Workshop on Privacy Security and Data Mining

Privacy vs Precision in Data Mining

Vladimir Estivill-Castro

Griffith UNIVERSITY
Australia

Workshop on Privacy Security and Data Mining

Elements of Discussion

- Privacy Threats
  - why privacy?
  - why a balance?
- Data Mining vs Statistical Databases
  - Protection Mechanisms in Statistical Databases
  - why data swapping/noise addition?
- Privacy of Organizations

Workshop on Privacy Security and Data Mining

Privacy Threats

- Vast amounts of personal data are being collected, processed and sold:
  - medical records
  - criminal records
  - bank balances and credit records
  - phone calls
  - shopping habits
  - rental histories
  - driving records

Mi-PAL

griffithuniversity

Workshop on Privacy Security and Data Mining

Commerce of Personal Data

- In US, yearly at least
  - 400 million credit records
  - 700 million annual drug prescription records
  - 100 million medical records
  - 600 million personal records

  are owned by 200 largest superbureaus, and billions of records are owned by federal, state and local governments (1996).

See The Economist May 1st,1999 "The End of Privacy"

5

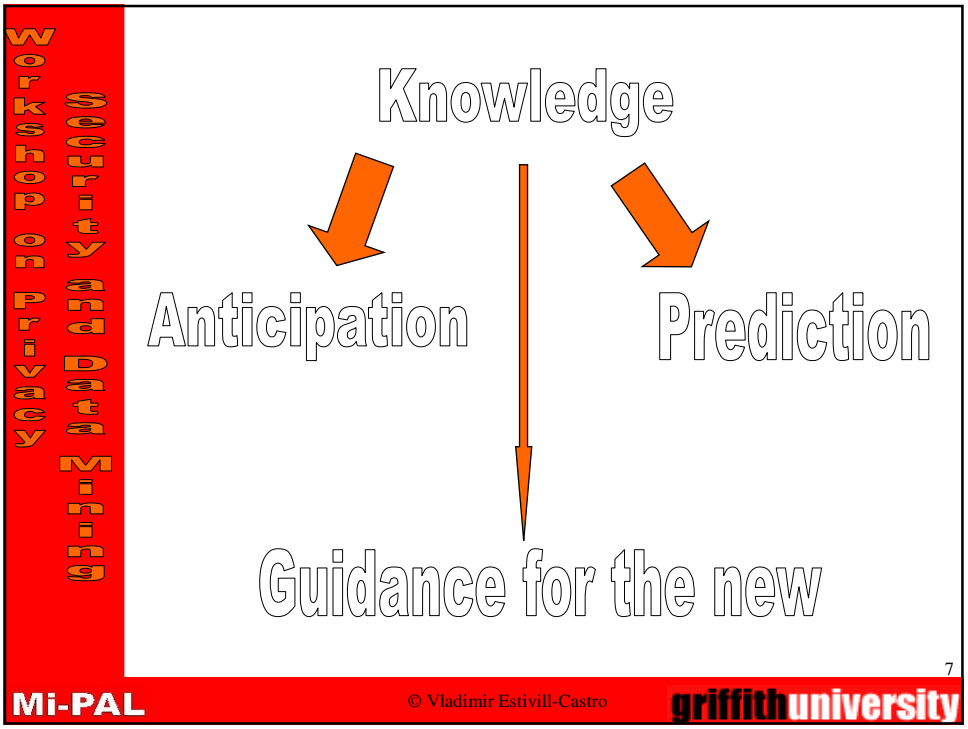
Workshop on Privacy Security and Data Mining

