Multi-party High-Dimensional Related Data Publishing via Probabilistic Principal Component Analysis and Differential Privacy

Zhen Gu[1,2(✉)], Guoyin Zhang[1], and Chen Yang[1]

[1] College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
{guzhen,zhangguoyin}@hrbeu.edu.cn
[2] The Department of Basic Education, East University of Heilongjiang, Harbin 150066, China

Abstract. In this paper, we study the problem of multi-party horizontal split high-dimensional related data publishing that satisfies differential privacy. The dataset held by each party contains sensitive personal information, directly aggregating and publishing the local dataset from multiple parties will leak personal privacy. Usually, high-dimensional data are correlated, adding noise directly to the data will cause repeated noise addition and reduce the utility of the released data. To solve this problem, we proposed a method that horizontally split data publishing via probabilistic principal component analysis and differential privacy, the data owners add noise to low-dimensional data to reduce noise intake, and collaborate with a semi-trusted curator to reduce the dimensionality, finally, the data owners use the generative model of probabilistic principal component analysis to generate a synthetic dataset for publishing. The experimental results show that the synthetic dataset can maintain more efficient under the guarantee of differential privacy.

Keywords: Data publishing · Differential privacy · High-dimensional · Multi-party · Probability model

1 Introduction

With the development of the artificial intelligence and Internet of Things (IoT), various forms of data have been increasingly collected and used. How to better share, describe, use and manage data has become the problem in the era of big

Supported by the Natural Science Foundation of Heilongjiang Province of China under Grant LH2019F011, and the Open Project of State Key Laboratory of Information Security under Grant 2019-ZD-05, and Natural Science Foundation of East University of Heilongjiang HDFKY210117.

W. Shi et al. (Eds.): SPNCE 2021, LNICST 423, pp. 117–131, 2022.
https://doi.org/10.1007/978-3-030-96791-8_9

data and Internet of Things (IoT). Analyzing and mining the information behind these data can improve the quality of various services or formulate commercial strategies. For example, in the smart grid, it is necessary to upload the data collected by the Internet of Things (IoT) devices and smart meters to the control center. However, while analyzing and using data, it also faces many challenges. For example, data contains personal privacy, and direct sharing or publishing will lead to leakage of grid data and user privacy [1], that is to say, data is facing serious privacy leakage risks in the process of data sharing, network transmission and storage [2]. Therefore, data security and privacy have become the focus of attention of users. It is very important to protect the privacy of shared data and weigh the security and availability of data [3–5]. High-dimensional and related data are usually stored by different owners, as if the dataset is split horizontally among multiple data owners. For example, in Table 1, the records 1 to 4 come from the data owner 1, the records 5 to 8 come from the data owner 2, and the records 9 to 11 come from the data owner 3. If all these records can be aggregated, data analysts will be able to better mine the information behind the data and provide people with better services and decision-making. However, the data contains sensitive personal information, simply integrating and publishing the local dataset will cause serious privacy leakage. Therefore, the data needs to be processed for privacy protection before publishing. In recent years, there have been some studies on privacy-preserving data publishing. The first type is traditional privacy models k-anonymity [6], but studies have shown that k-anonymity is vulnerable to attacks with background knowledge, and the second type is encryption technology [7–9], encryption technology can provide better privacy guarantees. However the computational performance of such encryption technology does not scale well with a large number of users. The third type is based on differential privacy, the principle of differential privacy is to add random noise to data, which makes the attacker unable to distinguish the original input data. Differential privacy can quantitatively measure the degree of privacy protection, and can resist attacks from attackers with background knowledge, so using differential privacy to protect the privacy of publishing data has become a research hotspot in recent years, such as [10–15]. However, the following two problems need to be considered, one is that a large amount of data is often stored by different data owners, directly aggregating the data and publishing it will lead to personal privacy leakage, the second is that when the data has high dimensionality and relevance, directly adding noise into the high-dimensional data will reduce the utility of the publishing data under the same degree of privacy protection, and even make the data unavailable. In view of the above two points, this paper proposes a horizontally split data publishing method based on probabilistic principal component analysis and differential privacy. The contributions of our work are as follows:

1) We propose a method of multi-party high-dimensional related data publishing based on probabilistic principal component analysis and differential private (PPCA-DP-MH). When high-dimensional related data is stored by different owners, the data owners and a semi-trusted curator collaborate to reduce

dimensionality, then the data owners use the generative model of probabilistic principal component analysis to generate a synthetic dataset for publishing.
2) We propose to add noise to low-dimensional data to reduce noise intake. Each data owner uses the Laplace mechanism to randomly perturb the local covariance matrix which will be sent to the semi-trusted curator, so the information of each dataset by different owners is fully utilized while ensuring privacy.
3) We conduct experiments on different real datasets. The experimental results show that the synthetic dataset released by the PPCA-DP-MH method proposed in this paper can maintain high utility in SVM classification.

Table 1. Aggregated dataset of each data owner.

