A deep neural network-based algorithm for safe release of big data under random noise disturbance

Jian Yu[1], Hui Wang[2,*]

1. Liuzhou Vocational and Technical College, School of Electronic Information Engineering, Liuzhou 545005, China

2. Liuzhou Vocational and Technical College, School of Art, Liuzhou 545005, China

Huiwang.liuzhou@gmail.com

ABSTRACT. *Despite its huge benefits, the release of big data is faced with the severe risk of privacy leakage. To solve the problem, this paper proposes a deep neural network (DNN)-based algorithm for safe release of big data under random noise disturbance. Specifically, a random noise of a certain probability distribution was added into the release of the big data, such that the public output will not change significantly whether an individual data record is in the dataset and that that the published data will be basically the same to the original dataset. The algorithm was then optimized in light of the attributes of the correlated datasets in big data. Finally, the proposed algorithm was proved better than the traditional algorithm in large-scale searches of correlated datasets, and capable of ensuring privacy at a lower privacy budget.*

RÉSUMÉ. *Malgré ses énormes avantages, la libération de Big Data est confrontée à un risque élevé de la divulgation de confidentialité. Pour résoudre ce problème, cet article propose un algorithme basé sur un réseau neuronal profond (DNN) pour la diffusion sécurisée de Big Data en cas de perturbation du bruit aléatoire. Plus précisément, un bruit aléatoire d'une certaine distribution de probabilité a été ajouté à la diffusion des données massives, de sorte que la sortie publique ne change pas de manière significative avec la présence d'un enregistrement de données individuel dans le jeu de données et que les données publiées seront fondamentalement identiques à l'ensemble de données d'origine. L'algorithme a ensuite été optimisé à la lumière des attributs des ensembles de données corrélés dans le Big Data. Enfin, l'algorithme proposé s'est avéré meilleur que l'algorithme traditionnel dans les recherches à grande échelle des ensembles de données corrélés, et il est capable de garantir la confidentialité avec un budget de confidentialité inférieur.*

KEYWORDS: *deep neural network (DNN), big data, privacy preserving, differential privacy.*

MOTS-CLÉS: *réseau neuronal profond (DNN), big data, préservation de la confidentialité, confidentialité différentielle.*

DOI:10.3166/ISI.23.6.189-200 © 2018 Lavoisier

1. Introduction

The number of Internet users in China has reached 751 million, about 54.3% of the total population [1]. Similar trends are observed across the globe. The popularity of the Internet has made it easy to acquire and share data, heralding the dawn of the big data era. In May, 2009, Data.gov was launched by the US government to improve public access to improve public access to high value, machine-readable datasets generated by the Executive Branch of the Federal Government. Nearly 40 countries and regions quickly followed suit by setting up their own open data portals [2]. The information made public on these websites often involves private data of government departments, enterprises and individual users. Against this backdrop, it is of great importance to hide individual data and private data in the release of big data (Fung *et al*., 2010; Wong *et al*., 2011). Otherwise, neither the traditional encryption strategies nor the access control of some fields could withstand the increasingly diverse hacker attacks (Kifer and Machanavajjhala, 2011; Kifer *et al*., 2012; Noman *et al*., 2011; Xiao *et al*., 2014).

In light of the above, this paper designs a transparency algorithm to preserving the sensitive information of individual records in big data, which adds a random noise of a certain probability distribution into the release of the big data, prevents the public output from changing significantly whether an individual data record is in the dataset, and ensures that the published data are similar to the original data within a certain threshold range. The transparency algorithm was then optimized in light of the attributes of the correlated datasets in big data. Finally, the proposed algorithm was proved effective through experiments.

2. Definition of privacy in the release of big data

For a dataset $\mathcal{D}$, its two sub-datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ differ by one record at the most. Let $\mathcal{M}$ be a set of random noises. Then, any output $O \subseteq Range(\mathcal{M})$ satisfies:

$$\Pr[\mathcal{M}(\mathcal{D}_1) \in O] \leq \exp(\epsilon) \times \Pr[\mathcal{M}(\mathcal{D}_2) \in O] \quad (1)$$

where $\mathcal{M}$ obeys a certain random probability distribution; ϵ is a real number falling in [0, 1] indicating the strength of privacy preserving; Pr[ ] is the privacy risk, i.e., the probability that the privacy is leaked. If dataset $\mathcal{D}_1$ is released to select counting statistics, then an aggregate search, Count(n) in the first n rows can be shown in Table 1 below.

