การจำแนกกลุ่มข้อมูลโดยอัลกอริทึม ANOVAID

นางสาวนวทิพย์ ไมตรี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2558

DATA CLASSIFICATION BY ANOVAID ALGORITHM

Miss Nawatip Maitree

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science Program in Statistics

Department of Statistics

Faculty of Commerce and Accountancy

Chulalongkorn University

Academic Year 2015

หัวข้อวิทยานิพนธ์	การจำแนกกลุ่มข้อมูลโดยอัลกอริทึม ANOVAID
โดย	นางสาวนวทิพย์ ไมตรี
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร. พสุ เดชะรินทร์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(อาจารย์ ดร. อัครินทร์ ไพบูลย์พานิช)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา)

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร. วิฐรา พึ่งพาพงศ์)

.....กรรมการภายนอกมหาวิทยาลัย

(อาจารย์ ดร. อรุณี กำลัง)

นวทิพย์ ไมตรี : การจำแนกกลุ่มข้อมูลโดยอัลกอริทึม ANOVAID (DATA CLASSIFICATION BY ANOVAID ALGORITHM) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: รศ. ดร. สุพล ดุรงค์วัฒนา, 48 หน้า.

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อศึกษากระบวนการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม ANOVAID ซึ่งเป็นส่วนผสมของการใช้การวิเคราะห์ความแปรปรวนทางเดียวและสถิติทดสอบ t สำหรับกลุ่มตัวอย่าง 2 กลุ่มที่เป็นอิสระกัน โดยตัวแปรตามเป็นตัวแปรเชิงปริมาณและตัวแปรอิสระเป็นตัวแปรเชิงคุณภาพ อัลกอริทึมนี้มีขั้นตอนในการทำงาน 2 ขั้นตอน คือ ขั้นตอนในการคัดเลือกตัวแปรอิสระและขั้นตอนในการรวมกลุ่มของตัวแปรอิสระนั้น โดยในการคัดเลือกตัวแปรอิสระนั้น จะพิจารณาจากค่า p-value น้อยสุด จากการวิเคราะห์ความแปรปรวนทางเดียว เมื่อเปรียบเทียบกันระหว่างตัวแปรอิสระทั้งหมด โดยที่ค่า p-value ต้องมีนัยสำคัญด้วยจึงจะเลือกตัวแปรอิสระนั้นเข้ามาในกระบวนการ จากนั้นจะใช้สถิติทดสอบ t สำหรับกลุ่มตัวอย่าง 2 กลุ่มที่เป็นอิสระกันในการรวมกลุ่มของตัวแปรอิสระที่ถูกเลือกเข้ามา โดยพิจารณาจากค่า p-value ที่ไม่มีนัยสำคัญถ้าไม่ตรงตามเงื่อนไขข้างต้นอัลกอริทึมจะหยุดทำงาน และสำหรับแต่ละกลุ่มที่จำแนกมาได้ ตัวแปรอิสระที่เหลือจะถูกจำแนกแยกกันและเป็นอิสระกัน จนกระทั่งไม่มีตัวแปรอิสระเหลือหรืออัลกอริทึมหยุดการทำงาน โดยข้อมูลที่ใช้ในการศึกษาจะจำลองภายใต้จำนวนกลุ่มของปัจจัยเท่ากับ 2, 3 และ 4, ขนาดข้อมูลเท่ากับ 6,000, 12,000 และ 24,000, ความแปรปรวนเท่ากับ 10,000 และ 40,000 และอัตราส่วนของค่าเฉลี่ยเท่ากับ 0.5, 1 และ 2 โดยทำการทดสอบที่ระดับนัยสำคัญเท่ากับ 0.05 และใช้เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มเป็นเกณฑ์ในการพิจารณาว่าอัลกอริทึมมีประสิทธิภาพในการจำแนกกลุ่มได้ดีหรือไม่

จากผลการศึกษาพบว่าเมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มเพิ่มขึ้น, เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มลดลง, เมื่ออัตราส่วนของค่าเฉลี่ยเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มลดลง และเมื่อจำนวนกลุ่มของปัจจัยเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มไม่แตกต่างกัน

ภาควิชา สถิติ ลายมือชื่อนิสิต .....

สาขาวิชา สถิติ ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ปีการศึกษา 2558

5681546626 : MAJOR STATISTICS
KEYWORDS: DATA CLASSIFICATION / ANOVAID ALGORITHM / ONE-WAY ANOVA / INDEPENDENT-SAMPLE T-TEST

NAWATIP MAITREE: DATA CLASSIFICATION BY ANOVAID ALGORITHM. ADVISOR: ASSOC. PROF. SUPOL DURONGWATANA, Ph.D., 48 pp.

The aim of this paper is to study the classification process of ANOVAID algorithm which is the mixture of one-way ANOVA and independent-sample t-test. The dependent variable is the quantitative variable and the independent variable is the fixed qualitative variable. There are 2 steps in this algorithm. Those are independent variable selection and merging steps. Each independent variable is selected using the least p-value of the one-way ANOVA when the least p-value of the selected independent variable shows the statistical significance to enter or to be selected, then the independent-sample t-test is used to merge the data by using the insignificance p-value otherwise the algorithm will be stopped. In each of merging group, the next hierarchy for the rest of independent variables will be classified separately and independently and so on until there is no independent variable to classify or the algorithm is stopped. The data are simulated under several situations. Each situation depends upon the numbers of levels in factor are 2, 3 and 4, the sample size of each set of data are 6,000, 12,000 and 24,000, the variance of random error in the one-way ANOVA model are 10,000 and 40,000, and lastly the ratio of means are 0.5, 1 and 2 at the hypothesis testing is 0.05. In the study, the percentage of misclassification is used as the measure how good the algorithm.

The results of the study show that when the value of variance for random error increases, the percentage of misclassification also increase; when the number of sample size increases, then the percentage of misclassification decreases; when the ratio of mean increases, then the percentage of misclassification decreases; and when the numbers of levels in factor increases, then the percentage of misclassification is indifferent.

Department: Statistics Student's Signature .....
Field of Study: Statistics Advisor's Signature .....
Academic Year: 2015

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์ลงได้ด้วยดี ด้วยความช่วยเหลือและเอาใจใส่จากรองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้วิจัยขอกราบขอบพระคุณท่านอาจารย์เป็นอย่างสูงที่กรุณาให้คำปรึกษา อบรมสั่งสอนและให้ข้อคิดเห็นต่างๆ ตลอดจนความช่วยเหลือคำแนะนำเพื่อปรับปรุงแก้ไขวิทยานิพนธ์และเป็นกำลังใจในการทำงาน จนกระทั่งวิทยานิพนธ์เสร็จสมบูรณ์ด้วยดี

ผู้วิจัยขอกราบขอบพระคุณ อาจารย์ ดร. อัครินทร์ ไพบูลย์พานิช ประธานกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร. วิฐรา พึ่งพาพงศ์ กรรมการสอบวิทยานิพนธ์ และอาจารย์ ดร. อรุณี กำลัง กรรมการภายนอกสอบวิทยานิพนธ์เป็นอย่างสูงที่ท่านอาจารย์ทั้งสามท่านได้เสียสละเวลาเพื่อสอบ ตรวจสอบและให้คำแนะนำเพื่อแก้ไขวิทยานิพนธ์ฉบับนี้ให้สมบูรณ์ยิ่งขึ้น อีกทั้งขอกราบขอบพระคุณคณาจารย์ประจำภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัยทุกท่านที่ให้โอกาสทางการศึกษาและอบรมสั่งสอนความรู้ ทั้งในการเรียนและการดำรงชีวิตให้แก่ผู้วิจัยเสมอมาจนกระทั่งสำเร็จการศึกษาในครั้งนี้

สุดท้ายนี้ผู้วิจัยขอกราบขอบพระคุณครอบครัว ที่ให้กำลังใจและความห่วงใย ส่งเสริมและสนับสนุนมาโดยตลอด และขอขอบคุณเพื่อนๆ ทุกคนที่คอยช่วยเหลือ ให้คำแนะนำและเป็นกำลังใจให้กับผู้วิจัยตลอดมา

สารบัญ

หน้า

หน้า

สารบัญตาราง

หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันการดำเนินธุรกิจที่เกี่ยวข้องกับตลาดผู้บริโภคนั้น ผู้ผลิตไม่อาจเข้าถึงหรือตอบสนองความต้องการของผู้บริโภคทุกคนในตลาดได้อย่างทั่วถึง เนื่องจากผู้บริโภคมีจำนวนมากและอยู่กันอย่างกระจัดกระจาย ความต้องการในผลิตภัณฑ์ ความชอบของผู้บริโภคแต่ละคนแตกต่างกันทั้งรูปแบบของสินค้า คุณสมบัติ ขนาด สี ลวดลาย ตรา ตลอดจนราคา วิธีการซื้อและตัดสินใจซื้อ ความต้องการดังกล่าวนั้นเปลี่ยนแปลงได้ตลอดเวลา ผู้ผลิตในปัจจุบันจึงมุ่งเข้าถึงหรือตอบสนองผู้บริโภคเพียงบางส่วนหรือแบ่งตลาดในภาพรวม (Total Market) ออกเป็นตลาดส่วนย่อย (Sub-market) จำนวนหลายๆ ส่วนเสียก่อน แล้วจึงเลือกส่วนแบ่งตลาดที่น่าสนใจ จากนั้นจึงพัฒนาผลิตภัณฑ์และวางแผนทางการตลาดให้กับส่วนแบ่งตลาดนั้น ดังนั้นงานวิจัยนี้จึงมุ่งเน้นไปที่การจำแนกกลุ่มข้อมูล (Data Classification) โดยใช้เกณฑ์ทางสถิติ เพื่อจำแนกกลุ่มผู้บริโภคที่มีอยู่ทั้งหมดออกเป็นกลุ่มๆ และสามารถนำไปสู่การวิเคราะห์และตัดสินใจวางแผนทางการตลาดที่มีประสิทธิภาพที่จะทำให้กลุ่มผู้บริโภคหันมาบริโภคสินค้าและบริการเพิ่มขึ้น ซึ่งได้มีการนำเสนอวิธีการหรือเทคนิคที่นำมาใช้ในการจำแนกกลุ่มข้อมูล คือ อัลกอริทึม ANOVAID

อัลกอริทึม ANOVAID (ANOVA Automatic Interaction Detection) เป็นอัลกอริทึมที่ใช้ในการจำแนกกลุ่มข้อมูลและสามารถสร้างแผนภาพต้นไม้การจำแนกแบบพหุภาค (Multi-branch Tree) ได้ โดยได้แนวคิดมาจากอัลกอริทึม CHAID (Chi-Square Automatic Interaction Detection) (Kass, 1980) ซึ่งเป็นอัลกอริทึมที่ใช้ในการจำแนกกลุ่มข้อมูลเช่นเดียวกัน แต่ตัวแปรตามและตัวแปรอิสระต้องเป็นเชิงคุณภาพจึงจะสามารถใช้งานอัลกอริทึมได้อย่างมีประสิทธิภาพ ดังนั้นงานวิจัยนี้จึงให้ความสนใจในกรณีที่ตัวแปรตามเป็นตัวแปรเชิงปริมาณและตัวแปรอิสระเป็นตัวแปรเชิงคุณภาพ โดยได้มีการนำการวิเคราะห์ความแปรปรวนทางเดียว (One-Way Analysis of Variance: One-Way ANOVA) และสถิติทดสอบ t สำหรับกลุ่มตัวอย่าง 2 กลุ่มที่เป็นอิสระกัน (Independent-Sample t-test) มาใช้ควบคู่กันในกระบวนการจำแนกกลุ่มข้อมูลของอัลกอริทึม ANOVAID เนื่องจากวิธีการทางสถิติทั้งสองมีความเป็นพื้นฐานและง่ายต่อการนำมาใช้ รวมทั้งยังมีการใช้งานอย่างแพร่หลายในด้านต่าง ๆ

จากการศึกษาพบว่า มีนักวิจัยหลายท่านได้นำการวิเคราะห์ความแปรปรวนทางเดียวหรือสถิติทดสอบ t มาใช้ในการแบ่งส่วนตลาด ดังนี้

(Lai & Wu, 2011) ได้ใช้ t-test สำหรับกลุ่มตัวอย่างที่อิสระกันและ One-Way ANOVA ในการวิเคราะห์ความแตกต่างระหว่างคุณลักษณะของผู้โดยสารและความต้องการในบริการของ KRTS ซึ่งเป็นระบบขนส่งมวลชนความเร็วสูงในไต้หวัน ผลที่ได้จากการใช้ t-test พบว่ามีความแตกต่างกันระหว่างเพศชายและเพศหญิง ส่วน One-Way ANOVA พบว่ามีความแตกต่างกันระหว่าง 6 กลุ่มอายุ 8 กลุ่มอาชีพ 5 กลุ่มการศึกษา 5 กลุ่มรายได้ 3 กลุ่มที่อยู่อาศัย 7 กลุ่มจุดประสงค์ในการใช้บริการ 4 กลุ่มความถี่ในการใช้บริการ 11 กลุ่มตั๋วที่ใช้และ 5 กลุ่มเครื่องมือการขนส่งแบบปกติ โดยตัวแปรที่มีนัยสำคัญมากที่สุด คือ เพศและจุดประสงค์ในการใช้บริการ จากนั้นจึงนำสองตัวแปรนี้มาใช้ในการแบ่งส่วนตลาดต่อไป

(Thachn & Olsen, 2015) ได้ใช้ One-Way ANOVA ในการระบุความแตกต่างระหว่างการใช้จ่ายในการซื้อไวน์ของผู้บริโภคไวน์ในสหรัฐอเมริกา 3 กลุ่ม โดยแบ่งออกเป็นใช้จ่ายสูง ใช้จ่ายปานกลางและใช้จ่ายต่ำ พิจารณาจากราคาที่ใช้จ่ายต่อไวน์ 1 ขวด สำหรับการบริโภคในครัวเรือน และมีตัวแปรที่ใช้ในการวิเคราะห์ คือ เพศ อายุ รายได้ ช่องทางในการซื้อและตัวแปรสำคัญอื่นๆ โดยทำการพิจารณาแยกเป็นแต่ละตัวแปรไป

จากงานวิจัยดังกล่าว พบว่าได้มีการนำ One-Way ANOVA หรือ t-test มาใช้ในการวิเคราะห์ความแตกต่างของตัวแปรที่ศึกษาเป็นตัวๆ ไปหรือใช้ในการตรวจสอบความถูกต้องในการจำแนกกลุ่มข้อมูลจากอัลกอริทึมอื่นเท่านั้น งานวิจัยนี้จึงได้ทำการผูกรวมวิธีการทางสถิติทั้งสองเข้าด้วยกันและนำมาใช้ในการจำแนกกลุ่มข้อมูลโดยตรง โดยสามารถใช้กับตัวแปรอิสระหลายตัวได้ แต่ในขั้นตอนการจำแนกกลุ่มนั้นจะพิจารณาจากตัวแปรอิสระทีละตัวแยกกัน โดยให้ชื่อใหม่ว่า อัลกอริทึม ANOVAID และทำการวัดประสิทธิภาพของอัลกอริทึมโดยพิจารณาจากเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระภายใต้กรณีที่ทำการศึกษา โดยทำซ้ำในแต่ละกรณี 1,000 รอบ