ID	Age	Job	Gender	Hours-per-week	Income
1	39	Shopkeeper	Male	40	>50K
2	55	Lawyer	Male	13	≤50K
3	38	Dancer	Male	20	≤50K
4	30	Dancer	Male	25	≤50K
5	28	Builder	Female	40	>50K
6	37	Dancer	Female	23	≤50K
7	49	Teacher	Female	16	≤50K
8	52	Builder	Male	45	>50K
9	31	Lawyer	Female	50	>50K
10	42	Builder	Male	40	>50K
11	37	Teacher	Male	55	>50K

2 Related Work

In recent years, there has been a number of studies on data security and privacy protection. Yang *et al.* [16] proposed that differential privacy may not guarantee privacy against arbitrary adversaries if the data are correlated. Jiang *et al.* [17] proposed that due to the relevance of data, if we add noise to each dimension of high-dimensional related data, the statistics of the data will be change drastically, which will reduce the utility of the publishing data, therefore, adding noise to fewer but more important part of data will improve the utility of publishing data. There have been some related studies, one type of algorithms are to add noise to the covariance matrix of data, such as [17–22]. Jiang *et al.* [17] proposed adding Laplace noise to the covariance matrix and projection matrix, and then use the noisy projection matrix to get the synthetic dataset for publishing. Blum *et al.* [18] proposed a Sub-Linear Query(SULQ) input perturbation framework, the algorithm adds noise to the covariance matrix, but this framework can

only be used for querying the projection subspace and cannot be used for data publishing. Chaudhury *et al.* [19] improved the SUQL algorithm and proposed the principal component analysis algorithm, this algorithm meet the differential privacy through the exponential mechanism, and the algorithm is suitable for data publishing. Kapralov *et al.* [20] pointed out that the principal component analysis algorithm lacks the guarantee of convergence time, which will affect the privacy guarantee, and they proposed a low-rank approximate matrix algorithm for differential privacy, however, the implementation of this algorithm is more complicated and difficult to process high-dimensional data. Dwork *et al.* [21] proposed adding Gaussian noise to the covariance matrix to obtain the optimal low-rank approximation of the covariance matrix, this algorithm satisfies (ε, δ) the differential privacy. Jiang *et al.* [22] proposed an algorithm to add a noise matrix that obeys the Wishart distribution, it can maintain the noise covariance matrix is positive semi-definite. Only the [17] and [19] algorithms can be used for data publishing. Another type of algorithms are suitable for data publishing, such as [23–27]. They builded a probabilistic graphical model, such as Bayesian network, Markov network or a tree model, and added noise to the low-dimensional marginal distribution, then generate a synthetic dataset based on the probabilistic graphical model. Zhang *et al.* [23] proposed the PrivBayes method, they used the relationship between attributes to construct a Bayesian network, and added Laplace noise to the low-dimensional marginal distribution, then generated a synthetic dataset for publishing. Chen *et al.* [24] proposed the Jtree method, firstly, they studied the relationship between attributes based on sparse vector sampling technology, and then constructed a Markov network, the joint distribution of all attributes is obtained through the joint tree algorithm. Zhang *et al.* [25] proposed the PrivHD method based on the Jtree method, this method used high-pass filtering technology to accelerate the construction of Markov network, and then used the maximum spanning tree method to build a better joint tree. Xu *et al.* [26] proposed the DPPro method, they randomly projected the original high-dimensional data into a low-dimensional space, and theoretically proved that the DPPro method generated a high-dimensional vector synthetic data sets with similar squared Euclidean distances. Zhang *et al.* [27] proposed the PrivMN algorithm, they constructed a Markov model to express the relationship of attributes, and then used the constructed model to generate a synthetic dataset for publishing.

From the above, we can see the studies are all about the privacy protection of data released by a single data owner, at present, there are fewer studies on the privacy protection of multi-party horizontal split data publishing, one type is multi-party data owners collaborate to reduce dimensionality and publish statistics of the data under differential privacy, such as [28–30], Ge *et al.* [28] proposed a distributed principal component analysis (DPS-PCA) algorithm with privacy protection. In a distributed environment, data owners collaborate to analyze the principal components while restricting the disclosure of private information, this algorithm can weigh the relationship between estimation accuracy and privacy protection, but this method only outputs low-dimensional subspaces of

high-dimensional sparse data. Wang *et al.* [29] designed an efficient and scalable distributed PCA protocol for privacy protection for horizontal split data, the data owner encrypts his shared data and sends them to a semi-trusted third party, the semi-trusted third party performs a private aggregation algorithm on the encrypted data, and then outputs the aggregated data to the data user, the data user calculates the principal component, the algorithm satisfies the (ε, δ) differential privacy. Imtiaz *et al.* [30] proposed a distributed principal component analysis (DPdisPCA) method that satisfies (ε, δ) differential privacy. This method used Gaussian noise to perturb the local covariance matrix, multi-party data owners collaborate to reduce dimensionality while ensuring local data privacy. The above algorithms publish statistical information of the data set, rather than publishing a data set of the same size as the original data set. Alhadidi *et al.* [31] proposed a two-party data publishing method that satisfies differential privacy, the dataset published by this method is suitable for data classification tasks. Hong *et al.* [32] proposed a collaborative sanitization framework for differential privacy search log publishing, the framework only satisfies (ε, δ) differential privacy, and their framework is not generic to handle other types of data. Cheng *et al.* [33] proposed a differential privacy sequential update of Bayesian network (DP-SUBN3) method, the parties and the semi-trusted curator collaboratively constructed the Bayesian network, the parties can treat the intermediate results as prior knowledge, and then used the constructed Bayesian network to synthesize the dataset for publishing. These algorithms are suitable for publishing a data set of the same size as the original data set, but the published data set does not support all types of data analysis.

