Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 2, pp. 673–681, DOI 10.12694/scpe.v26i2.3958

HYBRID DATA PUBLISHING BASED ON DIFFERENTIAL PRIVACY

TAO WANG*, KAINING SUN †, RUI YIN‡, TENG ZHANG§, AND LONGJUN ZHANG ¶

Abstract. The advent of the information and intelligence era has led to explosive growth of data. The author proposes a hybrid data model based on differential privacy. The main content of this model is based on the study of differential privacy, processing the data through a noise mechanism, using the calculation of tuple attribute differences and noise addition, and finally constructing a mixed data model based on differential privacy through experiments. The experimental results indicate that: as the value of k increases, the clustering results tend to be optimal, verifying that clustering the original data can reduce noise addition. However, ICMD-DP anonymizes the original dataset, resulting in much higher information loss than DCKPDP and prototype algorithms. A mixed data model based on differential privacy enables better clustering performance of the original dataset, thereby utilizing differential privacy to better protect the data.

Key words: Differential privacy, Mixed data, Information, Clustering

1. Introduction. In the era of big data, the release and utilization of data are key to promoting knowledge economy and social progress. Relevant research institutions will utilize these data resources for mining and analysis, in order to provide better services to the public. However, while providing significant benefits, publishing personal data to the public poses a significant threat to user privacy. In order to ensure user privacy and security, it is necessary to protect them. However, how to ensure that the published data is both usable and does not leak the privacy information contained in the data has become a major challenge in research on data publishing privacy protection.

The explosive growth of data, the release of which can provide scientific decision-making, predict market trends, and promote social development, truly promoting the flow of data value [1]. However, these data often contain a large amount of sensitive information, and direct publication will inevitably lead to user privacy leakage. Therefore, how to protect sensitive user information and maximize the availability of published data during the data publishing process has become an urgent problem to be solved. In recent years, some methods have been proposed to address privacy protection issues in data publishing, mainly based on data anonymity publishing methods and data distortion publishing methods. Although using such methods can to some extent protect sensitive information in the data, they require the assumption that the attacker does not have background knowledge, and therefore cannot resist background knowledge attacks and combination attacks. With the rapid development of the Internet, big data analysis technology and cloud computing, individuals, enterprises and institutions will generate a continuous stream of massive data every day. These massive amounts of data, when applied to research, can improve people's lives, promote development, and bring great convenience to their lives. However, while enjoying convenience, people are also facing the problem of personal privacy being violated. How to protect the privacy and security of user data while meeting the research needs of providing reasonable data is one of the hot topics of discussion and research in today's era (Figure 1.1).

*State Grid Xinjiang Electric Power Co., Ltd. Information and Communication Company, Urumqi, Xinjiang, 832000, China. Xinjiang Energy Internet Big Data Laboratory, Urumqi, Xinjiang, 832000, China (Corresponding author, TaoWang65@163.com)

†Xinjiang Energy Internet Big Data Laboratory,Urumqi, Xinjiang,832000, China. State Grid Xinjiang Electric Power Co., Ltd, Urumqi, Xinjiang, 832000, China (KainingSun6@126.com)

‡State Grid Xinjiang Electric Power Co., Ltd. Information and Communication Company, Urumqi, Xinjiang, 832000, China. Xinjiang Energy Internet Big Data Laboratory, Urumqi, Xinjiang, 832000, China (RuiYin17@163.com)

§State Grid Xinjiang Electric Power Co., Ltd. Information and Communication Company, Urumqi, Xinjiang, 832000, China. Xinjiang Energy Internet Big Data Laboratory, Urumqi, Xinjiang, 832000, China (TengZhang3@126.com)

¶State Grid Xinjiang Electric Power Co., Ltd. Information and Communication Company, Urumqi, Xinjiang, 832000, China. Xinjiang Energy Internet Big Data Laboratory, Urumqi, Xinjiang, 832000, China (LongjunZhang7@163.com)

