A Data Desensitization Algorithm for Privacy Protection
Electric Power Industry

Zhenying Tang*, Wei Zhao, Chenfei Wang, Zixing Yang, Yin Xu, Shenhao Cui
Information Operation and Maintenance Center of Customer Service Center of State Grid Co., Ltd.
*Corresponding author e-mail: gwkfzx_xxywzx@163.com

Abstract. In view of the data security problems faced by big data technology in the development of electric power industry, this paper proposes a power big data desensitization algorithm applied to privacy protection, that is, a binary K-clustering algorithm (BKC-LDA) based on K-anonymity and L diversity. First, in order to reduce the computational complexity, a classification attribute is determined to classify the data table initially, and the equivalent class number K and the sensitive attribute value category L are limited according to the number of the original ancestor in the source data table. Then, considering the influence of the change of the internal range of the attribute value on the clustering, the equation for calculating the distance between the original ancestors with weight is established, using the idea of greedy and binary K clustering to classify the data table initially sets are clustered and generalized. In addition, this big data desensitization method determines the security level policies of different permissions by adjusting the size of K and L. Our algorithm can adapt to the desensitization of power big data with different attributes in different scenarios. It can not only fully mine the value of data, but also effectively protect the privacy of users.

1. Introduction
With the rapid development of digital technology and the popularization of mobile terminals, data has grown explosively, and the era of big data has come quietly. With the great business value of big data being paid more and more attention, all walks of life are committed to the mining and analysis of big data. However, big data often contains a large number of sensitive information of users, and publishing or sharing the original data without processing will inevitably disclose the user's privacy information. Therefore, in the process of big data application, we should adhere to the principle of both security and development, give full play to the value of data, and strive to solve the problem of data security and personal information protection.

The K-anonymity model proposed by Sweeney et al. in 2002 is a classic privacy protection method [1]. Furthermore, machanavajhala et al. put forward the concept of L-diversity [2], and solved the homogeneity attack that K-anonymity model cannot solve [3]. Aggrawal and others first proposed to use clustering method to achieve data anonymous privacy protection [4], the core idea of which is to use a first-generation and later-test mechanism to discover frequent patterns that meet the minimum support. Since then, some research results on clustering anonymity have been seen [5-7], including the GAA-CP clustering anonymity algorithm [5] proposed by Jiang huowen et al., which uses the greedy
method and clustering idea to minimize the distance of N tuples, can realize the anonymity respectively, and ensure the minimum data loss. When calculating the distance between primitives, GAA-CP combines the numerical attribute data and the classification attribute data at the beginning, which causes the subtype data to be merged without differentiation during clustering, thus expanding the generalization degree of data, increasing the amount of data loss, and reducing the value of data mining and analysis.

According to the classification and classification of the data sensitivity of the power business system and the practical application requirements, the desensitization strategies and methods of each sensitive data are determined. On the premise of meeting the data availability, the sensitivity of sensitive data is reduced, so as to achieve the goal of protecting sensitive data. At present, State Grid Corporation of China is vigorously promoting the construction of smart home, improving the intelligent level of residents' electricity use, and collecting massive information of users, mainly including two aspects:

(1) Operation data of business system in work, including all kinds of work order data, real-time power consumption data of all kinds of instruments, on-site work tickets and other data information.

(2) Personal information of the user, including: name, mobile number, ID number, home address, unit and other information. When State Grid Corporation of China outsources project engineering, it needs to provide test data. If it is directly exported from the database, it will inevitably leak user information [8]. Not only that, many big data platforms, medical systems and so on also have the risk of privacy disclosure. At present, the issue of personal privacy protection has attracted wide public attention. Not only in China, but also in the European Union and the United States, there is much new legislation to protect personal information. It can be seen that if the problem of personal privacy disclosure caused by big data sharing can’t be solved, it will bring serious legal risks to the release and use of relevant data, thus hindering the application and development of big data technology.

