การจัดกลุ่มด้วยวิธีเอ็มพิลิคัลเบส์ผสมผสานกับเนียเรสเนเบอร์ เมื่อข้อมูลมีการแจกแจงเสถียร

The Hybrid Classification using Empirical Bayes and Nearest Neighbor with Stable Distribution

ณัฏฐินี ดีแท้

คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏพิบูลสงคราม จ.พิษณุโลก

E-mail: Natthineed@gmail.com

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการจัดกลุ่มด้วยวิธีเอ็มพิลิคัลเบส์ผสมผสานกับเนียเรสเนเบอร์ (EBNN) เมื่อข้อมูลมีการแจกแจงเสถียรปกติ การแจกแจงโคชี และการแจกแจงเลวี โดยกำหนดการแจกแจงเบื้องต้นที่ให้สารสนเทศที่เป็นประโยชน์คือกรณีที่ไม่ทราบค่าเฉลี่ยแต่ทราบความแปรปรวนของประชากรข้อมูล ในงานวิจัยนี้จะแบ่งออกเป็น 2 ส่วนเท่า ๆ กัน คือ ส่วนข้อมูลเรียนรู้และส่วนข้อมูลทดสอบที่มีขนาดตัวอย่างเป็น 100 และ 500 สำหรับการจำแนกข้อมูลออกเป็น 2 กลุ่ม ทำการจำลองข้อมูลโดยเทคนิคมอนติคาร์โลและกระทำซ้ำ 5,000 ครั้ง ในแต่ละสถานการณ์ที่กำหนด และใช้ค่าเฉลี่ยของเปอร์เซ็นต์การจัดกลุ่มถูกต้องเป็นเกณฑ์ในการเปรียบเทียบ ผลการวิจัยพบว่าวิธี EBNN ให้ผลการจัดกลุ่มดีกว่าวิธีเอ็มพิริคัลเบส์ทุกการแจกแจงที่ศึกษา ซึ่งเมื่อจำนวนค่าสังเกต (k) ที่อยู่ใกล้ค่าสังเกตค่าใหม่เพิ่มมากขึ้น เปอร์เซ็นต์การจัดกลุ่มข้อมูลถูกต้องมีแนวโน้มสูงขึ้นในทุกการแจกแจงของข้อมูล และกรณีข้อมูลมีการแจกแจงเสถียรปกติและการแจกแจงเลวี เมื่อขนาดตัวอย่างเพิ่มขึ้นเปอร์เซ็นต์การจัดกลุ่มข้อมูลถูกต้องมีแนวโน้มสูงขึ้นเช่นกัน

ABSTRACT

The propose of this research aimed to investigate a classification technique using Empirical Bayes in combination with Nearest Neighbor (EBNN) when data are distributed as Stable-normal, Cauchy and Levy distributions. The study is performed using informative priors, normal distributions with unknown mean but known variance. Data employed in this study were generated into two equal sets, consisting of training set and test set with the sample sizes 100

and 500 for the binary classification. In each situation, the data are simulated with Monte Carlo technique and repeated 5,000 times. The average percentage of correct classification is used as criteria for comparison. The results found that EBNN method exhibited an improved performance over Empirical Bayes method in all distributions under study. Increasing neighborhood (k) exhibited higher percentages of correct classification in all distributions. Stable-normal and Levy distribution with increasing sample size also exhibited higher percentages of correct classification.

คำสำคัญ: การจัดกลุ่ม เอ็มพิริคัลเบส์ การแจกแจงเสถียรปกติ การแจกแจงโคชี การแจกแจงเลวี
Keywords: Classification, Empirical Bayes, Stable-normal distribution, Cauchy distribution, Levy distribution

บทนำ

การจัดกลุ่ม (Classification) เป็นการจัดกลุ่มของข้อมูลหรือการเดาค่าของข้อมูลโดยที่ตัวแปรอิสระเป็นตัวแปรเชิงปริมาณหรือตัวแปรเชิงคุณภาพก็ได้ หลักของการจัดกลุ่มจะต้องกำหนดกลุ่มล่วงหน้าให้กับข้อมูล โดยจำแนกข้อมูลออกเป็นกลุ่มย่อยหลาย ๆ กลุ่มตั้งแต่ 2 กลุ่มขึ้นไป ซึ่งข้อมูลที่อยู่กลุ่มเดียวกันจะต้องมีความคล้ายคลึงกัน ส่วนข้อมูลที่อยู่ต่างกลุ่มกันจะมีความแตกต่างกัน การจัดกลุ่มจะจำแนกค่าสังเกตค่าใหม่ (New observation) ให้อยู่กลุ่มใดกลุ่มหนึ่งได้อย่างเหมาะสม (Johnson and Wichern, 2002; Hachicha and Ghorbel, 2012) ซึ่งในปัจจุบันได้มีแนวคิดใหม่เกี่ยวกับการจัดกลุ่มเกิดขึ้น คือเป็นการนำเอาข้อดีของแต่ละวิธีการจัดกลุ่มมาผสมผสานกันจนเกิดเป็นเทคนิคการจัดกลุ่มใหม่ ซึ่งให้ผลการจัดกลุ่มดีกว่าวิธีเดี่ยว ๆ แบบไม่ผสมผสาน เช่นงานวิจัยของ Guo and Chakraborty (2009) ได้ผสมผสานวิธีเบส์เซียนร่วมกับวิธีเนียเรสเนเบอร์ ซึ่งเรียกวิธีดังกล่าวว่าวิธีเบส์เซียนด้วยการปรับเนียเรสเนเบอร์ (BANN) Aci, Inan and Avci (2010) ได้ศึกษาวิธีการผสมผสานวิธีการจัดกลุ่มสามวิธีเข้าด้วยกันได้แก่ วิธีเค-เนียเรสเนเบอร์ วิธีเบส์เซียน และวิธีเจเนติกอัลกอรึทึม Ghosh and Godtliebsen (2011) ได้พัฒนาวิธีการผสมผสานการจัดกลุ่ม และทำการเปรียบเทียบประสิทธิภาพการจัดกลุ่มในกรณีพาราเมตริก (Parametric) และไม่ใช้พาราเมตริก (Nonparametric) Deetae et al. (2013) ได้ผสมผสานวิธีเอ็มพิลิคัลเบส์และวิธีเนียเรสเนเบอร์ ซึ่งเรียกวิธีดังกล่าวว่าวิธีเอ็มพิลิคัลเบส์ผสมผสานกับเนียเรสเนเบอร์ (EBNN) ซึ่งจากงานวิจัยจะเห็นว่าการจัดกลุ่มแบบผสมผสานแต่ละวิธีเข้าด้วยกันเป็นแนวคิดที่ดีและให้ผลการจัดกลุ่มดีกว่าวิธีเดี่ยว ๆ

