Computers & Security 136 (2024) 103542

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Correlated differential privacy based logistic regression for supplier data protection

Ming Liu, Xiao Song, Yong Li *, Wenxin Li

School of Cyber Science and Technology, Beihang University, Beijing 100019, China

ARTICLE INFO

Keywords:
Correlated differential privacy
Feature selection
Logistic regression
Supplier data
Privacy-preservation

ABSTRACT

As a crucial participant in the supply chain, the supplier's every action affects the supply chain's status, making predictions about whether a supplier will be listed or not essential. However, the large amount of sensitive data used in machine learning generates the problem of privacy leakage. Due to the data relevance, traditional differential privacy is prone to leakage of information of correlated data. To effectively tackle this problem, in the scenario of supplier listing prediction, we introduce the correlated differential privacy mechanism for the logistic regression model, propose the feature selection scheme DC-FBFS, and further explore different noise addition methods. The experiments show that the proposed scheme can improve the utility of data, increase the prediction accuracy, and reduce the error in data query while effectively protecting data.

1. Introduction

With the globalization of the manufacturing industry, the concept of the supply chain, first proposed in the late 1980s (Porter, 2004), has recently gained popularity as a management model in manufacturing. The supply chain is a complex, customer-focused, dynamic, and cross-cutting network structure that connects suppliers, manufacturers, distributors, and end customers. The complex supply chain network and fluctuating market environment can impact a player in the supply chain, which in turn affects the stability of the supply chain (Kamalahmadi and Parast, 2016; Pettit et al., 2019, 2010). For example, the COVID-19 outbreak severely affected the manufacturing and service sectors, resulting in significant losses (Belhadi et al., 2021; Rubbio et al., 2019; Soares et al., 2021). As a critical player in the supply chain, any modification made by the supplier will impact the entire supply chain. For instance, whether or not a supplier is publicly traded affects a company's reputation, transparency, market position, choice of partners, and the entire supply chain. Therefore, it is necessary to predict whether the supplier will go public.

Predictions are often made using machine learning techniques. However, machine learning often relies on massive amounts of data, most containing susceptible information. As a result, protecting data privacy has become a hotly debated topic in academia. Dwork et al. (2008) proposed a rigorous mathematical proof for privacy protection. Since then, differential privacy has become an emerging privacy protection mechanism. Differential privacy is now widely used to protect privacy in various industrial settings, such as location privacy protection (Yang et al., 2018; Yin et al., 2018), smart grid (Liu et al., 2019; Lyu et al., 2018), multi-intelligent body systems (Ye et al., 2020), and trajectory protection (Zhang et al., 2023).

The data in the dataset is assumed to be independent according to the original notion of differential privacy (Zhang et al., 2020). The conventional assumption of independent data distribution in differential privacy is unrealistic. In practical applications, data is often correlated due to temporal relationships, ethical relationships (family), geographical relationships (same province/city/region GDP), and others. Deleting one data record from a dataset connected to others may significantly influence other records, providing the adversary with more information than anticipated. The idea of correlated differential privacy is proposed, considering the correlation between the data. Chen et al. (2014) showed that differential privacy can be tuned to provide provable privacy guarantees even in the correlated setting by introducing an extra parameter, which measures the extent of correlation. Zhu et al. (2015) proposed an effective correlated differential privacy solution by defining the correlated sensitivity and designing a correlated data releasing mechanism focused on the private perturbation algorithms on correlated data to fill the gap. Chen et al. (2017) explored the perturbation mechanisms from two perspectives. Aiming at the privacy leakage problem of traditional differential privacy function in correlated datasets, a novel improved method based on machine learning and maximum information

* Corresponding author.
E-mail address: liyong_@buaa.edu.cn (Y. Li).

https://doi.org/10.1016/j.cose.2023.103542
Received 15 July 2023; Received in revised form 12 September 2023; Accepted 16 October 2023
Available online 18 October 2023

coefficient (MIC) was proposed, which improved the difference privacy of correlated datasets in big data (Lv and Zhu, 2018). Peng et al. (2019) proposed a method to protect multiuser location-correlated information under a strict privacy budget.

In the correlated dataset, the relationship between the features may be highly correlated, redundant, or irrelevant. Therefore, feature selection is performed to achieve optimal performance and eliminate redundant information. Liu et al. (2018) studied differential private ensemble feature selection. Privacy protection is combined with machine learning, in which logistic regression is adopted for local differential privacy protection to achieve classification by utilizing noise addition and feature selection (Yin et al., 2019). However, neither of these two studies considered the effect of data correlation. Zhang et al. (2020) proposed a correlation reduction scheme with differentially private feature selection, considering the issue of privacy loss when data have correlation in machine learning tasks. However, this method performs poorly in scenarios with few features and relies on threshold selection.

Targeted at reducing privacy leakage in the case of supplier listing prediction with few features, we implement supplier listing prediction with privacy protection, propose a feature selection method, and explore the impact of different noise addition methods.

Overall, the contributions of this paper can be summarized as follows:

- To protect suppliers' data privacy effectively, we propose applying the correlated differential privacy to the logistic regression algorithm to predict the listed suppliers.
- We propose a feature selection method called Data Correlation based Forward-Backward Feature Selection (DC-FBFS) based on feature importance and data relevance. This method can effectively improve data utility in scenarios with few features.
- We experimentally explore the impact of different noise addition methods on the Correlated Differential Privacy based Logistic Regression model (CDP-LR). We find that adding Laplace noise to the sample mean gradient for the same privacy budget usually leads to higher accuracy.

2. Preliminaries

2.1. Differential privacy

Differential privacy is a rigorous, mathematically provable privacy protection scheme tailored to data analysis problems and independent of prior knowledge. It aims to increase the accuracy of data queries while reducing the likelihood of identifying records when querying a dataset. Its relevant definitions and properties are as follows.

Definition 1. (ϵ-Differential Privacy (Dwork and Roth, 2014)). Suppose ϵ is a positive real number, $\mathscr{M}$ is a randomized algorithm, $Im(\mathscr{M})$ denotes the mapping of $\mathscr{M}$, and S is the set of all subsets of $Im(\mathscr{M})$. For any non-single-element neighbor set D and $D^{'}$, i.e., $|D\Delta D^{'}| \leq 1$, the algorithm $\mathscr{M}$ gives ϵ-Differential Privacy if it satisfies

$$\Pr[\mathscr{M}(D) \in S] \leq e^{\epsilon}\Pr[\mathscr{M}(D^{'}) \in S], \tag{1}$$

where ϵ is a privacy level metric called the privacy budget. A higher privacy budget means a higher probability of discrepancies in the randomized neighbor sets, i.e., not good enough to blur the differences among neighbor sets by the randomization algorithm $\mathscr{M}$. Hence, a higher privacy budget results in worse protection and a lower privacy level.

Definition 2. ((ϵ, δ)-Differential Privacy (Dwork and Roth, 2014)). Unlike ϵ-Differential Privacy, (ϵ, δ)-Differential Privacy introduces a parameter δ that allows the algorithm $\mathscr{M}$ to not satisfy pure ϵ-Differential Privacy with probability δ (preferably less than $1/|D|$), which can be expressed as

$$Pr[\mathscr{M}(D) \in S] \leq e^{\epsilon} Pr[\mathscr{M}(D^{'}) \in S] + \delta. \tag{2}$$

Definition 3. (Global Sensitivity (Dwork and Roth, 2014)). For any function $Q : D \rightarrow \mathbb{R}^d$, the global sensitivity of the function Q is defined as

$$GS = \max_{D,D^{'}:|D\Delta D^{'}|\leq 1} \| Q(D) - Q(D^{'}) \|_p, \tag{3}$$

where R is the real space mapped, d is the dimension of the function Q, p indicates the L_p distance, usually using the Manhattan distance L_1.

Global sensitivity measures the maximum difference between the original statistical task, i.e., the function Q in Eq. (3), on any pair of neighbor sets.

Definition 4. (Laplace mechanism (Dwork and Roth, 2014)). For a function $Q : D \rightarrow \mathbb{R}^d$, the randomized algorithm $\mathscr{M}(D)$ which can be written as

$$\mathscr{M}(D) = Q(D) + \left(Lap_1\left(\frac{GS}{\epsilon}\right), Lap_2\left(\frac{GS}{\epsilon}\right), \ \ldots, Lap_d\left(\frac{GS}{\epsilon}\right)\right)^T, \tag{4}$$

satisfies ϵ-Differential Privacy.

The Laplace mechanism adds Laplace noise, i.e., noise conforming to the Laplace distribution, which can be expressed as a probability density function $Lap_i(GS/\epsilon) \propto \exp(-\epsilon|Q_i(D)|/GS)$ with a mean of zero and a standard deviation of $\sqrt{2}GS/\epsilon$.