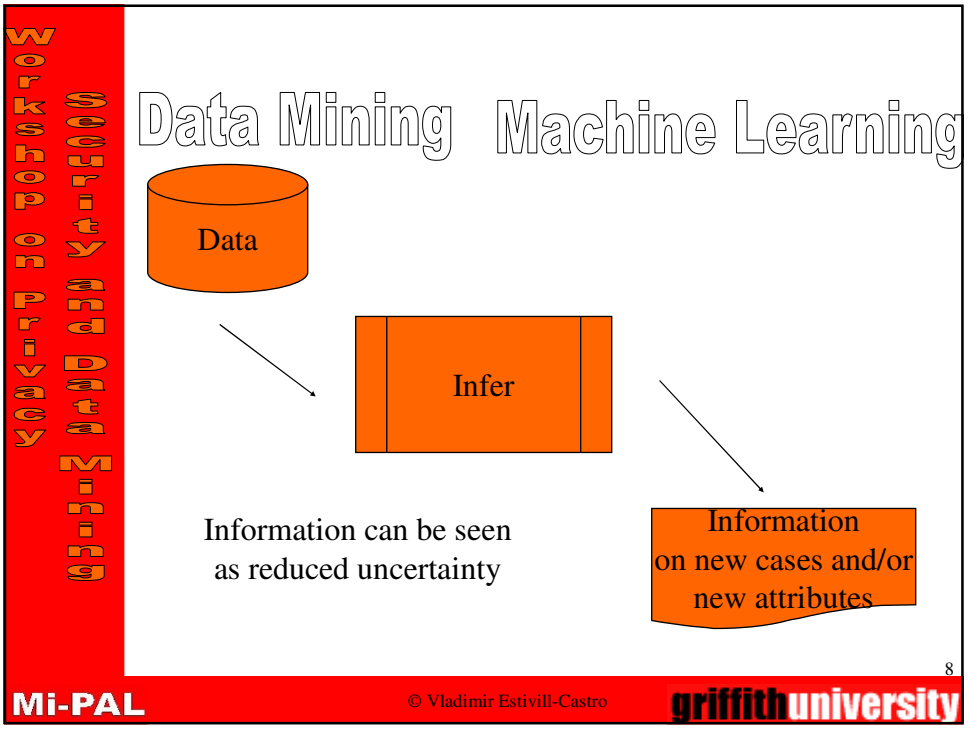
- ***Knowledge Discovery and Data Mining (KDDM)*** refers to techniques for extracting information from data and suggesting patterns in very large databases.
- KDDM facilitates research/ data exploration in many areas, including:
  - marketing
  - medicine
  - crime investigation (e.g., the Okalahoma City bombing)
  - fraud detection [Italy KDD-99 San Diego] but also Australian Taxation Office and HIC [PAKDD-99]

6

Mi-PAL

griffithuniversity


Workshop on Privacy Security and Data Mining
Knowledge
Anticipation
Prediction
Guidance for the new
7
Mi-PAL
© Vladimir Estivill-Castro
griffithuniversity


Workshop on Privacy Security and Data Mining
Data Mining
Machine Learning
Data
Infer
Information can be seen as reduced uncertainty
Information on new cases and/or new attributes
8
Mi-PAL
© Vladimir Estivill-Castro
griffithuniversity

Workshop on Privacy Security and Data Mining

Privacy Threats

- ***Privacy*** is the interest (right) of individuals to **control information** about themselves [Clarke].
- Question of ***ownership***: it is commonly assumed that the gathering organisation owns the data, and individuals have, at best, an 'interest' in information about themselves.
- The existing laws are far behind the developments in technology, and no longer offer adequate protection.

Workshop on Privacy Security and Data Mining

Dataveillance

- *Data Surveillance* .-systematic use of personal data systems in the investigation or monitoring of the actions and communications of one or more people
- *Behind the Scenes KDDM*.- Use of Data Mining technology without the consumer being aware [ACM SIGKDD Newsletter **1**]

Workshop on Privacy Security and Data Mining

Privacy Threats

- ***Illustration:***

  In late 1980's, the California Department of Motor Vehicles was selling driver-licence data about state residents. The data included:
  - addresses
  - dates of birth
  - driving records
  - number and type of vehicles

  19 year old Robert Brado 'bought' (for $1 USD) the address of actress Rebecca Schaeffer, and later killed her in her apartment.

With KDD technology is now easier to narrow down the possibilities for the address of a person?

Workshop on Privacy Security and Data Mining

Points of view

Data owners: 'What's the big deal?'

- often see privacy concern as unnecessary and unreasonable
- more moderate then 20 years ago ('too much privacy already')
- face increasing public opposition
  - In 1990, Lotus Development Corporation intended to make 120 million US consumers' data available for sale; the project was cancelled in Jan 1991, as a result of strong public opposition.
- 'externalise' some costs for their services or products [Pigou, 1989; Laudon, 1996 CACM]

Mi-PAL © Vladimir Estivill-Castro griffithuniversity

Workshop on Privacy Security and Data Mining

Points of view

- ***Individuals***: 'Where did you get my name…and why?'
  - 1995 and 1996 Equifax\Harris Consumer Privacy surveys show:
    - 80% believe that consumers have lost control over their personal information
    - 59% refused at some point to give an information to a business or a company
    - 24% have been the victim of improper invasion of privacy
    - 74% are concerned about the potential negative effects of computerised medical records

In 1970, only 33% consider computers as a threat to privacy

Workshop on Privacy Security and Data Mining

Points of view

- ***KDDM researchers***:
  - privacy regulations are inconsistent
  - data collection should not be restricted
  - different opinions about KDDM as a threat to privacy
    - [O'Leary IEEE Expert **10**(2) 1995]

How can we main the data if we can not see it?