1.2 วัตถุประสงค์การวิจัย

เพื่อศึกษากระบวนการจำแนกกลุ่มข้อมูลและเงื่อนไขต่างๆ ของอัลกอริทึม ANOVAID อย่างละเอียด และทำการวัดประสิทธิภาพของอัลกอริทึมเพื่อนำมาประยุกต์ใช้ในการจำแนกกลุ่มข้อมูลในด้านต่างๆ ต่อไป

1.3 ขอบเขตของการวิจัย

ในงานวิจัยครั้งนี้ได้ทำการจำลองข้อมูลตามตัวแบบ One-Way ANOVA โดยข้อมูลจำลองที่ได้จะประกอบไปด้วยตัวแปรตามและตัวแปรอิสระอย่างละตัว ซึ่งในความเป็นจริงแล้วอัลกอริทึม ANOVAID สามารถใช้กับตัวแปรอิสระมากกว่า 1 ตัวได้ แต่ในขั้นตอนการจำแนกกลุ่มนั้นจะพิจารณาจากตัวแปรอิสระทีละตัวแยกกัน เรียกว่า ปัจจัย (Factor) ดังนั้นในงานวิจัยนี้จึงได้ให้ความสำคัญที่ประสิทธิภาพในการจำแนกกลุ่มของตัวแปรอิสระแต่ละตัว โดยกำหนดสถานการณ์ต่างๆ ที่จะทำการศึกษา ดังนี้

1. ตัวแบบที่ใช้ในการศึกษา คือ ตัวแบบ One-Way ANOVA ซึ่งเขียนในรูปแบบถดถอยได้ดังนี้

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$
$$\mu_i = \mu + \tau_i$$

เมื่อ $i = 1,2,\dots,k$ และ $j = 1,2,\dots,n$

โดย Y_{ij} = ตัวแปรตามสำหรับปัจจัยกลุ่มที่ i และค่าสังเกตที่ j

μ = ค่าเฉลี่ยรวมทั้งหมด

τ_i = อิทธิพลจากปัจจัยกลุ่มที่ i

ε_{ij} = ค่าความคลาดเคลื่อนสุ่มสำหรับปัจจัยกลุ่มที่ i และค่าสังเกตที่ j

และเป็นตัวแบบ One-Way ANOVA ที่มีจำนวนค่าสังเกตเท่ากันในแต่ละกลุ่ม (Balanced Design)

2. การแจกแจงของค่าความคลาดเคลื่อนสุ่ม (ε_{ij}) ที่ทำการศึกษามีการแจกแจงปกติ กำหนดให้ ค่าเฉลี่ย $\mu_\varepsilon = 0$ และมีความแปรปรวน $\sigma_\varepsilon^2 =$ 10,000 และ 40,000 โดยค่าความแปรปรวนที่กำหนดพิจารณาจากค่าสัมประสิทธิ์ความแปรผัน (Coefficient of Variance, C.V.) เมื่อต้องการให้มีค่าเป็น 100% และ 200% โดยที่ค่าเฉลี่ยเท่ากับ 100

3. ศึกษาภายใต้จำนวนกลุ่มกับขนาดข้อมูล (k กับ n) $\mu_1 = 100$ ดังนี้

กรณีที่ 1: k = 2 กับ n = 6,000, 12,000 และ 24,000

กรณีที่ 1.1: $\mu_1 = \mu_2$

กรณีที่ 1.2.1: $\mu_2 = 0.5\mu_1$

กรณีที่ 1.2.2: $\mu_2 = 2\mu_1$

กรณีที่ 2: k = 3 กับ n = 6,000, 12,000 และ 24,000

กรณีที่ 2.1: $\mu_1 = \mu_2 = \mu_3$

กรณีที่ 2.2.1: $\mu_3 = \mu_2 = 0.5\mu_1$

กรณีที่ 2.2.2: $\mu_3 = \mu_2 = 2\mu_1$

กรณีที่ 2.3.1: $\mu_3 = 0.5\mu_2,\ \mu_2 = 0.5\mu_1$

กรณีที่ 2.3.2: $\mu_3 = 2\mu_2,\ \mu_2 = 2\mu_1$

กรณีที่ 3: k = 4 กับ n = 6,000, 12,000 และ 24,000

กรณีที่ 3.1: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

กรณีที่ 3.2.1: $\mu_4 = \mu_3 = \mu_2 = 0.5\mu_1$

กรณีที่ 3.2.2: $\mu_4 = \mu_3 = \mu_2 = 2\mu_1$

กรณีที่ 3.3.1: $\mu_1 = \mu_2,\ \mu_4 = \mu_3 = 0.5\mu_1$

กรณีที่ 3.3.2: $\mu_1 = \mu_2,\ \mu_4 = \mu_3 = 2\mu_1$

กรณีที่ 3.4.1: $\mu_4 = \mu_3 = 0.5\mu_2,\ \mu_2 = 0.5\mu_1$

กรณีที่ 3.4.2: $\mu_4 = \mu_3 = 2\mu_2,\ \mu_2 = 2\mu_1$

กรณีที่ 3.5.1: $\mu_4 = 0.5\mu_3,\ \mu_3 = 0.5\mu_2,\ \mu_2 = 0.5\mu_1$

กรณีที่ 3.5.2: $\mu_4 = 2\mu_3,\ \mu_3 = 2\mu_2,\ \mu_2 = 2\mu_1$

4. การวิจัยครั้งนี้ได้ทำการจำลองข้อมูลให้มีสถานการณ์ดังกล่าวข้างต้น ซึ่งผู้วิจัยจะประเมินผลด้วยโปรแกรม R เวอร์ชั่น 3.2.3 โดยทำการจำลองในแต่ละสถานการณ์ 1,000 รอบ

1.4 การวัดประสิทธิภาพการจำแนกกลุ่ม

วัดประสิทธิภาพของอัลกอริทึมโดยพิจารณาจากเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระภายใต้กรณีที่ทำการศึกษา โดยทำซ้ำในแต่ละกรณี 1,000 รอบ ยกตัวอย่างในกรณีที่ $\mu_1 = \mu_2$ หมายความว่าเมื่อทำการทดสอบโดยอัลกอริทึมแล้วค่าเฉลี่ยของทั้งสองกลุ่มต้องไม่แตกต่างกัน นั่นคือทั้งสองกลุ่มรวมกันเป็นกลุ่มเดียวได้ ฉะนั้นในการทำซ้ำ 1,000 รอบ ผลที่ออกมาคือทั้งสองกลุ่มต้องรวมกันได้ทั้ง 1,000 รอบ นั่นคือใน 1,000 รอบนั้นถ้ามีรอบไหนที่ผลออกมาแล้วทั้งสองกลุ่มไม่รวมกัน จึงจะนับรอบนั้นเป็นความผิดพลาดในการจำแนกกลุ่ม ดังนั้นเปอร์เซ็นต์ความผิดพลาดสามารถคำนวณได้ดังนี้

$$\text{เปอร์เซ็นต์ความผิดพลาด} = \frac{\text{จำนวนความผิดพลาดในการจำแนกกลุ่ม}}{1{,}000} \times 100$$

1.5 คำจำกัดความของงานวิจัย

1. การจำแนกกลุ่มข้อมูล (Data Classification) คือกระบวนการสร้างตัวแบบจัดการกับข้อมูลเพื่อทำนายกลุ่มของข้อมูลใหม่ โดยกลุ่มที่ได้จากการจำแนกกลุ่มข้อมูลที่อยู่กลุ่มเดียวกันจะมีลักษณะข้อมูลที่เหมือนหรือคล้ายคลึงกัน

2. อัลกอริทึม ANOVAID คือกระบวนการที่ใช้วิเคราะห์และแก้ปัญหาการจำแนกกลุ่มข้อมูลที่นิยมในการจำแนกกลุ่มข้อมูลที่มีลักษณะเป็นข้อมูลเชิงคุณภาพ โดยที่มีตัวแปรตามเป็นเชิงปริมาณจะใช้ One-Way ANOVA และ Independent-Sample t-test ในการจำแนกกลุ่ม โดยจะคัดเลือกตัวแปรพร้อมทั้งพิจารณาเกณฑ์การรวมและเกณฑ์การหยุดควบคู่กันเป็นขั้นตอน จนกระทั่งเสร็จสิ้นกระบวนการ

3. การวิเคราะห์ความแปรปรวนทางเดียว (One-Way ANOVA) คือกระบวนการศึกษาความสัมพันธ์ระหว่างตัวแปรตามที่เป็นตัวแปรเชิงปริมาณกับตัวแปรอิสระตัวแปรเดียวที่เป็นตัวแปรเชิงคุณภาพ

4. สถิติทดสอบ t สำหรับกลุ่มตัวอย่าง 2 กลุ่มที่เป็นอิสระกัน (Independent-Sample t-test) คือกระบวนการทดสอบความแตกต่างระหว่างค่าเฉลี่ยของกลุ่มตัวอย่าง 2 กลุ่มที่เป็นอิสระกัน

5. เกณฑ์การรวม (Merging Rules) คือเกณฑ์ในการวิเคราะห์ข้อมูลและเป็นแนวทางในการดำเนินการรวมประเภทตัวแปรของข้อมูลเพื่อจำแนกกลุ่มข้อมูล

6. เกณฑ์การหยุด (Stopping Rules) คือเกณฑ์ในการวิเคราะห์ข้อมูลและเป็นแนวทางในการดำเนินการหยุดการจำแนกกลุ่มและแผนภาพการตัดสินใจเพื่อจำแนกกลุ่มข้อมูล

1.6 วิธีดำเนินการวิจัย

1. ศึกษาเนื้อหาและทฤษฎีที่เกี่ยวข้อง พร้อมทั้งแนวทางการแก้ไขปัญหาและกระบวนการทำงานในการจำแนกกลุ่มข้อมูลโดยการใช้อัลกอริทึม ANOVAID

2. จำลองข้อมูลตามขอบเขตที่ต้องการศึกษา

2.1 สร้างข้อมูลที่มีขนาด (n) และตัวแปรอิสระเชิงคุณภาพที่มีจำนวนกลุ่ม (k) ตามที่กำหนด

2.2 กำหนดค่าเฉลี่ยของแต่ละกลุ่ม (μ_i) แบ่งตามแต่ละกรณีตามที่กำหนด โดยให้ $\mu_1 = 100$

2.3 สร้างข้อมูลตามตัวแบบ One-Way ANOVA

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$
$$\mu_i = \mu + \tau_i$$

เมื่อ $i = 1,2,\dots,k$ และ $j = 1,2,\dots,n$

โดย Y_{ij} = ตัวแปรตามสำหรับปัจจัยกลุ่มที่ i และค่าสังเกตที่ j

μ = ค่าเฉลี่ยรวมทั้งหมด

τ_i = อิทธิพลจากปัจจัยกลุ่มที่ i

ε_{ij} = ค่าความคลาดเคลื่อนสุ่มสำหรับปัจจัยกลุ่มที่ i และค่าสังเกตที่ j

และเป็นตัวแบบ One-Way ANOVA ที่มีจำนวนค่าสังเกตเท่ากันในแต่ละกลุ่ม (Balanced Design)

2.4 กำหนดการแจกแจงของค่าความคลาดเคลื่อนสุ่ม (ε_{ij}) ให้มีการแจกแจงปกติ โดยให้ ค่าเฉลี่ย $\mu_\varepsilon = 0$ และมีความแปรปรวน $\sigma_\varepsilon^2 = 10{,}000$ และ 40,000 โดยค่าความแปรปรวนที่กำหนดพิจารณาจากค่าสัมประสิทธิ์ความแปรผัน (Coefficient of Variance, C.V.) เมื่อต้องการให้มีค่าเป็น 100% และ 200% โดยที่ค่าเฉลี่ยเท่ากับ 100

3. กำหนดระดับนัยสำคัญ $\alpha = 0.05$ และนำข้อมูลจำลองมาทำการทดสอบด้วย One-Way ANOVA จากนั้นพิจารณาค่า p-value โดยถ้าค่า p-value น้อยกว่าระดับนัยสำคัญที่กำหนด หมายความว่าตัวแปรอิสระนั้นมีอย่างน้อย 1 คู่ที่แตกต่างกัน

4. จากขั้นตอนข้างต้น ถ้าได้ว่าตัวแปรอิสระนั้นมีอย่างน้อย 1 คู่ที่แตกต่างกัน จะใช้ t-test พิจารณาต่อว่าคู่ไหนบ้างที่สามารถนำมารวมกันได้ โดยจะทำการทดสอบทีละคู่และใช้ค่า p-value ในการพิจารณา โดยถ้าค่า p-value มากกว่าระดับนัยสำคัญที่กำหนด หมายความว่าสามารถรวมคู่นั้นเข้าด้วยกันได้

5. จาก 2 ขั้นตอนข้างต้น จะทำซ้ำในแต่ละกรณีที่ศึกษา 1,000 รอบ จากนั้นคำนวณค่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระเพื่อใช้พิจารณาประสิทธิภาพในการจำแนกกลุ่มของอัลกอริทึม ANOVAID

6. นำข้อมูลจริงมาทดลองใช้งานกับอัลกอริทึม ANOVAID

7. วิเคราะห์และสรุปผลการศึกษาพร้อมจัดทำรูปเล่มวิทยานิพนธ์

1.7 ประโยชน์ที่คาดว่าจะได้รับ

เพื่อให้ผู้วิจัยและผู้สนใจศึกษา เข้าใจเกี่ยวกับกระบวนการจำแนกกลุ่มข้อมูลและเงื่อนไขต่างๆ ของอัลกอริทึม ANOVAID มากขึ้น และสามารถนำไปประยุกต์ใช้ในด้านต่างๆ ที่สนใจต่อไปได้

