การรวมกลุ่มจำแนกข้อมูลด้วยเทคนิคต้นไม้ช่วยตัดสินใจ เทคนิคโครงข่ายประสาทเทียมและเทคนิคซัพพอร์ตเว็กเตอร์แมชชีน

เดช ธรรมศิริ[1] และ พยุง มีสัจ[2]

[1]ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ กรุงเทพฯ

[2]ภาควิชาครุศาสตร์ไฟฟ้า คณะครุศาสตร์อุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ กรุงเทพฯ

Emails: dech@npru.ac.th , pym@kmutnb.ac.th

บทคัดย่อ

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อนำเสนอเสนอวิธีการจำแนกข้อมูลโดยนำเสนอแนวคิดในการหาโมเดลที่เหมาะสมเน้นให้มีประสิทธิภาพความแม่นยำที่สูงขึ้นในการจำแนกข้อมูล ทำการทดสอบประสิทธิภาพด้วยข้อมูล Austrian Credit และ German Credit โดยนำเทคนิคซัพพอร์ตเวกเตอร์แมชชีน เทคนิคต้นไม้ตัดสินใจ และเทคนิคโครงข่ายประสาทเทียม ร่วมกันตัดสินใจด้วยเทคนิคการรวมกลุ่ม โดยใช้วิธีการโหวตเสียงข้างมาก พบว่าเทคนิคการรวมกลุ่มช่วยกันตัดสินใจ ข้างต้นให้ผลลัพธ์ที่ดีกว่าการใช้เทคนิคแบบโมเดลเดี่ยว นอกจากนั้นยังพบอีกว่าการกำหนดจำนวนตัวจำแนกข้อมูลที่ใช้ในการรวมกลุ่มเพื่อตัดสินใจที่เหมาะสมจะทำให้ได้ผลลัพธ์ที่มีประสิทธิภาพยิ่งขึ้น โดยผลการวิจัยพบว่า สำหรับข้อมูล Austrian Credit การใช้ตัวจำแนกข้อมูล 24 ตัวให้ผลลัพธ์ความแม่นยำที่ดีที่สุดที่ 87.86% และ ข้อมูล German Credit การใช้ตัวจำแนกข้อมูล 6 ตัว และ 18 ตัวจะให้ผลลัพธ์ความแม่นยำที่ดีที่สุดที่ 71.00%

คำสำคัญ—เทคนิคซัพพอร์ตเวกเตอร์แมชชีน; เทคนิคต้นไม้ตัดสินใจ; เทคนิคการรวมกลุ่ม; เทคนิคโครงข่ายประสาทเทียม

1. บทนำ

ปัจจุบันนักวิจัยจำนวนมากได้ให้ความสำคัญกับปัญหาที่ต้องการวิธีการจำแนกข้อมูลที่แม่นยำ โดยมีการนำเสนอวิธีการที่หลากหลาย วิธีการที่นำเอาวิธีการทางด้านปัญญาประดิษฐ์ (Artificial Intelligence) มาประยุกต์ใช้ เช่น เทคนิคต้นไม้ช่วยตัดสินใจ (Decision Tree Models) [1] เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Networks หรือ ANN) [2] เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine หรือ SVM) [3]

