Privacy-Preserving Data Mining

Ramakrishnan Srikant

IBM Almaden Research Center

www.almaden.ibm.com/cs/people/srikant/

Motivation

- Privacy & Data Mining: Adversaries forever?
- The primary task in data mining: development of models about aggregated data.
- Can we develop accurate models without access to precise information in individual data records?

Papers

- R. Agrawal and R. Srikant, "Privacy Preserving Data Mining", *Proc. of the ACM SIGMOD Conference on Management of Data*, Dallas, Texas, May 2000.
- Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining", *Crypto 2000*, August 2000.

Talk Overview

- Randomization protects information at the individual level.
- Algorithm to reconstruct the distribution of values.
- Use reconstructed distributions in data mining algorithms, e.g. to build decision-tree classifier.
- How well does it work?

Privacy Surveys

- [CRA99b] survey of web users:
  - 17% privacy fundamentalists
  - 56% pragmatic majority
  - 27% marginally concerned
- [Wes99] survey of web users:
  - 82% : privacy policy matters
  - 14% don't care
- Not equally protective of every field
  - may not divulge at all certain fields;
  - may not mind giving true values of certain fields;
  - may be willing to give not true values but modified values of certain fields.

Related Work: Statistical Databases

- Statistical Databases : provide statistical information without compromising sensitive information about individuals (surveys: [AW89] [Sho82])
- Query Restriction
  - restrict the size of query result (e.g. [FEL72][DDS79])
  - control overlap among successive queries (e.g. [DJL79])
  - keep audit trail of all answered queries (e.g. [CO82])
  - suppress small data cells (e.g. [Cox80])
  - cluster entities into mutually exclusive atomic populations (e.g. [YC77])
- Data Perturbation
  - replace the original database by a sample from the same distribution (e.g. [LST83][LCL85][Rei84])
  - sample the result of a query (e.g. [Den80])
  - swap values between records (e.g. [Den82])
  - add noise to the query result (e.g. [Bec80])
  - add noise to the values (e.g. [TYW84][War65])

Related Work: Statistical Databases (cont.)

- Negative results: cannot give high quality statistics and simultaneously prevent partial disclosure of individual information [AW89]

- Negative results not directly applicable to privacy-preserving data mining.
  - Also want to prevent disclosure of confidential information
  - But sufficient to reconstruct original distribution of data values, i.e. not interested in high quality point estimates

Using Randomization to protect Privacy

- Return $x_i + r$ instead of x_i, where r is a random value drawn from a distribution.
  - Uniform
  - Gaussian
- Fixed perturbation – not possible to improve estimates by repeating queries.
- Algorithm knows parameters of r's distribution.

Talk Overview

- Randomization protects information at the individual level.
- Algorithm to reconstruct the distribution of values.
- Use reconstructed distributions to build decision-tree classifier.
- How well does it work?

Reconstruction Problem

- Original values $x_1, x_2, \ldots, x_n$
  - realizations of iid random variables $X_1, X_2, \ldots, X_n$,
  - each with the same distribution as random variable X.
- To hide these values, we use $y_1, y_2, \ldots, y_n$
  - realizations of iid random variables $Y_1, Y_2, \ldots, Y_n$,
  - each with the same distribution as random variable Y.

Given

- $x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n$
- the density function f_Y for Y,

estimate the density function f_X for X.

Using Bayes' Rule

- Assume we know both f_X and f_Y.
- Let $w_i \equiv x_i + y_i$.

