มาตรวัดระยะห่างสำหรับข้อมูลแบบผสมกับการวิเคราะห์กลุ่ม

Distance Measures for Mixed Data with Application in Cluster Analysis

พิชญา บุตรขุนทอง[1*] และ อัครินทร์ ไพบูลย์พานิช[2]

[1,2]ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย

[*]pitchaya.bu@gmail.com

บทคัดย่อ

การศึกษาวิจัยนี้ได้เปรียบเทียบประสิทธิภาพการวิเคราะห์กลุ่มข้อมูลแบบผสม ซึ่งประกอบด้วยตัวแปรนามบัญญัติ ตัวแปรอันดับ และตัวแปรเชิงปริมาณ ด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ โดยใช้มาตรวัดระยะห่างแบบต่าง ๆ คือ ระยะห่างของ Kaufman and Rousseeuw (KR) ระยะห่างของ Podani (P) ซึ่งทั้งสองพัฒนามาจากความคล้ายของ Gower และมาตรวัดระยะห่างที่เสนอขึ้นใหม่โดยประยุกต์ระยะห่างสำหรับตัวแปรนามบัญญัติของ Noorbehbahani et al. (N) ร่วมกับ KR และ P ทำให้ระบุความต่างระหว่างข้อมูลได้ละเอียดยิ่งขึ้น โดยจำลองข้อมูลแบบผสมที่กำหนดให้ทราบกลุ่มแน่ชัด รวมถึงพิจารณากรณีที่ความถี่ของแต่ละประเภทหรืออันดับข้อมูลไม่แตกต่างและแตกต่างกัน ผลการศึกษาพบว่ากรณีที่ความถี่ของแต่ละประเภทหรืออันดับแตกต่างกัน การวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ที่ใช้ระยะห่างแบบผสม KR ร่วมกับ N มีประสิทธิภาพดีกว่าการวิเคราะห์กลุ่มด้วยระยะห่างแบบอื่น ๆ แต่กรณีที่ความถี่ของแต่ละประเภทหรืออันดับไม่แตกต่างกัน พบว่าการวิเคราะห์กลุ่มด้วยระยะห่างแบบต่าง ๆ มีประสิทธิภาพใกล้เคียงกัน

คำสำคัญ: ตัวแปรเชิงปริมาณ, ตัวแปรนามบัญญัติ, ตัวแปรอันดับ, ข้อมูลแบบผสม, ระยะห่าง

Abstract

This study presents comparison of performance of cluster analysis through Partitioning Around Medoids algorithm, for mixed data which contains numerical, nominal, and ordinal variables, using different types of distance measures: Kaufman and Rousseeuw distance (KR) and Podani distance (P) (both are applied from Gower's similarity), and two newly proposed distance measures: one is a combination between KR and Noorbehbahani et al. distance (KR&N) and the other is a combination between P and Noorbehbahani et al. distance (P&N). Mixed data were simulated with equal and unequal frequency of nominal and ordinal variables. In case of unequal frequency data, the clustering using KR&N distance gives better result. However, in case of equal frequency data, the clustering using different four distances shows similar efficiency.

Keywords: quantitative variable, nominal variable, ordinal variable, mixed data, distance

1. บทนำ

การวิเคราะห์กลุ่ม (Cluster analysis) เป็นการวิเคราะห์แบ่งกลุ่มข้อมูลหรือจำแนกข้อมูล ซึ่งข้อมูลที่มีลักษณะคล้ายคลึงกันจะถูกจัดให้อยู่ในกลุ่มเดียวกัน และข้อมูลที่มีลักษณะแตกต่างกันจะถูกจัดให้อยู่ต่างกลุ่มกัน การวิเคราะห์กลุ่มเป็นเทคนิคที่ถูกนำไปใช้ในหลายแขนงวิชา ทั้งในสังคมศาสตร์ วิทยาศาสตร์ และอื่น ๆ อีกมากมาย

ในทางปฏิบัติ ข้อมูลจริงที่ใช้ในการวิเคราะห์กลุ่ม อาจเป็นทั้งตัวแปรเชิงปริมาณ (Quantitative variable) และตัวแปรเชิงคุณภาพ (Qualitative variable) ซึ่งแบ่งได้อีกเป็น ตัวแปรนามบัญญัติ (Nominal variable) และตัวแปรอันดับ (Ordinal variable) เช่น การจำแนกกลุ่มผู้ตอบแบบสอบถามเกี่ยวกับการซื้อรถยนต์ โดยพิจารณาจากรายได้ อายุ ซึ่งเป็นตัวแปรเชิงปริมาณ อาชีพ ซึ่งเป็นตัวแปรนามบัญญัติ และระดับการศึกษา ซึ่งเป็นตัวแปรอันดับ เพื่อหาแนวทางการพัฒนารถยนต์ให้ตรงตามความต้องการเฉพาะกลุ่ม หากตัดตัวแปรตัวใดไป ผลที่ได้จากการวิเคราะห์กลุ่มจะขาดลักษณะของตัวแปรนั้น ๆ ส่งผลให้การกำหนดแนวทางการพัฒนารถยนต์ให้สอดคล้องกับลักษณะที่แท้จริงของกลุ่มและตรงตามความต้องการเฉพาะกลุ่มของลูกค้าเป็นไปได้ยาก นั่นคือการคัดเลือกตัวแปรในแต่ละชุดข้อมูลสามารถส่งผลต่อประสิทธิภาพที่ดีของการวิเคราะห์กลุ่ม ทำให้ในหลายสถานการณ์ไม่สามารถหลีกเลี่ยงการใช้ตัวแปรชนิดใดชนิดหนึ่งได้ ซึ่งข้อมูลที่ประกอบไปด้วยตัวแปรทั้งสามชนิดข้างต้นนี้ เรียกว่า ข้อมูลแบบผสม (Mixed data)

การพิจารณาว่าข้อมูลสองข้อมูลคล้ายคลึงกันหรือแตกต่างกัน ต้องอาศัยมาตรวัดความคล้าย (Similarity measure) หรือมาตรวัดระยะห่าง (Distance measure) สำหรับจุดข้อมูลแต่ละคู่ (Everitt et al., 2011: 43) ทั้งนี้มาตรวัดโดยส่วนใหญ่จะใช้ได้เฉพาะข้อมูลที่ประกอบไปด้วยตัวแปรเพียงชนิดใดชนิดหนึ่งเท่านั้น เช่น ระยะห่างแบบยูคลิเดียน (Euclidean distance) ใช้ได้กับตัวแปรเชิงปริมาณเท่านั้น หรือสัมประสิทธิ์แบบแจคการ์ด (Jaccard coefficient) ใช้ได้เฉพาะตัวแปรทวินาม (Binary variable) หรือตัวแปรนามบัญญัติที่ถูกแปลงเป็นตัวแปรดัมมี่ (Dummy variable) แล้ว สำหรับตัวแปรอันดับ ยังไม่มีมาตรวัดระยะห่างโดยเฉพาะ โดยทั่วไปจะใช้มาตรวัดเช่นเดียวกับตัวแปรเชิงปริมาณ ซึ่งไม่เหมาะสม เนื่องจากตัวแปรอันดับบอกเพียงลำดับของข้อมูลว่ามากหรือน้อยกว่ากัน แต่ไม่สามารถระบุระยะห่างที่แท้จริงของแต่ละระดับได้ ทั้งยังไม่สามารถใช้มาตรวัดเช่นเดียวกับตัวแปรนามบัญญัติได้ เนื่องจากทำให้เกิดการสูญเสียอันดับของข้อมูล ทั้งนี้แม้ว่าจะมีมาตรวัดระยะห่างสำหรับตัวแปรแต่ละประเภทแล้ว ในความเป็นจริง หากต้องวิเคราะห์กลุ่มข้อมูลที่ประกอบไปด้วยทั้งตัวแปรเชิงปริมาณ ตัวแปรนามบัญญัติ และตัวแปรอันดับ นั่นคือข้อมูลแบบผสมที่ได้กล่าวไว้ก่อนหน้านี้ มาตรวัดแบบใดที่สามารถวัดระยะห่างของข้อมูลแบบผสมนี้ได้ และทำให้การวิเคราะห์กลุ่มมีประสิทธิภาพมากที่สุด

Gower (1971) ได้เสนอมาตรวัดสำหรับข้อมูลที่ประกอบไปด้วยตัวแปรเชิงปริมาณและตัวแปรนามบัญญัติ โดยที่ระยะห่างสำหรับตัวแปรนามบัญญัติของ Gower จะมีค่าเป็นทวินาม (0 หากข้อมูลคู่นั้นอยู่ประเภทเดียวกัน หรือ 1 หากอยู่ต่างประเภทกัน) ขณะที่ Noorbehbahani et al. (2014) เล็งเห็นว่าระยะห่างสำหรับตัวแปรนามบัญญัติควรมีค่าขึ้นอยู่กับจำนวนความถี่ของแต่ละประเภทข้อมูลด้วย เพื่อให้ระยะห่างสามารถจำแนกข้อมูลได้ละเอียดยิ่งขึ้น อย่างไรก็ตาม ทั้ง Gower และ Noorbehbahani et al. ไม่ได้กล่าวถึงมาตรวัดสำหรับตัวแปรอันดับไว้

