IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 15, 2019, accepted March 23, 2019, date of publication April 3, 2019, date of current version April 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2909048

An Optimizing and Differentially Private Clustering Algorithm for Mixed Data in SDN-Based Smart Grid

ZEFANG LV[1], LIRONG WANG[2], ZHITAO GUAN[2], (Member, IEEE), JUN WU[3], XIAOJIANG DU[4], (Senior Member, IEEE), HONGTAO ZHAO[1], AND MOHSEN GUIZANI[5], (Fellow, IEEE)

[1]School of Mathematics and Physics, North China Electric Power University, Beijing 100026, China
[2]School of Control and Computer Engineering, North China Electric Power University, Beijing 100026, China
[3]College of Information Security Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[4]Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA
[5]Department of Computer Science and Engineering, Qatar University, Doha, Qatar

Corresponding author: Zhitao Guan (guan@ncepu.edu.cn)

This work was supported by the Beijing Natural Science Foundation under Grant 4182060.

ABSTRACT Software-defined network (SDN) is widely used in smart grid for monitoring and managing the communication network. Big data analytics for SDN-based smart grid has got increasing attention. It is a promising approach to use machine learning technologies to analyze a large amount of data generated in SDN-based smart grid. However, the disclosure of personal privacy information must receive considerable attention. For instance, data clustering in user electricity behavior analysis may lead to the disclosure of personal privacy information. In this paper, an optimizing and differentially private clustering algorithm named ODPCA is proposed. In the ODPCA, the differentially private K-means algorithm and K-modes algorithm are combined to cluster mixed data in a privacy-preserving manner. The allocation of privacy budgets is optimized to improve the accuracy of clustering results. Specifically, the loss function that considers both the numerical and categorical attributes between true centroids and noisy centroids is analyzed to optimize the allocation the privacy budget; the number of iterations of clustering is set to a fixed value based on the total privacy budget and the minimal privacy budget allocated to each iteration. It is proved that the ODPCA can meet the differential privacy requirements and has better performance by comparing with other popular algorithms.

INDEX TERMS Differential privacy, clustering, machine learning, SDN-based smart grid, big data.

I. INTRODUCTION

The emergence of SDN-based smart grid and the widespread use of big data analytics technology have led to much attention to big data analytics for SDN-based smart grids. The large amount of data generated in SDN-based smart grid is of great value. Advanced machine learning/deep learning techniques can be used to analyze data involved in SDN-based smart grid so that makes it data-driven and more intelligent [1]–[3]. Machine learning can be applied to user electricity behavior analysis power equipment monitoring and user classification in smart grid [4]–[6]. As shown in Fig.1, data generated from smart grid, such as user electricity information, power transmission and distribution data, has great value for promoting the development of smart grid. And big data technologies can be used to analyze power related data obtained from SDN-based smart grid and use analysis results to analyze user electricity behavior and improve equipment management in power systems [7], [8].

Cluster analysis is a typical unsupervised learning data mining method, which can be used to analyze user electricity behavior in smart grid, so that analyst can predict user behavior in a targeted manner and then manage and distribute power better [9]–[11]. The main idea is to divide the data into several clusters so that the distances among data items of the same cluster are as small as possible while the distances among data items of different clusters are as large as possible.

However, some data is very sensitive when linked with individual users. If not properly handled, the collection and analysis of the personal information may leak users' privacy.

The associate editor coordinating the review of this manuscript and approving it for publication was Zeeshan Kaleem.

