EGCO 425 – Project 1 (Association)

ตอนที่1

CES Dataset เป็นฐานข้อมูลที่ได้มาจากการสำรวจ 1540 ครอบครัวจาก 8 เขต ที่อยู่ในที่อยู่อาศัยในบราซิล ในการไปซื้อของในห้างสรรพสินค้า โดยเก็บข้อมูลเป็น เมือง,จำนวนสมาชิกในครอบครัว,เงินเดือน และรายการซื้อสินค้า โดยเรียกแบบสำรวจนี้ว่า Consumer Expenditure Survey

โดยจะมี Transaction id เป็นตัวแสดงถึงรายการ การซื่อของในห้างสรรพสินค้าของครอบครัวหนึ่ง โดย รายการแรกบอกถึงเมืองที่อาศัย รายการสองบอกถึงเงินเดือนที่เหลือและรายการสุดท้ายคือจำนวนสมาชิกในครอบครัว(ถ้าเป็นหนึ่งแสดงว่าอยู่คนเดียว) และรายการที่เหลือเป็นรายการในการซื้อสินค้า

ยกตัวอย่าง

10002	City_Belem
10002	Income_12_to_18
10002	Members_6
10002	amazon_papaya
10002	banana
10002	beef_breast
10002	beef_rump_cap
10002	beetroot
10002	black_beans
10002	canned_olives
10002	cauliflower
10002	chocolate_powder
10002	concentrate_guarana
10002	cooked_turkey_ham
10002	cream_cracker_biscuits
10002	creamy_white_cheese
10002	cupuacu_pulp
10002	egg
10002	french_fries
10002	garlic
10002	garlic_sauce
10002	gherkin_cucumber
10002	green_collard
10002	japanese_rice
10002	ketchup
10002	lime
10002	mayonnaise
10002	melon
10002	milk_powder
10002	milky_meal
10002	sliced_mozzarella_cheese
10002	soda_bottle_large
10002	soy_oil
10002	sugar
10002	tomato
10002	vinegar
10002	white_split_tin_bread

ครอบครัวนี้อาศัยอยู่ในเมือง Belem มีรายได้12 ถึง18 ต่อเดือน มีจำนวนสมาชิกในครอบครัว 6 คน จะทำการซื้อสินค้าพวก มะละกอ กล้วย เนื้อหน้าอก ไข่ และอื่นๆ

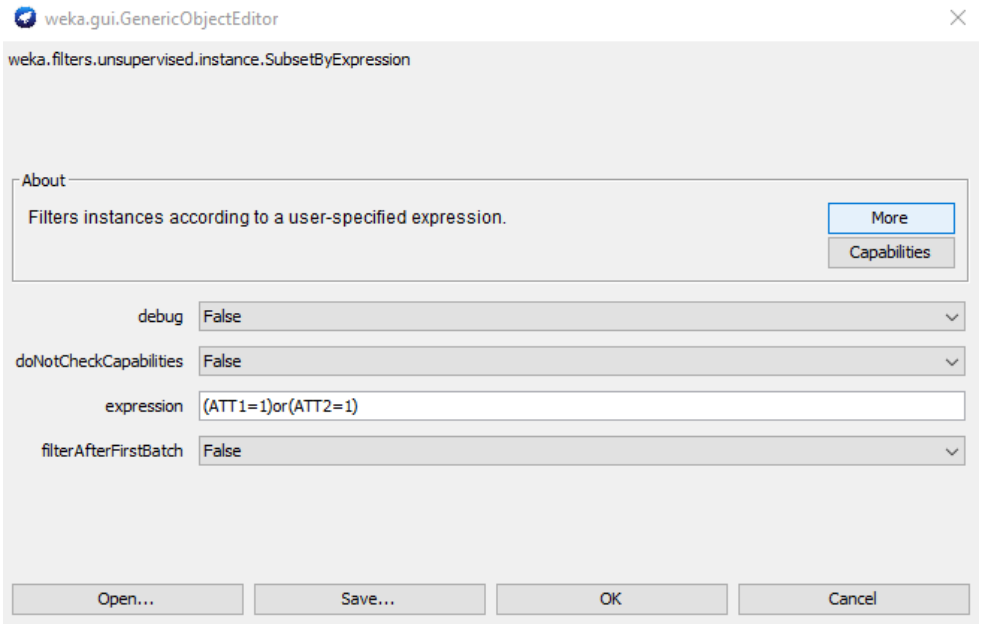
โดยตัวแปร city กับ สินค้า จะเป็น Nominal เพราะจะแยกความต่างจากตัวอักษร ตัวแปร income จะเป็นตัวแปร Ordinal เพราะมีการแยกค่าเป็นช่วงโดยที่ไม่ทราบค่าจริง เช่น Income_12to_18 จะรู้แค่ว่ารายต่อเดือนของครอบครัวนี้เป็น 12ถึง18 แต่ไม่ทราบค่าเงินเดือนจริงๆ อาจจะเป็น 13หรือ15ก็ได้ ตัวแปร member เป็น interval เพราะความต่างของจำนวนสมาชิกครอบครัวมีผลต่อการทำ association และไม่มีทางที่จำนวนสมาชิกครอบครัวเป็น 0

