Preserving Privacy Technique for Knowledge Discovery

Darshan .B. Patel[1], Dheeraj Kumar Singh[2]

[1]M.E (3rd Sem) CSE
Parul Institute of Technology
Gujarat, India
darshan6788@gmail.com

[2]Assistant Professor, I.T Department
Parul Institute of Engineering and Technology
Gujarat-India
dheerajsingh66@gmail.com

Abstract: ***Data mining techniques are becoming more and more important for assisting decision making processes and, more generally, to extract hidden knowledge from massive data collections in the form of patterns, models, and trends that hold in the data collections. During this extraction of hidden knowledge from this massive data collection, privacy of data is a big issue. PPDM (Privacy Preserving Data Mining) approaches protect data by modifying them to mask or erase the original sensitive data that should not be revealed. PPDM approaches based on principle- loss of privacy, measuring the capacity of estimating the original data from the modified data, and loss of information, measuring the loss of accuracy in the data. The main goal of these approaches is therefore to provide a trade-off between privacy and accuracy. In this paper we show that l-diversity has a number of limitations. In particular, it is neither necessary nor sufficient to prevent attribute disclosure. We propose a novel privacy notion called t-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t).***

Keywords: Data mining, PPDM, l-diversity, t-Closeness

1. Introduction

Data mining is the task of discovering interesting and hidden patterns from large amounts of data where the data can be stored in databases, data warehouses, OLAP (on line analytical process) or other repository information [1]. Data mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, neural networks, information retrieval, etc.

Data mining is a part of a process called KDD-knowledge discovery in databases. This process consists basically of steps that are performed before carrying out data mining, such as data selection, data cleaning, pre-processing, and data transformation [1].

Privacy preserving data mining aims at providing a trade-off between sharing information for data mining analysis, on the one side, and protecting information to preserve the privacy of the involved parties on the other side. PPDM approaches based on principle- loss of privacy, measuring the capacity of estimating the original data from the modified data, and loss of information, measuring the loss of accuracy in the data [4].

The main goal of these approaches is therefore to provide a trade-off between privacy and accuracy. The main problem of cryptography-based techniques is, however, that they are usually computationally expensive.

2. PPDM Techniques

Privacy preservation has become a major issue in many data mining applications. When a data set is released to other parties for data mining, some privacy-preserving technique is often required to reduce the possibility of identifying sensitive information about individuals. This is called the disclosure-control problem in statistics and has been studied for many years. Most statistical solutions concern more about maintaining statistical invariant of data. The data mining community has been studying this problem aiming at building strong privacy-preserving models and designing efficient heuristic solution.

A. k-Anonymity model

The k-anonymity model assumes a quasi-identifier [7], which is a set of attributes that may serve as an identifier in the data set. It is assumed that the dataset is a table and that each tuple corresponds to an individual. Let Q be the quasi-identifier. An equivalence class of a table with respect to Q is a collection of all tuples in the table containing identical values for Q. The size of an equivalence class indicates the strength of identification protection of individuals in the equivalent class. If the number of tuples in an equivalence class is greater, it will be more difficult to re-identify individual [11]. A data set D is k-anonymous with respect to Q if the size of every equivalence class with respect to Q is k or more. As a result, it is less likely that any tuple in the released table can be linked to an individual and thus personal privacy is preserved. K-Anonymity model can be implemented through two algorithms Local Recoding and Global Recoding [7].

Publishing data about individuals without revealing

sensitive information about them is an important problem. In recent years, a new definition of privacy called k-anonymity has gained popularity. In a k-anonymized dataset, each record is indistinguishable from at least k−1 other records with respect to certain "identifying" attributes.

Limitations

There are two simple attacks that a k-anonymized dataset has some subtle, but severe privacy problems. First, an attacker can discover the values of sensitive attributes when there is little diversity in those sensitive attributes. Second, attackers often have background knowledge, and we show that k-anonymity does not guarantee privacy against attackers using background knowledge. Therefore powerful privacy definition called ℓ-diversity was proposed [8].

B. l – Diversity

k-anonymity is susceptible to homogeneity and background knowledge attacks; thus a stronger definition of privacy is needed. In the remainder of the paper, we derive our solution. We start by introducing an ideal notion of privacy called Bayes-optimal for the case that both data publisher and the adversary have full (and identical) background knowledge [8]. Unfortunately in practice, the data publisher is unlikely to possess all this information, and in addition, the adversary may have more specific background knowledge than the data publisher. Hence, while Bayes-optimal privacy sounds great in theory, it is unlikely that it can be guaranteed in practice. To address this problem, we show that the notion of Bayes-optimal privacy naturally leads to a novel practical definition that we call ℓ-diversity. ℓ-Diversity provides privacy even when the data publisher does not know what kind of knowledge is possessed by the adversary [8]. The main idea behind ℓ-diversity is the requirement that the values of the sensitive attributes are well-represented in each group.