Kaufman and Rousseeuw (1990) และ Podani (1999) ต่างเสนอมาตรวัดสำหรับข้อมูลแบบผสม โดยนิยามมาตรวัดระยะห่างสำหรับตัวแปรอันดับเพิ่มเติมจากมาตรวัดของ Gower ทั้งนี้มาตรวัดระยะห่างของ Kaufman and Rousseeuw ให้ผลการวัดระยะห่างตัวแปรอันดับเช่นเดียวกับการวัดระยะห่างข้อมูลตัวแปรอันดับด้วยระยะห่างสำหรับตัวแปรเชิงปริมาณ ขณะที่ Podani เสนอมาตรวัดระยะห่างสำหรับตัวแปรอันดับที่พิจารณาการซ้ำกันของอันดับมาเกี่ยวข้อง ทำให้ระยะห่างระหว่างอันดับต่าง ๆ แตกต่างกัน อย่างไรก็ตาม ระยะห่างสำหรับข้อมูลแบบผสมของ Gower ที่ถูกพัฒนาโดย Kaufman and Rousseeuw เป็นที่นิยมในปัจจุบันในรูปแบบของฟังก์ชัน daisy ในแพ็กเกจ cluster และฟังก์ชัน gower.dist ในแพ็กเกจ StatMatch ในโปรแกรม R

การศึกษาวิจัยนี้ ได้เสนอมาตรวัดระยะห่างสำหรับข้อมูลแบบผสมที่ประกอบไปด้วยตัวแปรนามบัญญัติ ตัวแปรอันดับ และตัวแปรเชิงปริมาณ โดยการประยุกต์ใช้ระยะห่างของ Kaufman and Rousseeuw และระยะห่างของ Podani ซึ่งต่างมีพื้นฐานมาจากมาตรวัดของ Gower ร่วมกับระยะห่างสำหรับตัวแปรนามบัญญัติของ Noorbehbahani et al. เพื่อให้สามารถวัดระยะห่างข้อมูลตัวแปรนามบัญญัติและตัวแปรอันดับได้ละเอียดยิ่งขึ้น ซึ่งคาดว่าเมื่อนำไปใช้ในการวิเคราะห์กลุ่มจะทำให้การวิเคราะห์กลุ่มได้ผลที่มีประสิทธิภาพมากยิ่งขึ้น

นอกจากการเลือกใช้มาตรวัดระยะห่างสำหรับข้อมูลแบบผสมแล้ว ยังต้องพิจารณาเทคนิคการวิเคราะห์กลุ่มที่เหมาะสำหรับข้อมูลแบบผสม และสามารถปรับใช้กับมาตรวัดระยะห่างแบบต่าง ๆ ได้ ซึ่งก็คืออัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ (Partitioning around medoids algorithm, PAM algorithm) ซึ่งเป็นอัลกอริทึมจัดกลุ่มรูปแบบหนึ่งของอัลกอริทึมจัดกลุ่มแบบเคมีดอยด์ (K-medoids clustering algorithm) มีหลักการพื้นฐานคือการแบ่งข้อมูลทุกหน่วยออกเป็นกลุ่มย่อย โดยผลรวมของระยะห่างระหว่างข้อมูลภายในกลุ่มกับจุดศูนย์กลางของกลุ่มนั้นมีค่าน้อยที่สุด และจุดศูนย์กลางกลุ่มคือข้อมูลใด ๆ ในชุดข้อมูลเท่านั้น ซึ่งเป็นหลักการเดียวกับอัลกอริทึมจัดกลุ่มแบบเคมีน (K-Means clustering algorithm) (Madhulatha, 2011: 477) แต่เนื่องจากอัลกอริทึมการจัดกลุ่มแบบเคมีนจะกำหนดค่าของจุดศูนย์กลางกลุ่มใหม่ ซึ่งคำนวณจากค่าเฉลี่ยของข้อมูลภายในกลุ่มนั้น อัลกอริทึมการจัดกลุ่มแบบเคมีนจึงเหมาะสำหรับข้อมูลเชิงปริมาณเท่านั้น และไม่สามารถใช้กับข้อมูลแบบผสม เพราะไม่สามารถหาค่าเฉลี่ยของตัวแปรนามบัญญัติได้ ขณะที่อัลกอริทึมจัดกลุ่มแบบเคมีดอยด์ใช้ข้อมูลในชุดข้อมูลเท่านั้นเป็นจุดศูนย์กลางกลุ่ม จึงไม่มีปัญหาในการสร้างจุดศูนย์กลางกลุ่มใหม่ และเมทริกซ์ระยะห่างของข้อมูลจะคงที่ จึงปรับใช้การวิเคราะห์กลุ่มนี้กับเมทริกซ์ระยะห่างจากมาตรวัดระยะห่างแบบต่าง ๆ ได้

การศึกษาวิจัยนี้ ทำการวิเคราะห์กลุ่มข้อมูลแบบผสม ด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ โดยใช้มาตรวัดระยะห่าง 4 วิธี คือ ระยะห่างของ Kaufman and Rousseeuw (KR) ระยะห่างของ Podani (P) ระยะห่างแบบผสม Kaufman and Rousseeuw ร่วมกับ Noorbehbahani et al. (KR&N) และระยะห่างแบบผสม Podani ร่วมกับ Noorbehbahani et al. (P&N) และเปรียบเทียบประสิทธิภาพการวิเคราะห์กลุ่มด้วยมาตรวัดระยะห่างสำหรับข้อมูลแบบผสมแบบต่าง ๆ เพื่อเป็นแนวทางในการเลือกใช้มาตรวัดระยะห่างสำหรับการวิเคราะห์กลุ่มได้อย่างเหมาะสมกับลักษณะของข้อมูลแบบผสม

2. วัตถุประสงค์ในการวิจัย

1. เพื่อเสนอมาตรวัดระยะห่างสำหรับข้อมูลแบบผสม โดยประยุกต์ใช้ระยะห่างสำหรับตัวแปรนามบัญญัติของ Noorbehbahani et al. ร่วมกับระยะห่างของ Kaufman and Rousseeuw และระยะห่างของ Podani

2. เพื่อเปรียบเทียบประสิทธิภาพในการวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ ที่ใช้ระยะห่างของ Kaufman and Rousseeuw (KR) ระยะห่างของ Podani (P) ระยะห่างแบบผสม Kaufman and Rousseeuw ร่วมกับ Noorbehbahani et al. (KR&N) และระยะห่างแบบผสม Podani ร่วมกับ Noorbehbahani et al. (P&N) สำหรับข้อมูลแบบผสมที่ประกอบไปด้วยตัวแปร 9 ตัวแปร คือ ตัวแปรเชิงปริมาณ ตัวแปรนามบัญญัติ และตัวแปรอันดับ ชนิดละ 3 ตัวแปร

3. เอกสารและงานวิจัยที่เกี่ยวข้อง

Gower (1971) ได้เสนอมาตรวัดความคล้าย สำหรับข้อมูลที่ประกอบไปด้วยตัวแปรเชิงปริมาณและตัวแปรนามบัญญัติ อย่างไรก็ตาม การวิเคราะห์กลุ่มส่วนใหญ่ไม่นิยมวัดความคล้ายระหว่างข้อมูลในการวิเคราะห์ มาตรวัดความคล้ายของ Gower จึงถูกปรับให้อยู่ในรูปของมาตรวัดระยะห่าง นั่นคือ ระยะห่างระหว่างข้อมูลที่ i กับ j

$$D_{ij} = \frac{\sum_{l=1}^{p} d_{ijl}\delta_{ijl}}{\sum_{l=1}^{p} \delta_{ijl}} \qquad (1)$$

โดยที่ p แทน จำนวนตัวแปรของชุดข้อมูล

δ_{ijl} เท่ากับ 0 เมื่อไม่ทราบค่าของข้อมูลที่ i หรือ j ของตัวแปรที่ l

δ_{ijl} เท่ากับ 1 เมื่อทราบค่าของข้อมูลที่ i และ j ของตัวแปรที่ l

และ d_{ijl} แทน ระยะห่างระหว่างข้อมูลที่ i กับ j วัดโดยตัวแปรที่ l

จะเห็นว่า D_{ij} คือระยะห่างระหว่างข้อมูล โดยที่มีค่าขึ้นอยู่กับ d_{ijl} ซึ่งเป็นระยะห่างที่ถูกวัดโดยตัวแปร ดังนั้น d_{ijl} จึงถูกคำนวณด้วยวิธีที่แตกต่างกันขึ้นอยู่กับประเภทของตัวแปรที่ l นั้น ๆ โดย Gower ได้กล่าวถึงวิธีการวัดระยะห่างสำหรับตัวแปรเชิงปริมาณ และตัวแปรนามบัญญัติ ดังนี้

ระยะห่างสำหรับตัวแปรเชิงปริมาณของ Gower

$$d_{ijl} = \frac{\left|x_{il} - x_{jl}\right|}{R_l} \tag{2}$$

โดยที่ x_{hl} แทน ค่าของข้อมูลที่ h ของตัวแปรเชิงปริมาณที่ l เมื่อ $h = i, j$

และ R_l แทน พิสัยของข้อมูลตัวแปรตัวที่ l

ระยะห่างสำหรับตัวแปรนามบัญญัติของ Gower

d_{ijl} เท่ากับ 0 ถ้าข้อมูลที่ i กับ j ของตัวแปรนามบัญญัติที่ l อยู่ต่างประเภทกัน

d_{ijl} เท่ากับ 1 ถ้าข้อมูลที่ i กับ j ของตัวแปรนามบัญญัติที่ l อยู่ในประเภทเดียวกัน

อย่างไรก็ตาม Gower ไม่ได้กล่าวถึงความคล้ายสำหรับตัวแปรอันดับ ซึ่งอาจเป็นตัวแปรสำคัญในการวิเคราะห์กลุ่ม Kaufman and Rousseeuw (1990) จึงเสนอมาตรวัดความต่างสำหรับตัวแปรอันดับเพิ่มเติมจากระยะห่างของ Gower เพื่อให้สามารถวัดระยะห่างข้อมูลแบบผสมได้ โดยแปลงข้อมูลที่เป็นตัวแปรอันดับให้มีค่าอยู่ระหว่าง 0 ถึง 1

