A Spectral Clustering Algorithm Based on Differential Privacy Preservation

Yuyang Cui[1], Huaming Wu[2], Yongting Zhang[1], Yonggang Gao[1] and Xiang Wu[1*]

[1] School of Medical Information and Engineering, Xuzhou Medical University, Xuzhou 221000, Jiangsu, China
[2] Center for Applied Mathematics, Tianjin University, Tianjin 300072, China
wuxiang@xzhmu.edu.cn

Abstract. Spectral clustering is a widely used clustering algorithm based on the advantages of simple implementation, small computational cost, and good adaptability to arbitrarily shaped data sets. However, due to the lack of data protection mechanism in spectral clustering algorithm and the fact that the processed data often contains a large amount of sensitive user information, thus an existing risk of privacy leakage. To address this potential risk, a spectral clustering algorithm based on differential privacy protection is proposed in this paper, which uses the Laplace mechanism to add noise to the input data perturbing the original data information, and then perform spectral clustering, so as to achieve the purpose of privacy protection. Experiments show that the algorithm has both stability and usability, can correctly complete the clustering task with a small loss of accuracy, and can prevent reconstruction attacks, greatly reduce the risk of sensitive information leakage, and effectively protect the model and the original data.

Keywords: Differential privacy, Spectral clustering, Privacy preservation

1 Introduction

Machine learning clustering algorithms have made disruptive breakthroughs in recent years and are widely used in computer vision, data mining and remote sensing mapping. However, most machine learning clustering algorithms are designed without considering the security and privacy issues of data and models[1]. To protect the security of private information, many industry scholars have conducted extensive research.

There are three broad directions of privacy preservation currently combined with clustering algorithms: data transformation, data anonymization, and data perturbation represented by differential privacy. Among them, Oliverira et al. [2] proposed a new spatial data transformation method RT (Rotation-based Transformation) inspired by the rotational changes of geometry in space, whose advantage lies in being able to hide the original information while maintaining the validity of the data attributes

[*] Corresponding author: wuxiang@xzhmu.edu.cn

before and after the transformation. However, the computation is complex with higher dimensionality, and it is difficult to resist consistency attacks. Nayahi [3] proposed an anonymous data algorithm by distributing anonymous data in Hadoop Distributed File System (HDFS) based on the principle of clustering and resilient similarity attacks and probabilistic inference attacks, but it is difficult to resist emerging combinatorial attacks and foreground knowledge attacks. Current research on differential privacy combined with clustering algorithms focuses on the k-means algorithm. Blum et al [4] were the first to combine differential privacy techniques with k-means clustering algorithm in 2005, proposed the DPk-means algorithm, which pioneered the research of data perturbation represented by differential privacy. Dwork [5] proposed a method of privacy budget allocation and sensitivity calculation in view of the DPk-eams algorithm's shortcomings that the large sensitivity of query function and the privacy budget allocation method are not given. FU et al [6] improved the usability of clustering results by dynamically assigning the selection of initial centroids and iteratively updating the privacy budget during the operation of the algorithm, but they don't consider the effect of isolated points in the dataset on the clustering effect. NI et al [7] proposed the DP-KCCM algorithm to offset the effect of differential privacy techniques by merging adjacent clusters to join of noise on the clustering results. X. W et al. [8] proposed a DP-CFMF algorithm based on differential privacy protection, which guarantees the privacy of DNA data with high recognition accuracy and high utility. W. Wu et al. [9] proposed a DP-DBS can algorithm that realizes differential privacy protection by adding laplace noise, which was experimentally shown to be able to protect privacy while being both usable. In addition, W. Wu et al. [10] integrated various privacy protection schemes and innovatively designed a data-sharing platform that can guarantee data security, practicing the previous research results.

Compared with the traditional k-means algorithm, the spectral clustering algorithm is applicable to arbitrarily shaped data sets, and does not require prior assumptions about the probability distribution of the data, it is fast in computation and simple. In recent years, it has been widely used in computer vision, data mining, image processing and natural language processing[11-14]. For example, Yang Fan et al. [15] used the spectral clustering algorithm to mine the data of chemical reagents in stock of China Institute of Petrochemical Sciences. Guo Lei et al.[16] applied the spectral clustering algorithm to cognitive diagnostic assessment (CDA), and explored the possible effects of introducing the spectral clustering algorithm on CDA in terms of attribute hierarchy, number of attributes, sample size and failure rate, and achieved good results in specific experiments.