Even if dataset $\mathcal{D}_1$ rejects direct access and only offers the Count(n) search interface, a hacker (Attacker A) with certain background knowledge, e.g. the sorting position m of the user "Zhao" (the website data are usually ranked in such orders as

1. China Internet Network The 23 times Information Center. statistical report on Internet development in China. http://www.cac.gov.cn/2018-01/31/c_1122347026.htm
2. China's State Council. Action Plan on Promoting Big Data Development. http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm

the alphabetical order of user surnames), can acquire the private information of Zhao (whether the search result is satisfied) by the attack method Count(m)- Count(m-1).

Table 1. Data publishing

Users	Aggregation query
Lee	False
Mike	True
Green	True
Brown	True
…	…

According to equation (1), the dataset without the data record on Zhao can be considered as $\mathcal{D}_2$. The probability of Attacker A to acquire the Count() value from dataset $\mathcal{D}_1$ obeys a similar distribution as that of he/she to acquire the Count() value from dataset $\mathcal{D}_2$. Assuming that the probability of Count(m) is almost equal to that of Count(m-1), then the private information of Zhao in dataset $\mathcal{D}_1$ is protected.

As mentioned above, ϵ is the real number privacy preserving budget, that is, the strength of the differential privacy preserving. The value of ϵ is negatively correlated with the preserving strength. Hence, the value of privacy budget ε controls the similarity of the probability distribution. The smaller the ϵ, the closer the e^{ϵ} is to one. In other words, the privacy preserving model satisfies the differential privacy (Dwork, 2011a, 2011b; Dwork and Roth, 2014; Hall *et al*., 2013).

3. Random noise addition mechanism in the release of big data

In the previous case, it is assumed that, if $f(n)$ ensures the privacy of Count(n), then $f(n)$= Count(n)+ Noise, with Noise being a noise obeying certain random distribution.

If X = Noise, then we have:

$$\forall o, \frac{\Pr[\mathcal{M}(\mathcal{D}_1)=o]}{\Pr[\mathcal{M}(\mathcal{D}_2)=o]} = \frac{\Pr[\mathrm{Count}(\mathcal{D}_1)+\mathrm{X}_1=o]}{\Pr[\mathrm{Count}(\mathcal{D}_2)+\mathrm{X}_2=o]} \leq e^{\mathcal{E}} \tag{2}$$

Let d = Count ($\mathcal{D}_1$) - Count ($\mathcal{D}_2$). Then, we have:

$$\forall x, \frac{\Pr[\mathrm{X}_1=x]}{\Pr[\mathrm{X}_2=x+d]} \tag{3}$$

If $\mathcal{D}_1$ and $\mathcal{D}_2$ differ by one record at the most, Δf can be defined as the global sensitivity of any search f on dataset $\mathcal{D}$. Then, we have:

$$\Delta f = \max_{\mathcal{D}_1,\mathcal{D}_2} |\mathrm{Count}\ (\mathcal{D}_1)\ -\ \mathrm{Count}\ (\mathcal{D}_2)| \tag{4}$$

Here, d must be equal to or smaller than Δf to make equation (3) permanently established.

The Laplace mechanism and the exponential mechanism are two existing methods to obtain ϵ- differential privacy (Li *et al*., 2018). In the Laplace mechanism, the original function output is added with a noise $\mathrm{Lap}(\frac{\Delta f}{\epsilon})$, indicating that any search f on dataset $\mathcal{D}$ satisfies ϵ- differential privacy under:

$$\mathcal{M}(\mathcal{D}) = Count(\mathcal{D}) + \mathrm{Lap}(\frac{\Delta f}{\epsilon}) \quad (5)$$

The exponential mechanism depends on the function $Q(\mathcal{D}, \psi)$, where ψ is the candidate item. The function outputs the number of ψ items in $\mathcal{D}$. Thus, the probability of $\mathcal{M}(\mathcal{D})$ output $\psi \in \mathcal{D}$ is proportional to $\epsilon^{\frac{\epsilon Q(\mathcal{D},\psi)}{2\Delta q}}$, indicating that $\mathcal{M}(\mathcal{D})$ satisfies the exponential mechanism of ϵ-differential privacy.