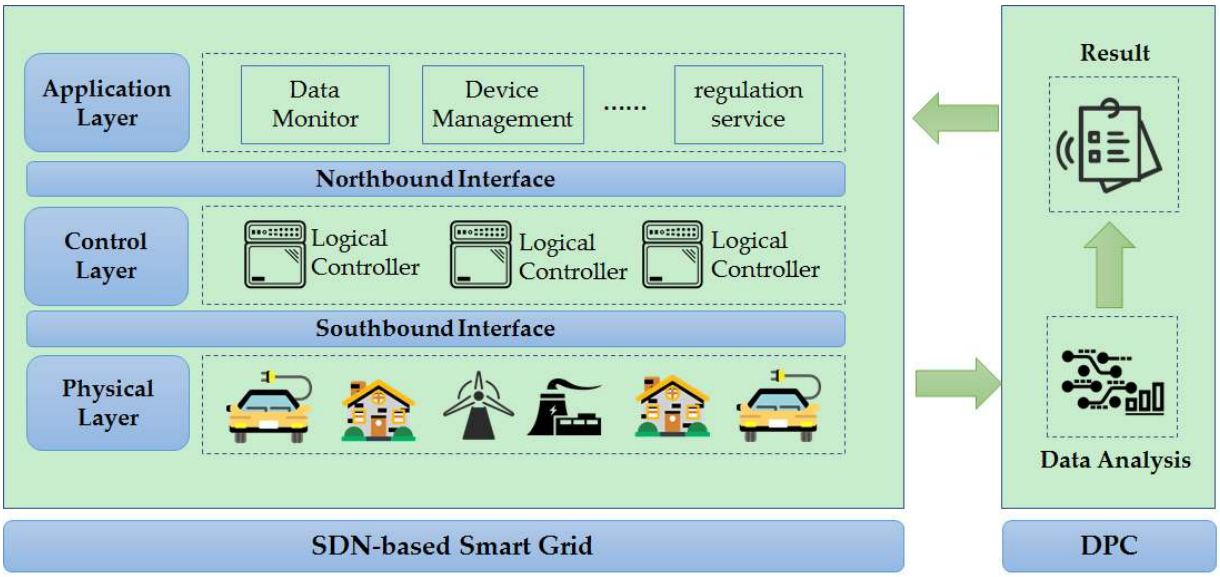

FIGURE 1. An overview of big data analytic in SDN-based smart grid.

Therefore, it is critical to achieve privacy-preserving data analysis in SDN-based smart grid. A definition called differential privacy proposed in [12] preserve privacy of all individual contributors in a dataset. Differential privacy is a privacy preserving method in which random noise that satisfies a specific distribution is added to distort the data [13]–[15].

A lot of differentially private cluster algorithms has been proposed. Most of the researches focus on the design of privacy preserving mechanisms and the tradeoff between privacy and utility. Reference [13] proposed a differentially private k-means algorithm that has a simple and efficient implementation. In each iteration, this algorithm computes noisy centroids by adding Laplace noise to the number of points in each cluster and the sum of those points. Reference [16] made a comprehensive comparison of the state-of-the-art including DPLloyd, GkM [17], PGkM [18]. They also proposed a technique to optimizing the number of iterations and privacy budget allocation by analyzing the mean square error between noisy centroids and true centroids. In [19], authors formulated the problem of differentially private k-modes for categorical data and design various schemes to tackle differentially private k-modes for both interactive and non-interactive settings.

In practice, however, most datasets are mixed data, including numerical and categorical data. So far, there is no clustering algorithm for mixed datasets that satisfies differential privacy. In this paper, we address the privacy preserving issue of clustering algorithms for mixed datasets.

We combine the differentially private k-means algorithm and k-modes algorithm to cluster mixed data in a privacy preserving manner. In case of the tradeoff between privacy and utility, we optimize the allocation of privacy budgets to improve the accuracy of clustering results by analyzing the loss function. And we set the number of iterations of clustering algorithm to a fixed value determined by the total privacy budget and the minimal privacy budget allocate to each iteration calculated by setting loss function to a threshold.

The contributions are summarized as follows:

1. We propose ODPCA, an optimizing and differentially private clustering algorithm for mixed data in SDN-based smart grid. ODPCA algorithm combines the differentially private k-means algorithm and k-modes algorithm to cluster mixed data in a privacy preserving manner and design a mechanism to make the algorithm satisfy differential privacy.
2. We optimize the allocation of privacy budgets to improve the accuracy of clustering results. Specifically, we analyze the loss function between true centroids and noisy centroids, and the number of iterations of clustering algorithm is set to a fixed value determined by the total privacy budget and the minimal privacy budget allocated to each iteration calculated by setting loss function to a threshold.
3. We prove that our proposed algorithm satisfies differential privacy and experiment with two datasets to illustrate performance of our proposed ODPCA by comparing the Normalized Intra-Cluster Variance (NICV) produced by our algorithm in different level of privacy preserving.

The rest of this paper is organized as follows. Section 2 introduces the related work. In section 3, some preliminaries are given. Section 4 shows the system model and design goals. In section 5, our proposed algorithm is stated. In section 6, privacy analysis of our proposed algorithm is given. In Section 7, the performance of our algorithm is evaluated. In Section 8, the paper is concluded.

II. RELATED WORK

The roles of clustering algorithm include power management, consumer behavior analysis, power equipment detection, etc., in smart grid. Some different clustering methods are intended to be used in smart grid to exploit useful information about customers' behavior, which can be used to promote the development and intelligence of smart grid [20]–[23].

