Privacy Preserving Data Mining Algorithms

(3 hours)

Lila Ghemri

Department of Computer Science

Texas Southern University

ghemri_lx@tsu.edu

This work was supported by NSF grants 1241772

Any opinions, findings, conclusions, or recommendations expressed herein are those of the authors and do not reflect the views of the National Science Foundation

Module Organization

- Audience: Senior computer science majors. A graduate level module can also be used with additional assignments.

- Prerequisites: Database, data structures, familiarity with data mining techniques

Privacy-Preserving Data Mining

- Data Mining is becoming more important for government and business alike
- Increase in amount of personal data that is being stored.
- Increased sophistication of data mining algorithms

Identifiers and Pseudo-Identifiers

- Identifiers: Those characteristics that can be used to uniquely identify a person, such as names, SSN, and other unique IDs
- Pseudo-identifiers: Do not uniquely identify a person, but could be used in conjunction with public records to identify a person.

Example of identifiers and pseudo-identifiers

Student	Gender	Grade	Course
ID1	M	B	CS111
ID2	M	C	CS112
ID3	F	A	CS113
ID3	F	C	CS113

Name	Email	Courses
Bernard Kump	kb@email.com	CS112, CS111
Alan Trump	TR@email.com	CS112, CS113
Alice Bump	BA@email.com	CS111, CS114
Bella Crump	CB@email.com	CS112,CS113

Privacy Preserving Data Publishing

- These techniques focus on modifying the data before the data is mined.
- Randomization techniques: uses distortion techniques (adding randomized noise, random projection)
- K-anonymity Method: the idea is that a given record cannot be distinguished from k-1 records.

Randomization techniques

- Data records $X=\{x_1,x_2,x_3,....x_n\}$ for $x_i \in X$
- We add a noise component which is denoted $y_1, y_2,....y_n$
- Thus the new set of distorted records as:

$$x_1+y_1, x_2+y_{2;}.....,....., x_n+y_n$$

This new set of records is denoted: $z_1, z_2,....,z_n$

Only relates to one attribute.

Randomization example

x	0	1	2	3	4	5
P[X=x]	0.05	0.1	0.2	0.4	.15	.10

We add a noise component y to x. For example if x=0, add y=1. If x=1, add y= 3, and keep the same distribution.

x	0	1	2	3	4	5
y	1	2	1	2	2	2
z	1	3	4	5	6	7
P[Z= z]	0.05	0.1	0.2	0.4	.15	.10

Micro aggregation

- The dataset is divided in *g* groups of *k* or more individuals.
- Confidentiality rules require replacing individual values with values computed on small aggregates prior to publication
- The optimal *k* partition is defined as the one that maximizes the group homogeneity.
- Works best with continuous data
- Group homogeneity is calculated as

$$\text{SSE}= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)'(x_{ij}-\bar{x}_i)$$

K anonymity

- A protected data set is said to satisfy *k*-anonymity for k >1 if for each combination of quasi-identifiers values, at least k records exist in the dataset sharing that combination.
- *k*-anonymity is achieved through generalizations
- Works best with categorical data.

k-anonymized table, k=2

Age	Race	Gender	Zip Code	Disease
*	White	*	21002	Common cold
*	White	*	21002	Flu
27	*	Female	92010	Flu
27	*	Female	92010	Hypertension

Utility Based Methods

- There is a natural tradeoff between privacy and accuracy of mining methods. Most methods will reduce the effectiveness of the data mining methods.
- The loss of specific information about certain individuals may affect data quality

Example

- Original Table

tide	Age	Education	Zip-code	Income	Target
T1	24	BSc	53711	40K	Y
T2	25	BSc	53712	50K	Y
T3	30	MSc	53713	50K	N
T4	30	MSc	53714	80K	N
T5	32	MSc	53715	50K	N
T6	32	PhD	53716	100K	N

2-Anonymized Tables

TABLE 1

gild	tide	Age	Education	Zip code	Income	Target
G1	T1	[24-25]	BSc	[53711-53712]	40K	Y
G1	T2	[24-25]	BSc	[53711-53712]	50K	Y
G2	T3	30	MSc	[53711-53714]	50K	N
G2	T4	30	MSc	[53711-53714]	80K	N
G3	T5	32	GradSchool	[53715-53716]	50K	N
G3	T5	32	GradSchool	[53715-53716]	100K	N