In the context of big data, user privacy information is distributed under multiple datasets. The possibility of privacy exposure will grow with the fusion of multi-source data. A preserving mechanism satisfies ϵ-differential privacy on a single dataset might not provide the same privacy guarantee on multiple datasets (Beimel *et al*., 2014). If multiple datasets, which are entirely different, have the same record on a user, then the direct correlation can satisfy ϵ-differential privacy. However, it is possible that different records between a user and those between the user's correlated users are also correlated, such as GPS records, sales records, or customer relationship network information. The correlated records of correlated datasets in big data provide additional, unpredictable information to the attacker. In fact, a key defect in traditional differential privacy lies in neglecting the correlation between records (Koufogiannis *et al.*, 201; Parra-Arnau et al., 2013). If this problem is solved simply by adding the number of correlated records and enhancing the corresponding sensitivity, the search results will contain a huge amount of redundant noises, hurting the validity of the dataset (Wang *et al*., 2016). The previous research has proved that the correlation between individual records will reduce the individual privacy (Kifer and Machanavajjhala, 2014). Therefore, it is imperative to protect the privacy of correlated datasets.

4. Privacy analysis of correlated datasets

This section defines the concepts and terms before handling the differential privacy of correlated datasets.

Definition 1: Correlated degree (CD)

If a record $\mathcal{R}_i \in \mathcal{D}$ is correlated to $k-1$ records ($k \leq |\mathcal{D}|$), then the record can be expressed as:

$$\mathrm{R_i} = \{\mathcal{R}_i, \mathcal{R}_j \in \mathcal{D} | \text{All } \mathcal{R}_j \text{ are correlated with } \mathcal{R}_i\} \quad (6)$$

When $k = 1$, the correlated dataset $\mathcal{D}$ is independent identically distributed (IID).

In big data scenarios, most records are partially correlated, and the deletion of a record may have varied degrees of impacts to the other records. Here, these impacts

are defined as the degree of correlation. For two correlated records $\mathcal{R}_i$ and $\mathcal{R}_j$, the degree of correlation $\vartheta_{ij} \in [-1, 1]$ and $\vartheta_{ij} \leq \vartheta_0$, with ϑ_0 being the threshold of the degree of correlation. If $\vartheta_{ij} < 0$, then $\mathcal{R}_i$ is negatively correlated with $\mathcal{R}_j$; if $\vartheta_{ij} > 0$, then $\mathcal{R}_i$ is positively correlated with $\mathcal{R}_j$; if $\vartheta_0 = 0$, then $\mathcal{R}_i$ is not correlated with $\mathcal{R}_j$; if $\vartheta_0 = 1$, then $\mathcal{R}_i$ is completely correlated with $\mathcal{R}_j$. The degree of correlation describes the degree of impacts of a record on other records. The greater the ϑ_{ij}, the weaker the correlation, i.e. $\mathcal{R}_j$ is not severely affected by the deletion of $\mathcal{R}_i$; otherwise, if the value of ϑ_{ij} approximates 1 or -1, the correlation between the record and the other record is strong, i.e. $\mathcal{R}_j$ is severely affected by the deletion of $\mathcal{R}_i$. Thus, the degree of correlation matrix Δ ($\vartheta \in \Delta$) of dataset $\mathcal{D}$ can be obtained as:

$$\Delta = \begin{pmatrix} \vartheta_{11} & \cdots & \vartheta_{1n} \\ \vdots & \ddots & \vdots \\ \vartheta_{n1} & \cdots & \vartheta_{nn} \end{pmatrix} \tag{7}$$

where ϑ_{ij} and ϑ_{ji} are symmetrical with each other. Their correlation is independent of the order of records, and the diagonal elements are all equal to 1. The degree of correlation can be filtered by adjusting the threshold ϑ: if $\vartheta_{ij} < \vartheta_0$ in Δ, then ϑ_{ij} is set to zero. In reality, it is very difficult for the attacker to obtain the entire Δ. To achieve the highest level of privacy guarantee, this paper assumes that the privacy mechanism can protect the privacy of individuals even if the attacker can obtain the entire Δ.