Definition 5. (Gaussian mechanism (Dwork and Roth, 2014)). For a function $Q : D \rightarrow \mathbb{R}^d$, the randomized algorithm $\mathscr{M}(D)$ which can be expressed as

$$\mathscr{M}(D) = Q(D) + \left(\mathscr{N}_1(\sigma^2), \mathscr{N}_2(\sigma^2), \ldots, \mathscr{N}_d(\sigma^2)\right)^T, \tag{5}$$

satisfies (ϵ, δ)-Differential Privacy, where $\sigma^2 = 2GS^2\log(1.25/\delta)/\epsilon^2$.

The Gaussian mechanism, which adds Gaussian noise instead of Laplacian noise, does not satisfy pure ε-Differential Privacy but satisfies (ϵ, δ)-Differential Privacy.

To bound the overall privacy cost of releasing numerous results of differentially private mechanisms, several composition theorems for differential privacy are proposed.

Theorem 1. *(Sequential Composition (Dwork and Roth, 2014)). Given a set of mechanisms $\mathscr{M} = \{\mathscr{M}_1, \mathscr{M}_2, \ldots, \mathscr{M}_n\}$, where each $\mathscr{M}_i$ satisfies ϵ_i-Differential Privacy and performs sequentially on the same dataset, then $\mathscr{M}$ satisfies $\sum_{i=1}^{n}\epsilon_i$-Differential Privacy. If each $\mathscr{M}_i$ satisfies (ϵ_i,δ_i)-Differential Privacy, then $\mathscr{M}$ satisfies ($\sum_{i=1}^{n}\epsilon_i$,$\sum_{i=1}^{n}\delta_i$)-Differential Privacy*

Theorem 2. *(Parallel Composition (Dwork and Roth, 2014)). For a dataset D which is split into n disjoint subsets such that $D_1 \cup D_2 \cup \ldots \cup D_n = D$, mechanism $\mathscr{M}$ gives ϵ-Differential Privacy, then the mechanism which releases all of the results $\mathscr{M}(D_1), \mathscr{M}(D_2), \ldots, \mathscr{M}(D_n)$ satisfies ϵ-Differential Privacy.*

2.2. Logistic regression

The logistic regression algorithm is a classic model for classification problems. This model introduces a nonlinear function $g : \mathbb{R}^D \rightarrow (0, 1)$ to predict the posterior probability of category labels $p(y = 1|x) = g(f(x; w))$, where $g(\cdot)$ is the activation function named logistic function. It compresses the range of the linear function into the interval $(0, 1)$, which can be used to represent probabilities (Qiu, 2020).

In logistic regression, the posterior probability of the label $y = 1$ is

$$p(y = 1|x) = \sigma\left(w^T x\right) = \frac{1}{1 + \exp(-w^T x)}, \tag{6}$$

and the posterior probability of the label $y = 0$ is

$$p(y = 0|x) = 1 - p(y = 1|x) = \frac{\exp(-w^T x)}{1 + \exp(-w^T x)}, \tag{7}$$

and then,

$$w^T x = \log \frac{p(y = 1|x)}{1 - p(y = 1|x)} = \log \frac{p(y = 1|x)}{p(y = 0|x)}. \tag{8}$$

Logistic regression uses cross-entropy as the loss function and gradient descent to optimize the parameters. Given N training samples $\{(x^{(n)}, y^{(n)})\}_{n=1}^{N}$, the logistic regression model is used to predict each sample $x^{(n)}$, and the posterior probability of its label being 1 is output as $\hat{y}^{(n)}$.

The risk function is as follows,

$$R(w) = -\frac{1}{N}\sum (p_r(y = 1|x)log\hat{y} + p_r(y = 0|x)log(1 - \hat{y})). \tag{9}$$

The partial derivative of the risk function $R(w)$ is given by

$$\frac{\partial R(w)}{\partial w} = -\frac{1}{N}\sum x(y - \hat{y}). \tag{10}$$

Using the gradient descent method, the training process for logistic regression is first to initialize $w_0 \leftarrow 0$ and then iteratively update the parameters by Eq. (11):

$$w_{t+1} \leftarrow w_t + \alpha \frac{1}{N}\sum x\left(y - \hat{y}_{w_t}\right). \tag{11}$$

3. Correlated differential privacy

The supplier dataset described in Section 5 has many correlations between suppliers. Hence, the disclosure of one record could simultaneously reveal additional hidden information. The exposure of sensitive data is a significant risk for unlisted suppliers. Therefore, we need to consider the impact of data correlation to protect suppliers' privacy effectively. Inspired by Zhu et al. (2015), we adopt the concept of correlated degree to measure the correlation between every two records in the dataset. Meanwhile, correlated sensitivity is introduced to reduce the leakage of correlated information on the one hand and the addition of redundant noise on the other hand. In this section, we introduce the concept of correlated degree, define correlated sensitivity, and then compare it with global sensitivity.

3.1. Correlated degree

The correlated coefficient δ_{ij} indicates the correlation between records i and j. This study used the Pearson correlation coefficient and the Mahalanobis distance, defined as follows.

Pearson Correlation Coefficient (Pearson, 1897): The Pearson correlation coefficient p_{ij} takes values in the range $[-1, 1]$. The correlation between records i and j is high when $|p_{ij}|$ approaches 1. Here, we use the absolute value of p_{ij} to represent the correlated degree.

Mahalanobis distance (Mahalanobis, 1936): The Mahalanobis distance was proposed by the Indian statistician P. C. Mahalanobis and represents the distance between a point and a distribution. It is an effective method for calculating the similarity of two unknown sample sets. It considers the association between different features and is scale-invariant, i.e., independent of the measurement scale. The Mahalanobis distance $d(v_1, v_2)$ of the vectors v_1 and v_2 is defined as

$$d(v_1, v_2) = \sqrt{(v_1 - v_2)^T \Sigma^{-1} (v_1 - v_2)}, \tag{12}$$

where v_1 and v_2 are two samples randomly selected from the population G, μ is the center of G and Σ represents the covariance matrix. It is not difficult to see $d(v_1, v_2) \geq 0$ by Eq. (12). Furthermore, we determine the correlated degree between the two records noted m_{ij} using the Mahalanobis distance as follows,

$$m_{ij} = \frac{1}{1 + d_{ij}}, \tag{13}$$

which belongs to the interval $(0, 1]$.

Either way, we can obtain the correlation matrix Λ, a non-negative real symmetric matrix with diagonal elements $\delta_{ii} = 1$ and can be expressed as

$$\Lambda = \begin{pmatrix} \delta_{11} & \cdots & \delta_{1n} \\ \vdots & \ddots & \vdots \\ \delta_{n1} & \cdots & \delta_{nn} \end{pmatrix}. \tag{14}$$

In addition, we set a threshold δ_0 to filter out weakly correlated degrees, which can be considered as irrelevant records:

$$\delta_{ij} = \begin{cases} 0, & \delta_{ij} < \delta_0 \\ \delta_{ij}, & \delta_{ij} \geq \delta_0 \end{cases}. \tag{15}$$

3.2. Correlated sensitivity

Global sensitivity considers only the number of related records and does not consider the degree of correlation between records. Furthermore, global sensitivity implies that the related records are fully related, i.e., $\{\delta_{ij} = 1 |\ i \text{ and } j \text{ are related records}\}$, which does not correspond to the real situation. Therefore, correlated sensitivity is introduced.

Definition 6. (Correlated sensitivity (Zhu et al., 2015)). For a query Q, where q is the set of responding records to the query, the correlated sensitivity can be expressed as

$$CS_q = \max_{i \in q} \sum_{j=0}^{n} |\delta_{ij}| \cdot \| Q(D^j) - Q(D^{-j}) \|_1. \tag{16}$$

As a parameter of the noise mechanism, sensitivity affects the distribution of noise. Correlated sensitivity can partially release noise according to the correlated degree. In contrast, global sensitivity introduces redundant noise due to a lack of correlation.

4. Algorithm description

In this section, we propose a feature selection method named DC-FBFS. After feature selection by DC-FBFS, a filtered dataset is formed based on the selected feature subset. This filtered dataset is then used to input the CDP-LR model presented later.

4.1. Data correlation based forward-backward feature selection

Based on data relevance, we propose a feature selection method named DC-FBFS for datasets with few features. In our proposed method, we select features based on two criteria: (1) feature importance and (2) data relevance. We hope to completely filter out redundant and unimportant features through a combination of forward and backward traversal. Additionally, by setting a data relevance threshold, we aim to eliminate features that have minimal or no impact on the relevance of the dataset. Thus, we obtain the final subset of features based on these two criteria.

In summary, DC-FBFS can be divided into five steps: (1) calculate the feature importance and data correlation; (2) perform forward traversal; (3) perform backward traversal; (4) select based on a data relevance threshold; and (5) obtain the optimal subset.

Step 1: Calculate feature importance and data correlation. With the random forest algorithm, we calculate the importance of each feature in dataset x, denoted as $FI_{f_k}(x)$. We also remove each feature and calculate its impact on the relevance of dataset x, denoted as $DC_{f_k}(x)$.