ลักษณะของข้อมูลก็มีความสำคัญอย่างมากในการเลือกวิธีการจัดกลุ่ม เพื่อให้ผลการจัดกลุ่มมีประสิทธิภาพที่ดีนั้นจึงมีการพิจารณาเลือกปัจจัยหรือตัวแปรอิสระต่าง ๆ ให้เหมาะสมและสอดคล้องกับสมมติฐานเบื้องต้นของวิธีการจัดกลุ่มนั้น ๆ และอีกปัจจัยหนึ่งที่มีความสำคัญอย่างมากคือลักษณะการแจงแจงของข้อมูล ซึ่งโดยทั่วไปการใช้วิธีการวิเคราะห์จำแนกกลุ่ม ตัวแปรอิสระจะต้องมีการแจกแจงแบบปกติเชิงพหุ (Multivariate Normal Distribution) แต่ในทางการปฏิบัติพบว่าบ่อยครั้งข้อมูลที่นำมาวิเคราะห์มาจากประชากรที่มีการแจกแจง

ลักษณะอื่นหรือข้อมูลอาจมีการแจกแจงที่เบ้ขวา (Right Skewed Distribution) หรือเบ้ซ้าย (Left Skewed Distribution) หรือข้อมูลอาจมีการแจกแจงที่มีลักษณะหางหนา (Heavy-tailed distribution) หรือลักษณะหางบาง (Light-tailed distribution) ซึ่งเราไม่สามารถใช้การวิเคราะห์จำแนกกลุ่มแบบดั้งเดิมได้เพราะอาจทำให้ผลของการจัดกลุ่มที่ได้ไม่มีประสิทธิภาพเพียงพอ

โดยการแจกแจงหางหนาเป็นอีกหนึ่งการแจกแจงที่น่าสนใจเนื่องจากการแจกแจงหางหนามีการแจกแจงย่อย (Sub class distribution) ทำให้เกิดข้อมูลที่มีการแจกแจงที่หลากหลาย ดังนั้นผู้วิจัยจึงสนใจศึกษาวิธีการจัดกลุ่มเมื่อข้อมูลมีการแจกแจงเสถียร (Stable distribution) ซึ่งเป็นการแจกแจงย่อยของการแจกแจงหางหนา และการแจกแจงเสถียรเป็นการแจกแจงเดียวที่ถูกรองรับโดยทฤษฎีบทลิมิตสู่ส่วนกลางวางนัยทั่วไป (Generalized central limit theorem) (Ravi and Butar, 2010) โดยการแจกแจงเสถียร ประกอบไปด้วยการแจกแจงเสถียรปกติ (Stable-normal distribution) การแจกแจงโคชี (Cauchy distribution) และการแจกแจงเลวี (Levy distribution) และใช้วิธีการจัดกลุ่มด้วยวิธี EBNN โดยงานวิจัยนี้แบ่งข้อมูลออกเป็น 2 ชุด คือ ชุดของหน่วยตัวอย่างที่ใช้ในการสร้างเกณฑ์การจำแนก เรียกว่า ชุดข้อมูลเรียนรู้ (Training sets) และชุดของหน่วยตัวอย่างที่ใช้ในการทดสอบเกณฑ์การจำแนกที่สร้างขึ้น เรียกว่า ชุดข้อมูลทดสอบ (Test sets) และเมื่อสร้างเกณฑ์การจำแนกจากชุดข้อมูลเรียนรู้แล้ว จึงนำเกณฑ์ที่ได้ไปใช้ในการจัดกลุ่มของค่าสังเกตค่าใหม่ที่อยู่ในชุดข้อมูลทดสอบว่าควรจัดอยู่ในกลุ่มใด โดยใช้เปอร์เซ็นต์การจัดกลุ่มถูกต้องเป็นเกณฑ์ในการเปรียบเทียบ

วัตถุประสงค์ของการวิจัย

เพื่อศึกษาการจัดกลุ่มด้วยวิธี EBNN โดยกำหนดการแจกแจงเบื้องต้นที่ให้สารสนเทศที่เป็นประโยชน์ (Informative priors) คือกรณีที่ไม่ทราบค่าเฉลี่ยแต่ทราบความแปรปรวนของประชากร สำหรับข้อมูลที่มีการแจกแจงเสถียร ดังนี้

1. เพื่อศึกษาการจัดกลุ่มข้อมูลด้วยวิธี EBNN เมื่อข้อมูลมีการแจกแจงเสถียรปกติ
2. เพื่อศึกษาการจัดกลุ่มข้อมูลด้วยวิธี EBNN เมื่อข้อมูลมีการแจกแจงโคชี
3. เพื่อศึกษาการจัดกลุ่มข้อมูลด้วยวิธี EBNN เมื่อข้อมูลมีการแจกแจงเลวี

