Postprint of Feasible Differential Privacy Protection Methods for Mixed-Attribute Data Tables

Authors: Ding Yongshan, Li Lixin

Date: 2018-05-20T00:00:00+00:00

Abstract

To enhance privacy protection and improve data utility, we propose a data protection method that can enforce differential privacy on mixed-attribute data tables. The method first employs the ICMD (insensitive clustering for mixed data) clustering algorithm to perform clustering-based anonymization on the dataset, and then applies #1;-differential privacy protection on this basis. The ICMD clustering algorithm employs different methods to calculate distances and centroids for categorical and numerical attributes in the data table, and introduces a total order function to satisfy the requirements for implementing differential privacy. Through clustering, it achieves the differentiation of query sensitivity from single records to group data, thereby reducing the risk of information loss and information disclosure. Finally, experimental results demonstrate the effectiveness of the proposed method.

Full Text

Preamble

Title: A Differential Privacy Protection Method for Mixed-Attribute Data Tables

Authors: Ding Yongshan, Li Lixin (The Third College, Information Engineering University, Zhengzhou 450000, China)

Abstract: To strengthen privacy protection and improve data utility, this paper proposes a differential privacy data protection method for mixed-attribute data tables. The method first employs the ICMD (Insensitive Clustering for Mixed Data) algorithm to perform cluster-based anonymization on the dataset, then applies -differential privacy protection on this foundation. The ICMD clustering algorithm uses different methods to calculate distances and centroids for categorical and numerical attributes in the data table, and introduces a total order function to satisfy the requirements for implementing differential privacy.

Through clustering, the query sensitivity is differentiated from single records to group records, reducing both information loss and the risk of information disclosure. Experimental results demonstrate the effectiveness of the proposed method.

Keywords: mixed data; clustering; differential privacy; sensitivity; privacy protection

0 Introduction

With the development of the Internet, the number of users on social networks has increased dramatically, and the user data and information contained therein possess tremendous value. Data miners can extract significant value from this data, but user privacy also faces the risk of leakage [1]. Ensuring network user privacy security requires effective protection mechanisms. The challenge lies in maximizing the protection of published user information from attackers while guaranteeing data usability. Privacy-Preserving Data Publishing (PPDP) should balance two considerations: (a) sufficient protection of sensitive information to eliminate user concerns about data sharing; and (b) minimizing information loss of non-sensitive information to ensure maximum data utility [2].

k-anonymity and its extensions are important methods for user information privacy protection. k-anonymity requires at least k indistinguishable records in quasi-identifiers, preventing attackers from identifying privacy information owners with high probability [3]. However, attackers can acquire increasingly more background knowledge, generating many attack variants such as background knowledge attacks and homogeneity attacks, making these models difficult to defend against. Attackers can often identify specific users by combining quasi-identifiers, thereby obtaining other private information. Therefore, privacy protection methods against arbitrary background knowledge have become a research hotspot [4]. Dwork [5] proposed the differential privacy protection model with rigorous privacy proofs, overcoming the limitation of being unable to resist arbitrary background knowledge attacks. However, the differential privacy model sacrifices considerable data utility when protecting data privacy.

To address these issues, this paper combines the characteristics of clustering and differential privacy to propose a privacy protection method for mixed-attribute data tables during publication. The method assumes that attackers possess complete background knowledge, overcoming the disadvantage of privacy protection models becoming ineffective as background knowledge expands. This method not only satisfies the privacy requirements of the differential privacy protection model but also ensures the quality of published data. The main contributions of this paper are: (a) improving MDAV to propose a clustering algorithm CMD for mixed-attribute data tables; (b) introducing a total order distance function into CMD to propose the insensitive clustering algorithm ICMD for better differential privacy implementation; (c) performing differential privacy operations

on the basis of clustering to propose the ICMD-DP differential privacy data publishing method; and (d) verifying through experiments the effectiveness of the proposed method in protecting data privacy and improving data utility.

