Privacy Preserving Data-Mining

Survey on **R. Agrawal** and **R. Srikant** paper:

“Privacy preserving data mining”

ACM SIGMOD Conference on Management of Data
Dallas, Texas, May 2000.

Xavier Rondé-Oustau

01 . 10 . 2003

Introduction

- Some data need to remain unrevealed
- We need to make statistics from these data

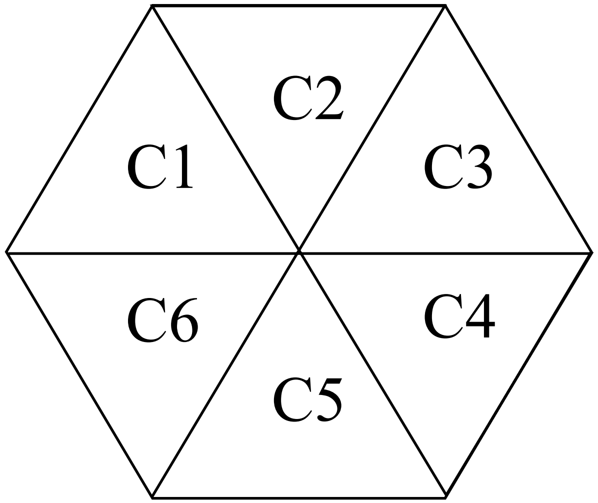
- Two methods are presented in this paper
  - Value Class Membership
  - Value Perturbation

- High accuracy can be reached with high privacy

Value Class Membership

- The hexagone is the set of value possible for an attribute

- C1... C6 are the 6 classes, that exclude each other, and that complete each other to form the whole set.



example of sensitive value: salary. from 0 € to 1 billion € for example. classes:

- 0 - 1000 €
- 1000 € - 2000 €
- 2000 € - 5000 €
- 5000 € - 15000€
- 15000 € - 50000 €
- 50000 € - 1 billion €

VALUE DISTORSION

- The global principle is to add random noise to the sensitive value: data=value+noise

    - Uniform noise:
    The added noise has a uniform distribution over an interval [-a a].

    - Gaussian noise:
    The added noise has a gaussian distribution with zero mean.

Reconstruction (1)

- The aim is to find the original distribution X from value perturbated data W=X+Y.

- We suppose we have enough data to make statistical approximations

- We suppose we have the computing facilities required to processed the data

RECONSTRUCTION (2)

$$F'_{X_1}(a) = \int_{-\infty}^{a} f_{X_1}(z | X_1 + Y_1 = w_1) dz$$

$$F'_{X_1}(a) = \frac{\int_{-\infty}^{a} f_Y(w_1 - z) f_X(z) dz}{\int_{-\infty}^{\infty} f_Y(w_1 - z) f_X(z) dz}$$

$$f'_X(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{f_Y(w_1 - z) f_X(z)}{\int_{-\infty}^{\infty} f_Y(w_1 - z) f_X(z) dz}$$

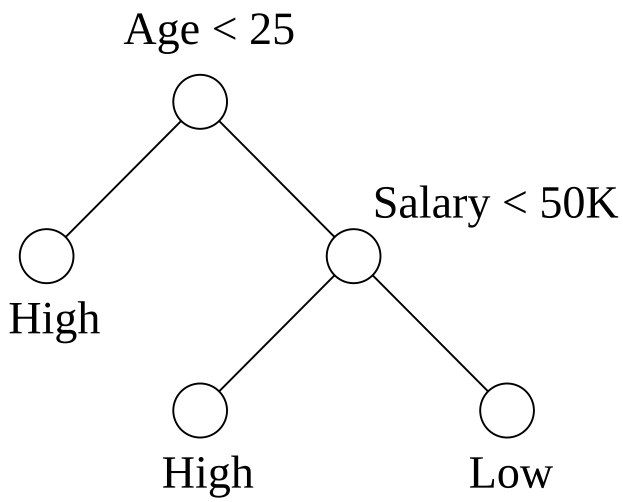
Reconstruction (3)

$$f_X^{j+1}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{f_Y(w_1 - z) f_X^j(z)}{\int_{-\infty}^{\infty} f_Y(w_1 - z) f_X^j(z) dz}$$

There is a method discribed in the original paper to improve the algorithm to a $O(n^2)$ complexity and the accuracy increases when n increses.

