CS 498
t-closeness: Privacy Preserving Data Mining
(Group based Anonymization methods)

Advisor : Dr Arnab Bhattacharya

Ritesh Gupta - Y6185

November 2 , 2009

Abstract

The t-closeness model [2] was introduced in order to provide a safeguard against the similarity attacks on published dataset.It requires that the earth mover's distance between the distribution of a sensitive attribute within each equivalence class does not differ from the overall earth movers distance of the sensitive attribute in the whole table by more than a predefined parameter t. The model uses uses Earth Mover's distance as there is a need to have a distance metric which takes into account semantic distance between two attributes.

Recently papers have tried to focus on various shortcomings of t-closeness and come up with a novel solution. None of them is able to provide a safeguard againt all known possible attacks on t-closeness and yet be efficient. We will try to evaluate other models which came in the last year and lookout for efficient implementation of t-closeness both in terms of time complexity and/or Data utility.

Contents

1 Introduction

Privacy-preserving data mining has emerged as a very important issue to be addressed in recent times. This is because of the ability to store data of users had increased , use of social networks helps in yielding personal information , sophisticated data mining algorithms and high computational powers available with the adversary. All this makes it possible to leverage this information. Although most of the applications first remove the records having sensitive information like name, social security numbers(or any other unique identification number), other kind of attributes like sex, age, pin codes, profession can be combined to form a pseudo-identifier and the sensitive information can then be retrieved from public data records like census which contain all records.

There can be two broad ways to achieve the goal of privacy. First is to release limited data such that personal information cannot be extracted out of it but the overall heuristics are still close to original dataset. And second is to pre-compute heuristics and release them instead of any data. Advantage of releasing some limited data instead of pre-computed heuristics is a increased flexibility and availability for the users. So in Privacy Preserving Data Mining we look for methods to transform the original data such that heuristics determined from the transformed data are close to original heuristics and the privacy of users is not endangered.

Most of the privacy preserving methods use generalization as techniques to achieve privacy goal. This reduction although makes it possible to succesfully apply constraints of the model but may lead to loss of effectiveness of mining algorithms.This is the natural trade-off between privacy and data utility . Many models for privacy have come up over the last few years.

1.1 Various Methods

These are the current methods used for Privacy Preserving Data Mining.

- *Statistical Methods* :
  - Randomization methods
  - Swapping
  - Micro Aggregation
  - Synthetic data generation
- *Group based anonymization methods*:
  - k-anonymity
  - ℓ-diversity
  - t-closeness

Definition 1 (Micro data): *The data to be released after applying anonymization methods on it is called the Micro Data.*

Definition 2 (Sensitive attribute): *Attribute which must not be disclosed in the released microdata.*

Definition 3 (Quasi Identifier) : *Attributes or combination of attributes within a dataset which on their own are non-sensitive but on combination with external data are capable of identifying records.*

Definition 4 (Equivalence Class): *All set of tuples which cannot be distinguished from each other with respect to Quasi-Identifier are called an Equivalence Class.*

2 k-anonymity

The k-anonymity[3] model requires that within any equivalence class of the microdata there are atleast k records. In other words we should not be able to make ANY query to the database which returns less than k matches.

2.1 Achieving k-anonymity

k-anonymity is provided by use of generalization relationships between domains and between values that attributes can assume. Suppression is an complementary approach to providing k-anonymity.

Definition 5 (Generalization): *Given two domains D_1 and D_2 , $D_1 \leq D_2$ describes the fact that values of attributes in D_2 are more generalized values.*

Definition 6 (Suppression): *Removing data(ie. rows) from table so that it is not released in the microdata is called suppression.*

Definition 7 (K-minimal Generalization with suppression): *Generalization T_1 is k-minimal iff it satisfies k-anonymity, it does not enforce more suppression than allowed(some predefined parameter), and there does not exist another generalization satisfying these conditions less general than T_1.*

Definition 8 (Identity disclosure): *An individual can be directly linked to a particular record in the released data.*

Definition 9 (Attribute disclosure): *The chances of guessing the sensitive attribute of an individual increase because of released microdata.*

2.2 K-anonymity : Attacks

k-anonymity cannot provide a safeguard against attribute disclosure in all cases. A simple case of attribute disclosure will be when all the sensitive attributes within a equivalence class have same value. Here we would have achieved k-anonymity but we can accurately predict the sensitive attribute of any person who we can match to this equivalence class by using information in public domain.

