การเปรียบเทียบวิธีการจัดกลุ่มกรณีข้อมูลมีการแจกแจงเสถียรปกติ การแจกแจงโคชี และการแจกแจงเลวี่

THE COMPARISON OF CLASSIFICATION WITH STABLE-NORMAL CAUCHY AND LE'VY DISTRIBUTIONS

ณัฏฐินี ดีแท้ *สุภาพร คลังเพ็ชร์*

Natthinee Deetae, Supaporn Klungpet*

คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏพิบูลสงคราม

Faculty of science and Technology, Pibulsongkram Rajabhat University.

***Corresponding author, E-mail**: *Natthineed@gmail.com*

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพวิธีการจัดกลุ่ม 2 วิธีคือวิธีซัพพอร์ตเวกเตอร์แมทชีน (Support Vector Machine) และวิธีเคเนียเรสเนเบอร์ (K-Nearest Neighbor) กรณีข้อมูลมีการแจกแจงแบบเสถียร ซึ่งมีหลายการแจกแจงย่อย คือ การแจกแจงเสถียรปกติ (Stable-Normal Distribution) การแจกแจงโคชี (Cauchy Distribution) และการแจกแจงเลวี่ (Le'vy Distribution) โดยข้อมูลแบ่งออกเป็น 2 ชุดคือชุดข้อมูลเรียนรู้ (Training Set) และชุดข้อมูลทดสอบ (Test Set) ที่มีขนาดตัวอย่างเป็น 100 300 และ 500 โดยจัดข้อมูลออกเป็น 2 กลุ่ม ซึ่งกำหนดอัตราส่วนระหว่างชุดข้อมูลเรียนรู้ : ชุดข้อมูลทดสอบ มีค่าเป็น 90:10, 80:20, 70:30, 60:40 และ 50:50 ทำการจำลองข้อมูลโดยเทคนิคมอนติคาร์โลและกระทำซ้ำ 1000 ครั้ง ในแต่ละสถานการณ์ที่กำหนด และใช้ค่าเฉลี่ยของเปอร์เซ็นต์การจัดกลุ่มถูกต้องเป็นเกณฑ์ในการเปรียบเทียบ ผลการวิจัยพบว่าวิธีซัพพอร์ตเวกเตอร์แมทชีนสามารถจัดกลุ่มข้อมูลได้ดีกว่าวิธีเคเนียเรสเนเบอร์ในทุกสถานการณ์ที่กำหนด

คำสำคัญ: ซัพพอร์ตเวกเตอร์แมทชีน เคเนียเรสเนเบอร์ การแจกแจงเสถียรปกติ การแจกแจงโคชี การแจกแจงเลวี่

Abstract

The purpose of this research aims to compare the efficiency of support vector machine (SVM) and K-nearest neighbor (K-NN) when data are distributed as Stable-normal, Cauchy and Le'vy distributions. Data employed in this research were generated into two sets, consisting of training set and test set with the sample sizes 100, 300 and 500 for the binary classification. The ratios of training sets: test sets are 90:10, 80:20, 70:30, 60:40 and 50:50. In each situation, the data are simulated with Monte Carlo technique and repeated 1,000 times. The average percentage of correct classification is used as criterion for comparison. Form the results, found that the SVM method produces higher percentages of correct classification than K-NN method in all situations under study.

Keywords: Support Vector Machine, K-nearest Neighbor, Stable-Normal Distribution, Cauchy Distribution, Le'vy Distribution

บทนำ

ในปัจจุบันการศึกษาวิจัยด้านต่างๆ เช่น ด้านวิทยาศาสตร์ ด้านการแพทย์ ด้านการศึกษา ด้านเศรษฐกิจ ด้านสังคมศาสตร์ และด้านสิ่งแวดล้อม ฯลฯ จำเป็นต้องอาศัยความรู้ทางด้านสถิติมาใช้ประโยชน์ โดยเฉพาะในกระบวนการวิเคราะห์ข้อมูล ซึ่งในทางปฏิบัติข้อมูลที่นำมาวิเคราะห์มักจะเกี่ยวข้องกับตัวแปรหลายตัวแปร ดังนั้นจึงต้องอาศัยการวิเคราะห์ตัวแปรเชิงพหุ (Multivariate Analysis) ซึ่งเป็นวิธีทางสถิติที่นิยมใช้กันอยู่ทั่วไปมาวิเคราะห์ข้อมูล เพื่อให้ข้อมูลที่ได้มีความถูกต้องและน่าเชื่อถือ โดยข้อมูลที่พบในชีวิตประจำวันส่วนมากจะเป็นข้อมูลที่มีการแจกแจงไม่ปกติ มักมีการแจกแจงลักษณะอื่นๆ หรือข้อมูลที่พบอาจมีการแจกแจงเบ้ขวา (Right Skewed Distribution) หรือเบ้ซ้าย (Left Skewed Distribution) หรือข้อมูลอาจมีการแจกแจงแบบเสถียร (Stable Distribution) โดยการแจกแจงแบบเสถียรสามารถนำไปประยุกต์ใช้ในหลายสาขา [1] อาทิ ทฤษฎีการสื่อสาร ฟิสิกส์ ชีววิทยา ดาราศาสตร์ การเงิน เศรษฐศาสตร์ และสังคมวิทยา เป็นต้น ซึ่งข้อมูลลักษณะนี้ต้องอาศัยวิธีทางนอนพาราเมตริกมาวิเคราะห์ เพราะวิธีทางนอนพาราเมตริกมีเงื่อนไขการใช้ภายใต้ข้อตกลงเบื้องต้นเพียงไม่กี่ข้อ และที่สำคัญไม่จำเป็นต้องทราบรูปแบบการแจกแจงของประชากร โดยสามารถใช้ได้กับข้อมูลที่มีระดับการวัดต่ำ ตั้งแต่มาตรานามบัญญัติที่สามารถนับเป็นความถี่ได้ มาตราเรียงอันดับหรือตัวเลขใดๆ ที่สามารถนำมาจัดอันดับที่ได้ เช่น วิธีเคเนียเรสเนเบอร์ (K-Nearest Neighbor: KNN) วิธีซัพพอร์ตเวกเตอร์แมทชีน (Support Vector Machine: SVM) วิธีกฎริปเปอร์ (Ripper Rules) วิธีนิวรอลเน็ตเวิร์ก (Neural Networks) และวิธีเนอีฟเบย์ (Naïve Bayes) เป็นต้น