However, sensitive data relevant to users' privacy information in smart grid may be involved when applying clustering algorithms to analyze the users' data. There are some existing solutions to the security and privacy issues in smart grid [24]–[28]. In the research on data privacy preservation, traditional methods, like k-anonymity [25] and l-diversity [29], cannot preserve privacy for all individual contributors in a dataset, but differential privacy technique can. And it has been increasingly adopted in data analysis to preserve individual privacy [30]–[34].

There is a lot of research for differentially private clustering algorithms. For categorical data, current research focus on the design of differential privacy mechanisms. Reference [19] addressed the privacy-preserving k-modes problem using differential privacy and ran the k-modes in private manners. The authors analyzed the challenges of differentially private k-modes with regard to the k-means counterpart and proposed several schemes in both interactive and non-interactive settings. In interactive setting same as our research, the authors used geometrical mechanism [32] and exponential mechanism [34] to design privacy-preserving k-modes algorithm respectively.

For numerical data, many researches mainly focus on differentially private k-means algorithms. And the availability of clustering results has been compromised due to the addition of noise. In order to increase the accuracy of differentially private k-means algorithm, current researches mainly focus on two directions, including improving the initial centroids selection method and the privacy budget allocation scheme.

In the research of initial centroids selection method, [35] proposed an improved initial centroids selection algorithm in the MapReduce framework by selecting a small portion of the dataset and performing rough clustering in advance to select the initial centroids. And they developed a method for selecting the initial centroid for a specified number of clusters k to solve the problem that the number of points outputs by the canopy algorithm is uncertain. Reference [36] proposed a DPLK-means algorithm based on differential privacy, which improved the selection of the initial center points through performing the differential privacy K-means algorithm to each subset divided by the original dataset.

For privacy budget allocation, there are generally two different methods for the allocation of privacy budgets in each iteration of the clustering algorithm, which correspond to two ways to determine the number of iterations including fixed iterations and unfixed iterations. One way is to fix the number of iterations. In some literature such as [37], the number of iterations is artificially determined with equal privacy budget allocation for each round. Another way is proposed in [38], in which the number of iterations is uncertain, each iteration consumes half of the remaining privacy budget. Reference [16] proposed an improved K-means clustering algorithm which satisfies differential privacy. The authors developed techniques to analyze MSE between the noisy centroids and the true centroids in one iteration and used this technique to determine the number of iterations and the budget allocation.

In reality, most of the dataset are mixed including both numerical and categorical attributes. Nowadays, there have been many researches focused on clustering algorithm for mixed datasets [39]–[43]. In [39], unsupervised feature learning (UFL) is applied to the mixed-type data to achieve a sparse representation, which makes it easier for clustering algorithms to separate the data. Reference [40] propose a novel framework for clustering of mixed data and find the cluster substructures that are common to both the categorical and numerical data. But, so far, there is no clustering algorithm for mixed datasets that satisfies differential privacy. Our proposed algorithm addressed the privacy preserving issue of clustering algorithm for mixed datasets.

III. PRELIMINARIES

In this section, differential privacy and two major algorithms are given for our proposed algorithm.

A. DIFFERENTIAL PRIVACY

Differential privacy protects individual privacy by adding noise to the query results, while maintaining the statistical characteristics and accuracy of the query results in an acceptable range.

Definition 1 (ε-Differential Privacy): A randomized mechanism M satisfies ε-differential privacy if for any pair of neighboring datasets D, D' that differ in at most one data record, and for any set of possible output $S \in Range(M)$,

$$\Pr(M(D) \in S) \leq e^{\varepsilon} \cdot \Pr(M(D') \in S).$$

The privacy budget ε represents the level of privacy guarantee - a lower privacy budget provides a stronger privacy guarantee.

B. K-MEANS CLUSTERING ALGORITHM

Cluster analysis is a very important topic in data analysis. The purpose of clustering is to classify the data into different classes. The k-means clustering algorithm is the simplest and most commonly used clustering algorithm. The fundamental principle is to divide the data into k clusters on the basis of minimizing the error function, with distance as the rating index of similarity. That is, the shorter the distance is between two objects, the greater of similarity they have. Given a d-dimensional dataset $D = \{x^1, x^2, \cdots, x^N\}$ (N is the total number of data points), the k-means algorithm divides the data points in D into k sets $O = \{O^1, O^2, \cdots, O^k\}$ so that

the mean square error (MSE) within the cluster is minimized

$$NICV = \sum_{i=1}^{k} \sum_{x^l \in O_i} \left\| x^l - o^i \right\|^2 \quad (1)$$

and

$$o_j^i = \frac{\sum_{x^l \in O^i} x_j^l}{\left| O^i \right|}, \quad j = 1, 2, \cdots d. \quad (2)$$

