

การวิเคราะห์การจำแนก (Discriminant Analysis)

น.อ.หญิง ยุวดี เปรมวิชัย
ผู้อำนวยการกองวิชาคณิตศาสตร์ ฝ่ายศึกษา โรงเรียนนายเรือ

ในปัจจุบันการวิจัยเริ่มมีการใช้วิธีการวิเคราะห์ข้อมูลโดยใช้สถิติขั้นสูง (Advance Statistics) กันมากขึ้น เช่น ใช้วิเคราะห์การถดถอยเชิงพหุ (Multiple Regression Analysis: MRA) สำหรับพยากรณ์ตัวแปรตามในกรณี มีตัวแปรอิสระหลายตัว เช่นพยากรณ์อัตราการเพิ่มของประชากรในเขตเมือง จากตัวแปรอิสระต่าง ๆ เช่นรายได้ประชาชาติ ความหนาแน่นของประชากร อัตราการย้ายถิ่น จำนวนธุรกิจในเมือง เป็นต้น ยังมีวิธีการทางสถิติขั้นสูงอีกวิธีหนึ่งที่น่าสนใจและมีการใช้ในการวิจัยมากขึ้น คือ “การวิเคราะห์การจำแนก (Discriminant Analysis : DA)” ซึ่งเป็นวิธีการที่ใช้พยากรณ์ได้ละเอียดยิ่งขึ้นกว่า การวิเคราะห์การถดถอยเชิงพหุ เพราะเป็นวิธีที่รวมวิธีการวิเคราะห์ความแปรปรวนเชิงพหุ (Multivariate Analysis of Variance) กับการวิเคราะห์การถดถอยเชิงพหุเข้าด้วยกัน อีกทั้งยังเหมาะสมกับข้อมูลที่มีตัวแปรตามมีมาตรวัดในระดับกลุ่ม (Nominal Scale) ที่เป็นข้อจำกัดที่การวิเคราะห์ การถดถอยเชิงพหุทำไม่ได้ เพราะต้องมีตัวแปรตามมีมาตรวัดในระดับช่วง (Interval Scale) เท่านั้น จึงทำให้ความนิยมในการใช้วิธีการวิเคราะห์การจำแนกมีมากขึ้น

แนวคิดและความจำเป็นในการใช้วิธีการวิเคราะห์การจำแนก

ข้อมูลที่ใช้การวิเคราะห์การจำแนก ต้องมีตัวแปรตามที่มีมาตรวัดในระดับกลุ่ม เช่น ธนาคารต้องการจำแนกลูกค้าออกเป็นลูกค้ากลุ่มที่มีเครดิตดี และกลุ่มเครดิตไม่ดีตัวแปรตามในกรณีนี้คือ ประเภทของลูกค้าธนาคาร เป็นตัวแปรประเภทกลุ่ม คือ ลูกค้ากลุ่มที่มีเครดิตดี และกลุ่มเครดิตไม่ดี จำนวน ๒ กลุ่ม

กลุ่มที่ ๑ คือ กลุ่มของลูกค้าธนาคารที่มีเครดิตดี อาจจะทำให้ค่าตัวแปรเป็น ๑

กลุ่มที่ ๒ คือ กลุ่มของลูกค้าธนาคารที่มีเครดิตไม่ดี อาจจะทำให้ค่าตัวแปรเป็น ๐ เป็นต้น

การตัดสินใจจัดกลุ่มลูกค้าเป็น ๒ กลุ่มแบบนี้ ธนาคารยังไม่มีหลักทางวิชาการ แต่ใช้ข้อมูลส่วนบุคคลของลูกค้า เช่น อายุ อาชีพ รายได้ ฯลฯ เป็นตัวแปรอิสระในการจัดกลุ่ม เพื่อประโยชน์ในการตัดสินใจใด ๆ เกี่ยวกับลูกค้ารายใหม่ที่เข้ามาใช้บริการ เพราะธนาคารย่อมอยากทราบว่ามีความเสี่ยงหรือไม่ในการให้บริการลูกค้ารายใหม่นี้ หากลูกค้ารายใหม่อยู่ในกลุ่มที่มีเครดิตไม่ดี ธนาคารก็จะรู้ถึงความเสี่ยง (Risk) ล่วงหน้าได้ ทำให้สามารถตัดสินใจอนุมัติหรือไม่อนุมัติการให้บริการได้อย่างมีประสิทธิภาพ

