CSC662 Data Mining, Data Warehouse and Visualization

บทที่ 11 เมตริกซ์และการวิเคราะห์การเกาะกลุ่ม

เตรียมโดย ผศ. ดร. กรุง สินอภิรมย์สราญ
ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

โครงร่าง

- ข้อมูลและชนิดของข้อมูล
- เมตริกซ์
- รูปแบบการกำหนดเมตริกซ์
- การคำนวณเมตริกซ์ตามประเภทของตัวแปร
- นิยามการเกาะกลุ่ม
- การวิเคราะห์การเกาะกลุ่ม
- การประยุกต์

ข้อมูลที่ผ่านขั้นตอนการเตรียมข้อมูลแล้ว

- ประกอบด้วยหลักและแถว
  - หลักคือลักษณะประจำ หรือเรียกว่าตัวแปร ซึ่งมีทั้งตัวแปรจำนวนและตัวแปรที่มีค่าที่ไม่ต่อเนื่อง
  - แถวคือระเบียนที่หมายถึงข้อมูลที่มีค่าในแต่ละลักษณะประจำที่สนใจ
- ไม่มีข้อมูลที่ขาดหายไป หรือข้อมูลที่ผิดปรกติ
- ในการทำเหมืองข้อมูล ลักษณะประจำและระเบียนต้องถูกเปลี่ยนเพื่อความเหมาะสมสำหรับแต่ละเทคนิคของการทำเหมืองข้อมูล

ชนิดของลักษณะประจำ หรือตัวแปร

- ตัวแปรไม่มีลำดับที่มีค่าไม่ต่อเนื่อง Nominal เป็นลักษณะประจำที่มีค่าไม่ต่อเนื่องในแต่ละระเบียน และไม่มีนิยามของลำดับที่ชัดเจน เช่น เพศ สี แผนก
- ตัวแปรมีลำดับที่มีค่าไม่ต่อเนื่อง Ordinal เป็นลักษณะประจำที่มีลำดับและมีค่าไม่ต่อเนื่อง เช่น กลุ่มอายุ = เด็ก วัยรุ่น ผู้ใหญ่
- ตัวแปรจำนวน Numerical (Continuous) เป็นลักษณะประจำที่มีค่าต่อเนื่อง เช่น ส่วนสูง ราคา น้ำหนัก

เมตริกซ์และการวัดระยะของข้อมูล

- การทำเหมืองข้อมูลแบบวิเคราะห์การเกาะกลุ่มของระเบียนข้อมูลใช้แนวคิดของระยะระหว่างระเบียน ตัววัดระยะระหว่างวัตถุเรียก เมตริกซ์ (Metric)
- สมบัติของเมตริกซ์ที่ถูกนำมาใช้
  - ระยะเมตริกซ์ที่สั้น สื่อถึงระเบียนที่คล้ายคลึงกัน (ไม่ต่างกันมาก)
  - ระยะเมตริกซ์ที่ไกล สื่อถึงระเบียนที่ต่างกันมาก
- การนิยามเมตริกซ์มักขึ้นกับชนิดของลักษณะประจำ Nominal, Ordinal, Numeric
- ผู้ใช้กำหนดค่าที่แยกว่าระเบียนคล้ายกันพอ (similar enough) หรือใกล้กันมากพอ (close enough) ได้

รูปแบบการกำหนดเมตริกซ์

- กำหนดจากเมทริกซ์ข้อมูล (Data matrix)
  - แบบ two modes $\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$
- กำหนดจากเมทริกซ์ความต่าง (Dissimilarity matrix)
  - แบบ one mode $\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$

การคำนวณเมตริกซ์ตามชนิดของตัวแปร

- ตัวแปรจำนวน (Numeric variable)
- ตัวแปรทวิภาค (Boolean variable)
- ตัวแปรไม่มีลำดับที่มีค่าไม่ต่อเนื่อง (Nominal variable)
- ตัวแปรมีลำดับที่มีค่าไม่ต่อเนื่อง (Ordinal variable)