แต่ถึงแม้ว่าเทคนิคข้างต้นจะให้ผลลัพธ์ในการจำแนกข้อมูลที่แม่นยำ แต่การเป็นโมเดลเดียว (Single Model) นั้นทำให้มีการกำหนดกลุ่มของข้อมูลที่ใช้ในการเรียนรู้ รวมทั้งมีการกำหนดค่าพารามิเตอร์ที่ตายตัว บางครั้งเกิดปัญหาความโน้มเอียง (Bias) ทำให้ได้ประสิทธิภาพที่ไม่ดีนัก หนทางหนึ่งที่จะสามารถลดค่าความโน้มเอียงได้คือการใช้วิธีการเรียนรู้แบบรวมกลุ่ม (Ensemble) ซึ่งวิธีการเรียนรู้แบบรวมกลุ่มนั้น สามารถสร้างความหลากหลายและลดค่าความผิดพลาดที่เกิดจากความแปรปรวนได้ แนวความคิดหลักของวิธีการเรียนรู้แบบรวมกลุ่มคือการรวมเอากลุ่มของตัวจำแนกข้อมูล เพื่อแก้ปัญหาเดียวกันและผลที่ได้จะมีความแม่นยำและความเที่ยงตรงมากกว่าการใช้โมเดลแบบเดี่ยว ดังเช่นงานวิจัยของนาย Yao และคณะ[4] ได้นำเสนอการรวมกลุ่มช่วยในการตัดสินใจโดยใช้เทคนิคต้นไม้ช่วยตัดสินใจ ร่วมกับวิธีการ Boosting โดยได้ทดสอบประสิทธิภาพกับข้อมูลจาก UCI ผลที่ได้แสดงให้เห็นว่าวิธีการดังกล่าวให้ประสิทธิภาพที่ดีมากกว่าโมเดลแบบเดี่ยว นาย Jian และคณะ [5] ได้ใช้เทคนิคการรวมกลุ่มตัดสินใจร่วมกับเทคนิคโครงข่ายประสาทเทียม เข้ามาใช้ในการแก้ปัญหาการวิเคราะห์ตัวเลขที่เขียนด้วยลายมือ ซึ่งผู้วิจัยพบว่ามันมีปัญหาทั้งในเรื่องของขนาด ความกว้าง ข้อมูลรบกวนและตำแหน่งที่แตกต่างกัน ผู้วิจัยได้นำเสนอการใช้ เทคนิคโครงข่ายประสาทเทียมแบบ Hopfield Neural Network เพื่อใช้ในการกำจัดข้อมูลรบกวน และใช้เทคนิค Bagging ทำการรวมกลุ่มตัวจำแนกข้อมูลที่สร้างขึ้นมาจากโครงข่ายประสาทเทียมแบบ Multilayer Feed Forward ผลที่ได้เมื่อเปรียบเทียบกับโมเดลแบบเดี่ยวทั้งในส่วนข้อมูลสอนและข้อมูลทดสอบพบว่า วิธีการนำเสนอให้ผลลัพธ์ที่แม่นยำกว่า นาย Chen และคณะ [6] ได้นำเสนอการใช้ตัวจำแนกข้อมูลด้วยเทคนิคซัพพอร์ตเว็กเตอร์แมชชีนที่มีการกำหนดค่า เคอร์แนลฟังก์ชัน รวมทั้งค่าพารามิเตอร์ที่หลากหลาย สำหรับสร้างตัวจำแนกข้อมูล ผลที่ได้มีประสิทธิภาพที่ดีกว่าการใช้โมเดลเทคนิคซัพพอร์ตเว็กเตอร์แมชชีนแบบเดี่ยว จะเห็นได้ว่าเทคนิคการรวมกลุ่มสามารถช่วยเพิ่มประสิทธิภาพให้กับโมเดล หากแต่วิธีการรวมกลุ่มถ้าจะให้มีประสิทธิภาพที่ดี ได้มีผู้วิจัยดังเช่น นาย Wang และคณะ [7] ให้คำแนะนำว่า *"โดยทั่วไปแล้วประสิทธิภาพของการรวมกลุ่มจะขึ้นอยู่กับ ความหลากหลาย และความแม่นยำของตัวแทนในการจำแนกข้อมูล"*

ดังนั้นสำหรับงานวิจัยในครั้งนี้ ผู้วิจัยได้ให้ความสำคัญด้านการแก้ปัญหาต่างๆ เหล่านี้ เพื่อทำให้การเรียนรู้แบบรวมกลุ่มมีประสิทธิภาพสูงขึ้นโดย การเลือกพารามิเตอร์ที่เหมาะสมสำหรับเพิ่มประสิทธิภาพสำหรับตัวจำแนกข้อมูล อีกทั้งการใช้ตัวจำแนกข้อมูลที่หลากหลายร่วมกันเรียนรู้ สำหรับนำไปสร้างโมเดลใหม่ที่มีประสิทธิภาพ เพื่อนำไปใช้สำหรับการจำแนกข้อมูล

2. ทฤษฎีที่เกี่ยวข้อง

2.1. เทคนิคซัพพอร์ตเวกเตอร์แมชชีน

เป็นเครื่องมือสำหรับการเรียนรู้แบบใหม่บนพื้นฐานของการเรียนรู้ทฤษฎีทางสถิติได้รับการนำเสนอครั้งแรกโดย Vapnik [8] โดยมีแนวคิดของทฤษฎีดังนี้

กำหนดข้อมูลทดลอง $D = \{x_i, y_i; i = 1, 2, ..., n\}$ ในขณะที่ $x_i = (x_{i1}, x_{i2}, ..., x_{in}) \in R^n$ นั้นเป็นข้อมูลนำเข้า และ $y_i \in \{+1, -1\}$ (+1="ข้อมูล Class 1", -1="ข้อมูล Class 0") ซึ่งเป็นการกำหนดกลุ่มเป้าหมาย ให้ SVM โดยที่ SVM นั้นมุ่งเป้าเพื่อหาฟังก์ชันการตัดสินใจที่สามารถแยกแยะค่าที่ไม่ทราบได้โดยใช้สมการที่ (1)