บทที่ 2

ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

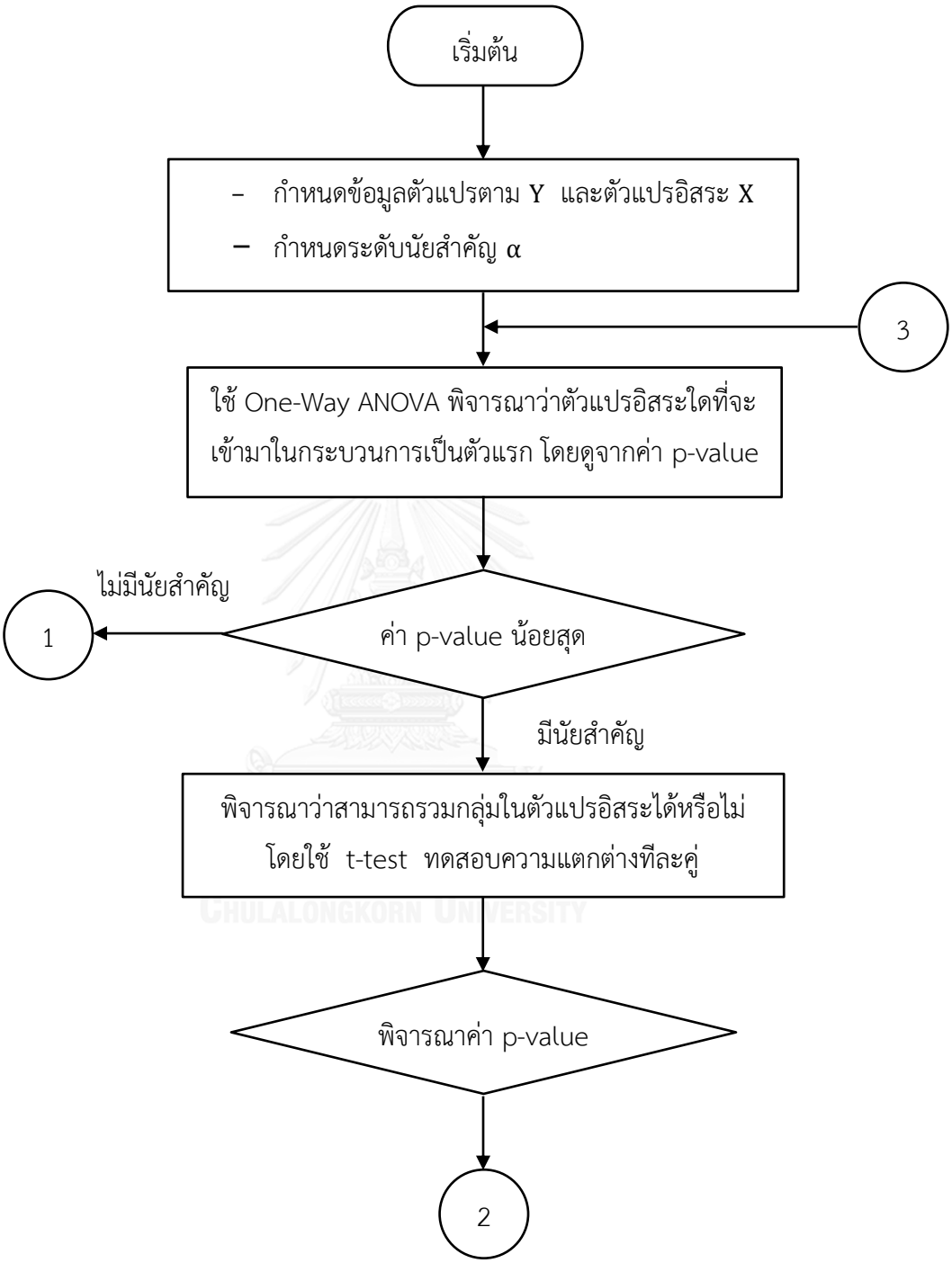
ในการวิจัยครั้งนี้ได้ทำการศึกษาเกี่ยวกับการจำแนกกลุ่มข้อมูล (Data Classification) โดยใช้การวิเคราะห์ความแปรปรวนทางเดียว (One-Way Analysis of Variance: One-Way ANOVA) และสถิติทดสอบ t สำหรับกลุ่มตัวอย่าง 2 กลุ่มที่เป็นอิสระกัน (Independent-Sample t test) มาใช้ควบคู่กัน โดยให้ชื่อกระบวนการในการจำแนกกลุ่มนี้ใหม่ว่า อัลกอริทึม ANOVAID สามารถใช้งานได้กับตัวแปรตามที่เป็นตัวแปรเชิงปริมาณและตัวแปรอิสระที่เป็นตัวแปรเชิงคุณภาพ โดยตัวแปรตามมีตัวเดียว ส่วนตัวแปรอิสระมีได้หลายตัว แต่ในขั้นตอนการจำแนกกลุ่มนั้นจะพิจารณาจากตัวแปรอิสระทีละตัวแยกกัน โดยมีขั้นตอนในการจำแนกกลุ่มและทฤษฎีที่เกี่ยวข้อง ดังนี้

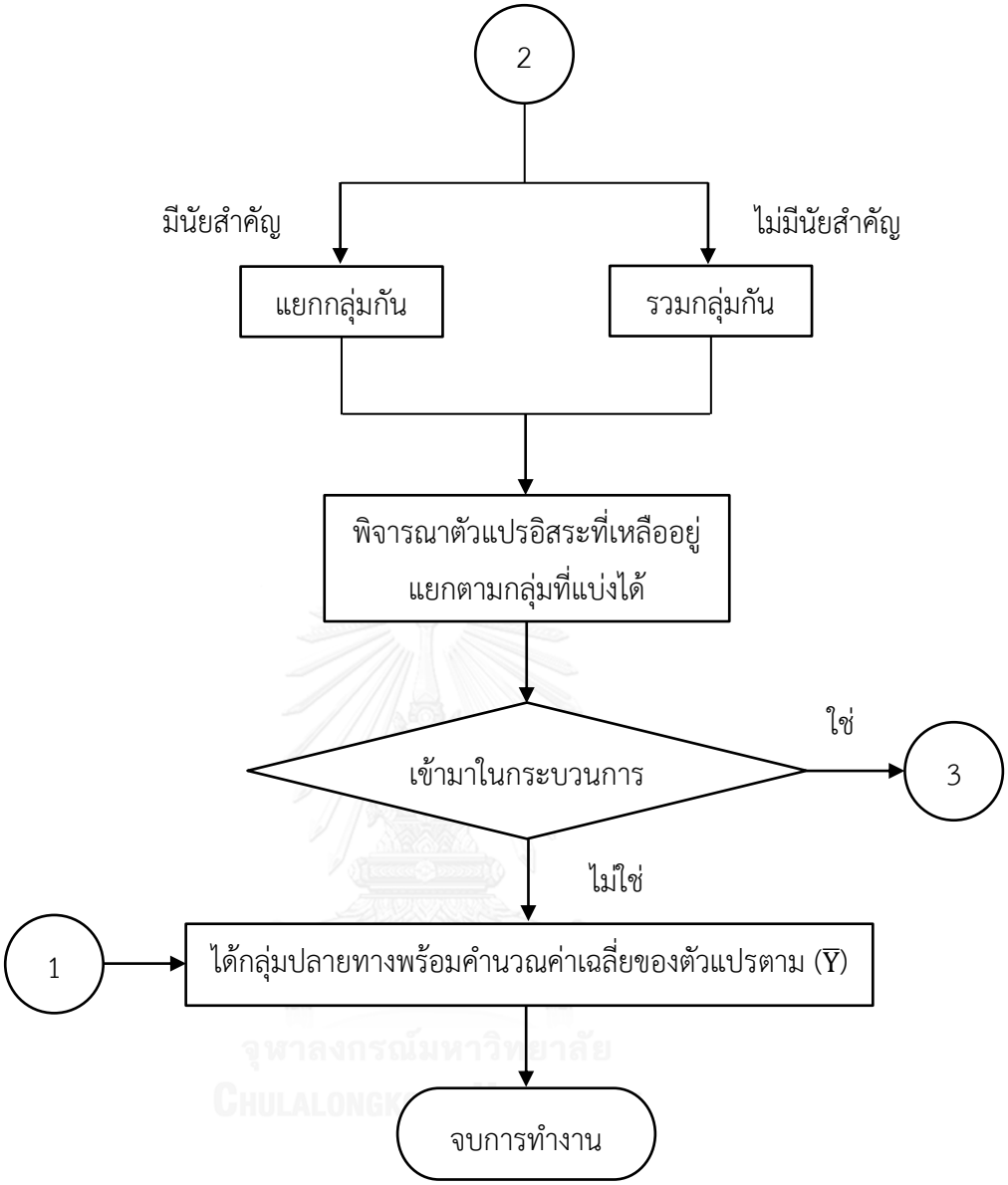
2.1 อัลกอริทึม ANOVAID

อัลกอริทึม ANOVAID เป็นอัลกอริทึมที่ใช้ในการจำแนกกลุ่มข้อมูล และสามารถสร้างแผนภาพต้นไม้การจำแนกแบบพหุภาค (Multi-branch Tree) ได้ โดยกระบวนการจำแนกกลุ่มข้อมูลของอัลกอริทึมได้นำการวิเคราะห์ความแปรปรวนทางเดียว (One-Way Analysis of Variance: One-Way ANOVA) และสถิติทดสอบ t สำหรับกลุ่มตัวอย่าง 2 กลุ่มที่เป็นอิสระกัน (Independent-Sample t test) มาใช้ควบคู่กัน โดยมีขั้นตอนการทำงานหลัก คือ การแยก (Splitting), การรวม (Merging) และการหยุด (Stopping)

โดยอัลกอริทึม ANOVAID จะใช้งานได้อย่างมีประสิทธิภาพในกรณีที่ตัวแปรตามเป็นตัวแปรเชิงปริมาณและตัวแปรอิสระเป็นตัวแปรเชิงคุณภาพ ถ้าตัวแปรตามเป็นตัวแปรเชิงคุณภาพจะต้องทำการแปลงให้เป็นตัวแปรหุ่น (Dummy Variable) เช่นเดียวกับตัวแปรอิสระ ถ้าเป็นตัวแปรเชิงปริมาณจะต้องทำการแปลงให้เป็นเชิงกลุ่มก่อน จึงจะเข้าสู่การทำงานในขั้นตอนต่อไป โดยเขียนแผนผังขั้นตอนในการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม ANOVAID ได้ดังนี้

ขั้นตอนในการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม ANOVAID





จากแผนผังข้างต้น สามารถอธิบายขั้นตอนในการจำแนกกลุ่มข้อมูลโดยอัลกอริทึม ANOVAID ได้ดังนี้

ขั้นตอนที่ 1 พิจารณาว่าตัวแปรอิสระใดที่จะเข้ามาในกระบวนการจำแนกกลุ่มข้อมูลก่อนตัวแปรอื่น หรือเรียกว่าขั้นตอนการแยก

1.1 พิจารณาชุดข้อมูลที่มีว่าตัวแปรไหนเป็นตัวแปรตามและตัวแปรไหนเป็นตัวแปรอิสระ พร้อมทั้งกำหนดค่าของระดับนัยสำคัญ (α) ที่จะใช้ในการทดสอบ

1.2 ใช้ One-Way ANOVA ในการพิจารณาว่าตัวแปรอิสระใดที่จะเข้ามาในกระบวนการจำแนกกลุ่มข้อมูลเป็นตัวแรก โดยพิจารณาจากค่า p-value น้อยสุดที่มีนัยสำคัญ (Significant)

การวิเคราะห์ความแปรปรวนทางเดียว (One-Way ANOVA)

(กัลยา วานิชย์บัญชา, 2551) กล่าวว่าการวิเคราะห์ความแปรปรวนทางเดียว เป็นการจำแนกข้อมูลด้วยตัวแปรหรือปัจจัยเพียงตัวเดียว นั่นคือวิเคราะห์ความแตกต่างของข้อมูลโดยพิจารณาจากปัจจัยที่มีผลต่อข้อมูลเพียงปัจจัยเดียว หรือเป็นการวิเคราะห์ความแตกต่างกันของระดับต่างๆ ของปัจจัยเพียงปัจจัยเดียวนั่นเอง ดังนั้นวัตถุประสงค์ของการวิเคราะห์ความแปรปรวนทางเดียว คือ การทดสอบความแตกต่างระหว่างค่าเฉลี่ยของประชากรที่ได้รับปัจจัยที่ต่างระดับกันนั่นเอง

ในงานวิจัยนี้ จะนำเสนอการวิเคราะห์ความแปรปรวนทางเดียวในแผนการทดลองแบบสุ่มโดยสมบูรณ์ (Completely Randomized Design: CRD) ซึ่งเป็นการทดลองเพื่อเปรียบเทียบ k ปัจจัยกลุ่ม โดยการสุ่มตัวอย่าง k กลุ่มอย่างเป็นอิสระกันแล้วกำหนดปัจจัยกลุ่มให้ตัวอย่างแต่ละกลุ่มอย่างสุ่ม ซึ่งมีข้อตกลงเบื้องต้นในการทำการวิเคราะห์ดังนี้

1. กลุ่มตัวอย่างถูกสุ่มมาจากประชากรที่มีการแจกแจงแบบปกติ (Normality)
2. กลุ่มตัวอย่างถูกสุ่มมาจากประชากรที่เป็นอิสระต่อกัน (Independent)
3. ความแปรปรวนของแต่ละประชากรต้องเท่ากัน คือ $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ (Homogeneity)

แผนการทดลองแบบสุ่มโดยสมบูรณ์มีตัวแบบดังนี้

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

$$\mu_i = \mu + \tau_i$$

เมื่อ $i = 1, 2, \dots, k$ และ $j = 1, 2, \dots, n$

โดย Y_{ij} = ตัวแปรตามสำหรับปัจจัยกลุ่มที่ i และค่าสังเกตที่ j

μ = ค่าเฉลี่ยรวมทั้งหมด

τ_i = อิทธิพลจากปัจจัยกลุ่มที่ i

ε_{ij} = ค่าความคลาดเคลื่อนสุ่มสำหรับปัจจัยกลุ่มที่ i และค่าสังเกตที่ j

และเป็นตัวแบบ One-Way ANOVA ที่มีจำนวนค่าสังเกตเท่ากันในแต่ละกลุ่ม (Balanced Design)

สมมติฐานสำหรับการทดสอบ สำหรับปัจจัยทดลองเป็นปัจจัยคงที่ (Fixed Factor)

$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$

$H_1:$ มี $\mu_i \neq \mu_j$ อย่างน้อย 1 คู่ ; $i \neq j$

สถิติที่ใช้ในการทดสอบ คือ สถิติทดสอบ F โดยสามารถคำนวณค่าได้ดังนี้

$F = \frac{MSTrt}{MSE}$ ซึ่งมีการแจกแจงแบบ F ด้วยองศาอิสระ $k-1, n-k$

เมื่อ $MSTrt = \frac{SSTrt}{k-1}$

$MSE = \frac{SSE}{n-k}$

$SST = \sum\sum(X_{ij} - \overline{X})^2$

$SSTrt = \sum n_i(\overline{X}_i - \overline{X})^2$

$SSE = SST - SSTrt = \sum\sum(X_{ij} - \overline{X}_i)^2$

โดยที่ X_{ij} คือ ข้อมูลของค่าสังเกตที่ j ในปัจจัยกลุ่มที่ i ; $i = 1,2, \dots, k$; $j = 1,2, \dots, n_i$

$n = n_1 + n_2 + \cdots + n_k$ คือ ขนาดของข้อมูลหรือหน่วยตัวอย่างทั้งหมด

$\overline{X} = \sum\sum X_{ij}/n$ คือ ค่าเฉลี่ยของข้อมูลทั้งหมด

$\overline{X}_i$ คือ ค่าเฉลี่ยของข้อมูลในปัจจัยกลุ่มที่ i

และจะสามารถคำนวณค่า p-value ได้ดังนี้

$$p - value = Pr(F_{\alpha,k-1,n-k} > F)$$

โดยมีเขตปฏิเสธสมมติฐานว่าง ดังนี้

จะปฏิเสธ H_0 ถ้า $F > F_{1-\alpha;k-1,n-k}$

หรือในงานวิจัยนี้เราจะพิจารณาจากค่า p-value โดยถ้า p-value มีค่าน้อยกว่าระดับนัยสำคัญที่กำหนด เราจะปฏิเสธสมมติฐานว่าง