However, the spectral clustering algorithm lacks privacy protection mechanism and there is a risk of privacy leakage. How to protect the privacy of the spectral clustering algorithm while ensuring the clustering accuracy becomes an urgent problem. While the relevant research on spectral clustering algorithms oriented to differential privacy protection are proposed. Zheng et al. [17] achieved differential privacy protection by adding laplace noise to perturb the objective function to hide the true weight values and the clustering results were more accurate compared with the spectral clustering algorithm without differential privacy protection, but the specific privacy budget allocation method was not described. The internal scale parameter and the number of clusters of the spectral clustering algorithm were optimized to further improve the accuracy of clustering[18]. The DP-CSC algorithm is proposed based on the improved spectral clustering algorithm CCL-S, and the noise is added to the compressed laplacian matrix using Wishart mechanism [19].

This paper will propose a new spectral clustering algorithm based on differential privacy to achieve privacy preservation, by adding noise conforming to the Laplace distribution to the input data to hiding the original data.

2 Theoretical foundation

2.1 Differential Privacy

Differential privacy (DP) is a privacy-preserving model with rigorous mathematical proof, first proposed by Microsoft's DWork team [20].

Definition 1 Differential privacy: assume that there exists a random function A such that A in any two neighbor data sets $Q,\ Q^{'}$(i.e.$||Q-Q^{'}||_1 \leq 1$) to obtain any identical set of outputs B with probability satisfying

$$\Pr[A(Q)\in B]\leq e^{\varepsilon}\Pr\left[A\left(Q'\right)\in B\right] \tag{1}$$

Then the random function is said to A satisfies $\varepsilon-differential\,privacy$, abbreviated as$\varepsilon-DP$. Where the neighbor datasets are the two datasets that differ by only 1 record and$||\ \ ||_1$ is theL_1 paradigm, Pr [ ] is the probability of occurrence of an event, andε is the privacy budget, the size of its value represents the degree of privacy protection, the smaller the value means the better the privacy protection.

Differential privacy is achieved mainly by adding noise to the data to perturb the data so that the output results are randomly different from the real results each time. The common mechanisms for adding noise are the Laplace mechanism for continuous data, the Exponential mechanism for discrete data, such as race, education, etc. and the Gaussian mechanism for image data. Since the data to be processed in this paper are of continuous type, the Laplace mechanism is adopted.

The Laplacian mechanism is implemented by adding random noise obeying the Laplacian distribution to the exact query result $\varepsilon-differential\,privacy$ protection. Let the location parameter be 0 and the scale parameter be b the Laplace distribution of $Lap(b)$, then its probability density function is

$$p(x)=\frac{1}{2b}e^{\left(-\frac{|x|}{b}\right)} \tag{2}$$

where e is the natural logarithm.

Its probability density function, as shown in Fig.1.

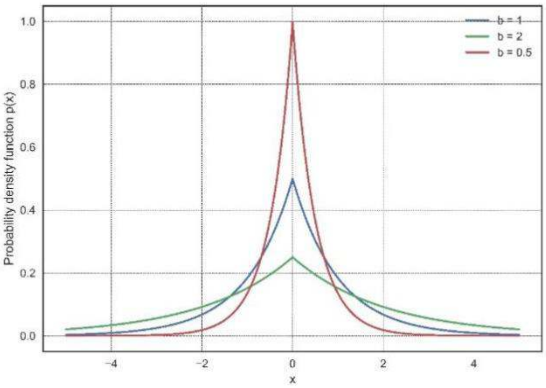


Fig. 1. Laplace probability density function

For the Laplace mechanism, there are the following definitions.

Definition 2 Laplace mechanism: given a data set D, there is a function $f: D -> Rd$, whose sensitivity is Δf, then the randomized algorithm $M(D) = f(D) + Y$ provides$\varepsilon - differential\,privacy$ protection, where $Y \sim Lap(\Delta f / \varepsilon)$ is the random noise, which obeys the scale parameter $\Delta f / \varepsilon$ of the Laplace distribution, and the random noise size depends mainly on the sensitivity of the functionΔf [21].

