การทดสอบประสิทธิภาพการแบ่งข้อมูลตัวแปรเดียวด้วยการใช้การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำ

นายวิชญ์ยุตม์ สุขแพทย์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564

A PERFORMANCE ASSESSMENT OF REPEATED JENKS NATURAL BREAKS CLASSIFICATION ON UNIVARIATE DATA

Mr. Vichayud Sukphaet

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Statistics
Department of Statistics
FACULTY OF COMMERCE AND ACCOUNTANCY
Chulalongkorn University
Academic Year 2021

หัวข้อวิทยานิพนธ์	การทดสอบประสิทธิภาพการแบ่งข้อมูลตัวแปรเดียวด้วยการใช้การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำ
โดย	นายวิชญ์ยุตม์ สุขแพทย์
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.นัท กุลวานิช

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.วิเลิศ ภูริวัชร)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(ผู้ช่วยศาสตราจารย์ ดร.นัท กุลวานิช)

..... กรรมการ

(อาจารย์ ดร.สาวิตรี บุญพัชรนนท์)

..... กรรมการภายนอกมหาวิทยาลัย

(ผู้ช่วยศาสตราจารย์ ดร.บุญสิทธิ์ ยิ้มวาสนา)

วิชญ์ยุตม์ สุขแพทย์ : การทดสอบประสิทธิภาพการแบ่งข้อมูลตัวแปรเดียวด้วยการใช้การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำ. (A PERFORMANCE ASSESSMENT OF REPEATED JENKS NATURAL BREAKS CLASSIFICATION ON UNIVARIATE DATA) อ.ที่ปรึกษาหลัก : ผศ. ดร.นัท กุลวานิช

การแบ่งช่วงธรรมชาติเจงค์เป็นวิธีการจัดกลุ่มข้อมูลที่ได้รับความนิยม งานวิจัยนี้ได้นำการแบ่งช่วงธรรมชาติเจงค์มาปรับใช้ด้วยการเพิ่มจำนวนกลุ่มที่ใช้แบ่งเรื่อย ๆ จนกว่าจุดแบ่งแรกของการแบ่งช่วงธรรมชาติเจงค์จะเปลี่ยนแปลงไปน้อยกว่าค่าร้อยละที่กำหนดและใช้จุดแบ่งแรกนั้นในการแบ่งข้อมูลออกเป็น 2 กลุ่ม จากการทดสอบประสิทธิภาพด้วยการจำลองข้อมูลตัวแปรเดียวที่มีการแจกแจงในรูปแบบการแจกแจงปกติแบบผสมและการแจกแจงล็อกปกติแบบผสม 2 กลุ่มและเปรียบเทียบกับวิธีการแบ่งกลุ่มข้อมูลอื่น ๆ พบว่าการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำนั้นไม่มีประสิทธิภาพในการแบ่งข้อมูลแจกแจงปกติแบบผสมเมื่อต้องการให้ได้ความแม่นยำสูงสุด และเหมาะสมกับการใช้ในข้อมูลแจกแจงล็อกปกติแบบผสมเมื่อข้อมูล 2 กลุ่มมีจำนวนใกล้เคียงกันหรือกลุ่มที่ค่าเฉลี่ยสูงกว่ามีจำนวนมากกว่า นอกจากนี้การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำใช้เวลาในการแบ่งกลุ่มกว่าวิธีอื่นมาก จึงไม่เหมาะสมที่จะนำมาใช้หากข้อมูลมีจำนวนมาก

สาขาวิชา	สถิติ	ลายมือชื่อนิสิต .....
ปีการศึกษา	2564	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6380315726 : MAJOR STATISTICS
KEYWORD: CLUSTERING METHOD, JENKS NATURAL BREAKS CLASSIFICATION, REPEATED JENKS NATURAL BREAKS CLASSIFICATION

Vichayud Sukphaet : A PERFORMANCE ASSESSMENT OF REPEATED JENKS NATURAL BREAKS CLASSIFICATION ON UNIVARIATE DATA. Advisor: Asst. Prof. NAT KULVANICH, Ph.D.

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 that are used for clustering until the change of the first break is less than the specified percentage. 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 other types of data clustering methods by using 2-group normal mixture distribution and 2-group log-normal mixture distribution univariate simulated data. The research found that repeated Jenks natural breaks classification is not suitable for maximizing the overall accuracy of the normal mixture distribution. Repeated Jenks natural breaks classification can be used for log-normal mixture distribution if the proportion of each group is relatively equal or higher-mean group leaning. Compare to other methods of clustering, repeated Jenks natural breaks classification has a relatively high computational time which might not be suitable for data with a high quantity of data points.

Field of Study: Statistics Student's Signature .....

Academic Year: 2021 Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สามารถดำเนินการและสำเร็จลุล่วงได้ ด้วยความช่วยเหลือจาก ผู้ช่วยศาสตราจารย์ ดร.นัท กุลวานิช อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ให้คำแนะนำ ปรึกษา และช่วยเหลือผู้วิจัยตลอดการวิจัย ผู้วิจัยขอขอบพระคุณเป็นอย่างยิ่ง

ผู้วิจัยขอขอบพระคุณ รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์ ประธานกรรมการสอบวิทยานิพนธ์ อาจารย์ ดร.สาวิตรี บุญพัชรนนท์ และ ผู้ช่วยศาสตราจารย์ ดร.บุญสิทธิ์ ยิ้มวาสนา กรรมการสอบวิทยานิพนธ์ ที่ให้คำแนะนำ ตรวจสอบ และข้อแก้ไขต่าง ๆ ให้วิทยานิพนธ์ฉบับนี้ให้สมบูรณ์ยิ่งขึ้น

สุดท้ายนี้ผู้วิจัยขอขอบคุณครอบครัวและเพื่อน ๆ ทุกคนที่คอยสนับสนุนผู้วิจัยเสมอมา จนสามารถทำการวิจัยจนสำเร็จลุล่วงได้

วิชญ์ยุตม์ สุขแพทย์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1

บทนำ

1.1. ที่มาและความสำคัญ

วิธีการแบ่งกลุ่มข้อมูล (data clustering method) เป็นเครื่องมือที่ใช้ในการแบ่งข้อมูลออกเป็นหลาย ๆ กลุ่ม โดยจุดมุ่งหมายหลักคือเพื่อจัดข้อมูลที่มีลักษณะใกล้เคียงกันไว้ในกลุ่มเดียวกัน วัตถุประสงค์ของการแบ่งนั้นหลากหลาย ขึ้นอยู่กับสาขาของงานที่ใช้ ตัวอย่างเช่น การแบ่งข้อมูลของระยะเวลาการรักษาในโรงพยาบาลเพื่อใช้ประกอบการตัดสินใจทางการแพทย์ (Zhang et al., 2019) เป็นต้น การใช้วิธีการแบ่งกลุ่มที่แสดงลักษณะของข้อมูลได้ดีหรือสามารถแบ่งกลุ่มของข้อมูลได้อย่างถูกต้องจึงมีความสำคัญ

การแบ่งช่วงธรรมชาติเจงค์ (Jenks natural breaks classification) เป็นหนึ่งในวิธีการแบ่งกลุ่มที่ได้รับความนิยมอย่างมากเพื่อใช้แบ่งข้อมูลในการสร้างแผนที่โคโรเพลท (choropleth map) (เช่น Chen et al. (2013) และ Baah et al. (2015)) เนื่องจากวิธีการแบ่งนี้อยู่ในโปรแกรมระบบสารสนเทศภูมิศาสตร์ (geographic information system software) หลายโปรแกรม และเป็นตัวเลือกเริ่มต้นในการแบ่งกลุ่มของบางโปรแกรม โดยการแบ่งข้อมูลด้วยวิธีนี้นั้นจะได้จุดแบ่งข้อมูลตามจุดแบ่งธรรมชาติของข้อมูลซึ่งเป็นจุดที่มีช่วงความห่างของข้อมูลมาก

งานวิจัยนี้ได้รับแรงบันดาลใจจากบริษัท e-commerce ที่ใช้การแบ่งช่วงธรรมชาติเจงค์ซ้ำหลาย ๆ ครั้งเพื่อแยกกลุ่มผู้ขายออกเป็น 2 กลุ่มระหว่างกลุ่มที่ใช้เว็บไซต์เป็นครั้งคราวเพื่อขายของกับกลุ่มที่ขายของในเว็บไซต์เป็นอาชีพโดยใช้ข้อมูลจำนวนครั้งการขายสินค้า ผู้วิจัยนั้นมองเห็นว่าวิธีนี้มีความน่าสนใจที่อาจมีประสิทธิภาพสูงในการแบ่งข้อมูลออกเป็น 2 กลุ่มได้ โดยเฉพาะข้อมูลที่มีลักษณะการแจกแจงเบ้ขวา (right-skewed distribution)

โดยปัญหาที่บริษัท e-commerce พบนั้นคือปัญหาที่มีความต้องการแบ่งข้อมูลออกเป็น 2 กลุ่มและรู้ว่าข้อมูลจริง ๆ นั้นควรจะแบ่งได้ออกเป็น 2 กลุ่มอยู่แล้ว ซึ่งจะใกล้เคียงกลับการเป็นปัญหาการจำแนกข้อมูล (classification) มากกว่า แต่เนื่องจากข้อมูลไม่สามารถระบุกลุ่มได้อย่างชัดเจนจึงไม่สามารถใช้การสร้างตัวจำแนกจากข้อมูลและทำนายได้และจำเป็นที่จะต้องใช้วิธีการแบ่งข้อมูลกลุ่ม

แทน และจากการที่จำเป็นที่จะต้องแบ่งข้อมูลออกเป็น 2 กลุ่ม จากข้อมูลที่ไม่สามารถระบุกลุ่มข้อมูลได้อย่างชัดเจนแต่ทราบว่ามี 2 กลุ่ม ผู้วิจัยจึงให้ความสำคัญต่อความแม่นยำในการแบ่งกลุ่ม (accuracy) มากกว่าการจัดกลุ่มข้อมูลให้มีความเหมือนกันอยู่ในกลุ่มเดียวกัน

ถึงแม้ว่าจะเคยมีงานวิจัยในอดีตที่ทดสอบประสิทธิภาพการแบ่งกลุ่มเช่น Qiu and Tamhane (2007) ที่ใช้การจำลองข้อมูลการแจกแจงปกติแบบผสม (normal หรือ gaussian mixture distribution) ซึ่งส่วนใหญ่นั้นแสดงถึงข้อได้เปรียบของการใช้ GMM (gaussian mixture model) ที่ใช้อัลกอริทึม EM (expectation-maximization algorithm) ในการประมาณค่าพารามิเตอร์เพื่อแบ่งกลุ่มที่มีเหนือกว่าการแบ่งด้วย K-means (ซึ่งมีการแบ่งใกล้เคียงกับการแบ่งช่วงธรรมชาติเจงค์เมื่อข้อมูลมีตัวแปรเดียว) แต่ยังไม่มีงานวิจัยใดได้ทดลองการใช้การแบ่งช่วงธรรมชาติเจงค์และการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำ (repeated Jenks natural breaks classification) ในการทดสอบประสิทธิภาพ และการทดสอบประสิทธิภาพมักจำกัดอยู่ในการการแจกแจงปกติแบบผสมเท่านั้น

งานวิจัยนี้จึงต้องการทดสอบประสิทธิภาพของการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำ และเปรียบเทียบกับการแบ่งรูปแบบอื่น ๆ โดยหากการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำสามารถใช้แบ่งข้อมูลได้ความแม่นยำสูงและดีกว่าการแบ่งข้อมูลด้วยวิธีอื่น จะเพิ่มตัวเลือกวิธีที่ใช้ในการแบ่งข้อมูลที่มีประสิทธิภาพและมีความแม่นยำได้

การแจกแจงปกติแบบผสมและ การแจกแจงล็อกปกติแบบผสม (log-normal mixture distribution) เป็น 2 รูปแบบการแจกแจงผสม (mixture distribution) ที่พบได้ในข้อมูลทั่วไป ตัวอย่างเช่นข้อมูลการปริมาณการใช้ไฟฟ้า (Li et al., 2018) ในกรณีการแจกแจงปกติแบบผสม นอกจากนี้การแจกแจงปกติแบบผสมยังเป็นที่นิยมใช้ในการแบ่งภาพ (image segmentation) อีกด้วย ตัวอย่างของการแจกแจงล็อกปกติแบบผสมสามารถพบได้ในข้อมูลรายได้ (Lubrano & Ndoye, 2016) และราคาสินทรัพย์ (Brigo & Mercurio, 2002) เป็นต้น เนื่องจากการพบเห็นได้ทั่วไปและความนิยมของการแจกแจง งานวิจัยนี้จึงใช้การจำลองข้อมูล 2 รูปแบบการแจงแจงนี้เพื่อเป็นข้อมูลที่ใช้ในการทดสอบการแบ่งข้อมูล

1.2. วัตถุประสงค์งานวิจัย

1. เพื่อเปรียบเทียบประสิทธิภาพของการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำกับวิธีการแบ่งข้อมูลด้วยวิธี การแบ่งช่วงธรรมชาติเจงค์, head/tail break, และ การจัดกลุ่มข้อมูลด้วยอัลกอริทึม EM ใน

ข้อมูลจำลองที่มีการแจกแจงปกติแบบผสม 2 กลุ่ม และข้อมูลจำลองที่มีการแจกแจงล็อกปกติแบบผสม 2 กลุ่ม

1.3. ประโยชน์ที่คาดว่าจะได้รับ

เป็นแนวทางในการเลือกวิธีการแบ่งกลุ่มข้อมูลตัวแปรเดียวสำหรับผู้ที่ต้องใช้การแบ่งกลุ่มในการวิเคราะห์ข้อมูล และเป็นการเพิ่มตัวเลือกวิธีแบ่งข้อมูลหากการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำมีประสิทธิภาพที่ดี

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1. การแจกแจงผสม (Mixture Distribution)

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

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

โดยที่

$g(x)$ คือฟังก์ชันการแจกแจงความน่าจะเป็นของการแจกแจงผสม

k คือจำนวนของการแจกแจงที่ผสม

ω_j คือความน่าจะเป็นของการแจกแจงที่ j ที่จะถูกเลือก โดยมีเงื่อนไข $\omega_j \geq 0$ และ $\sum_{j=1}^{k} \omega_j = 1$

$f_j(x)$ คือฟังก์ชันการแจกแจงความน่าจะเป็นที่ j

2.2. การแจกแจงปกติแบบผสม

การแจกแจงปกติแบบผสมคือการแจกแจงผสมที่เกิดจากการผสมกันของการแจกแจงปกติ (normal distribution) 2 การแจกแจงเป็นต้นไป โดยฟังก์ชันการแจกแจงความน่าจะเป็นของการแจกแจงปกติคือ

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

โดยที่

μ_j คือพารามิเตอร์ซึ่งเป็นค่าเฉลี่ยของการแจกแจงที่ j

σ_j คือพารามิเตอร์ซึ่งเป็นค่าส่วนเบี่ยงเบนมาตรฐานของการแจกแจงที่ j

โดยการแจกแจงผสมของของ 2 การแจกแจงปกติจะมีลักษณะเป็น unimodal หรือมีจุดยอดในการแจกแจงเพียงจุดเดียวก็ต่อเมื่อ

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

หรือถ้าหาก $\sigma_1 = \sigma_2 = \sigma$ การแจกแจงจะมีลักษณะเป็น unimodality ก็ต่อเมื่อ

$$|\mu_1 - \mu_2| \leq 2\sigma\sqrt{1 + \frac{|\log(p) - \log(q)|}{2}}$$

เมื่อ p และ q คือความน่าจะเป็นของการแจกแจงที่ 1 และ 2 ตามลำดับ โดยที่ $q = 1 - p$ (Behboodian, 1970)

2.3. การแจกแจงล็อกปกติแบบผสม

การแจกแจงล็อกปกติแบบผสมคือการแจกแจงผสมที่เกิดจากการผสมกันของการแจกแจงล็อกปกติ (log-normal distribution) 2 การแจกแจงเป็นต้นไป โดยฟังก์ชันการแจกแจงความน่าจะเป็นของการแจกแจงล็อกปกติคือ

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

โดยที่

μ_j คือพารามิเตอร์ซึ่งเป็นค่าเฉลี่ยของลอการิทึมของการแจกแจงที่ j

σ_j คือพารามิเตอร์ซึ่งเป็นค่าส่วนเบี่ยงเบนมาตรฐานของลอการิทึมของการแจกแจงที่ j

ค่าเฉลี่ยของการแจกแจงล็อกปกติมีค่าเท่ากับ

$$e^{\mu_j + \frac{\sigma_j^2}{2}}$$

และความแปรปรวนของการแจกแจงล็อกปกติมีค่าเท่ากับ

$$(e^{\sigma_j^2} - 1)e^{2\mu_j + \sigma_j^2}$$

โดยการแจกแจงผสมของของ 2 การแจกแจงปกติจะมีลักษณะเป็น unimodal ก็ต่อเมื่อการแจกแจงมีค่าพารามิเตอร์ σ เท่ากัน (Kayano & Shimizu, 1994)

2.4. การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำ (Repeated Jenks Natural Breaks Classification)

2.4.1. การแบ่งช่วงธรรมชาติเจงค์ (Jenks Natural Breaks Classification)