$$y = sign\left\{\sum_{j=1}^{n} w_j \phi_j(x) + b\right\} \quad (1)$$

$$\phi(x) = [\phi_1(x), \phi_2(x), ..., \phi_n(x)]^T \quad (2)$$

จากตัวอย่างสมการที่ (2) เป็นกลุ่มข้อมูล x ซึ่งไม่สามารถแบ่งแยกได้ด้วยสมการเส้นตรงให้อยู่ในรูปแบบที่สามารถใช้สมการเส้นตรงแบ่งแยกได้ โดยที่ ϕ() แทนฟังก์ชันสำหรับแปลงข้อมูลที่ไม่เป็นเชิงเส้นให้เป็นข้อมูลที่อยู่ในรูปแบบที่สมการเชิงเส้นสามารถแยกแยะได้ w_j แทนค่าน้ำหนัก (Weighting) ที่เชื่อมโยงจาก Feature Space ไปสู่ Output Space และ b นั้นเป็นค่าโน้มเอียง (Bias หรือ Threshold) ที่ตั้งไว้ทำได้ให้สมการที่ (3) สำหรับการจำแนกข้อมูล

$$f(x) = \sum_{j=1}^{n} w_j \phi_j(x) + b \quad (3)$$

หากแต่บางครั้งไม่สามารถแยกแยะข้อมูลได้ถูกต้องทั้งหมด ทำให้ต้องมีการกำหนดตัวแปรเพื่อยอมรับค่าความผิดพลาด โดยทำการเพิ่มตัวแปร ξ (Slack Variable) โดยเขียนแสดงรายละเอียดได้ดังสมการที่ (4) และ (5)

$$w^T \cdot x + b \geq +1 - \xi_i \ ; \ y_i = +1 \quad (4)$$

$$w^T \cdot x + b \leq -1 + \xi_i \ ; \ y_i = -1 \quad (5)$$

โดยที่ $\xi_i > 0$

ทำให้ได้โครงสร้าง ของ SVM ซึ่งประกอบด้วย 2 ส่วนหลัก คือ การเพิ่มระยะแยกแยะมากสุด และ การแก้ปัญหาด้วยการลดข้อผิดพลาดให้ต่ำที่สุด ดังแสดงในสมการที่ (6)

$$\underset{w,b,\xi}{Minimize} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i \quad (6)$$

โดยที่ : $y_i(w^T \phi(x_i) + b) + \xi_i - 1 \geq 0$

$\xi_i \geq 0, i = 1, 2, ..., N$

และมีเคอร์แนลฟังก์ชั่น (Kernel Function) ที่พบได้บ่อย 3 แบบด้วยกัน ดังแสดงในสมการที่ (7) (8) และ (9) ได้แก่

Polynomial Kernel:

$$k(x_i, x_j) = (1 + x_i \cdot x_j)^d \quad (7)$$

Radial Basis Function Kernel :

$$k(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) \quad (8)$$

Sigmoid Kernel :

$$k(x_i, x_j) = \tanh(kx_i \cdot x_j) - \delta) \quad (9)$$

2.2. เทคนิคโครงข่ายประสาทเทียม

มีพื้นฐานมาจากการจำลองการทำงานของสมองมนุษย์ ด้วยโปรแกรมคอมพิวเตอร์ จุดมุงหมายของโครงข่ายประสาทเทียมคือต้องการให้คอมพิวเตอร์มีความชาญฉลาดในการเรียนรู้เหมือนที่ มนุษย์มีการเรียนรู้สามารถฝึกฝนได้ และสามารถนำความรู้และทักษะ รวมทั้งสามารถนำไปประยุกต์ใช้ได้ดีกับปัญหา Classification, Regression และ Clustering เทคนิคนี้ มักถูกเรียกว่า "Black Box" เนื่องจากการทำงานมีความ ซับซ้อนมากกว่าเทคนิคอื่นๆ ค่อนข้างมาก

การเรียนรู้ของนิวรอลเน็ตเวิร์ก ทำได้โดยการส่งข้อมูลเข้ามายังส่วนที่เรียกว่าเพอร์เซ็ปตรอน (Perceptron) สามารถเทียบได้กับเซ็ลสมองของมนุษย์ โดยที่เพอร์เซ็ปตรอนทำการรับข้อมูลที่อยู่ในรูปของแมททริกซ์ซึ่งเป็นตัวเลข เข้ามาคำนวณ สำหรับฟังก์ชันผลรวม (Summation Function) มีการทำงานดังสมการที่ (10)

$$n = \sum_{i=1}^{z} x_i w_i + b \quad (10)$$

โดยที่ ตัวแปร n คือ ผลรวมที่ได้จากฟังก์ชันผลรวม

ตัวแปร x_i คือ ค่าข้อมูลเข้าตัวที่ i

ตัวแปร w_i คือ ค่าน้ำหนักของนิวรอนตัวที่ i

ตัวแปร z คือ จำนวนนิวรอนชั้นข้อมูลเข้า

ตัวแปร b คือ ค่าความโน้มเอียง

ตัวแปร i มีค่าตั้งแต่ 1 ถึง z

ขั้นตอนการเรียนรู้แบบเพอร์เซ็ปตรอนมีขั้นตอนดังนี้ [9]