Definition 2: Correlated Sensitivity (CS)

The correlated sensitivity of record $\mathcal{R}_i$ to the search $\mathcal{S}$ offered by the correlation degree matrix Δ of dataset $\mathcal{D}$ can be expressed as:

$$CS_i = \sum_{j=0}^{n} |\vartheta_{ij}| \left(\left\| \mathcal{S}(\mathcal{D}_j) - \mathcal{S}(\mathcal{D}_{-j}) \right\| 1 \right) \tag{8}$$

If $\mathcal{D}$ is IID, then the global sensitivity of $\mathcal{S}$ is equivalent to CS_i. Thus, the correlated sensitivity of $\mathcal{S}$ can be defined as the maximum value CS_i, i.e. $CS_{d_s} = \max_{i \in d_s}(CS_i)$. Let X_s be the results set of $\mathcal{S}$. When a search only covers independent or weak correlated sensitivity, the correlated sensitivity will not introduce additional noise.

The correlated sensitivity CS applies to various data distribution mechanisms. If the records in dataset $\mathcal{D}$ are independent, the correlated sensitivity CS will be equal to the global sensitivity Δf; for correlated dataset, the correlated sensitivity CS will be smaller than the global sensitivity Δf.

Definition 3: Semantics of correlated privacy features

First, the following parameters should be defined:

(1) $\mathcal{G} = (\mathcal{V}, \sum, \mathcal{S}, \mathcal{R})$, with $\mathcal{S} \in \mathcal{V}$ be the start symbol;

(2) $\mathcal{V} = \begin{Bmatrix} \mathcal{S};\ L, D, S; \\ N_1, \dots, N_n; \\ B_1, \dots, B_n; \\ A_1, \dots, A_n; \\ E_1, \dots, E_n; \\ P_1, \dots, P_n; \\ I_1, \dots, I_n; \\ \dots \end{Bmatrix}$ is a set of fields in the privacy feature, with N being the name, $N_1, \dots, N_n$ being the various descriptions methods of the name (e.g. N_1 is the full name or acronym), B being birthday, A being address, E being email, P being phone number, I being ID card number. The tag type in $\mathcal{V}$ can be added or removed as needed, and each type of tags can adapt to the increase or decrease of n;

(3) $\sum = \{ASCII \text{ characters}, NULL\}$ refers to all the characters that do not include $\mathcal{V}$ but $\mathcal{G}$;

(4) The finite set of $\mathcal{R}$: $\mathcal{A} \to \alpha, \mathcal{A} \in \mathcal{V}, \alpha \in \mathcal{V} \cup \Sigma$.

Then, the recognition of $\mathcal{R}$ in $\mathcal{G}$ can be expressed as $\mathcal{A} \to \alpha(\mathcal{A} \in \mathcal{V}, \alpha \in \mathcal{V} \cup \Sigma)$. This paper designs a suitable training algorithm for correlation recognition, which improves the accuracy of semantic recognition based on the deep neural network (DNN) (Deng and Yu, 2014).

Algorithm 1: Correlation relationship recognition algorithm

(1) Let $\mathcal{V} = \{\mathcal{V}_k | k = 1, 2, \dots, \mathcal{K}\}$ be the given fields, $\vartheta = \{\vartheta_k \epsilon Tag^D | k = 1, 2, \dots, \mathcal{N}\}$ be the correlation degree features of the classifier sample set and $Tag = \{tag_k | k = 1, 2, \dots, \mathcal{N}\}$ be the classification tags;

(2) Set up a DNN of *layer*+1 layers, with n_{layer} neurons in the *layer* hidden layers, and adopt ReLU as the activation function of neurons. Then, the j-th neuron in the ℓ-th hidden layer ($\ell \in (1, layer)$ can be expressed as:

$$\mathcal{H}_i^{(\ell,j)} = \max(0, \mathcal{Z}^{(\ell,j)}) \tag{9}$$

(3) Let $\mathcal{Z}^\ell$ be the output vector of the ℓ-th layer. To prevent local optimum trap or excessive gradient, the output vector should be adjusted as:

$$\mathcal{Z}^\ell = \mathcal{W}^\ell \mathcal{Z}^{(\ell-1)} + b^{(\ell)}, \ell = 1, 2, 3, \dots, \mathcal{L} \tag{10}$$

