บทความวิจัย

การจำแนกข้อมูลขนาดใหญ่โดยใช้การจัดกลุ่มด้วยวิธีเคมีนและวิธีการเรียนรู้เชิงลึก

นันทัชพร เสนาวงค์ สุภาวดี วิชิตชาญ และ อรวิชญ์ กุมพล*
ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยมหาสารคาม

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 08 4511 2886 อีเมล: bungon.k@msu.ac.th DOI: 10.14416/j.kmutnb.2021.03.012
รับเมื่อ 8 ตุลาคม 2563 แก้ไขเมื่อ 18 ธันวาคม 2563 ตอบรับเมื่อ 23 ธันวาคม 2563 เผยแพร่ออนไลน์ 24 มีนาคม 2564

บทคัดย่อ

ในการจำแนกประเภทข้อมูลที่มีขนาดใหญ่ ปัญหาที่พบคือเวลาที่ใช้ในการประมวลผลนาน และต้องใช้ข้อมูลฝึก (Training Data) เป็นจำนวนมากเพื่อให้การจำแนกประเภทมีประสิทธิภาพความแม่นยำสูง เพื่อแก้ไขปัญหานี้ผู้วิจัยจึงศึกษาวิธีการสำหรับการจำแนกข้อมูลขนาดใหญ่ เพื่อลดปัญหาการใช้ข้อมูลฝึกจำนวนมาก แต่ยังคงมีประสิทธิภาพในการจำแนกประเภทสูง โดยจะทำการลดขนาดข้อมูลฝึกด้วยการรวมเทคนิคการจัดกลุ่มของวิธีเคมีน (K-means) และวิธีการเรียนรู้เชิงลึก (Deep Learning) ในการศึกษาประสิทธิภาพของวิธีการที่นำเสนอพิจารณาจากค่าความแม่นยำและค่า AUC นอกจากนี้ได้ทำการเปรียบเทียบกับวิธีการเรียนรู้เชิงลึกแบบเดิมที่ใช้ข้อมูลฝึกขนาด 80% และ 90% ของข้อมูลทั้งหมด และกรณีที่ใช้ข้อมูลฝึกจำนวนเท่ากัน ผลการศึกษาพบว่า วิธีการที่นำเสนอสามารถลดขนาดของข้อมูลฝึกได้อย่างมาก โดยใช้ข้อมูลฝึกน้อยกว่า 1% ของขนาดข้อมูลทั้งหมด แต่ให้ค่าความแม่นยำเฉลี่ยและค่า AUC เฉลี่ยของการจำแนกประเภทสูง โดยในกรณีที่ข้อมูลมีการแจกแจงปรกติขนาด 1,000,000 × 5 (N × Feature) วิธีการที่นำเสนอให้ค่าความแม่นยำเฉลี่ยสูงถึง 97.4878% และให้ค่า AUC เฉลี่ยสูงถึง 0.9735 และเมื่อเปรียบเทียบกับผลการจำแนกประเภทโดยใช้วิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึก 80% และ 90% ของข้อมูลทั้งหมดพบว่า ประสิทธิภาพในการจำแนกประเภทสูงใกล้เคียงกัน โดยที่วิธีการที่นำเสนอใช้เวลาในการจำแนกประเภทน้อยกว่าวิธีการเรียนรู้เชิงลึกประมาณ 2–4 เท่า

คำสำคัญ: ข้อมูลขนาดใหญ่ การจัดกลุ่มของวิธีเคมีน การตรวจหาค่าผิดปกติ วิธีการเรียนรู้เชิงลึก การจำแนกประเภท

การอ้างอิงบทความ: นันทัชพร เสนาวงค์, สุภาวดี วิชิตชาญ และ อรวิชญ์ กุมพล, "การจำแนกข้อมูลขนาดใหญ่โดยใช้การจัดกลุ่มด้วยวิธีเคมีนและวิธีการเรียนรู้เชิงลึก," *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 32, ฉบับที่ 4, หน้า 957–965, ต.ค.–ธ.ค. 2565.

Research Article

Large-scale Data Classification based on K-means Clustering and Deep Learning

Nuntuschaporn Senawong, Supawadee Wichitchan and Orawich Kumphon*
Department of Mathematies, Faculty of Science, Mahasarakham University, Maha Sarakham, Thailand

* Corresponding Author, Tel. 08 4511 2886, E–mail: bungon.k@msu.ac.th
DOI: 10.14416/j.kmutnb.2021.03.012
Received 8 October 2020; Revised 18 December 2020; Accepted 23 December 2020; Published online: 24 March 2021

Abstract

Common problems in classifying large data are revealed as long processing time and a lot of training data in order to maintain high accuracy. To solve these problems, researchers study methods for classifying large data to reduce the use of large amounts of training data without sacrificing high classification efficiency. The proposed method reduces the size of the training data by combining K-means and deep learning. To study the effectiveness of the proposed method, the accuracy and AUC values were determined. In addition, it was compared with the original deep learning method using 80% and 90% training data out of the total data and was compared with the original deep learning using the same amount of training data. The results show that the proposed method can significantly reduce the size of the training data. Less than 1% of the total data size was used as training data, but the method yielded the high average percent of accuracy and the high average AUC of the classification. In the case of normal distribution and the size is 1,000,000 × 5 (N × Feature), the proposed method exhibits the average percent of accuracy as high as 97.4878% and the average AUC as 0.9735. When the proposed method was compared with the deep learning method using training data about 80% and 90% of the total data size, classification efficiency was relatively as high as that of the deep learning, but the classification time was 2–4 times less than the processing time of the deep learning method.

