P-ISSN 0858-4435
E-ISSN 2651-1231
วารสารวิทยาศาสตร์และเทคโนโลยี Homepage : *http://li01.tci-thaijo.org/index.php/tstj*
ปีที่ 30 ฉบับที่ 5 (กันยายน-ตุลาคม 2565) หน้า [1-15]

การทดสอบประสิทธิภาพการแบ่งข้อมูลตัวแปรเดียวด้วยการใช้การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำ

A Performance Assessment of Repeated Jenks Natural Breaks Classification on Univariate Data

วิชญ์ยุตม์ สุขแพทย์*, นัท กุลวานิช
ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย กรุงเทพมหานคร 10330
Vichayud Sukphaet*, Nat Kulvanich
Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University, Bangkok 10330
Received 16 May 2022; Received in revised 15 August 2022; Accepted 23 August 2022

บทคัดย่อ

วิธีการแบ่งช่วงธรรมชาติเจงค์เป็นวิธีการจัดกลุ่มข้อมูลที่ได้รับความนิยม งานวิจัยนี้ได้นำวิธีการแบ่งช่วงธรรมชาติเจงค์มาปรับใช้ด้วยการเพิ่มจำนวนกลุ่มที่ใช้แบ่งเรื่อย ๆ จนกว่าจุดแบ่งแรกของกลุ่มใหม่จะเปลี่ยนแปลงไปน้อยกว่าค่าร้อยละที่กำหนดเมื่อเทียบกับจุดแบ่งแรกของการแบ่งกลุ่มครั้งก่อนหน้า และใช้จุดแบ่งแรกนั้นในการแบ่งข้อมูลออกเป็น 2 กลุ่ม โดยทำการศึกษากรณีข้อมูลตัวแปรเดียวที่มีการแจกแจงในรูปแบบการแจกแจงปรกติแบบผสม 2 กลุ่มและการแจกแจงล็อกปรกติแบบผสม 2 กลุ่ม โดยเปรียบเทียบกับวิธีการแบ่งกลุ่มข้อมูลอื่น ๆ ได้แก่ วิธีการแบ่งช่วงธรรมชาติเจงค์, วิธี head/tail break และ วิธีจัดกลุ่มข้อมูลด้วยอัลกอริทึม EM การวัดประสิทธิภาพของวิธีการแบ่งกลุ่มพิจารณาจากค่าความแม่นในการจัดกลุ่ม ผลการวิจัยจากการจำลองข้อมูล พบว่าวิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำนั้นไม่มีประสิทธิภาพในการแบ่งข้อมูลแจกแจงปรกติแบบผสมแต่มีประสิทธิภาพในการแบ่งข้อมูลที่มีการแจกแจงล็อกปรกติแบบผสมเมื่อข้อมูล 2 กลุ่มมีจำนวนใกล้เคียงกันหรือกลุ่มที่ค่าเฉลี่ยสูงกว่ามีจำนวนมากกว่า นอกจากนี้วิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำยังสามารถในไปใช้ในการแบ่งข้อมูลเมื่อผู้ใช้ต้องการให้ความสำคัญด้านความแม่นของข้อมูลกลุ่มที่มีค่าเฉลี่ยสูงกว่าได้

คำสำคัญ: วิธีการจัดกลุ่ม; การแบ่งช่วงธรรมชาติเจงค์; การแจกแจงปรกติแบบผสม; การแจกแจงล็อกปรกติแบบผสม

***ผู้รับผิดชอบบทความ:** vichayud@gmail.com **doi:**

Abstract

Jenks natural breaks classification is a data clustering method that is widely used. This research uses a modified version of Jenks natural breaks classification by increasing the number of groups used for clustering until the change of the first break is less than the specified percentage compared to the previous clustering. The first break is then used to split the data into two groups. We perform a performance assessment of repeated Jenks natural breaks classification against Jenks natural breaks classification, head/tail break, and EM algorithm using 2-group normal mixture distribution and 2-group log-normal mixture distribution univariate simulated data. The performance is asserted by using clustering accuracy. The research found that repeated Jenks natural breaks classification is not suitable for maximizing the overall accuracy of the normal mixture distribution but can be used for log-normal mixture distribution if the proportion of each group is relatively equal or higher-mean group leaning. Repeated Jenks natural breaks classification can also be used if users need to prioritize the accuracy of the higher-mean group.

Keywords: Clustering method; Jenks natural breaks classification; Normal mixture distribution; Log-normal mixture distribution

1. บทนำ

วิธีการแบ่งกลุ่มข้อมูล (data clustering method) เป็นเครื่องมือที่ใช้ในการแบ่งข้อมูลออกเป็นหลายๆ กลุ่ม โดยจุดมุ่งหมายหลักคือเพื่อจัดข้อมูลที่มีลักษณะใกล้เคียงกันไว้ในกลุ่มเดียวกันหรือแบ่งข้อมูลออกเป็นกลุ่มที่ถูกต้อง การเลือกวิธีการแบ่งกลุ่มข้อมูลที่เหมาะสมกับข้อมูลจึงมีความสำคัญเพื่อให้ข้อมูลที่ถูกแบ่งนั้นแสดงลักษณะของข้อมูลได้ดีหรือสามารถแบ่งกลุ่มของข้อมูลได้อย่างถูกต้อง

วิธีการแบ่งช่วงธรรมชาติเจงค์ (Jenks natural breaks classification) เป็นวิธีการจัดกลุ่มข้อมูลเสนอโดย George F. Jenks [1] ซึ่งเป็นที่ได้รับความนิยมโดยเฉพาะอย่างยิ่งในการสร้างแผนที่โคโรเพลท (choropleth map) เนื่องจากวิธีการแบ่งนี้อยู่ในโปรแกรมระบบสารสนเทศภูมิศาสตร์ (geographic information system software) หลายโปรแกรม โดยการจัดกลุ่มรูปแบบนี้สามารถใช้ได้กับข้อมูลตัวแปรเดียว (univariate) และจะให้กลุ่มข้อมูลที่มีความแปรปรวนภายในกลุ่มต่ำที่สุด กลุ่มที่แบ่งนั้นจะมีความเหมือนกับการแบ่งกลุ่มด้วยวิธี k-means เมื่อใช้การแบ่งกลุ่มด้วยวิธี k-means เพื่อแบ่งกลุ่มข้อมูลตัวแปรเดียว

ในงานวิจัยซึ่งเปรียบเทียบการแบ่งกลุ่มด้วยวิธี k-means และวิธี gaussian mixture model (GMM) ที่ใช้อัลกอริทึม EM (expectation-maximization algorithm) นั้น วิธีอัลกอริทึม EM มีประสิทธิภาพในการแบ่งข้อมูลได้ดีกว่าในกรณีของข้อมูลการแจกแจงปรกติแบบผสม (gaussian หรือ normal mixture distribution) ในแทบทุกกรณี ยกเว้นเมื่อข้อมูลมีจำนวนน้อยหรือข้อมูลทั้งสองกลุ่มมีจำนวนใกล้เคียงกัน [2]

