

CSE 634/590 Data mining

Submitted By:
Moieed Ahmed

Original Data

Student	Grade	Income	Buys
CS	High	Low	Milk
CS	High	High	Bread
Math	Low	Low	Bread
CS	Medium	High	Milk
Math	Low	Low	Bread

Converted Data

Student = CS (I1)	Student =math (I2)	Grade = high (I3)	Grade =medium (I4)	Grade =low (I5)	Income =high (I6)	Income =low (I7)	Buys=milk (I8)	Buys =bread (I9)
+	-	+	-	-	-	+	+	-
+	-	+	-	-	+	-	-	+
-	+	-	-	+	-	+	-	+
+	-	-	+	-	+	-	+	-
-	+	-	-	+	-	+	-	+

Generating 1-itemset Frequent Pattern

Scan D for count of each candidate

Item Set	Support Count
{I1}	3
{I2}	2
{I3}	2
{I4}	1
{I5}	2
{I6,}	2
{I7}	3
{I8}	2
{I9}	3

Compare candidate support count with minimum support count

Item Set	Support Count
{I1}	3
{I2}	2
{I3}	2
{I5}	2
{I6,}	2
{I7}	3
{I8}	2
{I9}	3

C1

L1

Let, the minimum support count be 2.

Since we have 5 records \Rightarrow (Support) = $2/5 = 40\%$

Let, minimum confidence required is **70%**.

Generating 2-itemset Frequent Pattern

Generate C2 candidates from L1

Item Set
{11,12}
{11,13}
{11,14}
{11,15}
{11,16}
{11,17}
{11,18}
{11,19}
{12,13}
{12,14}
{12,15}
{12,16}
{12,17}
{12,18}
{12,19}
{13,14}
{13,15}
{13,16}
{13,17}
{13,18}
{13,19}
{14,15}
{14,16}
{14,17}
{14,18}
{14,19}
{15,16}
{15,17}
{15,18}
{15,19}
{16,17}
{16,18}
{16,19}
{17,18}
{17,19}
{18,19}

Scan D for count of each candidate

Item Set	Support Count
{11,12}	0
{11,13}	2
{11,14}	1
{11,15}	0
{11,16}	2
{11,17}	1
{11,18}	2
{11,19}	1
{12,13}	0
{12,14}	0
{12,15}	2
{12,16}	0
{12,17}	2
{12,18}	0
{12,19}	2
{13,14}	0
{13,15}	0
{13,16}	1
{13,17}	1
{13,18}	1
{13,19}	1
{14,15}	0
{14,16}	1
{14,17}	0
{14,18}	1
{14,19}	0
{15,16}	0
{15,17}	2
{15,18}	0
{15,19}	2
{16,17}	0
{16,18}	1
{16,19}	0
{17,18}	1
{17,19}	2
{18,19}	0

Compare candidate support count with minimum support count

Item Set	Support Count
{11,13}	2
{11,16}	2
{11,18}	2
{12,15}	2
{12,17}	2
{12,19}	2
{15,17}	2
{15,19}	2
{17,19}	2

L2

C2

C2

Joining and Pruning

1. **The join step:** To find L_k , a set of candidate k -itemsets is generated by joining L_{k-1} with itself. This set of candidates is denoted C_k .

L_k – Itemsets C_k – Candidates

Considering $\{I_2, I_5\}$, $\{I_7, I_9\}$ from L_2 to arrive at L_3 we Join $L_2 \times L_2$

And thus we have $\{I_2, I_5, I_7\}$, $\{I_2, I_5, I_9\}$ in the resultant candidates generated from L_2

Considering $\{I_1, I_3\}$, $\{I_1, I_6\}$ from L_2 we generate candidates $\{I_1, I_3, I_6\}$

