

CSE 634/590 Data mining

Extra Credit:

Classification by Association rules:

Example Problem

Muhammad Asiful Islam,

SBID: 106506983

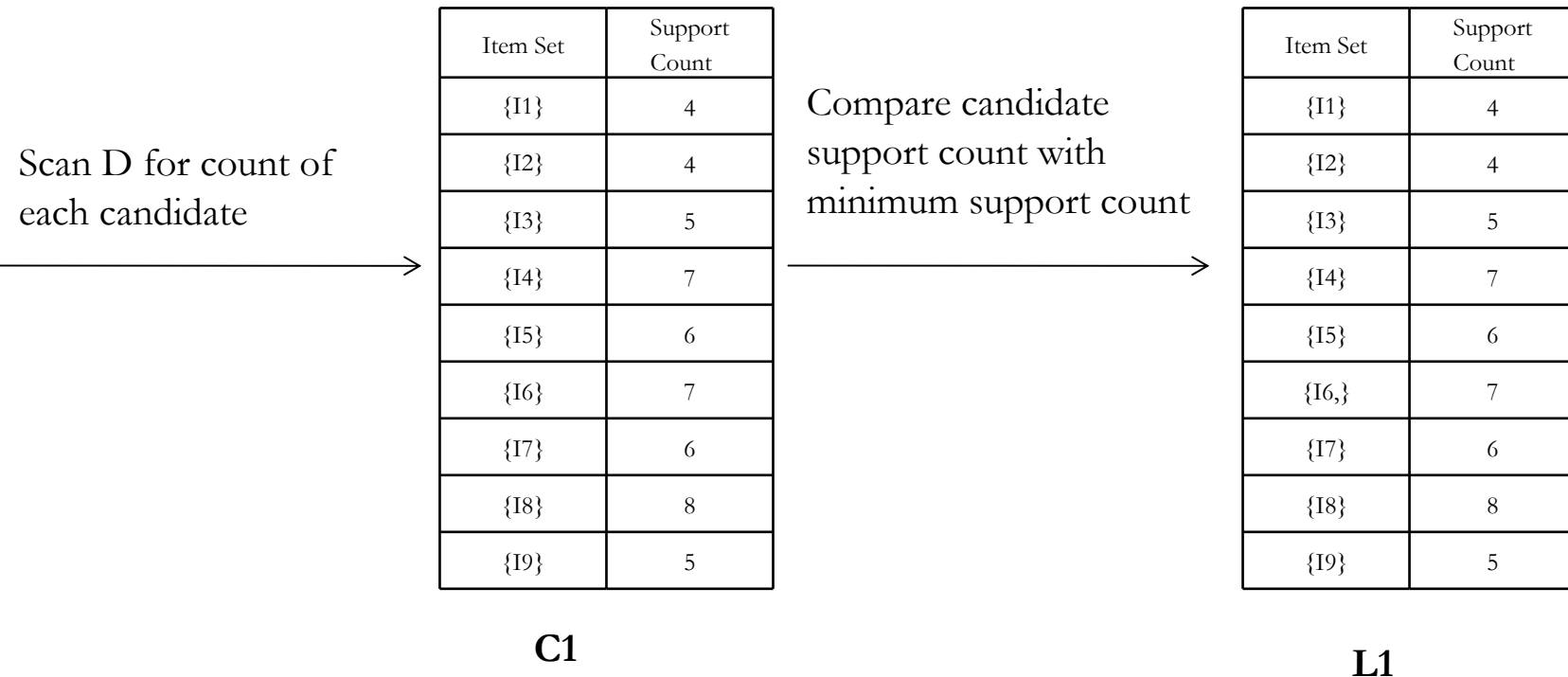
Original Data

Outlook	Humidity	Wind	PlayTennis
Sunny	High	Weak	No
Sunny	High	Strong	No
Overcast	Normal	Weak	Yes
Rain	High	Weak	Yes
Rain	Normal	Weak	Yes
Rain	High	Strong	No
Overcast	Normal	Strong	Yes
Sunny	High	Weak	No
Rain	Normal	Weak	Yes
Sunny	Normal	Strong	Yes
Overcast	High	Strong	Yes
Overcast	Normal	Weak	Yes
Rain	High	Strong	No

Converted Data

Outlook =Sunny (I1)	Outlook =Overcast (I2)	Outlook = Rain (I3)	Humidity =High (I4)	Humidity =Normal (I5)	Wind =Weak (I6)	Wind =Strong (I7)	PlayTennis =Yes (I8)	PlayTennis =No (I9)
+	-	-	+	-	+	-	-	+
+	-	-	+	-	-	+	-	+
-	+	-	-	+	+	-	+	-
-	-	+	+	-	+	-	+	-
-	-	+	-	+	+	-	+	-
-	-	+	+	-	-	+	-	+
-	+	-	-	+	-	+	+	-
+	-	-	+	-	+	-	-	+
-	-	+	-	+	+	-	+	-
+	-	-	-	+	-	+	+	-
-	+	-	+	-	-	+	+	-
-	+	-	-	+	+	-	+	-
-	-	+	+	-	-	+	-	+

Generating 1-itemset Frequent Pattern



Let, the minimum support count be 4.