ในงานวิจัยอื่นๆ เช่น Wang และคณะ [3] หรือ Patel และ Kushwaha [4] ซึ่งมีการเปรียบเทียบการแบ่งกลุ่มสองรูปแบบในข้อมูลตัวแปรเดียว งานวิจัยค่อนข้างสนใจในการแสดงลักษณะของข้อมูลมากกว่าการจัดกลุ่มข้อมูลที่ถูกต้อง นอกจากนี้ ยังไม่มีงานวิจัยที่ทดสอบเปรียบเทียบการแบ่งกลุ่มในการแจงแจงรูปแบบอื่น

งานวิจัยนี้ได้นำวิธีการแบ่งช่วงธรรมชาติเจงค์มาปรับใช้ โดยเพิ่มจำนวนกลุ่มที่แบ่งของวิธีการแบ่งช่วงธรรมชาติเจงค์ เรื่อยๆ จนกว่าจุดแบ่งแรกนั้นจะเปลี่ยนแปลงไปน้อยกว่าร้อยละที่กำหนดและใช้จุดแบ่งแรกในการแบ่งข้อมูลออกเป็นสองกลุ่ม โดยงานวิจัยนี้ได้รับแรงบัลดาลใจจากการใช้วิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำ (repeated Jenks natural breaks classification) เพื่อแบ่งกลุ่มลูกค้าออกเป็น 2 กลุ่มในบริษัท e-commerce แห่งหนึ่ง

การแบ่งกลุ่มข้อมูลด้วยวิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำนั้นมีความน่าสนใจและอาจสามารถนำไปใช้แบ่งข้อมูลได้ดี โดยเฉพาะในข้อมูลที่มีลักษณะการแจกแจงเบ้ขวา (right-skewed distribution) งานวิจัยนี้จึงได้ทดสอบประสิทธิภาพของวิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำด้วยการจำลองข้อมูลที่มีลักษณะการแจกแจงที่ต่างกัน ได้แก่การแจกแจงปรกติแบบผสม และการแจกแจงล็อกปรกติแบบผสม (log-normal mixture distribution) โดยเทียบกับการแบ่งกลุ่มรูปแบบอื่น 3 วิธี ได้แก่ วิธีการแบ่งช่วงธรรมชาติเจงค์, วิธี head/tail breaks และ วิธี mixture model ที่ใช้อัลกอริทึม EM ในการประมาณค่าพารามิเตอร์

2. วิธีการวิจัย

2.1 รูปแบบการแจกแจงที่ศึกษาและวิธีการแบ่งกลุ่มข้อมูลที่เกี่ยวข้อง

2.1.1. การแจกแจงผสม (mixture distribution)

การแจกแจงผสมคือการแจกแจงความน่าจะเป็นของตัวแปรสุ่มซึ่งเกิดจากการผสมกันของการแจกแจงความน่าจะเป็น 2 การแจกแจงเป็นต้นไป โดยตัวแปรสุ่มนั้นถูกเลือกจากการแจกแจงที่ผสมด้วยความน่าจะเป็น โดยสามารถเขียนในรูปแบบฟังก์ชันการแจกแจงความน่าจะเป็น (probability distribution function หรือ pdf) ได้คือ

$$g(x) = \sum_{j=1}^{k} \omega_j f_j(x)$$

เมื่อ $g(x)$ คือฟังก์ชันการแจกแจงความน่าจะเป็นของการแจกแจงผสม; k คือจำนวนของการแจกแจงที่ผสม; ω_j คือความน่าจะเป็นของการแจกแจงที่ j ที่จะถูกเลือก โดยมีเงื่อนไข $\omega_j \geq 0$ และ $\sum_{j=1}^{k} \omega_j = 1$; $f_j(x)$ คือฟังก์ชันการแจกแจงความน่าจะเป็นของตัวแปรสุ่มที่ j

2.1.2 การแจกแจงปรกติแบบผสม (normal mixture distribution)

การแจกแจงปรกติแบบผสมคือการแจกแจงผสมที่เกิดจากการผสมกันของการแจกแจงปรกติ (normal distribution) 2 การแจกแจงเป็นต้นไป โดยฟังก์ชันการแจกแจงความน่าจะเป็นของการแจกแจงปรกติคือ

$$f_j(x) = \frac{1}{\sigma_j\sqrt{2\pi}} e^{-\frac{(x_j-\mu_j)^2}{2\sigma_j^2}}$$

เมื่อ μ_j คือค่าเฉลี่ยของตัวแปรสุ่ม x_j; σ_j คือค่าส่วนเบี่ยงเบนมาตรฐานของตัวแปรสุ่ม x_j

โดยการแจกแจงผสมที่เกิดจากการแจกแจงปรกติ 2 การแจกแจง จะมีลักษณะเป็นฐานนิยมเดียวหรือมีจุดยอดในการแจกแจงเพียงจุดเดียวก็ต่อเมื่อ

$$\left|\mu_1 - \mu_2\right| \leq 2\min(\sigma_1, \sigma_2)$$

หรือถ้าหาก $\sigma_1 = \sigma_2 = \sigma$ การแจกแจงจะมีลักษณะเป็นฐานนิยมเดียว ก็ต่อเมื่อ

$$\left|\mu_1 - \mu_2\right| \leq 2\sigma\sqrt{1 + \frac{\left|\log(\omega_1) - \log(\omega_2)\right|}{2}}$$

เมื่อ ω_1 และ ω_2 คือความน่าจะเป็นของการแจกแจงที่ 1 และ 2 ตามลำดับ โดยที่ $\omega_2 = 1 - \omega_1$ [5]

2.1.3 การแจกแจงล็อกปรกติแบบผสม (log-normal mixture distribution)

การแจกแจงล็อกปรกติแบบผสมคือการแจกแจงผสมที่เกิดจากการผสมกันของการแจกแจงล็อกปรกติ (log-normal distribution) 2 การแจกแจงเป็นต้นไป โดยฟังก์ชันการแจกแจงความน่าจะเป็นของการแจกแจงล็อกปรกติคือ

$$f_j(x) = \frac{1}{x_j\sigma_j\sqrt{2\pi}} e^{-\frac{(\ln(x_j)-\mu_j)^2}{2\sigma_j^2}}$$

โดยที่ μ_j คือค่าเฉลี่ยของลอการิทึมตัวแปรสุ่ม x_j; σ_j คือค่าส่วนเบี่ยงเบนมาตรฐานของลอการิทึมตัวแปรสุ่ม x_j

ค่าเฉลี่ยของตัวแปรสุ่ม x_j มีค่าเท่ากับ $e^{\mu_j+\frac{\sigma_j^2}{2}}$ และความแปรปรวนของตัวแปรสุ่ม x_j มีค่าเท่ากับ $(e^{\sigma_j^2}-1)e^{2\mu_j+\sigma_j^2}$