	Department	Age	Course
1	ME	20	Mechanics
2	MME	21	Mechanics
3	ME	20	Mechanics
4	CHE	22	Algorithms
5	CHE	23	Psychology
6	CHM	22	Real Analysis
7	CSE	26	Algorithms
8	CSE	25	Algorithms
9	CSE	26	Mechanics

	Department	Age	Course
1	M*	[20-21]	Mechanics
2	M*	[20-21]	Mechanics
3	M*	[20-21]	Mechanics
4	CH*	[22-23]	Algorithms
5	CH*	[22-23]	Psychology
6	CH*	[22-23]	Real Analysis
7	CS*	[25-26]	Algorithms
8	CS*	[25-26]	Algorithms
9	CS*	[25-26]	Mechanics

Example 1 : The two tables show the original and anonymous version of the dataset. In the second table we have 3 equivalent classes. We have achieved 3-anonymity by generalization. The course attribute is sensitive. Let us assume that Alice can get from public information Bob's age,say 21 and department,say MME. Alice also knows that Bob's record is among one of the records in the original table. From second table, Alice can figure out that Bob's record is from first equivalence class and can thus figure out that he has taken Mechanics. This attack which predicts the sensitive attribute is the **homogeneity attack**.

Now consider second kind of attack, **background knowledge** attack.Lets say Alice knows Bob's department is CS from public information. She can conclude that Bob's record is one of the records in last equivalence class. Now if Alice knows that Bob is not really interested in Mechanics and would not take it , she can conclude that Bob must be studying algorithm. This attack, which happens because of the background knowledge with the adversary is called Background Knowledge attack. This also also be considered **probabilistic attack**. Suppose in beginning Bob was equally likely to study both mechanics and algorithms , his chances of studying algorithms were 0.5, but now are are increased to 0.66 (even without

background information).

2.3 Variants of K-anonymity

P-sensitive K-anonymity: It requires that within each equivalence class the number of distinct values for any confidential attribute is at least p.
Still insufficient to prevent att ribute disclosure.

3 ℓ-diversity

ℓ-diversity tries to put constraints on minimum number of distinct values seen within a equivalence class for any sensitive attribute. An equivalence class has ℓ-diversity[6] if there are ℓ or more well-represented values for the sensitive attribute. A table is said to be ℓ-diverse if each equivalence class of the table is ℓ-diverse.[6]

3.1 Variants of ℓ-diversity

- **Distinct ℓ-diversity** : This is the most general form of ℓ-diversity. Here well represented used in earlier definition of ℓ-diversity distinct means that there are at least ℓ-distinct values for the sensitive attributein any possible equivalence class. This prevents homogeneity attacks . But this does protect against probabilistic inference attacks. An example of the same will be when within an equivalence class some value appears much more frequently than the other values. Assuming that all values were equally likely at start adversary can now conclude that entity is more likely to have that more frequent value. This lead to the development of the following two stronger notions of ℓ-diversity.

- **Entropy ℓ-diversity** : For each equivalence class E , entropy is defined as Entropy(E) = $-\sum_{(s \epsilon S)}$ p(E, s) log p(E, s) where p(E, s) is the fraction of records in E that have sensitive value s and S is the domain of the sensitive attribute. The table has entropy ℓ-diversity if in each equivalence class E , entropy(E) $\geq$ log ℓ. The main problem with this model is that it is restricts the data a lot. For the entire table the entropy may be low if few values are very common.

- **Recursive (c, ℓ)-diversity** : Recursive (c, ℓ)-diversity restricts the less frequent values to not appear too rarely and similarly the most frequent value does not appear too frequently.

3.2 ℓ-diversity : Limitations

- ℓ-diversity is unnecessary and difficult to achieve for some cases
  Example : Let the original data have one sensitive attribute, pass or fail, the students who have failed or passed in a course.Further there are 1000 students enrolled and their corresponding records. Say 1% of the students have failed and rest have passed. We can see that both the values have sensitivity of different degree. A Student may not mind for others to know that he has passed but may not like others to know if he has failed.
  Here, 2-diversity is not desired for an equivalence class where there are records of

students who have passed and not failed. For achieving a 2-diverse table, there can be at most 1000 * 1% = 10 equivalence classes . This would mean large generalizations and large information loss.

- ℓ-diversity does not prevent attribute disclosure.

  Skewness Attack: Although identity disclosure is successfully handled by ℓ-diversity, it does not prevent attribute disclosure when the overall distribution is skewed.

  Example: Lets focus on again on the the example discussed above. Suppose one equivalent class has equal number of pass and fail records. Anyone belonging to that equivalent class would be considered to have 50% chance of having failed as compared with the 1% initially . This is thus a major privacy risk.
  Again consider an equivalence class with 49 fail records and 1 pass record. It would be satisfy 2-diverse but still there will be 98% chance of having failed for someone in that equivalence class, which is much more than 1% initially. This equivalence class has the same diversity as a class that has 1 failed and 49 pass records but we can clearly see that both have different levels of sensitity.