จากการทบทวนงานวิจัยพบว่า สุรศักดิ์ พรรณรักษา [2] ได้ศึกษาการเปรียบเทียบวิธีการจัดกลุ่มกรณีข้อมูลมีการแจกแจงแบบเสถียรที่มีลักษณะหางหนา โดยวิธีการจัดกลุ่มแบบฟิชเชอร์และวิธีเคเนียเรสเนเบอร์ กรณีที่ข้อมูลมีการแจกแจงแบบปกติเชิงพหุและแบบเสถียรที่มีลักษณะหางหนา โดยกำหนดขนาดตัวอย่างที่ใช้ในการศึกษาเป็น 100 300 และ 500 และอัตราส่วนของข้อมูลระหว่าง Training Data : Test data เป็น 90:10, 80:20, 70:30, 60:40 และ 50:50 ผลการวิจัยพบว่าเมื่อขนาดตัวอย่างและ Training Data เพิ่มขึ้น อัตราการจัดกลุ่มข้อมูลผิดพลาดของทั้ง 2 วิธีมีแนวโน้มลดลง โดยกรณีที่ข้อมูลมีการแจกแจงแบบปกติเชิงพหุ พบว่าวิธีการจัดกลุ่มแบบฟิชเชอร์ให้อัตราการจัดกลุ่มข้อมูลผิดพลาดต่ำกว่าวิธีเคเนียเรสเนเบอร์ และกรณีที่ข้อมูลมีการแจกแจงแบบเสถียรที่มีลักษณะหางหนา พบว่าวิธีเคเนียเรสเนเบอร์ให้อัตราการจัดกลุ่มข้อมูลผิดพลาดต่ำกว่าวิธีการจัดกลุ่มแบบฟิชเชอร์ในทุกกรณี นอกจากนี้ วาทินี นุ้ยเพียร และพยุง มีสัจ [3] ได้เปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะแบบการกรองและการควบรวมของการทำเหมืองข้อความเพื่อการจำแนกข้อความคัดเลือกคุณลักษณะการกรองแบบไคสแควร์ ใช้วิธีการจำแนกประเภทแบบวิธีซัพพอร์ตเวกเตอร์แมทชีน โดยใช้เคอร์เนลแบบโพลิโนเมียลและเรเดียลเบสิสฟังก์ชัน เนอีฟเบย์ เบย์เซี่ยนเน็ตเวิร์ก และเคเนียเรสเนเบอร์ ผลการวิจัยพบว่า วิธีซัพพอร์ตเวกเตอร์แมทชีนโดยใช้เคอร์เนลเรเดียลเบสิสฟังก์ชันให้ผลการวัดประสิทธิภาพโดยรวมสูงที่สุดคือ 92.2% และเคอร์เนลแบบโพลิโนเมียล 86.5% เนอีฟเบย์ 91.7% เบย์เซี่ยนเน็ตเวิร์ก 91.4% และเคเนียเรสเนเบอร์ 88.7% ตามลำดับ และนิเวศ จิระวิชิตชัย และคณะ (2555) [4] ได้นำเสนอวิธีการพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทย โดยนำเสนอแบบจำลองการจัดหมวดหมู่ของเอกสารภาษาไทยแบบอัตโนมัติ ทดสอบประสิทธิภาพแบบจำลองการจัดหมวดหมู่เอกสาร

ภาษาไทยกับอัลกอริทึมต้นไม้ตัดสินใจ ซัพพอร์ทเวกเตอร์แมชชีน เนอีฟเบย์ เครือข่ายฟังก์ชันฐานรัศมี เคเนียเรสเนเบอร์ และกฎริปเปอร์ โดยใช้วิธีการลดคุณลักษณะร่วมกับอัลกอริทึมเครื่องจักรการเรียนรู้เพื่อศึกษาวิธีการลดคุณลักษณะที่เหมาะสมและมีประสิทธิภาพในการจัดหมู่เอกสารข่าวภาษาไทย จากการวิจัยพบว่าการลดคุณลักษณะด้วยวิธีการเพิ่มของข้อมูล (Information Gain) เพื่อลดมิติของข้อมูลแล้วส่งเข้าเครื่องจักรการเรียนรู้ และวัดประสิทธิภาพจากการลดคุณลักษณะที่ทำให้ค่าวัดจากผลรวมของค่าเฉลี่ย (F-Measurement) สูงสุด สามารถสรุปได้ว่า อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีนให้ประสิทธิภาพสูงสุด คือ 94.3% รองลงมาเป็นอัลกอริทึมเนอีฟเบย์ให้ประสิทธิภาพสูงสุด คือ 86.2% อัลกอริทึมเครือข่ายฟังก์ชันฐานรัศมีให้ประสิทธิภาพสูงสุด คือ 86.1% อัลกอริทึมต้นไม้ตัดสินใจให้ประสิทธิภาพสูงสุด คือ 79.7% อัลกอริทึมกฎริปเปอร์ให้ประสิทธิภาพสูงสุด คือ 78.9% อัลกอริทึมเคเนียเรสเนเบอร์ให้ประสิทธิภาพสูงสุด คือ 69.5% ตามลำดับ

