ประสิทธิภาพการจัดกลุ่มของวิธีซัพพอร์ตเวกเตอร์แมทชีน และวิธีเคเนียเรสเนเบอร์ เมื่อข้อมูลมีการแจกแจงเบ้

ณัฏฐินี ดีแท้
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏพิบูลสงคราม
corresponding author e-mail: Natthineed@gmail.com

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพการจัดกลุ่มของวิธีซัพพอร์ตเวกเตอร์แมทชีน (SVM) และวิธีเคเนียเรสเนเบอร์ (K-NN) กรณีข้อมูลมีการแจกแจงใกล้เคียงปกติ (near Normal) การแจกแจงไม่สมมาตรแบบหางสั้น (asymmetric short-tailed) และการแจกแจงไม่สมมาตรแบบหางยาว (asymmetric long-tailed) โดยข้อมูลแบ่งออกเป็น 2 ชุด คือ ชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ ที่มีขนาดตัวอย่างเป็น 100 200 และ 500 โดยจัดข้อมูลออกเป็น 2 กลุ่ม ซึ่งกำหนดอัตราส่วนระหว่างชุดข้อมูลเรียนรู้: ชุดข้อมูลทดสอบมีค่าเป็น 90:10 80:20 70:30 60:40 และ 50:50 ทำการจำลองข้อมูลโดยเทคนิคมอนติคาร์โลและกระทำซ้ำ 5,000 ครั้งในแต่ละสถานการณ์ที่กำหนด และใช้ค่าเฉลี่ยของการจัดกลุ่มผิดพลาดเป็นเกณฑ์ในการเปรียบเทียบ โดยผลการวิจัยพบว่า เมื่อข้อมูลมีการแจกแจงใกล้เคียงปกติ วิธี K-NN ให้อัตราการจัดกลุ่มข้อมูลผิดพลาดต่ำกว่าวิธี SVM แต่เมื่อข้อมูลมีการแจกแจงไม่สมมาตรแบบหางยาวและไม่สมมาตรแบบหางสั้น วิธี SVM ให้อัตราการจัดกลุ่มข้อมูลผิดพลาดต่ำกว่าวิธี K-NN

คำสำคัญ: การจัดกลุ่ม ซัพพอร์ตเวกเตอร์แมทชีน เคเนียเรสเนเบอร์ การแจกแจงเบ้

PERFORMANCE OF CLASSIFICATION USING SUPPORT VECTOR MACHINE AND K-NEAREST NEIGHBOR IN SKEWED DISTRIBUTION

Natthinee Deetae
Faculty of Science and Technology, Pibulsongkram Rajabhat University
corresponding author e-mail: Natthineed@gmail.com

Abstract

The purposed of this research aimed to compare the efficiency of support vector machine and K-nearest neighbor when data are distributed as Near Normal, Asymmetric short-tailed and Asymmetric long-tailed. Data employed in this research were generated into two data sets, consisting of training set and test set with the sample sizes of 100, 200 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 were simulated with Monte Carlo technique and repeated 5,000 times. Rate of

misclassification is used as a criterion for comparison. The results found that the K-NN method produces lower misclassification rate than SVM method with Near Normal distribution, while the SVM method produces lower misclassification rate than K-NN method when data are distributed as Asymmetric long-tailed and Asymmetric short-tailed.

Keywords: classification, support vector machine, K-nearest neighbor, skewed distribution

