Privacy Preserving Data Mining
A Randomization Approach

Ramakrishnan Srikant
Data Mining and Privacy

• The primary task in data mining: development of models about aggregated data.
• What if we randomize individual data records to protect privacy?
• Can we still develop accurate models?
Talk Outline

- Classification
- Association Rules
- Open Problems

Randomization Approach Overview

Alice’s age

30 | 70K | ...
Randomizer

65 | 20K | ...
Reconstruct Distribution of Age

30 becomes 65 (30+35)

Add random number to Age

50 | 40K | ...
Randomizer

25 | 60K | ...
Reconstruct Distribution of Salary

Classification Algorithm

Model
Reconstruction Problem

- Original values $x_1, x_2, ..., x_n$
  - from probability distribution $X$ (unknown)
- To hide these values, we use $y_1, y_2, ..., y_n$
  - from probability distribution $Y$
- Given
  - $x_1 + y_1, x_2 + y_2, ..., x_n + y_n$
  - the probability distribution of $Y$

Estimate the probability distribution of $X$. 
Intuition (Reconstruct single point)

• Use Bayes' rule for density functions

Original distribution for Age

Probabilistic estimate of original value of V
Intuition (Reconstruct single point)

- Use Bayes' rule for density functions

Original Distribution for Age

- Probabilistic estimate of original value of V
Reconstructing the Distribution

• Combine estimates of where point came from for all the points:
  – Gives estimate of original distribution.

\[
f_X = \frac{1}{n} \sum_{i=1}^{n} \frac{f_Y((x_i + y_i) - a) f_X^j(a)}{\int_{-\infty}^{\infty} f_Y((x_i + y_i) - a) f_X^j(a)}
\]
Reconstruction: Bootstrapping

\[ f_X^0 := \text{Uniform distribution} \]
\[ j := 0 \quad \text{// Iteration number} \]
repeat
\[ f_X^{j+1}(a) := \frac{1}{n} \sum_{i=1}^{n} \frac{f_Y((x_i + y_i) - a) f_X^j(a)}{\int_{-\infty}^{\infty} f_Y((x_i + y_i) - a) f_X^j(a)} \quad \text{(Bayes' rule)} \]
\[ j := j + 1 \]
until (stopping criterion met)

- Converges to maximum likelihood estimate.
Works well
Recap: Why is privacy preserved?

• Cannot reconstruct individual values accurately.
• Can only reconstruct distributions.
Classification

- Naïve Bayes
  - Assumes independence between attributes.
- Decision Tree
  - Correlations are weakened by randomization, not destroyed.
### Decision Tree Example

<table>
<thead>
<tr>
<th>Age</th>
<th>Visitor?</th>
<th>Salary</th>
<th>Repeat</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>No</td>
<td>50K</td>
<td>Repeat</td>
</tr>
<tr>
<td>17</td>
<td>No</td>
<td>30K</td>
<td>Repeat</td>
</tr>
<tr>
<td>43</td>
<td>No</td>
<td>40K</td>
<td>Repeat</td>
</tr>
<tr>
<td>68</td>
<td>No</td>
<td>50K</td>
<td>Single</td>
</tr>
<tr>
<td>32</td>
<td>No</td>
<td>70K</td>
<td>Single</td>
</tr>
<tr>
<td>20</td>
<td>No</td>
<td>20K</td>
<td>Repeat</td>
</tr>
</tbody>
</table>

- **Age < 25**
  - If Yes, Repeat
  - If No, Salary < 50K
    - If Yes, Repeat
    - If No, Single
Randomization Level

- Add a random value between -30 and +30 to age.
- If randomized value is 60
  - know with 90% confidence that age is between 33 and 87.
- Interval width ∝ amount of privacy.
  - Example: (Interval Width : 54) / (Range of Age: 100) ⇒ 54% randomization level @ 90% confidence
Decision Tree Experiments

![Graph showing accuracy for different functions with 100% randomization level. The graph compares Original, Randomized, and Reconstructed models.](image-url)
Accuracy vs. Randomization Level

![Accuracy vs. Randomization Level](image_url)

- **Fn 3**

- **Accuracy**
  - Original
  - Randomized
  - ByClass
Talk Outline

- Motivation
- Association Rules
- Open Problems

S. Rizvi, J. Haritsa. Privacy-Preserving Association Rule Mining. VLDB 2002
Discovering Associations Over Privacy Preserved Categorical Data

• A transaction \( t \) is a set of items
• Support \( s \) for an itemset \( A \) is the number of transactions in which \( A \) appears
• Itemset \( A \) is frequent if \( s \geq s_{\text{min}} \)
• Task: Find all frequent itemsets, while preserving the privacy of individual transaction.
Uniform Randomization

- Given a transaction,
  - keep item with 20% probability,
  - replace with a new random item with 80% probability.

Is there a problem?
Example: \( \{x, y, z\} \)

10 M transactions of size 3 with 1000 items:

- 100,000 (1\%) have \( \{x, y, z\} \)
- 9,900,000 (99\%) have zero items from \( \{x, y, z\} \)

Uniform randomization: How many have \( \{x, y, z\} \)?