จากการที่ผู้วิจัยได้ทบทวนงานวิจัยที่เกี่ยวข้องและพบว่าการแจกแจงเสถียรเป็นอีกหนึ่งการแจกแจงที่น่าสนใจ เนื่องจากการแจกแจงเสถียรมีหลายการแจกแจงย่อย (Sub Class Distribution) ทำให้เกิดข้อมูลที่มีการแจกแจงที่หลากหลาย ดังนั้นผู้วิจัยจึงสนใจศึกษาวิธีการจัดกลุ่มเมื่อข้อมูลมีการแจกแจงเสถียรปกติ (Stable-Normal Distribution) การแจกแจงโคชี (Cauchy Distribution) และการแจกแจงเลวี่ (Le'vy Distribution) ซึ่งเป็นการแจกแจงย่อยของการแจกแจงเสถียร โดยเปรียบเทียบการจัดกลุ่มด้วยวิธีซัพพอร์ตเวกเตอร์แมทชีน และวิธีเคเนียเรสเนเบอร์ ซึ่งเป็นวิธีทางนอนพาราเมตริกเพื่อช่วยแก้ไขปัญหาการจัดกลุ่มข้อมูลทางสถิติได้อย่างถูกต้องชัดเจน ให้เกิดความคลาดเคลื่อนน้อยที่สุด ซึ่งจะส่งผลให้การจัดกลุ่มมีประสิทธิภาพมากยิ่งขึ้น โดยงานวิจัยนี้แบ่งข้อมูลออกเป็น 2 ชุด คือ ชุดของหน่วยตัวอย่างที่ใช้ในการสร้างเกณฑ์การจำแนก เรียกว่า ชุดข้อมูลเรียนรู้ (Training Sets) และชุดของหน่วยตัวอย่างที่ใช้ในการทดสอบเกณฑ์การจำแนกที่สร้างขึ้น เรียกว่า ชุดข้อมูลทดสอบ (Test Sets) และเมื่อสร้างเกณฑ์การจำแนกจากชุดข้อมูลเรียนรู้แล้ว จึงนำเกณฑ์ที่ได้ไปใช้ในการจัดกลุ่มของค่าสังเกตค่าใหม่ที่อยู่ในชุดข้อมูลทดสอบว่าควรจัดอยู่ในกลุ่มใด โดยใช้ค่าเฉลี่ยของเปอร์เซ็นต์การจัดกลุ่มถูกต้องเป็นเกณฑ์ในการเปรียบเทียบ

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาประสิทธิภาพการจัดกลุ่มของข้อมูลที่มีการแจกแจงแบบเสถียร โดยวิธีซัพพอร์ตเวกเตอร์แมทชีนและวิธีเคเนียเรสเนเบอร์
2. เพื่อเปรียบเทียบประสิทธิภาพการจัดกลุ่มของข้อมูลที่มีการแจกแจงแบบเสถียร โดยวิธีซัพพอร์ตเวกเตอร์แมทชีนและวิธีเคเนียเรสเนเบอร์

วิธีดำเนินการวิจัย

1. สร้างข้อมูลในการวิจัย

การแจกแจงแบบเสถียรมีหลายการแจกแจงย่อย ซึ่งงานวิจัยนี้สนใจที่จะศึกษาการแจกแจงแบบเสถียรปกติ การแจกแจงโคชี และการแจกแจงเลวี่ ซึ่งเหตุผลหลักในการเลือกการแจกแจงแบบเสถียรเพราะเป็นการแจกแจงที่ถูกสนับสนุนโดยทฤษฎีบทลิมิตสู่ส่วนกลางวางนัยทั่วไป (Generalized Central Limit Theorem) [5] โดยการแจกแจงแบบเสถียรประกอบด้วยพารามิเตอร์ 4 ตัว คือ α แทนรูปแบบของการแจกแจง, β แทนความเบ้, δ แทนลักษณะตำแหน่ง, γ แทนขนาดสัดส่วน ซึ่งการแจกแจงแบบเสถียรเป็นการใช้ความรู้พื้นฐานของการแจกแจงความน่าจะเป็นและมีฟังก์ชันลักษณะเฉพาะ ในการแจกแจงแบบเสถียรอาจมี

ลักษณะเบ้ (skewness) ข้อมูลอาจมีการแจกแจงเบ้ขวา (right skewed distribution) หรือเบ้ซ้าย (left skewed distribution) หรืออาจมีลักษณะโด่ง (kurtosis) และมีหางที่หนา (heavy-tailed) การแจกแจงแบบเสถียรเมื่อมีการปรับค่าพารามิเตอร์จะทำให้รูปแบบการแจกแจงเปลี่ยนไปเป็นการแจกแจงลักษณะอื่นๆ ได้ เช่น

การแจกแจงเสถียรปกติ มี 2 พารามิเตอร์ ประกอบด้วย $\alpha = 2$ และ $\beta = 0$ โดยมีฟังก์ชันความหนาแน่นความน่าจะเป็น คือ

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \quad ;-\infty < x < \infty$$

การแจกแจงโคชี มี 2 พารามิเตอร์ ประกอบด้วย $\alpha = 1$ และ $\beta = 0$ โดยมีฟังก์ชันความหนาแน่นความน่าจะเป็น คือ

$$f(x) = \frac{1}{\pi} \frac{\gamma}{\gamma^2 + (x-\delta)^2} \qquad ;-\infty < x < \infty$$

การแจกแจงเลวี่ มี 2 พารามิเตอร์ ประกอบด้วย $\alpha = 0.5$ และ $\beta = 1$ โดยมีฟังก์ชันความหนาแน่นความน่าจะเป็น คือ

$$f(x) = \sqrt{\frac{\gamma}{2\pi}} \frac{1}{(x-\delta)^{\frac{3}{2}}} \exp\left(-\frac{\gamma}{2(x-\delta)} \right); \delta < x < \infty$$