บทนำ

ปัจจุบันได้มีการนำข้อมูลที่ถูกจัดเก็บในระบบฐานข้อมูลด้านต่าง ๆ เช่น ด้านการศึกษา ด้านการแพทย์ ด้านเศรษฐกิจ ด้านสังคมศาสตร์ และด้านสิ่งแวดล้อม ฯลฯ มาใช้ประโยชน์ ซึ่งข้อมูลอาจมีหลายตัวแปรและจำแนกกลุ่มตั้งแต่ 2 กลุ่มขึ้นไป ดังนั้นจึงต้องอาศัยการวิเคราะห์ตัวแปรเชิงพหุ (multivariate analysis) ซึ่งเป็นวิธีทางสถิติที่นิยมใช้กันอยู่ทั่วไปมาวิเคราะห์ข้อมูล เพื่อให้ข้อมูลที่ได้มีความถูกต้องและน่าเชื่อถือ โดยข้อมูลที่พบในชีวิตประจำวันส่วนมากจะเป็นข้อมูลที่มีการแจกแจงไม่ปกติ มักมีการแจกแจงลักษณะอื่นๆ หรือข้อมูลที่พบอาจมีการแจกแจงที่เบ้ขวา (right skewed distribution) หรือเบ้ซ้าย (left skewed distribution) ซึ่งเมื่อพิจารณาในแง่ของการวิเคราะห์ จะแบ่งวิธีการทางสถิติได้เป็น 2 ประเภทตามลักษณะของข้อมูล คือ วิธีการพาราเมตริก (parametric methods) และวิธีการนอนพาราเมตริก (nonparametric methods) ซึ่งในทางปฏิบัติพบว่าวิธีการนอนพาราเมตริกมีใช้กันอย่างกว้างขวางในงานวิจัย เพราะวิธีการนอนพาราเมตริกมีเงื่อนไขการใช้ภายใต้ข้อตกลงเบื้องต้นเพียงไม่กี่ข้อและที่สำคัญไม่จำเป็นต้องทราบรูปแบบการแจกแจงของประชากร โดยสามารถใช้ได้กับข้อมูลที่มีระดับการวัดต่ำตั้งแต่มาตรานามบัญญัติที่สามารถนับเป็นความถี่ได้ มาตราเรียงอันดับหรือตัวเลขใดๆที่สามารถนำมาจัดอันดับที่ได้ เช่น วิธีเคเนียเรสเนเบอร์ (K-nearest neighbor: K-NN) วิธีซัพพอร์ตเวกเตอร์แมทชีน (support vector machine: SVM) วิธีกฎริปเปอร์ (ripper rules) วิธีนิวรอลเน็ตเวิร์ก (neural networks) และวิธีเนอีฟเบย์ (naïve bayes) เป็นต้น

นิเวศ และคณะ (2554) ได้นำเสนอวิธีการพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทย โดยนำเสนอแบบจำลองการจัดหมวดหมู่ของเอกสารภาษาไทยแบบอัตโนมัติ ทดสอบประสิทธิภาพแบบจำลองการจัดหมวดหมู่เอกสารภาษาไทยกับอัลกอริทึมต้นไม้ตัดสินใจ (decision tree) ซัพพอร์ทเวกเตอร์แมชชีน เนอีฟเบย์ เครือข่ายฟังก์ชันฐานรัศมี (radial basis function network) เคเนียเรสเนเบอร์ และกฎริปเปอร์ โดยใช้วิธีการลดคุณลักษณะร่วมกับอัลกอริทึมเครื่องจักรการเรียนรู้เพื่อศึกษาวิธีการลดคุณลักษณะที่เหมาะสมและมีประสิทธิภาพในการจัดหมู่เอกสารข่าวภาษาไทย จากการวิจัยพบว่า การลดคุณลักษณะด้วยวิธีการเพิ่มของข้อมูล (information gain) เพื่อลดมิติของข้อมูลแล้วส่งเข้าเครื่องจักรการเรียนรู้และวัดประสิทธิภาพจากการลดคุณลักษณะที่ทำให้ค่าวัดจากผลรวมของค่าเฉลี่ย (F-measurement) สูงสุด สามารถสรุปได้ว่า อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีนให้ประสิทธิภาพสูงสุด คือ 94.3% รองลงมาเป็นอัลกอริทึมเนอีฟเบย์ให้ประสิทธิภาพเท่ากับ 86.2% อัลกอริทึมเครือข่ายฟังก์ชันฐานรัศมีให้ประสิทธิภาพเท่ากับ 86.1% อัลกอริทึมต้นไม้ตัดสินใจให้ประสิทธิภาพเท่ากับ 79.7% อัลกอริทึมกฎริปเปอร์ให้

ประสิทธิภาพเท่ากับ 78.9% อัลกอริทึมเคเนียเรสเนเบอร์ให้ประสิทธิภาพเท่ากับ 69.5% ตามลำดับ นอกจากนี้วาทินี และพยุง (2556) ได้เปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะแบบการกรองและการควบรวมของการทำเหมืองข้อความเพื่อการจำแนกข้อความ โดยคัดเลือกคุณลักษณะการกรองแบบไคสแควร์ ใช้วิธีการจำแนกประเภทแบบวิธีซัพพอร์ตเวกเตอร์แมทชีนซึ่งใช้เคอร์เนลแบบโพลิโนเมียลและเรเดียลเบสิสฟังก์ชัน วิธีเนอีฟเบย์ วิธีเบย์เซียนเน็ตเวิร์ก และวิธีเคเนีย-เรสเนเบอร์ ผลการวิจัยพบว่า วิธีซัพพอร์ตเวกเตอร์แมทชีนโดยใช้เคอร์เนลเรเดียลเบสิสฟังก์ชัน (SVMR) ให้ผลการวัดประสิทธิภาพโดยรวมสูงที่สุดคือ 92.2% และเคอร์เนลแบบโพลิโนเมียลวัดประสิทธิภาพได้ 86.5% วิธีนาอีฟเบย์วัดประสิทธิภาพได้ 91.7% วิธีเบย์เซี่ยนเน็ตเวิร์ก วัดประสิทธิภาพได้ 91.4% และวิธีเคเนียเรสเนเบอร์วัดประสิทธิภาพ 88.7% ตามลำดับ