Inspired by the above research, when the data is stored by multiple data owners, we propose a horizontally split data publishing method based on probabilistic principal component analysis and differential privacy (PPCA-DP-MH). The data owners cooperate with the semi-trusted curator to reduce dimensionality. In order to protect the local data privacy, each data owner adds Laplace noise to the local covariance matrix, and then sends it to the semi-trusted curator to perform principal component analysis, and returns the principal components to the data owner, the data owner uses principal components and probability model to generate a dataset for publishing.

3 Preliminaries

3.1 Differential Privacy

Differential privacy provides a rigorous privacy protection for sensitive information, it can be quantified by mathematical formulas. The essence of differential privacy is to randomly perturb the data. There are Laplace mechanism and exponential mechanism, the Laplace mechanism is suitable for numerical queries and the exponential mechanism is suitable for non-numerical queries.

Definition 1 ***(Differential Privacy)*** [10]**.** *A randomized algorithm M satisfies ϵ differential privacy, if for any two neighboring databases D_1, D_2 and for any $S(S \in Rang(M))$ there is:*

$$P_r\{M(D_1 = S)\} \leq e^{\varepsilon} P_r\{M(D_2 = S)\} \tag{1}$$

where ε is privacy budget.

Definition 2 ***(Sensitivity)*** [10]**.** *Let f be a function that maps a database into a fixed size vector of real numbers, $f : D \rightarrow R^d$, for any neighboring databases D_1 and D_2, the sensitivity of f is defined as:*

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1 \tag{2}$$

where$\|\cdot\|_1$ denotes the L_1 norm.

Definition 3 ***(Laplace mechanism)*** [34]**.** *For any function $f : D \rightarrow R^d$, if the output of the algorithm M satisfies the equation:*

$$M(D) = f(D) + (Lap_1(\frac{\Delta f}{\varepsilon}), \cdots, Lap_d(\frac{\Delta f}{\varepsilon})) \tag{3}$$

then the algorithm satisfies differential privacy, where $Lap_1(\frac{\Delta f}{\varepsilon}), \cdots, Lap_d(\frac{\Delta f}{\varepsilon})$ are independent Laplace variables.

Theorem 1 ***(Sequential Composition)*** [34]**.** *Let $M_1, M_2, \cdots, M_n$ be a series of privacy algorithms, and their privacy budgets are $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$, for the same dataset D, the combined algorithm $M(M_1(D), M_2(D), \cdots, M_n(D))$ provides $\sum_{i=1}^{n} \varepsilon_i$ differential privacy.*

Theorem 2 ***(Parallel Composition)*** [34]**.** *Let $M_1, M_2, \cdots, M_n$ be a series of privacy algorithms, which privacy budgets are $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$, $D_1, D_2, \cdots, D_n$ are disjoint databases, the combined algorithm $M(M_1(D_1), M_2(D_2), \cdots, M_n(D_n))$ provides $\max_{1 \leq i \leq n} \varepsilon_i$ differential privacy.*

3.2 Probabilistic Principal Component Analysis(PPCA)

Principal component analysis (PCA) is a well technique for simplifying data in statistics, principal component analysis simplifies the original high-dimensional variables into fewer low-dimensional comprehensive hidden variables. Hidden variables are also called principal components, the principal components can retain most of the information of the original variables, and the principal components are not correlated. The covariance matrix is Σ, Eigenvalue decomposition of the matrix Σ, $\Sigma = U^T \Lambda U$, where $\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_p)$ is a diagonal matrix, the elements on the diagonal are the eigenvalues of the matrix Σ, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, and U is an orthogonal matrix consists of the eigenvectors, the eigenvectors are the principal components, the number of principal components retained is determined by the cumulative contribution rate $c = \sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i$.

However, principal component analysis (PCA) is a non-generative model, a notable feature of the definition of PCA is the absence of an associated probabilistic model for the observed data, therefor, Michael *et al.* [35] proposed the generative model called probabilistic principal component analysis (PPCA). A latent variable model can correlate high-dimensional observable variables with low-dimensional latent variables, the most common model is factor analysis where the relationship is $x = Ws + \mu + \xi$, where x is p dimensional observation vector, s is k dimensional latent variables vector, $\xi \sim N(0, \Psi)$, the matrix W relates the variables x and s, and the vector μ permits the model to have non-zero mean, the motivation is that, when $k < p$, the latent variable will provide a more parsimonious explanation of the dependence between the observed variables. Principal component analysis can be regarded as the maximum likelihood solution of a factor analysis model with an isotropic covariance matrix.

