Privacy-Preserving Data Mining

Rakesh Agrawal Ramakrishnan Srikant
IBM Almaden Research Center

Presented by Guiwen Hou

Motivation

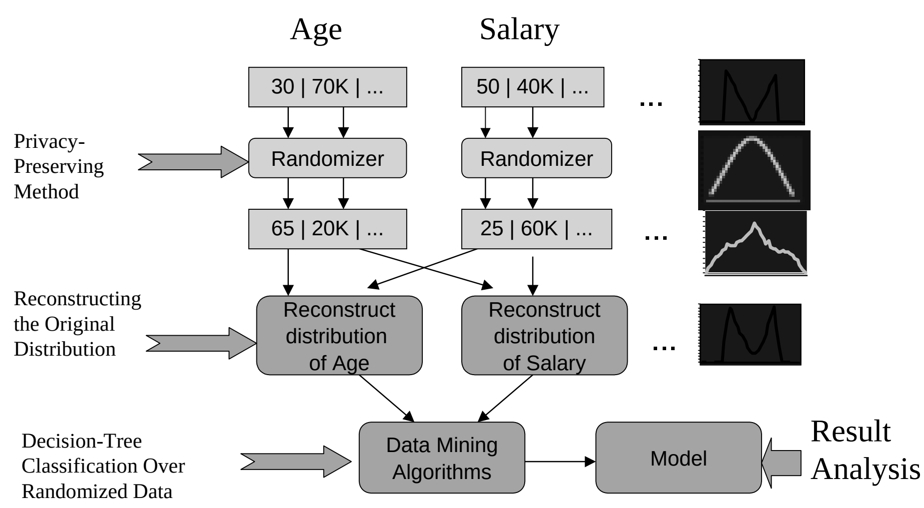
- Dramatic increase in digital data
- World Wide Web
- Growing Privacy Concerns

A Surveys of web users
  - 17% privacy fundamentalists, 56% pragmatic majority, 27% marginally concerned (Understanding net users' attitude about online privacy, April 99)
  - 82% said having privacy policy would matter (Freebies & Privacy: What net users think, July 99)

Technical Question

- The primary task in data mining: development of models about aggregated data.
- A Person
  - May not divulge at all the values of certain fields
  - May not mind giving true values of certain fields
  - May be willing to give not true values but modified values of certain data
- Can we develop accurate models without access to precise information in individual data records?
  - Randomization Approach
  - Cryptographic Approach

Randomized Approach Introduction



Based on "Privacy Preserving Data Mining: Challenges and Opportunities"

Talk Overview

- Introduction
- Privacy-Preserving Method
- Reconstructing the Original Distribution
- Decision-Tree Classification Over Randomized Data
- Experiment Result
- Conclusion and Future Work

Privacy-Preserving Method

- Value-Class Membership

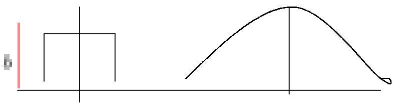
  Discretize continuous valued attributes. Values for an attribute are partitioned into a set of disjoint, mutually-exclusive classes. Instead of returning a true value, it returns the interval that the value lies.

- Value Distortion

  Add random component to data, return a value $x_i + r$ Instead of X_i

  - Uniform - Gaussian

$$P(x) = \begin{cases} \frac{1}{b-a} & \text{for } a < x < b \\ 0 & \text{for } x < a, x > b \end{cases}$$



$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2},$$

Based on R.Conway and D.Strip “select Partial Access to a Database”, In Proc, ACM Annual Conf.

Quantifying Privacy

- Measurement of how closely the original values of a modified attribute can be estimated.
- If it can be estimated with c% confidence that a value x lies in the interval [x1,x2], then the interval width (x2-x1) defines the amount of privacy at c% confidence level.