วิธีการแบ่งช่วงธรรมชาติเจงค์เป็นวิธีการจัดกลุ่มเสนอโดย George F. Jenks (1977) โดยแบ่งกลุ่มที่ให้ผลรวมของความเบี่ยงเบนจากค่าเฉลี่ยของกลุ่มกำลังสองของแต่ละกลุ่ม (squared deviations from the class means หรือ SDCM) น้อยที่สุดโดยเขียนเป็นสมการได้คือ

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

โดยที่

k คือจำนวนกลุ่มที่ต้องการจะแบ่ง

n_j คือจำนวนข้อมูลในกลุ่ม j

x_{ij} คือข้อมูลตัวที่ i ในกลุ่ม j

$\bar{x_j}$ คือค่าเฉลี่ยของข้อมูลในกลุ่ม j

โดยขั้นตอนในการหากลุ่มที่ให้ค่าน้อยที่สุดนั้นใช้วิธีการทดลองแบ่งข้อมูลทุกรูปแบบที่เป็นไปได้ และการแบ่งด้วยวิธีนี้ต้องกำหนดจำนวนกลุ่มที่ต้องการแบ่งเอง

ตัวอย่างการจัดกลุ่ม: กำหนดให้ข้อมูลมีค่าทั้งหมด 10 ค่า ได้แก่ 1, 2, 4, 5, 7, 9, 10, 15, 17 และ 21 และต้องการแบ่งข้อมูลออกเป็น 2 กลุ่ม

การแบ่งข้อมูลรูปแบบแรกคือการแบ่ง 1 อยู่ในกลุ่มแรก และ 2, 4, 5, 7, 9, 10, 15, 17, 21 อยู่ในกลุ่มที่ 2 ค่าเฉลี่ยของกลุ่มที่ 1 จึงมีค่าเท่ากับ 1 และค่าเฉลี่ยของกลุ่มที่ 2 มีค่าเท่ากับ 10 SDCM ของการแบ่งกลุ่มรูปแบบนี้จึงมีค่าเท่ากับ

$$\begin{aligned} SDCM = {} & [(1-1)^2] \\ & + [(2-10)^2 + (4-10)^2 + (5-10)^2 + (7-10)^2 + (9-10)^2 \\ & + (10-10)^2 + (15-10)^2 + (17-10)^2 + (21-10)^2] = 330 \end{aligned}$$

การแบ่งกลุ่มที่ให้ค่า SDCM น้อยที่สุดคือการแบ่ง 1, 2, 4, 5, 7, 9, 10 อยู่ในกลุ่มแรก และ 15, 17, 21 อยู่ในกลุ่มที่ 2 ซึ่งมีค่าเท่ากับ

$$SDCM = \left[\left(1-\frac{38}{7}\right)^2+\left(2-\frac{38}{7}\right)^2+\left(4-\frac{38}{7}\right)^2+\left(5-\frac{38}{7}\right)^2+\left(7-\frac{38}{7}\right)^2 + \left(9-\frac{38}{7}\right)^2+\left(10-\frac{38}{7}\right)^2\right] + \left[\left(15-\frac{53}{3}\right)^2+\left(17-\frac{53}{3}\right)^2+\left(21-\frac{53}{3}\right)^2\right] = 88.38095$$

รูปแบบการแบ่งกลุ่มทั้งหมดและค่า SDCM ของการแบ่งกลุ่มรูปแบบนั้นสามารถสรุปได้ดังนี้

กลุ่มที่ 1	กลุ่มที่ 2	SDCM
1	2, 4, 5, 7, 9, 10, 15, 17, 21	330
1, 2	4, 5, 7, 9, 10, 15, 17, 21	258.5
1, 2, 4	5, 7, 9, 10, 15, 17, 21	206.6667
1, 2, 4, 5	7, 9, 10, 15, 17, 21	154.8333
1, 2, 4, 5, 7	9, 10, 15, 17, 21	122
1, 2, 4, 5, 7, 9	10, 15, 17, 2	108.0833
1, 2, 4, 5, 7, 9, 10	15, 17, 21	*88.38095*
1, 2, 4, 5, 7, 9, 10, 15	17, 21	157.875
1, 2, 4, 5, 7, 9, 10, 15, 17	21	245.5556

ตารางที่ 1 สรุปรูปแบบกลุ่มที่เป็นไปได้และ SDCM ของกลุ่มตัวอย่าง (* แสดงถึงค่าต่ำที่สุด)

2.4.2. การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำ

การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำคือการใช้การแบ่งช่วงธรรมชาติเจงค์ซ้ำหลายครั้งโดยการเพิ่มจำนวนกลุ่มที่ต้องการแบ่งเพื่อให้ได้ช่วงการแบ่งสุดท้าย โดยการแบ่งจะหยุดก็ต่อเมื่อค่าจุดแบ่งแรกของกลุ่มใหม่เปลี่ยนแปลงน้อยกว่าร้อยละที่กำหนดของค่าจุดแบ่งแรกของการแบ่งกลุ่มครั้งที่แล้ว ซึ่งเขียนเป็นสมการได้คือ

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

โดยที่

$break_k$ คือจุดแบ่งแรกของการแบ่งข้อมูลเป็น k กลุ่ม

$perc$ คือค่าร้อยละที่ต้องการให้หยุดเมื่อน้อยกว่า

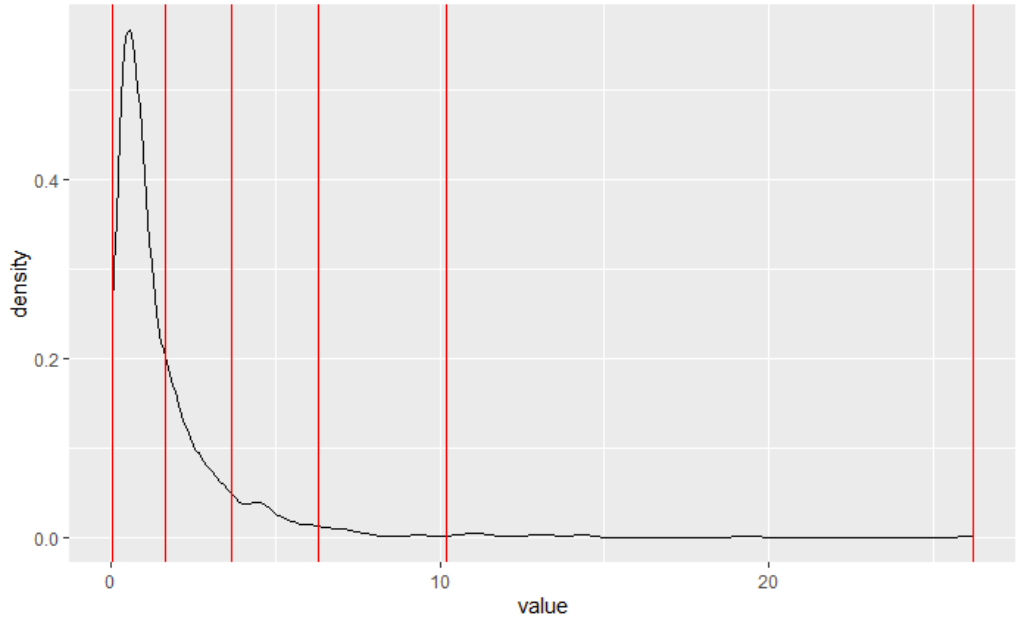
โดยหลังจากการแบ่งหยุดแล้ว ใช้จุดแบ่งแรกเท่านั้นเพื่อแบ่งข้อมูลออกเป็น 2 กลุ่ม วิธีการนี้จึงสามารถใช้ในการแบ่งข้อมูลออกเป็น 2 กลุ่มเท่านั้น

2.5. Head/Tail Breaks

head/tail breaks เป็นวิธีการแบ่งกลุ่มที่เสนอโดย Bin Jiang (2012) เพื่อแบ่งข้อมูลที่มีการแจกแจงเป็นลักษณะเบ้ขวาหรือข้อมูลที่มีการแจกแจงส่วนหางหนัก (heavy-tailed distribution) ซึ่งคือข้อมูลที่ส่วนหัว (ข้อมูลที่มีค่ามาก) มีจำนวนน้อย และข้อมูลส่วนหาง (ข้อมูลที่มีค่าน้อย) มีจำนวนมาก เช่นการแจกแจงล็อกปกติ

โดย head/tail breaks แบ่งข้อมูลด้วยการแบ่งข้อมูลที่ค่าเฉลี่ยของข้อมูลทั้งหมด ซึ่งจะแบ่งข้อมูลออกเป็น 2 กลุ่มคือข้อมูลที่มีค่าน้อยกว่าค่าเฉลี่ย (ส่วนหาง) และข้อมูลที่มีค่ามากกว่าค่าเฉลี่ย (ส่วนหัว) หากข้อมูลยังมีลักษณะหางหนักอยู่ (ข้อมูลส่วนหัวมีจำนวนน้อยกว่าร้อยละที่กำหนดโดยทั่วไปคือ 0.4) จึงนำข้อมูลส่วนหัวที่แบ่งแล้วนั้นมาแบ่งต่อที่ค่าเฉลี่ยของส่วนหัวต่อเรื่อย ๆ จนกว่าข้อมูลส่วนหัวที่แบ่งจะไม่มีลักษณะหางหนัก จากการแบ่งรูปแบบนี้จะเห็นได้ว่าวิธีแบ่งจะกำหนดจำนวนกลุ่มเอง ไม่จำเป็นต้องกำหนดจำนวนกลุ่มก่อนแบ่ง

วิธีการแบ่งนี้ถูกสร้างขึ้นเพื่อแบ่งข้อมูลที่มีลักษณะหางหนักโดยเฉพาะ เนื่องจากวิธีการแบ่งรูปแบบอื่น (รวมถึงการแบ่งช่วงธรรมชาติเจงค์) ไม่สามารถแสดงความแตกต่างของข้อมูลได้ดีพอ และข้อมูลที่มีลักษณะหางหนักนั้นสามารถพบได้ทั่วไป



รูปที่ 1 ความหนาแน่นของการแจกแจงล็อกปกติ $(\mu = 0, \sigma = 1)$ และการแบ่งด้วย head/tail breaks

2.6. การจัดกลุ่มข้อมูลด้วยอัลกอริทึม EM (Expectation-Maximization Algorithm)

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

1. สุ่มค่า $\theta^{(m=0)}$

2. E-step: ประมาณค่าความน่าจะเป็นที่ข้อมูลจะอยู่ในแต่ละกลุ่ม $p(x|y,\theta)$ โดยใช้พารามิเตอร์ครั้งที่ m $\theta^{(m)}$ 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) (Chen & Gupta, 2010)

อัลกอริทึม EM เป็น soft clustering ซึ่งคือการแบ่งกลุ่มที่ให้ค่าความน่าจะเป็นที่แต่ละจุดข้อมูลจะอยู่ในกลุ่มใดกลุ่มหนึ่ง ในงานวิจัยนี้จัดให้ข้อมูลที่มีอยู่ในกลุ่มที่มีค่าความน่าจะเป็นสูงสุด

2.7. การเปรียบเทียบระหว่างการแบ่งช่วงธรรมชาติเจงค์และ head/tail breaks

ในงานวิจัยที่เสนอการแบ่งแบบ head/tail breaks (Jiang, 2012) ได้มีการเปรียบเทียบกับการแบ่งช่วงธรรมชาติเจงค์ด้วยข้อมูลประชากรในแต่ละเมืองของประเทศสหรัฐอเมริกาปี 2000 และข้อมูลถนนในประเทศสวีเดนซึ่งข้อมูลมีลักษณะหางหนักทั้ง 2 ข้อมูล ในกรณีของข้อมูลประชากรนั้น head/tail breaks สามารถแบ่งข้อมูลให้เห็นข้อมูลได้แตกต่างชัดเจนกว่าการแบ่งช่วงธรรมชาติเจงค์ที่แบ่งข้อมูลส่วนใหญ่อยู่ในกลุ่มเดียวกันเกือบทั้งหมดทำให้ไม่เห็นความแตกแต่ง

ส่วนข้อมูลของถนนในสวีเดนนั้นได้แสดงถึงข้อได้เปรียบของ head/tail breaks โดยหากต้องการลดจำนวนกลุ่มที่ต้องการแบ่ง สามารถทำได้โดยรวมกลุ่มที่มีค่าข้อมูลมากที่สุดเข้าด้วยกันโดยไม่จำเป็นต้องคำนวณกลุ่มใหม่ซึ่งการแบ่งช่วงธรรมชาติเจงค์จำเป็นต้องคำนวณใหม่ และการลดจำนวนกลุ่มนั้นยังคงความง่ายในการแยกถนนยาวที่มีจำนวนน้อยจากถนนที่สั้นที่มีจำนวนมากดีกว่าการแบ่งช่วงธรรมชาติเจงค์ที่แสดงถนนยาวได้ไม่ดี

ในงานวิจัยที่ใช้แบบจำลองความสูงเชิงเลข (digital elevation model หรือ DEM) ของประเทศสหรัฐอเมริกาเพื่อแบ่งข้อมูล ได้ยืนยันข้อได้เปรียบของ head/tail breaks ที่แสดงลักษณะของข้อมูลได้ดีกว่าและความง่ายในการลดกลุ่มของข้อมูล แต่เมื่อข้อมูลในกลุ่มค่ามากของ head/tail breaks มีจำนวนน้อยการแบ่งช่วงธรรมชาติเจงค์สามารถแบ่งได้ดีกว่า (Lin, 2013)

งานวิจัยส่วนใหญ่นั้นสนใจในการแบ่งข้อมูลที่ไม่ได้เพื่อแสดงผลเพื่อแสดงความแตกต่างระหว่างกลุ่มมากกว่าการแบ่งข้อมูลให้เป็นกลุ่มที่ถูกต้อง ซึ่งแตกต่างจากงานวิจัยนี้

2.8. การเปรียบเทียบระหว่างวิธีแบ่งกลุ่มแบบ K-means และ EM

ในการเปรียบเทียบระหว่างวิธีแบ่งกลุ่มแบบ K-means และการแบ่งโดยใช้ EM ในข้อมูลที่มีการแจกแจงปกติแบบผสมตัวแปรเดียว 2 กลุ่ม EM สามารถแบ่งกลุ่มได้อย่างแม่นยำกว่าการการแบ่งกลุ่มแบบ K-means ยกเว้นในกรณีที่ค่าพารามิเตอร์ σ ของสองกลุ่มเท่ากันและความน่าจะเป็นของการเลือกข้อมูล 2 กลุ่มเท่ากัน นอกจากนี้ K-means ยังมีความแม่นยำกว่าเมื่อข้อมูลมีจำนวนน้อย (Qiu & Tamhane, 2007)

ในการใช้ K-means และ GMM ในการแบ่งข้อมูลการทำงานของ high speed machine (เครื่องจักรที่ใช้ในการผลิตส่วนประกอบเครื่องบินหรือใบพัด เป็นต้น) GMM สามารถจับรูปแบบการทำงานของเครื่องมือระหว่างกำลังเปลี่ยนความเร็วและใช้ความเร็วคงที่ได้ดีกว่า K-means โดยมีสาเหตุเนื่องจากข้อมูลกลุ่มหนึ่งนั้นมีจำนวนมากกว่าข้อมูลอีกกลุ่มอย่างมากทำให้ GMM มีความเหมาะสมกว่าในการแบ่งข้อมูลรูปแบบนี้ (Wang et al., 2019)

การแบ่งกลุ่มแบบ GMM โดยใช้อัลกอริทึม EM ยังสามารถแบ่งข้อมูลข้อมูลที่มีความละเอียดและซับซ้อนมากกว่าการแบ่งแบบ K-means ที่ให้การแบ่งข้อมูลที่ไม่เป็นประโยชน์มากนักจากการใช้การแบ่งกลุ่มทั้งสองรูปแบบเพื่อแบ่งกลุ่ม cloud workloads (Patel & Kushwaha, 2020)

บทที่ 3

ขอบเขตและวิธีการวิจัย

3.1. ขอบเขตงานวิจัย

ในงานวิจัยเพื่อทดสอบประสิทธิภาพการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำนั้นใช้การจำลองข้อมูลในรูปแบบการแจกแจงปกติแบบผสมและการแจกแจงล็อกปกติแบบผสมตัวแปรเดียวที่มีส่วนประกอบของ 2 การแจกแจง และจัดให้ข้อมูลที่มาจากแจกแจงที่มีค่าเฉลี่ยต่ำกว่าให้อยู่กลุ่ม 1 และข้อมูลที่มาจากแจกแจงที่มีค่าเฉลี่ยสูงกว่าให้อยู่กลุ่ม 2 โดยใช้ค่าความแม่นยำ (accuracy) ในการแบ่งกลุ่มของข้อมูลทั้งหมด และความแม่นยำในการแบ่งข้อมูลกลุ่ม 1 และกลุ่ม 2 ในการประเมิน โดยงานวิจัยนี้ใช้โปรแกรม R เวอร์ชัน 4.1.2 ในการทดสอบทั้งหมด โดยงานวิจัยมีรายละเอียดดังนี้

3.1.1. ข้อมูลที่ใช้ในงานวิจัย

1. การแจกแจงปกติแบบผสม

- ค่าเฉลี่ยของกลุ่ม 1 (กลุ่มที่มีค่าเฉลี่ยน้อยกว่า) (μ_1) มีค่าสุ่มมาจากการแจกแจงแบบยูนิฟอร์ม (uniform distribution) ในช่วง (0,10) ค่าเฉลี่ยของกลุ่ม 2 (กลุ่มที่มีค่าเฉลี่ยมากกว่า) (μ_2) มีทั้งหมด 3 ลักษณะ โดยมีค่าเท่ากับ μ_1 บวกกับ 0.5, 1, 2 เท่าของส่วนเบี่ยงเบนมาตรฐาน (σ) เพื่อให้ข้อมูลมีลักษณะ unimodal

