JIST Journal of Information Science and Technology
Volume 10, NO 1 | JAN – JUN 2020 | 10-23

วิธีการเลือกข้อมูลที่ไม่มีป้ายกำกับอย่างอัตโนมัติเพื่อการสร้างโมเดลการเรียนรู้ร่วมแบบกึ่งมีผู้สอน

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

ศิริขวัญ คีรีสุวรรณกุล และ เอกสิทธิ์ พัชรวงศ์ศักดา*

Sirikwan Kheereesuwannakul and Eakasit Pacharawongsakda*

สาขาวิศวกรรมข้อมูลขนาดใหญ่ วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์

มหาวิทยาลัยธุรกิจบัณฑิตย์

Big Data Engineering, College of Innovative Technology and Engineering, Dhurakit Pundit University

Received: November 30, 2019; **Revised:** February 02, 2020; **Accepted:** February 27, 2020; **Published:** June 23, 2020

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.

KEYWORDS: 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

**Corresponding Author: 585162020005@dpu.ac.th*

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 โคเทรนนิงมีการนำเสนอครั้งแรกโดย 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 ด้วยกฎการปรับโมเดลสุ่ม (Random Adaptive Model Rules) เปรียบเทียบผลจากสถานการณ์ที่ไม่นำเอาข้อมูลที่ไม่มีป้ายกำกับเข้าไปในโมเดล และสถานการณ์ที่นำเอาข้อมูลที่ไม่มีป้ายกำกับเข้าไปเพื่อปรับปรับการถดถอย 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 โดยจะทำการเลือกตัวอย่างที่ไม่มีป้ายกำกับมากำกับค่าผ่านโมเดลที่ให้ค่าความคาดเคลื่อนน้อยที่สุดทั้งสองโมเดล และการพยากรณ์ในขั้นสุดท้าย โดยการหาค่าเฉลี่ยของสมการถดถอยที่สร้างขึ้นทั้งสองตัว ซึ่งอัลกอริทึมนี้สามารถใช้ประโยชน์จากข้อมูลที่ไม่มีป้ายกำกับเพื่อปรับปรุงการพยากรณ์แบบถดถอย วิธีการนี้ได้มีการนำไปใช้อย่างแพร่หลายแต่ใช้เวลาการทำงานที่นานเนื่องจากในการเลือกข้อมูลที่ไม่มีป้ายกำกับจำเป็นต้องทดสอบกับข้อมูลที่ไม่มีป้ายกำกับทีละตัวอย่าง

เนื่องจากวิธีการของ COREG ใช้เวลาการทำงานนาน คณะผู้วิจัยจึงได้นำเสนอวิธีการปรับปรุงประสิทธิภาพของโมเดลพยากรณ์ COREG ด้วยวิธีการเลือกข้อมูลที่ไม่มีป้ายกำกับอย่างอัตโนมัติ (AU-COREG) เพื่อการสร้างโมเดลการ

เรียนรู้ร่วมแบบกึ่งมีผู้สอน 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 ซึ่งอยู่กลางระหว่างมินนิอาโปลิสและเซนต์พอลมินนิโซตา (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) จึงถูกนำมาใช้ ในงานวิจัยนี้เลือกใช้วิธี 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

(1) วิธีการเปรียบเทียบความคล้ายคลึงจะถูกกำหนดในรูปแบบของระยะทางในหลายๆ มิติ ตามขนาดของแอตทริบิวต์ในชุดข้อมูลการเรียนรู้

ขั้นตอนการหาเพื่อนบ้านเคตัว มีดังต่อไปนี้

1) กำหนดค่าเค (k) โดยปกติจะนิยมเป็นจำนวนคี่

2) คำนวณหาความคล้ายคลึงของข้อมูลที่ไม่มีป้ายกำกับกับข้อมูลที่มีป้ายกำกับ (ระยะทาง)