Definition 3 Sensitivity$\boldsymbol{\Delta f}$: With the function $f: D \rightarrow R\,d$, the input is the dataset and the output is d dimensional real vector, for two neighbor datasets D, D', the sensitivities are

$$\Delta f = \max D, D' f(D) - f\left(D'\right)_1 \tag{3}$$

Sensitivity measures the maximum change to the results caused by the deletion of anyone record from the dataset and is an important parameter in determining the amount of noise introduced into the data.

Spectral clustering algorithm

Compared with the traditional k-means algorithm and EM algorithm, the spectral clustering algorithm is applicable to data sets of arbitrary shapes that can do without prior assumptions about the probability distribution of the data, with fast computational speed and simple algorithmic ideas is simple and easy to implement.

The basic principle of the spectral clustering algorithm is that the sample data is considered as a vertex in the undirected graph, and then the similarity W_{ij} between the vertices is calculated based on the similarity function as the weight between the vertices, and finally the graph is divided according to different partitioning criteria to maximize the similarity within each subgraph and minimize the similarity between the subgraphs[22], each subgraph is equivalent to a cluster.

For a graph G to be divided, denote by A, B two subgraphs (where $A \cup B = V, A \cap B = \emptyset, V$ denotes the set of all vertices in the graph G), u, v denote the points in the graph A, B respectively, $w(u, v)$ denotes the similarity between the points u, v, and the division criteria are mainly as follows.

(1) Minimum cut [23]

$$\min\{Cut(A,B)=\sum_{u\in A,v\in B} w(u,v)\} \tag{4}$$

Where $Cut(A,B)$ is the cost function when dividing graph G into subgraphs A,B, which represents the sum of similarity between points u,v inside graphs A,B. The minimum cut set criterion divides graph G by minimizing the cost function, which proves to be effective for some image datasets in practice, but when the number of subgraphs divided exceeds 2, it is easy to cluster isolated points into a separate class. To address this situation, Shi and Malik proposed the canonical cut-set criterion and Hagen and Kahng proposed the proportional cut-set criterion, both of which solve the problem well.

(2) Normalized Cut [22]

$$\begin{gathered}\min\{NCut(A,B)=\frac{Cut(A,B)}{Vol(A,V)}+\frac{Cut(A,B)}{Vol(B,V)}\}\\ \text{among it,}Vol(A,V)=\sum_{i\in A,t\in V} W_{it}\end{gathered} \tag{5}$$

The minimization $NCut$ function is called the canonical cut-set criterion. This criterion measures not only the degree of similarity between samples within a class, but also the degree of dissimilarity between samples between classes

(3) Ratio Cut [24]

$$\min\{RCut(A,B)=\frac{Cut(A,B)}{\min(|A|,|B|)}\} \tag{6}$$

where $|A|,|B|$ denote the number of vertices in subgraphs A and B, respectively. The cut of graph G according to the case when the $RCut$ function of subgraphs A,B is minimized is the proportional cut set criterion, which can minimize the similarity between subgraphs and reduce the possibility of over-cutting, but the operation speed is slow.

(4) Average Cut [25]

$$\min\{AvCut=\frac{Cut(A,B)}{|A|}+\frac{Cut(A,B)}{|B|}\} \tag{7}$$

It can be seen that $Avcut$ uses the sum of the ratio of the cost function and the number of data points in the divided region, which can theoretically produce a more accurate division, but the same drawback is that it is easy to divide smaller subgraphs that contain only a few vertices. In addition, it is pointed out in the literature [22] that the experimental results of using Normalized cut criterion are better than Average cut criterion when dividing the same image.

(5) Minimum-Max Cut [26]

$$\min\{MCut = \frac{Cut(A,B)}{Vol(A,A)} + \frac{Cut(A,B)}{Vol(B,B)}\} \tag{8}$$

The central idea of the minimum-maximum cut-set criterion is to minimize Cut(A, B) while maximizing $Vol(A,A)$ and $Vol(B,B)$. Minimizing this function avoids dividing a smaller subgraph containing only a few vertices, so it tends to produce balanced cut sets, but is slower to implement. $MCut$ satisfies the same principle of small similarity between samples between classes and large similarity between samples within classes as $NCut$, and has similar behavior to $NCut$, but when the overlap between classes is large, $MCut$ is more efficient than $NCut$.