Two main approaches have been proposed for protecting the privacy of sensitive cells: data swapping and data suppression. The data swapping approach involves moving data entries from one cell to another in the contingency table in a manner that is consistent with the set of published marginal's. In the data suppression approach, cells with low counts are simply deleted, which in turn might lead to the deletion of additional cells. An alternate approach is to determine a safety range or protection interval for each cell, and publish only those marginal's which ensure that the feasibility intervals (i.e. upper and lower bounds on the values a cell may take) contain the protection intervals for all the cell entries. The above techniques, however, do not provide a strong theoretical guarantee of the privacy ensured.

The k-anonymity privacy requirement for publishing microdata requires that each equivalence class (i.e., a set of records that are indistinguishable from each other with respect to certain "identifying" attributes) contains at least k records. Recently, several authors have recognized that k-anonymity cannot prevent attribute disclosure [11]. The notion of l-diversity has been proposed to address this; l-diversity requires that each equivalence class has at least well-represented values for each sensitive attribute.

C. Reason for the Change

The protection k-anonymity provides is simple and easy to understand. If a table satisfies k-anonymity for some value k, then anyone who knows only the quasi-identifier values of one individual cannot identify the record corresponding to that individual with confidence greater than 1/k [9] [3]. While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure [8]. Two attacks were identified in k-anonymity: the homogeneity attack and the background knowledge attack [9].

Limitations

l-diversity may be difficult and unnecessary to achieve. Suppose that the original data has only one sensitive attribute: the test result for a particular virus. It takes two values: positive and negative. Further suppose that there are 10000 records, with 99% of them being negative, and only 1% being positive. Then the two values have very different degrees of sensitivity. One would not mind being known to be tested negative, because then one is the same as 99% of the population, but one would not want to be known/considered to be tested positive. In this case, 2-diversity is unnecessary for an equivalence class that contains only records that are negative. In order to have a distinct 2-diverse table, there can be at most 10000×1% = 100 equivalence classes and the information loss would be large. Also observe that because the entropy of the sensitive attribute in the overall table is very small, if one uses entropy _-diversity, _ must be set to a small value.[9] l-diversity is insufficient to prevent attribute disclosure. Below we present two attacks on l-diversity.

Skewness Attack: When the overall distribution is skewed, satisfying _-diversity does not prevent attribute disclosure. Suppose that one equivalence class has an equal number of positive records and negative records. It satisfies distinct 2-diversity, entropy 2-diversity, and any recursive (c, 2)-diversity requirement that can be imposed. However, this presents a serious privacy risk, because anyone in the class would be considered to have 50% possibility of being positive, as compared with the 1% of the overall population. Now consider an equivalence class that has 49 positive records and only 1 negative record. It would be distinct 2- diverse and has higher entropy than the overall table (and thus satisfies any Entropy l-diversity that one can impose), even though anyone in the equivalence class would be considered 98% positive, rather than 1% percent. In fact, this equivalence class has exactly the same diversity as a class that has 1 positive and 49 negative record, even though the two classes present very different levels of privacy risks [9].

Similarity Attack: When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information. Consider the following example.

Example: Table 1 is the original table, and Table 2 shows an anonymized version satisfying distinct and entropy 2-diversity. There are two sensitive attributes: Salary and Disease. Suppose one knows that Bob's record corresponds to one of the first three records, then one knows that Bob's salary is in the range [3K–5K] and can infer that Bob's salary is relatively low. This attack applies not only to numeric attributes like "Salary", but also to categorical attributes like "Disease".

This leakage of sensitive information occurs because while l-diversity requirement ensures "diversity" of sensitive values in each group, it does not take into account the semantical closeness of these values.

Sr.no	ZIPCODE	AGE	Salary	DISEASE
1	77624	33	2k	Flu
2	77644	36	4k	Flu
3	77654	32	6k	Flu
4	76217	23	11k	Blood Cancer
5	76218	25	8k	Heart Disease
6	76210	27	4k	Malaria
7	77434	41	5k	Heart Disease
8	77412	45	3k	Blood Cancer
9	77456	48	2k	Blood Cancer

Table 1: Original Salary/Disease Table

Sr.no	ZIPCODE	AGE	Salary	DISEASE
1	776**	3*	2k	Flu
2	776**	3*	4k	Flu
3	776**	3*	6k	Flu
4	7621*	≥20	11k	Blood Cancer
5	7621*	≥20	8k	Heart Disease
6	7621*	≥20	4k	Malaria
7	774**	4*	5k	Heart Disease
8	774**	4*	3k	Blood Cancer
9	774**	4*	2k	Blood Cancer

Table 2: A 3-diverse version of Table-1

C. t-closeness

Intuitively, privacy is measured by the information gain of an observer. Before seeing the released table, the observer has some prior belief about the sensitive attribute value of an individual. After seeing the released table, the observer has a posterior belief. Information gain can be represented as the difference between the posterior belief and the prior belief. The novelty of our approach is that we separate the information gain into two parts: that about the whole population in the released data and that about specific individuals.

