วิธีการเลือกข้อมูลที่ไม่มีป้ายกำกับอย่างอัตโนมัติเพื่อการสร้างโมเดลการเรียนรู้ร่วมแบบกึ่งมีผู้สอน

ศิริขวัญ คีรีสุวรรณกุล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
พ.ศ. 2563

An Automatic Unlabeled Selection for CO-training REGressors (AU-COREG)

Sirikwan Kheereesuwannakul

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering

Department of Big Data Engineering,

College of Innovative Technology and Engineering,

Dhurakij Pundit University

2020

ใบรับรองงานวิทยานิพนธ์

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

ปริญญา วิศวกรรมศาสตรมหาบัณฑิต

หัวข้อวิทยานิพนธ์ วิธีการเลือกข้อมูลที่ไม่มีป้ายกำกับอย่างอัตโนมัติเพื่อการสร้างโมเดลการเรียนรู้ร่วมแบบกึ่งมีผู้สอน

เสนอโดย ศิริขวัญ คีรีสุวรรณกุล

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

อาจารย์ที่ปรึกษาวิทยานิพนธ์ ดร.เอกสิทธิ์ พัชรวงศ์ศักดา

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบวิทยานิพนธ์แล้ว

.....ประธานกรรมการ

(ดร.สรรพฤทธิ์ มฤคทัต)

.....กรรมการและอาจารย์ที่ปรึกษาหลัก

(ดร.เอกสิทธิ์ พัชรวงศ์ศักดา)

.....กรรมการ

(ดร.ธนภัทร ฆังคะจิตร)

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น)

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์รับรองแล้ว

.....

(ผู้ช่วยศาสตราจารย์ ดร.ณรงค์เดช กีรติพรานนท์)

คณบดีวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์

วันที่ 12 เดือน มิถุนายน พ.ศ. 2563

หัวข้อวิทยานิพนธ์	วิธีการเลือกข้อมูลที่ไม่มีป้ายกำกับอย่างอัตโนมัติเพื่อการสร้างโมเดลการเรียนรู้ร่วมแบบกึ่งมีผู้สอน
ชื่อผู้เขียน	ศิริขวัญ คีรีสุวรรณกุล
อาจารย์ที่ปรึกษา	ดร.เอกสิทธิ์ พัชรวงศ์ศักดา
สาขาวิชา	วิศวกรรมข้อมูลขนาดใหญ่
ปีการศึกษา	2562

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อปรับปรุงประสิทธิภาพโมเดลพยากรณ์ ด้วยวิธีการเลือกข้อมูลที่ไม่มีป้ายกำกับอย่างอัตโนมัติ เพื่อสร้างโมเดลการเรียนรู้ร่วมแบบกึ่งมีผู้สอน ซึ่งเหมาะสำหรับข้อมูลที่มีป้ายกำกับปริมาณน้อยมาก โดยวิธีการที่นำเสนอนี้จะใช้ประโยชน์จากข้อมูลที่ไม่มีป้ายกำกับที่มีอยู่ปริมาณมากมาช่วยเพิ่มประสิทธิภาพของการสร้างโมเดลจำแนกประเภทข้อมูล หรือประมาณค่า วิธีการที่นำเสนอนี้เริ่มด้วยการใช้โมเดล 2 โมเดล ทำการกำกับค่าให้กับข้อมูลที่ไม่มีป้ายกำกับ จากนั้นนำข้อมูลเหล่านี้มาทำการจัดกลุ่มเพื่อให้ข้อมูลที่มีความคล้ายคลึงกันอยู่กลุ่มเดียวกัน และแยกข้อมูลที่ต่างกันออกให้อยู่ต่างกลุ่มกัน ถัดมาจึงเลือกตัวแทนแต่ละกลุ่มเพื่อหาข้อมูลที่ทำให้โมเดลพยากรณ์มีความคลาดเคลื่อนน้อยที่สุด เพื่อเข้าไปเป็นชุดข้อมูลสอนในรอบถัดไป และสร้างโมเดลพยากรณ์ใหม่ ทำซ้ำจนเพิ่มข้อมูลเข้าไปในชุดสอนได้ครบ จากนั้นการพยากรณ์ขั้นสุดท้ายทำได้โดยการหาค่าเฉลี่ยของค่าพยากรณ์จากทั้งสองโมเดลที่สร้างขึ้น จากการทดสอบด้วยข้อมูลจำนวน 3 ชุด แสดงให้เห็นว่าวิธีการที่นำเสนอ สามารถปรับปรุงประสิทธิภาพของโมเดลพยากรณ์ได้ และช่วยลดเวลาลงไปมากกว่า 84% เมื่อเทียบกับวิธีการเดิม

Thesis Title	An Automatic Unlabeled Selection for CO-training REGressors (AU-COREG)
Author	Sirikwan Kheereesuwannakul
Thesis Advisor	Dr.Eakasit Pacharawongsakda
Department	Big Data Engineering
Academic Year	2019

ABSTRACT

This research aims to improve the performance of semi-supervised learning by automatically select unlabeled data. The proposed method uses two regression models to estimate values for unlabeled data, then cluster the data into groups. Therefore, similar data are in the same group and the different data are into the different groups. After that, the method selects each group representative that have least error and append into training data. Then, we repeat until we have enough training data. From experimental results with three datasets, we found that the proposed method can improve performance and reduce computation time by 84%, comparing to previous work.

กิตติกรรมประกาศ

ผู้วิจัยขอกราบขอบพระคุณอย่างยิ่งในความกรุณาของท่าน และขอขอบพระคุณ ดร.ตวงทอง วัตรุจีกฤต ที่เป็นแรงบันดาลใจในการทำวิจัยครั้งนี้ นอกจากนี้ขอขอบคุณเจ้าหน้าที่และเพื่อน ๆ หลักสูตรวิศวกรรมศาสตร์มหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่ทุกท่าน สำหรับการช่วยเหลือและประสานงาน เพื่อให้การดำเนินการวิจัยเป็นไปอย่างราบรื่น และยังให้กำลังใจเสมอมาตลอดจนมหาวิทยาลัยที่ให้โอกาสในการศึกษาเรียนรู้ตามความสนใจของผู้วิจัย และสุดท้ายผู้วิจัยขอขอบพระคุณบิดามารดา และครอบครัว ซึ่งเป็นกำลังใจ และคอยสนับสนุนในทุกเรื่อง และทุก ช่วงชีวิต หากมีสิ่งใดที่ผู้วิจัยได้ทำผิดพลาดหรือบกพร่องประการใด ผู้วิจัยต้องกราบขออภัยเป็นอย่างสูงมา ณ โอกาสนี้ ผู้วิจัยหวังเป็นอย่างยิ่งว่าวิทยานิพนธ์ฉบับนี้จะเป็นพื้นฐานในการต่อยอดองค์ความรู้ของผู้ที่สนในศึกษาในงานด้านนี้ต่อไป

ศิริขวัญ คีรีสุวรรณกุล

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญภาพ

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา (Background and Significance of the Problem)

โลกปัจจุบันได้เข้าสู่ยุคดิจิทัล (digital) ดังเห็นได้จากอุปกรณ์ต่าง ๆ ที่มีการสร้างข้อมูลในเชิงดิจิทัล เพิ่มขึ้นอย่างเป็นจำนวนมาก โดยส่วนใหญ่เกิดจากการใช้งานอินเทอร์เน็ต จึงส่งผลให้พฤติกรรมของมนุษย์มีการเปลี่ยนแปลงจากอดีต กิจกรรมหลาย ๆ อย่างถูกแทนที่ด้วยแพลตฟอร์ม (platform) บนโลกออนไลน์ เช่น การซื้อสินค้า การประกาศรับสมัครงาน การดูหนังฟังเพลง การแชท (chat) หรือโซเชียลเน็ตเวิร์ค (Social Network) เป็นต้น แพลตฟอร์มเหล่านี้ได้สร้างข้อมูลจำนวนมหาศาลบนโลกออนไลน์ รายงานไอบีเอ็ม (IBM) ระบุว่า 90 % ของข้อมูลทั้งหมดในโลกออนไลน์เพิ่งถูกสร้างขึ้นในช่วง 2 ปีหลังนี้เอง โดยปัจจุบันมีข้อมูลเกิดขึ้นใหม่ราว 2,500 ล้านจิกะไบต์ (Gigabyte) ต่อวัน

หากพิจารณาข้อมูลเหล่านั้นข้อมูลที่มีป้ายกำกับ (Labeled Data) จะมีสัดส่วนน้อยมากเมื่อเทียบกับข้อมูลที่ไม่มีป้ายกำกับ (Unlabeled Data) เนื่องจากการกำกับค่าให้ข้อมูลนั้นมีต้นทุนที่สูงหรือค่าที่กำกับไม่เป็นจริง หรืออาจไม่สามารถกำกับค่าได้ เพราะความต้องการใช้ข้อมูลมีการเปลี่ยนแปลงไปอย่างรวดเร็ว ดังนั้นการสร้างโมเดลพยากรณ์จำเป็นต้องมีข้อมูลสอน (Training Data) ซึ่งข้อมูลสอนได้จากข้อมูลที่มีป้ายกำกับ ทว่าการที่ข้อมูลสอนนี้มีสัดส่วนน้อยมากอาจจะทำโมเดลการพยากรณ์ให้ความคลาดเคลื่อนสูง เมื่อเทียบกับการมีข้อมูลที่มีป้ายกำกับที่มีมากกว่า ดังนั้นการให้โมเดลเรียนรู้แบบกึ่งมีผู้สอน (Semi-supervised Learning) จากข้อมูลทั้งมีป้ายกำกับและข้อมูลที่ไม่มีป้ายกำกับจึงถูกนำมาประยุกต์ใช้ ซึ่งวิธีที่ได้รับความนิยมอย่างแพร่หลายคือวิธีการโค เทรนนิง (Co-Training)

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 ปรับปรุงประสิทธิภาพโมเดลพยากรณ์ จากการเรียนรู้ร่วมกึ่งมีผู้สอน สำหรับข้อมูลที่มีป้ายกำกับที่มีสัดส่วนน้อย เมื่อเทียบกับข้อมูลที่ไม่มีป้ายกำกับ

1.2.2 ลดระยะเวลาในการสร้างโมเดลพยากรณ์

1.3 ประโยชน์ที่คาดว่าจะได้รับ

1.3.1 เพิ่มประสิทธิภาพการพยากรณ์ให้มีความแม่นยำมากยิ่งขึ้น

1.3.2 ได้แนวทางในการพัฒนาโมเดลพยากรณ์ เพื่อนำไปประยุกต์ใช้ในการพยากรณ์กับข้อมูลที่มีป้ายกำกับจำนวนน้อย

1.4 ขอบเขตของงานวิจัย

1.4.1 ชุดข้อมูลที่มีแอททริบิวต์จำนวนไม่มาก

1.4.2 ชุดข้อมูลโฆษณาประกาศรับสมัครงาน เป็นข้อมูลประกาศสมัครงานในประเทศอังกฤษ จัดทำโดย Adzuna

1.4.3 ชุดข้อมูลปริมาณการจราจรด้วยรถไฟฟ้าใต้ดินรายชั่วโมง ของสถานี MN DoT ATR station 301

1.4.4 ชุดข้อมูลพลังงานความร้อนร่วม โดยข้อมูลที่รวบรวมจากโรงไฟฟ้าพลังความร้อนร่วมในช่วง 6 ปี (2549-2554)

บทที่ 2
แนวคิด ทฤษฎี และผลงานวิจัยที่เกี่ยวข้อง

2.1 บทนำ

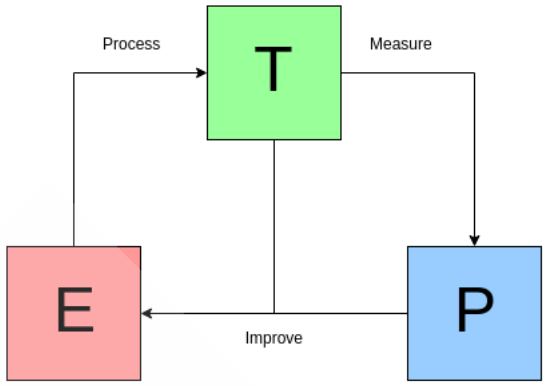
บทนี้ เป็นการทบทวนวรรณกรรม ที่เกี่ยวข้องกับการเรียนรู้ของเครื่องแบบกึ่งมีผู้สอนร่วม ซึ่งผู้วิจัยได้แบ่งการทบทวนวรรณกรรมออกเป็น 3 ตอน คือ ตอนที่หนึ่ง บทนำ ตอนที่สองเป็น แนวความคิดและทฤษฎีที่เกี่ยวข้องของคำว่าการเรียนรู้ของเครื่อง ตอนที่สามงานวิจัยที่เกี่ยวข้อง ดังต่อไปนี้

2.2 แนวคิดและทฤษฎีที่เกี่ยวข้อง

การวิจัยเรื่อง วิธีการเลือกข้อมูลที่ไม่มีป้ายกำกับอย่างอัตโนมัติ เพื่อการสร้างโมเดลการเรียนรู้ร่วมแบบกึ่งมีผู้สอน ผู้วิจัยได้ทำการศึกษา ค้นคว้า จากแหล่งความรู้ทางอินเตอร์เน็ต เช่น บทความวิชาการ บทความวิจัย และเนื้อหาด้านวิชาการ โดยงานวิจัยมีแนวคิดและทฤษฎีที่เกี่ยวข้องดังนี้

2.2.1 การเรียนรู้ของเครื่อง (Machine Learning)

Arthur Samuel ชาวอเมริกันในด้านการเล่นเกมคอมพิวเตอร์และปัญญาประดิษฐ์ได้บัญญัติศัพท์นี้ขึ้นในปี 1959 ไว้ว่า การเรียนรู้ของเครื่อง เป็นการฝึกฝนให้คอมพิวเตอร์มีความสามารถที่จะเรียนรู้โดยที่ไม่ต้องตั้งโปรแกรมให้ทำงานไว้อย่างชัดเจน จากนั้น Tom Michael Mitchell นักวิทยาศาสตร์คอมพิวเตอร์ชาวอเมริกัน และเป็นศาสตราจารย์มหาวิทยาลัย Carnegie Mellon (CMU) เป็นผู้มีชื่อเสียงในเรื่องการเรียนรู้ของเครื่อง และปัญญาประดิษฐ์ (Artificial Intelligence: AI) ได้ให้คำจำกัดความอย่างเป็นทางการในหนังสือการเรียนรู้ของเครื่องไว้ว่า การเรียนรู้ของเครื่อง คือโปรแกรมคอมพิวเตอร์ที่เรียนรู้จากประสบการณ์ (Experience: E) กับงานบางประเภท (Task: T) ได้โดยมีวัดประสิทธิผล (Performance: P) เมื่อโปรแกรมนั้นสามารถทำงาน T ที่วัดผลด้วย P แล้วพัฒนาขึ้นจากประสบการณ์ E

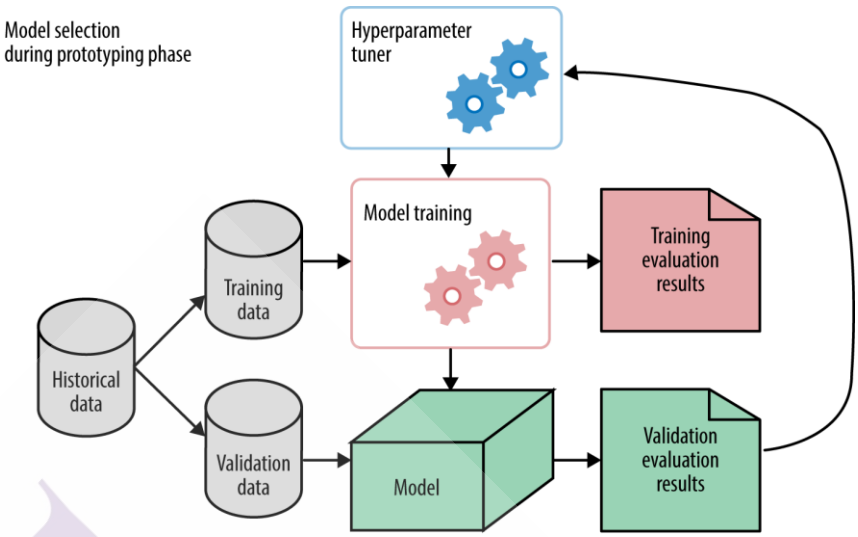


ภาพที่ 2.1 ภาพแสดงองค์ประกอบของการเรียนรู้ด้วยเครื่องตามคำจำกัดความของ Mitchell

ที่มา: A concise explanation of learning algorithms with the Mitchell paradigm
by Matthew Mayo, KDnuggets.

การสร้างโมเดลการเรียนรู้ของเครื่อง เริ่มด้วยการรวบรวมและจัดเตรียมข้อมูล โดยทำการแบ่งข้อมูลออกเป็น 2 ชุด ชุดข้อมูลสำหรับการสอน (Training Data) และชุดข้อมูลสำหรับใช้ในการตรวจสอบ (Validation Data) เพื่อเตรียมนำมาใช้ในการิเคราะห์สกัดหาความรู้ คัดเลือกและจัดการด้านคุณสมบัติของข้อมูล (Feature Selection and Feature Engineering) จากนั้นเลือกอัลกอริทึมและสร้างโมเดลการเรียนรู้ของเครื่อง (Model Training) ครั้งแรก ทำการวัดผลการประเมินการสอน (Training Evaluation Result) โดยการเลือกใช้ตัวชี้วัดที่เหมาะสมกับโมเดล จากนั้นนำโมเดลที่ได้ไปใช้ทดสอบกับชุดข้อมูลตรวจสอบ และวัดผลการทดสอบด้วยตัวชี้วัดเดียวกัน ทำการปรับแต่งโมเดล และพารามิเตอร์ต่าง ๆ และนำไปประยุกต์กับโมเดลการสอน ทำซ้ำจนกระทั่งได้โมเดลที่มีประสิทธิภาพที่ดีที่สุด

การเรียนรู้ของเครื่องมีการแบ่งประเภทการเรียนรู้ของเครื่อง แบ่งออกตามวัตถุประสงค์การใช้งาน ซึ่งประเภทหลัก ๆ มีดังนี้



ภาพที่ 2.2 ภาพแสดงขั้นตอนการสร้างโมเดลการเรียนรู้ของเครื่อง

ที่มา: Evaluating Machine Learning Models, By Alice Zheng (October 16, 2015).

2.2.1.1 ประเภทการเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised Learning)

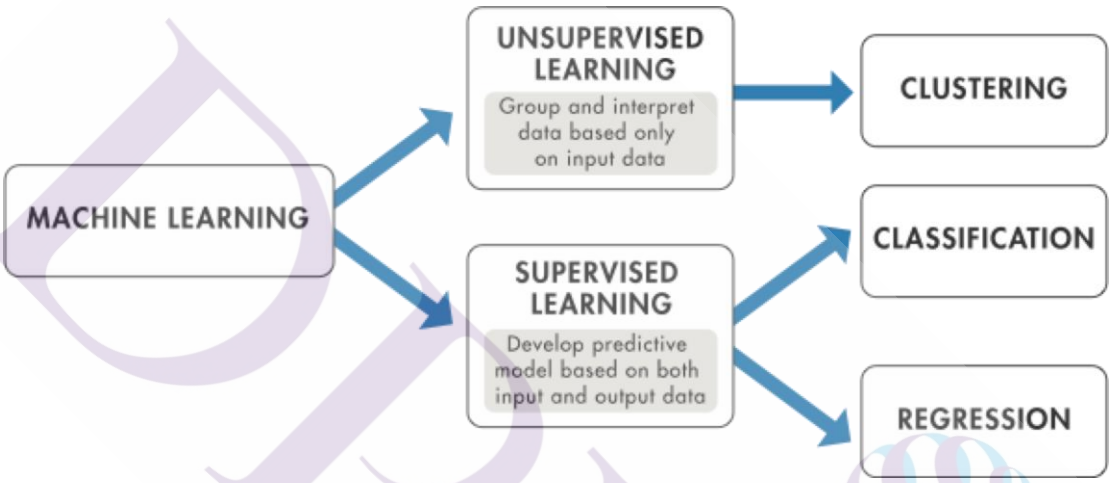
การเรียนรู้ของเครื่องประเภทการเรียนรู้แบบมีผู้สอน ถูกสร้างจากข้อมูลสอนที่มีป้ายกำกับ (Labeled Data) หรือเรียกว่าข้อมูลที่มีคำตอบ ซึ่งจะได้โมเดลการสอน แล้วนำโมเดลนั้นไปใช้ในการทำนายค่าหรือจำแนกประเภทข้อมูล และทำการทดสอบประสิทธิภาพของโมเดล (Test Model) ตัวอย่างเทคนิคการเรียนรู้ของเครื่องประเภทการเรียนรู้แบบมีผู้สอน ได้แก่ ตัวแบบการถดถอยเชิงเส้น (Linear Regression) วิธีการเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbors: kNN) โครงข่ายประสาทเทียม (Neural Networks) ต้นไม้ตัดสินใจ (Decision Tree) เป็นต้น

การทำนายค่า (Regression) วัตถุประสงค์เพื่อใช้โมเดลที่ทำงานกับข้อมูลที่มีตัวอย่างกำกับค่าด้วยข้อมูลเชิงปริมาณ (Quantitative Data) ตัวอย่างเช่น โมเดลการทำนายปริมาณน้ำฝนในช่วงเวลาต่าง ๆ โมเดลการประมาณการณ์ยอดใช้งาน Call Center ในแต่ละช่วงเวลา เป็นต้น

การจำแนกประเภท (Classification) วัตถุประสงค์เพื่อใช้โมเดลที่ทำงานกับข้อมูลที่มีตัวอย่างกำกับค่าด้วยข้อมูลเชิงคุณภาพ (Qualitative Data) ตัวอย่างเช่น โมเดลการจำแนกข้อมูลเพื่อทำนายสุขภาพของคน (ผลลัพธ์ คือ ป่วย หรือแข็งแรง) โมเดลการอนุมัติสินเชื่อโดยการวิเคราะห์ความเสี่ยงในการปล่อยกู้ (ผลลัพธ์ คือ อนุมัติ หรือปฏิเสธ) เป็นต้น

2.2.1.2 ประเภทการเรียนรู้ของเครื่องแบบไม่มีผู้สอน (Unsupervised Learning)

การเรียนรู้ของเครื่องประเภทการเรียนรู้แบบไม่มีผู้สอน การเรียนรู้ประเภทนี้แตกต่างจากการเรียนรู้แบบมีผู้สอน คือไม่มีการระบุผลที่ต้องการหรือประเภทไว้ก่อน การสร้างการเรียนรู้เกิดจากการค้นหารูปแบบของข้อมูลและจัดกลุ่มบนพื้นฐานความเหมือน (Similarities) และความแตกต่าง (Difference) ของข้อมูล โดยอยู่บนพื้นฐานของข้อมูลที่ใช้เท่านั้น นั่นคือไม่สามารถนำรูปแบบที่ได้ไปใช้ทำนายหรือจำแนกประเภทข้อมูลกับข้อมูลชุดใหม่ได้ ตัวอย่างเทคนิคการเรียนรู้ของเครื่องประเภทการเรียนรู้แบบไม่มีผู้สอน ได้แก่ การจัดกลุ่ม (Clustering) การหากฎความสัมพันธ์ (Association Rules) เป็นต้น

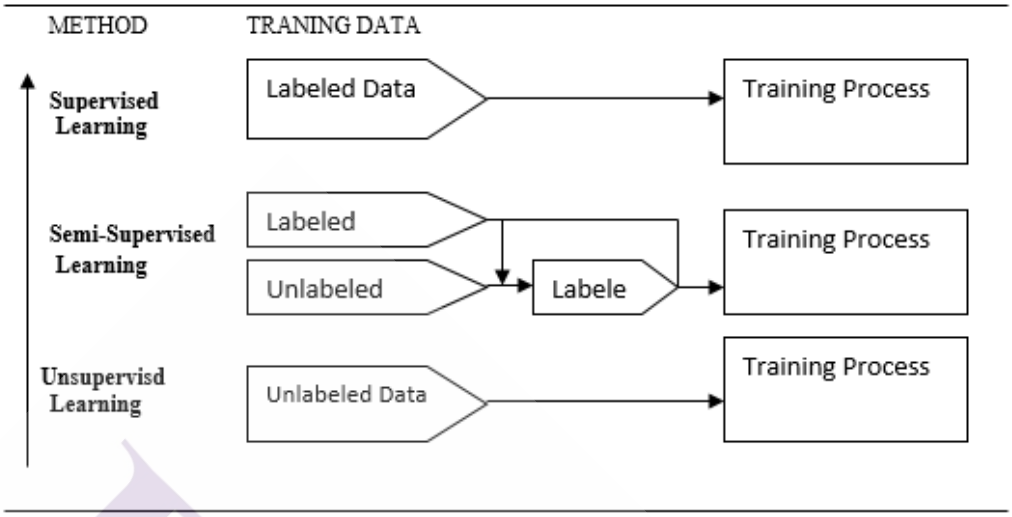


ภาพที่ 2.3 ภาพแสดงประเภทการเรียนรู้แบบมีผู้สอน และการเรียนรู้แบบไม่มีผู้สอน

ที่มา: What Is Machine Learning?, 3 things you need to know
www.mathworks.com/discovery/machine-learning

2.2.1.3 ประเภทการเรียนรู้ของเครื่องแบบกึ่งมีผู้สอน (Semi-Supervised Learning)

การเรียนรู้ของเครื่องประเภทการเรียนรู้แบบกึ่งมีผู้สอน เกิดจากการผสมผสานระหว่างการเรียนรู้ของเครื่องแบบมีผู้สอน และการเรียนรู้ของเครื่องแบบไม่มีผู้สอน โดยที่ข้อมูลสอนจะไม่สอนอย่างสมบูรณ์ นั่นคือสามารถเรียนรู้จากข้อมูลสอนที่มีคำตอบบางตัวอย่าง ตัวอย่างเทคนิคการเรียนรู้ของเครื่องประเภทการเรียนรู้แบบกึ่งมีผู้สอน คือ เทคนิคการสอนร่วม หรือโค เทรนนิง (Co-Training)



ภาพที่ 2.4 ภาพแสดงประเภทการเรียนรู้ของเครื่อง

ที่มา: Kulwinder Kaur, Machine Learning Techniques, Data Mining, Weka
http://www.e2matrix.com/blog/2018/01/24/machine-learning-techniques

2.2.2 เทคนิคการเรียนรู้ของเครื่อง (Machine Leaning Technique)

2.2.2.1 การวิเคราะห์การถดถอย (Regression Analysis)

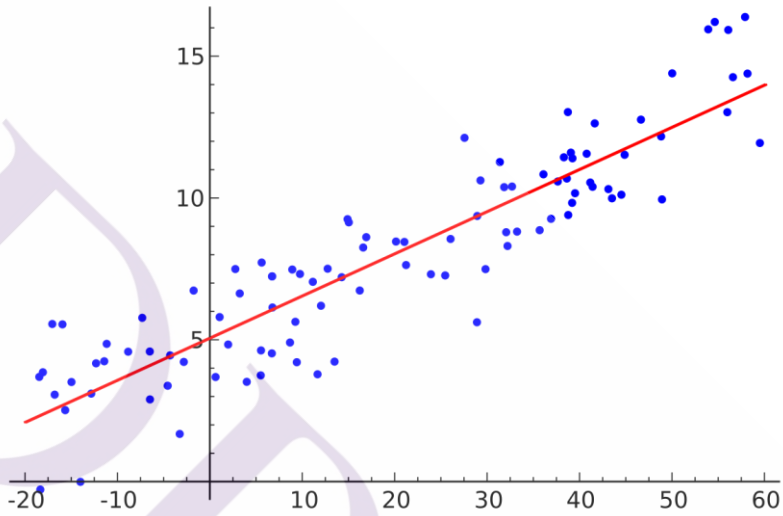
การวิเคราะห์การถดถอยเป็นเทคนิคการสร้างโมเดลจากความสัมพันธ์ระหว่างข้อมูล 2 ตัวเป็นต้นไป หรือทำนายข้อมูลตัวหนึ่งจากข้อมูลอีกตัว (หรือมากกว่า 1 ตัว) สามารถเขียนสมการอย่างง่ายได้ดังนี้

$$Y = \alpha + \beta X + \varepsilon \quad (2.1)$$

โดยที่ α เป็นค่าคงที่ที่ไม่ทราบค่า และ β เป็นสัมประสิทธิ์การถดถอย (Regression Coefficient) เป็นค่าแสดงความสัมพันธ์ระหว่างตัวแปร X และตัวแปร Y หากสัมประสิทธิ์การถดถอยมีค่ามากกว่าศูนย์ แสดงว่ามีความสัมพันธ์ไปในทิศทางเดียวกัน หรือหากสัมประสิทธิ์การถดถอยมีค่าน้อยกว่าศูนย์ แสดงว่ามีความสัมพันธ์ไปในทิศทางตรงกันข้าม แต่หากสัมประสิทธิ์การถดถอยมีค่าเป็นศูนย์ แสดงว่าไม่มีความสัมพันธ์ระหว่างตัวแปร

ตัววัดประสิทธิภาพของตัวแบบที่เป็นที่นิยม ได้แก่สัมประสิทธิ์ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Root Mean Square Error: RMSE) หรือคือค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนจากการทำนาย

$$RMSE = \sqrt{\sum(predition - actual)^2/n} \quad (2.2)$$



ภาพที่ 2.5 ภาพแสดงตัวแบบการถดถอยเชิงเส้นอย่างง่าย

ที่มา: Rohith Gandh, Introduction to Machine Learning Algorithms: Linear Regression (2018)

2.2.2.2 วิธีการเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbors: kNN)

วิธีการเพื่อนบ้านใกล้ที่สุด เป็นวิธีการใช้ในการจำแนก หรือทำนายข้อมูล ด้วยการเรียนรู้จากข้อมูลที่มีป้ายกำกับ โดยทำการเปรียบเทียบความคล้ายคลึงกับข้อมูลที่มีอยู่มากที่สุดเคตัว และกำหนดกลุ่มให้กับข้อมูลที่ไม่มีป้ายกำกับ ตามสมาชิกส่วนใหญ่ของกลุ่ม วิธีการเปรียบเทียบความคล้ายคลึงจะถูกกำหนดในรูปแบบของระยะทางในหลาย ๆ มิติ ตามคุณสมบัติในชุดข้อมูลสอน โดยขั้นตอนการหาเพื่อนบ้านมีดังต่อไปนี้

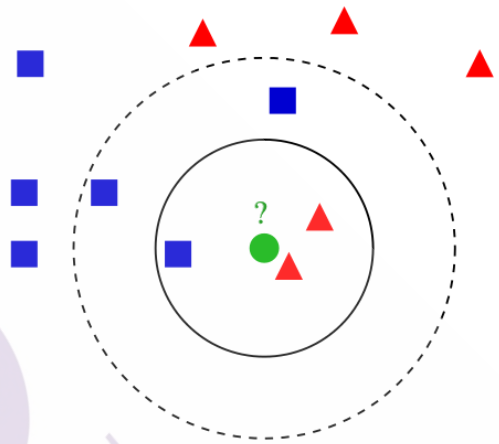
1) กำหนดค่าเค โดยปกติจะนิยมเป็นจำนวนคี่

2) คำนวณหาความคล้ายคลึงของข้อมูลที่ไม่มีป้ายกำกับ กับข้อมูลที่มีป้ายกำกับ ด้วยระยะทาง

3) เรียงลำดับความคล้ายคลึง โดยเลือกข้อมูลตัวอย่างที่มีความคล้ายคลึงมากที่สุดเคตัว

4) พิจารณาข้อมูลตัวอย่างทั้งเคตัว เพื่อจัดจำแนก หรือทำนายข้อมูลแต่ละตัวว่าถูกจัดเป็นกลุ่มใด

5) กำหนดกลุ่มใหม่ให้กับข้อมูลที่ไม่มีป้ายกำกับ ด้วยกลุ่มข้อมูลที่มีตัวอย่างมากที่สุดจากค่าเค



ภาพที่ 2.6 ภาพแสดงตัวอย่างการจำแนกข้อมูลด้วยวิธี kNN ที่มีเพื่อนบ้าน 3 ตัว และ 5 ตัว

ที่มา: Wikipedia, k-nearest neighbors algorithm

การวัดความคล้ายคลึง ด้วยวิธีการวัดระยะห่างยูคลิเดียน (Euclidean Distance) เป็นการวัดระยะห่างระหว่าง 2 จุดในแนวเส้นตรง ที่ได้มาจากทฤษฎีพีทาโกรัส ระยะห่างยูคลิเดียนระหว่างจุด p และ q คำนวณได้จาก

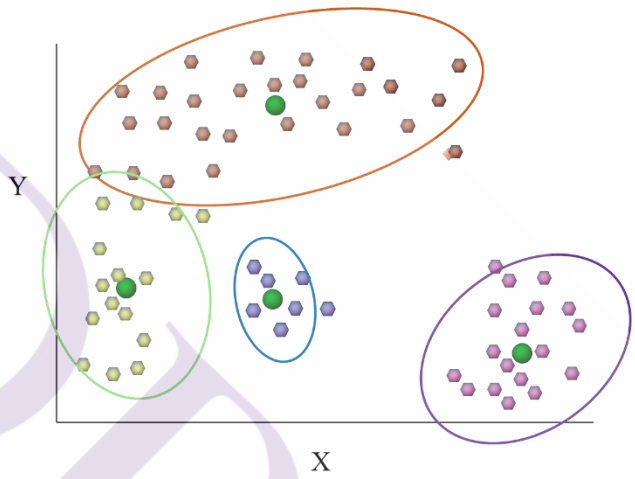
$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_3 - q_3)^2} \quad (2.3)$$

$$= \sum_{i=1}^{n} \sqrt{(p_i - q_i)^2} \quad (2.4)$$

หากคุณสมบัติของข้อมูลมีค่าประเภทข้อมูลแบบนามบัญญัติ (Nominal) หากเหมือนกันระยะทางจะมีค่าเป็นศูนย์ หากต่างกันจะเป็นค่าอย่างอื่น

2.2.2.3 วิธีการจัดกลุ่ม (Clustering)

การจัดกลุ่มข้อมูลจากความคล้ายคลึงกัน เป็นอัลกอริทึมเทคนิคการเรียนรู้แบบไม่มีผู้สอน โดยพยายามให้ระยะห่างของสิ่งที่อยู่ในกลุ่มเดียวกันอยู่ใกล้กันมากที่สุด (Minimize Intra-Cluster Distance) และสิ่งที่อยู่ต่างกลุ่มกันจะมีระยะห่างแตกต่างกันมากที่สุด (Maximize Inter-Cluster Distance) หรืออาจกล่าวได้ว่ากลุ่มข้อมูลที่มีคุณสมบัติและ/หรือคุณลักษณะที่คล้ายคลึงกันควรอยู่ในกลุ่มข้อมูลเดียวกัน และข้อมูลที่มีคุณสมบัติและ/หรือคุณสมบัติที่แตกต่างกันอย่างมากควรอยู่ต่างกลุ่มกัน



ภาพที่ 2.7 ภาพแสดงตัวอย่างการจัดกลุ่มด้วยวิธี K-Means

ที่มา: Understanding data mining clustering methods, SAS.

