่ทราบคลาสมีอยู่เป็นจำนวนน้อย

5.2 ข้อเสนอแนะ

งานวิจัยนี้ได้เสนอวิธีการเลือกเกณฑ์หยุดที่เหมาะสมที่ช่วยพัฒนาให้ตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบกึ่งมีผู้สอนได้ผลการจำแนกคลาสที่ดีมากยิ่งขึ้น อย่างก็ตาม ผู้เขียนยังมีข้อเสนอแนะบางประการที่สามารถช่วยเพิ่มประสิทธิภาพของตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบกึ่งมีผู้สอนให้มีดียิ่งขึ้น

1. สำหรับข้อมูลอนุกรมเวลาที่แปลงมาจากรูปภาพ หากข้อมูลนั้นได้รับวิธีการใช้วิธีการแปลงข้อมูลที่เหมาะสม หรือมีการเลือกคุณลักษณะที่เหมาะสมบางอย่างเพื่อแปลงเป็นข้อมูลอนุกรมเวลา จะทำให้ตัวจำแนกคลาสแบบกึ่งมีผู้สอนที่สร้างได้ ให้ค่าความถูกต้องของการจำแนกคลาสเพิ่มมากขึ้น
2. สำหรับการใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง การกำหนดค่าเงื่อนไขบังคับโดยรวม (Global Constraint) ที่เหมาะสมนั้น ส่งผลโดยตรงต่อความแม่นยำของการจำแนกคลาสของตัวจำแนกแบบกึ่งมีผู้สอนที่สร้างได้ ดังนั้นการเรียนรู้ขนาดของค่าเงื่อนไขบังคับโดยรวมก่อนทำการฝึกสอนจึงเป็นเรื่องที่ดี [9]
3. การหยุดฝึกสอนขณะที่กำลังฝึกสอนด้วยตนเอง ช่วยให้ความเร็วของการสร้างตัวจำแนกคลาสแบบกึ่งมีผู้สอนเพิ่มมากขึ้น
4. สำหรับตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบสองคลาส ค่าขีดแบ่งที่คำนวณได้อย่างเหมาะสมส่งผลให้การจำแนกคลาสมีความถูกต้องมากยิ่งขึ้น โดยงานวิจัยชิ้นนี้ยังไม่ได้กล่าวถึงการปรับค่าขีดแบ่งให้มีความยืดหยุ่นมากขึ้นสำหรับการจำแนกคลาสแต่อย่างใด
5. จำนวนข้อมูลและความถูกต้องของข้อมูลที่ทราบคลาสขณะเริ่มต้นทำการฝึกสอนมีผลต่อความแม่นยำของตัวจำแนกที่สร้างได้อย่างมาก

รายการอ้างอิง

[1] Ratanamahatana, C.A., and Keogh, E. (2005). Using Relevance Feedback to Learn Both the Distance Measure and the Query in Multimedia Databases. 9th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems: pp. 16-23.

[2] Xi, X., Keogh, E., Shelton, C., Wei, L., and Ratanamahatana, C.A. (2006). Fast Time Series Classification Using Numerosity Reduction. Proceedings of 23rd International Conference on Machine Learning (ICML 2006), Pittsburgh, PA.

[3] Chapelle, O., Schölkopf, B., and Zien, A. (2006). Semi-Supervised Learning. MIT Press, Cambridge, MA

[4] Zhu, X. (2005). Semi-Supervised Learning Literature Survey. Technical report, no.1530, Computer Sciences, University of Wisconsin-Madison.

[5] Cohen, I., Cozman, F.G., Sebe, N., Cirelo, M.C., and Huang, T.S. (2004). Semi-Supervised Learning of Classifiers: Theory,Algorithms, and Their Application to Human-Computer Interaction. IEEE Transaction on Pattern Analysis and Machine Intelligence, pp. 1553-1567.

[6] Li, M., and Zhou, Z.-H. (2005). SETRED: Self-Training with Editing. Proceedings of 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'05), pp. 611-621.

[7] Wei, L., and Keogh, E. (2006). Semi-Supervised Time Series Classification. Proceedings 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 748-753.

[8] Ratanamahatana, C.A., and Keogh, E. (2004). Everything you know about Dynamic Time Warping is wrong. Proceedings of 3rd Workshop on Mining Temporal and Sequential Data, In Conjunction with 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004).

[9] Ratanamahatana, C.A., and Keogh, E. (2004). Making Time-Series Classification More Accurate Using Learned Constraints. Proceedings of SIAM International Conference on Data Mining, pp. 11-22.

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

[11] Sakoe, H., and Chiba, S. (1990). Dynamic Programming Algorithm Optimization for Spoken Word Recognition. Morgan Kaufmann.

[12] Nigam, K., Mccallum, A.K., Thrun, S., and Mitchell, T. (2000). Text Classification from Labeled and Unlabeled Documents using EM. Machine Learning: pp. 103–134.

[13] Blum, A., and Lafferty, J. (2001). Learning from Labeled and Unlabeled Data using Graph Mincuts. Proceedings of 18th International Conference on Machine Learning, pp. 19-26.

[14] Bennett, K.P., and Demiriz, A. (1999). Semi-Supervised Support Vector Machines. Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II, pp. 368-374.

[15] Blum, A., and Mitchell, T. (1998). Combining Labeled and Unlabeled Data with Co-Training. Proceedings of 11th Annual Conference on Computational Learning Theory, pp. 92-100. Madison, Wisconsin, United States

[16] Shahshahani, B.M., and Landgrebe, D.A. (1994). The Effect of Unlabeled Samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon. IEEE Transactions on Geoscience and Remote Sensing, pp. 1087-1095.

[17] Zhang, R., and Alexander, I.R. (2006). a New Data Selection Principle for Semi-Supervised Incremental Learning. Proceedings of 18th International Conference on Pattern Recognition (ICPR'06), pp. 780-783.

[18] Wikipedia. (2008). Information Retrieval [Online]. Available from: http://en.wikipedia.org/wiki/Information_retrieval [cited 17 February 2008]

[19] UCR. (2008). The UCR Time Series Classification/Clustering Homepage [Online]. Available from: www.cs.ucr.edu/~eamonn/time_series_data/ [cited 1 January 2008]

[20] Wei, L. (2006). Self Training dataset [Online]. Available from: http://www.cs.ucr.edu/~wli/selfTraining/ [cited 1 May 2007]

[21] Keogh, E., Lin, J., and Truppel, W. (2005). Clustering of Time Series Subsequences is Meaningless: Implications for Past and Future Research. Proceedings of the 3rd IEEE International Conference on Data Mining, pp. 115-122.

ภาคผนวก

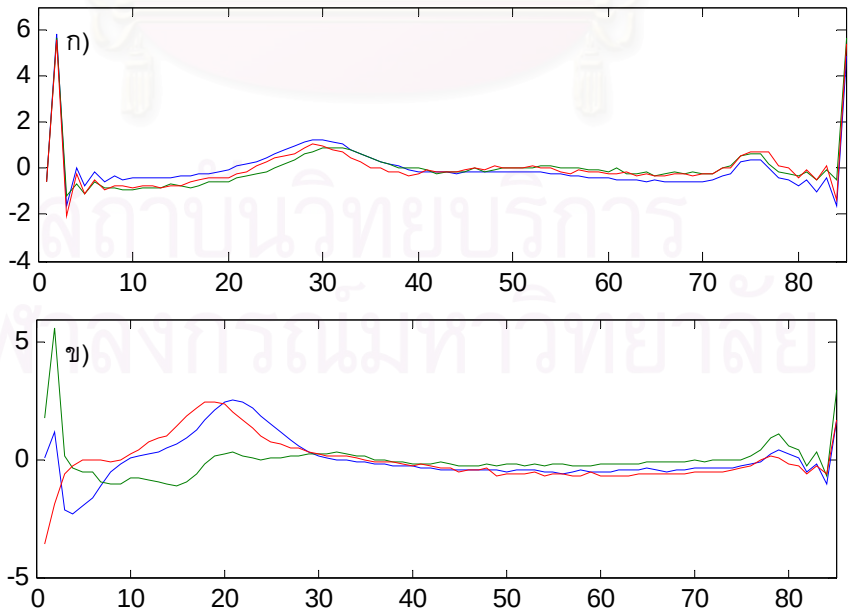
ภาคผนวก ก

ข้อมูลที่ใช้ในการทดลอง

ข้อมูลอนุกรมเวลาที่ใช้ในการทดลองนั้นมาจาก 2 แหล่ง คือ จากเว็บไซต์ของ University of California, Riverside's archive [19] และจากงานวิจัยชื่อ Semi Supervised Time Series Classification ของ Wei และ Keogh [20] ข้อมูลทั้งหมดมีจำนวน 10 ชุดข้อมูล โดยนำมาจากงานวิจัยของ Wei และ Keogh 4 ชุดข้อมูล คือ ข้อมูลคลื่นหัวใจ ข้อมูลลายมือ ข้อมูลโยคะ และข้อมูลปืน ส่วนข้อมูลอีก 6 ชุดข้อมูลที่เหลือนำมาจากเว็บไซต์ของ UCR คือ ข้อมูลถ้วยกาแฟ ข้อมูลน้ำมันมะกอก ข้อมูลซีบีเอฟ ข้อมูลสองรูปแบบ ข้อมูลนิวเคลียร์เทรซ ข้อมูลที่สังเคราะห์ขึ้นเพื่อการทดลอง

ก.1 ข้อมูลคลื่นหัวใจ (ECG dataset)

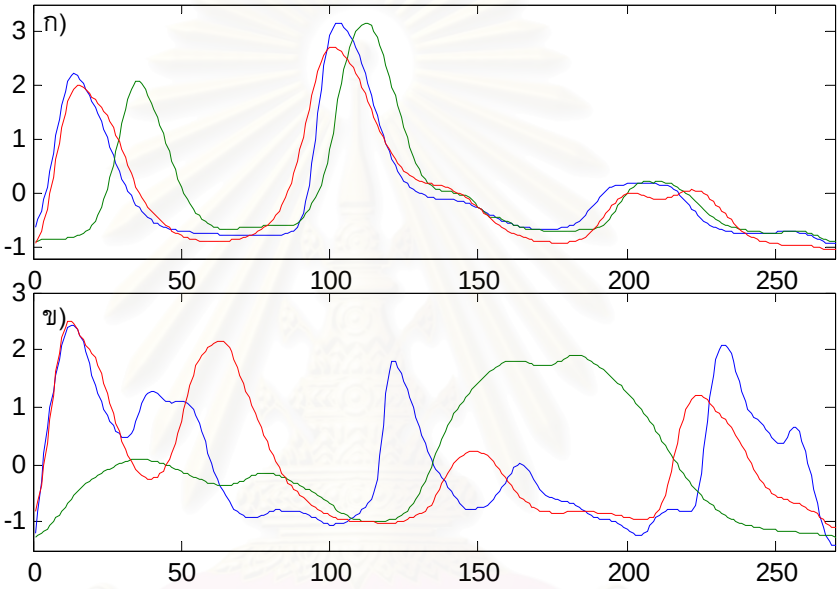
ข้อมูลคลื่นหัวใจเป็นข้อมูลที่อยู่ในโดเมนเกี่ยวกับการแพทย์ ข้อมูลที่บันทึกได้จะอยู่ในรูปของข้อมูลอนุกรมเวลา ข้อมูลชุดนี้ประกอบด้วยคลาส 2 คลาสคือ คลาสของคลื่นของหัวใจที่เต้นแบบเป็นปกติ ดังรูปที่ ก.1ก) และคลาสของคลื่นของหัวใจที่เต้นแบบไม่เป็นปกติ ดังรูปที่ ก.1ข) ข้อมูลชุดนี้ประกอบข้อมูลทั้งหมด 2,026 ตัว โดยข้อมูลแต่ละตัวมีความยาวของข้อมูล 87 จุด ลักษณะของข้อมูลคลื่นหัวใจในแต่ละคลาสเป็นดังรูปที่ ก.1



รูปที่ ก.1 ข้อมูลคลื่นหัวใจในแต่ละคลาส

ก.2 ข้อมูลลายมือ (Word spotting dataset)

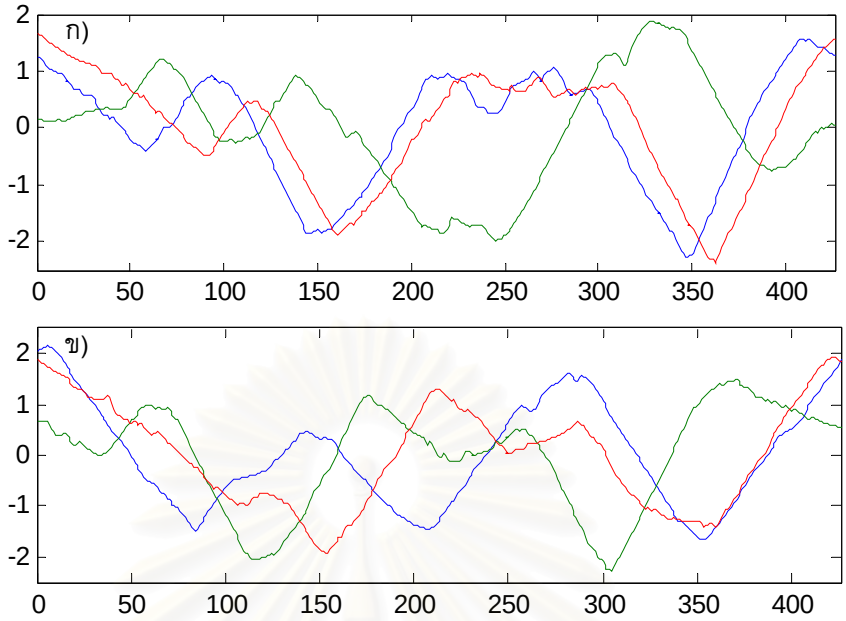
ข้อมูลลายมือได้รับการนำมาแปลงเป็นข้อมูลอนุกรมเวลา ซึ่งข้อมูลโดยดั้งเดิมนั้นเป็นรูปภาพลายมือเขียนของคน ข้อมูลชุดนี้ประกอบด้วยคลาส 2 คลาสคือ คลาสของลายมือที่เขียนคำว่า the ดังรูปที่ ก.2ก) และคลาสของลายมือที่เขียนถึงคำอื่น ดังรูปที่ ก.2ข) ข้อมูลชุดนี้ประกอบด้วยข้อมูลทั้งหมด 1,710 ตัว โดยข้อมูลแต่ละตัวมีความยาวของข้อมูล 272 จุด ลักษณะของข้อมูลลายมือในแต่ละคลาสเป็นดังรูปที่ ก.2



