CSC662 ปฏิบัติการ ๗

กฎเชื่อมโยงในซอฟต์แวร์ Weka

เขียนโดย ผศ. ดร. กรุง สินอภิรมย์สราญ

ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

เนื้อหาที่ครอบคลุม

- เกริ่นนำกฎเชื่อมโยงในซอฟต์แวร์ Weka
- การเตรียมแฟ้มข้อมูล
- ขั้นตอนวิธี Apriori
- การแปลผล
- การใช้ขั้นตอนวิธี Apriori กับข้อมูลที่ไม่ใช่ Transaction

การทำเหมืองข้อมูลแบบกฎเชื่อมโยง

- ใช้กับ Market Basket analysis
- กฎบ่งบอกพฤติกรรมการซื้อของลูกค้า
- ปรกติใช้กับฐานข้อมูลเชิงสัมพันธ์ที่บันทึกเป็น Transaction โดยที่แต่ละระเบียนคือการซื้อสินค้าในหนึ่งครั้ง
- ผลลัพธ์ที่ต้องการได้คือ กฎแสดงความสัมพันธ์ของการซื้อสินค้าต่างชนิดกันโดยไม่ขึ้นกับลูกค้าคนใดคนหนึ่ง

ข้อมูลที่นำมาใช้

รหัสการซื้อ	รายการสินค้า
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

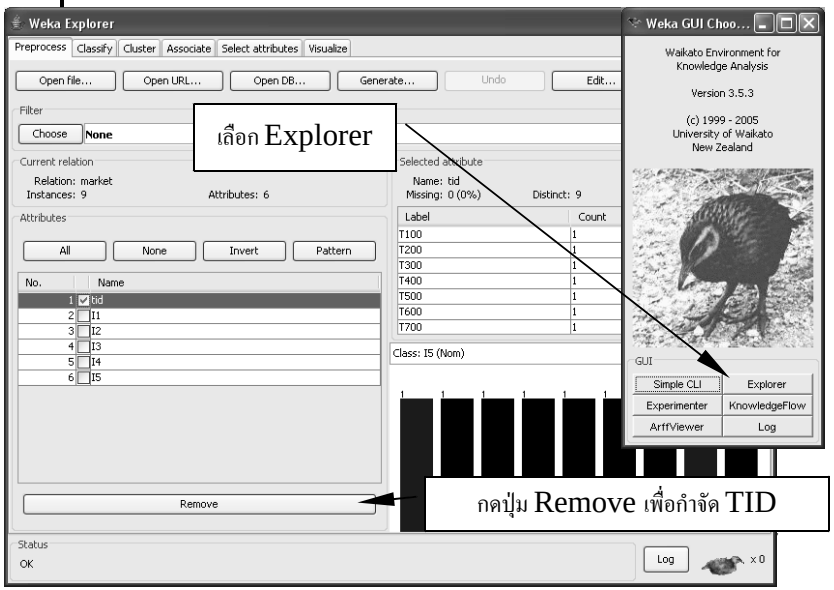
การเตรียมแฟ้มข้อมูล

- ลักษณะประจำของสินค้าคือชื่อสินค้าที่พิจารณา
- ลักษณะประจำตัวแรกคือรหัสการซื้อสินค้า TID ที่ไม่นำมาใช้ในการวิเคราะห์ ใช้เพื่อการเชื่อมโยงกลับไปยังฐานข้อมูลเริ่มต้นเท่านั้น
- ค่าในลักษณะประจำเป็น boolean เช่นกำหนดค่าที่เป็นไปได้คือ y แทนการใช้ตัวเลข 1
  - ตัวอย่าง การซื้อ T100, I1, I2 เขียนเป็น T100, 1, 1, ?, ?, ? ในซอฟต์แวร์ Weka สัญลักษณ์ ? แทนค่าที่หายไป (missing value)

แฟ้ม market.arff

```
@relation market
@attribute tid {T100, T200, T300, T400, T500, T600, T700, T800, T900}
@attribute I1 {y}
@attribute I2 {y}
@attribute I3 {y}
@attribute I4 {y}
@attribute I5 {y}
@data
T100,y,y,?,?,y
T200,?,y,?,y,?
T300,?,y,y,?,?
T400,y,y,?,y,?
T500,y,?,y,?,?
T600,?,y,y,?,?
T700,y,?,y,?,?
T800,y,y,y,?,y
T900,y,y,y,?,?
```