K-Means เป็นวิธีการจัดกลุ่มที่วิเคราะห์กลุ่มแบบไม่เป็นขั้นตอน หรือการแบ่งส่วน (Partition) ออกเป็นเคกลุ่ม และแทนค่าแต่ละกลุ่มด้วยค่าเฉลี่ยของกลุ่ม หรือเรียกว่าจุดศูนย์กลาง (Centroid) ของกลุ่ม ส่วนวิธีการจัดกลุ่มแบบ K-Medoids มีวิธีการเดียวกันกับ K-Means แต่จุดศูนย์กลางของกลุ่ม จะเป็นค่ากลางของข้อมูล (Median)

โดยขั้นตอนการจัดกลุ่มมีดังต่อไปนี้

1) กำหนดค่าเริ่มต้น จำนวนเคกลุ่ม และกำหนดจุดศูนย์กลางเริ่มต้นทั้งเคจุด

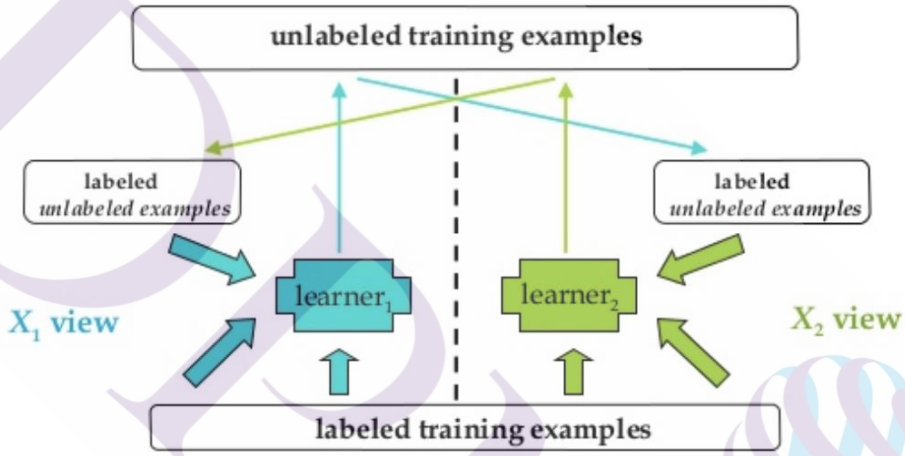
2) พิจารณาข้อมูลเพื่อจัดเข้ากลุ่ม โดยทำการหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลาง โดยหากข้อมูลใดใกล้ค่าจุดศูนย์กลางตัวไหน จะทำการจัดเข้ากลุ่มนั้น

3) หาค่าเฉลี่ย (Mean) ของแต่ละกลุ่มใหม่ หากเป็นวิธี K-Mean ส่วนวิธี K-Medoids จะหาค่ากลาง (Median) ของแต่ละกลุ่ม จากนั้นกำหนดให้เป็นจุดศูนย์กลางของกลุ่มใหม่

4) ทำซ้ำจนกระทั่งค่าเฉลี่ย หรือจุดศูนย์กลางใหม่ในแต่ละกลุ่มจะไม่มีการเปลี่ยนแปลง

2.2.2.4 วิธีการโค เทรนนิง (Co-Training)

วิธีการโค เทรนนิงเป็นอัลกอริทึมการเรียนรู้ของเครื่องที่ใช้เมื่อมีข้อมูลที่มีป้ายกำกับจำนวนเล็กน้อยและข้อมูลที่ไม่มีป้ายกำกับจำนวนมาก (Avrim Blum and Tom Mitchell, 1998) เป็นเทคนิคการเรียนรู้แบบกึ่งมีผู้สอน โดยแต่ละตัวอย่างถูกอธิบายด้วยแอททริบิวต์ที่แตกต่างกันสองมุมมอง การเรียนรู้ครั้งแรกจะแยกตัวจำแนก/ทำนายสำหรับแต่ละมุมมองโดยใช้ตัวอย่างที่มีป้ายกำกับ ผลของการจำแนก/ทำนายข้อมูลที่ไม่มีป้ายกำกับที่ให้ความเชื่อมั่นมากที่สุดจะถูกใช้เป็นข้อมูลสอนเพิ่มเติม (Co-Training, Wikipedia) ดังภาพที่ 2.8



ภาพที่ 2.8 ภาพแสดงวิธีการโค เทรนนิง

ที่มา: https://www.slideshare.net/butest/semisupervised-learning

โค เทรนนิงมีอัลกอริทึมดังนี้

1) ข้อมูลนำเข้า (Input) ได้แก่ ชุดตัวอย่างที่มีป้ายกำกับ (L) ชุดตัวอย่างที่ไม่มีป้ายกำกับ (U) ตัวจำแนก A ตัวจำแนก B

2) ขั้นตอนเริ่มต้น

2.1) แบ่งชุดแอททริบิวต์ของตัวอย่างที่มีป้ายกำกับ และไม่มีป้ายกำกับออกเป็น 2 ส่วน คือ X และ Y

2.2) สร้างชุดการฝึกจากตัวจำแนก A และ B โดยใช้เพียงข้อมูลที่มีป้ายกำกับด้วยแอททริบิวต์ X และ Y

2.3) สุ่มตัวอย่างด้วยแอททริบิวต์ X และ Y จากชุดตัวอย่างที่ไม่มีป้ายกำกับทั้งหมด เพื่อใช้เป็นชุดข้อมูลที่ไม่มีป้ายกำกับเพื่อสร้างโมเดล (U')

3) การวนซ้ำเพื่อเลือกตัวอย่างจาก U' เข้าไปในชุดข้อมูลสอน

3.1) สร้างตัวจำแนก A ด้วยชุดตัวอย่างสอน

3.2) สร้างตัวจำแนก B ด้วยชุดตัวอย่างสอน

3.3) ติดป้ายให้กับตัวอย่าง U' แต่ละตัว (C) โดยใช้ตัวจำแนก A และเลือกตัวอย่าง C ที่ให้ค่าความเชื่อมั่นสูงสุด และเพิ่มตัวอย่างนั้นเข้าไปชุดฝึก

3.4) ติดป้ายให้กับตัวอย่าง U' แต่ละตัว (C) โดยใช้ตัวจำแนก B และเลือกตัวอย่าง C ที่ให้ค่าความเชื่อมั่นสูงสุด และเพิ่มตัวอย่างนั้นเข้าไปชุดข้อมูลสอน

3.5) เติมตัวอย่างที่ไม่มีป้ายกำกับจาก U เข้าไปใน U'

4) ผลลัพธ์ (Output) การติดป้ายให้กับตัวอย่างที่ไม่มีป้ายกำกับนั้น สามารถนำมารวมกันด้วยการคูณเข้าด้วยกัน และทำการปรับค่าความน่าจะเป็น

2.3 งานวิจัยที่เกี่ยวข้อง

วิธีการโค เทรนนิง มีการนำเสนอครั้งแรกโดย Blum และ Mitchell ในปี 1998 ซึ่งเป็นการนำข้อมูลที่ไม่มีป้ายกำกับมาช่วยเพิ่มประสิทธิภาพของโมเดลพยากรณ์ การเลือกข้อมูลที่ไม่มีป้ายกำกับนี้ จะใช้การสร้างโมเดล 2 โมเดลและเลือกข้อมูลที่ไม่มีป้ายกำกับที่มีความน่าเชื่อถือมากที่สุดจากการพยากรณ์มาเพิ่มเข้าในชุดข้อมูลสอน และสร้างโมเดลพยากรณ์ใหม่วนเรื่อยไป

Luca Didaci, Giorgio Fumera และ Fabio Roli ได้ทำการวิจัยถึงผลกระทบของประสิทธิภาพของโมเดลที่เกิดจากการใช้วิธีโค เทรนนิงกับขนาดของชุดข้อมูลสอน นั่นคือลดขนาดของชุดข้อมูลสอนให้มีขนาดน้อยที่สุด จนกระทั่งไม่สามารถนำไปใช้ได้ โดยทดสอบกับข้อมูลทั้งหมด 24 ชุดข้อมูล พบว่าขนาดของข้อมูลที่มีป้ายกำกับเพียง 1 ตัวอย่างต่อชุดข้อมูลสอน ไม่ส่งผลกระทบต่อประสิทธิภาพการโค เทรนนิง

Ruiya Wang และ Li Li ได้พัฒนาอัลกอริทึมการปรับปรุงประสิทธิภาพของโค เทรนนิงโดยคณะกรรมการ (Co-Training by committee) เป็นวิธีการเรียนรู้แบบกึ่งกำกับซ้ำ ซึ่งในระหว่างการทำซ้ำ จะใช้หลาย ๆ โมเดลก่อนหน้านั้นทั้งหมดหลาย ๆ ชุด เพื่อใช้ในการทำนายตัวอย่างที่ไม่มีป้ายกำกับในแต่ละครั้ง ซึ่งสามารถเพิ่มความแม่นยำในการทำนายได้ถึง 10%

Ricardo Sousa และ Jao Gama ทำการเปรียบเทียบระหว่างการโค เทรนนิง และวิธีการเรียนรู้ด้วยตนเอง (Self Learning) สำหรับการถดถอยที่มีเป้าหมายเดียวในข้อมูลแบบสตรีมด้วยกฎการปรับโมเดลสุ่ม (Random Adaptive Model Rules) เปรียบเทียบผลการพยากรณ์ที่ไม่นำเอาข้อมูล

ที่ไม่มีป้ายกำกับเข้าไปเพื่อปรับปรุงการถดถอย ซึ่งผลลัพธ์แสดงหลักฐานที่ทำให้ประสิทธิภาพที่ดีขึ้น ในเรื่องช่วยลดความคลาดเคลื่อนในข้อมูลสตรีมระดับสูง

นอกจากนี้ยังมีงานวิจัยของ Fan Ma และคณะ ได้นำเสนออัลกอริทึมโค เทรนนิงแบบใหม่ที่ชื่อว่า SPaCo (Self-Paced Co-training) การเรียนรู้ร่วมด้วยตัวเอง แก้ไขปัญหาการกำกับค่าของตัวอย่างที่ไม่มีป้ายกำกับที่ไม่ถูกต้องในรอบการฝึกขั้นต้น โดยการแทนที่ของตัวอย่าง (เลือกตัวอย่างเข้าและออก) ซึ่งสามารถเพิ่มประสิทธิภาพของโมเดลได้ดียิ่งขึ้น

งานวิจัยที่ใช้เทคนิคโค เทรนนิงจะเน้นที่การจำแนกประเภทข้อมูล มากกว่าการประมาณค่า ซึ่งในหลาย ๆ งาน การประมาณค่าก็เป็นสิ่งที่จำเป็น ดังนั้น Zhi-Hua Zhou และ Ming Li ได้ทำการวิจัยและนำเสนอวิธีการ COREG (Co-Training Regressors) การเรียนรู้ร่วมแบบกึ่งถดถอย โดยจะทำการเลือกตัวอย่างที่ไม่มีป้ายกำกับมากำกับค่า ผ่านโมเดลที่ให้ค่าความคาดเคลื่อนน้อยที่สุดทั้งสองโมเดล และการพยากรณ์ในขั้นสุดท้าย โดยการหาค่าเฉลี่ยของสมการถดถอยที่สร้างขึ้นทั้งสองตัว ซึ่งอัลกอริทึมนี้สามารถใช้ประโยชน์จากข้อมูลที่ไม่มีป้ายกำกับ เพื่อปรับปรุงการพยากรณ์แบบถดถอย วิธีการนี้ได้มีการนำไปใช้อย่างแพร่หลายแต่ใช้เวลาในการทำงานที่นาน เนื่องจากในการเลือกข้อมูลที่มีป้ายกับจำเป็นต้องทดสอบกับข้อมูลที่ไม่มีป้ายกำกับทีละตัวอย่าง

บทที่ 3
ระเบียบวิธีวิจัย

งานวิจัยนี้เป็นงานวิจัยด้านการเรียนรู้ของเครื่อง เพื่อปรับปรุงประสิทธิภาพโมเดลพยากรณ์ จากการเรียนรู้ร่วมกึ่งมีผู้สอน สำหรับข้อมูลที่มีป้ายกำกับที่มีสัดส่วนน้อย เมื่อเทียบกับข้อมูลที่ไม่มีป้ายกำกับ โดยมีเป้าหมายคือเพิ่มประสิทธิภาพจากการพยากรณ์ และลดระยะเวลาในการสร้างโมเดล ซึ่งในบทนี้จะแบ่งออกเป็น 3 ตอน คือ ตอนที่หนึ่ง แนวทางการวิจัย ตอนที่สอง เครื่องมือที่ใช้ในการวิจัย ตอนที่สาม ขั้นตอนและวิธีการดำเนินงาน ดังต่อไปนี้

3.1 แนวทางการวิจัย

3.1.1 ศึกษาแนวความคิดและการวิจัยที่เกี่ยวข้อง

ศึกษาแนวความคิดและการวิจัยที่เกี่ยวข้องกับการเรียนรู้ร่วมกึ่งมีผู้สอน เพื่อให้ผู้วิจัยเข้าใจถึงแนวความคิด ขั้นตอนการสร้างโมเดล ประเมินผล ตลอดจนปรับปรุงโมเดลให้มีประสิทธิภาพมากยิ่งขึ้น

3.1.2 รวบรวมข้อมูลและศึกษาลักษณะของข้อมูล

ศึกษาลักษณะของข้อมูลที่นำมาใช้ในการฝึกโมเดล เพื่อให้ผู้วิจัยเข้าใจถึงลักษณะข้อมูลและสามารถจัดการข้อมูลให้พร้อมใช้ในการฝึกโมเดล

3.1.3 ศึกษาทฤษฎีที่เกี่ยวกับการเรียนรู้ของเครื่อง

ศึกษาทฤษฎีที่เกี่ยวกับการเรียนรู้ของเครื่องเพื่อใช้ในการใช้ในการสร้างโมเดลพยากรณ์จากการเรียนรู้ร่วมกึ่งมีผู้สอน สำหรับข้อมูลที่ป้ายกำกับที่มีสัดน้อยน้อย เมื่อเทียบกับข้อมูลที่ไม่มีป้ายกำกับ

3.1.4 พัฒนาขั้นตอนการฝึกโมเดล

พัฒนาขั้นตอนการฝึกโมเดลด้วยข้อมูลที่มีป้ายกำกับจำนวนน้อย โดยการเรียนรู้ร่วมกึ่งมีผู้สอน และปรับปรุงประสิทธิของโมเดล

3.1.5 ดำเนินการวัดผลของโมเดลและบันทึกผล

เมื่อพัฒนาแนวทางการฝึกโมเดลเสร็จสิ้น ผู้วิจัยจึงได้เตรียมเครื่องคอมพิวเตอร์ที่เหมาะสม สำหรับการประมวลผลข้อมูลที่ใช้ แล้วจึงนำข้อมูลมาประมวลผลและเก็บผลลัพธ์ไว้

3.1.6 วิเคราะห์และสรุปผลการทดลอง

หลังจากที่ได้ผลลัพธ์ของการทดลองทั้งหมดแล้วนั้น ผู้วิจัยรวบรวมผลการทดลองทั้งหมดมาวิเคราะห์ร่วมกัน แล้วสรุปผลของการเรียนรู้ร่วมแบบกึ่งมีผู้สอนว่าวิธีการเลือกข้อมูลที่ไม่มีป้ายกำกับสามารถเพิ่มประสิทธิภาพของโมเดล รวมถึงลดระยะเวลาการสร้างโมเดลได้จริงหรือไม่

3.2 เครื่องมือที่ใช้ในการวิจัย

3.2.1 ด้านฮาร์ดแวร์ (Hardware)

3.2.1.1 คอมพิวเตอร์ลูกข่าย (Client) สำหรับการสร้างโมเดล มีคุณสมบัติดังนี้

1) หน่วยประมวลผลกลาง (CPU) Intel Core i5 2.7 GHz
2) หน่วยความจำ (Memory) 8 GB
3) พื้นที่จัดเก็บ (Storage) 256 GB

3.2.1.2 คอมพิวเตอร์แม่ข่าย (Server) สำหรับการประมวลผล มีคุณสมบัติดังนี้

1) หน่วยประมวลผลกลาง (CPU) Intel Core (Skylake, IBRS) 2.3 GHz (4 Cores) HDD 120GB
2) หน่วยความจำ (Memory) 8 GB

3.2.2 ด้านซอฟท์แวร์ (Sotfware)

1) RapidMiner Studio เวอร์ชั่น 9.2.000
2) RapidMiner Server

3.3 ขั้นตอนและวิธีการดำเนินงาน

ในงานวิจัยนี้จะนำกระบวนการวิเคราะห์ข้อมูลด้วย CRISP-DM มาประยุกต์ใช้ในการดำเนินงาน ซึ่งมีขั้นตอนการวิจัยเป็นดังภาพต่อไปนี้

ภาพที่ 3.1 ภาพแสดงถึงขั้นตอนและวิธีการดำเนินงานวิจัย

3.3.1 รวบรวมข้อมูลและศึกษาแอททริบิวต์ของข้อมูล

รวบรวมข้อมูลจากอินเตอร์เน็ต โดยมี 2 แหล่งข้อมูล คือ Kaggle และ UCI Machine Learning Repository โดยใช้ข้อมูล 3 ชุดที่มีแอททริบิวต์ของข้อมูลที่แตกต่างกัน เพื่อวัดประสิทธิภาพของโมเดลในแต่ละแอททริบิวต์ของข้อมูล ซึ่งชุดข้อมูลทั้ง 3 ชุดข้อมูลมีดังนี้

3.3.1.1 ชุดข้อมูลโฆษณาประกาศรับสมัครงาน

ชุดข้อมูลที่ 1 ชุดข้อมูลโฆษณาประกาศรับสมัครงาน เป็นข้อมูลประกาศสมัครงานในประเทศอังกฤษ จัดทำโดย Adzuna เป็นข้อมูลที่ใช้ในการแข่งขันในหัวข้อการพยากรณ์เงินเดือนของ Kaggle ซึ่งมีจำนวน 244,768 ประกาศ ข้อมูลถูกจัดเก็บในรูปแบบของแฟ้มข้อมูลประเภท Comma Separate Value (.csv) โดยประกอบด้วยข้อมูลดังนี้

ตารางที่ 3.1 คำอธิบายแอททริบิวต์ชุดข้อมูลพยากรณ์เงินเดือนรวมถึงการใช้ข้อมูลทางสถิติ

แอททริบิวต์	**ประเภทแอททริบิวต์**	**คำอธิบายของข้อมูล**
Id	ตัวเลขจำนวนเต็ม	เลขที่ระบุประกาศการรับสมัครงาน
SalaryNormalized	ตัวเลขจำนวนเต็ม	เงินเดือน หน่วย: ปอนด์ ที่ผ่านการปรับหน่วยให้เป็นมาตรฐานเดียวกัน
Title	ตัวอักษร	ชื่อตำแหน่งงาน
FullDescription	ตัวอักษร	คำบรรยายลักษณะงาน
LocationRaw	ตัวอักษร	สถานที่ตั้ง
LocationNormalized	ตัวอักษร	สถานที่ตั้ง ที่ผ่านการปรับหน่วยให้เป็นมาตรฐานเดียวกัน
ContractType	ตัวอักษร	ประเภทการจ้างงาน
ContractTime	ตัวอักษร	ระยะเวลาการจ้างงาน
Company	ตัวอักษร	ชื่อบริษัท
Category	ตัวอักษร	ประเภทงาน
SalaryRaw	ตัวอักษร	เงินเดือน หน่วย: ปอนด์
SourceName	ตัวอักษร	แหล่งที่มาของข้อมูล

3.3.1.2 ชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน (Metro Interstate Traffic Volume Data Set)

ชุดข้อมูลที่ 2 ชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน เป็นข้อมูลปริมาณการจราจรระหว่างรัฐโดยรถไฟฟ้ารายชั่วโมง จากทิศตะวันตกของมินนิโซตา DoT ATR สถานี 301 ซึ่งอยู่กลางระหว่างมินนิอาโปลิสและเซนต์พอลมินนิโซตา เป็นข้อมูลจาก UCI Machine Learning Repository ซึ่งมีข้อมูลจำนวน 48,204 รายการ ข้อมูลถูกจัดเก็บในรูปแบบของแฟ้มข้อมูลประเภท Comma Separate Value (.csv) ประกอบด้วยข้อมูลดังนี้

ตารางที่ 3.2 คำอธิบายแอททริบิวต์ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

แอททริบิวต์	**ประเภทแอททริบิวต์**	**คำอธิบายของข้อมูล**
date_time	วันเวลา	วันเดือนปี เวลา ที่ทำการเก็บข้อมูล
traffic_volumn	ตัวเลขจำนวนเต็ม	ปริมาณจราจร หน่วย: คน
holiday	ตัวอักษร	วันหยุด
temp	ตัวเลขจำนวนจริง	อุณหภูมิโดยเฉลี่ย หน่วย: องศาเคลวิน
rain_1h	ตัวเลขจำนวนเต็ม	ปริมาณน้ำฝน หน่วย: มิลลิเมตร
snow_1h	ตัวเลขจำนวนเต็ม	ปริมาณหิมะ หน่วย: มิลลิเมตร
clouds_all	ตัวเลขจำนวนเต็ม	ร้อยละปริมาณหมอกที่ปกคลุม
weather_main	ตัวอักษร	ประเภทลักษณะอากาศ
weather_description	ตัวอักษร	ลักษณะอากาศ

3.3.1.3 ชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม (Combined Cycle Power Plant Data Set)

ชุดข้อมูลที่ 3 ชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม เป็นข้อมูลพลังงานความร้อนร่วม โดยข้อมูลที่รวบรวมจากโรงไฟฟ้าพลังความร้อนร่วมในช่วง 6 ปี (2549-2554) เป็นข้อมูลจาก UCI Machine Learning Repository ซึ่งมีข้อมูลจำนวน 9,568 รายการ ข้อมูลถูกจัดเก็บในรูปแบบของแฟ้มข้อมูลประเภท Excel (.xls) ประกอบด้วยข้อมูลดังนี้

ตารางที่ 3.3 คำอธิบายแอททริบิวต์ของชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วมจากโรงไฟฟ้า

แอททริบิวต์	ประเภทแอททริบิวต์	คำอธิบายของข้อมูล
PE	ตัวเลขจำนวนจริง	พลังงานไฟฟ้า
AT	ตัวเลขจำนวนจริง	อุณหภูมิ
AP	ตัวเลขจำนวนจริง	ความดันบรรยากาศ
RH	ตัวเลขจำนวนจริง	ความชื้นสัมพัทธ์
V	ตัวเลขจำนวนจริง	ไอเสีย

3.3.2 การประมวลผลข้อมูลเบื้องต้น

การประมวลผลข้อมูลเบื้องต้นเป็นการเตรียมความพร้อมของข้อมูลที่จะนำไปใช้ในการฝึกโมเดล เริ่มตั้งแต่การทำความเข้าใจข้อมูล การคัดเลือกแอททริบิวต์ของข้อมูลที่จะใช้ในการพัฒนาโมเดล การทำความสะอาดข้อมูล การปรับหน่วยข้อมูล การสุ่มเลือกตัวอย่าง

3.3.2.1 การทำความเข้าใจข้อมูล (Data Understanding)

การทำความเข้าใจข้อมูล รวมถึงการใช้ข้อมูลทางสถิติ เพื่อสามารถเลือกใช้โมเดลได้อย่างเหมาะสม โดยลักษณะผลการวิเคราะห์จะจัดเก็บดังตาราง 3.4, 3.5 และ 3.6 ตามลำดับ

ตารางที่ 3.4 ผลการวิเคราะห์ชุดข้อมูลการพยากรณ์เงินเดือน

แอททริบิวท์	บทบาท	มาตรวัด	ค่ากลางข้อมูล	ขอบเขตข้อมูล/ตัวอย่างข้อมูล	จำนวนข้อมูลที่ไม่ระบุค่า
Id	ตัวระบุ	นามบัญญัติ	-	ตัวเลขที่ไม่ซ้ำ	-
Salary Normalized	ป้ายกำกับ	อัตราส่วน	34,123	5,000 – 200,000	-
Title	ตัวแปร	นามบัญญัติ	Business Development Manager	Project Manager, Cleaner, Management Accountant, Account Manager, etc.	-
Full Description	ตัวแปร	ข้อความ	-	An expanding software and consultancy services … Send in your CV now, etc.	-

ตารางที่ 3.4 (ต่อ)

ข้อมูล	บทบาท	มาตรวัด	ค่ากลางข้อมูล	ขอบเขตข้อมูล/ตัวอย่างข้อมูล	จำนวนข้อมูลที่ไม่ระบุค่า
Location Raw	ตัวแปร	นามบัญญัติ	London	London South East, City London South East, City of London – London, Central London, etc.	-
Location Normalized	ตัวแปร	นามบัญญัติ	UK	London, South East London, The City, Manchester, Leeds, etc.	-
Contract Type	ตัวแปร	นามบัญญัติ	(blank)	full_time, part_time, (blank)	179,326
Contract Time	ตัวแปร	นามบัญญัติ	permanent	contract, permanent, (blank)	63,905
Company	ตัวแปร	นามบัญญัติ	(blank)	JAM Recruitment Ltd, London4Jobs, Hays, UKStaffsearch, etc.	32,430
Category	ตัวแปร	นามบัญญัติ	IT Jobs	Engineering Jobs, Accounting & Finance Jobs, Healthcare & Nursing Jobs, etc.	-
Salary Raw	ตัวแปร	นามบัญญัติ	50,000-74,999 yearly	30k - 40k, 6.19 per hour, 30000 - 35000/annum, etc.	-
Source Name	ตัวแปร	นามบัญญัติ	totaljobs.com	cv-library.co.uk, Jobcentre Plus, jobsite.co.uk, etc.	-

ตารางที่ 3.5 ผลการวิเคราะห์ชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐโดยรถไฟฟ้าใต้ดิน

แอททริบิวท์	บทบาท	มาตรวัด	ค่ากลางข้อมูล	ขอบเขตข้อมูล/ตัวอย่างข้อมูล	จำนวนข้อมูลที่ไม่ระบุค่า
date_time	ตัวระบุ	ช่วง	-	ช่วงวันเวลาที่ไม่ซ้ำ	-
traffic_volumn	ป้ายกำกับ	อัตราส่วน	3,260	0 - 7,280	-
holiday	ตัวแปร	นามบัญญัติ	None	Labor Day, Christmas Day, State Fair, Independence Day, etc.	-
temp	ตัวแปร	อัตราส่วน	281.21	0.00 - 310.07	-
rain_1h	ตัวแปร	อัตราส่วน	0.33	0.00 – 9,831.30	-
snow_1h	ตัวแปร	อัตราส่วน	0.0002	0.0000 – 0.5100	-
clouds_all	ตัวแปร	อัตราส่วน	49	0 - 100	-
weather_main	ตัวแปร	นามบัญญัติ	Clouds	Clear, Mist, Rain, Drizzle, etc.	-
weather_ description	ตัวแปร	นามบัญญัติ	Sky is clear	Mist, overcast clouds, broken clouds, drizzle, heavy snow, etc.	-

ตารางที่ 3.6 ผลการวิเคราะห์ชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วมจากโรงไฟฟ้า

แอททริบิวท์	บทบาท	มาตรวัด	ค่าเฉลี่ย	ค่าต่ำสุด	ค่ามากสุด	ค่าเบี่ยงเบนมาตรฐาน	จำนวนข้อมูลที่ไม่ระบุค่า
PE	ป้ายกำกับ	อัตราส่วน	454.37	420.26	495.76	17.07	-
AT	ตัวแปร	อัตราส่วน	19.65	1.81	37.11	7.45	-
AP	ตัวแปร	อัตราส่วน	1,013.26	992.89	1,033.30	5.94	-
RH	ตัวแปร	อัตราส่วน	73.31	25.56	100.16	14.60	-
V	ตัวแปร	อัตราส่วน	54.31	25.36	81.56	12.71	-

3.3.2.2 การคัดเลือกคุณลักษณะของข้อมูล (Feature Selection)

ชุดข้อมูลพยากรณ์เงินเดือนมีแอททริบิวต์ของข้อมูล 1 แอททริบิวต์ที่มีค่าไม่ซ้ำ (โดยไม่ใช่ตัวระบุ) คือ FullDescription และมี 2 แอททริบิวต์ที่มีค่าที่ยังไม่ได้ทำการปรับหน่วยข้อมูล โดยมีแอททริบิวต์ใช้แทนได้ คือ LocationRaw และ SalaryRaw โดยใช้ LocationNormalized และ SalaryNormalized แทนได้ ดังนั้น FullDescription LocationRaw และ SalaryRaw จะไม่นำไปใช้เพื่อฝึกโมเดล

3.3.2.3 การทำความสะอาดข้อมูล (Data Cleaning)

ชุดข้อมูลพยากรณ์เงินเดือนมีแอททริบิวต์ของข้อมูล 3 แอททริบิวต์ที่ไม่ได้ระบุค่า และมีมาตรวัดข้อมูลแบบนามบัญญัติ คือ ContractType มีตัวอย่างที่ไม่ได้ระบุค่าจำนวน 179,326 คิดเป็น 73% จากข้อมูลทั้งหมด ContractTime มีตัวอย่างที่ไม่ได้ระบุค่าจำนวน 63,905 คิดเป็น 26% จากข้อมูลทั้งหมด และ Company มีตัวอย่างที่ไม่ได้ระบุค่าจำนวน 32,430 คิดเป็น 13% จากข้อมูลทั้งหมด ดังนั้นผู้วิจัยจึงระบุค่าให้แอททริบิวต์ทั้ง 3 เป็น NA

3.3.2.4 การปรับหน่วยข้อมูล (Normalize)

ชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐโดยรถไฟฟ้าใต้ดิน และข้อมูลการพยากรณ์พลังงานความร้อนร่วมจากโรงไฟฟ้า มีบางแอททริบิวต์ที่มีมาตรวัดแบบอัตราส่วน และมีขนาดข้อมูลที่แตกต่างกัน เนื่องจากแอททริบิวต์ของข้อมูลแต่ละตัวไม่ได้มาจากเครื่องมือทดสอบการทำงานเดียวกัน ซึ่งไม่เหมาะต่อการนำไปพัฒนาโมเดล ดังนั้นการแปลงข้อมูลให้อยู่ในมาตรฐานเดียวกัน (Normalize) จึงถูกนำมาใช้ โดยการเลือกใช้วิธีปรับ Z-transformation ซึ่งทำให้เป็นค่ามาตรฐาน ด้วยการลบด้วยค่าเฉลี่ย แล้วหารด้วยค่าเบี่ยงเบนมาตรฐาน การกระจายของข้อมูลมีค่าเฉลี่ยเป็นศูนย์และความแปรปรวนเป็นหนึ่ง