2. ซัพพอร์ตเวคเตอร์แมชชีน

ซัพพอร์ตเวคเตอร์แมชชีนเป็นเทคนิคหนึ่งที่ได้รับความนิยมอย่างแพร่หลายในงานที่เกี่ยวข้องกับการจดจำรูปแบบตลอดจนการแก้ปัญหาการจัดกลุ่ม [6] โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกกลุ่มข้อมูลได้ดีที่สุด (optimal separating hyperplane) เมื่อเราพิจารณาข้อมูลที่ประกอบด้วยข้อมูล 2 กลุ่มดังสมการที่ 1

$$D = \{(\mathbf{x}_i, y_i) \; ; i = 1,2,...,n\} \qquad .....(1)$$

เมื่อ $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{im}) \in R^m$

$y_i \in \{1,-1\}$ โดย 1 คือ ข้อมูลกลุ่ม 1 และ -1 คือ ข้อมูลกลุ่ม 2

ซึ่งเป็นการกำหนดกลุ่มเป้าหมายให้ SVM โดยที่ SVM นั้นมุ่งเป้าเพื่อหาฟังก์ชันการตัดสินใจที่สามารถแบ่งแยกค่าที่ไม่ทราบได้ดังสมการที่ 2

$$f(\mathbf{x}) = sign\left\{ \sum_{k=1}^{n_v} w_k \varphi_k(\mathbf{x}) \varphi_k(\mathbf{x}_k) + b \right\} \quad .....(2)$$

$$\varphi(\mathbf{x}) = [\varphi_1(\mathbf{x}_1), \varphi_2(\mathbf{x}_2), ..., \varphi_n(\mathbf{x}_{n_v})]^T \quad .....(3)$$

กลุ่มข้อมูล $\mathbf{x}$ จากสมการที่ 3 ไม่สามารถแบ่งแยกได้ด้วยสมการเส้นตรงแต่จะถูกแปลงให้อยู่ในรูปแบบที่สามารถใช้สมการเส้นตรงแบ่งแยกได้ โดยใช้เคอร์เนลฟังก์ชัน (kernel function)

$$K(\mathbf{x},\mathbf{x}_k)=\varphi(\mathbf{x})\varphi(\mathbf{x}_k) \qquad .....(4)$$

เมื่อ $\varphi(\mathbf{x})$ แทน ฟังก์ชันสำหรับแปลงข้อมูลที่ไม่เป็นเชิงเส้นให้เป็นข้อมูลที่อยู่ในรูปเชิงเส้นสามารถแบ่งแยกได้

w_k แทน ค่าน้ำหนักที่เชื่อมโยงจาก feature space ไปสู่ output space
b แทน ค่าโน้มเอียง (bias)
$\mathbf{x}_k$ แทน ซัพพอร์ตเวกเตอร์ โดย
n_v แทน จำนวนซัพพอร์ตเวกเตอร์

วิธีการที่ใช้ในการหาเส้นแบ่งที่ดีที่สุดคือการเพิ่มเส้นขอบ (margin) ให้กับเส้นแบ่งทั้งสองข้างและสร้างเส้นขอบที่สัมผัสกับค่าข้อมูลใน feature space ที่ใกล้ที่สุดดังนั้นเส้นแบ่งที่มีเส้นขอบกว้างที่สุดจึงเป็นเส้นแบ่งที่ดีที่สุดและเรียกตำแหน่งการสัมผัสข้อมูลที่ใกล้ที่สุดจากการเพิ่มขอบนี้ว่า “ซัพพอร์ตเวกเตอร์” (support vector) เนื่องจากในบางกรณีการแบ่งแยกกลุ่มไม่สามารถทำได้ถูกต้องโดยสมบูรณ์

ดังนั้นจึงต้องมีการกำหนดตัวแปรสำหรับยอมรับค่าความผิดพลาดโดยการเพิ่มตัวแปร (slack variable) ดังสมการที่ 5 และ 6 ดังนี้

$$w^T\mathbf{x}+b\geq 1-\xi_i \qquad .....(5)$$
$$w^T\mathbf{x}+b\leq -1+\xi_i \qquad .....(6)$$

จากการกำหนดค่า $\xi_i > 0$ ทำให้โครงสร้างของซัพพอร์ตเวกเตอร์แมทชีนบรรลุวัตถุประสงค์ใน 2 ส่วนคือการเพิ่มระยะแบ่งแยกให้มากที่สุดและลดข้อผิดพลาดในการทำนายให้ต่ำที่สุด ดังสมการที่ 7

$$Minimize\ \frac{1}{2}\|\mathbf{w}\|^2+c\sum_{i=1}^{N}\xi_i \qquad .....(7)$$

โดยที่ : $y_i(\mathbf{w}^T\varphi(\mathbf{x})+b)+\xi_i-1\geq 0$
$\xi_i\geq 0, i=1,2,...,N$

และมีเคอร์เนลฟังก์ชัน ที่นิยมใช้อยู่ 3 ชนิดด้วยกันคือ
โพลิโนเมียล (polynomial) :

$$K(\mathbf{x}_i,\mathbf{x}_j)=(\mathbf{x}_i^T\mathbf{x}_j+r)^\gamma \ ;\gamma>0$$

โดยที่ r,γ คือพารามิเตอร์ของฟังก์ชันโพลิโนเมียล
เรเดียลเบสิสฟังก์ชัน (Radial Basis Function-RBF) :

$$K(\mathbf{x}_i,\mathbf{x}_j)=\exp(-\gamma\|\mathbf{x}_i-\mathbf{x}_j\|^2) \ ;\gamma>0$$

โดยที่ γ คือพารามิเตอร์ของฟังก์ชันเรเดียลเบสิสฟังก์ชันซิกมอยด์ (sigmoid) :

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma \mathbf{x}_i^T \mathbf{x}_j - r)$$

โดยที่ r, γ คือพารามิเตอร์ของฟังก์ชันซิกมอยด์

3. วิธีเคเนียเรสเนเบอร์

หลักการของวิธีการนี้จะจำแนกประเภทข้อมูลโดยขึ้นกับข้อมูลที่มีคุณสมบัติใกล้เคียงที่สุด k ตัว ซึ่งเป็นวิธีการจัดกลุ่มให้กับค่าสังเกตค่าใหม่ที่ใกล้กับชุดข้อมูลเรียนรู้ [7] โดยกำหนดให้ชุดข้อมูลเรียนรู้ ดังสมการที่ 8

$$T = \left\{ \left(\mathbf{x}_i, \mathbf{y}_j \right) \right\} \quad ; \; i = 1,2,3,...,n \; ; j = 1,2,3,...,t \qquad .....(8)$$