จากการที่ผู้วิจัยได้ทบทวนงานวิจัยที่เกี่ยวข้องและเห็นว่าในทางปฏิบัติเป็นไปได้ยากมากที่ข้อมูลจะมีการแจกแจงปกติ ส่วนใหญ่มักพบข้อมูลที่มีการแจกแจงที่เบ้ขวาหรือเบ้ซ้าย ดังนั้นผู้วิจัยจึงสนใจศึกษาวิธีการจัดกลุ่มเมื่อข้อมูลมีการแจกแจงใกล้เคียงปกติ การแจกแจงไม่สมมาตรแบบหางสั้น และการแจกแจงไม่สมมาตรแบบหางยาว ซึ่งเป็นการแจกแจงที่มีลักษณะเบ้ โดยเปรียบเทียบการจัดกลุ่มด้วยวิธีทางนอนพาราเมตริก 2 วิธี คือ วิธีซัพพอร์ตเวกเตอร์แมทชีน และวิธีเคเนียเรสเนเบอร์ เพื่อช่วยแก้ไขปัญหาการจัดกลุ่มข้อมูลทางสถิติได้อย่างถูกต้องชัดเจน ก่อให้เกิดความคลาดเคลื่อนน้อยที่สุด ซึ่งจะส่งผลให้การจัดกลุ่มมีประสิทธิภาพมากยิ่งขึ้น โดยงานวิจัยนี้แบ่งข้อมูลออกเป็น 2 ชุด คือชุดของหน่วยตัวอย่างที่ใช้ในการสร้างเกณฑ์การจำแนก เรียกว่า ชุดข้อมูลเรียนรู้ (training sets) และชุดของหน่วยตัวอย่างที่ใช้ในการทดสอบเกณฑ์การจำแนกที่สร้างขึ้น เรียกว่า ชุดข้อมูลทดสอบ (test sets) และเมื่อสร้างเกณฑ์การจำแนกจากชุดข้อมูลเรียนรู้แล้ว จึงนำเกณฑ์ที่ได้ไปใช้ในการจัดกลุ่มของค่าสังเกตค่าใหม่ที่อยู่ในชุดข้อมูลทดสอบว่าควรจัดอยู่ในกลุ่มใด โดยใช้อัตราความผิดพลาด (misclassification rate) เป็นเกณฑ์ในการเปรียบเทียบ ซึ่งวิธีที่ให้อัตราความผิดพลาดต่ำกว่าจะถือว่าวิธีนั้นสามารถจัดกลุ่มข้อมูลได้ดีกว่า

วิธีดำเนินการวิจัย

1. สร้างข้อมูลในการวิจัย

1.1 กำหนดลักษณะของตัวแปรอิสระให้เป็นข้อมูลเชิงปริมาณที่มีการแจงแจงแบบเบ้ ซึ่งแบ่งเป็นกลุ่มตามความเบ้และความโด่ง โดยใช้เกณฑ์ของ Shapiro et al. (1968) ดังนี้

ตารางที่ 1 กลุ่มความเบ้และความโด่ง โดยใช้เกณฑ์ของ Shapiro-Wilk และ Chen

การแจงแจงแบบเบ้	ความเบ้	ความโด่ง
การแจกแจงใกล้เคียงปกติ	0.25	2.80
การแจกแจงไม่สมมาตรแบบหางสั้น	0.75	2.80
การแจกแจงไม่สมมาตรแบบหางยาว	0.75	6.00

โดยการสร้างข้อมูลที่มีความเบ้นั้นจะใช้การแปลงฟังก์ชันในการสร้างข้อมูลของ Remberg et al. (1979) ซึ่งเรียกว่า generalized lambda distribution (GLD) ซึ่งการสร้างการแจงแจงแบบ GLD นั้นใช้การแปลงข้อมูลในลักษณะดังนี้

$$x = \frac{\lambda_1 + (p^{\lambda_3} - (1-p)^{\lambda_4})}{\lambda_2} \quad ;0 < p < 1 \qquad \text{..... (1)}$$