1) เริ่มจากการสุ่มค่าน้ำหนัก w_i
2) เทียบเพอร์เซ็ปตรอนกับข้อมูลที่นำมาสอน ทำการปรับแก้ไขน้ำหนัก เมื่อเพอร์เซ็ปตรอนแยกตัวอย่างผิดพลาด
3) ทำการวนซ้ำข้อมูลที่ใช้สำหรับการสอน จนกระทั่งเพอร์เซ็ปตรอน แยกตัวได้อย่างถูกต้องทั้งหมด
4) ในการแก้ไขน้ำหนัก น้ำหนักถูกปรับตาม

$Wi \longleftarrow Wi + \Delta Wi$

โดย $\Delta wi = \alpha \times (t - a) \times x_i = \alpha \times error \times input$

เมื่อ t เป็นผลลัพธ์ที่ถูกต้อง a เป็นผลลัพธ์ที่ได้จาก เพอร์เซ็ปตรอน และ α เป็นค่าที่แสดงอัตราการเรียนรู้

2.3. เทคนิคต้นไม้ช่วยตัดสินใจ

เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปของ ต้นไม้ช่วยตัดสินใจ ซึ่งต้นไม้ช่วยตัดสินใจนั้นทำงานแบบ Supervised Learning คือ สามารถสร้างแบบจำลองการจัดหมวดหมู่ได้จาก กลุ่มตัวอย่างของข้อมูลที่ได้กำหนดได้ก่อนล่วงหน้า ที่เรียกว่า Training Set ได้อัตโนมัติ และสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดหมวดหมู่ได้ด้วย

วิธีการที่ใช้สร้างต้นไม้ช่วยตัดสินใจมีขั้นตอนดังนี้

1) หา Attribute ที่สำคัญที่สุดมาแบ่งข้อมูลโดย Attribute นี้ถูกนำมาสร้างเป็น Root Node โดยมี Target Attribute เป็นผลลัพธ์ซึ่งเป็น Leaf Node ถูกกำหนดไว้ก่อน
2) นำค่าที่เป็นไปได้ใน Attribute ที่ถูกเลือกมาแตกออกเป็นกลุ่มของตัวเอง
3) แบ่งข้อมูลทั้งหมดตามกลุ่มที่แตกออกจาก Root Node
4) วนกลับไปทำที่ขั้นตอนแรก คือ หา Attribute ที่สำคัญที่สุดจากข้อมูลที่เข้ามาเพื่อหาตัวแบ่งต่อไป

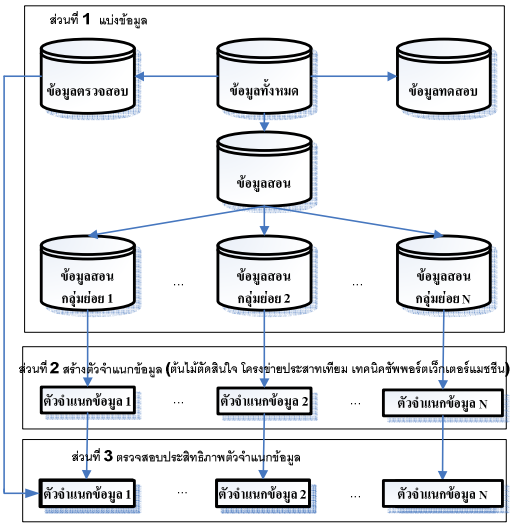
2.4. เทคนิคการเรียนรู้แบบรวมกลุ่ม

เป็นเทคนิคในการรวมเอากลุ่มของตัวจำแนกข้อมูลที่หลากหลาย เพื่อแก้ปัญหาเดียวกัน ด้วยวิธีการการนำเอาตัวจำแนกข้อมูลเหล่านั้นมาร่วมกันตัดสินใจ สำหรับงานวิจัยฉบับนี้เลือกใช้ใช้วิธีการ โหวตเสียงข้างมาก (Majority Vote) ซึ่งวิธีการนี้ง่ายต่อการใช้งาน มีวิธีการคือ กำหนดค่า f_k โดยที่ (k=1,2,…,K) ซึ่ง K คือ จำนวนของตัวจำแนกข้อมูลและ $f_k(x)=c, c\in\{+1,-1\}$ โดยที่ฟังก์ชั่นในการตัดสินใจจะตัดสินใจผลลัพธ์ที่ได้ว่าอยู่ในคลาส +1 หรือ คลาส -1 สุดท้ายนำค่าฟังก์ชัน $f_e(x)$ ที่ได้จากตัวจำแนกข้อมูลมาตัดสินใจร่วมกัน โดยดูที่เสียงข้างมากว่าตัดสินใจเป็นคลาส +1 หรือ คลาส -1 โดยมีสมการดังสมการที่ (11)

