บทที่ 1

บทนำ

1.1 ความสำคัญและความเป็นมาของปัญหา

การวิเคราะห์การแบ่งกลุ่ม (Discriminant Analysis) เป็นเทคนิคที่มุ่งจำแนกวัตถุ (object or individual) ออกเป็นกลุ่มๆ ตามธรรมชาติที่วัตถุนั้นอาจรวมเข้าเป็นหมู่พวกเดียวกันได้ เมื่อมีวัตถุใหม่เพิ่มเข้ามาก็เป็นภารกิจของการจัดวัตถุเข้ากลุ่ม (classification) ที่จะจัดการนำวัตถุนั้นเข้าพวกกับกลุ่มใดกลุ่มหนึ่ง ในทางปฏิบัติเหตุการณ์ที่สนใจศึกษามักจะเป็นการจำแนกวัตถุออกเป็น 2 กลุ่ม เช่น ผู้ที่มีเครดิตดีกับไม่ดี การซื้อและไม่ซื้อสินค้าของผู้บริโภค บริษัทที่มีแนวโน้มล้มละลายกับมั่นคง ผู้ที่เข้าข่ายเป็นโรคเอดส์กับไม่เป็น ผู้มีปัญหาทางกระเพาะอาหารกับไม่มีปัญหา ในแต่ละวิธีของการสร้างกฎจำแนกกลุ่มขึ้นมาใช้มีโอกาสที่จะเกิดความผิดผลาด 2 แบบคือ

1. ผู้วิจัยแบ่งกลุ่มวัตถุที่มาจากกลุ่มประชากรที่ 1 ให้กับกลุ่มประชากรที่ 2
2. ผู้วิจัยแบ่งกลุ่มวัตถุที่มาจากกลุ่มประชากรที่ 2 ให้กับกลุ่มประชากรที่ 1

ดังนั้นจึงมีวิธีการในการประเมินคุณภาพของกฎการจัดเข้ากลุ่มเพื่อวัดดูว่ากฎการจัดเข้ากลุ่มที่ได้มานี้สามารถนำเอาไปใช้ได้ด้วยความเสี่ยงมากน้อยเพียงใด ซึ่งปัญหาของการประมาณอัตราความผิดพลาด (error rate) ในการวิเคราะห์การจำแนกกลุ่มได้เป็นที่สนใจของนักวิจัยมาตั้งแต่ ค.ศ. 1930 โดยมีงานวิจัยที่เกี่ยวข้องดังนี้

- ค.ศ.1936 Fisher ได้พัฒนาวิธีการทางสถิติในเรื่องของการวิเคราะห์การจำแนกกลุ่ม (Discriminant Analysis) และ classification เป็นคนแรก โดยเสนอตัวประมาณอัตราความผิดผลาดในการจำแนกกลุ่มผิดตัวแรกคือ Plug-in estimator หรือ D method ซึ่งเป็นวิธีการที่ง่ายมีข้อตกลงเบื้องต้นคือการแจกแจงของประชากรเป็นการแจกแจงแบบปกติ แต่วิธีนี้มีความเอนเอียง (biased) มากถ้าขนาดตัวอย่างไม่ใหญ่พอ

- ค.ศ.1947 Smith ได้เสนอตัวประมาณอีกวิธีหนึ่งคือ Resubstitution estimator หรือ R method เป็นวิธีการที่ง่ายต่อการคำนวณและขั้นตอนไม่ยุ่งยาก จึงมีการใช้วิธี R กันมาก แต่อย่างไรก็ตามวิธี R มีความเอนเอียงเช่นเดียวกับวิธี D เมื่อขนาดตัวอย่างไม่ใหญ่พอ แต่ข้อดีของวิธี R คือไม่มีข้อจำกัดใดๆเกี่ยวกับการแจกแจงของประชากร

- ค.ศ.1967 Lachenbruch และ Mickey ได้เสนอตัวประมาณชื่อ Leave-one-out estimator หรือ U method เพื่อจะแก้ไขความเอนเอียงที่เกิดขึ้นในวิธี R นอกจากนี้ยังได้ปรับปรุงวิธี D ให้มีประสิทธิภาพมากขึ้น จึงเสนอ DS method แต่ต้องอยู่ภายใต้การแจกแจงแบบปกติ

