Data Mining: Transformed Data Sets

Andrew Kusiak
Intelligent Systems Laboratory
2139 Seamans Center
The University of Iowa
Iowa City, IA 52242 - 1527
andrew-kusiak@uiowa.edu
http://www.icaen.uiowa.edu/~ankusiak
Tel. 319-335 5934

Types of Data Sets

- Data sets with defined decisions (supervised learning)
- Data sets with undefined decisions (unsupervised learning)

“In-between Cases”

- For some objects the decisions may have alternative values, e.g., due to inability of assigning a correct decision value
- For some objects decision values may be difficult define, e.g., due to the lack of a proper measure

Data Sets with Ill-defined Decisions

- Action 1: Removal of objects with ill-defined decisions from the training data set.
- Action 2: Considering alternative decision values.
- Action 3: Action 1 and Action 2 applied simultaneously.

Data Set DS_0 (1)

- 18 features
- 29 objects
- Ill-defined decision D (e.g., alternative values)

Example 1

Data Set DS_0 (2)

CQ(F1) = … = CQ(F6) = 0, CQ(F7) = 0.034, CQ(F8) = 0.030, CQ(F9) = CQ(F10) = 0, CQ(F11) = 0.034, CQ(F12) = 0, CQ(F13) = 0.483, CQ(F14) = 0.069, CQ(F15) = 0.241, CQ(F16) = 0.310, CQ(F17) = 0.241, CQ(F18) = 0.172
Data Set DS_0 (3)

• Rule 1. (F1 = 4) AND (F13 in [1, 2, 4]) THEN (D1 = 1); [4, 4, 36.36%, 100.00%] [4, 0] [2, 5, 14, 20]
• Rule 2. (F18 in [3, 5]) THEN (D1 = 1); [5, 5, 45.45%, 100.00%] [5, 0] [1, 13, 24, 25, 29]
• Rule 3. (F10 = 3) AND (F16 = 2) THEN (D1 = 1); [7, 7, 63.64%, 100.00%] [7, 0] [2, 5, 11, 13, 18, 24, 29]
• Rule 4. (F13 = 3) THEN (D1 = 2); [13, 13, 72.22%, 100.00%] [0, 13] [3, 6, 7, 9, 10, 12, 16, 17, 21, 23, 26, 27, 28]
• Rule 5. (F1 in [2, 3]) AND (F18 = 1) THEN (D1 = 2); [12, 12, 66.67%, 100.00%] [0, 12] [4, 8, 9, 10, 15, 16, 17, 19, 21, 22, 23, 27]

Data Set DS_0 (4)

- Rule 1. (F1 = 4) AND (F13 in {1, 2, 4}) THEN (D1 = 1); [4, 4, 36.36%, 100.00%] [4, 0] [2, 5, 14, 20]
- Rule 2. (F18 in {3, 5}) THEN (D1 = 1); [5, 5, 45.45%, 100.00%] [5, 0] [1, 13, 24, 25, 29]
- Rule 3. (F10 = 3) AND (F16 = 2) THEN (D1 = 1); [7, 7, 63.64%, 100.00%] [7, 0] [2, 5, 11, 13, 18, 24, 29]
- Rule 4. (F13 = 3) THEN (D1 = 2); [13, 13, 72.22%, 100.00%] [0, 13] [3, 6, 7, 9, 10, 12, 16, 17, 21, 23, 26, 27, 28]
- Rule 5. (F1 in {2, 3}) AND (F18 = 1) THEN (D1 = 2); [12, 12, 66.67%, 100.00%] [0, 12] [4, 8, 9, 10, 15, 16, 17, 19, 21, 22, 23, 27]

Data Set DS_0 (5)

- Decision D = 1
  2, 5, 13, 18, 20
- Decision D = 2
  15, 22

Transformed Data Set DS_1 (1)

- Removal of all misclassified objects 2, 5, 13, 15, 18, 20, 22 (for D = 1, and D = 2)