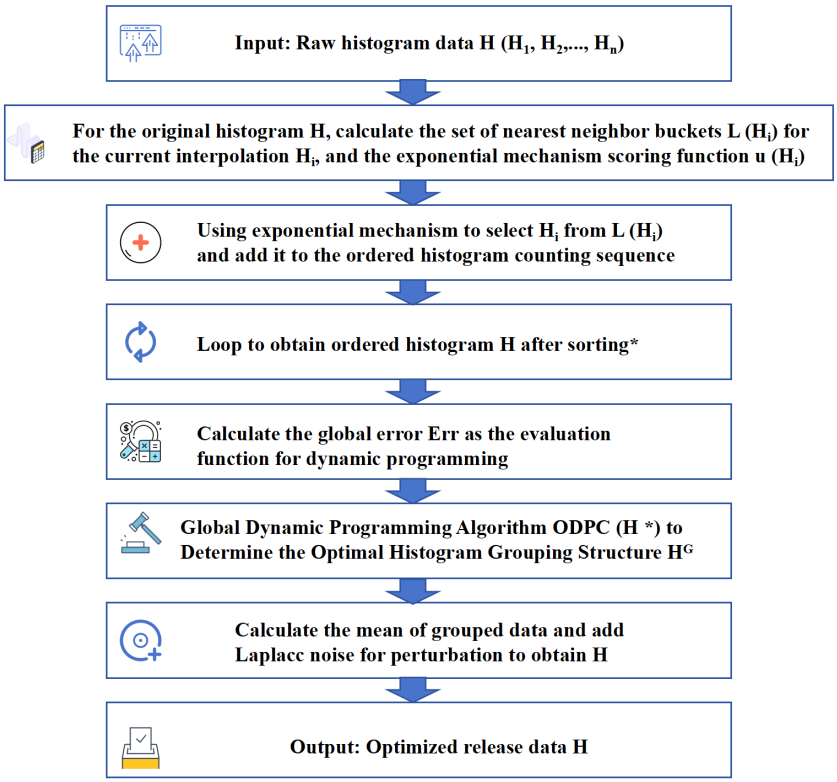

Fig. 1.1: Mixed data based on differential privacy

2. Literature Review. With the rapid development of mobile communication Internet and cloud computing technology, smart phones, wearable devices, sensors and other IP mobile devices with GPS chips can submit users' location information to location aware applications, providing consumers with convenient personalized services. But these mobile terminals can expose users' location tracks and personal information on the Internet in real time, which poses a great threat to people's privacy and security. Therefore, how to obtain valuable information from massive data while preventing privacy breaches is a current research hotspot in the field of data mining. With the rapid development of the Internet era, people find that deep learning models have great advantages in image processing, speech recognition and other fields. Deep neural networks have appeared in all aspects of human life. Among them, generative adversarial networks is one of the most promising deep learning models in the field of unsupervised learning. It consists of two parts: a generator and a discriminator. The generator generates "fake" data, which is then dynamically adjusted with the discriminator to ultimately generate the data that users need. But generating adversarial network models not only brings convenience to users, but also provides attackers with an opportunity to steal sensitive user data. For example, when a cancer diagnosis model is attacked by member inference, attackers can analyze the data information in the deep learning model to infer whether a patient has a certain type of cancer, and the user's privacy information is likely to be stolen by the attacker. Currently, there have been some research results on differential privacy data publishing methods, but these methods all have certain problems. Yang, J. proposed a new CMFD algorithm with the following workflow. Firstly, use the keypoint extraction method with the lowest contrast threshold to extract more keypoints from the input image. Secondly, a new technique, gradient hash matching, uses a hash table to quickly and effectively find similar pairs of key points, where the hash value is calculated using the gradient of the key points. Subsequently, a new method called simplified clustering filtering utilizes the density pattern of key points in the copy move region to remove mismatched key point pairs [2]. Huang, Z. et al.

believe that the accelerated mode matrix splitting method and the recently proposed generalized accelerated mode matrix splitting method are special cases.

Compared with existing methods, this method can use more information in each iteration, thereby improving computational efficiency. And the convergence of the method was studied, and the convergence of the method was proved under certain assumptions [3]. Zhang, P. proposed a data level fusion model that involves the integration of multiple information sources and unsupervised attribute selection of fused data [4]. Panfeng Zhang believes that excessive gradient perturbation noise in deep model differential privacy protection can lead to decreased usability, and proposes a differential privacy deep learning model based on particle swarm optimization algorithm. According to the particle swarm optimization strategy, the position of particles is mapped to network parameters to search for individual and global historical optimal positions. After perturbing the gradient obtained from the global optimal particle position, the model is re trained [5]. Liang, W proposed a differential privacy data publishing method based on DBSCAN clustering, but this method is also suitable for publishing numerical attribute data [6].

