CSC662 Data Mining, Data Warehouse and Visualization

บทที่ 6 การแปลงข้อมูลและการเปลี่ยนลักษณะประจำเป็นชนิดไม่ต่อเนื่อง

เตรียมโดย ผศ. ดร. กรุง สินอภิรมย์สราญ

ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

เนื้อหา

- ข้อมูลที่ผ่านขั้นตอนการทำความสะอาดมาแล้ว
- การทำเหมืองข้อมูลกับลักษณะประจำที่ใช้
- ขั้นตอนวิธีการแปลงค่าข้อมูลจำนวน
- ขั้นตอนวิธีการเปลี่ยนลักษณะประจำชนิดต่อเนื่องเป็นลักษณะประจำชนิดไม่ต่อเนื่อง
- การสร้างระดับชั้นของตัวแปรไม่ต่อเนื่องอย่างอัตโนมัติ

ข้อมูลที่ผ่านขั้นตอนการทำความสะอาดมาแล้ว

- ประกอบด้วยหลักและแถว
  - หลัก = ลักษณะประจำ = ตัวแปรในทางสถิติ ซึ่งมีทั้งตัวแปรจำนวน และตัวแปรที่มีค่าที่ไม่ต่อเนื่อง
  - แถว = ระเบียนที่หมายถึงข้อมูลที่กำหนดค่าในแต่ละลักษณะประจำ
- ไม่มีข้อมูลที่ขาดหายไป หรือข้อมูลที่ผิดปรกติ
- ในการทำเหมืองข้อมูล ชนิดของลักษณะประจำและระเบียนต้องถูกเปลี่ยนเพื่อความเหมาะสมสำหรับแต่ละเทคนิคในการทำเหมืองข้อมูล

ชนิดของลักษณะประจำ หรือตัวแปร

- ตัวแปรไม่มีลำดับที่มีค่าไม่ต่อเนื่องเรียก (Nominal) เป็นลักษณะประจำที่มีค่าไม่ต่อเนื่องในแต่ละระเบียน และไม่มีนิยามของลำดับที่ชัดเจน เช่น เพศ สี แผนก
- ตัวแปรที่มีลำดับที่มีค่าไม่ต่อเนื่อง (Ordinal) เป็นลักษณะประจำที่มีลำดับและมีค่าไม่ต่อเนื่อง เช่น กลุ่มอายุประกอบด้วย {เด็ก, วัยรุ่น, ผู้ใหญ่}
- ตัวแปรจำนวน Numerical (Continuous/Interval) เป็นลักษณะประจำที่มีค่าต่อเนื่อง เช่น ส่วนสูง ราคา น้ำหนัก

การทำเหมืองข้อมูลกับลักษณะประจำที่ใช้

- ขั้นตอนวิธีในการค้นหาความรู้หรือการทำเหมืองข้อมูลมีหลากหลายขึ้นกับปัญหา วัตถุประสงค์และความรู้ที่ได้ ซึ่งในแต่ละขั้นตอนวิธีเราจำเป็นต้องใช้ข้อมูลในรูปแบบที่แตกต่างกัน
- ผลลัพธ์ที่ได้จากการทำเหมืองข้อมูลแบบทำนายแบ่งได้เป็น
  - กฎเชื่อมโยง (Association rule)
  - ตัวแบบจัดจำแนกประเภท (Classifier)
  - ตัวแบบการเกาะกลุ่ม (Cluster)
- ลักษณะของข้อมูลที่ต้องการจะมีความแตกต่างไปตามขั้นตอนวิธีและผลลัพธ์

ลักษณะข้อมูลที่ต้องการของการหากฎเชื่อมโยง

- กฎเชื่อมโยง (Association rule) เป็นการวิเคราะห์หาความสัมพันธ์ระหว่างสิ่งมากกว่าหนึ่งสิ่งที่มีการบันทึกร่วมกันในหนึ่งแถวของข้อมูล
- ข้อมูลที่นำมาใช้วิเคราะห์มักมีลักษณะเป็น transaction กล่าวคือหนึ่งระเบียนข้อมูลคือความสัมพันธ์ของลักษณะประจำ (ตัวแปร) หนึ่งรูปแบบ
- การวิเคราะห์รูปแบบนี้ไม่มีลักษณะประจำเป้าหมาย
- ลักษณะประจำนำเข้าเป็นตัวแปรทวิภาค หรือตัวแปรที่มีค่าที่ไม่ต่อเนื่อง
- ในกรณีที่ลักษณะประจำมีค่าที่ต่อเนื่อง โปรแกรมจำเป็นต้องแปลงเป็นค่าไม่ต่อเนื่องก่อนการคำนวณ ซึ่งอาจทำโดยมนุษย์หรือใช้เครื่องคอมพิวเตอร์แปลงแบบอัตโนมัติ