$$f_e(x)=\arg\max_{c}\sum_{k:f_k(x)=c}1 \qquad (11)$$

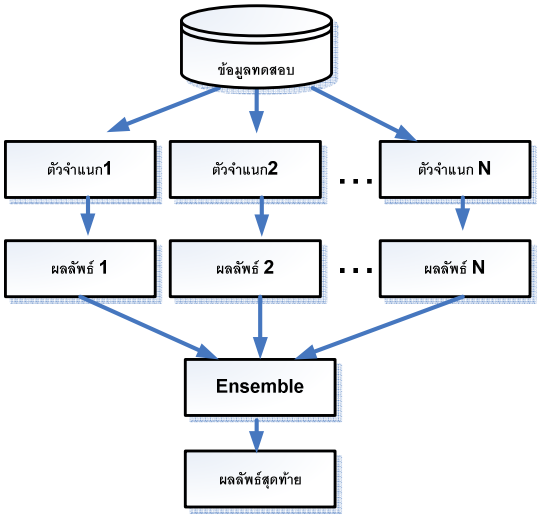
3. วิธีการและโมเดลที่นำเสนอ

ในงานวิจัยนี้นำเสนอวิธีการสร้างตัวจำแนกข้อมูลที่หลากหลายโดยใช้วิธีการสุ่มกลุ่มตัวอย่าง เพื่อสร้างความแตกต่างบน กลุ่มข้อมูลสำหรับการฝึกฝน (Training Sets) ซึ่งประกอบด้วย กลุ่มตัวอย่างจำนวน N ตามจำนวนของตัวจำแนกข้อมูลที่ได้กำหนดขึ้น โดยที่ข้อมูลถูกสุ่มและสร้างมาจากกลุ่มข้อมูลสำหรับการสอน โดยที่กลุ่มข้อมูลที่สุ่มสร้างขึ้นมาถูกเรียกว่า Bootstrap สำหรับการสุ่มจำนวนกลุ่มตัวอย่าง จากกลุ่มข้อมูลสำหรับการสอน มีค่าความน่าจะเป็นที่ $1-(1-\frac{1}{m})^m$ [10] โดยที่ m คือจำนวนของตัวอย่างทั้งหมด ที่อยู่ในกลุ่มข้อมูลสำหรับการฝึกฝน จะถูกนำไปสอนให้กับตัวจำแนกข้อมูล ซึ่งได้แก่การใช้เทคนิคต้นไม้ช่วยตัดสินใจ เทคนิคโครงข่ายประสาทเทียมและเทคนิคซัพพอร์ตเว็กเตอร์แมชชีน สำหรับเทคนิคโครงข่ายประสาทเทียมผู้วิจัยได้กำหนดเลือกใช้โครงสร้างที่แตกต่างกัน 6 โครงสร้าง และสำหรับเทคนิคซัพพอร์ตเว็กเตอร์แมชชีน ผู้วิจัยได้เลือกใช้เคอร์แนล Radial Basis Function เนื่องมาจากผู้วิจัยได้ทำการทดลองแล้วให้ผลที่ดีกับข้อมูลที่ผู้วิจัยได้นำมาใช้ สำหรับการกำหนดค่าพารามิเตอร์ให้กับโมเดลผู้วิจัยเลือกใช้วิธีการค้นหาแบบกริช (Grid Search) เพื่อช่วยในการกำหนดค่าพารามิเตอร์ที่เหมาะสมโดยกำหนดค่า C ซึ่งเป็นค่าถ่วงน้ำหนักสำหรับการให้ความสำคัญระหว่างระยะแยกแยะและค่าความผิดพลาดไว้ที่ [1,3,5,10,15,20,25,30] ส่วนค่าแกรมม่าเป็นพารามิเตอร์สำหรับเคอร์แนลฟังก์ชัน RBF ที่เลือกใช้ได้กำหนดค่าไว้ที่ [0.0001,0.001,0.01,0.1,0.5,1] โดยมีขั้นตอนการสร้างตัวจำแนกข้อมูลดังรูปที่ 1



รูปที่ 1. ขั้นตอนการสร้างตัวจำแนกข้อมูล

จากนั้นนำตัวจำแนกข้อมูลที่สร้างขึ้นในขั้นตอนก่อนหน้า มารวมกลุ่มร่วมกันตัดสินใจ โดยนำข้อมูลทดสอบมาใช้ในการประเมินประสิทธิภาพจากนั้น ใช้วิธีการโหวตเสียงข้างมากรวมผลลัพธ์ที่ได้เพื่อนำมาตัดสินใจเป็นผลลัพธ์สุดท้าย ดังรูปที่ 2



รูปที่ 2. โมเดลการเรียนรู้แบบรวมกลุ่ม (Ensemble)