3) เรียงลำดับความคล้ายคลึงและเลือกข้อมูลตัวอย่างที่มีความคล้ายคลึงมากที่สุดเคตัว

4) พิจารณาข้อมูลตัวอย่างทั้งเคตัว เพื่อจัดจำแนกหรือทำนายข้อมูลแต่ละตัวว่าถูกจัดเป็นกลุ่มใด

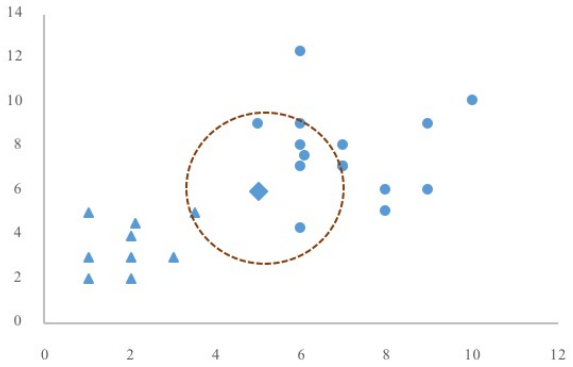
5) กำหนดกลุ่มใหม่ให้กับข้อมูลที่ไม่มีป้ายกำกับด้วยกลุ่มข้อมูลที่มีตัวอย่างมากที่สุดจากค่าเค

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

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

หากแอตทริบิวต์มีข้อมูลแบบอัตภาคชั้น (Nominal) การวัดระยะทางหากค่าเหมือนกันระยะทางจะเป็นศูนย์ หากต่างกันจะเป็นค่าเป็นอย่างอื่น

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

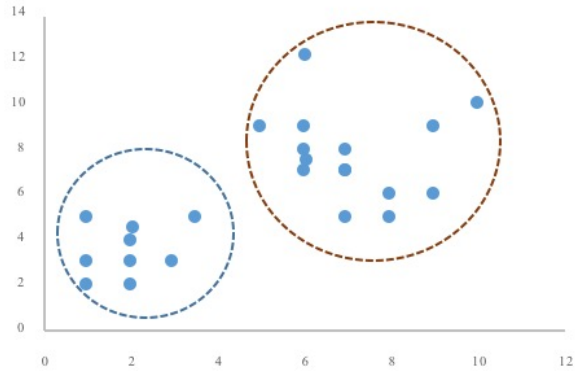
ขั้นตอนการจัดกลุ่มด้วยเทคนิค K-Mean และ K-Medoids

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)

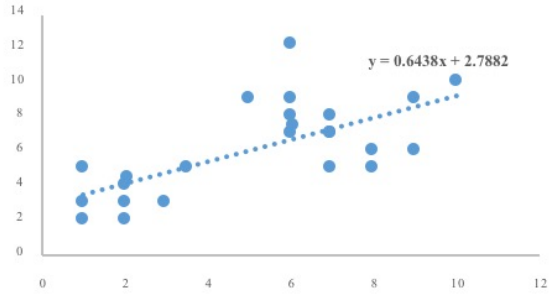
การวิเคราะห์การถดถอยเป็นเทคนิคการสร้างตัวแบบจากความสัมพันธ์ระหว่างข้อมูล 2 ตัวเป็นต้นไป หรือทำนายข้อมูลตัวหนึ่ง (ข้อมูลเชิงปริมาณ) จากข้อมูลอีกตัว (หรือมากกว่า) สามารถเขียนสมการอย่างง่ายได้ดังสมการที่ (2) ดังนี้

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

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 ด้วยสัมประสิทธิ์ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง RMSE (Root Mean Square Error) กับเพื่อนบ้านทั้ง 5 ตัว
- ทำตามขั้นตอนข้างต้นกับคลัสเตอร์ที่เหลือจนครบ หากพบ RMSE มากกว่า 0 ให้เลือกตัวแทนของคลัสเตอร์มีค่า RMSE มีค่ามากที่สุดออกจาก U' (π_1)