In response to the above issues, in order to ensure that the published data does not affect personal privacy and security, the author proposes a mixed data model research based on differential privacy. The differential privacy model, as a well-known privacy protection model, can provide privacy assurance by adding a certain amount of noise to the data query or analysis results without making any assumptions about the attacker's background knowledge. In a non interactive framework, data managers can publish datasets processed using differential privacy protection technology for researchers to mine and analyze.

The author reviews the quality related data of key links in electric energy meters and studies the method of extracting quality impact features; Compare various big data analysis technologies and establish a quality analysis model for smart energy meters; Use this model to predict and analyze potential quality hazards of smart energy meters, and conduct on-site verification. Continuously optimize the model based on the verification results.

3. Research Methods.

3.1. Differential privacy protection.

3.1.1. Definition of Differential Privacy. Differential privacy was initially used to limit the disclosure risk when returning query answers on a database, but its application in interactive scenarios strictly limits data analysis, because it only allows a limited number of queries to be answered, it promotes privacy protection research for data publishing in non cross five scenario scenarios.

Differential privacy is a model that provides strong privacy protection. In a non interactive framework, data managers can publish datasets processed using differential privacy protection techniques for researchers to conduct mining and analysis.

The main method for publishing differential privacy datasets in non interactive scenarios is based on histogram publishing. However, as the number of attributes increases, histogram based methods have serious limitations: For fixed attribute granularity, the number of histogram intervals increases exponentially with the number of attributes, which has a serious impact on computational cost and accuracy [7]. In addition, the histogram publishing method only provides approximate counts of partitioned data and cannot provide data details, thus limiting the utility of data analysis. Therefore, this limitation can be overcome by generating a universal dataset that satisfies differential privacy. The simplest method is to collect a set of query results that satisfy differential privacy, which requires querying each individual record in the original dataset. However, such queries require too much noise to meet the requirements of differential privacy, making it impossible for differential privacy datasets to maintain availability. The availability of differential privacy protection data can be improved by reducing query sensitivity and reducing the amount of noise added [8].

Differential privacy protection technology perturbs data by adding quantitative noise to the query results, ensuring that the insertion, modification, and deletion of records in any dataset will not affect the query results, thereby achieving privacy protection [9].

Differential privacy has a random algorithm K, as well as any adjacent datasets D1 and D2. If algorithm K satisfies differential privacy, it can be expressed as formula 3.1:

$$Pr[K(D_1 \in S)] \leqslant exp(\epsilon) Pr[K(D_2) \in S] \tag{3.1}$$

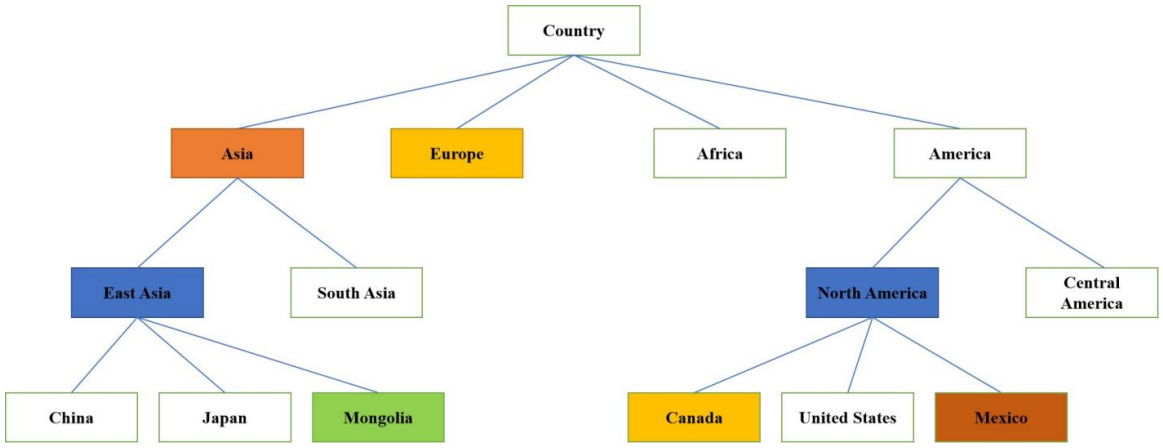

Fig. 3.1: Country Attribute Generalization Tree

3.1.2. Noise mechanism. Sensitivity refers to the maximum amount of change in the query result when the dataset changes and only one record changes. Differential privacy typically perturbs the return value of the query function with noise, and the magnitude of the added noise is closely related to the global sensitivity of the query function [10,11].