1	item	frequency	relative_frequency
2	eqq	1144	0.742857143
3	french_bread	1114	0.723376623
4	vinegar	1061	0.688961039
5	soy_oil	1043	0.677272727
6	garlic	975	0.633116883

จากค่าสถิติที่เราเห็น จะสามารถทำนายได้ว่าหลังจากการทำ association จะมี ไข่,ขนมปังฝรั่งเศล, น้ำส้มสายชู้,น้ำมันถั่วเหลืองออกมาเยอะ เพราะมีความถี่เยอะ

ตอนที่ 2

Parameter setup



ใช้ SubsetByExpression ในการ filter ข้อมูลที่จำเป็นออกมาโดยเราต้อง set ค่าใน expression โดยที่ ATT หมายถึง เลือก attribute ที่เราต้องการ filter ข้อมูล ตัวอย่างเช่น ATT1 = 1 หมายความว่า ทำการเลือก attribute ที่ 1 ว่าจะนำค่าที่เป็น 1 ออกมา และattributeที่นำมา expression ต้องเป็นชนิด numeric

Assosiation's parameter

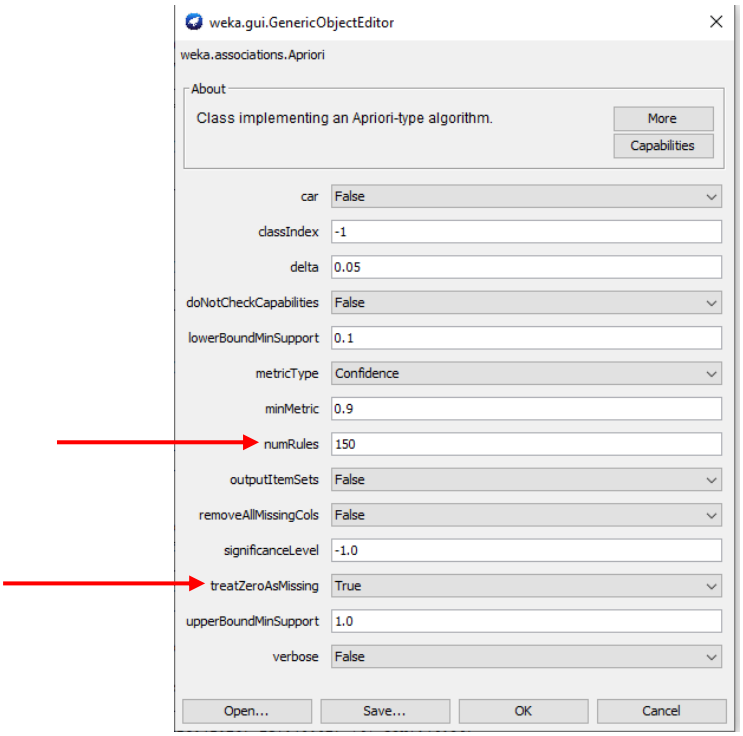
ค่าที่ได้หลังจาก association แล้วนำมาพิจรณาคือ