Since each layer receives the inputs from $\mathcal{K}$ function, the weight can be expressed as:

$$\mathcal{W}_k = \prod_{\ell=1}^{\mathcal{L}} \widehat{\mathcal{W}}_k^\ell \tag{11}$$

The weight should be corrected as:

$$b_k = \sum_{\ell=1}^{\mathcal{L}-1} \left(\prod_{n=\ell+1}^{\mathcal{L}} \widehat{\mathcal{W}}_k^n \right) \hat{b}_k^\ell + b^{\mathcal{L}} \tag{12}$$

(4) The probability that the sample vector $\mathcal{Z}$ belongs to the ℓ-th class can be expressed as:

$$\sigma\left(\mathcal{V}|\mathcal{Z}_k^{\mathcal{L}}\right) = \frac{exp\left(\mathcal{Z}_k^{\mathcal{L}}\right)}{\sum_{k=1}^{K} exp\left(\mathcal{Z}_k^{(\mathcal{L},k)}\right)} \tag{13}$$

(5) Exclude the neurons whose hidden layer output is 0 and the parameters connected to them and adjust the threshold or parameters. Then, the correlation probability of sample $\mathcal{V}_k$ can be judged as:

$$\xi(\mathcal{V}_k) = Tag_k \text{ iff } k = \arg \text{ReLU}\, \mathcal{Z}_k^{(\mathcal{L},j)}, 1 \le j \le K \tag{14}$$

5. Noise addition mechanism of correlated datasets

Definition 4: Correlated ϵ- differential privacy

Let $\mathcal{D} = \{D_1, D_2, \dots, D_n\}$ be a dataset and $\mathcal{M}$ be a mechanism satisfying the differential privacy. Then, any pair of datasets D_i^1 and $D_i^2 \in D_i (1 \le i \le n)$ differ by one record at the most. Assuming that any output $O \subseteq Range(\mathcal{M})$ in the domain of definition $\mathcal{D}$, then the correlated ϵ- differential privacy can be expressed as:

$$\mathcal{M}(\mathcal{D}) := \sup_{i,\mathcal{D}_{-i},D_i^1,D_i^2,O} \log \frac{\Pr(\mathcal{M}(D_i^1) \in O | D_i^1, \mathcal{D}_{-i})}{\Pr(\mathcal{M}(D_i^2) \in O | D_i^2, \mathcal{D}_{-i})} \tag{15}$$

Let $\mathcal{S}$ be the search provided by $\mathcal{D}$ and X_s be the results set of $\mathcal{S}$. If and only if $\mathcal{M}(\mathcal{D}) \le \epsilon, \mathcal{M}(\mathcal{D})$ provides the ϵ- differential privacy of the correlated datasets. From the above section, we have:

$$\mathcal{M}(\mathcal{D}) = \mathcal{S}(\mathcal{D}) + \text{Lap}(\frac{CS_{X_S}}{\epsilon}) \tag{16}$$

The correlated sensitivity is lower than the global sensitivity. However, the privacy budget must be divided into several small parts when multiple searches are performed in the big data environment, making it necessary to minimize the impact of the noises in the search results. When the records are closely correlated with the other data, the resulting noises will be obviously higher than those of independent datasets. Here, the noises in the search results are limited by the iterative mechanism.

Algorithm 2: Correlated data noise addition algorithm

Let $t, \mathcal{Q}, \mathcal{Q}_t, A_t$ and $\hat{A}$ be the round of iterations, the search set, the search of the *t*-th round, the actual search result, and the noise disturbed result, respectively. The data set in the process can be expressed as a histogram $x = \{x_0, x_1, \dots, x_t\}$ with the length of *N*:

$$A_t = \mathcal{Q}_t(x_t) \tag{17}$$

$$(A\_t\ )\hat{} = Q\_t(x_t) + \mathrm{Lap}(\frac{CS_{Q_t}}{\epsilon}) \tag{18}$$