Transformed Data Set DS_1 (2)

- Rule 1. (F2 = 1) AND (F12 = 4) THEN (D1 = 1); [6, 6, 100.00%, 100.00%] [6, 0] [1, 9, 11, 17, 18, 22]
- Rule 2. (F2 = 2) THEN (D1 = 2); [13, 13, 81.25%, 100.00%] [0, 13] [2, 3, 4, 5, 7, 8, 10, 12, 13, 15, 19, 20, 21]
- Rule 3. (F15 = 1) THEN (D1 = 2); [6, 6, 37.50%, 100.00%] [0, 6] [3, 6, 14, 16, 20, 21]
Transformed Data Set DS_2 (2)

• Rule 1. \((F_2 = 1) \text{ AND } (F_{12} = 4) \Rightarrow (D_1 = 1);\) \([6, 6, 100.00\%, 100.00\%] \{6, 0\} \{1, 9, 11, 19, 20, 24\}\)
• Rule 2. \((F_2 = 2) \Rightarrow (D_1 = 2); \{15, 15, 83.33\%, 100.00\%\} \{0, 15\} \{2, 3, 4, 5, 7, 8, 10, 12, 13, 14, 16, 17, 21, 22, 23\}\)
• Rule 3. \((F_{12} = 2) \Rightarrow (D_1 = 2); \{8, 8, 44.44\%, 100.00\%\} \{0, 8\} \{3, 4, 6, 10, 14, 15, 16, 18\}\)

Transformed Data Set DS_3 (2)

• Rule 1. \((F_1 = 4) \text{ AND } (F_{13} \in \{4, 2, 1\}) \Rightarrow (D_1 = 1);\) \([4, 4, 36.36\%, 100.00\%] \{4, 0\} \{2, 5, 14, 19\}\)
• Rule 2. \((F_10 \in \{2, 4\}) \text{ AND } (F_{12} \in \{4, 3\}) \Rightarrow (D_1 = 1);\) \([5, 5, 62.50\%, 100.00\%] \{5, 0\} \{1, 14, 15, 22, 25\}\)
• Rule 3. \((F_{13} = 3) \Rightarrow (D_1 = 2); \{13, 13, 81.25\%, 100.00\%\} \{0, 13\} \{3, 6, 7, 9, 10, 12, 15, 16, 20, 21, 24, 25, 26\}\)
• Rule 4. \((F_1 \in \{3, 2\}) \text{ AND } (F_{12} = 2) \Rightarrow (D_1 = 2); \{6, 6, 37.50\%, 100.00\%\} \{0, 6\} \{4, 8, 16, 18, 20, 21\}\)

Transformed Data Set DS_4 (2)

• Rule 1. \((F_1 = 2) \text{ AND } (F_2 = 1) \text{ AND } (F_{12} \in \{4, 3\}) \Rightarrow (D_1 = 1);\) \([4, 4, 50.00\%, 100.00\%] \{4, 0\} \{11, 24, 25, 29\}\)
• Rule 2. \((F_{10} \in \{2, 4\}) \text{ AND } (F_{12} \in \{4, 3\}) \text{ AND } (F_{13} \in \{2, 4\}) \Rightarrow (D_1 = 1);\) \([4, 4, 62.50\%, 100.00\%] \{4, 0\} \{11, 14, 15, 22, 25\}\)
• Rule 3. \((F_{13} \in \{3, 1\}) \Rightarrow (D_1 = 2); \{14, 14, 66.67\%, 100.00\%\} \{0, 14\} \{3, 6, 7, 9, 10, 12, 15, 16, 17, 20, 21, 23, 26, 27, 28\}\)
• Rule 4. \((F_{12} = 2) \Rightarrow (D_1 = 2); \{9, 9, 42.86\%, 100.00\%\} \{0, 9\} \{4, 6, 8, 12, 17, 19, 20, 21, 23\}\)
• Rule 5. \((F_1 = 4) \text{ AND } (F_{10} = 3) \Rightarrow (D_1 = 2); \{2, 2, 9.52\%, 100.00\%\} \{2, 5\}\)
• Rule 6. \((F_2 = 2) \text{ AND } (F_{10} = 3) \Rightarrow (D_1 = 2); \{3, 3, 14.29\%, 100.00\%\} \{0, 3\} \{9, 13, 18\}\)