Mi-PAL

griffithuniversity

Workshop on Privacy Security and Data Mining

The balance

- **Privacy advocates face considerable opposition, since Data Mining brings collective benefits in many contexts.**
  - **How could planning decisions be taken if census data were not collected?**
  - **How could epidemics be understood if medical records were not analysed?**
  - **Data Mining has been also instrumental in detecting money laundering operations, telephone fraud, and tax evasion schemes.**
  - **In some such domains, it can be argued that privacy issues are secondary in the light of a common good.**

15

Workshop on Privacy Security and Data Mining

Privacy Threats

- Revitalised privacy threats by KDDM:
  - secondary use of personal information
  - handling misinformation
  - granulated access to personal information
- New privacy threats posed by KDDM:
  - stereotypes
  - guarding personal data from KDDM researchers
  - individuals from training sets
  - combination of patterns
    - [Clifton & Marks ACM SIGMOD 1996]
    - The "Dark Side of KDD -- Causalty vs Correlation" [KDD-99 San Diego]

16

Mi-PAL

griffithuniversity

Workshop on Privacy Security and Data Mining

Secondary use of personal information

- refers to any use other than the one for which the information was originally collected.
  - Flying points / Purchasing Clubs
- A consumer attitudes survey (Culnan, 1993) shows that 96% of respondents believe that some types of personal information should never be shared without permission.

A business has collected large amounts of data on transactions by customers. KDDM offers analysis of this operational data

17

Workshop on Privacy Security and Data Mining

Misinformation

- can cause serious and long-term damage and individuals should be able to challenge the correctness of information about themselves.
- ***Example***:

  In early 1990's, District Cablevision in Washington, D.C., fired its employee James Russell Wiggins, because of Wiggins' drug conviction. The information was obtained from Equifax, Atlanta. However, the information was about James Ray Wiggins, and the case ended up in court.

In KDDM the issue of validity of the pattern is crucial

18

Workshop on Privacy Security and Data Mining

Granulated access

- access to personal data should be limited to relevant information only.
- New privacy law in Germany dramatically reduced the number of variables in census data.
  - Obstruct the social / collective effort of any data collection.

Workshop on Privacy Security and Data Mining

Stereotypes

- General patterns discovered by KDDM tools may lead to ***stereotypes*** and prejudices.
- If patterns are based on properties such as race, sex, religion, etc, they can be very sensitive and controversial.
- ***Example***: research by Murray and Herrnstein on average IQ of different races.

KDDMs inferences apply to groups

Workshop on Privacy Security and Data Mining

Guarding personal data from KDDM researches

- Three solutions:
  - restricting access to personal data - this can make KDDM task very difficult (even impossible)
  - providing KDDM researcher with perturbed data which contains similar general patterns as the original data
  - provide very small samples [Clifton, IFIP]
    - those in the sample are not protected
    - miners may collaborate to get larger sample
    - can not find patterns

Workshop on Privacy Security and Data Mining

Protecting privacy of individuals from training set

- one of the most common tasks in KDDM is the ***classification task***:
  - input:
    - set of classes
    - training set consisting of pre-classified cases
  - output:
    - classifier, I.e., an operator that assigns classes to unclassified cases
- KDDM classifiers are typically very accurate when applied to the cases to the training set, and should be modified to have the same accuracy when applied to the training set and new cases.

Statistical Databases near KDDM and OLAP

Statistical database models:

- Abstract model

Name	City	Age	Sex	Status	Child	HIV
Smith	Syd	33	F	M	2	1
Jones	Mel	24	M	W	3	0
Black	Ade	33	M	S	0	0
White	Syd	43	M	D	5	0
Adams	Bri	22	F	D	3	0
Brown	Per	51	F	M	5	0
Green	Dar	31	M	W	1	1
White	Mel	22	F	M	3	0
Baker	Mel	40	M	M	2	0
Ling	Syd	22	M	M	0	0

Workshop on Privacy Security and Data Mining

Statistical Databases vs KDDM

- Tabular model

Age HIV	20-29	30-39	40-49	50-59
0	4	1	2	1
1	0	2	0	0

- some information loss

Mi-PAL

griffithuniversity

Statistical Databases vs KDDM

- Multidimensional matrix model - essentially the same as the multidimensional cube in OLAP.