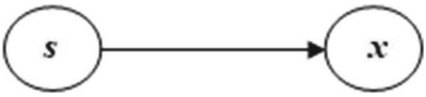


Fig. 1. Graphical model for principal component analysis

Theorem 3 [35]. *From Fig. 1 and the latent variable model* $x = Ws+\mu+\xi$, *when* $\xi \sim N(0, \sigma^2 I)$, $s \sim N(0, I_k)$, *then* $x|s \sim N(Ws + \mu, \sigma^2 I_p), \sigma > 0, W \in R^{p\times k}$, *where the maximum likelihood estimation of* μ, σ^2, *and* W *are:*

$$\hat{\mu} = \tilde{\mu} \tag{4}$$

$$\hat{\sigma}^2 = \frac{1}{p-k} \sum_{i=k+1}^{p} \lambda_i \tag{5}$$

$$\hat{W} = U_k(\Lambda_k - \hat{\sigma}^2 I)^{\frac{1}{2}} \tag{6}$$

where $\tilde{\mu}$ *is the sample mean vector, the column vectors in* U_k *is the eigenvectors corresponding to the first* k *eigenvalues of the sample covariance matrix.*

4 The PPCA-DP-MH Approach

4.1 Problem Statement

There exist $m(m \geq 2)$ local dataset owners, the i-th data owner holds a local dataset $X_{n_i\times p} = (x_1^T, \cdots, x_{n_i}^T)^T$, each row $x_i(1 \leq i \leq n_i)$ of the matrix $X_{n_i\times p}$ represents an individual, where n_i denotes the number of individuals owned by the i-th data owner, p denotes the number of attributes. All the local datasets have the same attributes, and do not intersect with each other. The datasets $X_{n_1\times p}, X_{n_2\times p}, \cdots, X_{n_m\times p}$ can be viewed as horizontally split the integrated dataset $X = \bigcup_{i=1}^{m} X_{n_i\times p}$ by m data owners. Our goal is that m data owners and semi-trusted curator collaborate to publish a synthetic dataset that satisfies ε differential privacy.

Algorithm 1. PPCA-DP-MH algorithm

Input: Data sets $X_{n_i \times p}(i = 1, 2, \cdots, m)$, privacy budget ε, cumulative contribution rate c

Output: Synthetic dataset $X' = \bigcup_{i=1}^{m} X'_{n_i \times p}$

1: **for** $i = 1$ to m **do**
2: generate noise matrices $L^1_{p\times 1}$ and $L^2_{p\times p}$, each element of the matrices obeys $Lap(\frac{2p}{n_i \varepsilon})$
3: compute:

$$\hat{E}(X^T_{n_i \times p}) = E(X^T_{n_i \times p}) + L^1_{p\times 1}$$
$$\hat{E}(X^T_{n_i \times p} X_{n_i \times p}) = E(X^T_{n_i \times p} X_{n_i \times p}) + L^2_{p\times p}$$
$$\hat{\Sigma}_i = \hat{E}(X^T_{n_i \times p} X_{n_i \times p}) - \hat{E}(X^T_{n_i \times p}) \hat{E}(X_{n_i \times p})$$

4: **end for**
5: compute: $\hat{\Sigma} = \frac{\sum_{i=1}^{m} n_i \hat{\Sigma}_i}{\sum_{i=1}^{m} n_i}$
6: perform eigenvalue decomposition of matrix $\hat{\Sigma}$, return eigenvalues and corresponding eigenvectors in descending order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, and $U = (u_1, u_2, \cdots, u_p)$
7: **for** $k = 1$ to p **do**
8: **if** $\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i \geq c$ **then**
9: $\Lambda_k = (\lambda_1, \lambda_2, \cdots, \lambda_k)$
10: $U_k = (u_1, u_2, \cdots, u_k)$
11: **end if**
12: **end for**
13: **return** Λ_k, U_k
14: **for** $i = 1$ to m **do**
15: compute $S_{n_i p} = X_{n_i p} \times U_k$
16: use the model defined in Theorem 3 to generate a synthetic data set $X'_{n_i \times p}$
17: **end for**
18: **return** $X' = \bigcup_{i=1}^{m} X'_{n_i \times p}$

4.2 Proposed Algorithm

In view of the above scenarios, we proposed a multi-party horizontal split data publishing method based on probabilistic principal component analysis and differential privacy (PPCA-DP-MH), the basic idea is that the data owners and the semi-trusted curator collaborate to reduce the dimensionality to obtain the principal components that satisfy the ε differential privacy, and then use the generative model of probabilistic principal component analysis to generate a synthetic dataset for publishing. Firstly, the data owners use the Laplace mechanism to perturb every local covariance matrix, and then send them to the semi-trusted curator, secondly, the semi-trusted curator aggregates the local noisy covariance matrices to obtain the covariance matrix of the overall data, and performs eigen-

value decomposition on the noisy covariance matrix of the overall data, then the semi-trusted curator sends the first k principal components to each data owner, lastly, the data owner uses the k principal components and Theorem 3 to generate a dataset $X'_{n_i \times p}(i = 1, 2, \cdots, m)$, and sends it to the semi-trusted curator, the semi-trusted curator obtains a synthetic dataset $X' = \bigcup_{i=1}^{m} X'_{n_i \times p}$ which satisfies ε differential privacy. Please see Algorithm 1 for details, assuming that the data has been normalized, that is, the value of the data is in the interval [0, 1].