Transformed Data Set DS_5 (1)

• Alternative decision values
  - \(D = 2\) to \(D = 1\) for objects 15 and 22

Transformed Data Set DS_5 (2)

• Alternative decision values
  - \(D = 2\) to \(D = 1\) for objects 15 and 22
Transformed Data Set DS_5 (2)

- Rule 1. \((F1 = 4) \text{ AND } (F13 \in \{1, 2, 4\}) \text{ THEN } (D1 = 1); \) 
  \([4, 4, 30.77\%, 100.00\%] \) \([4, 0]\) 
  \([2, 5, 14, 20]\)

- Rule 2. \((F12 \in \{3, 4\}) \text{ AND } (F13 \in \{2, 4\}) \text{ THEN } (D1 = 1); \) 
  \([12, 12, 92.31\%, 100.00\%] \) \([12, 0]\) 
  \([1, 2, 5, 11, 13, 14, 15, 18, 22, 24, 25, 29]\)

- Rule 3. \((F13 = 3) \text{ THEN } (D1 = 2); \) 
  \([3, 6, 7, 9, 10, 12, 16, 17, 21, 23, 26, 27, 28]\)

- Rule 4. \((F1 \in \{2, 3\}) \text{ AND } (F12 = 2) \text{ THEN } (D1 = 2); \) 
  \([6, 6, 37.50\%, 100.00\%] \) \([6, 0]\) 
  \([4, 8, 17, 19, 21, 23]\)

Transformed Data Set DS_6 (1)

- Removal of objects and alternative decisions
- removal of the misclassified objects objects 2, 5, 13, 18, and 20 with \(D = 1\) (data set DS_2)
- Alternative decision value for misclassified object 19 of DS_2 from \(D = 2\) to \(D = 1\)

Transformed Data Set DS_6 (2)

<table>
<thead>
<tr>
<th>Action 1 &amp; 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
</tbody>
</table>

Classification Accuracy

- Average 100.00 | 0.00 | 0.00
- 1 100.00 | 0.00 | 0.00
- 2 100.00 | 0.00 | 0.00

Transformed Data Set DS_6 (3)

- Rule 1. \((F2 = 1) \text{ AND } (F8 = 2) \text{ THEN } (D1 = 1); \) 
  \([7, 7, 100.00\%, 100.00\%] \) \([7, 0]\) 
  \([1, 9, 11, 15, 19, 20, 24]\)

- Rule 2. \((F2 = 2) \text{ THEN } (D1 = 2); \) 
  \([15, 15, 88.24\%, 100.00\%] \) \([0, 15]\) 
  \([2, 3, 4, 5, 7, 8, 10, 12, 13, 14, 16, 17, 21, 22, 23]\)

- Rule 3. \((F8 = 1) \text{ THEN } (D1 = 2); \) 
  \([9, 9, 52.94\%, 100.00\%] \) \([0, 9]\) 
  \([6, 7, 8, 12, 16, 18, 21, 22, 23]\)

- CQ(F1) = 0, CQ(F2) = 0.625, CQ(F3) = 0.167, CQ(F4) = 0, CQ(F5) = 0.250, CQ(F6) = CQ(F7) = 0, CQ(F8) = 0.375, CQ(F9) = 0, CQ(F10) = 0.083, CQ(F11) = 0.042, CQ(F12) = 0.208, CQ(F13) = 0.542, CQ(F14) = CQ(F15) = 0.292, CQ(F16) = CQ(F17) = 0, CQ(F18) = 0.167

Note

CQ(F2, F8) = 1 \(\Rightarrow\) \{F2, F8\} is a reduct.