- Discretization : Assumed that intervals are of equal width W
- Uniform: random variable between [-a,a], The mean of the random variables is 0
- Gaussian: The random variable has normal distribution, with mean u= 0 and stand deviation σ

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

Table 1: Privacy Metrics

Talk Overview

- Introduction
- Privacy-Preserving Method
- Reconstructing the Original Distribution
  - Problem
  - Reconstructing Procedure
  - Reconstruction Algorithm
  - How does it work
- Decision-Tree Classification Over Randomized Data
- Experiment Result
- Conclusion and Future Work

Reconstructing The Original Distribution

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 (known)
- Given
  - $x_1+y_1, x_2+y_2, ..., x_n+y_n$ (Perturbed Value)
  - the probability distribution of X+Y (known)

Estimate the probability distribution of X.

Reconstructing The Original Distribution (Procedure)

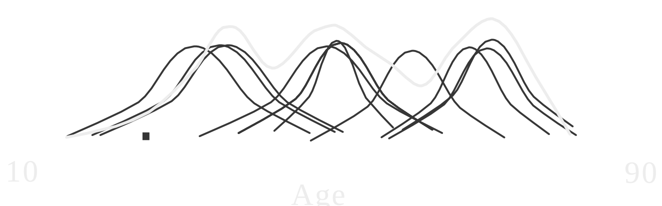
- Step1: Get single point density functions

Use Bayes' rule for density functions

Based on “Privacy Preserving Data Mining: Challenges and Opportunities”

Reconstructing The Original Distribution (Procedure)

- Step2 : Combine estimates of where point came from for all the points:



Based on “Privacy Preserving Data Mining: Challenges and Opportunities”

Reconstructing The Original Distribution (Bootstrapping)

f_X^0 := Uniform distribution
j := 0 // Iteration number
repeat

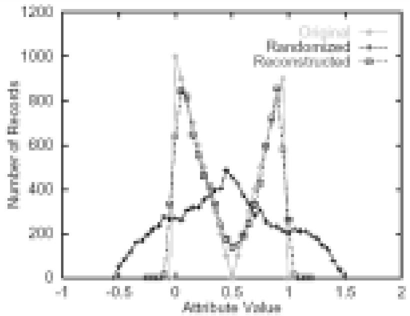
Compute new $f_X^{j+1}(a)$ based on $f_X^j(a)$
(Bayes' rule)

j := j+1
until (stopping criterion met)

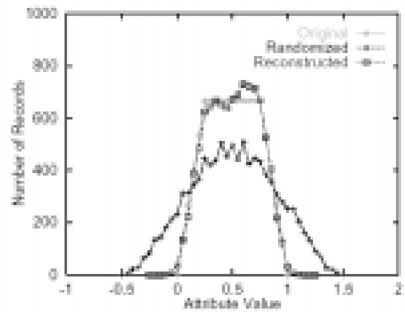
Stopping Criterion: Difference between successive estimates becomes very small

How well it works

- Uniform random variable [-0.5, 0.5]



Triangles



Plateau

Talk Overview

- Introduction
- Privacy-Preserving Method
- Reconstructing the Original Distribution
- Decision-Tree Classification Over Randomized Data
  - Decision Tree Algorithm
  - Demo a Decision Tree
  - Training using Randomized Data
  - Methods of building decision tree using Randomized Data
- Experiment Result
- Conclusion and Future Work