การเปรียบเทียบผลการทดลองผู้วิจัยได้ทำการเปรียบเทียบกับวิธีการที่เป็นแบบโมเดลเดี่ยวกับโมเดลที่ใช้เทคนิคการรวมกลุ่มอีกทั้งมีการกำหนดจำนวนตัวจำแนกข้อมูลที่นำมารวมกลุ่มที่แตกต่างกันได้แก่ 3,6,9,12,15,18,21,24,27,30 ตัว โดยแต่ละกลุ่มที่นำมารวมกันจะประกอบด้วย เทคนิคต้นไม้ช่วยตัดสินใจ เทคนิคโครงข่ายประสาทเทียม และ เทคนิคซัพพอร์ตเว็กเตอร์แมชชีน ในจำนวนที่เท่าๆกัน ตัวอย่างเช่น ถ้าใช้ตัวจำแนก 12 ตัวจะประกอบไปด้วย ตัวจำแนกที่ถูกสร้างด้วยเทคนิคต้นไม้ช่วยตัดสินใจ 4 ตัว ตัวจำแนกที่ถูกสร้างด้วยเทคนิคโครงข่ายประสาทเทียม 4 ตัว และ ตัวจำแนกที่ถูกสร้างด้วยเทคนิคซัพพอร์ตเว็กเตอร์แมชชีน 4 ตัว เช่นเดียวกัน เพราะต้องการเพิ่มประสิทธิภาพโดยให้มีโมเดลที่หลากหลาย

ข้อมูลและเครื่องมือที่ใช้ในการวิจัย สำหรับการวัดประสิทธิภาพของโมเดล งานวิจัยครั้งนี้ได้ใช้ข้อมูลจาก UCI ซึ่งเป็นข้อมูลเปิดสำหรับนำไปทดสอบประสิทธิภาพอัลกอริธึมทางด้าน AI ได้แก่ฐานข้อมูลAustrian Credit [11] และGerman Credit [12] รายละเอียดดังตารางที่ 1

ตารางที่ 1. รายละเอียดข้อมูลที่นำมาทดสอบประสิทธิภาพ

	ตัวอย่าง	แอตทริบิว	คลาส +1	คลาส -1
Austrian Credit	690	14	307	383
German Credit	1000	24	300	700

จากข้อมูลข้างต้นผู้วิจัยได้ทำการแบ่งกลุ่มข้อมูลออกเป็น กลุ่มข้อมูลสำหรับสอน (Training Set) กลุ่มข้อมูลสำหรับตรวจสอบ (Validate Set) และกลุ่มข้อมูลสำหรับทดสอบ (Test Set) โดยสำหรับกลุ่มข้อมูล Austrian Credit แบ่งข้อมูลสำหรับสอนไว้ที่ 450 เร็คคอร์ด แบ่งข้อมูลสำหรับตรวจสอบไว้ที่ 67 เร็คคอร์ด และ ข้อมูลสำหรับทดสอบไว้ที่ 173 เร็คคอร์ด ส่วนกลุ่มข้อมูล German Credit แบ่งข้อมูลสำหรับสอนไว้ที่ 700 เร็คคอร์ด แบ่งข้อมูลสำหรับตรวจสอบไว้ที่ 100 เร็คคอร์ด และ ข้อมูลสำหรับทดสอบไว้ที่ 200 เร็คคอร์ด การประเมินผลใช้การประเมินผลจากประสิทธิภาพการจำแนกข้อมูล โดยวัดค่าความถูกต้องแม่นยำ (Accuracy) ในการจำแนกข้อมูลในกลุ่มข้อมูลทดสอบ ดังสมการที่ 12

$$Accuracy = \left(\frac{TP + TN}{TP + FN + TN + FP} \right) 100 \qquad (12)$$

โดยที่

TP คือค่า True Positive
TN คือค่า True Negative
FP คือค่า False Positive
FN คือค่า False Negative

4. ผลการวิจัยและการอภิปรายผล

ผลการทดลองที่ได้จากงานวิจัยสำหรับฐานข้อมูล Austrian Credit พบว่า โมเดลแบบเดี่ยว (Single Model) ที่ทดลอง ทั้งสิ้น 30 โมเดลมีความแม่นยำที่สูงที่สุด คือ 82.08% และ ต่ำสุด 41.61% ส่วนค่าเฉลี่ยความแม่นยำอยู่ที่ 71.75% ดังตารางที่ 2

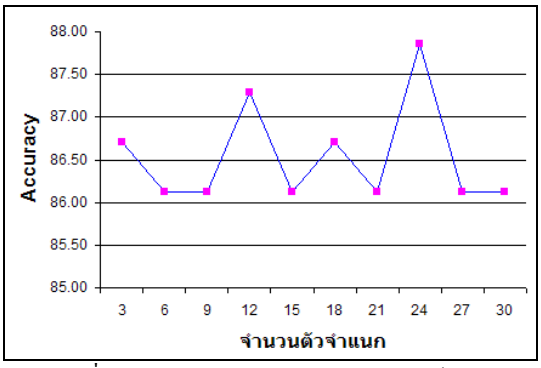
ตาราง 2. ผลการทดสอบกับข้อมูล Austrian Credit

ข้อมูล Austrian Credit (ผลทดสอบแบบโมเดลเดี่ยว)	
ค่า	ความแม่นยำ(%)
ต่ำสุด (Min)	41.61
สูงสุด (Max)	82.08
เฉลี่ย (Average)	71.75
การกระจายตัว (Std)	15.22

ส่วนผลลัพธ์ที่ได้จากการใช้เทคนิครวมกลุ่มกันด้วยวิธีการโหวตเสียงข้างมาก โดยการเลือกโมเดลแบบเดี่ยวที่สร้างขึ้นมา นำมารวมกันช่วยในการตัดสินใจ พบว่าให้ผลความแม่นยำสูงที่สุดที่ 87.86% โดยเป็นการรวมกลุ่มโดยใช้ตัวจำแนกข้อมูลทั้งสิ้น 24 ตัว ดังตารางที่ 3 และ รูปที่ 3.