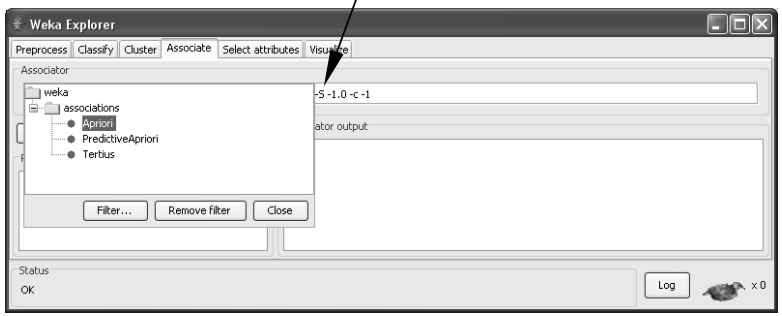

การเปิดแฟ้ม Market.arff



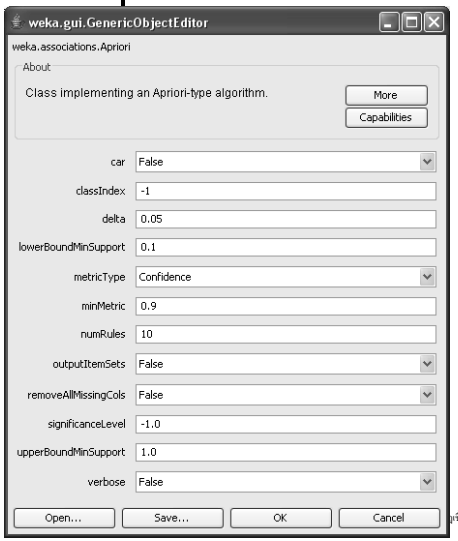
การเลือกขั้นตอนวิธี Apriori

- เลือกแถบ Associate
- ภายใต้ Associator เลือก Apriori

กดปุ่มในกล่อง Associator เพื่อปรับเปลี่ยนค่าพารามิเตอร์สำหรับ Apriori

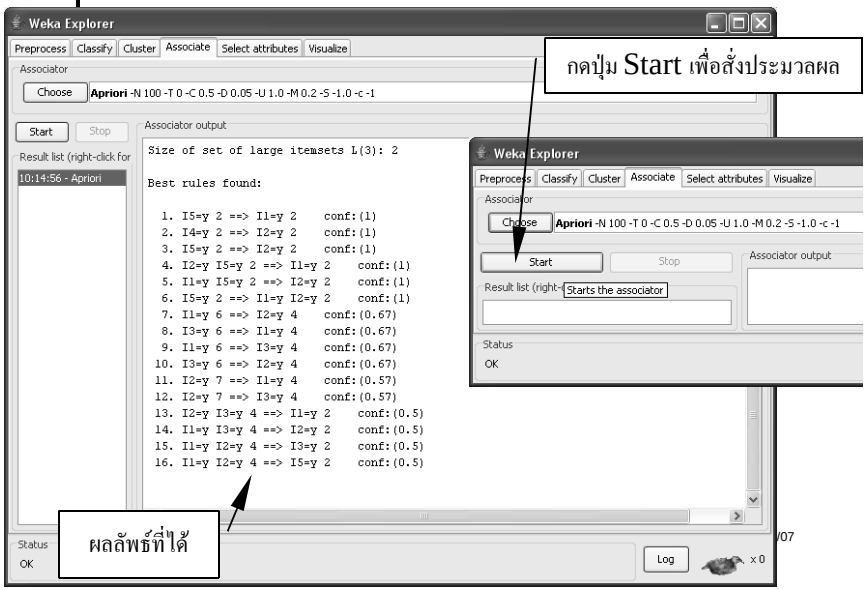


ตัวเลือกในขั้นตอนวิธี Apriori



- ปรับค่า min support ในกล่อง lowerBoundMinSupport เช่น 0.2 (หมายถึงค่าสนับสนุนต่ำสุด 20%)
- ปรับค่า min confidence ในกล่อง minMetric โดย metricType เป็น Confidence เช่น 0.5 (หมายถึงค่าความเชื่อมั่นต่ำสุด 50%)
- ปรับจำนวนกฎที่แสดงผลในกล่อง numRules เช่น 100

การประมวลผลของขั้นตอนวิธี Apriori



16 กฎที่ได้จาก market.arff

1. I5=y 2 ==> I1=y 2 conf:(1)
2. I4=y 2 ==> I2=y 2 conf:(1)
3. I5=y 2 ==> I2=y 2 conf:(1)
4. I2=y I5=y 2 ==> I1=y 2 conf:(1)
5. I1=y I5=y 2 ==> I2=y 2 conf:(1)
6. I5=y 2 ==> I1=y I2=y 2 conf:(1)
7. I1=y 6 ==> I2=y 4 conf:(0.67)
8. I3=y 6 ==> I1=y 4 conf:(0.67)
9. I1=y 6 ==> I3=y 4 conf:(0.67)
10. I3=y 6 ==> I2=y 4 conf:(0.67)
11. I2=y 7 ==> I1=y 4 conf:(0.57)
12. I2=y 7 ==> I3=y 4 conf:(0.57)
13. I2=y I3=y 4 ==> I1=y 2 conf:(0.5)
14. I1=y I3=y 4 ==> I2=y 2 conf:(0.5)
15. I1=y I2=y 4 ==> I3=y 2 conf:(0.5)
16. I1=y I2=y 4 ==> I5=y 2 conf:(0.5)