ลักษณะข้อมูลที่ต้องการของตัวแบบจัดจำแนกประเภท

- การจัดจำแนกประเภท (Classification) เป็นการสร้างตัวแบบเพื่อใช้จำแนกระเบียนที่พบในอนาคตว่าจัดอยู่ในประเภทใด
- ข้อมูลที่ใช้ต้องมีลักษณะประจำเป้าหมาย (Class) ที่มีค่าไม่ต่อเนื่อง โดยปรกติมักเป็นลักษณะประจำทวิภาค
- สำหรับขั้นตอนวิธีจัดจำแนกประเภทบางขั้นตอนวิธี (ID3) ต้องการลักษณะประจำนำเข้าที่มีค่าไม่ต่อเนื่องเท่านั้น และบางขั้นตอนวิธี (Logistic regression) ก็ต้องการลักษณะประจำที่มีค่าต่อเนื่องเท่านั้น
- ปัจจุบันขั้นตอนวิธีจัดจำแนกประเภทยอมให้ใช้ลักษณะประจำนำเข้ามีค่าต่อเนื่องหรือไม่ต่อเนื่องก็ได้

ลักษณะข้อมูลที่ต้องการของตัวแบบการเกาะกลุ่ม

- การวิเคราะห์การเกาะกลุ่ม (Cluster analysis) ต้องใช้ระยะระหว่างระเบียนในการบ่งบอกความ เหมือนหรือความแตกต่างระหว่างระเบียน
- การวัดระยะระหว่างระเบียนใช้ผลรวมของการคำนวณระยะระหว่างลักษณะประจำแต่ละตัว
  - ในกรณีที่ลักษณะประจำเป็นจำนวน เราสามารถใช้กำลังสองของผลต่าง
  - ในกรณีที่ลักษณะประจำเป็นค่าที่ไม่ต่อเนื่อง เราสามารถใช้การนับความแตกต่าง ระหว่างลักษณะประจำที่ต่างกันได้
- ลักษณะประจำนำเข้าอาจเป็นค่าต่อเนื่องหรือค่าไม่ต่อเนื่องก็ได้ ตราบที่การคำนวณระยะมีสเกลที่เหมาะสม

การแปลงลักษณะประจำประเภทจำนวน

- การแปลงให้ค่าที่ปรากฏในลักษณะประจำอยู่ในสเกลมาตรฐาน
  - ใช้วิธี max-min normalization
  - ใช้วิธี z-score normalization
  - ใช้วิธี decimal scaling
- การปรับค่าที่ไม่ให้ค่าลบหรือศูนย์ เพื่อนำไปใช้เป็นตัวหาร
  - ใช้วิธีบวก 1 กับตัวแปรทวิภาค
  - ใช้วิธีบวกด้วย 1-min{A} กับตัวแปรที่มีค่าลบ แต่ต้องการให้มีค่าเป็นบวก
- การคูณด้วยค่าถ่วงน้ำหนักที่เหมาะสม ใช้เมื่อต้องการให้ลักษณะประจำใหม่มีความสำคัญกว่าลักษณะประจำอื่น

การแปลงให้อยู่ในสเกลมาตรฐาน

- วิธี max-min normalization

$$v' = \frac{v - min_A}{max_A - min_A}(nmax_A - nmin_A) + nmin_A$$

เมื่อ min_A, max_A คือค่าต่ำสุดและค่าสูงสุดของลักษณะประจำปัจจุบัน $nmin_A$, $nmax_A$ คือค่าต่ำสุดและค่าสูงสุดของลักษณะประจำใหม่ โดยปรกติค่า $nmax_A$=1, min_A=0

