Zhao JZ, Wang XW, Mao KM *et al.* Correlated differential privacy of multiparty data release in machine learning. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 37(1): 231–251 Jan. 2022. DOI 10.1007/s11390-021-1754-5

Correlated Differential Privacy of Multiparty Data Release in Machine Learning

Jian-Zhe Zhao[1] (赵建喆), Xing-Wei Wang[2,3,*] (王兴伟), *Senior Member*, *CCF*, Ke-Ming Mao[1] (毛克明)
Chen-Xi Huang[1] (黄辰希), Yu-Kai Su[1] (苏昱恺), and Yu-Chen Li[1] (李宇宸)

[1] *Software College, Northeastern University, Shenyang 110169, China*
[2] *State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China*
[3] *School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China*

E-mail: zhaojz@swc.neu.edu.cn; wangxw@mail.neu.edu.cn; maokm@swc.neu.edu.cn; 20184901@stu.neu.edu.cn
suyukai2@huawei.com; 20185242@stu.neu.edu.cn

Received July 1, 2021; accepted November 12, 2021.

Abstract Differential privacy (DP) is widely employed for the private data release in the single-party scenario. Data utility could be degraded with noise generated by ubiquitous data correlation, and it is often addressed by sensitivity reduction with correlation analysis. However, increasing multiparty data release applications present new challenges for existing methods. In this paper, we propose a novel correlated differential privacy of the multiparty data release (MP-CRDP). It effectively reduces the merged dataset's dimensionality and correlated sensitivity in two steps to optimize the utility. We also propose a multiparty correlation analysis technique. Based on the prior knowledge of multiparty data, a more reasonable and rigorous standard is designed to measure the correlated degree, reducing correlated sensitivity, and thus improve the data utility. Moreover, by adding noise to the weights of machine learning algorithms and query noise to the release data, MP-CRDP provides the release technology for both low-noise private data and private machine learning algorithms. Comprehensive experiments demonstrate the effectiveness and practicability of the proposed method on the utilized Adult and Breast Cancer datasets.

Keywords correlated differential privacy, multiparty data release, machine learning

1 Introduction

With the development of information technology and its penetration into daily life, sensing devices connected to the Internet, such as smartphones and wearable devices, are widely utilized, which results in the collection and storage of a vast amount of personal data. When released, these collected data are valuable sources for machine learning applications that have created immense social benefits in fields such as healthcare, financial analysis, and law enforcement [1]. Meanwhile, the emergence of new computing paradigms has increased the possibility of collecting data from multiple sources on a large scale. In these collected data, different features of the same set of individuals are often possessed by different parties as if the data were vertically partitioned among multiple parties. In practice, these vertically partitioned data can often be integrated to enable data analysis tasks that lead to better decision-making or high-quality services. As a real-world example, loan company A and bank B observe different sets of features about the same set of individuals who are identified by a common identifier (ID) [2]. Let the local dataset of A and the local dataset of B be $T_A(ID, Job, Salary)$ and $T_B(ID, Age, Balance)$, respectively. Integrating their data can better profile cus-

Regular Paper

This work is supported by the National Natural Science Foundation of China under Grant Nos. 62102074 and 62032013, the Liaoning Revitalization Talents Program under Grant No. XLYC1902010, the Natural Science Foundation of Liaoning Province of China under Grant No. 2020-MS-091, and Fundamental Research Funds for the Central Universities of China under Grant No. N2017015.

*Corresponding Author

tomers and support better decision-making about loan or credit limit approval. We further investigate the training accuracy of parts of the Breast Censer dataset with disjoint features and the merged Breast Censer dataset with all features. The comparison results in Table 1 show that the merged dataset improves the accuracy of the training model, such as logistic regression (LR) and the support vector machine (SVM). The results in Table 1 also demonstrate that, as the number of parties increases, the number of features increases. Therefore, the merged multiparty data often have high dimensionality.

Table 1. Training Accuracy on Vertically Partitioned Dataset

Vertically Partitioned Dataset	SVM	LR
Merged dataset with all features	0.978 9	0.985 4
Party 1 with features (f_1–f_{10})	0.964 9	0.973 6
Party 2 with features (f_{11}–f_{20})	0.940 4	0.954 3
Party 3 with features (f_{21}–f_{30})	0.926 3	0.929 1

It benefits from multiparty high-dimensional data. The merging of multiparty datasets improves the training accuracy. However, the merged multiparty dataset also contains more sensitive information, and ill-formed releases may cause privacy leakage. Recently, private data release has aroused widespread concern in academia and industry. Performing specific processing on data or algorithms via privacy mechanisms prevents personal sensitive information leakage successfully [3]. In the multiparty data release scenario, when a trusted server exists, it can merge multiparty data and perform privacy operations. Compared with distributed model training such as federated learning, centralized data release can achieve better data utility while providing privacy guarantee and support users' customized requirements for data release services [4–7]. Providing a rigorous and quantifiable privacy protection method that is independent of attack background knowledge, differential privacy (DP) [8, 9] has been the most widely accepted private data release method. In recent years, differentially private data release has been widely applied in many areas. Differentially private single-party data (i.e., overall data belongs to the same data holder) release and application have achieved satisfactory results, in which much existing work focuses on differential privacy technologies in machine learning, such as private data release for machine learning [10, 11] and private machine learning algorithms (private ML algorithms) release [12–21]. However, in the context of big data, due to the driving of shared data, there is ubiquitous data correlation in the collected data, especially in the multiparty scenario. Correlated data makes additional privacy leakage caused by data association abound, and data correlation has become a new hotspot in single-party private data release [22, 23].

An example of hospital records shows how correlated data can degrade the level of privacy in medical applications. As shown in Table 2, there are a group of users' medical records of different features. It is observed that $user_1$ and $user_2$ take the same values of f_1 to f_4 (they may have social relationships). In this case, if one user were to change the values of $user_1$, the values of $user_2$ would also change. In this way, the records for $user_1$ and the records for $user_2$ are correlated. The last row in Table 2 shows the counts of different features ($Counts$). In terms of the Laplace mechanism, adding the amount of Lap$(1/\epsilon)$ noise to perturb each count can achieve ϵ-DP. However, the expected privacy guarantee may breach by correlated records in the dataset. With the background information about the social relationships of people, such as family members, an attack can infer the health information of $user_1$ and $user_2$ by the feature flu. Consequently, after releasing the private count of a user's feature values, the health information of $user_1$ and $user_2$ may not be ϵ-differentially private as expected. Instead, their health information is 2ϵ-differentially private since changing one user's feature values will change the counts. Studies show that although differential privacy provides a strong guarantee for private data release, the correlation between records produces an increment of sensitivity, which introduces additional noise at the same level of protection and reduces the data utility [24–26]. Many state-of-the-art correlated privacy technologies address the correlation measurement between two records by correlation analysis and reduce the sensitivity to improve the data utility, such as Group DP [27], Correlated DP [28], and Bayesian DP [29].

Table 2. Example of Correlated Users' Medical Records

	f_1: fever	f_2: cough	f_3: headache	f_4: flu
$user_1$	0	1	1	1
$user_2$	0	1	1	1
$user_3$	0	1	1	0
$user_4$	1	1	0	0
$Counts$	1	4	3	2

Correlated differential privacy technologies in single-party data release have accumulated a diversity of research results. However, the growing multiparty

data release presents new challenges for these technologies, especially for vertically partitioned data. Since different features of the same user increase, redundant features may exist, degrading the machine learning model's performance. Dimensionality reduction and important feature selection are dominant means to improve the accuracy. However, dimensionality reduction causes a further increment of the sensitivity due to data correlation, which reduces the utility of the released data. As shown in Table 2, if feature f_4 is the redundant feature, $user_1$, $user_2$ and $user_3$ are correlated after feature selection. The sensitivity will increase from 2 to 3. The current study seems to present a lack of correlation analysis in a dimensional way in the multiparty scenarios. In this work, we analyze the correlation variation caused by dimensionality changes and propose a multiparty correlation analysis method. According to the prior knowledge of multiparty data, we provide a more reasonable and rigorous standard to measure the correlated degree between two records, reducing the correlated sensitivity and improving data utility. The proposed method addresses the current deficiencies of correlated differential privacy and provides a multiparty data release method for machine learning.

The main contributions of this paper are summarized as follows.

• We propose a multiparty data release method (MP-CRDP). It effectively reduces the merged dataset's dimensionality and correlated sensitivity in two steps to optimize the utility. By adding noise to machine learning algorithms' weights and query noise to the release data, our method provides a release technology of low-noise private data and private ML algorithms.

• We also propose a multiparty correlation analysis technique. It reduces the correlated sensitivity by considering the correlated degree and defining a more reasonable and rigorous standard based on the prior knowledge of multiparty data, thereby reducing the DP noise injecting.

• We execute comprehensive experiments to present the data correlation variation caused by dimensionality changing. The experimental results also demonstrate the high practicability of our method in private ML algorithms and querying data release.

The paper is organized as follows. Section 2 introduces the related work specifically on differential private multiparty release, correlated differential privacy, and private data release in machine learning. Section 3 introduces the preliminary of this work, including the notation, the concepts of differential privacy and correlated sensitivity, the vertically partitioned multiparty scenario, and the problem statement. Section 4 describes the proposed multiparty correlation analysis by presenting the existing problems and our improvements. Section 5 shows the proposed method in detail. The key steps of MP-CRDP, as well as the privacy analysis in each step, are demonstrated. In Section 6, MP-CRDP is evaluated and compared with other methods based on the performance. The result analysis is conducted to conclude the advantages of our method. Lastly, Section 7 summarizes the paper.

2 Related Work

2.1 Differentially Private Multiparty Data Release

Differentially private multiparty data release mainly solves the problems of releasing data or statistical information of data as a whole based on differential privacy when the original data belong to multiparty data holders. In related literature, Alhadidi *et al.* designed a two-party protocol that is suitable for an exponential mechanism by using mathematical tools, such as the Taylor formula [30], to solve the problem of two-party, horizontally partitioned data release. However, the limitation of technology scalability makes the method only suitable for two-party scenarios. Aimed at the issue of horizontally partitioned multiparty search log release, Hong *et al.* proposed a joint search logs' release method that satisfies differential privacy based on sampling technology [31]. Limited by the sampling method, this method can only ensure that the released data meet the slack (ε, δ)-differential privacy.

Mohammed *et al.* solved the problem of releasing vertically partitioned relational datasets [32]. They designed a two-party protocol that is suitable for the exponential mechanism. This method has a higher data utility and ensures differential privacy. However, the method is designed only for two-party scenarios. Cheng *et al.* proposed a vertically partitioned multiparty data generation method that is based on naive Bayes [33]. The method extends the two-party scenario to the multiparty scenario and generates hidden features by modeling multiparty features to generate private data based on the same distribution to satisfy differential privacy. However, this method frequently measures the correlation strength between two features during the modeling process. Therefore, the communication cost and

the amount of introduced noise are relatively high. Goryczka and Xiong [34] investigated the problem of secure data aggregation in the multiparty setting while ensuring differential privacy of the result. Dangi and Santhi [35] introduced a privacy-preserving method in multiparty data release on the cloud for big data using fusion learning.

Many researchers have accumulated a diverse of results on the privacy protection of multiparty data. However, the existing studies do not consider the data correlation issues in the multiparty scenario.

