Research on Data Mining Technology of Internet Privacy Protection based on local differential privacy

Juan Yu

Maanshan Teacher's College

jolie_yuu@126.com

Project name and Number: Key Project of Natural Science Research in Anhui Province (2022AH052835)

Abstract. With the rapid development of social economy and science and technology, data sharing and analysis technology is one of the contents widely used in the Internet field. Collecting and distributing terminal data and mining analysis is the normal form of big data analysis. Under the condition of mutual distrust between the collector and the terminal data owner, privacy protection in data collection and analysis will directly affect the application results of big data analysis. Nowadays, localized differential privacy has been applied in data acquisition and data mining analysis. Although there are still many problems in relevant theories and technologies, with the continuous in-depth research of scientific researchers, relevant issues have been proposed according to practical cases, which not only effectively control the privacy budget loss, but also improve the accuracy of data application. Therefore, on the basis of understanding the concept of localized differential privacy and related models, and according to the current application status of Internet privacy protection data mining technology, this paper proposes a clustering method with LDP GMMC as the core. The final experimental results show that the improved data acquisition method can better meet the privacy protection requirements in the non-spherically distributed data scenario.

Keywords: Local differential privacy; The Internet; Privacy protection; Data mining; LDP GMMC.

1. Introducion

In the construction and development of modern society, widely used data information contains multiple values, attracting social organizations, researchers, enterprises and other companies to explore the potential content contained in it. Especially after entering the era of big data, with the continuous innovation of smart device technology, the nodes of information collection in the Internet of Things system have changed from traditional hardware devices to entity nodes represented by smart devices and users, and these contents have become more complex and personalized from traditional simple applications. [1-3]Even if the Internet of Things service is not actively accepted, there is also the risk of privacy disclosure. For example, in order to realize the requirements of energy conservation and environmental protection, an office building is equipped with an Internet of Things system with temperature control equipment. This system will control and adjust the temperature equipment by tracking the location of staff in the office building, and will collect all the information related to users during use. Therefore, it is likely to violate the privacy of employees. Generally, the architecture design of the Internet of Things system is divided into three levels, the first refers to the perception layer, the second refers to the network layer, and the last refers to the application layer. Wireless sensor network nodes are used to acquire all kinds of data, network devices such as sensor network video network reader are used to transmit and converge all information in the sensing layer, and various operations are completed according to system requirements. The network layer belongs to the middle layer, which is mainly used to transport and process the information required by the perception layer. The application layer belongs to the top layer, which is the interface between the Internet of Things system and the outside world. It can effectively combine the application layer with different user needs, so as to complete specific data information. From the perspective of practical application, the data collected by the Internet of Things system has two characteristics: on the one hand, it means dynamic, and on the other hand, it means data storage and release with set value.[4-6]

Nowadays, scholars from different countries have proposed two data privacy protection methods based on the Internet field, one refers to the anonymization technology, the other refers to the encryption method, both of which have corresponding research results and can effectively deal with data problems in different scenarios. At present, Chinese scholars focus on three aspects when studying the data privacy protection content of the Internet of Things. First, it refers to the anonymization technology, which can truly realize the basic goal of privacy protection by disconnecting the correspondence between sensitive data and individual identity. Secondly, it refers to the encryption technology. This method will use the security protocol and security group communication mode of encryption technology to ensure the data communication effect between devices. Finally, exponential data perturbation mechanism will protect user privacy by hiding data content. The most common way is divided into two types, one is exponential data aggregation, the other is differential privacy protection model.

In the process of studying how to release data and protect data privacy, domestic and foreign scholars mainly focus on two aspects: on the one hand, first add noise to the original data set, and then use different strategies to optimize the release results. For example, some scholars put forward the Boostl algorithm in their research, which will use the idea of consistency constraint and least square method to process the data set, so as to improve the availability of the data set, but it is only suitable for one-dimensional histogram and short query range. Some scholars also proposed the DPCube algorithm in their research, which mainly divides the original data set initially, adds Laplacian noise to the segmented set, uses the idea of kd tree to optimize the noisy results, and publishes the multidimensional V optimized histogram as the final result. On the other hand, the original data is transformed and analyzed first, and then noise is added to the transformed data. The publishing technology corresponding to this research not only includes histogram distribution method, but also involves distribution method with partition as the core. Based on the research status of local differential privacy and Internet privacy protection data mining technology, this paper proposes a clustering method based on LDP GMMC, and verifies the application value of local differential privacy clustering with practical cases, so as to provide an effective basis for data collection and data analysis in the new era.