1.1 Related Work

Differential privacy protection models achieve privacy protection by adding noise to original datasets or statistical results. The model provides rigorous and robust privacy protection proofs, ensuring that changing a single record in the dataset does not affect statistical results, thereby preserving the original statistical characteristics of the dataset. Additionally, the model can resist arbitrary background knowledge attacks [4].

Definition 1 Assume datasets D and D' differ by at most one record, and S is the value range of algorithm A. If algorithm A satisfies for any output result S S:

```
Pr[A(D)   S]   e^  × Pr[A(D')   S]
```

then algorithm A satisfies -differential privacy, where parameter is called the privacy budget. Smaller values provide higher privacy protection but introduce greater noise.

Differential privacy protection technology exhibits sequential composition and parallel composition properties [15]. Differential privacy can be achieved by adding Laplace noise to query results.

Definition 2 For query function f, if algorithm A satisfies:

```
A(D) = f(D) + Lap(Δf/ )
```

then algorithm A satisfies -differential privacy. Here, Δf represents the sensitivity of the query function, defined as the maximum difference in query results when applied to neighboring datasets. Literature [16] provides the error introduced by adding Laplace noise as: error = 2(Δf/)^2.

With the rise of big data and data mining, research on data privacy protection has gained increasing attention. Privacy protection techniques are roughly divided into noise perturbation, anonymous publication, and data encryption. Among them, anonymization technology plays a positive role in user data security. For example, Wang et al. [6] proposed using anonymization technology to protect user authentication information security. The anonymization involved in this paper is achieved through clustering algorithms.

Clustering-based anonymization plays an important role in data mining and data grouping, and has become a research hotspot in privacy protection. The purpose of clustering anonymization is to partition different objects into groups, maximizing intra-group similarity and minimizing inter-group similarity. k-anonymity algorithms are grouping algorithms ensuring at least k records per group. The MDAV (Maximum Distance to Average Vector) algorithm enables

clustering anonymization for numerical attribute data tables [7]. Based on k-anonymity, literature [8] proposed a k-degree anonymity privacy protection model that maintains network structure stability; literature [9] combined greedy methods with clustering partitioning to propose a greedy clustering anonymization method for optimizing information loss and time efficiency. However, as general methods, these struggle to cope with attack variants from growing background knowledge.

To better resist background knowledge attacks and homogeneity attacks, protecting specific sensitive values or all sensitive values, literature [10] proposed single-sensitive-value -anonymity and multi-sensitive-value -anonymity models; literature [11] proposed a new (p+,)-sensitive k-anonymity privacy protection model. However, they still cannot resist arbitrary background knowledge attacks. Dwork' s [5] differential privacy protection model addresses these issues to some extent; literature [12] introduced the theoretical foundation and latest research progress in differential privacy protection, providing an application framework PINQ (Privacy Integrated Queries).

Torra [13] proposed a k-modes-based microaggregation algorithm for categorical data, but it only handles single categorical or numerical attributes. The k-prototypes algorithm integrates k-means and k-modes to achieve clustering analysis for mixed data [14], but algorithm parameters are difficult to determine and cannot objectively reflect differences between data objects and cluster centers. Literature [13] proposed a differential privacy-preserving k-means clustering method that applies differential privacy to selected center points and within-cluster point sums, but the clustering utility depends not only on the privacy budget but also on dataset size; literature [2] proposed a differential privacy data protection method based on the DBSCAN clustering algorithm, but this method only works on numerical attribute datasets.

This paper describes distance and centroid calculation methods for mixed-attribute data tables, proposes a k-anonymity-satisfying insensitive clustering algorithm ICMD (k-anonymity by insensitive clustering for mixed data) by improving MDAV, and performs differential privacy protection on this basis.