2.2 Correlated Differential Privacy

Differential privacy provides a rigorous mathematical method for defining indistinguishability to preserve privacy and ensures that adding or removing any single record does not affect the analysis results. However, previous studies have shown that when multiple datasets are correlated, privacy leaks are increased. At the same level of privacy protection, the correlated data produces an increment of sensitivity, which increases the extra cost of noise and reduces the data utility [24, 25]. Balancing privacy and utility in correlated datasets has become a new research hotspot of correlated differential privacy technology. Some of the research focuses on correlation measurement to reduce sensitivity. When some users of different datasets have the same records, the datasets are considered directly correlated. The correlation is rigorously defined as two identical records. Unlike direct correlation, the indirect correlation is more complex and defined as two different records about a user or his/her correlated users. For example, the information flow of some user activities, such as GPS records and social network records, may be correlated. It is not easy to define and measure the correlation with different degrees. Zhu *et al.* [28, 36] applied a correlation matrix to express the correlation between correlated datasets. By converting the global sensitivity into the correlated sensitivity and setting it as the upper limit of sensitivity, the influence of correlation was limited. For data correlation measurement, some uncertain correlation models, such as the Gaussian correlation model in [29] and [37] and Markov chain model in [38], have been proposed. Song *et al.* [39] proposed two mechanisms based on Pufferfish privacy mechanisms using a Markov chain to measure the data correlation between adjacent states, which reduces the global sensitivity. Zhang *et al.* [40] reduced the global sensitivity by feature selection for datasets with correlated records based on correlated sensitivity. Recently, some research achievements have been made on differential privacy protection of correlated data with special formats, such as sequential data, tuple data, and trajectory data [41–44].

To balance the privacy and utility of an algorithm, the correlated differential privacy technology, which focuses on selecting privacy parameters of correlated datasets, was also proposed. For instance, Wu *et al.* [45] proposed a definition of correlated differential privacy, provided a vague correlation measurement method, dynamically adjusted the privacy protection level of correlated datasets by Nash equilibrium theory, and verified the impact of multiple datasets' privacy parameters on global data utility via experiments.

The technologies of correlated differential privacy have achieved good results. However, these methods seem to lack the consideration of correlated degree. An objective correlated threshold also needs to be defined in multiparty scenarios.

2.3 Private Data Release in Machine Learning

For sensitive information preserving, much work has addressed the privacy issue in machine learning with differential privacy. To preserve personal privacy, noise is added to the target dataset so that the statistical information of the released dataset and the original dataset satisfies the upper bound of the indistinguishable threshold, and the released private data can be employed for machine learning [10, 11]. Another way of privacy-preserving is to release private algorithms. The classification models with differential privacy have been proposed, such as DiffP-C4.5 [12] and DiffGen [13]. Differentially private naive Bayes models [14] and regression models [15–17] have been proposed. Some of the algorithms combine differential privacy and SVM, such as private SVM [18] and objective SVM [19]. Song *et al.* [20] deduced the stochastic gradient descent mechanism of differential privacy and conducted an empirical test in logistic regression. Abadi *et al.* [21] proposed a deep learning algorithm with differential privacy, which provides a differential privacy version based on stochastic gradient descent.

These methods combine differential privacy with specific machine learning algorithms by disturbing the weights of the algorithms. However, they do not consider the correlated data, which may cause additional noise injection.

3 Preliminary

3.1 Notation

Let D be the merged dataset with n records and d features, and let variable r represent a record sampled from a universe $\mathcal{X}$. Two datasets D and D' are neighboring if they have the same cardinality but differ in only one record. Let r_i be that record, then D^i represents the dataset with r_i and D^{-i} represents the dataset with r_i deleted from D. A query Q is a function that maps the dataset D to a real number: $Q : D \to \mathbb{R}$. A group of queries is denoted as $\mathcal{Q}$. The set of records q that are related to a query Q is referred to as the query's responding records. Differential privacy provides a randomization mechanism M to mask the difference of query Q on the neighboring datasets [4]. Normally, we use a "hat" on the notation to represent the randomized answer. For example, $\hat{Q}(D)$ denotes the randomized answer of querying Q on D.

3.2 Differential Privacy

Differential privacy is a concept of privacy that was proposed by Dwork *et al.* in 2006 to address the privacy leakage of statistical datasets [8, 9]. The technologies based on DP design the mechanisms to add noise to the target dataset so that the statistical information loss of the released dataset and the original dataset is in a small range as shown in (1). It ensures that the modification of an individual record in the dataset does not significantly impact the statistical results.

Definition 1 (Differential Privacy). *For any datasets D and D' differing on at most one record, and for any possible sanitized dataset $r \in Range(M)$, a random mechanism M satisfies ϵ-differential privacy if*

$$DP(M) = \sup_{D,D',S} \log \frac{\Pr\left(r \in S \,\middle|\, D\right)}{\Pr\left(r \in S \,\middle|\, D'\right)} \leqslant \epsilon, \tag{1}$$

where ϵ refers to the privacy budget that controls the privacy level of mechanism M. The lower ϵ represents the higher privacy level.

Definition 2 (Global Sensitivity). *For any function $Q : D \to \mathbb{R}$, for all D and D' differing on at most one record, the global sensitivity of Q is*

$$\Delta GS = \max_{D,D'} \left\| QD - QD' \right\|_1. \tag{2}$$

Mechanism M is associated with the global sensitivity as shown in (2). It measures the maximal change on the result of query Q when removing one record from the dataset D. Two mechanisms are usually utilized to satisfy the differential privacy definition: the Laplace mechanism and the Exponential mechanism. Studies have shown that the sequential combination is satisfied in DP mechanisms [4, 5].

The Laplace mechanism adds Laplace noise to the output of a function to achieve differential privacy as shown in (3). This mechanism takes the dataset D, function Q, and privacy budget ϵ as inputs; it is designed for functions whose outputs are real.

Definition 3 (Laplace Mechanism). *For any function $Q : D \to \mathbb{R}$, the following mechanism provides ϵ-differential privacy*

$$\hat{Q}(D) = Q(D) + \mathrm{Lap}\left(\frac{\Delta GS}{\varepsilon}\right), \tag{3}$$

in which $\epsilon = 1/\lambda$. Specifically, the Laplace noise is sampled from Laplace distribution $\mathrm{Lap}(\lambda)$.

McSherry and Talwar [46] proposed an exponential mechanism to choose an output $t \in T$ that is close to the optimum with respect to a utility function while achieving differential privacy. The exponential mechanism takes the dataset D, privacy parameter ϵ, and utility function $u : (D \times T) \to \mathbb{R}$ as inputs. The utility function assigns a real valued score to every output $t \in T$, where a higher score means a better utility. The mechanism mainly addresses the algorithms whose output is non-numerical.

Definition 4 (Exponential Mechanism). *For any function $u : (D \times T) \to \mathbb{R}$, an algorithm Ag that chooses an output t with the probability proportional to* $\exp\left(\frac{\epsilon u D,t}{2\Delta u}\right)$ *satisfies ϵ-differential privacy.*

Lemma 1 (Sequential Composition). *Supposing a set of privacy steps $\{M_1, \ldots, M_m\}$ are sequentially performed on a dataset, and each M_i provides ϵ_i privacy guarantee, M provides $m\epsilon_i$-differential privacy.*

3.3 Correlated Sensitivity

In the single-party scenario, the correlated sensitivity has been demonstrated to effectively solve the query data release issue based on differential privacy [28, 36, 40]. In the mechanism, since records are only partially correlated, deleting one record may have different effects on other records. The influence of different strengths is defined as the correlated degree of records in the framework of correlated sensitivity. From the perspective of correlation analysis, several methods can be employed

to calculate the correlated degree of records. There are many ways to measure the correlated degree between records, such as attribute analysis, time interval analysis, and Pearson's correlation coefficient [28, 40]. $|w_{ij}| \in [0, 1]$ indicates the correlated degree between record i and record j. When $|w_{ij}| > 0$, there is a certain correlation between record i and record j; $|w_{ij}| = 1$ indicates that record i and record j are completely correlated; and $|w_{ij}| = 0$ indicates that record i and record j are completely uncorrelated. The higher the correlated degree is, the stronger the correlation between record values is. As shown in (4), a correlated degree matrix $\boldsymbol{\Phi}$ can describe a set of records' correlation as

$$\boldsymbol{\Phi} = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & w_{nn} \end{pmatrix}. \tag{4}$$

To satisfy the requirements of correlation analysis, a correlation threshold w_0 is used to screen the correlation to a certain degree as

$$w_{ij} = \begin{cases} w_{ij}, & \text{if } \ w_{ij} \geqslant w_0, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The filtered correlated degree matrix describes the correlation among all records of a dataset that satisfy the correlation threshold w_0, by which the correlated sensitivity of the dataset is calculated as shown in (5).

Definition 5 (Record Sensitivity). *The record sensitivity measures the effect on all records in D when deleting a record i. The sensitivity of record i denoted as ΔCS_i can be calculated by the following formula*:

$$\Delta CS_i = \sum_{j=0}^{l} |w_{ij}| \left(\left\| Q(D^j) - Q(D^{-j}) \right\|_1 \right), \tag{6}$$

in which D^j represents the dataset with r_j and D^{-j} represents the dataset with r_j deleted from D.

Correlated sensitivity is related to query Q. It lists all the records q responding to Q and selects the maximal record sensitivity as the correlated sensitivity as shown in (6) and (7).

Definition 6 (Correlated Sensitivity). *The correlated sensitivity of the single-party dataset, which contains q data records, is denoted as*:

$$\Delta CS_q = \max_{i \in q} \left(\Delta CS_i \right). \tag{7}$$

Compared with the correlated sensitivity, the global sensitivity ΔGS only measures the maximum number of correlated records without considering the correlated degree of the records. Thus, by measuring the correlated degree, ΔCS reduces the global sensitivity to achieve noise reduction.

3.4 Vertically Partitioned Multiparty Dataset

The vertically partitioned relational dataset is defined as follows [33]: N parties of $P_1, P_2, \dots, P_N$ collaboratively release the merged dataset, and each party P_i $(1 \leqslant i \leqslant N)$ holds the local dataset $D_i(ID, F_i)$. N local datasets have the same group of identification feature ID, i.e., corresponding to the same group of individuals, and any two local datasets do not contain the same feature.

Differentially private data release of a vertically partitioned relational dataset is that for given N local datasets and privacy budget ϵ, N parties release the merged dataset that contains features $\bigcup_{i=1}^{N} F_i$ and collaboratively satisfies ϵ-differential privacy to ensure that the released merged dataset does not disclose any individual information about each party's local dataset.

3.5 Problem Statement

In multiparty scenarios, to solve the problem of differentially private data release, one rough solution is to directly merge the datasets and conduct differentially private operations. However, this solution produces immense computational costs due to the high-dimensional issue and hinders the accuracy of training models caused by redundant features [47]. In addition to redundant features and high dimensionality, adding noise to satisfy differential privacy degrades the performance of the model. In particular, data correlation introduces additional noise and reduces the likelihood that the correlation has a positive effect on improving the training accuracy. Previous studies indicate that the increment of the features' number eliminates the correlation of data to some extent and achieves the effect of noise reduction [40]. Therefore, how to merge the feature sets of different local datasets and how to make an appropriate feature selection have significance to our research.

The problem explored in this paper is defined as follows. Assume that N local datasets $D_i(ID, F_i)$ for $1 \leqslant i \leqslant N$ belong to different parties of data holders and the records with the same feature ID correspond

to the same individual. For any two local datasets D_m and D_n, where $1 \leqslant m < n \leqslant N$, the feature sets satisfy $F_m \cap F_n = \emptyset$. We consider the possible data correlation in multiparty datasets. Different data records might be partially correlated. There is a certain degree of correlation between two individuals with different id, such as $i, j \in id$, which is reflected in the partially equivalent feature values of record i and record j.

We attempt to design a method for multiparty datasets release in machine learning. According to the given privacy budget ϵ, the method should release the merged dataset $D_1\{f_1, \ldots, f_n\} \cup D_2\{f_{n+1}, \ldots, f_{n+l}\} \cup \ldots \cup D_N\{f_{n+l+1}, \ldots, f_{n+l+\ldots+m}\} = D'\{f'_1, \ldots, f'_n\}$ with the best features, which effectively achieves the optimal training accuracy and reduces the correlated sensitivity, or release private ML algorithms based on the best features. As shown in Fig.1, two vertically segmented multiparty datasets D_m and D_n have disjoint feature sets. We intend to perform features selection and differentially private operation on merged dataset $D = D_m\{f_m = f_1, f_2, f_3\} \cup D_n\{f_n = f_4, f_5, f_6\}$, and then output perturbed D' with f_{best}. Here, $f_{\text{best}} = \{f_1, f_3, f_5, f_6\}$ is the best feature set after selection to reduce correlation between records and feature redundancy.