โดยการแจกแจงผสมที่เกิดจากการแจกแจงล็อกปรกติ 2 การแจกแจงจะมีลักษณะเป็นเป็นฐานนิยมเดียวก็ต่อเมื่อการแจกแจงมีค่าพารามิเตอร์ σ เท่ากัน

2.1.4 การแบ่งกลุ่มด้วยวิธีการแบ่งช่วงธรรมชาติเจงค์ (Jenks natural breaks classification)

วิธีการแบ่งช่วงธรรมชาติเจงค์เป็นวิธีการแบ่งกลุ่มที่ให้กลุ่มซึ่งมีผลรวมของความเบี่ยงเบนจากค่าเฉลี่ยของกลุ่มกำลังสองของแต่ละกลุ่ม (squared deviations from the class means หรือ SDCM) น้อยที่สุดโดยเขียนเป็นสมการได้คือ

$$SDCM = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2$$

เมื่อ k คือจำนวนกลุ่มที่ต้องการจะแบ่ง; n_j คือจำนวนข้อมูลในกลุ่ม j; x_{ij} คือข้อมูลตัวที่ i ในกลุ่ม j; $\bar{x}_j$ คือค่าเฉลี่ยของข้อมูลในกลุ่ม j

โดยวิธีการแบ่งช่วงธรรมชาติเจงค์ใช้การทดลองแบ่งกลุ่มข้อมูลทุกรูปแบบที่เป็นไปได้เพื่อหาการแบ่งกลุ่มที่ให้ SDCM น้อยที่สุด นอกจากนี้ผู้ใช้จำเป็นต้องระบุจำนวนกลุ่มที่ต้องการแบ่งเอง

2.1.5 การแบ่งกลุ่มด้วยวิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำ (repeated Jenks natural breaks classification)

วิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำ เป็นวิธีการแบ่งกลุ่มที่ใช้วิธีการแบ่งช่วงธรรมชาติเจงค์ซ้ำหลายๆ ครั้ง โดยแต่ละครั้งจะเพิ่มจำนวนกลุ่มที่แบ่งเรื่อยๆ จนกว่าจุดแบ่งแรก (จุดแบ่งที่ใช้ในการแบ่งกลุ่มแรกและกลุ่มที่สอง) ของการแบ่งครั้งล่าสุดที่ใช้จำนวนกลุ่มในการแบ่งคือ k ($break_k$) จะแตกต่างจากจุดแบ่งแรกของการแบ่งครั้งที่แล้ว ($break_{k-1}$) ไม่เกินค่าร้อยละที่กำหนดเมื่อเทียบกับจุดแบ่งแรกของการแบ่งครั้งที่แล้ว ซึ่งเขียนเป็นสมการได้คือ

$$\frac{break_{k-1}-break_k}{\left|break_{k-1}\right|} < perc$$

เมื่อ $break_k$ คือจุดแบ่งแรกของการแบ่งข้อมูลเป็น k กลุ่ม; $perc$ คือค่าร้อยละที่กำหนดเพื่อให้หยุดแบ่งเมื่อความแตกแต่งของจุดแบ่งที่เปลี่ยนไปน้อยกว่า โดยหลังจากการหยุดแล้วจะใช้จุดแบ่งแรกของการแบ่งข้อมูลครั้งล่าสุดในการแบ่งข้อมูลออกเป็น 2 กลุ่มคือกลุ่มที่มีค่าน้อยกว่าจุดแบ่ง และกลุ่มที่มีค่ามากกว่าจุดแบ่ง วิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำจึงใช้ได้กับการแบ่งข้อมูลออกเป็น 2 กลุ่มเท่านั้น

2.1.6 การแบ่งกลุ่มด้วยวิธี head/tails breaks

วิธี head/tail breaks เป็นวิธีการแบ่งกลุ่มที่เสนอโดย Bin Jiang [6] การแบ่งวิธีนี้ถูกสร้างขึ้นเพื่อแบ่งข้อมูลที่มีลักษณะการแจกแจงหางหนา (heavy-tailed distribution) ซึ่งคือข้อมูลที่มีค่ามากมีจำนวนน้อย และข้อมูลที่มีค่าน้อยมีจำนวนมาก เนื่องจากวิธีการแบ่งรูปแบบอื่น (รวมถึงวิธีการแบ่งช่วงธรรมชาติเจงค์) ไม่สามารถแสดงความแตกต่างของข้อมูลได้ดีพอ

โดยวิธีแบ่งกลุ่มนี้จะเริ่มต้นจากการแบ่งข้อมูลทั้งหมดที่ค่าเฉลี่ยของข้อมูล ซึ่งจะแบ่งข้อมูลออกเป็น 2 กลุ่มคือข้อมูลที่มีค่าน้อยกว่าค่าเฉลี่ยและข้อมูลที่มีค่ามากกว่าค่าเฉลี่ย หากกลุ่มข้อมูลที่มีค่ามากกว่าค่าเฉลี่ยมีจำนวนน้อยกว่าร้อยละที่กำหนดเมื่อเทียบกับข้อมูลทั้งหมดหลังแบ่งข้อมูล โดยปกติจะใช้ร้อยละ 40 จะทำการแบ่งข้อมูลต่อ การแบ่งต่อจะใช้วิธีเดิมกับขั้นตอนแรกคือการหาค่าเฉลี่ยของข้อมูล แต่ในครั้งที่ 2 นั้นจะใช้ข้อมูลที่มีค่ามากกว่าค่าเฉลี่ยของการแบ่งครั้งแรกเท่านั้น และในครั้งที่ 3 จะใช้ข้อมูลที่มีค่ามากกว่าค่าเฉลี่ยของการแบ่งครั้งที่ 2 และทำการแบ่งเรื่อยๆ จนกว่าข้อมูลหลังการแบ่งในครั้งล่าสุดมีจำนวนข้อมูลส่วนที่มากกว่าค่าเฉลี่ยของการแบ่งมากกว่าร้อยละที่กำหนดเมื่อเทียบกับ

ข้อมูลทั้งหมดที่ใช้ในการแบ่งครั้งล่าสุด จุดแบ่งที่ได้จากการแบ่งรูปแบบนี้คือค่าเฉลี่ยที่ใช้ในการแบ่งครั้งต่างๆ

นอกจากนี้จะสังเกตได้ว่าวิธี head/tail breaks ไม่จำเป็นต้องกำหนดจำนวนกลุ่มที่ใช้ในการแบ่งเอง