2. Data Desensitization Technology

Data desensitization refers to the transformation of sensitive information existing in business data according to a data desensitization strategy, so as to hide the sensitive information in the data. The connotation of data desensitization is to use the data desensitization technology to shield the sensitive information in the data to meet the requirements of the masked data and retain its original data format and attributes to ensure the normal operation during the development and testing of desensitized data [9].

Data desensitization technology mainly desensitizes data through methods such as anonymous technology, pseudonymous replacement, removal of identification information, and reduction of data accuracy. The common methods of data anonymization are generalization and concealment. These two methods are different from general methods such as scrambling or randomization, which can keep the consistency and authenticity of data before and after publishing. Commonly used anonymity technologies include K-anonymity algorithm, L-diversity anonym ity algorithm, etc. The data desensitization scheme adopted in this paper is based on the above two methods.

2.1. Related definitions

(1) Identifier (ID)

Given an original data table \( T \), ID is used to uniquely identify the attribute or attribute combination of individual identity, such as ID card number, meter user number, etc.

(2) Quasi-Identifier (QI)

For an original data table \( T \), by linking with identifiers of other data tables, attributes or combinations of attributes that uniquely identify individuals, such as gender, age, power consumption, etc.

(3) Sensitive Attribute (SA)

For data table \( T \), sensitive attributes refer to the attributes involving personal privacy that cannot be obtained from external information in the data table, such as salary level, disease status, etc. For example, the sensitive attribute in this paper is account balance (Table 1).
(4) Generalization

Given a data table $T$, for one of the ID attributes, the operation method to achieve anonymization is to replace the specific value with the range of the expression attribute, which is called generalization.

2.2. K-Anonymity and L-Diversity algorithm

(1) K-Anonymity

For a given data table $T$, the ID value of each tuple is $Q_i (1 \leq i \leq n)$, $n$ is the number of tuples (i.e. the number of data records). When $T$ performs the generalization operation $G(T)$ is $\hat{T}$, and the ID value of each tuple is $\hat{Q}_i (1 \leq i \leq n)$. After generalization, $\hat{T}$ satisfies K-anonymity, that is, if and only if any $\hat{Q}_i (1 \leq i \leq n)$, at least K $\hat{Q}_j (1 \leq j \leq n)$ can be found in $\hat{T}$, so that $\hat{Q}_i = \hat{Q}_j$, and the set of these equal tuples $EQ$ is called an equivalence class.

(2) L-diversity

In order to solve the problem that there are few sensitive attributes of equivalent classes in K-anonymity, which is easy to cause homogenous attack and privacy disclosure, L-diversity principle is used to increase the diversity of sensitive attributes. The L-diversity algorithm requires a set of ID equivalence classes to contain at least L different sensitive attributes. If each equivalence class in the published data table meets this requirement, the data table meets L-diversity. The definitions are as follows:

For a given data table $T$ and a positive integer L, the ID $Q_i = \{A_1, A_2, ..., A_d\}$, each ID value is $Q_i (1 \leq i \leq n)$, the sensitive attribute is $SA$, and after the generalization of the equivalence classes $\{EQ_1, EQ_2, ..., EQ_t\}$ in $\hat{T}$, each equivalence class satisfies L-diversity, that is, for any $EQ_i (1 \leq i \leq t)$, the number of sensitive attribute values is not less than L.

3. BKC-LDA Algorithm

This paper provides a binary K-clustering L-diversity algorithm (BKC-LDA) for clustering and anonymity of numerical data and classified data. Based on K-anonymity and L-diversity algorithm, the relevant big data of power users are screened, mined and analyzed, and the sensitive data that needs privacy protection and the public data that needs generalization are classified. On the one hand, before generalizing the data, a classification attribute $C_x$ is determined to classify the data table $T$ initially, so as to reduce the computational complexity. On the other hand, when calculating the distance between the data, the proportion of the attribute value in the distance measurement is determined according to the variance of the current attribute value, so as to make the clustering more reasonable, reduce the degree of generalization as much as possible and reduce the amount of data loss. Here, the calculation method of data loss in this paper is referred to [5].