- σ มีค่าเท่ากันทั้งสองกลุ่มโดยมีค่าเท่ากับ 1, 4, 7 และ 10

- ความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 (p) มีค่าเท่ากับ 0.75, 0.5, 0.25 เพื่อจำลองข้อมูลที่มีกลุ่ม 1 มากกว่า, เท่ากัน, และน้อยกว่าตามลำดับ

- ใช้การจำลองข้อมูล 5000 จุดข้อมูลทั้งหมด 100 ครั้งในแต่ละการทดลอง โดยใน 100 ครั้งนั้นกำหนดให้ σ เท่ากับ 1, 4, 7 และ 10 อย่างละ 25 ครั้ง

การจำลองข้อมูลของการแจกแจงปกติแบบผสมสามารถสรุปได้ดังนี้

μ_1	μ_2	σ	p	q
(0,10)	$\mu_1 + 0.5\sigma$	1, 4, 7, 10	0.75	0.25
(0,10)	$\mu_1 + 0.5\sigma$	1, 4, 7, 10	0.5	0.5
(0,10)	$\mu_1 + 0.5\sigma$	1, 4, 7, 10	0.25	0.75
(0,10)	$\mu_1 + \sigma$	1, 4, 7, 10	0.75	0.25
(0,10)	$\mu_1 + \sigma$	1, 4, 7, 10	0.5	0.5
(0,10)	$\mu_1 + \sigma$	1, 4, 7, 10	0.25	0.75
(0,10)	$\mu_1 + 2\sigma$	1, 4, 7, 10	0.75	0.25
(0,10)	$\mu_1 + 2\sigma$	1, 4, 7, 10	0.5	0.5
(0,10)	$\mu_1 + 2\sigma$	1, 4, 7, 10	0.25	0.75

ตารางที่ 2 ตารางสรุปการจำลองข้อมูลของการแจกแจงปกติแบบผสม

2. การแจกแจงล็อกปกติแบบผสม

- ค่าเฉลี่ยของกลุ่ม 1 (m_1) มีค่าสุ่มมาจากการแจกแจงแบบยูนิฟอร์มในช่วง (0,10) ค่าเฉลี่ยของกลุ่ม 2 (m_2) มีทั้งหมด 3 ลักษณะ โดยมีค่าเท่ากับ m_1 บวกกับ 0.5, 1, 2 เท่าของส่วนเบี่ยงเบนมาตรฐานของกลุ่ม 1 (sd_1)

- sd_1 มีค่าเท่ากับ 1, 4, 7 และ 10 ส่วนเบี่ยงเบนมาตรฐานของกลุ่ม 2 (sd_2) มีค่าที่ทำให้พารามิเตอร์ σ ของการแจกแจงกลุ่ม 2 เท่ากับกลุ่ม 1 เพื่อให้ข้อมูลมีลักษณะ unimodal

- ความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 (p) มีค่าเท่ากับ 0.75, 0.5, 0.25

- ใช้การจำลองข้อมูล 5000 จุดข้อมูลทั้งหมด 100 ครั้งในแต่ละการทดลอง โดยใน 100 ครั้งนั้นกำหนดให้ sd_1 เท่ากับ 1, 4, 7 และ 10 อย่างละ 25 ครั้ง

การจำลองข้อมูลของการแจกแจงล็อกปกติแบบผสมสามารถสรุปได้ดังนี้

m_1	sd_1	m_2	μ_1	μ_2	σ	p	q
(0,10)	1, 4, 7, 10	$m_1 + 0.5sd_1$	$\log(\frac{{m_1}^2}{\sqrt{{sd_1}^2 + {m_1}^2}})$	$\log(\frac{{m_2}^2}{\sqrt{[(e^{{\sigma_1}^2} - 1){m_2}^2] + {m_2}^2}})$	$\sqrt{\log\left(1 + \frac{{sd_1}^2}{{m_1}^2}\right)}$	0.75	0.25
(0,10)	1, 4, 7, 10	$m_1 + 0.5sd_1$	$\log(\frac{{m_1}^2}{\sqrt{{sd_1}^2 + {m_1}^2}})$	$\log(\frac{{m_2}^2}{\sqrt{[(e^{{\sigma_1}^2} - 1){m_2}^2] + {m_2}^2}})$	$\sqrt{\log\left(1 + \frac{{sd_1}^2}{{m_1}^2}\right)}$	0.5	0.5
(0,10)	1, 4, 7, 10	$m_1 + 0.5sd_1$	$\log(\frac{{m_1}^2}{\sqrt{{sd_1}^2 + {m_1}^2}})$	$\log(\frac{{m_2}^2}{\sqrt{[(e^{{\sigma_1}^2} - 1){m_2}^2] + {m_2}^2}})$	$\sqrt{\log\left(1 + \frac{{sd_1}^2}{{m_1}^2}\right)}$	0.25	0.75
(0,10)	1, 4, 7, 10	$m_1 + sd_1$	$\log(\frac{{m_1}^2}{\sqrt{{sd_1}^2 + {m_1}^2}})$	$\log(\frac{{m_2}^2}{\sqrt{[(e^{{\sigma_1}^2} - 1){m_2}^2] + {m_2}^2}})$	$\sqrt{\log\left(1 + \frac{{sd_1}^2}{{m_1}^2}\right)}$	0.75	0.25
(0,10)	1, 4, 7, 10	$m_1 + sd_1$	$\log(\frac{{m_1}^2}{\sqrt{{sd_1}^2 + {m_1}^2}})$	$\log(\frac{{m_2}^2}{\sqrt{[(e^{{\sigma_1}^2} - 1){m_2}^2] + {m_2}^2}})$	$\sqrt{\log\left(1 + \frac{{sd_1}^2}{{m_1}^2}\right)}$	0.5	0.5
(0,10)	1, 4, 7, 10	$m_1 + sd_1$	$\log(\frac{{m_1}^2}{\sqrt{{sd_1}^2 + {m_1}^2}})$	$\log(\frac{{m_2}^2}{\sqrt{[(e^{{\sigma_1}^2} - 1){m_2}^2] + {m_2}^2}})$	$\sqrt{\log\left(1 + \frac{{sd_1}^2}{{m_1}^2}\right)}$	0.25	0.75
(0,10)	1, 4, 7, 10	$m_1 + 2sd_1$	$\log(\frac{{m_1}^2}{\sqrt{{sd_1}^2 + {m_1}^2}})$	$\log(\frac{{m_2}^2}{\sqrt{[(e^{{\sigma_1}^2} - 1){m_2}^2] + {m_2}^2}})$	$\sqrt{\log\left(1 + \frac{{sd_1}^2}{{m_1}^2}\right)}$	0.75	0.25
(0,10)	1, 4, 7, 10	$m_1 + 2sd_1$	$\log(\frac{{m_1}^2}{\sqrt{{sd_1}^2 + {m_1}^2}})$	$\log(\frac{{m_2}^2}{\sqrt{[(e^{{\sigma_1}^2} - 1){m_2}^2] + {m_2}^2}})$	$\sqrt{\log\left(1 + \frac{{sd_1}^2}{{m_1}^2}\right)}$	0.5	0.5
(0,10)	1, 4, 7, 10	$m_1 + 2sd_1$	$\log(\frac{{m_1}^2}{\sqrt{{sd_1}^2 + {m_1}^2}})$	$\log(\frac{{m_2}^2}{\sqrt{[(e^{{\sigma_1}^2} - 1){m_2}^2] + {m_2}^2}})$	$\sqrt{\log\left(1 + \frac{{sd_1}^2}{{m_1}^2}\right)}$	0.25	0.75

ตารางที่ 3 ตารางสรุปการจำลองข้อมูลของการแจกแจงล็อกปกติแบบผสม

3.1.2. วิธีการแบ่งข้อมูล

เพื่อแบ่งข้อมูลที่ระบุไว้ในข้อ 4.1. งานวิจัยใช้วิธีการแบ่งข้อมูล 7 รูปแบบเพื่อกำหนดจุดแบ่งข้อมูลเพื่อแบ่งข้อมูลออกเป็น 2 กลุ่ม ดังนี้

1. การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่หยุดเมื่อจุดแบ่งเปลี่ยนแปลงน้อยกว่าร้อยละ 0.05

2. การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่หยุดเมื่อจุดแบ่งเปลี่ยนแปลงน้อยกว่าร้อยละ 0.1

3. การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่หยุดเมื่อจุดแบ่งเปลี่ยนแปลงน้อยกว่าร้อยละ 0.2

4. การแบ่งช่วงธรรมชาติเจงค์โดยระบุจำนวนกลุ่มคือ 2

5. head/tail breaks ที่ใช้จุดแบ่งแรกในการแบ่ง

6. head/tail breaks ที่ใช้จุดแบ่งที่สองในการแบ่ง

7. การจัดกลุ่มข้อมูลด้วย EM

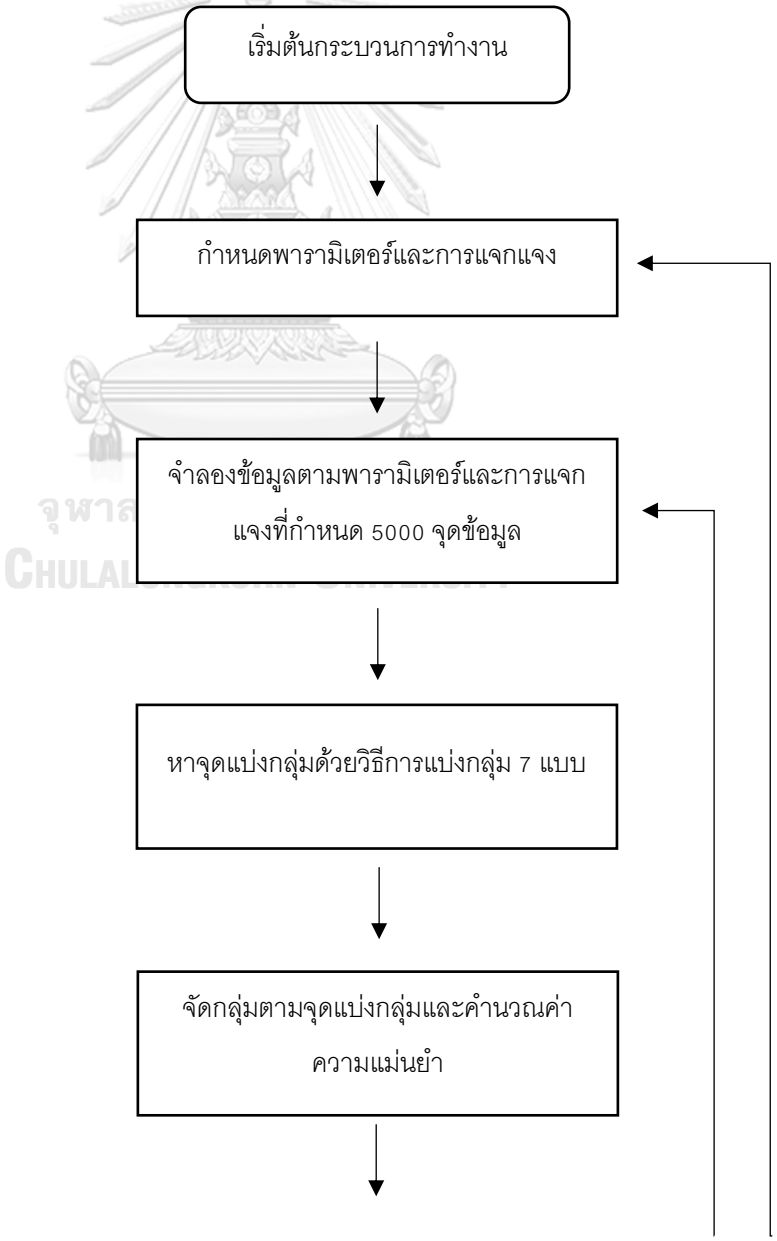
3.1.3. เกณฑ์การวัดประสิทธิภาพ

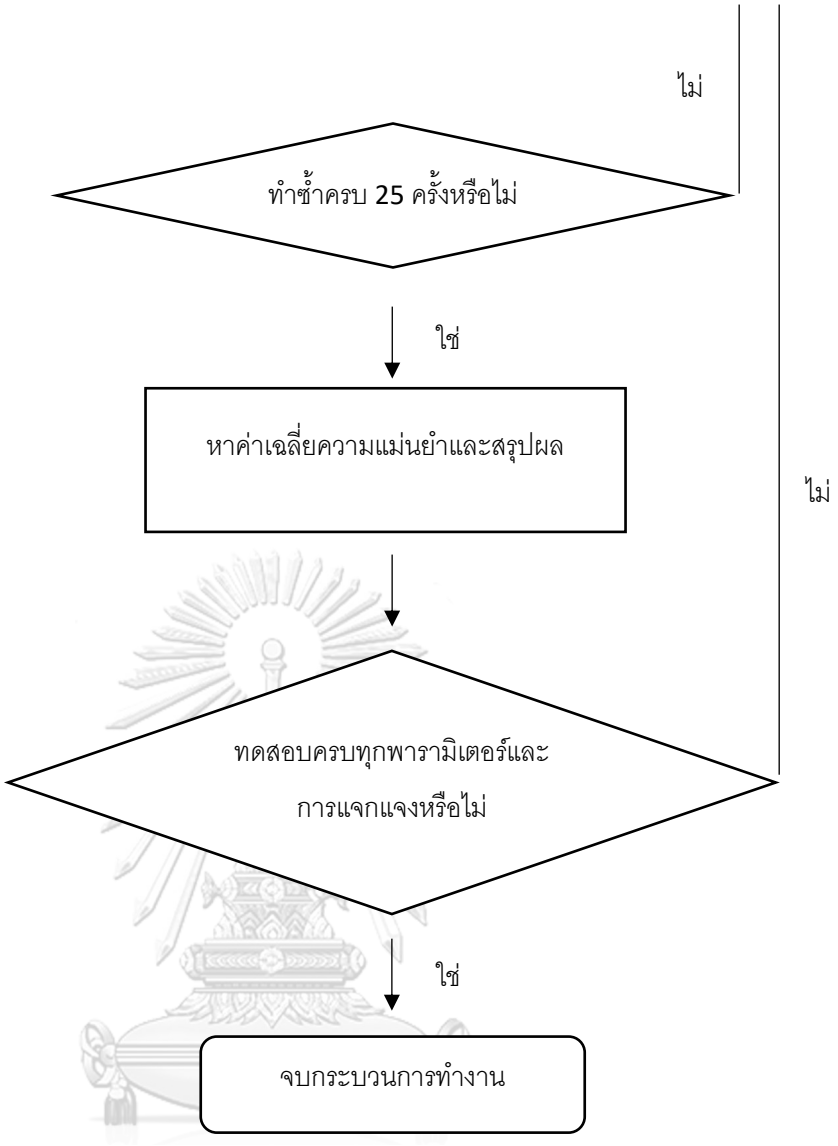
1. ความแม่นยำในการแบ่งกลุ่มข้อมูลทั้งหมดคือ จำนวนหน่วยข้อมูลที่แบ่งกลุ่มถูกต้องหารด้วยจำนวนหน่วยข้อมูลทั้งหมด

2. ความแม่นยำในการแบ่งกลุ่ม 1 คือ จำนวนหน่วยข้อมูลกลุ่ม 1 ที่แบ่งกลุ่มถูกต้องหารด้วยจำนวนหน่วยข้อมูลกลุ่ม 1

3. ความแม่นยำในการแบ่งกลุ่ม 2 คือ จำนวนหน่วยข้อมูลกลุ่ม 2 ที่แบ่งกลุ่มถูกต้องหารด้วยจำนวนหน่วยข้อมูลกลุ่ม 2

3.2. วิธีการดำเนินงานวิจัย





รูปที่ 2 ขั้นตอนการดำเนินงานวิจัย

ในงานวิจัยนั้นมีขั้นตอนการดำเนินงานวิจัยดังนี้

1. จำลองข้อมูลที่มีการแจกแจงปกติแบบผสม หรือการแจกแจงล็อกปกติแบบผสมตามพารามิเตอร์ที่กำหนดไว้จำนวน 5000 ข้อมูล

2. ใช้วิธีการแบ่งข้อมูลที่กำหนดไว้แบ่งข้อมูลเพื่อให้ได้จุดแบ่งเพื่อแบ่งข้อมูลออกเป็น 2 กลุ่ม

3. แบ่งข้อมูลด้วยจุดแบ่งที่ได้ โดยค่าที่น้อยกว่าจุดที่ใช้แบ่งอยู่ในกลุ่มที่ 1 และค่าที่มากกว่าอยู่ในกลุ่มที่ 2