รูปที่ ก.2 ข้อมูลลายมือในแต่ละคลาส

ก.3 ข้อมูลโยคะ (Yoga dataset)

ข้อมูลโยคะเป็นข้อมูลที่รวบรวมท่าทางของคนที่เล่นโยคะในท่าทางที่แตกต่างกัน ข้อมูลชุดนี้ประกอบด้วยคลาส 2 คลาสคือ คลาสของผู้เล่นโยคะที่เป็นเพศหญิง ดังรูป ก.3ก) และคลาสของผู้เล่นโยคะที่เป็นเพศชาย ดังรูป ก.3ข) ข้อมูลชุดนี้ประกอบด้วยข้อมูลทั้งหมด 612 ตัว โดยข้อมูลแต่ละตัวมีความยาวของข้อมูล 428 จุด ลักษณะของข้อมูลโยคะในแต่ละคลาสเป็นดังรูปที่ ก.3



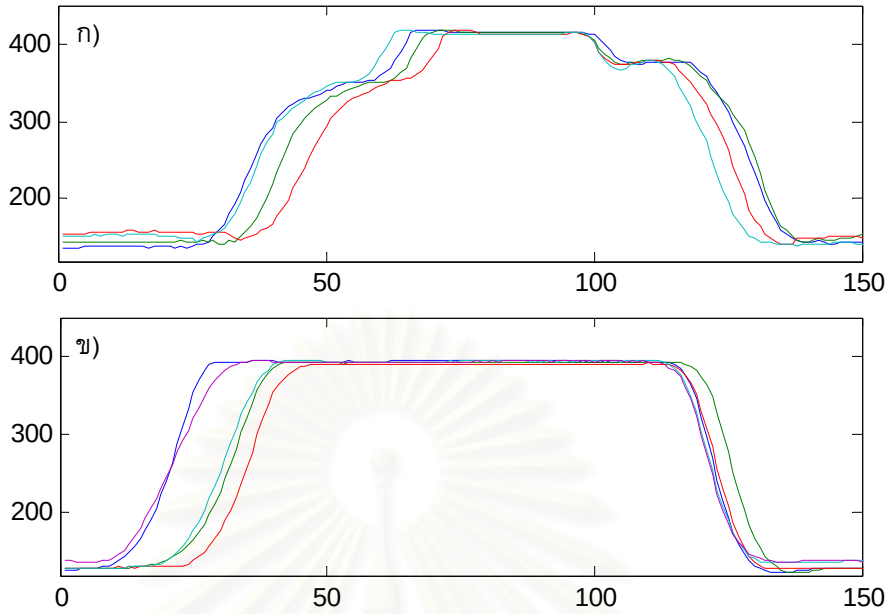
รูปที่ ก.3 ข้อมูลโยคะในแต่ละคลาส

ก.4 ข้อมูลปืน (Gun dataset)

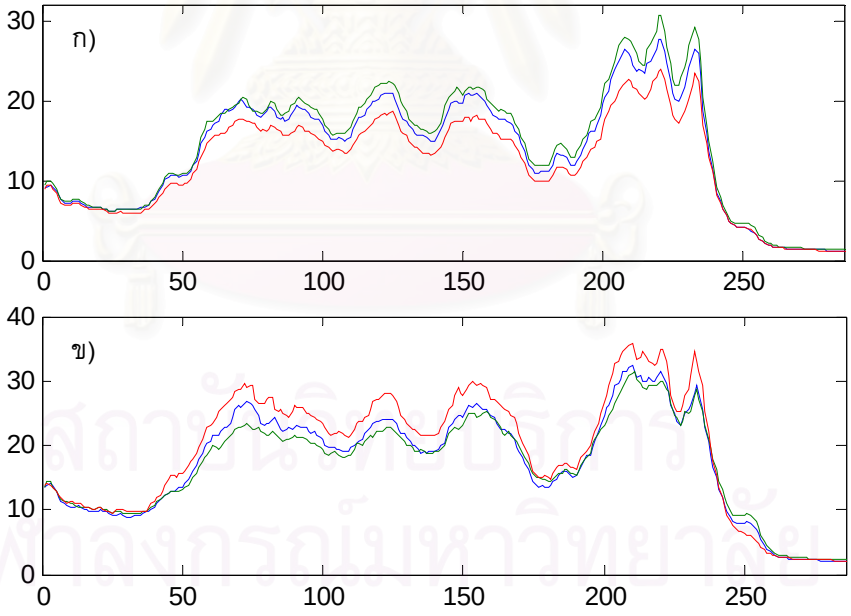
ข้อมูลอนุกรมเวลาชุดนี้ได้รับการแปลงมาจากการถ่ายวีดีโอของคนที่ใส่ถุงมือสีแดงที่มือแล้วยกมือขึ้นทำท่ายิงปืน โดยข้อมูลอนุกรมเวลาแสดงการเคลื่อนไหวของมือที่ทำท่ายิงปืน ข้อมูลชุดนี้ประกอบด้วยคลาส 2 คลาสคือ คลาสของคนที่ถือปืนจริง ๆ แล้วทำท่ายิงปืน ดังรูปที่ ก.4ก) และคลาสของคนที่ไม่ได้ถือปืนแต่ทำท่ายิงปืน ดังรูปที่ ก.4ข) ข้อมูลชุดนี้ประกอบด้วยข้อมูลทั้งหมด 247 ตัว โดยข้อมูลแต่ละตัวมีความยาวของข้อมูล 152 จุด ลักษณะของข้อมูลปืนในแต่ละคลาสเป็นดังรูปที่ ก.4

ก.5 ข้อมูลถ้วยกาแฟ (Coffee dataset)

ข้อมูลอนุกรมเวลาชุดนี้ได้รับการแปลงมาจากรูปภาพถ้วยกาแฟ ข้อมูลชุดนี้ประกอบด้วยคลาส 2 คลาสคือ คลาสของถ้วยกาแฟขนาดเล็ก ดังรูปที่ ก.5ก) และคลาสของถ้วยกาแฟขนาดใหญ่ ดังรูปที่ ก.5ข) ข้อมูลชุดนี้ประกอบด้วยข้อมูลทั้งหมด 56 ตัว โดยข้อมูลแต่ละตัวมีความยาวของข้อมูล 128 จุด ลักษณะของข้อมูลถ้วยกาแฟในแต่ละคลาสเป็นดังรูปที่ ก.5



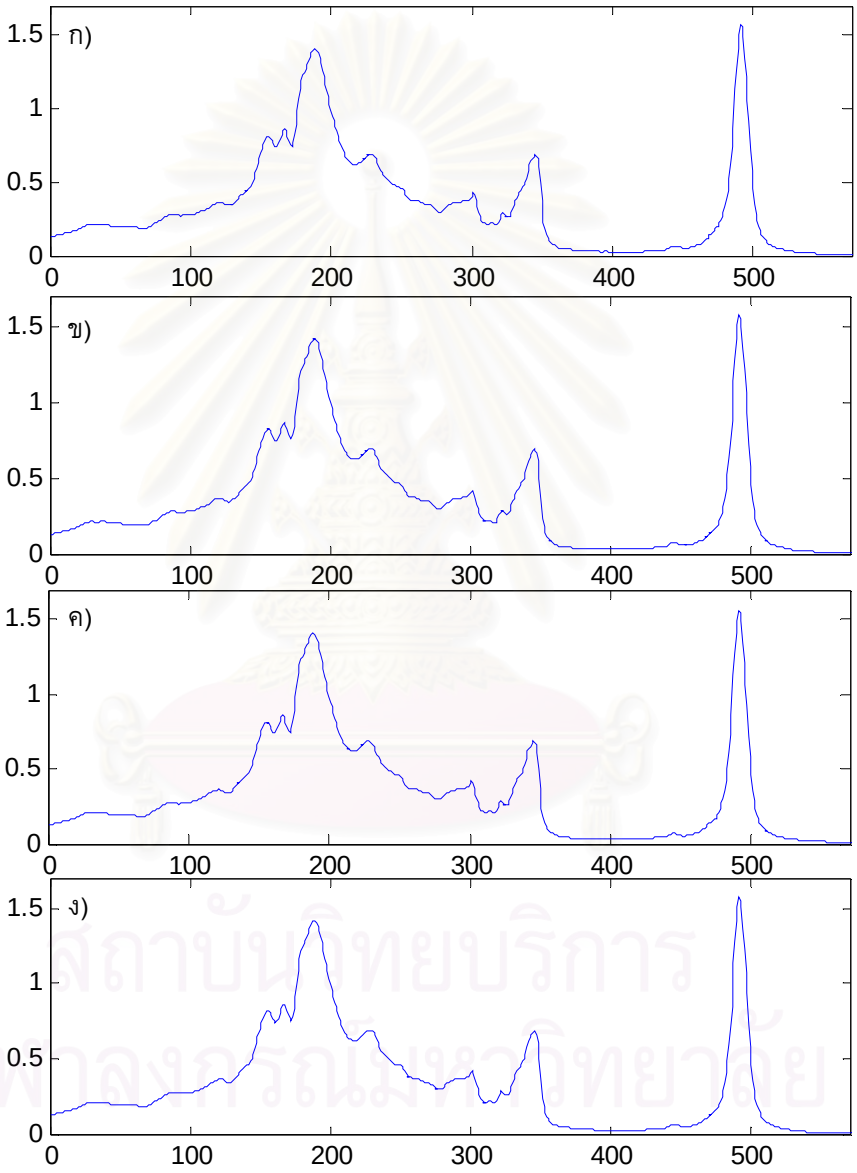
รูปที่ ก.4 ข้อมูลปืนในแต่ละคลาส



รูปที่ ก.5 ข้อมูลถ้วยกาแฟในแต่ละคลาส

ก.6 ข้อมูลน้ำมันมะกอก (Olive oil dataset)

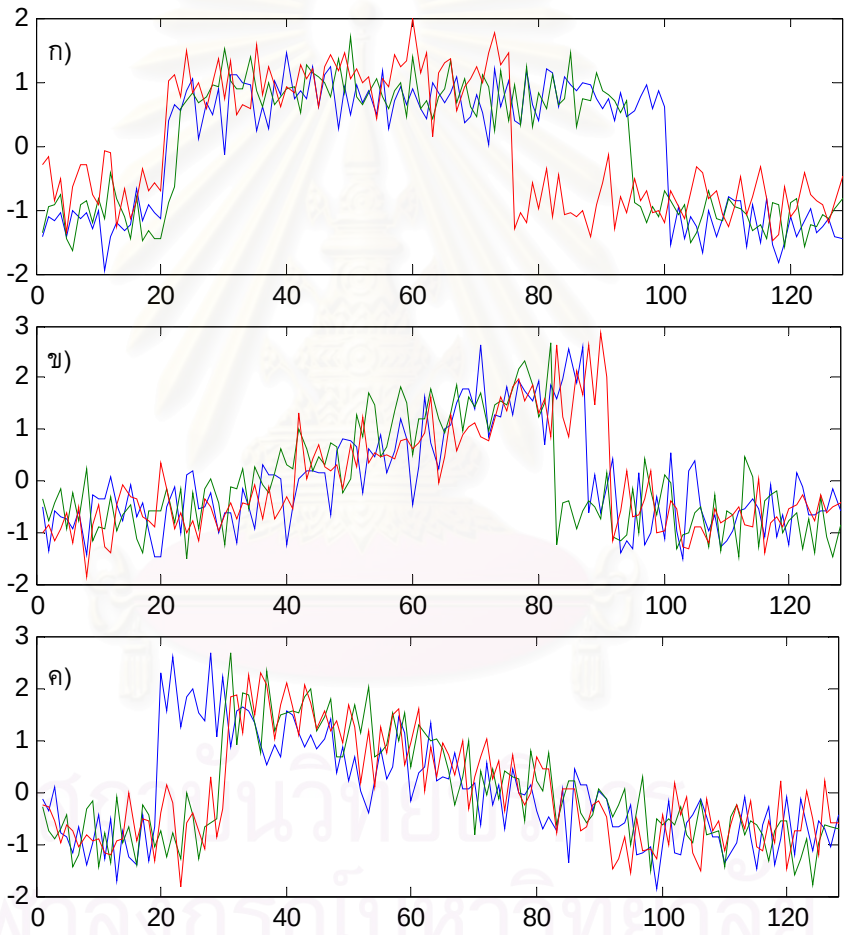
ข้อมูลอนุกรมเวลาชุดนี้เป็นประกอบด้วยคลาส 4 คลาส ข้อมูลชุดนี้ประกอบด้วยข้อมูลฝึกสอน 30 ตัว และข้อมูลสำหรับทดสอบ 30 ตัว โดยข้อมูลแต่ละตัวมีความยาวของข้อมูล 570 จุด ลักษณะของข้อมูลน้ำมันมะกอกในแต่ละคลาสเป็นดังรูปที่ ก.6



รูปที่ ก.6 ข้อมูลน้ำมันมะกอกในแต่ละคลาส

ก.7 ข้อมูลซีบีเอฟ (CBF dataset)

