9 Data Mining

Solutions

- Exercise 1.1: Enumerate and give an example of area to use for each of the **functionalities of data mining** presented in the lecture:
  - **Association** (correlation and causality)
    - age(X,"20..29") , income(X,"20..29K") ⟶ buys(X,"PC") [support = 2%, confidence = 60%]
  - **Classification** and **Prediction**
    - Classify cars based on gas mileage
  - **Cluster analysis**
    - Cluster vacation travel market: demanders, escapists, educationalists

9 Data Mining

Solutions

- Exercise 1.1
  - **Outlier analysis**
    - Fraud detection: did you just buy 3 LCD TVs and 4 laptops yesterday with your credit card?
  - **Trend and evolution analysis**
    - Stock market / FOREX investments

9 Data Mining

Solutions

- Exercise 1.2: Define association rules
  - Let $I = \{i_1, i_2, ..., i_m\}$ be a set of items.
    Let $T = \{t_1, t_2, ..., t_n\}$ be a set of transactions where each transaction t_i is a set of items such that $t_i \subseteq I$.
  - An **association rule** is an implication of the form:
    $X \longrightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$

9 Data Mining

Solutions

- Exercise 2: Multi MinSup
  - M={3, 2, 5, 4, 6, 1}
  - Read transactions:

Item	Count	SUP %	MIS %
1	8	80	70
2	2	20	17
3	2	20	15
4	2	20	30
5	3	30	30
6	6	60	35

  - L={3, 2, 5, 4, 6, 1}
  - F_1={3, 2, 5, 6, 1}

Transactions
1, 4, 6
1
1, 5, 6
1, 6
4, 6
1, 2, 3, 5
1, 2, 3, 5
6
1
1, 6

Item	MIS %
1	70
2	17
3	15
4	30
5	30
6	35

F	Item	SUP %	MIS %
F1	3	20	15
	2	20	17
	5	30	30
	6	60	35
	1	80	70

9 Data Mining

Solutions

Item	Count	SUP %	MIS %	Transactions
1	8	80	70	1, 4, 6
2	2	20	17	1
3	2	20	15	1, 5, 6
4	2	20	30	1, 6
5	3	30	30	4, 6
6	6	60	35	1, 2, 3, 5
				1, 2, 3, 5
				6
				1
				1, 6

– L={3, 2, 5, 4, 6, 1}

– Candidate gen., K=2

- {3, 2} : sup(2) = 20%
  20% > MIS(3) = 15 and
  |sup(3) – sup(2)| = |20 - 20| = 0 < φ = 20%
  so {3, 2} is a good candidate
- {3, 5}: is a good candidate
- {3, 4}: is a good candidate
- {3, 6}: is NOT a good candidate (> φ)
- {3, 1}: is NOT a good candidate (> φ)

$\varphi = 20\%$

9 Data Mining

Solutions

- L={3, 2, 5, 4, 6, 1}
  - {2, 5}: is a good candidate
  - {2, 4}: is a good candidate
  - {2, 6}: is NOT a good candidate (> φ)
  - {2, 1}: is NOT a good candidate (> φ)

$\varphi = 20\%$

Item	Count	SUP %	MIS %
1	8	80	70
2	2	20	17
3	2	20	15
4	2	20	30
5	3	30	30
6	6	60	35

9 Data Mining

Solutions

$\varphi = 20\%$

Transactions
1, 4, 6
1
1, 5, 6
1, 6
4, 6
1, 2, 3, 5
1, 2, 3, 5
6
1
1, 6

- L={3, 2, 5, 4, 6, 1}
  - {5, 4}: sup(4) = 20% < MIS(5) = 30% so {5, 4} is NOT a good candidate
  - {5, 6}: is NOT a good candidate
  - {5, 1}: is NOT a good candidate (> φ)
  - 4 can't be used as seed since sup(4) < MIS(4)
  - {6, 1}: is a good candidate
- C2={{3, 2}, {3, 5}, {3, 4}{2, 5}, {2, 4} {6, 1}}

