การวิเคราะห์จำแนกกลุ่ม (Cluster Analysis)

มนิต พลหลา
โปรแกรมวิชาสถิติประยุกต์

การวิเคราะห์จำแนกกลุ่ม (Cluster Analysis) เป็นการวิเคราะห์สถิติขั้นสูงถูกนำมาใช้ครั้งแรกโดย Tryon (1939) การวิเคราะห์จำแนกกลุ่มเป็นการจัดกลุ่มของวัตถุ (object) ให้อยู่เป็นกลุ่ม (cluster) [4] และมีวัตถุประสงค์เพื่อจัดกลุ่มของข้อมูล (cases) หรือตัวแปร (variables) โดยให้แต่ละกลุ่มมีลักษณะหรือคุณสมบัติที่เหมือนกันหรือมีความสัมพันธ์กัน (homogeneous groups of cases or variables)[5] และตัวแปรที่รวมอยู่ในกลุ่มเดียวกันจะต้องมีความสัมพันธ์ไปในทิศทางเดียวกันซึ่งต่างจากการวิเคราะห์ปัจจัย (Factor Analysis) ที่การรวมกลุ่มของตัวแปรจะใช้ความสัมพันธ์ของตัวแปรทั้งตัวแปรที่สัมพันธ์กันทั้งทางบวกและทางลบ[10] ทำให้สามารถจะลดจำนวนข้อมูล (reducing the number of cases)[4] ลดจำนวนตัวแปร (reducing the number of variables) หรือหาความสัมพันธ์ระหว่างตัวแปร (relationship between variables)ได้[5] แต่โดยส่วนใหญ่แล้วจะใช้เทคนิคการวิเคราะห์จำแนกกลุ่มในการจัดกลุ่มข้อมูลมากกว่าการจัดกลุ่มตัวแปร เพราะการจัดกลุ่มตัวแปรจะใช้กับเทคนิคการวิเคราะห์ปัจจัย (Factor Analysis)[11]

Hartigan (1975) ได้รวบรวมการนำเทคนิคการวิเคราะห์จำแนกกลุ่มไปใช้ในด้านต่างๆ เช่น แพทย์ใช้ในการจำแนกกลุ่มของโรค (clustering diseases) การจำแนกกลุ่มของวิธีรักษา (cures for diseases) หรือการจำแนกกลุ่มอาการของผู้ป่วย (symptoms of diseases) การจำแนกกลุ่มอาการของผู้ป่วยทางจิต หรือ Alan (2002) ใช้การวิเคราะห์จำแนกกลุ่มเพื่อแยกสปีชี่ส์ของสัตว์ (Alan, 2002) และการจำแนกกลุ่มของสัตว์ไม่มีกระดูกสันหลัง (Stribling, et al., 1998) เป็นต้น

หลักในการวิเคราะห์

1. คัดเลือกตัวแปรที่เหมาะสมสำหรับใช้จัดกลุ่มของข้อมูล
2. เลือกวิธีการวิเคราะห์ข้อมูลโดยวิธีที่ใช้กันอย่างแพร่หลายคือ การจัดกลุ่มแบบเชิงชั้น (Hierarchical Cluster Analysis)[10] โดยแบ่งออกเป็น 2 วิธีคือ วิธีการจัดกลุ่มโดยวิธีรวมเข้า (Agglomerative) และวิธีจัดกลุ่มโดยวิธีการแยกออก (Divisive) และสร้าง Dendogram เพื่อวัดระยะห่างระหว่างแต่ละชุดของข้อมูล ซึ่งชนิดของข้อมูลที่สามารถใช้เทคนิคการจัดกลุ่มแบบเชิงชั้นได้มี 3 ประเภทคือ

(1) ข้อมูลเป็นสเกลอันตรภาค (Interval scale) หรือสเกลอัตราส่วน (Ratio scale)

(2) ข้อมูลที่อยู่ในรูปความถี่ (Count Data)

(3) ข้อมูลอยู่ในรูป Binary นั่นคือ มีค่าได้ 2 ค่า คือ 0 กับ 1

กล่าวได้ว่าข้อมูลที่นำมาใช้กับเทคนิคการจัดกลุ่มแบบเชิงชั้นจะต้องเป็นข้อมูลชนิดตัวเลขหรือเป็นเชิงปริมาณหรือข้อมูลอยู่ในรูปของความถี่[11]

3. ตัดสินใจว่าจะจัดข้อมูลเป็นกี่กลุ่มโดยพิจารณาผลจากการจัดกลุ่มแบบเชิงชั้นในข้อที่ 2