4 Multiparty Correlation Analysis

The correlated sensitivity methods achieve satisfactory results in single-party scenarios but still present some weaknesses in multiparty scenarios as follows.

• The existing methods of correlated degree calculation seem to reveal poor interpretability for correlation analysis in multiparty scenarios. In addition, in the process of feature selection, Pearson's correlation coefficient cannot provide an accurate metric, that is, how much correlation is reduced by increasing features. [40] selects features by training the model repeatedly and comparing the precision, thereby involving excessively high computational complexity, especially in the context of high-dimensional data, where it would bring a greater burden.

• Correlation threshold plays an important role in generating correlated degree matrix and calculating correlated sensitivity, but the setting of the correlation threshold in existing studies is subjective to a certain extent.

To solve these problems, we propose the concept of multiparty correlated sensitivity via multiparty correlation analysis.

• We propose a feature-oriented correlated degree

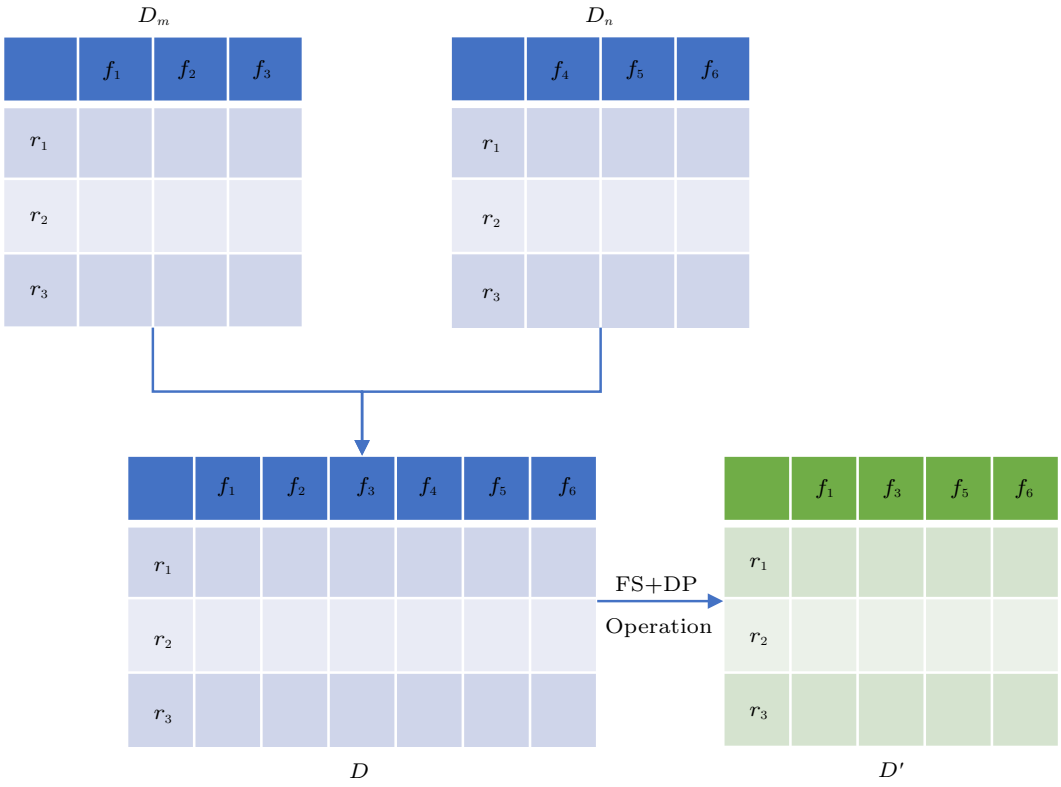


Fig.1. Example of the problem statement.

(FCD) calculation method as shown in (8) and (9), which extends the correlation analysis from single-party scenarios to multiparty scenarios.

• For choosing a more reasonable threshold in the multiparty case, we define the mean of multiparty correlated degree (MCD) as the correlated degree measurement standard. The standard not only reflects the correlation trend of overall multiparty data but also effectively reduces the correlated sensitivity of the merged dataset by defining more rigorous standards.

According to the requirements of multiparty datasets and the process of features selection, we calculate the correlated degree of records by measuring the matching degree of all feature values among data records. Obviously, when the merged multiparty data feature set shrinks, the correlated degree may increase, and vice versa. This change can be measured. But traditional methods such as Pearson's correlation coefficient cannot provide an objective measurement. Meanwhile, such methods as Pearson's correlation coefficient and other methods based on the distance can only measure the relationship of linear correlation. It cannot be measured if there is no linear correlation.

Definition 7 (Feature-Oriented Correlated Degree, FCD). *According to the matching degree of all feature values of record i and record j, the correlated degree w_{ij} of record i and record j is as follows*:

$$w_{ij} = \frac{match\,(i,j)}{l}, \tag{8}$$

$$match(i,j) = \begin{cases} 1, & if \;\; v_i^m = v_j^m \;\; for \;\forall m \in l, \\ 0, & otherwise. \end{cases} \tag{9}$$

Definition 8 (Mean of Multiparty Correlated Degree, MCD). *For an n-party dataset, the correlated sensitivity of one local dataset is denoted as ΔCS_n, and it is the sum of the correlated degree of K correlated records in the local dataset. Therefore, the mean correlated degree of the local dataset is denoted as $\overline{\Delta CS_n} = \frac{\Delta CS_n}{K}$, and then the mean of n-party correlated degree is*

$$MCD = \frac{1}{N}\sum_{k=1}^{N}\overline{\Delta CS_k}. \tag{10}$$

MCD measures the mean level of the correlated degree of multiple-feature datasets vertically partitioned in multiparty data scenarios. As $w_{ij} \in [0,1]$, and $\Delta CS_n \in [0,1]$, the value of MCD is also between 0 and 1, which reflects the trend of correlation of the merged dataset. Since the correlation threshold provided by traditional methods is subjective [40], it is reasonable to consider MCD as the correlation threshold. When the correlated degree of the records is higher than MCD, there is a correlation between two records in the merged dataset, and then the correlated degree w_{ij} between the records is marked according to (5); otherwise, the correlated degree is set to 0.

We then summarize the correlated sensitivity calculation by multiparty correlation analysis. For an n-party dataset, the correlated sensitivity ΔCS_n of all local datasets is calculated, and then MCD is calculated according to (10). Considering MCD as the correlation threshold and refreshing the correlated degree matrix, the multiparty correlated sensitivity of the merged dataset ΔCS_p can be calculated as follows:

$$\Delta CS_p = \max_{i\in p}\sum_{j=0}^{l}|w_{ij}|\left(\left\|Q(D^j) - Q(D^{-j})\right\|_1\right), \tag{11}$$

where D_p represents the merged n-party dataset, w_{ij} represents the correlated degree between record i and record j, D^j represents the dataset with r_j, and D^{-j} represents the dataset with r_j deleted from D_p.

When a dataset differs from its neighbor by only one record, ΔCS_p measures the maximum impact on all records in the merged dataset. For any query Q, the perturbed output of differential privacy based on the Laplace mechanism can be calculated as follows:

$$\hat{Q}(D_p) = Q(D_p) + \mathrm{Lap}\left(\frac{\Delta CS_p}{\varepsilon}\right). \tag{12}$$

MCD is a reasonable standard for measuring the correlated degree between two data records and a more rigorous threshold for calculating the correlated sensitivity. Therefore, MCD effectively reduces the correlated sensitivity, which reduces the noise introduced. According to existing research, such as [40], the correlation threshold in the Adult dataset is set to 0.9. By matching the values of the two records, if FCD is 0.9, 90% of the feature values are equivalent. MCD calculated by our method is approximately 0.92. Therefore, MCD provides a more rigorous threshold that reduces the correlated sensitivity for the same condition. Theorem 1 shows that the calculated multiparty correlated sensitivity proposed in this paper is not greater than the correlated sensitivity calculated by the existing method.

Theorem 1. *For any query Q, the calculated multiparty correlated sensitivity ΔCS_p is equal to or less than the correlated sensitivity ΔCS.*

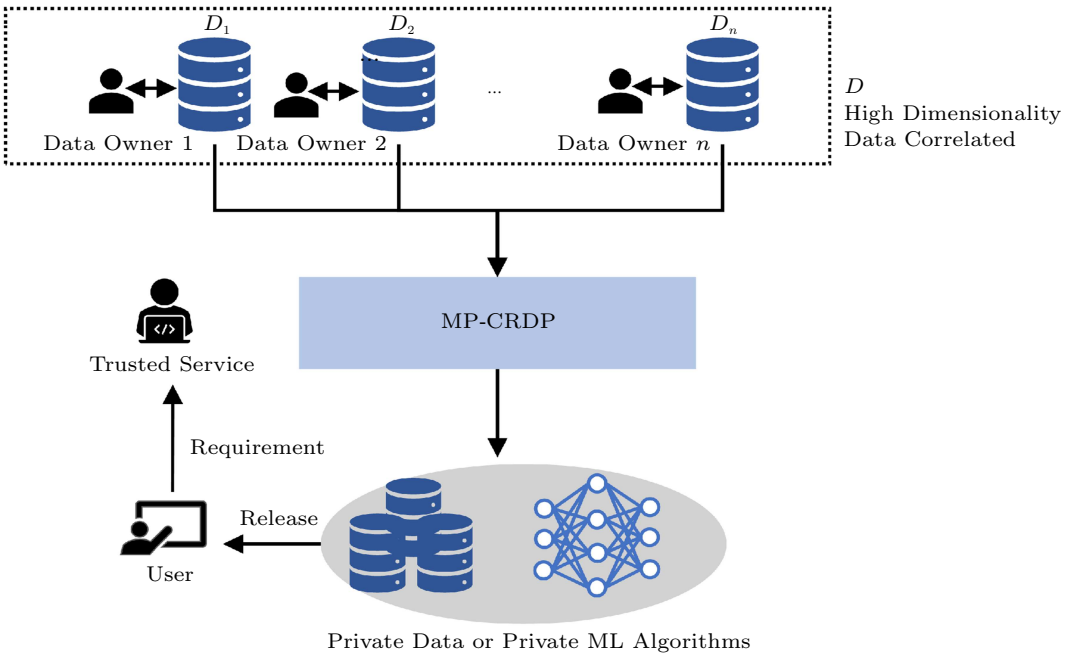
Proof. Assume that ΔCS and ΔCS_p are the correlation analysis results of the same dataset, which are calculated by the same correlated degree matrix $\boldsymbol{\Phi}$. Since

$MCD > w_0$, the number of correlated records $k \geqslant k_p$, and then $\Delta CS_p \leqslant \Delta CS$. □

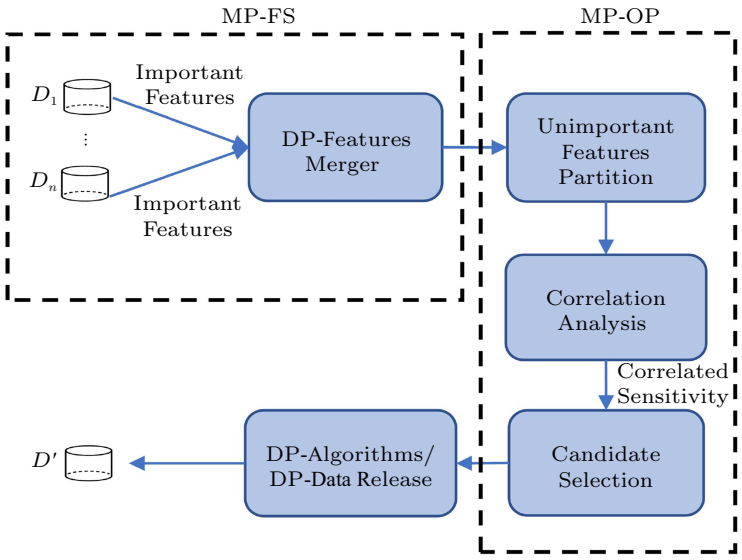
5 Multiparty Correlated Differential Privacy

5.1 Overview

In this paper, we propose a novel correlated differential privacy of the multiparty data release method. This method provides a solution of private data or private ML algorithms release for vertically partitioned multiparty datasets. As shown in Fig.2(a), our method consists of three roles: n-party data owners, each of which is derived from different data sources and owns a local dataset of a group of same individuals with non-overlapping features; a trusted service, which provides private ML algorithms or private data to obtain a better accuracy of data mining; and the users who have



(a)



(b)

Fig.2. (a) Overview of the proposed method MP-CRDP. (b) Key steps of MP-CRDP.

the special data requirements for use and enjoy private algorithms or data service. MP-CRDP performs best features selection and optimization with multiparty datasets and releases private data or private ML algorithms. In particular, MP-CRDP provides two distinct ways of noised data or noised algorithms according to the users' needs.