ความหมายของกฎที่ 1:การซื้อสินค้าของลูกค้าที่มีสินค้า I5 แล้วจะมีสินค้า I1 เสมอ

ความหมายของกฎที่ 2:การซื้อสินค้าของลูกค้าที่มีสินค้า I4 แล้วจะมีสินค้า I2 เสมอ

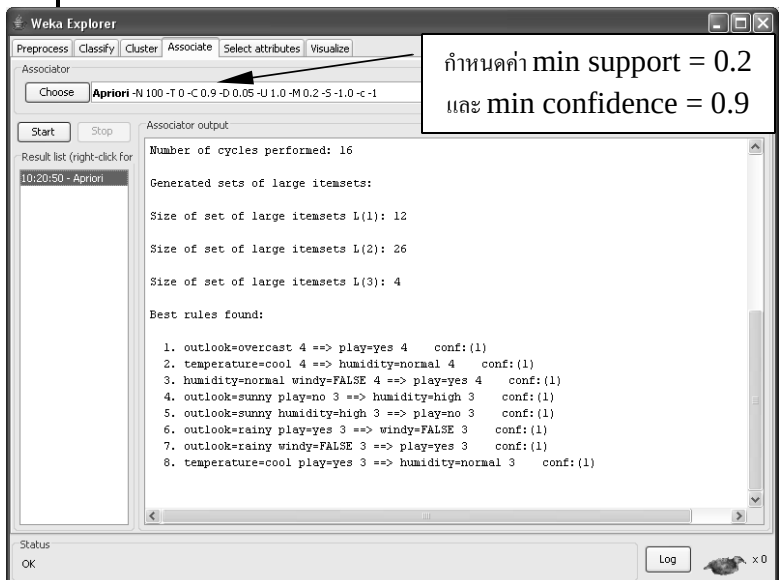
ลักษณะข้อมูลที่ไม่ใช่ตะกร้าซื้อ

- การทำเหมืองข้อมูลแบบกฎเชื่อมโยงสามารถนำไปใช้กับข้อมูลที่ไม่ใช่ transaction ได้ โดยใช้การเข้ารหัสของลักษณะประจำเป็นชนิด Nominal หรือ Ordinal
- ซอฟต์แวร์ Weka ใช้การเข้ารหัส dummy coding คือซอฟต์แวร์จะแปลงค่าของ Nominal หรือ Ordinal หนึ่งค่าแทนด้วยตัวแปรทวิภาคหนึ่งตัว เช่น
  - ลักษณะประจำ outlook มีค่าที่เป็นไปได้คือ overcast, sunny, rainy แล้วตัวแปรทวิภาคเขียนได้เป็น outlook = overcast, outlook = sunny, outlook = rainy

แฟ้ม weather.nominal.arff

```
@relation weather.symbolic
@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no
```


ผลลัพธ์ที่ได้จาก weather.nominal.arff



กำหนดค่า min support = 0.2
และ min confidence = 0.9

8 กฎที่ได้จาก weather.nominal.arff

1. outlook=overcast 4 ==> play=yes 4 conf:(1)
2. temperature=cool 4 ==> humidity=normal 4 conf:(1)
3. humidity=normal windy=FALSE 4 ==> play=yes 4 conf:(1)
4. outlook=sunny play=no 3 ==> humidity=high 3 conf:(1)
5. outlook=sunny humidity=high 3 ==> play=no 3 conf:(1)
6. outlook=rainy play=yes 3 ==> windy=FALSE 3 conf:(1)
7. outlook=rainy windy=FALSE 3 ==> play=yes 3 conf:(1)
8. temperature=cool play=yes 3 ==> humidity=normal 3 conf:(1)

ความหมายของกฎที่ 1:ถ้าสภาพอากาศเป็น overcast แล้ว play=yes เสมอ

ความหมายของกฎที่ 2:ถ้าอุณหภูมิเป็น cool แล้วความชื้นจะปรกติ (normal) เสมอ

ความหมายของกฎที่ 3:ถ้าความชื้นปรกติและไม่มีลม windy=FALSE แล้ว play=yes เสมอ

สรุป

- แฟ้มข้อมูลที่ถูกนำมาใช้ในการวิเคราะห์ต้องประกอบด้วยลักษณะประจำที่เป็น Nominal หรือ Ordinal เท่านั้น
- ข้อมูลในลักษณะ transaction เป็นข้อมูล Nominal และการไม่ซื้อใช้ ? (missing value) แทน

  TID, attri_1, attri_2, …, attri_n
  - เมื่อ TID แทนรหัสการซื้อและแต่ละ attri_i มีค่า y หรือ ?
- เลือก Associate และใช้ Apriori ใต้ Associator
- ปรับค่าพารามิเตอร์ min support กับ min confidence และ numRules ที่ต้องการแล้วสั่งให้ประมวลผล