ขั้นตอนที่ 2 พิจารณาว่าจะสามารถรวมกลุ่มข้อมูลในตัวแปรอิสระที่เข้ามาได้หรือไม่หรือเรียกว่าขั้นตอนการรวม

จากขั้นตอนที่ 1 นำตัวแปรอิสระที่ได้มาพิจารณาว่าจะสามารถรวมกลุ่มข้อมูลได้หรือไม่ โดยการใช้สถิติทดสอบ t สำหรับกลุ่มตัวอย่าง 2 กลุ่มที่เป็นอิสระกัน มาทดสอบความแตกต่างทีละคู่ พิจารณาจากค่า p-value โดยคู่ไหนที่ทดสอบแล้วไม่มีนัยสำคัญจะสามารถรวมกลุ่มกันได้

สถิติทดสอบ t สำหรับกลุ่มตัวอย่าง 2 กลุ่มที่เป็นอิสระกัน (Independent-Sample t-test)

(กัลยา วานิชย์บัญชา, 2551) กล่าวว่าเป็นการทดสอบความแตกต่างระหว่างค่าเฉลี่ยของปัจจัยกลุ่มที่สนใจของ 2 กลุ่มตัวอย่างว่าแตกต่างกันหรือไม่ โดยใช้ข้อมูลตัวอย่าง 2 ชุดที่เป็นอิสระกัน ซึ่งมีข้อตกลงเบื้องต้นในการทำการวิเคราะห์ คือ กลุ่มตัวอย่างทั้งสองต้องถูกสุ่มมาจากประชากรที่มีการแจกแจงแบบปกติ

ในงานวิจัยนี้ จะทำการศึกษาในกรณีที่ความแปรปรวนของทั้ง 2 กลุ่มเท่ากัน ($\sigma_1^2 = \sigma_2^2$) ซึ่งมีสมมติฐานสำหรับการทดสอบ ดังนี้

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

สถิติที่ใช้ในการทดสอบ คือ สถิติทดสอบ t โดยสามารถคำนวณค่าได้ดังนี้

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{MSE(\frac{1}{n_1} + \frac{1}{n_2})}}$$ ซึ่งมีการแจกแจงแบบ t ด้วยองศาอิสระ $n_1 + n_2 - 2$

โดยที่ MSE คือ ค่าเฉลี่ยกำลังสองภายในปัจจัยกลุ่มที่ได้จากขั้นตอนการทำ One-Way ANOVA

$\bar{x}_i$ คือ ค่าเฉลี่ยตัวอย่างของกลุ่มที่ i ; $i = 1, 2$

n_i คือ ขนาดตัวอย่างของกลุ่มที่ i ; $i = 1, 2$

และจะสามารถคำนวณค่า p-value ได้ดังนี้

$$p - value = 2Pr(t > t_{\alpha/2,\, n_1+n_2-2})$$

โดยมีเขตปฏิเสธสมมติฐานว่าง ดังนี้

จะปฏิเสธ H_0 ถ้า $|t| > t_{1-\alpha/2;\, n_1+n_2-2}$

หรือในงานวิจัยนี้เราจะพิจารณาจากค่า p-value โดยถ้า p-value มีค่าน้อยกว่าระดับนัยสำคัญที่กำหนด เราจะปฏิเสธสมมติฐานว่าง

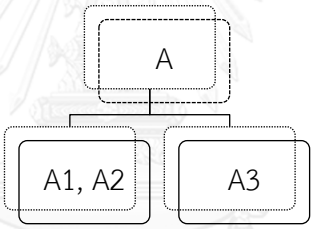
จาก 2 ขั้นตอนข้างต้น จะสามารถจำแนกกลุ่มของตัวแปรอิสระที่เข้ามาในกระบวนการได้ จากนั้นจึงมาพิจารณาตัวแปรอิสระที่เหลืออยู่ โดยทำซ้ำขั้นตอนที่ 1 และ 2 โดยที่ข้อมูลในตัวแปรอิสระนั้นๆ ต้องขึ้นอยู่กับข้อมูลในกลุ่มที่ทำการพิจารณาอยู่ด้วย โดยจะทำการแยกพิจารณาเป็นกลุ่มๆ ไป

สำหรับขั้นตอนการหยุดจะพิจารณาที่ขั้นตอนที่ 1 หรือขั้นตอนในการใช้ One-Way ANOVA เมื่อค่า p-value ทุกค่าที่ได้จากตัวแปรอิสระที่เหลืออยู่มีค่ามากว่าระดับนัยสำคัญที่กำหนดไว้ นั่นคือไม่มีนัยสำคัญจึงจะหยุดกระบวนการ คือไม่สามารถแบ่งกลุ่มข้อมูลต่อได้แล้ว

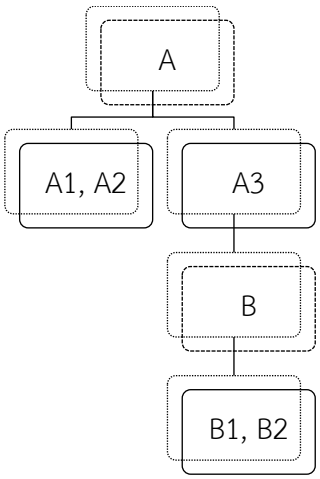
ขอยกตัวอย่างสำหรับการอธิบายขั้นตอนการจำแนกกลุ่มข้อมูลของอัลกอริทึม ANOVAID ให้เห็นภาพชัดเจนขึ้น โดยสมมติให้ชุดข้อมูลที่นำมาใช้ประกอบไปด้วยตัวแปรตาม 1 ตัวและตัวแปรอิสระ 2 ตัว ตัวแปรอิสระ A มีจำนวนกลุ่ม 3 กลุ่มและตัวแปรอิสระ B มีจำนวนกลุ่ม 2 กลุ่ม ทำการทดสอบที่ระดับนัยสำคัญเท่ากับ 0.05

ในขั้นตอนแรกเราจะใช้ One-Way ANOVA หาค่า p-value ของตัวแปรอิสระทั้งสองออกมา สมมติให้ค่า p-value ของตัวแปร A มีค่าเป็น 0.04 และของตัวแปร B มีค่าเป็น 0.06 จะได้ว่าค่า p-value ที่มีค่าน้อยกว่าคือ 0.04 และมีค่าน้อยกว่าระดับนัยสำคัญเท่ากับ 0.05 ที่กำหนดไว้ นั่นคือมีนัยสำคัญ เราจึงเลือกตัวแปร A เข้ามาในกระบวนการเป็นตัวแรก

ขั้นตอนต่อไปเราจะใช้ t-test มาทดสอบความแตกต่างระหว่างกลุ่มในตัวแปร A ทีละคู่ เนื่องจากตัวแปร A มีจำนวนกลุ่ม 3 กลุ่ม จะได้ว่าต้องทำการทดสอบ 3 ครั้ง คือ ทดสอบความแตกต่างระหว่างกลุ่ม A1 กับ A2, กลุ่ม A1 กับ A3 และกลุ่ม A2 กับ A3 สมมติให้ค่า p-value ของกลุ่ม A1 กับ A2 มีค่าเป็น 0.06, ของกลุ่ม A1 กับ A3 มีค่าเป็น 0.04 และของกลุ่ม A2 กับ A3 มีค่าเป็น 0.045 จะได้ว่าค่า p-value ที่ไม่มีนัยสำคัญคือค่า p-value ของกลุ่ม A1 กับ A2 นั่นคือสามารถรวมทั้งสองกลุ่มเข้าด้วยกันได้ จากตรงนี้จะได้ว่า ตัวแปร A สามารถแบ่งออกได้เป็น 2 กลุ่ม คือกลุ่มที่ 1 มาจากการรวมกลุ่ม A1 กับ A2 เข้าด้วยกัน และกลุ่มที่ 2 ซึ่งเป็นกลุ่ม A3 ที่เหลืออยู่ เขียนแสดงเป็นภาพได้ดังนี้



จากนั้นเราจะนำแต่ละกลุ่มที่แบ่งได้มาพิจารณากับตัวแปรอิสระที่เหลืออยู่ว่าจะทำการแยกต่อ หรือหยุดกระบวนการ โดยทำซ้ำขั้นตอน One-Way ANOVA แล้วพิจารณาจากค่า p-value ที่หาได้ ในที่นี้ตัวแปรอิสระที่เหลืออยู่มีแค่ตัวแปรเดียวคือตัวแปร B สมมติให้ในกลุ่มของ A1+A2 ค่า p-value ที่ได้มีค่าเป็น 0.07 หมายความว่า กลุ่มนี้จะไม่นำตัวแปร B เข้ามาพิจารณาต่อ นั่นคือการหยุดกระบวนการ และสมมติให้ในกลุ่มของ A3 ค่า p-value ที่ได้มีค่าเป็น 0.03 หมายความว่า กลุ่มนี้จะนำตัวแปร B เข้ามาพิจารณาต่อ โดยการทำซ้ำขั้นตอน t-test แล้วพิจารณาจากค่า p-value ว่ากลุ่มในตัวแปร B สามารถรวมกันได้หรือไม่ โดยตัวแปร B มีจำนวนกลุ่ม 2 กลุ่ม นั่นคือเราต้องทดสอบความแตกต่างระหว่างกลุ่ม B1 กับ B2 สมมติให้ค่า p-value ที่ได้มีค่าเป็น 0.08 หมายความว่าสามารถรวมกลุ่ม B1 กับ B2 เข้าด้วยกันได้ เขียนแสดงเป็นภาพได้ดังนี้



เมื่อเสร็จสิ้นกระบวนการจำแนกกลุ่มแล้ว สรุปได้ว่า กลุ่มปลายทางที่สามารถจำแนกได้มีอยู่ 2 กลุ่ม คือ กลุ่มที่มีคุณลักษณะของ A1 หรือ A2 และกลุ่มที่มีคุณลักษณะของ A3 และ B1 หรือ A3 และ B2

2.2 เกณฑ์ที่ใช้ในการวัดประสิทธิภาพการจำแนกกลุ่ม

วัดประสิทธิภาพของอัลกอริทึมโดยพิจารณาจากเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระภายใต้กรณีที่ทำการศึกษา โดยทำซ้ำในแต่ละกรณี 1,000 รอบ ยกตัวอย่างในกรณีที่ $\mu_1 = \mu_2$ หมายความว่าเมื่อทำการทดสอบโดยอัลกอริทึมแล้วค่าเฉลี่ยของทั้งสองกลุ่มต้องไม่แตกต่างกัน นั่นคือทั้งสองกลุ่มรวมกันเป็นกลุ่มเดียวได้ ฉะนั้นในการทำซ้ำ 1,000 รอบ ผลที่ออกมาคือทั้งสองกลุ่มต้องรวมกันได้ทั้ง 1,000 รอบ นั่นคือใน 1,000 รอบนั้นถ้ามีรอบไหนที่ผลออกมาแล้วทั้งสองกลุ่มไม่รวมกัน จึงจะนับรอบนั้นเป็นความผิดพลาดในการจำแนกกลุ่ม ดังนั้นเปอร์เซ็นต์ความผิดพลาดสามารถคำนวณได้ดังนี้

$$\text{เปอร์เซ็นต์ความผิดพลาด} = \frac{\text{จำนวนความผิดพลาดในการจำแนกกลุ่ม}}{\mathbf{1{,}000}} \times 100$$

บทที่ 3

วิธีดำเนินการวิจัย

การวิจัยครั้งนี้เป็นการศึกษาเกี่ยวกับการจำแนกกลุ่มข้อมูล (Data Classification) โดยใช้การวิเคราะห์ความแปรปรวนทางเดียว (One-Way Analysis of Variance: One-Way ANOVA) และสถิติทดสอบ t สำหรับกลุ่มตัวอย่าง 2 กลุ่มที่เป็นอิสระกัน (Independent-Sample t-test) มาใช้ควบคู่กัน โดยให้ชื่อกระบวนการในการจำแนกกลุ่มนี้ใหม่ว่า อัลกอริทึม ANOVAID สามารถใช้งานได้กับตัวแปรตามที่เป็นตัวแปรเชิงปริมาณและตัวแปรอิสระที่เป็นตัวแปรเชิงคุณภาพ โดยที่ตัวแปรตามมีตัวเดียว ส่วนตัวแปรอิสระมีหลายตัวได้ และในการวัดประสิทธิภาพของอัลกอริทึมจะทำโดยการวัดเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระภายใต้กรณีที่ทำการศึกษา โดยทำซ้ำในแต่ละกรณี 1,000 รอบ

การจำลองข้อมูลในแต่ละกรณีนั้นผู้วิจัยทำงานด้วยโปรแกรม R เวอร์ชั่น 3.2.3 ในบทนี้จะกล่าวถึงแผนการดำเนินการวิจัย ขั้นตอนในการดำเนินการวิจัย และขั้นตอนการดำเนินการของโปรแกรม ซึ่งมีรายละเอียด ดังนี้

3.1 แผนการดำเนินการวิจัย

ในงานวิจัยครั้งนี้ได้ทำการจำลองข้อมูลตามตัวแบบ One-Way ANOVA โดยข้อมูลจำลองที่ได้จะประกอบไปด้วยตัวแปรตามและตัวแปรอิสระอย่างละตัว ซึ่งในความเป็นจริงแล้วอัลกอริทึม ANOVAID สามารถใช้กับตัวแปรอิสระมากกว่า 1 ตัวได้ แต่ในขั้นตอนการจำแนกกลุ่มนั้นจะพิจารณาจากตัวแปรอิสระทีละตัวแยกกัน เรียกว่า ปัจจัย (Factor) ดังนั้นในงานวิจัยนี้จึงได้ให้ความสำคัญที่ประสิทธิภาพในการจำแนกกลุ่มของตัวแปรอิสระแต่ละตัว โดยกำหนดสถานการณ์ต่างๆ ที่จะทำการศึกษา ดังนี้

1. ตัวแบบที่ใช้ในการศึกษา คือ ตัวแบบ One-Way ANOVA ซึ่งเขียนในรูปแบบถดถอยได้ดังนี้

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

$$\mu_i = \mu + \tau_i$$

เมื่อ $i = 1,2,\dots,k$ และ $j = 1,2,\dots,n$

โดย Y_{ij} = ตัวแปรตามสำหรับปัจจัยกลุ่มที่ i และค่าสังเกตที่ j

μ = ค่าเฉลี่ยรวมทั้งหมด

τ_i = อิทธิพลจากปัจจัยกลุ่มที่ i

ε_{ij} = ค่าความคลาดเคลื่อนสุ่มสำหรับปัจจัยกลุ่มที่ i และค่าสังเกตที่ j

และเป็นตัวแบบ One-Way ANOVA ที่มีจำนวนค่าสังเกตเท่ากันในแต่ละกลุ่ม (Balanced Design)