ตาราง 3. ผลการทดสอบกับข้อมูล Austrian Credit

ข้อมูล Austrian Credit (ผลทดสอบแบบหลายโมเดลร่วมกัน)	
จำนวนตัวจำแนก	ความแม่นยำ (%)
3	86.71
6	86.13
9	86.13
12	87.28
15	86.13
18	86.71
21	86.13
24	87.86
27	86.13
30	86.13
เฉลี่ย (Average)	86.53



รูปที่ 3. กราฟแสดงประสิทธิภาพการจำแนกข้อมูล

ผลการทดลองที่ได้จากงานวิจัยสำหรับฐานข้อมูล German Credit พบว่า โมเดลแบบเดี่ยว (Single Model) ที่ทดลอง ทั้งสิ้น 30 โมเดลมีความ

แม่นยำที่สูงที่สุด คือ 70.00% และ ต่ำสุด 30.50% ส่วนค่าเฉลี่ยความแม่นยำอยู่ที่ 65.76% ดังตารางที่ 4

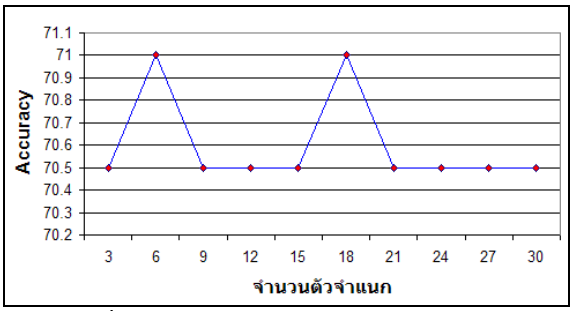
ตาราง 4. ผลการทดสอบกับข้อมูล German Credit

ข้อมูล German Credit (ผลทดสอบแบบโมเดลเดี่ยว)	
ค่า	ความแม่นยำ(%)
ต่ำสุด (Min)	30.50
สูงสุด (Max)	70.00
เฉลี่ย (Average)	65.76
การกระจายตัว (Std)	11.95

ส่วนผลลัพธ์ที่ได้จากการใช้เทคนิครวมกลุ่มกันด้วยวิธีการโหวตเสียงข้างมาก โดยการเลือกโมเดลแบบเดี่ยวที่สร้างขึ้นมา นำมารวมกันช่วยในการตัดสินใจ พบว่าให้ผลความแม่นยำสูงที่สุดที่ 71.00% โดยเป็นการรวมกลุ่มโดยใช้ตัวจำแนกข้อมูลทั้งสิ้น 6 ตัว และ 18 ตัว ดังตารางที่ 5 และ รูปที่ 4.

ตาราง 5. ผลการทดสอบกับข้อมูล German Credit

ข้อมูล German Credit (ผลทดสอบแบบหลายโมเดลร่วมกัน)	
จำนวนตัวจำแนก	ความแม่นยำ (%)
3	70.50
6	71.00
9	70.50
12	70.50
15	70.50
18	71.00
21	70.50
24	70.50
27	70.50
30	70.50
เฉลี่ย (Average)	70.60



รูปที่ 4. กราฟแสดงประสิทธิภาพการจำแนกข้อมูล

จากผลลัพธ์ที่ได้จากการทดลองจะพบว่า แนวทางในการนำตัวจำแนกข้อมูลที่หลากหลายมาร่วมกันตัดสินใจสามารถช่วยเพิ่มประสิทธิภาพความแม่นยำได้มากกว่าการใช้วิธีการแบบโมเดลเดี่ยว (Single Model) อันเนื่องมาจากการใช้เทคนิคการรวมกลุ่มสามารถลดการเกิดปัญหาความโน้มเอียง (Bias) ซึ่งปัญหานี้มักเกิดขึ้นจากการกำหนดกลุ่มของข้อมูลที่ใช้ในการเรียนรู้ รวมทั้งมีการกำหนดค่าพารามิเตอร์ที่ตายตัว อีกทั้งในการทดลองยังพบว่า จำนวนตัวจำแนกที่มากไม่ทำให้โมเดลที่ได้มีประสิทธิภาพเพิ่มขึ้นเสมอไป หากแต่การกำหนดจำนวนตัวจำแนกที่เหมาะสมจะช่วยทำให้โมเดลที่ได้มีประสิทธิภาพที่ดีที่สุด