Keywords: Large-scale Data, K-means Clustering, Outlier Detection, Deep Learning, Classification

Please cite this article as: N. Senawong, S. Wichitchan, and O. Kumphon, "Large-scale data classification based on K-means clustering and deep learning," *The Journal of KMUTNB*, vol. 32, no. 4, pp. 957–965, Oct.–Dec. 2022 (in Thai).

1. บทนำ

ในปัจจุบันนิยมใช้วิธีการเรียนรู้ของเครื่อง (Machine Learning) เพื่อสืบค้นความรู้จากฐานข้อมูลขนาดใหญ่ ซึ่งสามารถสกัดข้อมูลที่มีประโยชน์และน่าสนใจ อีกทั้งยังจดจำรูปแบบจากฐานข้อมูลขนาดใหญ่ได้ นั่นคือการทำเหมืองข้อมูล (Data Mining) ซึ่งเป็นเทคนิคที่ใช้จัดการกับข้อมูลขนาดใหญ่ โดยจะนำข้อมูลที่มีมาทำการวิเคราะห์แล้วดึงความรู้หรือสิ่งสำคัญออกมาเพื่อใช้ในการวิเคราะห์หรือทำนายสิ่งต่างๆ เป็นกระบวนการขุดค้นสิ่งที่น่าสนใจในข้อมูลที่มีอยู่ [1] ดังนั้นวิธีการเรียนรู้ของเครื่องจึงเป็นหัวใจหลักในการวิเคราะห์ข้อมูล เช่น การจำแนกประเภทข้อมูลขนาดใหญ่ โดยการใช้เทคนิคการจำแนกประเภทข้อมูล (Data Classification) เพื่อจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดจากกลุ่มตัวอย่างข้อมูลที่เรียกว่าข้อมูลฝึก ทั้งนี้ เมื่อต้องการจำแนกประเภทข้อมูลที่มีขนาดใหญ่ ปัญหาที่ตามมาคือการประมวลผลซึ่งต้องใช้เวลานาน และต้องใช้ข้อมูลฝึกเป็นจำนวนมาก โดยทั่วไปจะใช้ข้อมูลฝึกประมาณ 80% ถึง 90% ของข้อมูลทั้งหมด เพื่อได้ประสิทธิภาพความแม่นยำที่สูง ในการแก้ปัญหาดังกล่าว [2] ได้เสนอวิธีการสำหรับการจำแนกข้อมูลขนาดใหญ่โดยใช้การจัดกลุ่มด้วยวิธีเคมีนและวิธีมัลติเคอร์เนลซัพพอร์ตเวกเตอร์แมชชีน (Multi-kernel Support Vector Machine; Multi-kernel SVM) เพื่อลดขนาดของข้อมูลฝึกและลดเวลาในการฝึก (Train) ผลลัพธ์ที่ได้แสดงให้เห็นว่าวิธีการเลือกข้อมูลฝึกที่นำเสนอสามารถลดขนาดของข้อมูลฝึก อีกทั้งยังลดเวลาที่ใช้ในการฝึกและสามารถคงประสิทธิภาพความแม่นยำได้ดี

ปัจจุบันการจำแนกประเภทด้วยวิธีการเรียนรู้ของเครื่องในข้อมูลที่มีขนาดใหญ่มีหลายวิธี เช่น วิธีการเรียนรู้โดยการใช้โครงข่ายประสาทเทียม (Artificial Neural Network; ANN) โดย [3] ได้ศึกษาการประยุกต์ ANN หลายโครงข่ายบนข้อมูลขนาดใหญ่ แต่มักเกิดปัญหาในด้านเวลาที่ใช้ในการเรียนรู้ จึงได้เสนอการพัฒนาอัลกอริทึมในการรวมโหนด ผลการทดลองพบว่า วิธีการที่นำเสนอสามารถลดเวลาในการเรียนรู้ลงได้อย่างมาก และยังคงรักษาเปอร์เซ็นต์ความถูกต้องได้เหมือนกับการใช้ข้อมูลฝึก 80% ถึง 90% ทั้งนี้ วิธีการเรียนรู้โดยการใช้ ANN ได้พัฒนาเป็นวิธีการเรียนรู้เชิงลึก (Deep Learning) โดย [4] ได้เสนอการใช้เทคนิคการเรียนรู้เชิงลึกผ่านชุดโครงข่ายประสาทเทียม โดยศึกษาแบบจำลองทำนายผลคำตัดสินและประเด็นในคดีอาญาที่เรียนรู้จากคำพิพากษาศาลฎีกาไทย ผลการทดลองแสดงให้เห็นว่าแบบจำลองให้ประสิทธิภาพสูงกว่าแบบจำลองที่ใช้วิธีการเรียนรู้ของเครื่องแบบเดิม เช่น Naive Bayes และ SVM ในการพัฒนาแบบจำลองการดึงดูดการเดินทางโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ โดยทำการพัฒนาแบบจำลองโดยใช้วิธีการเรียนรู้เชิงลึก และเปรียบเทียบกับการวิเคราะห์ด้วยวิธีวิเคราะห์ความถดถอยและวิธีโครงข่ายประสาทเทียม ผลการวิจัยพบว่าข้อมูลจากการเช็คอินสามารถนำมาสร้างแบบจำลองการดึงดูดการเดินทางได้ เมื่อเปรียบเทียบแบบจำลองทั้งสามแบบพบว่า แบบจำลองที่ได้จากวิธีการเรียนรู้เชิงลึกให้ความถูกต้องในการพยากรณ์สูงที่สุด [5] เมื่อทำการประยุกต์ใช้เทคโนโลยีการเรียนรู้เชิงลึกในการจำแนกข้อมูลถนนจากภาพถ่ายโดรน (Drone) เพื่อการสำรวจถนนในเขตชนบทเพื่อปรับปรุง (Update) ข้อมูลถนน (Open Street Map; OSM) ซึ่งจะเลือกพื้นที่ที่ยังขาดข้อมูลในส่วนนี้ ซึ่งศักยภาพของระบบเทคโนโลยีในปัจจุบันที่น่าจะนำมาใช้ในการแก้ปัญหาในส่วนนี้ จากผลงานวิจัยนี้จึงทำให้สามารถจำแนกและวิเคราะห์ข้อมูลถนนในภาพถ่ายเพื่ออัปเดตข้อมูล OSM ได้อย่างถูกต้อง รวดเร็วและแม่นยำขึ้น [6] ความแม่นยำในการนำปัญญาประดิษฐ์ที่มีการเรียนรู้เชิงลึกมาทำหน้าที่แบ่งแยกพื้นที่บาดแผลเรื้องรังเป็นสิ่งที่จำเป็นสำหรับการประเมินและดูความก้าวหน้าในการฟื้นตัวของสภาพบาดแผล จากงานวิจัยพบว่า หากเพิ่มความหลากหลายของสีด้วยการขยายข้อมูลด้วยโมเดลรูปแบบสี จะมีความแม่นยำการแบ่งแยกพื้นที่บาดแผลใกล้เคียงวิธีการที่นำเสนอ ถึงแม้ว่าจะมีชุดข้อมูลการฝึกขนาดเล็กก็ตาม [7]