The Dendogram

Dendogram เป็นเครื่องมือที่สำคัญสำหรับการจำแนกกลุ่มของประชากร ซึ่งจะวัดระยะทาง (distance) ระหว่างกลุ่มของข้อมูลหรือตัวแปรโดยการวัดระยะห่างและความคล้ายจะขึ้นกับชนิดของข้อมูล และแบ่งวิธีการวัดระยะห่างออกเป็น 3 ประเภทคือ

1. **Interval หมายถึง ข้อมูลที่เป็นเสกลอันตรภาคชั้นหรืออัตราส่วน**

โดยใช้วิธีการต่อไปนี้คือ

(1) Euclidean distance d_{ij} [1]

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ip} - x_{jp})^2}$$

เมื่อ d_{ij} คือ ระยะทางระหว่างข้อมูลกลุ่มที่ i และกลุ่ม j

โดย $X_i = (x_{i1}, x_{i2}, ..., x_{ip})$ และ $X_j = (x_{j1}, x_{j2}, ..., x_{jp})$

ถ้าตัวแปรแต่ละตัวมีน้ำหนักไม่เท่ากันให้ปรับน้ำหนักของตัวแปรโดย

$$d_{ij} = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + ... + w_p(x_{ip} - x_{jp})^2}$$

เมื่อ $w_1, w_2, ..., w_p$ เป็นค่าถ่วงน้ำหนักของตัวแปรแต่ละตัว

โดย $w_i \geq 0$ และ $\sum_{i=1}^{p} w_i = 1$

(2) Mahalanobis distance[1]

$$d_{ij} = \sqrt{(x_i - x_j) S^{-1} (x_i - x_j)}$$

เมื่อ x_i และ x_j คือ P-dimensional vector ของตัวแปร i และ j ตามลำดับ

และ S^{-1} คือ Covariance metric ของเวกเตอร์

2. **Count ใช้กับข้อมูลที่อยู่ในรูปของความถี่**[11]

(1) Chi-square measure

$$d_{ij} = \sqrt{\sum \frac{(x_i - E(x_i))^2}{E(x_i)} + \sum \frac{(x_j - E(x_j))^2}{E(x_j)}}$$

(2) Phi-square measure

$$d_{ij} = \sqrt{\frac{\sum \frac{(x_i - E(x_i))^2}{E(x_i)} + \sum \frac{(x_j - E(x_j))^2}{E(x_j)}}{N}}$$

3. **Binary ใช้กับข้อมูลที่มีค่า 2 ค่าเท่านั้นคือ 0 และ 1** [11]

โดย SPSS จะสร้างตาราง 2x2 ของ case แต่ละคู่ให้

		case 2	
		Present	Absent
case 1	Present	a	b
	Absent	c	d

(1) Square Euclidean distance

$$d_{ij} = b + c$$

(2) Euclidean Distance

$$d_{ij} = \sqrt{b + c}$$

(3) Size Difference

$$d_{ij} = \frac{(b - c)^2}{(a + b + c + d)^2}$$

(4) Pattern Difference

$$d_{ij} = \frac{bc}{(a + b + c + d)^2}$$

การวิเคราะห์จำแนกกลุ่มด้วยโปรแกรมสำเร็จรูป SPSS for Windows

ปัจจุบันโปรแกรมสำเร็จรูป SPSS for Windows เป็นโปรแกรมที่ถูกนำมาใช้ในการวิเคราะห์ข้อมูลอย่างแพร่หลายและสามารถเรียนรู้ได้ง่าย

ตัวอย่าง Data are presented for six species, most having data for more than one strain and 16 phenotypic characters (0 = absent, 1 = present). The species are:

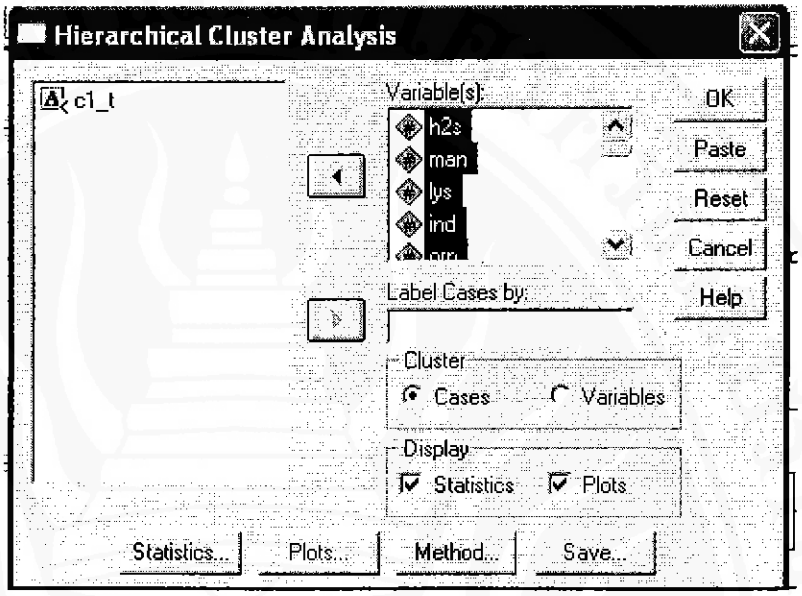
1. ecoli *Escherichia coli*
2. styphi *Salmonella typhi*
3. kpneu *Klebsiella pneumoniae*
4. pvul *Proteus vulgaris*
5. pmor *P. morganii*
6. smar *Serratia marcescens*