2. การแจกแจงของค่าความคลาดเคลื่อนสุ่ม (ε_{ij}) ที่ทำการศึกษามีการแจกแจงปกติ กำหนดให้ ค่าเฉลี่ย $\mu_\varepsilon = 0$ และมีความแปรปรวน $\sigma_\varepsilon^2 =$ 10,000 และ 40,000 โดยค่าความแปรปรวนที่กำหนดพิจารณาจากค่าสัมประสิทธิ์ความแปรผัน (Coefficient of Variance, C.V.) เมื่อต้องการให้มีค่าเป็น 100% และ 200% โดยที่ค่าเฉลี่ยเท่ากับ 100

3. ศึกษาภายใต้จำนวนกลุ่มกับขนาดข้อมูล (k กับ n) $\mu_1 = 100$ ดังนี้

กรณีที่ 1: k = 2 กับ n = 6,000, 12,000 และ 24,000

กรณีที่ 1.1: $\mu_1 = \mu_2$

กรณีที่ 1.2.1: $\mu_2 = 0.5\mu_1$

กรณีที่ 1.2.2: $\mu_2 = 2\mu_1$

กรณีที่ 2: k = 3 กับ n = 6,000, 12,000 และ 24,000

กรณีที่ 2.1: $\mu_1 = \mu_2 = \mu_3$

กรณีที่ 2.2.1: $\mu_3 = \mu_2 = 0.5\mu_1$

กรณีที่ 2.2.2: $\mu_3 = \mu_2 = 2\mu_1$

กรณีที่ 2.3.1: $\mu_3 = 0.5\mu_2,\ \mu_2 = 0.5\mu_1$

กรณีที่ 2.3.2: $\mu_3 = 2\mu_2,\ \mu_2 = 2\mu_1$

กรณีที่ 3: k = 4 กับ n = 6,000, 12,000 และ 24,000

กรณีที่ 3.1: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

กรณีที่ 3.2.1: $\mu_4 = \mu_3 = \mu_2 = 0.5\mu_1$

กรณีที่ 3.2.2: $\mu_4 = \mu_3 = \mu_2 = 2\mu_1$

กรณีที่ 3.3.1: $\mu_1 = \mu_2,\ \mu_4 = \mu_3 = 0.5\mu_1$

กรณีที่ 3.3.2: $\mu_1 = \mu_2,\ \mu_4 = \mu_3 = 2\mu_1$

กรณีที่ 3.4.1: $\mu_4 = \mu_3 = 0.5\mu_2,\ \mu_2 = 0.5\mu_1$

กรณีที่ 3.4.2: $\mu_4 = \mu_3 = 2\mu_2,\ \mu_2 = 2\mu_1$

กรณีที่ 3.5.1: $\mu_4 = 0.5\mu_3,\ \mu_3 = 0.5\mu_2,\ \mu_2 = 0.5\mu_1$

กรณีที่ 3.5.2: $\mu_4 = 2\mu_3,\ \mu_3 = 2\mu_2,\ \mu_2 = 2\mu_1$

4. การวิจัยครั้งนี้ได้ทำการจำลองข้อมูลให้มีสถานการณ์ดังกล่าวข้างต้น ซึ่งผู้วิจัยจะประเมินผลด้วยโปรแกรม R เวอร์ชั่น 3.2.3 โดยทำการจำลองในแต่ละสถานการณ์ 1,000 รอบ

3.2 ขั้นตอนในการดำเนินการวิจัย

สำหรับการดำเนินการวิจัย มีดังนี้

1. ศึกษาเนื้อหาและทฤษฎีที่เกี่ยวข้อง พร้อมทั้งแนวทางการแก้ไขปัญหาและกระบวนการทำงานในการจำแนกกลุ่มข้อมูลโดยการใช้อัลกอริทึม ANOVAID

2. จำลองข้อมูลตามขอบเขตที่ต้องการศึกษา

2.1 สร้างข้อมูลที่มีขนาด (n) และตัวแปรอิสระเชิงคุณภาพที่มีจำนวนกลุ่ม (k) ตามที่กำหนด

2.2 กำหนดค่าเฉลี่ยของแต่ละกลุ่ม (μ_i) แบ่งตามแต่ละกรณีตามที่กำหนด โดยให้ $\mu_1 = 100$

2.3 สร้างข้อมูลตามตัวแบบ One-Way ANOVA

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

$$\mu_i = \mu + \tau_i$$

เมื่อ $i = 1,2,\ldots,k$ และ $j = 1,2,\ldots,n$

โดย Y_{ij} = ตัวแปรตามสำหรับปัจจัยกลุ่มที่ i และค่าสังเกตที่ j

μ = ค่าเฉลี่ยรวมทั้งหมด

τ_i = อิทธิพลจากปัจจัยกลุ่มที่ i

ε_{ij} = ค่าความคลาดเคลื่อนสุ่มสำหรับปัจจัยกลุ่มที่ i และค่าสังเกตที่ j

และเป็นตัวแบบ One-Way ANOVA ที่มีจำนวนค่าสังเกตเท่ากันในแต่ละกลุ่ม (Balanced Design)

2.4 กำหนดการแจกแจงของค่าความคลาดเคลื่อนสุ่ม (ε_{ij}) ให้มีการแจกแจงปกติ โดยให้ ค่าเฉลี่ย $\mu_\varepsilon = 0$ และมีความแปรปรวน $\sigma_\varepsilon^2 = 10{,}000$ และ 40,000 โดยค่าความแปรปรวนที่กำหนดพิจารณาจากค่าสัมประสิทธิ์ความแปรผัน (Coefficient of Variance, C.V.) เมื่อต้องการให้มีค่าเป็น 100% และ 200% โดยที่ค่าเฉลี่ยเท่ากับ 100

3. กำหนดระดับนัยสำคัญ $\alpha = 0.05$ และนำข้อมูลจำลองมาทำการทดสอบด้วย One-Way ANOVA จากนั้นพิจารณาค่า p-value โดยถ้าค่า p-value น้อยกว่าระดับนัยสำคัญที่กำหนด หมายความว่าตัวแปรอิสระนั้นมีอย่างน้อย 1 คู่ที่แตกต่างกัน

4. จากขั้นตอนข้างต้น ถ้าได้ว่าตัวแปรอิสระนั้นมีอย่างน้อย 1 คู่ที่แตกต่างกัน จะใช้ t-test พิจารณาต่อว่าคู่ไหนบ้างที่สามารถนำมารวมกันได้ โดยจะทำการทดสอบทีละคู่และใช้ค่า p-value ในการพิจารณา โดยถ้าค่า p-value มากกว่าระดับนัยสำคัญที่กำหนด หมายความว่าสามารถรวมคู่นั้นเข้าด้วยกันได้

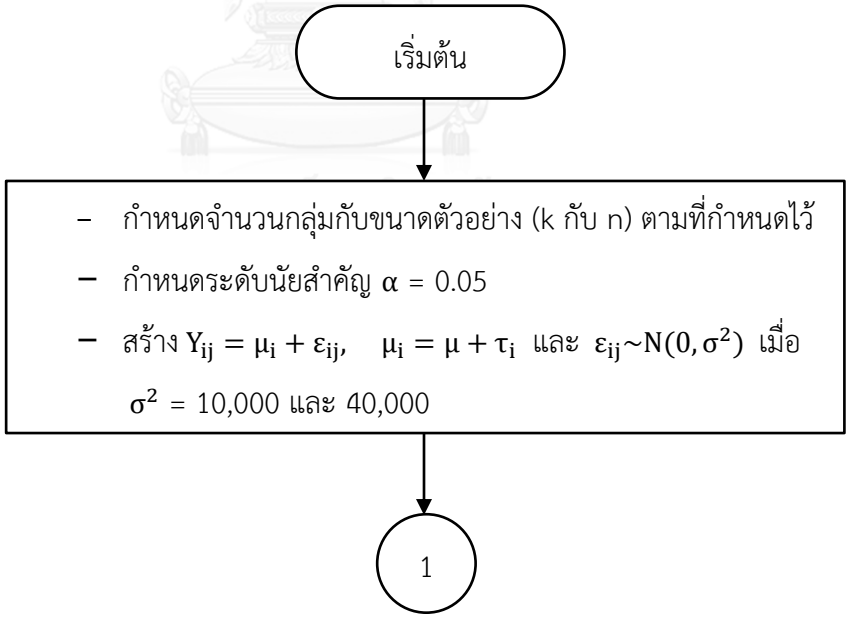
5. จาก 2 ขั้นตอนข้างต้น จะทำซ้ำในแต่ละกรณีที่ศึกษา 1,000 รอบ จากนั้นคำนวณค่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระเพื่อใช้พิจารณาประสิทธิภาพในการจำแนกกลุ่มของอัลกอริทึม ANOVAID

6. นำข้อมูลจริงมาทดลองใช้งานกับอัลกอริทึม ANOVAID

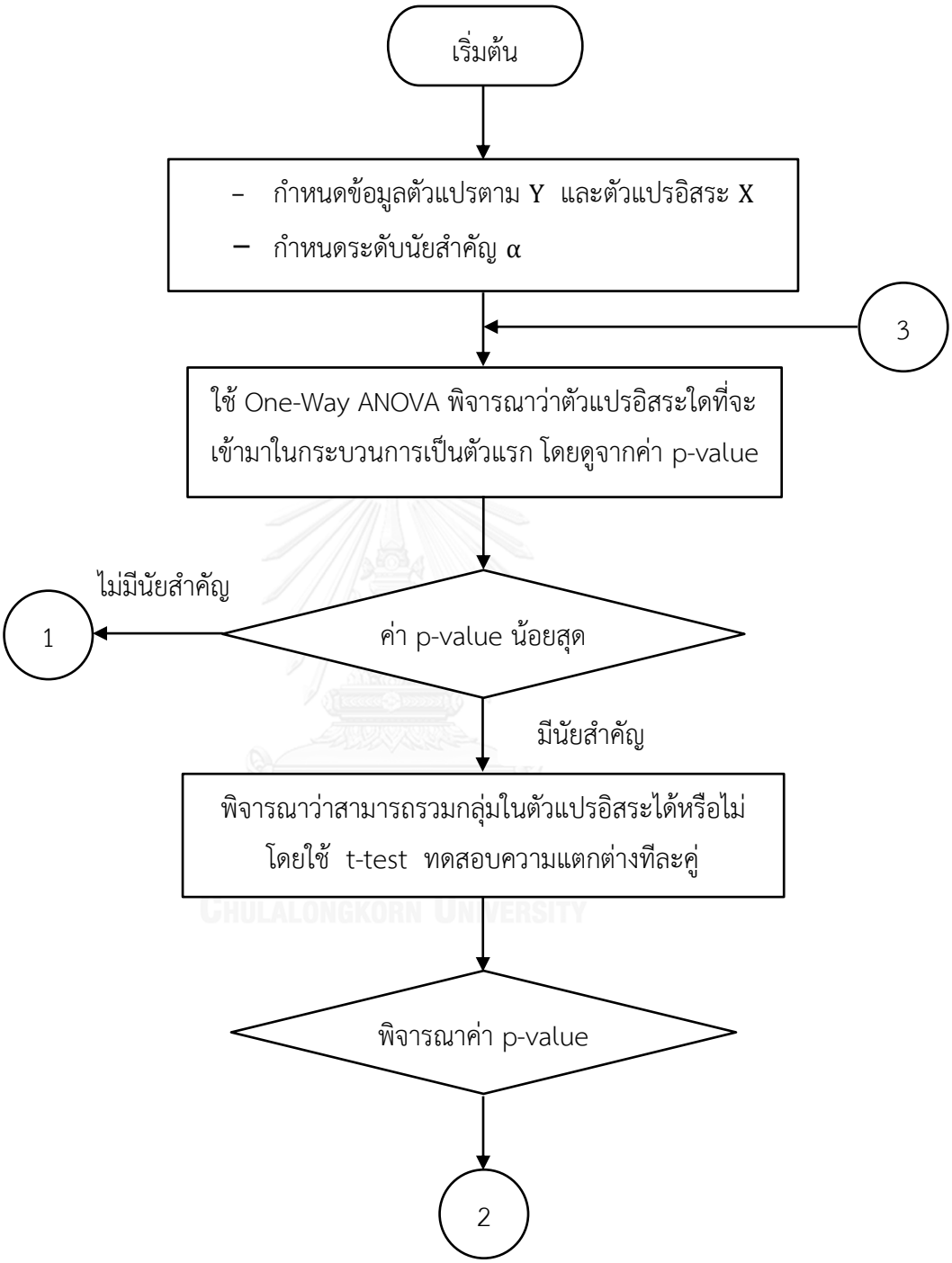
7. วิเคราะห์และสรุปผลการศึกษาพร้อมจัดทำรูปเล่มวิทยานิพนธ์