DECISION TREE CLASSIFIERS (1)

Classification of data into classes, at each not of the tree, there is a test.



Building a tree in 2 phases:
- growth phase
- pruned phase

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

DECISION TREE CLASSIFIERS (2)

The gini is used to determine the best split in a decision classifier tree, i.e. when the gini of a split is minimum.
Only distributions are needed to compute such trees

$$gini(S) = 1 - \sum p_j^2$$

$$gini_{split}(S) = \frac{n_1}{n} gini(S_1) + \frac{n_2}{n} gini(S_2)$$

DECISION TREE CLASSIFIERS (3)

Data are first divided into classes.

Place of the reconstruction in the process:

- **Global**: done at the beginning, first step.
- **ByClass**: done at the beginning, for each class.
- **Local**: same beginning as ByClass but the re-construction is done at each node of the tree.

Experimental Results (1)

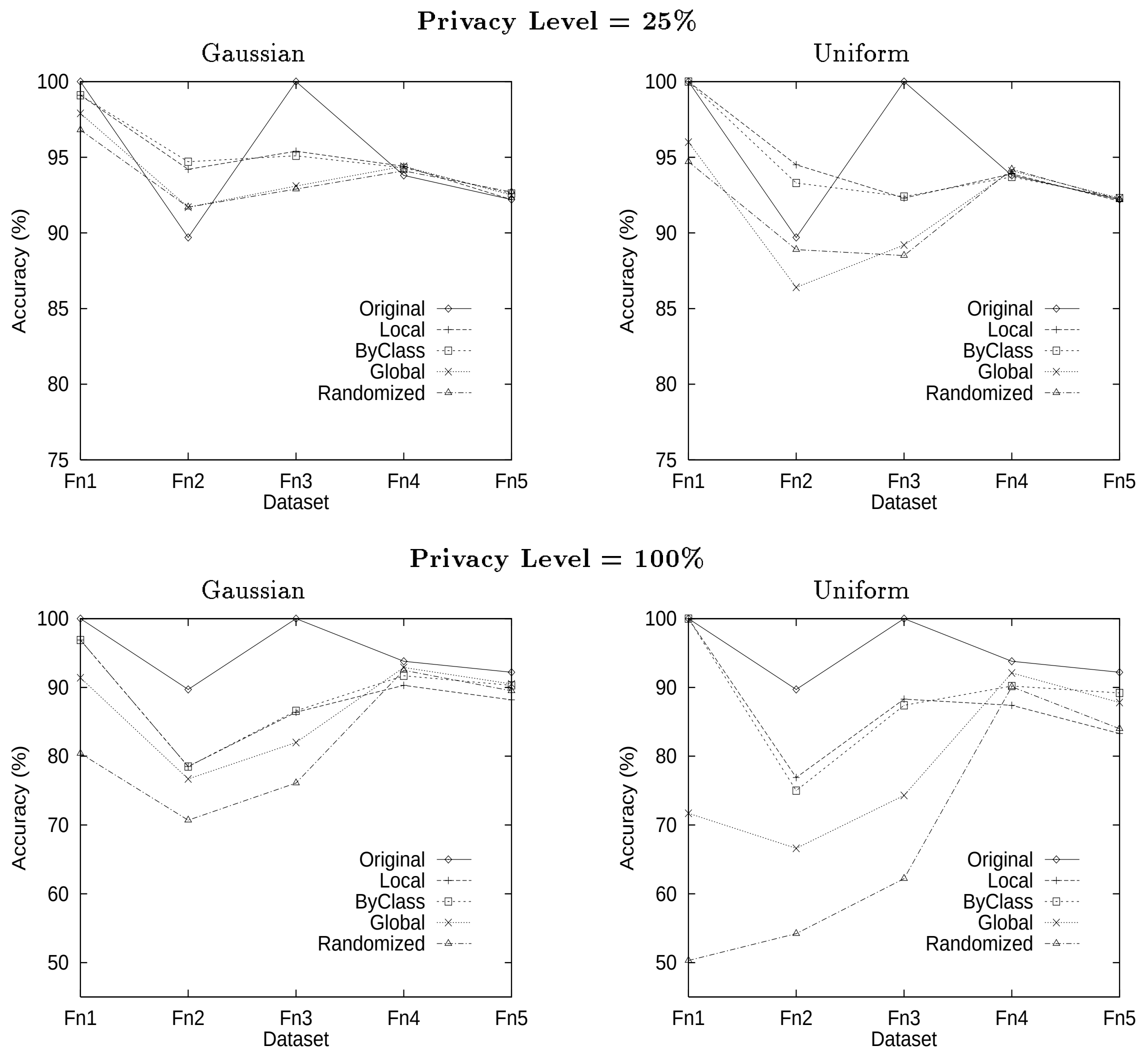
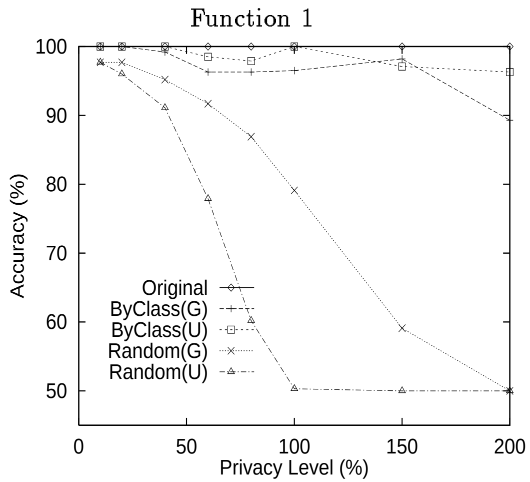
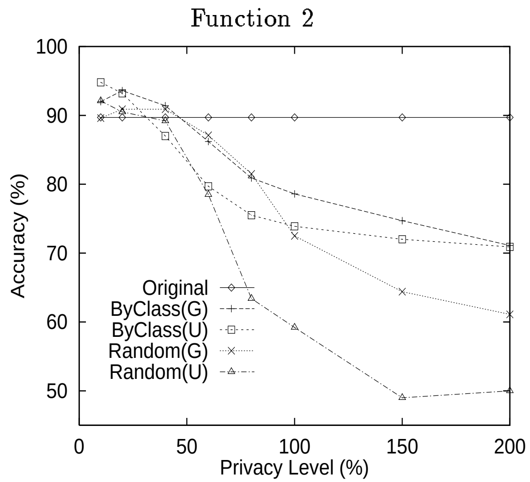
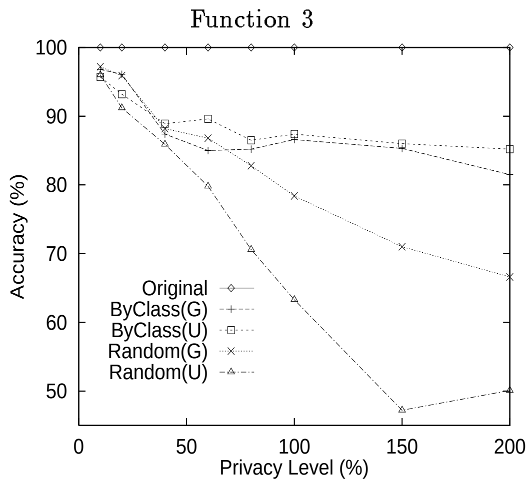


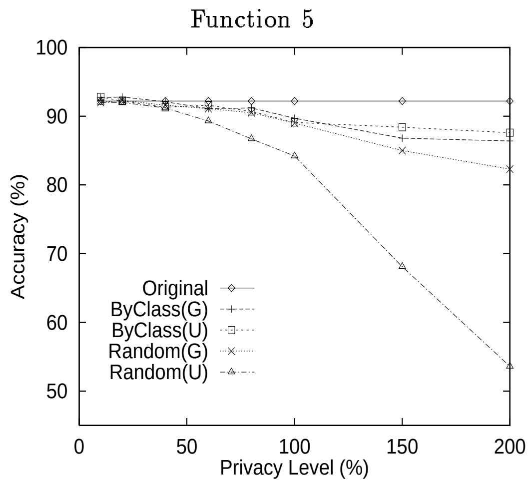
Figure 5: Classification Accuracy

EXPERIMENTAL RESULTS (2)









Conclusion

- Good accuracy for reconstruction at ByClass & Local schemes (for uniform & gaussian randomisation)

- Complexity a lot lower for ByClass compared to Local.

- Better privacy for gaussian randomisation, but difficult to figure out and explain the effects on data.