ขอบเขตของการวิจัย

การวิจัยครั้งนี้เป็นการศึกษาวิธีการจัดกลุ่มเมื่อข้อมูลมีการแจกแจงเสถียรปกติ การแจกแจงโคชี และการแจกแจงเลวี ด้วยวิธี EBNN และสร้างโปรแกรมจำลองทางคอมพิวเตอร์ สำหรับการประยุกต์ใช้ในกรณีศึกษา ซึ่งกำหนดขอบเขตวิจัยดังนี้

1. กำหนดตัวแปรอิสระเป็นข้อมูลเชิงปริมาณ
2. ทำการแบ่งกลุ่มตัวอย่างออกเป็น 2 กลุ่ม
3. ศึกษาการจัดกลุ่มด้วยวิธี EBNN ในกรณีที่ไม่ทราบค่าเฉลี่ยแต่ทราบความแปรปรวนของประชากร เมื่อข้อมูลมีการแจกแจงเสถียรปกติ การแจกแจงโคชี และการแจกแจงเลวี
4. ในแต่ละสถานการณ์จะแบ่งข้อมูลออกเป็นชุดข้อมูลเรียนรู้ และชุดข้อมูลทดสอบอย่างละเท่า ๆ กัน

5. กำหนดขนาดตัวอย่าง (n) ที่สนใจศึกษาเป็น 100 และ 500
6. กำหนดพารามิเตอร์ของการแจกแจงเสถียร $S(\alpha,\beta,\gamma,\delta)$ เมื่อ $0<\alpha\leq 2$, $-1\leq\beta\leq 1$, $\gamma\in \mathrm{R}$ และ $\delta>0$
7. จำลองข้อมูลโดยเทคนิคมอนติคาร์โล กระทำซ้ำ 5,000 ครั้งในแต่ละสถานการณ์ที่กำหนด แล้วคำนวณค่าเฉลี่ยของเปอร์เซ็นต์การจัดกลุ่มถูกต้องเพื่อนำมาเปรียบเทียบประสิทธิภาพของการจัดกลุ่ม

วิธีดำเนินการวิจัย

1. สร้างข้อมูลในการวิจัย

กำหนดลักษณะของตัวแปรอิสระให้เป็นข้อมูลเชิงปริมาณที่มีการแจงแจงเสถียรโดยประกอบไปด้วยพารามิเตอร์ทั้งหมด 4 พารามิเตอร์ (Sundaram and McDonald, 2010) ดังตารางที่ 1

ตารางที่ 1 พารามิเตอร์ของการแจกแจงเสถียร

พารามิเตอร์	ความหมายของพารามิเตอร์	ขอบเขตของพารามิเตอร์
α	รูปแบบของการแจกแจง	$0<\alpha\leq 2$
β	ความเบ้	$-1\leq\beta\leq 1$
δ	ลักษณะตำแหน่ง	$\delta\in \mathrm{R}$
γ	ขนาดสัดส่วน	$\gamma>0$

การแจกแจงเสถียรเป็นการใช้ความรู้พื้นฐานของทฤษฎีการแจกแจงประเภทหนึ่งของการแจกแจงความน่าจะเป็นและมีฟังก์ชันลักษณะเฉพาะ ซึ่งการแจกแจงเสถียรมีลักษณะหางหนาและอาจจะมีลักษณะเบ้ (Skewness) (Ravi and Butar, 2010; Deetae et al., 2013) เมื่อมีการปรับค่าพารามิเตอร์ของการแจกแจงเสถียรจะทำให้รูปแบบการแจกแจงเปลี่ยนแปลงไปเป็นการแจกแจงลักษณะอื่น ๆ ได้ โดยการแจกแจงเสถียรประกอบไปด้วยการแจกแจงที่สำคัญทั้งหมด 3 การแจกแจง ได้แก่

การแจกแจงเสถียรปกติ (Stable-normal distribution) $X\sim N\left(\delta,\gamma^2\right)$ มีฟังก์ชันความหนาแน่นความน่าจะเป็น

$$f\left(x\right)=\frac{1}{\sqrt{2\pi\gamma^2}}e^{-\frac{(x-\delta)^2}{2\gamma^2}};-\infty<x<\infty,\ \delta\in R,\ \gamma\in R^{+}$$

เมื่อ $\alpha=2,\beta=0$

การแจกแจงโคชี (Cauchy distribution) $X\sim Cauchy\left(\gamma,\delta\right)$ มีฟังก์ชันความหนาแน่นความน่าจะเป็น

$$f\left(x\right)=\frac{\gamma}{\pi\left(\gamma^2+\left(x-\delta\right)^2\right)};-\infty<x<\infty,\ \delta\in R,\ \gamma\in R^{+}$$

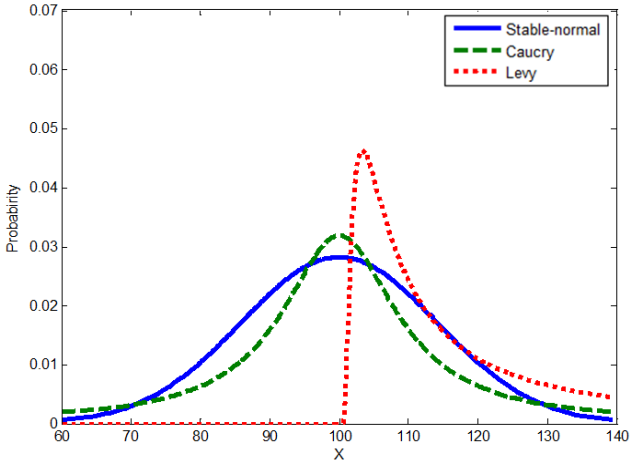
เมื่อ $\alpha=1,\beta=0$

การแจกแจงเลวี (Levy distribution) $X \sim Levy(\gamma,\delta)$ มีฟังก์ชันความหนาแน่นความน่าจะเป็น

$$f(x)=\left(\frac{\gamma}{2\pi}\right)^{\frac{1}{2}}\frac{1}{(x-\delta)^{\frac{3}{2}}}e^{\left(-\frac{\gamma}{2(x-\delta)}\right)};\delta<x<\infty,\ \delta\in R,\ \gamma\in R^{+}$$