5. สรุปผลการวิจัย

การวิจัยในครั้งนี้ มีวัตถุประสงค์เพื่อสร้างโมเดลที่มีประสิทธิภาพ สำหรับการจำแนกข้อมูล ผลที่ได้แสดงให้เห็นว่าการนำเอาเทคนิคซัพพอร์ตเวกเตอร์แมชชีน เทคนิคต้นไม้ตัดสินใจ และเทคนิคโครงข่ายประสาทเทียม ร่วมกันตัดสินใจด้วยเทคนิคการรวมกลุ่ม โดยใช้วิธีการโหวตเสียงข้างมาก พบว่าเทคนิคการรวมกลุ่มช่วยกันตัดสินใจ ข้างต้นให้ผลลัพธ์ที่ดีกว่าการใช้เทคนิคแบบโมเดลเดี่ยว นอกจากนั้นยังพบอีกว่าการกำหนดจำนวนตัวจำแนกข้อมูลที่ใช้ในการรวมกลุ่มเพื่อตัดสินใจที่เหมาะสมจะทำให้ได้ผลลัพธ์ที่มีประสิทธิภาพยิ่งขึ้น ทั้งนี้เพราะเมื่อนำตัวจำแนกข้อมูลที่หลากหลายมารวมกันด้วยวิธีการรวมกลุ่มของตัวจำแนกเป็นการลดปัญหาการเกิดไบแอสของข้อมูลและตัวจำแนกที่ดีแต่ละตัวช่วยกันเสริมประสิทธิภาพในการจำแนกข้อมูล ทำให้โมเดลที่ได้มีประสิทธิภาพมากยิ่งขึ้น หากแต่ปัญหาที่พบในงานวิจัยนี้คือ การกำหนดค่าเพื่อหาจำนวนของตัวจำแนกที่เหมาะสมต้องใช้การทดลองหลายครั้งเพื่อหาคำตอบ

ในอนาคตผู้วิจัยมีแนวคิดในการนำเอาวิธีการอื่นมาประยุกต์ใช้เพื่อให้เกิดความแม่นยำในการจำแนกข้อมูลที่เพิ่มขึ้น รวมทั้งลดขั้นตอนในการกำหนดค่าเพื่อหาจำนวนของตัวจำแนกที่อีกทั้งการเลือกเฉพาะตัวจำแนกที่เหมาะสมมาใช้งาน

เอกสารอ้างอิง

[1] Karine Zeitouni and Nadjim Chelghoum, "Spatial Decision Tree-Application to Traffic Risk Analysis," aiccsa, pp.0203, ACS/IEEE International Conference on Computer Systems and Applications (AICCSA'01), 2001.

[2] เดช ธรรมศิริ ณรงค์ โพธิ วาทินี นุ้ยเพียร ภัทราวุฒิ แสงศิริ ภรัณยา อำมฤครัตน์ และพยุง มีสัจ, “การให้คะแนนสินเชื่อโดยวิธีการทำเหมืองข้อมูลด้วยเทคนิคโครงข่ายประสาทเทียม แบบแพร่กระจายย้อนกลับ”, NCCIT2008, 2551.

[3] เดช ธรรมศิริ และ พยุง มีสัจ, “การจำแนกข้อมูลโดยวิธีการรวมกลุ่มของเทคนิคซัพพอร์ตเวกเตอร์แมชชีนโดยการปรับพารามิเตอร์ด้วยขั้นตอนวิธีเชิงพันธุกรรมรวมทั้งการเลือกใช้ลักษณะที่เหมาะสมด้วยวิธีการสัมประสิทธิ์สหสัมพันธ์”, NCCIT2010, 2553.

[4] Yao Yu, Fu Zhong-liang, Zhao Xiang-hui and Cheng Wen-fang, "Combining Classifier Based on Decision Tree," icie, vol. 2, pp.37-40, 2009 WASE International Conference on Information Engineering, 2009.

[5] Jian Wang, Jingfeng Yang, Shaofa Li, Qiufang Dai and Jiaxing Xie, "Number Image Recognition Based on Neural Network Ensemble," icnc, vol. 1, pp.237-240, Third International Conference on Natural Computation (ICNC 2007), 2007.

[6] S. Chen, W. Wang, and H. van Zuylen, "Construct support vector machine ensemble to detect traffic incident," *Expert Systems with Applications,* vol. 36, pp. 10976-10986, 2009.

[7] Wang Shi-jin, Mathew Avin, Chen Yan,Xi Li-feng, Ma Lin and Lee Jay, "Empirical analysis of support vector machine ensemble classifiers," *Expert Systems with Applications,* vol. 36, pp. 6466-6476, 2009.

[8] Vapnik, V, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.

[9] บุณเสริม กิจศิริกุล, "อัลกอลิทึมการทำเหมืองข้อมูล," ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวะกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย 2546.

[10] Breiman, L. “Bagging predictors”. Machine Learning, 24(2),: 123–140, 1996.

[11] Australian Credit Approval Data set. Retrieved February,10, 2008, From http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/australian/.

[12] German Credit Data set. Retrieved February,10, 2008, From http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/.