In addition, the algorithm can generalize the data into different levels according to the user authority, and implement different data desensitization strategies. Figure 1 shows that according to the different viewing authority of users, the Beijing area is divided into three user levels. The range of information viewed by users of each level is different, that is, the higher the level is, the more data information can be found. The implementation of BKC-LDA algorithm is as follows.

![Figure 1. User level classification](image-url)
3.1. Classify data table attributes

Collect and organize information in each database: user's district, street address information, user's real-time power consumption, account balance, name, age, ID number and other information to form a source data table (Table 1). The finishing process includes:

1. The data is divided into numerical data and classified data. All attribute is expressed as $QI = \{N_1, ..., N_n, C_1, ..., C_{n_2}\}$, where $\{N_1, ..., N_n\}$ represents numerical attribute and $\{C_1, ..., C_{n_2}\}$ represents subtype attribute.
2. Hide data that needs to be hidden and generalize data that needs to be public.

3.2. Calculate the distance between tuples

For each data record in initial data table $T$, that is, the tuple $(i, T(P, P', P_1, ..., P_i))$, select a ID $C_j$ with classification attribute for preliminary division, that is, $T = \{P, P_2, ..., P_i\}(P_1 \neq P_2 \neq ... \neq P_i \neq P_j)$. According to the formula of distance between primitives, cluster the tuples in $P_j$ to further determine the generalization class.

Suppose for any two tuples $r_j, r_k$, including numerical data $\{N_1, ..., N_n\}$ and classified data $\{C_1, ..., C_{n_2}\}$. For numerical attribute $N_i$, the distance between data value is $d_{N_i}(r_j, r_k)$, which is defined as follows:

$$d_{N_i}(r_j, r_k) = \frac{1}{\sigma / N_i} \frac{|r_j - r_k|}{|D_p|}$$

Where $\bar{N}$ is average value, $\sigma$ is variance of attributes, $|D_p|$ is the length of domain value of attribute $N_i$ in the preliminary classified set $P_i$. $\frac{\sigma}{\bar{N}}$ is equivalent to weighting numerical data. The larger $\sigma$ is, the easier to distinguish the data of this attribute. It is necessary to increase the generalization distance between data and reduce the importance of distance measurement between records, so as to effectively reduce the amount of data loss.

For the classified attribute $C_j$, let $D$ be the classification domain, and $T_o$ be the generalization tree on $D$. For any two classification values $r_j, r_k \in D$, the distance $d_{C_j}(r_j, r_k)$ between $r_j, r_k$ is defined as:

$$d_{C_j}(r_j, r_k) = W(\Delta(r_j, r_k))/W(T_o)$$

Where $\Delta(r_j, r_k)$ is the subtree whose root is the smallest common ancestor of nodes $j$ and $k$, and $W(T)$ is the sum of the hierarchical distance of the generalization tree $T_o$, so as to realize the standardization of distance.

Then the total distance between tuples is expressed as:

$$D(r_j, r_k) = \sum_{i=1}^{n} d_{N_i}(r_j, r_k) + \sum_{j=1}^{n} d_{C_j}(r_j, r_k)$$

3.3. Clustering tuples

Step 1: cluster the initial classification set $P_i$, that is, select any tuple $r_j$ from the data set $P_i$, find out the farthest tuple from $r_j$, and use these two tuples as the centroids to cluster. At this time, consider the principle of clustering: for any other data tuple $r_k$ in $P_i$, calculate the distance between them and the two centroids $D(r_m, r_j), D(r_m, r_k)$, and classify $r_m$ as the centroid close to them, so as to divide $P_i$ into two parts $\hat{P}_j, \hat{P}_k$. If the current data set $\hat{P}_j$ or $\hat{P}_k$ has met the limit of data number, add this equivalence class to the equivalence class set: (assuming $\hat{P}_j$ met the condition), stop the clustering of the data set $P_i$, otherwise repeat step 1.