Our method focuses on the following two aspects: private feature selection in multiparty datasets to improve the machine learning accuracy and data utility optimization by reducing the correlated sensitivity. In response to the previously mentioned problems and to improve the data training accuracy, MP-CRDP selects the best features and narrows the dimensionality of multiparty data feature sets. Based on the differential privacy technology, the multiparty data feature sets are merged. To reduce the noise introduced by the private operation, the multiparty data utility optimization operation is carried out to reduce the sensitivity. The critical steps of our method are shown in Fig.2(b). The private multiparty feature selection (MP-FS) selects the best features in the multiparty datasets. The private multiparty utility optimization (MP-OP) groups the adjusted features as candidates, analyzes the data correlation, selects candidates to adjust the features via the defined utility functions, and releases private ML algorithms and private data.

5.2 Private Multiparty Feature Selection

In traditional machine learning algorithms, feature selection is significant for reducing the data dimensionalities and improving accuracy. Therefore, we merge the vertically partitioned multiparty datasets by performing a selection operation on the feature sets. We choose the stability feature selection algorithm [40] in our method for overcoming the over-fitting problem in the process of feature selection. Stability selection is a commonly employed feature selection method based on a combination of sub-sampling and a selection algorithm. Since there are naturally multiple sub-datasets, the algorithm is the best choice in a multiparty environment. By repeatedly running a feature selection algorithm on different datasets and counting the frequency that a feature is considered to be an important feature, feature selection scores are calculated. It is the ratio of the times of the feature selected to be an important feature to the times tested, as shown in (13).

In the n-party dataset, the important feature score of a specific feature S_i is the ratio of the important features' frequency denoted as T_{freq} to the number of sub-datasets denoted as $N_{n\text{-party}}$ as follows:

$$S_i = \frac{T_{\text{freq}}}{N_{n\text{-party}}}. \tag{13}$$

The score is between 0 and 1. The score of an important feature is infinitely close to 1, and that of a useless feature is close to 0.

The weakness of the method is that it retains similar and correlated important features. To overcome the drawbacks of the method, we use Pearson's correlation coefficient calculation in MP-FS. It should be noticed that in this work, we use two concepts of correlation. One refers to the correlated degree between records, which is widely used in correlated differential privacy. The other is the term of Pearson's correlation coefficient, which is mainly used to measure the linear correlation between features. These two concepts are different. We do not use Pearson's correlation coefficient to measure the correlation between the records but use the proposed feature-oriented correlated degree, which is more applicable for the multiparty scenario. (14) calculates the Pearson's correlation coefficients of features f_m and f_n as follows:

$$p_{m,n} = \frac{E\left[\left(f_m - \mu_m\right)\left(f_n - \mu_n\right)\right]}{\sigma_m \sigma_n}, \tag{14}$$

where μ_m and μ_n denote the mean values of f_m and f_n respectively, and σ_m and σ_n denote the standard deviations of f_m and f_n respectively.

The proposed MP-FS algorithm, shown as Algorithm 1, is described as follows:

• sequentially cleaning up the features with a high probability of missing data and the features with a single data value;

• traversing each local dataset, calculating ΔCS_n and MCD, taking out the features in each local dataset in turn, running the feature selection algorithm, and calculating the stability scores;

• adding features with stability scores greater than the important feature threshold to the best feature set; otherwise adding it to the adjusted feature set;

• merging datasets according to the best feature set, calculating the Pearson's correlation between any two features, adding noise according to the sensitivity of correlation coefficient, picking one of the features exceeding the Pearson's correlation threshold randomly, and moving it to the adjusted feature set.

Algorithm 1 selects the best features to achieve the best training accuracy, initially merges multiparty data,

and returns the best feature set β and the adjusted feature set α. According to the multiparty data correlation analysis in Section 4, the correlation of the merged datasets introduces additional noise after the private operation. Therefore, we need to reduce the correlation and improve the data utility by properly adjusting the operation of features. Based on the differential privacy technology, our algorithm MP-CRDP divides the privacy budget ϵ into two parts: ϵ_1 is used for differentially private feature selection, while ϵ_2 is used for differentially private utility optimization.

Algorithm 1. Private Multiparty Feature Selection l

Input: n-party dataset D_n, ϵ_1, important feature threshold θ_{im}, Pearson's correlation threshold θ_{per}, initial w_0
Output: best feature set β, adjusted feature set α, and w_0

for $D_n, n = 1, 2, \ldots, N$ **do**
  Remove features with a high percentage of missing values and single values
  Calculate ΔCS_n and MCD; `// according to (10)`
  $w_0 = MCD$
end
for $D_n, n = 1, 2, \ldots, N$ **do**
  Create a temporary feature set $\tau \in \emptyset$
  for $f_m, m = 1, 2, \ldots, M$ **do**
    Add f_m and other features of D_n to temporary feature set τ
    Group data with temporary features and refresh the datasets
    Call feature selection algorithm on the refreshed datasets and D_n
    Calculate score of important S_m; `// according to (13)`
    if $S_m \geqslant \theta_{\text{im}}$ **then**
      Add f_m to best feature set β
    else
      Add f_m to adjusted feature set α
    end
  end
end
for $i \in \beta$ **do**
  Group data with best feature set as D_p
  Calculate $p^i_{m,n} \in P, \forall m, n \in \beta, i = 1, \cdots, l$
  Calculate sensitivity of Pearson's correlation coefficient ΔCS_{per}; `// according to (15) and (16)`
  Add Laplace noise to P according to $\hat{p}_{m,n} = p_{m,n} + \text{Lap}(\frac{\Delta CS_{\text{per}}}{\epsilon_1})$
  if $\hat{p}_{m,n} \geqslant \theta_{\text{per}}$ **then**
    Remove f_m or f_n from adjusted feature set α randomly
  end
end
Return best feature set β, adjusted feature set α, and w_0

In the last step of MP-FS, for private feature selection, we separately calculate the linear correlation between the features of the merged dataset and its neighboring dataset and obtain the groups of Pearson's correlation coefficients P and P', where $p^i_{m,n} \in P$ and ${p'}^i_{m,n} \in P'$, $\forall m, n \in best\ feature$, $i = 1, \cdots, l$. We introduce the concept of record sensitivity and sensitivity of Pearson's correlation coefficient as shown in (15) and (16) respectively.

Definition 9 (Record Sensitivity of Pearson's Correlation Coefficient). *For a query Q, the record sensitivity of Pearson's correlation coefficient of r_i can be defined as*

$$\Delta CS_{\text{per}_i} = \max_{\forall m,n \in p} \left\| p_{m,n} - {p'}_{m,n} \right\|_1, \tag{15}$$

where $p_{m,n}$ and ${p'}_{m,n}$ denote the correlation coefficients of features f_m and f_n in neighboring datasets, respectively.

Definition 10 (Sensitivity of Pearson's Correlation Coefficient). *For a query Q, the sensitivity of Pearson's correlation coefficient is determined by the maximal record sensitivity of Pearson's correlation coefficient*

$$\Delta CS_{\text{per}} = \max_{i \in D_\beta} \left(\Delta CS_{\text{per}_i} \right), \tag{16}$$

where Q denotes a query about a set of Pearson's correlation coefficients of records. It is easy to know that the sensitivity of the correlation coefficient $\Delta CS_{\text{per}} \leqslant 1$, because the correlation coefficient value ranges from 0 to 1.

5.3 Private Multiparty Utility Optimization

After feature selection according to MP-FS, we retain some features that are most relevant to the training accuracy of the model. However, considering the data correlation issue caused by multiparty data release, removing more features generally leads to a higher correlation. The differential privacy mechanism introduces additional noise and consequently reduces the data utility with the same privacy level. Therefore, the goal of the MP-OP algorithm is to add a certain number of features in the best feature set β to reduce the impact of data correlation. The adjustment coefficient b determines the relaxation degree of the best feature set. We can set the value of b according to the scales of the adjusted feature set of the specific dataset. According to the value of b, the adjusted feature set α is divided into the subsets of the feature combination to generate

several candidate schemes c. To explore the effects of additional noise and redundant features on the training accuracy of the model, we define two utility functions based on information gain and correlated sensitivity.

Information gain is an effective means for investigating the importance of features to model classification. The information gain utility function's design idea is to add features with high information gain while reducing data correlation and improving the training accuracy. We divide the feature set α according to the adjustment coefficient b, determining the number of features in c_i. The information gain of the candidates is the sum of b features' information gains. We obtain the utility function u_1 as follows

$$InfoGain\left(D_p, c_i\right) = \sum_{b} \left(H\left(D_p\right) - H_{p|f_i}\left(D_p\right)\right), \quad (17)$$

in which $H\left(D_p\right)$ is the initial information entropy of the merged dataset D_p, $H\left(D_p\right) = -\sum_{cls} \frac{\left|D_p^{cls}\right|}{|D_p|} \log_2 \frac{\left|D_p^{cls}\right|}{|D_p|}$, $H_{p|f_i}\left(D_p\right)$ is the conditional entropy of increasing features f_i in c_i, v is the value of f_i, and then $H_{p|f_i}\left(D_p\right) = -\sum_{v} \frac{\left|D_p^{v}\right|}{|D_p|} H\left(D_p^{v}\right)$.

Another way to define the utility function is based on the multiparty correlated sensitivity introduced in Section 4, which optimizes the data utility by maximally reducing the sensitivity and minimizing the amount of noise intake. According to the correlation matrix and (11), we can obtain the multiparty correlated sensitivity $\Delta CS_p^{c_i}$ of candidate scheme c_i. Therefore, the candidate set of the minimum multiparty correlated sensitivity defined by utility function u_2 is selected with a high probability. We obtain the utility function u_2 as follows:

$$u_2\left(D_p, c_i\right) = \frac{MCD}{\Delta CS_p^{c_i}}. \quad (18)$$

On the basis of the best feature set β, both two utility functions aim to improve utility by adding several features such as b features, where b is determined by the scale of the adjusted feature set of the dataset. The design idea of the information gain utility function, i.e., u_1, is to add b features with high information gain while reducing data correlation and improving the training accuracy, while another utility function, i.e., u_2, optimizes the data utility by maximally reducing the sensitivity and minimizing the amount of noise intake. We plan to explore the influence of the two utility functions on the training accuracy through experiments. According to the previous two utility functions, we select candidates with higher information gain and lower data correlation levels using the exponential mechanism of differential privacy as shown in (19). MP-OP simultaneously maintains excellent practicability for data release and analysis.

Given the utility scores of all candidate sets, the probability of selecting candidate c_i by the exponential mechanism is expressed as follows:

$$\frac{\exp\left(\frac{\epsilon u(D_p, c_i)}{2\Delta u}\right)}{\sum_{c_i \in c} \exp\left(\frac{\epsilon u(D_p, c)}{2\Delta u}\right)}. \quad (19)$$

The proposed MP-OP algorithm, shown as Algorithm 2, is described as follows:

- combining the features of the adjusted feature set according to the adjusted feature coefficients to generate candidates;
- calculating the scores of candidates according to the two utility functions;
- selecting high-scoring candidates using the exponential mechanism, adding the features of the candidates to the best feature set, and updating the dataset;
- performing differentially private operations, and calculating correlated sensitivity, and then adding noise to the queries value of released dataset or the weight of the corresponding machine learning algorithm;
- training the model to get the prediction result.