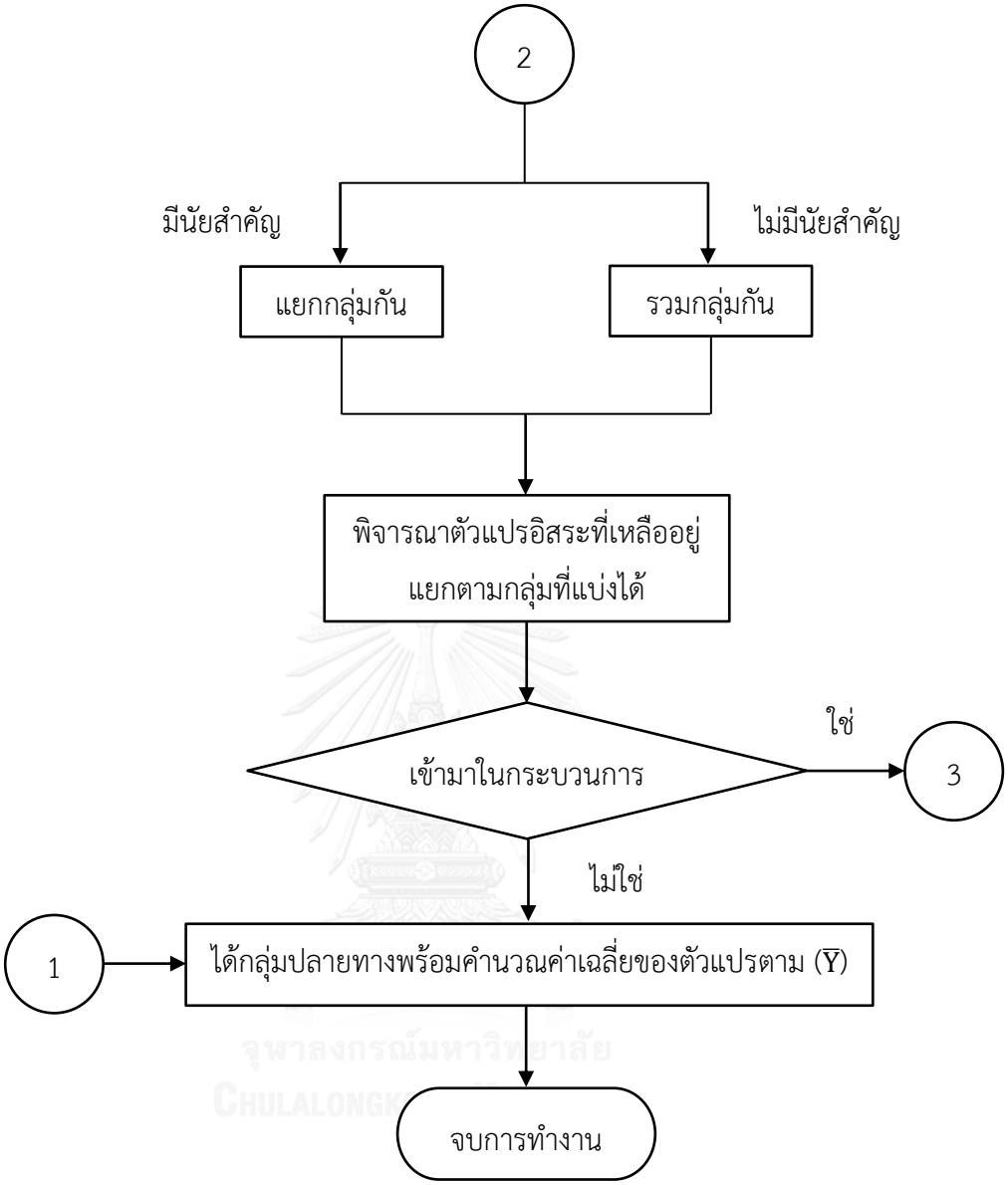
3.3 ขั้นตอนการดำเนินการของโปรแกรม

ภาพที่ 1 แสดงแผนผังขั้นตอนการวิจัย



ภาพที่ 2 แสดงแผนผังกระบวนการจำแนกกลุ่มข้อมูลของอัลกอริทึม ANOVAID





บทที่ 4

ผลการวิเคราะห์ข้อมูล

การศึกษางานวิจัยในครั้งนี้มีวัตถุประสงค์เพื่อศึกษากระบวนการจำแนกกลุ่มของอัลกอริทึม ANOVAID อย่างละเอียด โดยทำการศึกษาภายใต้จำนวนกลุ่ม (k) เท่ากับ 2, 3 และ 4, ขนาดข้อมูล (n) เท่ากับ 6,000, 12,000 และ 24,000, ความแปรปรวนของความคลาดเคลื่อนสุ่ม (ε_{ij}) เท่ากับ 10,000 และ 40,000 และกำหนดค่าเฉลี่ยของแต่ละกลุ่ม (μ_i) ตามแต่ละกรณีที่ศึกษา โดย $\mu_1 = 100$ ทำการทดสอบที่ระดับนัยสำคัญ (α) เท่ากับ 0.05 จากการศึกษาในครั้งนี้เกณฑ์การพิจารณาประสิทธิภาพการจำแนกกลุ่มของอัลกอริทึม คือ เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระ สมมติฐานโดยใช้ข้อมูลจำลอง

ในการนำเสนอผลการวิจัยจะแสดงในรูปแบบของตาราง โดยมีสัญลักษณ์ที่ใช้แทนความหมายต่าง ๆ ดังนี้

n	แทน	ขนาดของข้อมูล
k	แทน	จำนวนกลุ่มของตัวแปรอิสระ
μ_i	แทน	ค่าเฉลี่ยของแต่ละกลุ่ม
σ_ε^2	แทน	ความแปรปรวนของความคลาดเคลื่อนสุ่ม (ε_{ij})

การนำเสนอผลการวิจัยครั้งนี้ จะแบ่งการนำเสนอออกเป็น 2 ส่วนด้วยกัน โดยที่ส่วนแรกจะเป็นผลการวิเคราะห์เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระและส่วนสุดท้ายจะเป็นการนำอัลกอริทึมมาใช้งานกับข้อมูลจริง

4.1 ผลการวิเคราะห์เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระ

พิจารณาจากเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระภายใต้กรณีที่ทำการศึกษา โดยทำซ้ำในแต่ละกรณี 1,000 รอบ ยกตัวอย่างในกรณีที่ $\mu_1 = \mu_2$ หมายความว่าเมื่อทำการทดสอบโดยอัลกอริทึมแล้วค่าเฉลี่ยของทั้งสองกลุ่มต้องไม่แตกต่างกัน นั่นคือทั้งสองกลุ่มรวมกันเป็นกลุ่มเดียวได้ ฉะนั้นในการทำซ้ำ 1,000 รอบ ผลที่ออกมาคือทั้งสองกลุ่มต้องรวมกันได้ทั้ง 1,000 รอบ นั่นคือใน 1,000 รอบนั้นถ้ามีรอบไหนที่ผลออกมาแล้วทั้งสองกลุ่มไม่รวมกัน จึงจะนับรอบนั้นเป็นความผิดพลาดในการจำแนกกลุ่ม ดังนั้นเปอร์เซ็นต์ความผิดพลาดสามารถคำนวณได้ดังนี้

$$\text{เปอร์เซ็นต์ความผิดพลาด} = \frac{\text{จำนวนความผิดพลาดในการจำแนกกลุ่ม}}{\mathbf{1{,}000}} \times 100$$

สำหรับเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระในแต่ละกรณีนั้น จะนำเสนอดังนี้

ตารางที่ 4. 1 แสดงเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระในกรณีที่ค่าเฉลี่ยของแต่ละกลุ่มเท่ากันทั้งหมด

กรณีที่ศึกษา	ความแปรปรวน	ขนาดข้อมูล	เปอร์เซ็นต์ความผิดพลาด
$\mu_1 = \mu_2$	10,000	6,000	5.1%
		12,000	4.7%
		24,000	4.2%
	40,000	6,000	6.2%
		12,000	5.6%
		24,000	5.2%
$\mu_1 = \mu_2 = \mu_3$	10,000	6,000	4.7%
		12,000	4.5%
		24,000	4.1%
	40,000	6,000	5.9%
		12,000	5.3%
		24,000	4.7%
$\mu_1 = \mu_2 = \mu_3 = \mu_4$	10,000	6,000	4.5%
		12,000	4.4%
		24,000	4.3%
	40,000	6,000	5.6%
		12,000	5.1%
		24,000	5.1%

จากตารางที่ 4.1 เมื่อพิจารณาเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระในกรณีที่ค่าเฉลี่ยของแต่ละกลุ่มเท่ากันทั้งหมด สรุปผลได้ดังนี้

ในกรณีที่ $\mu_1 = \mu_2$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเฉลี่ยอยู่ที่ 5.2% มีค่าต่ำสุดอยู่ที่ 4.2% ในกรณีที่ความแปรปรวนเท่ากับ 10,000 และขนาดข้อมูลเท่ากับ 24,000 และมีค่าสูงสุดอยู่ที่ 6.2% ในกรณีที่ความแปรปรวนเท่ากับ 40,000 และขนาดข้อมูลเท่ากับ 6,000 และเมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มลดลง และที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มเพิ่มขึ้น

ในกรณีที่ $\mu_1 = \mu_2 = \mu_3$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเฉลี่ยอยู่ที่ 4.9% มีค่าต่ำสุดอยู่ที่ 4.1% ในกรณีที่ความแปรปรวนเท่ากับ 10,000 และขนาดข้อมูลเท่ากับ 24,000 และมีค่าสูงสุดอยู่ที่ 5.9% ในกรณีที่ความแปรปรวนเท่ากับ 40,000 และขนาดข้อมูลเท่ากับ 6,000 และเมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มลดลง และที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มเพิ่มขึ้น

ในกรณีที่ $\mu_1 = \mu_2 = \mu_3 = \mu_4$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเฉลี่ยอยู่ที่ 4.8% มีค่าต่ำสุดอยู่ที่ 4.3% ในกรณีที่ความแปรปรวนเท่ากับ 10,000 และขนาดข้อมูลเท่ากับ 24,000 และมีค่าสูงสุดอยู่ที่ 5.6% ในกรณีที่ความแปรปรวนเท่ากับ 40,000 และขนาดข้อมูลเท่ากับ 6,000 และเมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มลดลง และที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มเพิ่มขึ้น

สามารถสรุปได้ว่า เมื่อพิจารณาจากเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระในกรณีที่ค่าเฉลี่ยของแต่ละกลุ่มเท่ากันทั้งหมดค่าที่ได้จะอยู่ราว ๆ 5.0% และเมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มลดลง และที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มเพิ่มขึ้นในทุกกรณีที่ศึกษา

ตารางที่ 4. 2 แสดงเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระในกรณีที่ค่าเฉลี่ยแตกต่างกันเป็น 0.5 เท่า

กรณีที่ศึกษา	ความแปรปรวน	ขนาดข้อมูล	เปอร์เซ็นต์ความผิดพลาด
$\mu_2 = 0.5\mu_1$	10,000	6,000	0.0%
		12,000	0.0%
		24,000	0.0%
	40,000	6,000	0.0%
		12,000	0.0%
		24,000	0.0%
$\mu_3 = \mu_2 = 0.5\mu_1$	10,000	6,000	5.8%
		12,000	5.1%
		24,000	4.6%
	40,000	6,000	6.2%
		12,000	5.4%
		24,000	5.0%
$\mu_3 = 0.5\mu_2,$ $\mu_2 = 0.5\mu_1$	10,000	6,000	0.0%
		12,000	0.0%
		24,000	0.0%
	40,000	6,000	2.6%
		12,000	0.0%
		24,000	0.0%
$\mu_4 = \mu_3 = \mu_2 = 0.5\mu_1$	10,000	6,000	2.9%
		12,000	2.7%
		24,000	2.6%
	40,000	6,000	3.4%
		12,000	3.2%
		24,000	3.1%

กรณีที่ศึกษา	ความแปรปรวน	ขนาดข้อมูล	เปอร์เซ็นต์ความผิดพลาด
$\mu_1 = \mu_2$, $\mu_4 = \mu_3 = 0.5\mu_1$	10,000	6,000	9.7%
		12,000	8.4%
		24,000	8.4%
	40,000	6,000	11.1%
		12,000	10.9%
		24,000	10.3%
$\mu_4 = \mu_3 = 0.5\mu_2$, $\mu_2 = 0.5\mu_1$	10,000	6,000	4.9%
		12,000	4.8%
		24,000	4.6%
	40,000	6,000	12.0%
		12,000	5.2%
		24,000	5.1%
$\mu_4 = 0.5\mu_3$, $\mu_3 = 0.5\mu_2$, $\mu_2 = 0.5\mu_1$	10,000	6,000	7.2%
		12,000	0.2%
		24,000	0.0%
	40,000	6,000	62.8%
		12,000	31.0%
		24,000	7.1%

จากตารางที่ 4.2 เมื่อพิจารณาเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระในกรณีที่ค่าเฉลี่ยแตกต่างกันเป็น 0.5 เท่า สรุปผลได้ดังนี้

ในกรณีที่ $\mu_2 = 0.5\mu_1$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเท่ากับ 0.0%

ในกรณีที่ $\mu_3 = \mu_2 = 0.5\mu_1$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเฉลี่ยอยู่ที่ 5.4% มีค่าต่ำสุดอยู่ที่ 4.6% ในกรณีที่ความแปรปรวนเท่ากับ 10,000 และขนาดข้อมูลเท่ากับ 24,000 และมีค่าสูงสุดอยู่ที่ 6.2% ในกรณีที่ความแปรปรวนเท่ากับ 40,000 และขนาดข้อมูลเท่ากับ 6,000 และเมื่อพิจารณาที่

ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มลดลง และที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มเพิ่มขึ้น

ในกรณีที่ $\mu_3 = 0.5\mu_2,\ \mu_2 = 0.5\mu_1$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเท่ากับ 0.0% ยกเว้นกรณีที่ความแปรปรวนเท่ากับ 40,000 และขนาดข้อมูลเท่ากับ 6,000 ที่เปอร์เซ็นต์ความผิดพลาดจะมีค่าเท่ากับ 2.6%

ในกรณีที่ $\mu_4 = \mu_3 = \mu_2 = 0.5\mu_1$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเฉลี่ยอยู่ที่ 3.0% มีค่าต่ำสุดอยู่ที่ 2.6% ในกรณีที่ความแปรปรวนเท่ากับ 10,000 และขนาดข้อมูลเท่ากับ 24,000 และมีค่าสูงสุดอยู่ที่ 3.4% ในกรณีที่ความแปรปรวนเท่ากับ 40,000 และขนาดข้อมูลเท่ากับ 6,000 และเมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มลดลง และที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มเพิ่มขึ้น

ในกรณีที่ $\mu_1 = \mu_2,\ \mu_4 = \mu_3 = 0.5\mu_1$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเฉลี่ยอยู่ที่ 9.8% มีค่าต่ำสุดอยู่ที่ 8.4% ในกรณีที่ความแปรปรวนเท่ากับ 10,000 และขนาดข้อมูลเท่ากับ 12,000 และ 24,000 และมีค่าสูงสุดอยู่ที่ 11.1% ในกรณีที่ความแปรปรวนเท่ากับ 40,000 และขนาดข้อมูลเท่ากับ 6,000 และเมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มลดลง และที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มเพิ่มขึ้น

ในกรณีที่ $\mu_4 = \mu_3 = 0.5\mu_2,\ \mu_2 = 0.5\mu_1$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเฉลี่ยอยู่ที่ 6.1% มีค่าต่ำสุดอยู่ที่ 4.6% ในกรณีที่ความแปรปรวนเท่ากับ 10,000 และขนาดข้อมูลเท่ากับ 24,000 และมีค่าสูงสุดอยู่ที่ 12.0% ในกรณีที่ความแปรปรวนเท่ากับ 40,000 และขนาดข้อมูลเท่ากับ 6,000 และเมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มลดลง และที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มเพิ่มขึ้น

ในกรณีที่ $\mu_4 = 0.5\mu_3,\ \mu_3 = 0.5\mu_2,\ \mu_2 = 0.5\mu_1$ เมื่อพิจารณาที่ความแปรปรวนเท่ากับ 10,000 จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเฉลี่ยอยู่ที่ 2.5% มีค่าต่ำสุดอยู่ที่ 0.0% ในกรณีที่ขนาดข้อมูลเท่ากับ 24,000 และมีค่าสูงสุดอยู่ที่ 7.2% ในกรณีที่ขนาดข้อมูลเท่ากับ 6,000 และเมื่อพิจารณาที่ความแปรปรวนเท่ากับ 40,000 จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเฉลี่ยอยู่ที่ 33.6% มีค่าต่ำสุดอยู่ที่ 7.1% ในกรณีที่ขนาดข้อมูลเท่ากับ 24,000 และมีค่าสูงสุดอยู่ที่ 62.8% ในกรณีที่ขนาดข้อมูลเท่ากับ 6,000 และเมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มลดลง และที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มเพิ่มขึ้น

สามารถสรุปได้ว่า เมื่อพิจารณาจากเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระในกรณีที่ค่าเฉลี่ยแตกต่างกันเป็น 0.5 เท่า เปอร์เซ็นต์ความผิดพลาดเฉลี่ยที่ได้จะไม่เกิน 10.0% ยกเว้นกรณีที่ $\mu_4 = 0.5\mu_3,\ \mu_3 = 0.5\mu_2,\ \mu_2 = 0.5\mu_1$ ที่ระดับความแปรปรวนเท่ากับ 40,000 ที่เปอร์เซ็นต์ความผิดพลาดเฉลี่ยมีค่าเท่ากับ 33.6% และเมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มลดลง และที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มเพิ่มขึ้นในทุกกรณีที่ศึกษา