TABLE 2

gId	tId	Age	Education	Zip code	Income	Target
G1	T1	[24-30]	ANY	[53711-53714]	40K	Y
G2	T2	[25-32]	ANY	[53712-53716]	50K	Y
G3	T3	[30-32]	MSc	[53713-53715]	50K	N
G1	T4	[24-30]	ANY	[53711-53714]	80K	N
G3	T5	[30-32]	MSc	[53713-53715]	50K	N
G2	T5	[25-32]	ANY	[53712-53716]	100K	N

Example queries

- Q1: “How many customers under age 29 are in the data set?”
- Q2: Is an individual with age =25, Education= BSc, Zip Code 53712 a target customer?
- Table 1: Answer to Q1 = 2, Q2= Y
- Table 2: Answer to Q1= [0-4], Q2 =Yes and No with 50%
- Table 1 has better utility than Table 2 and returns more precise answers, although both tables have same k-privacy

Utility-based Privacy Preservation

- Has two goals:
  - Protecting the private information
  - Preserving the data utility
- Challenges:
  - Utility Measure: how to model it in different applications
  - Balance between Privacy and Utility: Sometimes conflicting goals
- Efficiency and scalability: Privacy is NP-hard, so when utility is added, this makes the problem even more computationally challenging

Privacy Models

- *K*- Anonymity : already introduced.
- Sometimes *K*- Anonymity is not enough.
- Example from Table2, T3 and T5 are generalized to the same class. However since their income is the same, an attacker can infer that T3 salary is $50K.
- With Table 1, the attacker has only 50% opportunity to know the real income of T3.

2-Anonymized Tables

TABLE 1

gild	tide	Age	Education	Zip code	Income	Target
G1	T1	[24-25]	BSc	[53711-53712]	40K	Y
G1	T2	[24-25]	BSc	[53711-53712]	50K	Y
G2	T3	30	MSc	[53711-53714]	50K	N
G2	T4	30	MSc	[53711-53714]	80K	N
G3	T5	32	GradSchool	[53715-53716]	50K	N
G3	T5	32	GradSchool	[53715-53716]	100K	N

TABLE 2

gId	tId	Age	Education	Zip code	Income	Target
G1	T1	[24-30]	ANY	[53711-53714]	40K	Y
G2	T2	[25-32]	ANY	[53712-53716]	50K	Y
G3	T3	[30-32]	MSc	[53713-53715]	50K	N
G1	T4	[24-30]	ANY	[53711-53714]	80K	N
G3	T5	[30-32]	MSc	[53713-53715]	50K	N
G2	T5	[25-32]	ANY	[53712-53716]	100K	N

L- Diversity

- *L*- Diversity: Complements *k*-anonymity by requiring certain diversity on the sensitive attribute.
- A table is *l*-diverse if each equivalence class contains *l* "well represented" sensitive values.
- Consider a table T =(A_1,…, A_n, S) and a constant *c* and *l* where (A_1,.. A_n) is a quasi-identifier and S is a sensitive attribute.
- Suppose an equivalence class *EC* contains value s_1,…,s_m with frequency f(s_1),.., f(s_m) of the sensitive attribute S.
- EC satisfies (c,l)-diversity with respect to S if :

$$f(s_1) < c \sum_{i=l}^{m} f(s_i)$$

Privacy Through Output Perturbation

- Sometimes, results of data mining applications can compromise the privacy of data,
- Results of data mining are modified to preserve privacy, such as association rule hiding methods.

Results Perturbation

- This method relies on a special program called a Sanitizer.
- A user communicates with the Sanitizer by issuing queries f_1, f_2,… and receiving answers a_1, a_2 …
- **A Sanitizer algorithm** may decide not to answer a query or to modify query results, by adding noise to query answers.
- The purpose is to mask individual records but leave global trends visible.

Sanitizer Algorithm

The answer given by the output perturbation sanitizer on a query *f* is distributed according to **San**(**x,** *f) =f(***x)** + Y, Y is referred to as noise:

The random variable Y is taken from a probability distribution.

Informally, a Sanitizer is private, if no adversary A gains significant knowledge about an individual entry of a database **x** beyond what A could have learnt by interacting with a similar database **x'** where that individual entry is arbitrarily modified.

Query Auditing

- Informally, auditing is the process of examining past actions to check whether they were conform to official policies.
- In the context of database, auditing is examining queries that were answered in the past to determine whether they have been used to divulgate confidential information.