(6) Multiway Normalized Cut [27]

The objective functions used in the above five partitioning criteria are all 2-way partitioning functions that partition the graph G into 2 subgraphs, and Meila proposes a partitioning function that can partition the graph G into k subgraphs k-ways at the same time,

$$\begin{aligned} MNCut = & \frac{Cut(A_1, V - A_1)}{Vol(A_1, V)} + \frac{Cut(A_2, V - A_2)}{Vol(A_2, V)} + \cdots \\ & + \frac{Cut(A_k, V - A_k)}{Vol(A_k, V)} \end{aligned} \tag{9}$$

The only difference between $Ncut$ and $MNcut$ is that the used spectral mapping is different and $MNcut$ is equivalent to $Ncut$ for $k = 2$. The multiplexed canonical cut-set criterion is reasonable and effective in practice, but its optimization problem is usually difficult to solve.

Although there are various classification criteria and implementations of spectral clustering algorithms, they can be summarized in the following general flow [28].

1) Calculate the similarity W_{ij} between the vertices, construct the similarity matrix W, and construct the matrix Z representing the sample set according to the different objective functions;

2) Calculate the first k eigenvectors of Z, and build the eigenvector space;

3) Clustering the eigenvectors in the eigenvector space by $K - means$ or other classical clustering algorithms.

Among them, the similarity function used to calculate the similarity between vertices W_{ij} often uses the Gaussian kernel function.

$$W_{ij} = e^{\left(-\frac{d\left(s_i, s_j\right)}{2\sigma^2}\right)} \tag{10}$$

Wheres_i, s_j is the data point in the sample, and$d\left(s_i, s_j\right)$ is the distance between index data points, generally referred to as the Euclidean distance. σ is the scale parameter, the σ The value taken affects the W_{ij} the computation of the algorithm, which indirectly affects the clustering results of the algorithm. In addition, the optimal σ values are not the same, so the value in practice needs to be determined after several experiments based on the specific dataset.

The significance of using Gaussian kernel function for the similarity function is to map the data points to a high-dimensional space, add more features, highlight the differences between data points, and thus calculate the similarity between data points more accurately.

Based on the general process above, the specific steps of a standard spectral clustering algorithm [29] are given below, the input dataset X is an arbitrarily shaped dataset consisting of n points $x_1, x_2, \cdots, x_n$ consisting of an arbitrarily shaped dataset, each point can be an arbitrary object, the similarity function uses a Gaussian kernel function, the division criterion is based on $Ncut$, and the final clustering method uses $K - means$ is as follows.

Input: sample set $X = (x_1, x_2, \cdots, x_n)$, scale parameter σ , number of clusters k.
Output: Cluster division result $C(c_1, c_2, \cdots, c_k)$.

Step1: Construct the similarity matrix $W_{n\times n}$. Where the element W_{ij} in $W_{n\times n}$ is the similarity between points i and j in the sample, calculated according to formula (10) .

Step2: Construct the degree matrix D, The element D_{ij} in D is the sum of the i row in the $W_{n\times n}$ matrix.

Step3: Based on $L = I - D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$, construct the normalized Laplacian matrix L_{sym}, where I is the unit matrix.

Step4: Calculate the eigenvectors f corresponding to each of the smallest k eigenvalues before L_{sym}.

Step5: Normalize the eigenvectors f to finally form an $n * k_1$ -dimensional eigenmatrix F.

Step6: for each row in F as a k_1-dimensional sample, a total of n samples, clustering by k-means or other clustering methods, the clustering dimension is k_2.

Step7: Get the cluster division result $C(c_1, c_2, \cdots, c_k)$.

3 Analysis of Spectral Clustering Algorithm Based on Differential Privacy Preservation

3.1 Algorithm description

This algorithm can be divided into two stages in the execution process, the first stage is to add noise conforming to the Laplace distribution to the training data to perturb the original data set; The second stage is to apply the standard spectral clustering algorithm in Table 1 to cluster the noise-added dataset. The specific algorithm is as follows.

Input: sample set $X = (x_1, x_2, \cdots, x_n)$, scale parameter σ, number of clusters k , privacy budget ε.