ระยะห่างสำหรับตัวแปรอันดับของ Kaufman and Rousseeuw

$$d_{ijl} = \frac{\left|z_{il} - z_{jl}\right|}{R_l} \tag{3}$$

โดยที่

$$z_{hl} = \frac{x_{hl} - 1}{M_l - 1} \tag{4}$$

เมื่อ x_{hl} แทน อันดับของข้อมูลที่ h ของตัวแปรอันดับที่ l สำหรับ $h = i, j$

M_l แทน ค่าอันดับสูงสุดของตัวแปรอันดับที่ l

และ R_l แทน พิสัยของข้อมูลตัวแปรอันดับที่ l ที่ถูกแปลงค่าเป็น z_{hl} แล้ว

นอกจากนี้ Podani (1999) ได้สร้างมาตรวัดความคล้ายสำหรับตัวแปรอันดับเพิ่มเติมจากวิธีเดิมของ Gower เช่นกัน โดยแปลงข้อมูลตัวแปรอันดับเป็นคะแนนอันดับ (Rank Score) ซึ่งเป็นค่าที่ได้จากการเรียงข้อมูลทั้งหมดของตัวแปรนั้นจากน้อยไปมาก และกำหนดอันดับใหม่ตั้งแต่ 1 ถึงจำนวนข้อมูลทั้งหมด จากนั้นบวกรวมค่าอันดับใหม่ที่เดิมมีอันดับเดียวกันทั้งหมด แล้วจึงหารด้วยจำนวนการซ้ำกันของอันดับนั้น ๆ เช่น ชุดข้อมูล 1, 1, 1, 2, 3 และ 3 มีข้อมูลทั้งสิ้น 6 ตัว เรียงเป็นข้อมูลชุดใหม่คือ 1, 2, 3, 4, 5 และ 6 เมื่อพิจารณาการซ้ำกันสามารถแปลงได้เป็นค่าอันดับใหม่ คือ 2, 2, 2, 4, 5.5 และ 5.5 กำหนดให้ใช้คะแนนอันดับแทนค่าอันดับเริ่มต้นและนำไปคำนวณด้วยความคล้ายซึ่ง Podani ได้สร้างขึ้นใหม่ 2 วิธี สำหรับวิธีแรก เมื่อแปลงเป็นค่าความต่างแล้ว จะได้เมทริกซ์ระยะห่าง (Distance matrix) ของข้อมูลที่มีคุณสมบัติเป็นเมทริก (Metric) ขณะที่วิธีที่สอง จะได้เมทริกซ์ระยะห่างที่ไม่มีคุณสมบัติเป็นเมทริก (Non-metric) วิธีนี้จึงไม่สามารถนำไปใช้กับเทคนิคการวิเคราะห์กลุ่มที่ต้องใช้เมทริกซ์ระยะห่างที่มีคุณสมบัติเป็นเมทริกได้ ทั้งนี้สนใจศึกษามาตรวัดของ Podani ที่เมื่อพิจารณาระยะห่างของข้อมูลทุกคู่แล้วได้เป็นเมทริกซ์ระยะห่าง ตามนิยามดังต่อไปนี้

ระยะห่างสำหรับตัวแปรอันดับของ Podani

$$d_{ijl} = \frac{|r_{il} - r_{jl}|}{R_l} \tag{5}$$

โดยที่ r_{hl} แทน คะแนนอันดับของข้อมูลที่ h ของตัวแปรอันดับที่ l เมื่อ $h = i, j$

และ R_l แทน พิสัยของข้อมูลตัวแปรอันดับที่ l ที่ถูกแปลงเป็นคะแนนอันดับแล้ว

Noorbehbahani et al. (2014) ได้นำเสนอมาตรวัดระยะห่างสำหรับข้อมูลที่ประกอบไปด้วยตัวแปรเชิงปริมาณและตัวแปรนามบัญญัติ ซึ่งแตกต่างจากวิธีของ Gower ทั้งนี้ Noorbehbahani et al. เล็งเห็นว่ามาตรวัดระยะห่างของตัวแปรนามบัญญัติไม่ควรมีค่าเพียง 0 และ 1 เท่านั้น แต่ควรมีค่าที่ขึ้นอยู่กับความถี่ของค่าของตัวแปรนั้น ๆ กล่าวคือสำหรับค่าที่ต่างกันของตัวแปรนามบัญญัติ ค่าของตัวแปรที่มีความถี่สูง (เกิดขึ้นบ่อย) สองค่าควรมีระยะห่างน้อย ขณะที่ค่าของตัวแปรที่มีความถี่ต่ำสองค่า หรือค่าของตัวแปรสองค่าที่ค่าหนึ่งมีความถี่ต่ำกับอีกค่ามีความถี่สูง ควรมีระยะห่างที่สูงขึ้น เนื่องจากโอกาสที่จะถูกจัดให้อยู่ในกลุ่มเดียวกันไม่มากนัก อย่างไรก็ตาม Noorbehbahani et al. พิจารณาเพียงกรณีที่ข้อมูลประกอบไปด้วยตัวแปรเชิงปริมาณและตัวแปรนามบัญญัติเท่านั้น และยังให้ค่าน้ำหนักในการวัดระยะห่างสำหรับตัวแปรสองประเภทนี้ไม่เท่ากันอีกด้วย แต่ผู้วิจัยเล็งเห็นว่าระยะห่างสำหรับตัวแปรนามบัญญัติของ Noorbehbahani et al. ที่มีค่าขึ้นกับจำนวนความถี่ของข้อมูลสามารถนำไปประยุกต์ใช้ร่วมกับระยะห่างของ Gower ได้ และจะทำให้ประสิทธิภาพในการวัดระยะห่างสำหรับข้อมูลแบบผสมแม่นยำขึ้น

ระยะห่างสำหรับตัวแปรนามบัญญัติของ Noorbehbahani et al.

ถ้าข้อมูลที่ i กับ j ของตัวแปรนามบัญญัติที่ l อยู่ในประเภทเดียวกัน กำหนดให้ $d_{ijl} = 0$

ถ้าข้อมูลที่ i กับ j ของตัวแปรนามบัญญัติที่ l อยู่ต่างประเภทกัน

$$d_{ijl} = \frac{|f_{il} - f_{jl}| + \min\{f_l\}}{\max\{f_{il}, f_{jl}\}} \tag{6}$$

โดยที่ f_{hl} แทน ความถี่ของข้อมูลที่มีประเภทเดียวกันกับข้อมูลตัวที่ h ของตัวแปรนามบัญญัติที่ l เมื่อ $h = i, j$

$\min\{f_l\}$ แทน ความถี่ที่น้อยที่สุดจากประเภทข้อมูลทั้งหมดของตัวแปรนามบัญญัติที่ l

และ $\max\{f_{il}, f_{jl}\}$ แทน ค่าที่มากที่สุดระหว่าง f_{il} และ f_{jl}

การวัดประสิทธิภาพในการวิเคราะห์กลุ่ม

สำหรับข้อมูลที่ทราบกลุ่มที่แท้จริง นั่นคือทราบอยู่แล้วว่าข้อมูลแต่ละตัวนั้นอยู่ในกลุ่มใด สามารถวัดประสิทธิภาพในการวิเคราะห์กลุ่มด้วย ค่าสถิติแบบแรนด์ (Rand statistic) ค่าสัมประสิทธิ์แบบแจคการ์ด (Jaccard coefficient) และค่าความบริสุทธิ์ (Purity)

กำหนดให้ I_i แทน ข้อมูลที่ i โดยที่ทราบว่า I_i อยู่ในกลุ่มที่ $L(I_i)$ และเมื่อทำการวิเคราะห์กลุ่ม พบว่า I_i ถูกจัดอยู่ในกลุ่มที่ $C(I_i)$

กำหนดให้ $SS = \{(I_i, I_j) \mid C(I_i) = C(I_j)$ และ $L(I_i) = L(I_j)\}$

$SD = \{(I_i, I_j) \mid C(I_i) = C(I_j)$ และ $L(I_i) \neq L(I_j)\}$

$DS = \{(I_i, I_j) \mid C(I_i) \neq C(I_j)$ และ $L(I_i) = L(I_j)\}$

$DD = \{(I_i, I_j) \mid C(I_i) \neq C(I_j)$ และ $L(I_i) \neq L(I_j)\}$

ค่าสถิติแบบแรนด์ (Rand statistic) คือ

$$R = \frac{|SS| + |DD|}{|SS| + |SD| + |DS| + |DD|} \tag{7}$$

ดังนั้น $0 \le R \le 1$ ซึ่งค่าสถิติแบบแรนด์จะพิจารณาการวิเคราะห์กลุ่มที่ให้ผลออกมาถูกต้องในภาพรวม คือพิจารณาทั้งกรณีข้อมูลแต่ละคู่ที่ทราบว่าอยู่กลุ่มเดียวกันและถูกจัดให้อยู่กลุ่มเดียวกัน และกรณีข้อมูลแต่ละคู่ควรอยู่ต่างกลุ่มกันและถูกจัดให้อยู่ต่างกลุ่มกัน ดังนั้น ถ้า R มีค่าใกล้ 1 นั่นคือในภาพรวมการวิเคราะห์กลุ่มมีความถูกต้องมาก และถ้า R มีค่าใกล้ 0 นั่นคือในภาพรวมการวิเคราะห์กลุ่มมีความถูกต้องน้อย