โดยที่ $\mathbf{x}_i$ เป็นตัวแปรอิสระและ $\mathbf{y}_j$ เป็นตัวแปรกลุ่ม ดังนั้นข้อมูลตัวแปรอิสระ $\mathbf{x}$ เขียนแทนได้ดังสมการที่ 9

$$\left\langle a_1(\mathbf{x}), a_2(\mathbf{x}), ..., a_m(\mathbf{x}) \right\rangle \quad .....(9)$$

โดยที่ $a_b(\mathbf{x})$ แทนค่าของ b^{th} ของตัวแปรอิสระ $\mathbf{x}$; $b = 1,2,...,m$

โดยมีคุณสมบัติของฟังก์ชันระยะห่างส่วนใหญ่พบว่าฟังก์ชันระยะห่างที่ใช้บ่อยเป็นฟังก์ชันระยะห่างแบบยูคลิเดียน (euclidean distance) ซึ่งเป็นฟังก์ชันที่นิยมใช้ในการหาค่าระยะห่างที่แท้จริงดังสมการที่ 10

$$d_{euclidean} = (\mathbf{u}, \mathbf{v}) = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2} \qquad ....(10)$$

โดยที่ $\mathbf{u} = \{u_1, u_2, ..., u_n\}$

$\mathbf{v} = \{v_1, v_2, ..., v_n\}$

แทนการบันทึกค่าคุณลักษณะของ m ทั้งสองกลุ่มดังนั้น ระยะห่างระหว่าง $\mathbf{x}_0$ ค่าสังเกตใหม่และ $\mathbf{x}_i$ ถูกกำหนดโดย

$$d(\mathbf{x}_0, \mathbf{x}_i) = \sqrt{\sum_{b=1}^{m} \left[a_b(\mathbf{x}_0) - a_b(\mathbf{x}_i) \right]^2} \quad .....(11)$$

โดยค่าสังเกตค่าใหม่ในชุดข้อมูลทดสอบจะจัดเข้ากลุ่มที่มีค่า $\hat{p}_q = (\mathbf{x}'_\ell \mid \mathbf{x}_i)$ สูงสุด ซึ่งถูกกำหนดโดย

$$\hat{p}_q = (\mathbf{x}'_\ell \mid \mathbf{x}_i) = (1/k) \sum_{i \sim \ell}^{k} \delta_{(q)(\mathbf{x}'_\ell)} \quad .....(12)$$

เมื่อ $\sum_{i \sim \ell}^{k} \delta_{(q)(\mathbf{x}'_\ell)}$ แทน จำนวน $\mathbf{x}_i$ ของกลุ่ม ภายในบริเวณใกล้เคียง k

$\delta_{(q)(\mathbf{x}'_\ell)}$ แทน ฟังก์ชันดิแรก (dirac function) ซึ่งถูกกำหนดโดย

$$\delta_{(q)(\mathbf{x}'_\ell)} = \begin{cases} 1, & q = \mathbf{x}'_\ell \\ 0, & otherwise. \end{cases}$$

4. เกณฑ์การตัดสินใจ

เปอร์เซ็นต์การจัดกลุ่มถูกต้อง จะพิจารณาประสิทธิภาพของการจัดกลุ่มโดยคำนวณจากเปอร์เซ็นต์การจัดกลุ่มถูกต้อง ดังนี้

$$\text{เปอร์เซ็นต์การจัดกลุ่มถูกต้อง} = \frac{Cr}{Tstotal} \times 100$$

เมื่อ Cr แทน จำนวนข้อมูลที่มีการจัดกลุ่มถูกต้องในชุดข้อมูลทดสอบ

$Tstotal$ แทน จำนวนข้อมูลทั้งหมดของชุดข้อมูลทดสอบ

ผลการวิจัย

ตารางที่ 1 เปอร์เซ็นต์การจัดกลุ่มถูกต้องด้วยวิธีซัพพอร์ตเวกเตอร์แมทชีน

การแจกแจงของข้อมูล	ขนาดตัวอย่าง (n)	อัตราส่วนระหว่าง Training Data : Test Data				
		50:50	60:40	70:30	80:20	90:10
การแจกแจงเสถียรปกติ	100	92.45	92.60	92.70	92.75	92.76
	300	92.90	92.87	92.96	92.83	93.25
	500	93.16	93.17	93.20	93.26	93.26
การแจกแจงโคชี	100	93.17	93.75	94.29	94.18	94.61
	300	94.61	94.84	94.95	94.97	95.00
	500	95.16	95.17	95.26	95.27	95.32
การแจกแจงเลวี่	100	89.50	89.59	90.11	89.67	89.62
	300	88.04	88.76	89.21	89.41	89.90
	500	90.01	90.63	91.19	91.57	91.84

เมื่อ ช่องทึบ แทน เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงสุดของแต่ละการแจกแจงและขนาดตัวอย่าง