Algorithm 1
DC-FBFS: **D**ata **C**orrelation based **F**orward-**B**ackward **F**eature **S**election.

Input: Dataset x, Feature set F, Correlation Threshold T_c;
Output: Selected feature subset $F^{'}$;
1: Initialize $F^{'} = \varnothing$, $F_{del} = \varnothing$, $F_1^{'} = \varnothing$, $F_2^{'} = F$, $F_{f-bdel} = \varnothing$, $F_{cdel} = \varnothing$;
2: $\forall f_k \in F$, calculate feature importance $FI_{f_k}(x)$ and data correlation difference by deleted 1 feature $DC_{f_k}(x)$;
3: Sort features by descending order of the feature importance $FI_{f_k}(x)$, obtain $f_1^{'}, f_2^{'}, \cdots, f_n^{'}$;
4: Initialize $acclist = \varnothing$;
5: **for** k=1, …, $Num(F)$: **do**
6: $F_1^{'} = F_1^{'} \cup \{f_k^{'}\}$;
7: Compute accuracy by using feature set $F_1^{'}$, obtain acc_k;
8: Add acc_k into $acclist$;
9: **end for**
10: Find the value of k corresponding to the maximum value of $acclist$, obtain k_1;
11: $F_{f-bdel} = \{f_{k_1+1}^{'}, \ldots, f_n^{'}\}$;
12: Initialize $acclist^{'} = \varnothing$;
13: **for** k=1, …, $Num(F)$: **do**
14: Compute accuracy by using feature set $F_2^{'}$, obtain $acc_k^{'}$;
15: Add $acc_k^{'}$ into $acclist^{'}$;
16: $F_2^{'} = F_2^{'} - \{f_{n+1-k}^{'}\}$;
17: **end for**
18: Find the value of k corresponding to the maximum value of $acclist^{'}$, obtain k_2;
19: $F_{f-bdel} = F_{f-bdel} \cup \{f_{n+2-k}^{'}, \ldots, f_n^{'}\}$;
20: $F_{cdel} = \{f_i | DC_{f_k}(x) < T_c\}$;
21: $F_{del} = F_{f-bdel} \cap F_{cdel}$;
22: $F^{'} = F - F_{del}$;

Step 2: Perform forward traversal. In the forward traversal, we discard features consecutively in the order of importance and input them into the model for training. This allows us to obtain the training accuracy for each iteration. When the accuracy reaches its highest point, the forward traversal identifies the optimal subset of features, while its complementary set represents the set of redundant features.

Step 3: Perform backward traversal. Like the forward traversal process, we can obtain the optimal subset of features and their corresponding redundant feature set through the backward traversal. Further, taking the two together yields F_{f-bdel}.

Step 4: Select by data correlation. Based on the data correlation threshold T_c, the feature subset F_{cdel} that has less impact on the dataset is filtered.

Step 5: Obtain the optimal subset. By combining forward traversal, backward traversal, and data correlation, we can obtain the final subset of redundant features F_{del}. The complement of this set is the optimal subset of features that will eventually be used for training.

4.2. Correlated differential based privacy logistic regression model

Based on the traditional logistic regression model, we have designed a CDP-LR model by incorporating correlated differential privacy.

Algorithm 2
CDP-LR: **C**orrelated **D**ifferential **P**rivacy based **L**ogistic **R**egression.

Input: Selected dataset $x^{'}$, Privacy budget ϵ, Epoch number N, Learning rate γ;
Output: Weight w;
1: Calculate the correlated sensitivity of new dataset
$CS_q = \max_{i \in q} \sum_{j=0}^{n} |\delta_{ij}| \cdot \|Q(D^j) - Q(D^{-j})\|_1$;
2: Initialize w;
3: **for** i = 1, …, N: **do**
4: $y_{pred} = logistic\ (x^{'})$;
5: Compute cross-entropy loss;
6: Compute sum of gradient dw; ← /*Add noise (ΔCS_q, ϵ)*/
7: Get sample size; ← /* Add noise (ΔCS_q, ϵ)*/
8: Compute the mean of gradient $\overline{dw} \leftarrow \frac{dw}{sample\ number}$; ← /*Add noise ($\Delta CS_q$, ϵ)*/
9: Update $w \leftarrow w - \gamma * dw$;
10: **end for**
11: Obtain weight w; ← /*Add noise (ΔCS_q, ϵ)*/

Algorithm 2 intuitively shows the process of CDP-LR, where /*Add noise (ΔCS_q, ϵ)*/ denotes the possible noise addition methods according to the sensitivity ΔCS_q under the same privacy budget, one of which is chosen in practice. To address the issue of noise addition methods, we design exploratory experiments in Section 6.2.4.

4.3. Theoretical analysis

The purpose of Algorithm 1 is feature selection, which serves as a preprocessing step for the inputs to Algorithm 2. The number of features affects the data correlation, which in turn affects the noise parameter sensitivity s, and even further, the correlated sensitivity ΔCS_q. After performing feature extraction, we demonstrate that our algorithm CDP-LR satisfies ϵ-Differential Privacy.

Consider the example of adding Laplace noise to the sample mean gradient $\overline{dw}$ in each epoch. The total privacy budget ϵ can be equally divided into N parts, each denoted by $\epsilon_0 = \epsilon/N$. We first analyze the privacy budget ϵ_0 in each epoch.

We add Laplace noise with correlated sensitivity. Thus, the Laplace mechanism can be expressed as follows,

$$\mathscr{M}(D) = Q(D) + \left(Lap_1\left(\frac{CS}{\epsilon_0}\right), Lap_2\left(\frac{CS}{\epsilon_0}\right), \ \ldots, Lap_d\left(\frac{CS}{\epsilon_0}\right)\right)^T. \tag{17}$$

D and $D^{'}$ are neighbor sets, and Q is a query. We can prove that

$$\begin{aligned} \Pr[Q(D) = t] &= \Pr\left[Q(D)_1 = t_1\right] \wedge \ldots \wedge \Pr\left[Q(D)_d = t_d\right] \\ &= \prod_1^d \frac{\epsilon_0}{2CS} \exp\left(\frac{-\epsilon_0 |t_i - Q(D)_i|}{CS}\right) \end{aligned} \tag{18}$$

$$\begin{aligned} \Pr[Q(D^{'}) = t] &= \Pr\left[Q(D^{'})_1 = t_1\right] \wedge \cdots \wedge \Pr\left[Q(D^{'})_d = t_d\right] \\ &= \prod_1^d \frac{\epsilon_0}{2CS} \exp\left(\frac{-\epsilon_0 |t_i - Q(D^{'})_i|}{CS}\right) \end{aligned} \tag{19}$$

$$\begin{aligned} \frac{\Pr[Q(D) = t]}{\Pr[Q(D^{'}) = t]} &= \exp \frac{-\epsilon_0 \left[\sum_1^d |t_i - Q(D)_i| - \sum_1^d |t_i - Q(D^{'})_i|\right]}{CS} \\ &\leq \exp \frac{-\epsilon_0 \sum_1^d |Q(D)_i - Q(D^{'})_i|}{CS} \\ &\leq \exp \frac{-\epsilon_0 \cdot \| Q(D) - Q(D^{'}) \|_1}{CS} \\ &\leq \exp(\epsilon_0) \end{aligned} \tag{20}$$

Therefore, each epoch is ϵ_0-Differential Privacy. Then, by applying Theorem 1 (sequential composition), the algorithm CDP-LR satisfies $N\epsilon_0$-Differential Privacy, i.e., ϵ-Differential Privacy.

5. Construction and preprocessing of the dataset

In this section, we introduce the process of constructing the dataset, followed by the details of the dataset's preprocessing.

5.1. Construction of the dataset

Our data is derived from Apsoto, a global automotive supply chain service platform, and Qichacha, a corporate credit inquiry website. Both of these databases are professional and authoritative. The supplier data includes the company name, date of establishment, region, number of insured persons, enterprise type, investment entity, whether it is listed or not, and information on the products offered. The supplier's company name, date of establishment, region, enterprise type, and product information are downloaded from Apsoto. The information on the number

of insured persons of suppliers was downloaded from Qichacha. To ensure the accuracy of the data, we cross-checked the common data (company name, date of incorporation, region, and type of enterprise) between both websites. Collecting and organizing the comprehensive information of numerous suppliers is a massive undertaking. To ensure the accuracy of our research, we meticulously filtered out duplicate information and standardized the company names.

5.2. Preprocessing of the dataset

Since the dataset was self-built, even though we have manually conducted preliminary filtering work, there may still be issues, such as data anomalies and missing data. Due to the direct impact of the dataset's quality on training effectiveness and the need to convert the original data into a suitable format, it is necessary to preprocess the dataset. In this study, the preprocessing of the dataset can be summarized as follows.