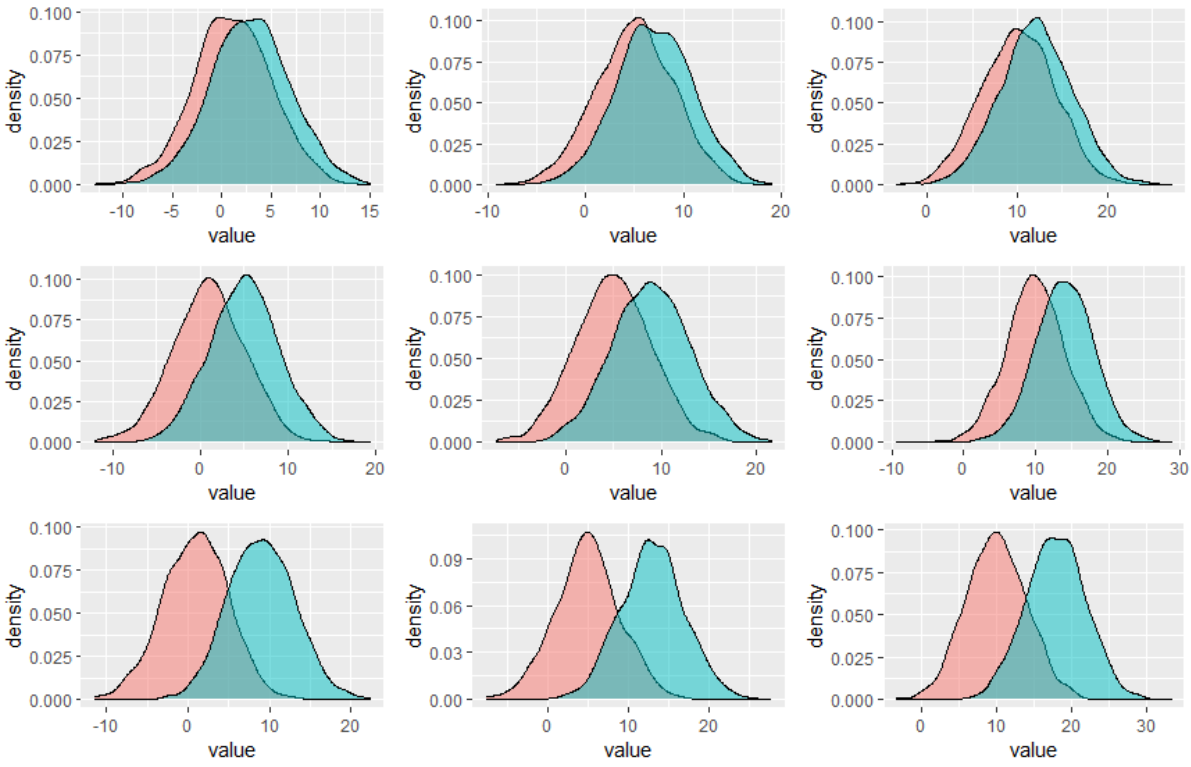
4. คำนวณค่าความแม่นยำทั้ง 3 แบบ

5. ทำซ้ำทั้งหมด 25 ครั้งและนำค่าความแม่นยำในแต่ละครั้งมาหาค่าเฉลี่ยและสรุปผล

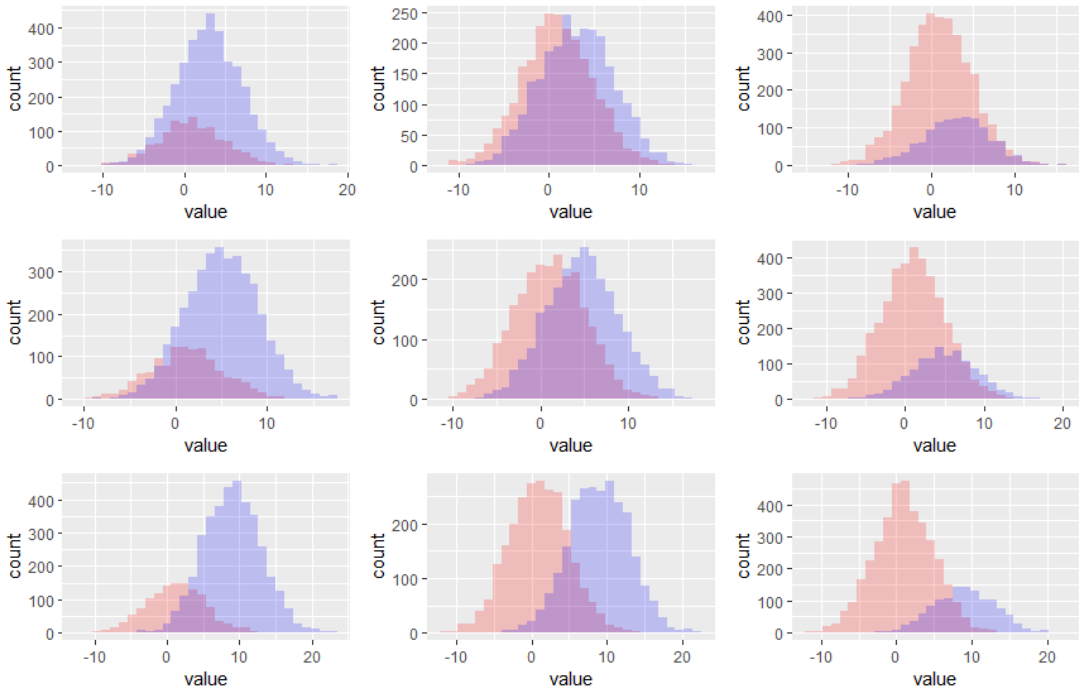
6. เริ่มทำจากข้อ 1 ใหม่โดยเปลี่ยนการแจกแจงและพารามิเตอร์ที่ใช้

3.3. ตัวอย่างการแจกแจงที่ใช้ในงานวิจัย

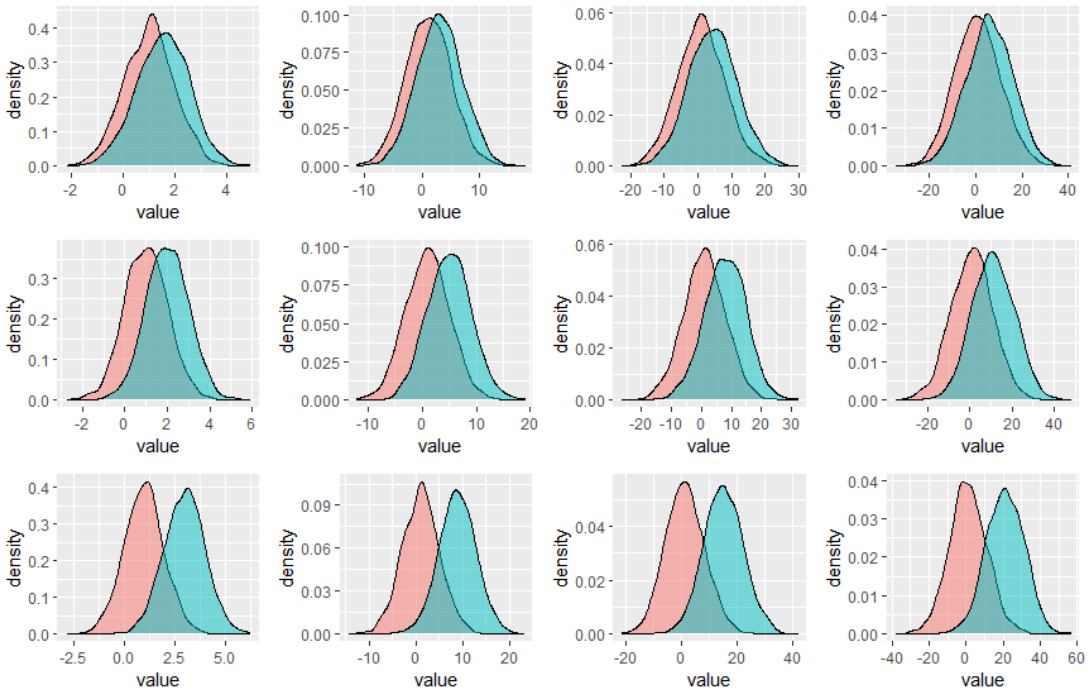
3.3.1. การแจกแจงปกติแบบผสม



รูปที่ 3 ตัวอย่างการแจกแจงปกติแบบผสมที่มีความห่าง (แนวตั้ง) และค่าเฉลี่ย (แนวนอน) แตกต่างกัน



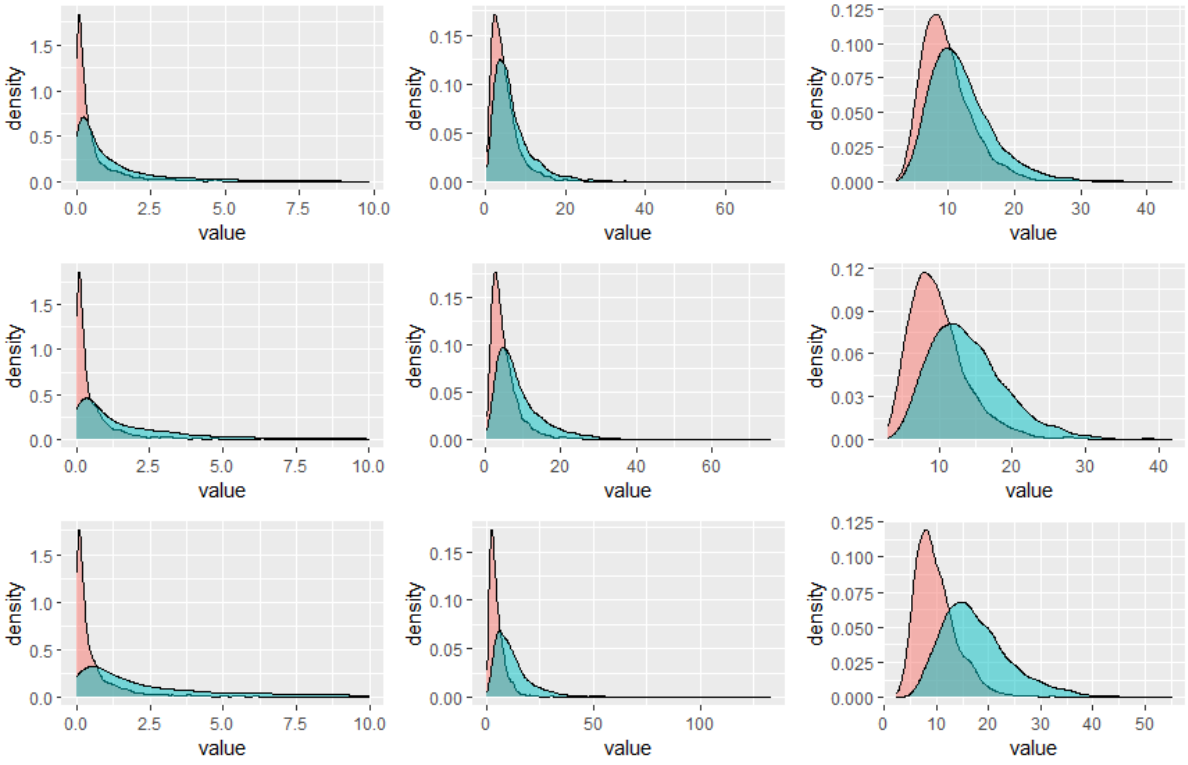
รูปที่ 4 ตัวอย่างการแจกแจงปกติแบบผสมที่มีความห่าง (แนวตั้ง) และค่าความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 (แนวนอน) แตกต่างกัน



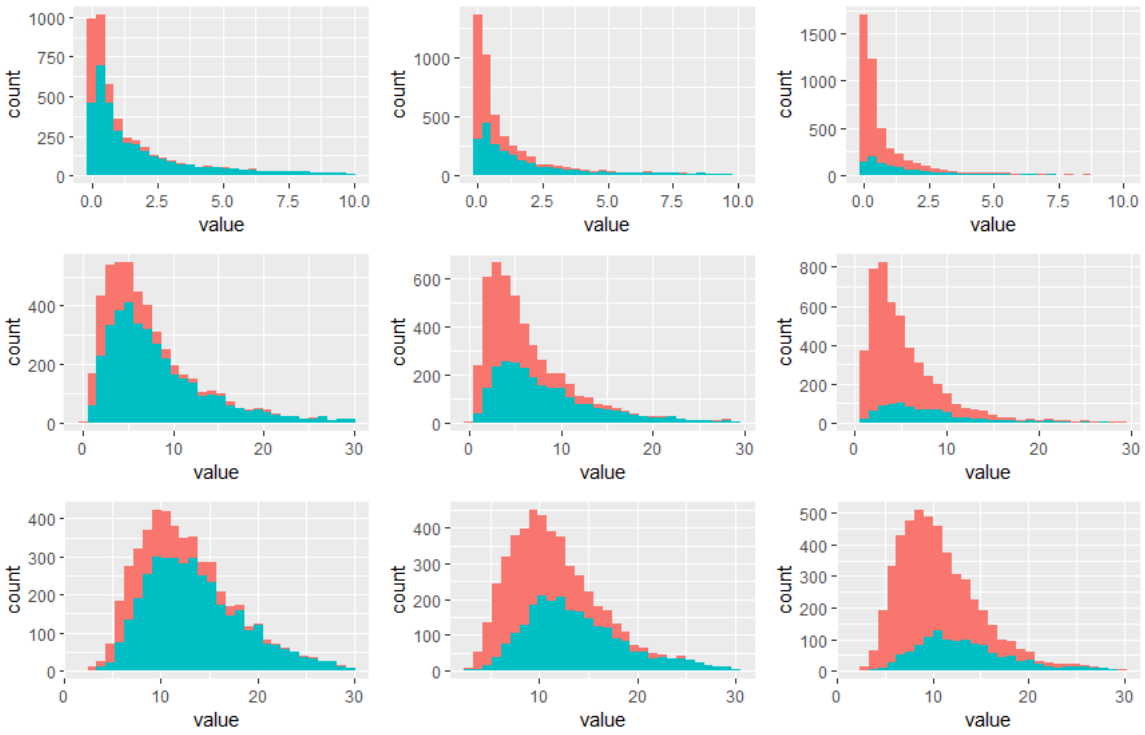
รูปที่ 5 ตัวอย่างการแจกแจงปกติแบบผสมที่มีความห่าง (แนวตั้ง) และค่าส่วนเบี่ยงเบนมาตรฐาน (แนวนอน) แตกต่างกัน

- เมื่อความห่างของค่าเฉลี่ยเพิ่มสูงขึ้นการแจกแจงของสองกลุ่มจะห่างกันมากขึ้นทำให้ส่วนที่ซ้อนทับกันลดลง
- ค่าความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 ส่งผลต่ออัตราส่วนของข้อมูลโดยเมื่อค่าความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 สูงจะส่งผลให้ข้อมูลกลุ่ม 1 มีมากกว่า และเช่นเดียวกันในทางตรงข้าม
- เมื่อค่าส่วนเบี่ยงเบนมาตรฐานเพิ่มขึ้นจะส่งผลให้ข้อมูลมีการกระจายมากขึ้น โดยสังเกตได้จากค่า density ที่ลดลงเมื่อค่าส่วนเบี่ยงเบนมาตรฐานเพิ่มขึ้น

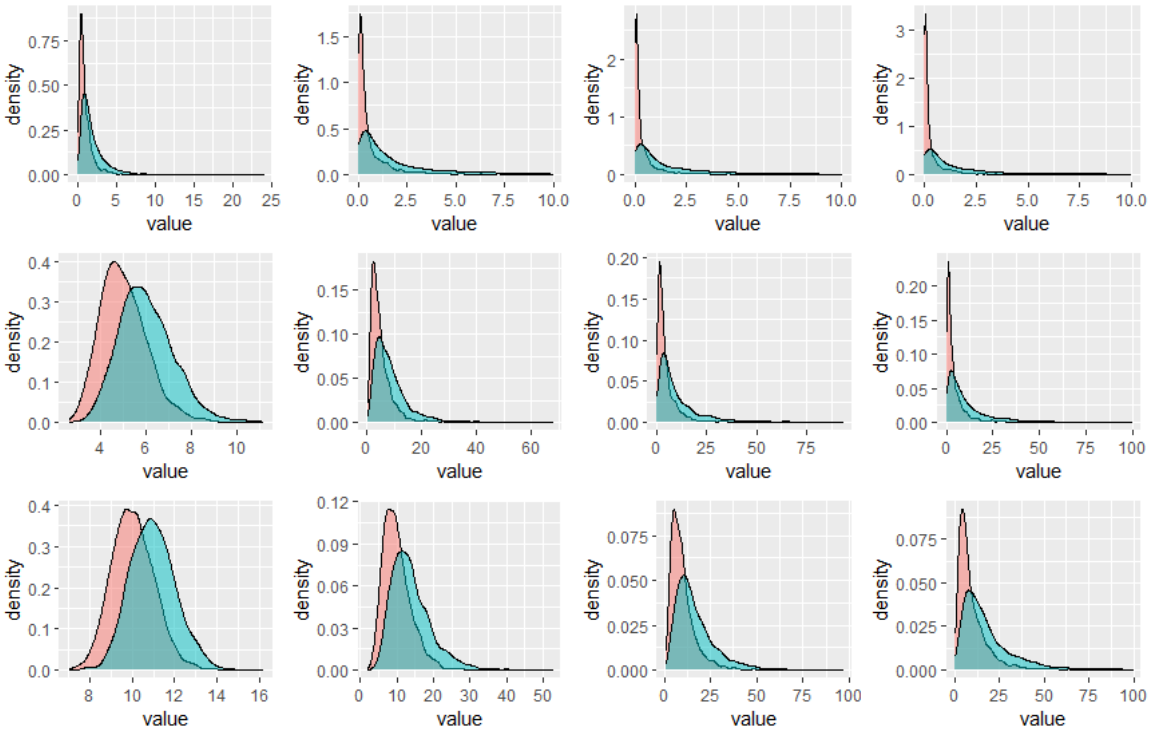
3.3.2. การแจกแจงล็อกปกติแบบผสม



รูปที่ 6 ตัวอย่างการแจกแจงล็อกปกติแบบผสมที่มีความห่าง (แนวตั้ง) และค่าเฉลี่ย (แนวนอน) แตกต่างกัน