Item	Count	SUP %	MIS %
1	8	80	70
2	2	20	17
3	2	20	15
4	2	20	30
5	3	30	30
6	6	60	35

9 Data Mining

Solutions

- C2={{3, 2}, {3, 5}, {3, 4}{2, 5}, {2, 4}, {6, 1}}
- Read Transactions to calculate F2
  - F2 ={{3, 2}, {3, 5}, {2, 5}, {6, 1}}

F	Item	SUP %	MIS %
F1	3	20	15
	2	20	17
	5	30	30
	6	60	35
	1	80	70
F2	{3, 2}	20	15
	{3, 5}	20	15
	{2, 5}	20	17
	{6, 1}	40	35

Transactions
1, 4, 6
1
1, 5, 6
1, 6
4, 6
1, 2, 3, 5
1, 2, 3, 5
6
1
1, 6

9 Data Mining

Solutions

- F2 ={{3, 2}, {3, 5}, {2, 5}, {6, 1}}; k = 3
- Join:
  - {3, 2, 5}: MIS(2) < MIS(5) and |sup(2) – sup(5)| = 10 < φ, so it can be joined
  - Nothing else can be joined
- Prune
  - {3, 2} and {3, 5} ∈ F2
  - Since {2, 5} ∈ F2 the head problem is avoided otherwise we should have recorded also sup({2, 5})
- C3 = {3, 2, 5}

Transactions
1, 4, 6
1
1, 5, 6
1, 6
4, 6
1, 2, 3, 5
1, 2, 3, 5
6
1
1, 6

Item	Count	SUP %	MIS %
1	8	80	70
2	2	20	17
3	2	20	15
4	2	20	30
5	3	30	30
6	6	60	35

9 Data Mining

Solutions

minconf = 60%

- Scan transactions, F3={3, 2, 5}
  - Sup({3, 2, 5}) = 20 % > MIS(3) = 15
- Step 2: rule generation from F3={3, 2, 5}
  - Non-empty subsets: {3, 2}, {3, 5}, {2, 5}, {3}, {2}, {5}
  - Possible rules derived from F_3:
    - {3, 2} ⟶ {5}, [sup = 20%, conf = 100%]
    - {3, 5} ⟶ {2}, [sup = 20%, conf = 100%]
    - {2, 5} ⟶ {3}, [sup = 20%, conf = 100%]
    - {3} ⟶ {2, 5}, [sup = 20%, conf = 100%]
    - {2} ⟶ {3, 5}, [sup = 20%, conf = 100%]
    - {5} ⟶ {3, 2}, [sup = 20%, conf = 67%]
  - All are valid since minconf = 60%

F	Item	SUP %	MIS %
F1	3	20	15
	2	20	17
	5	30	30
	6	60	35
	1	80	70
F2	{3, 2}	20	15
	{3, 5}	20	15
	{2, 5}	20	17
	{6, 1}	40	35
F3	{3, 2, 5}	20	15

9 Data Mining

Solutions

- Possible rules derived from F_2:
  - {3} ⟶ {2}, [sup = 20%, conf = 100%]
  - {2} ⟶ {3}, [sup = 20%, conf = 100%]
  - {3} ⟶ {5}, [sup = 20%, conf = 100%]
  - {5} ⟶ {3}, [sup = 20%, conf = 67%]
  - {2} ⟶ {5}, [sup = 20%, conf = 100%]
  - {5} ⟶ {2}, [sup = 20%, conf = 67%]
  - {6} ⟶ {1}, [sup = 40%, conf = 67%]
  - {1} ⟶ {6}, [sup = 40%, conf = 50%]
- Except {1} ⟶ {6}, all are valid

minconf = 60%

F	Item	SUP %	MIS %
F1	3	20	15
	2	20	17
	5	30	30
	6	60	35
	1	80	70
F2	{3, 2}	20	15
	{3, 5}	20	15
	{2, 5}	20	17
	{6, 1}	40	35
F3	{3, 2, 5}	20	15