Algorithm 2 takes the best feature set β, the adjusted feature set α, and w_0 obtained from Algorithm 1, b defined according to relevant research experience, and divided privacy budget ϵ_2 as input to achieve private multiparty utility optimization. The algorithm outputs the best feature set β, multiparty data D'_p, which obtains the best training precision or private ML algorithms based on differential privacy technology. Since MP-OP considers two schemes of maximum information gain and minimum noise respectively, the output term contains two schemes. On line 10 of Algorithm 2, the interactive mechanism can be employed to add noise to each query. The query Q here is a function that maps the dataset D to a real number. We provide a Laplace mechanism to add the noise to mask the difference on query Q between the neighboring datasets. For the private ML algorithms release, noise is added to the weights of the algorithms to satisfy differential privacy according to different algorithms.

5.4 Privacy Analysis

By analyzing the steps consuming the privacy budget, we prove that the proposed MP-CRDP satisfies ϵ-

differential privacy and analyze the sensitivity of each differentially private operation.

Algorithm 2. Private Multiparty Utility Optimization

Input: n-party dataset D_n, best feature set β, adjusted feature set α, w_0, ϵ_2, adjusting coefficient b

Output: best feature set β, optimal D_p, private ML

Divide α according to b to generate several candidate sets $c_i \in c$

for $c_i, i = 1, 2, \ldots, k$ **do**

  Calculate score of utility u_1^i and u_2^i; `// according to (17) and (18) respectively`

  Select $c_1^\beta \in c$ and $c_2^\beta \in c$ with probability $\propto \exp\left(\frac{\frac{\epsilon_2}{2} u(D,\, c_i)}{2\Delta u}\right)$

  Add features of c_1^β and c_2^β to β respectively

  Update D_1^p and D_2^p

end

for $D_n, n = 1, 2, \ldots, N$ **do**

  Calculate ΔCS_p according to (11)

  Add Laplace noise $\mathrm{Lap}(\frac{2\Delta CS_p}{\epsilon_2})$

  Train the datasets and get the predicted results

end

Return best feature set β_1 and β_2, optimal D_1' and D_2'; or private ML based on two kinds of schemes

In MP-FS, we perform a differentially private operation for private feature selection on the merged dataset. $Q_1(.)$ is the query of Pearson's correlation coefficient of any two features in both neighbors' datasets D and D', where D differs from D' by only one single record.

According to the Laplace mechanism, we have

$$M_1\left(x, Q_1\left(.\right), \epsilon_1\right) = Q_1\left(x\right) + \mathrm{Lap}\left(\frac{\Delta CS_{\mathrm{per}}}{\epsilon_1}\right). \quad (20)$$

Let x, y be two neighboring datasets. We compare two random points $z \in R$ and the ratio of two probability density can be presented as

$$\frac{p_x\left(z\right)}{p_y\left(z\right)} = \prod_{i=1}^{N} \frac{\exp\left(-\frac{\epsilon_1 |Q_1\left(x\right)_i - z_i|}{\Delta CS_{\mathrm{per}}}\right)}{\exp\left(-\frac{\varepsilon_1 |Q_1\left(y\right)_i - z_i|}{\Delta CS_{\mathrm{per}}}\right)} \leqslant \exp\left(\epsilon_1\right). \quad (21)$$

Therefore, MP-FS satisfies ϵ_1-differential privacy and only introduces a small amount of noise because the sensitivity $\Delta CS_{\mathrm{per}} \in [0, 1]$, as shown in (20) and (21).

In MP-OP for utility optimization, the first step is to select the candidate feature set by the exponential mechanism, and the second step is to release private data or ML algorithms, each of which consumes the privacy budget $\frac{\epsilon_2}{2}$.

For the first step of MP-OP, we allocate the $\frac{\epsilon_2}{2}$ privacy budget for candidate selection by the exponential mechanism. We perform further analysis of the sensitivity for the utility functions as follows.

For u_1, according to the concept of information gain, since $H\left(D_p\right) \in [0, \log_2\left(\pi\left(cls\right)\right)]$, and $H_{p|f_1}\left(D_p\right) \in [0, H\left(D_p\right)]$, the sensitivity $\Delta u_1 = \log_2\left(\pi\left(cls\right)\right)$, where $\pi\left(cls\right)$ is the range of the classified feature.

For u_2, according to the concept of MCD, since $u_2^i = 0$, when MCD is 1, the value of Δu_2 takes the maximum, and therefore, the sensitivity $\Delta u_2 = \frac{1}{K}$, where K is the number of correlated records in the merged dataset.

In the second step of MP-OP, when the private data is released, we add noise to the count queries $Q_2(.)$. When releasing a private ML algorithm, $Q_2(.)$ returns the weights of the specific ML algorithm, such as the LR and SVM model. We make the algorithm satisfy differential privacy by perturbing an objective function, such as FM-regression and Objective SVM.

According to the Laplace mechanism, for both situations we have

$$M_2\left(x, Q_2\left(.\right), \frac{\epsilon_2}{2}\right) = Q_2\left(x\right) + \mathrm{Lap}\left(\frac{2\Delta CS_p}{\epsilon_2}\right). \quad (22)$$

The ratio of two probability density can be presented as

$$\frac{p_x\left(z\right)}{p_y\left(z\right)} = \prod_{i=1}^{N} \frac{\exp\left(-\frac{\frac{\epsilon_2}{2} |Q_2\left(x\right)_i - z_i|}{\Delta CS_p}\right)}{\exp\left(-\frac{\frac{\varepsilon_2}{2} |Q_2\left(y\right)_i - z_i|}{\Delta CS_p}\right)} \leqslant \exp\left(\frac{\epsilon_2}{2}\right). \quad (23)$$

Therefore, for releasing private data or ML algorithms, the $\frac{\epsilon_2}{2}$ privacy budget is consumed in MP-OP, as shown in (22) and (23). The sensitivity is the multiparty correlated sensitivity (i.e., ΔCS_p) proposed in this paper.

It should be explained that the queries in feature selection and machine learning are from different persons. The DP in feature selection is to protect data from other parties inside the system, and the DP in machine learning is to protect data from adversaries outside the system. However, we still need to split the privacy budget because the allocation of the privacy budget is only related to the operands of the algorithm. The operand of the two algorithms in MP-CRDP is

D. According to the privacy analysis and Sequential Composition, MP-CRDP divides the privacy budget ϵ into ϵ_1, $\frac{\epsilon_2}{2}$ and $\frac{\epsilon_2}{2}$. Therefore, MP-CRDP satisfies ϵ-differential privacy. Explicitly, noise adding that impacts the accuracy of privacy-preserving models and data utility only happens in the second step. Thus, our method only consumes a small amount of noise while holding ϵ-differential privacy.

6 Experiments

We verify the effectiveness of MP-CRDP using two sets of experiments. The first set is aimed at the released private ML algorithms to verify the machine learning model's effectiveness of our method and other correlated differential privacy methods at the same privacy protection level, such as the training accuracy of LR and SVM. The second set is aimed at the data utility of the released private data to verify the average query accuracy of our method and the compared methods at the same privacy protection level.

6.1 Experimental Setup

Regarding the choice of compared methods, we focus on the efficiency of different correlated differential privacy methods in processing the correlation of multiparty datasets and analyze the impact of correlation processing mechanisms of different methods on the training accuracy and the utility of the released data. Different mechanisms and inappropriate feature selection operate higher sensitivity and introduce additional noise, which reduces the accuracy of the training model and the effectiveness of data release. Different correlated differential privacy methods include Group DP [27] and Correlated DP [28]. The former provides global sensitivity by calculating the number of correlated records, and the latter calculates the weighted correlated sensitivity by analyzing the correlated degree. Our method analyzes the multiparty data correlation, as shown in Section 4, and provides a rigorous correlation threshold, which reduces the multiparty correlated sensitivity in the weighted calculation.

With respect to dataset selection, we use the common machine learning datasets to arrange the experiment, such as the Adult dataset and the Breast Cancer dataset.

• *Adult.* The Adult dataset from the UCI Machine Learning Repository① originally has 48 842 records and 14 attributes with the label about whether a person's annual salary exceeds 50k. During the data preprocessing, we drop the duplicates and the records with missing or illegal values, and then perform dummy variable conversion and discretize continuous data. In the experiments of verifying the effectiveness of our methods in the ML model and query accuracy, we select 10 000 records with 12 features that are more correlated to other records to get closer to the real-life situation where ubiquitous data correlation exists, especially in multiparty data scenarios.

• *Breast Cancer.* The Breast Cancer dataset from UCI Machine Learning Repository① contains 569 records with 32 features. It is noted that the first column is the ID which provides the ID number and therefore is irrelevant to the diagnosis. Meanwhile, the second column is a diagnosis that provides the diagnosis of breast tissues (malignant or benign), and we take it as the label. Thus, after dropping the above two columns, we retain the 30 remaining features rather than 32 for our experiment.

To simulate the multiparty scenario presented in this paper, we expand and reconstruct the dataset to some extent. We divide the Adult and Breast Cancer datasets into three local datasets according to the number of features respectively. Each Adult local dataset contains four features, and each Breast Cancer local dataset contains 10 features. The Adult and Breast Cancer datasets are both used to solve binary classification issues. SVM can be directly applied to our datasets. In logistic regression, we map this continuous value to interval (0, 1) through the sigmoid function. Then we set a threshold, divide those mapped values greater than the threshold into one class and the others into the other class.

In terms of machine learning algorithm selection, we choose LR and SVM as the experimental algorithms. In particular, with regard to the utility verification of the private ML algorithms, we compare the sensitivity calculated by different methods and then uniformly perform the private operations for the machine learning algorithms. We choose classical private ML algorithms, such as FM-regression [17] and Objective SVM [19], by adding the noise generated by different correlated differential privacy methods to the weight of the objective functions and compare the training accuracy of the released private LR and SVM.

In terms of parameter selection, we set the initial $w_0 = 0.9$ to calculate the correlated sensitivity of

①http://www.ics.uci.edu/mlearn/MLRepository.html, July 2020.

the local dataset based on the experience of existing research [40]. Then we use the calculated MCD as the correlation threshold of the merged dataset. In Algorithm 2, the adjustment coefficient b determines the relaxation degree of the best feature set to reduce the correlation. According to the scales of the adjusted feature sets of the Adult dataset and the Breast Cancer dataset, we set the value of b to 2 and 4, respectively. Since the number of local datasets is 3, the important feature threshold θ_{im} is set to 0.6. The Pearson's correlation threshold θ_{per} is set to 0.9.

To present the details of the experiment, the processes of best features selection and features adjustment according to the utility function u_1 and u_2 are illustrated in Table 3. After MP-FS, seven best features are selected for the Adult dataset, and 21 best features are selected for the Breast Cancer dataset. MP-OP re-selects two different adjusted features for the Adult dataset and four different adjusted features for the Breast Cancer dataset by u_1 and u_2, respectively. The final best feature set contains the best features selected by MP-FS and the adjusted features selected by MP-OP.

We also need to note that some of the features in the Adult dataset are categorical. We could not use Pearson's correlation coefficient to measure the correlation between features. As a supplement, we use the information gain method to calculate the degree of correlation between features.

6.2 Experiments for ML Algorithm

To verify the effectiveness of MP-CRDP on private ML algorithms, we compare our method with Group DP and Correlated DP to conduct performance inspections. The classification accuracy in machine learning is utilized in this subsection as an indicator to reflect the algorithm's performance to inspect the comparison results and changing trends of the accuracy with different privacy budgets. The classification accuracy is the ratio of the number of samples correctly classified by the classifier to the total number of samples. Both the datasets used in this work are binary classification tasks with the explicit label. Therefore, the LR and SVM model should return the classification results, and then the accuracy is calculated sequentially.

Fig. 3 and Fig. 4 present the experimental results based on the LR and SVM on the Adult and the Breast Cancer datasets, respectively. The five curves in Fig.3 and Fig. 4 correspond to the following five situations: 1) merged dataset with feature selection, non-private-FS; 2) merged dataset without feature selection, non-private; 3) merged dataset with feature selection based on Group DP, Private GS; 4) merged dataset with feature selection based on Correlated DP, Private CS; 5) our method, Private CS_{p}.