ค่าสัมประสิทธิ์แบบแจคการ์ด (Jaccard coefficient) คือ

$$J = \frac{|SS|}{|SS| + |SD| + |DS|} \tag{8}$$

ดังนั้น $0 \le J \le 1$ ซึ่งค่าสัมประสิทธิ์แบบแจคการ์ดจะพิจารณาการวิเคราะห์กลุ่มที่ให้ผลออกมาถูกต้องเฉพาะกรณีข้อมูลแต่ละคู่ที่ทราบว่าอยู่กลุ่มเดียวกันและยังถูกจัดให้อยู่กลุ่มเดียวกัน ดังนั้น ถ้า J มีค่าใกล้ 1 นั่นคือการวิเคราะห์กลุ่มให้ผลถูกต้องใกล้เคียงกับกลุ่มที่แท้จริง และถ้า J มีค่าใกล้ 0 นั่นคือการวิเคราะห์กลุ่มให้ผลถูกต้องไม่ใกล้เคียงกับกลุ่มที่แท้จริง

ค่าความบริสุทธิ์ (Purity) คือ

$$purity(L,C) = \frac{\sum_k \max_h |L_h \cap C_k|}{N} \tag{9}$$

เมื่อ $L = \{L_1, L_2, \ldots, L_H\}$ แทน เซตของกลุ่มของข้อมูล H กลุ่มที่ทราบกลุ่มที่แท้จริง

$C = \{C_1, C_2, \ldots, C_K\}$ แทน เซตของกลุ่มของข้อมูล K กลุ่มที่ได้จากการวิเคราะห์กลุ่ม

N แทน จำนวนข้อมูลทั้งหมด

ดังนั้น $0 \le purity(L,C) \le 1$ โดยค่าความบริสุทธิ์จะพิจารณาส่วนที่ซ้อนทับกันของกลุ่มข้อมูลที่แท้จริงกับกลุ่มข้อมูลที่ได้จากการวิเคราะห์กลุ่ม ถ้าหากค่าความบริสุทธิ์ใกล้ 1 แสดงว่าการวิเคราะห์กลุ่มมีความถูกต้องมาก ถ้าหากค่าความบริสุทธิ์ใกล้ 0 แสดงว่าการวิเคราะห์กลุ่มมีความถูกต้องน้อย

ค่า Average Silhouette Width

ค่า Average Silhouette Width คือค่าเฉลี่ยของค่า Silhouette จากข้อมูลทั้งหมดในการวิเคราะห์กลุ่ม เป็นค่าที่บอกจำนวนกลุ่มที่เหมาะสมจากการวิเคราะห์กลุ่ม เมื่อไม่ทราบจำนวนกลุ่มที่แท้จริงหรือจำนวนกลุ่มที่ต้องการ โดยเปรียบเทียบผลการวิเคราะห์กลุ่มที่กำหนดจำนวนกลุ่มที่แตกต่างกัน (Rousseeuw, 1987: 53) หากการวิเคราะห์กลุ่มที่มีจำนวนกลุ่มแตกต่างกันแบบใดมีค่า Average Silhouette Width มากกว่า แสดงว่ากลุ่มที่กำหนดนั้นมีความเหมาะสมกับข้อมูลมากกว่า

กำหนด ค่า Silhouette ของข้อมูลที่ i คือ

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{10}$$

โดยที่ $a(i)$ แทน ระยะห่างเฉลี่ยระหว่างข้อมูลที่ i กับข้อมูลอื่น ๆ ทั้งหมดภายในกลุ่มเดียวกัน

$b(i)$ แทน ระยะห่างเฉลี่ยที่น้อยที่สุดของระยะห่างเฉลี่ยระหว่างข้อมูลที่ i กับข้อมูลทั้งหมดที่อยู่กลุ่มอื่น ๆ

ดังนั้น $-1 \le S(i) \le 1$ ซึ่งถ้า $S(i)$ มีค่าใกล้ 1 แสดงว่าข้อมูลที่ i ถูกจัดให้อยู่ในกลุ่มที่เหมาะสมแล้ว ถ้า $S(i)$ มีค่าใกล้ -1 แสดงว่าข้อมูลที่ i ควรอยู่ในกลุ่มอื่น และถ้า $S(i)$ มีค่าใกล้ 0 แสดงว่าอาจอยู่กลุ่มใดกลุ่มหนึ่งก็ได้

ในการวิจัยนี้จะคำนวณค่า Average Silhouette Width เพื่อสังเกตความแตกต่างของค่า Average Silhouette Width ของการวิเคราะห์กลุ่มด้วยมาตรวัดระยะห่างแบบต่าง ๆ เท่านั้น แต่ไม่สามารถนำมาเปรียบเทียบได้ว่าการวิเคราะห์กลุ่มด้วยมาตรวัดระยะห่างแบบใดมีประสิทธิภาพที่ดีกว่า

4. วิธีดำเนินการวิจัย

4.1 ขั้นตอนที่ 1

กำหนดข้อมูลแบบผสม ประกอบไปด้วยตัวแปร 9 ตัวแปร แบ่งเป็นตัวแปรนามบัญญัติ ตัวแปรอันดับ และตัวแปรเชิงปริมาณ ชนิดละ 3 ตัวแปร และกำหนดจำนวนกลุ่มเท่ากับ 3 และ 5 โดยที่จำนวนข้อมูลในแต่ละกลุ่มเท่ากับ 100 ทั้งนี้ ตัวแปรนามบัญญัติและตัวแปรอันดับ มีจำนวนประเภทหรืออันดับ เท่ากับ 5 ซึ่งมีค่าที่เป็นไปได้คือ 1 2 3 4 และ 5 โดยที่ข้อมูลตัวแปรนามบัญญัติและตัวแปรอันดับมีความน่าจะเป็นที่ข้อมูลจะอยู่ในประเภทหรืออันดับต่าง ๆ แตกต่างกันในแต่ละกลุ่มดังแสดงในตารางที่ 1

ตารางที่ 1 ความน่าจะเป็นของข้อมูลตัวแปรนามบัญญัติหรือตัวแปรอันดับ X_I ในแต่ละกลุ่ม ที่จะเกิดขึ้นในประเภทหรืออันดับต่าง ๆ 5 ประเภทหรืออันดับ

กลุ่มที่	$P(X_I=1)$	$P(X_I=2)$	$P(X_I=3)$	$P(X_I=4)$	$P(X_I=5)$	$\sum_{i=1}^{5}P(X_I=i)$
1	0.70	0.10	0.10	0.05	0.05	1.00
2	0.05	0.70	0.10	0.10	0.05	1.00
3	0.05	0.05	0.70	0.10	0.10	1.00
4	0.10	0.05	0.05	0.70	0.10	1.00
5	0.10	0.10	0.05	0.05	0.70	1.00

จากตารางที่ 1 จะพบว่า หากจำนวนกลุ่มเท่ากับ 3 จำนวนความถี่ของแต่ละประเภท/อันดับ 3 ประเภท/อันดับแรกจะมีโอกาสเกิดขึ้นเท่า ๆ กัน แต่จำนวนความถี่ของแต่ละประเภท/อันดับ 2 ประเภท/อันดับหลัง จะมีโอกาสเกิดขึ้นเท่า ๆ กัน แต่น้อยกว่า 3 ประเภท/อันดับแรก นั่นคือความถี่ของแต่ละประเภท/อันดับมีจำนวนแตกต่างกัน ขณะที่ หากจำนวนกลุ่มเท่ากับ 5 จำนวนความถี่ของแต่ละประเภท/อันดับ ทั้ง 5 ประเภท/อันดับ จะมีโอกาสเกิดขึ้นเท่า ๆ กัน

กำหนดตัวแปรเชิงปริมาณมีค่าเฉลี่ยแตกต่างกันระหว่างกลุ่ม ดังแสดงในตารางที่ 2

ตารางที่ 2 ค่าเฉลี่ยของตัวแปรเชิงปริมาณ

กลุ่มที่	ค่าเฉลี่ยของตัวแปรเชิงปริมาณที่		
	1	**2**	**3**
1	1	2	3
2	4	5	6
3	7	8	9
4	10	11	12
5	13	14	15

4.2 ขั้นตอนที่ 2

จำลองข้อมูลที่มีการแจกแจงแบบปกติหลายตัวแปร จำนวน 9 ตัวแปร และกำหนดให้ตัวแปรมีความสัมพันธ์กัน โดยค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.2 และ 0.8 เพื่อศึกษากรณีที่ข้อมูลมีความสัมพันธ์กันน้อยและมาก ตามลำดับ และแปลงข้อมูลที่จำลองนี้ให้เป็นข้อมูลแบบผสมตามลักษณะที่กำหนดในขั้นตอนที่ 1

4.3 ขั้นตอนที่ 3

จำลองข้อมูลแบบผสมกรณีต่าง ๆ กรณีละ 1,000 ชุด

4.4 ขั้นตอนที่ 4

นำข้อมูลแบบผสมที่จำลองขึ้น มาหาเมทริกซ์ระยะห่างจากการคำนวณด้วยมาตรวัดระยะห่างแบบต่าง ๆ ดังนี้

ระยะห่างของ Kaufman and Rousseeuw (KR) คือ ระยะห่างสำหรับข้อมูลแบบผสม โดยวัดระยะห่างข้อมูลที่เป็นตัวแปรเชิงปริมาณและตัวแปรนามบัญญัติตามนิยามของ Gower และวัดระยะห่างข้อมูลที่เป็นตัวแปรอันดับตามนิยามของ Kaufman and Rousseeuw

ระยะห่างของ Podani (P) คือ ระยะห่างสำหรับข้อมูลแบบผสม โดยวัดระยะห่างข้อมูลที่เป็นตัวแปรเชิงปริมาณและตัวแปรนามบัญญัติตามนิยามของ Gower และวัดระยะห่างข้อมูลที่เป็นตัวแปรอันดับตามนิยามของ Podani

ระยะห่างแบบผสม Kaufman and Rousseeuw ร่วมกับ Noorbehbahani et al. (KR&N) คือ ระยะห่างสำหรับข้อมูลแบบผสม โดยวัดระยะห่างตัวแปรข้อมูลที่เป็นตัวแปรเชิงปริมาณตามนิยามของ Gower วัดระยะห่างข้อมูลที่เป็นตัวแปรนามบัญญัติตามนิยามของ Noorbehbahani et al. และวัดระยะห่างข้อมูลที่เป็นตัวแปรอันดับตามนิยามของ Kaufman and Rousseeuw