Species	H2S	MAN	LYS	IND	ORN	CIT	URE	ONP	VPT	INO	LIP	PHE	MAL	ADO	ARA	RHA
ecoli1	0	1	1	1	0	0	0	1	0	0	0	0	0	0	1	1
ecoli2	0	1	0	1	1	0	0	1	0	0	0	0	0	0	1	0
ecoli3	1	1	0	1	1	0	0	1	0	0	0	0	0	0	1	1
Styphi1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0
Styphi2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Styphi3	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0
Kpneu1	0	1	1	1	0	1	1	1	1	1	0	0	0	1	1	1
Kpneu2	0	1	1	1	0	1	1	1	1	1	0	0	1	0	1	1
Kpneu3	0	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1
Kpneu4	0	1	1	1	0	1	1	1	0	1	0	0	1	1	1	1
Kpneu5	0	1	1	1	0	1	0	1	1	1	0	0	1	1	1	1
pvul1	1	0	0	1	0	1	1	0	0	0	0	1	0	0	0	0
pvul2	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
pvul3	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0
pmor1	0	0	1	1	1	0	1	0	0	0	0	1	0	0	0	0
pmor2	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0
smar	0	1	1	0	1	1	0	1	1	0	1	0	0	0	0	0

ที่มา: Rataj & Schindler (*1991, Binary*, **3**:159-164)[4]