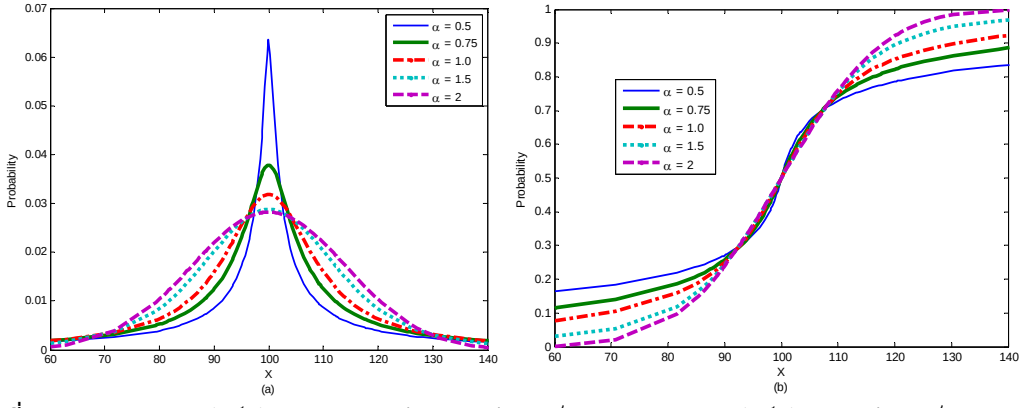
เมื่อ $\alpha=\frac{1}{2},\beta=1$

โดยลักษณะฟังก์ชันความหนาแน่นความน่าจะเป็นของการแจกแจงเสถียรปกติ การแจกแจงโคชี และการแจกแจงเลวี แสดงได้ดังรูปที่ 1

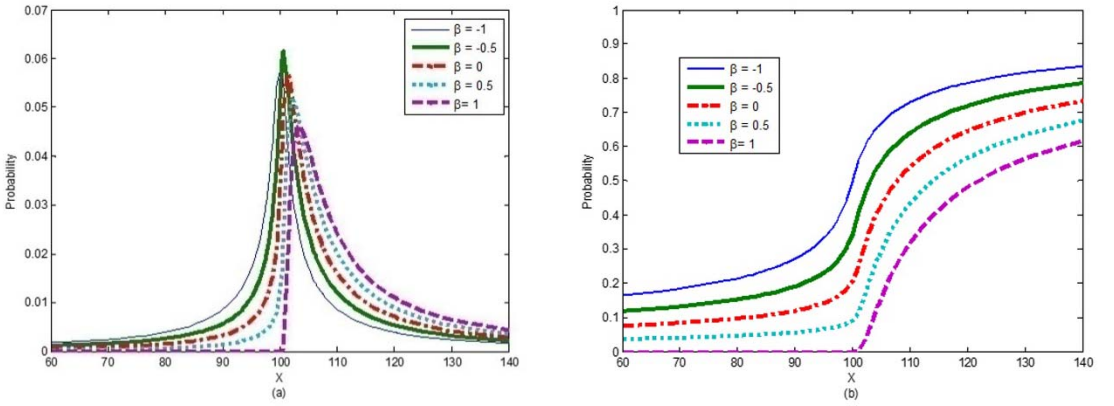


รูปที่ 1 ฟังก์ชันความหนาแน่นความน่าจะเป็นของการแจกแจงเสถียรปกติ การแจกแจงโคชี และการแจกแจงเลวี

เมื่อมีการปรับค่าพารามิเตอร์ α และ β ของการแจกแจงเสถียรจะทำให้รูปแบบฟังก์ชันความหนาแน่นความน่าจะเป็นและฟังก์ชันความน่าจะเป็นสะสมเปลี่ยนแปลงไป ดังรูปที่ 2 และ 3



รูปที่ 2 กราฟ (a) ฟังก์ชันความหนาแน่นความน่าจะเป็น และ (b) ฟังก์ชันความน่าจะเป็นสะสม เมื่อเปลี่ยนแปลงพารามิเตอร์ α



รูปที่ 3 กราฟ (a) ฟังก์ชันความหนาแน่นความน่าจะเป็น และ (b) ฟังก์ชันความน่าจะเป็นสะสม เมื่อเปลี่ยนแปลงพารามิเตอร์ β

2. ทฤษฎีเอ็มพิริคัลเบส์ (Empirical Bayes Theorem)

ทฤษฎีเอ็มพิริคัลเบส์ถูกนำมาประยุกต์ใช้ครั้งแรกโดย Robbins ใน ค.ศ. 1955 (Carlin and Louis, 2009) โดยหลักของทฤษฎีเอ็มพิริคัลเบส์มีความแตกต่างจากทฤษฎีเบส์ เนื่องจากทฤษฎีเบส์มักจะทราบค่าไฮเปอร์พารามิเตอร์ หรือถ้าไม่ทราบค่าไฮเปอร์พารามิเตอร์จะนำความรู้เดิมมาช่วยประเมินค่า (Assessment) ไฮเปอร์พารามิเตอร์ดังกล่าว โดยไม่เกี่ยวข้องกับข้อมูลที่ศึกษา ส่วนทฤษฎีเอ็มพิริคัลเบส์นั้นจะทำการประมาณค่าไฮเปอร์พารามิเตอร์ โดยนำข้อมูลที่ศึกษามาช่วยในการหาค่าประมาณไฮเปอร์พารามิเตอร์ ซึ่งจะต้องคำนวณหาฟังก์ชันการแจกแจงส่วนริมภายหลัง (Posterior marginal distribution function) ของข้อมูลที่ศึกษาก่อน แล้วจึงนำการแจกแจงส่วนริมภายหลังไปหาค่าประมาณไฮเปอร์พารามิเตอร์ด้วยวิธีแบบเดิม เช่น วิธีภาวะน่าจะเป็นสูงสุดหรือวิธีโมเมนต์ เป็นต้น แล้วจึงนำค่าประมาณไฮเปอร์พารามิเตอร์ที่ได้ไปแทนในฟังก์ชันการแจกแจงภายหลังที่คำนวณได้ตามวิธีการของเบส์