ค่า confident คือค่าความมั่นใจที่เกิด LHS แล้วจะเกิด RHS

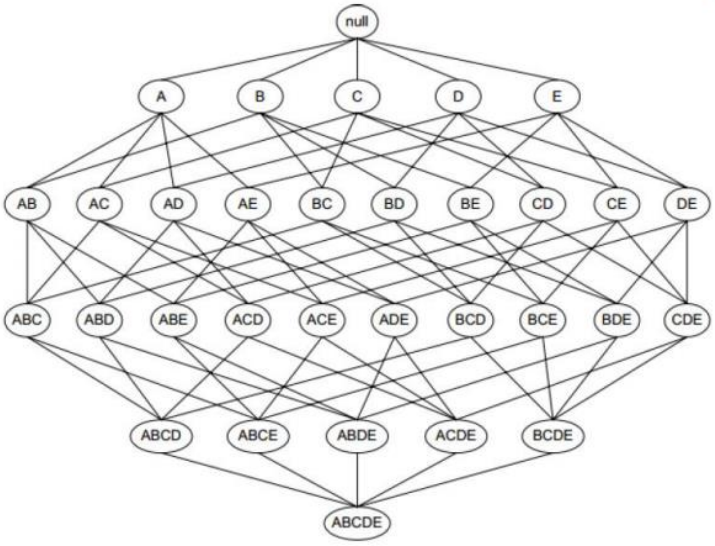
ค่า lift และ leverage คือค่าความเกี่ยวข้องของ LHS กับ RHS

ค่า conviction คือค่าความขัดแย้งที่เกิดLHS แต่จะไม่เกิด RHS

การตั้งค่า Apriori



1.ทำการเซ็ทค่าตัว treatZeroAsMissing ใหม่จาก false เป็น true เพื่อลดการนำค่า 0 มาคิดเนื่องจากจะทำให้ค่า Apriori ออกมามีค่ามากเกินไป เพื่อการลด runtime และการใช้งาน ram เพราะการทำ Apriori จะใช้ algorithm เป็นแบบ brute force ในการหาความสัมพันธ์ของ itemsets



2.เปลี่ยนค่า numRules ให้มีจำนวนมากขึ้นเพื่อที่เราจะมีค่า apriori มาใช้ในการวิเคราะห์ได้มากขึ้น (หา Bestrule)

ค่า Income ที่น้อยกว่า 5 หลังจากทำการ filter ข้อมูล

No.	1: Income_below_2.5 Numeric	2: Income_2.5_to_5 Numeric	3: acai_berry Nominal	4: acai_berry_pulp Nominal	5: ajinomoto_sauce Nominal	6: alphabet_pasta Nominal	7: amazon_papaya Nominal
4	0.0	1.0	0	0	0	0	0
5	1.0	0.0	0	0	0	0	0
6	1.0	0.0	0	0	0	0	0
7	0.0	1.0	0	0	0	0	0
8	0.0	1.0	0	0	0	0	1
9	1.0	0.0	0	0	0	0	0
10	0.0	1.0	0	0	0	0	0
11	1.0	0.0	0	0	0	0	0
12	0.0	1.0	0	0	0	0	0
13	0.0	1.0	0	0	0	0	0
14	0.0	1.0	0	0	0	0	0
15	1.0	0.0	0	0	0	0	0
16	1.0	0.0	0	0	0	0	0
17	1.0	0.0	0	0	0	0	0
18	1.0	0.0	0	0	0	0	0
19	0.0	1.0	0	0	0	0	1
20	0.0	1.0	0	0	0	0	0
21	0.0	1.0	0	0	0	0	0
22	0.0	1.0	0	0	1	0	1
23	0.0	1.0	0	0	0	0	1
24	1.0	0.0	0	0	0	0	0
25	0.0	1.0	0	0	1	0	1
26	0.0	1.0	0	0	0	0	0
27	0.0	1.0	0	0	0	0	0
28	0.0	1.0	0	0	0	0	0
29	1.0	0.0	0	0	0	0	0
30	0.0	1.0	0	0	0	0	0

ค่า best rule Income น้อยกว่า 5 ที่ทำการเลือกมา 3 ตัว ได้แก่

1. chocolate_powder=1 popping_corn=1 35 ==> egg=1 33

conf:(0.94) lift:(1.27) lev:(0.02) [7]