จากตารางที่ 1 เป็นการแสดงผลการจัดกลุ่มด้วยวิธีซัพพอร์ตเวกเตอร์แมทชีน ซึ่งมีการแจกแจงของข้อมูลทั้ง 3 การแจกแจง คือ การแจกแจงเสถียรปกติ การแจกแจงโคชี และการแจกแจงเลวี่ จะเห็นได้ว่า ทั้ง 3 การแจกแจง เมื่อกำหนดขนาดตัวอย่างคงที่ พบว่า เปอร์เซ็นต์การจัดกลุ่มถูกต้องมีแนวโน้มเพิ่มขึ้น เมื่ออัตราส่วนของข้อมูลใน Training Data เพิ่มมากขึ้น และเมื่อกำหนดอัตราส่วนระหว่าง Training Data : Test Data คงที่ พบว่าเปอร์เซ็นต์การจัดกลุ่มถูกต้องมีแนวโน้มเพิ่มขึ้นเมื่อขนาดตัวอย่างเพิ่มมากขึ้น โดยการแจกแจงเสถียรปกติ พบว่า เมื่อขนาดตัวอย่างเป็น 100 และ 300 เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงที่สุด คือ 92.76% และ 93.25% ตามลำดับ เมื่อใช้ Training Data : Test Data เป็น 90:10 และเมื่อขนาดตัวอย่างเป็น 500 เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงที่สุด คือ 93.26% เมื่อใช้ Training Data : Test Data เป็น 80:20 และ 90:10 ส่วนการแจกแจงโคชี พบว่า เมื่อขนาดตัวอย่างเป็น 100 300 และ 500 เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงที่สุด คือ 94.61% 95.00% และ 95.32% ตามลำดับ เมื่อใช้ Training Data : Test Data เป็น 90:10 และการแจกแจงเลวี่ พบว่า เมื่อขนาดตัวอย่างเป็น 100 เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงที่สุด คือ 90.11% เมื่อใช้ Training Data : Test Data เป็น 70:30 และเมื่อขนาดตัวอย่างเป็น 300 และ 500 เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงที่สุด คือ 89.90% และ 91.84% ตามลำดับ เมื่อใช้ Training Data : Test Data เป็น 90:10

ตารางที่ 2 เปอร์เซ็นต์การจัดกลุ่มถูกต้องด้วยวิธีเคเนียเรสเนเบอร์

การแจกแจงของข้อมูล	ขนาดตัวอย่าง (n)	k	อัตราส่วนระหว่าง Training Data : Test Data				
			50:50	60:40	70:30	80:20	90:10
การแจกแจงเสถียรปกติ	100	3	90.04	90.21	90.31	90.49	90.50
		5	90.57	90.78	90.93	91.04	91.11
		7	90.75	91.06	91.24	91.35	91.36
	300	3	90.52	90.66	90.76	90.80	90.94
		5	91.08	91.13	91.29	91.41	91.54
		7	91.41	91.46	91.47	91.72	91.66
	500	3	90.88	90.91	91.10	91.16	91.06
		5	91.35	91.36	91.42	91.54	91.40
		7	91.63	91.63	91.67	91.67	91.74
การแจกแจงโคชี	100	3	91.46	91.58	91.70	91.54	91.62
		5	93.38	93.67	93.78	93.91	93.91
		7	93.70	94.02	94.29	94.13	94.41
	300	3	93.89	94.03	94.02	94.11	93.63
		5	94.04	94.05	94.14	94.15	94.20
		7	94.41	94.54	94.57	94.57	94.67
	500	3	94.04	94.05	94.14	94.15	94.20
		5	94.16	94.17	94.34	94.37	94.49
		7	94.63	94.67	94.83	94.87	94.92
การแจกแจงเลวี่	100	3	85.74	86.41	87.21	87.79	87.78
		5	84.93	85.96	86.67	87.19	87.57
		7	84.07	85.16	85.77	86.31	86.73
	300	3	87.79	88.55	89.03	89.31	89.73
		5	87.19	88.12	88.41	88.82	89.30
		7	88.79	88.08	88.15	88.32	88.35
	500	3	90.53	90.83	90.86	90.92	90.97
		5	90.46	90.75	90.98	91.18	91.20
		7	90.68	91.07	91.44	91.66	91.67

เมื่อ ช่องทึบ แทน เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงสุดของแต่ละการแจกแจงและขนาดตัวอย่าง

จากตารางที่ 2 เป็นการแสดงผลการจัดกลุ่มด้วยวิธีเคเนียเรสเนเบอร์ ซึ่งมีการแจกแจงของข้อมูลทั้ง 3 การแจกแจง คือ การแจกแจงเสถียรปกติ การแจกแจงโคชี และการแจกแจงเลวี่ จะเห็นได้ว่า ทั้ง 3 การแจกแจง เมื่อกำหนดขนาดตัวอย่างและค่า k คงที่ พบว่าเปอร์เซ็นต์การจัดกลุ่มถูกต้องมีแนวโน้มเพิ่มขึ้นเมื่ออัตราส่วนของข้อมูลใน Training Data เพิ่มมากขึ้น เมื่อกำหนดอัตราส่วนระหว่าง Training Data : Test Data และค่า k คงที่ พบว่าเปอร์เซ็นต์การจัดกลุ่มถูกต้องมีแนวโน้มเพิ่มขึ้นเมื่อขนาดตัวอย่างเพิ่มมากขึ้น เมื่อกำหนดขนาดตัวอย่างและอัตราส่วนระหว่าง Training Data : Test Data คงที่ พบว่าเปอร์เซ็นต์การจัดกลุ่มถูกต้องมีแนวโน้มเพิ่มขึ้นเมื่อ k เพิ่มมากขึ้น โดยการแจกแจงเสถียรปกติ พบว่า เมื่อขนาดตัวอย่างเท่ากับ 100 เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงที่สุด คือ 91.36% เมื่อใช้ Training Data : Test Data เป็น 90:10 ที่ k=7 เมื่อขนาดตัวอย่างเท่ากับ 300 เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงที่สุด คือ 91.72% เมื่อใช้ Training Data : Test Data เป็น 80:20 ที่ k=7 เมื่อขนาดตัวอย่างเท่ากับ 500 เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงที่สุด คือ 91.74% เมื่อใช้ Training Data : Test Data เป็น 90:10 ที่ k=7 ส่วนการแจกแจงโคชี พบว่า เมื่อขนาดตัวอย่างเป็น 100 300 และ 500 เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงที่สุด คือ 94.41% 94.67% และ 94.92% ตามลำดับ เมื่อใช้ Training Data : Test Data เป็น 90:10 ที่ k=7 และการแจกแจงเลวี่ พบว่า เมื่อขนาดตัวอย่างเป็น 100 เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงที่สุด คือ 87.79% เมื่อใช้ Training Data : Test Data เป็น 80:20 ที่ k=3 เมื่อขนาดตัวอย่างเท่ากับ 300 เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงที่สุด คือ 89.73% เมื่อใช้ Training Data : Test Data เป็น 90:10 ที่ k=3 เมื่อขนาดตัวอย่างเท่ากับ 500 เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงที่สุด คือ 91.67% เมื่อใช้ Training Data : Test Data เป็น 90:10 ที่ k=7