Step 1: Data cleaning. Remove incorrect values, missing values, duplicate values, and outliers from the data to improve its cleanliness and reliability.

Step 2: Data integration. By assigning supplier IDs, data from multiple sources is merged to eliminate duplication and redundancy, resulting in a comprehensive and consistent supplier dataset.

Step 3: Data transformation. Convert product information into one-hot vectors for further processing. The other data should also be normalized and standardized.

6. Experiments

6.1. Experimental setup

In this paper, we conducted experiments on privacy protection in the data analysis and publishing phases, respectively. The data and models are protected by a combination of correlated differential privacy and machine learning in the analysis phase, and correlated differential privacy techniques in the publishing phase process the data.

In the data analysis experiments, we first clarified the relationship between data relevance, feature importance, and the number of features. Next, we presented intermediate results for the DC-FBFS algorithm in the feature selection phase. Then, we compared our proposed scheme with other differential privacy schemes. Finally, we explored the impact of different noise addition methods and selected the optimal one.

In the data publishing experiments, we first compared the errors introduced by our scheme with those introduced by other schemes for counting queries. Then, we introduced product information to observe its effect on the query error. Finally, we compared the query error under two different similarity measures.

6.1.1. Dataset

The experiments make use of a supplier dataset, which is self-constructed. The supplier dataset contains the following information: company name, date of establishment, region, number of insured persons, enterprise type, investment entity, whether listed or not, and information on products offered. After preprocessing the data, 1160 records with 6 features were obtained.

6.1.2. Evaluation index

Two evaluation indexes were applied to the experiments: accuracy for data analysis experiments and Mean Absolute Error (MAE) for data publishing experiments.

To measure the utility of data analysis, we evaluate according to the accuracy of the predicted results, which can be expressed as follows,

$$Accuracy = \frac{Total\ number\ of\ correctly\ predicted\ samples}{Total\ number\ of\ samples} \tag{21}$$

To measure the utility of published data, we evaluate it by counting queries, using Mean Absolute Error (MAE) as an evaluation metric. MAE is defined as

$$MAE = \frac{1}{|\mathscr{Q}|}\sum_{\mathscr{Q}_i \in \mathscr{Q}} |\widehat{\mathscr{Q}_i}(x) - \mathscr{Q}_i(x)| + |\mathscr{Q}_i(x) - \mathscr{Q}_i(x_0)| \tag{22}$$

where $\widehat{\mathscr{Q}_i}(x)$ represents the query result after noise addition, and $\mathscr{Q}_i(x)$ and $\mathscr{Q}_i(x_0)$ represent the actual query result. Meanwhile, $\widehat{\mathscr{Q}_i}(x)$ and $\mathscr{Q}_i(x)$ are the query results on the dataset after feature selection, while $\mathscr{Q}_i(x_0)$ is the query on the original dataset.

The MAE formula consists of two components: $|\widehat{\mathscr{Q}_i}(x) - \mathscr{Q}_i(x)|$ represents the error due to the noise added according to the sensitivity, and $|\mathscr{Q}_i(x) - \mathscr{Q}_i(x_0)|$ responds to the error introduced by the enhanced data correlation due to feature reduction after feature selection.

6.2. Experiment for data analysis

Improving the utility of data analysis is one of the goals of our proposed scheme. In the data analysis experiments, we first clarified the relationship among data relevance, feature importance, and number of features through experiments, which also laid the foundation for proposing our feature selection method DC-FBFS. Next, we showed the selection results of each stage of the DC-FBFS method and conducted comparative experiments on the GDP-LR and CDP-LR models. Then, to highlight the advantages of our scheme, we compared and analyzed our proposed scheme with three other schemes. Finally, since there are several methods of adding noise, we applied different methods to the CDP-LR model and obtained the optimal method of adding noise by comparison. All of the above experiments used the accuracy of the prediction result as the evaluation index. For correlation between records, we used the Pearson correlation coefficient to build the data correlation matrix with the threshold δ_0 set to 0.8.

6.2.1. Relation between data correlation, feature importance, and number of features

The experiment result shows that the data relevance gradually decreases as we add features consecutively. Furthermore, by calculating the data correlation and feature importance, we found that the features have different rankings according to these two criteria.

Fig. 1 shows the relationship between data correlation and the number of features. In general, the data correlation decreases as the number of features increases. Therefore, more features can effectively reduce data correlation. However, more features can also reduce model accuracy, increase model complexity, and increase training time.

As shown in the blue part of Fig. 2, where the dark blues indicate the correlation before the removal of the features and the light blues indicate the change in correlation after the removal of the current feature, we can see that the overall correlation of the dataset changes differently after the removal of different features, among which the removal of the feature "Date of Establishment" has the most significant impact and the feature "Enterprise Type" has the least.

Meanwhile, the impact of different features on data correlation varies widely. The effect of "Date of Establishment", "Region", and "Investment Entity" can be measured in the order of 10^2, while that of "Enterprise Type", "Registered Capital" and "Number of Insured Persons" is only in the teens.

The red part of Fig. 2 shows the importance of the features calculated by the random forest algorithm. It can be seen that the feature "Registered Capital" has the highest importance and dominates, while the

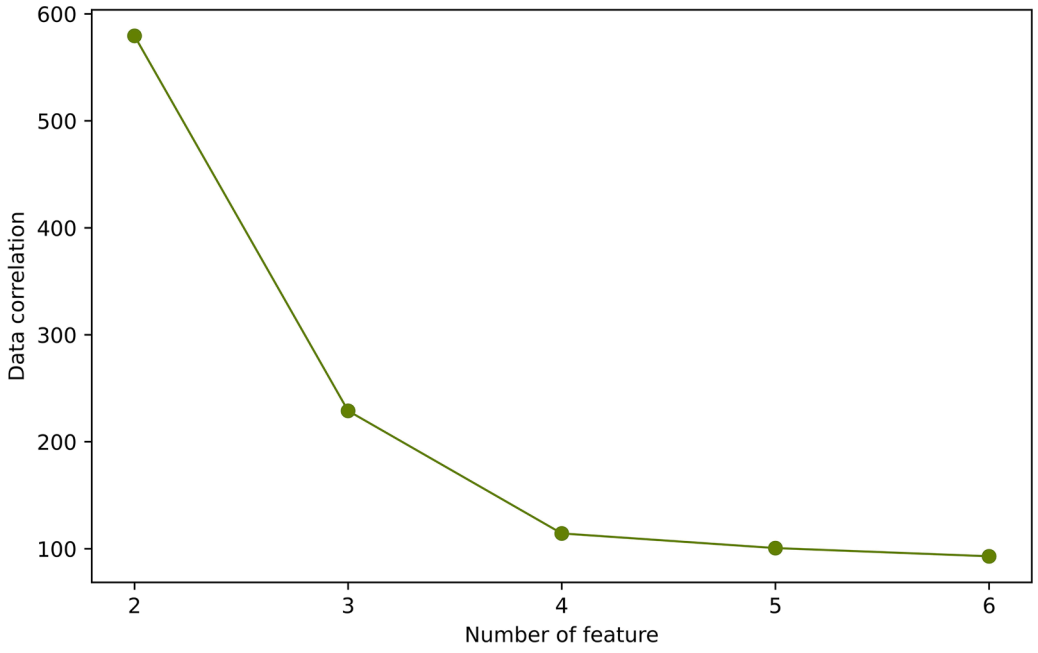

Fig. 1. Data correlation VS Number of features existed.