As shown explicitly, the accuracy after feature selection has been improved to varying degrees, and the private operations have led to varying degrees of accuracy decline. The feature selection can provide an improvement in accuracy compared with simple data merging. As shown in Fig. 3 and Fig. 4, the accuracy of Non-private-FS is better than the accuracy of Non-private which is the strategy without feature selection. As the privacy budget ϵ increases, the protection level reduces so that the accuracy tends to rise and then stabilizes.

In privacy protection schemes 3, 4, and 5, based on FM-regression and Objective SVM, under the same privacy budget, the accuracy of Private GS is slightly lower than that of Private CS, and they are lower than that of our method Private CS_{p} because the sensitivity reduction of our method reduces the noise introduced. In Fig.3 and Fig.4, the training results of different utility functions on different datasets show that the accuracy of Figs. 3(b) and 3(d) and Figs.4(b) and 4(d) is

Table 3. Process of Selecting and Adjusting Features by MP-CRDP

Dataset	MP-FS	MP-OP
Adult	Best features={'native-country', 'education', 'work-class', 'race', 'relationship', 'marital-status', 'occupation'}	Adjusted features by u_1 = {'education-num', 'age'} Adjusted features by u_2 = {'fnlwgt', 'age'}
Breast Cancer	Best features = {'concave points-mean', 'area-se', 'concavity-worst', 'compactness-se', 'perimeter-worst', 'concavity-mean', 'texture-mean', 'compactness-mean', 'smoothness-worst', 'symmetry-worst', 'symmetry-se', 'concave points-worst', 'perimeter-mean', 'concavity-se', 'radius-worst', 'radius-mean', 'fractal-dimension-worst', 'smoothness-mean', 'radius-se', 'concave points-se', 'texture-worst'}	Adjusted Features by u_1={'area-worst', 'area-mean', 'perimeter-se', 'compactness-worst'} Adjusted features by u_2 = {'texture-se', 'fractal-dimension-mean', 'area-mean', 'symmetry-mean'}

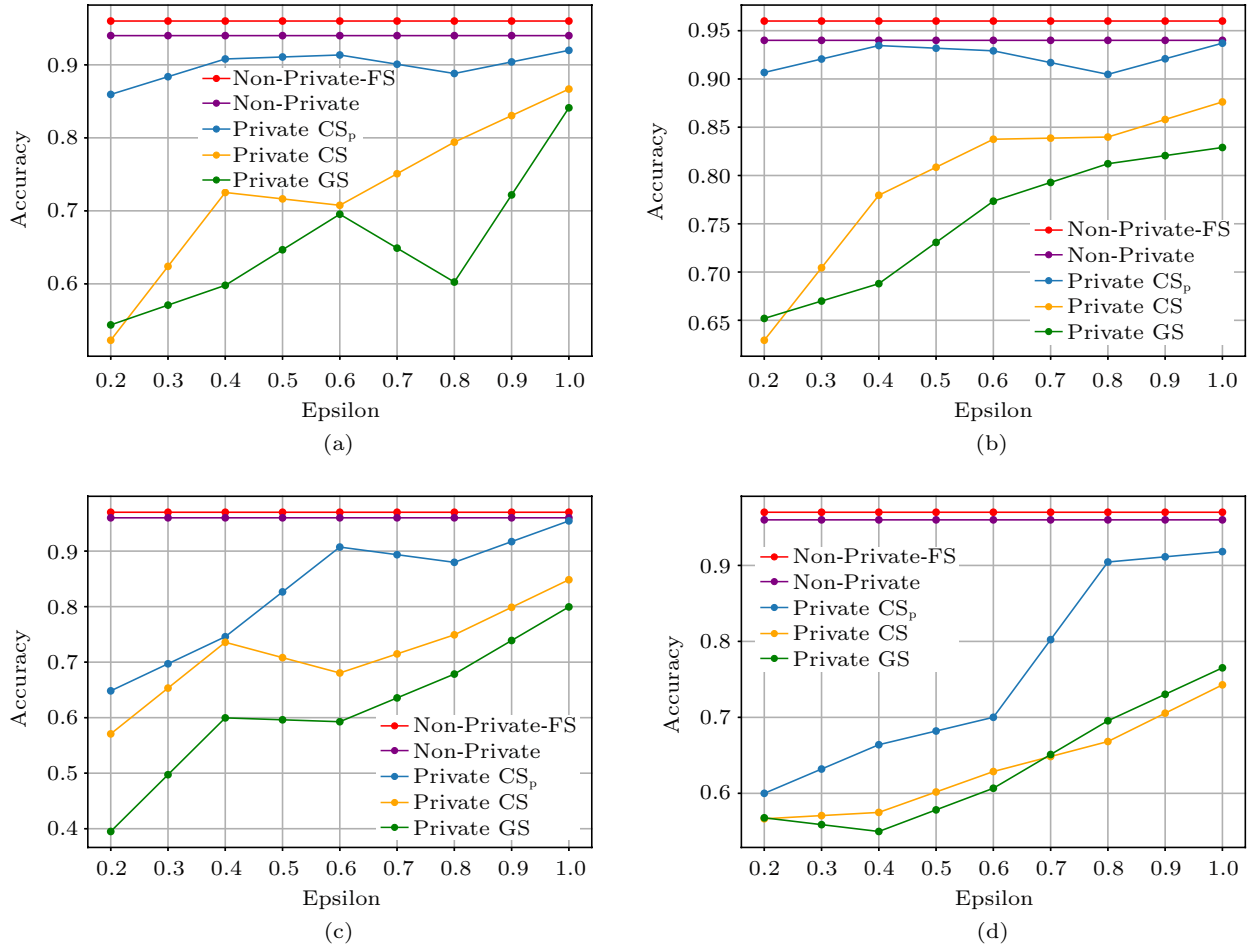


Fig.3. Privacy-accuracy trade-off in LR. (a) Accuracy of training LR model based on u_1 on Adult. (b) Accuracy of training LR model based on u_2 on Adult. (c) Accuracy of training LR model based on u_1 on Breast Cancer. (d) Accuracy of training LR model based on u_2 on Breast Cancer.

slightly better than the accuracy of Figs.3(a) and 3(c) and Figs.4(a) and 4(c) respectively. The results show that the performance of generating candidate sets and merged data based on the utility function u_2 is better than that based on u_1. It indicates that reducing noise intake has a greater impact on improving the accuracy.

6.3 Experiments for Data Release

Since our method provides both private algorithms and the private data release solution, the data utility after release is evaluated for verifying the performance of MP-CRDP. The mean absolute error (MAE) is applied to analyze the count results and the impact of varying the privacy budget. The data utility of D_p is measured by MAE, which is given as follows:

$$MAE = \frac{1}{|Q|} \sum_{Q_i \in Q} \left| \hat{Q}_i(x) - Q_i(x) \right|, \tag{24}$$

where $Q_i(x)$ is the true aggregation result, and $\hat{Q}_i(x)$ is the perturbed aggregation result. A low MAE indicates a low error, and thus, a better data utility. For each dataset, we generate the query set Q with 10 000 random linear queries, for which the value of each feature is randomly searched.

Fig.5 presents the data utility comparison of schemes 3, 4, and 5 in Subsection 6.2. As shown, MAE shows a downward trend as the privacy budget ϵ increases, which indicates that the private data utility increases as the level of privacy protection decreases. Similar to the results of released machine learning algorithms, MAE of our method Private CS$_p$ on different datasets is better than that of Private GS and Private CS. The experimental results also show that generating candidate sets and merged data based on the utility function u_2 is better than that based on u_1. It can be concluded that reducing noise intake also has a positive effect on improving the released data utility.

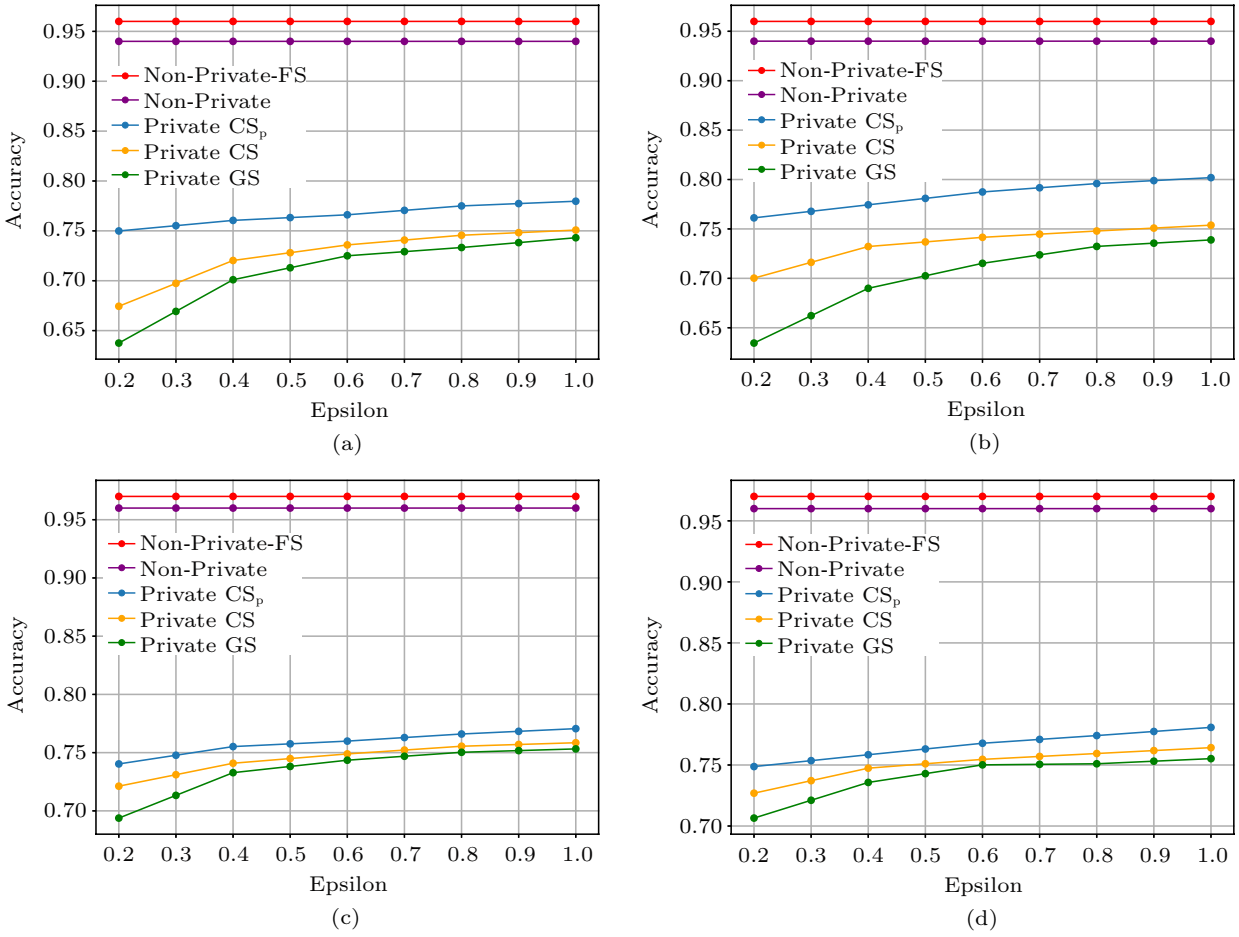


Fig.4. Privacy-accuracy trade-off in SVM. (a) Accuracy of training SVM model based on u_1 on Adult. (b) Accuracy of training SVM model based on u_2 on Adult. (c) Accuracy of training SVM model based on u_1 on Breast Cancer. (d) Accuracy of training SVM model based on u_2 on Breast Cancer.

6.4 Result Analysis

We analyze the experimental results and conclude several advantages of our method for the multiparty correlated data private operation.

• Our method performs feature selection on multiparty datasets, which improves the accuracy of machine learning algorithms. On the one hand, the selection of important features in machine learning significantly impacts the accuracy. On the other hand, redundant features bring performance degradation, and effective dimensionality reduction improves the accuracy in multiparty scenarios.