1.3 Distance and Centroid Calculation in Mixed-Attribute Data Tables

Most existing data tables are mixed-type, containing both numerical and categorical attributes. Different distance calculation and centroid solving methods are required for different attribute types. Using a single method often causes information loss, centroid deviation, and other issues. Therefore, this paper proposes a distance calculation and centroid solving method for mixed-attribute data tables.

Let mixed dataset D contain records X and Y, where each record has p-dimensional categorical attributes and q-dimensional numerical attributes. To calculate the distance between data records X and Y, we first calculate

their categorical attribute distance d_c(X,Y) and numerical attribute distance d_n(X,Y) separately.

Definition 3 Categorical Distance. For any records X and Y in a data table with p-dimensional categorical attributes, the categorical attribute distance is defined as:

d_c(X,Y) = Σ_{j=1}^p (x_j, y_j)

where (x_j, y_j) = 0 if x_j = y_j, and 1 otherwise. From this formula, each dimensional categorical attribute value ranges in [0,1].

For numerical attributes, if Hamming distance is used as the distance for each dimension, the categorical attribute distance would be overwhelmed by the numerical attribute distance. Therefore, we adopt the following definition for numerical attribute distance.

Definition 4 Numerical Distance [17]. First, standardize each dimension of the numerical attribute part of data records. For dimension q, the value is normalized as (x_q - min_q)/(max_q - min_q), where max_q is the maximum value and min_q is the minimum value for that dimension. The numerical distance is then:

d_n(X,Y) = Σ_{j=1}^q |x_j - y_j|

Definition 5 Mixed Distance. Based on Definitions 3 and 4, the distance between categorical and numerical attributes can be obtained by adding their distances [18]:

D(X,Y) = d_c(X,Y) + d_n(X,Y)

Definition 6 Centroid. Let T be an equivalence class of n-dimensional dataset D, with record t_i T. Let t_i^o be the numerical attribute part and t_i^c be the categorical attribute part of record t_i, i.e., t_i = {t_i^o, t_i^c}. Let _j be the mean of numerical attribute j, and G_j be the generalization of numerical attribute j. The centroid of equivalence class T is:

C(T) = { _1, _2, ..., _q, G_1, G_2, ..., G_p}

This paper uses mean and generalization to replace original numerical and categorical data in equivalence classes, avoiding one-sidedness and errors from single methods when clustering numerical and categorical data, thereby preserving more semantics.

2 Data Publishing Method

For mixed-attribute data tables, we first describe the distance and centroid calculation methods, propose a clustering method satisfying k-anonymity, then add noise to the clustered data to achieve differential privacy protection. Clustering operations reduce the sensitivity of query functions, thereby achieving the same privacy protection effect with smaller noise addition and improving data utility.

2.1 Clustering Method for Mixed-Attribute Data Tables

Based on MDAV [7] and using the mixed-attribute distance and centroid calculation methods from Section 1.3, this paper proposes a feasible clustering method CMD (Clustering for Mixed Data) for mixed-attribute data tables. According to the definition of k-anonymity, this method simultaneously satisfies the k-anonymity mechanism.

Algorithm 1 Clustering Algorithm CMD(D,k) - Input: Original dataset D with n records, minimum cluster size k - Output: Clustered dataset D' satisfying k-anonymity

1. Calculate the cluster center using the method from literature [19], and compute the farthest record r from this center and the farthest record s from r using Definition 5, as two initial cluster centers.
2. Calculate the k nearest records to r and s respectively, and assign them to clusters, adding them to dataset D' .
3. For the remaining m records, if m 2k, repeat step 1 for the remaining data records; if k m < 2k, form a separate class and add to dataset D' .
4. Otherwise, assign the remaining m records to their nearest clusters.
5. Calculate the centroid for each cluster and replace all records in the cluster with this centroid.
6. Return the replaced data table D' .

Algorithm 1 returns data table D' satisfying k-anonymity, where each group contains at least k records. Numerical and categorical attributes in each group are replaced with mean and generalized values respectively, reducing query function sensitivity.