หลักการของการวิเคราะห์การจำแนก จึงต้องเริ่มจากมีข้อมูลเดิมที่รู้กลุ่ม (Group) อยู่แล้ว นำมาศึกษาจนสามารถสร้างฟังก์ชันหรือสมการของตัวแปรตามและตัวแปรอิสระเพื่อเก็บไว้ใช้ตัดสินใจจัดกลุ่มให้

ข้อมูลใหม่หรือค่าสังเกตตัวใหม่ที่ยังไม่ทราบกลุ่ม ให้เข้าอยู่ในกลุ่มที่เหมาะสมได้ เพื่อประโยชน์ในการดำเนินการที่เกี่ยวข้องต่อไป ซึ่งการจัดกลุ่มนี้ก็คือการพยากรณ์ลักษณะข้อมูลใหม่ที่เรายังไม่ทราบแน่นอนนั่นเอง มีหลายเหตุผลที่ทำให้เราจำเป็นต้องพยากรณ์กลุ่มของข้อมูลใหม่ จนทำให้ต้องพยายามจัดกลุ่มหรือจำแนกกลุ่มข้อมูล ดังนี้

- (๑) เพื่อทำนายเหตุการณ์ในอนาคต เช่นกรณีตัวอย่างดังนี้
 - ธนาคารจำแนกลูกค้ากลุ่มที่เครดิตดี และกลุ่มที่เครดิตไม่ดี
 - การศึกษาชนิดของนกที่พบว่าเป็นนกชนิดใด ชนิดที่ ๑ หรือ ๒ หรือ ๓
 - การจัดนักศึกษาเข้าศึกษาในสาขาวิชาต่าง ๆ ที่เหมาะสมกับความสามารถเพราะคาดว่าในอนาคตนักศึกษาควรต้องสำเร็จการศึกษาในสาขานั้นกลุ่มของนักศึกษาคือ เรียนสำเร็จ และไม่สำเร็จ เป็นต้น
- (๒) เพื่อใช้ประโยชน์จากการพยากรณ์ขณะนั้น เพราะถ้าคอยผลลัพธ์ในอนาคต จะเกิดความสูญเสียเกิดความเสียหายต่อทรัพย์สินหรือชีวิตอย่างร้ายแรง เช่น
 - ยานชนิดอันตราย หรือไม่อันตราย
 - เครื่องยนต์ที่ประกอบด้วยระบบการทำงานนี้เมื่อใช้ไปจะระเบิดหรือไม่ระเบิด เป็นต้น
- (๓) เพื่อใช้ตัดสินใจว่าข้อมูลควรอยู่ในกลุ่มใดเนื่องจากไม่สามารถหาหลักฐานข้อเท็จจริงอื่นได้อีกแล้ว เช่น
 - การพิสูจน์โครงกระดูกของศพนิรนามที่พบว่าเป็นเพศหญิง หรือ ชาย
 - การค้นพบโบราณวัตถุว่าจะเป็นโบราณวัตถุสมัยสุโขทัย หรือสมัยกรุงศรีอยุธยา หรือสมัยรัตนโกสินทร์ เป็นต้น

วัตถุประสงค์ของการวิเคราะห์การจำแนก

การวิเคราะห์การจำแนกเป็นการวิเคราะห์เชิงพหุ (Multivariate Analysis) ที่จะศึกษาความสัมพันธ์ ของตัวแปรตามที่มีลักษณะแบ่งเป็นกลุ่ม จากหลาย ๆ ประชากร กับตัวแปรอิสระตั้งแต่ ๒ ตัวขึ้นไป เพื่อจัดกลุ่มข้อมูลที่มีอยู่แล้วให้แยกออกตามลักษณะที่สำคัญของแต่ละกลุ่ม และหากมีข้อมูลใหม่หรือค่าสังเกตใหม่เพิ่มเข้ามาก็ต้องมีสมการที่เหมาะสมในการจัดหรือจำแนก (Discriminant) ค่าสังเกตใหม่นั้นให้เข้ากับกลุ่มต่าง ๆ ที่มีอยู่ได้อย่างเหมาะสม ผู้คิดวิธีนี้คือ R.A.Fisher

