Unsupervised Data Mining

Association Rule Learning
Association Rule Analysis

- Popular in mining data bases
- Automated discovery of sets of variables that occur frequently or one(s) leading to other(s)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Demographic</th>
<th># values</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sex</td>
<td>2</td>
<td>categorical</td>
</tr>
<tr>
<td>2</td>
<td>marital status</td>
<td>5</td>
<td>categorical</td>
</tr>
<tr>
<td>3</td>
<td>age</td>
<td>7</td>
<td>ordinal</td>
</tr>
<tr>
<td>4</td>
<td>education</td>
<td>6</td>
<td>ordinal</td>
</tr>
<tr>
<td>5</td>
<td>occupation</td>
<td>9</td>
<td>categorical</td>
</tr>
<tr>
<td>6</td>
<td>income</td>
<td>9</td>
<td>ordinal</td>
</tr>
<tr>
<td>7</td>
<td>years in Bay Area</td>
<td>5</td>
<td>ordinal</td>
</tr>
<tr>
<td>8</td>
<td>dual incomes</td>
<td>3</td>
<td>categorical</td>
</tr>
<tr>
<td>9</td>
<td>number in household</td>
<td>9</td>
<td>ordinal</td>
</tr>
<tr>
<td>10</td>
<td>number of children</td>
<td>9</td>
<td>ordinal</td>
</tr>
<tr>
<td>11</td>
<td>householder status</td>
<td>3</td>
<td>categorical</td>
</tr>
<tr>
<td>12</td>
<td>type of home</td>
<td>5</td>
<td>categorical</td>
</tr>
<tr>
<td>13</td>
<td>ethnic classification</td>
<td>8</td>
<td>categorical</td>
</tr>
<tr>
<td>14</td>
<td>language in home</td>
<td>3</td>
<td>categorical</td>
</tr>
</tbody>
</table>
Association Rule Analysis (cont)

**Association rule 1:** Support 25%, confidence 99.7% and lift 1.03.

\[
\begin{align*}
\text{number in household} & = 1 \\
\text{number of children} & = 0 \\
\downarrow \\
\text{language in home} & = \text{English}
\end{align*}
\]

**Association rule 2:** Support 13.4%, confidence 80.8%, and lift 2.13.

\[
\begin{align*}
\text{language in home} & = \text{English} \\
\text{householder status} & = \text{own} \\
\text{occupation} & = \{\text{professional/managerial}\} \\
\downarrow \\
\text{income} & \geq 40,000
\end{align*}
\]

**Association rule 3:** Support 26.5%, confidence 82.8% and lift 2.15.

\[
\begin{align*}
\text{language in home} & = \text{English} \\
\text{income} & < 40,000 \\
\text{marital status} & = \text{not married} \\
\text{number of children} & = 0 \\
\downarrow \\
\text{education} & \notin \{\text{college graduate, graduate study}\}
\end{align*}
\]
Market Basket Analysis

- Retail outlets
  - Placement of merchandises (affinity positioning)
  - Cross advertising
- Banks
- Insurance
  - Link analysis for fraud
- Medical
  - Symptom analysis
Co-occurrence Matrix

Customer 1: beer, pretzels, potato chips, aspirin
Customer 2: diapers, baby lotion, grapefruit juice, baby food, milk
Customer 3: soda, potato chips, milk
Customer 4: soup, beer, milk, ice cream
Customer 5: soda, coffee, milk, bread
Customer 6: beer, potato chips

<table>
<thead>
<tr>
<th></th>
<th>Beer</th>
<th>Pot. Chips</th>
<th>Milk</th>
<th>Diap.</th>
<th>Soda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beer</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pot. Chips</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Milk</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Diapers</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Soda</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

- Interesting cases can have $10^4$ variables and $10^8$ of samples
- Co-occurrence gives only pair-wise association
Practical Solutions

- Run up against curse-of-dimensionality
  - With $10^4$ variables, each with many possible values, need very large number of samples to populate the space, “bump” hunting in fine scale is not possible
    - Look for regions in the probability spaces with high density
  - Even for binary variables, there are $2^k$ (e.g., $2^{1,000}$) possible 1,0-tuples, must have efficient search algorithms
Simplification

- Assuming binary variables
- If not, force them binaries
  - Instead of 6 different education levels, just 2 (college and above, or below)
- Change of variables
  - Initially \((X_1, \ldots, X_p)\)
  - Each with \((S_1, \ldots, S_p)\) possible values
  - \(K = S_1 + \ldots + S_p\)
  - Create \(Z_k\) binary variables
    - 1 if the corresponding variable \(X_i\) assuming value \(S_{ij}\)
    - 0 otherwise
Apriori Algorithm

- Threshold t
- 1st pass:
  - Single-variable set: must have occurrence larger than t
- 2nd pass:
  - Pair-wise variable sets: together must have occurrence larger than t
- ...
- mth pass:
  - Only those tuples in (m-1)th pass have probability higher than t are considered

To avoid combinatorial explosion, t cannot be too low
Tuples to Rules

- Tuples \{Z_k\} to A \Rightarrow B
  - A antecedent
  - B consequent
  - \(T(A \Rightarrow B)\): support, probability of simultaneously observing A and B \(P(A&B)\)
  - \(C(A \Rightarrow B) = \frac{T(A \Rightarrow B)}{T(A)}\): confidence, probability of \(P(B|A)\)
  - \(L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{T(B)}\): lift, probability of \(\frac{P(A&B)}{P(A)P(B)}\)
Examples

- K={peanut butter, jelly, bread}
- \{peanut butter, jelly\} => bread
- Support of 0.03: if \{peanut butter, jelly, bread\} appears in 3% of sample baskets
- Confidence of 82%: if peanut butter and jelly are purchased, then 82% time bread is also
- Lift of 1.9: If bread appear in 43% of sample baskets, then $0.82/0.43=1.9$
**FIGURE 14.2.** Market basket analysis: relative frequency of each dummy variable (coding an input category) in the data (top), and the association rules found by the Apriori algorithm (bottom).