รูปที่ 7 ตัวอย่างการแจกแจงล็อกปกติแบบผสมที่มีค่าความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 (แนวตั้ง) และค่าเฉลี่ย (แนวนอน) แตกต่างกัน



รูปที่ 8 ตัวอย่างการแจกแจงล็อกปกติแบบผสมที่มีค่าเฉลี่ย(แนวตั้ง) และส่วนเบี่ยงเบนมาตรฐาน (แนวนอน) แตกต่างกัน

- เมื่อความห่างของค่าเฉลี่ยเพิ่มสูงขึ้นการแจกแจงของกลุ่ม 2 จะมีการกระจายสูงขึ้น ในขณะที่การแจกแจงของกลุ่ม 1 มีลักษณะเดิม
- ค่าความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 ส่งผลต่ออัตราส่วนของข้อมูลโดยเมื่อค่าความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 สูงจะส่งผลให้ข้อมูลกลุ่ม 1 มีมากกว่า และเช่นเดียวกันในทางตรงข้าม
- เมื่อค่าส่วนเบี่ยงเบนมาตรฐานเพิ่มขึ้นจะส่งผลให้การแจกแจงกลุ่ม 1 มีความเบ้ขวาเพิ่มขึ้นและส่งผลให้การแจกแจงกลุ่ม 2 มีการกระจายสูงขึ้น
- เมื่อค่าเฉลี่ยเพิ่มขึ้นจะส่งผลให้การแจกแจงกลุ่ม 1 มีความเบ้ขวาลดลงและส่งผลให้การแจกแจงกลุ่ม 2 มีการกระจายลดลง ซึ่งจะตรงข้ามกับผลของส่วนเบี่ยงเบนมาตรฐาน

บทที่ 4

ผลงานวิจัย

จากการทดสอบประสิทธิภาพการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำ โดยเปรียบเทียบกับการแบ่งกลุ่มวิธีอื่น 3 วิธีได้แก่ การแบ่งช่วงธรรมชาติเจงค์, head/tail break, และ การจัดกลุ่มข้อมูลด้วยอัลกอริทึม EM ด้วยการจำลองข้อมูล 72 รูปแบบ รูปแบบละ 5000 หน่วยข้อมูล โดยจำลองข้อมูล 2 รูปแบบ ได้แก่การแจงแจงปกติแบบผสม 2 กลุ่มและการแจกแจงล็อกปกติแบบผสม 2 กลุ่ม โดยที่

- ความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่า 0.75, 0.5, และ 0.25 ตามลำดับ
- ความห่างระหว่างค่าเฉลี่ยสองกลุ่ม มีค่า 0.5, 1, และ 2 เท่าของส่วนเบี่ยงเบนมาตรฐานกลุ่ม 1 ตามลำดับ
- ส่วนเบี่ยงเบนมาตรฐานกลุ่ม 1 ได้แก่ 1, 4, 7, และ 10 ตามลำดับ

รวมเป็น 2 x 3 x 3 x 4 = 72 รูปแบบ โดยใช้ความแม่นยำในการแบ่งกลุ่มทั้งหมด, ความแม่นยำในการแบ่งกลุ่ม 1, และความแม่นยำในการแบ่งกลุ่ม 2 ในการวัดประสิทธิภาพ ได้ผลการทดสอบดังนี้

ในการนำเสนอผลการทดสอบได้แบ่งการนำเสนอตามรูปแบบการแจกแจง ตามด้วยความห่างระหว่างค่าเฉลี่ยสองกลุ่ม และใช้สัญลักษณ์ดังนี้

$RJ_{0.05}$ แทน การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.05

$RJ_{0.1}$ แทน การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.1

$RJ_{0.2}$ แทน การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.2

J แทน การแบ่งช่วงธรรมชาติเจงค์

HT_1 แทน head/tail breaks ที่ใช้จุดแบ่งแรกในการแบ่ง

HT_2 แทน head/tail breaks ที่ใช้จุดแบ่งที่สองในการแบ่ง

EM แทน การจัดกลุ่มข้อมูลด้วย EM

$no\ HT_2$ แทน จำนวนครั้งที่ head/tail breaks ไม่ได้ให้จุดแบ่งที่สอง

$no\ EM$ แทน จำนวนครั้งที่การจัดกลุ่มข้อมูลด้วย EM เกิดปัญหาไม่ลู่เข้า

-- แทน การจัดกลุ่มที่ให้ค่าความแม่นยำสูงสุด

โดยในตารางที่นำเสนอในส่วนนี้นำเสนอเฉพาะความแม่นยำในการแบ่งกลุ่มทั้งหมด การวัดประสิทธิภาพอื่นได้นำเสนอในภาคผนวกแต่จะมีการกล่าวถึง

4.1. การแจงแจงปกติแบบผสม

การทดสอบในการแจงแจงปกติแบบผสมนั้น head/tail breaks ไม่ได้ให้จุดแบ่งที่สอง เนื่องจากในทุกการแจกแจงนั้นเมื่อแบ่งครั้งแรกข้อมูลส่วนหัวมีจำนวนมากกว่าร้อยละ 40 ในทุกกรณี จึงไม่มีการนำเสนอในผลของการแจงแจงปกติแบบผสม

4.1.1. กรณีความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 0.5

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	EM
0.75	1	0.304	0.3564	0.4038	0.5746	*0.5748*	0.5192
0.75	4	0.2708	0.2783	0.293	*0.5758*	0.5746	0.5664
0.75	7	0.275	0.2808	0.2929	0.5752	*0.5753*	0.4785
0.75	10	0.2772	0.2862	0.3003	*0.5736*	0.5735	0.5494
0.5	1	0.5377	0.5554	0.5661	0.5986	*0.5988*	0.539
0.5	4	0.513	0.5171	0.5292	0.5972	*0.5977*	0.5318
0.5	7	0.5161	0.5184	0.524	0.5964	*0.5967*	0.5272
0.5	10	0.5161	0.5208	0.5242	*0.5983*	0.5982	0.536
0.25	1	*0.7412*	0.7182	0.6942	0.576	0.5744	0.5961
0.25	4	*0.7502*	0.7497	0.7422	0.5732	0.573	0.5363
0.25	7	*0.7489*	0.7483	0.7469	0.5722	0.5712	0.4957
0.25	10	*0.7489*	0.7482	0.7462	0.5753	0.5739	0.5315

ตารางที่ 4 ตารางสรุปความแม่นยำในการแบ่งกลุ่มทั้งหมด กรณีการแจกแจงปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 0.5

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.75

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำต่ำสุดทั้ง 3 รูปแบบ โดยความแม่นยำจะต่ำลงเมื่อค่าร้อยละที่ใช้ในการแบ่งลดลง
- การแบ่งช่วงธรรมชาติเจงค์และ head/tail breaks ที่ใช้จุดแบ่งแรกจะได้ผลลัพธ์ที่ใกล้เคียงกันและเป็นการแบ่งที่ให้ความแม่นยำสูงสุด
- การจัดกลุ่มข้อมูลด้วย EM มีความแม่นยำต่ำกว่าการแบ่งที่ให้ความแม่นยำสูงสุด โดยมีความแม่นยำต่ำกว่า 0.01 ถึง 0.1 ขึ้นอยู่กับค่าส่วนเบี่ยงเบนมาตรฐาน

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.5

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำต่ำสุดทั้ง 3 รูปแบบ โดยความแม่นยำจะต่ำลงเมื่อค่าร้อยละที่ใช้ในการแบ่งลดลง ยกเว้นในกรณี sd_1 เท่ากับ 1 ที่การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.1 และ 0.2 มีความแม่นยำกว่าการจัดกลุ่มข้อมูลด้วย EM
- การแบ่งช่วงธรรมชาติเจงค์และ head/tail breaks ที่ใช้จุดแบ่งแรกจะได้ผลลัพธ์ที่ใกล้เคียงกันและเป็นการแบ่งที่ให้ความแม่นยำสูงสุด
- การจัดกลุ่มข้อมูลด้วย EM มีความแม่นยำต่ำกว่าการแบ่งที่ให้ความแม่นยำสูงสุด โดยมีความแม่นยำต่ำกว่าการแบ่งที่ดีที่สุด 0.06 โดยเฉลี่ย
- การทดลองที่ส่วนเบี่ยงเบนมาตรฐานกลุ่ม 1 เท่ากับ 7 จากการทดลองทั้งหมด 25 ครั้ง การจัดกลุ่มข้อมูลด้วย EM เกิดปัญหาไม่ลู่เข้า 1 ครั้ง

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.25

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำสูงสุดทั้ง 3 รูปแบบ โดยความแม่นยำจะสูงขึ้นเมื่อค่าร้อยละที่ใช้ในการแบ่งลดลง
- การแบ่งช่วงธรรมชาติเจงค์และ head/tail breaks ที่ใช้จุดแบ่งแรกจะได้ผลลัพธ์ที่ใกล้เคียงกันและเป็นการแบ่งที่ให้ความแม่นยำต่ำสุดเมื่อส่วนเบี่ยงเบนมาตรฐานกลุ่ม 1 เท่ากับ 1 และต่ำสุดโดยสูงกว่าเพียงแค่การจัดกลุ่มข้อมูลด้วย EM ในกรณี sd_1 เป็นค่าอื่น
- การจัดกลุ่มข้อมูลด้วย EM มีความแม่นยำต่ำที่สุด ยกเว้นเมื่อ sd_1 เท่ากับ 1

4.1.2. กรณีความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 1

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	EM
0.75	1	0.3352	0.4313	0.4732	*0.6471*	0.6423	0.6209
0.75	4	0.2779	0.2845	0.3067	*0.6479*	0.6424	0.5685
0.75	7	0.2781	0.2849	0.2925	*0.6505*	0.6441	0.644
0.75	10	0.2811	0.2879	0.3086	*0.6454*	0.6402	0.5947
0.5	1	0.5708	0.6218	0.6371	0.6895	*0.6895*	0.5729
0.5	4	0.5199	0.5264	0.5634	*0.6883*	0.688	0.5732
0.5	7	0.5178	0.5228	0.5281	*0.6915*	0.6914	0.5926
0.5	10	0.5248	0.5315	0.5357	*0.6914*	0.6911	0.5683
0.25	1	*0.7719*	0.7624	0.7578	0.6482	0.6433	0.5661
0.25	4	0.7637	0.7654	*0.7688*	0.6469	0.6399	0.6432
0.25	7	0.7611	0.7622	*0.7656*	0.65	0.6448	0.5563
0.25	10	0.763	0.7655	*0.7677*	0.648	0.6427	0.6082

ตารางที่ 5 ตารางสรุปความแม่นยำในการแบ่งกลุ่มทั้งหมด กรณีการแจกแจงปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 1

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.75

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำต่ำสุดทั้ง 3 รูปแบบ โดยความแม่นยำจะต่ำลงเมื่อค่าร้อยละที่ใช้ในการแบ่งลดลง
- การแบ่งช่วงธรรมชาติเจงค์และ head/tail breaks ที่ใช้จุดแบ่งแรกจะได้ผลลัพธ์ที่ใกล้เคียงกันและเป็นการแบ่งที่ให้ความแม่นยำสูงสุด
- การจัดกลุ่มข้อมูลด้วย EM มีความแม่นยำต่ำกว่าการแบ่งที่ให้ความแม่นยำสูงสุด โดยมีความแม่นยำต่ำกว่าการแบ่งที่ดีที่สุด 0.04 โดยเฉลี่ย

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.5

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำต่ำสุดทั้ง 3 รูปแบบ โดยความแม่นยำจะต่ำลงเมื่อค่าร้อยละที่ใช้ในการแบ่งลดลง ยกเว้นในกรณี sd_1 เท่ากับ 1 ที่การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.1 และ 0.2 มีความแม่นยำกว่าการจัดกลุ่มข้อมูลด้วย EM

- การแบ่งช่วงธรรมชาติเจงค์และ head/tail breaks ที่ใช้จุดแบ่งแรกจะได้ผลลัพธ์ที่ใกล้เคียงกันและเป็นการแบ่งที่ให้ความแม่นยำสูงสุด
- การจัดกลุ่มข้อมูลด้วย EM มีความแม่นยำต่ำกว่าการแบ่งที่ให้ความแม่นยำสูงสุด โดยมีความแม่นยำต่ำกว่า 0.11 โดยเฉลี่ย

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.25

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำสูงสุดทั้ง 3 รูปแบบ โดยความแม่นยำจะสูงขึ้นเมื่อค่าร้อยละที่ใช้ในการแบ่งเพิ่มขึ้น ยกเว้นในกรณี sd_1 เท่ากับ 1 ที่ความแม่นยำจะสูงขึ้นเมื่อค่าร้อยละที่ใช้ในการแบ่งลดลง
- การแบ่งช่วงธรรมชาติเจงค์และ head/tail breaks ที่ใช้จุดแบ่งแรกจะได้ผลลัพธ์ที่ใกล้เคียงกันและเป็นการแบ่งที่ให้ความแม่นยำต่ำสุดโดยสูงกว่าเพียงแค่การจัดกลุ่มข้อมูลด้วย EM ยกเว้นในกรณี sd_1 เท่ากับ 4 ที่ head/tail breaks ที่ใช้จุดแบ่งแรกมีความแม่นยำต่ำสุดและการแบ่งช่วงธรรมชาติเจงค์ต่ำสุดเป็นอันดับ 3
- การจัดกลุ่มข้อมูลด้วย EM มีความแม่นยำต่ำที่สุด ยกเว้นในกรณี sd_1 เท่ากับ 4

4.1.3. กรณีความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 2

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	EM
0.75	1	0.3363	0.4136	0.4807	0.7968	0.7524	*0.8543*
0.75	4	0.2808	0.2947	0.3331	0.7952	0.7511	*0.8653*
0.75	7	0.2802	0.2896	0.2988	0.7969	0.7526	*0.865*
0.75	10	0.2902	0.3009	0.3229	0.7949	0.7538	*0.8537*
0.5	1	0.5944	0.6643	0.7129	0.8424	*0.8425*	0.8311
0.5	4	0.5308	0.5386	0.5681	*0.8414*	0.8412	0.8315
0.5	7	0.5346	0.5362	0.546	0.8416	*0.8418*	0.8338
0.5	10	0.5288	0.5351	0.5423	0.8394	*0.8394*	0.8275
0.25	1	0.8161	0.8562	*0.8664*	0.7926	0.7517	0.8586
0.25	4	0.778	0.7835	0.8077	0.7927	0.7501	*0.8663*
0.25	7	0.7707	0.7765	0.79	0.7947	0.7521	*0.867*
0.25	10	0.7721	0.7762	0.7804	0.7953	0.7524	*0.8627*

ตารางที่ 6 ตารางสรุปความแม่นยำในการแบ่งกลุ่มทั้งหมด กรณีการแจกแจงปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 2

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.75

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำต่ำสุดทั้ง 3 รูปแบบ โดยความแม่นยำจะต่ำลงเมื่อค่าร้อยละที่ใช้ในการแบ่งลดลง
- การแบ่งช่วงธรรมชาติเจงค์และ head/tail breaks ที่ใช้จุดแบ่งแรกจะได้ผลลัพธ์ที่ใกล้เคียงกันและเป็นการแบ่งที่ให้ความแม่นยำสูงสุดรองจากการจัดกลุ่มข้อมูลด้วย EM
- การจัดกลุ่มข้อมูลด้วย EM มีความแม่นยำสูงสุด

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.5

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำต่ำสุดทั้ง 3 รูปแบบ โดยความแม่นยำจะต่ำลงเมื่อค่าร้อยละที่ใช้ในการแบ่งลดลง
- การแบ่งช่วงธรรมชาติเจงค์และ head/tail breaks ที่ใช้จุดแบ่งแรกจะได้ผลลัพธ์ที่ใกล้เคียงกันและเป็นการแบ่งที่ให้ความแม่นยำสูงสุด
- การจัดกลุ่มข้อมูลด้วย EM มีความแม่นยำใกล้เคียงกับการแบ่งที่ให้ความแม่นยำสูงสุด โดยมีความแม่นยำต่ำกว่า 0.01 โดยเฉลี่ย

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.25

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.2 ให้ความแม่นยำสูงสุดในกรณี sd_1 เท่ากับ 1 เป็นอันดับ 2 ในกรณี sd_1 เท่ากับ 4 และเป็นอันดับ 3 ในกรณี sd_1 เท่ากับ 7 และ 10
- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.1 ให้ความแม่นยำสูงเป็นอันดับ 3 ในกรณี sd_1 เท่ากับ 1 เป็นอันดับ 4 ในกรณี sd_1 เท่ากับ 4, 7, และ 10
- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.05 ให้ความแม่นยำสูงเป็นอันดับ 4 ในกรณี sd_1 เท่ากับ 1 และต่ำเป็นอันดับ 2 ในกรณี sd_1 เท่ากับ 4, 7, และ 10
- การแบ่งช่วงธรรมชาติเจงค์ให้ความแม่นยำต่ำเป็นอันดับ 2 ในกรณี sd_1 เท่ากับ 1 และสูงเป็นอันดับ 3 ในกรณี sd_1 เท่ากับ 4 และเป็นอันดับ 2 ในกรณี sd_1 เท่ากับ 7 และ 10
- head/tail breaks ที่ใช้จุดแบ่งแรกมีความแม่นยำต่ำสุด
- การจัดกลุ่มข้อมูลด้วย EM มีความแม่นยำสูงที่สุด ยกเว้นในกรณี sd_1 เท่ากับ 1 ที่การจัดกลุ่มข้อมูลด้วย EM มีความแม่นยำสูงที่สุดเป็นอันดับ 2