เพื่อเป็นการลดปัญหาการใช้ข้อมูลฝึกจำนวนมากและลดระยะเวลาในการประมวลผลสำหรับการจำแนกข้อมูลขนาดใหญ่ ผู้วิจัยจึงเสนอวิธีการในการลดขนาดข้อมูลฝึก โดยการรวมเทคนิคการจัดกลุ่มของวิธีเคมีนและวิธีการเรียนรู้เชิงลึก โดยเริ่มจากการลดขนาดข้อมูลฝึกด้วยคุณสมบัติของวิธีเคมีน

และการตรวจหาค่าผิดปกติ จากนั้นจะนำวิธีการเรียนรู้เชิงลึกมาใช้ในการจำแนกประเภทเพื่อให้ได้ข้อมูลฝึกที่มีขนาดลดลง แต่ยังคงมีประสิทธิภาพและมีความแม่นยำสูง อีกทั้งยังทำการเปรียบเทียบประสิทธิภาพความแม่นยำจากวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่ศึกษา และข้อมูลฝึกขนาด 80% และ 90% ของจำนวนข้อมูลทั้งหมด

2. วัสดุ อุปกรณ์และวิธีการวิจัย

ในการจำแนกข้อมูลที่มีขนาดใหญ่ นั่นคือจำนวนข้อมูล (N) × คุณลักษณะ (Feature) 500,000 เพื่อทำให้ข้อมูลฝึกมีขนาดลดลง แต่ยังคงประสิทธิภาพและความแม่นยำสูง มีวิธีการดำเนินการดังนี้

2.1 ข้อมูลที่ใช้ในการศึกษา

2.1.1 ชุดข้อมูลขนาดใหญ่ที่ได้จากการสร้างข้อมูล (Generate Data)

การสร้างข้อมูล (N × Feature) ซึ่งข้อมูลแต่ละชุดมี 4 ขนาด ดังนี้ 1) ขนาด 1,000,000 × 5, 2) ขนาด 100,000 × 10, 3) ขนาด 30,000 × 30 และ 4) ขนาด 7,000 × 75 ข้อมูลที่สร้างขึ้นมีความสัมพันธ์ระหว่าง Feature ไม่เกิน ±0.5 จากข้อมูลที่มีการแจกแจง 3 การแจกแจง ดังนี้ 1) การแจกแจงปรกติมาตรฐาน [Standard Normal Distribution; N(0,1)] 2) การแจกแจงแบบเลขชี้กำลัง [Exponential Distribution; exp(1)] และ 3) การแจกแจงเอกรูป [Uniform Distribution; U(0,1)]

2.1.2 ชุดข้อมูลจริงที่มีขนาดใหญ่ ประกอบไปด้วย 2 ชุดข้อมูล ดังนี้

1) ชุดข้อมูล Skin Segmentation โดยมีขนาดข้อมูล 245,057 × 4 จากฐานเก็บข้อมูล UCI สืบค้นจาก https:// archive.ics.uci.edu

2) ชุดข้อมูล Coil 2000 โดยมีขนาดข้อมูล 9,822 × 85 จากฐานเก็บข้อมูล KEEL สืบค้นจาก https:// sci2s.ugr.es/keel

2.2 วิธีการที่ใช้ในการศึกษา

2.2.1 วิธีการจัดกลุ่มด้วยคุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติ

ขั้นตอนการจัดกลุ่มด้วยคุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติ

1) ตรวจสอบว่าข้อมูลมีค่าผิดปกติหรือไม่ตรวจสอบความสมบูรณ์ของข้อมูล เช่น กรณีมีข้อมูลสูญหาย (Missing Data) หรือมี NA ให้ทำการตัดข้อมูลแถวนั้นออก

2) กำหนดสัดส่วน (Proportion) ในการสุ่มข้อมูลเพื่อเป็นข้อมูลฝึกจากข้อมูลทั้งหมดสุ่มข้อมูลฝึกด้วยอัตราส่วนที่กำหนด เช่น กำหนดอัตราส่วน 5% ถึง 40% จากข้อมูลทั้งหมด

