Association Rules

- Unsupervised learning but complementary to data exploration in clustering.
- The goal is to find "weak implications" in the data that have "non-negligible coverage"
- Useful in marketing, in understanding application data, as feature generator for supervised learning.

Data Model

- Following the market-basket application
- We assume a table where
  - Each row is a "transaction"
  - Each column is an "item"
- Table entries are in \{0,1\} i.e., discrete
- A transaction can be seen to represent the corresponding set of items

\[\text{support}(X) = \text{#transactions including } X\]
\[\text{frequency}(X) = \frac{\text{support}(X)}{\text{#transactions}}\]
\[\text{confidence}(X \Rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)}\]
Association Rules

- Applications and data characteristics from some early papers:

<table>
<thead>
<tr>
<th>Data</th>
<th>#transactions</th>
<th>#items</th>
<th>transaction size</th>
<th>Avg size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market Basket</td>
<td>50K</td>
<td>13K</td>
<td>1-100</td>
<td>10</td>
</tr>
<tr>
<td>Web Clicks</td>
<td>50K</td>
<td>500</td>
<td>1-267</td>
<td>2.5</td>
</tr>
<tr>
<td>Census</td>
<td>30K</td>
<td>2000</td>
<td>70</td>
<td>70</td>
</tr>
</tbody>
</table>

- In more demanding applications data does not fit in memory

Association Rules

- More data characteristics

![Figure from [ZKM01]](http://www.almaden.ibm.com/cs/quest//syndata.html#assocSynData)

Association Rules

- Too many sets and rules ...

![Figure from [ZKM01]](http://www.almaden.ibm.com/cs/quest//syndata.html#assocSynData)

Association Rules

- A concrete task 1:
  - find all rules with frequency at least \( f \) and confidence at least \( c \).
  - How can we do this?
  - If \( (X \rightarrow Y) \) satisfies conditions then \( (X+Y) \) must also have frequency at least \( f \).

Association Rules

- From frequent sets to rules
  - Given frequent set \( Z \)
    - for example \( \{A,B,C,D,E\} \)
    - Remove potential conclusion \( W \)
    - for example \( D \)
    - And check the confidence of \( (Z\backslash W \rightarrow W) \) of \( (ABCE \rightarrow D) \)
Frequent Set Mining

• find all sets \( Z \) with frequency at least \( f \)

• How can we do this?

• Main insight: anti-monotonicity
  
  if set \( Z \) is frequent then all its subsets are also frequent

• Algorithmic ideas?

Lattice Structure of Freq Sets

• Notice the notions of
  
  – positive border
  – negative border

• that are implicit in the monotonicity property and in the view via the lattice

• The borders capture all the frequent sets. Some algorithms attempt to find these directly.

Level-wise (Apriori) Algorithm

• level=1
• \( \text{cands}[1] = \text{Sets with single items} \)
• While \( \text{cands}[\text{level}] \) not empty
  1. Calc support for \( \text{cands}[\text{level}] \)
  2. \( \text{freq}[\text{level}] = \text{cands}[\text{level}] \) with high support
  3. \( \text{cands}[\text{level}+1]=\text{generate from } \text{freq}[\text{level}] \)
  4. level=level+1

Calculating support for candidates

• Basic implementation of step 1

• For each row \( R \)
  • For each candidate \( X \)
    • if \( X \) subset of \( R \) then: count\([X]\)=1

• One pass over database
• Improve run time via trie data structure that captures set of candidates

Generating candidates

• Monotonicity can be used to generate and prune potential candidates

• Use trie or lexicographical ordering to identify potential candidates

• Prune via subset relation

• Prune via upper/lower bounds on frequency (we skip details of this idea)
Vertical View

- View each item as a set of transactions
- This directly captures support
- Support of a set is the intersection of support of parents
- This incurs large space cost for storing support
- DFS recursive exploration avoids this and leads to efficient algorithm (we skip the details of this)

Which Rules are Interesting?

- Confidence can often be misleading
  \[ \text{confidence}(X \Rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)} \]
- If \( p(B) \) is large
  \[ p(B|A) = p(B) \] \( i.e., \) independent
- Confidence \( (A \Rightarrow B) \) is still large

Positive and Negative Borders

- Lift measures dist from independence
  \[ \text{Lift}(X \Rightarrow Y) = \frac{\text{freq}(X \cup Y)}{\text{freq}(X) \text{freq}(Y)} \]
- Conviction aims at "implication"
  \[ \text{Conviction}(X \Rightarrow Y) = \frac{\text{freq}(X)(1 - \text{freq}(Y))}{\text{freq}(X) - \text{freq}(X \cup Y)} \]
- Interpret as inverse of \( \text{Lift}(X \text{ and Not } Y) \)

AMSS Alg: Find one maximal set

- Table 3: Sample Implication Rules From Census Data

<table>
<thead>
<tr>
<th>Implication Rule</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRE under 18 yrs</td>
<td>( \approx 0.5 )</td>
</tr>
<tr>
<td>Unemployed people don't earn income from work</td>
<td>( \approx 0.5 )</td>
</tr>
<tr>
<td>Men don't give birth</td>
<td>( \approx 0.5 )</td>
</tr>
<tr>
<td>People who are not in the military and are not looking for work and who have worked this year (1995, the year of the census) currently have civilian employment</td>
<td>( \approx 0.5 )</td>
</tr>
<tr>
<td>People who are not in the military and who worked two years ago are not limited in their work by a disability</td>
<td>( \approx 0.5 )</td>
</tr>
<tr>
<td>People 25-29; female; live in school and without personal care limitations have worked this year</td>
<td>( \approx 0.5 )</td>
</tr>
<tr>
<td>African American women who are not in the military</td>
<td>( \approx 0.5 )</td>
</tr>
<tr>
<td>African Americans reside in the same state they were born</td>
<td>( \approx 0.5 )</td>
</tr>
<tr>
<td>Unmarried people have moved in the past five years</td>
<td>( \approx 0.5 )</td>
</tr>
</tbody>
</table>
Dualize and Advance Algorithm

- ABC and BD are max set found so far
- AD and CD are the negative border of this set
- Any undiscovered frequent set must lie above the negative border.

Figure from [GKMSTS 2003]

Frequent Sets as Features

- One way to generate enriched features for supervised learning is to generate frequent sets (because they occur and thus have a chance of making a difference)
- Very successful in frequent sub-graph mining, which extends the topic of this lecture to graphs, and its application to classifying molecules

Summary

- Association rules: a novel form of data exploration with different goals from previous supervised and unsupervised learning
- Algorithmic/computational challenge
- Frequent set mining as an important subtask
- Property: Anti-monotonicity
- Level-wise algorithm
- Many alternative algorithms