  Similarity Attack : These attacks occur when the sensitive attributes are semantically similar. Similarity attacks are the main motivation to look forward for yet another kind of privacy preserving methods like t-closeness.[2] The following example exaplains similarity attacks.

	Department	Age	Course
1	ME	20	Mechanics
2	MME	21	Relativity
3	ME	20	Rotational
4	CHE	22	Algorithms
5	CHE	23	Psychology
6	CHM	22	Real Analysis
7	CSE	26	Algorithms
8	CSE	25	Architecture
9	CSE	26	Mechanics

	Department	Age	Course
1	M*	[20-21]	Mechanics
2	M*	[20-21]	Relativity
3	M*	[20-21]	Rotational
4	CH*	[22-23]	Algorithms
5	CH*	[22-23]	Psychology
6	CH*	[22-23]	Real Analysis
7	CS*	[25-26]	Algorithms
8	CS*	[25-26]	Architecture
9	CS*	[25-26]	Mechanics

The second table shows an anonymized version satisfying 3-distinct diversity. The sensitive attribute is Course. Suppose one knows that Bob's record corresponds to the one among first 3(by knowing his department to be ME(say)), then one knows that Bob's Courses are among Mechanics, Relativity, Rotational . Having the information that Bob's record belongs to the first equivalence class enables adversary to know that Bob does some course related to physics as all the 3 courses are of this nature.This semantic relationship between different values of sensitive attribute is one of the main reasons why to look for another approach which also incorporates the semantic distance.

4 t-closeness

The t-closeness model [2] was introducted to overcome attacks which were possible on ℓ-diversity(like similarity attack). ℓ-diversity mdoel uses all values of a given attribute in a similar way(as distinct) even if they are semantically related. Also not all values of an attribute are equally sensitive.

Definition 10 (The t-closeness Principle[2]): *It requires that the earth mover's distance between the distribution of a sensitive attribute within each equivalence class does not differ from the overall earth movers distance of the sensitive attribute in the whole table by more than a predefined parameter t*

Now a distance metric between 2 distributions is desired . There are metrics like Kullback-Leibler and variational distance distance but these dont take into account semantic distance. We thus want to take into account the ground distances(semantic distance) among these values.Thus Earth Mover's distance (EMD)[10] is used. *The EMD is based on the minimal amount of work which has to be done to transform one distribution to another by moving distribution mass between each other.*

4.1 Earth Movers Distance

Using the transportation problem EMD can be defined. Let A = $(a_1, a_2, \ldots, a_n)$, B=$(b_1, b_2, \ldots, b_n)$ be the rows of the dataset and d_{ij} be the ground distance between i^{th} element of A and j^{th} of B. Fina a flow F = $[f_{ij}]$ where f_{ij} represents flow of mass from element i of A to element j of B such that overall work is minimized subject to these constraints.

$$\begin{aligned} WORK(A,B,F) &= \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij} f_{ij} \\ f_{ij} \geq 0 &\quad , \quad 1 \leq i \leq n, 1 \leq j \leq n \\ p_i - \sum_{j=1}^{n} f_{ij} + \sum_{j=1}^{n} f_{ji} &= q_i, 1 \leq i \leq n \\ \sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij} &= \sum_{j=1}^{n} p_i = \sum_{j=1}^{n} q_i = 1 \end{aligned}$$

Example: Consider an example. Let S={ 3,4,5,6,7,8,9,10,11 } be the distribution of sensitve information, A_1 = {3, 4, 5} and A_2 = {6, 8, 11} be two equivalent classes. Now intutively we can see that in first equivalent class all values are at lower end and so reveal more information. Let us now apply t-closeness to it.
Finding EMD[A_1,S] and EMD[A_2,S] after defining our ground distance mod(i - j)/8 (so that max distance is 1) , we get D[A1,Q] = 0.37,and D[A2,Q] = 0.16 [2]. So if our t was 0.2 A1 equivalence class does not have 0.2-closeness and thus we can avoid similarity attacks.

4.2 Limitations of t-closeness

- There is no computational procedure to enforce t-closeness followed in [2].

- There is effective way till now of combining with generalizations and suppressions or slicing.
- Lost co-relation between different attributes : This is because each attribute is generalized separately and so we loose their dependence on each other.
- Utility of data is damaged if we use very small t.(And small t will result in increase in computational time.

5 (k,p,q,r) Anonymity Model

(k, p, q, r)- anonymity [7] is said to be achieved if:

- The data satisfies p-sensitivity for groups where confidential attributes appear very less frequently(less frequent than parameter q).
- For such groups (after p-sensitivity constraint), the ratio of variance within the group of sensitive attributes and variance within entire dataset is at least r.

