Privacy-Preserving Data Mining in Relational Databases

Xiaoxin Yin, Jiawei Han
University of Illinois at Urbana-Champaign

Data Mining in Relational Databases

- Most real-world data are stored in relational databases
- Information of different entities is stored in inter-connected relations
- Multiple relations need to be combined for data mining

Ask the backend database
Apply for loan

Data Mining across Databases

- Information is often distributed in different databases
- Information needs to be shared for data mining
- How to share information while preserving privacy?

Tuple ID Propagation

<table>
<thead>
<tr>
<th>Loan ID</th>
<th>Account ID</th>
<th>Amount</th>
<th>Duration</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>126</td>
<td>10000</td>
<td>12</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>103</td>
<td>-10000</td>
<td>24</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>145</td>
<td>10000</td>
<td>36</td>
<td>No</td>
</tr>
</tbody>
</table>

Possible features:
- Frequency=monthly: Yes
- Open date < 01/01/95: No
- Propagate tuple IDs of target relation to non-target relations
- Search for good features that distinguish positive examples from negative ones

Secure Cross-Database Classification

- Propagating across many links between databases
  - E.g. Link of SSN, card number, zip code, ...
- Each propagation uses a different encryption key
  - Credit bureau cannot know relationship among objects in the Bank
- Useful features are found from different propagations
  - E.g. Determine whether to approve a credit card application from
    - Salary
    - Unemployment rate
    - Num of previous applications
    - ...

Tuple ID Propagation across Databases

- To build a classifier, only tuple IDs and class labels are needed, which can be propagated across relations and databases
- From the IDs and labels, important features can be identified
  - E.g. Districts.unemployment_rate<3% => approve!
- Information is encrypted when propagated across databases
- Useful features are found in Credit Bureau database, and transferred back to Bank database

Scalability w.r.t. number of relations

Scalability w.r.t. number of tuples

Preliminary Results

Compared with previous approaches in single database:

- Our approach (CrossMine) is more accurate and much more efficient, because it propagates very small amounts of information