2. egg=1 garlic=1 milk_in_plastic_bag_category_C=1 34 ==> french_bread=1 32

conf:(0.94) lift:(1.35) lev:(0.03) [8]

3. egg=1 garlic=1 milk_in_plastic_bag_category_C=1 34 ==> soy_oil=1 32

conf:(0.94) lift:(1.19) lev:(0.02) [5]

จาก best rule ที่ได้เลือกมานั้น เราสามารถทำนายได้ว่า ถ้าเลือกซื้อ ผงช็อกโกแล็ต และป๊อปคอน จะต้องซื้อไข่ด้วย และ ถ้าเลือกซื้อ ไข่ กระเทียม และนมถุง จะต้องซื้อขนมปังฝรั่งเศส หรือน้ำมันถั่วเหลืองด้วย

ค่า Income ที่มากกว่า 25 หลังจากทำการ filter ข้อมูล

Viewer

Relation: ces_hybrid-weka.filters.unsupervised.attribute.Remove-R1-21-weka.filters.unsupervised.instance.SubsetByExpression-E(ATT1 = 1) or (ATT2 =

No.	1: Income_25_to_43 Numeric	2: Income_above_43 Numeric	3: acai_berry Nominal	4: acai_berry_pulp Nominal	5: ajinomoto_sauce Nominal	6: alphabet_pasta Nominal	7: amazon_papaya Nominal
1	1.0	0.0	0	0	0	0	0
2	1.0	0.0	0	0	0	0	1
3	1.0	0.0	0	0	0	0	0
4	1.0	0.0	0	0	0	0	0
5	0.0	1.0	0	0	0	0	1
6	1.0	0.0	1	0	0	0	0
7	1.0	0.0	0	0	0	0	0
8	0.0	1.0	0	0	0	0	1
9	1.0	0.0	0	0	0	0	0
10	1.0	0.0	0	0	0	0	0
11	1.0	0.0	0	0	0	0	0
12	1.0	0.0	0	0	0	0	0
13	1.0	0.0	0	0	0	0	1
14	1.0	0.0	0	0	0	0	0
15	0.0	1.0	0	0	0	0	0
16	1.0	0.0	0	0	0	0	1
17	1.0	0.0	1	0	0	0	0
18	1.0	0.0	1	0	0	0	0
19	0.0	1.0	0	0	0	0	0
20	1.0	0.0	0	0	0	0	0
21	0.0	1.0	0	0	0	0	0
22	1.0	0.0	0	0	0	1	1
23	1.0	0.0	0	0	0	0	0
24	0.0	1.0	0	0	0	0	0
25	1.0	0.0	0	0	0	0	0
26	0.0	1.0	0	0	0	0	0
27	1.0	0.0	0	0	0	0	0
28	0.0	1.0	0	0	0	0	0
29	1.0	0.0	0	0	0	0	0
30	1.0	0.0	0	0	0	0	0

ค่า best rule Income มากกว่า 25 ที่ทำการเลือกมา 3 ตัว ได้แก่

1.egg=1 french_bread=1 guava_candy=1 45 ==> mayonnaise=1 42

conf:(0.93) lift:(1.32) lev:(0.04) [10]

2. egg=1 guava_candy=1 mayonnaise=1 45 ==> french_bread=1 42

conf:(0.93) lift:(1.29) lev:(0.04) [9]

3. egg=1 mayonnaise=1 whipped_cream=1 46 ==> vinegar=1 42

conf:(0.91) lift:(1.3) lev:(0.04) [9]

จาก best rule ที่ได้เลือกมานั้น เราสามารถทำนายได้ว่า เมื่อผู้มีรายได้มากว่า 25 ทำการซื้อ ไข่ ขนมปังฝรั่งเศส และลูกอมรสฝรั้ง จะทำการซื้อมายองเนสด้วย และ ถ้าทำการเลือกซื้อ ไข่ ลูกอมรสฝรั้ง และมายองเนส จะทำการซื้อขนมปังฝรั่งเศสด้วย และ ถ้าทำการซื้อ ไข่ มายองเนส และวิปครีม จะทำการซื้อน้ำส้มสายชูด้วย