Step 2: On the basis of the previous step, perform generalization operation on the clustered data. In order to avoid homogenous attack, the sensitive attribute SA is considered. The set satisfying L-diversity is generalized. The set which does not satisfy the number of sensitive attributes is subdivided again. The number of tuples is increased and the size of equivalent class is expanded until the...
requirement that equivalence class category is not less than \( L \) is met. If \( P_i \) does not meet the limit of data number at this time, it will be classified as the current equivalent class. Although this may lead to the expansion of generalization scope and the increase of data loss, it reduces the risk of sensitive information disclosure and ensures information security. Thus, K-clustering and L-diversity generalization data are put into data table \( \hat{T} \).

Step 3: During step 2, the data is divided into three levels according to the user privilege. From level 1 to level 3, the user viewing privilege becomes higher. The larger the user privilege is, the smaller the data’s generalization degree is, and the more user information can be known. In order to describe BKC-LDA algorithm more clearly, the algorithm 1 is formed as follows:

\[
\text{Algorithm 1:}
\]

Input: Initial data \( T = \{P_1, P_2, \ldots, P_l \} (P_1 \neq P_2 \neq \ldots \neq P_l \neq P_i) \), \( P_i = \{r_{i1}, \ldots, r_{in} \} \) \( \text{threshold} \) \( K, L, QI = \{N_1, \ldots, N_n, C\} \) is \( QI \); Generalization tree of all classification attributes: \( \{T_{r_1}, T_{r_2}, \ldots, T_{r_n} \} \).

Output: \( \hat{T} \)
1: \( V \leftarrow \Phi \), where \( V \) is cluster set
2: \( EQ \leftarrow \Phi \), where \( EQ \) is the equivalence set
3: \( V + = T \);
4: \( \text{while } |V| > 0 \text{ do} \)
5: \( \text{For each } P_i (1 \leq i \leq l) \text{ in } V \text{ do:} \)
6: \( \text{randomly select } r_j \text{ from } P_i \)
7: \( \text{Get } r_j \text{ by } \arg \max d(r_j, r_i) (1 \leq j \leq l \leq P_i, 1 \leq k \leq l \leq P_i) \)
8: \( \text{Any record } r_m \text{ in } P_i: \)
9: \( \text{Compute } t = \min (d(r_m, r_j), d(r_m, r_i)) \text{?} k, j \)
10: \( \text{then } r_m \in P_j, P_i \text{ is divided into } P_j, P_i \)
11: \( \text{Suppose } P_j \text{ first satisfy } |P_j| \geq K \text{ do:} \)
12: \( \text{if } |S_{r_j}| \geq L \text{ do:} \)
13: \( \hat{T} + = E(P_j), EQ \leftarrow \{EQ, E(P_j)\} \)
14: \( V - = P_j, P_i \leftarrow P_i - P_j \)
15: \( \text{else} \)
16: \( \text{return (9)} \)
17: \( \text{end if} \)
18: \( \text{end if} \)
19: \( \text{if } P_i < K \)
20: \( \hat{T} + = E(P_i), EQ \leftarrow \{EQ, E(P_i)\}, V - = P_i \)
21: \( \text{end if} \)
22: \( \text{end while} \)
23: \( \text{return } \hat{T} \)