3) ทำการจัดกลุ่มด้วยวิธีเคมีน

- กำหนดจำนวนกลุ่ม (Km) และจำนวนรอบในการทำซ้ำ (RT)
- หาระยะห่างระหว่างจุดศูนย์กลางของแต่ละกลุ่มทีละคู่ ซึ่งเรียกว่าค่าเฉลี่ยหรือค่ากลางของแต่ละกลุ่ม ด้วยหลักเกณฑ์การรวมกลุ่มแบบ Centroid Clustering โดยหาระยะห่างแบบยูคลิค (Euclidean Distance) ระหว่างจุดศูนย์กลางของกลุ่มหากคู่ใดต่ำจะรวมกลุ่มคู่นั้นเข้าเป็นกลุ่มเดียวกัน
- เก็บ Case ที่ใกล้สุดและไกลสุดจากจุดศูนย์กลางของแต่ละกลุ่มในทุก Km และ RT
- ลดขนาดของข้อมูลโดยการลบข้อมูลที่ซ้ำออก

4) คำนวณค่าผิดปกติของชุดข้อมูลโดยใช้สมการที่ (1) [8] ดังนี้

$$KSE(p_j) = \frac{1}{n-1}\sum_{i=1}^{n} KS(p_j - p_i) \quad (1)$$

โดยที่ $KS(p_j - p_i) = sup_x \left| Fp_j(x) - Fp_i(x) \right|$

เมื่อ Fp_j คือ ระยะทางจากจุด j ไปยังอีกจุด i อื่นๆ ที่รวบรวม

จากการคำนวณ ทำการเรียงค่าจากน้อยไปมาก จากนั้นทำการตัดข้อมูลที่มีค่ามากกว่าตำแหน่งเปอร์เซ็นต์ไทล์ที่กำหนด [9] เช่น ตัดข้อมูลที่มีค่ามากกว่าตำแหน่งเปอร์เซ็นต์ไทล์ที่ 90 เพื่อกำจัดค่าผิดปกติสูงสุดและลดขนาดของข้อมูลฝึก

2.2.2 วิธีการเรียนรู้เชิงลึก

วิธีการเรียนรู้เชิงลึกจะทำการคำนวณผ่าน ANN โดย ANN [10] สามารถคำนวณตามสมการที่ (2) [11] ได้ดังนี้

$$y_j = f\left(\sum_{i=1}^{n} x_i w_{ij} + \theta_j\right) \quad (2)$$

เมื่อ y_i คือ ผลลัพธ์ในชั้นซ่อน หรือข้อมูลส่งออกในชั้นซ่อนโหนดที่ j

x_i คือ ข้อมูลนำเข้าโหนดที่ i ในชั้นอินพุต

x_{ij} คือ น้ำหนักบนเส้นเชื่อมระหว่างโหนดที่ i ในชั้นอินพุตและโหนดที่ j ในชั้นซ่อน

θ_i คือ ค่า Bias ของโหนดที่ j ในชั้นซ่อน

n คือ จำนวนโหนดทั้งหมดของชั้นอินพุต

ในการศึกษาครั้งนี้ได้ทำการกำหนดชั้นซ่อน (Hidden Layer) เป็น 3 ชั้นซ่อน โดยในแต่ละชั้นซ่อนกำหนดโหนดเป็น 5, 10 และ 20 ซึ่งรวมทั้งหมดเป็น 27 กรณี ดังตารางที่ 1

ตารางที่ 1 การกำหนดชั้นซ่อนและโหนด

ชั้นซ่อน	โหนด
1, 2, 3	(5, 5, 5), (5, 5, 10), (5, 5, 20), (5, 10, 5), (5, 10, 10), (5, 10, 20), (5, 20, 5), (5, 20, 10), (5, 20, 20), (10,5, 5), (10, 5, 10), (10, 5, 20), (10,10, 5), (10,10,10), (10, 10, 20), (10, 20, 5), (10,20, 10), (10,20, 20), (20, 5, 5), (20, 5, 10), (20,5, 20), (20,10, 5), (20, 10, 10), (20,10, 20), (20,20, 5), (20,20,10), (20, 20, 20)

ในการสุ่มข้อมูลฝึกเพื่อให้ได้ข้อมูลฝึกจากทุกส่วนของข้อมูลทั้งหมด จึงทำการสุ่มข้อมูลแบบ k-fold โดยกำหนด k = 10 fold หมายถึงการแบ่งข้อมูลออกเป็น 10 ส่วนเท่าๆ กัน จากนั้นนำแต่ละส่วนใช้เป็นข้อมูลฝึก หากกำหนดการใช้ข้อมูลฝึก 90% จะทำการฝึกโมเดลทั้งหมด 10 รอบ และให้ค่า Accuracy เฉลี่ยจากทั้งหมด 10 รอบ [12]

2.3 เกณฑ์ในการวัดประสิทธิภาพความแม่นยำ

ในการวัดประสิทธิภาพความแม่นในการศึกษาครั้งนี้ใช้เกณฑ์การวัดด้วยค่าความแม่น (Accuracy) และเกณฑ์ในการทำนาย Area Under Curve (AUC) โดยจะใช้ค่าเฉลี่ยของ Accuracy และ AUC จากวิธีการเรียนรู้เชิงลึกทั้ง 27 กรณี มีรายละเอียดดังนี้

2.3.1 เกณฑ์การวัดด้วยค่าความแม่นยำ

ในการทดสอบประสิทธิภาพความแม่นยำ โดยใช้เกณฑ์ในการวัดด้วยค่าความแม่นยำ ซึ่งจะระบุว่าโมเดลทำนายถูกทั้งหมดกี่เปอร์เซ็นต์ โดยคำนวณจากเส้นทแยงมุมของเมทริกซ์ความสับสน (Confusion Matrix) ได้ดังตารางที่ 2