Offline Auditing

Preliminary Definition:

- Let X=$\{x_1,\ldots,x_n\}$ be the set of private attribute values of n individuals in a database.
- An aggregate query $q=(\mathcal{Q}, f)$ specifies a set of the records $\mathcal{Q} \subseteq \{1,\ldots,n\}$ and an aggregate function f such as sum, max, min, or median. The result $f(\mathcal{Q})$, is f applied to the subset $\{x_i \mid i \in \mathcal{Q}\}$.
- We call $\mathcal{Q}$ the *query set* of q

Full Disclosure

- Full Disclosure:

Given the set of private values X and a set of aggregate queries $\mathcal{Q} = \{q_1,..,q_t\}$ posed over this data set with corresponding answers $\{a_1,...a_t\}$ the goal of an offline auditor is to determine if an individual private value can be deduced.

Definition: An element $x_i \in X$ is fully disclosed by a query set Q if it can be uniquely determined.

Example of full disclosure

If the query set consists of a single query asking for the sum of salaries of all female employees in the company and Alice is the only female employee in the company , then the answer to this query determines Alice's salary.

In general, the answers to many different queries can be stitched together by a user to determine an individual private value.

The goal of the auditor is to prevent that !!

Online Auditing

- Given a sequence of queries q_1,.., q_{t-1}that have already been posed and corresponding answers a_1,.., a_{t-1}that have already been supplied
- Given a new query q_t.
- The task of an online auditor is to determine if the new query should be answered or denied to prevent a privacy breach.
- Each answer a_i is either the true answer f_i (Q_i) to query q_i or a “denial”

Cryptographic Methods

- Data may be distributed across multiple sites.
- Some application may wish to compute a common function.
- Cryptographic protocols may be used in order to communicate the results from each site, without divulgating the data in each site.
- Secure multi party computation is used in those cases.

Secure Multiparty Computation

- "Millionaire Problem":

Two millionaires wish to know who is richer. However they do not want to find out inadvertently any additional information about each others wealth. How can they carry out such a conversation

Refer to "**Secure Multiparty Computation and Privacy"** by Yehuda Lindell (see biblograph**y)**

Measuring Privacy

Most important approaches:

- Statistical Measures of Anonymity:
  - Query restriction, Anonymity via random perturbation, etc.
- Probabilistic Measures of Anonymity:
  - Using Mutual Information, Distance between source and perturbed data distributions,

Suggested Exercise 1

- Consider the table below and Disease as the private attribute.
- Are there any queries that would violate the privacy. Redesign the table to mitigate these risks.

Age	Race	Gender	Zip Code	Disease	Drug	Dosage
24	White	Female	21004	Common cold	Aspirin	1000mg
26	White	Male	21002	Flu	Aspirin	1000mg
27	Black	Female	92010	Flu	Aspirin	1000mg
27	Black	Female	92010	Hypertension	Aspirin	81 mg

Suggested Exercise 2

Gender	Name	Zip Code	Income
M	John	11001	98k
F	Cathy	11001	62k
M	Ben	13010	36k
F	Laura	13010	115k
M	William	14384	44k
F	Lisa	15013	100k

- Consider the table on the left with salary being the sensitive attribute.
- Find the identifiers and remove them.
- Make this table 2-K anonymous
- Can you still infer Cathy's salary ? Lisa's

Suggested Exercise 3

Read the paper on “How to share a secret” at : http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.80.8910&rep=rep1&type=pdf

- And also “How to share a secret with cheaters” at https://www.christophedavid.org/w/c/files/ShareSecret/TompaHowToShareASecretWithCheaters.pdf

Describe with your own words the methods described in the paper on privacy.

Bibliography

-A General Survey of Privacy-Preserving Data Mining Models and Algorithms Charu Aggarwal, Philip Yu, Springer Verlag 2008

-Tutorial on Secure Multi-Party Computation and Privacy http://u.cs.biu.ac.il/~lindell/research-stat ements/tutorial-secure-computation.ppt.Yehuda Lindell. IBM T.J.Watson.

-Privacy-Preserving Data Publishing, B-C Chen, D Kifer, K. LeFevre and A. Machanavajjhala http://www.csd.uoc.gr/~hy558/papers/privacy_survey.pdf

-Providing k-Anonymity in Data Mining A. Friedman, R. Wolff, A. Schuster. The VLDB Journal — The International Journal on Very Large Data Bases Volume 17 Issue 4, July 2008, pages 789-804

-The Boundary Between Privacy and Utility in Data Publishing VLDB '07, September 2328, 2007, Vienna, Austria.