4.1.4. ผลความแม่นยำในการแบ่งกลุ่ม 1 และความแม่นยำในการแบ่งกลุ่ม 2 โดยภาพรวม

- ในทุกกรณี การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำในการแบ่งกลุ่ม 1 ต่ำสุดทั้ง 3 รูปแบบโดยความแม่นยำจะลดลงเมื่อค่าร้อยละที่ใช้ในการแบ่งลดลง
- ในทุกกรณีเช่นเดียวกัน การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำในการแบ่งกลุ่ม 2 สูงสุดทั้ง 3 รูปแบบโดยความแม่นยำจะเพิ่มขึ้นเมื่อค่าร้อยละที่ใช้ในการแบ่งลดลง
- การจัดกลุ่มข้อมูลด้วย EM มักให้ค่าความแม่นยำกลุ่ม 2 น้อยกว่าการแบ่งกลุ่มรูปแบบอื่น

4.1.5. ระยะเวลาที่ใช้ในการแบ่งกลุ่ม

Gap between Mean	$RJ_{0.05}$		$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	EM
0.5	204.68		163.12	121.45	13.82	0	23.45
1	252.87		205.97	154.81	16.82	0	23.98
2	262.83		214.74	167.5	16.53	0	2.21

ตารางที่ 7 ตารางสรุปค่าเฉลี่ยของเวลาที่ใช้ในการแบ่งกลุ่มข้อมูลการแจกแจงปกติแบบผสม

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำใช้เวลานานกว่าการแบ่งข้อมูลรูปแบบอื่นมาก โดยการลดค่าร้อยละที่ใช้ในการแบ่งลง 0.05 จะเพิ่มระยะเวลาที่ใช้ประมาณ 50 วินาทีโดยเฉลี่ยในการแบ่งข้อมูล 5000 ตัว และใช้เวลานานขึ้นเมื่อความห่างของค่าเฉลี่ยข้อมูลมีค่ามาก
- การแบ่งช่วงธรรมชาติเจงค์ใช้เวลาใกล้เคียงกันในทุกความห่างของค่าเฉลี่ยข้อมูลโดยใช้เวลาน้อยกว่าการจัดกลุ่มข้อมูลด้วย EM ยกเว้นเมื่อความห่างของค่าเฉลี่ยเท่ากับ 2
- head/tail breaks ใช้เวลาใกล้เคียง 0 ในทุกกรณี
- การจัดกลุ่มข้อมูลด้วย EM ใช้เวลาลดลงอย่างมากเมื่อข้อมูลนั้นมีความห่างของค่าเฉลี่ยเท่ากับ 2

4.2. การแจงแจงล็อกปกติแบบผสม

4.2.1. กรณีความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 0.5

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	HT_2	EM	$no\ HT_2$	$no\ EM$
0.75	1	0.4168	0.4848	0.5366	0.6604	0.6132	*0.7324*	0.5329	16	1
0.75	4	0.5548	0.5855	0.6353	0.7264	0.6665	*0.7418*	0.682	5	1
0.75	7	0.6218	0.653	0.6862	0.7403	0.6915	*0.7434*	0.725	0	2
0.75	10	0.6193	0.6369	0.6678	*0.7455*	0.6876	0.7436	0.7103	0	0
0.5	1	0.5667	0.5853	0.5887	0.5798	*0.5929*	0.5413	0.5207	20	0
0.5	4	0.573	0.5768	0.5798	0.5438	*0.5883*	0.5402	0.5087	5	1
0.5	7	0.5745	0.5729	0.5686	0.5218	*0.5838*	0.5341	0.5049	0	1
0.5	10	0.5849	0.5816	0.5687	0.5209	*0.5876*	0.5358	0.5088	0	3
0.25	1	*0.6947*	0.6437	0.6197	0.4989	0.5454	0.3391	0.3862	22	1
0.25	4	*0.6665*	0.6268	0.5573	0.3942	0.4975	0.3319	0.296	13	0
0.25	7	*0.4747*	0.4386	0.4089	0.2945	0.4316	0.3077	0.2521	0	1
0.25	10	*0.478*	0.4535	0.4096	0.2782	0.4319	0.3054	0.2855	0	4

ตารางที่ 8 ตารางสรุปความแม่นยำในการแบ่งกลุ่มทั้งหมด กรณีการแจกแจงล็อกปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 0.5

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.75

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำต่ำสุดทั้ง 3 รูปแบบ โดยความแม่นยำจะต่ำลงเมื่อค่าร้อยละที่ใช้ในการแบ่งลดลงยกเว้นในกรณี sd_1 เท่ากับ 1 ที่การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.2 มีต่ำสุดเป็นอันดับ 4
- head/tail breaks ที่ใช้จุดแบ่งที่สองเป็นวิธีแบ่งที่ให้ความแม่นยำสูงสุดยกเว้นในกรณี sd_1 เท่ากับ 10 ที่การแบ่งช่วงธรรมชาติเจงค์ให้ความแม่นยำสูงสุด แต่ head/tail breaks ที่ใช้จุดแบ่งที่สองมีความใกล้เคียงกัน

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.5

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำใกล้เคียงกับการแบ่งที่ให้ความแม่นยำสูงสุดทั้ง 3 รูปแบบ โดยความแม่นยำน้อยกว่า 0.01 โดยเฉลี่ย
- head/tail breaks ที่ใช้จุดแบ่งแรกเป็นวิธีแบ่งให้ความแม่นยำสูงสุด

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.25

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.05 ให้ความแม่นยำสูงสุด โดยการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.1 และ 0.2 ให้ความแม่นยำต่ำกว่า 0.04 และ 0.08 โดยเฉลี่ยตามลำดับ
- head/tail breaks ที่ใช้จุดแบ่งแรก นั้นมีความแม่นยำใกล้เคียงกับการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.05 ในกรณี sd_1 เท่ากับ 7 และ 10 นอกเหนือจากนั้นการแบ่งด้วยวิธีอื่นให้ความแม่นยำต่ำกว่าอย่างชัดเจน

4.2.2. กรณีความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 1

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	HT_2	EM	$no\ HT_2$	$no\ EM$
0.75	1	0.4372	0.5346	0.5774	0.6981	0.6633	*0.7651*	0.6941	22	0
0.75	4	0.548	0.5866	0.6495	0.7502	0.7018	*0.7622*	0.7096	4	0
0.75	7	0.6711	0.6984	0.727	0.7596	0.7354	*0.7665*	0.6915	0	2
0.75	10	0.6746	0.7048	0.7345	0.7579	0.7416	*0.7647*	0.72	0	2
0.5	1	0.6092	0.6463	0.6599	0.657	*0.6743*	0.5242	0.5618	23	0
0.5	4	0.6091	0.6187	0.6238	0.5738	*0.6334*	0.5608	0.5311	3	3
0.5	7	0.6304	0.6167	0.6064	0.5443	*0.6307*	0.555	0.5263	0	1
0.5	10	0.6095	0.601	0.5829	0.5176	*0.6173*	0.5415	0.504	0	1
0.25	1	*0.7543*	0.7086	0.6793	0.5271	0.5919	0.3537	0.524	21	0
0.25	4	*0.5776*	0.5541	0.5074	0.3476	0.4801	0.3159	0.323	6	3
0.25	7	*0.5671*	0.5344	0.4948	0.3251	0.4791	0.3298	0.3055	0	1
0.25	10	*0.5481*	0.5174	0.4505	0.2908	0.4655	0.318	0.3209	0	1

ตารางที่ 9 ตารางสรุปความแม่นยำในการแบ่งกลุ่มทั้งหมด กรณีการแจกแจงล็อกปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 0.1

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.75

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำต่ำสุดทั้ง 3 รูปแบบ โดยความแม่นยำจะต่ำลงเมื่อค่าร้อยละที่ใช้ในการแบ่งลดลงยกเว้นในกรณี sd_1 เท่ากับ 7 ที่การจัดกลุ่มข้อมูลด้วย EM ให้ความแม่นยำต่ำกว่าการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.1 และ sd_1 เท่ากับ 10 ที่การจัดกลุ่มข้อมูลด้วย EM ให้ความแม่นยำต่ำกว่าการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.2
- head/tail breaks ที่ใช้จุดแบ่งที่สองเป็นวิธีแบ่งที่ให้ความแม่นยำสูงสุด

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.5

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำใกล้เคียงกับการแบ่งที่ให้ความแม่นยำสูงสุดทั้ง 3 รูปแบบ โดยความแม่นยำน้อยกว่า 0.02 โดยเฉลี่ย
- head/tail breaks ที่ใช้จุดแบ่งแรกเป็นวิธีแบ่งให้ความแม่นยำสูงสุด

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.25

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.05 ให้ความแม่นยำสูงสุด โดยการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.1 และ 0.2 ให้ความแม่นยำต่ำกว่า 0.03 และ 0.08 โดยเฉลี่ยตามลำดับ
- การแบ่งด้วยวิธีอื่นให้ความแม่นยำต่ำกว่าอย่างชัดเจน

4.2.3. กรณีความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 2

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	HT_2	EM	$no\ HT_2$	$no\ EM$
0.75	1	0.5429	0.6318	0.6797	*0.8031*	0.7669	0.7915	0.8014	16	1
0.75	4	0.5762	0.6305	0.7122	0.8051	0.7712	*0.808*	0.7036	0	0
0.75	7	0.7341	0.7491	0.7703	0.7729	*0.7923*	0.7875	0.7595	0	0
0.75	10	0.7825	0.7833	0.7862	0.7612	*0.7981*	0.7805	0.7591	0	0
0.5	1	0.6776	0.7439	0.7751	0.7605	*0.7874*	0.5859	0.7627	21	1
0.5	4	0.6799	0.6906	0.7082	0.6063	*0.7115*	0.5852	0.614	3	0
0.5	7	0.6799	0.676	0.6633	0.5505	*0.6805*	0.5681	0.5946	0	0
0.5	10	*0.6685*	0.6572	0.6344	0.5334	0.6685	0.5595	0.5798	0	0
0.25	1	*0.8011*	0.794	0.7615	0.5958	0.6521	0.3137	0.6287	22	0
0.25	4	*0.6393*	0.6269	0.5572	0.353	0.5223	0.334	0.4316	3	0
0.25	7	*0.6335*	0.6047	0.5394	0.3354	0.5193	0.3405	0.4243	0	1
0.25	10	*0.4803*	0.4495	0.4054	0.2774	0.4483	0.3062	0.358	0	4

ตารางที่ 10 ตารางสรุปความแม่นยำในการแบ่งกลุ่มทั้งหมด กรณีการแจกแจงล็อกปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 2

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.75

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำต่ำสุดทั้ง 3 รูปแบบ โดยความแม่นยำจะต่ำลงเมื่อค่าร้อยละที่ใช้ในการแบ่งลดลงยกเว้นในกรณี sd_1 เท่ากับ 4 และ 7 ที่การจัดกลุ่มข้อมูลด้วย EM ให้ความแม่นยำต่ำกว่าการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.2 และ sd_1 เท่ากับ 10 ที่การจัดกลุ่มข้อมูลด้วย EM ให้ความแม่นยำสูงเป็นอันดับ 2, 3, และ 4 สำหรับการใช้ค่าร้อยละ 0.2, 0.1 และ 0.05 ตามลำดับ

- การแบ่งช่วงธรรมชาติเจงค์เป็นวิธีแบ่งที่ให้ความแม่นยำสูงสุดในกรณี sd_1 เท่ากับ 1 head/tail breaks ที่ใช้จุดแบ่งที่สองเป็นวิธีแบ่งที่ให้ความแม่นยำสูงสุดในกรณี sd_1 เท่ากับ 4 และ head/tail breaks ที่ใช้จุดแบ่งที่หนึ่งเป็นวิธีแบ่งที่ให้ความแม่นยำสูงสุดในกรณี sd_1 เท่ากับ 7 และ 10
- การแบ่งช่วงธรรมชาติเจงค์, head/tail breaks ที่ใช้จุดแบ่งที่หนึ่ง, และ head/tail breaks ที่ใช้จุดแบ่งที่สองให้ความแม่นยำใกล้เคียงกัน โดยมีความแม่นยำต่างกันไม่เกิด 0.037

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.5

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำใกล้เคียงกับการแบ่งที่ให้ความแม่นยำสูงสุดทั้ง 3 รูปแบบ โดยความแม่นยำน้อยกว่า 0.02 โดยเฉลี่ย แต่เป็นที่สังเกตว่าการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.05 นั้นที่ความแม่นยำต่ำกว่าการแบ่งที่มีความแม่นยำสูงสุด 0.11 ในกรณี sd_1 เท่ากับ 4
- head/tail breaks ที่ใช้จุดแบ่งแรกเป็นวิธีแบ่งให้ความแม่นยำสูงสุด

กรณีความน่าจะเป็นที่ข้อมูลจะเป็นกลุ่ม 1 มีค่าเท่ากับ 0.25

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.05 ให้ความแม่นยำสูงสุด โดยการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.1 และ 0.2 ให้ความแม่นยำต่ำกว่า 0.02 และ 0.07 โดยเฉลี่ยตามลำดับ
- การแบ่งด้วยวิธีอื่นให้ความแม่นยำต่ำกว่าอย่างชัดเจนยกเว้น head/tail breaks ที่ใช้จุดแบ่งที่หนึ่งในกรณี sd_1 เท่ากับ 10 ที่ต่ำกว่าเพียง 0.03

4.2.4. ผลความแม่นยำในการแบ่งกลุ่ม 1 และความแม่นยำในการแบ่งกลุ่ม 2 โดยภาพรวม

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำในการแบ่งกลุ่ม 1 ต่ำสุดทั้ง 3 รูปแบบโดยความแม่นยำจะลดลงเมื่อค่าร้อยละที่ใช้ในการแบ่งลดลง แต่จะมีบางกรณีที่มีวิธีการแบ่งอื่นมีค่าต่ำกว่าการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำทั้งที่ต่ำกว่าทั้งหมดและต่ำกว่าแค่การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ค่าร้อยละบางค่า โดยส่วนใหญ่จะเป็นการจัดกลุ่มข้อมูลด้วย EM
- การแบ่งช่วงธรรมชาติเจงค์และ head/tail breaks ที่ใช้จุดแบ่งที่สอง ให้ความแม่นยำในการแบ่งกลุ่ม 1 สูงในทุกกรณี
- ในทางกลับกัน การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำให้ความแม่นยำในการแบ่งกลุ่ม 2 สูงสุดทั้ง 3 รูปแบบโดยความแม่นยำจะลดลงเมื่อลดค่าร้อยละที่ใช้ในการแบ่ง แต่จะมีบางกรณีที่มี

head/tail breaks ที่ใช้จุดแบ่งที่หนึ่งมีค่าต่ำกว่าการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำทั้งที่ต่ำกว่าทั้งหมดและต่ำกว่าแค่การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ค่าร้อยละบางค่า

- การแบ่งช่วงธรรมชาติเจงค์และ head/tail breaks ที่ใช้จุดแบ่งที่สอง ให้ความแม่นยำในการแบ่งกลุ่ม 1 ต่ำในทุกกรณี

4.2.5. ระยะเวลาที่ใช้ในการแบ่งกลุ่ม

Gap between Mean	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	HT_2	EM
0.5	141.97	104.6	72.23	14.76	0	0	64.89
1	150.36	108.94	73.19	15.28	0	0	60.38
2	185.15	143.19	99.84	15.13	0	0	25.95

ตารางที่ 11 ตารางสรุปค่าเฉลี่ยของเวลาที่ใช้ในการแบ่งกลุ่มข้อมูลการแจกแจงปกติแบบผสม

- การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำใช้เวลานานกว่าการแบ่งข้อมูลรูปแบบอื่นแต่ใช้เวลาน้อยกว่าการแบ่งข้อมูลการแจกแจงปกติแบบผสม โดยการลดค่าร้อยละที่ใช้ในการแบ่งลง 0.05 จะเพิ่มระยะเวลาที่ใช้ประมาณ 40 วินาทีโดยเฉลี่ยในการแบ่งข้อมูล 5000 ตัว และใช้เวลานานขึ้นเมื่อความห่างของค่าเฉลี่ยข้อมูลมีค่ามาก
- การแบ่งช่วงธรรมชาติเจงค์ใช้เวลาใกล้เคียงกันในทุกความห่างของค่าเฉลี่ยข้อมูลโดยใช้เวลาน้อยกว่าการจัดกลุ่มข้อมูลด้วย EM และใช้เวลาน้อยกว่าการแบ่งข้อมูลการแจกแจงปกติแบบผสม
- head/tail breaks ใช้เวลาใกล้เคียง 0 ในทุกกรณี
- การจัดกลุ่มข้อมูลด้วย EM นั้นใช้เวลานานขึ้นมากเมื่อเทียบกับการแบ่งข้อมูลการแจกแจงปกติแบบผสมและใช้เวลาใกล้เคียงกับการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.2 แต่เวลาที่ใช้ลดลงอย่างมากเมื่อข้อมูลนั้นมีความห่างของค่าเฉลี่ยเท่ากับ 2