ระยะห่างแบบผสม Podani ร่วมกับ Noorbehbahani et al. (P&N) คือ ระยะห่างสำหรับข้อมูลแบบผสม โดยวัดระยะห่างข้อมูลที่เป็นตัวแปรเชิงปริมาณตามนิยามของ Gower วัดระยะห่างข้อมูลที่เป็นตัวแปรนามบัญญัติตามนิยามของ Noorbehbahani et al. และวัดระยะห่างข้อมูลที่เป็นตัวแปรอันดับตามนิยามของ Podani

4.5 ขั้นตอนที่ 5

ทำการวิเคราะห์กลุ่มข้อมูลแบบผสมที่จำลองขึ้น ด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ โดยใช้เมทริกซ์ระยะห่างจากการคำนวณด้วยมาตรวัดระยะห่างแบบต่าง ๆ ในขั้นตอนที่ 4

4.6 ขั้นตอนที่ 6

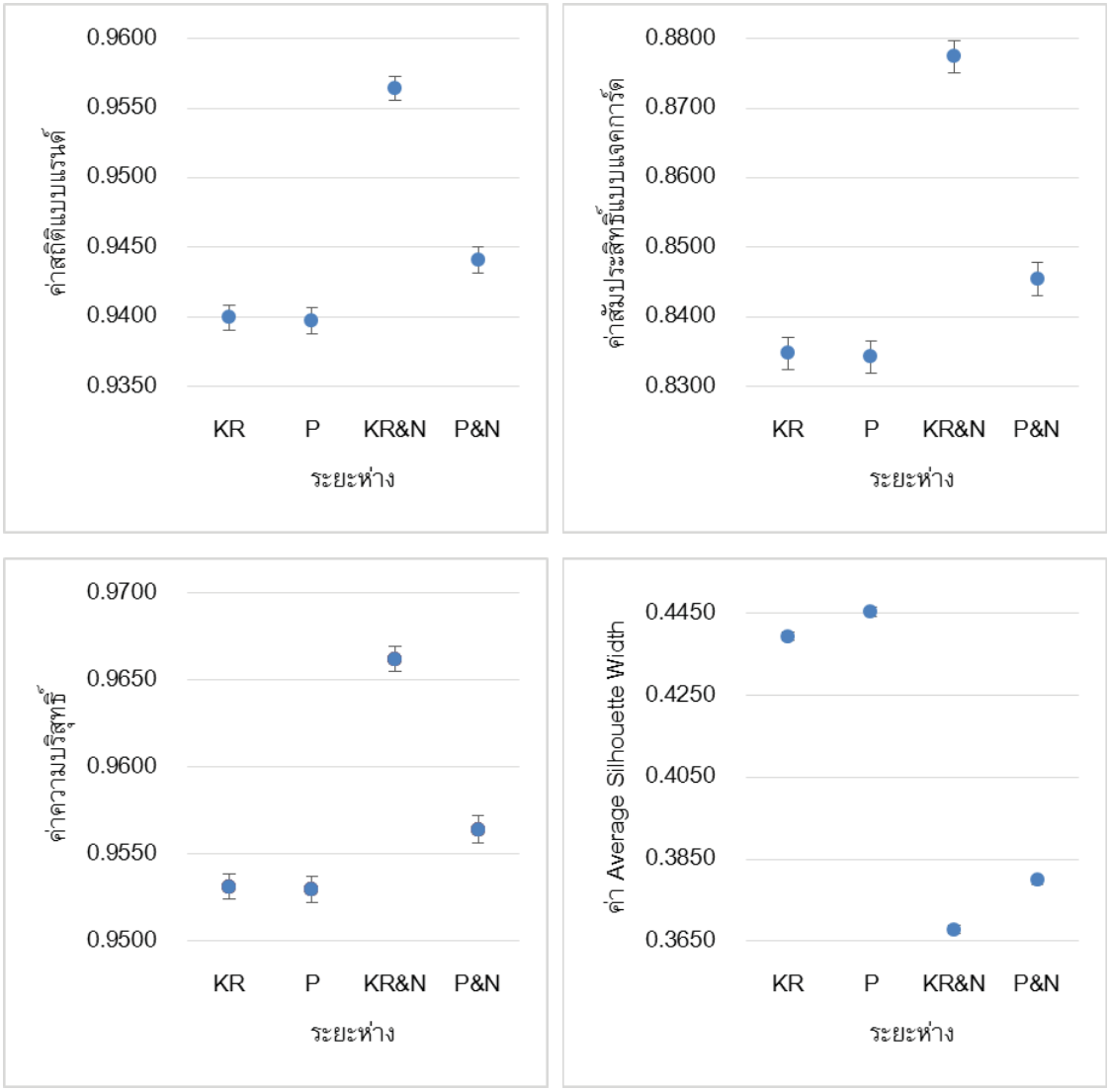
เปรียบเทียบประสิทธิภาพในการวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ ที่ใช้ระยะห่างแบบต่าง ๆ โดยคำนวณค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน ของค่าความบริสุทธิ์ ค่าสถิติแบบแรนด์ และค่าสัมประสิทธิ์แบบแจคการ์ด และยังคำนวณค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน ของค่า Average Silhouette Width เพื่อสังเกตความแตกต่างของมาตรวัดระยะห่างแบบต่าง ๆ เพื่อเป็นกรณีศึกษาหากมีการวิเคราะห์กลุ่มด้วยมาตรวัดระยะห่างแบบต่าง ๆ กับข้อมูลจริง ซึ่งไม่ทราบจำนวนกลุ่มที่แท้จริงของข้อมูล

5. ผลการวิจัย

5.1 กรณีตัวแปรนามบัญญัติ/ตัวแปรอันดับ มีจำนวนความถี่ของแต่ละประเภท/อันดับ แตกต่างกัน (จำนวนกลุ่มเท่ากับ 3 กลุ่ม)

ตารางที่ 3 ค่าเฉลี่ย (ส่วนเบี่ยงเบนมาตรฐาน) จากการวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ ที่ใช้มาตรวัดระยะห่างแบบต่าง ๆ สำหรับข้อมูลแบบผสมที่ตัวแปรมีความสัมพันธ์กันน้อย จำนวนกลุ่มเท่ากับ 3 กลุ่ม

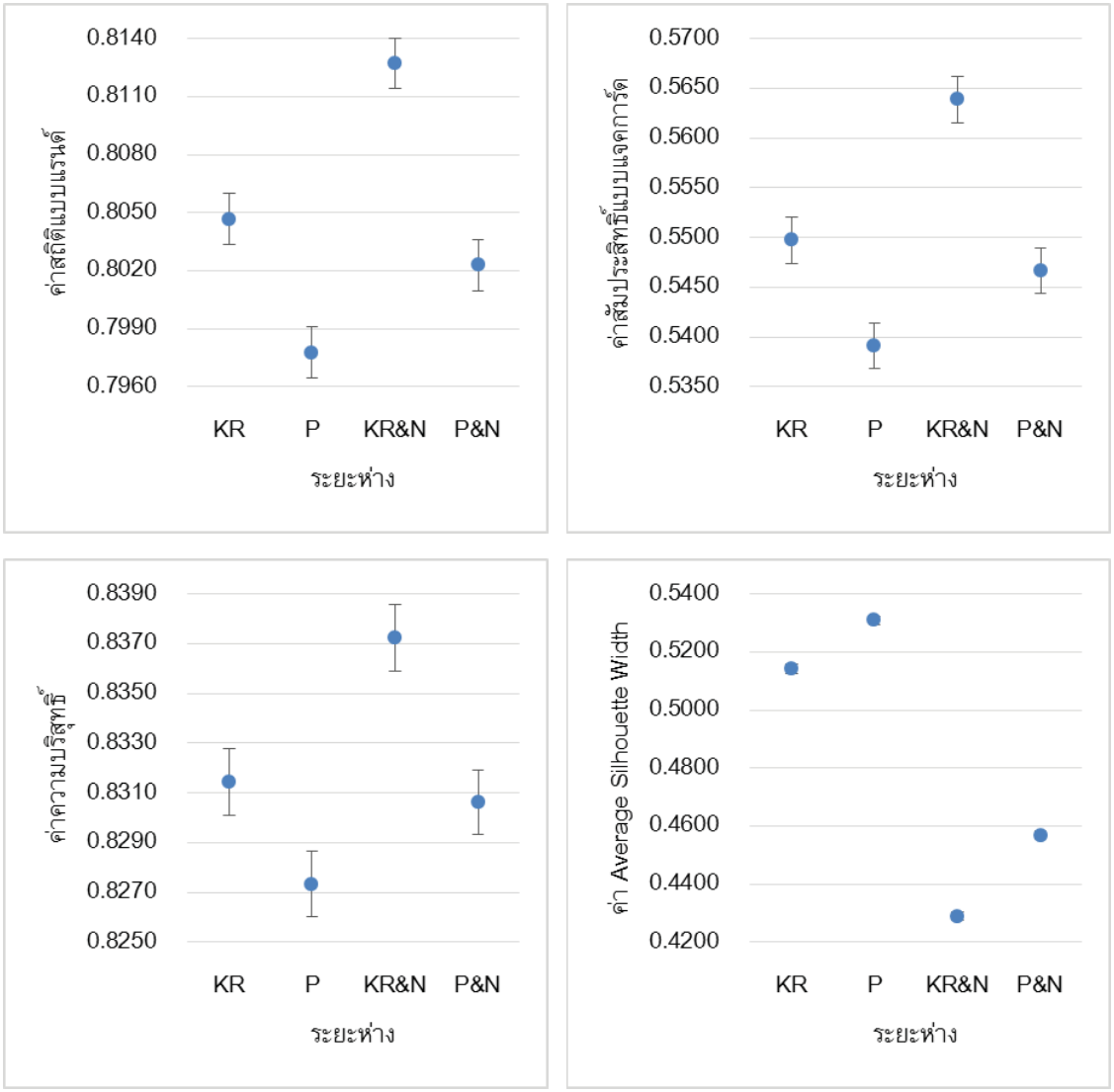
ระยะห่าง	KR	P	KR&N	P&N
ค่าสถิติแบบแรนด์	0.9400 (0.0144)	0.9398 (0.0150)	**0.9564** (0.0142)	0.9441 (0.0152)
ค่าสัมประสิทธิ์แบบแจคการ์ด	0.8348 (0.0365)	0.8343 (0.0377)	**0.8774** (0.0374)	0.8455 (0.0388)
ค่าความบริสุทธิ์	0.9531 (0.0117)	0.9530 (0.0122)	**0.9662** (0.0114)	0.9564 (0.0124)
ค่า Average Silhouette Width	0.4394 (0.0183)	**0.4453** (0.0184)	0.3678 (0.0151)	0.3798 (0.0163)