เมื่อ Θ เป็นตัวแปรสุ่มชนิดไม่ต่อเนื่อง การหาค่าประมาณไฮเปอร์พารามิเตอร์ด้วยวิธีเอ็มพิริคัลเบส์หาได้จากฟังก์ชันการแจกแจงส่วนริมภายหลัง คือ

$$m\left(\underline{x}\,|\,\delta\right)=\sum_{\theta} f\left(x\,|\,\theta\right)h\left(\theta\,|\,\delta\right)$$

และเมื่อ Θ เป็นตัวแปรสุ่มชนิดต่อเนื่อง การหาค่าประมาณไฮเปอร์พารามิเตอร์ด้วยวิธีเอ็มพิริคัลเบส์หาได้จากฟังก์ชันการแจกแจงส่วนริมภายหลัง คือ

$$m\left(\underline{x}\,|\,\delta\right)=\int_{\theta} f\left(x\,|\,\theta\right)h\left(\theta\,|\,\delta\right)d\theta$$

โดยที่ $m\left(\theta\,|\,\underline{x},\delta\right)$ แทน ฟังก์ชันการแจกแจงส่วนริมภายหลัง

3. วิธีเอ็มพิลิคัลเบส์ผสมผสานกับเนียเรสเนเบอร์ (EBNN method)

ให้ $X_1, X_2, ..., X_n$ เป็นตัวแปรสุ่มที่มีการแจกแจงแบบปกติ ที่ไม่ทราบค่าเฉลี่ยแต่ทราบความแปรปรวน โดย $X_i \sim N\left(\theta,\sigma_0^2\right)$ เมื่อ θ และ σ_0^2 ถูกกำหนดให้ไม่ทราบค่าเฉลี่ยแต่ทราบความแปรปรวน

ตามลำดับ และกำหนดให้ข้อมูลก่อน (Informative prior) คือ θ โดย $\theta \sim N(\mu, \tau^2)$ เมื่อ μ และ τ^2 แทนไฮเปอร์พารามิเตอร์ และจะได้ฟังก์ชันการแจกแจงส่วนริมภายหลัง ดังนี้

$$\begin{aligned} m(x|\mu,\tau^2) &= \int_{-\infty}^{\infty} f(x|\theta)\pi(\theta)\,d\theta \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(X-\theta)^2} \cdot \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(\theta-\mu)^2} d\theta \\ &= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{\tau^2+\sigma_0^2}} e^{-\frac{1}{2(\tau^2+\sigma_0^2)}(x-\mu)^2} \end{aligned}$$

โดยคำนวณฟังก์ชันการแจกแจงภายหลัง ดังนี้

$$\pi(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\pi(\theta)}{\int_{-\infty}^{\infty} f(\underline{x}|\theta)\pi(\theta)d\theta}$$

$$\pi(\theta|\underline{x},\mu,\tau^2) = \frac{\frac{1}{(2\pi)^{(n+1)/2}} \cdot \frac{1}{(\sigma_0^2)^{n/2}(\tau^2)^{1/2}} e^{-\frac{1}{2\tau^2\sigma_0^2}\left[\tau^2\sum_{i=1}^{n}x_i^2+\sigma_0^2\mu^2-\frac{\left(n\tau^2\bar{x}+\mu\sigma_0^2\right)^2}{n\tau^2+\sigma_0^2}+\left(n\tau^2+\sigma_0^2\right)\left(\theta-\frac{n\tau^2\bar{x}+\mu\sigma_0^2}{n\tau^2+\sigma_0^2}\right)^2\right]}}{\int_{-\infty}^{\infty} \frac{1}{(2\pi)^{(n+1)/2}} \cdot \frac{1}{(\sigma_0^2)^{n/2}(\tau^2)^{1/2}} e^{-\frac{1}{2\tau^2\sigma_0^2}\left[\tau^2\sum_{i=1}^{n}x_i^2+\sigma_0^2\mu^2-\frac{\left(n\tau^2\bar{x}+\mu\sigma_0^2\right)^2}{n\tau^2+\sigma_0^2}+\left(n\tau^2+\sigma_0^2\right)\left(\theta-\frac{n\tau^2\bar{x}+\mu\sigma_0^2}{n\tau^2+\sigma_0^2}\right)^2\right]}d\theta}$$

$$= \left[\frac{e^{-\frac{1}{2\tau^2\sigma_0^2}\left[\tau^2\sum_{i=1}^{n}x_i^2+\sigma_0^2\mu^2-\frac{\left(n\tau^2\bar{x}+\mu\sigma_0^2\right)^2}{n\tau^2+\sigma_0^2}+\left(n\tau^2+\sigma_0^2\right)\left(\theta-\frac{n\tau^2\bar{x}+\mu\sigma_0^2}{n\tau^2+\sigma_0^2}\right)^2\right]}}{(2\pi)^{(n+1)/2}\cdot(\sigma_0^2)^{n/2}(\tau^2)^{1/2}}\right]$$

$$\times\left[\frac{(2\pi)^{(n+1)/2}\cdot(\sigma_0^2)^{n/2}(\tau^2)^{1/2}}{\left(2\pi\left(\frac{\tau^2\sigma_0^2}{n\tau^2+\sigma_0^2}\right)\right)^{1/2}\cdot e^{-\frac{\tau^2\sum_{i=1}^{n}x_i^2+\sigma_0^2\mu^2-\frac{\left(n\tau^2\bar{x}+\mu\sigma_0^2\right)^2}{n\tau^2+\sigma_0^2}}{2\tau^2\sigma_0^2}}}\right]$$