เมื่อ p แทน ตัวเลขสุ่ม มีค่าอยู่ระหว่าง 0 ถึง 1

λ_1 แทน พารามิเตอร์ที่กำหนดตำแหน่ง

λ_2 แทน พารามิเตอร์ที่กำหนดสเกล

λ_3, λ_4 แทน พารามิเตอร์ที่กำหนดรูปแบบการแจกแจง

ในการวิจัยครั้งนี้จะสร้างการแจกแจงแบบ GLD ซึ่งมีความเบ้และความโด่งตามต้องการ โดยมีค่าเฉลี่ยเป็น 0 และความแปรปรวนเป็น 1 ก่อน ถ้าต้องการให้ข้อมูลมีค่าเฉลี่ย μ และความแปรปรวน σ จะต้องแปลงค่า λ_1 และ λ_2 ในตารางของ Ramberg 1979 ก่อนที่จะแทนในสมการ 1 ดังนี้

$$\lambda_1(EX, STD) = \lambda_1(0,1)\sigma + \mu$$
$$\lambda_2(EX, STD) = \lambda_2(0,1)/\sigma$$

เมื่อ EX แทน ค่าคาดหวัง

STD แทน ส่วนเบี่ยงเบนมาตรฐาน

ส่วน λ_3 และ λ_4 จะกำหนดค่าต่าง ๆ กันตามความเบ้และความโด่งของการแจกแจงที่ต้องการ

1.2 กำหนดอัตราส่วนระหว่างชุดข้อมูลเรียนรู้: ชุดข้อมูลทดสอบมีค่าเป็น 90:10 80:20 70:30 60:40 และ 50:50

1.3 กำหนดขนาดตัวอย่าง (n) ที่สนใจศึกษาเป็น 100 200 และ 500

1.4 ทำการแบ่งกลุ่มตัวอย่างออกเป็น 2 กลุ่ม

1.5 จำลองข้อมูลโดยเทคนิคมอนติคาร์โล กระทำซ้ำ 5,000 ครั้งในแต่ละสถานการณ์ที่กำหนด

2. ซัพพอร์ตเวคเตอร์แมชชีน

ซัพพอร์ตเวคเตอร์แมชชีนเป็นเทคนิคหนึ่งที่ได้รับความนิยมอย่างแพร่หลายในงานที่เกี่ยวข้องกับการจดจำรูปแบบตลอดจนการแก้ปัญหาการจัดกลุ่ม (classification problem) (Wang et al. 2009; Chen et al. 2009) โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการ เพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกกลุ่มข้อมูลได้ดีที่สุด (optimal separating hyperplane) เมื่อเราพิจารณาข้อมูลที่ประกอบด้วยข้อมูล 2 กลุ่มดังสมการที่ 2

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

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 นั้นมุ่งเป้าเพื่อหาฟังก์ชันการตัดสินใจที่สามารถแบ่งแยกค่าที่ไม่ทราบได้ ดังสมการที่ 3

$$f(\mathbf{x}) = sign\left\{\sum_{k=1}^{n_v} w_k \varphi_k(\mathbf{x})\varphi_k(\mathbf{x}_k) + b\right\} \quad \text{.....(3)}$$

$$\varphi(\mathbf{x}) = [\varphi_1(\mathbf{x}_1), \varphi_2(\mathbf{x}_2), ..., \varphi_n(\mathbf{x}_{n_v})]^T \quad \text{.....(4)}$$

กลุ่มข้อมูล $\mathbf{x}$ จากสมการที่ 4 ไม่สามารถแบ่งแยกได้ด้วยสมการเส้นตรงแต่จะถูกแปลงให้อยู่ในรูปแบบที่สามารถใช้สมการเส้นตรงแบ่งแยกได้ โดยใช้เคอร์เนลฟังก์ชัน (kernel function)

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

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

w_k แทน ค่าน้ำหนักที่เชื่อมโยงจาก feature space ไปสู่ out put space

b แทน ค่าโน้มเอียง (bias)