ตารางที่ 2 เมทริกซ์ความสับสน แบบ 2 × 2 [13]

	Predicted positive	Predicted negative
Positive class	True positive (TP)	False negative (FN)
Negative class	False positive (FP)	True negative (TN)

สามารถคำนวณได้โดยใช้สมการที่ (3) ดังนี้

$$Accuracy = \frac{(TP + TN)}{(TP + FN + FP + TN)} \quad (3)$$

โดยที่ $Accuracy$ คือ ค่าอยู่ระหว่าง 0–1 เมื่อค่าเข้าใกล้ 1 นั่นคือโมเดลสามารถจำแนกประเภทได้ดีมาก

2.3.2 เกณฑ์ในการวัดผล Area Under Curve (AUC)

ในการวัดผลด้วยค่า AUC นิยมใช้วัดผลโมเดลในทุกงานโดย AUC มีค่าอยู่ระหว่าง 0–1 เมื่อค่าเข้าใกล้ 1 นั่นหมายความว่าโมเดลในภาพรวมสามารถจำแนกประเภทได้ดีมาก ซึ่งสามารถคำนวณได้โดยใช้สมการที่ (4) ดังนี้ [14]

$$AUC = \frac{Sensitivity + Specificity}{2} \quad (4)$$

โดยที่ $Sensitivity$ คือ $\frac{TP}{(TP + FN)}$

$Specificity$ คือ $\frac{TN}{(TN + FP)}$

สามารถสรุปได้ดังเกณฑ์ต่อไปนี้ [12]

$AUC = 0.50$ คือ โมเดลมีประสิทธิภาพต่ำ

$0.70 \le AUC < 0.80$ คือ เกณฑ์มาตรฐานสำหรับโมเดลส่วนใหญ่

$0.80 \le AUC < 0.90$ คือ โมเดลทำงานได้ดี

$AUC > 0.90$ คือ โมเดลทำงานได้ดีมาก

3. ผลการทดลอง

ผลการวิจัยของวิธีการเรียนรู้เชิงลึกโดยใช้ข้อมูลฝึกที่ผ่านคุณสมบัติของวิธีเคมีนและการตรวจหาค่าผิดปกติเป็นดังนี้

3.1 การเปรียบเทียบประสิทธิภาพของวิธีการที่ศึกษากับวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่ศึกษาจากข้อมูลที่สร้างขึ้นทั้งจากข้อมูลที่สร้างขึ้นและข้อมูลจริง

ประสิทธิภาพความแม่นยำของวิธีการที่ศึกษา และวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่ศึกษา โดยกำหนดการตัดข้อมูลที่มีค่ามากกว่าตำแหน่งเปอร์เซ็นต์ไทล์ที่ 90 จากชุดข้อมูลที่ได้จากการสร้างข้อมูล (ตารางที่ 3) สำหรับชุดข้อมูลจริงจากชุดข้อมูล Skin Segmentation และชุดข้อมูล Coil2000 (ตารางที่ 4) พบว่าประสิทธิภาพความแม่นยำยังคงสูงกว่าวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่ศึกษา

3.2 การเปรียบเทียบประสิทธิภาพของการจำแนกประเภทโดยใช้ข้อมูลฝึกขนาด 80% และ 90%

ในส่วนนี้จะแสดงให้เห็นถึงขนาดของข้อมูลฝึก และประสิทธิภาพความแม่นยำระหว่างวิธีการเรียนรู้เชิงลึกแบบเดิมและวิธีการที่ศึกษา (ตัวหนา) ในตารางที่ 5 เป็นกรณีของชุดข้อมูลที่มีขนาด 7,000 × 75 ทำการสุ่มแบบ k-fold โดยกำหนด k = 10 fold หมายถึงการแบ่งข้อมูลออกเป็น 10 ส่วนเท่าๆ กัน จากนั้นนำแต่ละส่วนใช้เป็นข้อมูลฝึก

ตารางที่ 3 ประสิทธิภาพของวิธีการที่ศึกษาจากการสร้างข้อมูลโดยแสดงขนาดของข้อมูลฝึกและประสิทธิภาพความแม่นยำเมื่อเทียบกับวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่ศึกษา

ชุดข้อมูล	ขนาดข้อมูล (N×Feature)	พารามิเตอร์ สัดส่วน	ขนาดข้อมูลฝึก Case (%)	ค่าความแม่นยำเฉลี่ย	
				Accuracy %	AUC
Normal Distribution	1,000,000×5	0.05	104 (0.01%)	97.4878 (94.9744)	0.9735 (0.9659)
	100,000×10	0.10	90 (0.09%)	95.7757 (92.5282)	0.92284 (0.9075)
	30,000×30	0.10	50 (0.17%)	90.0225 (89.2700)	0.8280 (0.8333)
	7,000×75	0.10	48 (0.69%)	90.1968 (89.8467)	0.8026 (0.8351)
Exponential Distribution	1,000,000×5	0.05	109 (0.01%)	95.6735 (95.5576)	0.9259 (0.9568)
	100,000×10	0.10	93 (0.09%)	86.8219 (88.9000)	0.8231 (0.8722)
	30,000×30	0.10	79 (0.26%)	69.5191 (71.5839)	0.6938 (0.6775)
	7,000×75	0.10	61 (0.87%)	61.3046 (58.5077)	0.5946 (0.5986)
Uniform Distribution	1,000,000×5	0.05	55 (0.01%)	**95.7708 (89.6742)**	**0.9434 (0.9184)**
	100,000×10	0.10	99 (0.10%)	90.4185 (93.0459)	0.8663 (0.9098)
	30,000×30	0.10	78 (0.26%)	81.4312 (80.9128)	0.7843 (0.7817)
	7,000×75	0.10	60 (0.86%)	75.2455 (72.9768)	0.7069 (0.7151)