3.3.2.5 การสุ่มเลือกตัวอย่าง (Sampling)

การวิจัยครั้งนี้ใช้วิธีการสุ่มข้อมูลแบบชั้นภูมิ (Stratified Random Sampling) ซึ่งเป็นการสุ่มตัวอย่างจากประชากรที่มีจำนวนมาก โดยประชากรจะถูกแบ่งออกเป็นชั้นภูมิตามลักษณะอย่างใดอย่างหนึ่ง โดยไม่ให้มีหน่วยซ้ำกัน ซึ่งในชั้นภูมิเดียวกันจะประกอบด้วยหน่วยที่มีลักษณะคล้ายคลึงกันมากที่สุด และแตกต่างระหว่างชั้นภูมิมากที่สุด โดยข้อมูลทั้ง 3 ชุดทำการสุ่มตัวอย่างจำนวน 5,000 ตัวอย่าง เพื่อเป็นชุดทดสอบเปรียบเทียบประสิทธิภาพโมเดล AU-COREG กับโมเดล COREG และสุ่มตัวอย่าง อีก 2 กลุ่มตัวอย่าง (กลุ่มตัวอย่างละ 5,000 ตัวอย่าง) เพื่อเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG ที่เกิดจากการสุ่มตัวอย่างที่แตกต่างกันทั้ง 3 ชุด

3.3.2.6 การแบ่งข้อมูลเพื่อสร้างโมเดล (Data Split)

ทำการแบ่งข้อมูลออกเป็น 2 ส่วน 1) ข้อมูลที่กำหนดให้มีป้ายกำกับ (Labeled Data) เพื่อใช้ในการสร้างและทดสอบโมเดล 2) ข้อมูลที่กำหนดให้ไม่มีป้ายกำกับ (Unlabeled Data) โดยการใช้การสุ่มข้อมูลแบบชั้นภูมิ ดังตารางที่ 3.7

ตารางที่ 3.7 การแบ่งข้อมูลเพื่อพัฒนาโมเดลขั้นต้น

ส่วนของข้อมูล	สัดส่วน	ขนาดข้อมูล
ตัวอย่างทั้งหมด: S	100%	5,000
- ข้อมูลสำหรับการทดสอบโมเดลสุดท้าย: F	25.0%	1,250
- ข้อมูลสำหรับการสร้างโมเดลขั้นแรก (Initial Training Data): L	7.5%	375
- ข้อมูลที่ไม่มีป้ายกำกับ เพื่อเป็นส่วนที่เลือกข้อมูลเข้าใช้ในการสร้างโมเดลเพิ่มเติม (Unlabeled Data): U'	2.0%	100
- ข้อมูลที่ไม่มีป้ายกำกับส่วนที่เหลือ: U	65.5%	3,275

3.3.3 การพัฒนาโมเดล

3.3.3.1 การกำหนดพารามิเตอร์

การวิจัยนี้จะทำการเปรียบเทียบการเลือกตัวอย่างจากข้อมูลที่ไม่มีป้ายกำกับ มาใช้ในการพัฒนาโมเดล เพื่อเพิ่มความแม่นยำจากการพยากรณ์ค่าที่ได้จากโมเดล ดังนั้นผู้วิจัยจึงกำหนดพารามิเตอร์ ดังตารางที่ 3.8

ตารางที่ 3.8 การกำหนดพารามิเตอร์

พารามิเตอร์	ค่า
จำนวนรอบ (จำนวนตัวอย่างที่ถูกเลือกเพื่อสร้างโมเดล จากข้อมูลไม่มีป้ายกำกับ): T	100, 200
จำนวนคลัสเตอร์ (จำนวนตัวอย่างที่เป็นตัวแทนของข้อมูลที่ไม่มีป้ายกำกับ): k	5,6,7,8,9,10
วิธีการจัดคลัสเตอร์และวิธีการเลือกตัวแทนคลัสเตอร์	- K-Medoid โดยเลือกสมาชิกที่เป็นค่ากึ่งกลางของคลัสเตอร์ (Median)

	- K-Mean ที่มี Seed ได้แก่ 1992, 100, 3645 โดยวิธีการสุ่มเลือกสมาชิก
จำนวนเพื่อนบ้านเพื่อหาตัวอย่างที่ให้ค่า RMSE ที่น้อยที่สุด (Neighbors): K	5

3.3.3.2 การเลือกโมเดลการสอน (Training Model)

หลังจากรวบรวมข้อมูลและศึกษาแอททริบิวต์ข้อมูล และประมวลผลข้อมูลเบื้องต้นแล้วนั้น นำข้อมูลเหล่านั้นมาฝึกโมเดลสำหรับการพยากรณ์ โดยการเลือกข้อมูลที่ไม่มีป้ายกำกับด้วยวิธีการเรียนรู้ร่วมแบบกึ่งมีผู้สอน ซึ่งโมเดลเพื่อการพยากรณ์ที่ใช้ คือ kNN และวิธีการวัดระยะทางจะแตกต่างกันในแต่ละชุดข้อมูล ดังตารางที่ 3.9

ตารางที่ 3.9 โมเดลและวิธีการวัดระยะทางในแต่ละชุดข้อมูล

ชุดข้อมูล	**โมเดล kNN_1**		**โมเดล kNN_2**	
	วิธีการวัดระยะทาง p_1	k	**วิธีการวัดระยะทาง p_2**	k
1	Mix-Measure: Mix Euclidean Distance	3	Nominal-Measure: Nominal Distance	3
2	Mix-Measure: Mix Euclidean Distance	5	Mix-Measure: Mix Euclidean Distance	9
3	Numerical-Measure: Euclidean Distance	5	Numerical-Measure: Correlation Similarity	5

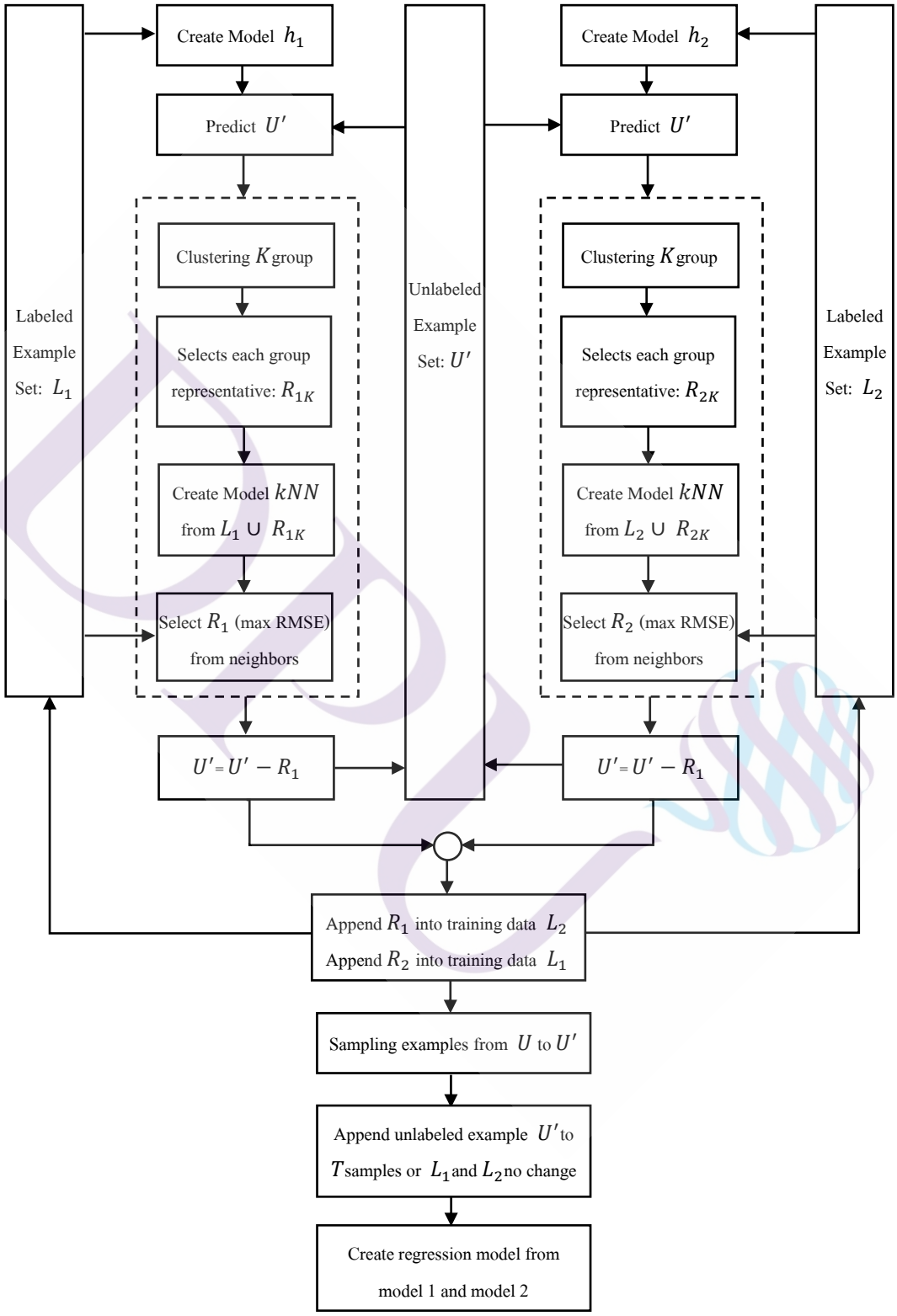
ข้อมูลชุดที่ 1 มีแอททริบิวต์ของข้อมูลมีมาตรวัดเป็นนามบัญญัติทั้งหมด ผู้วิจัยใช้วิธีการวัดระยะทางกับทั้ง 2 โมเดลที่ต่างกัน คือ Mix-Measure และ Nominal-Measure โดยที่ค่า k เท่ากัน คือมีค่าเท่ากับ 3

ข้อมูลชุดที่ 2 มีแอททริบิวต์ข้อมูลมีมาตรวัดทั้งแบบอัตราส่วนและแบบนามบัญญัติ ผู้วิจัยใช้วิธีการวัดระยะทางเดียวกันกับทั้ง 2 โมเดล คือ Mix-Measure แต่ต่างกันที่ค่า k คือมีค่าเท่ากับ 5 และ 9 ตามลำดับ

ข้อมูลชุดที่ 3 มีแอททริบิวต์ข้อมูลที่มีมาตรวัดเป็นอัตราส่วนทั้งหมด ผู้วิจัยใช้วิธีการวัดระยะทางเดียวกันกับทั้ง 2 โมเดล คือNumerical-Measure และที่ค่า k เท่ากัน แต่ต่างกันด้วยตัววัดระยะทาง คือ Euclidean Distance และ Correlation Similarity ตามลำดับ

เมื่อทำการแบ่งชุดข้อมูล กำหนดพารามิเตอร์ และเลือกโมเดล ขั้นตอนในการพัฒนาอัลกอริทึมเพื่อใช้ในเลือกตัวอย่างที่ไม่มีป้ายกำกับใช้เป็นข้อมูลที่ใช้ในการพัฒนาโมเดลการพยากรณ์ เพื่อเพิ่มประสิทธิภาพของโมเดล และลดเวลาการประมวลผล ตามขั้นตอนดังนี้

3.3.3.3 อัลกอริทึมในการสร้างโมเดล



ภาพที่ 3.2 ภาพแสดงอัลกอริทึมในการเลือกข้อมูลที่ไม่มีป้ายกำกับเพื่อสร้างโมเดลการพยากรณ์

โดยที่

T	หมายถึง	จำนวนรอบ (จำนวนตัวอย่างที่ถูกเลือกเพื่อสร้างโมเดล จากข้อมูลไม่มีป้ายกำกับ)
L_1, L_2	หมายถึง	ตัวอย่างที่มีป้ายกำกับเพื่อสร้างโมเดลที่ 1 และ 2 ตามลำดับ
U’	หมายถึง	ตัวอย่างที่ไม่มีป้ายกำกับ เพื่อเป็นส่วนที่เลือกข้อมูลเข้า ใช้ในการสร้างโมเดลเพิ่มเติม (Unlabeled Data)
U	หมายถึง	ข้อมูลที่ไม่มีป้ายกำกับส่วนที่เหลือ
$h_1, h_2,$	หมายถึง	โมเดลพยากรณ์โมเดลที่ 1 และ 2 ตามลำดับ
K	หมายถึง	จำนวนคลัสเตอร์
R_{1k}, R_{2k}	หมายถึง	ตัวแทนของคลัสเตอร์ที่ k ของโมเดลที่ 1 และ 2 ตามลำดับ
R_1, R_2	หมายถึง	ตัวแทนของคลัสเตอร์ที่ให้ค่า RMSE ที่มากที่สุดของโมเดลที่ 1 และ 2 ตามลำดับ
F	หมายถึง	ข้อมูลสำหรับการทดสอบโมเดลสุดท้าย

1) กำหนดให้ L_1 และ L_2 มีค่าเท่ากับ L จากนั้นสร้างโมเดลเริ่มต้น ทั้งสองโมเดล (kNN_1 และ kNN_2) จากการเทรนข้อมูลที่มีป้ายกำกับ L_1 และ L_2 โดยใช้เทคนิคตามตารางที่ 3.9 ที่ได้กล่าวไว้ข้างต้น

$$L_1,\ L_2 \leftarrow L \tag{3.1}$$

$$h_1 \leftarrow kNN(L_1, k, p_1) \tag{3.2}$$

$$h_2 \leftarrow kNN(L_2, k, p_2) \tag{3.3}$$

เมื่อค่า p_1, p_2 วัดระยะทางตามวิธีที่กำหนดในตารางที่ 3.9

2) ทำซ้ำตั้งแต่ 1-100 รอบ (T = 100)

2.1) สำหรับโมเดลที่ j ตั้งแต่ 1-2 (j =1,2)

2.1.1) นำโมเดลที่ h_j มาพยากรณ์ข้อมูลที่อยู่ใน U’ ทั้งหมด (x_u) จนครบ

$$\hat{y}_u \leftarrow h_j(x_u) \tag{3.4}$$

2.1.2) จากนั้นทำการจัด Cluster ของข้อมูลที่ถูกพยากรณ์ $\hat{y}_u$ โดยใช้เทคนิคและจำนวน Cluster ที่กำหนดไว้ในตารางที่ 3.8

2.1.3) ทำซ้ำกับ Cluster ที่ 1 ถึง Cluster สุดท้าย

สุ่มเลือกตัวแทนของ Cluster (R_{jK}) จำนวน 1 ตัวอย่าง และเลือกตัวอย่างจาก L_j ที่มีระยะห่างน้อยที่สุดจำนวน 5 ตัวอย่าง จากตัวแทนของ Cluster (R_{jK}) ดังกล่าว จากนั้นสร้างโมเดลใหม่จากตัวอย่าง L_j และ R_{jK} และทดสอบประสิทธิภาพของโมเดลด้วยสัมประสิทธ์ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง RMSE (Root Mean Square Error) กับเพื่อนบ้านทั้ง 5 ตัวอย่าง

$$h_j \leftarrow kNN(L_j \cup R_{jK}, k, p_1) \tag{3.5}$$

$$RMSE = \sqrt{\frac{\sum_{K=1}^{5}(predition-actual)^2}{5}} \tag{3.6}$$

2.1.4) เลือกตัวแทนของ Cluster (R_{jK}) ที่ทำให้ RMSE มากกว่า 0 และมีค่ามากที่สุด (R_j) เพื่อจะนำเข้าไปเป็นข้อมูลเทรนนิ่ง จากนั้นนำ R_j ออกจาก U'

$$U' \leftarrow U' - R_j \tag{3.7}$$

2.1.5) สุ่มเลือกตัวอย่างจาก U เพิ่มเข้ามาใน U' ให้ครบจำนวน

2.2) เมื่อทำตามขั้นตอนที่ 2.1) ครบทั้ง 2 โมเดล จากนั้นเพิ่ม R_1 เข้าไปใน L_2 และเพิ่ม R_2 เข้าไปใน L_1

$$L_1 \leftarrow L_1 \cup R_2 \tag{3.8}$$

$$L_2 \leftarrow L_2 \cup R_1 \tag{3.9}$$

2.3) สร้างโมเดล 1 และ 2 จากข้อมูล L_1 และ L_2 ใหม่ และหากพบว่า L_1 และ L_2 ไม่เปลี่ยนแปลง ให้หยุดดำเนินการ

3) สร้างโมเดล Regression ได้ดังนี้

$$h^*(x) = \frac{h_1(x) + h_2(x)}{2} \tag{3.10}$$

4) กำหนดจำนวนรอบการทำซ้ำเป็น 200 รอบ (T=200) และทำตามขั้นตอนข้างต้นทั้งหมด เพื่อสร้างโมเดลใหม่มาเปรียบเทียบ

3.3.4 ทดสอบประสิทธิภาพ

เมื่อได้โมเดลจากการเลือกข้อมูลที่ไม่มีป้ายกำกับเพิ่มเข้าไปในข้อมูลที่ Training ตามพารามิเตอร์ที่กำหนดไว้ตอนต้น จากนั้นทดสอบประสิทธิภาพด้วย RMSE กับข้อมูลสำหรับการทดสอบโมเดลสุดท้าย (F) และเปรียบเทียบประสิทธิภาพและระยะเวลาในการประมวลผลที่ได้จากการกำหนดพารามิเตอร์ที่ต่างกัน

บทที่ 4
ผลการศึกษา

จากการวิจัยวิธีการเลือกข้อมูลที่ไม่มีป้ายกำกับอย่างอัตโนมัติเพื่อการสร้างโมเดลการเรียนรู้ร่วมแบบกึ่งมีผู้สอน เพื่อปรับปรุงประสิทธิภาพโมเดลพยากรณ์ จากการเรียนรู้ร่วมกึ่งมีผู้สอนสำหรับข้อมูลที่มีป้ายกำกับที่มีสัดส่วนน้อย เมื่อเทียบกับข้อมูลที่ไม่มีป้ายกำกับ และลดระยะเวลาในการสร้างโมเดลพยากรณ์ กับชุดข้อมูลทั้งหมด 3 ชุดข้อมูล ได้แก่ ชุดข้อมูลโฆษณาประกาศรับสมัครงาน ชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐโดยรถไฟฟ้าใต้ดิน และชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม โดยการใช้วิธีการเรียนรู้ของเครื่อง ได้แก่ วิธีการจัดกลุ่ม (Clustering) วิธีการเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbors: kNN) วิธีการวิเคราะห์การถดถอย (Regression Analysis) และวิธีการโค เทรนนิง (Co-Training) เพื่อสร้างโมเดลจากการเลือกข้อมูลที่ไม่มีป้ายกำกับเพื่อเพิ่มเข้าไปในข้อมูลฝึกสอน จากนั้นทำการเปรียบเทียบประสิทธิภาพของโมเดลและระยะเวลาที่ใช้ในการสร้างโมเดลที่ได้จากการกำหนดพารามิเตอร์ที่ต่างกัน ตามชุดของข้อมูลได้ 3 หัวข้อ ดังนี้

4.1 จะนำเสนอผลการเปรียบเทียบประสิทธิภาพของโมเดลและระยะเวลาที่ใช้ในการสร้างโมเดล AU-COREG ที่ได้จากการกำหนดพารามิเตอร์ที่ต่างกัน กับการสร้างโมเดลด้วยวิธี Self-Training และ COREG ของข้อมูลชุดที่ 1 ข้อมูลโฆษณาประกาศรับสมัครงาน

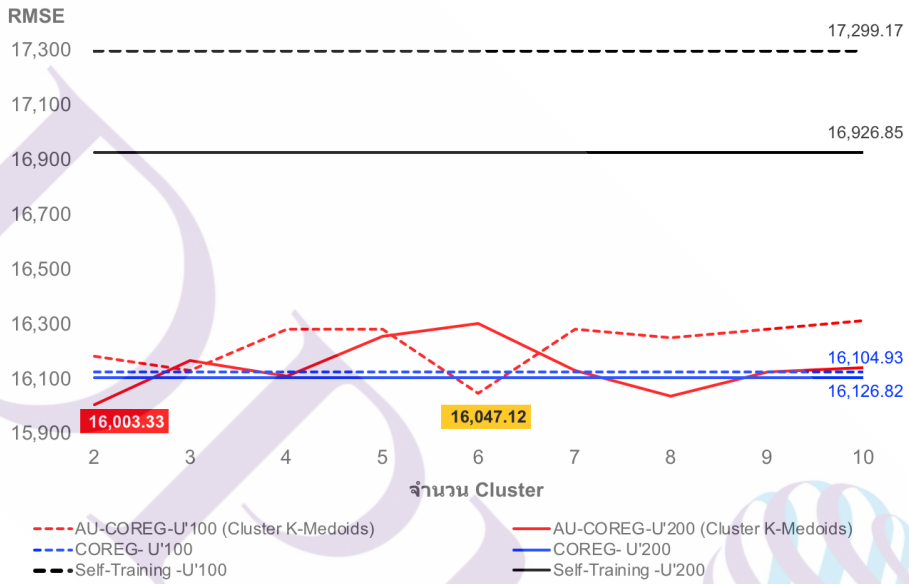
4.2 จะนำเสนอผลการเปรียบเทียบประสิทธิภาพของโมเดลและระยะเวลาที่ใช้ในการสร้างโมเดล AU-COREG ที่ได้จากการกำหนดพารามิเตอร์ที่ต่างกัน กับการสร้างโมเดลด้วยวิธี Self-Training และ COREG ของข้อมูลชุดที่ 2 ข้อมูลปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

4.3 จะนำเสนอผลการเปรียบเทียบประสิทธิภาพของโมเดลและระยะเวลาที่ใช้ในการสร้างโมเดล AU-COREG ที่ได้จากการกำหนดพารามิเตอร์ที่ต่างกัน กับการสร้างโมเดลด้วยวิธี Self-Training และ COREG ของข้อมูลชุดที่ 3 ข้อมูลพลังงานความร้อนร่วม

4.4 จะนำเสนอผลการเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG ที่ใช้กลุ่มตัวอย่างจากการสุ่มที่แตกต่างกัน 3 กลุ่ม ด้วยวิธีการจัดกลุ่มแบบ K-Medoids และได้เพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ100 ตัวอย่าง (U'=100) กับข้อมูลทั้ง 3 ชุดข้อมูล

4.1 ผลการเปรียบเทียบประสิทธิภาพของโมเดลและระยะเวลาที่ใช้ในการสร้างโมเดล AU-COREG กับการสร้างโมเดลด้วยวิธี Self-Training และ COREG ของข้อมูลชุดที่ 1 ข้อมูลโฆษณาประกาศรับสมัครงาน

จากผลการทดลองสามารถสรุปผลการเปรียบเทียบประสิทธิภาพของโมเดลและระยะเวลาในการสร้างโมเดลพยากรณ์เงินเดือน (Dollar) ด้วยโมเดล AU-COREG กับวิธี Self-Training และ COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงานได้ดังนี้

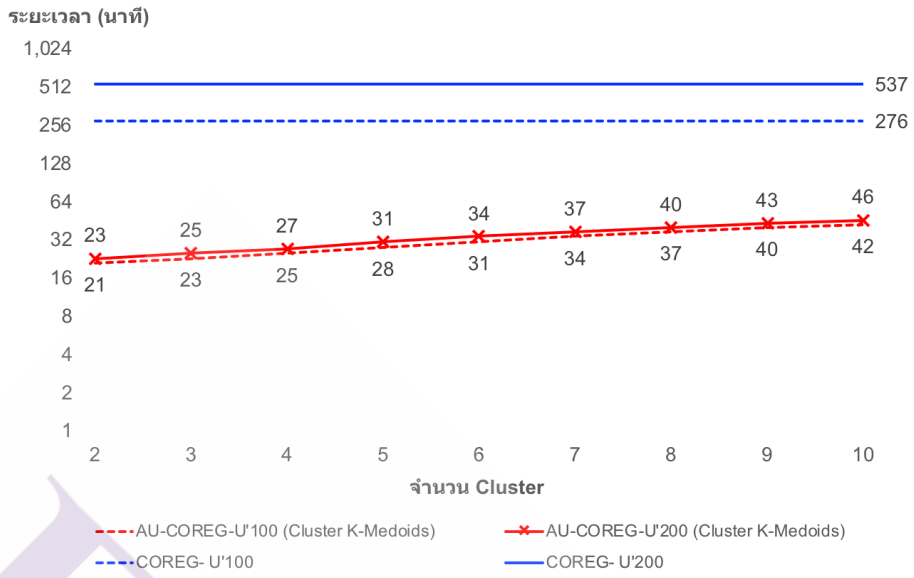


ภาพที่ 4.1 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์เงินเดือน (Dollar) ด้วยโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Medoids กับวิธี Self-Training และ COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน

4.1.1 การสร้างโมเดลด้วยวิธี Self-Training และวิธี COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน

4.1.1.1 วิธี Self-Training ด้วยโมเดล kNN ที่ได้จากการเรียนรู้จากชุดข้อมูลฝึกขนาด 100 และ 200 ตัวอย่าง จะได้ค่า RMSE เท่ากับ 17,299.17 และ 16,926.85 ตามลำดับ

4.1.1.2 โมเดล COREG ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ ขนาด 100 และ 200 ตัวอย่าง COREG(U'=100 และ U'=200) จะได้ค่า RMSE เท่ากับ 16,126.82 และ 16,104.93 ตามลำดับ



ภาพที่ 4.2 ภาพแสดงการเปรียบเทียบระยะเวลา (นาที) การสร้างโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Medoids กับ COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน

4.1.2 การเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Medoids กับวิธี Self-Training และ COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน

จากภาพที่ 4.1 และ 4.2 แสดงให้เห็นว่า

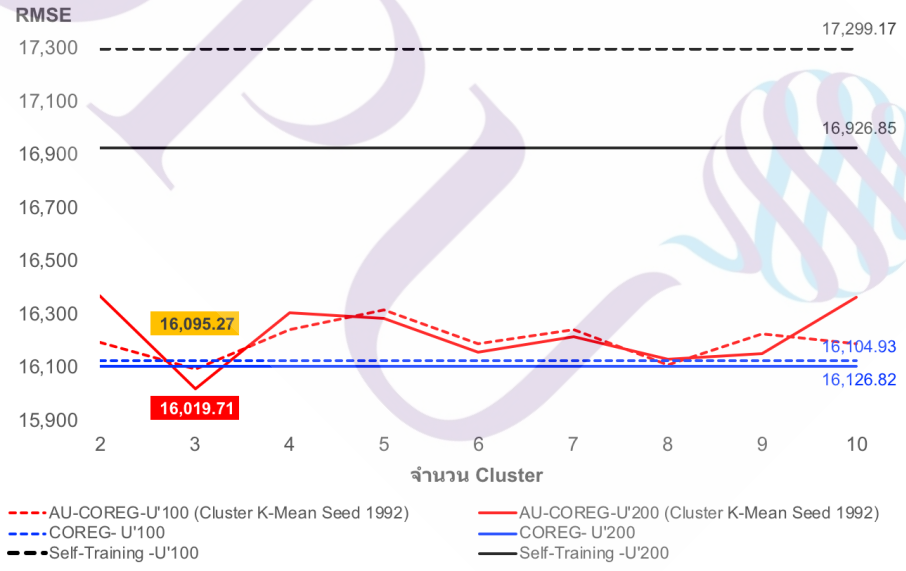
4.1.2.1 โมเดล AU-COREG (U'= 100) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Medoids) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 16,047.12 ซึ่งได้จากการทำ Cluster จำนวน 6 Cluster ลดลง 0.49% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 276 นาที เป็น 31 นาที หรือลดลงถึง 89% ดังตารางที่ 4.1

4.1.2.2 โมเดล AU-COREG (U'= 200) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Medoids) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 16,003.33 ซึ่งได้จากการทำ Cluster จำนวน 2 Cluster ลดลง 0.14% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 527 นาที เหลือเพียง 23 นาที หรือลดลงถึง 96% ดังตารางที่ 4.1

ตารางที่ 4.1 ผลการเปรียบเทียบประสิทธิภาพโมเดล AU-COREG กับโมเดล COREG ด้วยค่า RMSE ที่มีค่าน้อยที่สุดและระยะเวลา (นาที) โดยการจัดกลุ่มด้วยวิธี K-Medoids ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน

Model	U'100		U'200	
	RMSE	**Time (min)**	**RMSE**	**Time (min)**
COREG	16,126.82	276	16,104.93	527
AU-COREG โดยการจัดกลุ่มด้วยวิธี K-Medoids	16,047.12 (6 Cluster)	31	16,003.33 (2 Cluster)	23
ลดลง	0.49%	89%	0.63%	96%

4.1.3 การเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Mean (Seed 1992) กับโมเดล COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน

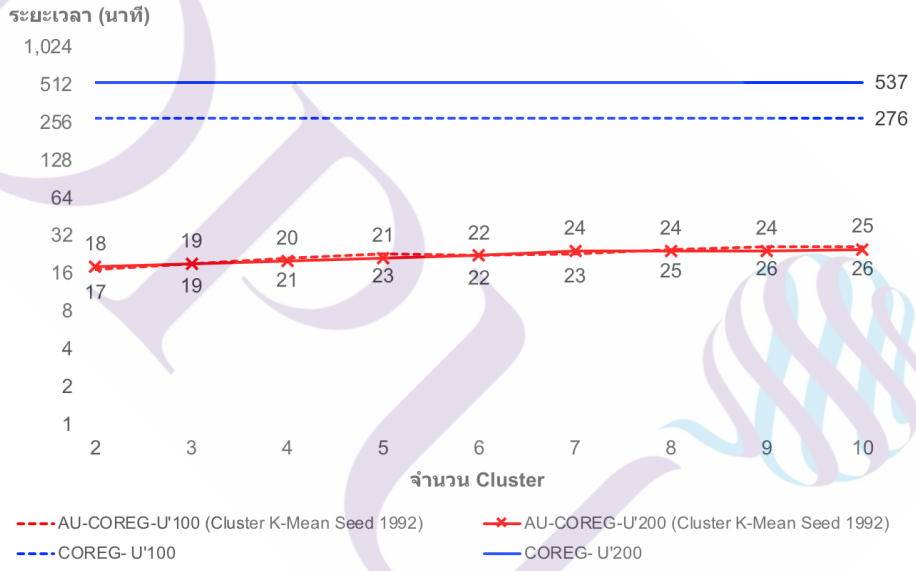


ภาพที่ 4.3 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์เงินเดือน (Dollar) ด้วยโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Mean (Seed 1992) กับวิธี Self-Training และ COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน

จากภาพที่ 4.3 และ 4.4 แสดงให้เห็นว่า

4.1.3.1 โมเดล AU-COREG (U'= 100) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 1992) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 16,095.27 ซึ่งได้จากการทำ Cluster จำนวน 3 Cluster ลดลง 0.06% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 276 นาที เหลือเพียง 19 นาที หรือคิดเป็น 93% ดังตารางที่ 4.2

4.1.3.2 โมเดล AU-COREG (U'= 200) ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 1992) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 16,019.71 ซึ่งได้จากการทำ Cluster จำนวน 3 Cluster ลดลง 0.67% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 537 เหลือเพียง 19 นาที หรือคิดเป็น 96% ดังตารางที่ 4.2

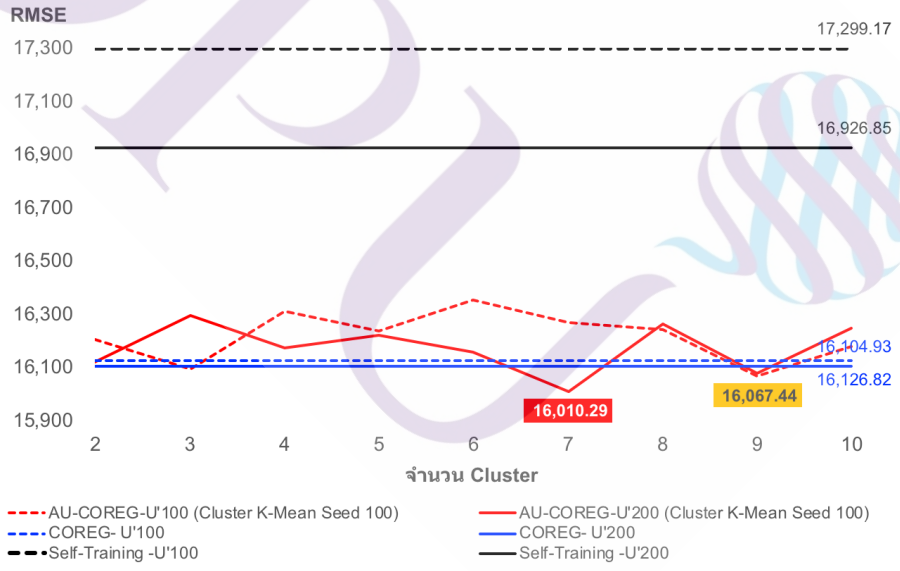


ภาพที่ 4.4 ภาพแสดงการเปรียบเทียบระยะเวลา (นาที) การสร้างโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Mean (Seed 1992) กับวิธี Self-Training และ COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน

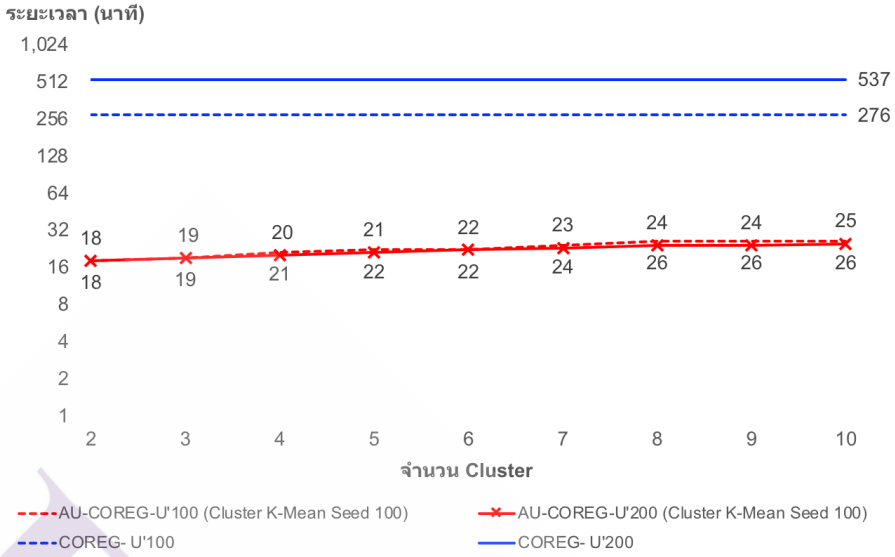
ตารางที่ 4.2 ผลการเปรียบเทียบประสิทธิภาพโมเดล AU-COREG กับโมเดล COREG ด้วยค่า RMSE ที่มีค่าน้อยที่สุดและระยะเวลา (นาที) โดยการจัดกลุ่มด้วยวิธี K-Mean (Seed 1992) ที่มีค่าน้อยที่สุดของชุดข้อมูลโฆษณาประกาศรับสมัครงาน

Model	U'100		U'200	
	RMSE	Time (min)	RMSE	Time (min)
COREG	16,126.82	276	16,104.93	527
AU-COREG โดยการจัดกลุ่มด้วยวิธี K-Mean (Seed 1992)	16,095.27 (3 Cluster)	19	16,019.71 (3 Cluster)	19
ลดลง	0.20%	92%	0.53%	97%

4.1.4 การเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Mean (Seed 100) กับโดเมล COREG



ภาพที่ 4.5 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์เงินเดือน (Dollar) ด้วยโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Mean (Seed 100) กับวิธี Self-Training และ COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน



ภาพที่ 4.6 ภาพแสดงการเปรียบเทียบระยะเวลา (นาที) การสร้างโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Mean (Seed 100) กับวิธี Self-Training และ COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน

จากภาพที่ 4.5 และ 4.6 แสดงให้เห็นว่า

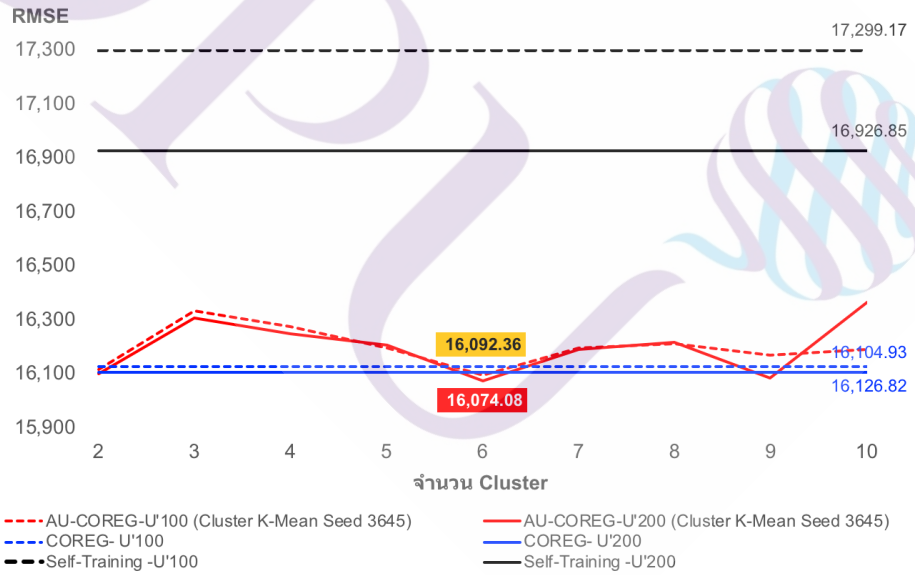
4.1.4.1 โมเดล AU-COREG (U'= 100) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 100) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 16,067.44 ซึ่งได้จากการทำ Cluster จำนวน 9 Cluster ลดลง 0.37% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 276 นาที เหลือเพียง 24 นาที หรือคิดเป็น 92% ดังตารางที่ 4.3

4.1.4.2 โมเดล AU-COREG (U'= 200) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 100) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 16,010.29 ซึ่งได้จากการทำ Cluster จำนวน 7 Cluster ลดลง 0.59% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 537 เหลือเพียง 24 นาที หรือคิดเป็น 97% ดังตารางที่ 4.3

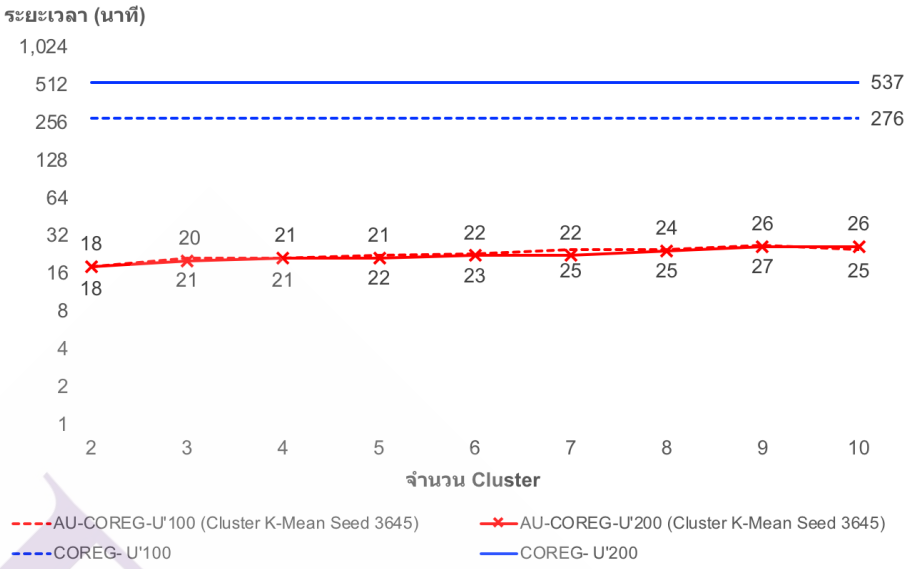
ตารางที่ 4.3 ผลการเปรียบเทียบประสิทธิภาพโมเดล AU-COREG กับโมเดล COREG ด้วยค่า RMSE ที่มีค่าน้อยที่สุดและระยะเวลา (นาที) โดยการจัดกลุ่มด้วยวิธี K-Mean (Seed 100) ที่มีค่าน้อยที่สุดของชุดข้อมูลโฆษณาประกาศรับสมัครงาน

Model	U'100		U'200	
	RMSE	Time (min)	RMSE	Time (min)
COREG	16,126.82	276	16,104.93	527
AU-COREG โดยการจัดกลุ่มด้วยวิธี K-Mean (Seed 100)	16,067.44 (9 Cluster)	24	16,019.71 (7 Cluster)	24
ลดลง	0.37%	92%	0.59%	97%

4.1.5 การเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Mean (Seed 3645) กับโดเมล COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน



ภาพที่ 4.7 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์เงินเดือน (Dollar) ด้วยโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Mean (Seed 3645) กับวิธี Self-Training และ COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน



ภาพที่ 4.8 ภาพแสดงการเปรียบเทียบระยะเวลา (นาที) การสร้างโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Mean (Seed 3645) กับวิธี Self-Training และ COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน

จากภาพที่ 4.7 และ 4.8 แสดงให้เห็นว่า

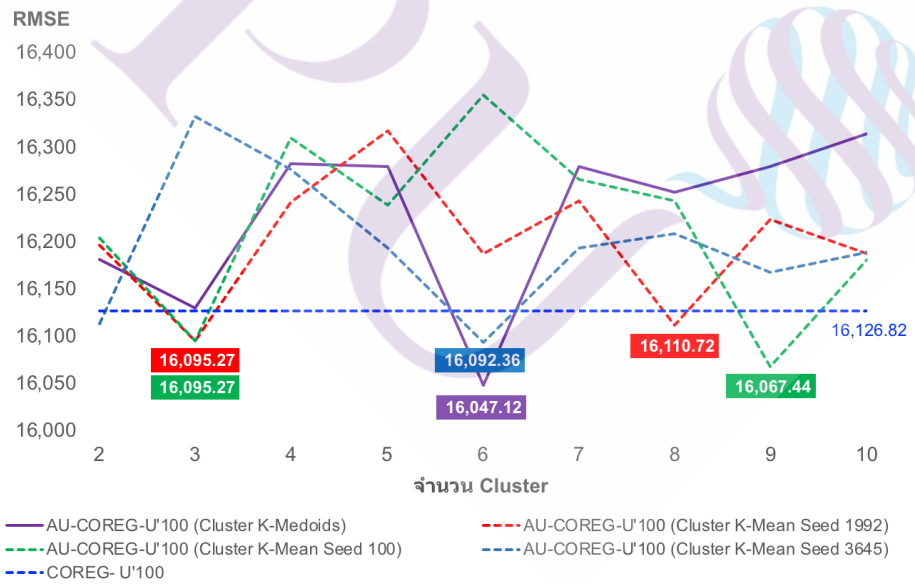
4.1.5.1 โมเดล AU-COREG (U'= 100) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 3645) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 16,092.36 ซึ่งได้จากการทำ Cluster จำนวน 6 Cluster ลดลง 0.21% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 276 นาที เหลือเพียง 22 นาที หรือคิดเป็น 92% ดังตารางที่ 4.4

4.1.5.2 โมเดล AU-COREG (U'= 200) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 3645) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 16,074.08 ซึ่งได้จากการทำ Cluster จำนวน 6 Cluster ลดลง 0.19% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 537 เหลือเพียง 23 นาที หรือคิดเป็น 96% ดังตารางที่ 4.4

ตารางที่ 4.4 ผลการเปรียบเทียบประสิทธิภาพโมเดล AU-COREG กับโมเดล COREG ด้วยค่า RMSE ที่มีค่าน้อยที่สุดและระยะเวลา (นาที) โดยการจัดกลุ่มด้วยวิธี K-Mean (Seed 3649) ที่มีค่าน้อยที่สุดของชุดข้อมูลโฆษณาประกาศรับสมัครงาน

Model	U'100		U'200	
	RMSE	Time (min)	RMSE	Time (min)
COREG	16,126.82	276	16,104.93	527
AU-COREG โดยการจัดกลุ่มด้วยวิธี K-Mean (Seed 3645)	16,092.36 (6 Cluster)	22	16,074.08 (6 Cluster)	23
ลดลง	0.21%	92%	0.19%	97%

4.1.6 การเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG (U'=100) ด้วยวิธีการจัดกลุ่มแบบ K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645 ตามลำดับ กับโมเดล COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน



ภาพที่ 4.9 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์เงินเดือน (Dollar) ด้วยโมเดล AU-COREG (U'=100) โดยวิธีการจัดกลุ่มแบบ K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645 ตามลำดับ กับโมเดล COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน

จากภาพที่ 4.9 แสดงให้เห็นว่าโมเดล AU-COREG (U'= 100) ในทุก ๆ วิธีการจัดกลุ่ม (K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645) ให้ค่า RMSE น้อยกว่าหรือเท่ากับค่า RMSE ที่ได้จากโมเดล COREG ดังนี้

4.1.6.1 การจัดกลุ่ม 3, 6, 8 และ 9 Cluster ทำให้ค่า RMSE ของโมเดล AU-COREG น้อยกว่าค่า RMSE ที่ได้จากโมเดล COREG

4.1.6.2 โมเดล AU-COREG โดยวิธีการจัดกลุ่ม 4 วิธี (K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645) จำนวนกลุ่มที่ให้ค่า RMSE น้อยกว่าหรือเท่ากับ RMSE จากโมเดล COREG ที่ดีที่สุดคือ จำนวน 3 และ 6 Cluster

4.1.6.3 โมเดล AU-COREG โดยวิธีการจัดกลุ่มที่ให้ค่า RMSE น้อยกว่าหรือเท่ากับ RMSE ที่ได้จากโมเดล COREG ที่ดีที่สุด คือ วิธี K-Mean ที่กำหนด Seed เท่ากับ 1992 และ 100

4.1.6.4 โมเดล AU-COREG โดยวิธีการจัดกลุ่มที่ให้ค่า RMSE น้อยที่สุด ได้แก่ วิธีการจัดกลุ่มโดยวิธี K-Medoids จำนวน 6 Cluster

4.1.7 การเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG (U'=200) ด้วยวิธีการจัดกลุ่มแบบ K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645 ตามลำดับ กับโมเดล COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน

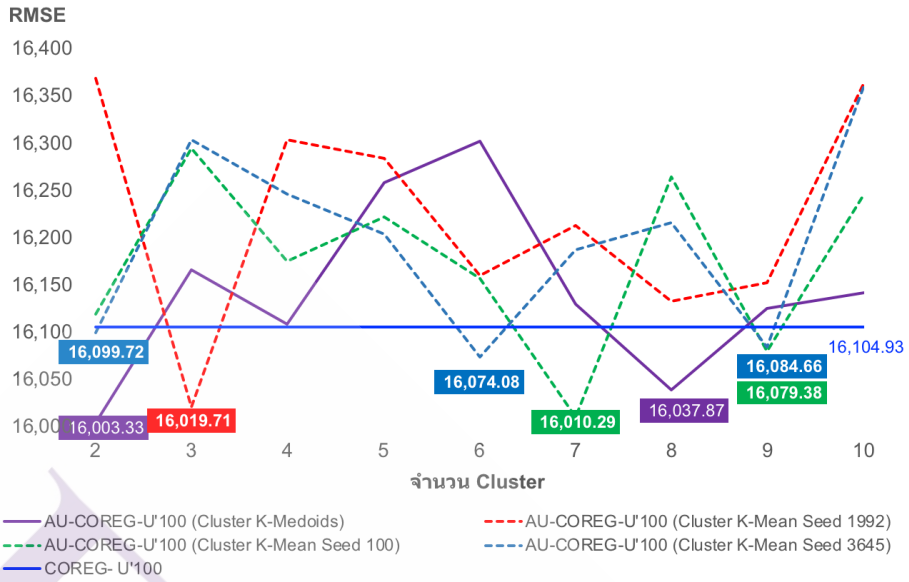
จากภาพที่ 4.10 แสดงให้เห็นว่าโมเดล AU-COREG (U'= 200) ในทุก ๆ วิธีการจัดกลุ่ม (K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645) ให้ค่า RMSE น้อยกว่าหรือเท่ากับค่า RMSE ที่ได้จากโมเดล COREG ดังนี้

4.1.7.1 การจัดกลุ่ม 2, 3, 6, 7, 8 และ 9 Cluster ทำให้ค่า RMSE ของโมเดล AU-COREG น้อยกว่าค่า RMSE ที่ได้จากโมเดล COREG

4.1.7.2 โมเดล AU-COREG โดยวิธีการจัดกลุ่ม 4 วิธี (K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645) จำนวนกลุ่มที่ให้ค่า RMSE น้อยกว่าหรือเท่ากับ RMSE จากโมเดล COREG ที่ดีที่สุดคือ จำนวน 2 และ 9 Cluster

4.1.7.3 โมเดล AU-COREG โดยวิธีการจัดกลุ่มที่ให้ค่า RMSE น้อยกว่าหรือเท่ากับ RMSE ที่ได้จากโมเดล COREG ที่ดีที่สุด คือ วิธี K-Mean ที่กำหนด Seed เท่ากับ 3645

4.1.7.4 โมเดล AU-COREG โดยวิธีการจัดกลุ่มที่ให้ค่า RMSE น้อยที่สุด ได้แก่ วิธีการจัดกลุ่มโดยวิธี K-Medoids จำนวน 2 Cluster



ภาพที่ 4.10 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์เงินเดือน (Dollar) ด้วยโมเดล AU-COREG (U'=200) โดยวิธีการจัดกลุ่มแบบ K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645 ตามลำดับ กับโมเดล COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน

4.1.8 การเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG จากการสุ่มข้อมูลในขั้นแรกด้วยวิธีการจัดกลุ่มแบบ K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645 ตามลำดับ กับโมเดล COREG ของชุดข้อมูลโฆษณาประกาศรับสมัครงาน

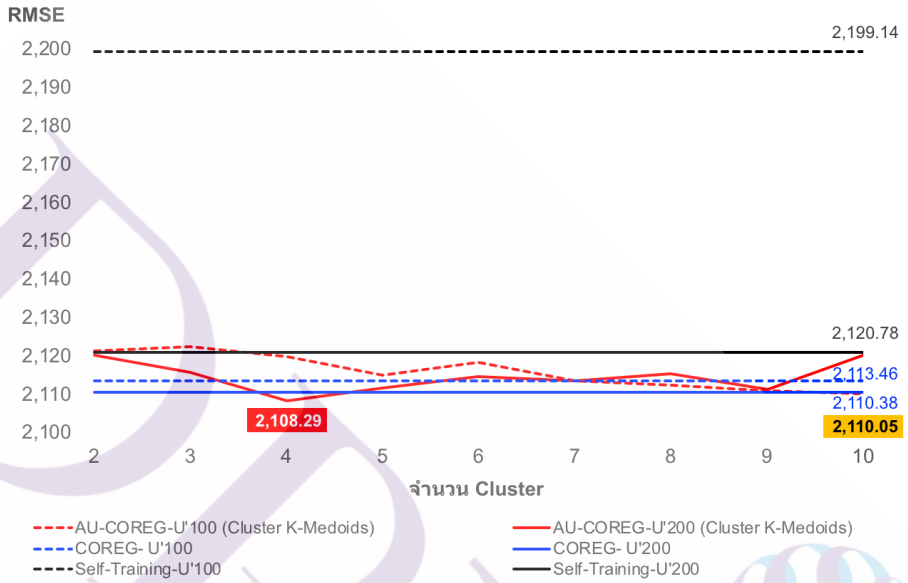
4.2 ผลการเปรียบเทียบประสิทธิภาพของโมเดลและระยะเวลาที่ใช้ในการสร้างโมเดล AU-COREG กับการสร้างโมเดลด้วยวิธี Self-Training และ COREG ของข้อมูลชุดที่ 2 ข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

จากผลการทดลองสามารถสรุปผลการเปรียบเทียบประสิทธิภาพของโมเดลและระยะเวลาในการสร้างโมเดลพยากรณ์จำนวนคนที่ใช้บริการต่อชั่วโมงด้วยโมเดล AU-COREG กับวิธี Self-Training และ COREG ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดินได้ดังนี้

4.2.1 การสร้างโมเดลด้วยวิธี Self-Training และวิธี COREG ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

4.2.1.1 วิธี Self-Training ด้วยโมเดล kNN ที่ได้จากการเรียนรู้จากชุดข้อมูลฝึกขนาด 100 และ 200 ตัวอย่าง จะได้ค่า RMSE เท่ากับ 2,199.14 และ 2,120.78 ตามลำดับ

4.2.1.2 โมเดล COREG ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ ขนาด 100 และ 200 ตัวอย่าง COREG(U'=100 และ U'=200) จะได้ค่า RMSE เท่ากับ 2,112.46 และ 2,110.38 ตามลำดับ



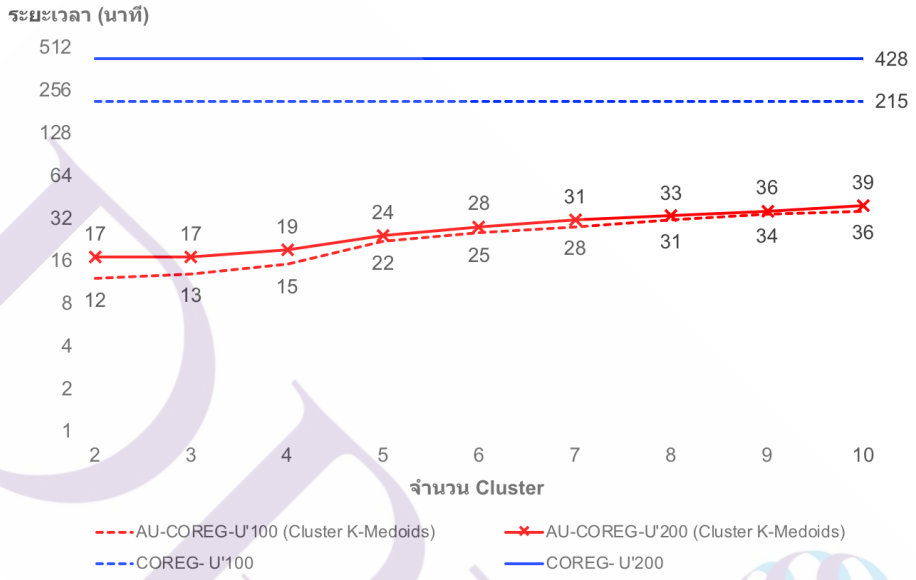
ภาพที่ 4.11 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์จำนวนคนที่ใช้บริการต่อชั่วโมง ด้วยโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Medoids กับวิธี Self-Training และ COREG ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

4.2.2 การสร้างโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Medoids ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

จากภาพที่ 4.11 และ 4.12 แสดงให้เห็นว่า

4.2.2.1 โมเดล AU-COREG (U'= 100) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Medoids) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 2,110.05 ซึ่งได้จากการทำ Cluster จำนวน 10 Cluster ลดลง 0.16% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 215 นาที เหลือเพียง 36 นาที หรือคิดเป็น 83% ดังตารางที่ 4.5

4.2.2.2 โมเดล AU-COREG (U'= 200) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Medoids) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 2,108.29 ซึ่งได้จากการทำ Cluster จำนวน 4 Cluster ลดลง 0.10% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 428 นาที เหลือ 24 นาที หรือคิดเป็น 96% ดังตารางที่ 4.5

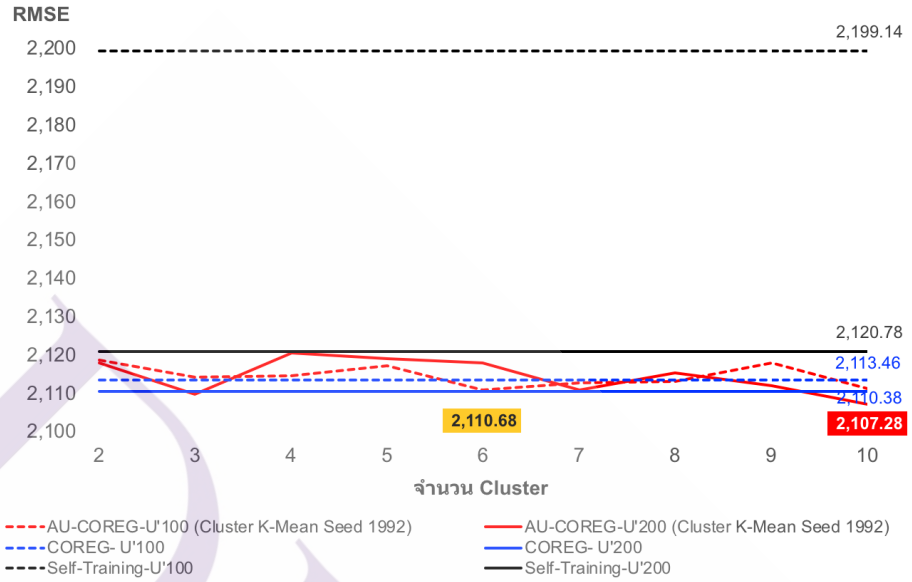


ภาพที่ 4.12 ภาพแสดงการเปรียบเทียบระยะเวลา (นาที) การสร้างโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Medoids กับวิธี Self-Training และ COREG ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

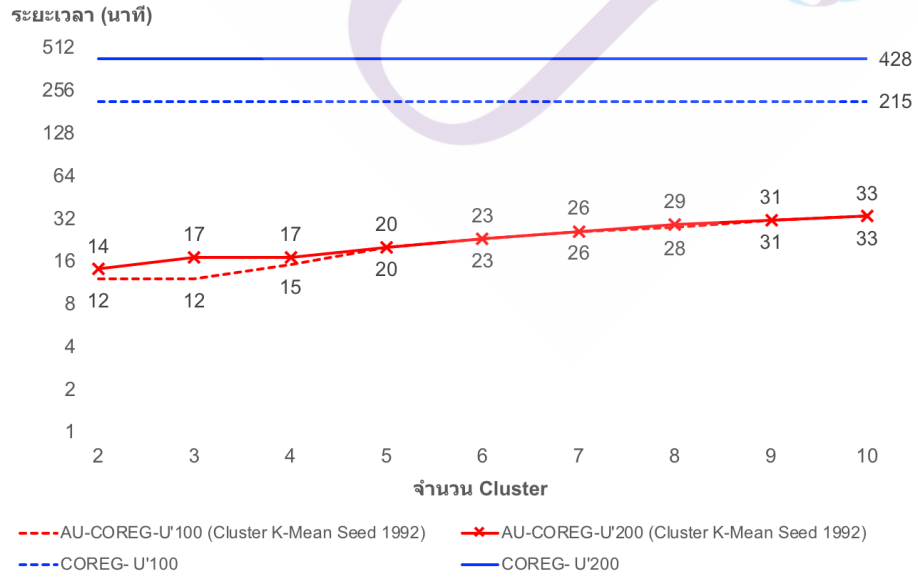
ตารางที่ 4.5 ผลการเปรียบเทียบประสิทธิภาพโมเดล AU-COREG กับโมเดล COREG ด้วยค่า RMSE ที่มีค่าน้อยที่สุดและระยะเวลา (นาที) โดยการจัดกลุ่มด้วยวิธี K-Medoids ที่มีค่าน้อยที่สุดของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

Model	**U'100**		**U'200**	
	RMSE	**Time (min)**	**RMSE**	**Time (min)**
COREG	2,113.46	215	2,110.38	428
AU-COREG โดยการจัดกลุ่มด้วยวิธี K-Medoids	2,110.05 (10 Cluster)	36	2,108.29 (4 Cluster)	24
ลดลง	0.16%	83%	0.10%	96%

4.2.3 การสร้างโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Mean Seed 1992 ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน



ภาพที่ 4.13 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์จำนวนคนที่ใช้บริการต่อชั่วโมง ด้วยโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Mean (Seed 1992) กับวิธี Self-Training และ COREG ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐโดยรถไฟฟ้าใต้ดิน



ภาพที่ 4.14 ภาพแสดงการเปรียบเทียบระยะเวลา (นาที) การสร้างโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Mean (Seed 1992) กับวิธี Self-Training และ COREG ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

จากภาพที่ 4.13 และ 4.14 แสดงให้เห็นว่า

4.2.3.1 โมเดล AU-COREG (U'= 100) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 1992) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 2,110.68 ซึ่งได้จากการทำ Cluster จำนวน 6 Cluster ลดลง 0.13% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 215 นาที เหลือเพียง 23 นาที หรือคิดเป็น 89% ดังตารางที่ 4.6

4.2.3.2 โมเดล AU-COREG (U'= 200) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 1992) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 2,107.28 ซึ่งได้จากการทำ Cluster จำนวน 10 Cluster ลดลง 0.15% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 428 นาที เหลือเพียง 33 นาที หรือคิดเป็น 92% ดังตารางที่ 4.6

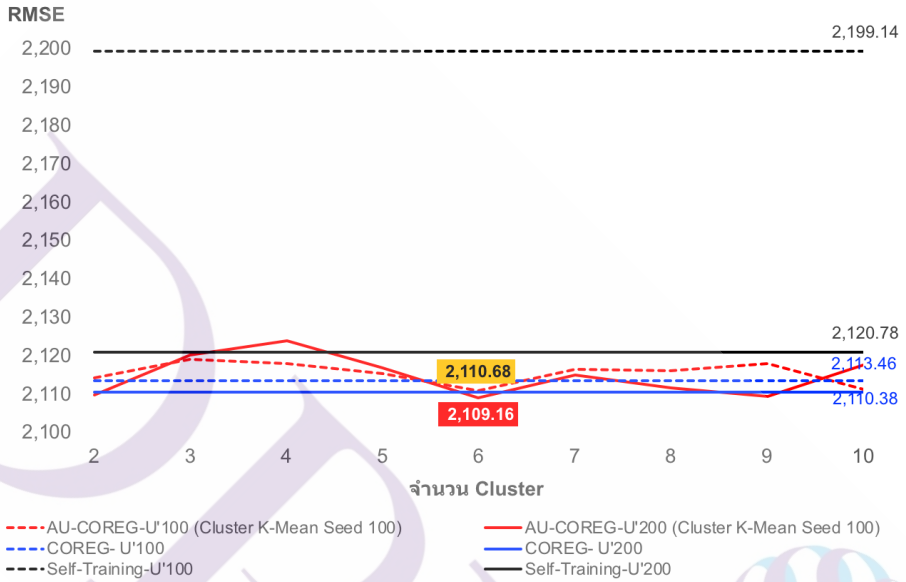
ตารางที่ 4.6 ผลการเปรียบเทียบประสิทธิภาพโมเดล AU-COREG กับโมเดล COREG ด้วยค่า RMSE ที่มีค่าน้อยที่สุดและระยะเวลา (นาที) โดยการจัดกลุ่มด้วยวิธี K-Mean Seed 1992 ที่มีค่าน้อยที่สุดของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

Model	**U'100**		**U'200**	
	RMSE	**Time (min)**	**RMSE**	**Time (min)**
COREG	2,113.46	215	2,110.38	428
AU-COREG โดยการจัดกลุ่มด้วยวิธี K-Mean Seed 1992	2,110.68 (6 Cluster)	23	2,107.28 (10 Cluster)	33
ลดลง	0.13%	89%	0.10%	92%

4.2.4 การสร้างโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Mean Seed 100 ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

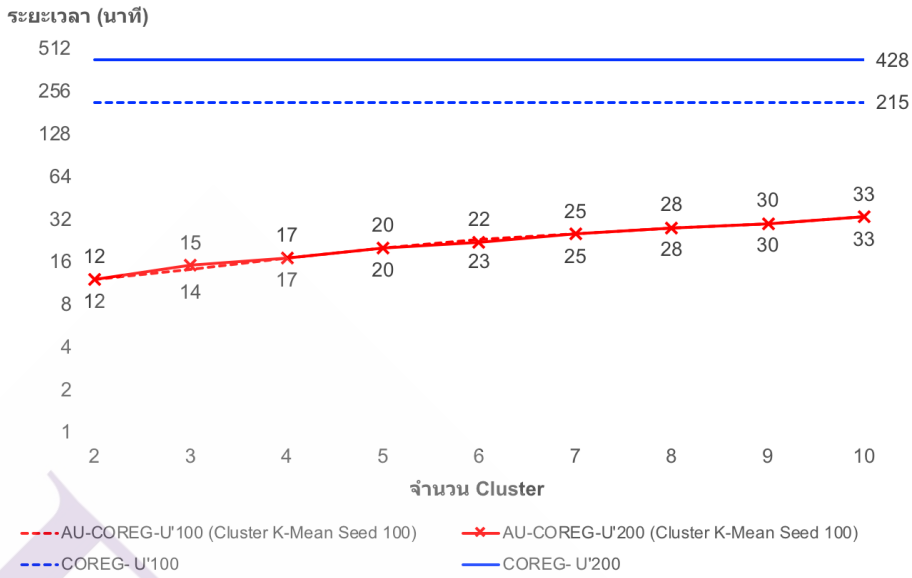
จากภาพที่ 4.15 และ 4.16 แสดงให้เห็นว่า

4.2.4.1 โมเดล AU-COREG (U'= 100) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 100) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 2,110.68 ซึ่งได้จากการทำ Cluster จำนวน 6 Cluster ลดลง 0.13% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 215 นาที เหลือเพียง 23 นาที หรือคิดเป็น 89% ดังตารางที่ 4.7



