International Journal of Science Engineering and Advance Technology, IJSEAT, Vol 2, Issue 11, November - 2014 | ISSN 2321-6905

International Journal of Science Engineering and Advance Technology

Privacy and Classification Of Analyzed Data Using EMD

[1]T.Sravya, [2]G.Appa Rao
Dept of **CSE**, S.R.K.Institute Of Technology, Enikepadu, Krishna dist, AP, India
email:thota.sravya@gmail.com;appu.prem05@gmail.com

Abstract:

In recent years many researchers issued on data publishing with recommended settings .But privacy is a key issue here. Existing techniques such as K-anonymity and L-diversity should not provide effective and sufficient results for privacy preserving in data publishing. So in this paper we propose tree base algorithm for providing security, In this technique we arrange the data in tree based format for closeness of a data publishing and for retrieving data in sequential order. Our techniques also improved more security to micro data publishing and retrieving relevant information from micro data using attribute disclosure.

Keywords - Privacy, data publishing, K-anonymity, L-diversity

I Introduction:

When releasing micro data, it is necessary to prevent the sensitive information of the individuals from being disclosed. Two types of information disclosure have been identified: identity disclosure and attribute disclosure. Identity disclosure occurs when an individual is linked to a particular record in the released table. Government agencies and other organizations often need to publish micro data, e.g., medical data or census data, for research and other purposes.

It has been recognized that even disclosure of false attribute information may cause harm. To effectively limit disclosure, we need to measure the disclosure risk of an anonymized table. While k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. A new notion of privacy, called l- diversity, which requires that the distribution of a sensitive attribute in each equivalence class has at least '"well represented" in general, is that they effectively assume all attributes to be categorical; the adversary either does or does not learn something sensitive. For privacy in Micro data publishing a base model called t-closeness and a more flexible privacy model called (n, t)-closeness were proposed. These closeness measures require Probability distributions that are assessed using Earth Mover's Distance (EMD) measurement. Side effects of using EMD include large number of unknown variables that are to be resolved and have high time complexities. We Examine the formulation of Tree-EMD is much simpler than the original EMD formulation. Tree EMD exploits the fact that a basic feasible solution of the simplex algorithm-based solver forms a spanning tree. It has only $O(N)$ unknown variables, which is significantly less than the $O(N2)$ variables required in the original EMD. The efficiency of this algorithm enables its application to handle problems that were previously prohibitive due to high time complexities. In order to reduce the computation times of the original distance, the proposed method uses the lower bounding distance.

In this paper, we propose a new fast algorithm, *i.e.*, Tree-EMD Algorithm to compute EMD between histograms with $L1$ ground distance.

II Related Work:

A number of information disclosure limitation techniques have been designed for data publishing, including Sampling, Cell Suppression, Rounding, and Data Swapping and Perturbation. The first category of work aims at devising privacy requirements. The k-anonymity model assumes that the adversary has access to some publicly available databases and the adversary knows who is and who is not in the table.

L-Diversity assumes an adversity who has knowledge of the form "Carl does not have heart disease," while our closeness measures consider an adversary who knows the distributional information of the sensitive attributes. Our goal is to propose an alternative technique for data publishing that remedies the limitations of l-diversity in some applications. Privacy-preserving data publishing has been extensively studied in several other aspects.

First, background knowledge presents additional challenges in defining privacy requirements. Several recent studies have aimed at modeling and integrating background knowledge in

data Anonymization. Second, several works considered continual data publishing, i.e., republication of the data after it has been updated. Nergiz et al. proposed σ-presence to prevent membership disclosure, which is different from identity/attribute disclosure. Wong et al. showed that knowledge of the anonymization algorithm for data publishing can leak extra sensitive information. Our goal is to propose an alternative technique for data publishing that remedies the limitations of ldiversity in some applications. Privacy-preserving data publishing has been extensively studied in several other aspects.