บทที่ 5

สรุปผลงานวิจัย อภิปรายผล และข้อเสนอแนะ

งานวิจัยนี้มีจุดประสงค์เพื่อเปรียบเทียบประสิทธิภาพของการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำในการแบ่งกลุ่มข้อมูลตัวแปรเดียว โดนเปรียบเทียบการแบ่งกลุ่มอื่น ๆ 3 วิธี ได้แก่ การแบ่งช่วงธรรมชาติเจงค์, head/tail break, และ การจัดกลุ่มข้อมูลด้วยอัลกอริทึม EM ในข้อมูลที่มีการแจกแจงแบบการแจงแจงปกติแบบผสม 2 กลุ่ม และการแจกแจงล็อกปกติแบบผสม 2 กลุ่มที่มีค่าพารามิเตอร์ที่แตกต่างกัน ในงานวิจัยได้ใช้การจำลองข้อมูลผ่านโปรแกรม R เพื่อทดสอบประสิทธิภาพ โดยใช้เกณฑ์วัดประสิทธิภาพได้แก่ ความแม่นยำในการแบ่งกลุ่มทั้งหมด และความแม่นยำในการแบ่งของแต่ละกลุ่ม โดยสามารถสรุปผลการศึกษาได้ดังนี้

5.1. สรุปผลงานวิจัย และอภิปรายผล

5.1.1. การแบ่งกลุ่มข้อมูลที่มีการแจกแจงแบบการแจกแจงปกติแบบผสม

ในการแบ่งข้อมูลการแจงแจงปกติแบบผสมนั้นเป็นที่ชัดเจนว่าการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำนั้นไม่เหมาะสมกับการนำมาใช้แบ่งข้อมูลในที่มีการแจกแจงรูปแบบนี้ไม่ว่าจะใช้ค่าร้อยละในการหยุดเท่าไหร่ก็ตามเนื่องจากความแม่นยำในการแบ่งนั้นต่ำกว่าวิธีการแบ่งรูปแบบอื่น ยกเว้นในกรณีที่ข้อมูลส่วนใหญ่นั้นอยู่ในกลุ่ม 2 ซึ่งคือกลุ่มที่มีค่าเฉลี่ยสูงกว่า ซึ่งความแม่นยำที่สูงกว่าวิธีอื่นนั้นเกิดจากความแม่นยำในการแบ่งกลุ่ม 2 ที่มีจำนวนมากเป็นส่วนใหญ่

ในการแบ่งข้อมูลที่กลุ่ม 1 มีจำนวนมากกว่าหรือข้อมูลแต่ละกลุ่มมีจำนวนเท่ากัน การแบ่งช่วงธรรมชาติเจงค์และการใช้จุดแบ่งแรกของ head/tail break ซึ่งคือการใช้ค่าเฉลี่ยของข้อมูลในการแบ่งจะมีประสิทธิภาพใกล้เคียงกันและดีที่สุดหากข้อมูลนั้นมีความห่างของค่าเฉลี่ยกลุ่มไม่มาก หากข้อมูลนั้นมีความห่างของค่าเฉลี่ยกลุ่มสูงการแบ่งกลุ่มด้วย EM จะมีประสิทธิภาพมากที่สุดหรือใกล้เคียงไม่ว่าข้อมูลในแต่ละกลุ่มจะมีเท่ากันหรือไม่ก็ตาม

5.1.2. การแบ่งกลุ่มข้อมูลที่มีการแจกแจงแบบการแจกแจงล็อกปกติแบบผสม

ในการแบ่งกลุ่มข้อมูลการแจกแจงล็อกปกติแบบผสมนั้นประสิทธิภาพของการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำนั้นมีความหลากหลายโดยขึ้นอยู่กับอัตราส่วนของกลุ่มข้อมูลเป็นหลัก

ในกลุ่มที่มีข้อมูลกลุ่ม 1 จำนวนมากกว่า การแบ่งโดยใช้จุดแบ่งที่ 2 ของ head/tail break มีประสิทธิภาพที่สุดหรือมีประสิทธิภาพใกล้เคียงสูงสุดโดยความแม่นยำนั้นเกิดขึ้นจากกการแบ่งข้อมูลกลุ่ม 1 ถูกต้องเป็นส่วนใหญ่ส่งผลความแม่นยำกลุ่ม 2 น้อย ในขณะที่การแบ่งกลุ่มด้วยการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำมีประสิทธิภาพต่ำ ไม่เหมาะสมกับการนำมาแบ่งข้อมูลที่มีกลุ่ม 1 จำนวนมาก

ในกลุ่มที่มีข้อมูลแต่ละกลุ่มจำนวนใกล้เคียงกัน การแบ่งโดยใช้จุดแบ่งที่ 1 ของ head/tail break มีประสิทธิภาพที่สุด แต่การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำมีประสิทธิภาพไม่ต่างกันมากจึงสามารถนำมาใช้ได้ โดยการใช้ค่าร้อยละที่ 0.1 มีความเหมาะสมที่สุด โดยเฉพาะเมื่อต้องการให้ความสำคัญกับกลุ่ม 2 มากกว่า เนื่องจากการแบ่งรูปแบบอื่นนั้นให้ความแม่นยำกลุ่ม 1 แต่ความแม่นยำกลุ่ม 2 ต่ำ

ในกลุ่มที่มีข้อมูลกลุ่ม 2 จำนวนมากกว่า การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำที่ใช้ค่าร้อยละ 0.05 มีประสิทธิภาพที่สุดและเหมาะสมกับการใช้แบ่งข้อมูลลักษณะนี้

การแบ่งด้วย EM นั้นไม่มีในการทดลองไหนเลยที่ให้ประสิทธิภาพสูงที่สุดและส่วนมากให้ค่าต่ำกว่าการแบ่งกลุ่มที่มีประสิทธิภาพสูงสุดมาก จึงไม่เหมาะสมกับการนำมาแบ่งกลุ่มข้อมูลที่มีการแจกแจงล็อกปกติแบบผสม

5.1.3. เวลาที่ใช้ในการแบ่งกลุ่ม

ในระหว่างจำลองข้อมูลผู้วิจัยได้เก็บข้อมูลระยะเวลาที่ใช้ในการแบ่งกลุ่มของแต่ละวิธีด้วย เป็นที่สังเกตว่าการแบ่งช่วงธรรมชาติเจงค์แบบซ้ำใช้เวลานานกว่าการแบ่งกลุ่มวิธีอื่นมาก เนื่องจากเป็นการใช้การแบ่งช่วงธรรมชาติเจงค์หลาย ๆ ครั้งซึ่งใช้เวลาประมาณครั้งละ 15 วินาทีในการแบ่งกลุ่มข้อมูล 5000 จุดข้อมูล และการแบ่งช่วงธรรมชาติเจงค์นั้นใช้เวลานานขึ้นเรื่อย ๆ หากมีจำนวนข้อมูลที่เพิ่มมากขึ้นเนื่องจากต้องคำนวณค่า SDCM ในทุกกลุ่มที่เป็นไปได้ การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำจึงใช้เวลานานขึ้นเป็นหลายเท่าจากการแบ่งช่วงธรรมชาติเจงค์ การแบ่งช่วงธรรมชาติเจงค์แบบซ้ำจึงไม่เหมาะสมเมื่อต้องการแบ่งข้อมูลที่มีจำนวนมาก

5.2. ข้อจำกัดของงานวิจัยและข้อเสนอแนะ

ในงานวิจัยนี้ผู้วิจัยได้ใช้ความแม่นยำในการแบ่งกลุ่มข้อมูลในการวัดประสิทธิภาพของวิธีการแบ่ง โดยไม่ได้สนใจการวัดประสิทธิภาพรูปแบบอื่นที่ใช้ทั่วไปในปัญหาการแบ่งกลุ่มเช่นความแตกต่างของข้อมูลในกลุ่มเดียวกันหรือความแตกต่างระหว่างกลุ่ม งานวิจัยต่อไปจึงอาจนำการวัดประสิทธิภาพ

ที่ใช้ในปัญหาการแบ่งกลุ่มมาทดสอบประสิทธิภาพเพิ่มเติมได้ นอกจากนี้ในงานวิจัยได้จำกัดการทดสอบการแบ่งข้อมูลใน 2 รูปแบบการแจกแจง มีปรับเปลี่ยนจำนวนข้อมูลในแต่ละกลุ่มและปรับเปลี่ยนค่าส่วนเบี่ยงเบนมาตรฐานโดยยังคงให้ค่าส่วนเบี่ยงเบนมาตรฐานเท่ากัน งานวิจัยในอนาคตจึงอาจทดสอบในการแจกแจงรูปแบบอื่นหรือทดสอบในข้อมูลที่ค่าส่วนเบี่ยงเบนมาตรฐานไม่เท่ากันได้

บรรณานุกรม

Baah, K., Dubey, B., Harvey, R., & McBean, E. (2015). A risk-based approach to sanitary sewer pipe asset management. *Science of The Total Environment*, 505, 1011-1017. https://doi.org/https://doi.org/10.1016/j.scitotenv.2014.10.040

Behboodian, J. (1970). On the Modes of a Mixture of Two Normal Distributions. *Technometrics*, 12(1), 131-139. https://doi.org/10.2307/1267357

Brigo, D., & Mercurio, F. (2002). Lognormal-mixture Dynamics and Calibration to Market Volatility Smiles. *International Journal of Theoretical and Applied Finance*, 05(04), 427-446. https://doi.org/10.1142/s0219024902001511

Chen, J., Yang, S., Li, H., Zhang, B., & Lv, J. (2013). Research on Geographical Environment Unit Division Based on the Method of Natural Breaks (Jenks). *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XL-4/W3*, 47-50. https://doi.org/10.5194/isprsarchives-XL-4-W3-47-2013

Chen, Y., & Gupta, M. R. (2010). EM Demystified: An Expectation-Maximization Tutorial.

Jenks, G. F., & University of Kansas Department of Geography. (1977). *Optimal Data Classification For Choropleth Maps*. University of Kansas. https://books.google.co.th/books?id=HvAENQAACAAJ

Jiang, B. (2012). Head/tail Breaks: A New Classification Scheme for Data with A Heavy-Tailed Distribution. *Professional Geographer - PROF GEOGR*, 65. https://doi.org/10.1080/00330124.2012.700499

Kayano, K., & Shimizu, K. (1994). Optimal Thresholds for a Mixture of Lognormal Distributions as the Continuous Part of the Mixed Distribution. *Journal of Applied Meteorology and Climatology*, 33(12), 1543-1550. https://doi.org/10.1175/1520-0450(1994)033<1543:Otfamo>2.0.Co;2

Li, K., Ma, Z., Robinson, D., & Ma, J. (2018). Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering. *Applied Energy*, 231, 331-342. https://doi.org/https://doi.org/10.1016/j.apenergy.2018.09.050

Lin, Y. (2013). *A Comparison Study on Natural and Head/tail Breaks Involving Digital Elevation Models* [Student thesis, DiVA. http://urn.kb.se/resolve?urn=urn:nbn:se:hig:diva-15609

Lubrano, M., & Ndoye, A. A. J. (2016). Income inequality decomposition using a finite mixture of log-normal distributions: A Bayesian approach. *Computational Statistics & Data Analysis*, 100, 830-846. https://doi.org/https://doi.org/10.1016/j.csda.2014.10.009

Patel, E., & Kushwaha, D. S. (2020). Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model. *Procedia Computer Science*, 171, 158-167. https://doi.org/https://doi.org/10.1016/j.procs.2020.04.017

Qiu, D., & 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. https://doi.org/10.1016/j.jspi.2007.03.045

Wang, Z., Da Cunha, C., Ritou, M., & Furet, B. (2019). Comparison of K-means and GMM methods for contextual clustering in HSM. *Procedia Manufacturing*, 28, 154-159. https://doi.org/https://doi.org/10.1016/j.promfg.2018.12.025

Zhang, X., Barnes, S., Golden, B., Myers, M., & Smith, P. (2019). Lognormal-based mixture models for robust fitting of hospital length of stay distributions. *Operations Research for Health Care*, 22, 100184. https://doi.org/https://doi.org/10.1016/j.orhc.2019.04.002

ภาคผนวก

ตารางแสดงข้อมูลความแม่นยำแยกกลุ่มของการแจกแจงปกติแบบผสม

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	EM
0.75	1	0.0808	0.1668	0.2461	*0.55*	0.55	0.4974
0.75	4	0.0313	0.0428	0.0657	0.5521	0.5503	*0.5863*
0.75	7	0.0368	0.0459	0.0643	0.5506	*0.5507*	0.4374
0.75	10	0.0396	0.0532	0.0749	0.5491	0.5489	*0.5592*
0.5	1	0.1297	0.2294	0.2799	*0.6013*	0.5999	0.5682
0.5	4	0.0386	0.0528	0.0993	0.5973	*0.598*	0.4117
0.5	7	0.046	0.0537	0.0731	*0.6004*	0.5982	0.574
0.5	10	0.0449	0.061	0.0742	0.5953	*0.5968*	0.5184
0.25	1	0.139	0.2384	0.341	0.6423	*0.6457*	0.462
0.25	4	0.0505	0.0612	0.1231	0.6439	*0.6446*	0.534
0.25	7	0.0445	0.0567	0.0781	0.6383	*0.6397*	0.5771
0.25	10	0.0544	0.0657	0.0935	0.6436	*0.6448*	0.5336

ตารางที่ 12 ตารางสรุปความแม่นยำในการแบ่งกลุ่ม 1 กรณีการแจกแจงปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 0.5

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	EM
0.75	1	0.1192	0.2608	0.3232	0.6065	0.5978	*0.6613*
0.75	4	0.0389	0.048	0.0791	*0.6085*	0.5987	0.5735
0.75	7	0.0388	0.0483	0.0588	0.612	0.6005	*0.6719*
0.75	10	0.042	0.0515	0.0805	*0.6068*	0.5971	0.5966
0.5	1	0.1722	0.3275	0.3796	0.6892	*0.6895*	0.5165
0.5	4	0.0406	0.0558	0.1508	*0.6926*	0.6898	0.685
0.5	7	0.0421	0.0537	0.0661	0.6902	*0.6905*	0.6538
0.5	10	0.0537	0.0686	0.0781	0.6905	*0.6909*	0.6154
0.25	1	0.2194	0.3547	0.4269	0.7674	*0.7751*	0.6273
0.25	4	0.0755	0.0952	0.2076	0.7581	*0.7677*	0.4108
0.25	7	0.0681	0.0758	0.106	0.7682	*0.7769*	0.62
0.25	10	0.0791	0.0994	0.1158	0.7676	*0.7743*	0.5806

ตารางที่ 13 ตารางสรุปความแม่นยำในการแบ่งกลุ่ม 1 กรณีการแจกแจงปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 1

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	EM
0.75	1	0.1142	0.2191	0.3102	0.7618	0.692	*0.9092*
0.75	4	0.0418	0.0604	0.1118	0.7601	0.6906	*0.9321*
0.75	7	0.0423	0.0548	0.0671	0.761	0.6919	*0.9277*
0.75	10	0.0488	0.0632	0.0926	0.7583	0.6934	*0.9005*
0.5	1	0.1925	0.3421	0.4475	0.84	0.8412	*0.8482*
0.5	4	0.0613	0.0769	0.1378	0.841	*0.8412*	0.8238
0.5	7	0.0694	0.0726	0.0924	*0.8424*	0.8423	0.8415
0.5	10	0.0642	0.0768	0.0911	*0.8378*	0.8375	0.8275
0.25	1	0.2858	0.563	0.6559	0.9027	*0.9329*	0.7088
0.25	4	0.1111	0.1342	0.2482	0.9042	*0.9348*	0.6723
0.25	7	0.0849	0.1091	0.1695	0.8989	*0.9321*	0.697
0.25	10	0.0914	0.1089	0.1266	0.9016	*0.9328*	0.6942

ตารางที่ 14 ตารางสรุปความแม่นยำในการแบ่งกลุ่ม 1 กรณีการแจกแจงปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 2

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	EM
0.75	1	*0.9705*	0.9226	0.8747	0.6484	0.649	0.5839
0.75	4	*0.9912*	0.9866	0.9765	0.6469	0.6479	0.5045
0.75	7	*0.9882*	0.9843	0.9773	0.6492	0.6489	0.6072
0.75	10	*0.9874*	0.9824	0.9745	0.6471	0.6472	0.517
0.5	1	*0.9438*	0.88	0.851	0.5958	0.5978	0.5074
0.5	4	*0.9876*	0.9818	0.959	0.5971	0.5975	0.6549
0.5	7	*0.9855*	0.9824	0.9743	0.5924	0.5952	0.4825
0.5	10	*0.9862*	0.9795	0.9734	0.6014	0.5997	0.551
0.25	1	*0.9409*	0.8773	0.8112	0.554	0.5507	0.6418
0.25	4	*0.9835*	0.9793	0.9487	0.5496	0.5491	0.5354
0.25	7	*0.9852*	0.9803	0.9712	0.55	0.5482	0.4682
0.25	10	*0.9813*	0.9766	0.9646	0.5526	0.5502	0.531

ตารางที่ 15 ตารางสรุปความแม่นยำในการแบ่งกลุ่ม 2 กรณีการแจกแจงปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 0.5

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	EM
0.75	1	*0.9846*	0.9443	0.9243	0.7692	0.7762	0.4983
0.75	4	*0.9971*	0.996	0.9919	0.7662	0.7737	0.5536
0.75	7	*0.9968*	0.9955	0.9944	0.7664	0.7751	0.5612
0.75	10	*0.9966*	0.9953	0.9913	0.7612	0.7696	0.5828
0.5	1	*0.9696*	0.9161	0.8947	0.69	0.6896	0.6325
0.5	4	*0.9969*	0.9948	0.9737	0.6841	0.6864	0.4642
0.5	7	*0.9968*	0.9953	0.9935	0.6929	0.6924	0.5317
0.5	10	*0.9946*	0.9931	0.992	0.6923	0.6915	0.5218
0.25	1	*0.9572*	0.8989	0.8686	0.6084	0.5991	0.547
0.25	4	*0.9918*	0.9876	0.9545	0.61	0.5976	0.7194
0.25	7	*0.9931*	0.992	0.9866	0.6107	0.6006	0.5372
0.25	10	*0.9922*	0.9886	0.9859	0.608	0.5987	0.6179