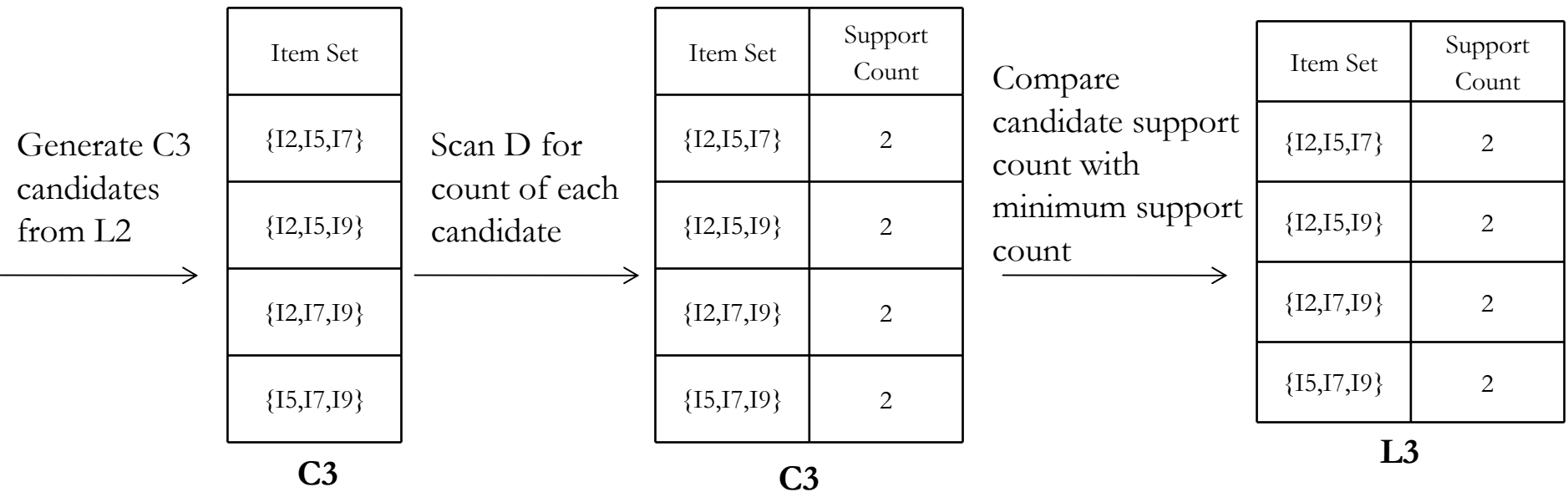
2. **The prune step:**

C_k is a superset of L_k , that is, its members may or may not be frequent

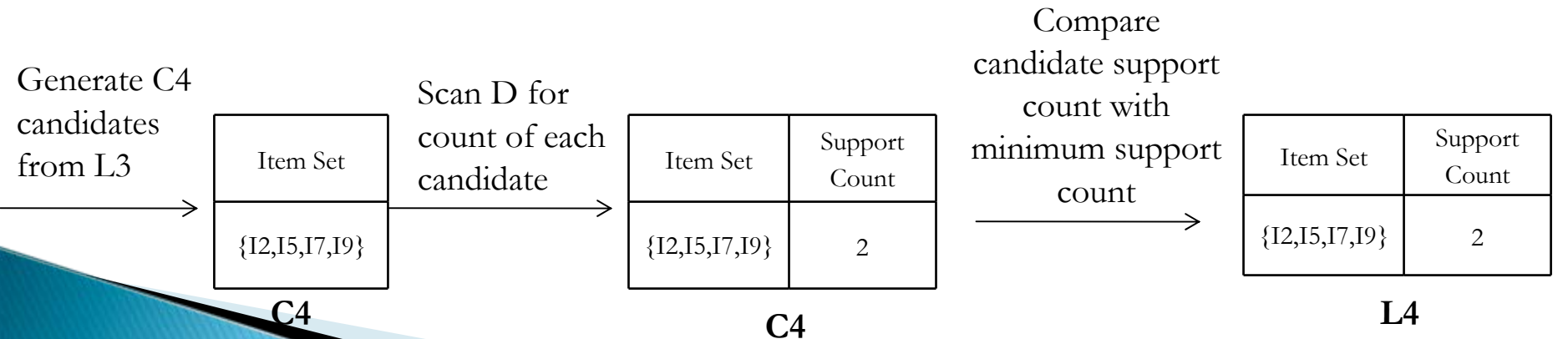
All candidates having a count no less than the minimum support count are frequent by definition, and therefore belong to L_k . C_k , however, can be huge

Thus, $\{I_2, I_5, I_7\}$, $\{I_2, I_5, I_9\}$ from join step are considered since they have minimum support but $\{I_1, I_3, I_6\}$ is discarded since it does not meet the support count needed.