$\mathbf{x}_k$ แทน ซัพพอร์ตเวกเตอร์ โดย $k = 1,2,...,n_v$

n_v แทน จำนวนซัพพอร์ตเวกเตอร์

วิธีการที่ใช้ในการหาเส้นแบ่งที่ดีที่สุดคือการเพิ่มเส้นขอบ (margin) ให้กับเส้นแบ่งทั้งสองข้างและสร้างเส้นขอบที่สัมผัสกับค่าข้อมูลใน feature space ที่ใกล้ที่สุด ดังนั้นเส้นแบ่งที่มีเส้นขอบกว้างที่สุดจึงเป็นเส้นแบ่งที่ดีที่สุดและเรียกตำแหน่งการสัมผัสข้อมูลที่ใกล้ที่สุดจากการเพิ่มขอบนี้ว่า “ซัพพอร์ตเวกเตอร์” (support vector) เนื่องจากในบางกรณีการแบ่งแยกกลุ่มไม่สามารถทำได้ถูกต้องโดยสมบูรณ์ ดังนั้นจึงต้องมีการกำหนดตัวแปรสำหรับยอมรับค่าความผิดพลาดโดยการเพิ่มตัวแปร ξ (slack variable) ดังสมการที่ 5 และ 6 ดังนี้

$$w^T\mathbf{x} + b \geq 1 - \xi_i \quad \text{.....(5)}$$

$$w^T\mathbf{x} + b \leq -1 + \xi_i \quad \text{.....(6)}$$

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

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

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) :

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

เรเดียลเบสิสฟังก์ชัน (Radial Basis Function-RBF) :

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

ซิกมอยด์ (sigmoid) :

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

2. วิธีเคเนียเรสเนเบอร์

หลักการของวิธีการนี้จะจำแนกประเภทข้อมูลโดยขึ้นกับข้อมูลที่มีคุณสมบัติใกล้เคียงที่สุด k ตัว ซึ่งเป็นวิธีการจัดกลุ่มให้กับค่าสังเกตค่าใหม่ที่ใกล้กับชุดข้อมูลเรียนรู้ (Duda et al. 2001) โดยกำหนดให้ชุดข้อมูลเรียนรู้ ดังสมการที่ 8

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

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

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

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

โดยมีคุณสมบัติของฟังก์ชันระยะห่าง ส่วนใหญ่พบว่าฟังก์ชันระยะห่างที่ใช้บ่อยเป็นฟังก์ชันระยะห่างแบบยูคลิเดียน (euclidean distance) ซึ่งเป็นฟังก์ชันที่นิยมใช้ในการหาค่าระยะห่างที่แท้จริงดังสมการที่ 10

$$\begin{aligned} d_{euclidean}(\mathbf{u}, \mathbf{v}) &= \sqrt{\mathbf{u}^{\mathrm{T}} \mathbf{v}} \\ &= \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2} \end{aligned} \qquad \text{.....(10)}$$

โดยที่ $\mathbf{u} = \{u_1, u_2, ..., u_n\}$ และ $\mathbf{v} = \{v_1, v_2, ..., v_n\}$ แทนการบันทึกค่าคุณลักษณะของ m ทั้งสองกลุ่ม ดังนั้นระยะห่างระหว่างค่าสังเกตใหม่ ($\mathbf{x}_0$) และ $\mathbf{x}_i$ ถูกกำหนดโดย

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

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

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

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

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

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

4. เกณฑ์การตัดสินใจ

อัตราการจัดกลุ่มข้อมูลผิดพลาดจะพิจารณาประสิทธิภาพของการจัดกลุ่มโดยคำนวณจากการจัดกลุ่มผิดพลาด ดังนี้