ตารางที่ 16 ตารางสรุปความแม่นยำในการแบ่งกลุ่ม 2 กรณีการแจกแจงปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 1

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	EM
0.75	1	*0.9986*	0.9948	0.9905	0.9015	0.9328	0.6919
0.75	4	*0.9999*	0.9998	0.999	0.9013	0.9331	0.6634
0.75	7	*0.9999*	0.9998	0.9997	0.9054	0.9363	0.6755
0.75	10	*0.9998*	0.9997	0.9996	0.9023	0.9313	0.7158
0.5	1	*0.9968*	0.9863	0.9781	0.8449	0.8439	0.8136
0.5	4	*0.9998*	0.9997	0.9981	0.8418	0.8412	0.8392
0.5	7	*0.9997*	0.9997	0.9994	0.841	0.8413	0.8264
0.5	10	*0.9996*	0.9995	0.9994	0.841	0.8415	0.8278
0.25	1	*0.9937*	0.9546	0.937	0.7557	0.691	0.9086
0.25	4	*0.9991*	0.9987	0.9931	0.7558	0.6889	0.9305
0.25	7	*0.9996*	0.9992	0.9974	0.76	0.6922	0.9238
0.25	10	*0.9995*	0.9992	0.9989	0.7599	0.6922	0.9184

ตารางที่ 17 ตารางสรุปความแม่นยำในการแบ่งกลุ่ม 2 กรณีการแจกแจงปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 2

ตารางแสดงข้อมูลความแม่นยำแยกกลุ่มของการแจกแจงล็อกปกติแบบผสม

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	HT_2	EM
0.75	1	0.2772	0.3923	0.4879	0.7414	0.6334	*0.911*	0.5527
0.75	4	0.5321	0.5898	0.6831	0.907	0.7395	*0.9373*	0.859
0.75	7	0.6672	0.7303	0.8032	*0.9516*	0.7924	0.9458	0.9442
0.75	10	0.6541	0.689	0.7532	*0.9635*	0.7841	0.9451	0.9219
0.5	1	0.3082	0.4397	0.4935	0.7536	0.6657	*0.9464*	0.8265
0.5	4	0.5149	0.5914	0.6884	0.9094	0.7667	*0.9464*	0.8266
0.5	7	0.751	0.772	0.8354	0.9706	0.8368	0.9623	*0.972*
0.5	10	0.7411	0.7768	0.8426	*0.9772*	0.8379	0.9632	0.8991
0.25	1	0.2962	0.4709	0.5336	0.7563	0.6912	*0.9481*	0.7923
0.25	4	0.3883	0.4914	0.6406	0.8807	0.7557	*0.9503*	0.9155
0.25	7	0.7523	0.8012	0.851	0.9728	0.8639	0.9702	*0.9995*
0.25	10	0.7882	0.823	0.872	*0.9849*	0.8701	0.9725	0.9435

ตารางที่ 18 ตารางสรุปความแม่นยำในการแบ่งกลุ่ม 1 กรณีการแจกแจงล็อกปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 0.5

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	HT_2	EM
0.75	1	0.2767	0.4284	0.5022	0.7333	0.6528	*0.9552*	0.7903
0.75	4	0.4799	0.543	0.6592	0.9046	0.7488	*0.9402*	0.902
0.75	7	0.716	0.7705	0.8389	*0.9764*	0.8279	0.9635	0.8532
0.75	10	0.7258	0.7831	0.8533	*0.9806*	0.8417	0.9673	0.9135
0.5	1	0.3562	0.5042	0.5833	0.8	0.7376	*0.9925*	0.6979
0.5	4	0.5826	0.6446	0.7523	0.9406	0.8329	*0.9666*	0.8505
0.5	7	0.7016	0.7836	0.8591	*0.9746*	0.863	0.9737	0.8417
0.5	10	0.8687	0.8842	0.9311	*0.9943*	0.9096	0.985	0.9255
0.25	1	0.405	0.5761	0.6555	0.8653	0.8053	*0.9741*	0.6021
0.25	4	0.6785	0.742	0.8184	0.9611	0.8891	*0.9832*	0.9003
0.25	7	0.7275	0.7963	0.8481	0.9775	0.8906	*0.9803*	0.9201
0.25	10	0.7855	0.8365	0.9	*0.9906*	0.9076	0.9844	0.9153

ตารางที่ 19 ตารางสรุปความแม่นยำในการแบ่งกลุ่ม 1 กรณีการแจกแจงล็อกปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 1

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	HT_2	EM
0.75	1	0.4402	0.5682	0.6394	0.8725	0.7652	*0.976*	0.8712
0.75	4	0.4869	0.5712	0.7053	0.9446	0.7953	*0.9642*	0.7329
0.75	7	0.7929	0.839	0.8858	*0.9892*	0.8775	0.9823	0.911
0.75	10	0.907	0.9156	0.9463	*0.9969*	0.9074	0.9886	0.9162
0.5	1	0.4185	0.5847	0.7118	0.9025	0.8533	*0.9877*	0.7364
0.5	4	0.6156	0.7007	0.8321	0.9751	0.8976	*0.9869*	0.7281
0.5	7	0.8156	0.8678	0.9122	*0.993*	0.9224	0.99	0.91
0.5	10	0.8605	0.869	0.94	*0.9961*	0.936	0.9926	0.8417
0.25	1	0.56	0.6957	0.8034	0.9473	0.9266	*0.9936*	0.7532
0.25	4	0.7722	0.8133	0.8961	0.9896	0.9472	*0.9939*	0.839
0.25	7	0.8048	0.8648	0.9198	0.9935	0.9468	*0.994*	0.8509
0.25	10	0.9321	0.9531	0.9735	*0.9986*	0.9663	0.9967	0.9033

ตารางที่ 20 ตารางสรุปความแม่นยำในการแบ่งกลุ่ม 1 กรณีการแจกแจงล็อกปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 2

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	HT_2	EM
0.75	1	*0.8401*	0.7657	0.6853	0.4168	0.5529	0.1945	0.4716
0.75	4	*0.6216*	0.5706	0.4898	0.1794	0.4453	0.1507	0.1509
0.75	7	*0.4836*	0.4201	0.3347	0.1062	0.3883	0.136	0.0691
0.75	10	*0.5142*	0.4805	0.4123	0.0941	0.399	0.1414	0.0806
0.5	1	*0.8262*	0.7316	0.6844	0.4053	0.5198	0.1353	0.2115
0.5	4	*0.6326*	0.5631	0.4732	0.1797	0.4107	0.1358	0.1893
0.5	7	*0.3983*	0.3739	0.3025	0.0727	0.331	0.1059	0.0368
0.5	10	*0.4261*	0.3834	0.2912	0.0612	0.3354	0.1055	0.1112
0.25	1	*0.8295*	0.7022	0.6487	0.4117	0.4959	0.136	0.2486
0.25	4	*0.7584*	0.6715	0.5294	0.2326	0.4116	0.1269	0.0894
0.25	7	*0.3803*	0.316	0.26	0.0678	0.2869	0.0864	0.0026
0.25	10	*0.3754*	0.331	0.2566	0.0444	0.2869	0.0847	0.0679

ตารางที่ 21 ตารางสรุปความแม่นยำในการแบ่งกลุ่ม 2 กรณีการแจกแจงล็อกปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 0.5

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	HT_2	EM
0.75	1	*0.9192*	0.8532	0.8032	0.5928	0.6951	0.1932	0.4077
0.75	4	*0.7505*	0.7152	0.6199	0.2877	0.5607	0.2302	0.1395
0.75	7	*0.5358*	0.4801	0.389	0.1015	0.4548	0.1679	0.2028
0.75	10	*0.5167*	0.467	0.3772	0.0919	0.4412	0.159	0.1436
0.5	1	*0.8616*	0.7886	0.7374	0.5164	0.6121	0.0741	0.4276
0.5	4	*0.6341*	0.5915	0.4938	0.2054	0.4328	0.1537	0.2097
0.5	7	*0.5585*	0.4503	0.354	0.1134	0.398	0.1357	0.2104
0.5	10	*0.3491*	0.3171	0.232	0.0372	0.3232	0.0946	0.0819
0.25	1	*0.8718*	0.7532	0.6874	0.4135	0.5202	0.1447	0.4976
0.25	4	*0.542*	0.4895	0.4023	0.1412	0.3425	0.0907	0.1268
0.25	7	*0.5153*	0.4481	0.3777	0.1069	0.3419	0.1122	0.0994
0.25	10	*0.4691*	0.411	0.3006	0.0571	0.3178	0.0955	0.1229

ตารางที่ 22 ตารางสรุปความแม่นยำในการแบ่งกลุ่ม 2 กรณีการแจกแจงล็อกปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 1

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT_1	HT_2	EM
0.75	1	*0.8447	0.8177	0.7969	0.5941	0.7702	0.2352	0.5956
0.75	4	*0.849	0.813	0.7377	0.388	0.7009	0.34	0.6187
0.75	7	*0.5567	0.4754	0.4199	0.1212	0.535	0.2006	0.299
0.75	10	0.4079	0.3848	0.304	0.0524	*0.469*	0.1544	0.2873
0.5	1	*0.9376*	0.9036	0.8392	0.6189	0.7218	0.1841	0.7885
0.5	4	*0.7449*	0.6809	0.5843	0.237	0.5254	0.1831	0.503
0.5	7	*0.5428*	0.4815	0.4103	0.1024	0.4355	0.141	0.2757
0.5	10	*0.476*	0.4444	0.3265	0.0667	0.3991	0.1229	0.3153
0.25	1	*0.8809*	0.8264	0.7475	0.4785	0.5605	0.0841	0.5873
0.25	4	*0.5946*	0.5645	0.4441	0.1417	0.3811	0.1146	0.2945
0.25	7	*0.5766*	0.5176	0.412	0.1144	0.3758	0.121	0.2821
0.25	10	*0.3302*	0.2824	0.217	0.0383	0.2764	0.0773	0.1773

ตารางที่ 23 ตารางสรุปความแม่นยำในการแบ่งกลุ่ม 2 กรณีการแจกแจงล็อกปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 2

ตารางสรุปเวลาที่ใช้ในการแบ่งกลุ่ม

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT	EM
0.75	1	162.24	104.73	62.02	13.66	0	25.83
0.75	4	243.86	197.54	154.81	13.5	0	18.61
0.75	7	204.94	170.27	137.95	13.41	0	45.52
0.75	10	190.88	156.06	116.42	13.33	0	19.39
0.5	1	139.43	85.7	63.64	14.61	0	20.61
0.5	4	268.9	208.56	155.41	14.34	0	25.02
0.5	7	210.8	185.93	146.29	13.91	0	20.64
0.5	10	211.37	167.25	140.83	13.67	0	23.26
0.25	1	131.63	86.56	46.35	14.41	0	21.15
0.25	4	245.8	206.73	132.88	13.95	0	20.64
0.25	7	234.1	207.62	164.33	13.67	0	21.15
0.25	10	212.27	180.42	136.45	13.37	0	19.61

ตารางที่ 24 ตารางสรุปเวลาที่ใช้โดยเฉลี่ยในการแบ่งกลุ่ม กรณีการแจกแจงปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 0.5

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT	EM
0.75	1	147.2	75.48	52.09	17.18	0	31.12
0.75	4	285.51	251.73	175.6	16.82	0	27.93
0.75	7	281.58	234.72	198.32	16.88	0	17.38
0.75	10	249.7	208.13	154.55	16.85	0	20.17
0.5	1	142.66	89.93	55.83	17.18	0	29.35
0.5	4	323.05	262.85	147.68	16.83	0	21.47
0.5	7	314.18	265.37	228.39	16.94	0	29.8
0.5	10	264.03	206.37	177.54	15.77	0	20.63
0.25	1	150.59	100.13	70.64	17.47	0	24.7
0.25	4	289.9	256.53	157.91	16.82	0	26.91
0.25	7	319.15	290.21	234.88	16.82	0	22.25
0.25	10	266.87	230.25	204.28	16.32	0	16.11

ตารางที่ 25 ตารางสรุปเวลาที่ใช้โดยเฉลี่ยในการแบ่งกลุ่ม กรณีการแจกแจงปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 1

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT	EM
0.75	1	190.36	131.39	95.47	17.26	0	2.74
0.75	4	331.75	257.06	166.05	16.74	0	2.85
0.75	7	302.2	252.12	213.77	16.7	0	2.52
0.75	10	261.68	214.92	167.63	16.37	0	2.54
0.5	1	178.66	115.25	80.95	17.59	0	1.45
0.5	4	302.76	266.11	206.31	16.79	0	1.27
0.5	7	281.67	269.48	230.1	16.81	0	1.37
0.5	10	293.97	254.01	216.32	16.85	0	1.62
0.25	1	135.8	63.01	46.04	16.66	0	2.88
0.25	4	293.33	247.99	174.8	16.52	0	2.67
0.25	7	309.71	266	209.85	16.16	0	2.66
0.25	10	272.06	239.54	202.67	13.91	0	2.01

ตารางที่ 26 ตารางสรุปเวลาที่ใช้โดยเฉลี่ยในการแบ่งกลุ่ม กรณีการแจกแจงปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 2

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT	EM
0.75	1	114.61	75.99	47.79	15.03	0	100.87
0.75	4	154.14	116.73	81.87	14.7	0	56.3
0.75	7	163.29	117.88	83.12	14.44	0	67.02
0.75	10	149.51	128.97	93.96	14.43	0	15.26
0.5	1	112.35	65.76	50.15	15.49	0	31.79
0.5	4	161.88	114.22	74.62	14.84	0	61.54
0.5	7	131.72	116.45	85.84	14.6	0	47.5
0.5	10	146.94	120.21	82.07	14.6	0	104.53
0.25	1	100.63	49.95	37.14	15.35	0	73.28
0.25	4	166.03	107.35	56.42	14.83	0	26.65
0.25	7	146.54	110.34	82.06	14.52	0	39.87
0.25	10	156.03	131.3	91.65	14.26	0	154.06

ตารางที่ 27 ตารางสรุปเวลาที่ใช้โดยเฉลี่ยในการแบ่งกลุ่ม กรณีการแจกแจงล็อกปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 0.5

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT	EM
0.75	1	110.06	68.69	45.39	16.62	0	32.7
0.75	4	159.18	120.92	75.07	15.88	0	25.25
0.75	7	181.45	137.76	92.13	15.68	0	81.72
0.75	10	169.47	128.18	87.08	14.58	0	79.91
0.5	1	94.66	55.83	41.68	16.47	0	43.06
0.5	4	159.21	119.02	74.84	15.75	0	121.61
0.5	7	189.51	133.45	86.38	15.72	0	50.3
0.5	10	142.54	123.95	87.3	14.17	0	47.05
0.25	1	115.84	62.48	43.82	15.25	0	40.28
0.25	4	159.65	117.69	75.98	14.63	0	100.48
0.25	7	151.55	108.69	79.26	14.51	0	59.12
0.25	10	171.23	130.63	89.41	14.14	0	43.06

ตารางที่ 28 ตารางสรุปเวลาที่ใช้โดยเฉลี่ยในการแบ่งกลุ่ม กรณีการแจกแจงล็อกปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 1

p	sd_1	$RJ_{0.05}$	$RJ_{0.1}$	$RJ_{0.2}$	J	HT	EM
0.75	1	190.36	131.39	95.47	17.26	0	2.74
0.75	4	331.75	257.06	166.05	16.74	0	2.85
0.75	7	302.2	252.12	213.77	16.7	0	2.52
0.75	10	261.68	214.92	167.63	16.37	0	2.54
0.5	1	90.74	59.27	37.58	13.49	0	32.7
0.5	4	153.29	102.3	62.89	12.94	0	13.51
0.5	7	134.66	98.77	77.21	12.74	0	15.31
0.5	10	143.44	126.64	75.79	12.69	0	18.8
0.25	1	110.21	75.37	46.33	16.47	0	10.23
0.25	4	177.27	148.96	82.57	15.74	0	16.11
0.25	7	181.36	135.73	86	15.71	0	48.79
0.25	10	144.78	115.78	86.78	14.64	0	145.34

ตารางที่ 29 ตารางสรุปเวลาที่ใช้โดยเฉลี่ยในการแบ่งกลุ่ม กรณีการแจกแจงล็อกปกติแบบผสมที่ความห่างระหว่างค่าเฉลี่ยสองกลุ่มเท่ากับ 2

ประวัติผู้เขียน

ชื่อ-สกุล	วิชญ์ยุตม์ สุขแพทย์
วัน เดือน ปี เกิด	28 กุมภาพันธ์ 2541
สถานที่เกิด	กรุงเทพมหานคร
วุฒิการศึกษา	สำเร็จการศึกษาหลักสูตรเศรษฐศาสตรบัณฑิต(ศ.บ.) คณะเศรษฐศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2562 และ เข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาสถิติ ภาควิชาสถิติคณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2563
ที่อยู่ปัจจุบัน	1021/29 เพชรเกษม 106 แขวงหนองค้างพลู เขตหนองแขม กรุงเทพมหานคร 10160