หมายเหตุ: กำหนดพารามิเตอร์ Km และพารามิเตอร์ RT = 10 ในการหาข้อมูลฝึก ค่า % ในวงเล็บ (.) คือ ขนาดข้อมูลฝึกจากข้อมูลทั้งหมด และค่าในวงเล็บ (.) หมายถึง ผลของวิธีการเรียนรู้เชิงลึก

ตารางที่ 4 ประสิทธิภาพของวิธีการที่ศึกษาจากชุดข้อมูลจริง โดยแสดงขนาดของข้อมูลฝึกและประสิทธิภาพความแม่นยำเมื่อเทียบกับวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่ศึกษา

ข้อมูล	ขนาดข้อมูล (N×Feature)	พารามิเตอร์			ขนาดข้อมูลฝึก Case (%)	ค่าความแม่นยำเฉลี่ย	
		สัดส่วน	Km	RT		Accuracy %	AUC
Skin Segmentation	245,057 × 4	0.40	10	10	90 (0.04%)	97.8902 (96.7434)	0.9864 (0.9379)
Coil2000	9,822 × 85	0.10	10	20	67 (0.68%)	93.7354 (92.3439)	0.9404 (0.9405)

ตารางที่ 5 ประสิทธิภาพของวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึกขนาด 80% และ 90% และวิธีการที่ศึกษา

ข้อมูล	วิธี	ขนาดข้อมูลฝึก Cases (%)	ค่าความแม่นยำเฉลี่ย		เวลา (วินาที)
			Accuracy %	AUC	
Standard Normal Distribution : N(0,1)	วิธีการเรียนรู้เชิงลึก	6,300 (90%)	96.5513	0.9413	209.41
		5,600 (80%)	96.5587	0.9425	129.25
สัดส่วน = 0.30; Km = 50; RT = 30	**วิธีการที่ศึกษา**	**524 (7.49%)**	**95.0804**	**0.9049**	**54.34**
Exponential Distribution : exp(1)	วิธีการเรียนรู้เชิงลึก	6,300 (90%)	94.6989	0.9190	977.30
		5,600 (80%)	94.3476	0.9129	851.58
สัดส่วน = 0.40; Km = 200; RT = 100	**วิธีการที่ศึกษา**	**2,079 (29.70%)**	**90.3053**	**0.8612**	**402.05**
Uniform Distribution : U(0,1)	วิธีการเรียนรู้เชิงลึก	6,300 (90%)	96.0540	0.9459	1432.68
		5,600 (80%)	95.7487	0.9429	1428.89
สัดส่วน = 0.30; Km = 200; RT = 30	**วิธีการที่ศึกษา**	**1,689 (24.13%)**	**92.0375**	**0.8819**	**440.16**

4. อภิปรายผลและสรุป

จากตารางที่ 3 พบว่า วิธีการที่ศึกษาสามารถลดขนาดของข้อมูลฝึกได้อย่างมากในทุกการแจกแจงที่ศึกษา โดยใช้ขนาดข้อมูลฝึกน้อยกว่า 1% ของจำนวนข้อมูลทั้งหมด ในกรณีที่ชุดข้อมูลมีขนาด N จำนวนมาก และจำนวนคุณลักษณะน้อย (1,000,000 × 5) ใช้ข้อมูลฝึกเพียง 104 เคส (Cases) หรือคิดเป็น 0.01% โดย 104 เคส นั้น ได้จากขั้นตอนแรกที่ทำการสุ่มข้อมูลให้มีสัดส่วน 0.05 หรือ 5% ของข้อมูลทั้งหมด จากนั้นนำข้อมูลที่สุ่มมาทั้ง 50,000 เคส ไปจัดกลุ่มด้วยวิธีเคมีนและการตรวจหาค่าผิดปกติ ทำให้เหลือจำนวนข้อมูลฝึกเพียง 104 เคส ทั้งนี้ กรณีที่ชุดข้อมูลมีขนาด N จำนวนมาก และจำนวนคุณลักษณะน้อย (1,000,000 × 5) ให้ค่าความแม่นยำเฉลี่ยสูงมาก (> 95%) และให้ค่า AUC สูงมาก (> 0.90) ในทุกการแจกแจง โดยเฉพาะอย่างยิ่งกรณีชุดข้อมูลที่มีการแจกแจงปรกติ ให้ค่าความแม่นยำเฉลี่ยสูงถึง 97.4878% อีกทั้งยังให้ค่า AUC สูงถึง 0.9735 ในขณะที่วิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่ศึกษา ซึ่งมีประสิทธิภาพความแม่นยำเฉลี่ย 94.9744% และค่า AUC คือ 0.9659 ตามลำดับ กรณีที่ Feature มากขึ้น (7,000 × 75) ค่าความแม่นยำเฉลี่ยและค่า AUC ของการจำแนกประเภทจะน้อยลง [15] แต่ยังคงใกล้เคียงกับผลจากวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่ศึกษา หากพิจารณาตามการแจกแจงพบว่า ประสิทธิภาพความแม่นยำของวิธีการที่ศึกษานั้นจะสูงเมื่อข้อมูลมีการแจกแจงปรกติและการแจกแจงเอกรูป และจะน้อยลงเมื่อข้อมูลมีการแจกแจงแบบเลขชี้กำลัง โดยเฉพาะในกรณีที่ข้อมูลมีการแจกแจงแบบเลขชี้กำลังซึ่งมีความเบ้ ขนาด 7,000 × 75 จะให้ค่าความแม่นยำเฉลี่ยและ AUC เฉลี่ยที่น้อยกว่ากรณีอื่นๆ ทั้งนี้ สามารถเพิ่มค่าความแม่นยำของการจำแนกประเภทขึ้นได้ โดยการปรับค่าพารามิเตอร์ สัดส่วน km และ RT ดังที่แสดงในตารางที่ 5 ว่าเมื่อกำหนดสัดส่วนเป็น 0.40 กำหนด km=200 และ RT=100 ค่าความแม่นยำเฉลี่ยจะสูงถึง 90.3053% และค่า AUC เฉลี่ยสูงถึง 0.8612