$$\text{อัตราการจัดกลุ่มข้อมูลผิดพลาด} = \frac{Mc}{Tstotal}$$

เมื่อ Mc แทน จำนวนข้อมูลที่มีการจัดกลุ่มผิดในชุดข้อมูลทดสอบ

$Tstotal$ แทน จำนวนข้อมูลทั้งหมดของชุดข้อมูลทดสอบ

ผลการวิจัย

ตารางที่ 1 ผลของอัตราการจัดกลุ่มข้อมูลผิดพลาด เมื่อขนาดตัวอย่างเท่ากับ 100

การแจกแจงแบบเบ้	วิธีการจัดกลุ่ม	อัตราส่วนของข้อมูลระหว่างชุดข้อมูลเรียนรู้ : ชุดข้อมูลทดสอบ				
		50:50	60:40	70:30	80:20	90:10
การแจกแจงใกล้เคียงปกติ	SVM	0.1879	0.1857	0.1848	0.1842	0.1823
	3-NN	0.1821	0.1704	0.1642	0.1580	0.1525*
	5-NN	0.1996	0.1856	0.1748	0.1688	0.1622
	7-NN	0.2236	0.2054	0.1909	0.1859	0.1774
การแจกแจงไม่สมมาตรแบบหางยาว	SVM	0.1050	0.1041	0.0989*	0.1033	0.1038
	3-NN	0.1426	0.1359	0.1279	0.1221	0.1222
	5-NN	0.1507	0.1404	0.1333	0.1281	0.1243
	7-NN	0.1593	0.1484	0.1423	0.1369	0.1327
การแจกแจงไม่สมมาตรแบบหางสั้น	SVM	0.0683	0.0625	0.0571	0.0582	0.0539*
	3-NN	0.0854	0.0842	0.0830	0.0846	0.0838
	5-NN	0.0662	0.0633	0.0622	0.0609	0.0609
	7-NN	0.0630	0.0598	0.0571	0.0587	0.0559

หมายเหตุ * แทน อัตราการจัดกลุ่มข้อมูลผิดพลาดของแต่ละการแจกแจง

จากตารางที่ 1 จะเห็นว่าในแต่ละการแจกแจงแบบเบ้ เมื่อชุดข้อมูลเรียนรู้เพิ่มมากขึ้น พบว่าอัตราการจัดกลุ่มข้อมูลผิดพลาดโดยวิธี SVM และวิธี KNN ที่เปลี่ยนแปลงค่า k $(k = 3, 5, 7)$ มีแนวโน้มลดลง โดยการแจกแจงแบบใกล้เคียงปกติพบว่าวิธี 3-NN สามารถจัดกลุ่มข้อมูลได้ดีที่สุด ที่อัตราการจัดกลุ่มข้อมูลผิดพลาดเท่ากับ 0.1525 เมื่อใช้ชุดข้อมูลเรียนรู้: ชุดข้อมูลทดสอบ เป็น 90:10 การแจกแจงไม่สมมาตรแบบหางยาว พบว่าวิธี SVM สามารถจัดกลุ่มข้อมูลได้ดีที่สุด ที่อัตราการจัดกลุ่มข้อมูลผิดพลาดเท่ากับ 0.0989 เมื่อใช้ชุดข้อมูลเรียนรู้: ชุดข้อมูลทดสอบ เป็น 70:30 และเมื่อข้อมูลมีการแจกแจงไม่สมมาตรแบบหางสั้น พบว่าวิธี SVM สามารถจัดกลุ่มข้อมูลได้ดีที่สุด ที่อัตราการจัดกลุ่มข้อมูลผิดพลาดเท่ากับ 0.0539 เมื่อใช้ชุดข้อมูลเรียนรู้: ชุดข้อมูลทดสอบเป็น 90:10

ตารางที่ 2 ผลของอัตราการจัดกลุ่มข้อมูลผิดพลาด เมื่อขนาดตัวอย่างเท่ากับ 200

การแจกแจงแบบเบ้	วิธีการจัดกลุ่ม	อัตราส่วนของข้อมูลระหว่างชุดข้อมูลเรียนรู้ : ชุดข้อมูลทดสอบ				
		50:50	60:40	70:30	80:20	90:10
การแจกแจงใกล้เคียงปกติ	SVM	0.1421	0.1392	0.1390	0.1388	0.1385
	3-NN	0.1473	0.1407	0.1346	0.1302*	0.1303
	5-NN	0.1550	0.1463	0.1380	0.1312	0.1311
	7-NN	0.1658	0.1562	0.1464	0.1360	0.1377
การแจกแจงไม่สมมาตรแบบหางยาว	SVM	0.1196	0.1124	0.1079	0.1059	0.1010*****
	3-NN	0.1221	0.1145	0.1097	0.1069	0.1027
	5-NN	0.1281	0.1188	0.1159	0.1118	0.1070
	7-NN	0.1121	0.1192	0.1185	0.1168	0.1165
การแจกแจงไม่สมมาตรแบบหางสั้น	SVM	0.0539	0.0516	0.0505	0.0503	0.0500*****
	3-NN	0.0611	0.0597	0.0598	0.0589	0.0637
	5-NN	0.0596	0.0595	0.0586	0.0585	0.0580
	7-NN	0.0559	0.0546	0.0543	0.0543	0.0533