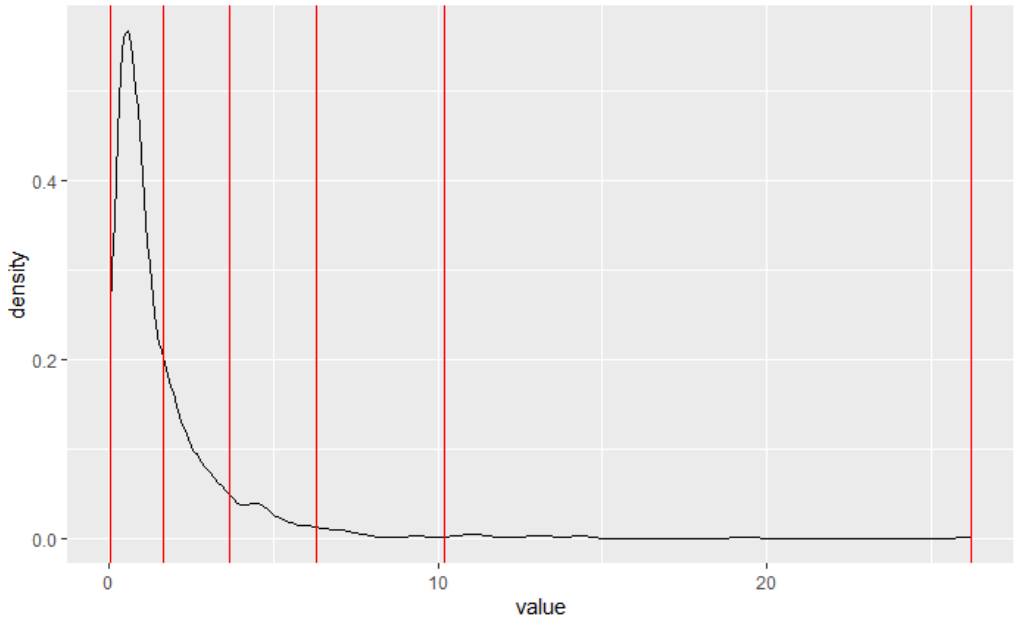

Figure 1 Example of heavy-tailed distribution and result of head/tail breaks on heavy-tailed distribution.

2.1.7 การจัดกลุ่มข้อมูลด้วยวิธีอัลกอริทึม EM

วิธีอัลกอริทึม EM เป็นการจัดกลุ่มข้อมูลโดยอาศัยการประมาณพารามิเตอร์ของการแจกแจงในแต่ละกลุ่ม โดยวิธีการประมาณค่าพารามิเตอร์นั้นเกิดจากการทำสลับกันระหว่าง 2 ขั้นตอนคือ การกำหนดกลุ่มให้กับจุดข้อมูล (expectation step หรือ E-step) และการประมาณค่าพารามิเตอร์จากจุดข้อมูลที่อยู่ในกลุ่มที่จัดในขั้นตอนที่แล้ว (maximization step หรือ M-step) หรือสามารถเขียนเป็นสมการได้ขั้นตอนได้ดังนี้

1. สุ่มค่า $\theta^{(m=0)}$

2. E-step: ประมาณค่าความน่าจะเป็นที่ข้อมูลจะอยู่ในแต่ละกลุ่ม $p\left(x\middle|y,\theta\right)$ โดยใช้พารามิเตอร์ครั้งที่ $\theta^{(m)}$ ในการประมาณค่า โดยจะได้ฟังก์ชัน Q คือ

$$Q\left(\theta\middle|\theta^{(m)}\right)=expected\log p(X|\theta)=E_{X|y,\theta^{(m)}}[\log p(X|\theta)]=\int\log p(x|\theta)p\left(x\middle|y,\theta^{(m)}\right)dx$$

เมื่อ y คือ ข้อมูลที่เก็บได้และ x คือข้อมูลที่สมบูรณ์

3. M-step: ประมาณค่า θ ที่ให้ค่าฟังก์ชัน Q สูงสุด

$$\theta^{(m+1)}=argmax_{\theta}Q\left(\theta\middle|\theta^{(m)}\right)$$

4. วนซ้ำขั้นตอนที่ 2 และ 3 จนกระทั่งลู่เข้า (converge) [7]

วิธีอัลกอริทึม EM เป็น soft clustering ซึ่งคือการแบ่งกลุ่มที่ให้ค่าความน่าจะเป็นที่แต่ละจุดข้อมูลจะอยู่ในกลุ่มใดกลุ่มหนึ่ง ในการวิจัยนี้จัดให้ข้อมูลที่มีอยู่ในกลุ่มที่มีค่าความน่าจะเป็นสูงสุด

2.2 การศึกษาเปรียบเทียบประสิทธิภาพด้วยวิธีการจำลองข้อมูล

งานวิจัยนี้ได้ใช้การจำลองข้อมูลเพื่อจำลองรูปแบบของข้อมูลต่างๆ ที่อาจพบเจอได้ในข้อมูลจริง โดยใช้โปรแกรม R ในการจำลองข้อมูลโดยมีวิธีดำเนินงานวิจัยดังนี้

2.2.1 สร้างข้อมูลตัวแปรเดียวจำนวน 5000 ข้อมูล จากรูปแบบพารามิเตอร์ที่กำหนด โดยมีรูปแบบการสร้าง 72 รูปแบบ ได้แก่ 36 รูปแบบการแจกแจงปรกติแบบผสม และ 36 รูปแบบการแจกแจงล็อกปรกติแบบผสม โดยในแต่รูปแบบจะทำการทดลองซ้ำ 25 รอบเพื่อหาค่าเฉลี่ยของความแม่น

โดยในการแจกแจงปรกติแบบผสมนั้น ค่าส่วนเบี่ยงเบนมาตรฐานที่ใช้ในการสร้างข้อมูลจะใช้ 4 ค่า ได้แก่ 1, 4, 7 และ 10 ค่าเฉลี่ยของตัวแปรสุ่มกลุ่ม 1 (กลุ่มที่มีค่าเฉลี่ยน้อยกว่า) หรือ μ_1 จะถูกสุ่มค่าทุกครั้งที่ทำการทดลองซ้ำด้วยการใช้การแจกแจงเอกรูประหว่าง 0 ถึง 10 ค่าเฉลี่ยของตัวแปรสุ่มกลุ่ม 2 (กลุ่มที่มีค่าเฉลี่ยมากกว่า) จะมีค่าเท่ากับค่าเฉลี่ยของตัวแปรสุ่มกลุ่ม 1 บวกกับ 0.5, 1 หรือ 2 เท่าของค่าส่วนเบี่ยงเบนมาตรฐานเพื่อจำลองความห่างของข้อมูล 2 กลุ่ม ความน่าจะเป็นที่ข้อมูลจะมาจากการแจกแจงกลุ่ม 1 (ω_1) และกลุ่ม 2 (ω_2) จะใช้ 3 ค่าได้แก่ (0.75, 0.25), (0.5, 0.5) และ (0.25, 0.75) เพื่อจำลองสัดส่วนของข้อมูลที่แตกต่างกัน จากรูปแบบค่าส่วนเบี่ยงเบนมาตรฐาน 4 ค่า รูปแบบค่าเฉลี่ยกลุ่ม 2 3 ค่า และความน่าจะเป็น 3 ค่า เมื่อคูณกันจึงได้รูปแบบการทดลองของการแจกแจงปรกติแบบผสมทั้งหมด 36 รูปแบบ โดยสามารถสรุปค่าพารามิเตอร์ที่ใช้ในการทดลองได้ดังนี้