สำหรับชุดข้อมูลจริงจากชุดข้อมูล Skin Segmentation โดยทำการสุ่มข้อมูล 0.40 หรือ 40% ของข้อมูลทั้งหมด (98,023 เคส) เมื่อทำการจำแนกประเภทด้วยวิธีที่นำเสนอจะได้ขนาดข้อมูลฝึก 90 เคส คิดเป็น 0.04% ของข้อมูลทั้งหมด สำหรับชุดข้อมูล Coil2000 ทำการสุ่มข้อมูล 0.10 หรือ 10% ของข้อมูลทั้งหมด (982 เคส) เมื่อทำการจำแนกประเภทด้วยวิธีที่นำเสนอจะได้ขนาดข้อมูลฝึก 67 เคส คิดเป็น 0.68% ของข้อมูลทั้งหมด จากผลการศึกษาในตารางที่ 4 พบว่าประสิทธิภาพความแม่นยำยังคงสูงกว่าวิธีการเรียนรู้เชิงลึกที่ใช้การสุ่มข้อมูลฝึกขนาดเท่ากับวิธีการที่ศึกษา

จากตารางที่ 5 วิธีการเรียนรู้เชิงลึกใช้ข้อมูลฝึก

มากถึง 90% หรือ 6,300 เคส ของขนาดข้อมูลทั้งหมด มีประสิทธิภาพความแม่นยำเฉลี่ยในการจำแนกประเภทสูงในทุกการแจกแจง (> 94%) โดยค่าความแม่นยำเฉลี่ยและค่า AUC ของชุดข้อมูลที่มีการแจกแจงแบบเลขชี้กำลังให้ค่าน้อยกว่าการแจกแจงปรกติและการแจกแจงเอกรูปเล็กน้อย ในส่วนของวิธีการที่ศึกษาเมื่อข้อมูลมีการแจกแจงปรกติใช้ข้อมูลฝึกเพียง 7.49% หรือ 524 เคส ให้ค่าความแม่นยำเฉลี่ยสูงถึง 95.0804% แต่ใช้เวลาเพียง 54.34 วินาที ซึ่งใช้เวลาน้อยกว่าวิธีการเรียนรู้เชิงลึกลึกเกือบ 4 เท่า สำหรับข้อมูลมีการแจกแจงเอกรูปและการแจกแจงแบบเลขชี้กำลังต้องใช้ข้อมูลฝึกมากขึ้นเพื่อให้ยังคงประสิทธิภาพในการจำแนกประเภท โดยใช้ข้อมูลฝึกประมาณ 20% ถึง 30% แต่ยังให้ค่าความแม่นยำเฉลี่ยสูง (> 90%) ซึ่งเกือบเท่ากับวิธีการเรียนรู้เชิงลึก โดยเวลาที่ใช้ในการประมวลผลนั้นยังน้อยกว่าวิธีการเรียนรู้เชิงลึกอย่างเห็นได้ชัด

จากการศึกษาการจำแนกข้อมูลขนาดใหญ่ เพื่อลดปัญหาของเวลาในการประมวลผลซึ่งต้องใช้เวลานานและต้องใช้ข้อมูลฝึกเป็นจำนวนมาก แต่ยังคงประสิทธิภาพความแม่นยำที่สูง จึงทำการลดขนาดข้อมูลฝึกด้วยการรวมเทคนิคการจัดกลุ่มของวิธีเคมีนและวิธีการเรียนรู้เชิงลึก จากผลการศึกษาพบว่า วิธีการที่ศึกษาสามารถลดขนาดข้อมูลฝึกได้อย่างมาก โดยเฉพาะอย่างยิ่งในกรณีของชุดข้อมูลที่มีขนาด N จำนวนมาก และจำนวนคุณลักษณะน้อย สามารถใช้ข้อมูลฝึกน้อยกว่า 1% ของจำนวนข้อมูลทั้งหมด แต่ยังคงให้ค่าความแม่นยำเฉลี่ยและค่า AUC สูงมาก เมื่อเปรียบเทียบกับชุดข้อมูลที่มีการแจกแจงปรกติและชุดข้อมูลที่มีการแจกแจงเอกรูป กรณีของวิธีการที่ศึกษานั้นมีประสิทธิภาพความแม่นยำเฉลี่ยที่สูงกว่าอีกทั้งยังให้ค่า AUC อยู่ในเกณฑ์ที่โมเดลทำงานได้ดีถึงดีมาก ในขณะที่ชุดข้อมูลที่มีการแจกแจงแบบเลขชี้กำลังมีประสิทธิภาพความแม่นยำเฉลี่ยที่เทียบเท่ากัน แต่ยังคงให้ค่า AUC ที่อยู่ในเกณฑ์มาตรฐานสำหรับโมเดลส่วนใหญ่ เมื่อพิจารณาผลที่ได้จากวิธีการที่ศึกษาเทียบกับวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึกมากถึง 90% ของข้อมูลทั้งหมด โดยประสิทธิภาพความแม่นยำของวิธีการที่ศึกษายังคงสูงอีกทั้งยังให้ค่า AUC อยู่ในเกณฑ์ที่โมเดลทำงานได้ดีถึงดีมาก รวมถึงใช้เวลาในการประมวลผลน้อยกว่าวิธีการเรียนรู้เชิงลึกที่ใช้ข้อมูลฝึกขนาด 80% และ 90% อย่างมาก