ดังนั้นจะได้ฟังก์ชันการแจกแจงภายหลังคือ

$$\theta|\underline{X} \sim N\left(\frac{n\hat{\tau}^2\bar{X}+\hat{\mu}\sigma_0^2}{n\hat{\tau}^2+\sigma_0^2}, \frac{\hat{\tau}^2\sigma_0^2}{n\hat{\tau}^2+\sigma_0^2}\right)$$

โดยคำนวณความน่าจะเป็นพยากรณ์ภายหลัง (Posterior Predictive Probability) ดังนี้

$$p(y_j \mid \underline{x}, \theta) = \int f(y_j \mid \theta)\pi(\theta \mid \underline{x})\, d\theta$$

และประมาณค่า $p(y_j \mid \underline{x}, \theta)$ ด้วยเทคนิคมาร์คอฟ เชน มอนติคาร์โล (Markov Chain Monte Carlo: MCMC) (Everson and Fieldsend, 2004; Guo and Chakraborty, 2009) ได้ดังนี้

$$\hat{p}(y_j \mid \underline{x}, \theta) \approx \frac{1}{T}\sum_{i=1}^{T} p(y_j \mid \underline{x}, \theta^{(t)})$$

ดังนั้นความน่าจะเป็นพยากรณ์ภายหลังคือ

$$\hat{p}_q = \hat{p}_q(y_j \mid \underline{x}, \theta);\ q = 1, 2, \ldots, Q$$

วิธีเค-เนียเรสเนเบอร์จะถูกประยุกต์ใช้เมื่อต้องการจัดค่าสังเกตค่าใหม่ในชุดข้อมูลทดสอบ โดยพิจารณาจากค่า $\overline{P}_q = (s_j \mid L, k)$ ที่มากที่สุด ดังนี้

$$\overline{P}_q = (s_j \mid L, k) = (1/k)\sum\nolimits_{i \sim j}^{k} \delta_{(q)(s_j)}$$

เมื่อ $\sum_{i \sim j}^{k} \delta_{(q)(s_j)}$ แทน จำนวน s_i ของกลุ่ม q ภายในบริเวณใกล้เคียง k เมื่อ $\delta_{(q)(s_j)}$ แทนฟังก์ชันดิแรก (Dirac function) ซึ่งถูกกำหนดโดย

$$\delta_{(q)(s_j)} = \begin{cases} 1, & q = s_j \\ 0, & \text{กรณีอื่น} \end{cases}$$

ซึ่งในงานวิจัยนี้จะแสดงการหาฟังก์ชันการแจกแจงภายหลังของข้อมูลที่มีการแจกแจงเสถียรปกติเท่านั้น ส่วนในกรณีที่ข้อมูลมีการแจกแจงโคชีและการแจกแจงเลวีจะใช้โปรแกรมในการหาค่าประมาณ

4. อัลกอริทึมของกระบวนการ (Algorithm of the Process)

โดยกระบวนการทั้งหมดที่ใช้ในการวิจัย แสดงได้ดังตารางที่ 2

ตารางที่ 2 อัลกอริทึมของกระบวนการวิจัย

อัลกอริทึม

กำหนดชุดข้อมูลเรียนรู้ $(\underline{x})$

กำหนดชุดข้อมูลทดสอบ $(\underline{y})$

กำหนดจำนวนตัวอย่าง (n)

กำหนดจำนวนรอบกระทำซ้ำ (N)

กำหนดจำนวนค่าสังเกต (k) ที่อยู่ใกล้ค่าสังเกตค่าใหม่

For $n = 100, 500$ กระทำซ้ำ

สุ่มตัวอย่างจากชุดข้อมูลเรียนรู้

สุ่มตัวอย่างจากชุดข้อมูลทดสอบ

For $i = 1, 2, ..., N$ กระทำซ้ำ

หาค่าประมาณไฮเปอร์พารามิเตอร์ (ผลการคำนวณ)

หาค่าฟังก์ชันการแจกแจงภายหลัง (ผลการคำนวณ)

สร้างคลังข้อมูลของพารามิเตอร์ (ผลการคำนวณ)

เรียนรู้โดยใช้เทคนิคมาร์คอฟ เชน มอนติคาร์โล

อัลกอรีทึมสำหรับวิธี EB (ผลการคำนวณ)

อัลกอรีทึมสำหรับวิธี EBNN (ผลการคำนวณ)

จัดกลุ่มให้กับค่าสังเกตค่าใหม่ในชุดข้อมูลทดสอบโดยจัดให้อยู่กลุ่มที่มีความ

น่าจะเป็นพยากรณ์ภายหลังสูงที่สุด (ผลการคำนวณ)

End

คำนวณค่าเฉลี่ยของเปอร์เซ็นต์การจัดกลุ่มถูกต้อง (ผลการคำนวณ)

End

5. เกณฑ์การตัดสินใจ (Decision Criteria)

5.1 เปอร์เซ็นต์การจัดกลุ่มถูกต้อง จะพิจารณาประสิทธิภาพของการจัดกลุ่มโดยคำนวณจากเปอร์เซ็นต์การจัดกลุ่มถูกต้อง ดังนี้