ตารางที่ 4. 3 แสดงเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระในกรณีที่ค่าเฉลี่ยแตกต่างกันเป็น 2 เท่า

กรณีที่ศึกษา	ความแปรปรวน	ขนาดข้อมูล	เปอร์เซ็นต์ความผิดพลาด
$\mu_2 = 2\mu_1$	10,000	6,000	0.0%
		12,000	0.0%
		24,000	0.0%
	40,000	6,000	0.0%
		12,000	0.0%
		24,000	0.0%
$\mu_3 = \mu_2 = 2\mu_1$	10,000	6,000	5.0%
		12,000	4.8%
		24,000	3.6%
	40,000	6,000	5.6%
		12,000	5.1%
		24,000	4.4%
$\mu_3 = 2\mu_2, \mu_2 = 2\mu_1$	10,000	6,000	0.0%
		12,000	0.0%
		24,000	0.0%
	40,000	6,000	0.0%
		12,000	0.0%
		24,000	0.0%
$\mu_4 = \mu_3 = \mu_2 = 2\mu_1$	10,000	6,000	3.0%
		12,000	2.8%
		24,000	2.6%
	40,000	6,000	3.6%
		12,000	3.5%
		24,000	3.0%

กรณีที่ศึกษา	ความแปรปรวน	ขนาดข้อมูล	เปอร์เซ็นต์ความผิดพลาด
$\mu_1 = \mu_2$, $\mu_4 = \mu_3 = 2\mu_1$	10,000	6,000	9.3%
		12,000	8.6%
		24,000	7.4%
	40,000	6,000	11.4%
		12,000	10.5%
		24,000	9.4%
$\mu_4 = \mu_3 = 2\mu_2$, $\mu_2 = 2\mu_1$	10,000	6,000	4.6%
		12,000	4.3%
		24,000	4.0%
	40,000	6,000	5.2%
		12,000	4.8%
		24,000	4.6%
$\mu_4 = 2\mu_3$, $\mu_3 = 2\mu_2$, $\mu_2 = 2\mu_1$	10,000	6,000	0.0%
		12,000	0.0%
		24,000	0.0%
	40,000	6,000	0.0%
		12,000	0.0%
		24,000	0.0%

จากตารางที่ 4.3 เมื่อพิจารณาเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระในกรณีที่ค่าเฉลี่ยแตกต่างกันเป็น 2 เท่า สรุปผลได้ดังนี้

ในกรณีที่ $\mu_2 = 2\mu_1$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเท่ากับ 0.0%

ในกรณีที่ $\mu_3 = \mu_2 = 2\mu_1$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเฉลี่ยอยู่ที่ 4.8% มีค่าต่ำสุดอยู่ที่ 3.6% ในกรณีที่ความแปรปรวนเท่ากับ 10,000 และขนาดข้อมูลเท่ากับ 24,000 และมีค่าสูงสุดอยู่ที่ 5.6% ในกรณีที่ความแปรปรวนเท่ากับ 40,000 และขนาดข้อมูลเท่ากับ 6,000 และเมื่อพิจารณาที่

ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มลดลง และที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มเพิ่มขึ้น

ในกรณีที่ $\mu_3 = 2\mu_2,\ \mu_2 = 2\mu_1$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเท่ากับ 0.0%

ในกรณีที่ $\mu_4 = \mu_3 = \mu_2 = 2\mu_1$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเฉลี่ยอยู่ที่ 3.1% มีค่าต่ำสุดอยู่ที่ 2.6% ในกรณีที่ความแปรปรวนเท่ากับ 10,000 และขนาดข้อมูลเท่ากับ 24,000 และมีค่าสูงสุดอยู่ที่ 3.6% ในกรณีที่ความแปรปรวนเท่ากับ 40,000 และขนาดข้อมูลเท่ากับ 6,000 และเมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มลดลง และที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มเพิ่มขึ้น

ในกรณีที่ $\mu_1 = \mu_2,\ \mu_4 = \mu_3 = 2\mu_1$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเฉลี่ยอยู่ที่ 9.4% มีค่าต่ำสุดอยู่ที่ 7.4% ในกรณีที่ความแปรปรวนเท่ากับ 10,000 และขนาดข้อมูลเท่ากับ 24,000 และมีค่าสูงสุดอยู่ที่ 11.4% ในกรณีที่ความแปรปรวนเท่ากับ 40,000 และขนาดข้อมูลเท่ากับ 6,000 และเมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มลดลง และที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มเพิ่มขึ้น

ในกรณีที่ $\mu_4 = \mu_3 = 2\mu_2,\ \mu_2 = 2\mu_1$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเฉลี่ยอยู่ที่ 4.6% มีค่าต่ำสุดอยู่ที่ 4.0% ในกรณีที่ความแปรปรวนเท่ากับ 10,000 และขนาดข้อมูลเท่ากับ 24,000 และมีค่าสูงสุดอยู่ที่ 5.2% ในกรณีที่ความแปรปรวนเท่ากับ 40,000 และขนาดข้อมูลเท่ากับ 6,000 และเมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มลดลง และที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มเพิ่มขึ้น

ในกรณีที่ $\mu_4 = 2\mu_3,\ \mu_3 = 2\mu_2,\ \mu_2 = 2\mu_1$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระมีค่าเท่ากับ 0.0%

สามารถสรุปได้ว่า เมื่อพิจารณาจากเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระในกรณีที่ค่าเฉลี่ยแตกต่างกันเป็น 2 เท่า เปอร์เซ็นต์ความผิดพลาดเฉลี่ยที่ได้จะไม่เกิน 5.0% ยกเว้นกรณีที่ $\mu_1 = \mu_2,\ \mu_4 = \mu_3 = 2\mu_1$ ที่เปอร์เซ็นต์ความผิดพลาดเฉลี่ยมีค่าเท่ากับ 9.4% และเมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มลดลง และที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดจะมีแนวโน้มเพิ่มขึ้นในทุกกรณีที่ศึกษา

4.2 การนำอัลกอริทึม ANOVAID มาใช้งานกับข้อมูลจริง

โดยข้อมูลที่นำมาใช้จะประกอบไปด้วยคุณลักษณะและราคาของเพชร ซึ่งตัวแปรตาม คือ ราคา (price) มีหน่วยเป็นดอลลาร์สิงคโปร์ ส่วนตัวแปรอิสระ คือ สี (colour) = (D, E, F, G, H, I), ความบริสุทธิ์ (clarity) = (IF, VVS1, VVS2, VS1, VS2) และใบรับรองคุณภาพ (certification) = (GIA, IGI, HRD)

จากผลการศึกษา สามารถแบ่งกลุ่มของเพชรได้ 5 กลุ่ม ซึ่งมีคุณลักษณะ, จำนวนข้อมูลและราคาเฉลี่ย ดังนี้

กลุ่มที่ 1 เป็นเพชรที่มีใบรับรองคุณภาพของบริษัท GIA โดยมีจำนวนข้อมูลเท่ากับ 151 และมีราคาเฉลี่ย 5,310.42 ดอลลาร์สิงคโปร์

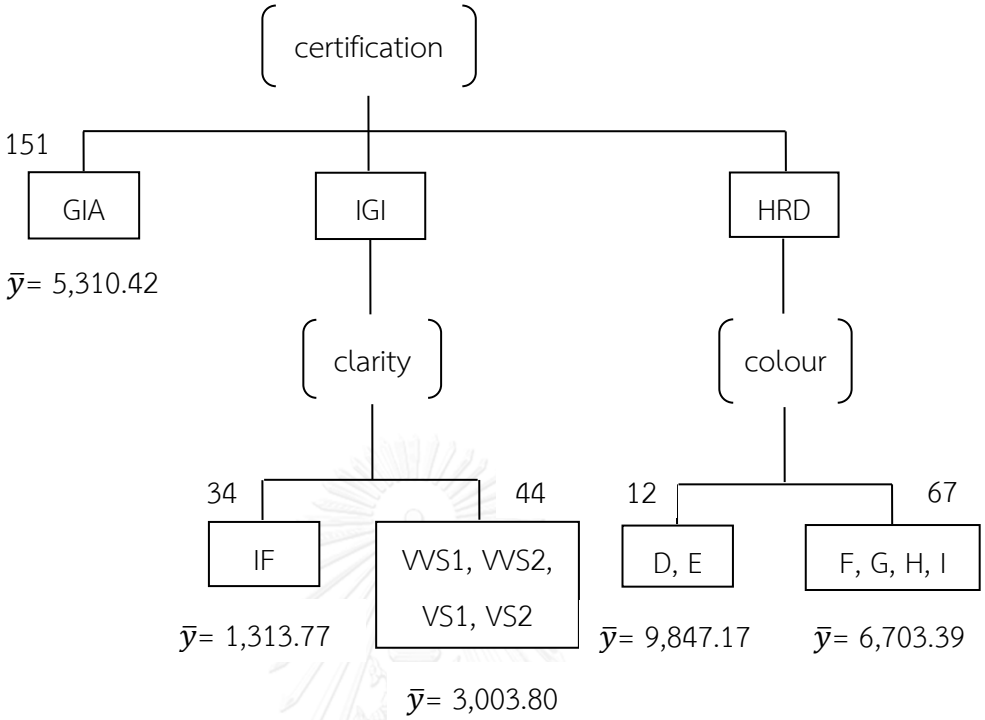
กลุ่มที่ 2 เป็นเพชรที่มีใบรับรองคุณภาพของบริษัท IGI และมีความบริสุทธิ์อยู่ที่ระดับ IF โดยมีจำนวนข้อมูลเท่ากับ 34 และมีราคาเฉลี่ย 1,313.77 ดอลลาร์สิงคโปร์

กลุ่มที่ 3 เป็นเพชรที่มีใบรับรองคุณภาพของบริษัท IGI และมีความบริสุทธิ์อยู่ที่ระดับ VVS1, VVS2, VS1, VS2 โดยมีจำนวนข้อมูลเท่ากับ 44 และมีราคาเฉลี่ย 3,003.80 ดอลลาร์สิงคโปร์

กลุ่มที่ 4 เป็นเพชรที่มีใบรับรองคุณภาพของบริษัท HRD และมีสีอยู่ที่ระดับ D, E โดยมีจำนวนข้อมูลเท่ากับ 12 และมีราคาเฉลี่ย 9,847.17 ดอลลาร์สิงคโปร์

กลุ่มที่ 5 เป็นเพชรที่มีใบรับรองคุณภาพของบริษัท HRD และมีสีอยู่ที่ระดับ F, G, H, I โดยมีจำนวนข้อมูลเท่ากับ 67 และมีราคาเฉลี่ย 6,703.39 ดอลลาร์สิงคโปร์

โดยสามารถแสดงแผนผังการจำแนกกลุ่มข้อมูลได้ดังนี้



บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

จากการศึกษากระบวนการจำแนกกลุ่มข้อมูลของอัลกอริทึม ANOVAID พร้อมทั้งทำการวัดประสิทธิภาพในการจำแนกกลุ่ม โดยทำการศึกษาภายใต้จำนวนกลุ่มของตัวแปรอิสระเป็น 2, 3 และ 4, ขนาดข้อมูลเป็น 6,000, 12,000 และ 24,000, ความแปรปรวนเป็น 10,000 และ 40,000 และค่าเฉลี่ยของแต่ละกลุ่มในสถานการณ์ต่างๆ โดยพิจารณาจากเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระ เพื่อหาข้อสรุปว่าอัลกอริทึม ANOVAID มีประสิทธิภาพเพียงพอต่อการใช้งานหรือไม่ สามารถสรุปผลการวิจัยได้ดังนี้

5.1 สรุปผลการวิจัย

ผลการวิเคราะห์เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระ

เมื่อพิจารณาเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระในกรณีที่ค่าเฉลี่ยของแต่ละกลุ่มเท่ากันทั้งหมด ได้ผลสรุปดังนี้

1. เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มมีค่าเฉลี่ยอยู่ที่ราวๆ 5.0% ในทุกกรณีที่ศึกษา

2. กรณีที่จำนวนกลุ่มเป็น 2 และ $\mu_1 = \mu_2$ เมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มลดลง และเมื่อพิจารณาที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มเพิ่มขึ้น

3. กรณีที่จำนวนกลุ่มเป็น 3 และ $\mu_1 = \mu_2 = \mu_3$ เมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มลดลง และเมื่อพิจารณาที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มเพิ่มขึ้น

4. กรณีที่จำนวนกลุ่มเป็น 4 และ $\mu_1 = \mu_2 = \mu_3 = \mu_4$ เมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มลดลง และเมื่อพิจารณาที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มเพิ่มขึ้น

เมื่อพิจารณาเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระในกรณีที่ค่าเฉลี่ยแตกต่างกัน ได้ผลสรุปดังนี้

1. กรณีที่ค่าเฉลี่ยแตกต่างกันเป็น 0.5 เท่า เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มมีค่าเฉลี่ยไม่เกิน 10.0% ในทุกกรณีที่ศึกษา ยกเว้นกรณีที่ $\mu_4 = 0.5\mu_3,\ \mu_3 = 0.5\mu_2,\ \mu_2 = 0.5\mu_1$ ที่ระดับความแปรปรวนเท่ากับ 40,000 ที่เปอร์เซ็นต์ความผิดพลาดเฉลี่ยมีค่าเท่ากับ 33.6%

2. กรณีที่ค่าเฉลี่ยแตกต่างกันเป็น 2 เท่า เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มมีค่าเฉลี่ยไม่เกิน 5.0% ในทุกกรณีที่ศึกษา ยกเว้นกรณีที่ $\mu_1 = \mu_2,\ \mu_4 = \mu_3 = 2\mu_1$ ที่เปอร์เซ็นต์ความผิดพลาดเฉลี่ยมีค่าเท่ากับ 9.4%

3. กรณีที่จำนวนกลุ่มเป็น 2 และ $\mu_2 = 0.5\mu_1$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มมีค่าเท่ากับ 0.0% เช่นเดียวกับในกรณีที่ $\mu_2 = 2\mu_1$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มมีค่าเท่ากับ 0.0% เช่นเดียวกัน

4. กรณีที่จำนวนกลุ่มเป็น 3 และ $\mu_3 = \mu_2 = 0.5\mu_1$ เมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มลดลง และเมื่อพิจารณาที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มเพิ่มขึ้น เช่นเดียวกับในกรณีที่ $\mu_3 = \mu_2 = 2\mu_1$ เมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มลดลง และเมื่อพิจารณาที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มเพิ่มขึ้นเช่นเดียวกัน

5. กรณีที่จำนวนกลุ่มเป็น 3 และ $\mu_3 = 0.5\mu_2,\ \mu_2 = 0.5\mu_1$ เมื่อพิจารณาที่ความแปรปรวนเท่ากับ 10,000 เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มมีค่าเท่ากับ 0.0% ในทุกขนาดข้อมูล ในขณะที่ความแปรปรวนเท่ากับ 40,000 เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มลดลง และเมื่อพิจารณาที่ขนาดข้อมูลเท่ากับ 6,000 เมื่อความแปรปรวน

เพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มเพิ่มขึ้น ในขณะที่ขนาดข้อมูลเท่ากับ 12,000 และ 24,000 เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มมีค่าเท่ากับ 0.0% ในทุกความแปรปรวน ส่วนในกรณีที่ $\mu_3 = 2\mu_2,\ \mu_2 = 2\mu_1$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มมีค่าเท่ากับ 0.0%

6. กรณีที่จำนวนกลุ่มเป็น 4 และ $\mu_4 = \mu_3 = \mu_2 = 0.5\mu_1$ เมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มลดลง และเมื่อพิจารณาที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มเพิ่มขึ้น เช่นเดียวกับในกรณีที่ $\mu_4 = \mu_3 = \mu_2 = 2\mu_1$ เมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มลดลง และเมื่อพิจารณาที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มเพิ่มขึ้นเช่นเดียวกัน

7. กรณีที่จำนวนกลุ่มเป็น 4 และ $\mu_1 = \mu_2,\ \mu_4 = \mu_3 = 0.5\mu_1$ เมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มลดลง และเมื่อพิจารณาที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มเพิ่มขึ้น เช่นเดียวกับในกรณีที่ $\mu_1 = \mu_2,\ \mu_4 = \mu_3 = 2\mu_1$ เมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มลดลง และเมื่อพิจารณาที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มเพิ่มขึ้นเช่นเดียวกัน

8. กรณีที่จำนวนกลุ่มเป็น 4 และ $\mu_4 = \mu_3 = 0.5\mu_2,\ \mu_2 = 0.5\mu_1$ เมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มลดลง และเมื่อพิจารณาที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มเพิ่มขึ้น เช่นเดียวกับในกรณีที่ $\mu_4 = \mu_3 = 2\mu_2,\ \mu_2 = 2\mu_1$ เมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มลดลง และเมื่อพิจารณาที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มเพิ่มขึ้นเช่นเดียวกัน

9. กรณีที่จำนวนกลุ่มเป็น 4 และ $\mu_4 = 0.5\mu_3,\ \mu_3 = 0.5\mu_2,\ \mu_2 = 0.5\mu_1$ เมื่อพิจารณาที่ความแปรปรวนเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มลดลง และเมื่อพิจารณาที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มเพิ่มขึ้น ส่วนในกรณีที่ $\mu_4 = 2\mu_3,\ \mu_3 = 2\mu_2,\ \mu_2 = 2\mu_1$ เมื่อพิจารณาจากทุกความแปรปรวนและขนาดข้อมูล จะได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มมีค่าเท่ากับ 0.0%

สรุป

จากการวัดประสิทธิภาพการจำแนกกลุ่มของอัลกอริทึม ANOVAID โดยพิจารณาจากเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มของตัวแปรอิสระในกรณีศึกษาต่างๆ พบว่าในกรณีที่ค่าเฉลี่ยของแต่ละกลุ่มเท่ากันทั้งหมด เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มมีค่าเฉลี่ยอยู่ที่ราวๆ 5.0%, ในกรณีที่ค่าเฉลี่ยแตกต่างกันเป็น 0.5 เท่า เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มมีค่าเฉลี่ยไม่เกิน 10.0% ในทุกกรณีที่ศึกษา ยกเว้นกรณีที่ $\mu_4 = 0.5\mu_3,\ \mu_3 = 0.5\mu_2,\ \mu_2 = 0.5\mu_1$ ที่ระดับความแปรปรวนเท่ากับ 40,000 ที่เปอร์เซ็นต์ความผิดพลาดเฉลี่ยมีค่าเท่ากับ 33.6% และในกรณีที่ค่าเฉลี่ยแตกต่างกันเป็น 2 เท่า เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มมีค่าเฉลี่ยไม่เกิน 5.0% ในทุกกรณีที่ศึกษา ยกเว้นกรณีที่ $\mu_1 = \mu_2,\ \mu_4 = \mu_3 = 2\mu_1$ ที่เปอร์เซ็นต์ความผิดพลาดเฉลี่ยมีค่าเท่ากับ 9.4% โดยรวมแล้วกล่าวได้ว่าเปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มมีค่าอยู่ในช่วงที่สามารถยอมรับได้ โดยเมื่อจำนวนกลุ่มของตัวแปรอิสระเป็น 2, 3 และ 4 เมื่อพิจารณาที่ความแปรปรวนมีค่าเท่ากัน เมื่อขนาดข้อมูลเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มลดลง, เมื่อพิจารณาที่ขนาดข้อมูลเท่ากัน เมื่อความแปรปรวนเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มเพิ่มขึ้น และเมื่อพิจารณาที่จำนวนเท่าของความแตกต่างของค่าเฉลี่ย เมื่อความแปรปรวนและขนาดข้อมูลมีค่าเท่ากัน พบว่าเมื่อจำนวนเท่าเพิ่มขึ้น เปอร์เซ็นต์ความผิดพลาดในการจำแนกกลุ่มจะมีแนวโน้มลดลง

5.2 ข้อเสนอแนะ

เนื่องจากงานวิจัยนี้มีขอบเขตการวิจัยที่จำกัด เพื่อให้เกิดประสิทธิภาพสูงสุดในการวิจัยหรือเพื่อหาข้อเท็จจริงที่ให้ผลที่ละเอียดมากยิ่งขึ้น สามารถปรับเปลี่ยนขอบเขตของการวิจัยในส่วนของความแตกต่างของค่าเฉลี่ย (μ_i) ในแต่ละกลุ่ม, จำนวนกลุ่ม (k), ขนาดข้อมูล (n), ความแปรปรวน (σ_ε^2) หรือจำนวนตัวแปรอิสระได้ เพื่อให้ครอบคลุมในขอบเขตต่าง ๆ และได้ข้อมูลที่ต้องการและเป็นประโยชน์ในการศึกษาต่อไป

รายการอ้างอิง

ภาษาไทย

กัลยา วานิชย์บัญชา. (2551). หลักสถิติ. กรุงเทพมหานคร: สำนักพิมพ์ธรรมสาร.

ภาษาต่างประเทศ

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics, 29*, 119-127.

Lai, H.-J., & Wu, H.-H. (2011). A Case Study of Applying Kano's Model and ANOVA Technique in Evaluating Service Quality. *Information Technology Journal, 10*, 89-97.

Thachn, L., & Olsen, J. (2015). Profiling the High Frequency Wine Consumer by Price Segmentation in the US Market. *Wine Economics and Policy, 4*, 53–59.

ภาคผนวก

โปรแกรม R ที่ใช้ในการจำลอง

ในการศึกษาครั้งนี้จะจำลองข้อมูลตามขอบเขตงานวิจัยข้างต้น ซึ่งผู้วิจัยจะประมวลผลโดยใช้โปรแกรม R เวอร์ชั่น 3.2.3 โดยวิธีดังกล่าวจะกำหนดการจำลองซ้ำของข้อมูลแต่ละกรณีไว้ที่จำนวน 1,000 รอบ

สำหรับข้อมูลจำลอง จะขอยกตัวอย่างในกรณีที่จำนวนกลุ่มเป็น 2 โดยที่ $\mu_1 = \mu_2$, ขนาดข้อมูลเท่ากับ 6,000 และความแปรปรวนเท่ากับ 10,000

```
anovaid_sim<-function(data){
  colofy<-which(names(data)=="DV")
  colofx<-which(names(data)!="DV")
  p_anova<-oneway.test(DV ~ ., data, var.equal=TRUE)$p.value
  if(p_anova<0.05){
    x_in<-as.character(data[,colofx])
    xall<-levels(factor(x_in))
    p_ttest<-1
    while(length(xall)>2&any(p_ttest>0.05)){
      p_ttest<-c()
      xpair<-combn(xall,2)
      for(i in 1:ncol(xpair))
      p_ttest[i]<-t.test(data$DV[x_in==xpair[1,i]],data$DV[x_in==xpair[2,i]], data=data,
var.equal=TRUE)$p.value
      names(p_ttest)<-apply(xpair,2,toString)
      if(any(p_ttest>0.05)){
        pairall<-which(p_ttest>0.05)
        pair_pmost<-which.max(p_ttest)
        pairmost<-xpair[,pair_pmost]
        pairall<-pairall[pairall!=pair_pmost]
        if(length(pairall)>0&nlevels(factor(x_in))>3){
          if(length(pairall)>1){
            xpairnew<-xpair[,pairall]
```

```
            torf<-apply(xpairnew,2,function(x)any(x%in%pairmost))
            gpfalse<-as.numeric(xpairnew[,!torf])
            gptnotf<-apply(xpairnew,2,function(x)!any(x%in%gpfalse))
            pairmost<-sort(unique(c(pairmost,as.vector(xpairnew[,gptnotf]))))
          } else if(any(xpair[,pairall]%in%pairmost)) pairmost<-
sort(unique(c(pairmost,xpair[,pairall])))
        }
        x_in[x_in%in%pairmost]<-toString(pairmost)
        xall<-levels(factor(x_in))
        p_ttest<-p_ttest[apply(xpair,2,function(x)!any(x%in%pairmost))]
      }
    }
    return(toString(paste(1:length(xall),xall,sep="-")))
  } else return("not diff")
}
k <- 2
Ni <- c(3000,3000)
N<-sum(Ni)
MUi <- rep(c(100,100),Ni)
sd <- sqrt(10000)
IV<-factor(rep(LETTERS[1:k],Ni))
result0<-matrix(,1000,k)
colnames(result0)<-LETTERS[1:k]
result<-c()
for (i in 1:1000){
  data <- data.frame(IV,DV=MUi+rnorm(N,0,sd))
  result0[i,]<-tapply(data$DV,data$IV,mean)
  result[i]<-anovaid_sim(data)
}
result0<-data.frame(result0,result)
table(result)
```

สำหรับข้อมูลจริง

```
anovaid<-function(data){
  colofy<-which(names(data)=="price")
  colofx<-which(names(data)!="price")
  p_anova<-c()
  for(i in colofx)
  p_anova<-c(p_anova,oneway.test(price ~ ., data=data[,c(i,colofy)],
var.equal=TRUE)$p.value)
  names(p_anova)<-names(data)[colofx]
  print(p_anova)
  if(any(p_anova<0.05)){
    xbest<-names(p_anova)[which.min(p_anova)]
    x_in<-as.character(data[,xbest])
    xall<-levels(factor(x_in))
    p_ttest<-1
    while(length(xall)>2&any(p_ttest>0.05)){
      p_ttest<-c()
      xpair<-combn(xall,2)
      for(i in 1:ncol(xpair))
      p_ttest[i]<-t.test(data$price[x_in==xpair[1,i]],data$price[x_in==xpair[2,i]],
data=data, var.equal=TRUE)$p.value
      names(p_ttest)<-paste(xbest,apply(xpair,2,toString),sep="_")
      if(any(p_ttest>0.05)){
        pairall<-which(p_ttest>0.05)
        pair_pmost<-which.max(p_ttest)
        pairmost<-xpair[,pair_pmost]
        pairall<-pairall[pairall!=pair_pmost]
        if(length(pairall)>0&nlevels(factor(x_in))>3){
          if(length(pairall)>1){
            xpairnew<-xpair[,pairall]
```

```
            torf<-apply(xpairnew,2,function(x)any(x%in%pairmost))
            gpfalse<-as.numeric(xpairnew[,!torf])
            gptnotf<-apply(xpairnew,2,function(x)!any(x%in%gpfalse))
            pairmost<-sort(unique(c(pairmost,as.vector(xpairnew[,gptnotf]))))
          } else if(any(xpair[,pairall]%in%pairmost)) pairmost<-
sort(unique(c(pairmost,xpair[,pairall])))
        }
        x_in[x_in%in%pairmost]<-toString(pairmost)
        xall<-levels(factor(x_in))
        p_ttest<-p_ttest[apply(xpair,2,function(x)!any(x%in%pairmost))]
      }
    }
    x_in<-paste(xbest,x_in,sep=":")
    print(tapply(data$price,x_in,function(x)c(n=length(x),mean=mean(x))))
    return(split(data[,which(names(data)!=xbest)],x_in))
  } else {
    print(rbind(n=nrow(data),mean=mean(data$price)))
    return(NULL)
  }
}
result<-anovaid(diamond)
if(class(result)=="list"){
  terminalnode<-result
  finaldv<-c()
  for(i1 in 1:length(terminalnode)){
    result<-anovaid(terminalnode[[i1]])
    if(class(result)=="list"){
      terminalnode[[i1]]<-result
      for(i2 in 1:length(terminalnode[[i1]])){
        result<-anovaid(terminalnode[[i1]][[i2]])
        if(class(result)=="list"){
```

```
          terminalnode[[i1]][[i2]]<-result
          for(i3 in 1:length(terminalnode[[i1]][[i2]])){
            result<-anovaid(terminalnode[[i1]][[i2]][[i3]])
            if(class(result)=="list"){
              terminalnode[[i1]][[i2]][[i3]]<-result
              for(i4 in 1:length(terminalnode[[i1]][[i2]][[i3]])){
                result<-anovaid(terminalnode[[i1]][[i2]][[i3]][[i4]])
                if(class(result)=="list"){
                  terminalnode[[i1]][[i2]][[i3]][[i4]]<-result
                  for(i5 in 1:length(terminalnode[[i1]][[i2]][[i3]][[i4]]))
                  finaldv<-rbind(finaldv,data.frame(y=terminalnode[[i1]][[i2]][[i3]][[i4]][[i5]]
[,"price"],infogp=paste(names(terminalnode)[i1],names(terminalnode[[i1]])[i2],
names(terminalnode[[i1]][[i2]])[i3],names(terminalnode[[i1]][[i2]][[i3]])[i4],
names(terminalnode[[i1]][[i2]][[i3]][[i4]])[i5],sep=":")))
                } else finaldv<-rbind(finaldv,data.frame(y=terminalnode[[i1]][[i2]][[i3]][[i4]]
[,"price"],infogp=paste(names(terminalnode)[i1],names(terminalnode[[i1]])[i2],
names(terminalnode[[i1]][[i2]])[i3],names(terminalnode[[i1]][[i2]][[i3]])[i4],sep=":")))
              }} else finaldv<-rbind(finaldv,data.frame(y=terminalnode[[i1]][[i2]][[i3]]
[,"price"],infogp=paste(names(terminalnode)[i1],names(terminalnode[[i1]])[i2],
names(terminalnode[[i1]][[i2]])[i3],sep=":")))
          }} else finaldv<-rbind(finaldv,data.frame(y=terminalnode[[i1]][[i2]][,"price"],
infogp=paste(names(terminalnode)[i1],names(terminalnode[[i1]])[i2],sep=":")))
      }} else finaldv<-rbind(finaldv,data.frame(y=terminalnode[[i1]][,"price"],
infogp=paste(names(terminalnode)[i1],sep=":")))
}}
tapply(finaldv[,1],finaldv[,2],mean)
table(finaldv[,2])
```

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวนวทิพย์ ไมตรี เกิดวันอังคารที่ 7 สิงหาคม พ.ศ. 2533 สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิชาคณิตศาสตร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2555 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาสถิติ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2556