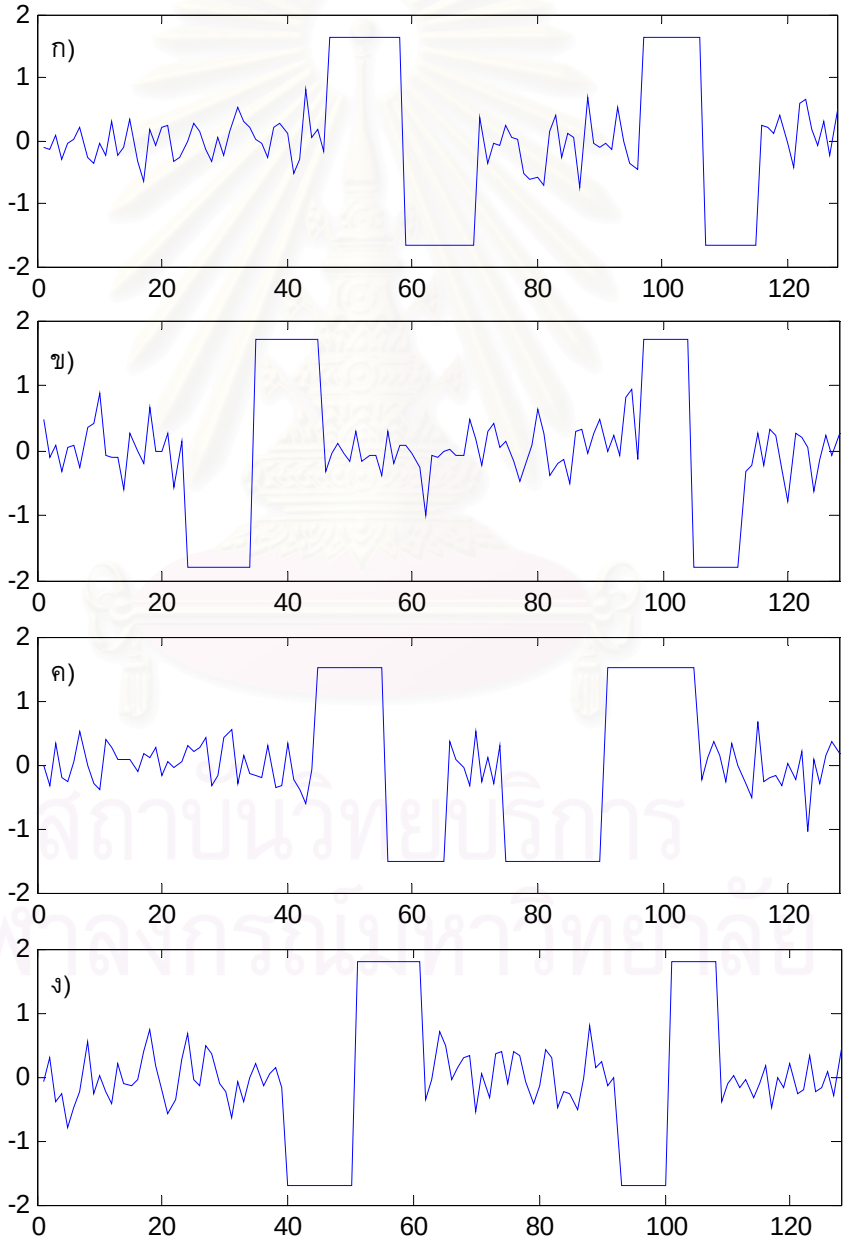
ข้อมูลอนุกรมเวลาชุดนี้เป็นข้อมูลที่สังเคราะห์ขึ้นเป็นรูปแบบสามรูปแบบที่แตกต่างกัน ข้อมูลชุดนี้ประกอบด้วยคลาส 3 คลาสคือ คลาสของข้อมูลอนุกรมเวลาที่มีรูปร่างเป็นรูปกระบอก (Cylinder) ดังรูปที่ ก.7ก) คลาสที่มีรูปร่างเป็นรูประฆัง (Bell) ดังรูปที่ ก.7ข) และคลาสที่มีรูปร่างเป็นรูปกรวย (Funnel) ดังรูปที่ ก.7ค) ข้อมูลชุดนี้ประกอบด้วยข้อมูลทั้งหมด 930 ตัว โดยข้อมูลแต่ละตัวมีความยาวของข้อมูล 128 จุด ลักษณะของข้อมูลกระบอก ระฆัง และกรวย ในแต่ละคลาสเป็นดังรูปที่ ก.7



รูปที่ ก.7 ข้อมูลซีบีเอฟในแต่ละคลาส

ก.8 ข้อมูลสองรูปแบบ (Two pattern dataset)

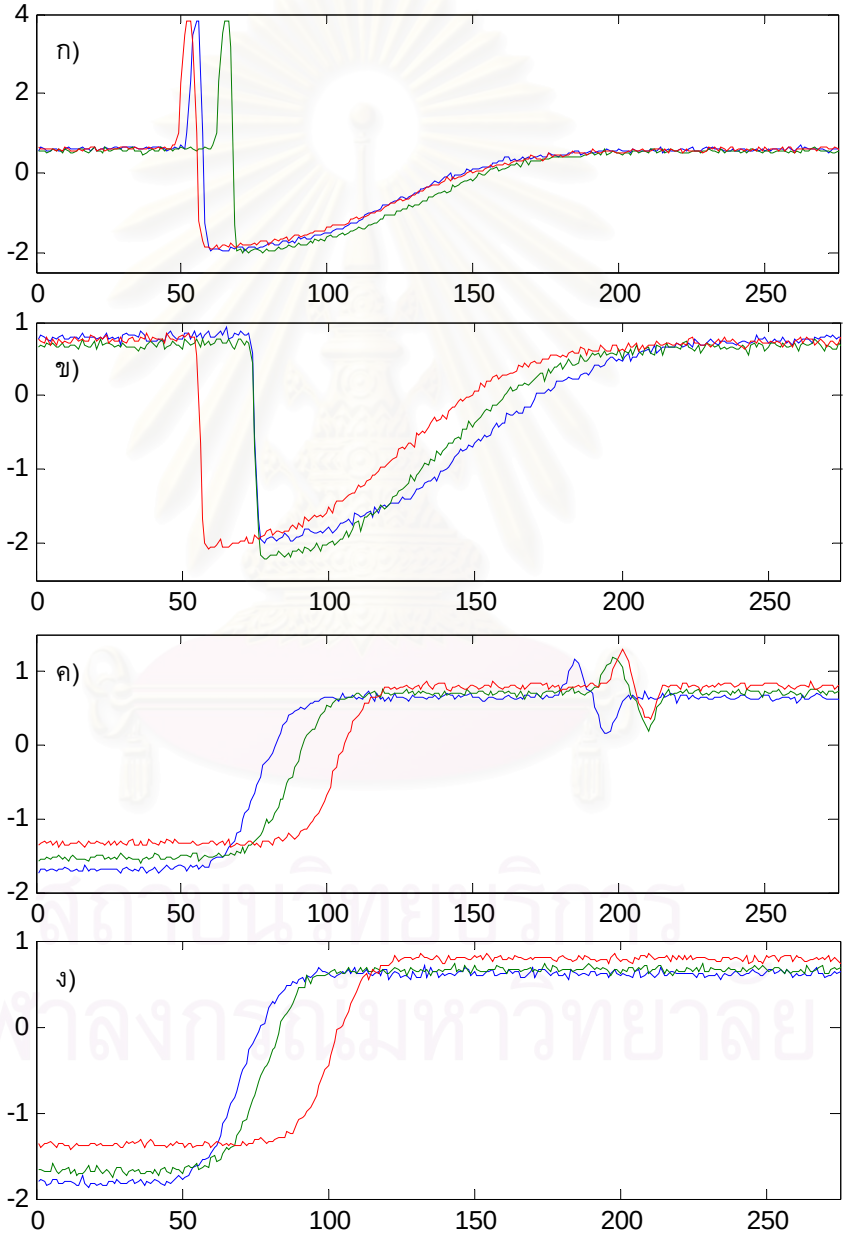
ข้อมูลอนุกรมเวลาชุดนี้เป็นข้อมูลที่มีรูปแบบสองรูปแบบอยู่ในข้อมูลอนุกรมเวลาตัวเดียวกัน ข้อมูลชุดนี้ประกอบด้วยคลาส 4 คลาสคือ คลาสที่มีรูปแบบของคลื่นสี่เหลี่ยม (Square wave) เป็นแบบลง-ลง ดังรูปที่ ก.8ก) ขึ้น-ลง ดังรูปที่ ก.8ข) ลง-ขึ้น ดังรูปที่ ก.8ค) และขึ้น-ขึ้น ดังรูปที่ ก.8ง) ข้อมูลชุดนี้ประกอบด้วยข้อมูลทั้งหมด 5,000 ตัว โดยข้อมูลแต่ละตัวมีความยาวของข้อมูล 128 จุด ลักษณะของข้อมูลสองรูปแบบในแต่ละคลาสเป็นดังรูปที่ ก.8



รูปที่ ก.8 ข้อมูลสองรูปแบบในแต่ละคลาส

ก.9 ข้อมูลนิวเคลียร์เทรซ (Trace dataset)

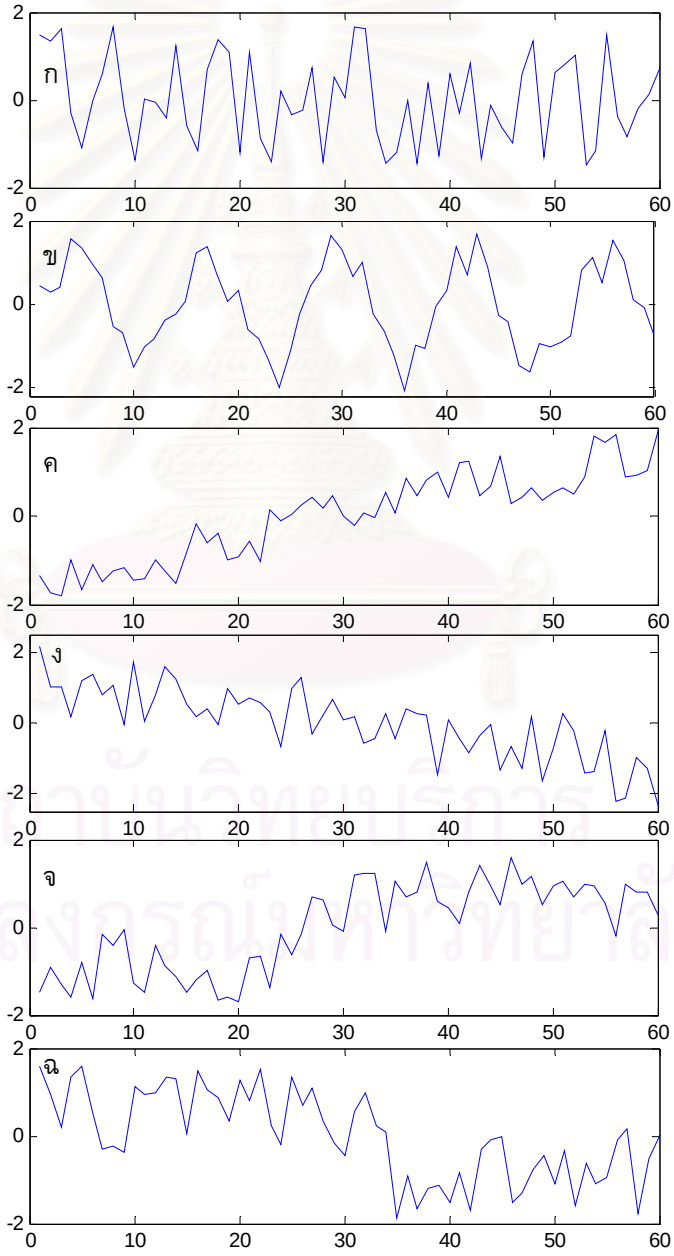
ข้อมูลอนุกรมเวลาชุดนี้เป็นข้อมูลที่ได้รับการสังเคราะห์ขึ้นที่ออกแบบมาเพื่อจำลองความผิดพลาดของเครื่องมือในโรงงานไฟฟ้านิวเคลียร์ ข้อมูลชุดนี้ประกอบด้วยคลาส 4 คลาส มีจำนวนทั้งหมด 200 ตัว โดยข้อมูลแต่ละตัวมีความยาวของข้อมูล 275 จุด ลักษณะของข้อมูลความผิดพลาดในโรงงานไฟฟ้านิวเคลียร์ในแต่ละคลาสเป็นดังรูปที่ ก.9



รูปที่ ก.9 ข้อมูลนิวเคลียร์เทรซในแต่ละคลาส

ก.10 ข้อมูลสังเคราะห์เพื่อการทดลอง (Synthetic control dataset)

ข้อมูลอนุกรมเวลาชุดนี้สังเคราะห์ขึ้นโดยที่มีรูปแบบ 6 รูปแบบที่แตกต่างกัน ข้อมูลชุดนี้ประกอบด้วยคลาส 6 คลาสคือ ข้อมูลที่มีรูปแบบปกติ (Normal) ดังรูปที่ ก.8ก) เป็นวงกลม (Cyclic) ดังรูปที่ ก.8ข) มีแนวโน้มเพิ่มขึ้น (Increasing trend) ดังรูปที่ ก.8ค) มีแนวโน้มลดลง (Decreasing trend) ดังรูปที่ ก.8ง) เพิ่มระดับขึ้น (Upward shift) ดังรูปที่ ก.8จ) และลดระดับลง (Downward shift) ดังรูปที่ ก.8ฉ) ข้อมูลชุดนี้ประกอบด้วยข้อมูลทั้งหมด 600 ตัว โดยข้อมูลแต่ละตัวมีความยาวของข้อมูล 60 จุด ลักษณะของข้อมูลในแต่ละคลาสเป็นดังรูปที่ ก.10



รูปที่ ก.10 ข้อมูลสังเคราะห์เพื่อการทดลองในแต่ละคลาส

ภาคผนวก ข

ผลงานตีพิมพ์

งานประชุมวิชาการ “11th National Computer Science and Engineering Conference (NCSEC 2007)” ซึ่งจัดขึ้น ณ โรงแรมมิราเคิลแกรนด์ กรุงเทพมหานคร ประเทศไทย ระหว่างวันที่ 19-21 พฤศจิกายน 2550 ในหัวเรื่อง “การหาเกณฑ์หยุดสำหรับตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบกึ่งมีผู้สอน” หน้า 588-595 โดย เดชาวุฒิ วานิชสรรพ์ และโชติรัตน์ รัตนามหัทธนะ