- ค.ศ.1979 Efron ได้เสนอวิธีการที่เรียกว่า bootstrap นำมาใช้ในการประมาณค่าความน่าจะเป็นในการจำแนกกลุ่มผิด วิธีนี้ไม่มีเงื่อนไขเกี่ยวกับการแจกแจงของประชากร

- ค.ศ.1985 Page ได้ศึกษาตัวประมาณในกลุ่มพาราเมตริกและได้สรุปว่าตัวประมาณอัตราความผิดพลาดวิธี DS จะนำเอาไปใช้ได้ดีเมื่อขนาดของตัวแปรมีขนาดเล็กและขนาดกลางและขนาดตัวอย่างต้องมีขนาดใหญ่พอ

ปัญหาที่พบในการประมาณอัตราความผิดผลาดในการวิเคราะห์การจำแนกกลุ่มก็คือตัวประมาณแต่ละวิธีมีข้อจำกัดในการนำเอาไปใช้ ซึ่งปัจจัยที่เกี่ยวข้องกับประสิทธิภาพของตัวประมาณคือขนาดตัวอย่างบางครั้งมีปัญหาในเรื่องจำนวนข้อมูลหรือขนาดของข้อมูลมีขนาดเล็ก ลักษณะการแยกจากกันของกลุ่มประชากรบางครั้งประชากรมีส่วนที่คาบเกี่ยว (overlap) กันมาก ขนาดของตัวแปรอิสระที่ใช้อธิบายลักษณะกลุ่มประชากรเป็นต้น ส่วนลักษณะการแจกแจงของกลุ่มประชากรนั้นโดยส่วนใหญ่นักวิจัยจะให้อยู่ภายใต้ข้อตกลงเบื้องต้นของการแจกแจงปกติ จากการที่ได้มีนักวิจัยเสนอวิธีการประมาณอัตราความผิดพลาดในการวิเคราะห์การจำแนกกลุ่มขึ้นมามากมายหลายตัวด้วยกันทั้งตัวประมาณแบบพาราเมตริกและตัวประมาณแบบนอนพาราเมตริก ดังนั้นในงานวิจัยนี้จึงสนใจศึกษาตัวประมาณที่มีการใช้กันมากและเป็นที่สนใจของนักวิจัยโดยทั่วๆ ไปดังนี้

1. วิธี R หรือ Resubstitution Estimator
2. วิธี U หรือ Leave-one-out Estimator
3. วิธี B หรือ Bootstrap Estimator
4. วิธี DS หรือ Shrunken-D Estimator

1.2 วัตถุประสงค์ของการวิจัย

เพื่อเปรียบเทียบตัวประมาณอัตราความผิดพลาดในการวิเคราะห์การจำแนกกลุ่ม จำแนกตามขนาดตัวอย่าง ลักษณะการแยกจากกันของกลุ่มประชากร และ ขนาดของตัวแปรอิสระ เมื่อใช้วิธี

- Resubstitution estimator หรือ วิธี R
- Leave-one-out estimator หรือ วิธี U
- Bootstrap estimator หรือ วิธี B
- Shrunken-D Estimator หรือ วิธี DS

โดยพิจารณาจากค่าความคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุด

1.3. สมมติฐานของการวิจัย

โดยทั่วไปตัวประมาณอัตราความผิดพลาดเมื่อใช้วิธี DS ดีที่สุดยกเว้นในกรณีที่ค่าเฉลี่ยของประชากรแต่ละกลุ่มแตกต่างกันน้อย ตัวประมาณอัตราความผิดพลาดเมื่อใช้วิธี B จะดีที่สุด

1.4. ข้อตกลงเบื้องต้น

1.4.1 ในการวิจัยนี้จะถือว่า ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเป็นเกณฑ์ในการเลือกตัวประมาณ

1.4.2 การแจกแจงของประชากรเป็นการแจกแจงแบบปกติ

1.4.3 ข้อมูลที่ใช้ศึกษาเป็นข้อมูลเชิงปริมาณ

1.5 ขอบเขตการวิจัย

1.5.1 ขนาดตัวอย่างที่ใช้ในการวิจัย กำหนดให้ $n_1 = n_2$ มีขนาดดังนี้ 10, 20, 25, 50, 100