$$
\begin{aligned}
& f_{X_1}(a \mid X_1 + Y_1 = w_1) \\
= \quad & \frac{f_{X_1+Y_1}(w_1 \mid X_1 = a)\, f_{X_1}(a)}{f_{X_1+Y_1}(w_1)} \\
& \text{(using Bayes' rule for density functions)} \\
= \quad & \frac{f_{X_1+Y_1}(w_1 \mid X_1 = a)\, f_{X_1}(a)}{\int_{-\infty}^{\infty} f_{X_1+Y_1}(w_1 \mid X_1 = z)\, f_{X_1}(z)\, dz} \\
= \quad & \frac{f_{Y_1}(w_1 - a)\, f_{X_1}(a)}{\int_{-\infty}^{\infty} f_{Y_1}(w_1 - z)\, f_{X_1}(z)\, dz} \qquad (Y_1 \text{ independent of } X_1) \\
= \quad & \frac{f_Y(w_1 - a)\, f_X(a)}{\int_{-\infty}^{\infty} f_Y(w_1 - z)\, f_X(z)\, dz} \qquad (f_{X_1} \equiv f_X,\ f_{Y_1} \equiv f_Y)
\end{aligned}
$$

$$
\begin{aligned}
f'_X(a) \quad & \approx \quad \frac{1}{n} \sum_{i=1}^{n} f_{X_i}(a \mid X_i + Y_i = w_i) \\
& = \quad \frac{1}{n} \sum_{i=1}^{n} \frac{f_Y(w_i - a)\, f_X(a)}{\int_{-\infty}^{\infty} f_Y(w_i - z)\, f_X(z)\, dz}
\end{aligned}
$$

Reconstruction Method: Algorithm

$f_X^0 :=$ Uniform distribution
$j := 0$ // Iteration number
`repeat`
    Use equation to compute a new estimate f_X^{j+1}.
    $j := j + 1$
`until` (stopping criterion met)

Stopping Criterion: Stop when difference between successive estimates of the original distribution becomes very small (1% of the threshold of the χ^2 test).

Using Partitioning to Speed Computation

- distance(z, w_i) $\approx$ distance between the mid-points of the intervals in which they lie, and
- density function $f_X(a)$ $\approx$ the average of the density function over the interval in which a lies.

$$f'_X(a) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \frac{f_Y(w_i - a)\, f_X(a)}{\int_{-\infty}^{\infty} f_Y(w_i - z)\, f_X(z)\, dz}$$

becomes

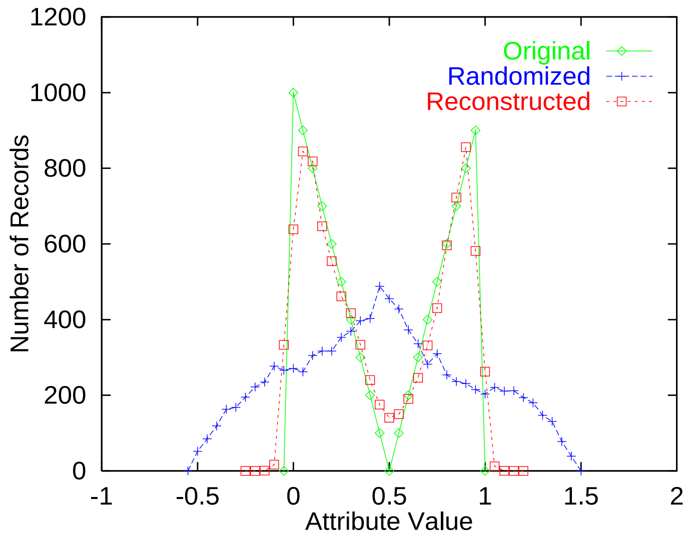
$$\Pr{}'(X \in I_p) =$$

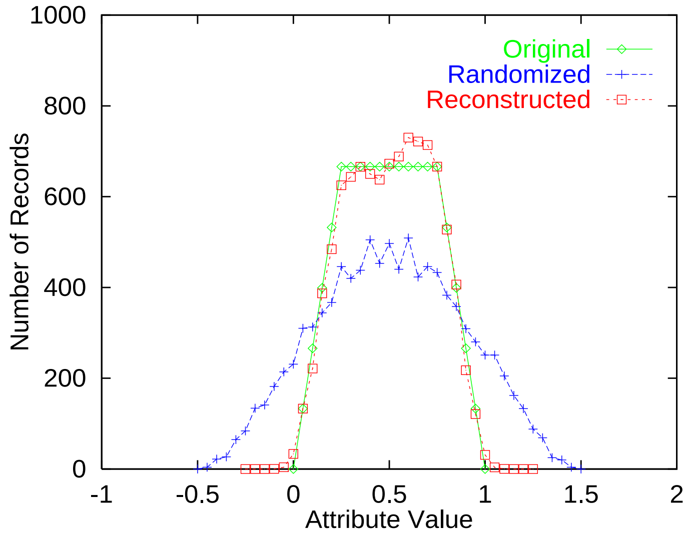
$$\frac{1}{n}\sum_{s=1}^{k} N(I_s) \times \frac{f_Y(m(I_s) - m(I_p))\ \Pr(X \in I_p)}{\sum_{t=1}^{k} f_Y(m(I_s) - m(I_t))\ \Pr(X \in I_t)}$$