- วิธี z-score normalization

$$v' = \frac{v - \bar{A}}{std_A}$$

เมื่อ$\bar{A}$ คือค่าเฉลี่ยเลขคณิตและ std_A ส่วนเบี่ยงเบนมาตรฐานของลักษณะประจำ A

- วิธี decimal scaling

$$v' = \frac{v}{10^j}$$ เมื่อ j คือจำนวนเต็มที่น้อยที่สุดที่ $\max(|v'|) < 1$

การแปลงลักษณะประจำประเภทจำนวนให้เป็นค่าไม่การแปลงให้เป็นค่าไม่ต่อเนื่อง

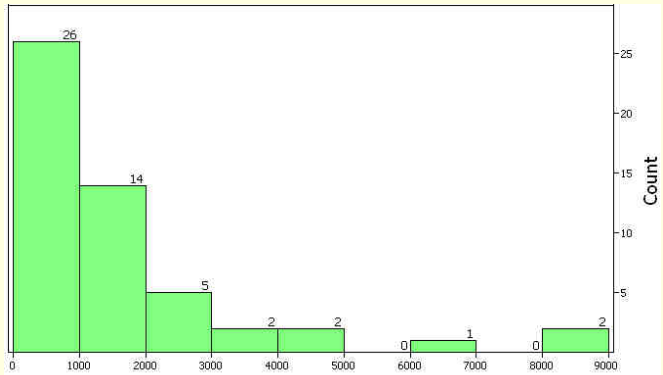
- การใช้กล่อง (Binning) ดูบทการเตรียมข้อมูล
- การใช้ฮีสโตแกรม (Histogram) ดูจากวิธีการทางสถิติ
- การใช้กลุ่มที่ได้จากการวิเคราะห์การเกาะกลุ่ม (Clustering) ใช้ขั้นตอนวิธีการวิเคราะห์การเกาะกลุ่มที่จะศึกษาในอนาคต
- การใช้เอ็นโทรปี (Entropy-based discretization)
- การใช้ผลแบ่งกั้นที่เข้าใจง่าย (Natural partitioning)

วิธีการใช้กล่อง

- เรียงข้อมูลและแบ่งข้อมูลลงกล่อง
  - การแบ่งกล่องที่มีขนาดเท่ากัน (equal width)
  - การแบ่งกล่องที่มีปริมาณข้อมูลเท่ากัน (equal depth)
- ตัวแทนกล่องสามารถใช้
  - ค่าเฉลี่ยเลขคณิต (bin means)
  - ค่ามัธยฐานของกล่อง (bin median)
  - ค่าฐานนิยมของกล่อง (bin mode)
  - ค่าขอบกล่อง (bin boundaries)

ฮีสโตแกรม (Histogram)

- เป็นวิธีทางสถิติที่ใช้ในการพิจารณาการกระจายของข้อมูลที่เป็นจำนวน มีการกำหนดจำนวนช่วงในการแบ่ง ความสูงคือผลรวมของความถี่
- สามารถคำนวณได้เร็ว



การใช้การวิเคราะห์การเกาะกลุ่ม

- กำหนดเมตริกซ์ที่ใช้ในการวิเคราะห์การเกาะกลุ่ม ซึ่งขึ้นกับชนิดของลักษณะประจำ และลักษณะข้อมูลที่ต้องการวิเคราะห์
- เลือกใช้ขั้นตอนวิธีการเกาะกลุ่มในรูปแบบต่าง ๆ
- ผลลัพธ์ที่ได้อาจเป็น
  - ตัวแทนกลุ่ม
  - ขอบเขตของกลุ่ม
  - ฟังก์ชันในการกำหนดกลุ่ม

วิธีการแบ่งค่าต่อเนื่องออกเป็นสองกลุ่มโดยใช้เอ็นโทรปี

- กำหนดข้อมูลในเซต S แบ่งโดยเลือกค่า T ที่ข้อมูลในหลักที่ $x \le T$ อยู่ในกลุ่ม S_1 และข้อมูลในหลักที่ $x > T$ อยู่ในกลุ่ม S_2 ค่าการแบ่ง

$$E_T(S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

เมื่อ $Ent(S_i)$ คือค่าเฉลี่ยแบบถ่วงน้ำหนักที่ S_i แบ่งข้อมูลเป็นพวกที่ i

- ทำซ้ำกับกลุ่มข้อมูลที่แบ่งกั้นจนกระทั่งกลุ่มข้อมูลสอดคล้องกับเงื่อนไขการหยุด

$$Ent(S) - E_T(S) > \delta$$

- จากประสบการณ์การแบ่งตามคลาสดังกล่าวให้ผลลัพธ์ที่ดี เมื่อนำไปใช้กับการจัดจำแนกประเภท