So, $\text{min_sup} = 4/13 = 30.76\%$

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}

C2

Scan D for count of each candidate

Item Set	Support Count
{11,12}	0
{11,13}	0
{11,14}	3
{11,15}	1
{11,16}	2
{11,17}	2
{11,18}	1
{11,19}	3
{12,13}	0
{12,14}	1
{12,15}	3
{12,16}	2
{12,17}	2
{12,18}	4
{12,19}	0
{13,14}	3
{13,15}	2
{13,16}	3
{13,17}	2
{13,18}	3
{13,19}	2
{14,15}	0
{14,16}	3
{14,17}	4
{14,18}	2
{14,19}	5
{15,16}	4
{15,17}	2
{15,18}	6
{15,19}	0
{16,17}	0
{16,18}	5
{16,19}	2
{17,18}	3
{17,19}	3
{18,19}	0

C2

Compare candidate support count with minimum support count

Item Set	Support Count
{12,18}	4
{14,17}	4
{14,19}	5
{15,16}	4
{15,18}	6
{16,18}	5

L2

Join Operation

$L2 = \{\{I2, I8\}, \{I4, I7\}, \{I4, I9\}, \{I5, I6\}, \{I5, I8\}, \{I6, I8\}\}$.

-To find **C3**, we compute $L2 \text{ join } L2$

-Similar to natural join operation in Data Base

Join Operation illustrated:

-Consider an itemset $\{I2, I8\}$ from **L2**,

-Now search for other itemsets containing I2 or I8 in **L2**;

-this gives us itemsets $\{I5, I8\}$ and $\{I6, I8\}$

$$\boxed{\{I2, I8\}} \text{ join } \boxed{\{I5, I8\}} = \boxed{\{I2, I5, I8\}}$$

So, it's like joining two database tuples with common column value I8

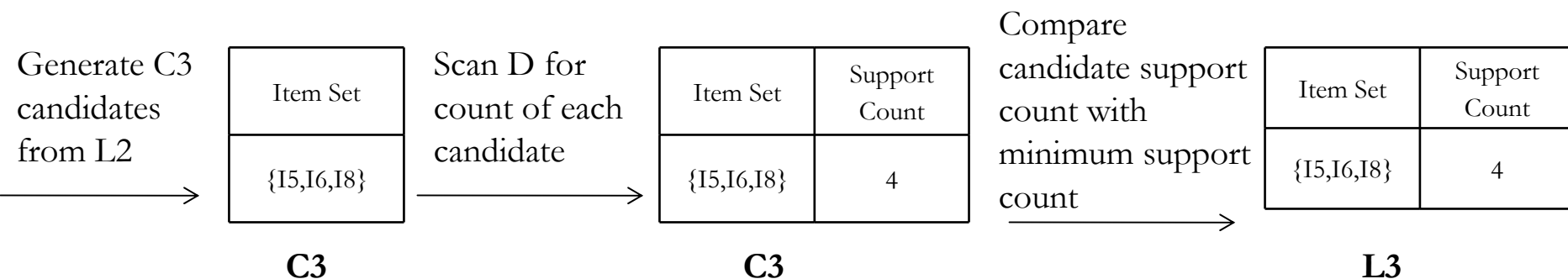
So, $\{I2, I8\}$ joining with $\{I5, I8\}$ results $\{I2, I5, I8\}$

and, $\{I2, I8\}$ joining with $\{I6, I8\}$ results $\{I2, I6, I8\}$

Similarly, using the same technique for every itemset in L2 we get;-

-**C3** = $L2 \text{ Join } L2 = \{\{I2, I5, I8\}, \{I2, I6, I8\}, \{I4, I7, I9\}, \{I5, I6, I8\}\}$.

Generating 3-itemset Frequent Pattern



-Use Apriori Property -- all subsets of a frequent itemset must also be frequent

-C3= L2 Join L2 = $\{\{I2, I5, I8\}, \{I2, I6, I8\}, \{I4, I7, I9\}, \{I5, I6, I8\}\}$.