III Back Ground:

The first category of work aims at devising privacy requirements. The k-anonymity model assumes that the adversary has access to some publicly-available databases and the adversary Knows who is and who is not in the table. A few subsequent works recognize that the adversary has also knowledge of the distribution of the sensitive

Definition K-Anonymity: Let T (A1….. Am) be a table, and QI be quasi-identifier associated with it. T is said to satisfy k-anonymity with respect to QI iff each sequence of values in T[QI] appears at least with k occurrences in T[QI] (T[QI] denotes the projection, maintaining duplicate tuples, of attributes QI in T.

The L-Diversity Principle: *An* Equivalence class is said to have L-diversity if there are at least "well-represented" values for the sensitive attribute. A table is said to have L- diversity if every equivalence class of the table has L-diversity. Machanavajjhala et al. gave a number of interpretations of the term "well represented" in this principle:

Distinct L-Diversity: The simplest understanding of "well represented" would be to ensure that there are at least 'distinct values for the sensitive attribute in each equivalence class. Distinct L-diversity does not prevent probabilistic inference attacks. An equivalence class may have one value appear much more frequently than other values, enabling an adversary to conclude that an entity in the equivalence class is very likely to have that value. This motivated the development of the following stronger notions of L-diversity.

Probabilistic L-Diversity: *An* anonymized table satisfies probabilistic L-diversity if the frequency of a sensitive value in each group is at most 1/l. This guarantees that an observer cannot infer the sensitive value of an individual with probability greater than 1/l.

Entropy L-Diversity: The entropy of an equivalence class E is defined to being which S is the domain of the sensitive attribute and p(E,s) is the fraction of records in E that have sensitive values.

A table is said to have entropy L-diversity if for every equivalence class E, Entropy (E) ≥ log L. Entropy L-diversity is stronger than distinct Ldiversity. As pointed out in [23], in order to have entropy 'L-diversity for each equivalence class, the entropy of the entire table must be at least log (l). Sometimes, this may too restrictive, as the entropy of the entire table may be low if a few values are very common. This leads to the following less conservative notion of L-diversity. L-diversity and Limitations. L-diversity requires that each equivalence class contains at least l "well represented" values for the sensitive attribute. This is in contrast to the above definition of utility where the homogeneous distribution of the sensitive attribute preserves the most amount of data utility. Emphasize that L-diversity is still a useful measure for data publishing. L-diversity and our closeness measures make different assumptions about the adversary.

While the 'L-diversity principle represents an important step beyond k-anonymity in protecting against attribute disclosure, it has several shortcomings that we now discuss. 'L-diversity may be difficult to achieve and may not provide sufficient privacy protection. The goal is to propose an alternative technique for data publishing that remedies the limitations of L-diversity in some application.

t-closeness. We show that t-closeness substantially limits the amount of useful information that the released table preserves. t-closeness requires. (n,t)-closeness. The (n, t) closeness model allows better data utility than t-closeness. Given an anonymized table {E1,……,Ep} where each Ei($1 \le I \le P$) is an equivalence class and another anonymized table {G1,. . .,Gd} where each Gj ($1 \le j \le d$) is the union of a set of equivalence classes in {E1, . . . ,Ep} and contains at least n records. The anonymized table {E1, . . .,Ep} satisfies the (n, t)-closeness requirement if the distribution of the sensitive attribute in each Ei ($1 \le i \le p$) is close to that in Gj containing Ei.

We are thus able to separate the utility of the anonymized table into two parts: 1) the first part U{G1; . . .;Gd} is the sensitive information about the large groups {G1; . . .;Gd}and 2) the second part is

further sensitive information about smaller groups. By requiring the distribution of the sensitive attribute in each Ei to be close to that in the corresponding Gj containing Ei, the (n,t)-closeness principle only limits the second part of the utility function and does not limit the first part.

IV Earth Mover's Distance:

EMD for categorical attributes: a total order often does not exist. Two distance measures are considered. Equal Distance: The ground distance between any two values of a categorical attribute is defined to be 1. It is easy to verify that this is a metric. As the distance between any two values is 1, for each point that $pi - qi > 0$, one just needs to move the extra to some other points. Hierarchical Distance: The distance between two values of a categorical attribute is based on the minimum level to which these two values are generalized to the same value according to the domain hierarchy.