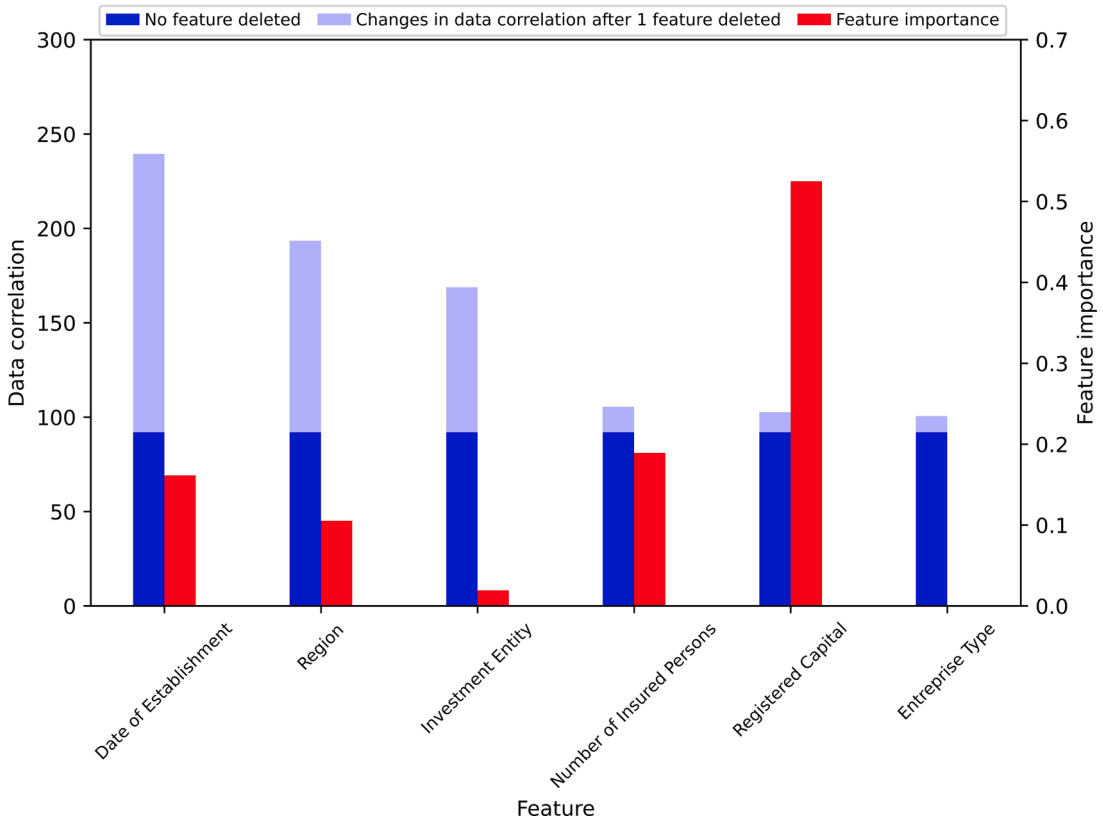

Fig. 2. Change in data relevance after deletion of current feature & Feature importance.

feature "Enterprise Type" has the lowest importance.

If we compare these two parts, we can see that we can get two different ranking results if we simultaneously rank the features according to data correlation and feature importance. In other words, in addition to feature importance, the influence of data correlation should also be considered as a reference for feature selection. In addition, Fig. 1 shows that reducing the number of features increases the data correlation, making extracting relevant information more accessible. As a result, we need to consider the impact of the change in data correlation after feature selection and the effect of feature importance, which is the central idea of DC-FBFS.

6.2.2. *Feature selection by DC-FBFS*

This section describes the processing of our proposed method DC-

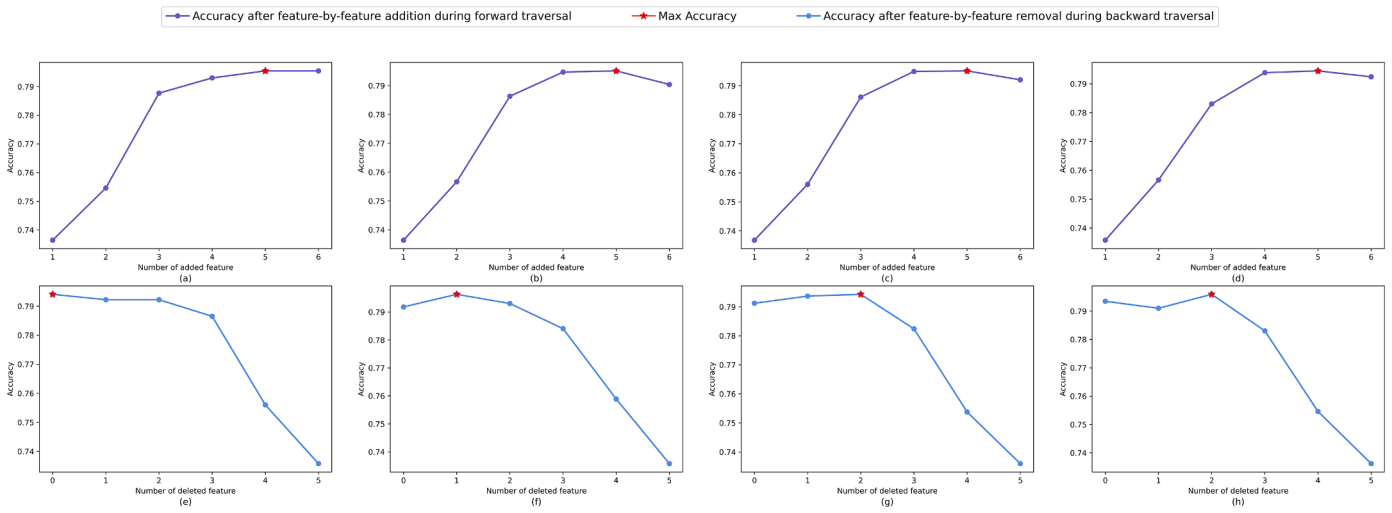

Fig. 3. Performance of forward-backward traversal.

FBFS and gives the intermediate results.

Fig. 3 shows the performance of forward and backward traversal. The results show that forward traversal filters out 1 removable feature, while backward traversal filters out 0, 1, or 2 removable features in different set experiments. In this study, the correlation threshold T_c is set to 40. Based on T_c, we filter out some features with low correlation influence. The number of feature deletions at different stages is summarized in Table 1. Finally, the feature selection is completed according to the DC-FBFS scheme.

Table 2 shows the sensitivities and accuracies of the GDP-LR and CDP-LR models under different feature selection stages. The results show that the correlation sensitivities are all lower than the global sensitivities. Thus, the accuracies of correlated differential privacy are all higher than the accuracies of general differential privacy. In addition, the results of DC-FBFS (i.e., the removed features are "Enterprise Types") show that DC-FBFS effectively filters out the redundant features even though the reduction of features increases the relevance of the data, which means that the sensitivity increases and the noise increases, there is almost no loss of accuracy. The sensitivity visually shows that the DC-FBFS results are very little different from the sensitivity before selection. In other words, the addition of redundant noise is reduced. This is a tradeoff between accuracy and complexity.

6.2.3. Comparison of different schemes

For the comparison with our proposed scheme DC-FBFS (denoted by Proposed scheme), we consider a traditional machine learning approach where there is no privacy protection (denoted by NonPrivate scheme) and the other two differential privacy schemes: the Group Differential Privacy scheme, which multiplies the number of correlated records to introduce noise (Chen et al., 2014) (denoted by Group scheme) and the CR-FS scheme where noise is added by correlated differential privacy after feature selection (Zhang et al., 2020) (denoted by Zhang's scheme), using the accuracy as an index to evaluate the classification effectiveness.

Table 1
Number of deleted features.

	F_{f-bdel}	F_{cdel}	$F_{del} = F_{f-bdel} \cap F_{cdel}$
N_{del}	1	3	1
	2	3	1

N_{del}: Number of features to delete.
F_{f-bdel}: features to be removed during forward-backward traversal.
F_{cdel}: features to be removed during selection by data correlation threshold.
F_{del}: features to be removed by DC-FBFS.

Table 2
Sensitivities and accuracies of the LR model under different feature selection stages.

Features Deleted	Sensitivity		Accuracy	
	GS	ΔCS_q	GDP-LR	CDP-LR
∅	45.86	20.55	**49.97 %**	50.09 %
F_{f-bdel}	49.53	43.06	48.95 %	49.74 %
F_{cdel}	41.84	39.18	49.45 %	50.18 %
F_{del}	45.86	21.64	49.91 %	**50.60 %**

∅: No feature deleted.
F_{f-bdel}: {Enterprise Type, Investment Entity}.
F_{cdel}: {Enterprise Type, Number of Insured Persons, Registered Capital}.
F_{del}: {Enterprise Type}.

Fig. 4 shows the experimental results of LR. From Fig. 4, we can see that the overall accuracy of our scheme is higher than Zhang's scheme and Group's scheme. Since this classification result is the result after adding the perturbation, in some cases, the accuracy of our scheme is not the highest. Therefore, we need to minimize the sensitivity and, thus, the perturbation to improve the utility.

It can be seen that the classification results of the three differential privacy schemes are similar when the privacy budget is small (such as $\epsilon < 3$), i.e., when more noise is added and the privacy level is high, because the random noise interference occupies a lot at that time.

As the privacy budget increases, our scheme emphasizes the advantages. This is because our method considers both the importance of the features in the machine learning algorithm and the relevance of the data, more accurately measures the role of the features in the dataset, and achieves effective feature screening. When $\epsilon = 7$, the accuracy of our scheme suddenly decreases, which should be caused by the randomness of the noise. After $\epsilon > 7$, the results gradually stabilize.

In addition, there is still a large gap in accuracy compared to the NonPrivate scheme. However, compared to the other two differential privacy schemes, our scheme has narrowed a small gap.