ภาพที่ 4.15 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์จำนวนคนที่ใช้บริการต่อชั่วโมง ด้วยโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Mean (Seed 100) กับวิธี Self-Training และ COREG ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐโดยรถไฟฟ้าใต้ดิน

4.2.4.2 โมเดล AU-COREG (U'= 200) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 100) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 2,109.16 ซึ่งได้จากการทำ Cluster จำนวน 6 Cluster ลดลง 0.06% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 428 นาที เหลือเพียง 22 นาที หรือคิดเป็น 95% ดังตารางที่ 4.7



ภาพที่ 4.16 ภาพแสดงการเปรียบเทียบระยะเวลา (นาที) การสร้างโมเดล AU-COREG ด้วยโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Mean (Seed 100) กับวิธี Self-Training และ COREG ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

ตารางที่ 4.7 ผลการเปรียบเทียบประสิทธิภาพโมเดล AU-COREG กับโมเดล COREG ด้วยค่า RMSE ที่มีค่าน้อยที่สุดและระยะเวลา (นาที) โดยการจัดกลุ่มด้วยวิธี K-Mean Seed 100 มีค่าน้อยที่สุดของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

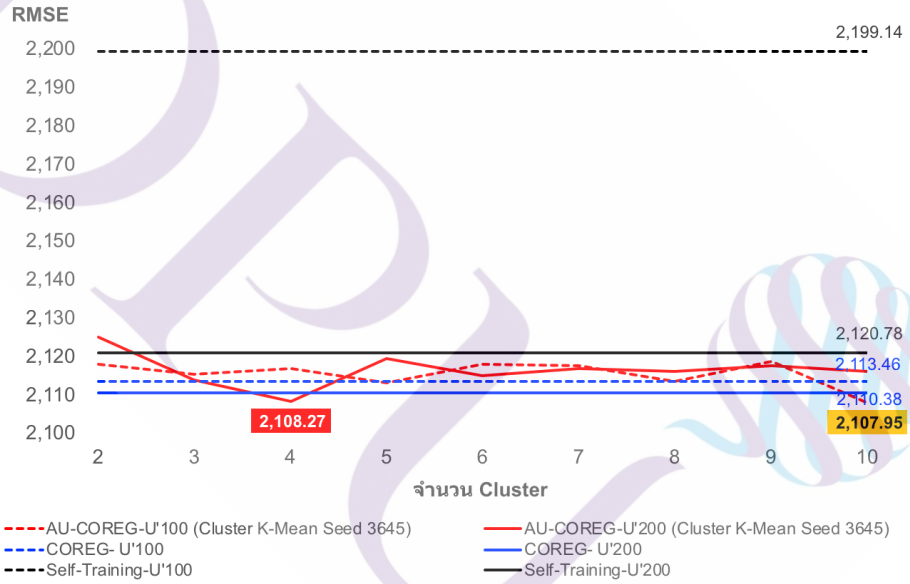
Model	**U'100**		**U'200**	
	RMSE	**Time (min)**	**RMSE**	**Time (min)**
COREG	2,113.46	215	2,110.38	428
AU-COREG โดยการจัดกลุ่มด้วยวิธี K-Mean Seed 100	2,110.68 (6 Cluster)	23	2,109.16 (6 Cluster)	22
ลดลง	0.13%	89%	0.06%	95%

4.2.5 การสร้างโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Mean Seed 3645 ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

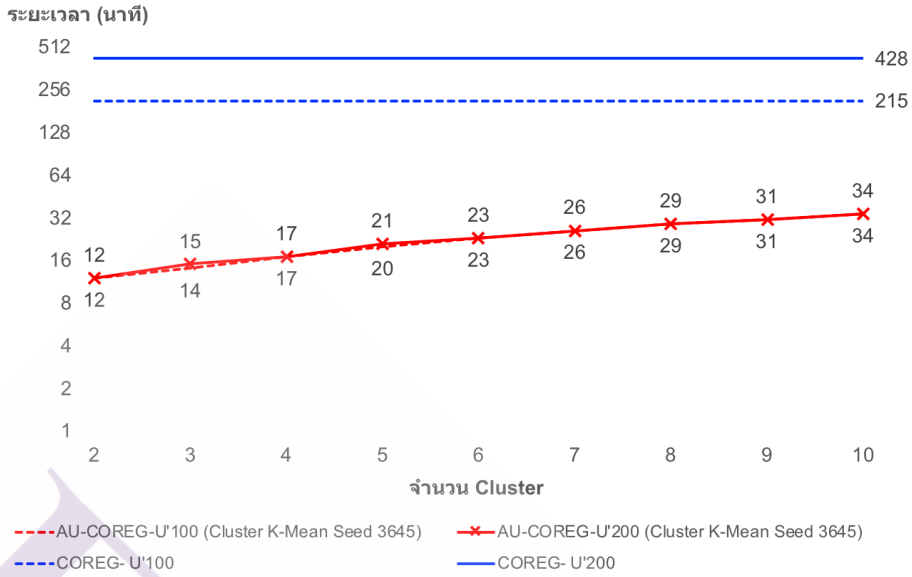
จากภาพที่ 4.17 และ 4.18 แสดงให้เห็นว่า

4.2.5.1 โมเดล AU-COREG (U'= 100) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 3645) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 2,107.95 ซึ่งได้จากการทำ Cluster จำนวน 10 Cluster ลดลง 0.26% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 215 นาที เหลือเพียง 34 นาที หรือคิดเป็น 84% ดังตารางที่ 4.8

4.2.5.2 โมเดล AU-COREG (U'= 200) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 3645) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 2,108.27 ซึ่งได้จากการทำ Cluster จำนวน 4 Cluster ลดลง 0.10% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 428 นาที เหลือเพียง 17 นาที หรือคิดเป็น 96% ดังตารางที่ 4.8



ภาพที่ 4.17 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์จำนวนคนที่ใช้บริการต่อชั่วโมง ด้วยโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Mean (Seed 3645) กับวิธี Self-Training และ COREG ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐโดยรถไฟฟ้าใต้ดิน



ภาพที่ 4.18 ภาพแสดงการเปรียบเทียบระยะเวลา (นาที) การสร้างโมเดล AU-COREG ด้วยโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Mean (Seed 3645) กับวิธี Self-Training และ COREG ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

ตารางที่ 4.8 ผลการเปรียบเทียบประสิทธิภาพโมเดล AU-COREG กับโมเดล COREG ด้วยค่า RMSE ที่มีค่าน้อยที่สุดและระยะเวลา (นาที) โดยการจัดกลุ่มด้วยวิธี K-Mean Seed 3645 ที่มีค่าน้อยที่สุดของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

Model	**U'100**		**U'200**	
	RMSE	**Time (min)**	**RMSE**	**Time (min)**
COREG	2,113.46	215	2,110.38	428
AU-COREG โดยการจัดกลุ่มด้วยวิธี K-Mean Seed 3645	2,107.95 (10 Cluster)	34	2,108.27 (4 Cluster)	22
ลดลง	0.26%	84%	0.10%	97%

4.2.6 การเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG (U'=100) ด้วยวิธีการจัดกลุ่มแบบ K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645 ตามลำดับ กับโมเดล COREG ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

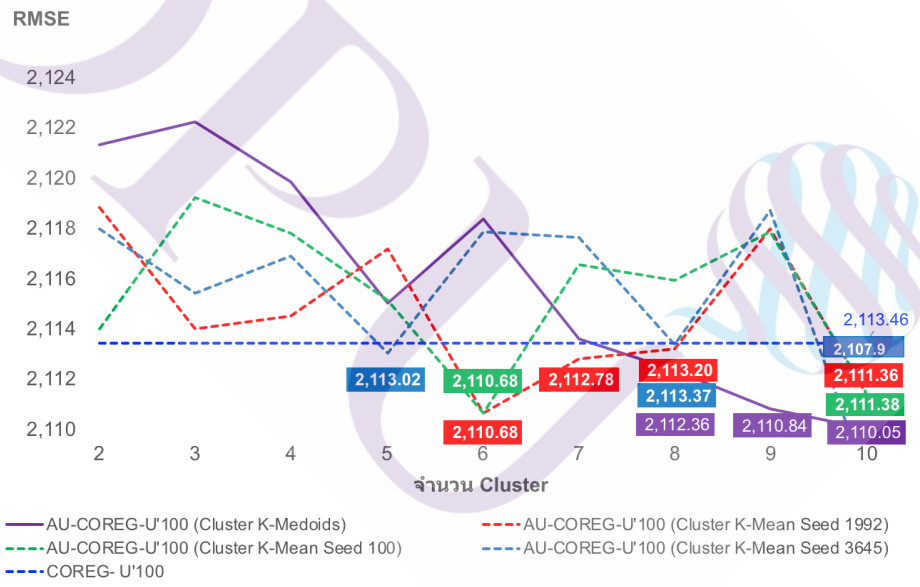
จากภาพที่ 4.19 แสดงให้เห็นว่าโมเดล AU-COREG (U'= 100) ในทุก ๆ วิธีการจัดกลุ่ม (K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645) ให้ค่า RMSE น้อยกว่าหรือเท่ากับค่า RMSE ที่ได้จากโมเดล COREG ดังนี้

4.2.6.1 การจัดกลุ่ม 5, 6, 7, 8, 9 และ 10 Cluster ทำให้ค่า RMSE ของโมเดล AU-COREG น้อยกว่าค่า RMSE ที่ได้จากโมเดล COREG

4.2.6.2 โมเดล AU-COREG โดยวิธีการจัดกลุ่ม 4 วิธี (K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645) จำนวนกลุ่มที่ให้ค่า RMSE น้อยกว่าหรือเท่ากับ RMSE จากโมเดล COREG ที่ดีที่สุดคือ จำนวน 10 Cluster

4.2.6.3 โมเดล AU-COREG โดยวิธีการจัดกลุ่มที่ให้ค่า RMSE น้อยกว่าหรือเท่ากับ RMSE ที่ได้จากโมเดล COREG ที่ดีที่สุด คือ วิธี K-Mean ที่กำหนด Seed เท่ากับ 1992

4.2.6.4 โมเดล AU-COREG โดยวิธีการจัดกลุ่มที่ให้ค่า RMSE น้อยที่สุด ได้แก่ วิธีการจัดกลุ่มโดยวิธี K-Mean ที่กำหนด Seed เท่ากับ 3645 จำนวน 10 Cluster



ภาพที่ 4.19 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์จำนวนคนที่ใช้บริการต่อชั่วโมง ด้วยโมเดล AU-COREG (U'=100) โดยวิธีการจัดกลุ่มแบบ K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645 ตามลำดับ กับโมเดล COREG ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

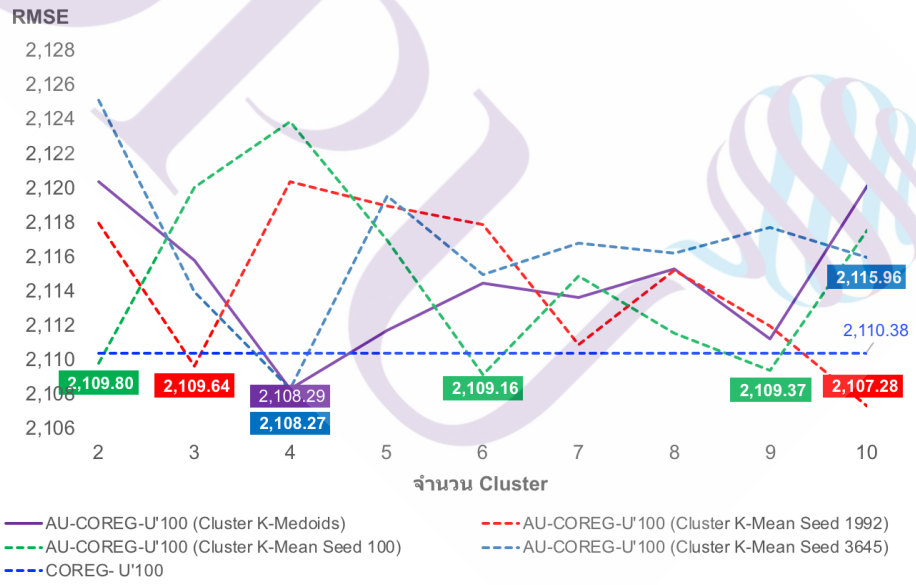
4.2.7 การเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG (U'=200) ด้วยวิธีการจัดกลุ่มแบบ K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645 ตามลำดับ กับโมเดล COREG ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

จากภาพที่ 4.20 แสดงให้เห็นว่าโมเดล AU-COREG (U'= 100) ในทุก ๆ วิธีการจัดกลุ่ม (K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645) ให้ค่า RMSE น้อยกว่าหรือเท่ากับค่า RMSE ที่ได้จากโมเดล COREG ดังนี้

4.2.7.1 การจัดกลุ่ม 2, 3, 4, 6 และ9 Cluster ทำให้ค่า RMSE ของโมเดล AU-COREG น้อยกว่าค่า RMSE ที่ได้จากโมเดล COREG

4.2.7.2 โมเดล AU-COREG โดยวิธีการจัดกลุ่ม 4 วิธี (K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645) จำนวนกลุ่มที่ให้ค่า RMSE น้อยกว่าหรือเท่ากับ RMSE จากโมเดล COREG ที่ดีที่สุดคือ จำนวน 4 Cluster

4.2.7.3 โมเดล AU-COREG โดยวิธีการจัดกลุ่มที่ให้ค่า RMSE น้อยกว่าหรือเท่ากับ RMSE ที่ได้จากโมเดล COREG ที่ดีที่สุด คือ วิธี K-Mean ที่กำหนด Seed เท่ากับ 100



ภาพที่ 4.20 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์จำนวนคนที่ใช้บริการต่อชั่วโมง ด้วยโมเดล AU-COREG (U'=200) โดยวิธีการจัดกลุ่มแบบ K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645 ตามลำดับ กับโมเดล COREG ของชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน

4.2.7.4 โมเดล AU-COREG โดยวิธีการจัดกลุ่มที่ให้ค่า RMSE น้อยที่สุด ได้แก่ วิธีการจัดกลุ่มโดยวิธี K-Mean ที่กำหนด Seed เท่ากับ 1992 จำนวน 10 Cluster

4.3 ผลการเปรียบเทียบประสิทธิภาพของโมเดลและระยะเวลาที่ใช้ในการสร้างโมเดล AU-COREG กับการสร้างโมเดลด้วยวิธี Self-Training และ COREG ของข้อมูลชุดที่ 3 ข้อมูลพลังงานความร้อนร่วม

จากผลการทดลองสามารถสรุปผลการเปรียบเทียบประสิทธิภาพของโมเดลและระยะเวลาในการสร้างโมเดลพยากรณ์พลังงานไฟฟ้าต่อชั่วโมง ด้วยโมเดล AU-COREG กับวิธี Self-Training และ COREG ของชุดข้อมูลพลังงานความร้อนร่วมได้ดังนี้

4.3.1 การสร้างโมเดลด้วยวิธี Self-Training และวิธี COREG ของชุดข้อมูลพลังงานความร้อนร่วม

4.3.1.1 วิธี Self-Training ด้วยโมเดล kNN ที่ได้จากการเรียนรู้จากชุดข้อมูลฝึกขนาด 100 และ 200 ตัวอย่าง จะได้ค่า RMSE เท่ากับ 10.76 และ 10.67 ตามลำดับ

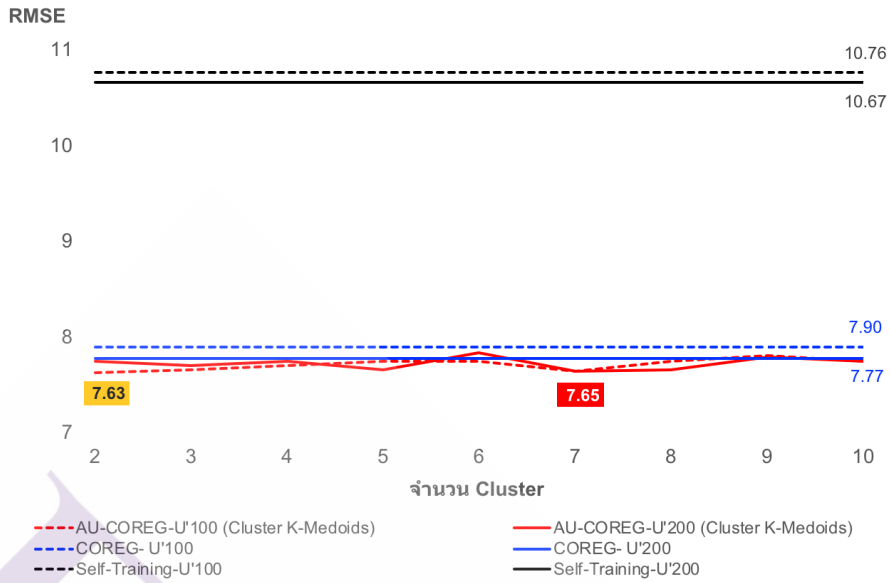
4.3.1.2 โมเดล COREG ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ ขนาด 100 และ 200 ตัวอย่าง COREG(U'=100 และ U'=200) จะได้ค่า RMSE เท่ากับ 7.90 และ 7.77 ตามลำดับ

4.3.2 การสร้างโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Medoids ของชุดข้อมูลพลังงานความร้อนร่วม

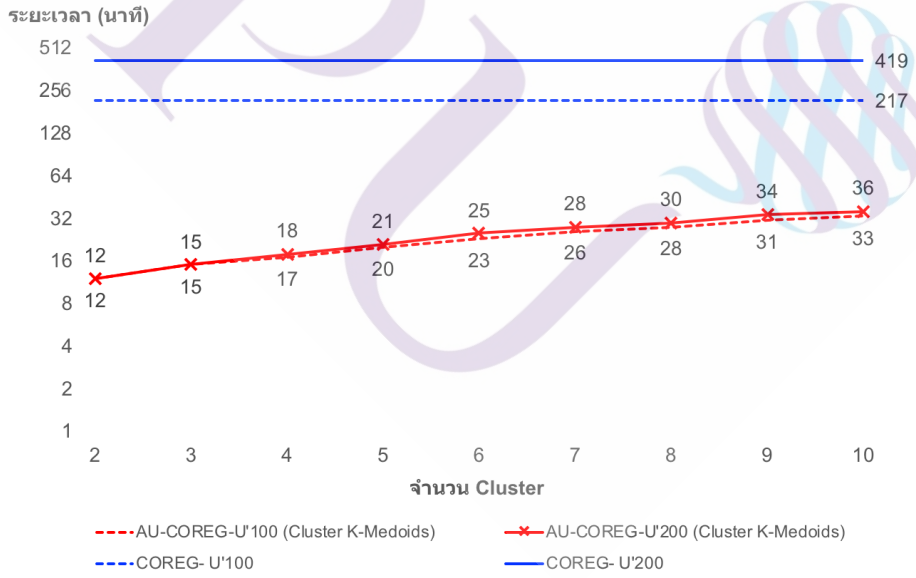
จากภาพที่ 4.21 และ 4.22 แสดงให้เห็นว่า

4.3.2.1 โมเดล AU-COREG (U'= 100) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Medoids) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 7.63 ซึ่งได้จากการทำ Cluster จำนวน 2 Cluster ลดลง 1.80% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 217 นาที เหลือเพียง 12 นาที หรือคิดเป็น 94% ดังตารางที่ 4.9

4.3.2.2 โมเดล AU-COREG (U'= 200) ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Medoids) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 7.65 ซึ่งได้จากการทำ Cluster จำนวน 7 Cluster ลดลง 3.21% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 419 นาที เหลือเพียง 28 นาที หรือคิดเป็น 93% ดังตารางที่ 4.9



ภาพที่ 4.21 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์พลังงานไฟฟ้าต่อชั่วโมง ด้วยโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Medoids กับวิธี Self-Training และ COREG ของชุดข้อมูลพลังงานความร้อนร่วม



ภาพที่ 4.22 ภาพแสดงการเปรียบเทียบระยะเวลา (นาที) การสร้างโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Medoids กับวิธี Self-Training และ COREG ของชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม

ตารางที่ 4.9 ผลการเปรียบเทียบประสิทธิภาพโมเดล AU-COREG กับโมเดล COREG ด้วยค่า RMSE ที่มีค่าน้อยที่สุดและระยะเวลา (นาที) โดยการจัดกลุ่มด้วยวิธี K-Medoids ที่มีค่าน้อยที่สุดของชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม

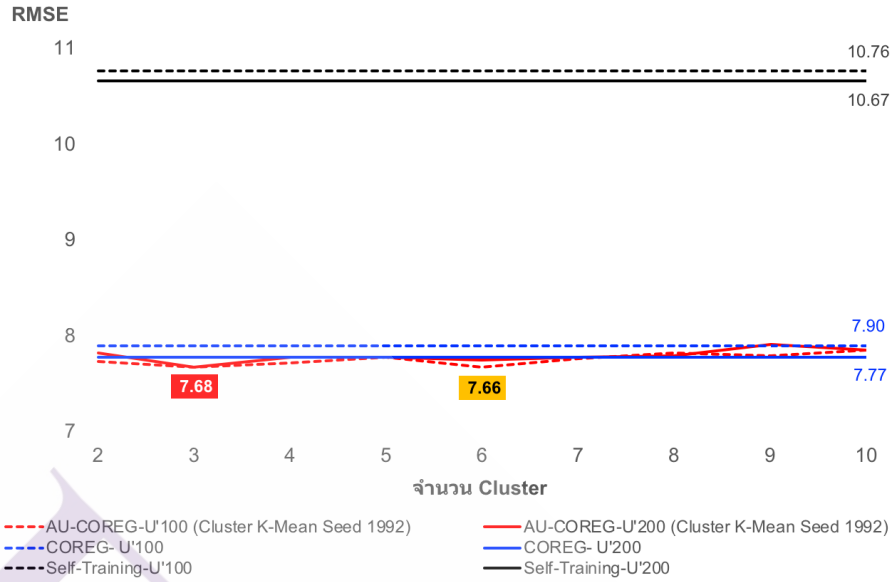
Model	U'100		U'200	
	RMSE	Time (min)	RMSE	Time (min)
COREG	7.90	217	7.77	419
AU-COREG โดยการจัดกลุ่มด้วยวิธี K-Medoids	7.63 (2 Cluster)	14	7.65 (7 Cluster)	28
ลดลง	1.80%	94%	3.21%	93%

4.3.3 การสร้างโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Mean Seed 1992 ของชุดข้อมูลพลังงานความร้อนร่วม

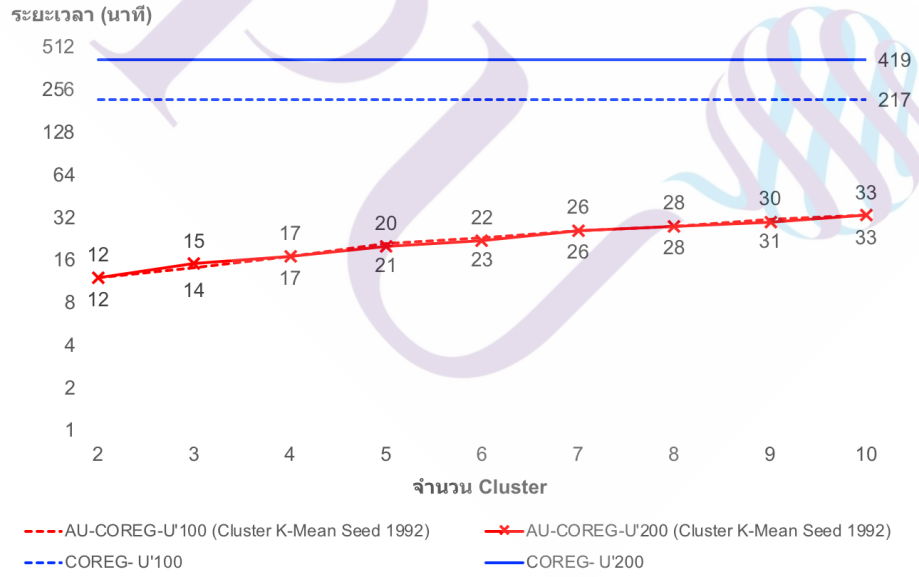
จากภาพที่ 4.23 และ 4.24 แสดงให้เห็นว่า

4.3.3.1 โมเดล AU-COREG (U'= 100) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 1992) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 7.66 ซึ่งได้จากการทำ Cluster จำนวน 6 Cluster ลดลง 1.40% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 217 นาที เหลือเพียง 23 นาที หรือคิดเป็น 89% ดังตารางที่ 4.10

4.3.3.2 โมเดล AU-COREG (U'= 200) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 1992) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 7.68 ซึ่งได้จากการทำ Cluster จำนวน 3 Cluster ลดลง 2.82% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 419 นาที เหลือเพียง 15 นาที หรือคิดเป็น 96% ดังตารางที่ 4.10



ภาพที่ 4.23 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์พลังไฟฟ้าต่อชั่วโมง ด้วยโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Mean (Seed 1992) กับวิธี Self-Training และ COREG ของชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม



ภาพที่ 4.24 ภาพแสดงการเปรียบเทียบระยะเวลา (นาที) การสร้างโมเดล AU-COREG ด้วยโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Mean (Seed 1992) กับวิธี Self-Training และ COREG ของชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม

ตารางที่ 4.10 ผลการเปรียบเทียบประสิทธิภาพโมเดล AU-COREG กับโมเดล COREG ด้วยค่า RMSE ที่มีค่าน้อยที่สุดและระยะเวลา (นาที) โดยการจัดกลุ่มด้วยวิธี K-Mean Seed 1992 ที่มีค่าน้อยที่สุดของชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม

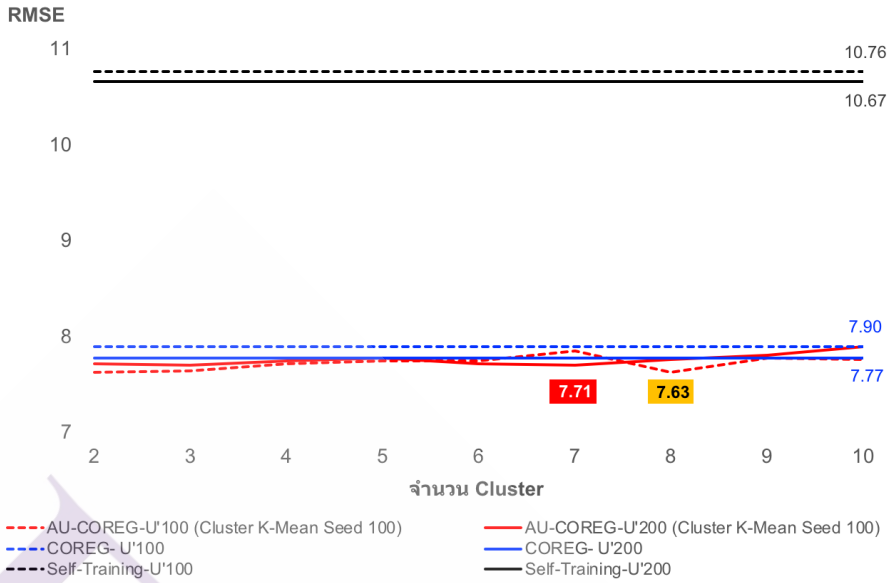
Model	U'100		U'200	
	RMSE	Time (min)	RMSE	Time (min)
COREG	7.90	217	7.77	419
AU-COREG โดยการจัดกลุ่มด้วยวิธี K-Mean Seed 1992	7.66 (6 Cluster)	23	7.68 (3 Cluster)	15
ลดลง	1.40%	89%	2.82%	96%

4.3.4 การสร้างโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Mean Seed 100 ของชุดข้อมูลพลังงานความร้อนร่วม

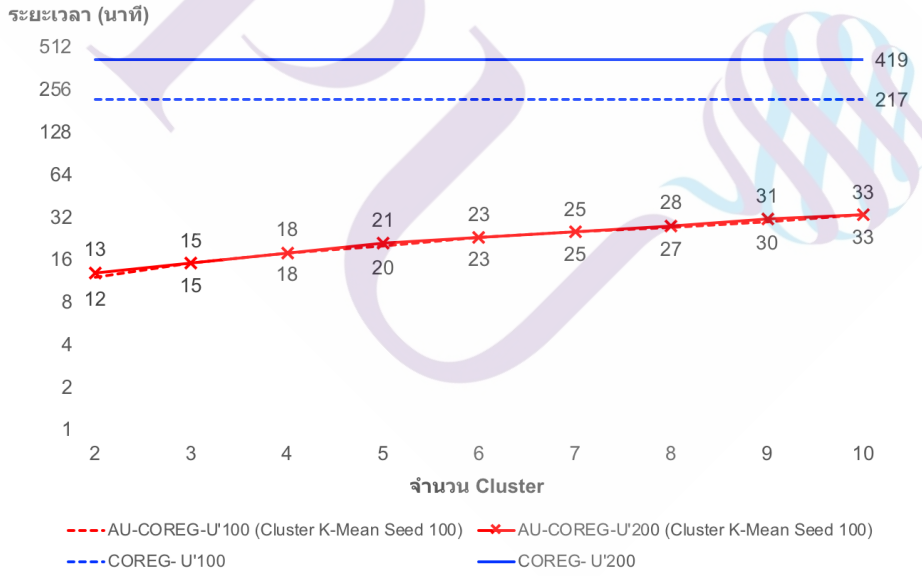
จากภาพที่ 4.25 และ 4.26 แสดงให้เห็นว่า

4.3.4.1 โมเดล AU-COREG (U'= 100) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 100) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 7.63 ซึ่งได้จากการทำ Cluster จำนวน 8 Cluster ลดลง 1.85% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 217 นาที เหลือเพียง 27 นาที หรือคิดเป็น 88% ดังตารางที่ 4.11

4.3.4.2 โมเดล AU-COREG (U'= 200) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 100) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 7.71 ซึ่งได้จากการทำ Cluster จำนวน 7 Cluster ลดลง 2.55% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 419 นาที เหลือเพียง 25 นาที หรือคิดเป็น 95% ดังตารางที่ 4.11



ภาพที่ 4.25 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์พลังไฟฟ้าต่อชั่วโมง ด้วยโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Mean (Seed 100) กับวิธี Self-Training และ COREG ของชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม



ภาพที่ 4.26 ภาพแสดงการเปรียบเทียบระยะเวลา (นาที) การสร้างโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Mean (Seed 100) กับวิธี COREG ของชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม

ตารางที่ 4.11 ผลการเปรียบเทียบประสิทธิภาพโมเดล AU-COREG กับโมเดล COREG ด้วยค่า RMSE ที่มีค่าน้อยที่สุดและระยะเวลา (นาที) โดยการจัดกลุ่มด้วยวิธี K-Mean Seed 100 ที่มีค่าน้อยที่สุดของชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม

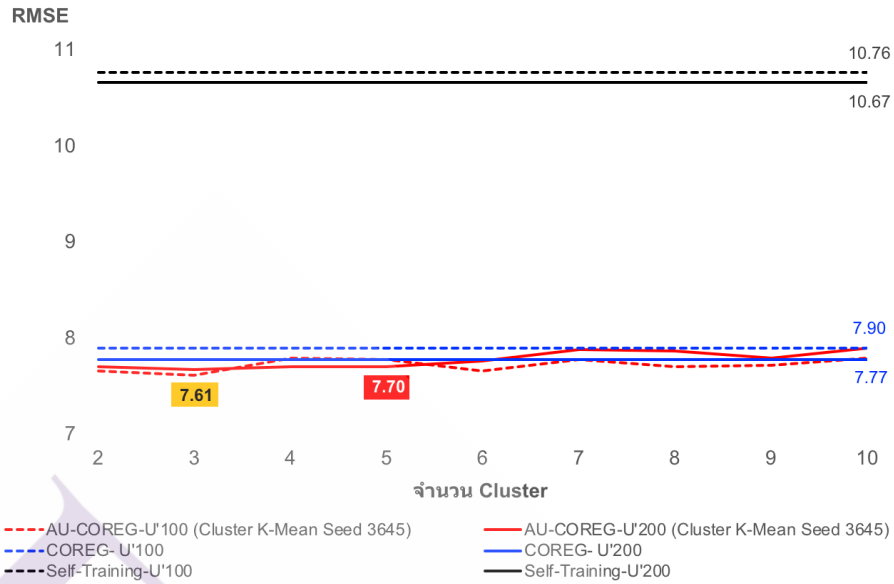
Model	U'100		U'200	
	RMSE	Time (min)	RMSE	Time (min)
COREG	7.90	217	7.77	419
AU-COREG โดยการจัดกลุ่มด้วยวิธี K-Mean Seed 100	7.63 (8 Cluster)	27	7.71 (7 Cluster)	25
ลดลง	1.85%	88%	2.55%	95%