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

เมื่อ Cr แทน จำนวนข้อมูลที่มีการจัดกลุ่มถูกต้องในชุดข้อมูลทดสอบ

$Tstotal$ แทน จำนวนข้อมูลทั้งหมดของชุดข้อมูลทดสอบ

5.2 ประสิทธิภาพสัมพัทธ์ (Relative Efficiency; RE) จะพิจารณาจากเปอร์เซ็นต์การจัดกลุ่มถูกต้องของวิธี
EBNN เปรียบเทียบกับวิธี EB ดังนี้

$$RE = \frac{EBNN}{EB}$$

เมื่อ $EBNN$ แทน ค่าเฉลี่ยของเปอร์เซ็นต์การจัดกลุ่มถูกต้องโดยใช้วิธี EBNN

EB แทน ค่าเฉลี่ยของเปอร์เซ็นต์การจัดกลุ่มถูกต้องโดยใช้วิธี EB

เกณฑ์ที่ใช้วัดคือ

ถ้า RE มีค่ามากกว่า 1 หมายความว่า วิธี EBNN มีประสิทธิภาพมากกว่าวิธี EB

ถ้า RE มีค่าน้อยกว่า 1 หมายความว่า วิธี EBNN มีประสิทธิภาพน้อยกว่าวิธี EB

ถ้า RE มีค่าเท่ากับ 1 หมายความว่า วิธี EBNN และวิธี EB มีประสิทธิภาพเท่ากัน

ผลการวิจัย

จากวิธีดำเนินการวิจัย ผู้วิจัยได้ศึกษาประสิทธิภาพวิธีการจัดกลุ่มในแต่ละสถานการณ์ที่กำหนดตามขอบเขตของการวิจัย ซึ่งผลการจัดกลุ่มด้วยวิธี EB และวิธี EBNN จะแสดงในตารางที่ 3 และตารางที่ 4 ตามลำดับ โดยตารางที่ 5 จะแสดงค่าประสิทธิภาพสัมพัทธ์ระหว่างวิธี EB และวิธี EBNN

ตารางที่ 3 ค่าเฉลี่ยของเปอร์เซ็นต์การจัดกลุ่มถูกต้องด้วยวิธี EB

การแจกแจงของข้อมูล	**ขนาดตัวอย่าง (n)**	
	100	**500**
การแจกแจงเสถียรปกติ	95.62	95.79
การแจกแจงโคชี	79.91	81.43
การแจกแจงเลวี	60.15	60.02

จากตารางที่ 3 เป็นการแสดงผลการจัดกลุ่มข้อมูลด้วยวิธี EB ซึ่งจากตารางพบว่าข้อมูลที่มีการแจกแจงเสถียรปกติให้ผลการจัดกลุ่มที่ดีทั้งสองขนาดตัวอย่าง และเมื่อขนาดตัวอย่างเพิ่มขึ้นส่งผลให้การจัดกลุ่มดีขึ้นเมื่อข้อมูลมีการแจกแจงเสถียรปกติและการแจกแจงโคชี

ตารางที่ 4 ค่าเฉลี่ยของเปอร์เซ็นต์การจัดกลุ่มถูกต้องด้วยวิธี EBNN

การแจกแจงของข้อมูล	k	ขนาดตัวอย่าง (n)	
		100	500
การแจกแจงเสถียรปกติ	3	97.01	97.34
	5	97.06	97.37
	7	97.13	97.47
	9	97.16	97.57
การแจกแจงโคชี	3	83.48	83.05
	5	84.55	84.34
	7	84.95	84.59
	9	84.85	84.96
การแจกแจงเลวี	3	61.69	62.33
	5	62.13	62.50
	7	62.44	62.63
	9	62.84	62.81

จากตารางที่ 4 เป็นการแสดงผลการจัดกลุ่มข้อมูลด้วยวิธี EBNN ซึ่งเมื่อขนาดตัวอย่างคงที่พบว่าเปอร์เซ็นต์การจัดกลุ่มข้อมูลถูกต้องมีแนวโน้มสูงขึ้นเมื่อค่า k เพิ่มมากขึ้นในทุกการแจกแจงของข้อมูล และเมื่อค่า k คงที่พบว่าเปอร์เซ็นต์การจัดกลุ่มข้อมูลถูกต้องมีแนวโน้มสูงขึ้นเมื่อขนาดตัวอย่างเพิ่มขึ้นในกรณีที่ข้อมูลมีการแจกแจงเสถียรปกติและการแจกแจงเลวี แต่เมื่อข้อมูลมีการแจกแจงโคชีพบว่าเปอร์เซ็นต์การจัดกลุ่มข้อมูลถูกต้องมีแนวโน้มลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้นในกรณีที่ k = 3, 5, 7 และเมื่อ k = 9 พบว่าเปอร์เซ็นต์การจัดกลุ่มข้อมูลถูกต้องมีแนวโน้มสูงขึ้นเมื่อขนาดตัวอย่างเพิ่มขึ้นเช่นเดียวกับกรณีข้อมูลมีการแจกแจงเสถียรปกติและการแจกแจงเลวี เมื่อข้อมูลมีการแจกแจงเสถียรปกติพบว่าเปอร์เซ็นต์การจัดกลุ่มข้อมูลถูกต้องสูงสุดมีค่าเท่ากับ 97.57 เมื่อใช้ขนาดตัวอย่างเท่ากับ 500 เมื่อข้อมูลมีการแจกแจงโคชีพบว่าเปอร์เซ็นต์การจัดกลุ่มข้อมูลถูกต้องสูงสุดมีค่าเท่ากับ 84.96 เมื่อใช้ขนาดตัวอย่างเท่ากับ 500 และเมื่อข้อมูลมีการแจกแจงเลวีพบว่าเปอร์เซ็นต์การจัดกลุ่มข้อมูลถูกต้องสูงสุดมีค่าเท่ากับ 62.84 เมื่อใช้ขนาดตัวอย่างเท่ากับ 100 และเมื่อพิจารณาค่าประสิทธิภาพสัมพัทธ์ในตารางที่ 5 พบว่าวิธี EBNN ให้ผลการจัดกลุ่มดีกว่าวิธี EB ในทุกสถานการณ์ที่ศึกษา