งานประชุมวิชาการ “11th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (KES 2007)” ซึ่งจัดขึ้นที่เมือง Vietri sul Mare ประเทศอิตาลี ระหว่างวันที่ 12-14 กันยายน 2550 ในหัวเรื่อง “Hand Geometry Verification using Time Series Representation” หน้า 824-831 โดย วิชญ์ เนียรนาทตระกูล เดชาวุฒิ วานิชสรรพ์ และ โชติรัตน์ รัตนามหัทธนะ

การหาเกณฑ์หยุดสำหรับตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบกึ่งมีผู้สอน
Finding Stopping Criterion for Semi-Supervised Time Series Classification

เดชาวุฒิ วานิชสรรพ์ และโชติรัตน์ รัตนามหัทธนะ
ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ถนนพญาไท แขวงวังใหม่ เขตปทุมวัน กรุงเทพมหานคร 10330 ประเทศไทย
Email: {g49dwn, ann}@cp.eng.chula.ac.th

บทคัดย่อ

การสร้างตัวจำแนกคลาสสำหรับข้อมูลอนุกรมเวลาให้สามารถจำแนกคลาสได้อย่างมีประสิทธิภาพจะต้องอาศัยข้อมูลที่เราทราบคลาสเป็นจำนวนมาก ซึ่งข้อมูลประเภทนี้มีอยู่อย่างจำกัด ในขณะที่ข้อมูลที่ไม่ทราบคลาสนั้นมีอยู่ทั่วไป จึงได้มีงานวิจัยที่นำเสนอการเรียนรู้แบบกึ่งมีผู้สอนด้วยการเรียนรู้ด้วยตนเองที่สามารถสร้างตัวจำแนกคลาสที่ดีแม้ว่าจะใช้ข้อมูลที่ทราบคลาสจำนวนไม่มาก อย่างไรก็ตามการเรียนรู้ประเภทนี้มีข้อจำกัดเกี่ยวกับการหาเกณฑ์หยุด ทำให้ได้ผลการจำแนกคลาสที่ไม่ดีเท่าที่ควร งานวิจัยนี้ได้พัฒนาการหาเกณฑ์หยุดโดยใช้ค่าระยะทางที่เปลี่ยนแปลงสำหรับการจำแนกคลาสข้อมูลอนุกรมเวลาด้วยการเรียนรู้แบบกึ่งมีผู้สอน ซึ่งจากผลการทดลองแสดงให้เห็นว่าตัวจำแนกคลาสที่สร้างจากเกณฑ์หยุดที่เสนอนั้นสามารถจำแนกคลาสได้ด้วยความถูกต้องแม่นยำมากกว่าการใช้เกณฑ์หยุดแบบเดิม

คำสำคัญ : การเรียนรู้แบบกึ่งมีผู้สอน วิธีการฝึกสอนด้วยตนเอง ข้อมูลอนุกรมเวลา การจำแนกคลาสข้อมูล ไดนามิกไทม์วอร์ปปิง

Abstract

Building a good time series classifier necessarily requires a large amount of labeled data. In reality, labeled training data may be difficult to obtain and unlabeled data is plentiful. Many researchers proposed Semi-Supervised learning Methods with Self training, which can build satisfactory classifiers by using only small amount of labeled data. However, the main limitation of the current method is the way to determine optimal stopping criterion. In this work, we propose a novel stopping criterion for semi-supervised time series classification. The experimental results show that our approach can build a better classifier that has higher classification accuracy than the current approach.

Key Words: Semi-Supervised Learning, Self-Training Method, Time Series, Classification, Dynamic Time Warping

1. บทนำ

ข้อมูลอนุกรมเวลา (Time Series) นั้นมีอยู่ในทุกหนทุกแห่ง สามารถพบได้ทั่วไปในชีวิตประจำวัน และเป็นที่สนใจในหลาย ๆ วงการ เช่น วงการแพทย์ (ความดันโลหิต คลื่นหัวใจ) และวงการธุรกิจ (ข้อมูลดัชนีหุ้นในตลาดหลักทรัพย์) นอกจากนี้ยังมีการนำสื่อประสม (Multimedia) เช่น รูปภาพ และวีดีโอ มาแปลงเป็นข้อมูลอนุกรมเวลาอีกด้วย [12] และปัญหาที่น่าสนใจสำหรับข้อมูลอนุกรมเวลาในการทำเหมืองข้อมูล (Data Mining) คือ ปัญหาการ

จำแนกคลาสข้อมูล (Classification) โดยมีจุดประสงค์เพื่อจำแนกคลาสที่ถูกต้องให้กับข้อมูลที่ยังไม่ทราบคลาสมาก่อน ด้วยตัวจำแนกคลาส (Classifier) ซึ่งใช้ข้อมูลที่ทราบคลาส (Labeled Data) จำนวนหนึ่งเพื่อนำมาฝึกสอน (Training) ด้วยเกณฑ์การเรียนรู้บางอย่าง ตัวอย่างของการจำแนกคลาสข้อมูลอนุกรมเวลาที่พบในชีวิตประจำวัน เช่น การจำแนกคลาสข้อมูลคลื่นหัวใจ โดยจำแนกเป็นคลาสของคลื่นหัวใจที่เต้นเป็นปกติและผิดปกติ

การจำแนกคลาสข้อมูลโดยทั่วไปเป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) ซึ่งใช้เฉพาะข้อมูลที่ทราบคลาสในการสร้างตัวจำแนกคลาส แต่บางครั้งข้อมูลที่เราทราบคลาสนั้นมีจำนวนน้อยเกินไปและข้อมูลส่วนใหญ่ที่มี เป็นข้อมูลที่เราไม่ทราบคลาส (Unlabeled Data) ทำให้ได้ผลการจำแนกที่ไม่ดีพอ จึงได้มีการพัฒนาการเรียนรู้แบบกึ่งมีผู้สอน (Semi-Supervised Learning) [5][7][19] เพื่อแก้ปัญหากรณีข้อมูลที่ทราบคลาสมีจำนวนไม่มาก

การเรียนรู้แบบกึ่งมีผู้สอน สร้างตัวจำแนกคลาสโดยใช้ข้อมูลที่ทราบและไม่ทราบคลาสมาทำการฝึกสอนร่วมกัน หากได้รับการฝึกสอนด้วยวิธีที่เหมาะสมกับชนิดของข้อมูลแล้ว จะทำให้ผลการจำแนกคลาสมีความถูกต้องแม่นยำสูงกว่าการที่เราทำการฝึกสอนด้วยข้อมูลที่ทราบคลาสเพียงอย่างเดียว [5][6][19] การเรียนรู้แบบนี้มีจุดเด่นอยู่ที่การใช้ข้อมูลที่ทราบคลาสจำนวนไม่มาก ในปัจจุบันนี้มีเทคนิคการเรียนรู้แบบกึ่งมีผู้สอนหลายวิธี สำหรับข้อมูลอนุกรมเวลาแล้ว วิธีที่น่าสนใจคือ วิธีการฝึกสอนด้วยตนเอง (Self Training) วิธีการนี้จะทำการฝึกสอนด้วยการเพิ่มจำนวนข้อมูลที่ทราบคลาส ด้วยข้อมูลที่ไม่ทราบคลาสที่มีความคล้ายคลึงมากที่สุด หากมีการเพิ่มข้อมูลที่ผิดคลาสระหว่างทำการฝึกสอน จะส่งผลให้ตัวจำแนกที่ได้มานั้นจำแนกคลาสผิดพลาดได้ [8] Li Wei et al. [15] ได้เสนอวิธีการฝึกสอนด้วยตนเองกับข้อมูลอนุกรมเวลา โดยวัดความคล้ายคลึงของข้อมูลด้วยวิธีวัดระยะทางแบบยุคลิด (Euclidean Distance) และหาเกณฑ์หยุด (Stopping Criterion) เพื่อเลือกข้อมูลที่จะนำมาสร้างตัวจำแนกคลาสที่เหมาะสม แต่ในบางกรณี เกณฑ์หยุดที่หาได้นั้นไม่สามารถสร้างตัวจำแนกที่สามารถจำแนกคลาสได้อย่างถูกต้อง และการวัดความคล้ายคลึงของข้อมูลอนุกรมเวลาด้วยวิธีวัดระยะทางแบบยุคลิดนั้น อาจทำให้เกิดการเพิ่มข้อมูลที่ผิดคลาสขณะทำการฝึกสอนได้ นอกจากนี้แล้วยังมีกรณีที่พบเกณฑ์หยุดหลายเกณฑ์ทำให้ไม่สามารถเลือกได้ว่าเกณฑ์ใดควรเป็นเกณฑ์หยุดที่เหมาะสมสำหรับสร้างตัวจำแนกคลาส

จากการที่เกณฑ์หยุดสร้างตัวจำแนกที่ไม่ดีนัก งานวิจัยนี้จึงได้เสนอวิธีการหาเกณฑ์หยุดแบบใหม่ เพื่อให้ได้เกณฑ์หยุดที่เหมาะสมสำหรับสร้างตัวจำแนกคลาสที่ดี ซึ่งจะช่วยแก้ปัญหาในกรณีที่พบเกณฑ์หยุดหลายเกณฑ์อีกด้วย นอกจากนี้ การวัดความคล้ายคลึงระหว่างข้อมูลอนุกรมเวลาจะใช้วิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง (Dynamic Time Warping Distance) ซึ่งวิธีนี้จะให้ผลการจำแนกคลาสที่ถูกต้องกว่าวิธีวัดระยะทางแบบยุคลิด [2][10][11] ซึ่งจะช่วยแก้ปัญหาการเลือกข้อมูลระหว่างทำการฝึกสอนได้

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ข้อมูลอนุกรมเวลา (Time Series Data) หมายถึง เซตของข้อมูลที่มีค่าเป็นจำนวนจริงที่เกิดขึ้นตามเวลาที่เปลี่ยนแปลงไป เช่น ข้อมูลคลื่นหัวใจ รายรับรายจ่ายต่อเดือน ข้อมูลเสียง และสต็อกสินค้า นอกเหนือจากข้อมูลเหล่านี้แล้ว ยังมีการนำสื่อประสมมาแปลงเป็นข้อมูลอนุกรมเวลา [12] เช่น ลายมือ รูปภาพ และวีดีโอ อีกด้วย การวัดความคล้ายคลึงระหว่างข้อมูลอนุกรมเวลานั้นสามารถทำได้ด้วยการใช้วิธีวัดระยะทาง เช่น การวัดระยะทางแบบยุคลิด และการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง ซึ่งเป็นวิธีที่นำมาใช้งานวิจัยนี้ เป็นที่รู้กันว่าการวัดระยะทางแบบยุคลิดเป็นการคำนวณค่าระยะทางระหว่างข้อมูลในตำแหน่งที่ตรงกัน แต่การวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงจะอนุญาตให้คำนวณค่าระยะทางแบบไม่เป็นเชิงเส้น (Non-Linear Alignment)

และเลือกเส้นทางการวอร์ปที่ดีที่สุดเพื่อให้ได้ค่าระยะทางที่น้อยที่สุด สมมติให้ข้อมูลอนุกรมเวลา $Q = q_1, q_2, \ldots, q_m$ และ $C = c_1, c_2, \ldots, c_n$ ซึ่งมีความยาวของข้อมูล m และ n ตามลำดับ การคำนวณหาเส้นทางการวอร์ปจะเริ่มจากการสร้างเมทริกซ์ระยะทางขนาด m x n ขึ้น โดยที่เมทริกซ์ (i,j) คือค่าระยะทางสะสม ซึ่งคำนวณจากผลรวมระหว่าง ค่าระยะทางของเมทริกซ์ (i,j) และค่าระยะทางที่น้อยที่สุดของเซลล์ที่อยู่ข้างเคียงเมทริกซ์ (i,j) ดังสมการที่ (1)

$$e(i,j) = d(i,j) + \min\{e(i-1,j), e(i,j-1), e(i-1,j-1)\} \quad \textbf{(1)}$$

โดยที่ $e(i,j)$ คือ ผลรวมของค่าระยะทางสะสม $d(i,j) = (q_i - c_j)^2$ ซึ่งเส้นทางที่เหมาะสมที่สุด (Optimal Path) คือเส้นทางที่ทำให้เกิดค่าสะสมในเมทริกซ์เซลล์ (i,j) ใด ๆ ที่ให้ค่าระยะทางที่น้อยที่สุด ดังสมการที่ (2)

$$DTW(Q,C) = \min_{\forall w \in P} \left\{ \sqrt{\sum_{k=1}^{K} d_{w_k}} \right. \quad \textbf{(2)}$$

โดยที่ w คือ เส้นทางการวอร์ปเส้นทางหนึ่ง P คือเซตของเส้นทางที่เป็นไปได้ทั้งหมด K คือความยาวของเส้นทางที่ทำการวอร์ป และ d_{w_k} คือ ค่าระยะทางของเส้นทางการวอร์ป w ในลำดับที่ k แต่อย่างไรก็ตาม เส้นทางการวอร์ปที่ทำให้เกิดค่าระยะทางที่น้อยที่สุดนั้นอาจรวมเส้นทางการวอร์ปที่ไม่มีสมเหตุสมผลเข้าไปด้วย ซึ่งเป็นผลทำให้การจำแนกคลาสไม่ถูกต้อง ดังนั้นเราควรบังคับเส้นทางการวอร์ปให้อยู่ภายในขอบเขตที่เรากำหนดไว้ โดยขอบเขตนั้นเรียกว่าเงื่อนไขบังคับโดยรวม ซึ่งทำให้ความถูกต้องในการจำแนกคลาสข้อมูลสูงขึ้นด้วย [11]