4.3.5 การสร้างโมเดล AU-COREG ด้วยวิธีการจัดกลุ่มแบบ K-Mean Seed 3645 ของชุดข้อมูลพลังงานความร้อนร่วม

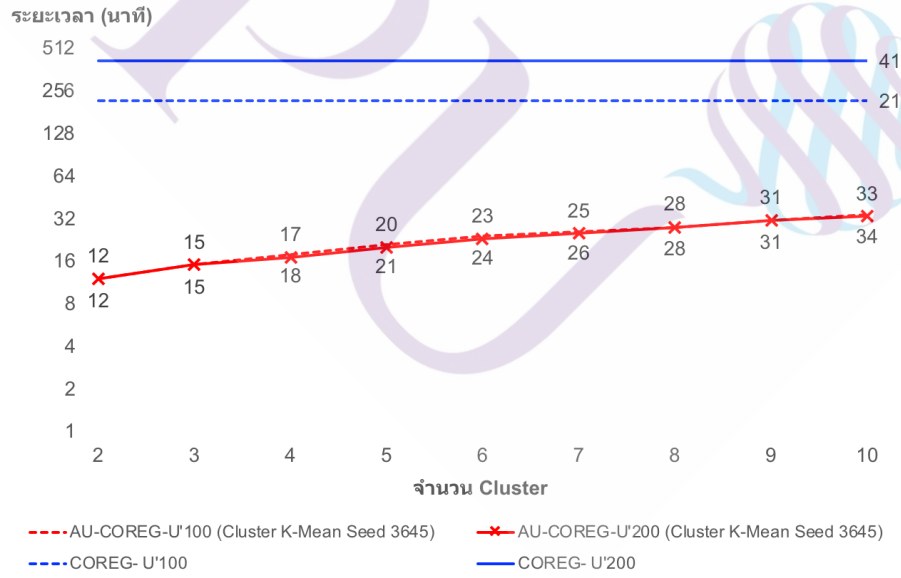
จากภาพที่ 4.27 และ 4.28 แสดงให้เห็นว่า

4.3.5.1 โมเดล AU-COREG (U'= 100) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 3645) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 7.61 ซึ่งได้จากการทำ Cluster จำนวน 3 Cluster ลดลง 2.07% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 217 นาที เหลือเพียง 15 นาที หรือคิดเป็น 93% ดังตารางที่ 4.12

4.3.5.2 โมเดล AU-COREG (U'= 200) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean Seed 3645) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 7.70 ซึ่งได้จากการทำ Cluster จำนวน 5 Cluster ลดลง 2.88% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 419 นาที เหลือเพียง 20 นาที หรือคิดเป็น 95% ดังตารางที่ 4.12



ภาพที่ 4.27 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์พลังไฟฟ้าต่อชั่วโมง ด้วยโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Mean (Seed 3645) กับวิธี Self-Training และ COREG ของชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม

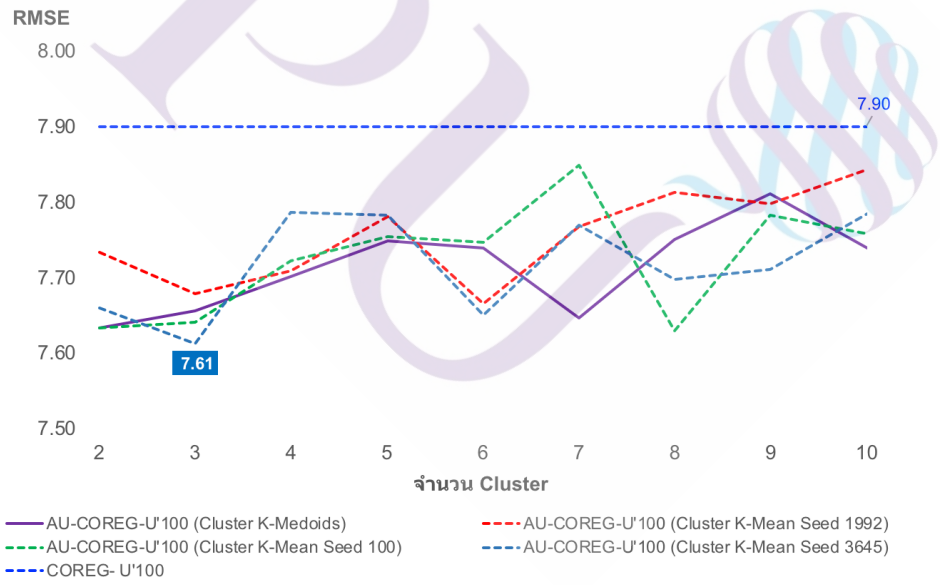


ภาพที่ 4.28 ภาพแสดงการเปรียบเทียบระยะเวลา (นาที) การสร้างโมเดล AU-COREG โดยวิธีการจัดกลุ่มแบบ K-Mean (Seed 3645) กับวิธี COREG ของชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม

ตารางที่ 4.12 ผลการเปรียบเทียบประสิทธิภาพโมเดล AU-COREG กับโมเดล COREG ด้วยค่า RMSE ที่มีค่าน้อยที่สุดและระยะเวลา (นาที)โดยการจัดกลุ่มด้วยวิธี K-Mean Seed 3645 ที่มีค่าน้อยที่สุดของชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม

Model	U'100		U'200	
	RMSE	Time (min)	RMSE	Time (min)
COREG	7.90	217	7.77	419
AU-COREG โดยการจัดกลุ่มด้วยวิธี K-Mean Seed 3645	7.61 (3 Cluster)	15	7.70 (5 Cluster)	20
ลดลง	2.07%	93%	2.88%	95%

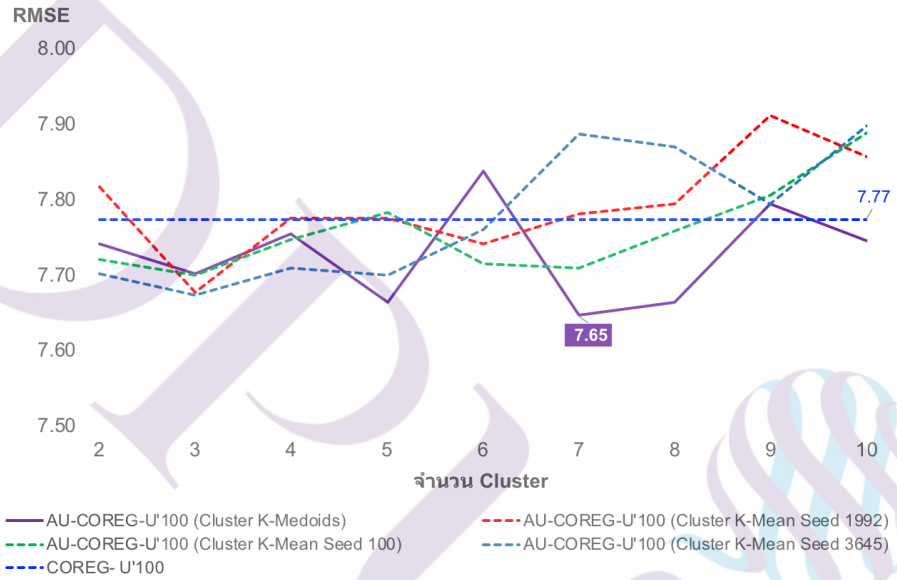
4.3.6 การเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG (U'=100) ด้วยวิธีการจัดกลุ่มแบบ K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645 ตามลำดับ กับโมเดล COREG ของชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม



ภาพที่ 4.29 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์จำนวนคนที่ใช้บริการต่อชั่วโมง ด้วยโมเดล AU-COREG (U'=100) โดยวิธีการจัดกลุ่มแบบ K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645 ตามลำดับ กับโมเดล COREG ของชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม

จากภาพที่ 4.29 แสดงให้เห็นว่าโมเดล AU-COREG (U'= 100) ในทุก ๆ วิธีการจัดกลุ่ม (K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645) และจำนวนกลุ่มตั้งแต่ 2 ถึง 10 กลุ่มให้ค่า RMSE น้อยกว่าหรือเท่ากับค่า RMSE ที่ได้จากโมเดล COREG โดยวิธีการจัดกลุ่มที่ให้ค่า RMSE ที่น้อยที่สุด ได้แก่ วิธีการจัดกลุ่มโดยวิธี K-Mean ที่กำหนด Seed เท่ากับ 3645 จำนวน 3 Cluster

4.3.7 การเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG (U'=200) ด้วยวิธีการจัดกลุ่มแบบ K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645 ตามลำดับ กับโมเดล COREG ของชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม



ภาพที่ 4.30 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์จำนวนคนที่ใช้บริการต่อชั่วโมง ด้วยโมเดล AU-COREG (U'=200) โดยวิธีการจัดกลุ่มแบบ K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645 ตามลำดับ กับโมเดล COREG ของชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม

จากภาพที่ 4.30 แสดงให้เห็นว่าโมเดล AU-COREG (U'= 200) ในทุก ๆ วิธีการจัดกลุ่ม (K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645) ให้ค่า RMSE น้อยกว่าหรือเท่ากับค่า RMSE ที่ได้จากโมเดล COREG ดังนี้

4.3.7.1 การจัดกลุ่ม 5, 6, 7, 8 และ 10 Cluster ทำให้ค่า RMSE ของโมเดล AU-COREG น้อยกว่าค่า RMSE ที่ได้จากโมเดล COREG

4.3.7.2 โมเดล AU-COREG โดยวิธีการจัดกลุ่ม 4 วิธี (K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645) จำนวนกลุ่มที่ให้ค่า RMSE น้อยกว่าหรือเท่ากับ RMSE จากโมเดล COREG ที่ดีที่สุดคือ จำนวน 3 Cluster

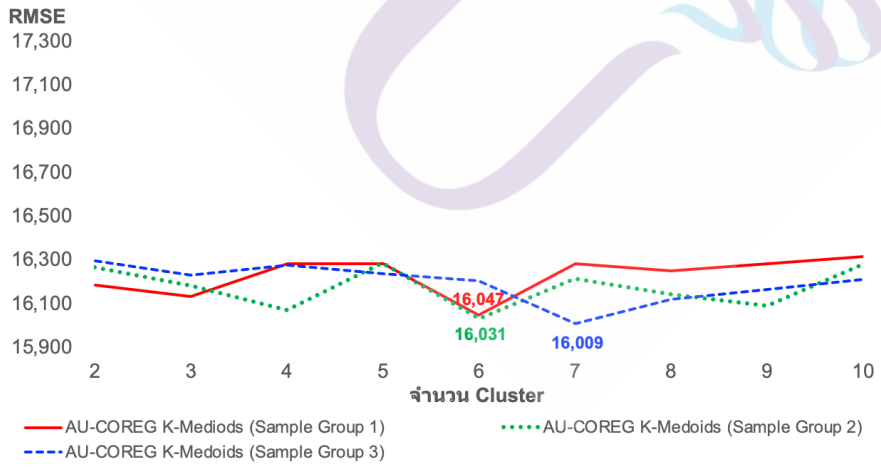
4.3.7.3 โมเดล AU-COREG โดยวิธีการจัดกลุ่มที่ให้ค่า RMSE น้อยกว่าหรือเท่ากับ RMSE ที่ได้จากโมเดล COREG ที่ดีที่สุด คือ วิธี K-Medoids

4.3.7.4 โมเดล AU-COREG โดยวิธีการจัดกลุ่มที่ให้ค่า RMSE น้อยที่สุด ได้แก่ วิธีการจัดกลุ่มโดยวิธี K-Medoids จำนวน 7 Cluster

4.4 ผลการเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG ที่ใช้กลุ่มตัวอย่างจากการสุ่มที่แตกต่างกัน 3 กลุ่ม ด้วยวิธีการจัดกลุ่มแบบ K-Medoids และได้เพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ100 ตัวอย่าง (U'=100) กับข้อมูลทั้ง 3 ชุดข้อมูล

จากผลการทดลองสามารถสรุปผลการเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG ที่ใช้กลุ่มตัวอย่างจากการสุ่มที่แตกต่างกัน 3 กลุ่ม ด้วยวิธีการจัดกลุ่มแบบ K-Medoids กับข้อมูลทั้ง 3 ชุดข้อมูล ได้ดังนี้

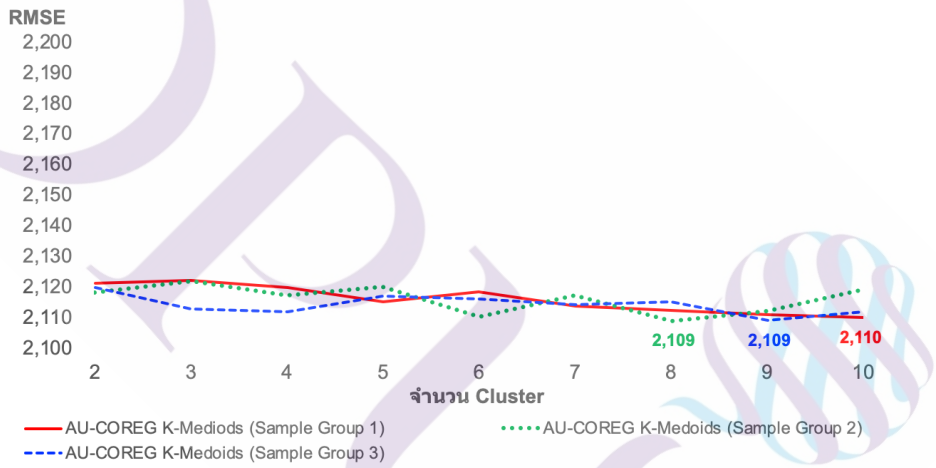
4.4.1 ผลการเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG ที่ใช้กลุ่มตัวอย่างจากการสุ่มที่แตกต่างกัน 3 กลุ่ม ด้วยวิธีการจัดกลุ่มแบบ K-Medoids และได้เพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ100 ตัวอย่าง (U'=100) กับข้อมูลชุดที่ 1 ข้อมูลการพยากรณ์เงินเดือน ดังภาพที่ 4.31



ภาพที่ 4.31 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากการพยากรณ์เงินเดือน (Dollar) ด้วยโมเดล AU-COREG (U'=100) ที่ใช้กลุ่มตัวอย่างจากการสุ่มที่แตกต่างกัน 3 กลุ่ม ด้วยวิธีการจัดกลุ่มแบบ K-Medoids

จากภาพที่ 4.31 แสดงให้เห็นว่าโมเดล AU-COREG (U'= 100) ที่ใช้กลุ่มตัวอย่างจากการสุ่มที่แตกต่างกัน 3 กลุ่ม ด้วยวิธีการจัดกลุ่มแบบ K-Medoids โดยจากกลุ่มตัวอย่างในกลุ่มที่ 1 และ 2 จำนวน Cluster ที่ให้ค่า RMSE น้อยที่สุดคือ จำนวน 6 Cluster คือ 16,047 และ 16,031 ตามลำดับ และกลุ่มตัวอย่างกลุ่มที่ 3 จำนวน Cluster ที่ให้ค่า RMSE น้อยที่สุด คือจำนวน 7 Cluster มีค่าเท่ากับ 16,009

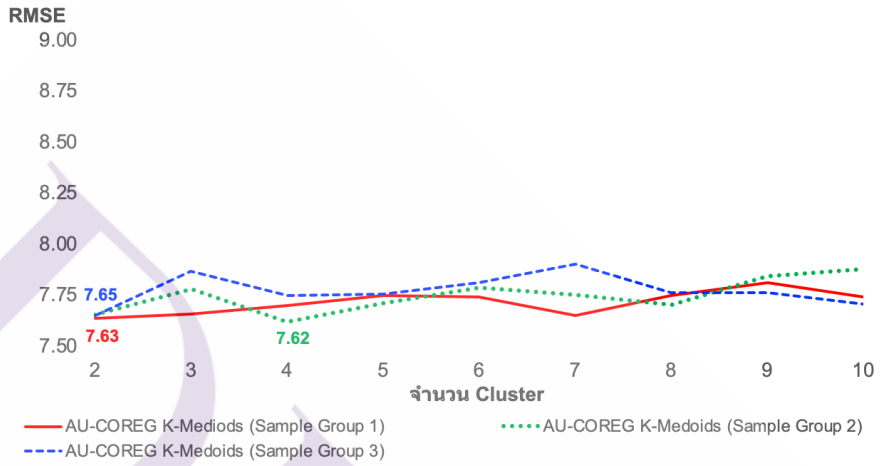
4.4.2 ผลการเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG ที่ใช้กลุ่มตัวอย่างจากการสุ่มที่แตกต่างกัน 3 กลุ่ม ด้วยวิธีการจัดกลุ่มแบบ K-Medoids และได้เพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ100 ตัวอย่าง (U'=100) กับข้อมูลชุดที่ 2 ชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน ดังภาพที่ 4.32



ภาพที่ 4.32 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดิน ด้วยโมเดล AU-COREG (U'=100) ที่ใช้กลุ่มตัวอย่างจากการสุ่มที่แตกต่างกัน 3 กลุ่ม ด้วยวิธีการจัดกลุ่มแบบ K-Medoids

จากภาพที่ 4.32 แสดงให้เห็นว่าโมเดล AU-COREG (U'= 100) ที่ใช้กลุ่มตัวอย่างจากการสุ่มที่แตกต่างกัน 3 กลุ่ม ด้วยวิธีการจัดกลุ่มแบบ K-Medoids โดยจากกลุ่มตัวอย่างกลุ่มที่ 1 ให้ค่า RMSE น้อยที่สุดคือ จำนวน 10 Cluster มีค่าเท่ากับ 2,110 และจากกลุ่มตัวอย่างกลุ่มที่ 2 และ 3ให้ค่า RMSE น้อยที่สุดคือ จำนวน 8 Cluster มีค่าเท่ากับ 2,109 และ 9 Cluster มีค่าเท่ากับ 2,109 ตามลำดับ

4.4.3 ผลการเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG ที่ใช้กลุ่มตัวอย่างจากการสุ่มที่แตกต่างกัน 3 กลุ่ม ด้วยวิธีการจัดกลุ่มแบบ K-Medoids และได้เพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ100 ตัวอย่าง (U'=100) กับข้อมูลชุดที่ 3 ชุดชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม ดังภาพที่ 4.33



ภาพที่ 4.33 ภาพแสดงการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE จากชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม ด้วยโมเดล AU-COREG (U'=100) ที่ใช้กลุ่มตัวอย่างจากการสุ่มที่แตกต่างกัน 3 กลุ่ม ด้วยวิธีการจัดกลุ่มแบบ K-Medoids

จากภาพที่ 4.33 แสดงให้เห็นว่าโมเดล AU-COREG (U'= 100) ที่ใช้กลุ่มตัวอย่างจากการสุ่มที่แตกต่างกัน 3 กลุ่ม ด้วยวิธีการจัดกลุ่มแบบ K-Medoids โดยจากกลุ่มตัวอย่างกลุ่มที่ 1 และกลุ่มที่ 3 ให้ค่า RMSE น้อยที่สุดคือ จำนวน 2 Cluster มีค่าเท่ากับ 7.63 และ 7.65 และจากกลุ่มตัวอย่างกลุ่มที่ 2 ให้ค่า RMSE น้อยที่สุดคือ จำนวน 4 Cluster มีค่าเท่ากับ 7.62

บทที่ 5
บทสรุปและข้อเสนอแนะ

งานวิจัยในวิทยานิพนธ์เล่มนี้ได้กล่าวถึงการวิธีการเลือกข้อมูลที่ไม่มีป้ายกำกับอย่างอัตโนมัติเพื่อการสร้างโมเดลการเรียนรู้ร่วมแบบกึ่งมีผู้สอน เพื่อปรับปรุงประสิทธิภาพโมเดลพยากรณ์ จากการเรียนรู้ร่วมกึ่งมีผู้สอน สำหรับข้อมูลที่มีป้ายกำกับที่มีสัดส่วนน้อย เมื่อเทียบกับข้อมูลที่ไม่มีป้ายกำกับและลดระยะเวลาในการสร้างโมเดลพยากรณ์ กับชุดข้อมูลที่มีคุณสมบัติที่แตกต่างกันทั้งหมด 3 ชุดข้อมูล ได้แก่

1) ข้อมูลที่ 1 ชุดข้อมูลโฆษณาประกาศรับสมัครงาน เป็นข้อมูลประกาศสมัครงานในประเทศอังกฤษ จัดทำโดย Adzuna เป็นข้อมูลที่ใช้ในการแข่งขันในหัวข้อการพยากรณ์เงินเดือนของ Kaggle เพื่อทำการพยากรณ์เงินเดือน

2) ชุดข้อมูลที่ 2 ชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐ โดยรถไฟฟ้าใต้ดินเป็นข้อมูลปริมาณการจราจรระหว่างรัฐโดยรถไฟฟ้ารายชั่วโมง จากทิศตะวันตกของมินนิโซตา DoT ATR สถานี 301 ซึ่งอยู่กลางระหว่างมินนิอาโปลิสและเซนต์พอลมินนิโซตา เพื่อทำการพยากรณ์ปริมาณผู้ใช้บริการรถไฟฟ้ารายชั่วโมง

3) ชุดข้อมูลที่ 3 ชุดข้อมูลการพยากรณ์พลังงานความร้อนร่วม เป็นข้อมูลพลังงานความร้อนร่วม โดยข้อมูลที่รวบรวมจากโรงไฟฟ้าพลังความร้อนร่วมในช่วง 6 ปี (2549-2554) เพื่อทำการพยากรณ์พลังงานความร้อนต่อชั่วโมง

ข้อมูลทั้ง 3 ชุดข้อมูลได้ผ่านกระบวนการประมวลผลข้อมูลเบื้องต้น และทำการแบ่งข้อมูลเพื่อใช้ในการสร้างโมเดลและทดสอบประสิทธิภาพของโมเดล จากนั้นกำหนดพารามิเตอร์ต่างๆ และทำการสร้างโมเดล AU-COREG ซึ่งอัลกอริทึมในการสร้างโมเดลได้ใช้เทคนิคการเรียนรู้ของเครื่อง 4 วิธี ได้แก่ วิธีการเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbors: kNN) วิธีการจัดกลุ่ม (Clustering) การวิเคราะห์การถดถอย (Regression Analysis) และวิธีการโค เทรนนิง (Co-Training) และการทดสอบประสิทธิภาพของโมเดลใช้สัมประสิทธิ์ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Root Mean Square Error: RMSE) จากนั้นนำผลลัพธ์ที่ได้ (RMSE และเวลาในการสร้างโมเดล AU-COREG) มาเปรียบเทียบกับโมเดล COREG

5.1 สรุปผลการศึกษา

ในการเปรียบเทียบประสิทธิภาพของโมเดล AU-COREG ด้วยเทคนิคการจัดกลุ่มแบบ K-Medoids และ K-Mean ที่กำหนด Seed 1992, 100 และ 3645 ตามลำดับ กับโมเดล COREG สรุปผลการศึกษาได้ดังนี้

5.1.1 ชุดข้อมูลที่ 1 ข้อมูลโฆษณาประกาศรับสมัครงาน ซึ่งข้อมูลชุดนี้มีแอททริบิวต์ทั้งหมดเป็นประเภทนามบัญญัติ (Nominal) โดยโมเดล AU-COREG ที่ U'=100 ที่ให้ค่า RMSE มีน้อยที่สุดมาจากวิธีการจัดกลุ่มด้วยเทคนิค K-Medoids ทั้งหมดจำนวน 6 Cluster มีค่าเท่ากับ 16,047.12 ลดลงจากโมเดล COREG 16,126.82 คิดเป็น 0.49% และระยะเวลาลดลงจาก 276 นาที เป็น 31 นาที คิดเป็น 89% และโมเดล AU-COREG ที่ U'=200 ที่ให้ค่า RMSE มีน้อยที่สุดมาจากวิธีการจัดกลุ่มด้วยเทคนิค K-Medoids ทั้งหมดจำนวน 2 Cluster มีค่าเท่ากับ 16,003.33 ลดลงจากโมเดล COREG 16,104.93 คิดเป็น 0.14% และระยะเวลาลดลงจาก 527 นาที เป็น 23 นาที คิดเป็น 96%

5.1.2 ชุดข้อมูลที่ 2 ชุดข้อมูลการพยากรณ์ปริมาณการจราจรระหว่างรัฐโดยรถไฟฟ้าใต้ดิน ซึ่งข้อมูลชุดนี้มีแอททริบิวต์มีทั้งประเภทนามบัญญัติ (Nominal) และตัวเลข (Numerical) โดยโมเดล AU-COREG ที่ U'=100 ที่ให้ค่า RMSE มีน้อยที่สุดมาจากวิธีการจัดกลุ่มด้วยเทคนิค K-Mean ที่กำหนด Seed เท่ากับ 3645 ทั้งหมดจำนวน 10 Cluster มีค่าเท่ากับ 2,107.95 ลดลงจากโมเดล COREG 2,113.46 คิดเป็น 0.26% และระยะเวลาลดลงจาก 215 นาที เป็น 34 นาที คิดเป็น 84% และโมเดล AU-COREG ที่ U'=200 ที่ให้ค่า RMSE มีน้อยที่สุดมาจากวิธีการจัดกลุ่มด้วยเทคนิค K-Mean ที่กำหนด Seed เท่ากับ 1992 ทั้งหมดจำนวน 10 Cluster มีค่าเท่ากับ 2,107.28 ลดลงจากโมเดล COREG 2,110.38 คิดเป็น 0.10% และระยะเวลาลดลงจาก 428 นาที เป็น 33 นาที คิดเป็น 92%

5.1.3 ชุดข้อมูลที่ 3 ชุดข้อมูลพลังงานความร้อนร่วม ซึ่งข้อมูลชุดนี้มีแอททริบิวต์ทั้งหมดเป็นประเภทตัวเลข (Numerical) โดยโมเดล AU-COREG ที่ U'=100 ที่ให้ค่า RMSE มีน้อยที่สุดมาจากวิธีการจัดกลุ่มด้วยเทคนิค K-Mean ที่กำหนด Seed เท่ากับ 3645 ทั้งหมดจำนวน 3 Cluster มีค่าเท่ากับ 7.61 ลดลงจากโมเดล COREG 7.9 คิดเป็น 2.07% และระยะเวลาลดลงจาก 217 นาที เป็น 15 นาที คิดเป็น 93% และโมเดล AU-COREG ที่ U'=200 ที่ให้ค่า RMSE มีน้อยที่สุดมาจากวิธีการจัดกลุ่มด้วยเทคนิค K-Medoids จำนวน 7 Cluster มีค่าเท่ากับ 7.65 ลดลงจากโมเดล COREG 7.77 คิดเป็น 3.21% และระยะเวลาลดลงจาก 419 นาที เป็น 28 นาที คิดเป็น 93%

5.2 อภิปรายผลการศึกษา

การสร้างโมเดลจากการเลือกข้อมูลที่ไม่มีป้ายกำกับอย่างอัตโนมัติแบบกึ่งถดถอย (AU-COREG) ถูกพัฒนามาจากโมเดล COREG คือลดขนาดตัวอย่างที่ไม่มีป้ายกำกับเพื่อคัดเลือกเข้าโมเดล ด้วยการทำ Cluster เพื่อช่วยลดจำนวนรอบในการคำนวณได้อย่างมีนัยสำคัญ และยังทำให้ประสิทธิภาพของโมเดลไม่ลดลงเมื่อเทียบกับโมเดล COREG ซึ่ง

จำนวนรอบในการคำนวณ วิธี COREG

$$N = T * i * n * nU' \quad (5.1)$$

จำนวนรอบในการคำนวณ วิธี AU-COREG

$$N = T * i * n * nC \quad)5.2($$

โดยที่ N คือจำนวนรอบในการคำนวณ

T คือจำนวนรอบในการทำซ้ำ

i คือจำนวนโมเดล

n คือจำนวนเพื่อนบ้าน

nU' คือจำนวนตัวอย่างที่ไม่มีป้ายกำกับ

nC คือจำนวนคลัสเตอร์

หากเพิ่มจำนวน Cluster ให้มากขึ้น จะเพิ่มจำนวนรอบในการคำนวณมากขึ้นอย่างไม่มีนัยสำคัญ ดังนั้นผู้วิจัยจึงได้ทำการทดลองเพิ่มจำนวน Cluster ตั้งแต่ 2 จนถึง 10 Cluster และเพิ่ม U' ให้มากขึ้นจาก 100 เป็น 200 ตัวอย่าง เพื่อให้สามารถทำการจัดกลุ่มให้ดียิ่งขึ้น

โมเดล AU-COREG ต้องการเลือกตัวแทนของแต่ละ Cluster ที่ดีที่สุด (ลดค่าคาดเคลื่อนจากการพยากรณ์ให้มากที่สุด) เพื่อนำเข้าชุดข้อมูลเพื่อฝึกการเรียนรู้ให้กับโมเดล ดังนั้นผู้วิจัยจึงได้ใช้เทคนิคการทำ Cluster 2 วิธี ได้แก่ K-Medoids และ K-Mean โดยวิธี K-Medoids ผู้วิจัยเลือกสมาชิกที่อยู่จุดกึ่งกลางของข้อมูลใน Cluster (Centroid) เป็นตัวแทนของ Cluster และวิธี K-Mean ผู้วิจัยทำการระบุค่าสุ่ม ในการเลือกสมาชิก 3 ค่า เพื่อเปรียบเทียบประสิทธิภาพของโมเดล จากการเลือกตัวแทนของ Cluster ที่แตกต่างกัน

จากผลการทดลอง พบว่าโมเดล AU-COREG เมื่อทำการเพิ่มจำนวนข้อมูลที่ไม่มีป้ายกำกับเพิ่มจาก 100 เป็น 200 โดยส่วนใหญ่ทำให้ค่า RMSE ลดลง

การเพิ่มจำนวนของ Cluster แต่ไม่มากจนเกินไป (ถ้าจำนวน Cluster มากจนเกินไป จะทำให้บาง Cluster มีสมาชิกจำนวนน้อย) จะช่วยทำให้ค่า RMSE ลดลงได้ โดยใช้ระยะเวลาไม่แตกต่างไปจากเดิม

การเลือกตัวแทนของ Cluster โดยใช้เทคนิคการจัด Cluster ที่แตกต่างกัน (K-Medoids และ K-Mean) มีผลต่อค่า RMSE ซึ่งพบกว่าส่วนใหญ่การจัด Cluster ด้วยวิธี K-Medoids กับข้อมูลประเภท Nominal จะให้ค่า RMSE ที่ต่ำกว่า

นอกจากนั้นผู้วิจัยยังทำการเปรียบเทียบประสิทธิภาพจากโมเดล AU-COREG ที่ใช้กลุ่มตัวอย่างจากการสุ่มที่แตกต่างกัน 3 กลุ่ม ด้วยวิธีการจัดกลุ่มแบบ K-Medoids และได้เพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ100 ตัวอย่าง (U'=100) กับข้อมูลทั้ง 3 ชุดข้อมูล เพื่อดูความแตกต่างของประสิทธิภาพของโมเดลที่เกิดจากการสุ่มตัวอย่างที่แตกต่างกัน ซึ่งจากผลการทดสอบพิสูจน์ได้ว่า การสุ่มตัวอย่างที่แตกต่างกันไม่ได้ทำให้ประสิทธิภาพของโมเดล AU-COREG แตกต่างกันอย่างมีนัยสำคัญ

5.3 ข้อเสนอแนะ

ผู้วิจัยยังขาดการปรับพารามิเตอร์ให้เหมาะสมกับคุณลักษณะของข้อมูล เช่น จำนวนข้อมูลที่ไม่มีป้ายกำกับที่จะเพิ่มเข้าไปในข้อมูลการเรียนรู้ จำนวนเพื่อนบ้านที่ใช้ทดสอบความคลาดเคลื่อน จำนวนรอบการทำซ้ำ เป็นต้น ซึ่งอาจช่วยปรับปรุงโมเดลให้มีประสิทธิภาพมากยิ่งขึ้น

บรรณานุกรม

บรรณานุกรม

ภาษาต่างประเทศ

Alice Zheng. (2015). Evaluating Machine Learning Models, A Beginner's Guide to Key Concepts and Pitfalls, 19, Retrieved June 16, 2019 from http://www.pindex.com/uploads/post_docs/evaluating-machine-learning-models(PINDEX-DOC-6950).pdf

Bizibl Marketing. 10 Key Marketing Trends for 2017 and Ideas for Exceeding Customer Expectations, Retrieved June 16, 2019 from https://bizibl.com/marketing/download/10-key-marketing-trends-2017-and-ideas-exceeding-customer-expectations

Didaci L., Fumera, G., Roli, F. (2012). Analysis of co-training algorithm with very small training sets. Gimel'farb, G., et al. (eds.) SSPR/SPR 2012. LNCS, vol. 7626. Springer, Heidelberg.