2. Methods

2.1 Local differential privacy model

As one of the variation types of differential privacy, localized differential privacy is mainly used to solve the problem that there is no trusted data acquisition party in the actual scenario, so it is very suitable for application in the distributed data collection scenario. Combined with the comparative framework analysis shown in FIG. 1 and FIG. 2 below, it can be seen that FIG. 1 represents the differential privacy framework and FIG. 2 represents the localized differential privacy framework. After the user perturbates the information locally and transmits it to a third-party data collection institution, it is difficult for the data collection party to access the original data, so the security of user privacy data can be ensured on a fundamental basis. Compared with differential privacy, localized differential privacy fully considers the possibility of privacy leakage of the data collector, so it can effectively avoid the direct exposure of user privacy in the process of data analysis and mining when transmitting the data set after processing.[7-9]

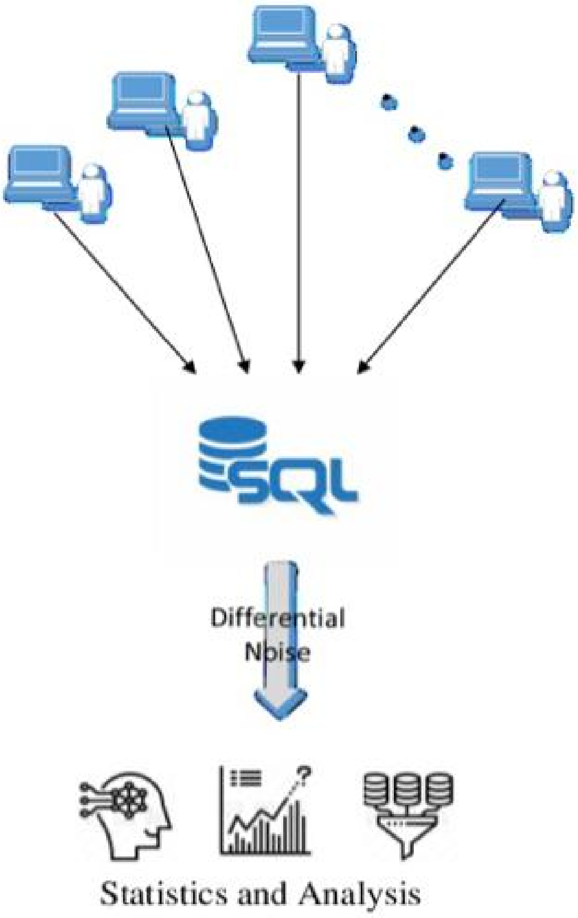

Figure 1 Frame diagram of differential privacy

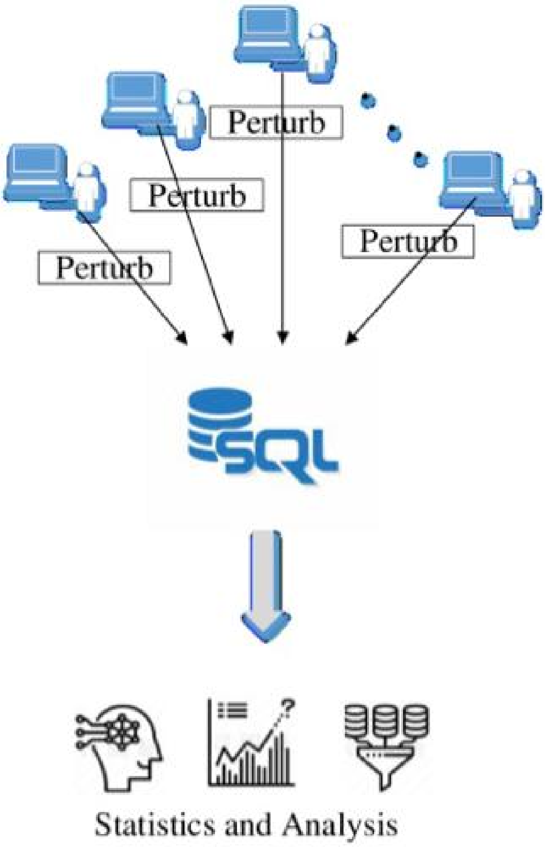

Figure 2 Frame diagram of localized differential privacy

2.2 Localized differential Privacy clustering method based on LDP GMMC

Based on the current knowledge of localized differential privacy clustering, we can see that the method should be used to improve the data collection method and set corresponding measures and mechanisms, so as to improve the quality of privacy protection clustering mining and information application security in the non-spherically distributed data scenario.[10-13]

Suppose you have a set of N user $U=\{u_1,u_2,...,u_N\}$, all users have a data set of size d $X_i=\{x_1,x_2,...,x_d\},1\leq i\leq N$, then the data set of all users is:

$$D=\{X_1,X_2,...,X_N\},1\leq i\leq N$$

Under the condition of conforming to the privacy of localized difference, the data collector will deeply mine the user data information and divide it into different clusters. Then, for the user data set D, the cluster with a given size k can be divided into:

$$C=\{C_1,C_2,...,C_k\}$$

The clustering division should meet the following conditions:

$$\arg\min_c \sum_{1\leq i\leq k}\frac{\left|C_i\cup\hat{C}_l\right|-\left|C_i\cap\hat{C}_l\right|}{\left|C_i\cup\hat{C}_l\right|}$$

In the above formula $\hat{C}_l$ Represents the actual cluster classification. From a theoretical point of view, the clustering quality can be optimized at the minimum of the above formula.

Assuming random variable X, there will be a Gaussian mixture model composed of K Gaussian models. The specific formula is shown as follows:

$$p(X)=\sum_{k=1}^{K}\pi_k N\left(X\middle|\mu_k,\sum{}_k\right)$$

In the above formula $N\left(X\middle|\mu_k,\sum{}_k\right)$ Represents a component of the mixed model, μ_k represents the mean vector $\sum{}_k$ Represents the covariance matrix, π_k represents the mixing coefficient, which conforms to $0\leq\pi_k\leq 1 and \sum\pi_k=1$ This condition.

Point P represents the probability of different clusters, and the specific calculation formula is as follows:

$$Belong(P,i)=\frac{\pi_k N\left(P\middle|\mu_k,\sum{}_k\right)}{\sum_{i=1}^{K}\pi_k N\left(P\middle|\mu_k,\sum{}_k\right)}$$

Compared with K-means method, Gaussian mixture model is easier to fit the scene with uneven cluster distribution, can quickly grasp the correlation between various attributes, does not need to rely on distance calculation, and can avoid the influence of disturbance noise on the composition of distance calculation.

From the perspective of practical application, the research method in this paper should first consider the conversion of complex operations into single-value operations supported by the existing perturbation framework, reduce the privacy budget after the completion of the approximate calculation, and improve the feasibility of the method by balancing errors and privacy budget consumption. At the same time, the Gaussian mixture model should be used for cluster analysis, and the sub-cluster merging mechanism based on model overlap degree should be used to optimize the clustering results, so as to improve the availability of the algorithm. According to this method, the interactive framework as shown in Figure 3 below is constructed:

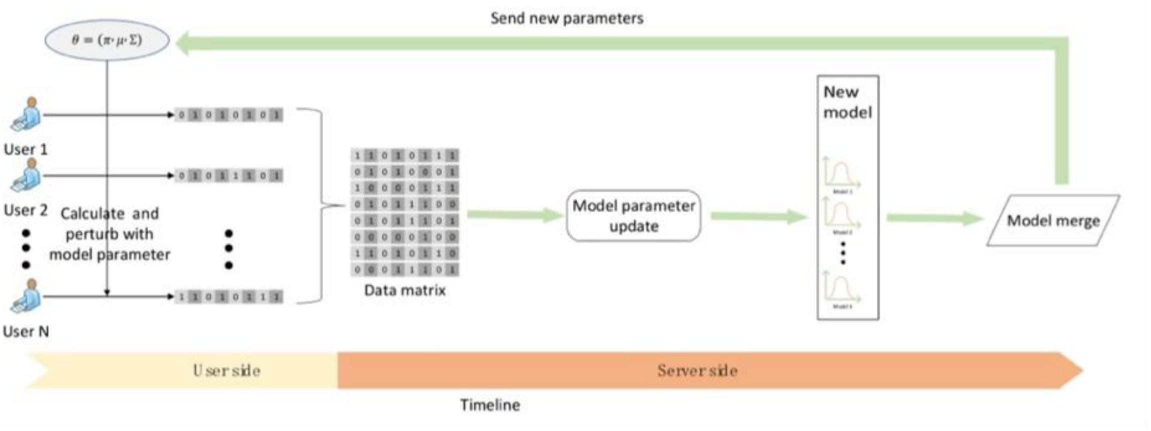

Figure 3 Method framework based on LDP GMMC

Based on the above figure, we can see that the overall system framework is divided into two parts: on the one hand, it refers to the client side, which will calculate the parameter model sent by the corresponding server side, collect and analyze the disturbance of data encoding, and transmit it directly to the server side; On the other hand, the server side. The content of user data collection will be updated, and the cluster will be merged when the corresponding conditions are met. At the same time, the model parameters will be provided to the client after the update, and the algorithm will be finished and the result will be returned after the label is stable.

Generally, the localized differential privacy clustering method is divided into two stages, one is the exponential data disturbance, the other is the calculation of the corresponding clustering parameters, so after the construction of the method framework as shown in the figure above, the user and server should choose the following methods as shown in Table 1 and Table 2, the former is mainly used to protect user privacy. The latter can realize the clustering analysis of Gaussian mixture model under localized differential privacy, and finally complete the clustering process through iterative update.[14.15]

Table 1 Data disturbance based on the client

input：Privacy budget ε ，Server parameters θ=（π，μ，Σ），code base λ
output：Disturbed user data $\tilde{X}$, model calculation median $\tilde{Y}$ Disturbance mode flag.
1.Allocate each round of privacy budget ε_{i}
2.Using coding parameters λ Code the user.
3.for each bit in user data X
4.$\tilde{b}=\begin{cases} b, with \quad probability \quad p \\ 1-b, with \quad probability \quad q \end{cases}$
5.Collect disturbance data $\tilde{b}$ and add it to $\tilde{X}$
6.for each component patamenters θ_i *in* θ
7.calculate $\gamma_k=\pi_k N\left(X \mid \mu_k,\sum_k\right)$
8.flag=perturbation_selection （ε_{i}）
9.if flag=0
10.Perturbation $X,\tilde{Y}$ and product.
11.else if flag=1
12.Direct disturbance X *and* $\tilde{Y}$

13.return perturded data and flag

Table 2 Iterative clustering based on server side

input：Disturbed user data $\widetilde{X}$, median value of model calculation $\widetilde{Y}$ Disturbance mode flag
output：Cluster result c
1.Generate initial parameters
2.do
3.Send parameter θ To each user
4.Collecting disturbance data $\widetilde{X}$ and disturbance mode label flag.
5.Determine the data acquisition mode through flag
6. $\theta^{new} = \left(\pi^{new}, \mu^{new}, \sum^{new}\right)$/* Calculate and update parameters using $\widetilde{Y}$ and $\widetilde{X}$ aggregation.*/
7.while θ converge or halt_condition is satisfied/*Result convergence*/
8.for each user μ_i
9.for each model
10. $label = \arg\max_{j}\left(Belong\left(u_i, \mathrm{mod}\, el_j\right)\right)$
11. $C_{label} = C_{label \cap u_i}$