4.3 Privacy Analysis

For the PPCA-DP-MH algorithm proposed in this paper, there is a risk of privacy leakage only when the data owner sends the covariance matrix of the local data to the semi-trusted curator, therefore, the data owner uses the Laplace mechanism of differential privacy to perturb the local covariance matrix. Because the local datasets do not intersect each other, according to the parallel combination theorem of differential privacy, as long as the local covariance matrix satisfies ε differential privacy, the covariance matrix of the overall data satisfies ε differential privacy, that is, the PPCA-DP-MH algorithm satisfies ε differential privacy.

Theorem 4. *The PPCA-DP-MH algorithm satisfies ε differential privacy.*

Proof. The normalized dataset is still denoted as $X_{n_i \times p}$, since the data is normalized to $[0, 1]$, that is, each entry in $X_{n_i \times p}$ is bounded to $[0, 1]$, so the sensitivity of $E(X^T_{n_i \times p})$ and $E(X^T_{n_i \times p} X_{n_i \times p})$ are $\frac{p}{n_i}$, because

$$\hat{E}(X^T_{n_i \times p}) = E(X^T_{n_i \times p}) + L^1_{p \times 1}$$

$$\hat{E}(X^T_{n_i \times p} X_{n_i \times p}) = E(X^T_{n_i \times p} X_{n_i \times p}) + L^2_{p \times p}$$

and each element of the matrices $L^1_{p \times 1}$ and $L^2_{p \times p}$ obeys $Lap(\frac{2p}{n_i \varepsilon})$, so the stage of calculating $\hat{E}(X^T_{n_i \times p})$ and $\hat{E}(X^T_{n_i \times p} X_{n_i \times p})$ satisfies $\frac{\varepsilon}{2}$ differential privacy, and by the sequential composition theorem of differential privacy, the local covariance matrix $\hat{\Sigma}_i = \hat{E}(X^T_{n_i \times p} X_{n_i \times p}) - \hat{E}(X^T_{n_i \times p}) \hat{E}(X_{n_i \times p})$ satisfies ε differential privacy, due to the local datasets do not intersect each other and the parallel combination theorem of differential privacy, the covariance matrix of the overall data $\hat{\Sigma}$ satisfies ε differential privacy, so the PPCA-DP-MH algorithm satisfies ε differential privacy.

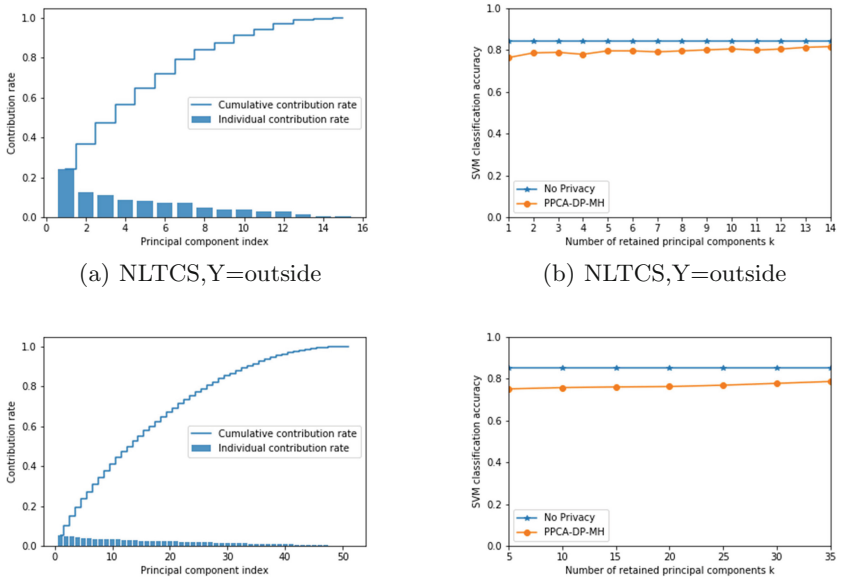
5 Experiment

In this section, we conducted simulation experiments on real datasets to demonstrate the effectiveness of our PPCA-DP-MH method, we used two datasets: NLTCS and Adult. NLTCS dataset is extracted from the National Long Term Care Survey, and recorded the daily activities of 21574 disabled persons at different time periods, each individual has 16 attributes. Adult dataset is extracted

from the 1994 US Census, it contains 45222 individuals, each individual has 15 attributes. In order to compare with the DP-SUBN3 algorithm proposed in [33], we preprocessed data similar to [33]. We use SVM classification accuracy to evaluate the utility of our PPCA-DP-MH method and the DP-SUBN3 method.