จากผลการศึกษาจะเห็นได้ว่าวิธีการที่นำเสนอเหมาะสำหรับกรณีที่ชุดข้อมูลมีขนาด N จำนวนมาก และจำนวนคุณลักษณะน้อย โดยเป็นชุดข้อมูลที่คุณลักษณะมีการแจกแจงปรกติรวมทั้งการแจกแจงเอกรูป หากชุดข้อมูลมีขนาด N จำนวนน้อย และจำนวนคุณลักษณะมาก ยังถือว่าอยู่ในเกณฑ์มาตรฐานสำหรับตัวแบบส่วนใหญ่ เว้นแต่ชุดข้อมูลที่มีการแจกแจงแบบเลขชี้กำลังโดยสามารถปรับพารามิเตอร์ km และ RT เพิ่มขึ้นได้ เพื่อประสิทธิภาพในการจำแนกประเภทที่ดีขึ้น และแน่นอนว่าจำนวนข้อมูลฝึกและเวลาในการประมวลผลจะเพิ่มมากขึ้นด้วยเช่นกัน ทั้งนี้ ขึ้นอยู่กับผู้ใช้ว่ายอมรับประสิทธิภาพรวมถึงจำนวนข้อมูลฝึกและเวลาที่เพิ่มมากขึ้นได้มากน้อยเพียงใด ในส่วนของเรื่องประสิทธิภาพในการจำแนกประเภทถึงแม้วิธีการที่นำเสนอจะมีประสิทธิภาพที่น้อยกว่าวิธีการเรียนรู้เชิงลึก แต่ถ้าหากพิจารณาในเรื่องของเวลาในการประมวลผล โดยเฉพาะชุดข้อมูลที่มีขนาด N จำนวนมาก และจำนวนคุณลักษณะน้อย อย่างชุดข้อมูลที่มีการแจกแจงปรกติ วิธีการที่นำเสนอก็เป็นทางเลือกที่น่าสนใจเพราะนอกจากลดระยะเวลาในการประมวลผลลงแล้วประสิทธิภาพในการจำแนกประเภทก็ยังสูงมากอีกด้วย

6. กิตติกรรมประกาศ

คณะผู้วิจัยขอขอบคุณภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยมหาสารคามที่ได้อนุเคราะห์คอมพิวเตอร์เพื่อใช้ในการวิเคราะห์ข้อมูล และสถานที่ในการทำวิจัยครั้งนี้

เอกสารอ้างอิง

[1] N. Suradet and W. Yathongkhum, “Supervised learning for demospongiae identification using graph mining technique,” *UTK Research Journal*, vol. 13, no. 1, pp. 167–179, 2019 (in Thai).

[2] T. Tang, S. Chen, M. Zhao, W. Huang, and J. Luo, “Very large-scale data classification based

on K-means clustering and multi-kernel SVM," *Soft Computing*, vol. 23, no. 11, pp. 3793–3801, 2018.

[3] K. Boonkiatpong and S. Sinthupinyo "Applying multiple neural networks on large scale data," M.S. thesis, Graduate School, Chulalongkorn University, 2011 (in Thai).

[4] K. Kowsrihawat, "A criminal case outcome and issue prediction model on Thai supreme court cases using deep learning techniques," M.S. thesis, Graduate School, Chulalongkorn University, 2018 (in Thai).

[5] W. Hirun and T. Pobutdee, "Trip attraction model using social network data and deep learning," *Sripatum Review of Science and Technology*, vol. 10, pp. 146–157, 2019 (in Thai).

[6] W. Boonpook, Y. Tan, Y. Ye, P. Torteeka, K. Torsri, and S. Dong, "A deep learning approach on road detection from unmanned aerial vehicle-based images in rural road monitoring," *Sensors*, vol. 18, no. 11, pp. 3921, 2018.

[7] N. Pholberdee and P. Taeprasartsiit, "Wound-region segmentation from image by using deep learning and various data augmentation methods," M.S. thesis, Graduate School, Silpakorn University, 2018 (in Thai).

[8] M. S. Kim, "Robust, scalable anomaly detection for large collections of images," presented at International Conference on Social Computing, Alexandria, VA, USA, September 8–14, 2013.

[9] T. Tang, S. Chen, M. Zhao, W. Huang, and J. Luo, "Very large-scale data classification based on K-means clustering and multi-kernel SVM," *Soft Computing*, vol. 23, no. 1, pp. 3793–3801, 2018.

[10] Y. Yoru and T. Hikmet Karakoc, "Application of artificial neural network (ANN) method to exergy analysis of thermodynamic systems," presented at International Conference on Machine Learning and Applications, Miami Beach, FL, USA, 2009.

[11] S. Nissen. (2003, October). Implementation of a Fast Artificial Neural Network. Department of Computer Science, University of Copenhagen. [Online]. Available: http://fann.sourceforge.net/report/

[12] D.W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. John Wiley & Sons, Inc., 2013, pp. 162.

[13] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[14] A. I. Marqués, V. García, and J. S. Sánchez, "On the suitability of resampling techniques for the class imbalance problem in credit scoring," *Journal of the Operational Research Society*, vol. 64, pp. 1060–1070, 2013.

[15] P. Wiriyathammabhum, "An approach to basis selection for dimensional reduction techniques," M.S. thesis, Graduate School, Chulalongkorn University, 2009 (in Thai).