Table 1 Summary of parameters used for normal mixture distribution data simulations.

u_1	u_2	σ	ω_1	ω_2
(0,10)	$u_1 + 0.5\sigma$	1, 4, 7, 10	0.75	0.25
(0,10)	$u_1 + 0.5\sigma$	1, 4, 7, 10	0.5	0.5
(0,10)	$u_1 + 0.5\sigma$	1, 4, 7, 10	0.25	0.75
(0,10)	$u_1 + \sigma$	1, 4, 7, 10	0.75	0.25
(0,10)	$u_1 + \sigma$	1, 4, 7, 10	0.5	0.5
(0,10)	$u_1 + \sigma$	1, 4, 7, 10	0.25	0.75
(0,10)	$u_1 + 2\sigma$	1, 4, 7, 10	0.75	0.25
(0,10)	$u_1 + 2\sigma$	1, 4, 7, 10	0.5	0.5
(0,10)	$u_1 + 2\sigma$	1, 4, 7, 10	0.25	0.75

ในการสร้างข้อมูลแจกแจงล็อกปรกติแบบผสมจะใช้การสร้างข้อมูลที่ใกล้เคียงกับการสร้างการแจกแจงปรกติแบบผสม แต่จะกำหนดค่าเฉลี่ยและค่าส่วนเบี่ยงเบนมาตรฐานของการแจงแจง และคำนวณค่าพารามิเตอร์ที่จำเป็นต้องใช้เพื่อให้ได้การแจกแจงที่มีค่าเฉลี่ยและค่าส่วนเบี่ยงเบนมาตรฐานตามที่ต้องการ โดยค่าส่วนเบี่ยงเบนมาตรฐานของกลุ่ม 1 (sd_1) จะใช้ 4 ค่า ได้แก่ 1, 4, 7 และ 10 ค่าเฉลี่ยของตัวแปรสุ่มกลุ่ม 1 (m_1) จะถูกสุ่มค่าทุกครั้งที่ทำการทดลองซ้ำด้วยการใช้การแจกแจงเอกรูประหว่าง 0 ถึง 10 ค่าเฉลี่ยของตัวแปรสุ่มกลุ่ม 2 (m_2) จะมีค่าเท่ากับค่าเฉลี่ยของตัวแปรสุ่มกลุ่ม 1 บวกกับ 0.5, 1 หรือ 2 เท่าของค่าส่วนเบี่ยงเบนมาตรฐานเพื่อจำลองความห่างของข้อมูล 2 กลุ่ม ความน่าจะเป็นที่ข้อมูลจะมาจากการแจกแจงกลุ่ม 1 และกลุ่ม 2 จะใช้ 3 ค่าได้แก่ (0.75, 0.25), (0.5, 0.5) และ (0.25, 0.75) ซึ่งการคำนวณค่าพารามิเตอร์ u_1, u_2 และ σ มีวิธีคำนวณดังนี้

$$\mu_1 = \log\left(\frac{m_1^2}{\sqrt{sd_1^2 + m_1^2}}\right)$$

$$\mu_2 = \log\left(\frac{m_2^2}{\sqrt{\left[\left(e^{\sigma^2} - 1\right)m_2^2\right] + m_2^2}}\right)$$

$$\sigma = \sqrt{\log\left(1 + \frac{sd_1^2}{m_1^2}\right)}$$

โดยค่าส่วนเบี่ยงเบนมาตรฐานของกลุ่ม 2 จะไม่กำหนดเนื่องจากต้องการให้ใช้ค่าพารามิเตอร์ σ เดียวกันเพื่อให้ข้อมูลมีลักษณะเป็นฐานนิยมเดียว โดยรูปแบบการทดลองของการแจกแจงล็อกปรกติแบบผสมจะมีทั้งหมด 36 รูปแบบเช่นเดียวกันกับการแจกแจงปรกติแบบผสม โดยสามารถสรุปค่าที่พารามิเตอร์ใช้ในการทดลองได้ดังนี้

Table 2 Summary of parameters used for log-normal mixture distribution data simulations.

m_1	m_2	sd_1	ω_1	ω_2
(0,10)	$m_1 + 0.5sd_1$	1, 4, 7, 10	0.75	0.25
(0,10)	$m_1 + 0.5sd_1$	1, 4, 7, 10	0.5	0.5
(0,10)	$m_1 + 0.5sd_1$	1, 4, 7, 10	0.25	0.75
(0,10)	$m_1 + sd_1$	1, 4, 7, 10	0.75	0.25
(0,10)	$m_1 + sd_1$	1, 4, 7, 10	0.5	0.5
(0,10)	$m_1 + sd_1$	1, 4, 7, 10	0.25	0.75
(0,10)	$m_1 + 2sd_1$	1, 4, 7, 10	0.75	0.25
(0,10)	$m_1 + 2sd_1$	1, 4, 7, 10	0.5	0.5
(0,10)	$m_1 + 2sd_1$	1, 4, 7, 10	0.25	0.75

โดยในการจำลองข้อมูลได้มีการเก็บข้อมูลว่าจุดข้อมูลนั้นมาจากการแจกแจงกลุ่ม 1 หรือกลุ่ม 2 เพื่อใช้ในการวัดความแม่น แต่วิธีการแบ่งที่ใช้ในการแบ่งข้อมูลนั้นจะไม่ทราบว่าข้อมูลนั้นมาจากการแจกแจงใด

2.2.2 ใช้วิธีการแบ่งข้อมูลได้แก่ วิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่หยุดเมื่อจุดแบ่งเปลี่ยนแปลงน้อยกว่าร้อยละ 0.05, 0.1, และ 0.2, วิธีการแบ่งช่วงธรรมชาติเจงค์โดยระบุจำนวนกลุ่มคือ 2 , วิธี head/tail breaks ที่ใช้จุดแบ่งแรกและจุดแบ่งที่ 2 ในการแบ่ง และวิธีจัดกลุ่มข้อมูลด้วย EM ในการแบ่งข้อมูลเพื่อให้ได้จุดแบ่งเพื่อแบ่งข้อมูลออกเป็น 2 กลุ่ม โดยใช้จุดแบ่งข้อมูลให้ค่าที่น้อยกว่าจุดที่ใช้แบ่งอยู่ในกลุ่ม 1 และค่าที่มากกว่าอยู่ในกลุ่ม 2

2.2.3 คำนวณค่าความแม่นโดยใช้ค่าความแม่น 3 รูปแบบเพื่อวัดประสิทธิภาพของวิธีแบ่ง

(1) ความแม่นในการแบ่งกลุ่มข้อมูลทั้งหมดคือ จำนวนข้อมูลที่แบ่งกลุ่มถูกต้องหารด้วยจำนวนข้อมูลทั้งหมด

(2) ความแม่นในการแบ่งกลุ่ม 1 คือ จำนวนข้อมูลกลุ่ม 1 ที่แบ่งกลุ่มถูกต้องหารด้วยจำนวนข้อมูลกลุ่ม 1