การคำนวณค่าเอ็นโทรปี

- ฟังก์ชันเอ็นโทรปี Ent ที่ใช้ได้เช่น
  - เลือกค่าสูงสุดของ Information gain วิธีการคือกำหนดเซต S แบ่งออกเป็น m ค่าตามคลาสที่สนใจ สมมติเราสนใจคลาส P กับ N ให้ p แทนจำนวนข้อมูลใน S ที่เป็นคลาส P และ n แทนจำนวนข้อมูลใน S ที่เป็นคลาส N

$$Ent(S) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

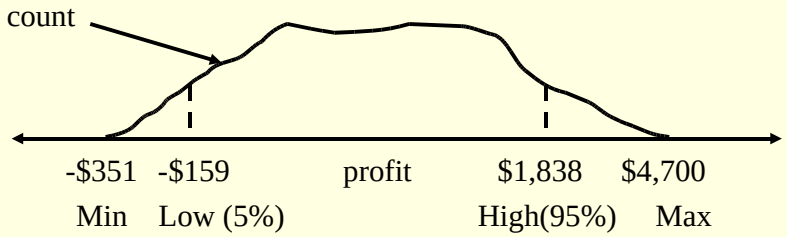
  - เลือกค่าน้อยสุดของ Gini index วิธีการคือกำหนดเซต S แบ่งออกเป็น m ค่าตามคลาสที่สนใจ แต่ละส่วนมีความถี่สัมพัทธ์คือ p_j

$$Ent(S) = 1 - \sum_{j=1}^{n} p_j^2$$

การใช้ผลแบ่งกั้นที่เข้าใจง่าย

- เป็นการแบ่งกั้นข้อมูลที่ใช้มนุษย์เป็นหลัก โดยใช้ระดับนัยสำคัญของตัวเลขและกำหนดการแบ่งออกเป็น 3, 4 หรือ 5 กลุ่มเท่านั้น
- กฎ 3-4-5 เลือกระดับนัยสำคัญสูงสุดจากจำนวนที่มีค่าน้อยที่สุดกับมากที่สุด
  - ถ้าค่าในลักษณะประจำมีเลขนัยสำคัญเป็น 3, 6, 7 หรือ 9 แบ่งข้อมูลออกเป็น 3 ช่วงคือ 3:1-1-1, 6:2-2-2, 7:2-3-2, 9:3-3-3
  - ถ้าค่าในลักษณะประจำมีเลขนัยสำคัญเป็น 2, 4 หรือ 8 แบ่งข้อมูลออกเป็น 4 ช่วงคือ 2:½-½-½-½, 4:1-1-1-1, 8:2-2-2-2
  - ถ้าค่าในลักษณะประจำมีเลขนัยสำคัญเป็น 1, 5 หรือ 10 แบ่งข้อมูลเป็น 5 ช่วงคือ 1:1/5-1/5-1/5-1/5-1/5, 5:1-1-1-1-1, 10:2-2-2-2-2

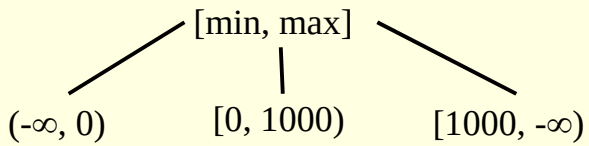
ตัวอย่างการใช้กฎ 3-4-5



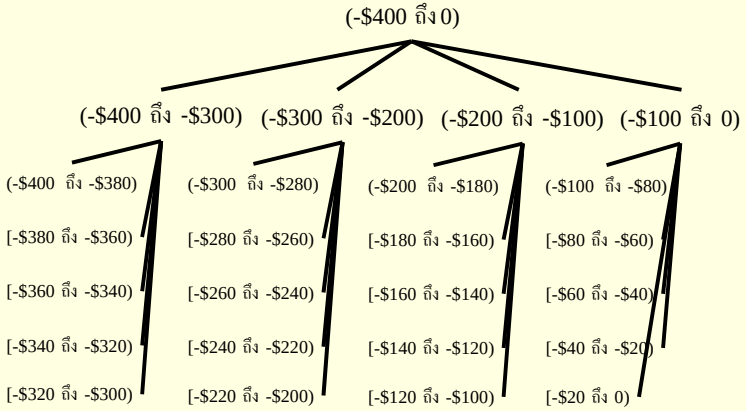
- ขั้นแรกหาค่าต่ำสุดและสูงสุด Min = -351, Max = 4700 เพื่อลดผลจากข้อมูลที่ผิดปรกติมาก (Outlier) เราพิจารณาเพียง 90% กลางของข้อมูล กล่าวคือเราตัด 5 percentile = -159 และ 95 percentile = 1838