3. Result analysis

In the research experiment of this paper, different privacy budgets and initial cluster numbers are set respectively, and the influence of privacy operation on method composition is judged in the comparative study. From the perspective of experimental research, with the increase of privacy budget, the feasibility of the algorithm will become higher and higher. Faced with the problem of single application of localized differential privacy clustering method in data scenarios, the localized differential privacy Gaussian mixture model clustering method proposed in this paper can truly meet the collection requirements of multi-attribute multiplication in the Gaussian mixture model and comprehensively improve the availability of data analysis by improving the data collection mechanism. The clustering and merging mechanism can reduce the influence of initial values on the application method composition and improve the usability of the algorithm. The final experimental results show that the LDP GMMC-based method can show high availability and effectiveness in non-data sets on the basis of protecting user data and film, which is closely related to the protection of user privacy in the new era of big data environment. Therefore, on the basis of mastering the local differential privacy requirements, Chinese scholars should continue to put forward frequent item set mining methods and clustering methods with localized differential privacy as the core in combination with practice cases. Only in this way can they master more high-quality technical theories and practical results and provide effective basis for solving user data privacy problems in the era of big data. It should be noted that the data mining and data clustering activities in the localized differential privacy scenario studied in this paper still have certain limitations. Therefore, in future experimental discussions, scholars should pay attention to grasp practical research problems and actively explore data disturbance and data acquisition strategies of data flow, so as to fully demonstrate the application value of local differential privacy.

4. Conclusion

To sum up, in the big data environment, user privacy protection has been widely concerned by the whole society, and localized differential privacy technology, as a new content proposed based on self-differential privacy technology, has a strong level of privacy protection, so relevant experimental topics have obtained rich results. At present, scholars at home and abroad pay more attention to the analysis of data mining technical means for Internet privacy protection based on local differential privacy, and master relevant theoretical technologies from practical research, which provides an effective guarantee for data mining and cluster analysis in the new era, and can fully meet the basic needs of data mining in various fields.

References

[1] Meishan Wang, Lan Yao,Fuxiang Gao, et al. Research on differential privacy protection technology for medical set-valued data [J]. Computer Science, 2022, 49(4):7.

[2] Lin Sun, Guolou Ping, Xiaojun Ye. Key-value data association analysis based on localized difference privacy [J]. Computer Science, 2021, 48(8):6.

[3] Min Fan,Geng Yang. Research on Information Security, 2021, 007(011):1007-1016.

[4] Linyu Wang. Research scheme of Social network user attribute differential Privacy Protection based on blockchain [J]. Software, 2022(004):043.

[5] Yurong Hao,Chunhui Pu,Jiaqi Yan, Jiang Xuehong. Research on privacy protection algorithm of government data Sharing based on localized difference privacy [J]. Chinese Journal of Information, 2021, 040(002):169-175,137.

[6] K-modes Clustering Data Privacy Protection Method Based on Local Differential Privacy [J]. Acta Electronica Sinica, 2022, 50(9):2181-2188.

[7] Wenjuan Yang. Privacy protection in online social networking based on differential privacy [J]. Wireless Internet Technology, 2021, 18(22):28-30.

[8] Shaobo Zhang, Liu Jie Yuan,Gengming Zhu. A Novel Privacy Protection Method for K-prototypes Clustering Data Based on Local Difference Privacy [J]. Acta Electronica Sinica, 2022, 50(9):8.

[9] Long Zhao, Dragon Worker. Crowdsourcing privacy protection Method based on local differential privacy [J]. Computer and Modernization, 2021(7):6. (in Chinese)

[10] Meiqi Zhu, Geng Yang, Baiyunlu. Frequent item Mining Algorithm based on Localized Differential Privacy Protection [J]. Computer Technology and Development, 2021, 031(008):92-99,168.

[11] Shunyong Li,Jiaxuan Zhang, Ruixuan Zhang, et al. Research on Differential Privacy Protection Algorithm for Athletes Based on Spectral Clustering [J]. Network Security Technology and Application, 2022(12):5.

[12] Jian Zheng, Licong Yang. Differential privacy Algorithm for large social networks based on singular Value decomposition [J]. Computer Technology and Development, 2022, 32(3):126-131.

[13] Yunlu Yang,Yajian Zhou,Hua Ning. Research on image data mining method supporting differential privacy [J]. Data Acquisition and Processing, 2021, 36(1):10.

[14] Xian Cao, Xuekun Zhao. Application analysis of Privacy Protection Based on Statistical Machine Learning Algorithm in Data Publishing and Data Mining [J]. Computer Application Abstracts, 2022(004):038.

[15] Chunchun Peng,Yanli Chen,Yanmei Xun. k-modes clustering method supporting localized differential privacy Protection [J]. Computer Science, 2021, 48(2):9.