การวิเคราะห์การจำแนกจึงมีวัตถุประสงค์คือ

- (๑) Separating หรือ Grouping คือการจัดกลุ่มข้อมูลเดิมที่มีอยู่แล้วให้สามารถอธิบายมิติทางคณิตศาสตร์ (Graphically หรือ Algebraically) ของค่าสังเกตที่แตกต่าง กันจากหลาย ๆ ประชากร โดยสร้างสมการเพื่อจำแนกค่าทางคณิตศาสตร์ของแต่ละกลุ่ม อยู่ในรูป Linear Combination

$$Y_k = v_1 X_1 + v_2 X_2 + \dots + v_p X_p$$

เมื่อ Y เป็น ตัวแปรอิสระระดับกลุ่มที่มี k กลุ่ม ตัวแปรอิสระ p ตัว
 $X_1, X_2, \dots, X_p$ เป็น ตัวแปรอิสระหรือ Predictor หรือลักษณะข้อมูล p
 ลักษณะที่นำมาศึกษา
 v_i เป็น ค่าสัมประสิทธิ์หรือน้ำหนัก (Weight) ของ x_i เมื่อ

$$i = 1, 2, \dots, p$$

จึงเรียกขั้นตอนนี้ว่าการจำแนก (Discrimination)

- (๒) Classification หรือ Allocation คือการจัดให้ข้อมูลหรือค่าสังเกต (Observation) ที่เข้ามาใหม่แต่ไม่รู้ค่า ค่าสังเกตนั้นควรอยู่ในกลุ่มใด ให้เข้าอยู่ในกลุ่มที่เหมาะสมโดยใช้ Discriminant function ที่ได้จากขั้นตอนที่ ๑ เป็นเครื่องมือในการจัดกลุ่มค่าสังเกตใหม่ ขั้นตอนนี้จะเกี่ยวข้อง กับความน่าจะเป็นด้วย

ในทางสถิติให้ความสำคัญกับขั้นตอน Separating มาก เพราะหากได้ Discriminant function ที่ไม่ถูกต้อง การนำไป Classification ต่อไปก็จะเกิดผลเสียตามมา การทดสอบทางสถิติที่เกี่ยวข้องในขั้นตอน Separating สรุปได้ดังนี้

- สร้าง Discriminant functions ในรูป Linear Combination ของตัวแปรอิสระหรือ Predictor จากข้อมูลเดิมที่มีอยู่แล้ว
- ทดสอบและคัดเลือกตัวแปรอิสระเข้ามาในสมการ Discriminant functions ว่าตัวแปรที่ใช้เหมาะสมหรือไม่ และประมาณค่าสัมประสิทธิ์ให้มีอำนาจการจำแนกสูงสุด โดยใช้วิธี Stepwise ที่มีหลักการเหมือนการวิเคราะห์การถดถอยเชิงพหุ เพื่อคัดเลือกตัวแปรอิสระตัวใดจำแนกกลุ่มได้ดีที่สุด
- ทดสอบความแตกต่างระหว่างกลุ่ม (Between Group) อย่างมีนัยสำคัญ เพื่อแสดงว่า Discriminant functions ที่ได้สามารถจำแนกกลุ่มออกจากกันได้อย่างมีนัยสำคัญ นิยมใช้ค่าสถิติที่ใช้เรียก Wilks' lambda (λ) และมีการกระจายแบบ chi square (χ^2)
- ตรวจสอบการจำแนกข้อมูลเดิม ว่าข้อมูลเดิมควรอยู่ในกลุ่มใดจึงดีที่สุด ผลการตรวจสอบจะแสดงในรูปร้อยละ (%) ของการจำแนกได้ถูกต้อง เรียกว่า ดัชนีประสิทธิภาพ ซึ่งควรมีร้อยละของการจำแนกได้ถูกต้องตั้งแต่ ๘๕ % ขึ้นไป
- เมื่อมีข้อมูลใหม่ สามารถจำแนกเข้ากลุ่มได้อย่างเหมาะสมมีเหตุผลเชิงสถิติ
- การทดสอบทั้งหมดของการวิเคราะห์การจำแนกอยู่ภายใต้ Assumptions ว่า ทุกประชากรมีการแจกแจงแบบปกติ (Multivariate Normally distribution) ที่มีความแปรปรวนร่วม (Population Covariance Matrices) เท่ากัน