ตัวแปรจำนวน (Numeric variable)

- ปรับให้อยู่ในรูปมาตรฐาน z-score $z_{if} = \frac{x_{if} - m_f}{s}$
  - ใช้ mean absolute deviation:
    $$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \dots + x_{nf})$$
  - ใช้ standard deviation:
    $$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$
- ในการวิเคราะห์การเกาะกลุ่ม การใช้ z-score ด้วย mean absolute deviation มักให้ผลที่แม่นยำกว่า standard deviation
  $$s_d = \frac{1}{n-1}\sqrt{(x_{1f} - m_f)^2 + (x_{2f} - m_f)^2 + \dots + (x_{nf} - m_f)^2}$$

ความคล้ายคลึงและความแตกต่าง d(*i* , *j*)

- ระยะทาง มีความสัมพันธ์ผกผันกับตัววัดความคล้ายคลึง (similarity มาก ค่าระยะจะน้อย แต่ถ้า similarity ค่าระยะจะมาก)
- การคำนวณระยะที่ใช้ในปัจจุบันคือ *Minkowski distance*:

$$d(i,j)=\sqrt[q]{(|x_{i1}-x_{j1}|^q+|x_{i2}-x_{j2}|^q+\ldots+|x_{ip}-x_{jp}|^q)}$$

เมื่อ $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ และ $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ เป็นระเบียนข้อมูลขนาด p มิติและ q เป็นจำนวนเต็มบวก

- ถ้า $q = 1$ แล้ว d คือ Manhattan distance

$$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+\ldots+|x_{ip}-x_{jp}|$$

สมบัติของเมตริกซ์

- ถ้า $q = 2$ แล้ว d คือ Euclidean distance:

$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\ldots+|x_{ip}-x_{jp}|^2)}$$

  - สมบัติของเมตริกซ์
    - $d(i, j) \geq 0$
    - $d(i, i) = 0$
    - $d(i, j) = d(j, i)$
    - $d(i, j) \leq d(i, k) + d(k, j)$
- นอกจากนี้เราอาจใช้ระยะแบบถ่วงน้ำหนัก หรือ parametric Pearson product moment correlation หรือตัววัดอื่น ๆ

ตัวแปรทวิภาค (Binary variables)

- รวมตัวแปรทวิภาคทั้งหมด มาสร้างตารางการจร (contingency table)

		Object *j*		
		1	0	*sum*
Object *i*	1	*a*	*b*	*a*+*b*
	0	*c*	*d*	*c*+*d*
	sum	*a*+*c*	*b*+*d*	*p*

- ระยะที่ได้จากสัมประสิทธิ์ที่คิดแบบสมมาตร (symmetric coefficient) ระหว่างระเบียน i กับระเบียน j

$$d(i,j)=\frac{b+c}{a+b+c+d}$$

- ระยะที่ได้จากสัมประสิทธิ์ที่คิดแบบไม่สมมาตร (Jaccard coefficient)

$$d(i,j)=\frac{b+c}{a+b+c}$$

ตัวอย่างการคำนวณเมตริกซ์ของตัวแปรทวิภาค

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- พิจารณาเฉพาะลักษณะประจำที่ไม่สมมาตรคือ Fever, Cough, Test-1, Test-2, Test-3, Test-4
- ให้ Y และ P มีค่าเป็น 1 และ N มีค่าเป็น 0

$$d(jack, mary)=\frac{0+1}{2+0+1}=0.33$$

$$d(jack, jim)=\frac{1+1}{1+1+1}=0.67$$

$$d(jim, mary)=\frac{1+2}{1+1+2}=0.75$$

ตัวแปรไม่มีลำดับที่มีค่าไม่ต่อเนื่อง (Nominals)