Output: Cluster division result $C(c_1, c_2, \cdots, c_k)$.

Add the Laplacian noise under a given privacy budget to the points in the sample set X to obtain the disturbed sample set X'.

Step1: Construct the similarity matrix $W_{n \times n}$. Where the element W_{ij} in $W_{n \times n}$. is the similarity between points i and j in the sample, calculated according to formula (10).

Step2: Construct the degree matrix D, The element D_{ij} in D is the sum of the i row in the $W_{n \times n}$ matrix.

Step3: Based on $L = I - D^{-\frac{1}{2}}SD^{-\frac{1}{2}}$, construct the normalized Laplacian matrix L_{sym}, where I is the unit matrix.

Step4: Calculate the eigenvectors f corresponding to each of the smallest k eigenvalues before L_{sym}.

Step5: Normalize the eigenvectors f to finally form an $n * k_1$-dimensional eigenmatrix F.

Step6: for each row in F as a k_1-dimensional sample, a total of n samples, clustering by k-means or other clustering methods, the clustering dimension is k_2.

Step7: Get the cluster division result $C(c_1, c_2, \cdots, c_k)$.

3.2 Algorithm Analysis

Scheme for input perturbations

Among several existing machine learning differential privacy protection schemes, the algorithm proposed in this paper uses an input perturbation scheme to protect privacy security by adding noise conforming to the Laplace distribution to the original dataset before clustering. Compared with other schemes, the input perturbation scheme has the advantages of easy implementation and less loss of clustering accuracy [30], which enables the algorithm to achieve privacy protection at source under the condition of guaranteeing clustering accuracy; from the specific implementation of this scheme in this algorithm, the disturbance of input data will lose dataset reconstruction on the one hand so that the trained model can resist model inversion attacks [31]-[32] and model theft attacks[30] [33], greatly reducing the risk of model information leakage; on the other hand, the datasets contacted by the model are perturbed, hiding the true intimacy between points in the original dataset, so that this algorithm can also solve the traditional spectral clustering algorithm concerned in the literature [16], it is easy to reveal the problem of intimacy between samples.·

Sensitivity

Regarding the sensitivity, this algorithm is sensitive in the neighboring data set D, D' When any record is modified on the neighboring data set, the data sensitivity is 1 for each dimension, so the global sensitivity is n.

Privacy Budget

This algorithm adds noise that fits the Laplace distribution to each data set before the model is learned, so the total privacy budget of the algorithm ε satisfies the $\varepsilon - differential\ privacy$ the defined measure of the differential privacy model.

4 Simulation experiments

4.1 Experimental design

The experimental design session includes the selection of evaluation metrics for the clustering algorithm, the introduction and pre-processing of the data set, and the hardware and software environment for running the experiments.

we choose to use the Adjusted Rand Index (ARI)[34] as the evaluation index of the clustering algorithm to measure how well the algorithm clustering results match with the actual situation.

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \tag{11}$$

Where RI is the Rand Index (RI, equation 12), $E[RI]$ is the mathematical expectation of the Rand Index, and $max(RI)$ is the maximum value of the Rand Index. The Rand coefficient is calculated by the following formula:

$$RI = \frac{a+b}{C_{n\_samples}^{2}} \tag{12}$$

where a is the number of correct similar pairs, b is the number of correct dissimilar pairs, C is the combinatorial number symbol, and $n\_samples$ is the total number of data points. RI takes a range of $[0,1]$, and a larger value means that the clustering results match the real situation.

The Adjusted Rand Index is an improvement of the Rand Index (RI), which overcomes the shortcomings of the original Rand index for "random clustering does not guarantee that the score is close to zero", it also has a higher degree of discrimination.

(1) Experimental data set and pre-processing

The datasets used in this paper include two artificially synthetic-sized two-dimension datasets, Moon and R2, and two datasets wine and iris from the UCI Machine Learning Repository [34]. Detailed information is shown in Table 1.