หมายเหตุ * แทน อัตราการจัดกลุ่มข้อมูลผิดพลาดของแต่ละการแจกแจง

จากตารางที่ 2 จะเห็นว่าในแต่ละการแจกแจงแบบเบ้ เมื่อชุดข้อมูลเรียนรู้เพิ่มมากขึ้น พบว่าอัตราการจัดกลุ่มข้อมูลผิดพลาดโดยวิธี SVM และวิธี KNN มีแนวโน้มลดลงโดยการแจกแจงใกล้เคียงปกติ พบว่าวิธี 3-NN สามารถจัดกลุ่มข้อมูลได้ดีที่สุด ที่อัตราการจัดกลุ่มข้อมูลผิดพลาดเท่ากับ 0.1302 เมื่อใช้ชุดข้อมูลเรียนรู้: ชุดข้อมูลทดสอบ เป็น 80:20 การแจกแจงไม่สมมาตรแบบหางยาว พบว่าวิธี SVM สามารถจัดกลุ่มข้อมูลได้ดีที่สุด ที่อัตราการจัดกลุ่มข้อมูลผิดพลาดเท่ากับ 0.1010 เมื่อใช้ชุดข้อมูลเรียนรู้: ชุดข้อมูลทดสอบ เป็น 90:10 และเมื่อข้อมูลมีการแจกแจงไม่สมมาตรแบบหางสั้น พบว่าวิธี SVM สามารถจัดกลุ่มข้อมูลได้ดีที่สุด ที่อัตราการจัดกลุ่มข้อมูลผิดพลาดเท่ากับ 0.0500 เมื่อใช้ชุดข้อมูลเรียนรู้: ชุดข้อมูลทดสอบ เป็น 90:10

ตารางที่ 3 ผลของอัตราการจัดกลุ่มข้อมูลผิดพลาด เมื่อขนาดตัวอย่างเท่ากับ 500

การแจกแจงแบบเบ้	วิธีการจัดกลุ่ม	อัตราส่วนของข้อมูลระหว่างชุดข้อมูลเรียนรู้ : ชุดข้อมูลทดสอบ				
		50:50	60:40	70:30	80:20	90:10
การแจกแจงใกล้เคียงปกติ	SVM	0.1225	0.1298	0.1283	0.1280	0.1246
	3-NN	0.1216	0.1155	0.1148	0.1105	0.1066
	5-NN	0.1210	0.1143	0.1127	0.1075	0.1026*****
	7-NN	0.1242	0.1159	0.1137	0.1085	0.1038
การแจกแจงไม่สมมาตรแบบหางยาว	SVM	0.0999	0.0937	0.0881	0.0843	0.0816*
	3-NN	0.0947	0.0917	0.0914	0.0908	0.0903
	5-NN	0.0954	0.0925	0.0902	0.0882	0.0880
	7-NN	0.0932	0.0893	0.0856	0.0834	0.0833
การแจกแจงไม่สมมาตรแบบหางสั้น	SVM	0.0484	0.0483	0.0474	0.0473	0.0468*****
	3-NN	0.0596	0.0595	0.0586	0.0585	0.0580
	5-NN	0.0584	0.0583	0.0566	0.0563	0.0551
	7-NN	0.0537	0.0533	0.0517	0.0513	0.0508

หมายเหตุ * แทน อัตราการจัดกลุ่มข้อมูลผิดพลาดของแต่ละการแจกแจง

จากตารางที่ 3 จะเห็นว่าในแต่ละการแจกแจงแบบเบ้ เมื่อชุดข้อมูลเรียนรู้เพิ่มมากขึ้น พบว่าอัตราการจัดกลุ่มข้อมูลผิดพลาดโดยวิธี SVM และวิธี KNN มีแนวโน้มลดลง โดยการแจกแจงใกล้เคียงปกติ พบว่าวิธี 5-NN สามารถจัดกลุ่มข้อมูลได้ดีที่สุด ที่อัตราการจัดกลุ่มข้อมูลผิดพลาดเท่ากับ 0.1026 เมื่อใช้ชุดข้อมูลเรียนรู้: ชุดข้อมูลทดสอบ เป็น 90:10 การแจกแจงไม่สมมาตรแบบหางยาว พบว่าวิธี SVM สามารถจัดกลุ่มข้อมูลได้ดีที่สุด ที่อัตราการจัดกลุ่มข้อมูลผิดพลาดเท่ากับ 0.0816 เมื่อใช้ชุดข้อมูลเรียนรู้: ชุดข้อมูลทดสอบ เป็น 90:10 และเมื่อข้อมูลมีการแจกแจงไม่สมมาตรแบบหางสั้น พบว่าวิธี SVM สามารถจัดกลุ่มข้อมูลได้ดีที่สุด ที่อัตราการจัดกลุ่มข้อมูลผิดพลาดเท่ากับ 0.0468 เมื่อใช้ชุดข้อมูลเรียนรู้: ชุดข้อมูลทดสอบ เป็น 90:10