4. Simulation results
The above is a data desensitization algorithm for user privilege. For different user privilege, our paper adjusts \( K \) and \( L \) values, changes the number of tuples in the class, expands the generalization degree of ID attributes, and determines 1-3 levels of user privilege according to the different degree of data generalization.
Table 1. Initial data $T$

| ID   | Address                        | age | Monthly electricity consumption (kwh) | Account balance |
|------|--------------------------------|-----|--------------------------------------|-----------------|
| 1    | Street, Haidian district, Beijing | 23  | 170.4                               | 0.12            |
| 2    | Street, Haidian district, Beijing | 34  | 334.8                               | 0.25            |
| 3    | Street, Haidian district, Beijing | 26  | 289.2                               | 0.12            |
| 4    | Street, Haidian district, Beijing | 38  | 270.4                               | 0.08            |
| 5    | Street, Haidian district, Beijing | 46  | 192.6                               | 0.23            |
| 6    | Town, Haidian district, Beijing  | 32  | 128.9                               | 0.53            |
| 7    | Town, Haidian district, Beijing  | 27  | 99.2                                | 0.37            |
| 8    | District, Haidian district, Beijing | 36  | 329.4                               | 0.32            |
| 9    | District, Haidian district, Beijing | 29  | 209.1                               | 0.32            |
| 10   | District, Haidian district, Beijing | 48  | 289.7                               | 0.49            |
| 11   | District, Haidian district, Beijing | 39  | 167.9                               | 0.32            |

Table 2. Clustering results of BKC-LDA algorithm

| ID   | Address                        | age | Monthly electricity consumption (kwh) | Account balance |
|------|--------------------------------|-----|--------------------------------------|-----------------|
| 1    | Street, Haidian district, Beijing | 23[23-46] | 170.4[170-193]                  | 0.12            |
| 2    | Street, Haidian district, Beijing | 46[23-46] | 192.6[170-193]                  | 0.23            |
| 3    | Street, Haidian district, Beijing | 34[26-38] | 334.8[270-335]                  | 0.25            |
| 4    | Street, Haidian district, Beijing | 26[26-38] | 289.2[270-335]                  | 0.12            |
| 5    | Street, Haidian district, Beijing | 38[26-38] | 270.4[270-335]                  | 0.08            |
| 6    | Town, Haidian district, Beijing  | 32[27-32] | 128.9[99-129]                   | 0.53            |
| 7    | Town, Haidian district, Beijing  | 27[27-32] | 99.2[99-129]                    | 0.37            |
| 8    | District, Haidian district, Beijing | 36[29-48] | 329.4[167-330]                  | 0.32            |
| 9    | District, Haidian district, Beijing | 29[29-48] | 209.1[167-330]                  | 0.32            |
| 10   | District, Haidian district, Beijing | 48[29-48] | 289.7[167-330]                  | 0.49            |
| 11   | District, Haidian district, Beijing | 39[29-48] | 167.9[167-330]                  | 0.32            |

Table 1 is the initial data $T$. Table 2 is the result of executing BKC-LDA algorithm on the source data table of user information, where $K = 3$, $L = 2$, corresponding to the third level of user authority. Similarly, the data in Table 1 is implemented with the GAA-CP algorithm, and the data desensitization results are shown in Table 3. It is easy to know that since the GAA-CP algorithm does not carry out initial classification when calculating the distance between, clustering based tuples on the distance formula alone will lead to the expansion of the generalization degree of equivalence classes. As shown in Table 3, $r_1 \sim r_2$ has been generalized into Haidian District of Beijing City, expanding the degree of generalization so as to increase the loss of data information, which is not conducive to data analysis and further mining.
Table 3. Clustering results of GAA-CP algorithm