ตารางที่ 5 ค่าประสิทธิภาพสัมพัทธ์ระหว่างวิธี EB และวิธี EBNN

การแจกแจงของข้อมูล	k	ขนาดตัวอย่าง (n)	
		100	500
การแจกแจงเสถียรปกติ	3	1.0145	1.0162
	5	1.0151	1.0165
	7	1.0158	1.0175
	9	1.0161	1.0186
การแจกแจงโคชี	3	1.0447	1.0199
	5	1.0581	1.0357
	7	1.0631	1.0388
	9	1.0618	1.0434
การแจกแจงเลวี	3	1.0256	1.0385
	5	1.0329	1.0413
	7	1.0381	1.0435
	9	1.0447	1.0465

อภิปรายผล

จากผลการวิจัยพบว่า เมื่อพิจารณาทั้ง 3 การแจกแจง วิธี EBNN ให้ผลการจัดกลุ่มดีกว่าวิธี EB ในทุกขนาดตัวอย่าง ทั้งนี้อาจเป็นผลมาจากวิธี EBNN เกิดจากการผสมผสานวิธีเอ็มพิลิคัลเบส์ร่วมกับวิธีเค-เนียเรสเนเบอร์ ซึ่งวิธีเค-เนียเรสเนเบอร์เป็นวิธีการจัดกลุ่มแบบนอนพาราเมตริกที่ไม่มีข้อกำหนดเกี่ยวกับการแจกแจงของประชากร นั่นคือประชากรมีการแจกแจงแบบใดก็ได้ เมื่อเรานำมาจัดกลุ่มข้อมูลที่มีการแจกแจงแบบเสถียร วิธี EBNN จึงมีเปอร์เซ็นต์การจัดกลุ่มถูกต้องสูงกว่าวิธี EB ในทุกการแจกแจงที่ศึกษา นอกจากนี้ยังพบว่าเปอร์เซ็นต์การจัดกลุ่มข้อมูลถูกต้องมีแนวโน้มสูงขึ้นเมื่อค่า k เพิ่มขึ้นในทุกการแจกแจงของข้อมูล และกรณีข้อมูลมีการแจกแจงเสถียรปกติและการแจกแจงเลวี เมื่อขนาดตัวอย่างเพิ่มขึ้นเปอร์เซ็นต์การจัดกลุ่มข้อมูลถูกต้องมีแนวโน้มสูงขึ้นเช่นกัน

สรุปผลการวิจัย

เมื่อขนาดตัวอย่างเท่ากับ 100 และ 500 วิธี EB และวิธี EBNN ให้ผลการจัดกลุ่มถูกต้องค่อนข้างสูงเมื่อข้อมูลมีการแจกแจงเสถียรปกติและการแจกแจงโคชี แต่เมื่อข้อมูลมีการแจกแจงเลวีพบว่าเปอร์เซ็นต์การจัดกลุ่มข้อมูลถูกต้องอยู่ในระดับปานกลาง นอกจากนี้ยังพบว่าการจัดกลุ่มข้อมูลด้วยวิธี EBNN ที่ k = 9 จะให้ผลการจัด

กลุ่มถูกต้องสูงที่สุดในทุกการแจกแจงของข้อมูล และกรณีข้อมูลมีการแจกแจงเสถียรปกติและการแจกแจงเลวี พบว่าเปอร์เซ็นต์การจัดกลุ่มข้อมูลถูกต้องมีแนวโน้มสูงขึ้นเมื่อขนาดตัวอย่างเพิ่มขึ้น

กิตติกรรมประกาศ

ขอขอบคุณมหาวิทยาลัยราชภัฏพิบูลสงคราม ที่ให้การสนับสนุนทุนวิจัยนี้

เอกสารอ้างอิง

Aci, M., Inan, C. and Avci, M. (2010). A hybrid classification method of k nearest neighbor, Bayesian methods and genetic algorithm. Expert Systems with Applications 37: 5061-5067.

Carlin, B.P. and Louis, T.A. (2009). Bayesian Methods for Data Analysis. United States of America: Chamman & Hall.

Deetae, N., Sukparungsee, S., Areepong, Y. and Jampachaisri, K. (2013). The Combination of Empirical Bayes and Nearest Neighbor in Classification of Heavy tailed distribution. Far East Journal of Mathematical Sciences 77(2): 255-266.

Everson, R.M. and Fieldsend, J.E. (2004). A variable metric probabilistic k-nearest-neighbours classifier. In: Intelligent Data Engineering and Automated Learning-IDEAL. 654-659.

Ghosh, A.K. and Godtliebsen, F. (2011). On hybrid classification using model assisted posterior estimates. Pattern Recognition 45: 2288-2298

Guo, R. and Chakraborty, S. (2009). Bayesian Adaptive Nearest Neighbor. Statistical Analysis and Data Mining. 92-105.

Hachicha, W. and Ghorbel, A. (2012). A survey of control-chart pattern-recognition literature (1991–2010) based on a new conceptual classification scheme. Computers & Industrial Engineering 63: 204–222.

Johnson, R. and Wichern, D. (2002). Applied multivariate statistical analysis. United States of America: Prentice–Hall.

Ravi, A. and Butar, F. B. (2010). An Insight Into Heavy-Tailed Distribution. Journal of Mathematical Science & Mathematics Education 5: 19-31.

Sundaram, S. and McDonald, K. (2010). Stable Distributions for Heavy-Tailed Data and Their Application in Asset Health Monitoring. In: The Seventh International Conference on Condition Monitoring and Machinery Failure Prevention Technologies. 1-10.

❑❑❑❑❑