รูปภาพที่ 1 กราฟช่วงความเชื่อมั่น 95% ของค่าต่าง ๆ จากการวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ ที่ใช้มาตรวัดระยะห่างแบบต่าง ๆ สำหรับข้อมูลแบบผสมที่ตัวแปรมีความสัมพันธ์กันน้อย จำนวนกลุ่มเท่ากับ 3 กลุ่ม

ตารางที่ 4 ค่าเฉลี่ย (ส่วนเบี่ยงเบนมาตรฐาน) จากการวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ ที่ใช้มาตรวัดระยะห่างแบบต่าง ๆ สำหรับข้อมูลแบบผสมที่ตัวแปรมีความสัมพันธ์กันมาก จำนวนกลุ่มเท่ากับ 3 กลุ่ม

ระยะห่าง	KR	P	KR&N	P&N
ค่าสถิติแบบแรนด์	0.8046 (0.0214)	0.7978 (0.0215)	**0.8127** (0.0210)	0.8023 (0.0213)
ค่าสัมประสิทธิ์แบบแจคการ์ด	0.5497 (0.0373)	0.5391 (0.0365)	**0.5639** (0.0376)	0.5467 (0.0367)
ค่าความบริสุทธิ์	0.8314 (0.0214)	0.8273 (0.0212)	**0.8372** (0.0214)	0.8306 (0.0210)
ค่า Average Silhouette Width	0.5140 (0.0256)	**0.5309** (0.0252)	0.4289 (0.0216)	0.4568 (0.0228)

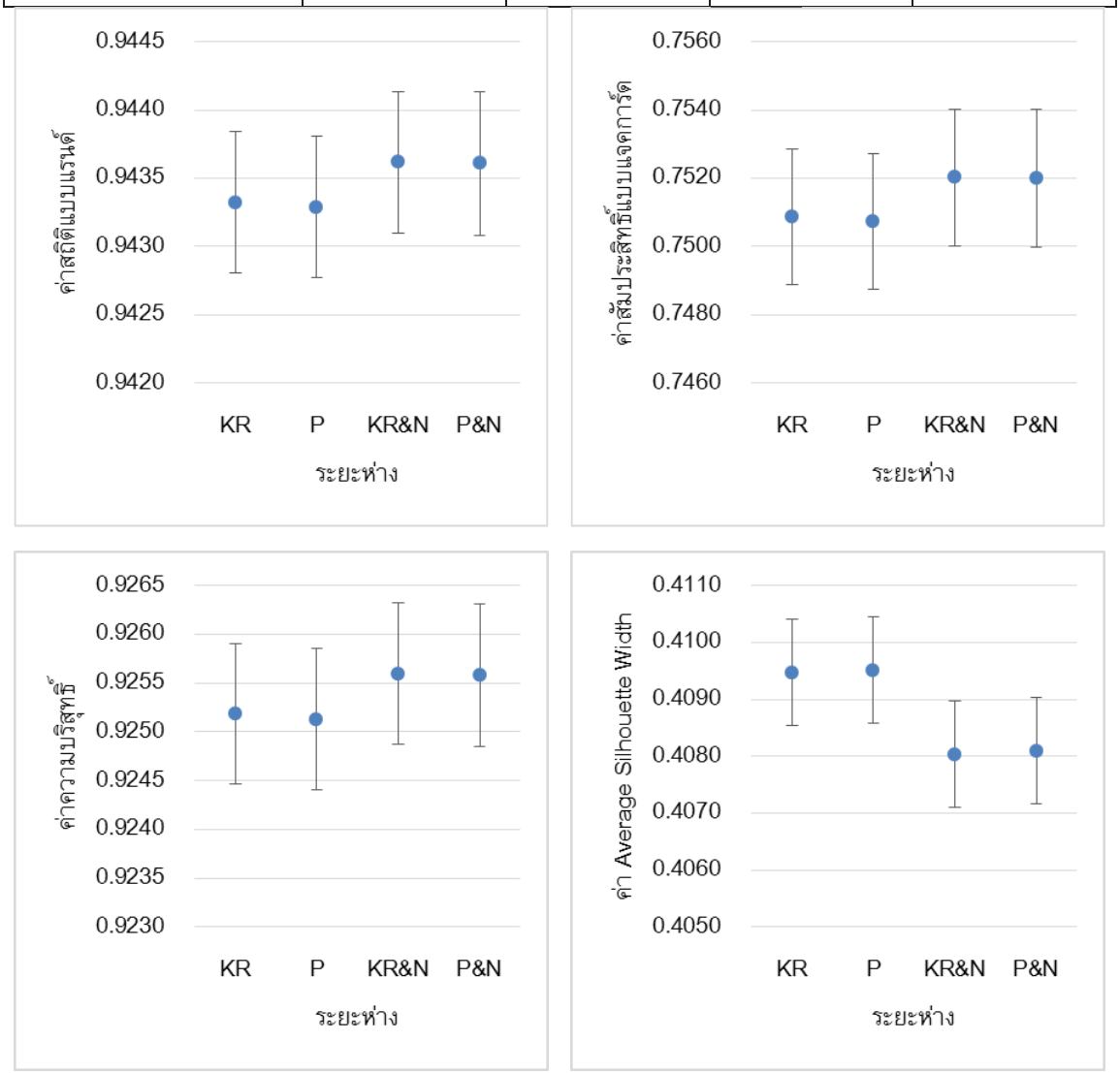


รูปภาพที่ 2 กราฟช่วงความเชื่อมั่น 95% ของค่าต่าง ๆ จากการวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ที่ใช้มาตรวัดระยะห่างแบบต่าง ๆ สำหรับข้อมูลแบบผสมที่ตัวแปรมีความสัมพันธ์กันมาก จำนวนกลุ่มเท่ากับ 3 กลุ่ม

5.2 กรณีตัวแปรนามบัญญัติ/ตัวแปรอันดับ มีจำนวนความถี่ของแต่ละประเภท/อันดับ ไม่แตกต่างกัน (จำนวนกลุ่มเท่ากับ 5 กลุ่ม)

ตารางที่ 5 ค่าเฉลี่ย (ส่วนเบี่ยงเบนมาตรฐาน) จากการวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ ที่ใช้มาตรวัดระยะห่างแบบต่าง ๆ สำหรับข้อมูลแบบผสมที่ตัวแปรมีความสัมพันธ์กันน้อย จำนวนกลุ่มเท่ากับ 5 กลุ่ม