2.2 Improved Clustering Method for Differential Privacy

Differential privacy and clustering algorithms provide different information disclosure protections. Using clustering algorithms can reduce the noise required in differential privacy, achieving differentiation of query function sensitivity, while differential privacy protection can compensate for clustering algorithms' inability to resist arbitrary background knowledge attacks. Their combination achieves better privacy protection results while maintaining good data utility.

Let M be the clustering function and f be the query function. To effectively reduce the sensitivity of f, M should satisfy that for datasets D and D' (where D' is generated by modifying one record in D), the cluster centers remain basically stable. This requires that all clusters generated from dataset D and the corresponding clusters from D' differ by only one record. Such clustering algorithms are called insensitive clustering, and only insensitive clustering functions can perform differential privacy protection [20].

Definition 7 Insensitive Clustering. Assume dataset D and clustering function M, with clustering result {C_1, C_2, ···, C_n} = M(D). Let D' be a dataset obtained by modifying only one record in D, with clustering result {C' _1, C'

_2, ···, C' _n} = M(D'). If the corresponding clusters differ by only one data record, the clustering algorithm M is called insensitive clustering.

To make clustering method CMD satisfy insensitive clustering for differential privacy protection, the distance function D must be changed to a total order function [20]. For mixed-type data tables, a distance function satisfying total order relations can be constructed as follows:

Assume data table D contains n attributes, with p categorical attributes and q numerical attributes. Let X,Y be any records in D, and Z be the cluster center of D. Compute the farthest record from Z using the distance formula from Definition 5, denoted as X_b, and compute the farthest record from X_b, denoted as X_t. Define the anonymization method for data table D as:

By introducing the above distance function into clustering algorithm CMD, we construct the insensitive clustering algorithm ICMD (Insensitive CMD).

Algorithm 2 Insensitive Clustering Algorithm ICMD(D,k) - Input: Original dataset D with n records, minimum cluster size k - Output: Clustered dataset D' that can perform differential privacy protection

1. Calculate the boundary [X_b, X_t] of the original dataset.
2. Calculate the k nearest records to X_b and X_t respectively, and assign them to clusters, adding them to dataset D' .
3. For the remaining m records, if m 2k, repeat step 1 for the remaining data records; if k m < 2k, form a separate class and add to dataset D' .
4. Otherwise, assign the remaining m records to their nearest clusters.
5. Calculate the centroid for each cluster and replace all records in the cluster with this centroid.
6. Return the replaced data table D' .

Using the distance calculation method from equation (4) in Algorithm 2, ICMD satisfies the definition of insensitive clustering algorithms, enabling differential privacy protection on its results. From literature [20], for query function f_i (returning the i-th record of the dataset), the sensitivity Δf_i(ICMD) Δf_i(k). This shows that the original dataset, after clustering, achieves record hiding and differentiates query sensitivity from single records to group records.

2.3 Differential Privacy-Preserving Data Publishing Method

Clustering anonymization based on k-anonymity cannot resist background knowledge attacks and homogeneity attacks. For further protection, noise is added to data records on the basis of clustering to achieve differential privacy protection. Using the method from literature [20], Laplace noise is added to implement a data protection method ICMD-DP (Clustering for Mixed Data with Differential Privacy) that perturbs mixed-attribute data tables.

Algorithm 3 Differential Privacy Protection Algorithm ICMD-DP(D,) - Input: Original dataset D with n records, privacy budget - Output:

Dataset D' satisfying -differential privacy

1. Perform clustering on dataset D using ICMD(D,k) to obtain dataset D' .
2. For query function f_i returning the i-th record' s attributes, add Laplace noise to the query result and add it to dataset D' .
3. Return dataset D' .

Each query function result satisfies -differential privacy, and since each query targets disjoint records, according to the parallel composition principle [15], the final dataset D' satisfies -differential privacy.