• In the multiparty scenario, the sensitivity of correlated data increases, and additional noise is introduced, which reduces data utility and model training accuracy. MP-OP effectively reduces the correlation by relaxing the number of features. We perform an experiment to expose the trend of correlation and correlated records on different datasets with the number of features. The results are shown in Fig.6, where the correlation shows a downward trend as the number of features increases.

• By correlation analysis of different correlated differential privacy methods, our method reduces the data correlation more effectively in the case of the same number of features so that the released algorithms and data maintain good utility. Fig.6(a) and Fig.6(c) show the comparison of the correlation between ΔGS, ΔCS, and our method ΔCS_p with the same number of features on different datasets. Fig.6(b) and Fig.6(d) show the comparison of the number of correlated records K. Since ΔGS does not consider the correlated degree, the calculated correlation of ΔGS is higher than that of ΔCS, which is the addition of the weighted correlated degree, but the number of related records K is the same for the two methods. Our method not only considers the correlated degree but also provides more rigorous standards for the correlation threshold. In the multiparty dataset, the correlation of the local dataset pro-

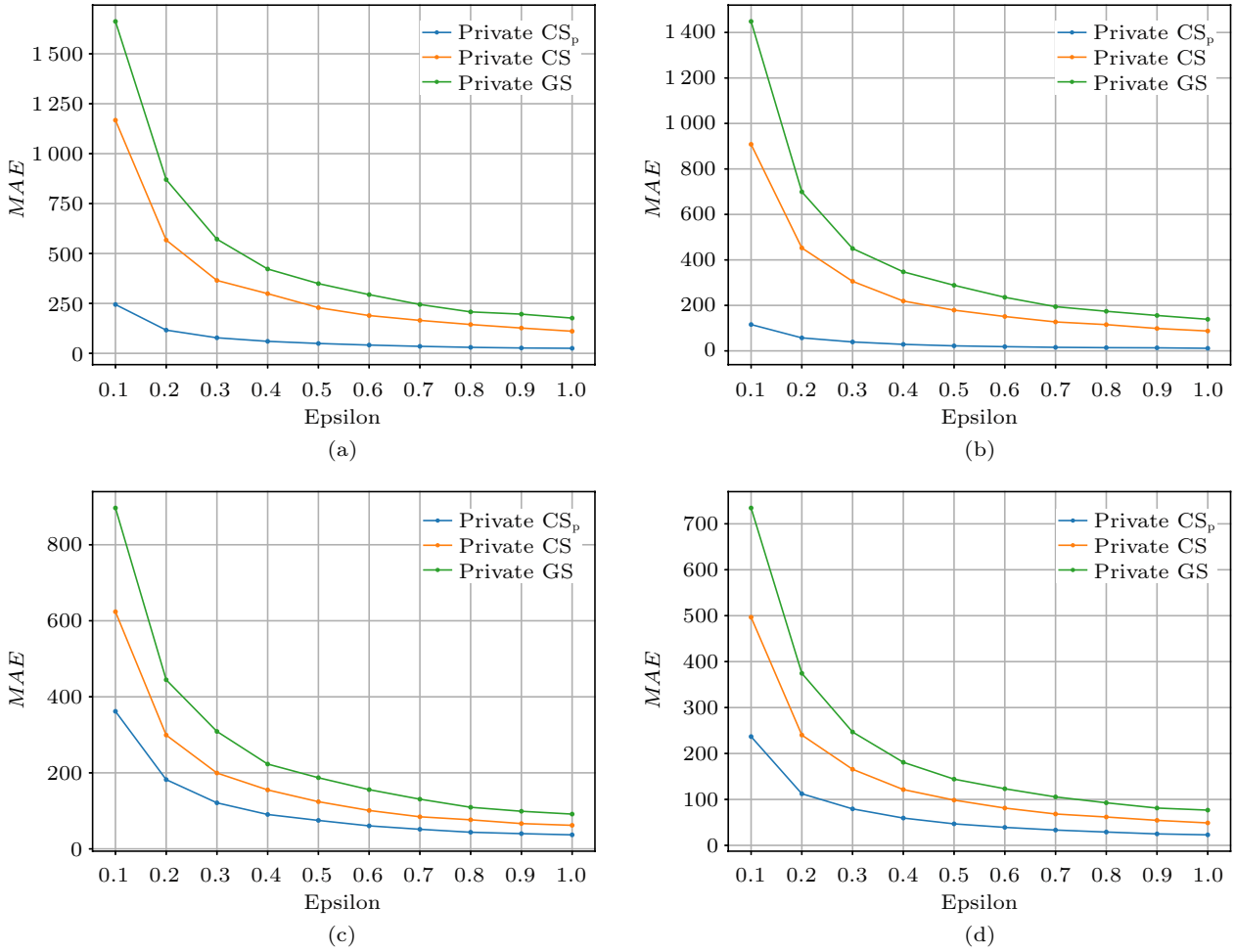


Fig.5. MAE for u_1 and u_2 on different datasets. (a) u_1 on Adult. (b) u_2 on Adult. (c) u_1 on Breast Cancer. (d) u_2 on Breast Cancer.

vides prior knowledge for determining the correlation threshold of the merged dataset, thereby the selection of ΔCS_p is more objective than that of ΔCS. While reflecting the global correlation trend, the correlation threshold of ΔCS_p is more rigorous than that of ΔCS. As shown in Fig.6(b) and Fig.6(d), the number of correlated records K of ΔCS_p is less than that of ΔCS. Therefore, our method more effectively reduces the correlation.

7 Conclusions

In this work, we studied the inherent problem of private machine learning algorithms, which is balancing privacy and utility theoretically and empirically. Concerning reduced data utility due to privacy protection operations in the consolidated high-dimensional and correlated data, MP-CRDP was proposed to optimize data utility by private feature selection and correlated sensitivity reduction operations. Compared with the existing correlated differentially private method, MP-CRDP provides the private querying data or the private ML algorithm to meet the data analysis requirement of users by the design of releasing mechanism and improved the ML model training accuracy effectively.

We also proposed a multiparty correlation analysis technique, which reduces the correlated sensitivity, thereby reducing the DP noise injecting. The existing methods lack considering the correlated degree and defining an objective correlated threshold causes high data correlation. Our method considers the correlated degree and prior knowledge about the correlation of local datasets and provides a more rigorous standard for determining the correlation threshold. Therefore, it effectively reduces the correlated sensitivity.

Comprehensive experiments on different datasets demonstrated that the classification accuracy with different privacy budgets of the proposed method in machine learning was superior to that of other compared algorithms. Moreover, the proposed got low MAE val-

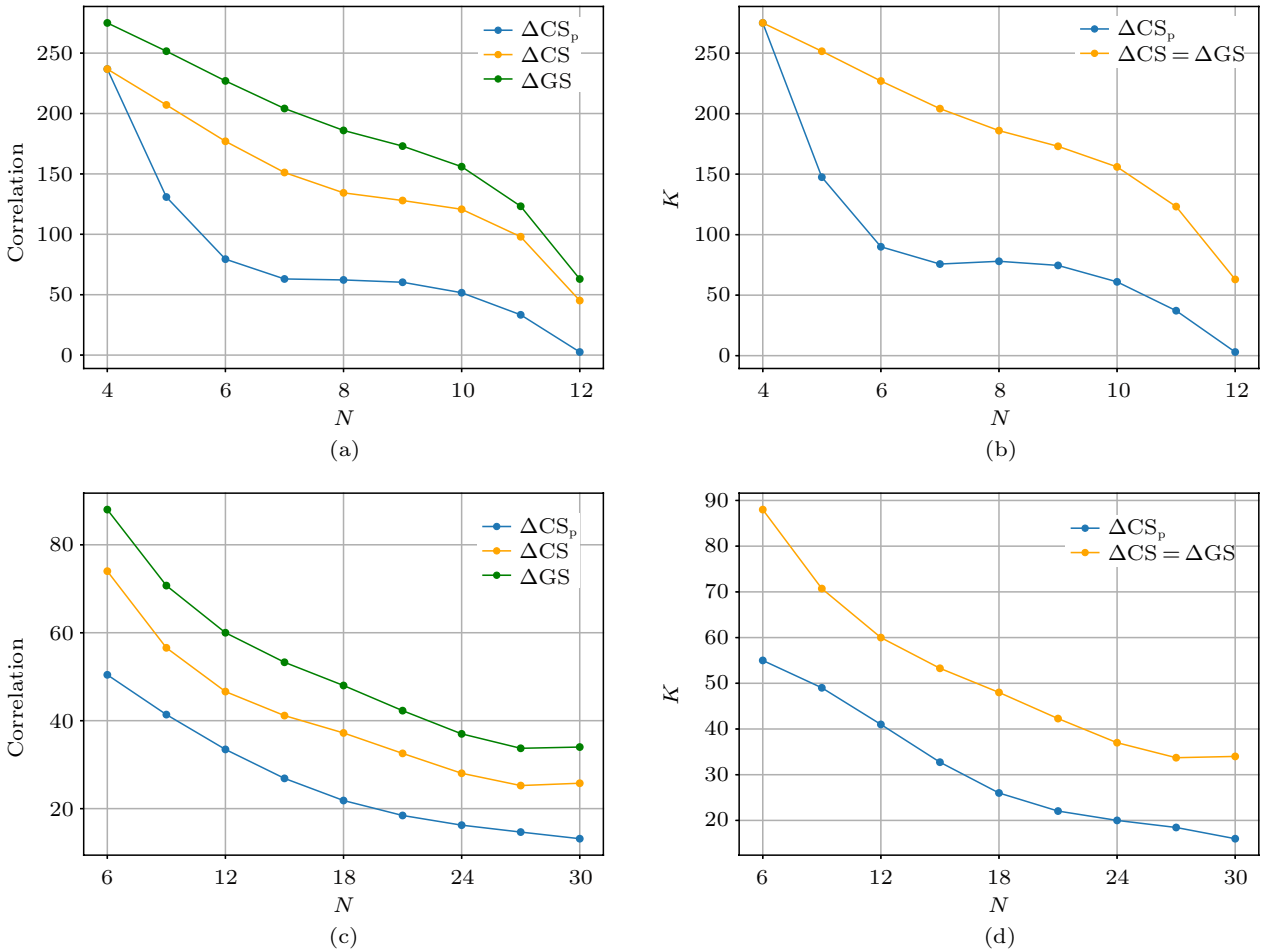


Fig.6. Correlation and number of correlated records K with the number of features N on different datasets. (a) Correlation with N on Adult. (b) Number of correlated records K with N on Adult. (c) Correlation with N on Breast Cancer. (d) Number of correlated records K with N on Breast Cancer.

ues for better data utility.

It can be concluded that the proposed MP-CRDP could improve data utility by optimal feature set and reducing correlated sensitivity via proposed multiparty correlation analysis. MP-CRDP is an effective and practical method for multiparty data release with differential privacy protection.

Our method assumes existing trusted servers and centralizes data for training and private operation in a multiparty data release scenario. Correlation analysis in federated learning scenarios is an interesting direction for our future research.

References

[1] Shanthamallu U S, Spanias A, Tepedelenlioglu C, Stanley M. A brief survey of machine learning methods and their sensor and IoT applications. In *Proc. the 8th Int. Conf. Information, Intelligence, Systems & Applications*, Aug. 2017. DOI: 10.1109/IISA.2017.8316459.

[2] Mohammed N, Fung B C M, Debbabi M. Anonymity meets game theory: Secure data integration with malicious participants. *The VLDB Journal*, 2011, 20(4): 567-588. DOI: 10.1007/s00778-010-0214-6.

[3] Fung B C M, Wang K, Chen R, Yu P S. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 2010, 42(4): Article No. 14. DOI: 10.1145/1749603.1749605.

[4] Kim H, Ben-Othman J, Mokdad L. UDiPP: A framework for differential privacy preserving movements of unmanned aerial vehicles in smart cities. *IEEE Trans. Veh. Technol.*, 2019, 68(4): 3933-3943. DOI: 10.1109/TVT.2019.2897509.