In practical applications, commonly used noise mechanisms include Laplace mechanism and exponential mechanism. The amount of noise can affect data security and availability, and is closely related to global sensitivity. Global sensitivity can be expressed as formula 3.2:

$$\Delta f = max_{D_1,D_2}||f(D_1) - f(D_2)|| \tag{3.2}$$

3.2. Differential Privacy Hybrid Data Publishing Algorithm.

3.2.1. Calculation of Tuple Attribute Differences. In the context of privacy data dissemination, data dissemination can be viewed as the collection of answers to continuous queries for each record in the dataset. The author proposes a hybrid data publishing algorithm based on differential privacy: using the k-prototype clustering algorithm, the initial class center is randomly selected first . Cluster the dataset based on an improved method for calculating the difference in tuple attributes, and then calculate the cluster center for each numerical attribute in each cluster based on the best clustering result. Generate a set of attribute values for categorical attributes. Next, traverse each data record and determine its clustering category. Replace numerical attributes with cluster center values and use Laplace mechanism to add noise. Use exponential mechanism to select categorical attributes. Finally, generate a differential privacy dataset. Due to the sensitivity of the query function being differentiated into k records in each set of data, it can reduce the amount of noise added and improve data availability.

Most existing data tables are mixed data Tables, which means that the data attributes in the tables are divided into numerical and categorical types. There are different methods for calculating attribute differences for data with different types of attributes [12].

Unlike numerical attributes, categorical attributes require the establishment of a generalized hierarchical tree to calculate attribute differences. Each subtyping attribute needs to establish a generalized hierarchical tree [13]. Figure 3.1 shows the generalized hierarchical tree of the Country attribute, with leaf nodes representing the values of each attribute on the Country attribute.

3.2.2. Noise addition method. For numerical attributes, the Laplace mechanism is used to add noise to the cluster center, which can be expressed as formula 3.3:

$$Centroid'(C_m(A_i^q)) = Centroid(C_m(A_i^q)) + Lap(\frac{\Delta f}{\epsilon}) \tag{3.3}$$

Table 3.1: adult Dataset

Attribute	Attribute type	Number of attribute values	Attribute	Attribute type	Number of attribute values
Age	Numerical type	75	Education level	Classification	15
Weekly working hours	Numericaltype	88	Gender	Classification	3
Education duration	Numerical type	18	Occupation	Classification	15
Marital status	Classification	8	Original nationality	Classification	42

Unlike numerical attributes, categorical attributes obtain values from a limited set of categories. Since adding Laplacian noise to cluster centers is meaningless, another method of obtaining differential privacy output is to select cluster centers in a probabilistic manner, which can be achieved through an exponential mechanism [14]. This mechanism selects the closest optimal center point based on input data, differential privacy parameters, and quality standards. In this case, the quality standard is the probability of each subtype attribute value appearing.

The differential privacy static data publishing algorithm based on k-prototype clustering is mainly used to publish mixed datasets containing numerical and subtype attributes. The algorithm is divided into a clustering and grouping stage and a data publishing stage. In the first stage, an improved k-prototype clustering algorithm is used to cluster and partition the data. In the second stage, differential privacy data publishing is achieved. If it is a numerical attribute, replace it with the cluster center value, and then use the Laplace mechanism to independently add noise to each attribute value; If it is a categorical attribute, the output attribute value is selected using an exponential mechanism based on the selection criteria of the central candidate of the attribute in the cluster attribute value set to which it belongs [15].

3.2.3. Algorithm Description. The DCKPDP algorithm is designed for publishing datasets containing mixed attributes, and the process mainly includes two parts: a) Cluster the original dataset using an improved k-prototype algorithm; b) Using differential privacy technology to add noise to the clustered dataset and output a dataset that satisfies differential privacy.

3.3. Experimental research.

3.3.1. Experimental Environment and Datasets. The experimental dataset used the adult dataset from the UCI machine learning database, which contains a total of 48943 data records. After deleting records with missing attributes, a total of 31257 records were obtained, due to the author's publication on a mixed attribute dataset, three numerical attributes and five categorical attributes were selected from the adult dataset as experimental attributes. The adult dataset is shown in Table 3.1.