- Can be computed in O(k^2) time, where k is the number of intervals.

How well does this work?

- Uniform random variable [-0.5, 0.5]





Talk Overview

- Randomization protects information at the individual level.
- Algorithm to reconstruct the distribution of values.
- Use reconstructed distributions to build decision-tree classifier.
- How well does it work?

Decision Tree Classification

Classification: Given a set of classes, and a set of records in each class, develop a model that predicts the class of a new record.

```
Partition(Data S)
begin
   if (most points in S are of the same class) then
      return;
   for each attribute A do
      evaluate splits on attribute A;
   Use best split to partition S into S1 and S2;
   Partition(S1);
   Partition(S2);
end
```

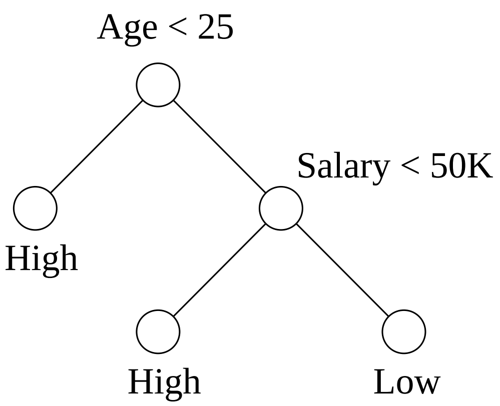
Initial call: Partition(TrainingData)

Example: Decision Tree Classification

Training Set:

Age	Salary	Credit Risk
23	50K	High
17	30K	High
43	40K	High
68	50K	Low
32	70K	Low
20	20K	High

Decision Tree:



Training using Randomized Data

- Need to modify two key operations:
  - Determining a split point.
  - Partitioning the data.
- When and how do we reconstruct the original distribution?
  - Reconstruct using the whole data (globally) or reconstruct separately for each class?
  - Reconstruct once at the root node or reconstruct at every node?

Training using Randomized Data (cont.)

- Determining split points:
  - Candidate splits are interval boundaries.
  - Use statistics from the reconstructed distribution.
- Partitioning the data:
  - Reconstruction gives estimate of number of points in each interval.
  - Associate each data point with an interval by sorting the values.

Algorithms

Global:

- Reconstruct for each attribute once at the beginning.
- Induce decision tree using reconstructed data.

ByClass:

- For each attribute, first split by class, then reconstruct separately for each class.
- Induce decision tree using reconstructed data.

Local:

- As in ByClass, split by class and reconstruct separately for each class.
- However, reconstruct at each node (not just once).

Talk Overview

- Randomization protects information at the individual level.
- Algorithm to reconstruct the distribution of values.
- Use reconstructed distributions to build decision-tree classifier.
- How well does it work?

Methodology

- Compare accuracy of Global, ByClass and Local against
  - Original: unperturbed data without randomization.
  - Randomized: perturbed data but without making any corrections for randomization.
- Synthetic data generator from [AGI+92].
- Training set of 100,000 records, split equally between the two classes.

Quantifying Privacy

If it can be estimated with $c\%$ confidence that a value x lies in the interval $[x_1,\ x_2]$, then the interval width $(x_2 - x_1)$ defines the amount of privacy at $c\%$ confidence level.