การคำนวณค่าระยะทางระหว่างข้อมูลอนุกรมเวลาด้วยวิธีการต่าง ๆ ได้รับการนำมาใช้ในการทำเหมืองข้อมูลอนุกรมเวลาในด้านต่าง ๆ เช่น การจำแนกคลาส (Classification) การจัดกลุ่ม (Clustering) การสืบหาค่าผิดปกติ (Anomaly Detection) สำหรับการจำแนกคลาสข้อมูลแล้วข้อมูลที่นำมาฝึกสอนสำหรับสร้างตัวจำแนกคลาสเป็นข้อมูลที่ทราบคลาส (Labeled Data) ซึ่งเป็นข้อมูลที่หาได้ยาก ในขณะที่ข้อมูลที่เราไม่ทราบคลาส (Unlabeled Data) นั้นมีอยู่ทั่วไป จึงได้มีเทคนิคการเรียนรู้แบบกึ่งมีผู้สอนขึ้น (Semi-Supervised Learning) ซึ่งจะสร้างตัวจำแนกคลาสด้วยการใช้ทั้งข้อมูลที่ทราบคลาสและไม่ทราบคลาสมาทำการฝึกสอนร่วมกัน ปัจจุบันนี้มีเทคนิคการเรียนรู้แบบกึ่งมีผู้สอนหลายวิธี และแบ่งเป็นห้าประเภทหลัก [5][7][19] คือ โมเดลแบบเพิ่มพูน (Generative Model) [9] วิธีการเชิงกราฟ (Graph Based Method) [3] วิธีการการแบ่งบริเวณที่ความหนาแน่น (Density Based Approach) [1] วิธีการฝึกสอนร่วมกัน (Co-Training) [4] และวิธีการฝึกสอนด้วยตนเอง (Self Training) [8][14][18]

งานวิจัยนี้จะสนใจการเรียนรู้แบบกึ่งมีผู้สอนด้วยวิธีการฝึกสอนด้วยตนเองเพราะเป็นวิธีที่พิจารณาความคล้ายคลึงระหว่างข้อมูลซึ่งเหมาะสมกับข้อมูลอนุกรมเวลาที่หาความคล้ายคลึงระหว่างข้อมูลได้ด้วยวิธีการวัดระยะทาง วิธีการฝึกสอนด้วยตนเองเป็นการเรียนรู้ที่มีความตรงไปตรงมา และมีอัลกอริทึมดังนี้

อัลกอริทึม : การฝึกสอนด้วยตนเอง

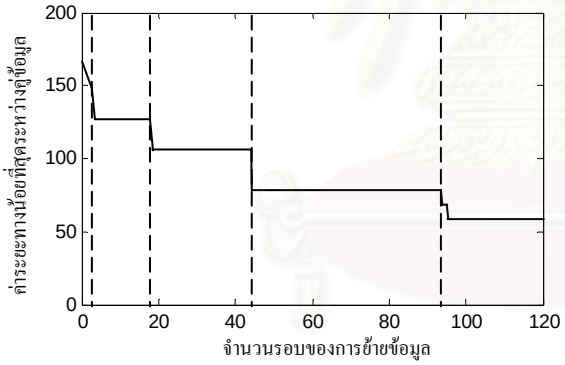
1. สร้างตัวจำแนกคลาสจากกลุ่มข้อมูลที่ทราบคลาส
2. ใช้ตัวจำแนกคลาสที่ได้จากขั้นตอนที่ 1 จำแนกคลาส กลุ่มข้อมูลที่ไม่ทราบคลาส
3. เลือกข้อมูลที่มีความคล้ายมากที่สุดจากการจำแนกในขั้นตอนที่ 2 แล้วเพิ่มเข้ากับกลุ่มข้อมูลที่ทราบคลาส
4. สร้างตัวจำแนกคลาส โดยรวมข้อมูลจากขั้นตอนที่ 3
5. ทำขั้นตอนที่ 2 ถึง 4 ซ้ำ จนกระทั่งได้ตัวจำแนกคลาสที่เหมาะสมสำหรับการจำแนกข้อมูล

รูปที่ 1 อัลกอริทึมทั่วไปของวิธีการฝึกสอนด้วยตนเอง

การเริ่มฝึกสอนช่วงแรกนั้นจะสร้างตัวจำแนกคลาสด้วยการใช้ข้อมูลที่ทราบคลาสซึ่งมีอยู่จำนวนไม่มาก และใช้ตัวจำแนกที่ได้ไปจำแนกข้อมูลที่ไม่ทราบคลาส จากนั้นย้ายข้อมูลที่ไม่ทราบคลาสที่มีความคล้ายกับข้อมูลที่ทราบคลาสมากที่สุด ไปรวมกับข้อมูลชุดที่ฝึกสอนก่อนหน้า การที่เราจะทราบว่าตัวจำแนกคลาสที่เราได้มานั้นมีความเหมาะสมแล้วหรือยังนั้น จำเป็นต้องคำนวณหาค่าเกณฑ์หยุดซึ่งเป็นค่าที่ใช้เพื่อบอกว่าควรทำการฝึกสอนเพื่อเพิ่มจำนวนของเซตข้อมูลที่ทราบคลาสกี่รอบเพื่อให้ได้ตัวจำแนกคลาสที่ดี อย่างไรก็ตามวิธีการฝึกสอนด้วยตนเองมี

ข้อควรระวังสองประการคือ การเลือกข้อมูลที่ไม่ทราบคลาสเข้ามาเพื่อสร้างตัวจำแนกระหว่างที่ทำการฝึกสอน หากเลือกข้อมูลที่ผิดคลาส จะมีผลต่อการฝึกสอนในรอบต่อไปด้วย ทำให้ตัวจำแนกที่สร้างได้ให้ความแม่นยำในการจำแนกคลาสลดลง [8][14] ข้อควรระวังอีกข้อหนึ่งคือ เกณฑ์หยุดที่ไม่ดี ซึ่งจะทำให้ข้อมูลที่นำมาสร้างตัวจำแนกอาจรวมข้อมูลคลาสที่ผิดเข้ามามากเกินไปจนส่งผลให้ตัวจำแนกที่ได้ให้ผลการจำแนกคลาสที่ไม่ดีนัก

สำหรับข้อมูลอนุกรมเวลา Li Wei et al. [15] ได้เสนอวิธีการสร้างตัวจำแนกคลาสข้อมูลด้วยวิธีการฝึกสอนด้วยตนเองและวัดความคล้ายคลึงระหว่างข้อมูลด้วยการวัดระยะทางแบบยุคลิด วิธีการนี้จะทำการฝึกสอนด้วยตนเองและในขณะเดียวกัน จะบันทึกค่าระยะทางระหว่างข้อมูลที่ถูกย้ายกลุ่มในแต่ละรอบไว้สำหรับนำไปคำนวณหาเกณฑ์หยุด โดยค่าที่ถูกบันทึกไว้จะถูกนำมาหาค่าระยะทางที่น้อยที่สุดที่เกิดขึ้นตั้งแต่รอบแรกจนถึงรอบปัจจุบัน โดยกราฟค่าระยะทางที่น้อยที่สุดแสดงดังรูปที่ 2



รูปที่ 2 กราฟแสดงค่าระยะทางที่มีค่าน้อยที่สุดที่เกิดขึ้นขณะทำการฝึกสอนด้วยตัวเอง โดยแกน *y* แสดง ค่าระยะทางต่ำสุดระหว่างคู่ข้อมูล และแกน *x* แสดงจำนวนรอบของการย้ายข้อมูล

ค่าบนกราฟในรูปที่ 2 คือค่าระยะทางที่น้อยที่สุดตั้งแต่ตำแหน่งแรกจนถึงตำแหน่งปัจจุบัน และหากพิจารณาจากแนวโน้มของกราฟแล้ว บริเวณเส้นประบนกราฟคือบริเวณที่กราฟมีค่าระยะทางลดลง ซึ่งจะนำมาพิจารณาเป็นเกณฑ์หยุดต่อไป

อย่างไรก็ตาม การหาเกณฑ์หยุดด้วยการวิธีนี้มีข้อจำกัดหลายกรณี เช่น เกณฑ์หยุดสามารถเกิดขึ้นหลายจุด ดังบริเวณเส้นประในรูปที่ 2 ทำให้ไม่สามารถทราบได้ว่าจุดใดเป็นจุดที่ทำให้ตัวจำแนกคลาสให้ผลการจำแนกที่ถูกต้อง และข้อจำกัดของเกณฑ์หยุดอีกข้อหนึ่งคือ การที่กราฟระยะทางมีค่าคงที่ ซึ่งเกิดขึ้นขณะข้อมูลในคลาสที่เราไม่ต้องการถูกย้ายเข้ามายังคลาสของข้อมูลที่ทราบคลาส หากมีการฝึกสอนต่อไปเรื่อย ๆ ข้อมูลคลาสที่เราไม่ต้องการ จะถูกย้ายเข้ามาเป็นส่วนหนึ่งของตัวจำแนก หากนำตัวจำแนกคลาสไปใช้ จะได้ผลการจำแนกคลาสที่มีความผิดพลาดสูง นอกจากนี้การวัดความคล้ายคลึงด้วยวิธีการวัดระยะทางแบบยุคลิด อาจทำให้เกิดปัญหาการเลือกข้อมูลที่ผิดคลาสระหว่างทำการฝึกสอนซึ่งจะส่งผลต่อประสิทธิภาพของการฝึกสอนในรอบต่อไปด้วย

3. อัลกอริทึม

จากข้อจำกัดที่กล่าวมา งานวิจัยนี้มีแนวคิดที่จะปรับปรุงกระบวนการหาเกณฑ์หยุดด้วยวิธีการใหม่เพื่อแก้ปัญหากรณีที่เกณฑ์หยุดยังไม่สามารถใช้แบ่งข้อมูลสำหรับสร้างตัวจำแนกประเภทที่ถูกต้อง และกรณีที่เกณฑ์หยุดที่หาได้มีหลายเกณฑ์ ในงานวิจัยนี้ยังนำวิธีการวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงมาใช้วัดความคล้ายคลึงระหว่างข้อมูลอนุกรมเวลาเพื่อลดปัญหาการเลือกข้อมูลที่ผิดคลาสระหว่างทำการฝึกสอนอีกด้วย

3.1 การสร้างตัวจำแนกคลาสข้อมูลอนุกรมเวลา

การสร้างตัวจำแนกคลาสจะใช้วิธีการฝึกสอนด้วยตนเอง โดยอัลกอริทึมการทำงานแสดงในรูปที่ 3 โดยกำหนดให้ P เป็นเซตของกลุ่มข้อมูลที่ทราบคลาสและเป็นกลุ่มที่เราให้ความสนใจ เรียกได้อีกแบบว่าคลาสบวก และ U เป็นเซตของกลุ่มข้อมูลที่ยังไม่ทราบคลาส เรียกได้อีกแบบว่าคลาสลบ ซึ่งจากรูปจะเห็นว่า การฝึกสอนจะเริ่มที่การแบ่งข้อมูลออกเป็นเซต P และ U จากนั้นจะมีการคำนวณหาความคล้ายระหว่างข้อมูล แต่เนื่องจากการคำนวณค่าระยะทางขณะทำการฝึกสอนด้วยตนเองนั้นมี

การคำนวณค่าระยะทางระหว่างข้อมูลคู่เดิมซ้ำกันหลายครั้ง งานวิจัยนี้จึงทำการคำนวณค่าระยะทางของข้อมูลระหว่างทั้งสองเซตแล้วทำการบันทึกค่าลงในเมทริกซ์ระยะทาง (Distance Matrix) ก่อนที่จะทำการฝึกสอนด้วยตนเอง เพื่อช่วยลดเวลาในการฝึกสอนด้วยตัวเองลง แสดงได้ดังตารางที่ 1

อัลกอริทึม : การสร้างตัวจำแนกคลาสข้อมูลอนุกรมเวลา

1. แบ่งข้อมูลออกเป็น P และ U
2. คำนวณความคล้ายระหว่างคู่ข้อมูลจาก P และ U ด้วยวิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิง และบันทึกค่าระยะทางลงในเมทริกซ์ระยะทาง
3. ย้ายข้อมูลรายการที่มีความคล้ายมากที่สุดจาก U มายัง P พร้อมทั้งบันทึกข้อมูลที่จำเป็นต่อการนำไปหาเกณฑ์หยุด
4. ตรวจสอบจำนวนข้อมูลใน U ถ้ายังมีข้อมูลอยู่ให้กลับไปทำขั้นตอนที่ 2 ซ้ำ แต่ถ้าไม่มีข้อมูลแล้วให้ไปทำขั้นตอนที่ 5
5. คำนวณหาเกณฑ์หยุดที่เหมาะสม
6. นำข้อมูลที่ย้ายเข้ามาในเซต P ที่อยู่ในลำดับก่อนเกณฑ์หยุดไปใช้เป็นตัวจำแนกคลาส

รูปที่ 3 อัลกอริทึมการสร้างตัวจำแนกคลาสข้อมูลอนุกรมเวลาด้วยวิธีการฝึกสอนด้วยตนเอง

ตารางที่ 1 : แสดงจำนวนครั้งในการคำนวณค่าระยะทางของข้อมูลระหว่างเซต *P* และ *U* ด้วยวิธีการที่ต่างกัน

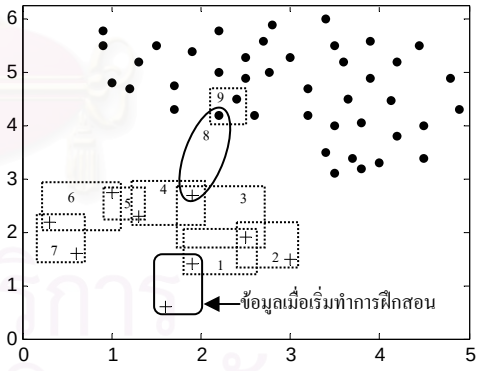
จำนวนข้อมูล	จำนวนครั้งในการคำนวณ		สัดส่วนการคำนวณเป็น % ของทั้งสองวิธี
	การฝึกสอนแบบดึงค่าจากเมทริกซ์ระยะทาง	การฝึกสอนด้วยตนเองแบบปกติ	
50	1,225	20,825	5.88%
100	4,950	166,650	2.97%
150	11,175	562,475	1.98%
200	19,900	1,333,300	1.49%