(3) ความแม่นในการแบ่งกลุ่ม 2 คือ จำนวนข้อมูลกลุ่ม 2 ที่แบ่งกลุ่มถูกต้องหารด้วยจำนวนข้อมูลกลุ่ม 2

3. ผลการวิจัย

จากการทดสอบประสิทธิภาพวิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำ โดยเปรียบเทียบกับการแบ่งกลุ่ม 3 วิธีได้แก่ วิธีการแบ่งช่วงธรรมชาติเจงค์,วิธี head/tail break, และ วิธีจัดกลุ่มข้อมูลด้วยอัลกอริทึม EM ด้วยการจำลองข้อมูล 72 รูปแบบ รูปแบบละ 5000 หน่วยข้อมูล โดยจำลองข้อมูลการแจกแจง 2 รูปแบบ ได้แก่ การแจงแจงปรกติแบบผสม 2 กลุ่มและการแจกแจงล็อกปรกติแบบผสม 2 กลุ่ม จากการทดลองได้ค้นพบว่าค่าส่วนเบี่ยงเบนมาตรฐานไม่ส่งผลต่อความแม่นและให้ผลการทดลองที่ใกล้เคียงกัน จึงเลือกนำเสนอในรูปแบบค่าเฉลี่ยความแม่นของการทดลองทั้ง 4 ค่าส่วนเบี่ยงเบนมาตรฐาน โดยยังคงแยกความห่างระหว่างค่าเฉลี่ยของกลุ่มและความน่าจะเป็นของที่ข้อมูลจะเป็นกลุ่ม 1 และกลุ่ม 2 ในการนำเสนอ

3.1 การแจกแจงปรกติแบบผสม

การทดสอบในการแจงแจงปรกติแบบผสมนั้นวิธี head/tail breaks ไม่ได้ให้จุดแบ่งที่สองเนื่องจากในทุกการแจกแจงนั้นเมื่อแบ่งครั้งแรกข้อมูลส่วนหัวมีจำนวนมากกว่าร้อยละ 40 ในทุกกรณี จึงไม่มีการนำเสนอในผลของการแจงแจงปรกติแบบผสม

Table 3 Average overall accuracy for normal mixture distribution.

Mean Gap	ω_1	Repeated Jenks 0.05	Repeated Jenks 0.1	Repeated Jenks 0.2	Jenks	Head-tail 1st break	EM
0.5	0.75	0.2817	0.3004	0.3225	**0.5748**	0.5745	0.5284
0.5	0.5	0.5207	0.5279	0.5359	0.5976	**0.5979**	0.5335
0.5	0.25	**0.7473**	0.7411	0.7323	0.5742	0.5731	0.5399
1	0.75	0.2931	0.3222	0.3452	**0.6477**	0.6422	0.607
1	0.5	0.5333	0.5506	0.5661	**0.6902**	0.69	0.5768
1	0.25	**0.7649**	0.7639	0.765	0.6483	0.6427	0.5935
2	0.75	0.2969	0.3247	0.3589	0.796	0.7525	**0.8596**
2	0.5	0.5472	0.5685	0.5923	**0.8412**	**0.8412**	0.831
2	0.25	0.7842	0.7981	0.8111	0.7938	0.7516	**0.8636**

ในการแบ่งข้อมูลการแจงแจงปรกติแบบผสมนั้นเป็นที่ชัดเจนว่าวิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำนั้นไม่เหมาะสมกับการนำมาใช้แบ่งข้อมูลที่มีการแจกแจงรูปแบบนี้ไม่ว่าจะใช้ค่าร้อยละในการหยุดเท่าไหร่ก็ตามเนื่องจากความแม่นในการแบ่งนั้นต่ำกว่าวิธีการแบ่งรูปแบบอื่น ยกเว้นในกรณีที่ข้อมูลส่วนใหญ่นั้นอยู่ในกลุ่ม 2 ซึ่งคือกลุ่มที่มีค่าเฉลี่ยสูงกว่า

โดยกรณีที่ข้อมูลมีความน่าจะเป็นกลุ่ม 1 เท่ากับ 0.75 และ 0.5 วิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำทุกค่าร้อยละให้ความแม่นต่ำกว่าการแบ่งกลุ่มที่ให้ความ

แม่นสูงสุดถึง 0.27 โดยเฉลี่ย แต่กรณีที่ข้อมูลมีความน่าจะเป็นกลุ่ม 1 เท่ากับ 0.25 วิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นสูงสุดยกเว้นเมื่อความห่างระหว่างค่าเฉลี่ยข้อมูลเท่ากับ 2

ในการแบ่งข้อมูลที่กลุ่ม 1 มีจำนวนมากกว่าหรือข้อมูลแต่ละกลุ่มมีจำนวนเท่ากัน วิธีการแบ่งช่วงธรรมชาติเจงค์และการใช้จุดแบ่งแรกของวิธี head/tail break ซึ่งคือการใช้ค่าเฉลี่ยของข้อมูลในการแบ่งจะมีประสิทธิภาพใกล้เคียงกับวิธีแบ่งที่ดีที่สุดหรือเป็นวิธีดีที่สุดหากข้อมูลนั้นมีความห่างของค่าเฉลี่ยกลุ่มไม่มาก

หากข้อมูลนั้นมีความห่างของค่าเฉลี่ยกลุ่มสูง วิธีแบ่งกลุ่มด้วย EM จะมีประสิทธิภาพมากที่สุดหรือใกล้เคียงไม่ว่าข้อมูลในแต่ละกลุ่มจะมีเท่ากันหรือไม่ก็ตาม

Table 4 Average group 1 accuracy for normal mixture distribution.

Mean Gap	ω_1	Repeated Jenks 0.05	Repeated Jenks 0.1	Repeated Jenks 0.2	Jenks	Head-tail 1st break	EM
0.5	0.75	0.0471	0.0772	0.1127	**0.5504**	0.5499	0.52
0.5	0.5	0.0648	0.0992	0.1316	**0.5986**	0.5982	0.5181
0.5	0.25	0.0721	0.1055	0.1589	0.642	**0.6437**	0.5267
1	0.75	0.0597	0.1021	0.1354	0.6085	0.5985	**0.6258**
1	0.5	0.0771	0.1264	0.1686	**0.6906**	0.6902	0.6177
1	0.25	0.1105	0.1563	0.2141	0.7653	**0.7735**	0.5597
2	0.75	0.0618	0.0994	0.1454	0.7603	0.692	**0.9174**
2	0.5	0.0968	0.1421	0.1922	0.8403	**0.8406**	0.8352
2	0.25	0.1433	0.2288	0.3001	0.9019	**0.9331**	0.6931

Table 5 Average group 2 accuracy for normal mixture distribution.