Table 1. Data set

Dataset	Number of samples	Number of attributes	Number of categories
moon	600	2	2
R2	358	2	4
wine	178	13	3
iris	150	4	3

In the preprocessing step, this article first normalizes all the data sets so that their attribute values all fall within the interval [0,1]. In addition, considering that this algorithm uses a Gaussian kernel function for the calculation of the similarity between samples (Equation 4), the selection of the scale parameter ∂ affects the final clustering result. In order to eliminate this influence, the value of ∂_g in the following experiment of the differential privacy spectral clustering algorithm is the value of ∂ that makes the spectral clustering effect of the four data sets optimal. The process of determining the specific value of ∂_g is as follows: adjust the value of ∂, such as 0.1, 0.2, 0.5, 2, 6, 9, 11 and 12, perform multiple spectral clustering on the normalized data set, and cluster Calculate the ARI coefficient from the class result, and select the ∂ value corresponding to the largest coefficient as the value of ∂_g. It can be seen from Fig. 2 that for the data set moon and R2, the optimal ∂ value of the clustering effect should be selected 12. For the data set iris and wine, the optimal ∂ value of the clustering effect is maintained at about 8.

4.2 Experimental results and analysis

Experiment 1 compared the clustering results of spectral clustering algorithm and differential privacy spectral clustering algorithm.

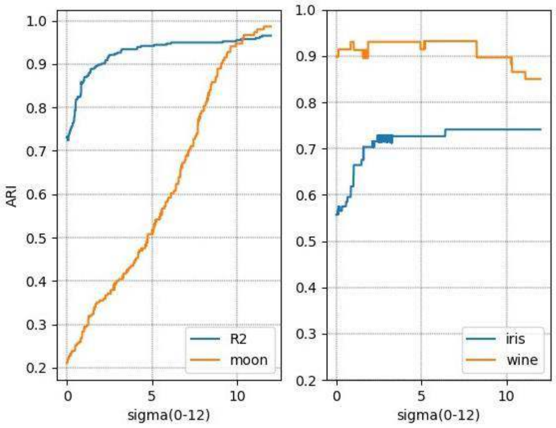
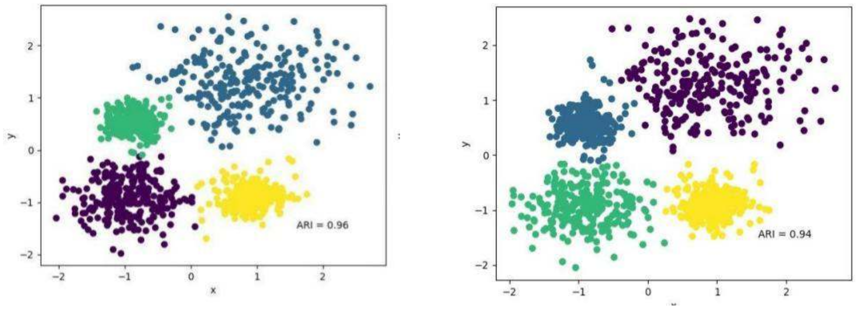


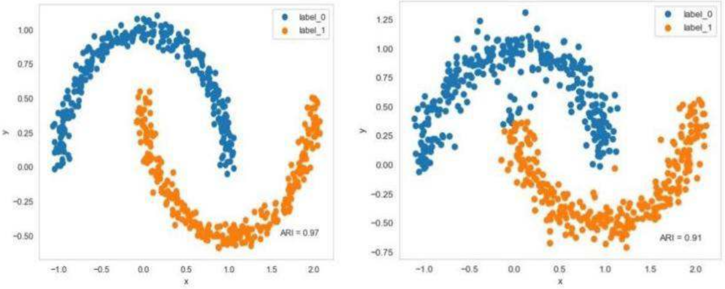
Fig. 2. Relationship between parameter ∂ and clustering effect

Fig. 3 and 4 show the clustering effects on the two datasets R2 and moon, respectively. The ARI index are marked in the lower right corner of each image. Among them, Fig. 3(a) and Fig. 4(a) depict the clustering effect without the effect of differential privacy protection, while Fig. 3(b) and Fig. 4(b) correspond to the clustering effect of the differential privacy spectrum clustering. From the experimental results, the differential privacy spectral clustering algorithm proposed in this paper is able to identify the correct clustering classes with less loss of accuracy.