To motivate our approach, let us perform the following thought experiment: First an observer has some prior belief B0 about an individual's sensitive attribute. Then, in a hypothetical step, the observer is given a completely generalized version of the data table where all attributes in a quasi-identifier are removed (or, equivalently, generalized to the most general values) [9]. The observer's belief is influenced by Q, the distribution of the sensitive attribute value in the whole table, and changes to B1 [9]. Finally, the observer is given the released table. By knowing the quasi-identifier values of the individual, the observer is able to identify the equivalence class that the individual's record is in, and learns the distribution P of sensitive attribute values in this class. The observer's belief changes to B2.

The l-diversity requirement is motivated by limiting the difference between B0 and B2 (although it does so only indirectly, by requiring that P has a level of diversity). We choose to limit the difference between B1 and B2. In other words, we assume that Q, the distribution of the sensitive attribute in the overall population in the table, is public information. We do not limit the observer's information gain about the population as a whole, but limit the extent to which the observer can learn additional information

about specific individuals.

To justify our assumption that Q should be treated as public information, we observe that with generalizations, the most one can do is to generalize all quasi-identifier attributes to the most general value. Thus as long as a version of the data is to be released, a distribution Q will be released.1 We also argue that if one wants to release the table at all, one intends to release the distribution Q and this distribution is what makes data in this table useful. In other words, one wants Q to be public information. A large change from B0 to B1 means that the data table contains a lot of new information, e.g., the new data table corrects some widely held belief that was wrong. In some sense, the larger the difference between B0 and B1 is, the more valuable the data is. Since the knowledge gain between B0 and B1 is about the whole population, we do not limit this gain.

We limit the gain from B1 to B2 by limiting the distance between P and Q. intuitively, if P = Q, then B1 and B2 should be the same. If P and Q are close, then B1 and B2 should be close as well, even if B0 may be very different from both B1 and B2.

The t-closeness Principle: An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness [9].

Of course, requiring that P and Q to be close would also limit the amount of useful information that is released, as it limits information about the correlation between quasi identifier attributes and sensitive attributes. However, this is precisely what one needs to limit. If an observer gets too clear a picture of this correlation, then attribute disclosure occurs. The t parameter in t-closeness enables one to trade off between utility and privacy [9].

Now the problem is to measure the distance between two probabilistic distributions. There are a number of ways to define the distance between them. Given two distributions P = (p1, p2... pm), Q = (q1, q2... qm), two well-known distance measures are as follows. The variation distance is defined as:

$$D[P,Q] = \sum_{i=1}^{m} \frac{1}{2}|Pi - Qi|$$

And the Kullback-Leibler (KL) distance [8] is defined as:

$$D[P,Q] = \sum_{i=1}^{m} \frac{1}{2} Pilog\frac{Pi}{qi} = H(P) - H(P,Q)$$

where H(P) = $\sum_{i=1}^{m} \frac{1}{2} Pilog\frac{Pi}{qi}$ is the entropy of **P** and H(**P**,**Q**) =$\sum_{i=1}^{m} \frac{1}{2} PilogQi$ is the cross-entropy of **P** and **Q** **[9]**.

These distance measures do not reflect the semantic distance among values. Recall (Tables 1 and 2), where the overall distribution of the Income attribute is Q = {3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k}.2 The first equivalence class in Table 4 has distribution P1 = {3k, 4k, 5k} and the second equivalence class has distribution P2 = {6k, 8k, 11k}. Our intuition is that P1 results in more information leakage than P2, because the values in P1 are all in the lower end; thus we would like to have D [P1,Q] >D[P2,Q]. The distance measures mentioned above would not be able to do so, because from their point of view values such as 3k and 6k are just different points and have no other semantic meaning [9].

In short, we have a metric space for the attribute values so that a ground distance is defined between any pair of values. We then have two probability distributions over these values and we want the distance between the two probability distributions to be dependent upon the ground distances among these values.

The EMD is based on the minimal amount of work needed to transform one distribution to another by moving distribution mass between each other. Intuitively, one distribution is seen as a mass of earth spread in the space and the other as a collection of holes in the same space. EMD measures the least amount of work needed to fill the holes with earth. A unit of work corresponds to moving a unit of earth by a unit of ground distance.