สรุปจาก best rule ทั้งของ Income น้อยกว่า 5 และ Income มากกว่า 25 จะสังเกตได้ว่าทั้ง 2 กลุ่มจะนิยมซื้อไข่ และขนมปังฝรั่งเศส แต่ที่ต่างกันคือกลุ่มที่มี Income น้อยกว่า 5 จะนิยมเลือกซื้อน้ำมันถั่วเหลือง และกระเทียม ส่วนกลุ่มที่มี Income มากกว่า 25 จะนิยมเลือกซื้อน้ำส้มสายชู และมายองเนส

ตอนที่ 3

ค่า Members <=2 หลังจากทำการ filter ข้อมูล

No.	1: Members_1 Numeric	2: Members_2 Numeric	3: acai_berry Nominal	4: acai_berry_pulp Nominal	5: ajinomoto_sauce Nominal	6: alphabet_pasta Nominal	7: amazon_papaya Nominal
1	0.0	1.0	0	0	0	0	0
2	0.0	1.0	0	0	0	0	0
3	0.0	1.0	1	0	1	0	0
4	0.0	1.0	0	0	0	0	0
5	0.0	1.0	0	0	0	0	0
6	0.0	1.0	0	0	0	0	0
7	0.0	1.0	0	0	0	0	1
8	0.0	1.0	0	0	0	0	1
9	0.0	1.0	0	0	0	0	0
10	0.0	1.0	0	0	0	0	0
11	0.0	1.0	0	0	0	0	0
12	0.0	1.0	0	0	0	0	0
13	0.0	1.0	0	0	0	0	0
14	0.0	1.0	0	0	0	0	0
15	0.0	1.0	1	0	0	0	1
16	0.0	1.0	0	0	0	0	0
17	0.0	1.0	0	0	0	0	0
18	0.0	1.0	0	0	0	0	0
19	0.0	1.0	0	0	0	0	0
20	0.0	1.0	0	0	0	0	0
21	1.0	0.0	0	0	1	0	0
22	0.0	1.0	0	0	0	0	0
23	0.0	1.0	0	0	0	0	0
24	0.0	1.0	0	0	0	0	1
25	0.0	1.0	0	0	0	0	0
26	0.0	1.0	0	0	0	0	0
27	0.0	1.0	0	0	0	0	0
28	0.0	1.0	0	0	0	0	0
29	0.0	1.0	0	0	0	0	0
30	0.0	1.0	0	0	0	0	0

ค่า best rule ค่า Members <=2 ที่ทำการเลือกมา 3 ตัว ได้แก่

1. chocolate_powder=1 french_bread=1 garlic=1 mayonnaise=1 34 ==> vinegar=1 33

conf:(0.97) lift:(1.54) lev:(0.04) [11]

2. chocolate_powder=1 mayonnaise=1 vinegar=1 wheat_flour_special=1 33 ==> egg=1 32

conf:(0.97) lift:(1.39) lev:(0.03) [8]

3. french_bread=1 green_beans=1 wheat_flour_special=1 37 ==> egg=1 36

conf:(0.97) lift:(1.39) lev:(0.03) [10]

จาก best rule ที่ได้เลือกมานั้น เราสามารถทำนายได้ว่า ถ้าเลือกซื้อ ผงซ็อกโกแล็ต ขนมปังฝรั่งเศส หัวหอม และมายองเนส จะทำการซื้อน้ำส้มสายชูด้วย และถ้าเลือกซื้อ ผงซ็อกโกแล็ต มายองเนส น้ำส้มสายชู และแป้งสาลีพิเศษ จะทำการซื้อไข่ด้วย และถ้าเลือกซื้อขนมปังฝรั่งเศส ถั่วเขียว และแป้งสาลีพิเศษ จะทำการซื้อไข่ด้วย

ค่า Members >= 6 หลังจากทำการ filter ข้อมูล

Viewer

Relation: ces_hybrid-weka.filters.unsupervised.attribute.Remove-R1-13,16-23-weka.filters.unsupervised.instance.SubsetByExpression-E(ATT1 =