วิธีการวิเคราะห์การจำแนก

วัตถุประสงค์ที่สำคัญคือ ต้องสร้าง Discriminant functions ที่มีความแตกต่างของความแปรปรวนระหว่างกลุ่ม (Between Group) ความแปรปรวนภายในกลุ่ม (Within Group) ให้มากที่สุด จากหลักการพื้นฐาน ว่าสมการ Discriminant functions ต้องจำแนกกลุ่มได้ดีที่สุด เราจึงต้องการหาค่า λ ที่มีนัยสำคัญทางสถิติ โดยที่

$$\lambda = \frac{\text{sum square between Group of } Y}{\text{sum square within Group of } Y}$$

$$= \frac{ss_b(Y)}{ss_w(Y)}$$

เรียก λ ว่า Eigenvalue ซึ่งคือ Variance ของ Y นั้นเอง นิยมเรียก ว่า Wilks' Lambda ซึ่งสามารถใช้วิธีการของ matrice หาค่าได้ ดังนี้

สมมติมีตัวแปรอิสระ p ตัว คือ $x_1, x_2, \dots, x_p$ ต้องจัดกลุ่ม k กลุ่ม

เช่น ลูกค้าธนาคาร เป็นตัวแปรประเภทกลุ่ม คือ ลูกค้ากลุ่มที่มีเครดิตดี และกลุ่มเครดิตไม่ดี $k = 2$

สังเกตลูกค้าจากรายได้ ขนาดครอบครัว ระดับการศึกษา จำนวนปีที่ทำงานมาแล้ว

$$X = \begin{bmatrix} x_1 = \text{รายได้} \\ x_2 = \text{ขนาดครอบครัว} \\ x_3 = \text{ระดับการศึกษา} \\ x_4 = \text{จำนวนปีที่ทำงานมาแล้ว} \end{bmatrix} \quad \text{ได้ } p = 4$$

จะได้จำนวนสมการ Discriminant functions จำนวน ๑ สมการ จาก $r = 1$ คือ

$$r = \min[k - 1, p] = \min[1, 4] = 1$$

$$\text{ดังนั้น matrix ของ } x \text{ คือ } X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_p \end{bmatrix}_{(p \times 1)}$$

$$V = [v_1, v_2, v_3, \dots, v_p] \quad , \quad X' = [x_1, x_2, x_3, \dots, x_p]$$

ต้องทำการหา matrix V_m ที่ทำให้ $Y_m = V_m X$ เมื่อ $m = 1, 2, \dots, r$

$$\text{จากค่าสังเกต } Y_i = x_i v_i$$

โดย V ยังไม่ทราบค่า และ $i = 1, 2, \dots, n$

เมื่อ n คือจำนวนข้อมูลที่ทำการศึกษา

$SS(X)$ หรือ Sum Square $x = (x - \bar{x})(x - \bar{x})$ หรือ $XX - \bar{X}\bar{X}$

$SS(Y)$ หรือ Sum square $Y = (Y - \bar{Y})(Y - \bar{Y})'$

แทนค่า $Y = XV$ และ $\bar{Y} = \bar{X}V$ จะได้

$$\begin{aligned} SS(Y) &= V'X'XV - V'\bar{X}'\bar{X}V \\ &= V'(X'X - \bar{X}'\bar{X})V \\ &= V'SS(X)V \\ &= \text{Sum Square Total ของ } Y \end{aligned}$$

ถ้า $SS_w(x) = s_1 + s_2 + \dots + s_k = (x - \bar{x}_g)(x - \bar{x}_g)' = w$

$SS_b(x) = (\bar{x}_g - \bar{x} \dots)(\bar{x}_g - \bar{x} \dots)' = B$

$SS_T(x) = (x - \bar{x} \dots)(\bar{x} - \bar{x} \dots)' = T$

ทำนองเดียวกันสามารถหา $SS_w(Y)$ ได้จากที่มี k กลุ่ม

$$\begin{aligned} SS_w(Y) &= SS_1(Y) + SS_2(Y) + \dots + SS_k(Y) \\ &= VS_1(x)V + V'S_2(x)V + \dots + V'S_k(x)V \\ &= V'(S_1 + S_2 + S_3 + \dots + S_k)V \\ &= V'SS_w(x)V = V'WV \end{aligned}$$

และ $SS_b(Y) = V'BV$

จะสามารถหา $\lambda = \frac{SS_b(Y)}{SS_w(Y)}$ ได้

ต่อจากนั้น ต้องทดสอบว่า λ ที่ได้จะมีค่ามากน้อยอย่างมีนัยสำคัญหรือไม่ ทดสอบต่อไปโดยให้

$$\begin{aligned} \lambda &= \frac{SS_b(Y)}{SS_w(Y)} \\ &= \frac{V'BV}{V'WV} \end{aligned}$$