Let $\widehat{d_t}$ be the difference between the actual search result $X_{s,t-1}$ and the noise disturbed result $X_{s,t}$. This difference can be adopted to control the update in each iteration, such that the $x_0, x_1, \dots, x_t$ in each round approximates the original dataset x:

$$\widehat{d_t} = Q_t(x_{t-1}) - \widehat{A_t} \tag{19}$$

First, the privacy budget can be divided into several parts. Assuming that $\epsilon_0 = \frac{\epsilon\mu\vartheta_0}{\log N}$, the histogram should be initialized as a uniform distribution x_0. In each iteration, $Q_t(Q_t \in Q)$ is executed in x_t, yielding the result $A_t = Q_t(x_t)$. Then, the A_t can be disturbed as $\widehat{A_t}$. After that, the $Q_t(x_{t-1})$ and the distance $\widehat{d_t}$ of $\widehat{A_t}$ can be calculated for the previous round. If $\widehat{d_t} < T$ (T is the given threshold), then x_{t-1} is very similar to x during the search Q_t. In this case, $Q_t(x_{t-1})$ and x_{t-1} should be released directly and the next round of iteration should begin; otherwise, if $\widehat{d_t} > T$, then the x_{t-1} must be corrected.

Let $\mathbb{q}_t = \{b_0, b_1, \dots\}$ be a superset of all the results on x_{t-1} and its correlated records. Then, the superset should be corrected by:

$$x_t = F(x_{t-1}) \tag{20}$$

$$F(x_{t-1}) = x_t(b_i) \tag{21}$$

$$F(x_{t-1}) = x_{t-1}(b_i)\exp(-\mu\theta_{Q_t}\gamma(x_{t-1})) \tag{22}$$

If $\widehat{d_t} > 0$, then $\gamma(x_{t-1}) = Q_t(x_{t-1})$; otherwise, $\gamma(x_{t-1}) = 1 - Q_t(x_{t-1})\mu$, with μ being the adjustment parameter.

As above, a search result on the noise perturbation can be generated in each iteration by accessing the histogram $x = \{x_0, x_1, \dots, x_t\}$. The result only verifies if the current histogram is the correct result of the current round of search. In most cases, no corrected distributed result is released, leaving the privacy budget untouched. The correction and release of distributed result only occur when the current histogram is inaccurate. Therefore, the privacy budget is only consumed in the iterative round of corrections, in which the privacy analysis is severely constrained.

6. Experiments and results analysis

The experiments were carried out in the following environment: Pentium Xeon E5-2620 V3, 32GB memory and one NVIDIA GeForce GTX 980 GPU, using the Adult dataset from the UCI machine learning library. The initial dataset contains 48,842 records and 15 attributes. After removing unfavorable data, there are 35,561 records and 14 attributes in experimental dataset. The Salary field was selected for

our experiments and the attacks aim to determine whether the annual income of an individual user surpasses 5,000 USD. The experimental results shed light on the privacy preserving methods for individual users, enterprises, public institutions or government during the release of big data.

No data contains pre-defined correlation information. The elements in the correlation degree matrix Δ in our simulation all fall between [-1, 1] and are linearly correlated. The matrix was generated by Algorithm 1. The threshold was set to 0.6. About one fourth of the records in the dataset are correlated with each other. The size of the $k-1$ group is about 10. A total of 10,000 linear searches were generated randomly, and the search results fell in $[0, 1]$. Let DR be the deviation rate. Then, the algorithm accuracy can be verified by:

$$\mathrm{AD} = \frac{1}{|A|}\sum_{A_i \in A}\left|\left(\widehat{A}_i(x) - A_i(x)\right)\right| \tag{23}$$

where A is the search result; $\hat{A}$ is the search result distributed by the additive noise.

To verify the accuracy, the Laplace mechanism and CS mechanism were contrasted in privacy budget and prediction accuracy, under the privacy budget ϵ of [0.1, 1], correlation degree threshold T of 0.3 and μ of 0.7. The DRs of the two mechanisms are shown in Figure 1 below. It can be seen that the Laplace mechanism is, as expected, less accurate than the CS mechanism. This is because the privacy budget is only consumed in the correction iteration rounds of the CS mechanism, but in every release of search result in the Laplace mechanism.