No.	1: Members_6 Numeric	2: Members_above_6 Numeric	3: acai_berry Nominal	4: acai_berry_pulp Nominal	5: ajinomoto_sauce Nominal	6: alphabet_pasta Nominal	7: amazon_papaya Nominal
1	1.0	0.0	0	0	0	0	1
2	1.0	0.0	1	0	0	0	0
3	1.0	0.0	1	0	0	0	0
4	1.0	0.0	0	0	0	0	0
5	1.0	0.0	0	0	0	0	0
6	1.0	0.0	0	0	0	0	0
7	1.0	0.0	0	0	0	0	0
8	1.0	0.0	0	0	0	0	0
9	0.0	1.0	0	0	0	0	0
10	1.0	0.0	0	0	1	0	0
11	0.0	1.0	0	0	0	0	0
12	0.0	1.0	1	0	0	0	0
13	0.0	1.0	0	0	0	0	1
14	1.0	0.0	0	0	0	0	0
15	1.0	0.0	1	0	0	0	0
16	1.0	0.0	1	0	0	0	0
17	0.0	1.0	0	0	0	0	0
18	0.0	1.0	0	0	1	0	0
19	1.0	0.0	0	0	1	0	0
20	0.0	1.0	0	0	0	0	1
21	0.0	1.0	0	0	0	0	0
22	0.0	1.0	0	0	1	1	1
23	1.0	0.0	0	0	0	0	0
24	1.0	0.0	0	0	0	0	0
25	1.0	0.0	0	0	0	0	0
26	1.0	0.0	0	0	0	0	1
27	1.0	0.0	0	0	0	0	1
28	0.0	1.0	0	0	0	0	0
29	0.0	1.0	0	0	0	0	0
30	1.0	0.0	0	0	0	0	0

ค่า best rule ค่า Members >= 6 ที่ทำการเลือกมา 3 ตัว ได้แก่

1.egg=1 mayonnaise=1 popping_corn=1 vinegar=1 47 ==> french_bread=1 44

conf:(0.94) lift:(1.22) lev:(0.03) [8]

2.chocolate_powder=1 french_bread=1 garlic=1 mayonnaise=1 soy_oil=1 46 ==> vinegar=1 44

conf:(0.96) lift:(1.32) lev:(0.04) [10]

3.chocolate_powder=1 ketchup=1 soy_oil=1 vinegar=1 42 ==> mayonnaise=1 39

conf:(0.93) lift:(1.52) lev:(0.05) [13]

จาก best rule ที่ได้เลือกมานั้น เราสามารถทำนายได้ว่า ถ้าเลือกซื้อ ไข่ ป็อปคอน และน้ำส้มสายชู จะทำการซื้อขนมปังฝรั่งเศสด้วยด้วย และถ้าเลือกซื้อ ผงซ็อกโกแล็ต มายองเนส ขนมปังฝรั่งเศส และหัวหอม จะทำการซื้อน้ำส้มสายชูด้วย และถ้าเลือกซื้อผงซ็อกโกแล็ต ซอสมะเขือเทศ น้ำมันถั่วเหลือง และน้ำส้มสายชู จะทำการซื้อมายองเนสด้วย

สรุปจาก best rule ทั้งของ Members <=2 และ Members >= 6 จะสังเกตได้ว่าทั้ง 2 กลุ่มจะนิยมซื้อผงซ็อกโกแล็ต ขนมปังฝรั่งเศส และหัวหอม ส่วนกลุ่มที่ Members >= 6 จะนิยมซื้อน้ำส้มสายชู และมายองเนส ส่วนกลุ่มที่ Members <=2 จะนิยมซื้อถั่วเขียว และแป้งข้าวสาลี

ตอนที่ 4

สรุปไม่ว่าจะมีรายได้ในช่วงไหน หรือจำนวนสมาชิกเท่าไหร่ก็จะนิยมซื้อขนมปังฝรั่งเศสกับไข่ ส่วนถ้าดูตามรายได้จะนิยมซื้อน้ำส้มสายชู ไวน์และมายองเนส ส่วนถ้าดูตามจำนวนสมาชิกในครอบครัวจะนิยมซื้อถั่วเขียว เนย โยเกิร์ตผลไม้ และแป้งข้าวสาลี

อ้างอิง

Gonçalves, E. C. (2014). A Human-Centered Approach for Mining Hybrid-Dimensional Association Rules. Proceedings of the 17th International Conference on Information Fusion, (FUSION 2014), Salamanca, Spain.