- ตัวแปรไม่มีลำดับที่มีค่าไม่ต่อเนื่องมักมีค่ามากกว่าสองค่า ตัวอย่างเช่น สีที่อาจมีค่าเป็นสีแดง (red) สีเหลือง (yellow) สีน้ำเงิน (blue) สีเขียว (green)
- วิธีการวัดแบบที่ 1: นับการเข้าคู่ทั้งหมดทุก Nominal คำนวณโดย
  - m:จำนวนที่เข้าคู่ และ p:จำนวนทั้งหมด

$$d(i,j)=\frac{p-m}{p}$$

- วิธีการวัดแบบที่ 2: แปลงเป็นตัวแปรทวิภาค (Dummy coding) สำหรับแต่ละค่าที่เป็นไปได้ แล้วใช้การวัดระยะสำหรับตัวแปรทวิภาคแบบสมมาตร เช่น ตัวแปรสี (color) ที่เป็นไปได้สี่สีข้างต้น สร้างเป็นตัวแปรทวิภาคสี่ตัวคือ red, yellow, blue, green

เมตริกซ์รวม

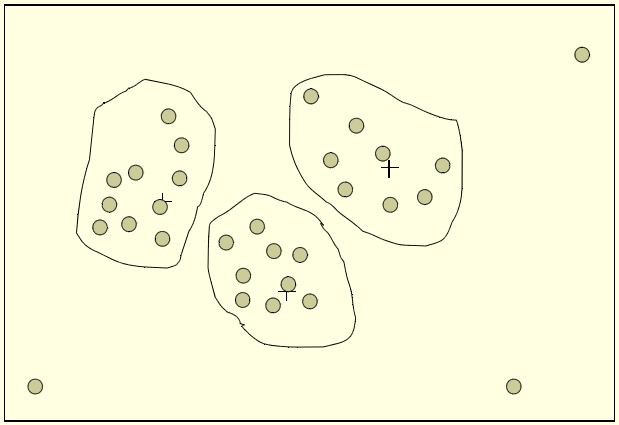
- การทำเหมืองข้อมูลแบบวิเคราะห์การเกาะกลุ่มต้องพิจารณาตัวแปรทั้ง 5 รูปแบบ symmetric binary, asymmetric binary, nominal, ordinal, numeric
- ระยะระหว่างระเบียนคำนวณจากตัวแปรทั้งหมดที่เกี่ยวข้อง โดยอาจมีการกำหนดค่าถ่วงน้ำหนักให้กับแต่ละตัวแปร
  - ตัวแปรทวิภาคหรือตัวแปรไม่มีลำดับที่มีค่าไม่ต่อเนื่อง
    $d_{ij}^{(t)} = 0$ ถ้า x_{if} มีค่าที่ต้องการ, $d_{ij}^{(t)} = 1$ กรณีอื่น
  - ตัวแปรจำนวนให้แปลงให้อยู่ในสเกลมาตรฐานก่อนนำมารวม
  - ตัวแปรมีลำดับที่มีค่าไม่ต่อเนื่อง
    - คำนวณค่าลำดับที่ r_{if} (rank) $z_{if}=\frac{r_{if}-1}{M_f-1}$
    - แล้วหา z_{if} คล้ายกับการจัดการกับตัวแปรจำนวน

$$d(i,j)=\frac{\sum_{t=1}^{p}\delta_{ij}^{(t)}d_{ij}^{(t)}}{\sum_{t=1}^{p}\delta_{ij}^{(t)}}$$

