Privacy-preserving statistical analysis

Liina Kamm

liina@cyber.ee
http://sharemind.cyber.ee/
Rmind: a tool for cryptographically secure statistical analysis
Dan Bogdanov, Liina Kamm, Sven Laur and Ville Sokk
ePrint report 2014/512
Sharemind

Input parties

Computing parties

Result parties

Step 1: secret sharing of inputs

Step 2: secure multiparty computation

Step 3: reconstruction of results
Necessary functionality

- Classification, declassification and publishing of values
- Protected storage of a private value
- Support for vectors and matrices
- Integer, Boolean, floating-point arithmetic
- Division, square root
- Shuffling
- Linking
- Sorting
Filtering

### Database

<table>
<thead>
<tr>
<th>Attribute 1</th>
<th>Attribute 2</th>
<th>...</th>
<th>Attribute m</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>D2</td>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Di</td>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Dn</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$n$ elements

### Filtered attribute in the usual setting

- **Attribute $j$**
- $k$ elements

### Filtered attribute in the privacy-preserving setting

- **Attribute $j$**
- **Mask vector**: $1\ 0\ \cdots\ 1\ \cdots\ 0$
- $n$ elements
Quantiles (1)

\[ Q(p, [\tilde{a}]) = (1 - \gamma) \cdot \lfloor a_j \rfloor + \gamma \cdot \lceil a_{j+1} \rceil \]

\[ j = \lfloor (n - 1)p \rfloor + 1 \]

\[ \gamma = np - \lfloor (n - 1)p \rfloor - p. \]
Quantiles (2)

Algorithm 2: Privacy-preserving algorithm for finding the five-number summary of a vector that leaks the size of the selected subset

Data: Input data vector $[\bar{a}]$ and corresponding mask vector $[\bar{m}]$.
Result: Minimum $[min]$, lower quartile $[lq]$, median $[me]$, upper quartile $[uq]$, and maximum $[max]$ of $[\bar{a}]$ based on the mask vector $[\bar{m}]$

1. $[\bar{x}] \leftarrow \text{cut}([\bar{a}], [\bar{m}])$
2. $[\vec{b}] \leftarrow \text{sort}([\bar{x}])$
3. $[min] \leftarrow [b_1]$
4. $[max] \leftarrow [b_n]$
5. $[lq] \leftarrow Q(0.25, [\vec{b}])$
6. $[me] \leftarrow Q(0.5, [\vec{b}])$
7. $[uq] \leftarrow Q(0.75, [\vec{b}])$
8. return $([min], [lq], [me], [uq], [max])$
Quantiles (3)

**Algorithm 3**: Privacy-preserving algorithm for finding the five-number summary of a vector that hides the size of the selected subset.

**Data**: Input data vector \([\bar{a}]\) of size \(N\) and corresponding mask vector \([\bar{m}]\).

**Result**: Minimum \([\text{min}]\), lower quartile \([\text{lq}]\), median \([\text{me}]\), upper quartile \([\text{uq}]\), and maximum \([\text{max}]\) of \([\bar{a}]\) based on the mask vector \([\bar{m}]\)

1. \(([\bar{b}], [\bar{m}'])\) ← \text{sort}^\ast([\bar{a}], [\bar{m}])
2. \([n]\) ← \text{sum}([\bar{m}])
3. \([\text{os}]\) ← \(N - [n]\)
4. \([\text{min}]\) ← \([b_{[1+\text{os}]}}\)
5. \([\text{max}]\) ← \([b_N]\)
6. \([\text{lq}]\) ← \(Q^\ast(0.25, [\bar{a}], [\text{os}])\)
7. \([\text{me}]\) ← \(Q^\ast(0.5, [\bar{a}], [\text{os}])\)
8. \([\text{uq}]\) ← \(Q^\ast(0.75, [\bar{a}], [\text{os}])\)
9. \text{return} \(([\text{min}], [\text{lq}], [\text{me}], [\text{uq}], [\text{max}])\)
Descriptive statistics

- Five number summary and boxplot
- Histogram, frequency table, heatmap
- Mean, variance, standard deviation, covariance
Statistical testing

Option 1  Private data

Data → Test statistic → p-value → Comparison → Threshold

Option 2  Private data

Data → Test statistic → Comparison → Critical test statistic → Threshold

Option 3  Private data

Data → Test statistic → p-value → Comparison → Threshold
Statistical tests

- t-test, paired t-test
- Wilcoxon rank sum test, signed rank test
- chi-square test
- Multiple testing correction
  - Bonferroni correction
  - Benjamini-Hochberg procedure
Linear regression (1)

- \( k \) independent variables \( x_k \), one dependent variable \( y \)
- Want to find \( b_i \) such that
  \[
y_j = \beta_k x_{j,k} + \ldots + \beta_1 x_{j,1} + \beta_0 x_{j,0} + \varepsilon_j
  \]
  \[
  \bar{\varepsilon} = X\bar{\beta} - \bar{y}.
  \]
- Minimise the square of residuals \( ||\bar{\varepsilon}||^2 = ||\bar{y} - X\bar{\beta}||^2 \)
- Convert the task to its equivalent characterisation in terms of linear equations
  \[
  X^T X \bar{\beta} = X^T \bar{y}
  \]
Linear regression (2)

- Simple linear regression (one variable)
- Matrix inversion (up to four variables)
- Gaussian elimination with back substitution
- LU decomposition

- Conjugate gradient method
Algorithm 10: maxLoc: Finding the first maximum element and its location in a vector in a privacy-preserving setting

Data: A vector $\vec{a}$ of length $n$

Result: The maximum element $[b]$ and its location $[l]$ in the vector

1. Let $\pi(j)$ be a permutation of indices $j \in \{1, \ldots, n\}$
2. $[b] \leftarrow [a_{\pi(1)}]$ and $[l] \leftarrow \pi(1)$
3. for $i \in \{\pi(2), \ldots, \pi(n)\}$ do
   4. $[c] \leftarrow ([a_{\pi(i)}] > [b])$
   5. $[b] \leftarrow [b] - [c] \cdot [b] + [c] \cdot [a_{\pi(i)}]$
   6. $[l] \leftarrow [l] - [c] \cdot [l] + [c] \cdot \pi(i)$
4. end
5. return $([b], [l])$
Rmind demo