Mean Gap	ω_1	Repeated Jenks 0.05	Repeated Jenks 0.1	Repeated Jenks 0.2	Jenks	Head-tail 1st break	EM
0.5	0.75	**0.9844**	0.969	0.9508	0.6479	0.6483	0.5532
0.5	0.5	**0.9758**	0.9559	0.9394	0.5967	0.5976	0.5489
0.5	0.25	**0.9727**	0.9534	0.9239	0.5516	0.5496	0.5441
1	0.75	**0.9938**	0.9828	0.9755	0.7658	0.7737	0.549
1	0.5	**0.9895**	0.9748	0.9635	0.6898	0.69	0.5376
1	0.25	**0.9836**	0.9668	0.9489	0.6093	0.599	0.6054
2	0.75	**0.9996**	0.9986	0.9972	0.9027	0.9334	0.6866
2	0.5	**0.999**	0.9963	0.9937	0.8422	0.842	0.8268
2	0.25	**0.9979**	0.9879	0.9816	0.7578	0.6911	0.9203

ในทุกกรณีวิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นในการแบ่งกลุ่ม 1 ต่ำสุดทั้ง 3 รูปแบบโดยความแม่นจะลดลงเมื่อค่าร้อยละที่ใช้ในการแบ่งลดลง และวิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นในการแบ่งกลุ่ม 2 สูงสุดทั้ง 3 รูปแบบโดยความแม่นจะเพิ่มขึ้นเมื่อค่าร้อยละที่ใช้ในการแบ่งลดลง

ถึงแม้ว่าวิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำจะไม่มีประสิทธิภาพในการแบ่งกลุ่มในกรณีทั่วไป แต่หากผู้ใช้ให้ความสำคัญต่อความแม่นในการแบ่งกลุ่ม 2 มากกว่าความแม่นทั้งหมด สามารถนำวิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำมาใช้เป็นวิธีการแบ่งข้อมูลได้ โดยการใช้ค่าร้อยละในการหยุดที่ 0.2 จะเหมาะสมที่สุด เนื่องจากให้ความแม่นกลุ่ม 2 สูงโดยไม่เสียความแม่นของกลุ่ม 1 มากเมื่อเทียบกับการใช้ค่าร้อยละที่น้อยกว่า

เป็นที่สังเกตว่าวิธีจัดกลุ่มด้วยอัลกอริทึม EM นั้นเกิดการไม่ลู่เข้า 1 ครั้ง ในการทดลองเมื่อความห่างระหว่างค่าเฉลี่ยเท่ากับ 0.5 ความน่าจะเป็นกลุ่ม 1 เท่ากับ 0.75 และค่าส่วนเบี่ยงเบนมาตรฐานของกลุ่มที่ 1 เท่ากับ 1

3.2 การแจกแจงล็อกปรกติแบบผสม

Table 6 Average overall accuracy for log-normal mixture distribution.

Mean Gap	ω_1	Repeated Jenks	Repeated Jenks 0.1	Repeated Jenks 0.2	Jenks	Head-tail 1st break	Head-tail 2nd	EM
0.5	0.75	0.5532	0.5901	0.6315	0.7182	0.6647	**0.7403**	0.6625
0.5	0.5	0.5748	0.5792	0.5764	0.5416	**0.5881**	0.5379	0.5108
0.5	0.25	**0.5785**	0.5407	0.4989	0.3665	0.4766	0.321	0.3049
1	0.75	0.5827	0.6311	0.6721	0.7414	0.7105	**0.7646**	0.7038
1	0.5	0.6145	0.6207	0.6183	0.5732	**0.6389**	0.5454	0.5308
1	0.25	**0.6118**	0.5786	0.533	0.3726	0.5041	0.3294	0.3683
2	0.75	0.6589	0.6987	0.7371	0.7856	0.7821	**0.7919**	0.7559
2	0.5	0.6765	0.6919	0.6952	0.6126	**0.712**	0.5747	0.6378
2	0.25	**0.6386**	0.6188	0.5659	0.3904	0.5355	0.3236	0.4606

ในวิธีแบ่งกลุ่มข้อมูลการแจกแจงล็อกปรกติแบบผสมนั้นประสิทธิภาพของวิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำนั้นมีความหลากหลายโดยขึ้นอยู่กับอัตราส่วนของกลุ่มข้อมูลเป็นหลัก

ในกลุ่มที่มีข้อมูลกลุ่ม 1 จำนวนมากกว่า การแบ่งโดยใช้จุดแบ่งที่ 2 ของวิธี head/tail break มีประสิทธิภาพที่สุดโดยความแม่นนั้นเกิดขึ้นจากกการแบ่งข้อมูลกลุ่ม 1 ถูกต้องเป็นส่วนใหญ่โดยเสียเปรียบที่ความแม่นกลุ่ม 2 น้อย ในขณะวิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำมีประสิทธิภาพต่ำ ไม่เหมาะสมกับการนำมาแบ่งข้อมูลที่มีกลุ่ม 1 จำนวนมาก

ในกลุ่มที่มีข้อมูลแต่ละกลุ่มจำนวนใกล้เคียงกัน การแบ่งโดยใช้จุดแบ่งที่ 1 ของวิธี head/tail break มีประสิทธิภาพที่สุด แต่วิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำมีประสิทธิภาพไม่ต่างกันมากจึงสามารถนำมาใช้ได้โดยการใช้ค่าร้อยละที่ 0.1 มีความเหมาะสมที่สุด

ในกลุ่มที่มีข้อมูลกลุ่ม 2 จำนวนมากกว่า วิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.05 มีประสิทธิภาพที่สุดและเหมาะสมกับการใช้แบ่งข้อมูลลักษณะนี้

วิธีแบ่งด้วย EM นั้นไม่มีในการทดลองไหนเลยที่ให้ประสิทธิภาพสูงที่สุด ส่วนมากให้ความแม่นต่ำกว่าการแบ่งกลุ่มที่มีประสิทธิภาพสูงสุดมากและเกิดการไม่ลู่เข้าบ่อยครั้ง จึงไม่เหมาะสมกับการนำมาแบ่งกลุ่มข้อมูลที่มีการแจกแจงล็อกปรกติแบบผสม

Table 7 Average group 1 accuracy for log-normal mixture distribution.

Mean Gap	ω_1	Repeated Jenks	Repeated Jenks 0.1	Repeated Jenks 0.2	Jenks	Head-tail 1st break	Head-tail 2nd	EM
0.5	0.75	0.5326	0.6004	0.6818	0.8909	0.7373	**0.9348**	0.8194
0.5	0.5	0.5788	0.645	0.715	0.9027	0.7768	**0.9546**	0.881
0.5	0.25	0.5562	0.6466	0.7243	0.8987	0.7952	**0.9603**	0.9127
1	0.75	0.5496	0.6313	0.7134	0.8987	0.7678	**0.9565**	0.8647
1	0.5	0.6273	0.7041	0.7814	0.9273	0.8358	**0.9795**	0.8289
1	0.25	0.6491	0.7377	0.8055	0.9486	0.8731	**0.9805**	0.8344
2	0.75	0.6567	0.7235	0.7942	0.9508	0.8364	**0.9778**	0.8578
2	0.5	0.6775	0.7556	0.849	0.9667	0.9023	**0.9893**	0.8041
2	0.25	0.7673	0.8317	0.8982	0.9822	0.9467	**0.9946**	0.8366

Table 8 Average group 2 accuracy for log-normal mixture distribution.