- Example: Privacy Level for Age[10,90]
  - Given a perturbed value 40
  - 95% confidence that true value lies in [30,50]
  - $\frac{\text{Interval Width : 20}}{\text{Range : 80}} \Rightarrow$ 25% privacy level

- Uniform: between $[-\alpha,\ +\alpha]$
- Gaussian: mean $\mu = 0$ and standard deviation σ

	Confidence		
	50%	95%	99.9%
Uniform	$0.5 \times 2\alpha$	$0.95 \times 2\alpha$	$0.999 \times 2\alpha$
Gaussian	$1.34 \times \sigma$	$3.92 \times \sigma$	$6.8 \times \sigma$

Synthetic Data Functions

- Class A if function is true, Class B otherwise.

F1 $(\texttt{age} < 40) \lor ((60 \le \texttt{age})$

F2 $((\texttt{age} < 40) \land (50K \le \texttt{salary} \le 100K)) \lor$
$((40 \le \texttt{age} < 60) \land (75K \le \texttt{salary} \ge 125K)) \lor$
$((\texttt{age} \ge 60) \land (25K \le \texttt{salary} \le 75K))$

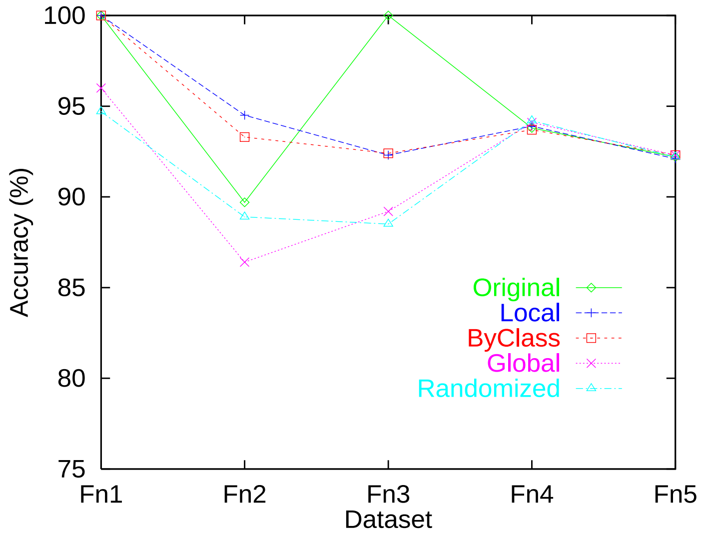
F3 $((\texttt{age} < 40) \land$
$(((\texttt{elevel} \in [0..1]) \land (25K \le \texttt{salary} \le 75K)) \lor$
$((\texttt{elevel} \in [2..3]) \land (50K \le \texttt{salary} \le 100K)))) \lor$
$((40 \le \texttt{age} < 60) \land$
$(((\texttt{elevel} \in [1..3]) \land (50K \le \texttt{salary} \le 100K)) \lor$
$(((\texttt{elevel} = 4)) \land (75K \le \texttt{salary} \le 125K)))) \lor$
$((\texttt{age} \ge 60) \land$
$(((\texttt{elevel} \in [2..4]) \land (50K \le \texttt{salary} \le 100K)) \lor$
$((\texttt{elevel} = 1)) \land (25K \le \texttt{salary} \le 75K))))$

F4 $(0.67 \times (\texttt{salary} + \texttt{commission}) - 0.2 \times \texttt{loan} - 10K) > 0$