1.5.2 ความน่าจะเป็นที่วัตถุนั้นจะเป็นสมาชิกของประชากรกลุ่มที่ i (prior probability) คือ q_i และความสูญเสียในการจำแนกกลุ่มผิด (cost of misclassification) ในกลุ่มที่ 1 และกลุ่มที่ 2 คือ c(1|2) และ c(2|1) ตามลำดับ มีค่าเท่ากันทั้งสองกลุ่ม

1.5.3 ในการวิจัยนี้ใช้ Anderson's Classification Function (Statistic) ดังนี้

$$W(x) = \{ X-(\bar{X}_1+\bar{X}_2)/2 \}' S^{-1}(\bar{X}_1-\bar{X}_2)$$

กฎที่ใช้จำแนกกลุ่ม คือ

จัด X ให้กลุ่ม 1 ถ้า $W(x) \geq K$
จัด X ให้กลุ่ม 2 ถ้าเป็นอย่างอื่น

เมื่อ

$$K = \ln[\, c(1|2).q_2 \,/\, c(2|1).q_1 \,]$$

1.5.4 ในการวิจัยครั้งนี้ ศึกษาการประมาณอัตราความผิดพลาดชนิดที่มีเงื่อนไข (Conditional Error Rate) คือความน่าจะเป็นที่ตัวอย่างสุ่มจากประชากรกลุ่มที่ 1 ถูกจัดกลุ่มให้ผิดเมื่อใช้กฎการจำแนกกลุ่มที่สร้างขึ้น แสดงได้ดังนี้

$$P(2|1) = Pr\{ W(x) \leq 0 \mid \theta \}$$

$$= \Phi\left[\frac{-\{\mu_1 - (\bar{x}_1+\bar{x}_2)/2\}' s^{-1}(\bar{x}_1-\bar{x}_2)}{\{(\bar{x}_1-\bar{x}_2)' s^{-1} \Sigma\, s^{-1} (\bar{x}_1-\bar{x}_2)\}^{1/2}} \right]$$

$\Phi(t)$ หมายถึง ฟังก์ชั่นการแจกแจงปกติมาตราฐาน ที่ t (Standard Normal Distribution Function)

1.5.6 ค่า Square Root of Mahalanobis distance (Δ)
มีค่าดังนี้ 0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0

1.5.7 ตัวแปรอิสระ X ที่ศึกษามีขนาด 3 , 5 , 7 , 9

1.5.8 การแจกแจงของประชากรกลุ่มที่ 1 เป็นดังนี้ $N(\mu_1, \Sigma)$
การแจกแจงของประชากรกลุ่มที่ 2 เป็นดังนี้ $N(\mu_2, \Sigma)$
กำหนดให้ $\mu_1 = (0, 0, 0, \ldots, 0)$
$\mu_2 = (\Delta, 0, 0, \ldots, 0)$
และ
$\Sigma = I$

1.5.9 การวิจัยนี้สร้างแบบจำลองข้อมูลให้มีสถานการณ์ตามต้องการ โดยใช้เทคนิคมอนติคาร์โลซิมูเลชั่นจากเครื่องคอมพิวเตอร์ IBM 370/3031 เขียนโปรแกรมด้วยภาษาFortran 77 โดยจะทำการทดลองซ้ำ 500 ครั้งในแต่ละสถานการณ์

1.6 ประโยชน์ที่คาดว่าจะได้รับ

เพื่อเป็นแนวทางให้นักวิจัยได้เลือกใช้ตัวสถิติที่เหมาะสม สำหรับการประมาณอัตราความผิดพลาดในการวิเคราะห์การจำแนกกลุ่ม

1.7 ตัวประมาณอัตราความผิดพลาด

1.7.1 Resubstitution estimstor หรือ วิธี R

สัญลักษณ์ที่ใช้คือ $\hat{\alpha}^R$ กำหนดให้ x_j $(j = 1,2,..,n_1)$ แทนตัวอย่างสุ่มจากประชากรกลุ่มที่ 1 และให้ $h(x)$ เป็นฟังก์ชั่นนับ (counting function) ดังนี้

$$h(x) = \begin{cases} 1 & \text{ถ้า } W(x) \leqslant 0 \\ 0 & \text{ถ้าเป็นอย่างอื่น} \end{cases}$$