3.3.2. Algorithm performance evaluation criteria. The author improved the k-prototype clustering algorithm to achieve better clustering performance on the original dataset, thereby incorporating less noise and improving data availability when using differential privacy protection. Therefore, the main purpose of this experiment is to demonstrate that the algorithm proposed in the article can improve data availability while ensuring a lower risk of data leakage [16].

Regarding the DCKPDP algorithm, adjust the privacy budgetϵ values are set to {0.02, 0.2, 2, 6}, and the number of attributes q is taken as 3 and 7 for comparative experiments, among them, when q is set to 3, two numerical attributes (age, age, weekly working hours) and two subtype attributes (original nationality, education level) are taken, and the information loss SSE caused by the DCKPDP algorithm is shown in Figure 3.2 (a) (b).

As shown in Figure 3.2, when q=3, the value of SSE is much smaller than when q=7. This is because as the number of attributes increases, more and more noise is added to the original data, resulting in more information loss in the data and a larger corresponding SSE value. Whenϵ taking 0.02, although SSE shows a downward trend, the change is not significant because a large amount of noise is added when the privacy budget is too small. Using the author's improved k-prototype clustering algorithm to process the original dataset resulted in very low data availability; Whenϵ taking 0.2, the change in SSE is most significant; Whenϵ taking 2 and 6,

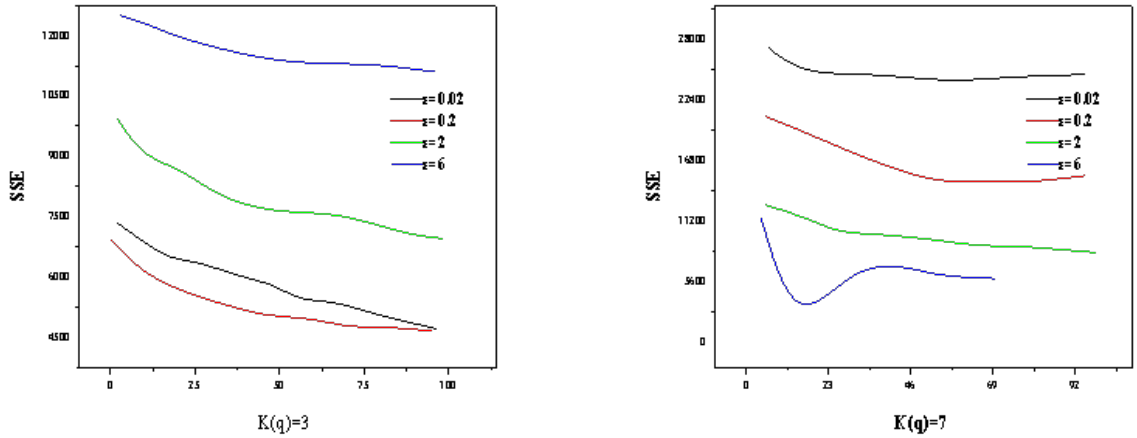

Fig. 3.2: Changes in SSE when q takes different values

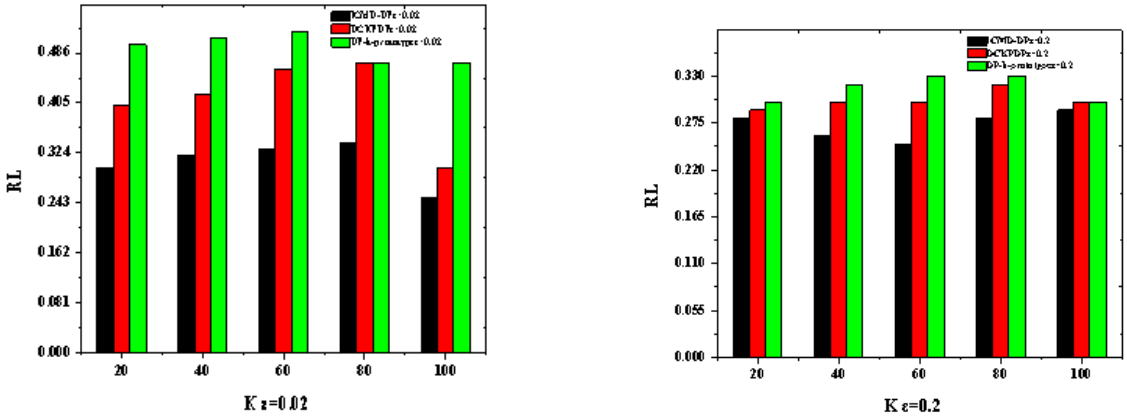

Fig. 3.3: When ϵ taking 0.002 and 0.2, the variation of RL with k value

the SSE values of the two are lower and the difference is small, because whenϵ taking a larger value, the added noise is small and has little impact on the SSE value of the data.

