PURDUE UNIVERSITY

When do data mining results violate privacy?

Chris Clifton

March 17, 2004

This is joint work with Jiashun Jin and Murat Kantarcıoğlu

CERIAS Center for Education and Research in Information Assurance and Security

Indiana Center for Database Systems

Individual Privacy: Protect the "record"

- Individual item in database must not be disclosed
- Not necessarily a person
  - Information about a corporation
  - Transaction record
- Disclosure of parts of record may be allowed
  - Individually identifiable information

Privacy-Preserving Data Mining to the Rescue!

- Methods to let us mine data without disclosing it
  - Data obfuscation: value swapping, noise addition, …
  - Secure Multiparty Computation
  - ?
- Nobody sees (real) individual records
- *Is this enough?*

What is Missing: Do Results Violate Privacy?

- The approaches discussed give results without revealing data items
  - *Maybe the results violate privacy!*
- Example: (Privately) learn a regression model to estimate salary from public data
  - Privacy preserving data mining ensures salaries of "training samples" not revealed
  - But model can be used to estimate those salaries

Doesn't this violate privacy?

Does a Classifier Violate Privacy?

- Goal: Develop a classifier to predict likelihood of early-onset Alzheimer's
  - Make it available on the web so people can use it and prepare themselves…
- Problem: Don't want Insurance companies to use it
  - But that's okay, since not all the input attributes are known to insurers
- Can't the insurance company just fix knowns and try several values for unknowns?
  - *Should improve insurer's estimate!*

Formal Problem Definition

- $X=(P,U)^T$ distributed as $N(0,\Sigma)$

$$\Sigma = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$$

$-1<r<1$ is the correlation between P and U

- Let $s_i = C_0(x_i) = \begin{cases} 1 & \text{if } p_i \geq u_i \\ 0 & \text{otherwise} \end{cases}$

But the Insurer (adversary?) has Prior Knowledge

- Adversary likely to have training data
  - Causes of death public
  - Likely as complete in public and sensitive as our training set
- Gives adversary $\Pr[S=1 \mid P=p] = \Phi\left(\frac{1-r}{\sqrt{1-r^2}}\, p\right)$

  where $\Phi(\cdot)$ is the cdf of $N(0,1)$ $= \begin{cases} \geq 1/2, \text{if } p \geq 0, \\ < 1/2, \text{otherwise} \end{cases}$
- Adversary's classifier: $s_i = \begin{cases} 1 & \text{if } p \geq 0, \\ 0 & \text{otherwise} \end{cases}$

Classifier Doesn't Hurt Privacy!

- What if we make *our* classifier public?

$$s_i = \begin{cases} 1 & \text{if } \Pr[U \leq P \mid P = p_i] > \frac{1}{2}, \\ 0 & \text{otherwise} \end{cases}$$

$$\Pr[U \leq P \mid P = p_i] = \Phi\left(\frac{1-r}{\sqrt{1-r^2}}\, p_i\right)$$

Challenge: Define Metrics and Evaluate Tradeoffs

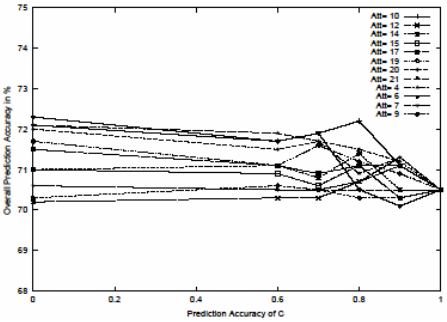
- Public à Sensitive

$$\sup_i \left(\Pr[C(X) \neq Y \mid Y = i] - \frac{1}{n_i} \right)$$

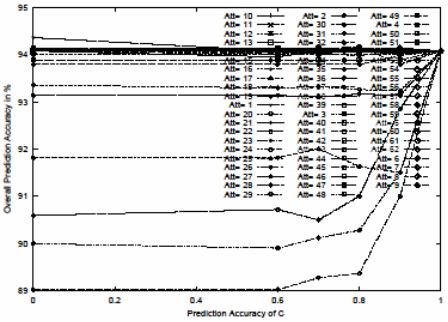
- Public+Unknown à Sensitive
- Public+Sensitive à Sensitive
- Assume adversary has access to Sensitive data for some individuals:
  - Public à Sensitive
  - Public à Unknown

Does Estimating an Unknown Help?

- Examples from UCI
  - Altered values of an attribute
  - Did it make a difference?



Credit-G dataset



Splice dataset

Another Issue: Limitations on Results

- Data mining results may violate privacy
  - Must restrict results to prevent such violations
- Some results may be unacceptable

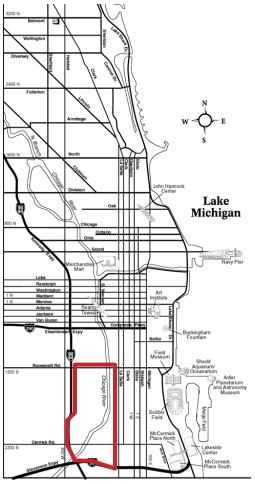
  Need not violate privacy of "training data"
  - Particular uses of data proscribed
  - Data mining only allowed for prearranged purpose

Regulatory Examples

- Use of Call Records for Fraud Detection vs. Marketing
  - FCC § 222(c)(1) restricted use of individually identifiable information

    Until overturned by US Appeals Court
  - 222(d)(2) allows use for fraud detection
- Mortgage Redlining
  - Racial discrimination in home loans prohibited in US
  - Banks drew lines around high risk neighborhoods!!!
  - These were often minority neighborhoods
  - Result: Discrimination (redlining outlawed)

  What about data mining that "singles out" minorities?