จากตารางที่ 1 จะเห็นว่าการคำนวณค่าระยะทางด้วยวิธีการฝึกสอนแบบดึงค่าจากเมทริกซ์นั้นจะช่วยลดจำนวนครั้งในการคำนวณค่าระยะทางลงอย่างมากเมื่อเทียบกับการฝึกสอนด้วยตนเองแบบปกติ เมื่อสังเกตสัดส่วนการคำนวณเป็นเปอร์เซ็นต์ของวิธีการที่เราเสนอกับวิธีการเดิม พบว่ามีสัดส่วนของการคำนวณมีค่าน้อยลงเมื่อข้อมูลมีจำนวนมากขึ้น

วิธีการฝึกสอนด้วยตนเองจะทำการฝึกสอนเพื่อเพิ่มจำนวนข้อมูลในเซต P ด้วยการย้ายข้อมูลรายการที่มีความคล้ายคลึงมากที่สุดจากเซต U และระหว่างย้ายข้อมูลนั้น จะมีการบันทึกค่าที่จำเป็นต่อการนำไปคำนวณเกณฑ์หยุด ซึ่งคือค่าระยะทางที่มากที่สุด ระยะทางที่น้อยที่สุด และระยะทางระหว่างคู่ข้อมูล การฝึกสอนจะได้รับการทำซ้ำในขั้นตอนที่ 2 ถึงขั้นตอนที่ 4 จนกระทั่งไม่มีข้อมูลเหลืออยู่ในเซต U จากนั้นจะมีการนำค่าที่บันทึกไว้ระหว่างฝึกสอนมาคำนวณหาเกณฑ์หยุดเพื่อแบ่งข้อมูลสำหรับสร้างตัวจำแนกคลาสต่อไป

3.2 การหาเกณฑ์หยุด

ในงานวิจัยนี้เราจะเสนอวิธีการหาเกณฑ์หยุดแบบใหม่ ซึ่งมีแนวคิดที่ว่า ข้อมูลที่อยู่ในคลาสเดียวกันควรมีค่าระยะทางระหว่างข้อมูลน้อย และข้อมูลที่อยู่ต่างคลาสกันควรมีค่าระยะทางระหว่างข้อมูลมาก และเมื่อพิจารณาถึงลำดับของการย้ายข้อมูลขณะทำการฝึกสอนแล้ว ช่วงที่ค่าผลต่างระหว่างค่าระยะทางของข้อมูลกับค่าระยะทางของข้อมูลในลำดับใกล้เคียง มีค่าเปลี่ยนแปลงมาก แสดงว่าข้อมูลคลาสที่เราไม่สนใจเริ่มถูกย้ายเข้ามาในคลาส P รูปที่ 4 แสดงตัวอย่างของค่าระยะทางระหว่างข้อมูลที่อยู่คลาสเดียวกันและต่างคลาสกัน



รูปที่ 4 แสดงกระบวนการฝึกสอนด้วยตนเอง โดยสัญลักษณ์บวกเป็นข้อมูลที่เราให้ความสนใจ สัญลักษณ์จุดสีดำทึบเป็นข้อมูลประเภทที่เราไม่สนใจ กรอบสี่เหลี่ยมลายเส้นประคือคู่ข้อมูลที่มีความคล้ายกันมากที่สุดที่ถูกย้ายเซตขณะทำการฝึกสอนแต่ละรอบ และข้อมูลกรอบรูปวงรีคือคู่ของข้อมูลต่างคลาสที่ถูกย้ายขณะทำการฝึกสอน

จากรูปที่ 4 บริเวณที่ถูกล้อมด้วยรูปสี่เหลี่ยมมุมมนคือข้อมูลในเซต P ขณะเริ่มทำการฝึกสอน ขณะทำการฝึกสอนในแต่ละรอบ ข้อมูลที่เราไม่ทราบคลาสจะถูกเพิ่มในเซต P ซึ่งแสดงบริเวณที่ถูกล้อมด้วยกรอบสี่เหลี่ยมลายเส้นประ โดยตัวเลขบนกรอบหมายถึงรอบที่ทำการฝึกสอน และรอบที่แปดของการฝึกสอนแสดงในบริเวณที่ถูกล้อมด้วยกรอบวงรีแสดงค่าระยะระหว่างข้อมูลที่ถูกย้ายจากเซต U ไปยังเซต P ซึ่งไม่ใช่ข้อมูลในคลาสที่เราสนใจจึงมีค่าระยะทางมาก โดยค่าระยะทางระหว่างข้อมูลที่บันทึกได้ในแต่ละรอบของการย้ายข้อมูลแสดงในตารางที่ 2

ตารางที่ 2 : ค่าระยะทางระหว่างคู่ข้อมูลที่ถูกบันทึกตามลำดับของการย้ายข้อมูลจากเซต U ไปเซต P

	รอบของการย้ายข้อมูล (Iteration)								
	1	2	3	4	5	6	7	8	9
ระยะทาง	0.7	0.5	0.7	0.6	0.3	0.8	0.5	1.2	0.3

จากตารางที่ 2 จะเห็นว่าจุดที่ควรนำมาพิจารณาเป็นเกณฑ์หยุดมีอยู่สองช่วง คือ ช่วงการย้ายข้อมูลระหว่างลำดับที่ 7-8 เพราะให้ค่าผลต่างของระยะทางระหว่างข้อมูลในลำดับที่ติดกันมีค่าเปลี่ยนแปลงไปมากคือ 1.2-0.5 = 0.7 และช่วงการย้ายข้อมูลระหว่างลำดับที่ 8-9 ที่ให้ค่าระยะทางที่เปลี่ยนแปลงไปคือ 1.2-0.3 = 0.9

อย่างไรก็ตาม การหาเกณฑ์หยุดจากค่าระยะทางที่เปลี่ยนแปลงเพียงอย่างเดียวไม่อาจทำให้ได้ค่าเกณฑ์หยุดที่ดี เพราะค่าระยะทางที่เปลี่ยนแปลงจะมีค่ามากขึ้นเรื่อย ๆ เมื่อเซต P มีขนาดใหญ่ขึ้นเสมอ จึงต้องมีการหาขอบเขตบางอย่างเพื่อใช้เป็นฐานเปรียบเทียบกับค่าระยะทางที่เปลี่ยนแปลงในบริเวณนั้นว่าควรจะนำมาใช้สำหรับเป็นเกณฑ์หยุดหรือไม่ ค่าขอบเขตนั้นเราจะคำนวณจากค่าผลต่างระหว่างระยะทางที่มากที่สุดและค่าระยะทางที่น้อยที่สุด โดยเราสามารถหาค่าเกณฑ์หยุดได้จากสมการที่ (3)

$$Confidence(i) = \frac{\left|Move(i) - Move(i-1)\right|}{Maximum(i) - Minimum(i)} \times 100 \quad \textbf{(3)}$$

โดยที่ i คือค่าลำดับที่ของการย้ายข้อมูล ค่า $Confidence(i)$ คือค่าเปอร์เซ็นต์ของความเชื่อมั่นที่บอกว่าบริเวณลำดับที่ i มีเหมาะสมที่จะใช้เป็นเกณฑ์หยุดหรือไม่ ค่า $Move(i)$ คือค่าระยะทางระหว่างคู่ข้อมูลที่ถูกย้ายเข้ามาในรอบที่ i โดยที่ $Move(0) = Move(1)$ ค่า $Maximum(i)$ และ $Minimum(i)$ คือค่าระยะทางระหว่างคู่ข้อมูลที่มีค่ามากที่สุดและน้อยที่สุดที่เกิดตั้งแต่ตำแหน่งแรกจนถึงตำแหน่งที่ i บริเวณที่เหมาะสมที่ควรนำมาพิจารณาเป็นเกณฑ์หยุดคือ ลำดับของการย้ายของมูลที่มีค่า *Confidence* สูง

3.3 การใช้ตัวจำแนกคลาส

เมื่อได้เกณฑ์หยุดที่เหมาะสมแล้ว เราจะนำข้อมูลที่อยู่ในลำดับก่อนเกิดเกณฑ์หยุดไปสร้างตัวจำแนกคลาสข้อมูล โดยจะจำแนกคลาสด้วยการใช้วิธี one-nearest-neighbor และใช้ค่าขีดแบ่ง (Threshold) เพื่อตัดสินว่าข้อมูลที่ไม่ทราบคลาสควรถูกจำแนกคลาสเป็นประเภทเดียวกับข้อมูลในตัวจำแนกคลาสหรือไม่ ค่าขีดแบ่งสามารถคำนวณได้จากค่าที่มากที่สุดของค่าระยะทางของคู่ข้อมูลภายในตัวจำแนกคลาสดังสมการที่ (4)

$$Threshold = \text{Max}(\text{Distance}(a,b)) \mid a,b \in P,\ a \neq b \quad \textbf{(4)}$$

เมื่อ P คือเซตของข้อมูลที่อยู่ในลำดับก่อนเกิดเกณฑ์หยุด a และ b คือ สมาชิกตัวใด ๆ ที่อยู่ในเซต P โดยที่ a และ b ไม่ใช่สมาชิกตัวเดียวกัน Distance คือฟังก์ชันหาค่าระยะทาง และ Max คือฟังก์ชันหาค่ามากที่สุด

4. การทดลอง

งานวิจัยนี้ทำการทดลองโดยใช้ข้อมูล 4 ชุดข้อมูล ดังแสดงในตารางที่ 3

ตารางที่ 3 : รายละเอียดข้อมูลที่ใช้ทำการทดลอง

ข้อมูล		คลาส		รวม
		บวก	ลบ	
คลื่นหัวใจ	เซตฝึกสอน	208	602	810
	เซตทดสอบ	312	904	1,216
ลายมือ	เซตฝึกสอน	109	796	905
	เซตทดสอบ	109	796	905
ปืน	เซตฝึกสอน	27	95	122
	เซตทดสอบ	30	95	125
โยคะ	เซตฝึกสอน	156	150	306
	เซตทดสอบ	156	150	306

ตารางที่ 3 แสดงจำนวนข้อมูลในแต่ละชุดข้อมูลที่ใช้ในการทดลอง โดยข้อมูล 4 ชุด คือ ข้อมูลคลื่นหัวใจ ข้อมูลลายมือ ข้อมูลปืนและข้อมูลโยคะ ซึ่งสามารถดาวน์โหลดได้จาก [17] การประเมินผลของงานวิจัยชิ้นนี้จะใช้ค่า *F-measure* ในการวัดประสิทธิภาพของการจำแนกคลาสข้อมูล ซึ่งสามารถคำนวณได้ดังสมการที่ (5)

$$F\text{-}measure = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad \textbf{(5)}$$

ค่า *Precision* และ *Recall* ซึ่งคำนวณได้ดังสมการที่ (6) และ (7)

$$\text{Precision} = \frac{\text{จำนวนข้อมูลที่จำแนกได้ในคลาสที่สนใจ}}{\text{จำนวนข้อมูลที่จำแนกได้ทั้งหมด}} \quad \textbf{(6)}$$

$$\text{Recall} = \frac{\text{จำนวนข้อมูลที่จำแนกได้ในคลาสที่สนใจ}}{\text{จำนวนข้อมูลทั้งหมดในคลาสที่สนใจ}} \quad \textbf{(7)}$$

ในงานวิจัยนี้เราเปรียบเทียบผลการทดลองระหว่างวิธีการที่เรานำเสนอคือ การหาเกณฑ์หยุดจากค่าระยะทางที่เปลี่ยนแปลง โดยการฝึกสอนด้วยการวัดระยะทางระหว่างข้อมูลอนุกรมเวลาด้วยวิธีการแบบไดนามิกไทม์วอร์ปปิงกับวิธีการเดิม โดยข้อมูลที่นำมาทำการทดลองมีดังนี้

ข้อมูลคลื่นหัวใจเป็นข้อมูลที่อยู่ในโดเมนเกี่ยวกับการแพทย์ ข้อมูลที่บันทึกได้จะอยู่ในรูปของข้อมูลอนุกรมเวลา โดยทั่วไปแล้วการจำแนกคลาสจะให้ความสนใจในการแบ่งคลาสของข้อมูลเป็นสองคลาส คือ คลาสคลื่นของหัวใจที่เต้นแบบปกติ และคลื่นของหัวใจที่เต้นแบบไม่ปกติ

ข้อมูลลายมือเป็นข้อมูลลายมือเขียนที่เป็นรูปภาพที่นำมาแปลงเป็นข้อมูลอนุกรมเวลา ซึ่งข้อมูลโดยดั้งเดิมนั้นเป็นรูปภาพลายมือเขียนของคน โดยข้อมูลชุดนี้จะแบ่งข้อมูลเป็นสองคลาสคือ คลาสของลายมือที่เขียนคำว่า the กับคลาสของลายมือที่เขียนคำอื่น

ข้อมูลปืนเป็นข้อมูลอนุกรมเวลาที่แปลงมาจากวีดีโอของคนที่ติดเซนเซอร์ไว้ที่มือแล้วยกมือขึ้นทำท่ายิงปืน คลาสของข้อมูลชุดนี้แบ่งออกเป็นสองคลาสคือ คลาสของคนที่ถือปืนจริง ๆ แล้วทำท่ายิงปืน กับคลาสของคนที่ไม่ได้ถือปืนแต่ทำท่ายิงปืน

ข้อมูลโยคะเป็นข้อมูลที่รวบรวมท่าทางของคนที่เล่นโยคะในท่าทางต่างกัน แต่คลาสของข้อมูลชุดนี้แบ่งออกเป็นสองคลาสคือ คลาสของผู้เล่นโยคะที่เป็นเพศชายและคลาสของผู้เล่นโยคะที่เป็นเพศหญิง

เราจะเลือกสมาชิกเริ่มต้นสำหรับสร้างตัวแบบด้วยการสุ่มข้อมูล สำหรับข้อมูลคลื่นหัวใจและข้อมูลลายมือ เราทำการทดลองโดยสุ่มเลือก 10 ข้อมูลจากคลาสบวก และสำหรับข้อมูลปืนและข้อมูลโยคะนั้น เราทำการสุ่มเลือกข้อมูลเพียง 1 ข้อมูลจากคลาสบวก จากนั้นทำการทดลอง 20 รอบ แล้วจึงนำตัวแบบที่ได้ไปจำแนกข้อมูลเซตทดสอบและคำนวณค่า *Precision Recall* และ *F-measure* ซึ่งแสดงในตารางที่ 4