HIV / Age	0	1
22	3	0
24	1	0
31	0	1
33	1	1
40	1	0
43	1	0
51	1	0

Results (theoretical)

- OLAP multidimensional data cubes are equivalent to the abstract model of statistical databases
  - one can reconstruct the other
  - [Shoshani 97, Brankovic and Estivill-Castro99]

Workshop on Privacy Security and Data Mining

Protection Mechanisms in Statistical Databases

- Query restriction:
  - query size control
  - query set overlap control
  - maximum order control
  - partitioning
  - cell suppression
  - auditing
- Noise addition:
  - probability distribution data perturbation
  - fixed data perturbation
  - random sample
  - varying output perturbation
  - rounding

Workshop on Privacy Security and Data Mining

Query restriction

provide exact answers to some queries, and reject others that may lead to compromise

- Advantages
  - statistical quality of released information is high
- Disadvantages
  - overly restrictive
  - inadequate against skilled users or previous knowledge
  - require high initial implementation effort
  - deny information / obscure patterns

Workshop on Privacy Security and Data Mining

Noise addition

introducing errors either to data or to results of queries

- Advantages
  - resistant to users' supplementary knowledge
  - answers to all queries
- Disadvantages
  - allow partial dis-clousure
  - may produce low statistical quality

Workshop on Privacy Security and Data Mining

Noise addition

introducing errors either to data or to results of queries

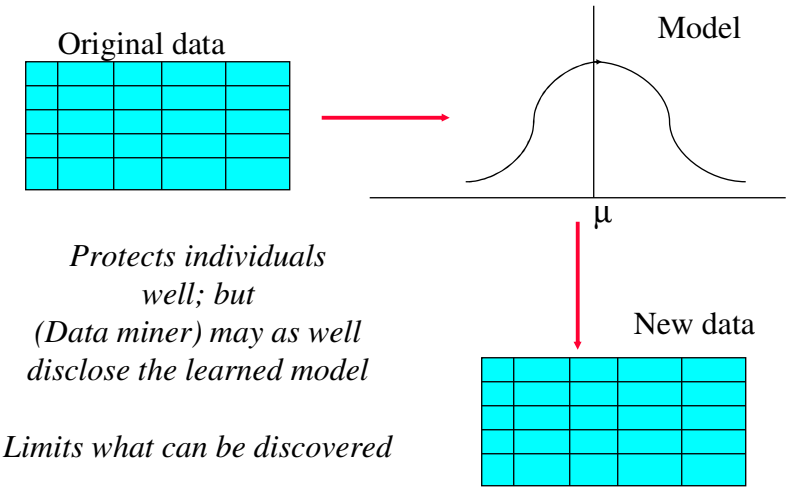
- Noise addition:
  - probability distribution data perturbation
    - learn parameters of a model and generate data from the model
      - may as well disclose what I have learned
  - fixed data perturbation
  - random sample
  - varying output perturbation
  - rounding

Mi-PAL

griffithuniversity

Workshop on Privacy Security and Data Mining

Problem with data generation



Protects individuals well; but (Data miner) may as well disclose the learned model

Limits what can be discovered

Workshop on Privacy Security and Data Mining

A Balance

- Statistics involving confidential attributes reveal some information about individual values
  - example
    - if $50 million is the answer to the statistical query "What is the average gross income of all small business in town X?" one learns that
      - one business has gross income at least $50 million
- All methods trade privacy of individual values for statistics (pattern) distortion.
  - To protect all patterns requires to know them all
    - infinite CPU time / infinite data

Workshop on Privacy Security and Data Mining

Protection Mechanisms in Statistical Databases

- Data swapping interchanges the values in the database in such a way that low-order statistics are preserved.
  - Those involving a few (k) attributes

D			D '		
Sex	Age	HIV	Sex	Age	HIV
F	20	1	F	20	0
F	30	0	F	30	1
M	20	0	M	20	1
M	30	1	M	30	0

Workshop on Privacy Security and Data Mining

Data swapping

- Finding a data swap is considered intractable
  - statistical parameters of the perturbation are generally publishable without compromising compromising privacy
  - allowing parametric statistical inference to proceed
    - for example, the average, the maximum, the range, the standard deviation of a population can still be known

What is the maximum amount of noise that ensures a minimum of privacy and a maximum of information loss?

Likely to be NP-hard/NP-complete

Mi-PAL

griffithuniversity

Workshop on Privacy Security and Data Mining

The corporate world

- (Public data about a competitor
  ^
- my data of operation with suppliers and customers)
  +
- KDDM tools
  =
- competitive advantage

Chris Clifton leads the debate

Workshop on Privacy Security and Data Mining

Summary

- Proposal
  - noise addition is best
  - increases un-certanity on individual data
  - General patterns are obtainable
    - parameters of noise can be made public
    - parametric statisticians can recover original model
      - but not individual values
  - other ways to find the (relaxed) swap
    - rough sets
- A balance may be possible between privacy and KDDM


Workshop on Privacy Security and Data Mining
THANK YOU
37
Mi-PAL
© Vladimir Estivill-Castro
griffithuniversity