5.1 Limitations of (k,p,q,r) Anonymity Model

It does not provide any defence against skewness attack which was one of the main reasons for using distance metric like EMD in t-closeness model.

6 Slicing: Another Approach to Privacy Preserving Data Publishing

Slicing [9] is an alternative approach to generalization and suppressions to achieve a anonymized data.In high dimensional data most data points have similar distances with each other, forcing a great amount of generalizations to satisfy k-anonymity. Also there is an assumption of uniform distribution in every generalized interval. This significantly reduces the data utility of the generalized data set.

Definition 11 (Slicing): *Data is vertically partitioned into sensitive attributes and Quasi - Identifiers and then horizontally partition it into group of tuples. Among a horizontal group data is randomly perturbated.*

Department	Age	Course	Marks
ME	20	Mechanics	34
MME	21	Relativity	54
ME	20	Rotational	87
CHE	22	Algorithms	39
CHE	23	Psychology	65
CHM	22	Real Analysis	71
CSE	26	Algorithms	91
CSE	25	Architecture	11

(Department,Age)	(Course, Marks)
ME , 20	Relativity , 54
MME , 21	Algorithms , 39
ME , 20	Mechanics , 34
CHE , 22	Rotational , 87
CHE , 23	Architecture , 11
CHM , 22	Algorithms , 91
CSE , 26	Psychology , 65
CSE , 25	Real Analysis , 71

Example: Slicing first partitions attributes into **columns**. Each column is defined by a subset of attributes. This is vertically partitioning of the table. For example, the sliced table contains 2 columns: the first one contains Department, Age and the second column contains Course,Marks.Slicing also partitions tuples into **buckets**. Each bucket contains a subset of tuples. This horizontally partitions the table. For example, sliced tables contains 2 buckets, each containing 4 tuples. Within each bucket, values in each column are randomly permutated to hide the linking between different columns.
Take for example, in the top bucket the values (Relativity , 54) , (Algorithms , 39), (Mechanics , 34) , (Rotational , 87) are randomly permutated so that the linking between the two columns within one bucket is hidden.

7 Current and Future Work

In this semester we were able to go through most recent publications(three of them in 2009) in this area. We found out the drawbacks in t-closeness paper and the possible approaches and from here on. In the starting of the semester the randomization techniques were our focus but they did not look very promising as they dont gurantee privacy and have many known adversary attacks against them . We narrowed down to t-closeness because it provided a safeguard against most of the attacks possible. Since none of the recent methods also has been able to serve as a substitute for t-closeness and defend against skewness attacks we will focus on finding a computational procedure to enforce t-closeness and compare it's privacy preserving efficiency(data utility and efficient running time) with current methods. Also Slicing is a relatively new concept and we want to find an effective way of using slicing to achieve t-closeness. Most of the earlier approaches only depended on data generalizations and suppressions for achieving the privacy goal. We also plan to analyze and compare limited t-closeness, in which only certain records(with high sensitivity) will be required to fulfill the t-closeness criteria.

References

[1] Models and Algorithms : Privacy-Preserving Data Mining
Springer 2008 : Charu Aggarwal , Philip Yu

[2] t-Closeness: Privacy Beyond k-Anonymity and l -Diversity
ICDE Conference, 2007, Ninghui Li , Tiancheng Li , Suresh Venkatasubramanian

[3] Protecting Respondents Identities in Microdata Release.
IEEE Trans. Knowl. Data Eng. , Ciriani V., De Capitiani di Vimercati S., Foresti S., Samarati P

[4] k-Anonymity. Security in Decentralized Data Management
Springer,2006. :Jajodia , Yu T , Bayardo R. J., Agrawal R

[5] Data Privacy through Optimal k-Anonymization.
Machanavajjhala A., Gehrke J., Kifer D., and Venkitasubramaniam M.

[6] ℓ diversity: Privacy Beyond k-Anonymity. ICDE, 2006.

[7] An Anonymity Model Achievable Via Microaggregation
Josep Domingo-Ferrer, Francesc Sebe, Agusti Solanas (Secure Data Management 2008)

[8] Modeling and Integrating Background Knowledge in Data Anonymization.
Tiancheng Li, Ninghui Li, Jian Zhan

[9] Slicing: A new Approach to Privacy Preserving Data Publishing
Tiancheng Li, Ninghui Li, Jian Zhang, Ian Molloy

[10] Indexing Spatially Sensitive Distance Measures Using Multi-resolution Lower Bounds
Vebjorn Ljosa, Arnab Bhattacharya, Ambuj K. Singh