We trained multiple SVM classifiers on the synthetic dataset. Each classifier predicts one attribute based on all other attributes in the dataset. Two classifiers are trained on NLTCS, one is to predict whether a person is unable to get outside, and the other is to predict whether a person is unable to manage money. Two classifiers are trained on Adult, one is to predict whether a person holds a post-secondary degree and the other is to predict whether a person earns more than 50K. For each classification task, we use 80% of the tuples in the dataset as the training set, and the remaining 20% as the testing set. We run each experiment 5 times, and the average results are reported. In order to better measure the effectiveness of our PPCA-DP-MH method, the same SVM classifier is also trained on the original data set. In the figures, we use "No Privacy" to represent the SVM classification accuracy on the original dataset.

For the parameter k which is the number of retained principal components, it is determined by the cumulative contribution rate c of the principal components, in our experiments, the cumulative contribution rate c for NLTCS and Adult are set to 0.85 and 0.9, respectively.



(a) NLTCS,Y=outside (b) NLTCS,Y=outside

(c) Adult,Y=salary (d) Adult,Y=salary

Fig. 2. Relationship between SVM classification accuracy and number of principal components

5.1 Relationship Between SVM Classification Accuracy and Number of Principal Components Retained

In order to study the relationship between the SVM classification accuracy and the number of principal components retained k, we trained two classifiers outside on NLTCS and salary on Adult separately, in this set of experiments, the number of data owners m is set to 3, the privacy budget ε is set to 0.1.

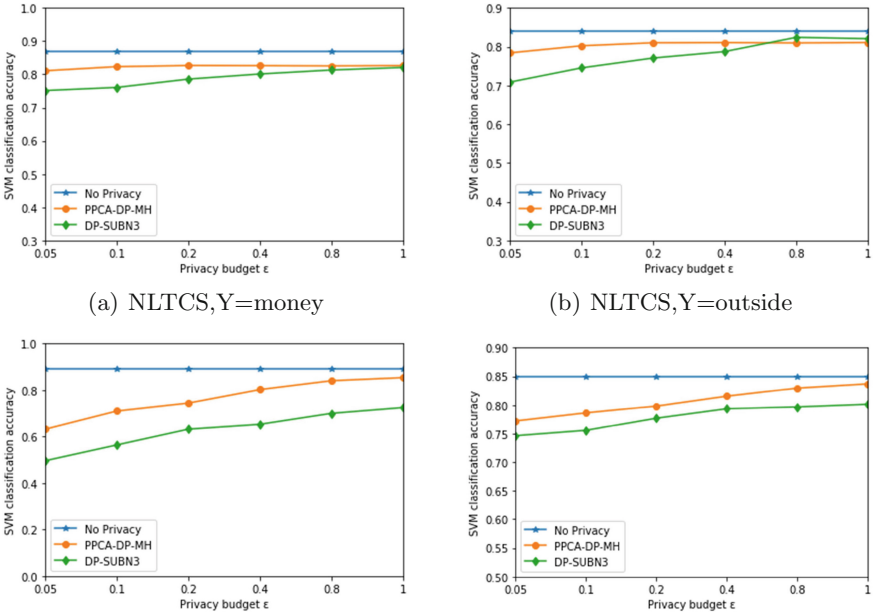
For the NLTCS dataset, Fig. 2(a) shows the cumulative contribution rate and individual contribution rate of the principal component, Fig. 2(b) shows the relationship between the SVM classification accuracy of the synthetic dataset and the number of principal components retained k. The results show that the number of principal components retained k increases (the cumulative contribution rate increases), the SVM classification accuracy also increases. It can be seen from Fig. 2(a) that the contribution rate of only the first principal component has reached more than 27%, and the cumulative contribution rate of the first 8 principal components can reach 85%, at the same time from Fig. 2(b) we can see the corresponding SVM classification accuracy can reach nearly 80%.

For the Adult dataset, Fig. 2(c) shows the cumulative contribution rate and individual contribution rate of the principal component, Fig. 2(d) shows the relationship between the SVM classification accuracy of the synthetic dataset and the number of retained principal components k. Because the Adult dataset has many attributes, we only marked the corresponding SVM classification accuracy when the number of retained principal components k is $5, 10, 15, 20, 25, 30,$ and 35 in Fig. 2 (d), it can be seen that as the number of retained principal components k increases, the SVM classification accuracy also increases.

The experimental conclusions on the two datasets are similar, that is, the SVM classification accuracy increases as the number of retained principal components k increases (the cumulative contribution rate increases), this is because each principal component contains the information of the original dataset and is not related to each other, as the number of retained principal components increases, the information of the original dataset retained increases, and the synthetic dataset contains more information about the original dataset accordingly.

5.2 Relationship Between SVM Classification Accuracy and Privacy Budget

In this set of experiments, we set the number of data owners to 3, and privacy budget ε takes different values. Figure 3 shows the SVM classification accuracy of each method on NLTCS and Adult under different privacy budgets, Fig. 3(a) and Fig. 3(b) show the results of the two classifiers money and outside on the NLTCS, respectively. Figure 3(c) and Fig. 3(d) show the results of the two classifiers education and salary on the Adult, respectively. We can observe that our PPCA-DP-MH method clearly outperforms DP-SUBN3 method, only in Fig. 3(b) when the privacy budget ε is greater than 0.5, the SVM classification accuracy of the synthetic dataset released by PPCA-DP-MH method is slightly



(a) NLTCS,Y=money (b) NLTCS,Y=outside

(c) Adult,Y=education (d) Adult,Y=salary

Fig. 3. Relationship between SVM classification accuracy and privacy budget

lower than DP-SUBN3 method, in other cases, the SVM classification accuracy of the synthetic dataset released by PPCA-DP-MH method is higher than DP-SUBN3 method. Especially, in Fig. 3(c), for the education classifier on the Adult, the SVM classification accuracy of the synthetic dataset released by PPCA-DP-MH method is nearly 10% higher than DP-SUBN3 method.