For a dataset with cluster size k, the single query sensitivity is less than n/k. With n mutually independent queries, to ensure the query sensitivity of differential privacy-protected data is less than that of the original dataset, we need $n/k < n$, i.e., $k > 1$. Although clustering algorithms cause information loss, this loss can be compensated by the gain from reduced sensitivity.

3 Experiments

This chapter experimentally analyzes the proposed method in terms of time consumption, information loss, and privacy leakage risk.

3.1 Experimental Data and Environment

The experimental data uses the UCI Adult dataset (http://archive.ics.uci.edu/ml/datasets/Adult), commonly used to evaluate privacy protection methods. This mixed-attribute dataset contains 6 numerical attributes (e.g., age, hours-per-week) and 8 categorical attributes (e.g., occupation, native-country), with 48,842 records total. Records with missing values are removed, and 30,000 complete records are selected for experiments.

3.2 Experimental Methods

Similar to k-anonymity evaluation, the proposed combination scheme is assessed from two aspects: information loss (affecting data utility) and information disclosure (revealing privacy protection level).

3.2.1 Information Loss Information loss refers to the difference between the anonymized dataset and the original dataset, typically measured by SSE (Sum of Squared Errors) [20]. SSE represents the sum of squared attribute distances between all records in the anonymized dataset and the original dataset:

```
SSE = Σ_{i X} Σ_{j A} dist(a_{ij}, a'_{ij})^2
```

where a_{ij} is the j-th attribute of the j-th record in the original dataset, and a' _{ij} is the corresponding attribute in the anonymized dataset. For categorical and numerical attributes, distance functions dist($\cdot$) use equations (2) and (3) respectively. Larger SSE values indicate more severe information loss and poorer data utility.

3.2.2 Information Disclosure Information disclosure is measured by the probability of successfully matching records from the anonymized dataset to the original dataset, also called Record Linkages (RL):

```
RL = (1/n) × Σ_{j=1}^n Pr(x_j → x'_j) × 100
```

where:

```
Pr(x_j → x'_j) = { 1/|G| if x'_j   G
                 { 0 otherwise
```

Here, G is the set where record x_j resides. If record x' _j is also in set G, it is considered to cause information disclosure with probability 1/|G|; otherwise, the probability is 0.

To better demonstrate the effectiveness of the proposed method, we calculate SSE and RL for clustering algorithm CMD and standard -differential privacy as baselines, and compare them with ICMD and ICMD-DP. Privacy budget uses common values 0.01, 0.1, 1, 10, and k ranges from 2 to 500.

3.3 Experimental Results

Using clustering algorithm CMD and standard -differential privacy algorithm as baselines, comparative experiments are conducted by adjusting k values, with results shown in [Figure 1: see original paper] and [Figure 2: see original paper].

As shown in Figures 1 and 2, the insensitive clustering algorithm ICMD causes greater information loss than the original clustering algorithm CMD but correspondingly reduces information disclosure risk. Performing differential privacy on top of ICMD effectively reduces information disclosure risk, providing better privacy protection. Smaller privacy budgets yield more obvious privacy protection effects but also cause greater information loss.

Figure 3 [Figure 3: see original paper] shows that as k increases, the information loss of differential privacy after clustering gradually decreases, and when k is near n/ , it has similar information loss to standard differential privacy. However, when k > n/ , its information loss gradually becomes smaller than standard differential privacy. Figure 4 [Figure 4: see original paper] shows that differential privacy after clustering has smaller information disclosure and better data protection effects than standard differential privacy.

4 Conclusion

This paper proposes a feasible data publishing method ICMD-DP for mixed-attribute data tables, which combines clustering and differential privacy to balance the contradiction between data utility and privacy protection. ICMD-DP performs differential privacy after cluster anonymization, reducing information disclosure risk while increasing data utility. First, based on MDAV, we propose the clustering algorithm CMD for mixed data tables. Second, to better implement differential privacy, we introduce a total order distance function into CMD,

propose the insensitive clustering algorithm ICMD, and use ICMD' s results as input for differential privacy protection. Finally, experiments analyze the effectiveness of the proposed method in protecting user privacy and improving data utility on mixed-attribute data tables.