-For example , lets take $\{I5, I6, I8\}$.The 2-item subsets of it are $\{I5, I6\}$, $\{I5, I8\}$ & $\{I6, I8\}$. Since all 2-item subsets of $\{I5, I6, I8\}$ are members of L2, We will keep $\{I5, I6, I8\}$ in C3.

-Lets take another example of $\{I2, I5, I8\}$ which shows how the pruning is performed. The 2-item subsets are $\{I2, I5\}$, $\{I2, I8\}$ & $\{I5,I8\}$.

-BUT, $\{I2, I5\}$ is not a member of L2and hence it is not frequent, violating Apriori Property. Thus We will have to remove $\{I2, I5, I8\}$ from C3.

Therefore, C3= $\{\{I5, I6, I8\}\}$ after checking for all members of result of Join operation for Pruning.

Generating 4-itemset Frequent Pattern

- The algorithm uses L3 Join L3 to generate a candidate set of 4-itemsets, C4.
- Thus, $C4 = \varnothing$, and algorithm terminates, having found all of the frequent items. This completes our Apriori Algorithm.
- Next ?
 - These frequent itemsets will be used to generate strong association rules (where strong association rules satisfy both minimum support & minimum confidence).
- We had $L = \{\{I1\}, \{I2\}, \{I3\}, \{I4\}, \{I5\}, \{I6\}, \{I7\}, \{I8\}, \{I9\}, \{I2,I8\}, \{I4,I7\}, \{I4,I9\}, \{I5,I6\}, \{I5,I8\}, \{I6,I8\}, \{I5,I6,I8\}\}$.
- Now, to generate classification rules, we need to consider only the frequent item sets which contains I8 or I9.

Generating Association Rules from Frequent Itemsets

Consider only the frequent item sets which contains I8 or I9: {I2,I8}, {I4,I9}, {I5,I8}, {I6,I8}, {I5,I6,I8}

- Consider {I2,I8}
- R1: $I2 \rightarrow I8$
 - Confidence = $sc\{I2,I8\}/sc\{I2\} = 4/4 = 100\%$ selected.
- Consider {I4,I9}
- R2: $I4 \rightarrow I9$
 - Confidence = $sc\{I4,I9\}/sc\{I4\} = 5/7 = 71.42\%$ selected.
- Consider {I5,I8}
- R3: $I5 \rightarrow I8$
 - Confidence = $sc\{I5,I8\}/sc\{I5\} = 6/6 = 100\%$ selected.
- Consider {I6,I8}
- R4: $I6 \rightarrow I8$
 - Confidence = $sc\{I6,I8\}/sc\{I6\} = 5/7 = 71.42\%$ selected.
- Consider {I5,I6,I8}
- R5: $I5 \wedge I6 \rightarrow I8$
 - Confidence = $sc\{I5,I6,I8\}/sc\{I5,I6\} = 4/4 = 100\%$ selected.

Classification Rules

- Rule (A→B) [support, confidence]
- R1: I2 → I8 [30.76%, 100%]
 - R2: I4 → I9 [38.46%, 71.42%]
 - R3: I5 → I8 [46.15%, 100%]
 - R4: I6 → I8 [38.46%, 71.42%]
 - R6: I5 ^ I6 → I8 [30.76%, 100%]

Alternatively:-

1. Outlook = overcast → PlayTennis = yes [30.76%, 100%]
2. Humidity = high → PlayTennis = no [38.46%, 71.42%]
3. Humidity = normal → PlayTennis = yes [46.15%, 100%]
4. Wind = weak → PlayTennis = yes [38.46%, 71.42%]
5. Humidity = normal AND Wind = weak → PlayTennis = yes [30.76%, 100%]

Test Data Set

Outlook	Humidity	Wind	PlayTennis
Overcast	Normal	Strong	Yes
Rain	Normal	Weak	Yes
Rain	Normal	Strong	No
Sunny	Normal	Weak	Yes
Sunny	High	Strong	No

Given this test data, let's classify each of them and determine the accuracy:-

1: *Outlook = overcast*, And using rule 1, we get *PlayTennis = Yes*, So, this example is **correctly** classified. Note that rule 3 also applies.

2: *Humidity = normal ^ Wind = weak*, And using rule 5, we get *PlayTennis = Yes*, So, this example is **correctly** classified.

3: *Humidity = normal*, And using rule 3, we get *PlayTennis = Yes*. But actual class is No. So, this example is **incorrectly** classified.

4: *Humidity = normal ^ Wind = weak*, And using rule 5, we get *PlayTennis = Yes*, So, this example is **correctly** classified.

5: *Humidity = high*, And using rule 2, we get *PlayTennis = No*, So, this example is **correctly** classified.

So, the accuracy of our association rule based classifier is = $4/5 = 80\%$

Thank You