As the value of k increases, the number of clustering clusters increases. Data records with lower dissimilarity are divided into the same cluster, and the clustering effect is close to optimal. The less noise is added, so the overall trend of SSE is decreasing, which also proves the feasibility of the algorithm [17].

Compare the RL values of DCKPDP, ICMD-DP, and DP k-prototype algorithms, as shown in Figures 3.3 (a) (b) and 3.4 (a) (b).

From Figures 3.3 and 3.4, it can be seen that the privacy budget whenϵ taking different values, the RL value of ICMD-DP is lower compared to DCKPDP and DP k-prototype, indicating a lower risk of privacy leakage. This is because ICMD-DP anonymizes the dataset and applies differential privacy protection to the anonymized dataset, which will inevitably provide stronger protection for the data. The RL values of DCKPDP and DP-k-prototype are not significantly different, with a difference of about 3%. However, from the experimental results,

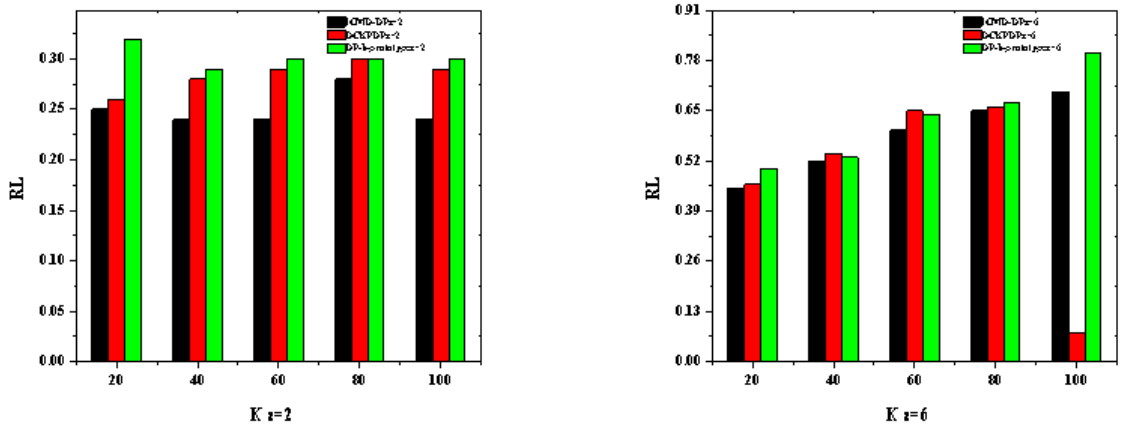

Fig. 3.4: When ϵ taking 2 and 6, the variation of RL with k value

even if ICMD-DP uses anonymization and differential privacy to process the original dataset, the difference in RL values between them and DCKPDP is still controlled within 7%. Therefore, from the perspective of data leakage risk alone, the dataset processed by DCKPDP can still meet the requirement of ensuring data privacy is not leaked. When privacy budgetϵ value of RL is 0.02, and the value of RL is the smallest, which means the risk of privacy leakage is minimized. This is because a large amount of noise is added to the data, and the data availability is also the lowest at this time; When privacy budget ϵ values are 0.2 and 2, it can be seen that the RL values of both DCKPDP and DP k-prototype are not significantly different, with a difference of about 4% and a difference of 7% compared to ICMD-DP; When privacy budgetϵ value of k is 6, as the value of k increases, the RL value increases significantly, and the risk of data leakage also increases. Therefore, considering the DCKPDP comprehensively whenϵ is taken 2, the algorithm performance is optimal [18].