6.2.4. Optimal method for adding noise to the CDP-LR model

As described in Section 4.2, for the same privacy budget ϵ and the same ΔCS_q, there may be multiple ways to add noise. To investigate which method of adding noise is better, we designed the control experiments in Table 3, where the parameters are set as follows: the

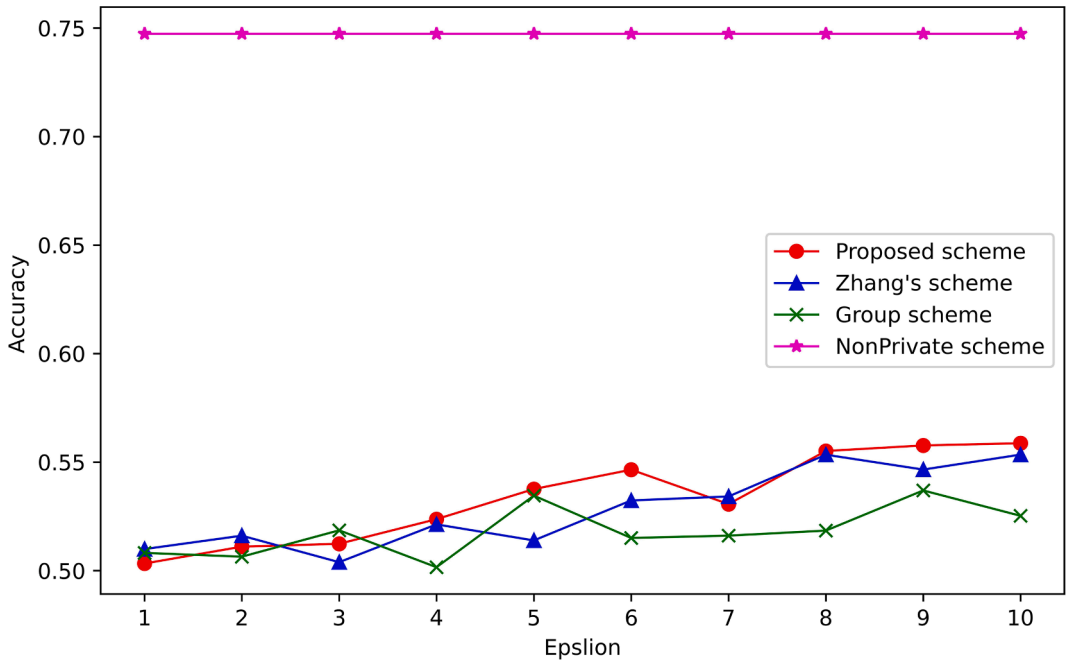

Fig. 4. Privacy-accuracy tradeoff in LR compared with different private schemes.

Table 3
Different methods of adding noise (Abadi et al., 2016).

Method	Description
0	No noise added
1	Add Laplace noise to sample average gradient $\overline{dw}$
2	Add Gaussian noise to sample average gradient $\overline{dw}$
3	Add Laplace noise to sample gradient dw
4	Add Laplace noise to sample size N
5	Add Laplace noise to sample gradient dw and to sample size N
6	Add Laplace noise to weight w

number of training rounds (epoch) is 100, the learning rate is 10^{-6}, and the relaxation term for Gaussian noise is 0.2.

Method 0 is a blank control with no noise added. Method 1 and method 2 add Laplace and Gaussian noise to the sample average gradient $\overline{dw}$ respectively. Method 3 and method 4 add noise to the sample gradient dw, sample size N, respectively. Method 5 adds noise to both the sample gradient dw and the sample size N. Method 6 adds noise directly to the training-derived weights w.

Similar to the analysis in Section 4.3, by applying Theorem 1, we can show that method 2 satisfies (ϵ, δ)-Differential Privacy and the rest satisfy ϵ-Differential Privacy.

The experimental results are shown in Fig. 5, where "meth" refers to methods in Table 3. It can be seen that the accuracy of method 2 is lower than that of method 1 in most cases under the same privacy budget. Methods 1 and 5 are two ways of noise addition on the average gradient of samples, and the results show that there is almost no difference in the effect of the two ways under the same privacy budget. Still, the effect of method 1 is slightly better than that of method 5 in the case of a high privacy budget. Methods 3 and 4 are noise addition on the sum of sample gradients and the number of samples, respectively, and it is noted that both method 4 outperforms method 3 under the same privacy budget. The effect of method 6, which directly adds noise to the weight w, is the

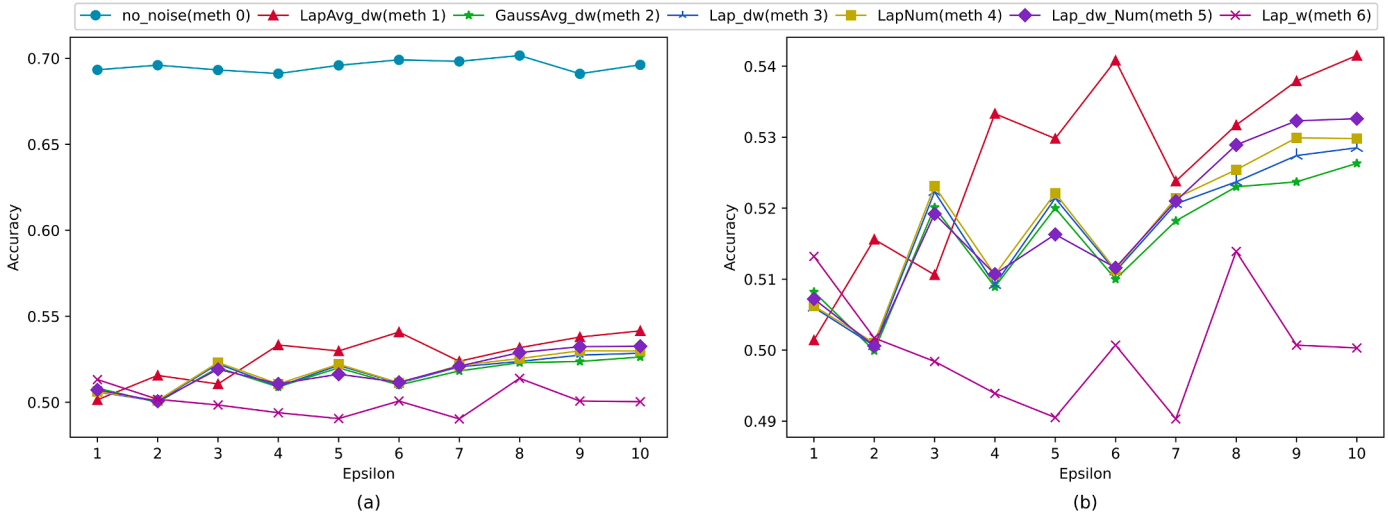

Fig. 5. Accuracy for different noise addition methods.

worst in most cases, indicating that by adding noise to specific intermediate parameters during training, such as methods 1 through 5, the training process actually corrects the weights w continuously in each epoch, and the noise during the training process is a perturbation to the direction of gradient descent. Therefore, adding noise directly to the training results will have the most direct and prominent effect. Compared with the case of method 0 without noise addition, methods 1 to 6 have different degrees of accuracy degradation, i.e., the effect of noise.

Fig. 5 also shows the trend of accuracy under different privacy budgets. As the privacy budget increases and the noise addition decreases, the accuracy of all noise mechanisms shows an overall increasing trend. For the same privacy budget, adding Laplace noise to the average gradient of the samples $\overline{dw}$ usually leads to higher accuracy.

6.3. Experiments for data publishing

We are also committed to providing a solution for publishing private data via our proposed scheme. In the data publishing experiments, we conducted three sets of experiments using count queries as an example: first, we compared the MAE differences between Zhang's scheme, the Group scheme, and our scheme; then, we introduced supplier product information into the dataset and explored its MAE variation; finally, we compared the MAE results under two different similarity measures, Pearson correlation coefficient and Mahalanobis distance. In this series of experiments, we successively used the Pearson correlation coefficient and the Mahalanobis distance to construct the data correlation matrix, with both thresholds δ_0 set to 0.8.

6.3.1. Comparison of different schemes

The performances of Zhang's scheme, the Group scheme, and our scheme are shown in Fig. 6. In Fig. 6, by comparing the Group scheme and our scheme, it can be seen that the errors introduced by the correlated sensitivity are all smaller than the global sensitivities. Moreover, the error caused by feature selection is likewise not negligible by comparing Zhang's scheme and our scheme. Meanwhile, feature selection has an essential impact on the utility of the published data.

Fig. 6 also illustrates that at higher levels of privacy protection, i.e., when ϵ is small, our scheme reflects a noticeable advantage. As ϵ increases, the MAEs of all three schemes show a trend of rapid decrease followed by a slow decrease, indicating that higher privacy protection implies a higher loss of data utility.

6.3.2. Effects of adding product information

To make the information more comprehensive and to better reflect the correlation between suppliers, we added the supplier product information to the dataset. As shown in Fig. 7, the errors in the counting query after adding the product information are lower than the other three. Although this may have an element of randomness, it shows that the appropriate addition of data can effectively reduce the MAE.

6.3.3. Comparison of different similarity measures

Furthermore, we compared the effect of the Pearson correlation coefficient and Mahalanobis distance on data release.