How do we Constrain Results?

- Need to specify what is:
  - Acceptable
  - Forbidden
- Can't we just say what is/isn't allowed?
  - *If it were this easy, we wouldn't need to mine the data in the first place!*
- Idea: Constraint-based mining *(KDD Explorations 4(1))*
  - Specify bounds on what we can (can't?) learn
  - Privacy-preserving data mining *enforces* those constraints
- How do we know if privacy is good enough?
  - *Metrics*

Need to Know

We have a good reason for anything we learn

- Good criteria for Secure Multiparty Computation
  - Results can be justified
  - Nothing outside of results is learned
- Likely real-world acceptability
  - Legal precedents
  - Social norms

Okay, it isn't a metric…

Need to Know: Legally/Socially Meaningful

- Access to U.S. Government classified data requires:
  - Clearance
  - *Need to Know*
- Antitrust law
  - Collaboration generally suspect
  - But okay *when it benefits the consumer*

Antitrust Example: Airline Pricing

- Airlines share real-time price and availability with reservation systems
  - Eases consumer comparison shopping
  - Gives airlines access to each other's prices

  Ever noticed that all airlines offer the same price?
- Shouldn't this violated price-fixing laws?
  - *It did!*

Antitrust Example: Airline Pricing

- Airlines used to post "notice of proposed pricing"
  - If other airlines matched the change, the prices went up
  - If others kept prices low, proposal withdrawn
  - This violated the law
- Now posted prices effective immediately
  - If prices not matched, airlines return to old pricing
- Prices are still all the same
  - *Why is it legal?*

The Difference: *Need to Know*

- Airline prices easily available
  - Enables comparison shopping
- Airlines can change prices
  - Competition results in lower prices
- *These are needed to give desired consumer benefit*
  - "Notice of proposed pricing" wasn't

Need to Know: How do we use it?

- Secure Multiparty Computation approach
  - "Need to know" data defined as results
  - Prove nothing else shared
- Potentially privacy-damaging values could be inferred from results
  - Need to know trumps this
- To be determined: How to specify need to know
  - Domain specific?

Bounded Knowledge

We can't violate privacy very well

- Metric for data obscuration techniques
  - Example: Add random value from [-1,1]
  - Can't rely on observed data if exact value needed
- How do we capture this in general?

Quantification of Privacy
Agrawal and Aggarwal '01

- Intuition: A random variable distributed uniformly between [0,1] has half as much privacy as if it were in [0,2]
- Also: if a sequence of random variable A_n, n=1, 2, … converges to random variable B, then privacy inherent in A_n should converge to the privacy inherent in B

Differential entropy

- Based on differential entropy:

$h(A) = -\int_{\Omega A} f_A(a) \log_2 f_A(a) da$ where Ω_A is the domain of A

- Random variable U distributed between 0 and a, h(U)=$\log_2$(a). For a=1, h(U)=0
- Random variables with less uncertainty than uniform distribution on [0,1] have negative differential entropy, more uncertainty à positive differential entropy

Proposed metric

- Propose $\Pi(A)=2^{h(A)}$ as measure of privacy for attribute A
- Uniform *U* between 0 and a: $\Pi(U)=2^{\log_2(a)}=a$
- General random variable A, $\Pi(A)$ denotes length of interval over which a uniformly distributed random variable has equal uncertainty as A
- Ex: $\Pi(A)=2$ means A has as much privacy as a random variable distributed uniformly in an interval of length 2

Anonymity

We may know what, but we don't know who

- Goal is to preserve individual privacy
  - Individual privacy is preserved if we can not distinguish people on any basis
- Idea: Okay if individuals indistinguishable
  - You know that Joe is above 60
  - You would like to learn which data entries might be about Joe
  - If for every data entry $\Pr\{Age > 60 \mid X_i\} = 0.3$ each is equally likely to belong to Joe
- *Haven't gained any information!*

Anonymity: Formal Definitions

- Two records $(X_1, X_2 \in X)$ that belongs to different indiviuals are p - indistinguishable if for every function $f : X \rightarrow \{0,1\}$ that can be evaluated in polnomial - time $|\Pr\{f(X_1) = 1\} - \Pr\{f(X_2) = 1\}| \leq p$ where $0 < p < 1$
- Definition: A data mining process is said to be *p-individual privacy preserving* if at every step of the process, any two individual records are p-indistinguishable.

Conclusions

- Privacy Preserving Data Mining techniques emerging
- Many challenges for the next generation of data mining research
- Progress needs a vocabulary
  - Need to define “privacy preserving”
  - Metrics for privacy