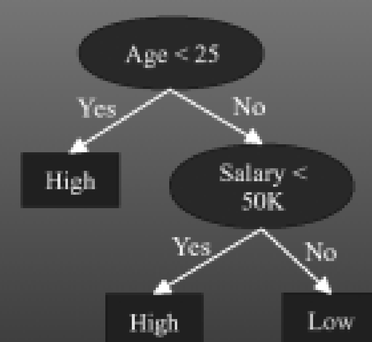
Decision Tree Classification

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)
```

An Example of Decision Tree

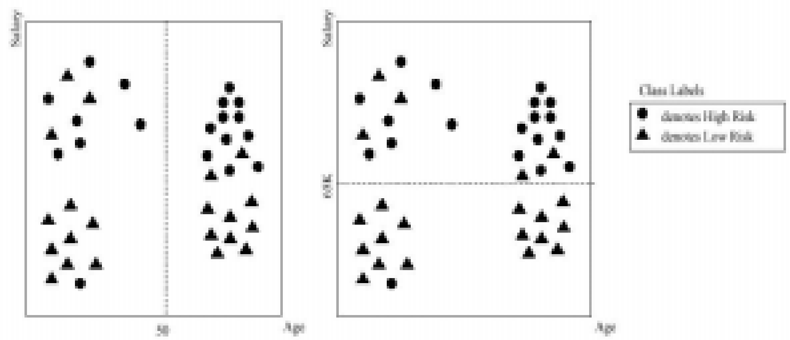
Age	Salary	Credit Risk
23	50K	High
17	30K	High
43	40K	High
68	50K	Low
32	70K	Low
20	20K	High



Selecting split point using “gini index”

We use the gini index to determine the goodness of a split. For a data set S containing examples from m classes, gini(S) = 1 - $\sum_j p_j^2$ where p_j is the relative frequency of class j in S. If a split divides S into two subsets S_1 and S_2 , the index of the divided data $gini_{split}$ (S) is given by $gini_{split}$ (S) = n1/n x gini(S_1) + n2/n x gini(S_2).

Note that calculating this index requires only the distribution of the class values in each of the partitions.



Selecting split point using “gini index”(cont.)

SPLIT: Age <= 50

	High	Low	Total
S1 (left)	8	11	19
S2 (right)	11	10	21

For S1: P(high) = 8/19 = 0.42 and P(low) = 11/19 = 0.58
For S2: P(high) = 11/21 = 0.52 and P(low) = 10/21 = 0.48
Gini(S1) = 1-[0.42x0.42 + 0.58x0.58] = 1-[0.18+0.34] = 1-0.52 = 0.48
Gini(S2) = 1-[0.52x0.52 + 0.48x0.48] = 1-[0.27+0.23] = 1-0.5 = 0.5
Gini-Split(Age<=50) = 19/40 x 0.48 + 21/40 x 0.5 = 0.23 + 0.26 = 0.49

SPLIT: Salary <= 65K

	High	Low	Total
S1 (top)	18	5	23
S2 (bottom)	1	16	17

For S1: P(high) = 18/23 = 0.78 and P(low) = 5/23 = 0.22
For S2: P(high) = 1/17 = 0.06 and P(low) = 16/17 = 0.94
Gini(S1) = 1-[0.78x0.78 + 0.22x0.22] = 1-[0.61+0.05] = 1-0.66 = 0.34
Gini(S2) = 1-[0.06x0.06 + 0.94x0.94] = 1-[0.004+0.884] = 1-0.89 = 0.11
Gini-Split(Age<=50) = 23/40 x 0.34 + 17/40 x 0.11 = 0.20 + 0.05 = 0.25

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 of Building Decision Tree

- "Global" Algorithm
  - Reconstruct for each attribute once at the beginning
- "By Class" Algorithm
  - For each attribute, first split by class, then reconstruct separately for each class.
- "Local" Algorithm
  - As in By Class, split by class and reconstruct separately for each class.
  - However, reconstruct at each node (not just once).

Talk Overview

- Introduction
- Privacy-Preserving Method
- Reconstructing the Original Distribution
- Experiment Result
  - Experimental Methodology
  - Synthetic Data Functions
  - Classification Accuracy
  - Accuracy vs. Randomization Level
- Conclusion and Future Work

Experimental Methodology