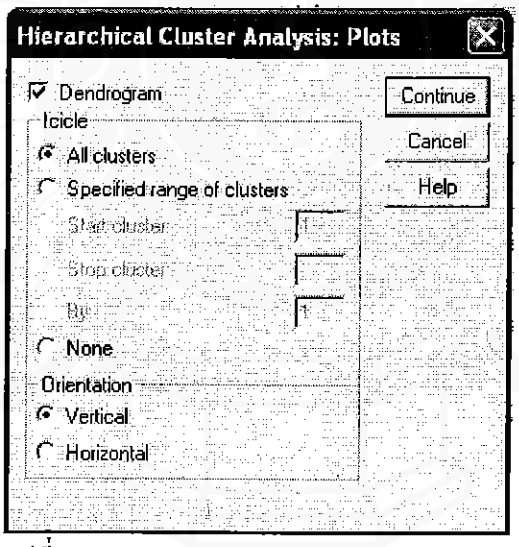
ขั้นตอนการวิเคราะห์

1. เลือก Analyze → Classify → Hierarchical Cluster … จะได้



รูปที่ 1 แสดง Hierarchical Cluster Analysis

2. เลือก Plots… จะได้

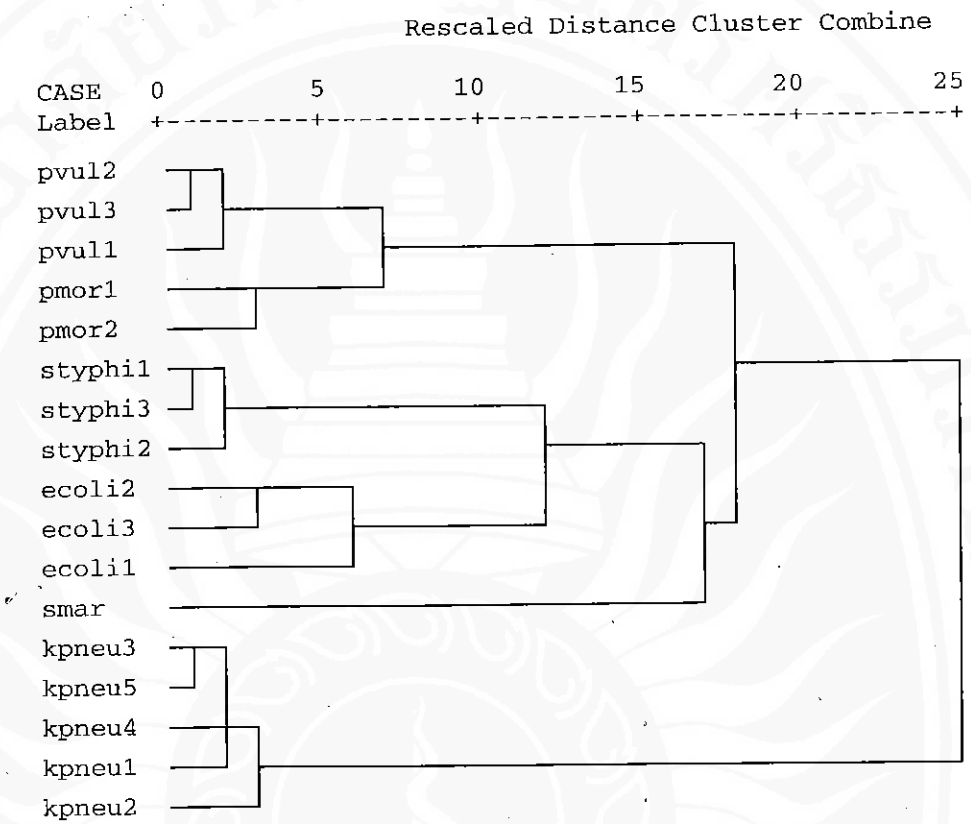


รูปที่ 2 แสดง Hierarchical Cluster Analysis: Plots

3. จากรูปที่ 1 กดปุ่ม OK … จะได้

HIERARCHICAL CLUSTER ANALYSIS

Dendrogram using Average Linkage (Between Groups)



จาก Dendogram ตัวอย่างแบ่งออกเป็น 6 กลุ่ม (Cluster) คือ

กลุ่มที่ 1 กลุ่ม *Proteus vulgaris* ประกอบด้วยหน่วยตัวอย่างที่ 13 14 และ 12

กลุ่มที่ 2 กลุ่ม *P. morganii* ประกอบด้วยหน่วยตัวอย่างที่ 15 และ 16

กลุ่มที่ 3 กลุ่ม *Salmonella typhi* ประกอบด้วยหน่วยตัวอย่างที่ 4 6 และ 5

กลุ่มที่ 4 กลุ่ม *Escherichia coli* ประกอบด้วยหน่วยตัวอย่างที่ 2 3 และ 1

กลุ่มที่ 5 กลุ่ม *Serratia marcescens* ประกอบด้วยหน่วยตัวอย่างที่ 17

กลุ่มที่ 6 กลุ่ม *Klebsiella pneumoniae* ประกอบด้วยหน่วยตัวอย่างที่ 9 11 10 7 และ 8

จะเห็นว่าวิธีการทางสถิติขั้นสูงสามารถจำแนกกลุ่มสปีชี่ส์ได้เช่นเดียวกับการจำแนกกลุ่มโดยใช้วิธีการทางด้านชีววิทยาและปัจจุบันวิธีการทางสถิติขั้นสูงมีการใช้อย่างแพร่หลายในหลายสาขาวิชา

เอกสารอ้างอิง

กัลยา วานิชย์บัญชา.(2544). **การวิเคราะห์ตัวแปรหลายตัวด้วย SPSS for Windows.** โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพมหานคร.

ปรีชา วิจิตรธรรมรส.(2544). **เอกสารประกอบการอบรมเรื่อง การวิเคราะห์ข้อมูลขั้นสูงด้วยโปรแกรมสำเร็จรูป**. สถาบันบัณฑิตพัฒนบริหารศาสตร์, กรุงเทพมหานคร.

Alan, F.(2004). **Cluster Analysis** [Online]. Available HTTP: http://obelia.jde.aca.mmu.ac.uk/multivar/ca.html

Everitt, B. S., Landau, S. and Leese, M. (2001). **Cluster Analysis (4th edition)**. Edward Arnold.

Jongman, R. H. G., ter Braak, C. J. F. and van Tongeren, O. F. R.(1995). **Data analysis in community and landscape ecology**. Cambridge University Press.

Legendre, P. and Legendre, L.(1998). **Numerical Ecology (2nd English edition)**. Elservier, Amsterdam.

Sloan School of Management Massachusetts Institute of Technology.(2004). **Cluster Analysis** [Online]. Available HTTP: http://ocw.mit.edu/NR/rdonlyres/Sloan-School-of-Management/15-062Data-MiningSpring2003/E31DD7A9-4B2E-48BF-AB3F-6F4CCAD7F24F/0/lec11.pdf

Statistical Programming for Social Sciences. (2004). Advanced Model [Online]. Available HTTP: http://www.spss.com

Statsoft.(2004). **Cluster Analysis** [Online]. Available HTTP: http://www.statsoftinc.com/textbook/stcluan.html#d

United State Environmental Protection Agency. (2004). **Multivariate Methods: Cluster Analysis** [Online]. Available HTTP: http://www.epa.gov/bioindicators/primer/cluster.html