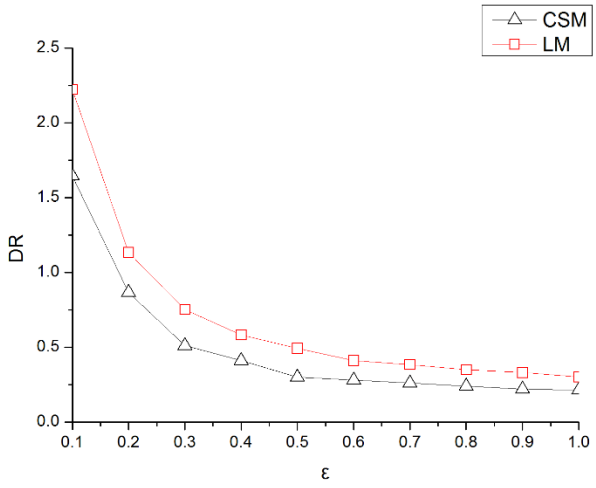

Figure 1. The DRs of Laplace mechanism and CS mechanism

T is the threshold that determines whether the search result needs to be corrected in iteration. It directly bears on the execution of the correction function and search

result. Thus, the relationship between the number of corrections and the threshold T was verified under the privacy budget $\epsilon = 1$. The results in Figure 2 show that the number of corrections decreased with the growth in T. When T was small, the number of all update rounds was 9,928; when T increased to 1, the number dropped to 62. As shown in Figure 3, the accuracy surged up with the increase of T, reached the maximum when T surpassed a threshold, and remained stable ever since.

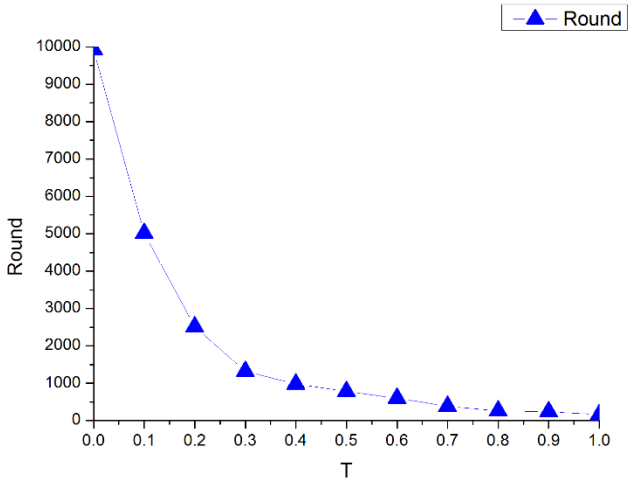

Figure 2. The relationship between the number of corrections and the threshold T

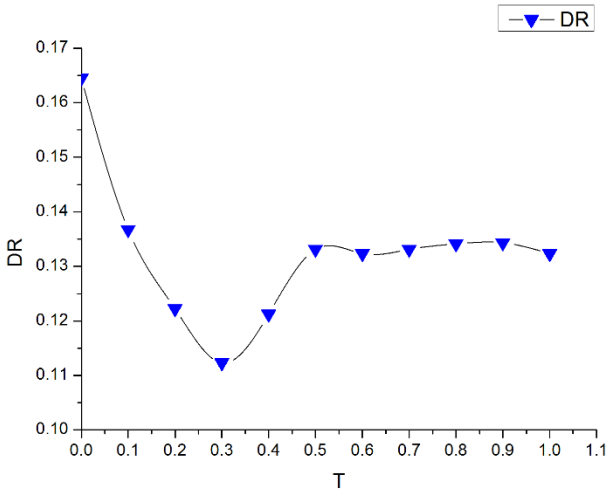

Figure 3. The effects of threshold T on accuracy

7. Conclusions

The traditional differential privacy mainly focuses on IID datasets, failing to ensure the privacy of correlated datasets. To solve the problem, this paper designs an algorithm capable of recognizing the differential privacy issue of the correlation and providing a data release mechanism for large-scale searches to reduce the loss of privacy budget and enhance data validity. The proposed algorithm was proved robust and effective through experiments. The future research will further investigate the differential privacy preserving of correlated datasets, including but not limited to effectively expressing the correlation degree matrix in big data, judging the importance of dataset released in each round under large-scale searches, and finding ways to improve the effect of the proposed algorithm.