Limitations

t-closeness protects against attribute disclosure, but does not deal with identity disclosure. Thus, it may be desirable to use both t-closeness and k-anonymity at the same time. Further, it should be noted that t-closeness deals with the homogeneity and background knowledge attacks on k-anonymity not by guaranteeing that they can never occur, but by guaranteeing that if such attacks can occur, then similar attacks can occur even with a fully-generalized table. As we argued earlier, this is the best one can achieve if one is to release the data at all. [9]

3. Conclusion and Future Work

While k-anonymity is favourable against identity disclosure, it does not stand as an alternative against attribute disclosure. l-diversity, an enhanced technique for privacy preserving in data mining remedies this situation by requiring that each equivalence class has at least l well-represented values for each sensitive attribute. l-diversity has a number of limitation so that has led to a proposition of a new privacy preserving mode for data mining called t-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be less than threshold t). One key feature of our approach is that we separate the information gain an observer can get from a released data table into two parts: that about all population in the released data and that about specific individuals. This enables us to limit only the second kind of information gain. Some of the open research issues for

privacy preserving are:

Multiple Sensitive Attributes Multiple sensitive attributes present additional challenges. Suppose we have two sensitive attributes X and Y. One can consider the two attributes separately, i.e., an equivalence class E has t-closeness if E has t-closeness with respect to both X and Y. Another approach is to consider the joint distribution of the two attributes. To use this approach, one has to choose the ground distance between pairs of sensitive attribute values. A simple formula for calculating EMD may be difficult to derive, and the relationship between t and the level of privacy becomes more complicated.

Other Anonymization Techniques t-closeness allows us to take advantage of anonymization techniques other than generalization of quasi-identifier and suppression of records. For example, instead of suppressing a whole record, one can hide some sensitive attributes of the record; one advantage is that the number of records in the anonymized table is accurate, which may be useful in some applications. Because this technique does not affect quasi identifiers, it does not help achieve k-anonymity and hence has not been considered before. Removing a value only decreases diversity; therefore, it does not help to achieve l-diversity. However, in t-closeness, removing an outlier may smooth a distribution and bring it closer to the overall distribution. Another possible technique is to generalize a sensitive attribute value, rather than hiding it completely. An interesting question is how to effectively combine these techniques with generalization and suppression to achieve better data quality.

Limitations of using EMD in t-closeness the t-closeness principle can be applied using other distance measures. While EMD is the best measure we have found so far, it is certainly not perfect. In particular, the relationship between the value t and information gain is unclear. For example, the EMD between the two distributions (0.01, 0.99) and (0.11, 0.89) is 0.1, and the EMD between (0.4, 0.6) and (0.5, 0.5) is also 0.1. However, one may argue that the change between the first pair is much more significant than that between the second pair. In the first pair, the probability of taking the first value increases from 0.01 to 0.11, a 1000% increase. While in the second pair, the probability increase is only 25%. In general, what we need is a measure that combines the distance-estimation properties of the EMD with the probability scaling nature of the KL distance.

References

[1] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni. Association rule hiding. IEEE Transactions on Knowledge and Data Engineering, 16(4):434–447, 2004

[2] T. M. Truta and B. Vinay, Privacy protection: p-sensitive k-anonymity property. In Proceedings of the 22nd International Conference on Data Engineering Workshops, the Second International Workshop on Privacy Data Management (PDM'06), page 94, 2006

[3] K. LeFevre, D. DeWitt, and R. Ramakrishnan, Incognito: Efficient full-domain k-anonymity. In Proc. ACM SIGMOD International Conference on Management of Data (SIGMOD'05), pages 49–60, 2005

[4] I. Dinur and K. Nissim. Revealing information while preserving privacy, In PODS, pages 202–210, 2003

[5] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. Toward privacy in public databases, In TCC, 2005

[6] X. Xiao and Y. Tao. Personalized privacy preservation, In Proceedings of ACM zonference on Management of Data (SIGMOD'06), pages 229–240, June 2006

[7] R. Bayardo and R. Agrawal, Data privacy through optimal k-anonymization, In Proc. of the 21st Int'l Conf. on Data Engineering, 2005

[8] A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam, l-Diversity: Privacy Beyond k-Anonymity, In Proc. 22nd ICDE, 2006

[9] N. Li, T. Li and S. Venkatasubramanian, t-closeness:Privacy beyond k-anonymity and l-diversity, IEEE 23rd International Conference on Data Engineering,2007

[10] S.R.M. Oliveira and O. R. Zaiane, Privacy preserving clustering by data transformation, 18th Brazilian Symposium on Databases (SBBD 2003), 2003

[11] S. Xiaoxun, H. Wang, J. Li and T. M. Truta, Enhanced P-Sensitive K-Anonymity Models for Privacy Preserving Data Publishing. In: Transaction on Data Privacy, 2008

Author Profile

Darshan Patel is currently doing research work in data mining as an M-tech student in computer science & Engineering from Gujarat Technological University. He pursued his Bachelors in computer engineering from VNSGU. His research interest includes Steganography, Linux Clustering and data mining.