จะได้ $V'BV = \lambda V'WV$

$$\begin{aligned} BV &= \lambda WV \\ BV - \lambda WV &= 0 \\ [B - \lambda W]V &= 0 \end{aligned}$$

$$[W^{-1}B - \lambda I]V = 0$$

เรียก $|W^{-1}B - \lambda I|$ ว่า Characteristic Equation

ซึ่ง $|W^{-1}B - \lambda I| = 0$ เสมอ

นำ Characteristic Equation ไปหาค่า λ เมื่อรู้ค่า λ นำไปหา V ได้ต่อไป

matrix ของ V จะมีค่า V อยู่ m ตัว โดย $m = k - 1$

โดย V_m คือค่าสถิติในการทดสอบความแตกต่างระหว่างกลุ่มของ Discriminant functions ที่ต้องนำไปทดสอบว่ามีนัยสำคัญหรือไม่ ซึ่งมีหลายวิธี ในที่นี้จะกล่าวถึงวิธี Bartlett

ให้ $v_m = \text{Bartlett Statistics}$ มีการแจกแจง χ^2

โดยมีสูตร $v_m = [N - 1 - (p + k) / 2] \ln(1 + \lambda_m)$

เมื่อทดสอบได้ว่า V_m มีนัยสำคัญทางสถิติ ก็ะหมายถึงสมการที่ m ของ Discriminant functions นั้นมีความสามารถในการจำแนกได้อย่างมีนัยสำคัญ

ปัจจุบันในการคำนวณของการวิเคราะห์การจำแนกมีโปรแกรมสำเร็จรูปทางสถิติให้เลือกใช้ เช่น โปรแกรม SPSS หรือ โปรแกรม SAS ซึ่งสามารถกำหนดรายละเอียดได้ถึงความน่าจะเป็นของข้อมูลในการเป็นสมาชิกของแต่ละกลุ่ม (Prior Probability) ตลอดจนสามารถเลือกวิธีทดสอบได้หลายวิธีแตกต่างกันไป

ตัวอย่างงานวิจัยที่ใช้วิธีการวิเคราะห์การจำแนก

ชื่อเรื่อง บริการธนาคารผ่านอินเทอร์เน็ต ลักษณะสำคัญของผู้สมัครและปัจจัยที่มีผลต่อการสมัครใช้บริการอินเทอร์เน็ตแบงกิ้ง (Internet banking service)

ผู้วิจัย ดร.ชัชพงศ์ ตั้งมณี และ กิตติศักดิ์ ตันฤดี

วัตถุประสงค์การวิจัย

๑. เพื่อรายงานลักษณะสำคัญ (Profile) ของผู้สมัครและผู้ไม่สมัครใช้บริการอินเทอร์เน็ตแบงกิ้ง
๒. เพื่อนำเสนอความคาดหวัง (Expectation) ของ ผู้สมัครใช้บริการและผู้ไม่สมัครใช้บริการ
๓. เพื่อวิเคราะห์การจำแนก (Discriminant analysis) ของผู้สมัครและผู้ไม่สมัครใช้อินเตอร์เน็ตแบงกิ้งด้วยข้อมูลส่วนบุคคล (Demographic factors)

การวิเคราะห์ข้อมูล

- ใช้สถิติพรรณนา (Descriptive Statistics) เพื่อรายงานลักษณะสำคัญ (Profile) ของผู้สมัครและผู้ไม่สมัคร
- ใช้การวิเคราะห์การจำแนกเพื่อจำแนกลูกค้าว่าจะสมัครหรือไม่สมัครใช้บริการอินเทอร์เน็ตแบงกิ้ง โดยมีตัวแปรตามคือ การสมัครและไม่สมัครใช้บริการ และชุดตัวแปรอิสระคือตัวแปรข้อมูลส่วนบุคคลที่ธนาคารมีอยู่

ผลการวิเคราะห์ข้อมูลในส่วนของการวิเคราะห์การจำแนก

- Profile ของผู้สมัครใช้บริการอินเทอร์เน็ตแบงกิ้งประมาณครึ่งหนึ่งของผู้สมัครใช้บริการอินเทอร์เน็ตแบงกิ้งเป็นหญิงโสด มีอายุไม่เกิน ๒๕ ปี เพิ่งจบการศึกษาระดับปริญญาตรี ทำงานบริษัทเอกชน มีรายได้ไม่เกิน ๒๐,๐๐๐ บาทต่อเดือน ใช้อินเตอร์เน็ตมาแล้ว ไม่เกิน ๒ ปี และเคยแวะชมเว็บไซต์ของธนาคารมาแล้ว