Kulwinder Kaur. Machine Learning Techniques, Data Mining, Weka, Retrieved June 16, 2019 from http://www.e2matrix.com/blog/2018/01/24/machine-learning-techniques

F Ma, D Meng., Q Xie., Z Li., X Don. (2017). Self-paced co-training, Proceedings of the 34th International Conference on Machine Learning-Volume 70, 2017.

Matthew Mayo, KDnuggets. A concise explanation of learning algorithms with the Mitchell paradigm, Retrieved June 16, 2019 from https://www.kdnuggets.com/2018/10/mitchell-paradigm-concise-explanation-learning-algorithms.html

Rohith Gandh. (2018). Introduction to Machine Learning Algorithms: Linear Regression, Retrieved June 16, 2019 from https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a

R Wang, L Li. (2016). The performance improvement algorithm of co-training by committee 5th international conference on computer science and network technology.

SAS. Understanding data mining clustering methods. Retrieved June 16, 2019 from

https://blogs.sas.com/content/subconsciousmusings/2016/05/26/data-mining-clustering/

Semi-supervised Learning, Retrieved June 16, 2019 from https://www.slideshare.net/butest/semisupervised-learning

Sousa R., Gama J. (2017). Co-training Semi-supervised Learning for Single-Target Regression in Data Streams Using AMRules. International Symposium on Methodologies for Intelligent Systems.

What Is Machine Learning?, 3 things you need to know, Retrieved June 16, 2019 from https://www.mathworks.com/discovery/machine-learning.html

Wikipedia, k-nearest neighbors algorithm, Retrieved June 16, 2019 from https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

Zhi Hua., Ming Li. (2005). Semi-supervised regression with co-training IJCAI'05 proceeding of the 19th international joint conference on artificial intelligence.

ภาคผนวก

วิธีการเลือกข้อมูลที่ไม่มีป้ายกำกับอย่างอัตโนมัติเพื่อการสร้างโมเดลการเรียนรู้ร่วมแบบกึ่งมีผู้สอน

An Automatic Unlabeled Selection for CO-training REGressors (AU-COREG)

ศิริขวัญ คีรีสุวรรณกุล
Sirikwan Kheereesuwannakul
E-mail: 585162020005@dpu.ac.th
สาขาวิศกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
Big Data Engineering, College of Innovative Technology and Engineering, Dhurakit Pundit University
Bangkok, Thailand

เอกสิทธิ์ พัชรวงศ์ศักดา
Eakasit Pacharawongsakda
E-mail: eakasit.pac@dpu.ac.th
สาขาวิศกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
Big Data Engineering, College of Innovative Technology and Engineering, Dhurakit Pundit University
Bangkok, Thailand

ABSTRACT

This research aims to improve the performance of semi-supervised learning by automatically select unlabeled data. The proposed method uses two regression models to estimate values for unlabeled data, then cluster the data into groups. Therefore, similar data are assigned in the same group and the different data are assigned into the different groups. After that, the method selects each group representative that have least error and append into training data. Then, we repeat until we have enough training data. From experimental results with three datasets, we found that the proposed method can improve performance and reduce computation time by 84%, comparing to previous work.

KEY WORDS Co-Training, Simi-supervised Learning

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อปรับปรุงประสิทธิภาพโมเดลพยากรณ์ ด้วยวิธีการเลือกข้อมูลที่ไม่มีป้ายกำกับอย่างอัตโนมัติ เพื่อสร้างโมเดลการเรียนรู้ร่วมแบบกึ่งมีผู้สอน (semi-supervised Learning) ซึ่งเหมาะสำหรับข้อมูลที่มีป้ายกำกับ (labeled data) ปริมาณน้อยมาก โดยวิธีการที่นำเสนอนี้จะใช้ประโยชน์จากข้อมูลที่ไม่มีป้ายกำกับ (unlabeled data) ที่มีอยู่ปริมาณมากมาช่วยเพิ่มประสิทธิภาพของการสร้างโมเดลจำแนกประเภทข้อมูล (classification) หรือการประมาณค่า (regression) วิธีการที่นำเสนอนี้เริ่มด้วยการใช้โมเดล 2 โมเดลทำการกำกับค่าให้กับข้อมูลที่ไม่มีป้ายกำกับ จากนั้นนำข้อมูลเหล่านี้มาทำการจัดกลุ่ม (clustering) ให้ข้อมูลที่มีความคล้ายคลึงกันอยู่ในกลุ่มเดียวกัน และแยกข้อมูลที่ต่างกันออกให้อยู่ต่างกลุ่มกัน ถัดมาจึงเลือกตัวแทนแต่ละกลุ่มเพื่อหาข้อมูลที่ทำให้โมเดลการพยากรณ์มีความคลาดเคลื่อนน้อยที่สุดเข้าไปเป็นชุดข้อมูลสอน (training data) ในรอบถัดไปและสร้างโมเดลพยากรณ์ใหม่ ทำซ้ำจนเพิ่มข้อมูลเข้าไปในชุดสอนได้ครบ จากนั้นการพยากรณ์ขั้นสุดท้ายทำได้โดยการหาค่าเฉลี่ยของการพยากรณ์จากทั้งสองโมเดลที่สร้างขึ้น จากการทดสอบด้วยข้อมูลจำนวน 3 ชุดแสดงให้เห็นว่าวิธีการที่นำเสนอ (AU-COREG) สามารถปรับปรุงประสิทธิภาพของโมเดลได้อย่างมีนัยสำคัญและช่วยลดเวลาลงไปมากกว่า 84% เมื่อเทียบกับวิธีการเดิม

คำสำคัญ การเรียนรู้ร่วม, เรียนรู้แบบกึ่งมีผู้สอน

1. บทนำ

โลกปัจจุบันได้เข้าสู่ยุคดิจิทัล ดังเห็นได้จากอุปกรณ์ต่างๆ ที่มีการสร้างข้อมูลในเชิงดิจิทัลเพิ่มขึ้นอย่างเป็นจำนวนมาก โดยส่วนใหญ่เกิดจากการใช้งานอินเตอร์เน็ต จึงส่งผลให้พฤติกรรมของมนุษย์มีการเปลี่ยนแปลงจากอดีต กิจกรรมหลายๆอย่าง ถูกแทนที่ด้วยแพลตฟอร์มบนโลกออนไลน์ เช่น การซื้อสินค้า การประกาศรับสมัครงาน การดูหนังฟังเพลง การแชทหรือโซเชียลเน็ตเวิร์ค เป็นต้น แพลตฟอร์มเหล่านี้ได้สร้างข้อมูลจำนวนมหาศาลบนโลกออนไลน์ รายงานของ IBM ระบุว่า 90% ของข้อมูลทั้งหมดในโลกออนไลน์ ถูกสร้างขึ้นในช่วง 2 ปีหลังนี้เอง โดยปัจจุบันมีข้อมูลเกิดขึ้นใหม่ราว 2,500 ล้านกิกะไบต์ต่อวัน [1]

หากพิจารณาข้อมูลเหล่านั้น ข้อมูลที่มีป้ายกำกับ (Labeled Data) จะมีสัดส่วนน้อยมาก เมื่อเทียบกับข้อมูลที่ไม่มีป้ายกำกับ (Unlabeled Data) เนื่องจากการกำกับค่าให้ข้อมูลนั้น มีต้นทุนที่สูง หรือค่าที่กำกับไม่เป็นจริง หรืออาจไม่สามารถกำกับค่าได้ เพราะความต้องการใช้ข้อมูลที่เปลี่ยนแปลงไปอย่างรวดเร็ว ดังนั้นการสร้างโมเดลพยากรณ์จำเป็นต้องมีข้อมูลสอน (Training Data) ซึ่งข้อมูลสอนได้จากข้อมูลที่มีป้ายกำกับ ทว่าการที่ข้อมูลสอนนี้มีสัดส่วนน้อยมากอาจจะทำให้ความคลาดเคลื่อนสูง เมื่อเทียบกับการมีข้อมูลที่มีป้ายกำกับที่มีมากกว่า ดังนั้นการให้โมเดลเรียนรู้แบบกึ่งมีผู้สอน (Semi-supervised Learning) จากข้อมูลทั้งมีป้ายกำกับและข้อมูลที่ไม่มีป้ายกำกับจึงถูกนำมาประยุกต์ใช้ ซึ่งวิธีที่ได้รับความนิยมอย่างแพร่หลายคือ วิธีการโค เทรนนิง (Co-Training)

วิธีการ โค เทรนนิงมีการนำเสนอครั้งแรกโดย Blum และ Mitchell ในปี 1998 [2] ซึ่งเป็นการนำข้อมูลที่ไม่มีป้ายกำกับมาช่วยเพิ่มประสิทธิภาพของโมเดลพยากรณ์ การเลือกข้อมูลที่ไม่มีป้ายกำกับนี้จะใช้การสร้างโมเดล 2 โมเดลและเลือกข้อมูลที่ไม่มีป้ายกำกับที่มีความเชื่อมั่นมากที่สุดจากการพยากรณ์มาเพิ่มเข้าไปในข้อมูลสอน และสร้างโมเดลพยากรณ์ใหม่วนเรื่อยไป หลังจากนั้น Luca Didaci, Giorgio Fumera และ Fabio Roli [3] ได้ทำการวิจัยถึงผลกระทบของประสิทธิภาพของโมเดลที่เกิดจากการใช้โค เทรนนิงกับขนาดของชุดข้อมูลสอน นั่นคือลดขนาดของชุดข้อมูลสอน ให้มีขนาดน้อยที่สุด จนกระทั่งไม่สามารถนำไปใช้ได้ โดยทดสอบกับข้อมูลทั้งหมด 24 ชุดข้อมูล พบว่าขนาดของข้อมูลที่มีป้ายกำกับเพียง 1 ตัวอย่างต่อชุดข้อมูลสอน ไม่ส่งผลต่อประสิทธิภาพการโค เทรนนิง ถัดมา Ruiya Wang และ Li Li [4] ได้ทำพัฒนาอัลกอริทึมการปรับปรุงประสิทธิภาพของโค เทรนนิงโดยคณะกรรมการ (Co-Training by committee) เป็นวิธีการเรียนรู้แบบกึ่งกำกับซ้ำ ซึ่งในระหว่างการทำซ้ำ จะใช้หลายๆ โมเดล (committee) ก่อนหน้านั้นทั้งหมดหลายๆ ชุด เพื่อใช้ในการทำนายตัวอย่างที่ไม่มีป้ายกำกับในแต่ละครั้ง ซึ่งสามารถเพิ่มความแม่นยำในการทำนายได้ถึง 10% จากนั้น Ricardo Sousa และ Joao Gama [5] ทำการเปรียบเทียบระหว่างวิธีการเรียนรู้ร่วมและวิธีการเรียนรู้ด้วยตนเอง (self-learning) สำหรับการถดถอยที่มีเป้าหมายเดียวในข้อมูลแบบสตรีม ด้วยกฎการปรับ โมเดลสุ่ม (Random Adaptive Model Rules) เปรียบเทียบผลจากสถานการณ์ที่ไม่นำเอาข้อมูลที่ไม่มีป้ายกำกับเข้าไปในโมเดล และสถานกรณ์ที่นำเอาข้อมูลที่ไม่มีป้ายกำกับเข้าไปเพื่อปรับปรับการถดถอย ซึ่งผลลัพธ์แสดงหลักฐานที่ทำให้ประสิทธิภาพที่ดีขึ้น ในเรื่องช่วยลดความคลาดเคลื่อนในข้อมูลแบบสตรีมระดับสูง นอกจากนี้ยังมีงานวิจัยของ Fan Ma และคณะ [6] ได้นำเสนออัลกอริทึมโค เทรนนิงแบบใหม่ที่ชื่อว่า SPaCo (Self-Paced Co-training) การเรียนรู้ร่วมด้วยตัวเอง แก้ไขปัญหาการกำกับค่าของตัวอย่างที่ไม่มีป้ายกำกับที่ไม่ถูกต้องในรอบการฝึกขั้นต้น โดยการแทนที่ของตัวอย่าง (เลือกตัวอย่างเข้าและออก) ซึ่งสามารถเพิ่มประสิทธิภาพของโมเดลได้ดียิ่งขึ้น

งานวิจัยที่ใช้เทคนิคโค เทรนนิงจะเน้นที่การจำแนกประเภทข้อมูล (classification) มากกว่าการประมาณค่า (regression) ซึ่งในหลายๆ งานการประมาณค่าก็เป็นสิ่งจำเป็น ดังนั้น Zhi-Hua Zhou และ Ming Li [7] ได้ทำการวิจัยและนำเสนอวิธีการ COREG (Co-Training Regressors) การเรียนรู้ร่วมแบบกึ่งถดถอย โดยจะทำการเลือกตัวอย่างที่ไม่มีป้ายกำกับมากำกับค่า ผ่านโมเดลที่ให้ค่าความคาดเคลื่อนน้อยที่สุดทั้งสองโมเดล และการพยากรณ์ในขั้นสุดท้าย โดยการหาค่าเฉลี่ยของสมการถดถอยที่สร้างขึ้นทั้งสองตัว ซึ่งอัลกอริทึมนี้สามารถใช้

ประโยชน์จากข้อมูลที่ไม่มีป้ายกำกับเพื่อปรับปรุงการพยากรณ์แบบถดถอย วิธีการนี้ได้มีการนำไปใช้อย่างแพร่หลายแต่ใช้เวลาการทำงานที่นานเนื่องจากในการเลือกข้อมูลที่ไม่มีป้ายกำกับจำเป็นต้องทดสอบกับข้อมูลที่ไม่มีป้ายกำกับทีละตัวอย่าง

เนื่องจากวิธีการของ COREG ใช้เวลาการทำงานนาน คณะผู้วิจัยจึงได้นำเสนอวิธีการปรับปรุงประสิทธิภาพของโมเดลพยากรณ์ COREG ด้วยวิธีการเลือกข้อมูลที่ไม่มีป้ายกำกับอย่างอัตโนมัติ (AU-COREG) เพื่อการสร้างโมเดลการเรียนรู้ร่วมแบบกึ่งมีผู้สอน สำหรับข้อมูลที่มีป้ายกำกับที่มีสัดส่วนน้อย เมื่อเทียบกับข้อมูลที่ไม่มีป้ายกำกับ โดยเพื่อลดความคลาดเคลื่อนจากการพยากรณ์ และลดระยะเวลาในการสร้างโมเดล เพื่อรองรับกับข้อมูลที่หลากหลายและมีจำนวนมาก

2. แนวคิด / วิธีการที่นำเสนอ

2.1. เครื่องมือที่ใช้ในการวิจัย

เครื่องมือการสร้างโมเดล คือ Rapid Miner Studio เวอร์ชั่น 9.2.000 บนเครื่องลูกข่าย (Client) Processor 2.7 GHz Intel Core i5 RAM 8 GB 1867 MHz DDR3 และ Rapid Miner Server CPU 4 cores RAM 8GB และ HDD 120 GB

2.2. ศึกษาและรวบรวมข้อมูล

ข้อมูลที่ใช้ในการวิจัยมีทั้งหมด 3 ชุดข้อมูล

2.2.1. ชุดข้อมูลการพยากรณ์เงินเดือน (Job Salary Prediction)

ข้อมูลโฆษณาประกาศสมัครงานในประเทศอังกฤษ จัดทำโดย Adzuna (ข้อมูลที่ใช้ในการแข่งขัน Job Salary Prediction : Kaggle) ซึ่งเป็นข้อมูลที่ซึ่งประกอบด้วยข้อมูลดังนี้

- เงินเดือน (salary: US dollar)
- ชื่อตำแหน่ง (title)
- สถานที่ตั้งบริษัท (location)
- ประเภทการจ้างงาน (contact_type)
- สัญญาการจ้างงาน (contact time)
- บริษัท (company)
- ประเภทธุรกิจ (business_case)
- แหล่งข้อมูล (source)

2.2.2. ชุดข้อมูลการพยากรณ์ปริมาณการจราจร (Metro Interstate Traffic Volume Data Set)

ข้อมูลปริมาณการจราจรระหว่างรัฐโดยรถไฟฟ้ารายชั่วโมง จากทิศตะวันตกของมินนิโซตา DoT ATR สถานี 301 ซึ่งอยู่กลางระหว่างมินนิอาโปลิสและเซนต์พอล มินนิโซตา (UCI Machine Learning Repository) โดยมีข้อมูลดังนี้

- ปริมาณการจราจร (traffic_volume)
- วันหยุด (holiday)
- อุณหภูมิโดยเฉลี่ย (temp เป็น องศาเคลวิน)
- ปริมาณน้ำฝน (rain_1h (mm))
- ปริมาณหิมะ (snow_1h (mm))
- ร้อยละปริมาณหมอกที่ปกคลุม (clouds_all)
- ประเภทลักษณะอากาศ (weather_main)
- ลักษณะอากาศ (weather_description)

2.2.3. ชุดข้อมูลการพยากรณ์พลังงานไฟฟ้า (Combined Cycle Power Plant Data Set)

ข้อมูลที่รวบรวมจากโรงไฟฟ้าพลังความร้อนร่วม ในช่วง 6 ปี (2549-2554) (UCI Machine Learning Repository) โดยมีข้อมูลดังนี้

- พลังงานไฟฟ้า (EP)
- อุณหภูมิ (AT)
- ความดันบรรยากาศ (AP)
- ความชื้นสัมพัทธ์ (RH)
- ไอเสีย (V)

ตารางที่ 1. แสดงลักษณะของชุดข้อมูลและการสุ่มข้อมูล

ชุดข้อมูล	จำนวนแอตทริบิวต์	ประเภทข้อมูล	ขนาดข้อมูล (แถว)	ขนาดข้อมูลที่ใช้ (แถว)
1	8	Nominal	244,768	5,000
2	8	Nominal, Real	48,204	5,000
3	5	Real	9,568	5,000

2.3. วิธีการสุ่มข้อมูล

การวิจัยครั้งนี้ใช้วิธีการสุ่มข้อมูล 2 วิธีดังนี้

2.3.1 วิธีการสุ่มตัวอย่างแบบชั้นภูมิ (Stratified Random Sampling) ซึ่งเป็นการสุ่มตัวอย่างจากประชากรที่มีจำนวนมาก โดยประชากรจะถูกแบ่งออกเป็นชั้นภูมิตามลักษณะอย่างใดอย่างหนึ่ง โดยไม่ให้มีหน่วยซ้ำกัน ซึ่งในชั้นภูมิเดียวกันจะประกอบด้วยหน่วยที่มีลักษณะคล้ายคลึงกันมากที่สุด และแตกต่างระหว่างชั้นภูมิมากที่สุด

2.3.2 วิธีการสุ่มตัวอย่างแบบง่าย (Simple random sampling) เป็นการสุ่มตัวอย่างโดยถือว่าทุกๆ หน่วยหรือทุกๆ สมาชิกในประชากรมีโอกาสจะถูกเลือกเท่าๆ กัน

2.4. การจัดการข้อมูล

ชุดข้อมูลที่ 2 และ 3 มีข้อมูลประเภทตัวเลขและแต่ละแอตทริบิวต์มีขนาดข้อมูลที่แตกต่างกัน ดังนั้นการแปลงข้อมูลให้อยู่ในสเกลเดียวกัน (Normalize) จึงถูกนำมาใช้ในงานวิจัยนี้เลือกใช้วิธี Z-transformation ซึ่งทำให้เป็นค่ามาตรฐานและการกระจายของข้อมูลมีค่าเฉลี่ยเป็นศูนย์และความแปรปรวนเป็นหนึ่ง ดังแสดงในสมการต่อไปนี้

$$Z_i = (X_i - Mean) / Standard\ Deviation$$

ตารางที่ 2. แสดงค่าสถิติของข้อมูลชุดที่ 2

แอตทริบิวต์	ค่าน้อยที่สุด	ค่ามากที่สุด	ค่าเฉลี่ย
Temp (K)	245.62	308.43	281.24
rain_1h (mm)	0.00	25.57	0.147
snow_1h (mm)	0.00	0.44	0.00
clouds_all (%)	0.00	100.00	48.88

ตารางที่ 3. แสดงค่าสถิติของข้อมูลชุดที่ 3

แอตทริบิวต์	ค่าน้อยที่สุด	ค่ามากที่สุด	ค่าเฉลี่ย
AT (C)	1.81	35.56	19.58
V(cm Hg)	25.36	81.56	54.25
AP (milibar)	992.90	1,033.30	1,013.36
RH (MW)	25.56	100.16	48.88

2.5. ทฤษฎีที่เกี่ยวข้อง

2.5.1. การจำแนกข้อมูลด้วยวิธีการเพื่อนบ้านใกล้ที่สุดเคตัว (K-Nearest Neighbors :K-NN)

เป็นวิธีการใช้จำแนกหรือทำนายข้อมูลด้วยการเรียนรู้จากข้อมูล (Supervised Learning) ที่มีป้ายกำกับ (Labeled Data) โดยทำการเปรียบเทียบความคล้ายคลึงกับข้อมูลที่มีอยู่มากที่สุดเคตัว และกำหนดกลุ่มให้กับข้อมูลที่ไม่มีป้ายกำกับตามสมาชิกส่วนใหญ่ของกลุ่ม

วิธีการเปรียบเทียบความคล้ายคลึงจะถูกกำหนดในรูปแบบของระยะทางในหลายๆ มิติ ตามขนาดของแอตทริบิวส์ในชุดข้อมูลการเรียนรู้

ขั้นตอนการหาเพื่อนบ้านเคตัว มีดังต่อไปนี้

1) กำหนดค่าเค (k) โดยปกติจะนิยมเป็นจำนวนคี่

2) คำนวณหาความคล้ายคลึงของข้อมูลที่ไม่มีป้ายกำกับกับข้อมูลที่มีป้ายกำกับ (ระยะทาง)

3) เรียงลำดับความคล้ายคลึงและเลือกข้อมูลตัวอย่างที่มีความคล้ายคลึงมากที่สุดเคตัว

4) พิจารณาข้อมูลตัวอย่างทั้งเคตัว เพื่อจัดจำแนกหรือทำนายข้อมูลแต่ละตัวว่าถูกจัดเป็นกลุ่มใด (y)

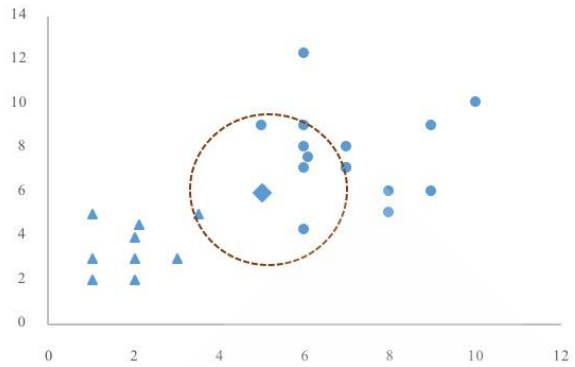
5) กำหนดกลุ่มใหม่ให้กับข้อมูลที่ไม่มีป้ายกำกับด้วยกลุ่มข้อมูลที่มีตัวอย่างมากที่สุดจากค่าเค

การวัดความคล้ายคลึงด้วยวิธีการวัดระยะห่างยูคลิเดียน (Euclidean Distant) เป็นการวัดระยะห่างระหว่าง 2 จุดในแนวเส้นตรงที่ได้มาจากทฤษฎีพีทาโกรัส ระยะห่างยูคลิเดียนระหว่างจุด p และ q คำนวณได้จาก

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \qquad (2)$$

หากแอตทริบิวส์มีข้อมูลแบบอัตภาคชั้น (Nominal) การวัดระยะทางหากค่าเหมือนกันระยะทางจะเป็นศูนย์หากต่างกันจะเป็นค่าเป็นอย่างอื่น

ตัวอย่างการจำแนกข้อมูลด้วยเพื่อนบ้านทั้ง n ตัว ดังรูปที่ 1



รูปที่ 1. แสดงตัวอย่างจำแนกข้อมูลด้วยเพื่อนบ้าน 7 ตัว

2.5.2. การจัดกลุ่ม (Clustering)

การจัดกลุ่มข้อมูลจากความคล้ายคลึงกัน (Clustering) เป็นเทคนิคการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) โดยพยายามให้ระยะห่างของสิ่งที่อยู่ในกลุ่มเดียวกันให้อยู่ใกล้กันมากที่สุด (Minimize Intra-Cluster Distance) และสิ่งที่อยู่ต่างกลุ่มกันจะมีระยะห่างแตกต่างกันมากที่สุด (Maximize Inter-Cluster Distance) หรืออาจกล่าวได้ว่ากลุ่มข้อมูลที่มีคุณสมบัติและ/หรือคุณลักษณะที่คล้ายคลึงกันควรอยู่ในกลุ่มข้อมูลเดียวกัน และข้อมูลที่มีคุณสมบัติและ/หรือคุณสมบัติที่แตกต่างกันอย่างมากควรอยู่ต่างกลุ่มกัน ดังตัวอย่างการจัดกลุ่ม n กลุ่มดังรูปที่ 2

2.5.2.1. การจัดกลุ่มด้วยเทคนิค K-Mean

เป็นวิธีการจัดกลุ่มที่วิเคราะห์กลุ่มแบบไม่เป็นขั้นตอนหรือการแบ่งส่วน (Partitioning) ออกเป็นเคกลุ่ม และแทนค่าแต่ละกลุ่มด้วยค่าเฉลี่ยของกลุ่ม หรือเรียกว่าจุดศูนย์กลาง (Centroid) ของกุล่ม

2.5.2.2. การจัดกลุ่มด้วยเทคนิค K-Medoids

เป็นวิธีการจัดกลุ่มที่วิเคราะห์กลุ่มที่เหมือนกับเทคนิค K-Mean แต่การคำนวณจุดศูนย์กลางของกลุ่มจะแทนที่ด้วยค่าของข้อมูลจริงๆ ที่อยู่ในกลุ่มนั้น

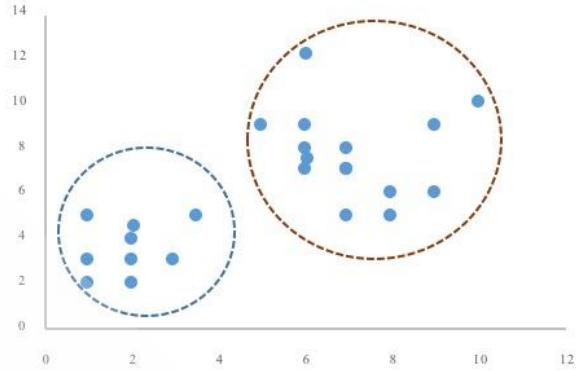
ขั้นตอนการจัดกลุ่มด้วยเทคนิค K-Mean และ K-Medoids

1) กำหนดค่าเริ่มต้นจำนวนเคกลุ่ม และกำหนดจุดศูนย์กลางเริ่มต้นทั้งเคจุด

2) พิจารณาข้อมูลที่เหลือเพื่อจัดเข้ากลุ่ม โดยการหาระยะห่างระหว่างข้อมูลกับจุดศูนย์กลาง โดยหากข้อมูลใดใกล้ค่าจุดศูนย์กลางตัวไหน จะทำการจัดเข้ากลุ่มนั้น

3) หาจุดศูนย์กลางของแต่ละกลุ่มโดย

3.1) เทคนิค K-Mean จะทำการหาค่าเฉลี่ย (Mean) ของแต่ละกลุ่มใหม่ และกำหนดให้เป็นจุดศูนย์กลางของกลุ่มใหม่



รูปที่ 2. แสดงตัวอย่างการจัดกลุ่ม 2 กลุ่ม

3.2) เทคนิค K-Medoids จะทำการหาค่ากลางค่าใหม่ของกลุ่ม แล้วเปรียบเทียบค่าความหนาแน่น เพื่อเลือกข้อมูลที่เป็นค่ากลางที่ทำให้ค่าความหนาแน่นต่ำที่สุด

4) ทำซ้ำข้อ 2) จนกระทั่งค่าเฉลี่ยหรือจุดศูนย์กลางใหม่ในแต่ละกลุ่มจะไม่มีการเปลี่ยนแปลง

2.5.3. การวิเคราะห์การถดถอย (Regression Analysis)

การวิเคราะห์การถดถอยเป็นเทคนนิคการสร้างตัวแบบจากความสัมพันธ์ระหว่างข้อมูล 2 ตัวเป็นต้นไป หรือทำนายข้อมูลตัวหนึ่ง (ข้อมูลเชิงปริมาณ) จากข้อมูลอีกตัว (หรือมากกว่า) สามารถเขียนสมการอย่างง่ายได้ดังสมการที่ (2) ดังนี้

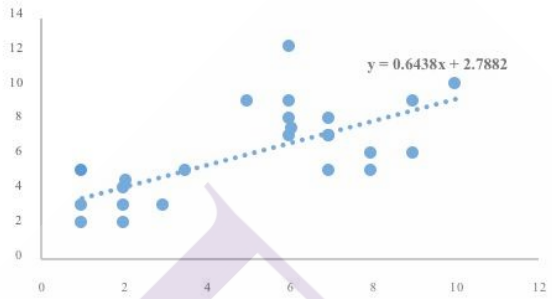
$$y = \alpha + \beta X + \varepsilon \qquad (2)$$

โดยที่ α เป็นค่าคงที่ที่ไม่ทราบค่าของสมการถดถอย β เป็นสัมประสิทธิ์ถดถอย (Regression Coefficient) เป็นอัตราการเปลี่ยนแปลงของค่า X ต่อค่า y และ ε เป็นค่าความคลาดเคลื่อนระหว่างค่าพยากรณ์ และค่าจริง y

ตัวอย่างการวิเคราะห์การถดถอยอย่างง่าย ดังรูปที่ 3 ซึ่งเรียกว่าตัวแบบถดถอยเชิงเส้นอย่างง่าย และมีตัววัดประสิทธิภาพของตัวแบบที่เป็นที่นิยม ได้แก่ สัมประสิทธิ์ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Root Mean Square

Error: RMSE) ดังสมการที่ (3) ซึ่งหมายถึงค่าความเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนจากการทำนาย

$$RMSE = \sqrt{\sum_{i=1}^{n}(prediction - actual)^2} \quad (3)$$



รูปที่ 3. แสดงการวิเคราะห์การถดถอยอย่างง่าย

2.6. ขั้นตอนการสร้างโมเดล AU-COREG

2.6.1. การแบ่งข้อมูล

ทำการแบ่งข้อมูลออกเป็น 2 ส่วน 1) ข้อมูลที่กำหนดให้มีป้ายกำกับ (Labeled Data) เพื่อใช้ในการสร้างและทดสอบโมเดล 2) ข้อมูลที่กำหนดให้ไม่มีป้ายกำกับ (Unlabeled Data) โดยการใช้การสุ่มข้อมูลแบบชั้นภูมิ ดังตารางที่ 4

ตารางที่ 4. แสดงการแบ่งข้อมูลเพื่อสร้างโมเดลขั้นแรก

ส่วนของข้อมูล	สัดส่วน	ขนาดข้อมูล (แถว)
ข้อมูลสำหรับการทดสอบโมเดลสุดท้าย (Testing Data)	25.0%	1,250
ข้อมูลสำหรับการสร้างโมเดลขั้นแรก (Initial Training Data) : L	7.5%	375
ข้อมูลที่ไม่มีป้ายกำกับ เพื่อเป็นส่วนที่เลือกข้อมูลเข้าใช้ในการสร้างโมเดลเพิ่มเติม : U'	2.0%	100 หรือ200
ข้อมูลที่ไม่มีป้ายกำกับ: U	65.5%	3,275

2.6.2. ขั้นตอนสร้างโมเดลขั้นต้นจากข้อมูลที่มีป้ายกำกับ (Labeled Data)

กำหนดให้ L_1 เป็นข้อมูลสำหรับสร้างโมเดลที่ 1 ในขั้นแรก และ L_2 เป็นข้อมูลสำหรับสร้างโมเดลที่ 2 ในขั้นแรก ซึ่งให้ L_1 และ L_2 มีค่าเท่ากับ L และสร้างโมเดล 2 โมเดลสมการที่ (4) และ (5)

$$h_1 \leftarrow kNN\,(L_1,k,p_1) \quad (4)$$

$$h_2 \leftarrow kNN\,(L_2,k,p_2) \quad (5)$$

2.6.3. กำหนดจำนวนรอบการทำซ้ำ (Iteration)