Pearson correlation coefficients are used in Zhang's scheme, the Group scheme, and our scheme. For comparison, we applied Mahalanobis distance to Zhang's scheme and our scheme. The Group scheme counts the relevant records, so no similarity measure is involved.

Fig. 8 illustrates the error of the counting query using the Pearson correlation coefficient and the Mahalanobis distance matrix. As shown in Fig. 8, the use of the Mahalanobis distance leads to less error than using the Pearson correlation coefficient. This is mainly because the data correlation of the Mahalanobis distance matrix is lower than that of the Pearson correlation coefficient matrix, thereby reducing the addition of noise.

Also, by comparing Zhang's scheme (Mahalanobis) and our proposed scheme (Pearson), it can be seen that the error due to feature selection is greatly reduced by using Mahalanobis distance, which deserves further investigation.

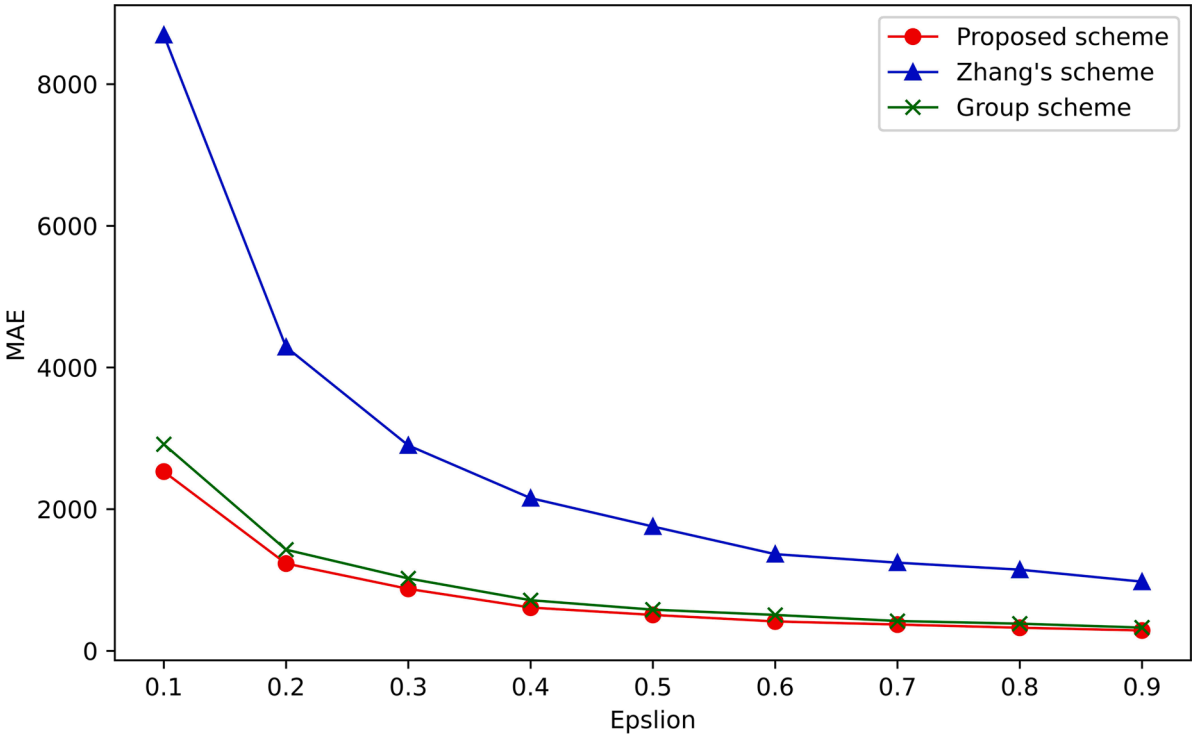

Fig. 6. MAE performance for different schemes.

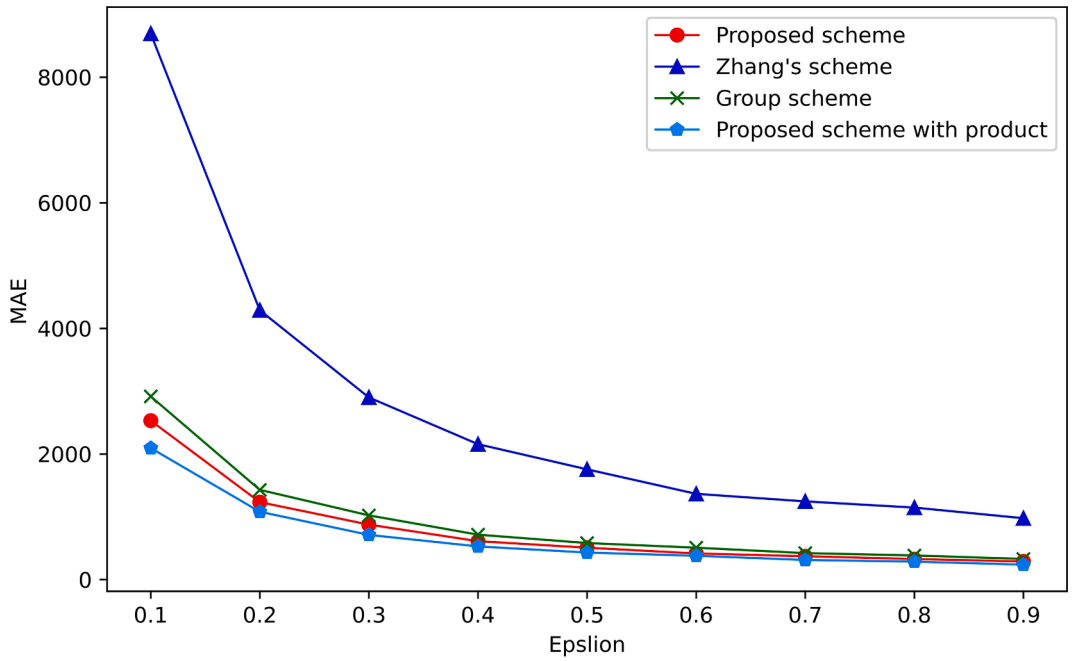

Fig. 7. MAE performance considering product information.

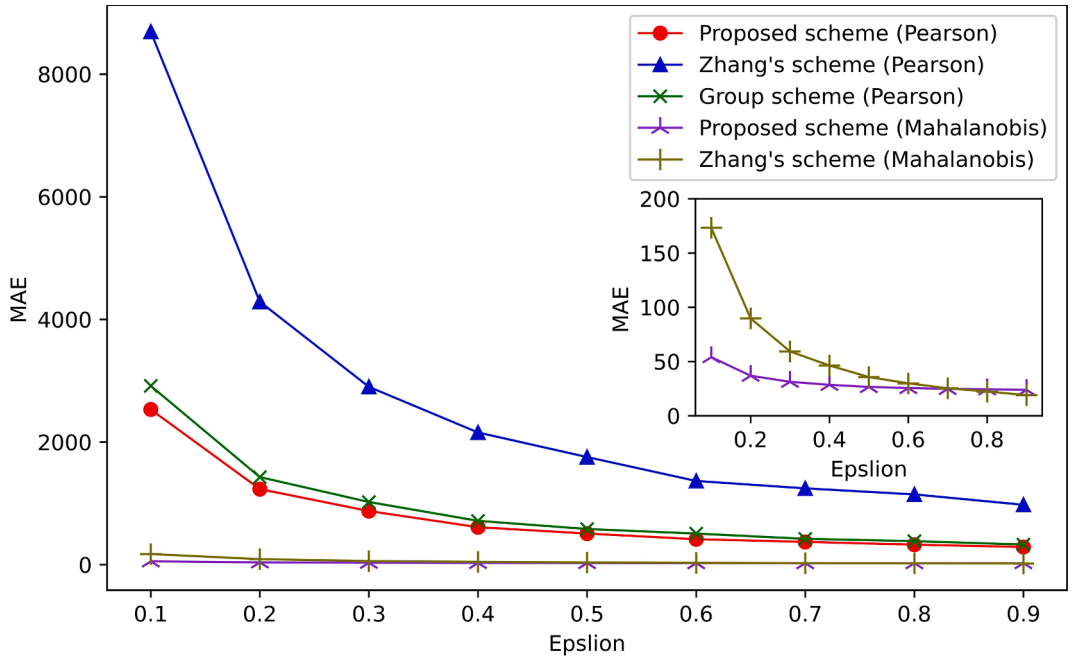

Fig. 8. MAE performance for different similarity metrics.

7. Conclusion and future work

In this paper, we introduce correlated differential privacy to logistic regression to predict whether a supplier will be listed or not. Meanwhile, for the characteristics of few-feature dimensionality and data correlation in the supplier scenario, this paper proposes the feature selection method DC-FBFS, which performs feature selection by two criteria of feature importance and data correlation. In addition, this paper investigates the effect of different noise-adding methods. This paper further studies the effects of adding product information and various similarity measures for this supplier dataset.