ดังนั้น

$$\hat{\alpha}^{R} = \sum_{i=1}^{n_1} h(x_j)/n_1$$

1.7.2 Leave-one-out estimator หรือ วิธี U

สัญลักษณ์ที่ใช้คือ $\hat{\alpha}^{U}$ วิธีนี้ แนะนำให้ตัดค่าสังเกตหนึ่งออกจากกลุ่มตัวอย่างกลุ่มที่ 1 แล้วสร้าง classification rule จากข้อมูลที่เหลือคือ $n_1 - 1, n_2$ จากนั้นจึงนำค่าสังเกตที่กักไว้หรือที่ตัดออกไป (holdout) นำมาจัดเข้ากลุ่ม ถ้าเข้ากลุ่มผิดให้บันทึกไว้ ดำเนินการเช่นนี้จนครบทุก ๆ หน่วย กำหนดฟังก์ชั่นนับเป็น ดังนี้

$$h(x_j) = \begin{cases} 1 & \text{ถ้า } W(x) < 0 \\ 0 & \text{ถ้าเป็นอย่างอื่น} \end{cases}$$

ดังนั้น

$$\hat{\alpha}^{U} = \sum_{i=1}^{n_1} h^{(j)}(x_j)/n_1$$

1.7.3 Bootstrap estimator หรือ วิธี B

เนื่องจาก Efron ได้เสนอตัวประมาณอัตราความผิดพลาดโดยวิธี B ดังนี้คือ $\hat{\alpha}^{B} = \hat{\alpha}^{R} + b_1$ ซึ่ง b_1 คือค่าความเอนเอียงที่ผิดพลาด (bias correction) ไปจากวิธี R ดังนั้นจึงใช้วิธีการของ bootstrap มาประมาณค่า b_1

ขั้นตอนการประมาณค่า b_1 มีดังนี้

1. สุ่มตัวอย่างใหม่จากกลุ่มตัวอย่างเดิม 2 กลุ่ม ให้สัญลักษณ์เป็น H^{*}_1 และ H^{*}_2 โดยแต่ละค่าสังเกตใหม่มีโอกาสถูกเลือกเท่าๆ กันคือ $1/n_1$ ซึ่งเป็นการสุ่มแบบใส่คืน (with replacement)

2. นำข้อมูลที่ได้ใหม่มาคำนวนหาดิสคริมิแนนท์ฟังก์ชั่น และกฎการจำแนกกลุ่มได้ $W^*(x)$ จาก H^*_1 และ H^*_2 ซึ่งขั้นตอนการหา $W^*(x)$ เหมือนกับการหา $W(x)$

3. หาผลต่างของ $d = A^{**}_1 - A^*_1$

เมื่อ

A^{**}_1 คือ อัตราส่วนระหว่างจำนวนสมาชิกในกลุ่มตัวอย่างเดิมที่ถูกจำแนกกลุ่มผิดโดยใช้ $W^*(x)$ กับขนาดตัวอย่างของกลุ่มที่ 1 (เนื่องจากพิจารณากลุ่มที่ 1)

A^*_1 คือ อัตราส่วนระหว่างจำนวนสมาชิกในกลุ่มตัวอย่างใหม่ที่ถูกจำแนกกลุ่มผิดโดยใช้ $W^*(x)$ กับขนาดตัวอย่างของกลุ่มที่ 1

4. ค่าคาดหวังของ d คือค่าประมาณของ b_1 ดังนั้น $\hat{b}_1 = \bar{d}$ การหา $\bar{d}$ ก็คือการเฉลี่ยค่า d ประมาณ 50 - 200 ครั้ง นั่นคือต้องทำขั้นตอนที่ 1 - 3 เป็นจำนวน 50-200 ครั้ง ดังนั้นตัวประมาณโดยวิธี B ที่ได้คือ

$$\alpha^B = \alpha^R + \hat{b}_1$$

1.7.4 Shrunken-$\bar{D}$ Estimator หรือ วิธี DS

$$\alpha^{DS} = \Phi\{ (- DS/2)(DS)^{1/2} \}$$

โดยที่

$$DS = (n_1+n_2-k-3)D^2/(n_1+n_2-2)$$
$$D^2 = (\bar{x}_1-\bar{x}_2)'S^{-1}(\bar{x}_1-\bar{x}_2)$$
$$k = \text{จำนวนตัวแปรอิสระ}$$