4. Result analysis. The experimental settings for q are 3 and 7, when ϵ value is set to 2, the experimental results are shown in Figure 4.1(a) (b) to compare the changes in information loss of DCKPDP, ICMD-DP, DP k-prototype algorithm, and standard difference privacy algorithm on the adult dataset.

As shown in Figure 4.1, when ϵ taking 2, as the value of k increases, the clustering results tend to be optimal. The information loss of DCKPDP, ICMD-DP, and DP k-prototype algorithms gradually decreases, and the information loss is much lower than that of the standard difference privacy algorithm. This verifies that clustering the original data can reduce noise addition, however, ICMD-DP anonymizes the original dataset, resulting in much higher information loss than DCKPDP and DP k-prototype algorithms. The clustering algorithm proposed by the author adaptively selects the initial center point and improves the dissimilarity calculation formula compared to the DP-k-prototype algorithm. The clustering effect is improved, and the information loss caused by adding noise to it through differential privacy is also reduced, resulting in improved data availability [19,20]. Therefore, from the experimental results, it can be concluded that compared to ICMD-DP and DP kprotype algorithms, DCKPDP can reduce information loss and significantly improve data availability while ensuring a lower risk of information leakage, proving the superiority of the DCKPDP algorithm.

5. Conclusion. The author studied the privacy protection issue of mixed attribute data publishing and proposed a new data publishing protection method. In response to the research question, the author first improved the traditional k-prototype clustering algorithm's dissimilarity calculation method and proposed a method that can adaptively select the initial clustering center point, improving the accuracy and stability of clustering. Finally, differential privacy was applied to the classified dataset to ensure data privacy was not leaked. Through experimental verification, the DCKPDP algorithm can improve the availability of data while ensuring a lower risk of data leakage compared to similar algorithms. However, the author used a lower data

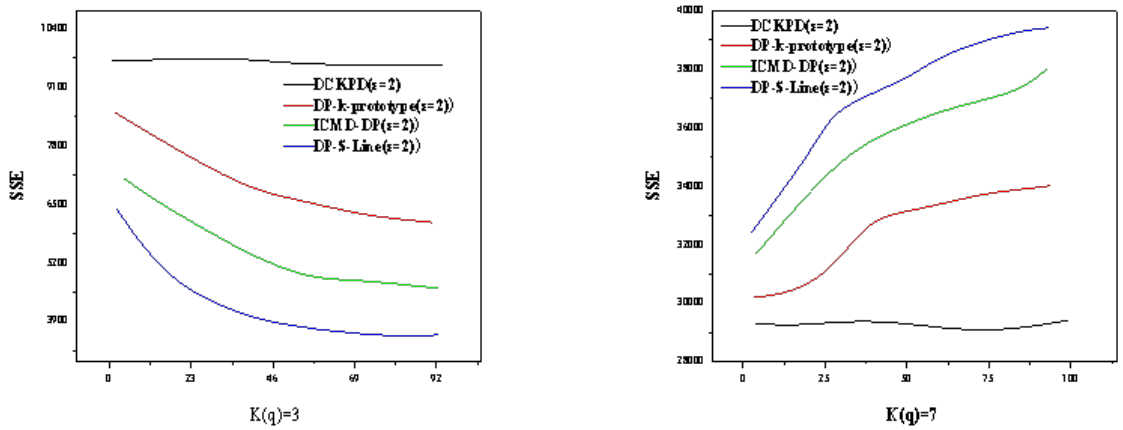

Fig. 4.1: When ϵ taking 2 and q with different values, the variation of SSE

dimension during the experiment, which may lead to efficiency issues when publishing high-dimensional data; And the author has adopted the principle of equal distribution for the allocation of privacy budget, which may result in waste of privacy budget and loss of data information, reducing the utility of data.

REFERENCES

[1] Li, L., Guo, L., Zhong, D., Huang, X., & Zhang, J. . (2023). Reliability analysis of deep-water explosion test vessel based on fuzzy interval. International Journal of Pattern Recognition and Artificial Intelligence, 37(04).

[2] Yang, J., Liang, Z., Li, J., Gan, Y., & Zhong, J. . (2023). A novel copy–move forgery detection algorithm via gradient-hash matching and simplified cluster-based filtering. International journal of pattern recognition and artificial intelligence(6), 37.

[3] Huang, Z., & Cui, J. . (2024). Accelerated relaxation two-sweep modulus-based matrix splitting iteration method for linear complementarity problems. International Journal of Computational Methods, 21(02).