ตารางที่ 3 เปอร์เซ็นต์การจัดกลุ่มถูกต้องด้วยวิธีซัพพอร์ตเวกเตอร์แมทชีนและวิธีเคเนียเรสเนเบอร์ เมื่อข้อมูลมีการแจกแจงเสถียรปกติ

วิธีการจัดกลุ่ม	ขนาดตัวอย่าง		
	100	300	500
SVM	92.76	93.25	93.26
KNN	91.36	91.72	91.74

เมื่อ ช่องทึบ แทน เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงสุดของแต่ละขนาดตัวอย่าง

จากตารางที่ 3 จะเห็นได้ว่าเมื่อขนาดตัวอย่างเพิ่มมากขึ้น พบว่าเปอร์เซ็นต์การจัดกลุ่มถูกต้องของวิธีซัพพอร์ตเวกเตอร์แมทชีนและวิธีเคเนียเรสเนเบอร์เมื่อข้อมูลมีการแจกแจงเสถียรปกติจะมีแนวโน้มเพิ่มขึ้น และเมื่อพิจารณาแต่ละขนาดตัวอย่างของทั้ง 2 วิธี พบว่าวิธีซัพพอร์ตเวกเตอร์แมทชีนสามารถจัดกลุ่มข้อมูลได้ดีกว่าวิธีเคเนียเรสเนเบอร์ในทุกขนาดตัวอย่าง

ตารางที่ 4 เปอร์เซ็นต์การจัดกลุ่มถูกต้องด้วยวิธีซัพพอร์ตเวกเตอร์แมทชีนและวิธีเคเนียเรสเนเบอร์ เมื่อข้อมูลมีการแจกแจงโคชี

วิธีการจัดกลุ่ม	ขนาดตัวอย่าง		
	100	300	500
SVM	94.61	95.00	95.32
KNN	94.41	94.67	94.92

เมื่อ ช่องทึบ แทน เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงสุดของแต่ละขนาดตัวอย่าง

จากตารางที่ 4 จะเห็นได้ว่าเมื่อขนาดตัวอย่างเพิ่มมากขึ้น พบว่าเปอร์เซ็นต์การจัดกลุ่มถูกต้องของวิธีซัพพอร์ตเวกเตอร์แมทชีนและวิธีเคเนียเรสเนเบอร์เมื่อข้อมูลมีการแจกแจงโคชีจะมีแนวโน้มเพิ่มขึ้น และเมื่อพิจารณาแต่ละขนาดตัวอย่างของทั้ง 2 วิธี พบว่า วิธีซัพพอร์ตเวกเตอร์แมทชีนสามารถจัดกลุ่มข้อมูลได้ดีกว่าวิธีเคเนียเรสเนเบอร์ในทุกขนาดตัวอย่าง

ตารางที่ 5 เปอร์เซ็นต์การจัดกลุ่มถูกต้องด้วยวิธีซัพพอร์ตเวกเตอร์แมทชีนและวิธีเคเนียเรสเนเบอร์ เมื่อข้อมูลมีการแจกแจงเลวี่

วิธีการจัดกลุ่ม	ขนาดตัวอย่าง		
	100	300	500
SVM	90.11	89.90	91.84
KNN	87.79	89.73	91.67

เมื่อ ช่องทึบ แทน เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงสุดของแต่ละขนาดตัวอย่าง

จากตารางที่ 5 จะเห็นได้ว่าเมื่อขนาดตัวอย่างเพิ่มมากขึ้น พบว่าเปอร์เซ็นต์การจัดกลุ่มถูกต้องของวิธีซัพพอร์ตเวกเตอร์แมทชีนและวิธีเคเนียเรสเนเบอร์เมื่อข้อมูลมีการแจกแจงเลวี่จะมีแนวโน้มเพิ่มขึ้น และเมื่อพิจารณาแต่ละขนาดตัวอย่างของทั้ง 2 วิธี พบว่า วิธีซัพพอร์ตเวกเตอร์แมทชีนสามารถจัดกลุ่มข้อมูลได้ดีกว่าวิธีเคเนียเรสเนเบอร์ในทุกขนาดตัวอย่าง

ตารางที่ 6 การวิเคราะห์เปรียบเทียบวิธีซัพพอร์ตเวกเตอร์แมทชีนและวิธีเคเนียเรสเนเบอร์โดยสถิติทดสอบที

การแจกแจง	$\bar{x}$		t-test
	SVM	KNN	
การแจกแจงเสถียรปกติ	93.090	91.607	7.197*
การแจกแจงโคชี	94.977	94.667	1.227
การแจกแจงเลวี่	90.617	89.730	0.694

จากตารางที่ 6 จะเห็นได้ว่า เมื่อข้อมูลมีการแจกแจงเสถียรปกติ พบว่าเปอร์เซ็นต์การจัดกลุ่มถูกต้องของวิธีซัพพอร์ตเวกเตอร์แมทชีนและวิธีเคเนียเรสเนเบอร์แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 และเมื่อข้อมูลมีการแจกแจงโคชีและการแจกแจงเลวี่ พบว่าเปอร์เซ็นต์การจัดกลุ่มถูกต้องของวิธีซัพพอร์ตเวกเตอร์แมทชีนและวิธีเคเนียเรสเนเบอร์ไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05

สรุปและอภิปรายผล

1. กรณีจัดกลุ่มข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมทชีน จะเห็นได้ว่า เมื่อกำหนดอัตราส่วนระหว่าง Training Data : Test Data คงที่ พบว่าเปอร์เซ็นต์การจัดกลุ่มถูกต้องมีแนวโน้มเพิ่มขึ้น เมื่อขนาดตัวอย่างเพิ่มมากขึ้น และเมื่อกำหนดขนาดตัวอย่างคงที่ พบว่าเปอร์เซ็นต์การจัดกลุ่มถูกต้องมีแนวโน้มเพิ่มขึ้น เมื่ออัตราส่วนระหว่าง Training Data : Test Data เพิ่มมากขึ้น ทั้ง 3 การแจกแจง