References

[1] Cao Chunping, Zheng Xia. Research on privacy protection anonymity model in social networks [J]. Small Microcomputer Systems, 2016, 37(8): 1821-1825. [2] Liu Xiaoqian, Li Qianmu. Differential privacy preserving data publishing method based on clustering anonymization [J]. Journal on Communications, 2016, 37(5): 125-129. [3] Liu Xiangyu, Wang Bin, Yang Xiaochun. Survey on privacy preserving techniques for publishing social network data [J]. Journal of Software, 2014, 25(3): 576-590. [4] Li Hongcheng, Wu Xiaoping, Chen Yan. K-means clustering method supporting differential privacy protection under MapReduce framework [J]. Journal on Communications, 2016, 37(2): 124-130. [5] Dwork C. Differential privacy [J]. Lecture Notes in Computer Science, 2006, 26(2): 1-12. [6] Wang D, He D, Wang P, et al. Anonymous two-factor authentication in distributed systems: certain goals are beyond attainment [J]. IEEE Trans on Dependable & Secure Computing, 2015, 12(4): 428-442. [7] Domingo-Ferrer J, Sánchez D, Hajian S. Database privacy [M]. Springer International Publishing, 2015. [8] Gong Weihua, Lan Xuefeng, Pei Xiaobing, et al. Social network privacy protection method based on K-degree anonymity [J]. Acta Electronica Sinica, 2016, 44(6): 1437-1444. [9] Jiang Huowen, Zeng Guosun, Ma Haiying. Greedy clustering anonymous method for privacy protection of tabular data publishing [J]. Journal of Software, 2017, 28(2): 341-351. [10] Yang Gaoming, Yang Jing, Zhang Jianpei. Clustering-based (, k)-anonymous data publishing [J]. Acta Electronica Sinica, 2011, 39(8): 1941-1946. [11] Huang Shiping, Gu Jinyuan. Enhanced privacy protection model based on (p+,)-sensitive k-anonymity [J]. Application Research of Computers, 2014, 31(11): 3465-3468. [12] Li Yang, Wen Wen, Xie Guangqiang. Survey on differential privacy protection [J]. Application Research of Computers, 2012, 29(9): 3201-3205, 3211. [13] Torra V. Microaggregation for categorical variables: a median based approach [C]// Proc of Privacy in Statistical Database. 2004: 162-174. [14] Zhao Xingwang, Liang Jiye. Mixed data attribute weighted clustering algorithm based on information entropy [J]. Journal of Computer Research and Development, 2016, 53(5): 1018-1028. [15] Li Yang, Hao Zhifeng, Wen Wen, et al. Research on differential privacy preserving K-means clustering method [J]. Computer Science, 2013, 40(3): 287-290. [16] Zhang Xiaojian, Meng Xiaofeng. Differential privacy for data publication and analysis [J]. Chinese Journal of Computers, 2014, 37(4): 927-949. [17] Xiong Ping, Zhu Tianqing, Wang Xiaofeng. Differential privacy protection and its applications [J]. Chinese Journal of Computers, 2014, 37(1): 101-122. [18] Xia Zanzhu. Research on privacy protection anonymization algorithms for microdata publishing [D]. Jinhua: Zhejiang Normal University, 2011. [19] Zhang Huizhe, Wang Jian. Improved FCM clustering algorithm based on initial cluster center selection [J]. Computer Science, 2009, 36(6): 206-209. [20]

Soria C J, Domingo F J, Nchez D, et al. Enhancing data utility in differential privacy via microaggregation-based k-anonymity [J]. VLDB Journal, 2014, 23(5): 771-794.

Note: Figure translations are in progress. See original paper for figures.