การวิเคราะห์การเกาะกลุ่ม

- ข้อมูลเกาะกลุ่ม หมายถึงลักษณะของกลุ่มข้อมูลที่จับตัวอยู่ร่วมกัน โดย
  - ข้อมูลที่อยู่ในกลุ่มเดียวกันมีความคล้ายกัน (ระยะใกล้)
  - ข้อมูลที่อยู่ต่างกลุ่มมีความแตกต่างกัน (ระยะไกล)
- การวิเคราะห์การเกาะกลุ่มคือ การหาวิธีการแบ่งข้อมูลออกเป็นกลุ่มตามลักษณะประจำที่มีของแต่ละข้อมูลโดยไม่มีการกำหนดลักษณะกลุ่มหรือการจัดกลุ่ม
- การเกาะกลุ่มมีลักษณะ unsupervised classification กล่าวคือไม่มีการกำหนดคลาสที่ต้องการ การเรียนรู้ได้มาจากค่าลักษณะประจำของข้อมูล
- นำไปประยุกต์ใช้กับการวิเคราะห์การกระจายข้อมูล และถูกใช้ในกระบวนการเตรียมข้อมูล เพื่อนำไปใช้กับวิธีการอื่น

การวิเคราะห์การเกาะกลุ่ม



ตัวอย่างการนำไปประยุกต์

- ใช้สำหรับการรู้จำรูปแบบ (Pattern Recognition)
- ใช้สำหรับวิเคราะห์ข้อมูลที่เกี่ยวกับปริภูมิ (Spatial Data Analysis)
  - การสร้าง thematic maps ใน GIS โดยการเกาะ feature space
  - การตรวจจับกลุ่มที่เกี่ยวข้องกับระยะ สิ่งรอบข้าง ความสูง ความลึก
- ใช้สำหรับการประมวลผลภาพ (Image Processing)
- ใช้ในวิทยาศาสตร์เศรษฐศาสตร์ (โดยเฉพาะการวิจัยตลาด)
- ใช้กับ WWW
  - การแบ่งกลุ่มเอกสาร (Document classification)
  - การเกาะกลุ่มข้อมูลใน Weblog เพื่อดูรูปแบบการเข้าถึงข้อมูล

ศาสตร์ที่นำการวิเคราะห์การเกาะกลุ่มไปใช้

- ทางการตลาด: ช่วยนักการตลาดค้นพบการจับกลุ่มของลูกค้าที่มีพฤติกรรมคล้ายกัน ซึ่งนำมาช่วยออกแบบการขายตามลักษณะกลุ่มเป้าหมายที่ต้องการ
- ทางพื้นที่: ช่วยค้นหาบริเวณที่มีการใช้สอยที่คล้ายกัน ตามลักษณะผืนดิน
- ทางการประกัน: ตรวจจับกลุ่มผู้ถือประกันรถยนต์ที่มีการเรียกร้องค่าประกันที่ผิดปรกติ
- ทางการวางผังเมือง: วิเคราะห์การจับตัวกันของบ้าน ตามชนิด ราคา สถานที่
- ทางการศึกษาแผ่นดินไหว: การวิเคราะห์ศูนย์กลางการไหว ที่นำไปตรวจจับแผ่นดินไหวตามแนวแยกของขอบโลก

ลักษณะการเกาะกลุ่มที่ดี

- การเกาะกลุ่มที่ดี ต้องรวมข้อมูลที่คล้ายกันไว้ด้วยกัน และแยกข้อมูลที่ต่างกันออกเป็นกลุ่มย่อย
- ผู้ใช้มักสนใจคุณภาพของการเกาะกลุ่มซึ่งขึ้นกับตัววัดความคล้ายคลึงและขั้นตอนวิธีที่ใช้ พร้อมกับลักษณะกลุ่ม (Clustering profiles)
- การเกาะกลุ่มที่น่าสนใจคือการแยกข้อมูลเป็นกลุ่ม และมีการบ่งบอกลักษณะที่ซ่อนไว้ในกลุ่มที่น่าสนใจได้
- วิธีการเกาะกลุ่มที่มีประสิทธิภาพควรใช้ได้กับข้อมูลขนาดใหญ่ และไม่ใช้เวลานานมากในการสร้างตัวแบบ