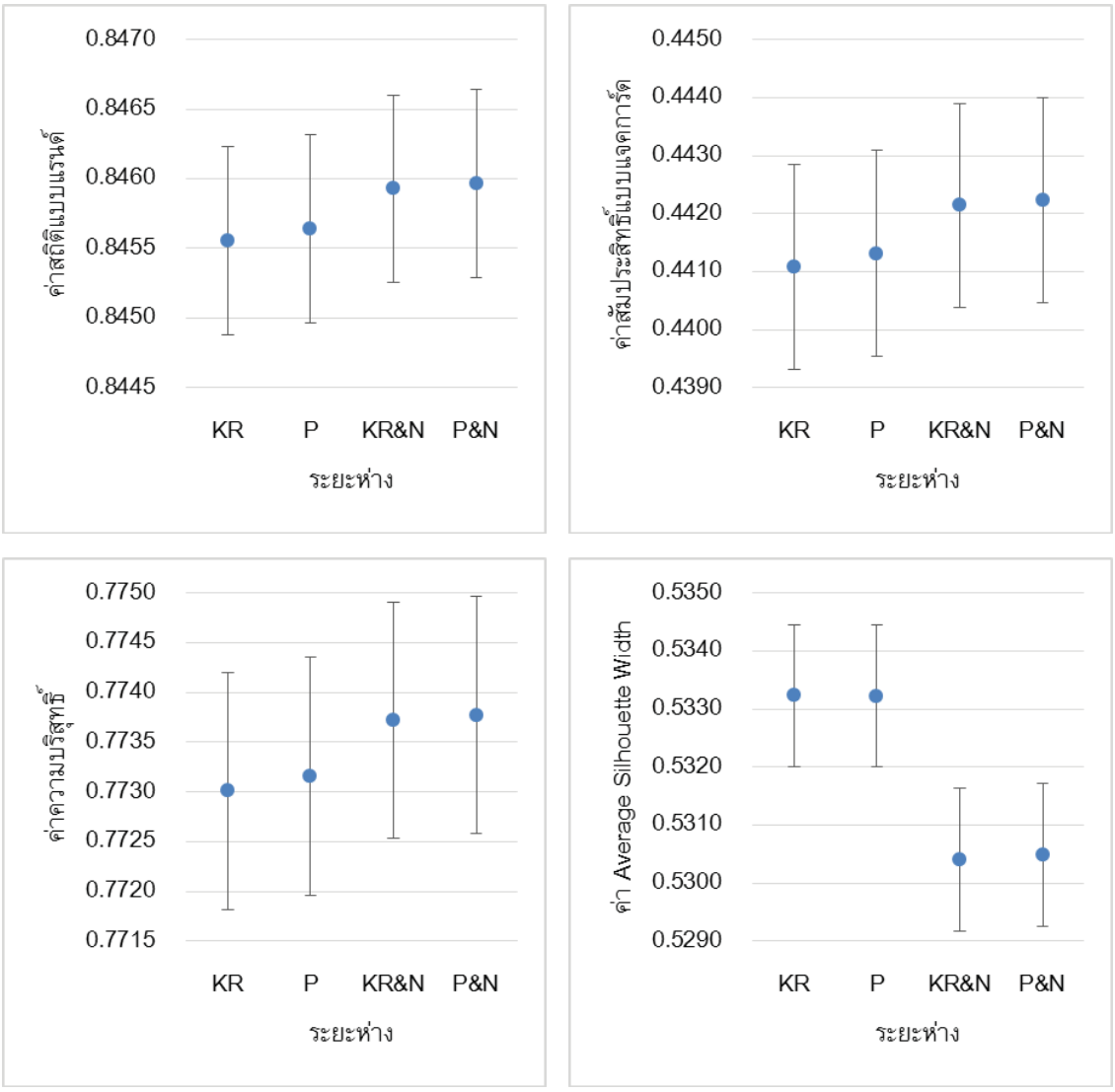
ระยะห่าง	KR	P	KR&N	P&N
ค่าสถิติแบบแรนด์	0.9433 (0.0083)	0.9433 (0.0083)	**0.9436** (0.0084)	**0.9436** (0.0084)
ค่าสัมประสิทธิ์แบบแจคการ์ด	0.7509 (0.0319)	0.7507 (0.0320)	**0.7520** (0.0322)	**0.7520** (0.0324)
ค่าความบริสุทธิ์	0.9252 (0.0116)	0.9251 (0.0116)	**0.9256** (0.0117)	**0.9256** (0.0117)
ค่า Average Silhouette Width	**0.4095** (0.0151)	**0.4095** (0.0151)	0.4080 (0.0151)	0.4081 (0.0151)



รูปภาพที่ 3 กราฟช่วงความเชื่อมั่น 95% ของค่าต่าง ๆ จากการวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ ที่ใช้มาตรวัดระยะห่างแบบต่าง ๆ สำหรับข้อมูลแบบผสมที่ตัวแปรมีความสัมพันธ์กันน้อย จำนวนกลุ่มเท่ากับ 5 กลุ่ม

ตารางที่ 6 ค่าเฉลี่ย (ส่วนเบี่ยงเบนมาตรฐาน) จากการวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ ที่ใช้มาตรวัดระยะห่างแบบต่าง ๆ สำหรับข้อมูลแบบผสมที่ตัวแปรมีความสัมพันธ์กันมาก จำนวนกลุ่มเท่ากับ 5 กลุ่ม

ระยะห่าง	KR	P	KR&N	P&N
ค่าสถิติแบบแรนด์	0.8456 (0.0109)	0.8456 (0.0109)	0.8459 (0.0109)	**0.8460** (0.0109)
ค่าสัมประสิทธิ์แบบแจคการ์ด	0.4411 (0.0284)	0.4413 (0.0286)	0.4421 (0.0284)	**0.4422** (0.0285)
ค่าความบริสุทธิ์	0.7730 (0.0192)	0.7732 (0.0193)	0.7737 (0.0192)	**0.7738** (0.0192)
ค่า Average Silhouette Width	**0.5332** (0.0197)	**0.5332** (0.0197)	0.5304 (0.0198)	0.5305 (0.0198)



รูปภาพที่ 4 กราฟช่วงความเชื่อมั่น 95% ของค่าต่าง ๆ จากการวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ ที่ใช้มาตรวัดระยะห่างแบบต่าง ๆ สำหรับข้อมูลแบบผสมที่ตัวแปรมีความสัมพันธ์กันมาก จำนวนกลุ่มเท่ากับ 5 กลุ่ม

6. สรุปผลการวิจัย

6.1 กรณีตัวแปรนามบัญญัติ/ตัวแปรอันดับ มีจำนวนความถี่ของแต่ละประเภท/อันดับ แตกต่างกัน (จำนวนกลุ่มเท่ากับ 3 กลุ่ม)

ในการวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ ที่ใช้มาตรวัดระยะห่างแบบต่าง ๆ สำหรับข้อมูลแบบผสม ที่ตัวแปรนามบัญญัติ/ตัวแปรอันดับ มีจำนวนความถี่ของแต่ละประเภท/อันดับ แตกต่างกัน นั่นคือ จำนวนกลุ่มที่จำลองเท่ากับ 3 กลุ่ม พบว่าค่าสถิติแบบแรนด์ ค่าสัมประสิทธิ์แบบแจคการ์ด และค่าความบริสุทธิ์ ซึ่งเป็นค่าที่ใช้วัดประสิทธิภาพในการวิเคราะห์กลุ่ม โดยเฉลี่ย เป็นไปในทิศทางเดียวกัน และพบว่าการวิเคราะห์กลุ่มด้วยระยะห่างแบบผสม KR&N มีประสิทธิภาพดีที่สุด ทั้งกรณีที่ตัวแปรมีความสัมพันธ์กันน้อยและตัวแปรมีความสัมพันธ์กันมาก เนื่องจากให้ค่าสถิติแบบแรนด์ ค่าสัมประสิทธิ์แบบแจคการ์ด และค่าความบริสุทธิ์ สูงที่สุดโดยเฉลี่ย

นอกจากนี้ กรณีที่ตัวแปรมีความสัมพันธ์กันน้อย การวิเคราะห์กลุ่มด้วยระยะห่างแบบผสม P&N ระยะห่างของ KR และระยะห่างของ P มีประสิทธิภาพรองลงมาตามลำดับ อย่างไรก็ตามการวิเคราะห์กลุ่มด้วยระยะห่างของ KR และระยะห่างของ P มีประสิทธิภาพไม่แตกต่างกันอย่างมีนัยสำคัญ ดังแสดงในตารางที่ 3 และรูปภาพที่ 1

กรณีที่ตัวแปรมีความสัมพันธ์กันมาก พบว่าการวิเคราะห์กลุ่มด้วยระยะห่างของ KR และระยะห่างแบบผสม P&N มีประสิทธิภาพรองลงมาอย่างใกล้เคียงกัน ขณะที่การวิเคราะห์กลุ่มด้วยระยะห่างของ P มีประสิทธิภาพดีน้อยที่สุด ดังแสดงในตารางที่ 4 และรูปภาพที่ 2

เมื่อพิจารณาค่า Average Silhouette Width ทั้งกรณีที่ตัวแปรมีความสัมพันธ์กันน้อยและมาก พบว่าการวิเคราะห์กลุ่มด้วยระยะห่างของ P ระยะห่างของ KR ระยะห่างแบบผสม P&N และระยะห่างแบบผสม KR&N ให้ค่า Average Silhouette Width สูงที่สุดโดยเฉลี่ย ตามลำดับ ดังแสดงในตารางที่ *3* และ 4 และรูปภาพที่ 1 และ 2

6.2 กรณีตัวแปรนามบัญญัติ/ตัวแปรอันดับ มีจำนวนความถี่ของแต่ละประเภท/อันดับ ไม่แตกต่างกัน (จำนวนกลุ่มเท่ากับ 5 กลุ่ม)

ในการวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ ที่ใช้มาตรวัดระยะห่างแบบต่าง ๆ สำหรับข้อมูลแบบผสม ที่ตัวแปรนามบัญญัติ/ตัวแปรอันดับ มีจำนวนความถี่ของแต่ละประเภท/อันดับ ไม่แตกต่างกัน นั่นคือ จำนวนกลุ่มที่จำลองเท่ากับ 5 กลุ่ม ยังคงพบว่าค่าสถิติแบบแรนด์ ค่าสัมประสิทธิ์แบบแจคการ์ด และค่าความบริสุทธิ์ โดยเฉลี่ยเป็นไปในทิศทางเดียวกัน และพบว่า ทั้งกรณีที่ตัวแปรมีความสัมพันธ์กันน้อยและตัวแปรมีความสัมพันธ์กันมาก การวิเคราะห์กลุ่มด้วยมาตรวัดระยะห่างทั้ง 4 วิธี มีประสิทธิภาพไม่แตกต่างกัน ทั้งนี้ค่าเฉลี่ยของ ค่าสถิติแบบแรนด์ ค่าสัมประสิทธิ์แบบแจคการ์ด และค่าความบริสุทธิ์ ในการวิเคราะห์กลุ่มด้วยระยะห่างแบบผสม KR&N และระยะห่างแบบผสม P&N มากกว่าค่าเฉลี่ยจากการวิเคราะห์กลุ่มด้วยระยะห่างของ KR และระยะห่างของ P เพียงเล็กน้อย ดังแสดงในตารางที่ 5 และ 6 และรูปภาพที่ 3 และ 4

เมื่อพิจารณาค่า Average Silhouette Width ทั้งกรณีที่ตัวแปรมีความสัมพันธ์กันน้อยและมาก พบว่าการวิเคราะห์กลุ่มด้วยระยะห่างของ KR และด้วยระยะห่างของ P ให้ค่า Average Silhouette Width สูงที่สุดโดยเฉลี่ยใกล้เคียงกัน และการวิเคราะห์กลุ่มด้วยระยะห่างแบบผสม KR&N และด้วยระยะห่างแบบผสม P&N ให้ค่า Average Silhouette Width โดยเฉลี่ยใกล้เคียงกัน และน้อยกว่ามาตรวัดระยะห่างสองวิธีแรก

7. อภิปรายผลการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพในการวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ ที่ใช้ระยะห่างของ KR ระยะห่างของ P ระยะห่างแบบผสม KR&N และระยะห่างแบบผสม P&N สำหรับข้อมูลแบบผสม ที่ประกอบไปด้วยตัวแปรเชิงปริมาณ ตัวแปรนามบัญญัติ และตัวแปรอันดับ ชนิดละ 3 ตัวแปร โดยทำการจำลองข้อมูล

ที่กำหนดให้ทราบกลุ่มแน่ชัด และมีการพิจารณากรณีที่ตัวแปรมีความสัมพันธ์กันน้อยและตัวแปรมีความสัมพันธ์กันมาก ทั้งยังพิจารณากรณีที่จำนวนความถี่ของแต่ละประเภท/อันดับแตกต่างกัน และไม่แตกต่างกัน ด้วยการกำหนดจำนวนกลุ่มของข้อมูลที่แตกต่างกัน ผลการศึกษาโดยเปรียบเทียบ ค่าสถิติแบบแรนด์ ค่าสัมประสิทธิ์แบบแจคการ์ด และค่าความบริสุทธิ์ พบว่าทั้งสามค่าให้ผลการวิเคราะห์กลุ่มเป็นไปในทิศทางเดียวกัน

ในการวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ สำหรับข้อมูลแบบผสมที่ตัวแปรนามบัญญัติหรือตัวแปรอันดับ มีจำนวนความถี่ของแต่ละประเภท/อันดับ แตกต่างกัน ไม่ว่าตัวแปรจะมีความสัมพันธ์กันมากหรือน้อย การวิเคราะห์กลุ่มด้วยระยะห่างแบบผสม KR&N มีประสิทธิภาพดีที่สุด อาจเป็นเพราะว่าข้อมูลมีความแตกต่างของจำนวนความถี่ของแต่ละประเภทของตัวแปรนามบัญญัติ การวัดระยะห่างสำหรับตัวแปรนามบัญญัติของ Noorbehbahani et al. จึงทำให้ระยะห่างระหว่างข้อมูลมีความละเอียดมากกว่าการวัดระยะห่างสำหรับตัวแปรนามบัญญัติของ Gower สามารถแบ่งแยกความแตกต่างระหว่างข้อมูลตัวแปรนามบัญญัตินั้น ๆ ที่อยู่ต่างประเภทกันได้ โดยเฉพาะอย่างยิ่ง เมื่อจำนวนประเภทของตัวแปรนามบัญญัติมากกว่าจำนวนกลุ่มข้อมูล ทำให้ในหนึ่งกลุ่มข้อมูลอาจมีข้อมูลตัวแปรนั้นมากกว่าหนึ่งประเภท ขณะที่พบว่าการวิเคราะห์กลุ่มด้วยระยะห่างแบบผสม P&N มีประสิทธิภาพด้อยกว่า แม้ว่าจะวัดระยะห่างสำหรับตัวแปรนามบัญญัติด้วยวิธีของ Noorbehbahani et al. เช่นเดียวกัน อาจเป็นเพราะความแตกต่างของจำนวนความถี่ของอันดับของตัวแปรอันดับ ทำให้เมื่อวัดระยะห่างสำหรับตัวแปรอันดับตามนิยามของ Podani เกิดระยะห่างระหว่างข้อมูลอันดับที่มีความถี่สูงกับข้อมูลอันดับที่มีความถี่ต่ำมากเกินไป ทำให้การวัดระยะห่างสำหรับตัวแปรอันดับตามนิยามของ Podani มีความผิดพลาดมากกว่าการวัดระยะห่างสำหรับตัวแปรอันดับของ Kaufman and Rousseeuw

ในการวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ สำหรับข้อมูลแบบผสมที่ตัวแปรนามบัญญัติหรือตัวแปรอันดับ มีจำนวนความถี่ของแต่ละประเภท/อันดับ ไม่แตกต่างกัน ไม่ว่าตัวแปรจะมีความสัมพันธ์กันมากหรือน้อย พบว่า การวิเคราะห์กลุ่มด้วยระยะห่างของ KR ระยะห่างของ P ระยะห่างแบบผสม KR&N และระยะห่างแบบผสม P&N มีประสิทธิภาพดีที่สุดใกล้เคียงกัน อย่างไรก็ตามการวิเคราะห์กลุ่มด้วยระยะห่างแบบผสม KR&N และระยะห่างแบบผสม P&N มีประสิทธิภาพดีกว่าอีกสองวิธีเพียงเล็กน้อย อาจเป็นเพราะว่าจำนวนความถี่ของแต่ละประเภท/อันดับไม่แตกต่างกันมากนัก ทำให้การวัดระยะห่างสำหรับตัวแปรนามบัญญัติของ Noorbehbahani et al. ดีกว่าของ Gower เพียงเล็กน้อย และการวัดระยะห่างสำหรับตัวแปรอันดับของ Podani และของ Kaufman and Rousseeuw ให้ระยะห่างที่ใกล้เคียงกัน นอกจากนี้จำนวนประเภท/อันดับของข้อมูลแบบผสมชุดนี้ ยังมีค่าเท่ากับจำนวนกลุ่มข้อมูลอีกด้วย

เมื่อพิจารณาค่า Average Silhouette Width พบว่าการวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ที่ใช้ระยะห่างของ P ระยะห่างของ KR ระยะห่างแบบผสม P&N และระยะห่างแบบผสม KR&N ให้ค่า Average Silhouette Width สูงที่สุดโดยเฉลี่ยตามลำดับ และมีค่าอยู่ระหว่าง 0.36 ถึง 0.54 แสดงว่าหากวิเคราะห์กลุ่มด้วยระยะห่างข้างต้นกับข้อมูลจริง ซึ่งเป็นข้อมูลแบบผสมและประกอบไปด้วยตัวแปรเชิงปริมาณ ตัวแปรนามบัญญัติ และตัวแปรอันดับ ชนิดละ 3 ตัวแปร โดยตัวแปรนามบัญญัติและตัวแปรอันดับมีจำนวนประเภทและอันดับเท่ากับ 5 อาจเป็นไปได้ว่าค่า Average Silhouette Width จะไม่สูงมากนัก แม้ว่าการวิเคราะห์กลุ่มนั้นจะถูกกำหนดจำนวนกลุ่มที่เหมาะสมแล้วก็ตาม

ข้อจำกัดของการศึกษาวิจัยนี้ คือ ไม่สามารถจำลองข้อมูลแบบผสม ซึ่งประกอบไปด้วยตัวแปรเชิงปริมาณ ตัวแปรนามบัญญัติ และตัวแปรอันดับ ได้ครอบคลุมทุกกรณีที่เป็นไปได้ เนื่องจากในความเป็นจริงข้อมูลมีความหลากหลายสูง ทั้งด้านจำนวนตัวแปรแต่ละชนิดที่แตกต่างกัน จำนวนประเภทหรืออันดับที่มีค่าที่เป็นไปได้มากมาย จำนวนกลุ่มของข้อมูล และความสัมพันธ์ระหว่างตัวแปรที่เป็นไปได้หลายค่า การวัดประสิทธิภาพในการวิเคราะห์กลุ่มด้วยอัลกอริทึมจัดกลุ่มโดยรอบมีดอยด์ ที่ใช้มาตรวัดระยะห่างแบบต่าง ๆ ในการศึกษาวิจัยนี้ จึงสามารถสรุปได้แน่ชัดในบางกรณีเท่านั้น และอาจเป็นแนวทางการศึกษาข้อมูลแบบผสมกรณีอื่น ๆ ต่อไป อย่างไรก็ตามระยะห่างแบบผสม KR&N และระยะห่างแบบผสม P&N ที่นำเสนอขึ้นใหม่ ให้ผลการวิเคราะห์กลุ่มที่ดีในกรณีที่จำนวนความถี่ของแต่ละประเภทของข้อมูลตัวแปรนามบัญญัติแตกต่างกัน ซึ่งเป็นลักษณะทั่วไปของข้อมูลจริง ที่ไม่สามารถกำหนดหรือควบคุม

ได้ว่าข้อมูลตัวแปรนามบัญญัตินั้น ๆ จะมีความถี่ของแต่ละประเภทเท่าใด ดังนั้นระยะห่างสำหรับข้อมูลแบบผสมที่นำเสนอขึ้นใหม่ทั้งสองนี้ จึงเป็นอีกทางเลือกที่สามารถนำไปใช้ในการวิเคราะห์กลุ่มข้อมูลจริงได้อย่างมีประสิทธิภาพ

8. เอกสารอ้างอิง

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). **Cluster Analysis.** 5th ed. London: A John Wiley and Sons, Ltd., Publication.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. **Biometrics**, 27 (4), 857-871.

Kaufman, L. & Rousseeuw, P. J. (1990). **Finding groups in data: An introduction to cluster analysis**. USA: A Wiley-Interscience Publication.

Madhulatha, T. S. (2011). Comparison between K-Means and K-Medoids Clustering Algorithms. **Communications in Computer and Information Science**, 198, 472-481.

Noorbehbahani, F. Mousavi, S. R., & Mirzaei, A. (2014). An incremental mixed data clustering method using a new distance measure. **Springer-Verlag Berlin Heidelberg**,

Podani, J. (1999). Extending Gower's general coefficient of similarity to ordinal characters, **International Association for Plant Taxonomy**, 48 (2), 331-340.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, (20), 53-65.