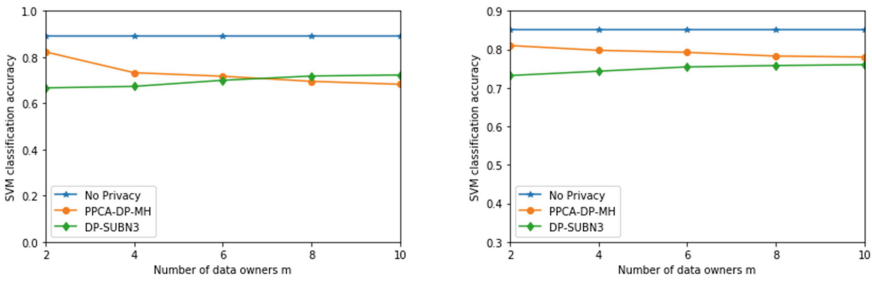
In Fig. 3, we can also observe a commonality, the SVM classification accuracy increases with the increase of the privacy budget ε both on the synthetic datasets released by PPCA-DP-MH method and DP-SUBN3 method, and this phenomenon is consistent with the theory that as the privacy budget ε increases, privacy protection will weaken and the availability of data will increase.

5.3 Relationship Between SVM Classification Accuracy and Number of Data Owners

In this section, the experiment studied the relationship between SVM classification accuracy and the number of data owners m. The number of data owners m is set to 2, 4, 6, 8, 10, and the privacy budget ε is set to 0.2, We trained two classifiers, education classifier and salary classifier on the Adult dataset.

The results in Fig. 4 show that the SVM classification accuracy of the synthetic dataset by PPCA-DP-MH method decreases as the number of data owners m increases, however, the SVM classification accuracy of the synthetic dataset by

DP-SUBN3 method increases as the number of data owners m increases, this is because, for DP-SUBN3 method, with the number of data owners increases, the number of update iterations increases when constructing the Bayesian network, and the Bayesian network constructed is closer to the distribution of the original data. For PPCA-DP-MH method, the number of individuals in the overall data set is fixed, generally, the more data owners, the less the number of individuals owned by each data owner, the more noise added by each data owner, so with the increase of data owners, the SVM classification accuracy of the synthetic dataset by PPCA-DP-MH method decreases. However, in Fig. 4(a), for education classifier, when there are no more than 6 data owners, the SVM classification accuracy of the synthetic data set by PPCA-DP-MH method can still be higher than DP-SUBN3 method, and in Fig. 4(b), for salary classifier, when there are no more than 10 data owners, the SVM classification accuracy of the synthetic data set released by PPCA-DP-MH method is still higher than DP-SUBN3 method.



(a) Adult,Y=education (b) Adult,Y=salary

Fig. 4. Relationship between SVM classification accuracy and number of data owners

6 Conclusion

In this paper, we proposed a method for multi-party horizontally split data publishing based on probabilistic principal component analysis and differential privacy (PPCA-DP-MH). The data owners and the semi-trusted curator cooperate with each other to reduce the dimensionality of high-dimensional related data and generate a synthetic data set for publishing. We used the Laplacian mechanism to add noise to less but more important data in order to increase the utility of the published data. The experimental results show that the synthetic data set released by our PPCA-DP-MH method can maintain high utility in SVM classification. In the future, we will study the vertically split data publishing based on differential privacy.