อภิปรายและสรุปผลการวิจัย

เมื่อพิจารณาแต่ละระดับของอัตราส่วนระหว่างชุดข้อมูลเรียนรู้: ชุดข้อมูลทดสอบ โดยส่วนใหญ่แล้วพบว่าในแต่ละการแจกแจงแบบเบ้ เมื่อขนาดตัวอย่างเพิ่มมากขึ้น อัตราการจัดกลุ่มข้อมูลผิดพลาดโดยวิธี SVM และวิธี K-NN มีแนวโน้มลดลงและเมื่อพิจารณาในแต่ละการแจกแจงแบบเบ้ เมื่อชุดข้อมูลเรียนรู้เพิ่มมากขึ้น พบว่าอัตราการจัดกลุ่มข้อมูลผิดพลาดโดยวิธี SVM และวิธี K-NN มีแนวโน้มลดลงทุกขนาดตัวอย่าง นอกจากนี้เมื่อพิจารณาวิธี KNN ในแต่ละการแจกแจง โดยกำหนดให้อัตราส่วนของข้อมูลระหว่าง ชุดข้อมูลเรียนรู้: ชุดข้อมูลทดสอบคงที่ โดยส่วนใหญ่แล้วพบว่าอัตราการจัดกลุ่มข้อมูลผิดพลาดมีแนวโน้มลดลง เมื่อค่า k เพิ่มมากขึ้น เมื่อข้อมูลมีการแจกแจงไม่สมมาตรแบบหางสั้น ที่ขนาดตัวอย่างเท่ากับ 100 และ 200 และเมื่อข้อมูลมีการแจกแจงไม่สมมาตรแบบหางยาวและไม่สมมาตรแบบหางสั้น ที่ขนาดตัวอย่างเท่ากับ 500 โดยเมื่อเปรียบเทียบอัตราการจัดกลุ่มข้อมูลผิดพลาดด้วยวิธี SVM และวิธี K-NN เมื่อพิจารณาภายใต้ขอบเขตของการศึกษาแต่ละการแจกแจงแบบเบ้ จะเห็นได้ว่าเมื่อข้อมูลมีการแจกแจงใกล้เคียงปกติ วิธี K-NN ให้อัตราการจัดกลุ่มข้อมูลผิดพลาดต่ำกว่าวิธี SVM เมื่อข้อมูลมีการแจกแจงไม่สมมาตรแบบหางยาวและไม่สมมาตรแบบหางสั้น โดยส่วนใหญ่แล้ววิธี SVM ให้อัตราการจัดกลุ่มข้อมูลผิดพลาดต่ำกว่าวิธี K-NN

กิตติกรรมประกาศ

ขอขอบคุณมหาวิทยาลัยราชภัฏพิบูลสงคราม ที่ให้การสนับสนุนทุนวิจัยนี้

เอกสารอ้างอิง

นิเวศ จิระวิชิตชัย, ปริญญา สงวนสัตย์ และพยุง มีสัจ. (2554). การพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ. **วารสารพัฒนบริหารศาสตร์**, 51(3). 187-205.

วาทินี นุ้ยเพียร และพยุง มีสัจ. (2556). การเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะแบบการกรองและการควบรวมของการทำเหมืองข้อความเพื่อการจำแนกข้อความ. **วารสารวิชาการเทคโนโลยีอุตสาหกรรม**, 9(3), 118-129.

Chen, S., Wang, W., & Zuylen, H. van. (2009). Construct support vector machine ensemble to detect traffic incident. **Journal of Expert Systems with Applications**, 36, 10976-10986.

Duda, R.O., Hart, P.E. & Stork, D.G. (2001). **Pattern Classification**. United States of America: John Wiley & Sons.

Ramberg, J.S., Tadikamalla, P.R., Dudewicz, E.J. & Mykytka, E.F. (1979). A probability distribution and its uses in fitting data. **Journal of the American Statistical Association**, 21(2), 201-214.

Shapiro, S.S., Wilk, M.B. & Chen, H.J. (1968). A comparative study of various tests for normality. **Journal of the American Statistical Association**, 63(324), 1343-1372.

Wang, S.J, Mathew, A., Chen, Y., Xi, L.F., Ma, L., & Lee, J. (2009). Empirical analysis of support vector machine ensemble classifiers. **Journal of Expert Systems with Applications**, 36, 6466-6476.