[5] Du M, Wang K, Xia Z, Zhang Y. Differential privacy preserving of training model in wireless big data with edge computing. *IEEE Trans. Big Data*, 2020, 6(2): 283-295. DOI: 10.1109/TBDATA.2018.2829886.

[6] Kim S, Shin H, Baek C H, Kim S, Shin J. Learning new words from keystroke data with local differential privacy. *IEEE Trans. Knowl. Data Eng.*, 2020, 32(3): 479-491. DOI: 10.1109/TKDE.2018.2885749.

[7] Li D, Yang Q, Yu W, An D, Zhang Y, Zhao W. Towards differential privacy-based online double auction for smart

grid. *IEEE Trans. Inf. Forensics Secur.*, 2020, 15: 971-986. DOI: 10.1109/TIFS.2019.2932911.

[8] Dwork C. Differential privacy. In *Proc. the 33rd International Colloquium on Automata, Languages and Programming*, July 2006, pp.1-12. DOI: 10.1007/11787006_1.

[9] Dwork C, McSherry F, Nissim K, Smith A D. Calibrating noise to sensitivity in private data analysis. In *Proc. the 3rd Theory of Cryptography Conference*, March 2006, pp.265-284. DOI: 10.1007/11681878_14.

[10] Ji Z, Lipton Z C, Elkan C. Differential privacy and machine learning: A survey and review. arXiv:1412.7584, 2014. https://arxiv.org/abs/1412.7584, May 2020.

[11] Mir D J. Differentially-private learning and information theory. In *Proc. the 2012 EDBT/ICDT Workshops*, March 2012, pp.206-210. DOI: 10.1145/2320765.2320823.

[12] Friedman A, Schuster A. Data mining with differential privacy. In *Proc. the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July 2010, pp.493-502. DOI: 10.1145/1835804.1835868.

[13] Mohammed N, Chen R, Fung B C M, Yu P S. Differentially private data release for data mining. In *Proc. the17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2011, pp.493-501. DOI: 10.1145/2020408.2020487.

[14] Vaidya J, Shafiq B, Basu A, Hong Y. Differentially private naive Bayes classification. In *Proc. the 2013 IEEE/WIC/ACM International Conferences on Web Intelligence*, November 2013, pp.571-576. DOI: 10.1109/WI-IAT.2013.80.

[15] Chaudhuri K, Monteleoni C. Privacy-preserving logistic regression. In *Proc. the 22nd Annual Conference on Neural Information Processing Systems*, December 2008, pp.289-296.

[16] Lei J. Differentially private M-estimators. In *Proc. the 25th Annual Conference on Neural Information Processing Systems*, December 2011, pp.361-369.

[17] Zhang J, Zhang Z, Xiao X, Yang Y, Winslett M. Functional mechanism: Regression analysis under differential privacy. *Proceedings of the VLDB Endowment*, 2012, 15(11): 1364-1375. DOI: 10.14778/2350229.2350253.

[18] Rubinstein B I P, Bartlett P L, Huang L, Taft N. Learning in a large function space: Privacy-preserving mechanisms for SVM learning. arXiv:0911.5708, 2009. https://arxiv.org/abs/0911.5708, May 2020.

[19] Chaudhuri K, Monteleoni C, Sarwate A D. Differentially private empirical risk minimization. *Machine Learning Research*, 2011, 12: 1069-1109.

[20] Song S, Chaudhuri K, Sarwate A D. Stochastic gradient descent with differentially private updates. In *Proc. the 2013 IEEE Global Conf. Signal Inf. Process.*, December 2013, pp.245-248. DOI: 10.1109/GlobalSIP.2013.6736861.

[21] Abadi M, Chu A, Goodfellow I J, McMahan H B, Mironov I, Talwar K, Zhang L. Deep learning with differential privacy. In *Proc. the 2016 ACMSIGSAC Conf. Comput. Commun. Secur.*, October 2016, pp.308-318. DOI: 10.1145/2976749.2978318.

[22] Xiao Y, Xiong L. Protecting locations with differential privacy under temporal correlations. In *Proc. the 22nd ACM Conference on Computer and Communications Security*, October 2015, pp.1298-1309. DOI: 10.1145/2810103.2813640.

[23] Lv D, Zhu S. Achieving correlated differential privacy of big data publication. *Computers & Security*, 2019, 82: 184-195. DOI: 10.1016/j.cose.2018.12.017.

[24] Kifer D, Machanavajjhala A. No free lunch in data privacy. In *Proc. the 2011 ACM SIGMOD International Conference on Management of Data*, June 2011, pp.193-204. DOI: 10.1145/1989323.1989345.

[25] He X, Machanavajjhala A, Ding B. Blowfish privacy: Tuning privacy-utility trade-offs using policies. In *Proc. the 2014 ACM SIGMOD International Conference on Management of Data*, June 2014, pp.1447-1458. DOI: 10.1145/2588555.2588581.

[26] Kifer D, Machanavajjhala A. Pufferfish: A framework for mathematical privacy definitions. *ACM Trans. Database Syst.*, 2014, 39(1): Article No. 3. DOI: 10.1145/2514689.

[27] Chen R, Fung B C M, Yu P S, Desai B C. Correlated network data publication via differential privacy. *The VLDB Journal*, 2014, 23(4): 653-676. DOI: 10.1007/s00778-013-0344-8.

[28] Zhu T, Xiong P, Li G, Zhou W. Correlated differential privacy: Hiding information in Non-IID data set. *IEEE Trans. Info. Fore. and Secur.*, 2015, 10(2): 229-242. DOI: 10.1109/TIFS.2014.2368363.

[29] Yang B, Sato I, Nakagawa H. Bayesian differential privacy on correlated data. In *Proc. the 2015 ACM SIGMOD International Conference on Management of Data*, May 31-June 4, 2015, pp.747-762. DOI: 10.1145/2723372.2747643.

[30] Alhadidi D, Mohammed N, Fung B C M, Debbabi M. Secure distributed framework for achieving ϵ-differential privacy. In *Proc. the 12th International Symposium on Privacy Enhancing Technologies*, July 2012, pp.120-139. DOI: 10.1007/978-3-642-31680-7_7.

[31] Hong Y, Vaidya J, Lu H, Karras P, Goel S. Collaborative search log sanitization: Toward differential privacy and boosted utility. *IEEE Trans. Dependable Secur. Comput.*, 2015, 12(5): 504-518. DOI: 10.1109/TDSC.2014.2369034.

[32] Mohammed N, Alhadidi D, Fung B C M, Debbabi M. Secure two-party differentially private data release for vertically partitioned data. *IEEE Trans. Dependable Secur. Comput.*, 2014, 11(1): 59-71. DOI: 10.1109/TDSC.2013.22.

[33] Cheng X, Tang P, Su S, Chen R, Wu Z, Zhu B. Multi-party high-dimensional data publishing under differential privacy. *IEEE Trans. Knowl. Data Eng.*, 2020, 32(8): 1557-1571. DOI: 10.1109/TKDE.2019.2906610.

[34] Goryczka S, Xiong L. A comprehensive comparison of multiparty secure additions with differential privacy. *IEEE Transactions on Dependable and Secure Computing*, 2017, 14(5): 463-477. DOI: 10.1109/TDSC.2015.2484326.

[35] Dangi D, Santhi G. Secured multi-party data release on cloud for big data privacy-preserving using fusion learning. *Turkish Journal of Computer and Mathematics Education*, 2021, 12(3): 4716-4725. DOI: 10.17762/turcomat.v12i3.1893.

[36] Zhu T, Xiong P, Li G, Zhou W. Answering differentially private queries for continual datasets release. *Future Gener. Comput. Syst.*, 2018, 87: 816-827. DOI: 10.1016/j.future.2017.05.007.

[37] Chen J, Ma H, Zhao D, Liu L. Correlated differential privacy protection for mobile crowdsensing. *IEEE Trans. Big Data*, 2021, 7(4): 784-795. DOI: 10.1109/TBDATA.2017.2777862.

[38] Cao Y, Yoshikawa M, Xiao Y, Xiong L. Quantifying differential privacy in continuous data release under temporal correlations. *IEEE Trans. Knowl. Data Eng.*, 2019, 31(7): 1281-1295. DOI: 10.1109/TKDE.2018.2824328.

[39] Song S, Wang Y, Chaudhuri K. Pufferfish privacy mechanisms for correlated data. In *Proc. the 2017 ACM International Conference on Management of Data*, May 2017, pp.1291-1306. DOI: 10.1145/3035918.3064025.

[40] Zhang T, Zhu T, Xiong P, Huo H, Tari Z, Zhou W. Correlated differential privacy: Feature selection in machine learning. *IEEE Trans. Industrial Informatics*, 2020, 16(3): 2115-2124. DOI: 10.1109/TII.2019.2936825.

[41] Wang H, Wang H. Correlated tuple data release via differential privacy. *Inf. Sci.*, 2021, 560: 347-369. DOI: 10.1016/j.ins.2021.01.058.

[42] Wang H, Xu Z, Jia S, Xia Y, Zhang X. Why current differential privacy schemes are inapplicable for correlated data publishing? *World Wide Web*, 2021, 24(1): 1-23. DOI: 10.1007/s11280-020-00825-8.

[43] Ou L, Qin Z, Liao S, Hong Y, Jia X. Releasing correlated trajectories: Towards high utility and optimal differential privacy. *IEEE Trans. Dependable Secur. Comput.*, 2020, 17(5): 1109-1123. DOI: 10.1109/TDSC.2018.2853105.

[44] Tang P, Chen R, Su S, Guo S, Ju L, Liu G. Differentially private publication of multi-party sequential data. In *Proc. the 37th IEEE International Conference on Data Engineering*, April 2021, pp.145-156, DOI: 10.1109/ICDE51399.2021.00020.

[45] Wu X, Dou W, Ni Q. Game theory based privacy preserving analysis in correlated data publication. In *Proc. the Australasian Computer Science Week Multiconference*, January 31-February 3, 2017, Article No. 73. DOI: 10.1145/3014812.3014887.

[46] McSherry F, Talwar K. Mechanism design via differential privacy. In *Proc. the 48th Annu. IEEE Symp. Found. Comput. Sci.*, October 2007, pp.94-103. DOI: 10.1109/FOCS.2007.66.

[47] Chandrashekar G, Sahin F. A survey on feature selection methods. *Comput. Elect. Eng.*, 2014, 40(1): 16-28. DOI: 10.1016/j.compeleceng.2013.11.024.

Jian-Zhe Zhao received her B.E. and M.S. degrees in management science and engineering from Beijing Institute of Technology, Beijing, in 2005 and 2009, respectively, and her Ph.D. degree in business management from Northeastern University, Shenyang, in 2015. Since 2009, she has been a lecturer with the Software College, Northeastern University, Shenyang. Her research interests include big data, data privacy and machine learning.

Xing-Wei Wang received his B.E., M.S., and Ph.D. degrees in computer science from Northeastern University, Shenyang, in 1989, 1992, and 1998, respectively. He is currently a professor with the School of Computer Science and Engineering, Northeastern University, Shenyang. His research interests include cloud computing and future Internet. He has published more than 100 journal articles, book chapters, and refereed conference papers.

Ke-Ming Mao received his B.E. and Ph.D. degrees from Northeastern University, Shengyang, in 2003 and 2009, respectively. He is currently an associate professor in Northeastern University, Shenyang. His research interests include computer vision, deep learning, reinforcement learning and incremental learning.

Chen-Xi Huang is currently pursuing her Bachelor's degree in software engineering with the Software College, Northeastern University, Shenyang. Her research interests include data privacy and machine learning.

Yu-Kai Su received his B.E. degree in software engineering from Northeastern University, Shenyang, in 2021. He is currently a technician with Quality and Process IT Management Department, Huawei Technologies Co., Ltd., Shenzhen. His research interests include big data and machine learning.

Yu-Chen Li is currently pursuing his Bachelor's degree in software engineering with the Software College, Northeastern University, Shenyang. His research interests include big data management and software architecture.