2) ทำตามขั้นตอนข้างต้นกับโมเดลที่ 2 และเลือกตัวแทนจากคลัสเตอร์ที่ได้จากการพยากรณ์ด้วยโมเดลที่ 2 ที่ทำ

ให้ค่า 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

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 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

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

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

MODEL		U'100		U'200	
AU-COREG		RMSE	TIME	RMSE	TIME
K-MEDOIDS	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 ที่ได้จากการเรียนรู้จากชุดข้อมูลฝึกขนาด 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) ในขณะที่ระยะเวลาในการสร้างโมเดล ลดลงจาก 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) ในขณะที่ระยะเวลาในการสร้างโมเดลลดลงจาก 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

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

MODEL		U'100		U'200	
AU-COREG		RMSE	TIME	RMSE	TIME
K-MEDOIDS	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 ที่ได้จากการเรียนรู้จากชุดข้อมูลฝึกขนาด 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

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

จำนวนรอบในการคำนวณ วิธี AU-COREG

$$N = T * i * n * nC \quad (10)$$

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 ผู้วิจัยทำการระบุค่าสุ่ม 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'_{200}	
	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'_{200}	
	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'_{200}	
	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 ที่มีคุณลักษณะของ

ข้อมูลที่แตกต่างกัน และใช้โมเดลพยากรณ์ kNN 2 โมเดล ที่มีวิธีการวัดระยะทางเหมือนกันแต่ต่างกันที่ค่าพารามิเตอร์เค หรือวิธีการวัดระยะทางที่แตกต่างกันขึ้นอยู่กับลักษณะของข้อมูล

ในการเรียนรู้ซ้ำในแต่ละรอบ มีการจัดกลุ่มข้อมูลที่ไม่มีป้ายกำกับ ที่ผ่านการทำนายจากโมเดลพยากรณ์ โดยและเลือกตัวแทนกลุ่มเพื่อหาตัวแทนกลุ่ม ที่ทำให้ความคลาดเคลื่อนน้อยที่สุด โดยการวิจัยนี้ ยังมีการเพิ่มข้อมูลที่ไม่มีป้ายกำกับ เพื่อให้ข้อมูลเหมาะสมกับการจัดกลุ่มที่เพิ่มขึ้น และเพิ่มวิธีการคัดเลือกตัวแทนของกลุ่ม เพื่อเพิ่มประสิทธิภาพของโมเดลให้ดียิ่งขึ้น

การทำนายขั้นสุดท้าย โดยหาค่าเฉลี่ยของการทำนายของทั้งสองโมเดล การวิจัยนี้แสดงให้เห็นว่าวิธี AU-COREG สามารถปรับปรุงประสิทธิภาพของโมเดลโดยช่วยลด RMSE ได้มากกว่า 0.14% และยังช่วยลดเวลามากกว่า 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,” Bizibl Marketing, June 16, 2019. [Online]. Available: https://bizibl.com/marketing/download/10-key-marketing-trends-2017-and-ideas-exceeding-customer-expectations. [Accessed: June 16, 2020].

[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, July, 1998, pp. 92-100.

[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 and L. Li, "The performance improvement algorithm of co-training by committee," 2016 5th International Conference on Computer Science and Network Technology (ICCSNT), Changchun, 2016, pp. 407-412, doi: 10.1109/ICCSNT.2016.8070190.

[5] Sousa R., Gama J., “Co-training Semi-Supervised Learning for Single-Target Regression in Data Streams Using AMRules,” In: Kryszkiewicz M., Appice A., Ślęzak D., Rybinski H., Skowron A., Raś Z. (eds) Foundations of Intelligent Systems, ISMIS 2017, Lecture Notes in Computer Science, Vol. 10352, 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, Vol. 70, pp. 2275-2284, 2017.

[7] Zhi Hua., Ming Li., “Semi-supervised regression with co-training,” IJCAI’05 proceeding of the 19th international joint conference on artificial intelligence, July, 2005, pp. 908–913.