ในงานวิจัยกำหนดจำนวนรอบการทำซ้ำเป็น 100 รอบ (T=100) เพื่อเลือกข้อมูลในส่วน U' เข้าไปเป็นข้อมูล Training รอบละ 1 ตัวอย่าง โดยหลักการเลือกตัวอย่างเข้าไปนั้น จะทำการเลือกตัวอย่างที่ช่วยลดความคาดเคลื่อนจากการเพิ่มข้อมูลเข้าไปที่มากที่สุด จนกระทั่งครบ 100 รอบ หรือข้อมูลที่เพิ่มเข้าไปนั้นไม่สามารถช่วยลดความคาดเคลื่อนได้ หลังจากการเลือกข้อมูลเพิ่มเข้าไปทุกครั้ง (จาก U' ไป L) ต้องทำการสุ่มข้อมูลที่ไม่มีป้ายกำกับ (U) เข้าไปในส่วนของกลุ่มข้อมูล (U') ทุกครั้ง

1) พยากรณ์ข้อมูล U' ด้วยโมเดลที่ 1 (h_1) แล้วทำการจัดคลัสเตอร์ (Cluster) ข้อมูลที่ถูกกำกับค่า ด้วยเทคนิค K-Medoids โดยกำหนดจำนวนคลัสเตอร์เท่ากับ 2 (ในการทดลองจะทำการปรับค่าจำนวนคลัสเตอร์ตั้งแต่ 2 จนถึง 10 คลัสเตอร์)

- คลัสเตอร์ 1 ทำการเลือกสมาชิกที่เป็นตัวแทนคลัสเตอร์จากนั้นหาข้อมูลที่มีป้ายกำกับ (L_1) ที่มีระยะห่างกับตัวแทนของคลัสเตอร์ที่น้อยที่สุด (เพื่อนบ้าน) 5 ตัวอย่าง (มีความคล้ายคลึงกันมากที่สุด) โดยวัดระยะทางตามวิธีที่กำหนดในตารางที่ 5
- เพิ่มตัวแทนของคลัสเตอร์ที่ 1 เข้าไปใน L_1 เพื่อสร้างโมเดลใหม่ นำโมเดลที่ได้ไปทดสอบประสิทธิภาพของโมเดล ด้วยสัมประสิทธ์ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง RMSE (Root Mean Square Error) กับเพื่อนบ้านทั้ง 5 ตัว
- ทำตามขั้นตอนข้างต้นกับคลัสเตอร์ที่เหลือจนครบ หากพบ RMSE มากกว่า 0 ให้เลือกตัวแทนของคลัสเตอร์มีค่า RMSE มีค่ามากที่สุดออกจาก U' (π_1)

2) ทำตามขั้นตอนข้างต้นกับโมเดลที่ 2 และเลือกตัวแทนจากคลัสเตอร์ที่ได้จากการพยากรณ์ด้วยโมเดลที่ 2 ที่ทำให้ค่า RMSE มีค่ามากที่สุดออกจาก U' (π_2)

3) นำ π_1 เพิ่มเข้าไปใน L_1และนำ π_2 เพิ่มเข้าไปใน L_1 ดังสมการที่ (6) และ (7)

$$L_1 \leftarrow L_1 \cup \pi_2 \quad (6)$$

$$L_2 \leftarrow L_2 \cup \pi_1 \quad (7)$$

4) สร้างโมเดล 1 และ 2 จากข้อมูล L_1 และ L_2 ใหม่ และหากพบว่า L_1 และ L_2 ไม่เปลี่ยนแปลง ให้หยุดดำเนินการ

5) สุ่มเลือกตัวอย่างจาก U เพิ่มเข้ามาใน U' ให้ครบจำนวน

6) ทำตามขั้นตอนข้างต้น จนครบรอบจำนวนการทำซ้ำ (100 รอบ(

7) สร้างโมเดล AU-COREG และทดสอบประสิทธิภาพของโมเดล ดังสมการที่ (8)

$$h^*(x) = (h_1(x) + h_2(x)) / 2 \quad (8)$$

2.6.4. กำหนดจำนวนรอบการทำซ้ำ (Iteration) 200 รอบ

กำหนดจำนวนรอบการทำซ้ำเป็น 200 รอบ (T =200) และทำตามขั้นตอน 2.6.3. ทั้งหมด เพื่อสร้างโมเดลใหม่มาเปรียบเทียบ

2.6.5. ทำตามขั้นตอนที่ 2.6.3. โดยเปลี่ยนเทคนิคการทำ Cluster จาก K-Medoids เป็น K-Mean และทำการเลือกสมาชิกใน Cluster โดยการระบุค่า เพื่อเป็นตัวแทน Cluster และทำการทดสอบประสิทธิภาพของโมเดลที่ได้ บันทึกผลที่ได้ จากนั้นให้ดำเนินตามขั้นตอนเดิม และเลือกสมาชิกใหม่ใน Cluster ให้เป็นตัวแทนกลุ่ม จนครบ 3 รอบ บันทึกผลเพื่อนำมาเปรียบเทียบ

2.6.6. ทำตามขั้นตอนทั้งหมดกับชุดข้อมูลทั้ง 3 โดยโมเดลที่ใช้คือ kNN โดยวิธีการวัดระยะทาง และ K ให้เหมาะสมกับประเภทของข้อมูล ดังตารางที่ 5

ตารางที่ 5. แสดงโมเดลและวิธีการวัดระยะทาง

ชุดข้อมูล	โมเดล	วิธีวัดระยะทาง (p_i)	K
1	kNN1	Mix Euclidean Distance	3
	kNN2	Nominal Distance	3
2	kNN1	Mix Euclidean Distance	5
	kNN2	Mix Euclidean Distance	9
3	kNN1	Euclidean Distance	5
	kNN2	Correlation Similarity	5

3. ผลการทดลองและคำอธิบายรายละเอียด

ผลการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE ของโมเดล Self-training, COREG และ AU-COREG ที่มีการจัดกลุ่มตั้งแต่ 2 จนถึง 10 กลุ่ม และที่มี U' เท่ากับ 100 และ 200 และระยะเวลาที่ใช้ในการสร้างโมเดลของข้อมูลชุดที่ 1 ดังตารางที่ 6-10

ตารางที่ 6. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล SELF-TRAINNING และ COREG ของข้อมูลชุดที่ 1

MODEL	Training Set = 100		Training Set = 200	
	RMSE	TIME	RMSE	TIME
SELF TRAINING	17,299.17		16,926.85	
COREG	16,126.82	276	16,104.93	537

ตารางที่ 7. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล AU-COREG ที่จัด Cluster ด้วยวิธี K-Medoids ของข้อมูลชุดที่ 1

MODEL		U'100		U'200	
AU-COREG		RMSE	TIME	RMSE	TIME
K-MEDOIDS	2 CLUSTER	16,180.82	21	16,003.33	23
	3 CLUSTER	16,129.74	23	16,165.40	25
	4 CLUSTER	16,282.24	25	16,107.52	27
	5 CLUSTER	16,279.32	28	16,257.53	31
	6 CLUSTER	16,047.12	31	16,301.17	34
	7 CLUSTER	16,279.32	34	16,128.76	37
	8 CLUSTER	16,251.33	37	16,037.87	40
	9 CLUSTER	16,279.38	40	16,124.75	43
	10 CLUSTER	16,313.28	42	16,141.95	46

ตารางที่ 8. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล AU-COREG ที่จัด Cluster ด้วยวิธี K-Mean และ Seed = 1992

ของข้อมูลชุดที่ 1

MODEL		U'100		U'200	
AU-COREG		RMSE	TIME	RMSE	TIME
K-MEAN , SEED = 1992	2 CLUSTER	16,195.76	17	16,368.06	18
	3 CLUSTER	16,095.27	19	16,019.71	19
	4 CLUSTER	16,241.59	21	16,303.79	20
	5 CLUSTER	16,317.01	23	16,284.17	21
	6 CLUSTER	16,186.70	22	16,159.21	22
	7 CLUSTER	16,243.52	23	16,212.69	24
	8 CLUSTER	16,110.72	25	16,132.22	24
	9 CLUSTER	16,223.59	26	16,152.23	24
	10 CLUSTER	16,186.64	26	16,363.86	25

ตารางที่ 9. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล AU-COREG ที่จัด Cluster ด้วยวิธี K-Mean และ Seed = 100 ของข้อมูลชุดที่ 1

MODEL		U'100		U'200	
AU-COREG		RMSE	TIME	RMSE	TIME
K-MEAN , SEED = 100	2 CLUSTER	16,203.94	18	16,119.15	18
	3 CLUSTER	16,095.27	19	16,293.71	19
	4 CLUSTER	16,309.88	21	16,174.86	20
	5 CLUSTER	16,238.26	22	16,221.03	21
	6 CLUSTER	16,354.49	22	16,156.71	22
	7 CLUSTER	16,265.92	24	16,010.29	23
	8 CLUSTER	16,242.97	26	16,264.19	24
	9 CLUSTER	16,067.44	26	16,079.38	24
	10 CLUSTER	16,180.36	26	16,246.13	25

ตารางที่ 10. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล AU-COREG ที่จัด Cluster ด้วยวิธี K-Mean และ Seed = 3645 ของข้อมูลชุดที่ 1

MODEL		U'100		U'200	
AU-COREG		RMSE	TIME	RMSE	TIME
K-MEAN , SEED = 3745	2 CLUSTER	16,112.53	18	16,099.72	18
	3 CLUSTER	16,331.73	21	16,303.51	20
	4 CLUSTER	16,275.75	21	16,245.74	21
	5 CLUSTER	16,193.15	22	16,202.93	21
	6 CLUSTER	16,092.36	23	16,074.08	22
	7 CLUSTER	16,193.15	25	16,187.52	22
	8 CLUSTER	16,208.41	25	16,216.09	24
	9 CLUSTER	16,166.59	27	16,084.66	26
	10 CLUSTER	16,188.76	25	16,361.12	26

3.1. ผลการพยากรณ์ข้อมูลชุดที่ 1

เมื่อทำการพยากรณ์ข้อมูลชุดทดสอบ จากตารางที่ 6-10 แสดงให้เห็นว่า

- โมเดล kNN ที่ได้จากการเรียนรู้จากชุดข้อมูลฝึก ขนาด 100 และ 200 ตัวอย่าง จะได้ค่า RMSE เท่ากับ 17,299.17 และ 16,926.85 ตามลำดับ
- โมเดล COREG ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ ขนาด 100 และ 200 ตัวอย่าง จะได้ค่า RMSE เท่ากับ 16,126.82 และ 16,104.93 ตามลำดับ
- โมเดล AU-COREG (U'= 100) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ ขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Medoids) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 16,047.12 ซึ่งได้จากการทำ Cluster จำนวน 6 Cluster ลดลง 0.49% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 276 นาที เหลือเพียง 31 นาที หรือลดลงถึง 89% และโมเดล AU-COREG (U'= 200) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 16,003.33 ซึ่งได้จากการทำ Cluster จำนวน 2 Cluster ลดลง 0.14% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 527 นาที เหลือเพียง 23 นาที หรือลดลงถึง 96%
- โมเดล AU-COREG (U'= 200) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ ขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean โดยเลือกตัวแทน Cluster จาก Seed 1992 100 และ 3645) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 16,067.44 ซึ่งได้จากการทำ Cluster จำนวน 9 Cluster (Seed 100) ลดลง 0.37% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 276 นาที เหลือเพียง 26 นาที หรือลดลงถึง 91% และโมเดล AU-COREG (U'= 200) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 16,010.29 ซึ่งได้จากการทำ Cluster จำนวน 7 Cluster (Seed 100) ลดลง 0.59% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 527 นาที เหลือเพียง 23 นาที หรือลดลงถึง 96%

3.2. ผลการพยากรณ์ข้อมูลชุดที่ 2

ผลการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE ของโมเดล Self-training, COREG และ AU-COREG ที่มีการจัดกลุ่มตั้งแต่ 2 จนถึง 10 กลุ่ม และที่มี *U'* เท่ากับ 100 และ 200 และระยะเวลาที่ใช้ในการสร้างโมเดลของข้อมูลชุดที่ 2 ดังตารางที่ 11-15 สรุปได้ดังนี้

ตารางที่ 11. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล SELF-TRAINNING และ COREG ของข้อมูลชุดที่ 2

MODEL	Training Set = 100		Training Set = 200	
	RMSE	TIME	RMSE	TIME
SELF TRAINING	2,199.14		2,120.78	
COREG	2,113.46	215	2,110.38	428

ตารางที่ 12. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล AU-COREG ที่จัด Cluster ด้วยวิธี K-Medoids ของข้อมูลชุดที่ 2

MODEL		U'100		U'200	
AU-COREG		RMSE	TIME	RMSE	TIME
K-MEDOIDS	2 CLUSTER	2,121.32	12	2,120.32	17
	3 CLUSTER	2,122.24	13	2,115.78	17
	4 CLUSTER	2,119.87	15	2,108.29	19
	5 CLUSTER	2,115.01	22	2,111.66	24
	6 CLUSTER	2,118.39	25	2,114.48	28
	7 CLUSTER	2,113.61	28	2,113.61	31
	8 CLUSTER	2,112.36	31	2,115.24	33
	9 CLUSTER	2,110.84	34	2,111.19	36
	10 CLUSTER	2,110.05	36	2,120.08	39

ตารางที่ 13. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล AU-COREG ที่จัด Cluster ด้วยวิธี K-Mean และ Seed = 1992 ของข้อมูลชุดที่ 2

MODEL		U'100		U'200	
AU-COREG		RMSE	TIME	RMSE	TIME
K-MEAN , SEED = 1992	2 CLUSTER	2,118.82	12	2,117.95	14
	3 CLUSTER	2,114.02	12	2,109.64	17
	4 CLUSTER	2,114.52	15	2,120.36	17
	5 CLUSTER	2,117.16	20	2,118.93	20
	6 CLUSTER	2,110.68	23	2,117.87	23
	7 CLUSTER	2,112.78	26	2,110.86	26
	8 CLUSTER	2,113.20	28	2,115.21	29
	9 CLUSTER	2,117.96	31	2,111.94	31
	10 CLUSTER	2,111.36	33	2,107.28	33

ตารางที่ 14. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล AU-COREG ที่จัด Cluster ด้วยวิธี K-Mean และ Seed = 100 ของข้อมูลชุดที่ 2

MODEL		U'100		U'200	
AU-COREG		RMSE	TIME	RMSE	TIME
K-MEAN , SEED = 100	2 CLUSTER	2,114.02	12	2,109.80	12
	3 CLUSTER	2,119.22	14	2,120.06	15
	4 CLUSTER	2,117.78	17	2,123.85	17
	5 CLUSTER	2,115.14	20	2,116.97	20
	6 CLUSTER	2,110.68	23	2,109.16	22
	7 CLUSTER	2,116.55	25	2,114.88	25
	8 CLUSTER	2,115.95	28	2,111.49	28
	9 CLUSTER	2,117.86	30	2,109.37	30
	10 CLUSTER	2,111.38	33	2,117.51	33

ตารางที่ 15. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล AU-COREG ที่จัด Cluster ด้วยวิธี K-Mean และ Seed = 3645 ของข้อมูลชุดที่ 2

MODEL		U'100		U'200	
AU-COREG		RMSE	TIME	RMSE	TIME
K-MEAN , SEED = 3645	2 CLUSTER	2,114.02	12	2,125.11	12
	3 CLUSTER	2,119.22	14	2,113.98	15
	4 CLUSTER	2,117.78	17	2,108.27	17
	5 CLUSTER	2,115.14	20	2,119.51	21
	6 CLUSTER	2,110.68	23	2,114.91	23
	7 CLUSTER	2,116.55	25	2,116.81	26
	8 CLUSTER	2,115.95	28	2,116.16	29
	9 CLUSTER	2,117.86	30	2,117.70	31
	10 CLUSTER	2,111.38	33	2,115.96	34

- โมเดล kNN ที่ได้จากการเรียนรู้จากชุดข้อมูลฝึก ขนาด 100 และ 200 ตัวอย่าง จะได้ค่า RMSE เท่ากับ 2,199.14 และ 2,120.78 ตามลำดับ

- โมเดล COREG ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ ขนาด 100 และ 200 ตัวอย่าง จะได้ค่า RMSE เท่ากับ 2,113.46 และ 2,110.38 ตามลำดับ

- โมเดล AU-COREG (*U'*= 100) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ ขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Medoids) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 2,110.05 ซึ่งได้จากการทำ Cluster จำนวน 10 Cluster ลดลง 0.16% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 215 นาที เหลือเพียง 36 นาที หรือลดลงถึง 83% และโมเดล AU-COREG

(*U'*= 200) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 2,108.29 ซึ่งได้จากการทำ Cluster จำนวน 4 Cluster ลดลง 0.10% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 428 นาที เหลือเพียง 19 นาที หรือลดลงถึง 96%

- โมเดล AU-COREG (*U'*= 200) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ ขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean โดยเลือกตัวแทน Cluster จาก Seed 1992 100 และ 3645) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 2,107.95 ซึ่งได้จากการทำ Cluster จำนวน 10 Cluster (Seed 3645) ลดลง 0.26% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 215 นาที เหลือเพียง 34 นาที หรือลดลงถึง 84% และโมเดล AU-COREG (*U'*= 200) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 2,107.28 ซึ่งได้จากการทำ Cluster จำนวน 10 Cluster (Seed 1992) ลดลง 0.15% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 428 นาที เหลือเพียง 33 นาที หรือลดลงถึง 92%

3.3. ผลการพยากรณ์ข้อมูลชุดที่ 3

ผลการเปรียบเทียบประสิทธิภาพของโมเดลด้วยค่า RMSE ของโมเดล Self-training, COREG และ AU-COREG ที่มีการจัดกลุ่มตั้งแต่ 2 จนถึง 10 กลุ่ม และที่มี *U'* เท่ากับ 100 และ 200 และระยะเวลาที่ใช้ในการสร้างโมเดลของข้อมูลชุดที่ 3 ดังตารางที่ 16-20 สรุปได้ดังนี้

ตารางที่ 16. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล SELF-TRAINNING และ COREG ของข้อมูลชุดที่ 3

MODEL	Training Set = 100		Training Set = 200	
	RMSE	TIME	RMSE	TIME
SELF TRAINING	10.76		10.67	
COREG	7.90	217	7.70	419

ตารางที่ 17. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล AU-COREG ที่จัด Cluster ด้วยวิธี K-Medoids ของข้อมูลชุดที่ 3

MODEL		U'100		U'200	
AU-COREG		RMSE	TIME	RMSE	TIME
K-MEDOIDS	2 CLUSTER	7.63	12	7.74	12
	3 CLUSTER	7.66	15	7.70	15
	4 CLUSTER	7.70	17	7.75	18
	5 CLUSTER	7.75	20	7.66	21
	6 CLUSTER	7.74	23	7.84	25
	7 CLUSTER	7.65	26	7.65	28
	8 CLUSTER	7.75	28	7.66	30
	9 CLUSTER	7.81	31	7.79	34
	10 CLUSTER	7.74	33	7.75	36

ตารางที่ 18. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล AU-COREG ที่จัด Cluster ด้วยวิธี K-Mean และ Seed = 1992 ของข้อมูลชุดที่ 3

MODEL		U'100		U'200	
AU-COREG		RMSE	TIME	RMSE	TIME
K-MEAN , SEED = 1992	2 CLUSTER	7.73	12	7.82	12
	3 CLUSTER	7.68	14	7.68	15
	4 CLUSTER	7.71	17	7.78	17
	5 CLUSTER	7.78	21	7.78	20
	6 CLUSTER	7.66	23	7.74	22
	7 CLUSTER	7.77	26	7.78	26
	8 CLUSTER	7.81	28	7.79	28
	9 CLUSTER	7.80	31	7.91	30
	10 CLUSTER	7.84	33	7.86	33

ตารางที่ 19. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล AU-COREG ที่จัด Cluster ด้วยวิธี K-Mean และ Seed = 100 ของข้อมูลชุดที่ 3

MODEL		U'100		U'200	
AU-COREG		RMSE	TIME	RMSE	TIME
K-MEAN , SEED = 100	2 CLUSTER	7.63	12	7.72	13
	3 CLUSTER	7.64	15	7.70	15
	4 CLUSTER	7.72	18	7.75	18
	5 CLUSTER	7.75	20	7.78	21
	6 CLUSTER	7.75	23	7.71	23
	7 CLUSTER	7.85	25	7.71	25
	8 CLUSTER	7.63	27	7.76	28
	9 CLUSTER	7.78	30	7.81	31
	10 CLUSTER	7.76	33	7.89	33

ตารางที่ 20. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล AU-COREG ที่จัด Cluster ด้วยวิธี K-Mean และ Seed = 3645 ของข้อมูลชุดที่ 3

MODEL	U'100		U'200	
AU-COREG	RMSE	TIME	RMSE	TIME

K-MEAN , SEED = 3645	2 CLUSTER	7.66	12	7.70	12
	3 CLUSTER	7.61	15	7.67	15
	4 CLUSTER	7.79	18	7.71	17
	5 CLUSTER	7.78	21	7.70	20
	6 CLUSTER	7.65	24	7.76	23
	7 CLUSTER	7.77	26	7.89	25
	8 CLUSTER	7.70	28	7.87	28
	9 CLUSTER	7.71	31	7.79	31
	10 CLUSTER	7.78	34	7.90	33

- โมเดล kNN ที่ได้จากการเรียนรู้จากชุดข้อมูลฝึก ขนาด 100 และ 200 ตัวอย่าง จะได้ค่า RMSE เท่ากับ 10.76 และ 10.67 ตามลำดับ

- โมเดล COREG ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ ขนาด 100 และ 200 ตัวอย่าง จะได้ค่า RMSE เท่ากับ 7.90 และ 7.70 ตามลำดับ

- โมเดล AU-COREG (U'= 100) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ ขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Medoids) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 7.63 ซึ่งได้จากการทำ Cluster จำนวน 2 Cluster ลดลง 1.80% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 217 นาที เหลือเพียง 12 นาที หรือลดลงถึง 94% และโมเดล AU-COREG (U'= 200) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 7.65 ซึ่งได้จากการทำ Cluster จำนวน 7 Cluster ลดลง 3.21% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 419 นาที เหลือเพียง 28 นาที หรือลดลงถึง 93%

- โมเดล AU-COREG (U'= 200) ที่ได้จากการเพิ่มตัวอย่างที่ถูกกำกับค่าจากข้อมูลที่ไม่มีป้ายกำกับ ขนาด 2 ตัวอย่าง จนกระทั่งถึง 10 ตัวอย่าง (ตัวแทนของ Cluster ด้วยเทคนิค K-Mean โดยเลือกตัวแทน Cluster จาก Seed 1992 100 และ 3645) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 7.61 ซึ่งได้จากการทำ Cluster จำนวน 3 Cluster (Seed 3645) ลดลง 1.58% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 217 นาที เหลือเพียง 24 นาที หรือลดลงถึง 89% และโมเดล AU-COREG (U'= 200) จะได้ค่า RMSE ที่มีค่าต่ำสุดเท่ากับ 7.67 ซึ่งได้จากการทำ Cluster จำนวน 3 Cluster (Seed 3645) ลดลง 2.88% (เทียบกับ COREG) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 419 นาที เหลือเพียง 15 นาที หรือลดลงถึง 97%

4. สรุปและอภิปรายผล

การสร้างโมเดลจากการเลือกข้อมูลที่ไม่มีป้ายกำกับอย่างอัตโนมัติแบบกึ่งถดถอย (AU-COREG) ถูกพัฒนามาจากโมเดล COREG โดยลดขนาดตัวอย่างที่ไม่มีป้ายกำกับเพื่อคัดเลือกเข้าโมเดล ด้วยการทำ Cluster เพื่อช่วยลดจำนวนรอบในการคำนวณ และยังทำให้ประสิทธิภาพของโมเดลไม่ลดลง ซึ่ง

จำนวนรอบในการคำนวณ วิธี COREG

$$N = T * i * n * nU' \qquad (9$$

จำนวนรอบในการคำนวณ วิธี AU-COREG

$$N = T * i * n * nC \qquad (10$$

โดยที่ N คือจำนวนรอบในการคำนวณ

T จำนวนรอบในการทำซ้ำ

i จำนวนโมเดล

n จำนวนเพื่อนบ้าน

nU' จำนวนตัวอย่างที่ไม่มีป้ายกำกับ

nC จำนวนคลัสเตอร์

หากเพิ่มจำนวน Cluster ให้มากขึ้น จะเพิ่มจำนวนรอบในการคำนวณมากขึ้นอย่างไม่มีนัยสำคัญ ดังนั้นผู้วิจัยจึงได้ทำการทดลองเพิ่มจำนวน Cluster ตั้งแต่ 2 จนถึง 10 Cluster และเพิ่ม U' ให้มากขึ้นจาก 100 เป็น 200 ตัวอย่าง เพื่อให้สามารถทำการจัดกลุ่มให้ดียิ่งขึ้น

โมเดล AU-REG ต้องการเลือกตัวแทนของแต่ละ Cluster ที่ดีที่สุด (ลดค่าคาดเคลื่อนจากการพยากรณ์ให้มากที่สุด) เพื่อนำเข้าชุดข้อมูลเพื่อฝึกการเรียนรู้ให้กับโมเดล ดังนั้นผู้วิจัยจึงได้ใช้เทคนิคการทำ Cluster 2 วิธี ได้แก่ K-Medoids และ K-Mean โดยวิธี K-Medoids ผู้วิจัยเลือกสมาชิกที่อยู่จุดกึ่งกลางของข้อมูลใน Cluster (Centroid) เป็นตัวแทนของ Cluster และวิธี K-Mean ผู้วิจัยทำการระบุ

ค่าสุ่ม ในการเลือกสมาชิก 3 ค่า เพื่อเปรียบเทียบประสิทธิภาพของโมเดล จากการเลือกตัวแทนของ Cluster ที่แตกต่างกัน

จากผลการทดลอง ดังตารางที่ 21-23 พบว่าโมเดล AU-COREG

- เมื่อทำการเพิ่มจำนวนข้อมูลที่ไม่มีป้ายกำกับเพิ่มจาก 100 เป็น 200 โดยส่วนใหญ่ทำให้ค่า RMSE ลดลง
- การเพิ่มจำนวนของ Cluster (แต่ไม่มากจนเกินไป) จะช่วยทำให้ค่า RMSE ลดลงได้ โดยใช้ระยะเวลาไม่แตกต่างไปจากเดิม
- การเลือกตัวแทนของ Cluster โดยใช้เทคนิคการจัด Cluster ที่แตกต่างกัน (K-Medoids และ K-Mean) มีผลต่อค่า RMSE ซึ่งพบกว่าส่วนใหญ่การจัด Cluster ด้วยวิธี K-Medoids กับข้อมูลประเภท Nominal จะให้ค่า RMSE ที่ต่ำกว่า

ตารางที่ 21. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล AU-COREG ที่มีค่าน้อยที่สุดของข้อมูลชุดที่ 1

Model	*U'*100		*U'2*00	
	RMSE	Time (m)	RMSE	Time (m)
COREG	16,126.82	276	16,104.93	276
AU-COREG	16,047.12	31	16,003.33	31
ลดลง	0.49%	89%	0.14%	96%

ตารางที่ 22. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล AU-COREG ที่มีค่าน้อยที่สุดของข้อมูลชุดที่ 2

Model	*U'*100		*U'2*00	
	RMSE	Time (m)	RMSE	Time (m)
COREG	2,113.46	215	2,110.38	428
AU-COREG	2,107.95	34	2,107.28	33
ลดลง	0.26%	84%	0.15%	92%

ตารางที่ 23. แสดงค่า RMSE และระยะเวลา (นาที) จากโมเดล AU-COREG ที่มีค่าน้อยที่สุดของข้อมูลชุดที่ 3

Model	*U'*100		*U'2*00	
	RMSE	Time (m)	RMSE	Time (m)
COREG	7.77	217	7.90	419
AU-COREG	7.61	15	7.65	28
ลดลง	1.58%	93%	3.25%	93%

การวิจัยครั้งนี้นำเสนอขั้นตอนการสร้างโมเดลจากการเลือกข้อมูลที่ไม่มีป้ายกำกับอย่างอัตโนมัติแบบกึ่งถดถอย (AU-COREG) ซึ่งทำการทดลองกับข้อมูล 3 ชุด ที่มีคุณลักษณะของข้อมูลที่แตกต่างกัน และใช้โมเดลพยากรณ์ kNN 2 โมเดล ที่มีวิธีการวัดระยะทางเหมือนกันแต่ต่างกันที่ค่าพารามิเตอร์เค หรือวิธีการวัดระยะทางที่แตกต่างกันขึ้นอยู่กับลักษณะของข้อมูล

ในการเรียนรู้ซ้ำในแต่ละรอบ มีการจัดกลุ่มข้อมูลที่ไม่มีป้ายกำกับ ที่ผ่านการทำนายจากโมเดลพยากรณ์ โดยและเลือกตัวแทนกลุ่มเพื่อหาตัวแทนกลุ่ม ที่ทำให้ความคลาดเคลื่อนน้อยที่สุด โดยการวิจัยนี้ ยังมีการเพิ่มข้อมูลที่ไม่มีป้ายกำกับ เพื่อให้ข้อมูลเหมาะสมกับการจัดกลุ่มที่เพิ่มขึ้น และเพิ่มวิธีการคัดเลือกตัวแทนของกลุ่ม เพื่อเพิ่มประสิทธิภาพของโมเดลให้ดียิ่งขึ้น

การทำนายขั้นสุดท้าย โดยหาค่าเฉลี่ยของการทำนายของทั้งสองโมเดล การวิจัยนี้แสดงให้เห็นว่าวิธี AU-COREG สามารถปรับปรุงประสิทธิภาพของโมเดล โดยช่วยลด RMSE ได้มากกว่า 0.14และยังช่วยลดเวลา % มากกว่า 84%

ผู้วิจัยยังขาดการปรับพารามิเตอร์ให้เหมาะสม กับคุณลักษณะของข้อมูล เช่น จำนวนข้อมูลที่ไม่มีป้ายกำกับที่จะเพิ่มเข้าไปในข้อมูลการเรียนรู้ จำนวนเพื่อนบ้านที่ใช้ทดสอบความคลาดเคลื่อน จำนวนรอบการทำซ้ำ เป็นต้น ซึ่งอาจช่วยปรับปรุงโมเดลให้มีประสิทธิภาพมากยิ่งขึ้น

เอกสารอ้างอิง

[1] Bizibl Marketing, “10 Key Marketing Trends for 2017 and Ideas for Exceeding Customer Expectations” Retrieved June 16, 2019 from https://bizibl.com/marketing/download/10-key-

marketing-trends-2017-and-ideas-exceeding-customer-expectations

[2] Blum A., Mitchell T., "Combining labeled and unlabeled data with co-training," COLT' 98 Proceedings of the eleventh annual conference on Computational learning theory, pp 92-100, July. 1998.

[3] Didaci L., Fumera, G., Roli, F., "Analysis of co-training algorithm with very small training sets," Gimel'farb, G., et al. (eds.) SSPR/SPR 2012. LNCS., Springer, Heidelberg., vol. 7626, 2012.

[4] R Wang, L Li., "The performance improvement algorithm of co-training by committee," 5th international conference on computer science and network technology, 2016.

[5] Sousa R., Gama J., "Co-training Semi-supervised Learning for Single-Target Regression in Data Streams Using AMRules," International Symposium on Methodologies for Intelligent Systems, 2017.

[6] F Ma, D Meng, Q Xie, Z Li, X Don, "Self-paced co-training," Proceedings of the 34th International Conference on Machine Learning-Volume 70, 2017.

[7] Zhi Hua., Ming Li., "Semi-supervised regression with co-training," IJCAI'05 proceeding of the 19th international joint conference on artificial intelligence, 2005.

ประวัติผู้เขียน

ชื่อ-นามสกุล	นางสาวศิริขวัญ คีรีสุวรรณกุล
ประวัติการศึกษา	วิทยาศาสตร์บัณฑิต
	สาขาสถิติ
	มหาวิทยาลัยขอนแก่น
	ปีการศึกษา 2544
ตำแหน่งและสถานที่ทำงานปัจจุบัน	นักวิเคราะห์ข้อมูลขนาดใหญ่,
	ธนาคารทหารไทย จำกัด (มหาชน)

ผลงานทางวิชาการ

- **ศิริขวัญ คีรีสุวรรณกุล** และ รัชนี ภูวพัฒนะพันธุ์. (2555). ระบบสนับสนุนการตัดสินใจสำหรับการจับคู่รถบรรทุกวิ่งเที่ยวเปล่ากับงานขนส่ง: กรณีศึกษาเว็บไซต์ dxplace.com. การประชุมวิชาการด้านการวิจัยดำเนินงานแห่งชาติ, 2555, 239-245.

- **ศิริขวัญ คีรีสุวรรณกุล** และ เอกสิทธิ์ พัชรวงศ์ศักดา. (2562). วิธีการเลือกข้อมูลที่ไม่มีป้ายกำกับอย่างอัตโนมัติเพื่อการสร้างโมเดลการเรียนรู้ร่วมแบบกึ่งมีผู้สอน การประชุม National Conference on Information (NCIT2019) ครั้งที่ 11, 2562.