The main findings of the paper can be summarized as follows:

- Compared with the original differential privacy, correlated differential privacy can effectively reduce the addition of noise and improve the utility of data.
- Our proposed feature selection method DC-FBFS effectively selects features and removes redundant information. Comparative experiments have shown that the method effectively improves the utility of data analysis and reduces the error in counting queries on published data.

- By exploring different noise-adding methods, we find that adding Laplace noise to the sample mean gradient usually leads to higher accuracy for the same privacy budget.
- The experimental results show that appropriately adding product information improves the utility of data, and using Mahalanobis distance for the similarity measure can effectively reduce the error of data query results.

There are several promising directions for future work. First, we will model the supplier relationship network and construct a robust supplier relationship network that can be further explored and solved for its privacy issues from the perspective of graph neural networks. Second, our proposed feature selection method DC-FBFS can be further extended and combined with other privacy-preserving methods. In addition, we will further investigate other privacy-preserving methods in supplier scenario, such as data encryption and combination with blockchain technology.

CRediT authorship contribution statement

Ming Liu: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Data curation, Writing – original draft. **Xiao Song:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Yong Li:** Resources, Writing – review & editing, Supervision, Project administration. **Wenxin Li:** Resources, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2020YFB1712203).

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., et al., 2016. Deep learning with differential privacy. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. Vienna Austria: ACM, pp. 308–318. https://doi.org/10.1145/2976749.2978318.

Belhadi, A., Kamble, S., Jabbour, C.J.C., Gunasekaran, A., Ndubisi, N.O., Venkatesh, M., 2021. Manufacturing and service supply chain resilience to the COVID-19 outbreak: lessons learned from the automobile and airline industries. Technol. Forecasting Soc. Change 163, 120447. https://doi.org/10.1016/j.techfore.2020.120447.

Chen, J., Ma, H., Zhao, D., Liu, L., 2017. Correlated differential privacy protection for mobile crowdsensing. IEEE Trans. Big Data 784–795. https://doi.org/10.1109/TBDATA.2017.2777862.

Chen, R., Fung, B.C.M., Yu, P.S., Desai, B.C., 2014. Correlated network data publication via differential privacy. VLDB J. 23, 653–676. https://doi.org/10.1007/s00778-013-0344-8.

Dwork, C., 2008. Differential privacy: a survey of results. In: Agrawal, M., Du, D., Duan, Z., Li, A. (Eds.), Theory and Applications of Models of Computation. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–19. https://doi.org/10.1007/978-3-540-79228-4_1 vol. 4978.

Dwork, C., Roth, A., others, 2014. The algorithmic foundations of differential privacy. Foundations and trends® in theoretical. Comput. Sci. 9, 211–407.

Kamalahmadi, M., Parast, M.M., 2016. A review of the literature on the principles of enterprise and supply chain resilience: major findings and directions for future research. Int. J. Prod. Econ. 171, 116–133. https://doi.org/10.1016/j.ijpe.2015.10.023.

Liu, Y., Guo, W., Fan, C.I., Chang, L., Cheng, C., 2019. A practical privacy-preserving data aggregation (3PDA) scheme for smart grid. IEEE Trans. Ind. Inf. 15, 1767–1774. https://doi.org/10.1109/TII.2018.2809672.

Liu, Z., Li, Y., Ji, W, 2018. Differential private ensemble feature selection. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN). Rio de Janeiro: IEEE, pp. 1–6. https://doi.org/10.1109/IJCNN.2018.8489308.

Lv, D., Zhu, S., 2018. Correlated differential privacy protection for big data. In: Proceedings of the IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA). Krakow: IEEE, pp. 1011–1018. https://doi.org/10.1109/AINA.2018.00147.

Lyu, L., Nandakumar, K., Rubinstein, B., Jin, J., Bedo, J., Palaniswami, M, 2018. PPFA: privacy preserving fog-enabled aggregation in smart grid. IEEE Trans. Ind. Inf. 14, 3733–3744. https://doi.org/10.1109/TII.2018.2803782.

Mahalanobis, P.C., 1936. On the generalised distance in statistics. Proc. Natl. Inst. Sci. India 2, 49–55.

Pearson, K., 1897. Mathematical contributions to the theory of evolution. –On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc. R. Soc. Lond. 60, 489–498. https://doi.org/10.1098/rspl.1896.0076.

Peng, Z., An, J., Gui, X., Wang, Z., Zhang, W., Gui, R., et al., 2019. Location correlated differential privacy protection based on mobile feature analysis. IEEE Access 7, 54483–54496. https://doi.org/10.1109/ACCESS.2019.2912006.

Pettit, T.J., Croxton, K.L., Fiksel, J., 2019. The evolution of resilience in supply chain management: a retrospective on ensuring supply chain resilience. J. Bus. Logist. 40, 56–65. https://doi.org/10.1111/jbl.12202.

Pettit, T.J., Fiksel, J., Croxton, K.L., 2010. Ensuring supply chain resilience: development of a conceptual framework. J. Bus. Logistics 31, 1–21. https://doi.org/10.1002/j.2158-1592.2010.tb00125.x.

Porter, M.E, 2004. Competitive Advantage: Creating and Sustaining Superior Performance. Free Press, New York.

Qiu, X., 2020. Neural networks and deep learning. Neural Networks and Deep Learning. China Machine Press, Beijing.

Rubbio, I., Bruccoleri, M., Pietrosi, A., Ragonese, B., 2019. Digital health technology enhances resilient behaviour: evidence from the ward. IJOPM 40, 34–67. https://doi.org/10.1108/IJOPM-02-2018-0057.

Soares, M.C., Ferreira, C.V., Murari, T.B., 2021. Supply chain resilience and industry 4.0: a evaluation of the Brazilian northeast automotive OEM scenario post COVID-19. AI Perspect. 3, 3. https://doi.org/10.1186/s42467-021-00010-1.

Yang, M., Zhu, T., Xiang, Y., Zhou, W., 2018. Density-based location preservation for mobile crowdsensing with differential privacy. IEEE Access 6, 14779–14789. https://doi.org/10.1109/ACCESS.2018.2816918.

Ye, D., Zhu, T., Zhou, W., Yu, P.S., 2020. Differentially private malicious agent avoidance in multiagent advising learning. IEEE Trans. Cybern. 50, 4214–4227. https://doi.org/10.1109/TCYB.2019.2906574.

Yin, C., Xi, J., Sun, R., Wang, J., 2018. Location privacy protection based on differential privacy strategy for big data in industrial internet of things. IEEE Trans. Ind. Inf. 14, 3628–3636. https://doi.org/10.1109/TII.2017.2773646.

Yin, C., Zhou, B., Yin, Z., Wang, J., 2019. Local privacy protection classification based on human-centric computing. Hum. Cent. Comput. Inf. Sci. 9, 33. https://doi.org/10.1186/s13673-019-0195-4.

Zhang, J., Huang, Q., Huang, Y., Ding, Q., 2023. Tsai P-W. DP-TrajGAN: a privacy-aware trajectory generation model with differential privacy. Future Gener. Comput. Syst. 142, 25–40. https://doi.org/10.1016/j.future.2022.12.027.

Zhang, T., Zhu, T., Xiong, P., Huo, H., Tari, Z., Zhou, W., 2020. Correlated differential privacy: feature selection in machine learning. IEEE Trans. Ind. Inf. 16, 2115–2124. https://doi.org/10.1109/TII.2019.2936825.

Zhu, T., Xiong, P., Li, G., Zhou, W., 2015. Correlated differential privacy: hiding information in non-IID data set. IEEE Trans. Inform. Forensic Secur. 10, 229–242. https://doi.org/10.1109/TIFS.2014.2368363.

Ming LIU received B.S. degree in Sino-French Engineer School from Beihang University, Beijing, China, in 2022. She is currently pursuing a master's degree in Sino-French Engineer School, Beihang University. Her research interests include data security, privacy protection and knowledge graph. Email: liu_ming@buaa.edu.cn.

Xiao SONG received Ph.D. degree in Electrical Engineering from Beihang University, Beijing, China, in 2006. He is full professor of School of Cyber Science and Technology, Beihang University. His research interests include industrial internet security, differential privacy and cloud manufacturing. Email: songxiao@buaa.edu.cn

Yong LI received his Bachelor's degree and Master degree from University of Science and Technology Beijing. He is a Ph.D. candidate of the School of Cyber Science and Technology, Beihang University. His research interests include digital twin, co-simulation and knowledge graph. Email: liyong_@buaa.edu.cn

Wenxin LI is currently studying for a Ph.D. at the School of Cyber Science and Technology, Beihang University. She received her B.S. degree from Beijing University of Technology, Beijing, China, in 2020. Her research interests include graph neural networks, big data and artificial intelligence. Email: BY2039107@buaa.edu.cn