| ID (name) | Address                      | Age  | Electricity consumption (kwh/Month) | Account balance |
|-----------|------------------------------|------|------------------------------------|-----------------|
| 1 ***     | Haidian district, Beijing    | 23[23-32] | 170.4[99-170] | 0.12            |
| 2 ***     | Haidian district, Beijing    | 32[23-32] | 128.9[99-170] | 0.53            |
| 3 ***     | Haidian district, Beijing    | 27[23-32] | 99.2[99-170]  | 0.37            |
| 4 ***     | Street, Haidian district, Beijing | 34[26-46] | 334.8[192-335] | 0.25            |
| 5 ***     | Street, Haidian district, Beijing | 26[26-46] | 289.2[192-335] | 0.12            |
| 6 ***     | Street, Haidian district, Beijing | 38[26-46] | 270.4[192-335] | 0.08            |
| 7 ***     | Street, Haidian district, Beijing | 46[26-46] | 192.6[192-335] | 0.23            |
| 8 ***     | District, Haidian district, Beijing | 36[29-48] | 329.4[167-330] | 0.32            |
| 9 ***     | District, Haidian district, Beijing | 29[29-48] | 209.1[167-330] | 0.32            |
| 10 ***    | District, Haidian district, Beijing | 48[29-48] | 289.7[167-330] | 0.49            |
| 11 ***    | District, Haidian district, Beijing | 39[29-48] | 167.9[167-330] | 0.32            |

Figure 2 is a comparison of data loss of each record under three levels of privilege when implementing BKC-LDA algorithm and GAA-CP algorithm. By adjusting the values of K and L, we can change the size of the generalization class, thus forming different levels of generalization results. It is easy to know that the user with greater privilege can see more user information. With the increase of privilege level, the BKC-LDA algorithm of the invention always maintains a lower amount of data loss under the condition of protecting the user's privacy. When the data processed by the invention is disclosed, the data has strong availability, can effectively protect the user's privacy, and has significant effect on data desensitization.

![Figure 2. Comparison of data loss](image_url)

5. Conclusions

Aiming at the data security problems faced by big data technology in the development of power industry, according to the data desensitization classification method under the typical application scenarios of power big data, this paper proposes a data desensitization algorithm for privacy protection. Based on K-anonymity and L-diversity algorithm, we propose a binary K-clustering algorithm (BKC-LDA). First of all, the data will be classified initially by one of the classification attributes, and then the data will be clustered according to the proposed equation of the distance between the data tuples, and then the clustering tuples will be generalized. The generalization results need to meet the
conditions of K-anonymity and L-diversity. The simulation results show that the BKC-LDA algorithm can not only fully mine the data value, but also ensure the user’s privacy information security, and effectively reduce the calculation cost.

References

[1] Sweeney, Latanya. k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(05):557-570.

[2] Machanavajjhala A, Gehrke J, Kifer D, et al. l-Diversity: Privacy Beyond k-Anonymity[C]// Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3-8 April 2006, Atlanta, GA, USA. IEEE, 2006.

[3] Li J, Wong C W, Fu W C, et al. Achieving k-Anonymity by Clustering in Attribute Hierarchical Structures[C]// Proceedings of the 8th international conference on Data Warehousing and Knowledge Discovery, Springer Berlin Heidelberg, 2006.

[4] Aggarwal, G., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D., Zhu, A.: Achieving anonymity via clustering. In: ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 153–162 (2006)

[5] Jiang huowen, Zeng guosun, Ma Haiying. Greedy clustering anonymous method for privacy protection of table data publishing [J]. Journal of software, 2017,28 (02): 341-351

[6] Zheng W., Wang Z., Lv T., Ma Y., Jia C. (2018) K-Anonymity Algorithm Based on Improved Clustering. In: Vaidya J., Li J. (eds) Algorithms and Architectures for Parallel Processing. ICA3PP 2018. Lecture Notes in Computer Science, vol 11335. Springer, Cham

[7] Widodo, E. K. Budiardjo, W. C. Wibowo and H. T. Y. Achsan, "An Approach for Distributing Sensitive Values in k-Anonymity," 2019 International Workshop on Big Data and Information Security (IWBIS), Bali, Indonesia, 2019, pp. 109-114.

[8] Huang Lingyu, Cong Zhongfang, Zhao Cheng, et al. Important data protection in power big data application construction [C]// 2018 smart grid information construction seminar.

[9] Liu Minghui, Zhang Ni, Zhang Yunyong, et al. Research on sensitive data protection technology in cloud environment [J]. Telecommunication science, 2014 (11): 2-8.