Acknowledgement

This work is supported by {2018,2019} Foundation of Improving Academic Ability in University for Young Scholars of Guangxi.

References

Beimel A., Nissim K., Stemmer U. (2014). Private learning and sanitization: Pure vs. approximate differential privacy. *APPROX 2013, RANDOM 2013. Lecture Notes in Computer Science, Springer*, Berlin, Heidelberg, Vol. 8096, pp. 363-378. https://doi.org/10.1007/978-3-642-40328-6-26

Deng L., Yu D. (2014). Deep learning: Methods and applications. *Foundations and Trends® in Signal Processing*, Vol. 7, No. 3-4, pp. 197-387. http://dx.doi.org/10.1561/2000000039

Dwork C. (2011a). A firm foundation for private data analysis. *Communications of the ACM*, Vol. 54, No. 1, pp. 86-95. https://doi.org/10.1145/1866739.1866758

Dwork C. (2011b). The promise of differential privacy: a tutorial on algorithmic techniques. *Proc of the 52nd Annual IEEE Symposium on Foundations of Computer Science*, USA, pp. 1-2. https://doi.org/10.1109/FOCS.2011.88

Dwork C., Roth A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, Vol. 9, No. 3-4, pp. 211-407. https://doi.org/10.1561/0400000042

Fung B. C. M., Wang K., Chen R., Yu P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (CSUR)*, Vol. 42, No. 4, pp. 1-53. https://doi.org/10.1145/1749603.1749605

Hall R., Rinaldo A., Wasserman L. (2013). Differential privacy for functions and functional data. *J. Mach. Learn. Res*, Vol. 14, No. 1, pp. 703-727. https://doi.org/10.1109/MCS.2012.2225913

Kifer D., Machanavajjhala A. (2011). No free lunch in data privacy. *SIGMOD '11 Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, Athens, Greece*, pp. 193-204. https://doi.org/10.1145/1989323.1989345

Kifer D., Machanavajjhala A. (2014). Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems*, Vol. 39, No. 1, pp. 1-36. https://doi.org/10.1145/2514689

Kifer D., Smith A. D., Thakurta A. (2012). Private convex optimization for empirical risk minimization with applications to high-dimensional regression. *In COLT, Edinburgh, United Kingdom Duration*, pp.1-40. https://doi.org/10.1109/FOCS.2014.56

Koufogiannis F., Han S., Pappas G. J. (2016). Gradual release of sensitive data under differential privacy. *Privacy and Confidentiality*, Vol. 7, No. 2, pp. 1-22. https://doi.org/10.29012/jpc.v7i2.649

Li X. G., Li H., Li F. H., Zhu H. (2018). A survey on differential privacy. *Journal of Cyber Security*, Vol. 3, No. 5, pp. 92-104. http://dx.doi.org/10.19363/J.cnki.cn10-1380/tn.09.08

Noman M., Chen R., Fung B. C. M., Yu S. (2011). Differentially private data release for data mining. *Proceeding KDD '11 Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, California, USA*, pp. 493-501. https://doi.org/10.1145/2020408.2020487

Parra-Arnau J., Perego A., Ferrari E., Forne J., Rebollo-Monedero D. (2013). Privacy-preserving enhanced collaborative tagging. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 1, pp. 180-193. https://doi.org/10.1109/tkde.2012.248

Wang Y., Wang Y., Singh A. (2016). A theoretical analysis of noisy sparse subspace clustering on dimensionality-reduced data. *CoRR, eprint arXiv*, Vol. 1610, No. 07650, pp. 99. http://dx.doi.org/10.1109/TIT.2018.2879912

Wong R. C. W., Fu A. W., Wang K., Xu Y., Yu P. S. (2011). Can the utility of anonymized data be used for privacy breaches. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 5, No. 3, pp. 1-24. https://doi.org/10.1145/1993077.1993080

Xiao Q., Chen R., Tan K. (2014). Differentially private network data release via structural inference. *Proceeding KDD '14 Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, USA, pp. 911-920. https://doi.org/10.1145/2623330.2623642