2. กรณีจัดกลุ่มข้อมูลด้วยวิธีเคเนียเรสเนเบอร์ จะเห็นได้ว่าทั้ง 3 การแจกแจง เมื่อกำหนดขนาดตัวอย่างและค่า k คงที่ พบว่าเปอร์เซ็นต์การจัดกลุ่มถูกต้องมีแนวโน้มเพิ่มขึ้น เมื่ออัตราส่วนของข้อมูลใน Training Data เพิ่มมากขึ้น เมื่อกำหนดอัตราส่วนระหว่าง Training Data : Test Data และค่า k คงที่ พบว่าเปอร์เซ็นต์การจัดกลุ่มถูกต้องมีแนวโน้มเพิ่มขึ้น เมื่อขนาดตัวอย่างเพิ่มมากขึ้น และเมื่อกำหนดขนาดตัวอย่างและอัตราส่วนระหว่าง Training Data : Test Data คงที่ พบว่าเปอร์เซ็นต์การจัดกลุ่มถูกต้องมีแนวโน้มเพิ่มขึ้น เมื่อค่า k เพิ่มมากขึ้น

3. เปรียบเทียบเปอร์เซ็นต์การจัดกลุ่มถูกต้องด้วยวิธีซัพพอร์ตเวกเตอร์แมทชีนและวิธีเคเนียเรสเนเบอร์ เมื่อพิจารณาจากภายใต้ขอบเขตของการศึกษาแต่ละขนาดตัวอย่าง แต่ละการแจกแจงของข้อมูล และแต่ละระดับของอัตราส่วนของข้อมูลระหว่าง Training Data : Test Data จะเห็นได้ว่าวิธีซัพพอร์ตเวกเตอร์แมทชีนให้เปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงกว่าวิธีเคเนียเรสเนเบอร์ในทุกกรณี

4. การวิเคราะห์เปรียบเทียบวิธีซัพพอร์ตเวกเตอร์แมทชีนและวิธีเคเนียเรสเนเบอร์โดยสถิติทดสอบที ซึ่งมีการแจกแจงของข้อมูลทั้ง 3 การแจกแจง จะเห็นได้ว่าเมื่อข้อมูลมีการแจกแจงเสถียรปกติ พบว่าเปอร์เซ็นต์การจัดกลุ่มถูกต้องของวิธีซัพพอร์ตเวกเตอร์แมทชีนและวิธีเคเนียเรสเนเบอร์แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 และเมื่อข้อมูลมีการแจกแจงโคชีและการแจกแจงเลวี่ พบว่าเปอร์เซ็นต์การจัดกลุ่มถูกต้องของวิธีซัพพอร์ตเวกเตอร์แมทชีนและวิธีเคเนียเรสเนเบอร์ไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05

ข้อเสนอแนะ

งานวิจัยนี้สามารถนำไปประยุกต์ใช้ได้อย่างมีประสิทธิภาพในทุกลักษณะที่ใกล้เคียง สามารถใช้เป็นเครื่องมือที่สำคัญในการช่วยเหลือนักศึกษาและอาจารย์ในการศึกษางานวิจัยทางด้านนี้สืบเนื่องต่อไป และสามารถนำไปประยุกต์ใช้ในด้านต่างๆ เช่น ด้านอุตสาหกรรม ด้านการเงิน ด้านหุ้น ด้านเศรษฐศาสตร์ ฯลฯ ได้ เนื่องจากข้อมูลทางด้านนี้ส่วนใหญ่มีการแจกแจงแบบเสถียรปกติ การแจกแจงโคชี หรืออาจมีการแจกแจงเลวี่ จึงเป็นประโยชน์และสามารถใช้เป็นแนวทางให้กับผู้ที่สนใจต่อไป

กิตติกรรมประกาศ

ขอขอบคุณมหาวิทยาลัยราชภัฏพิบูลสงคราม ที่ให้การสนับสนุนทุนวิจัยนี้

เอกสารอ้างอิง

[1] ณัฐวุฒิ คูวัฒนเธียรชัย. (2555). *การวิจัยทางการเงิน (FinancialResearch).* สืบค้นเมื่อ 8 ตุลาคม 2558, จาก http://fin.bus.ku.ac.th

[2] สุรศักดิ์ พรรณรักษา. (2558). *การเปรียบเทียบวิธีการจัดกลุ่มกรณีข้อมูลมีการแจกแจงแบบเสถียรที่มีลักษณะหางหนา.* วิทยานิพนธ์ วท.บ. (สถิติประยุกต์). พิษณุโลก: มหาวิทยาลัยราชภัฏพิบูลสงคราม.

[3] วาทินี นุ้ยเพียร; และ พยุง มีสัจ. (2556). เปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะแบบการกรองและการควบรวมของการทำเหมืองข้อความเพื่อการจำแนกข้อความ. *วารสารวิชาการเทคโนโลยีอุตสาหกรรม.* 9(3): 118-129.

[4] นิเวศ จิระวิชิตชัย; และคณะ. (2555, มีนาคม). วิธีการพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ. *วารสารพัฒนบริหารศาสตร์.* 51(3): 193-203.

[5] Ravi, A.; and Butar, F. B. (2010). An Insight Into Heavy-Tailed Distribution. *Journal of Mathematical Science & Mathematics Eduction.* 19-31.

[6] Wang, S.-j, Mathew, A., Chen, Y., Xi, L.-f., Ma, L., and Lee, J. (2009). Empirical analysis of support vector machine ensemble classifiers. *Journal of Expert Systems with Applications.* 36: 6466-6476.

[7] Duda, R.O., et al. (2001). *Pattern Classification.* United States of America : John Wiley & Sons.