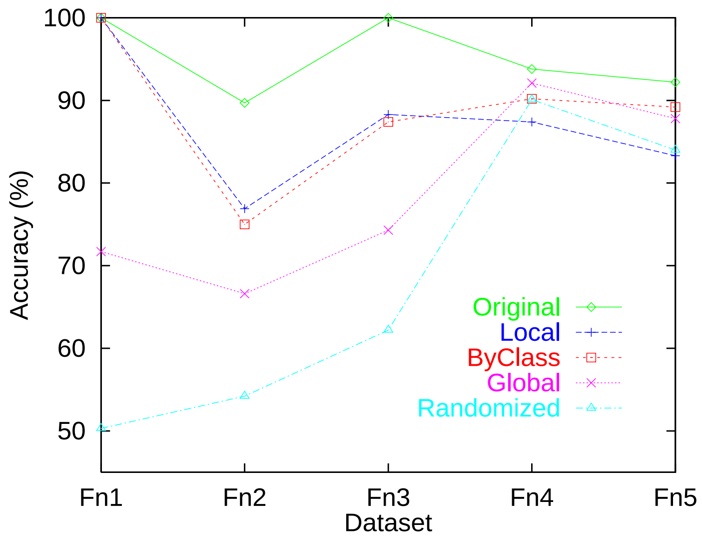
F5 $(0.67 \times (\texttt{salary} + \texttt{commission}) - 0.2 \times \texttt{loan} +$
$0.2 \times \texttt{equity} - 10K) > 0$
where $\texttt{equity} = 0.1 \times \texttt{hvalue} \times \max(\texttt{hyears} - 20, 0)$

Classification Accuracy

Privacy Level: 25% of Attribute Range

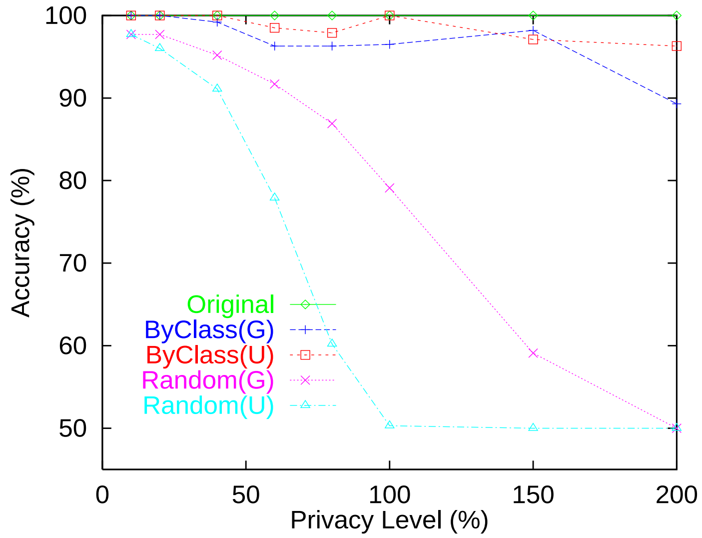


Privacy Level: 100% of Attribute Range

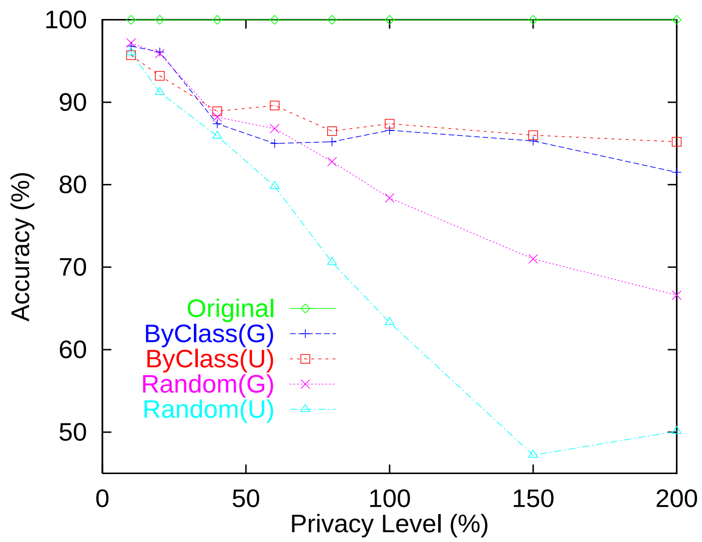


Change in Accuracy with Privacy

Fn 1



Fn 3



Conclusions

- Have your cake and mine it too!
  - Preserve privacy at the individual level, but still build accurate models.
- Future work:
  - other data mining algorithms,
  - characterize loss in accuracy,
  - other randomization functions,
  - better reconstruction techniques, ...