(a) Effect of spectral clustering algorithm (b) Effect of differential privacy spectral clustering algorithm

Fig. 3. Clustering effect on R2 dataset

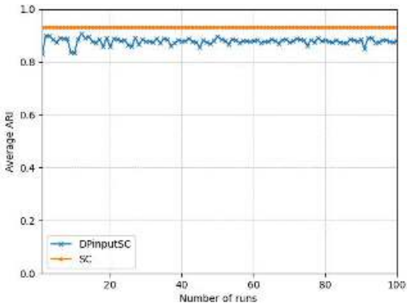


(a) Effect of spectral clustering algorithm (b) Effect of differential privacy spectral clustering algorithm

Fig. 4. Clustering effect on moon dataset

The spectral clustering algorithm and the differential privacy spectral clustering algorithm were run several times on two datasets from UCI, wine and iris, and draw a trend graph, where the x-axis is the number of runs of the algorithm was run n (each run of the algorithm was based on the original dataset), and the y-axis is the average of the n ARI index for n runs of the algorithm (for distinction, it is denoted as "Average ARI"). The aim is to observe the usability and stability of the algorithm by the trend of the average ARI with the number of runs.

As can be seen from Fig.5 and 6, the Average ARI coefficient of the differential privacy spectral clustering algorithm is slightly lower than that of the spectral clustering algorithm, because a certain amount of perturbation is generated after adding noise to the original dataset, which causes the original dataset to lose a portion of its accuracy. From the average ARI coefficient, the scale of the reduction of the value is not large, indicating that the clustering results do not produce a large change, so the differential privacy spectrum is clustered class algorithms are still available. In addition, the average ARI index of the differential privacy spectrum clustering algorithm gradually stabilizes with the increase of the number of runs, which also reflects the stability of the algorithm.



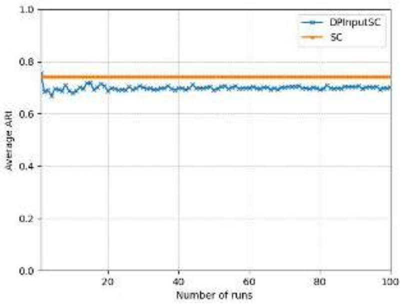


Fig. 5. Wine Data Clustering Classes **Fig. 6.** iris dataset clustering

5 Conclusion

In order to solve the problem of privacy leakage in traditional clustering algorithms, we design a spectral clustering algorithm with differential privacy mechanism. By adding noise that conforms to the laplacian distribution to the original data, to reduce the risk of sensitive information leakage and prevent reconstruction attacks against the model to achieve the purpose of protecting privacy and security. The experimental results show that the spectral clustering algorithm based on differential privacy protection proposed in this paper can not only achieve privacy protection, but also has stability and usability. Since the perturbed data can affect the clustering accuracy, the next step will be to investigate how to guarantee the differential privacy implementation while improving the accuracy of the algorithm clustering as much as possible and increasing the algorithm usability.

References

1. Wu X, Wang H ,Shi M , et al. DNA Motif finding Method without Protection can leak user Privacy[J]. IEEE Access, 2019, PP(99):1-1.
2. Achieving Privacy Preservation when Sharing Data for Clustering[C]// Springer Berlin Heidelberg. Springer Berlin Heidelberg, 2004.
3. Nayahi J, Kavitha V . Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop[J]. Future Generation Computer Systems, 2016, 74(SEP.):393-408.
4. Practical privacy: the SuLQ framework[C]// Proceedings of the Twenty-fourth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 13-15, 2005, Baltimore, Maryland, USA. acm, 2005.
5. Dwork C . A Firm Foundation for Private Data Analysis[J]. Communications of the Acm, 2011, 54(1):p.86-95.
6. Yanming F U, Zhenduo L I . Research on k-means++ Clustering Algorithm Based on Laplace Mechanism for Differential Privacy Protection[J]. Netinfo Security, 2019.
7. Ni T, Qiao M, Chen Z, et al. Utility-efficient differentially private K-means clustering based on cluster merging[J]. Neurocomputing, 2020.