ตารางที่ 4 : ประสิทธิภาพของตัวจำแนกคลาส

		ข้อมูล			
		คลื่นหัวใจ	ลายมือ	ปืน	โยคะ
วิธีการเดิม	***Precision***	0.7417	0.6806	0.6849	0.6244
	Recall	0.7578	0.6278	0.6961	0.8378
	F-measure	0.7497	0.6531	0.6905	0.7155
วิธีการที่นำเสนอ	***Precision***	0.9835	0.7935	0.9630	0.8408
	Recall	0.7790	0.7726	0.8500	0.7589
	F-measure	**0.8694**	**0.7829**	**0.9030**	**0.7978**

ตารางที่ 4 แสดงค่าเฉลี่ยของ *Precision Recall* และ *F-measure* ที่คำนวณจากการทดลองทุกรอบของข้อมูลแต่ละชุด ค่าในตารางแสดงให้เห็นว่าการทดลองที่ได้จากวิธีการที่นำเสนอให้ค่า *Precision* ที่สูงกว่าในขณะที่ค่า *Recall* สูงกว่า และให้ค่า *F-measure* ที่สูงกว่ากว่าวิธีการเดิมทั้ง 4 ชุดข้อมูล ส่วนค่า *Recall* ของวิธีการที่นำเสนอให้ค่าดีกว่าวิธีการเดิมทุกชุดข้อมูลยกเว้นข้อมูลโยคะ ซึ่งได้ค่า *Recall* ที่น้อยกว่า แต่เมื่อสังเกตจากค่า *F-measure* ของข้อมูลโยคะแล้ว วิธีการที่นำเสนอยังให้ค่า *F-measure* ที่สูงกว่าวิธีการเดิม แสดงให้เห็นว่าวิธีสร้างตัวจำแนกจากเกณฑ์หยุดที่งานวิจัยนี้นำเสนอช่วยให้ตัวจำแนกสามารถทำการจำแนกประเภทได้ดียิ่งขึ้น

5. บทสรุป

การเรียนรู้แบบกึ่งมีผู้สอนด้วยวิธีการฝึกสอนด้วยตนเองนั้นจะสร้างตัวจำแนกคลาสที่ดีได้ เมื่อเกณฑ์หยุดที่ใช้สำหรับสร้างตัวจำแนกมีความเหมาะสม และการเลือกข้อมูลระหว่างทำการฝึกสอนสามารถทำได้อย่างถูกต้อง ในงานวิจัยนี้ได้นำเสนอการวิธีการหาเกณฑ์หยุดจากค่าระยะทางระหว่างข้อมูลจากค่าระยะทางที่เปลี่ยนแปลงอย่างทันที เพื่อพัฒนาเกณฑ์หยุดให้ดียิ่งขึ้น นอกจากนี้งานวิจัยนี้ยังนำวิธีวัดระยะทางแบบไดนามิกไทม์วอร์ปปิงมาใช้เพื่อพัฒนาการเลือกข้อมูลระหว่างทำการฝึกสอนให้ถูกต้องมากยิ่งขึ้น ซึ่งจากผลการทดลองแสดงว่า วิธีที่เสนอในงานวิจัยนี้สามารถสร้างตัวจำแนกที่สามารถจำแนกคลาสข้อมูลอนุกรมเวลาได้ดีขึ้น

6. เอกสารอ้างอิง

[1] Bennett, K.P. and Demiriz, A., “Semi-supervised support vector machines”, Proc. of the 1998 Conf. on Advances in neural information processing systems II, 1999, pp.368-374.

[2] Berndt, D.J. and Clifford, J., “Using dynamic time warping to find patterns in time series”, Proc. of the AAAI Workshop on Knowledge Discovery in Databases, 1994, pp.229-248.

[3] Blum, A. and Lafferty, J., “Learning from labeled and unlabeled data using graph mincuts”, Proc. of 18th Int. Conf. on Machine Learning, 2001, pp.19-26.

[4] Blum, A. and Mitchell, T., “Combining labeled and unlabeled data with co-training”, Proc. of 11th annual Conf. on Computational learning theory Madison, Wisconsin, United States, 1998, pp.92-100.

[5] Chapelle, O., Schölkopf, B. and Zien, A., “Semi-Supervised Learning”, MIT Press, Cambridge, MA, 2006.

[6] Cohen, I., Cozman, F.G., Sebe, N., Cirelo, M.C. and Huang, T.S., “Semisupervised learning of classifiers: theory,algorithms and their application to human-computer interaction”, IEEE Trans. on Pattern Analysis and Machine Intelligence, 2004, pp.1553-1567.

[7] Huang, T.-M., Kecman, V. and Kopriva, I., “Kernel Based Algorithms for Mining Huge Data Sets”, Springer-Verlag, Berlin, Heidelberg, 2006.

[8] Li, M., Zhou and Z.-H., “SETRED: self-training with editing”, Proc. of 9th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'05), 2005, pp.611-621.

[9] Nigam, K., Mccallum, A.K., Thrun, S. and Mitchell, T., “Text classification from labeled and unlabeled documents using EM”, Machine Learning, 2000, pp.103–134.

[10] Ratanamahatana, C.A. and Keogh, E., “Everything you know about Dynamic Time Warping is wrong”, Proc. of 3rd Workshop on Mining Temporal and Sequential Data, In conjunction with 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD-2004), 2004.

[11] Ratanamahatana, C.A. and Keogh, E., “Making Time-series Classification More Accurate Using Learned Constraints”, Proc. of SIAM Int. Conf. on Data Mining, 2004, pp.11-22.

[12] Ratanamahatana, C.A. and Keogh, E., “Using Relevance Feedback to Learn Both the Distance Measure and the Query in Multimedia Databases”, 9th Int. Conf. on Knowledge-Based & Intelligent Information & Engineering Systems, 2005, pp.16-23.

[13] Shahshahani, B.M. and Landgrebe, D.A., “The Effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon”, IEEE Trans. on Geoscience and Remote Sensing, 1994, pp.1087-1095.

[14] Sakoe, H. and Chiba, S. “Dynamic programming algorithm optimization for spoken word recognition”, Morgan Kaufmann, 1990.

[15] Wei, L. and Keogh, E., “Semi-supervised time series classification”, Proc. of 12th ACM SIGKDD, 2006, pp.748-753.

[16] Wei, L., Keogh, E., Herle, H.V. and Mafra-Neto, A., “Atomic Wedgie: Efficient Query Filtering for Streaming Time Series”, Proc. of 5th IEEE Int. Conf. on Data Mining, 2005, pp.490-497.

[17] Wei, L. (2006). www.cs.ucr.edu/~wli/selfTraining/

[18] Zhang, R. and Alexander, I.R. “A New Data Selection Principle for Semi-Supervised Incremental Learning”, Proc. of 18th Int. Conf. on Pattern Recognition (ICPR'06), 2006, pp.780-783.

[19] Zhu, X., “Semi-supervised learning literature survey”, Technical report, no.1530, Computer Sciences, University of Wisconsin-Madison, 2005.

Hand Geometry Verification Using Time Series Representation

Vit Niennattrakul, Dachawut Wanichsan, and Chotirat Ann Ratanamahatana

Department of Computer Engineering, Chulalongkorn University
Phayathai Rd., Pathumwan, Bangkok 10330 Thailand
{g49vnn, g49dwn, ann}@cp.eng.chula.ac.th

Abstract. Biometric authentication based on human physical traits has recently been heavily studied; these biometric sources include face, hand geometry, voice, fingerprint, iris, retina, etc. The hand geometry is one of the most conventional biometric since it is fairly easy to implement and acquire the data, comparing to other biometrics such as retina, iris, or DNA sequences. In this work, we propose a novel time series representation for hand geometry system by converting raw images into time series data, where this representation can gracefully handle variability of hand's position, translation, and rotation, especially in a peg-free system with the help of a Dynamic Time Warping similarity measure. We demonstrate the utility of our approach by implementing the real hand geometry verification/identification system, and it has proven to work effectively and competitively with low false acceptance and false rejection rates.

Keywords: Biometric, Hand geometry, Verification System, Time Series

1 Introduction

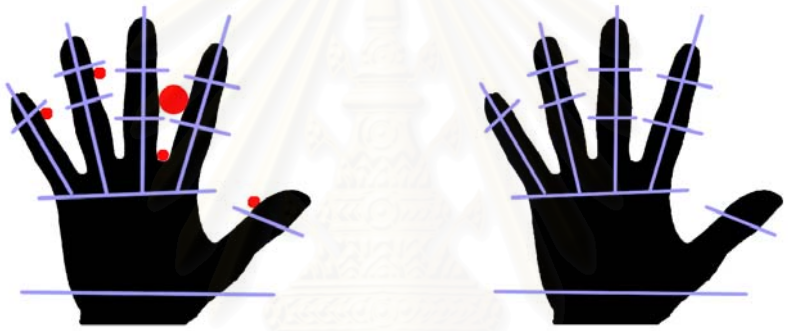
At present, biometric authentication has been widely accepted and used in place of well-known traditional methods, such as password authentication and a use of key because passwords can be forgotten and keys can be misplaced or stolen. Therefore, many biometric authentication systems have been implemented [1], e.g., face, retina, DNA, voice, iris, fingerprint, hand geometry, etc. Each biometric has different strengths and weaknesses with characteristics categorized into seven categories [3], i.e., universality, uniqueness, permanence, collectability, performance, acceptability, and circumvention. Uniqueness is perhaps one of the most important factors in a good verification and identification system; however, it also comes with a cost. Hand geometry, on the other hand, may not achieve the highest score on uniqueness, but its strong advantages include the ease to acquire data, low-cost hardware, and convenience.

In general, most hand geometry verification systems have been designed and proposed to work in either one of the two environments:- a peg-fixed system [9][13] and a peg-free system [5][6][7][11][12]. The main difference between the peg-fixed system and the peg-free system is that in the peg-fixed system, users have to align

their hand to abut against pins that are fixed on a plate, but in the peg-free system, users can place their hand freely on a surface. These systems usually utilize extracted geometric features from the hand, e.g., hand width, finger width, finger length, and fingertip regions [6][7][9][12] (shown in Figure 1).

In this paper, we propose a new representation of hand geometry; instead of several extracted features. We consider two techniques in converting shapes to time series, an angle-based technique [2] and a centroid-based technique [4]. To compare the similarity of two time series, we use Dynamic Time Warping distance measure with global constraint to control the warping path.

The rest of the paper is organized as follows. Section 2 describes our approach of converting raw images into time series data, and describes our similarity measurement and verification method. Section 3 discusses our evaluation method and shows the experimental results. Finally, in Section 4, we conclude our work and provide some suggestions.



(a) (b)

Fig. 1. (a) An image acquired from the peg-fixed system, and (b) an image acquired from the peg-free system.

2 Methodology

This section will first describe techniques we use to transform raw hand images into time series representation, and then discuss our authentication process.

2.1 Time Series Conversion

To convert an image to time series data, we need three main pre-processing steps, i.e., brightness and contrast adjustment, binarization, and edge extraction, before time series data can be achieved. Figure 2 shows the block diagram describing the processes and the output in each stage.

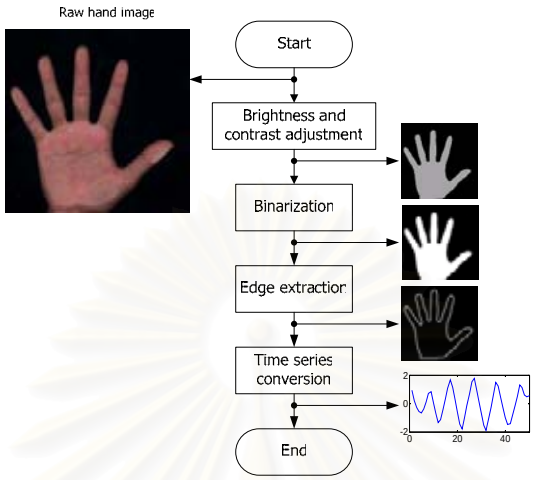


Fig. 2. A diagram describes process steps which converting raw images to time series.

Brightness and Contrast Adjustment. After an original color raw image is obtained, it will be transformed into a grayscale image. Then, the system will adjust its brightness and contrast to the quality that is suitable for the binarization step.

Binarization. After brightness and contrast adjustment, some image pixels may not be pure black or white. So, we binarize each pixel into solid black or white, represented by '0' and '1', respectively. The binarization function is defined as follows

$$B(x,y) = \begin{cases} 1 & if\ I_{xy} \geq t \\ 0 & \end{cases} \quad (1)$$

where I_{xy} is the intensity, which ranges from 0 to 1, at pixel (x, y) and t is the specific threshold for binarization. The default threshold value is set to 0.5.

Edge Extraction. The goal in this step is to find an edge sequence from a binarized hand image by using boundary extraction algorithm [2]. We consider two different approaches in this work. The traditional algorithm precisely starts the scanning of each pixel from the top left of the image to the bottom right of the image. Once it finds a first black pixel, it stops scanning, and then traces along the edge in a clockwise direction until it returns to the starting pixel. An example of an extracted edge is shown in Figure 3(a). Unlike the traditional algorithm, our proposed algorithm will not include the edge of the wrist at the bottom border of the image; our scanning starts from the bottom left to the right. After it has found the first pixel, it travels along the edge in a clockwise direction, and the algorithm stops when it touches the image boundaries again. Figure 3(b) shows an example of our extracted edge.

Fig. 3. Examples of extracted edge using (a) the original algorithm and (b) our proposal algorithm.