[4] Zhang, P., Li, T., Yuan, Z., Luo, C., Wang, G., & Liu, J., et al. (2022). A data-level fusion model for unsupervised attribute selection in multi-source homogeneous data. Information Fusion(80-), 80.

[5] Panfeng ZHANG, Danhua WU, & Minggang DONG. (2023). Differential privacy deep learning model based on particle swarm optimization. Computer Engineering, 49(9), 144-157.

[6] Liang, W., Chenyang, H., Jiangning, S., & Jianhua, Y. . (2024). Ctec: a cross-tabulation ensemble clustering approach for single-cell rna sequencing data analysis. Bioinformatics(4), 4.

[7] Ju, H., Ding, W., & Gu, P. Y. X. . (2023). Bi-directional adaptive neighborhood rough sets based attribute subset selection. International journal of approximate reasoning, 160(9), 1.1-1.18.

[8] (2022). Investigators from university of texas austin release new data on machine learning (3d microseismic monitoring using machine learning). Robotics & Machine Learning Daily News(4), 6-7.

[9] Bristow, N. R., Best, J., Wiggs, G. F. S., Nield, J. M., Baddock, M. C., & Delorme, P., et al. (2022). Topographic perturbation of turbulent boundary layers by low-angle, early-stage aeolian dunes. Earth Surface Processes and Landforms: The journal of the British Geomorphological Research Group(6), 47.

[10] Jeon-Young Kang, Michels, A., Crooks, A., Aldstadt, J., & Wang, S. . (2022). An integrated framework of global sensitivity analysis and calibration for spatially explicit agent-based models. Transactions in GIS: TG(1), 26.

[11] Akshay, V., Sunil, K., Raj, G. P., Tarique, R., & Arvind, K. . (2023). Enhanced cost and sub-epoch based stable energy-efficient clustering algorithm for heterogeneous wireless sensor networks. Wireless personal communications: An Internaional Journal(4), 131.

[12] Zheng, T., QiGe, Xiong, F., Li, G., Xue, Y., & Deng, X. . (2023). Study of the flexural performance and a novel calculation formula for the degree of composite action for precastc oncrete sandwich panels. The structural design of tall and special buildings(18), 32.

[13] Kim, H., Strang, A., & Sanz-Alonso, D. . (2023). Hierarchical ensemble kalman methods with sparsity-promoting generalized gamma hyperpriors. Foundations of Data Science, 5(3), 366-388.

[14] Viviana Elizabeth Zárate-Mirón, & Serrano, R. M. . (2023). The impact of smart specialization strategies on sub-cluster efficiency: simulation exercise for the case of mexico. Competitiveness Review: An International Business Journal , 33(2),

364-394.
[15] Sun, Z., & Zhao, J. . (2023). Comprehensive performance evaluation of landing gear retraction mechanism in a certain model of aircraft based on rpca method. Journal of Circuits, Systems and Computers, 32(11).
[16] Benrhouma, O., Alzahrani, A., Alkhodre, A., Namoun, A., & Bhat, W. A. . (2022). To sell, or not to sell: social media data-breach in second-hand android devices. Information & computer sceurity(1), 30.
[17] Chumnangoon, P., Chiralaksanakul, A., & Chintakananda, A. . (2023). How closeness matters: the role of geographical proximity in social capital development and knowledge sharing in smes. Competitiveness Review: An International Business Journal , 33(2), 280-301.
[18] Zhou, T., Hu, Z., Su, Q., & Xiong, W. . (2023). A clustering differential evolution algorithm with neighborhood-based dual mutation operator for multimodal multiobjective optimization. Expert Systems with Applications, 216, 119438-.
[19] Wang, W., Li, G., Wang, Y., Wu, F., Zhang, W., & Li, L. . (2022). Clearing-based multimodal multi-objective evolutionary optimization with layer-to-layer strategy. Swarm and Evolutionary Computation(68-), 68.
[20] Kaho, T., Watanabe, S., & Sakakibara, K. . (2022). Multi-objective branch and bound based on decomposition. IEEJ Transactions on Electronics, Information and Systems, 142(3), 373-381.

Edited by: Bradha Madhavan
Special issue on: High-performance Computing Algorithms for Material Sciences
Received: May 17, 2024
Accepted: Jun 15, 2024