V Tree- Earth Mover's Distance:

We introduce Tree-EMD with $L1$, a novel efficient formulation of EMD. We first show that, by using the $L1$ (Manhattan) distance as the ground distance. We designed a tree-based algorithm as an efficient discrete optimization solver, which extends the original simplex algorithm.

VI Performance:

The Tree-EMD algorithm is presented in several issues:

(1) *The root of a BFT*: The root *r* is heuristically set to be the center of the graph. This is to make the tree as balanced as possible. Once *r* is fixed, the *u* value at *r* is fixed to 0.

2) *Build the initial BFT*: The nodes are considered sequentially, in a left-to-right and bottom-to-top order, i.e., starting from bottom-left node. When processing node *q*, all the flows connecting its lower and left neighbors are fixed. As a result, only one BV flow needs to be chosen between *q* and either its upper or right neighbor such that the flow makes the weight at *q* vanish.

The Tree-EMD algorithm can also be generalized to solve the original EMD problem (i.e. beyond histograms) for speedup. This is because the tree structure used in Tree-EMD is also true for the transportation simplex used in the original EMD. In addition, as indicated EMD can also be modeled as a network flow problem. This raises interest in the underlying relationship between the tree-based algorithm and network flow algorithms. It may be a key to find more efficient solutions the original EMD.

VII Conclusion:

Another privacy measure l- diversity attempts to solve this problem. But it is neither necessary nor sufficient to prevent attribute disclosures and fails at data utilization. So a base model called t-closeness and a more flexible privacy model called (n, t)-closeness were developed that achieves a better balance between privacy and utility.). (n, t)- Closeness offers higher utility. Existing privacy measures such as k- anonymity protects against identity disclosures, but it does not provide sufficient protection against attribute disclosures. We propose to use an efficient tree-based algorithm, Tree-EMD. Tree-EMD exploits the fact that a basic feasible solution of the simplex algorithm-based solver forms a spanning tree. The formulation of Tree-EMD is much simpler than the original EMD formulation.

VIII References:

[1[1] Tiancheng Li, Ninghui Li, Senior Member, IEEE, Jia Zhang, Member, IEEE, and Ian Molloy "Slicing: A New Approach for Privacy Preserving Data Publishing" Proc. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 3 MARCH 2012.

[2] J.-W. Byun, Y. Sohn, E. Bertino, and N. Li, "Secure Anonymization for Incremental Datasets," Proc. VLDB Workshop Secure Data Management (SDM), pp. 48-63, 2006.

[3] B.-C. Chen, K. LeFevre, and R. Ramakrishnan, "Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 770- 781, 2007.

[4]G. T. Duncan, S. E. Fienberg, R. Krishnan, R. Padman, and S.

[5] G. T. Duncan and D. Lambert. Disclosure limited data dissemination. *J. Am. Stat. Assoc.*, pages 10–28, 1986

ISSN 2321-6905

Miss. Thota Sravya is a student of S.R.K Institute Of Technology, Enikepadu .Presently she is pursuing her M.Tech [Computer Science Engineering] from this college and she received his M.C.A from Nova College of Engineering,& Tech, Jupudi, Ibraheempatnam affiliated to JNT University, Kakinada in the year 2012. Her area of interest includes Information Security, Computer Networks and Data Warehouse & Data Mining, all current trends and techniques in Computer Science.

Mr.G.Appa Rao, well known Author and excellent teacher Received B.Tech from Sarada Institute of Tech&Management, Srikakulam, Affliated by JNTUH and M.E (NIE) from Karunya University, Coimbatore is working as Engineering Technician Gr.2 for 2 years. He has 3 years of teaching experience in St.Theressa Institute of Engineering &Technology, Garividi, Vizayanagaram Dist. He is working presently as Asst.Professor in Dept. of Computer Science &Engineering. In S.R.K Institute Of Technology, Vijayawada, Krishna Dist. His area of Interest includes Data Warehouse and Data Mining, Networking information security, flavors of Unix Operating systems and other advances in computer Applications.