Time Series Conversion. In this step, we calculate the edge sequence, and transform hand's shape into time series data by using two techniques, i.e., an angle-based technique [2] and a centroid-based technique [4]. In the angle-based technique, for each pixel index *i*, we create two tangent lines – forward and backward tangents. The forward line is created by drawing the straight line from the pixel index *i* to the pixel index $i+\delta$, and the backward line is create by drawing the straight line from the pixel index *i* to the pixel index $i-\delta$. Note that the δ value depends on the size of the image; the larger the δ value the smoother the time series, and vice versa. In our experiment, the default value of δ is set to 10. After that we record the angle formed by these two lines as time series amplitude, as shown in Figure 4 (a). In the centroid-based technique, we first locate the hand's centroid. Once the centroid is obtained, we slightly plot the Euclidean distance from each pixel position to the centroid position. The general idea of centroid-based conversion is shown in Figure 4 (b).

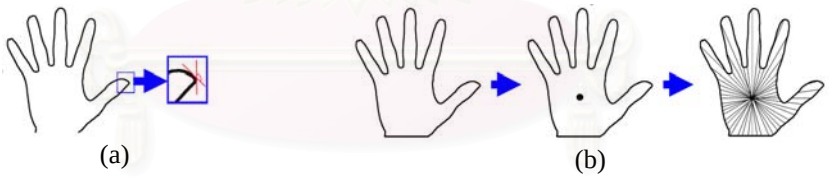


Fig. 4. The images show the general idea of (a) angle-based conversion technique and (b) controid-based conversion technique.

2.2 Finding User's Threshold and Making Decision for Authentication

Finding User's Threshold (Training Phase). To set the threshold parameter to accept or reject a new hand image inquiry, we start collecting several hand images from a user. The Dynamic Time Warping distance measure is used to calculate the distance between the new hand image inquiry and every possible pairing of the user's

templates. Once all the distance calculation is completed, the maximum value will be stored as the user's threshold.

Making Decision (Testing Phase). In verification step, once the system obtains a new hand image, it will transform the input image to time series, and will retrieve a set of that particular user's templates that he/she claims to be. After that, it will calculate the best distance between the converted time series and the set of time series templates. If this best distance is smaller than the weighted user's threshold, the access will be granted, or rejected otherwise.

Dynamic Time Warping. The Dynamic Time Warping (DTW) distance measure is used to determine the distance between two time series. Unlike the Minkowski distance function, DTW breaks the limitation of one-to-one alignment. It first finds all possible paths, and then selects the one that yields a minimum distance between the two time series using a distance matrix, where each element in the matrix is a cumulative distance of the minimum of the three surrounding neighbors. Suppose we have two time series, a sequence $Q = q_1, q_2, \ldots, q_i, \ldots, q_n$ and a sequence $C = c_1, c_2, \ldots, c_j, \ldots, c_m$. We first create an n-by-m matrix where every (i, j) element of the matrix is the cumulative distance of the distance at (i, j) and the minimum of the three elements neighboring the (i, j) element, where $0<i\leq n$ and $0<j\leq m$. We can define the (i, j) element as:

$$e_{ij} = d_{ij} + \min\{e_{(i-1)(j-1)}, e_{(i-1)j}, e_{i(j-1)}\} \quad (2)$$

where $d_{ij} = (c_i - q_j)^2$ and e_{ij} is (i, j) element of the matrix which is the summation between the squared distance of q_i and c_j, and the minimum cumulative distance of three elements surrounding the (i, j) element. Then, to find an optimal path, we have to choose the path that gives minimum cumulative distance at (n, m). The distance is defined as:

$$D_{DTW}(Q,C) = \min_{\forall w \in P}\left\{\sqrt{\sum_{k=1}^{K} d_{w_k}}\right. \quad (3)$$

where P is a set of all possible warping paths, and w_k is (i, j) at k^{th} element of a warping path and K is the length of the warping path. Due to space limitations, we refer interest readers to consult [8] for more details on DTW.

In reality, DTW may not give the best alignment that fits our purpose because it will try its best to find the minimum distance, and may generate unwanted paths. We can resolve this problem by limiting the warping path using a global constraint. A large number of researches have been adopting the Sakoe-Chiba Band [10] for the global constraint with the 10% warping window size (percentage of time series length) as a typical value. However, recent work has demonstrated that it is not always the case that using 10% band will result in best accuracies [8]. In this work, we allow the warping window size to vary from 0% to 100%.

3 Experimental Evaluation

To evaluate our proposed system, we collect hand images by using a color scanner from 22 people with 6 to 7 images for each person. The image resolution is set to 120 dpi, and its size is 1,200×1,200 pixels, as shown in Figure 5. After that, we transform all images to four time series types, i.e., two extraction methods and two conversion techniques. The length of the transformed data is between 3,000-5,000 data points, which appears to be quite oversampled. To reduce the processing time while maintaining the crucial verification features, we could downsample our time series to as short as 50 data points. Figure 6 shows examples of different combinations of transformation methods.

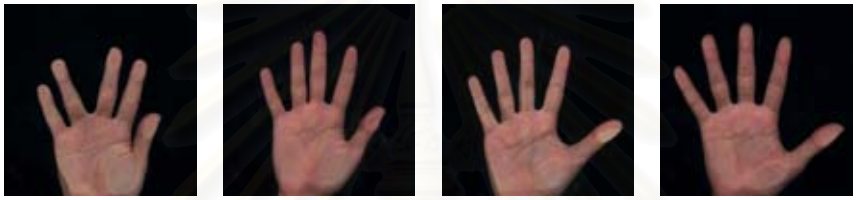


Fig. 5. Examples of four hand images from four different persons.

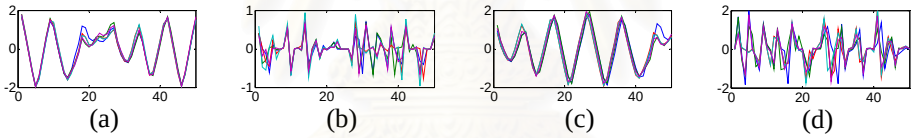


Fig. 6. Examples of four different combinations of time series transformation methods (downsampled to 50 data points) (a) traditional edge extraction with centroid-based conversion technique (b) traditional edge extraction with angle-based conversion technique (c) our proposed edge extraction with centroid-based conversion technique (d) our proposed edge extraction with angle-based conversion technique.

To evaluate the performance, we use three following error measures criteria:- False Acceptance Rate (FAR), False Rejection Rate (FRR), and Total Success Rate (TSR).

$$FRR = \frac{\#\text{RejectGenuineClaims}}{\text{Total}\#\text{GenuineAccess}} \times 100\% \quad (4)$$

$$FAR = \frac{\#\text{AcceptImposterClaims}}{\text{Total}\#\text{ImposterAccess}} \times 100\% \quad (5)$$

$$TSR = \left(1 - \frac{FAR + FRR}{\text{Total}\#\text{Access}}\right) \times 100\% \quad (6)$$

To evaluate this system, we vary three parameters, i.e., the time series transformation method, the width of the global constraint, and the weighted user's threshold. Table 1 shows the performance among four transforming combinations at EER point (Equal Error Rate - FAR≈FRR) and at the point where TSR is maximum. To further illustrate our overall system performance, we also show their ROC (Receiver Operating Characteristics) curves (shown in Figure 7.)

Table 1. The comparison of FAR, FRR, and TSR among several approaches.

	Window size (%)	Weighted user's threshold	FAR (%)	FRR (%)	TSR (%)
Traditional Extraction + Centroid-based Technique (EER)	2	0.6	17.03	16.41	98.76
Traditional Extraction + Centroid-based Technique (max TSR)	14	0.55	5.39	21.09	99.01
Proposed Extraction + Centroid-based Technique (EER)	0	0.35	24.65	28.91	98.25
Proposed Extraction + Centroid-based Technique (max TSR)	8	0.5	5.39	41.41	98.26
Traditional Extraction + Angle-based Technique (EER)	2	0.65	26.45	31.25	98.25
Traditional Extraction + Angle-based Technique (max TSR)	0	0.6	0.78	99.22	97.85
Proposed Extraction + Angle-based Technique (EER)	8	0.7	29.14	30.47	97.78
Proposed Extraction + Angle-based Technique (max TSR)	0	0.6	25.23	28.91	97.99

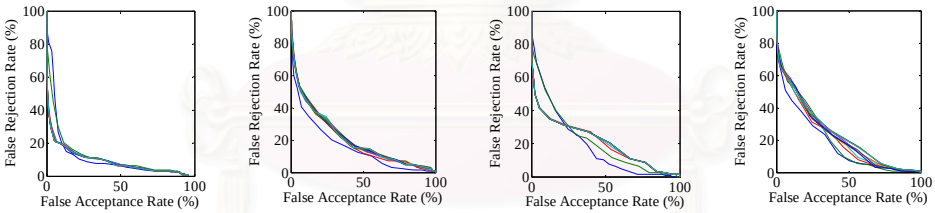


Fig. 7. ROC of (a) traditional edge extraction with centroid-based conversion technique (b) traditional edged extraction with angle-based conversion technique (c) our proposed edge extraction with centroid-based conversion technique (d) our proposed extraction with angle-based conversion technique.

For the best TSR, our system achieves 99.01% by using the traditional edge extraction and the centroid-based technique with 2% warping window and a weighted user's threshold of 0.55. An EER, the best FAR and FRR is approximately 16.50% by using the traditional edge extraction and the centroid-based technique with 2% warping window and a weighted user's threshold of 0.6. Note that the weighted user's threshold and the warping window size depends on the system environments such as image resolution, image size, and hand shapes in the database.

4 Conclusion

In this paper, we have demonstrated the utility of our novel time series representation for hand geometry system by converting raw images into time series data, where this representation can gracefully handle variability of hand's position, translation, and rotation, especially in a peg-free system with the help of a Dynamic Time Warping similarity measure. We implement the real hand geometry verification system, and it has proven to work effectively and competitively with low false acceptance and false rejection rate.

We do hope that this work will be a good motivation for exploiting time series representation in a much wider range in biometric authentication such as face recognition, fingerprinting, and especially hand geometry which allows a large reduction in computational effort and storage space.

References

1. Jain, A., Bolle, R., Pankanti, S.: Biometrics Personal Identification in Networked Society: Personal Identification in Networked Society. (1998)
2. Gandhi, A.: Content-based Image Retrieval: Plant species Identification. MS Thesis, Oregon State U. (2002)
3. Jain, A., Hong, L., Pankanti S.: Biometric identification. Communication of the ACM 43 (2000) 90-98
4. Keogh, E., Wei, L., Xi, X., Lee, S.-H., Vlachos, M.: LB_Keogh Supports Exact Indexing of Shapes under Rotation Invariance with Arbitrary Representations and Distance Measures. In Proceedings of the 32nd VLDB. (2006) 882–893
5. Kumar, A., Wong, D. C., Shen, H. C., Jain, A.K.: Personal Verification Using Palm print and Hand Geometry Biometric. In Proceedings of 4th International Conference, AVBPA 2003. (2003) 668–678
6. Kumar, A., Wong, D. C., Shen, H. C., Jain, A.K.: Personal authentication using hand images. Pattern Recognition Letters 27, 13 (2006) 1478–1486
7. Ong, M. G.-K., Connie, T., Jin, A. T.-B., Ling, D.N.-C.: A Single-Sensor Hand Geometry and Palmprint Verification System. In Proceedings of the 2003 ACM SIGMM Workshop on Biometrics Methods and Applications. (2003) 100–106
8. Ratanamahatana, C.A., Keogh, E.: Everything you know about Dynamic Time Warping is Wrong. In Proceedings of 3rd SIGKDD Workshop on Mining Temporal and Sequential Data, at 10th ACM SIGKDD (2004)
9. Sanchez-Reillo, R., Sanchez-Avila, C., Gonzalez-Marcos, A.: Biometric identification through hand geometry measurements. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.22, Iss.10, (2000) 1168- 1171
10. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing. (1978) 43-49.
11. Toh, K.A., Xiong, W., Yau, W.-Y., Jiang, X.: Combining Fingerprint and Hand-Geometry Verification Decisions. In Proceedings of 4th International Conference, AVBPA 2003. (2003) 688-696
12. Wong, A. L.-N., Shi, P.: Peg-Free Hand Geometry Recognition Using Hierarchical Geometry and Shape Matching. IAPR Workshop on Machine Vision Applications. (2002)
13. Zunkel, L. R.: Hand Geometry Based Verification. In Biometrics: Personal identification in networked society. Kluwer Academic Publishers (1999) 87–101

ประวัติผู้เขียนวิทยานิพนธ์

นายเดชาวุฒิ วานิชสรรพ์ เกิดเมื่อวันที่ 4 ธันวาคม พ.ศ.2526 ที่จังหวัดชลบุรี เป็นนักศึกษาทุนโครงการส่งเสริมการผลิตครูที่มีความสามารถพิเศษทางด้านวิทยาศาสตร์และคณิตศาสตร์ (สควค.) จนสำเร็จการศึกษาระดับปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) เกียรตินิยมอันดับ 1 จากคณะวิทยาศาสตร์และเทคโนโลยี สาขาวิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยราชภัฏพระนครศรีอยุธยา ในปีการศึกษา 2547 และสำเร็จการศึกษาประกาศนียบัตรสาขาวิชาการสอนวิทยาศาสตร์ จากมหาวิทยาลัยมหิดล ในปีการศึกษา 2548 และได้เข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2549 ระหว่างที่กำลังศึกษาอยู่ได้มีโอกาสไปนำเสนอผลงานเรื่อง “การหาเกณฑ์หยุดสำหรับตัวจำแนกคลาสข้อมูลอนุกรมเวลาแบบกึ่งมีผู้สอน” ในงานประชุมวิชาการ “11th National Computer Science and Engineering Conference (NCSEC 2007)” ซึ่งจัดขึ้น ณ โรงแรมมิราเคิลแกรนด์ กรุงเทพมหานคร ประเทศไทย ระหว่างวันที่ 19-21 พฤศจิกายน 2550 และผลงานเรื่อง “Hand Geometry Verification using Time Series Representation” ในงานประชุมวิชาการ “11th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (KES 2007)” ซึ่งจัดขึ้นที่เมือง Vietri sul Mare ประเทศอิตาลี ระหว่างวันที่ 12-14 กันยายน 2550