References

1. Kolter, J.Z., Jaakkola, T.S.: Approximate inference in additive factorial HMMs with application to energy disaggregation (2012)
2. Wang, D., Zhang, X., Zhang, Z., Wang, P.: Understanding security failures of multi-factor authentication schemes for multi-server environments. Comput. Secur. **88**, 1–13 (2020)
3. Tsou, Y.T., Lin, B.C.: PPDCA: privacy-preserving crowdsourcing data collection and analysis with randomized response. IEEE Access **6**, 76970–76983 (2018)
4. Ren, X., et al.: High-dimensional crowdsourced data publication with local differential privacy. IEEE Trans. Inf. Forensics Secur. **13**, 2151–2166 (2018)
5. Qiu, S., Wang, D., Xu, G., Kumari, S.: Practical and provably secure three-factor authentication protocol based on extended chaotic-maps for mobile lightweight devices. IEEE Trans. Dependable Secure Comput. **17**, 1–14 (2020)
6. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowl. Based Syst. **10**, 557–570 (2002)
7. Lu, R., Liang, X., Xu, L., Lin, X., Shen, X.: EPPA: an efficient and privacy-preserving aggregation scheme for secure smart grid communications. IEEE Trans. Parallel Distrib. Syst. **23**(9), 1621–1631 (2012)
8. Wang, C., Wang, D., Xu, G., He, D.: Efficient privacy-preserving user authentication scheme with forward secrecy for industry 4.0. Sci. China Inf. Sci. **65**(1), 1–15 (2020)
9. Wang, C., Wang, D., Tu, Y., Xu, G., Wang, H.: Understanding node capture attacks in user authentication schemes for wireless sensor networks. IEEE Trans. Dependable Secure Comput. **19**, 507–523 (2020)
10. Dwork, C., Mcsherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. J. Priv. Confidentiality **7**(3), 17–51 (2017)
11. Han, C., Wang, K.: Sensitive disclosures under differential privacy guarantees. In: IEEE International Congress on Big Data, pp. 110–117 (2015)
12. Wang, Q., Zhang, Y., Xiao, L., Wang, Z., Ren, K.: RescueDP: real-time spatio-temporal crowd-sourced data publishing with differential privacy. In: IEEE Infocom the IEEE International Conference on Computer Communications (2016)
13. Hao, W., Xu, Z.: CTS-DP: publishing correlated time-series data via differential privacy. Knowl.-Based Syst. **122**, 167–179 (2017)
14. Wang, H., Wang, H.: Correlated tuple data release via differential privacy. Inf. Sci. **560**, 347–369 (2021)
15. Chen, S., Fu, A., Yu, S., Ke, H., Su, M.: DP-QIC: a differential privacy scheme based on quasi-identifier classification for big data publication. Soft Comput. **25**(3), 7325–7339 (2021)
16. Yang, B., Sato, I., Nakagawa, H.: Bayesian differential privacy on correlated data. In: SIGMOD/PODS (2015)
17. Jiang, X., Ji, Z., Wang, S., Mohammed, N., Cheng, S., Ohno-Machado, L.: Differential-private data publishing through component analysis. Trans. Data Priv. **6**(1), 19 (2013)
18. Nissim, K., Mcsherry, F.D., Dwork, C., Blum, A.L.: Practical privacy: the SuLQ framework. In: Proceedings of the Twenty-Fourth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Baltimore, Maryland, USA, 13–15 June 2005 (2005)
19. Chaudhuri, K., Sarwate, A.D., Sinha, K.: Near-optimal differentially private principal components. In: Advances in Neural Information Processing Systems, vol. 2, pp. 989–997 (2012)

20. Kapralov, M., Talwar, K.: On differentially private low rank approximation. In: Soda, pp. 1395–1414 (2013)
21. Dwork, C., Talwar, K., Thakurta, A., Zhang, L.: Analyze gauss: optimal bounds for privacy-preserving PCA. In: Proceedings of the Annual ACM Symposium on Theory of Computing, pp. 11–20 (2014)
22. Jiang, W., Xie, C., Zhang, Z.: Wishart mechanism for differentially private principal components analysis. Comput. Sci. **9285**, 458–473 (2015)
23. Zhang, J., Cormode, G., Procopiuc, C.M., Srivastava, D., Xiao, X.: PrivBayes: private data release via Bayesian networks. ACM Trans. Database Syst. **42**(4), 1–41 (2014)
24. Rui, C., Qian, X., Yu, Z., Xu, J.: Differentially private high-dimensional data publication via sampling-based inference. In: The 21th ACM SIGKDD International Conference (2015)
25. Zhang, X., Chen, L., Jin, K., Meng, X.: Private high-dimensional data publication with junction tree. J. Comput. Res. Dev. **55**, 2794 (2018)
26. Xu, C., Ren, J., Zhang, Y., Qin, Z., Ren, K.: DPPro: differentially private high-dimensional data release via random projection. IEEE Trans. Inf. Forensics Secur. **PP**(99), 1 (2017)
27. Zhang, W., Zhao, J., Wei, F., Chen, Y.: Differentially private high-dimensional data publication via Markov network. Secur. Saf. **6**(19), 159626 (2019)
28. Ge, J., Wang, Z., Wang, M., Han, L.: Minimax-optimal privacy-preserving sparse PCA in distributed systems (2018)
29. Wang, S., Chang, J.M.: Differentially private principal component analysis over horizontally partitioned data. In: 2018 IEEE Conference on Dependable and Secure Computing (DSC) (2018)
30. Imtiaz, H., Sarwate, A.D.: Differentially private distributed principal component analysis. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2018 (2018)
31. Alhadidi, D., Mohammed, N., Fung, B.C.M., Debbabi, M.: Secure distributed framework for achieving ε-differential privacy. In: Fischer-Hübner, S., Wright, M. (eds.) PETS 2012. LNCS, vol. 7384, pp. 120–139. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31680-7_7
32. Hong, Y., Vaidya, J., Lu, H., Karras, P., Goel, S.: Collaborative search log sanitization: toward differential privacy and boosted utility. IEEE Trans. Dependable Secure Comput. **12**(5), 504–518 (2015)
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. **32**, 1557–1571 (2019)
34. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**, 211–407 (2013)
35. Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. J. Roy. Stat. Soc. **61**(3), 611–622 (2010)