8. Xiang W, Wei Y, Mao Y, et al. A differential privacy DNA motif finding method based on closed frequent patterns[J]. Cluster Computing, 2018, 22(21).
9. Wu, W. M., Huang, H. K. Research on DP-DBScan clustering algorithm based on differential privacy preservation[J]. Computer Engineering and Science, 2015, 37(4):830-834.
10. Wu X, Zhang Y, Wang A , et al. MNSSp3: Medical big data privacy protection platform based on Internet of things[J]. Neural Computing and Applications, 2020(4).
11. Tian Wang, Yucheng Lu, Jianhuang Wang, Hong-Ning Dai, Xi Zheng, Weijia Jia, EIHDP: Edge-Intelligent Hierarchical Dynamic Pricing Based on Cloud-Edge-Client Collaboration for IoT Systems, IEEE Transactions on Computers, 2021,70(8): 1285-1298.
12. Youke Wu, Haiyang Huang, Ningyun Wu, Yue Wang, Md Zakirul Alam Bhuiyan and Tian Wang, An Incentive-Based Protection and Recovery Strategy for Secure Big Data in Social Networks, Information Sciences, 2020, 508: 79-91.
13. Xiang Wu, Yongting Zhang, Minyu Shi, Pei Li, Ruirui Li, Neal N. Xiong, An adaptive federated learning scheme with differential privacy reserving, Future Generation Computer Systems,2022, 127,pp: 362-372, https://doi.org/10.1016/j.future.2021.09.015.
14. Tian Wang, Yan Liu, Xi Zheng, Hong-Ning Dai, Weijia Jia, Mande Xie. Edge-based Communication Optimization for Distributed Federated Learning. IEEE Transactions on Network Science and Engineering, 2021, 10.1109/TNSE.2021.3083263.
15. Yang Fan, Zhou Xiang, Ma Li. Application of spectral clustering algorithm in chemical reagent library preparation optimization[J]. Computers and Applied Chemistry, 2019, 36(5):3.
16. Guo L, Yang J, Song NQ. Application of spectral clustering algorithm in the diagnostic assessment of different attribute hierarchical structures[J]. Psychological Science, 2018, 41(3):8.
17. Zheng Xiaoyao, Chen Dongmei, Liu Yuqing, et al. A spectral clustering algorithm based on differential privacy preservation[J]. Computer Applications, 38(10):5.
18. Hu, B.. Research on clustering algorithm for differential privacy protection[D]. Nanjing University of Posts and Telecommunications, 2019.
19. Liu W , Li J , Wei J , et al. Privacy-preserving Constrained Spectral Clustering Algorithm for Large-scale Data Sets[J]. IET Information Security, 2019, 14(1).
20. C D work. Differential Privacy[J]. Lecture notes in computer science, 2006.
21. Liu JX, Meng SF. A review of privacy-preserving research on machine learning[J]. Computer Research and Development, 2020, 057(002):346-362.
22. Shi J, Malik J. Normalized cuts and image segmentation. IEEE Tranactions on Pattern Analysis and Machine Intelligence, 2000, 22(8) :888-905.
23. Wu Z , Leahy R . An optimal graph theoretic approach to data clustering: theory and its application to image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1993, 15(11):1101-1113.
24. Hagen L , Kahng A B . New spectral methods for ratio cut partitioning and clustering[J]. IEEE Transactions on Computer Aided Design of Integrated Circuits & Systems, 1992, 11(9):P.1074-1085.
25. Sarkar, Sudeep, Soundararajan, et al. Supervised Learning of Large Perceptual Organization: Graph Spectral Partitioning and Learning Automata.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2000.
26. Ding C , He X , Zha H , et al. Spectral min-max cut for graph partitioning and data clustering. 2001.
27. Meila M , Xu L . Multiway cuts and spectral clustering. U .Washingt on Tech Report .2003

28. Cai X Y, Dai G Z, Yang L B. Survey on spectral clustering algorithms[J]. Computer Science. 2008, 35(7): 14-18.
29. Bai Lu, Zhao Xin, Kong Yuting, et al. A review of spectral clustering algorithms[J]. Computer Engineering and Applications, 57(14):12.
30. Zhao ZD, Chang XL, Wang YX. A review of privacy protection in machine learning[J]. Journal of Information Security, 2019, v.4(05):5-17.
31. Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. USENIX Association, 2014.
32. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures[C]// the 22nd ACM SIGSAC Conference. ACM, 2015.
33. Stealing Machine Learning Models via Prediction APIs[C]// 25th USENIX Security Symposium, USENIX Security 16, Austin, TX, USA, August 10-12, 2016. 2016.
34. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance[M]. JMLR.org, 2010.
35. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.