ตัวอย่างการใช้กฎ 3-4-5

- ขั้นที่สองหาระดับนัยสำคัญที่ต้องการใช้ ในกรณีนี้คือ 3 เพราะค่าที่ใช้ตัดคือ -159 ปัดลงได้ -1000 กับ 1838 หาปัดขึ้นได้ 2000
- ขอบล่างและขอบบนที่ครอบคลุมค่าดังกล่าวคือ -1000 ถึง 2000 มีผลต่างเป็น 3000 ดังนั้นเราแบ่งช่วงออกเป็นสามช่วง (-∞, 0), [0, 1000), [1000, ∞)



ตัวอย่างการใช้กฎ 3-4-5

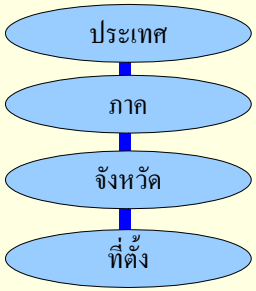


การสร้างระดับชั้นของตัวแปรไม่ต่อเนื่องอย่างอัตโนมัติ

- ลักษณะประจำที่มีค่าไม่ต่อเนื่องอาจมีค่าที่ต่างกันปริมาณมาก โดยไม่มีอันดับกำหนดให้ แต่เรายังต้องการสร้างระดับชั้นเพื่อนำไปใช้วิเคราะห์
  - ผู้ใช้หรือผู้เชี่ยวชาญกำหนดกลุ่มของลักษณะประจำที่นำมาใช้สร้างระดับชั้น พร้อมกับอันดับ เช่น street < city < province_or_state < country
  - ผู้ใช้หรือผู้เชี่ยวชาญกำหนดการสร้างกลุ่มบางส่วน เช่น {เชียงใหม่, เชียงราย, แม่ฮ่องสอน} ⊂ ภาคเหนือ
  - ผู้ใช้หรือผู้เชี่ยวชาญอาจกำหนดกลุ่มของลักษณะประจำที่นำมาใช้สร้างระดับชั้นแต่ไม่กำหนดอันดับ แล้วใช้โปรแกรมสร้างระดับชั้นแบบอัตโนมัติ

การสร้างระดับชั้นอัตโนมัติ

การสร้างระดับชั้นอัตโนมัติทำได้โดยใช้จำนวนค่าที่ต่างกันในการตัดสิน ลักษณะประจำที่มีปริมาณค่าที่แตกต่างกันมากที่สุดจะถูกจัดอยู่ในระดับล่างสุด ในขณะที่ลักษณะประจำที่มีปริมาณค่าที่แตกต่างกันน้อยจะถูกจัดอยู่ในระดับที่สูงขึ้นไป



ค่าต่างกัน 4 ค่า

ค่าต่างกัน 20 ค่า

ค่าต่างกัน 3042 ค่า

ค่าต่างกัน 487623 ค่า

สรุป

- การแปลงข้อมูลและการเปลี่ยนให้เป็นข้อมูลไม่ต่อเนื่อง เป็นหนึ่งในขั้นการเตรียมข้อมูล ซึ่งนำไปใช้กับ คลังข้อมูล (Data warehouse) และการทำเหมืองข้อมูล (Data mining)
- การแปลงข้อมูลขึ้นกับเทคนิคการทำเหมืองข้อมูล
- การแปลงชนิดของลักษณะประจำเป็นค่าที่ไม่ต่อเนื่อง เราสามารถใช้ กล่อง ฮีสโตแกรม วิธีการเอ็นโทรปี กฎ 3-4-5
- การสร้างระดับชั้นแบบอัตโนมัติทำได้โดยอาศัยจำนวนค่าที่แตกต่างกัน

เอกสารอ้างอิง

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Communications of ACM, 42:73-78, 1999.
- Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), December 1997.
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999.
- T. Redman. Data Quality: Management and Technology. Bantam Books, New York, 1992.
- 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.