- 800 transactions 99.99\%
- \(0.2^3 = 0.008\)
- 9,900,000 transactions 0.01\%
- \(6 \times \left(\frac{0.8}{999}\right)^3 = 3 \times 10^{-9}\)
Our Solution

“Where does a wise man hide a leaf? In the forest.
But what does he do if there is no forest?”
“He grows a forest to hide it in.”

G.K. Chesterton

• Insert many false items into each transaction
• Hide true itemsets among false ones
Cut and Paste Randomization

- Given transaction $t$ of size $m$, construct $t'$:
  - Choose a number $j$ between 0 and $K_m$ (cutoff);
  - Include $j$ items of $t$ into $t'$;
  - Each other item is included into $t'$ with probability $p_m$.

The choice of $K_m$ and $p_m$ is based on the desired level of privacy.

$t = [a, b, c, u, v, w, x, y, z]$

$t' = [b, v, x, z, \alpha, \dot{a}, \beta, \xi, \psi, \xi, \kappa, \eta, \ldots]$

$j = 4$
Partial Supports

To recover original support of an itemset, we need randomized supports of its subsets.

• Given an itemset $A$ of size $k$ and transaction size $m$,
• A vector of partial supports of $A$ is

$$\bar{s} = (s_0, s_1, \ldots, s_k), \text{ where}$$

$$s_l = \frac{1}{|T|} \cdot \# \{ t \in T \mid \# (t \cap A) = l \}$$

– Here $s_k$ is the same as the support of $A$.
– Randomized partial supports are denoted by $\bar{s}'$. 
Transition Matrix

• Let $k = |A|$, $m = |t|$.

• **Transition matrix** $P = P(k, m)$ connects randomized partial supports with original ones:

\[
E \bar{s}' = P \cdot \bar{s}, \quad \text{where}
\]

\[
P_{l', l} = \Pr \left[ \#(t' \cap A) = l' \mid \#(t \cap A) = l \right]
\]
**The Estimators**

- Given randomized partial supports, we can estimate original partial supports:
  \[
  \vec{s}_{\text{est}} = Q \cdot \vec{s}', \quad \text{where} \quad Q = P^{-1}
  \]
- Covariance matrix for this estimator:
  \[
  \text{Cov} \vec{s}_{\text{est}} = \frac{1}{|T|} \sum_{l=0}^{k} s_l \cdot Q D[l] Q^T,
  \]
  where \( D[l]_{i,j} = P_{i,l} \cdot \delta_{i=j} - P_{i,l} \cdot P_{j,l} \)
- To estimate it, substitute \( s_l \) with \((s_{\text{est}})_l\).
  - Special case: estimators for support and its variance
Privacy Breach Analysis

- How many added items are enough to protect privacy?
  - Have to satisfy \( \Pr [z \in t | A \subseteq t'] < \rho \) (\( \Leftrightarrow \) no privacy breaches)
  - Select parameters so that it holds for all itemsets.
  - Use formula (\( s^+_l = \Pr [(t \cap A) = l, z \in t], \ s^+_0 = 0 \)):

\[
\Pr[z \in t | A \subseteq t'] = \sum_{l=0}^{k} s^+_l \cdot P_{k,l} / \sum_{l=0}^{k} s_l \cdot P_{k,l}
\]

- Parameters are to be selected in advance!
  - Enough to know maximal support of an itemset for each size.
  - Other parameters chosen for worst-case impact on privacy breaches.
Can we still find frequent itemsets?

Privacy Breach level = 50%.

### Soccer:

- $s_{\text{min}} = 0.2\%$

<table>
<thead>
<tr>
<th>Itemset Size</th>
<th>True Itemsets</th>
<th>True Positives</th>
<th>False Drops</th>
<th>False Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>266</td>
<td>254</td>
<td>12</td>
<td>31</td>
</tr>
<tr>
<td>2</td>
<td>217</td>
<td>195</td>
<td>22</td>
<td>45</td>
</tr>
<tr>
<td>3</td>
<td>48</td>
<td>43</td>
<td>5</td>
<td>26</td>
</tr>
</tbody>
</table>

### Mailorder:

- $s_{\text{min}} = 0.2\%$

<table>
<thead>
<tr>
<th>Itemset Size</th>
<th>True Itemsets</th>
<th>True Positives</th>
<th>False Drops</th>
<th>False Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>65</td>
<td>65</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>228</td>
<td>212</td>
<td>16</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>18</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
Talk Outline

- Classification
- Association Rules
- Open Problems
Privacy Breaches

• We know how to control privacy breaches for boolean data (associations) – what about quantitative data?

• Example: 80% of the people whose randomized value of age is in [80,90] and whose randomized value of income is [...] have their true age in [70,80].

• Challenge: How do you limit privacy breaches without prior knowledge of data distributions?
Clustering

• Classification: Partitioned the data by class & then reconstructed attributes.
  – Assumption: Attributes independent given class attribute.
• Clustering: Don’t know the class label.
  – Assumption: Attributes independent.
  – Latter assumption is much worse!
• Can we reconstruct a set of attributes together?
  – Amount of data needed increases exponentially with number of attributes.
Summary

• Can have our cake and mine it too!
  – Randomization is an interesting approach for building data mining models while preserving user privacy.
• Algorithms for privacy-preserving classification and association rules.
• Lots of interesting open problems.
Slides available from ...

www.almaden.ibm.com/u/srikant/talks.html