รายการสรุป การวิเคราะห์การเกาะกลุ่ม

- ใช้ได้กับข้อมูลปริมาณมาก
- จัดการกับลักษณะประจำแยกชนิด
- ไม่ควรถูกจำกัดการเกาะตัวในลักษณะทรงกลมหรือวงรีเท่านั้น
- พารามิเตอร์ที่มีควรปรับแต่งโดยอัตโนมัติ
- การเกาะกลุ่มไม่ควรเปลี่ยนไปเมื่อเจอข้อมูลขยะและ/หรือข้อมูลที่ผิดปรกติ
- ลำดับของข้อมูลเข้าไม่ควรมีผลต่อขั้นตอนวิธีที่ใช้
- ใช้กับข้อมูลที่มีมิติสูง คือมีลักษณะประจำจำนวนมากได้
- สามารถเพิ่มเงื่อนไขที่ผู้ใช้ต้องการได้
- มีการแปลความที่ง่ายและสะดวกในการนำไปใช้

ประเภทของขั้นตอนวิธีการวิเคราะห์การเกาะกลุ่ม

- ประเภทใช้วิธีแบ่งกั้น (Partitioning): ข้อมูลถูกแบ่งกั้นเป็นกลุ่มที่ไม่มีสมาชิกร่วมกันเลย แล้วใช้ตัววัดทดสอบว่ากลุ่มที่แบ่งเหมาะสมหรือไม่
- ประเภทระดับชั้น (Hierarchy): สร้างกลุ่มโดยการใช้ระดับชั้น (hierarchical decomposition) ตามเงื่อนไขที่ต้องการ
- ประเภทความหนาแน่น (Density-based): เน้นการเชื่อมกันและความหนาแน่น ระหว่างระเบียนข้อมูล
- ประเภทแบ่งกริด (Grid-based): ใช้ความละเอียดโดยแบ่งกริด
- ประเภทตัวแบบ (Model-based): สร้างตัวแบบ แล้วเลือกตัวแบบที่ดีที่สุด

สรุป

- ชนิดของข้อมูลมีผลต่อการคำนวณระยะระหว่างระเบียน
- การคำนวณระยะใช้เมตริกซ์ซึ่งนิยามกับข้อมูลห้าชนิด Numeric variable, Symmetric boolean variable, Asymmetric boolean variable, Nominal variable, Ordinal variable
- ในการวิเคราะห์การเกาะกลุ่ม เราใช้ตัววัดระยะรวมซึ่งเกิดจากการใช้ค่าถ่วงน้ำหนักคูณกับระยะของการวัดข้อมูลทั้งห้าชนิด
- ขั้นตอนวิธีการเกาะกลุ่มคือการวิเคราะห์เพื่อแบ่งแยกข้อมูลออกเป็นส่วน ๆ โดยใช้ระยะที่กำหนดให้

เอกสารอ้างอิง ๑

- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999.
- Y. Wand and R. Wang. Anchoring data quality dimensions ontological foundations. Communications of ACM, 39:86-95, 1996.
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. IEEE Trans. Knowledge and Data Engineering, 7:623-640, 1995.
- R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. SIGMOD'98
- M. R. Anderberg. Cluster Analysis for Applications. Academic Press, 1973.
- M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure, SIGMOD'99.
- P. Arabie, L. J. Hubert, and G. De Soete. Clustering and Classification. World Scietific, 1996

เอกสารอ้างอิง ๒

- M. Ester, H.-P. Kriegel, and X. Xu. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. SSD'95.
- D. Gibson, J. Kleinberg, and P. Raghavan. Clustering categorical data: An approach based on dynamic systems. In Proc. VLDB'98.
- A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Printice Hall, 1988.
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- G. J. McLachlan and K.E. Bkasford. Mixture Models: Inference and Applications to Clustering. John Wiley and Sons, 1988.
- P. Michaud. Clustering techniques. Future Generation Computer systems, 13, 1997.
- R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. VLDB'94.