Mean Gap	ω_1	Repeated Jenks	Repeated Jenks 0.1	Repeated Jenks 0.2	Jenks	Head-tail 1st break	Head-tail 2nd	EM
0.5	0.75	**0.6149**	0.5592	0.4805	0.1991	0.4464	0.1557	0.1931
0.5	0.5	**0.5708**	0.513	0.4378	0.1797	0.3992	0.1206	0.1372
0.5	0.25	**0.5859**	0.5052	0.4237	0.1891	0.3703	0.1085	0.1021
1	0.75	**0.6806**	0.6289	0.5473	0.2685	0.5379	0.1876	0.2234
1	0.5	**0.6008**	0.5369	0.4543	0.2181	0.4415	0.1145	0.2324
1	0.25	**0.5996**	0.5255	0.442	0.1797	0.3806	0.1108	0.2117
2	0.75	**0.6646**	0.6228	0.5646	0.2889	0.6188	0.2326	0.4501
2	0.5	**0.6753**	0.6276	0.5401	0.2562	0.5204	0.1578	0.4706
2	0.25	**0.5956**	0.5477	0.4552	0.1932	0.3984	0.0993	0.3353

เป็นที่สังเกตว่าวิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำไม่ว่าจะใช้ค่าร้อยละเท่าใดจะให้ความแม่นในกลุ่ม 2 สูง และให้ความแม่นในกลุ่ม 1 ต่ำ เหมือนในการแบ่งข้อมูลการแจกแจงปรกติแบบผสม หากผู้ใช้ต้องการให้ความสำคัญต่อการแบ่งข้อมูลกลุ่ม 2 ให้ถูกต้องวิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำนั้นเหมาะสมกับการใช้งานรูปแบบนี้ที่สุด

4. สรุปผลการวิจัย

งานวิจัยนี้เปรียบเทียบประสิทธิภาพของวิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำซึ่งคือการนำวิธีการแบ่งช่วงธรรมชาติเจงค์มาทำการแบ่งกลุ่มซ้ำหลายๆ ครั้ง และเพิ่มจำนวนกลุ่มเรื่อยๆ จนกว่าจุดแบ่งจุดแรกจะเปลี่ยนแปลงน้อยกว่าค่าร้อยละที่กำหนด โดยได้ทดสอบในการแบ่งกลุ่มข้อมูลตัวแปรเดียว

ในงานวิจัยได้มีการเปรียบเทียบกับการแบ่งกลุ่มอื่นๆ 3 วิธีได้แก่ วิธีการแบ่งช่วงธรรมชาติเจงค์, วิธี head/tail break และวิธีจัดกลุ่มข้อมูลด้วยอัลกอริทึม EM โดยใช้การจำลองข้อมูลการแจกแจงปรกติแบบผสม 2 กลุ่มและการแจกแจงล็อกปรกติแบบผสม 2 กลุ่ม ในงานวิจัยนั้นมีการจำลองข้อมูลด้วยพารามิเตอร์หลายรูปแบบเพื่อจำลองลักษณะข้อมูลที่แตกต่างกัน การวัดประสิทธิภาพในการแบ่งกลุ่มใช้ความแม่นในการแบ่งกลุ่มของข้อมูลทั้งหมดและความแม่นในการแบ่งกลุ่มข้อมูลกลุ่ม 1 และกลุ่ม 2

โดยภาพรวมนั้นวิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำไม่มีความเหมาะสมในการใช้แบ่งกลุ่มข้อมูลการแจกแจงปรกติแบบผสมเมื่อต้องการความแม่นทั้งหมดสูงสุด เนื่องจากให้ความแม่นต่ำกว่าการแบ่งรูปแบบอื่นมาก แต่หากผู้ใช้ต้องการให้ความสำคัญต่อการแบ่งข้อมูลกลุ่ม 2 วิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำนั้นสามารถนำมาใช้ได้ โดยแนะนำใช้ร้อยละการแบ่ง 0.2

การเลือกใช้วิธีแบ่งกลุ่มข้อมูลการแจกแจงล็อกปรกติแบบผสมนั้นขึ้นอยู่กับความสัดส่วนของข้อมูล หากข้อมูลที่มีค่าเฉลี่ยสูงกว่ามีจำนวนมากกว่า วิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ร้อยละการแบ่งที่ 0.05 นั้นมีความเหมาะสมที่จะใช้แบ่งข้อมูล แต่หากข้อมูลมีสัดส่วนอื่นควรใช้การแบ่งรูปแบบอื่นมากกว่า เช่นเดียวกับกรณีการแบ่งกลุ่มข้อมูลการแจกแจงปรกติแบบผสม วิธีการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำสามารถนำมาใช้เมื่อต้องการให้ความสำคัญต่อกลุ่ม 2 มากกว่า

ในงานวิจัยนี้ได้ทดสอบการแบ่งข้อมูลใน 2 รูปแบบการแจกแจง มีการปรับเปลี่ยนจำนวนข้อมูลในแต่ละกลุ่มและปรับเปลี่ยนค่าส่วนเบี่ยงเบนมาตรฐานโดยยังคงให้ค่าส่วนเบี่ยงเบนมาตรฐานเท่ากันระหว่างกลุ่ม งานวิจัยในอนาคตจึงอาจทดสอบในการแจกแจงรูปแบบอื่นหรือทดสอบในข้อมูลที่ค่าส่วนเบี่ยงเบนมาตรฐานไม่เท่ากันได้

5. References

[1] Jenks, G.F. and University of Kansas Department of Geography, Optimal Data Classification for Choropleth Maps. 1977: University of Kansas, Kansas, 24 p.

[2] Qiu, D., and Tamhane, A., 2007, A comparative study of the K-means algorithm and the normal mixture model for clustering: Univariate case. Journal of Statistical Planning and Inference. 137: 3722-3740.

[3] Wang, Z., Da Cunha, C., Ritou, M., and Furet, B., 2019, Comparison of K-means and GMM methods for contextual clustering in HSM. Procedia Manufacturing. 28: 154-159.

[4] Patel, E., and D.S. Kushwaha, 2020, Clustering cloud workloads: K-Means vs gaussian mixture model. Procedia Computer Science, 2020. 171: 158-167.

[5] Behboodian, J., 1970, On the modes of a mixture of two normal distributions. Technometrics. 12(1): 131-139.

[6] Jiang, B., 2012, Head/tail Breaks: A new classification scheme for data with A heavy-tailed distribution. Professional Geographer - PROF GEOGR. 65(3): 482-494.

[7] Chen, Y., and Gupta, M.R., 2010, EM Demystified: An Expectation-Maximization Tutorial. Electrical Engineering.