- Compare accuracy against
  - Original: unperturbed data without randomization.
  - Randomized: perturbed data but without making any corrections for randomization.
- Test data not randomized.
- Synthetic data generator from [AGI+92].
- Training set of 100,000 records, a test set of 5,000 records. split equally between the two classes.

Synthetic Data Functions

Class A if function is true, Class B otherwise

- F1

  (age < 40) or ((60 <= age)
- F2

  ((age < 40) and (50K <= salary <= 100K)) or
  ((40 <= age < 60) and (75K <= salary <= 125K)) or
  ((age >= 60) and (25K <= salary <= 75K))
- F3

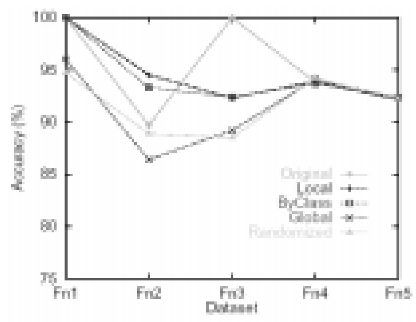
  ((age < 40) and (((elevel in [0..1]) and (25K <= salary <= 75K)) or
  ((elevel in [2..3]) and (50K <= salary <= 100K))) or
  ((40 <= age < 60) and ...
- F4

  (0.67 x (salary+commission) - 0.2 x loan - 10K) > 0
- F5

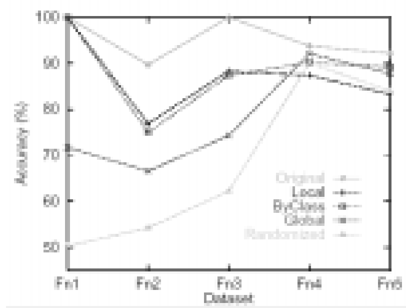
  (0.67 x (salary+commission) - 0.2 x loan +0.2 x equity - 10K) > 0
  Where equity = 0.1 x hvalue x max(hyears – 20.0)

Classification accuracy

Uniform



Privacy Level: 25% of Attribute Range



Privacy Level: 100% of Attribute Range

Privacy Level

Example:

Privacy Level for Age[10,90]

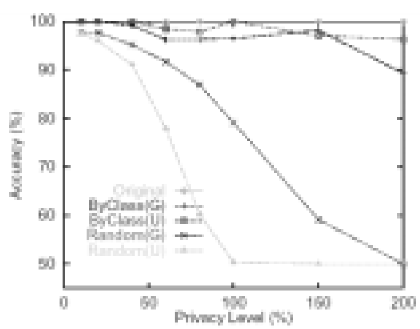
Given a perturbed value 40

95% confidence that true value lies in [30,50]

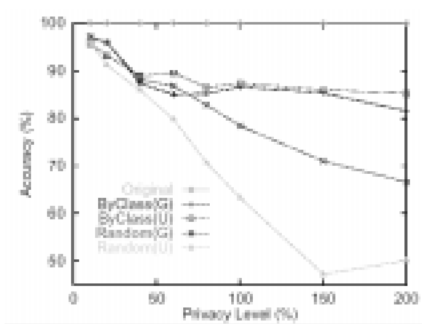
(Interval Width : 20) /(Range : 80) = 25% privacy level

25% privacy level @ 95% confidence

Accuracy vs. Privacy Level

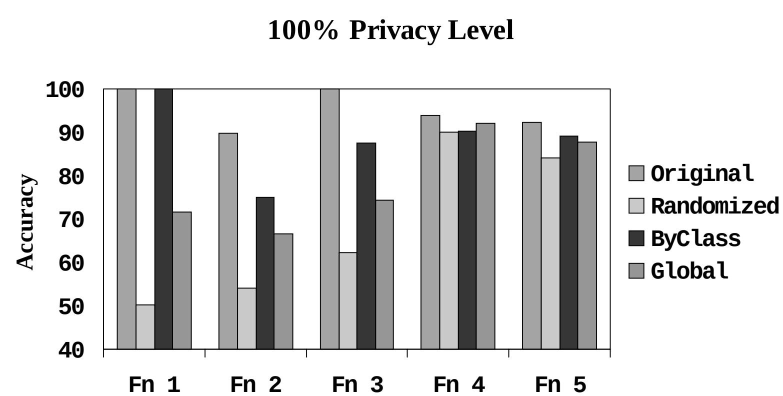


Fn 1



Fn 3

Acceptable loss in accuracy



Conclusions and Future Work

Conclusions

- Preserve privacy at the individual level, but still build accurate models
- By class and Local are both effective in correcting for the effects of perturbation
- Local performed better than By class but required more computation
- For same privacy level, Uniform perturbation did slightly worse than Gaussian.

Future work

- Other data mining algorithms,
- Guard against potential privacy breaches
  - Some randomized values are only possible from a given range.

    Example: Add U[-50,+50] to age and get 125 , True age is 75.

  - Most randomized values in a given interval come from a given interval.

    Example: 60% of the people whose randomized value is in [120,130] have their true age in [70,80].
- Find approach to process categorical and boolean type data

Thank You

?