Generating 3-itemset Frequent Pattern



Generating 4-itemset Frequent Pattern



Generating Association Rules by Classification

- When mining association rules for use in classification, we are only interested in association rules of the form
- $(p_1 \wedge p_2 \wedge \dots \wedge p_l) \rightarrow \text{Aclass} = C$ where the rule antecedent is a conjunction of items, p_1, p_2, \dots, p_l ($l \leq n$), associated with a class label, C .
- ▶ In our example Aclass would be either (I8 or I9 on RHS) that is to predict whether a student with given characteristics buys Milk / Bread.
- ▶ Let, minimum confidence required be 70%
- ▶ Considering, $l = \{I_2, I_5, I_7, I_9\}$
- ▶ It's nonempty subsets are $\{\{2\}, \{5\}, \{7\}, \{9\}, \{2,5\}, \{2,7\}, \{2,9\}, \{5,7\}, \{5,9\}, \{7,9\}, \{2,5,7\}, \{2,5,9\}, \{2,7,9\}, \{5,7,9\}\}$

Generating Association Rules by Classification

- ▶ $R1 : I2 \wedge I5 \wedge I7 \rightarrow I9$ [40%,100%]
 - Confidence = $sc\{I2,I5,I7,I9\} / sc\{I2,I5,I7\} = 2/2 = 100\%$
 - R1 is **Selected**
- ▶ **Considering 3 itemset Frequent Pattern**
- ▶ $R2 : I5 \wedge I7 \rightarrow I9$ [40%,100%]
 - Confidence = $sc\{I5,I7,I9\} / sc\{I5,I7\} = 2/2 = 100\%$
 - R2 is **Selected**
- ▶ $R3 : I2 \wedge I7 \rightarrow I9$ [40%,100%]
 - Confidence = $sc\{I2,I7,I9\} / sc\{I2,I7\} = 2/2 = 100\%$
 - R3 is **Selected**
- ▶ $R4 : I2 \wedge I5 \rightarrow I9$ [40%,100%]
 - Confidence = $sc\{I2,I7,I9\} / sc\{I2,I7\} = 2/2 = 100\%$
 - R4 is **Selected**

Generating Association Rules by Classification

Considering 2 itemset Frequent Pattern

- ▶ R5 : I5 → I9 [40%,100%]
 - Confidence = $sc\{I5,I9\} / sc\{I9\} = 2/2 = 100\%$
 - R5 is **Selected**
- ▶ R6 : I2 → I9 [40%,100%]
 - Confidence = $sc\{I2,I9\} / sc\{I9\} = 2/2 = 100\%$
 - R6 is **Selected**
- ▶ R7 : I7 → I9 [40%,100%]
 - Confidence = $sc\{I7,I9\} / sc\{I9\} = 2/2 = 100\%$
 - R7 is **Selected**
- ▶ R8 : I1 → I8 [40%, 66%]
 - Confidence = $sc\{I1,I8\} / sc\{I1\} = 2/3 = 66.66\%$
 - R8 is **Rejected**

List of Selected Rules by Classification

- ▶ $I2 \wedge I5 \wedge I7 \rightarrow I9$ [40%,100%]
- ▶ $I2 \wedge I5 \rightarrow I9$ [40%,100%]
- ▶ $I2 \wedge I7 \rightarrow I9$ [40%,100%]
- ▶ $I5 \wedge I7 \rightarrow I9$ [40%,100%]
- ▶ $I5 \rightarrow I9$ [40%,100%]
- ▶ $I7 \rightarrow I9$ [40%,100%]
- ▶ $I2 \rightarrow I9$ [40%,100%]

- ▶ We reduce the confidence to 66% to include I8 on R.H.S
- ▶ $I1 \rightarrow I8$ [40%,66%]

Test Data

Student	Grade	Income	Buys
Math	Low	Low	Bread
CS	Low	Low	Milk
Math	Low	Low	Milk
Math	Low	Low	Bread
CS	Medium	High	Milk

- **First Tuple:**

Can be written as $I2 \wedge I5 \wedge I7 \rightarrow I9$ **[Success]**

The above rule is correctly classified

And hence the Math student with low grade and low income buys bread

- **Second Tuple:**

Can be written as $I1 \rightarrow I8$ **[Success]**

The above rule is not correctly classified

- **Third Tuple:**

Can be written as $I2 \wedge I5 \wedge I7 \rightarrow I8$ **[Error]**

The above rule is not classified

Test Data

Student	Grade	Income	Buys
Math	Low	Low	Bread
CS	Low	Low	Milk
Math	Low	Low	Milk
Math	High	Low	Bread
CS	Medium	High	Bread

- **Fourth Tuple:**

Can be written as $I_2 \wedge I_7 \rightarrow I_9$ [Success]

The above rule is correctly classified

And hence the Math student with low grade and low income buys bread

- **Fifth Tuple:**

Can be written as $I_1 \rightarrow I_9$ [Success]

The above rule is correctly classified

Hence we have **80%** predictive accuracy.

And 20% Error rate