C. K-MODES CLUSTERING ALGORITHM

Let $D = \{x^1, x^2, \cdots, x^N\}$ be a categorical dataset with N data points. Each data point has M categorical attributes from the set $A = \{A_1, A_2, \cdots, A_M\}$. We use $|A_j|$ to denote the cardinality of the j-th attribute and

$$|A| = \prod_{j=1}^{M} |A_j|$$

to denote the cardinality of full domain A. The k-modes clustering algorithm [42] is an extension of k-means clustering algorithm for clustering categorical data by using a simple dissimilarity measure. It adopts a frequency-related strategy to update modes in the clustering to minimize the clustering costs. The simplest matching dissimilarity measure between two data points x and y is defined by Hamming distance:

$$Dis(x, y) = d_H(x, y) = \sum_{j=1}^{M} (1 - \delta(x_j, y_j)),$$

where x_j denotes the j-th attribute of x and

$$\delta(x_j, y_j) = \begin{cases} 1, & x_j = y_j \\ 0, & x_j \neq y_j. \end{cases}$$

The original k-modes clustering algorithm tries to minimize the following cost function

$$Cost(X, Z) = \sum_{k=1}^{K} \sum_{i=1}^{N} \omega_{ik} d_H(X_i, Z_k),$$

where Z is the set of K modes of dataset X, $\omega_{ik} \in \{0, 1\}$ and $\sum_{k=1}^{K} \omega_{ik} = 1, \forall i = 1, 2, \cdots, N, \omega_{ik} = 1$ denotes that X_i belongs to k-th mode.

IV. MODEL AND GOALS

A. AN OVERVIEW

SDN-based smart grid has been proposed to manage the overall network and communication entities for the future smart grid system to improve the efficiency and resiliency of the entire system. As shown in Fig.2, user electricity information, power transmission and distribution data generated from

SDN-based smart grid has great value for promoting the development of smart grid. And the data processing center (DPC) can apply big data technologies in analyzing power related data obtained from SDN-based smart grid and use analysis results to analyze user electricity behavior and improve equipment management in power systems, etc. However, the analysis process and the release of the analysis results may lead to the leakage of users' privacy information. In our proposed algorithm, we apply differential privacy in clustering algorithm to achieve privacy-preserving big data analysis in SDN-based smart grid to preserve sensitive information invovled.

B. MODEL GOALS

In order to solve the privacy preservation issue for mixed data in SDN-based smart grid, the design goal of our algorithm can be roughly divided into two aspects:

1) Privacy preservation: deleting or adding a data point in the dataset will not reveal personal sensitive information. In other words, a malicious analyst cannot obtain any private information of a single record by mining a similar dataset, compared with original dataset.
2) Accuracy: achieve a tradeoff between accuracy of cluster results and privacy preservation by optimizing the privacy budget allocation.

C. SECURITY MODEL

In this subsection, we introduce the security model of our system. We assume that DPC are trusted. But an adversary may obtain the data analysis results when the result is transmitted to smart grid system. Differential privacy guarantees a strong privacy that deleting or adding a particular record in a dataset will not significantly change the output of any function on a dataset. Therefore, adversary will just obtain approximate information about any individual record rather than specific information.

V. DESCRIPTION OF PROPOSED ALGORITHM

A. PROBLEM DEFINITION

Suppose that the d-dimensional mixed dataset $D = \{x^1, x^2, \cdots, x^N\}$ have N data points and each point have d attributes $A = \{A_1, A_2, \cdots, A_d\}$, among which are p numerical attributes and q categorical attributes, i.e., $p + q = d$. For $x_i \in D$, x_{ij} denotes the value on the j-th attribute of x_i. Generally, we assume x_{ij} that is a numerical data for $j = 1, 2, \ldots, p$ and x_{ij} is a categorical data for $j = p + 1, \ldots, d$. For simplicity, the values of attribute in $A = \{A_1, A_2, \cdots, A_p\}$ are normalized to [0], [1]. We use $\left|A_j\right|$ to denote the cardinality of categorical attribute, $j = p+1, p+2, \ldots, d$, i.e., the j-th attribute of each point has $\left|A_j\right|$ values (generally, we assume that $\left|A_j\right| \geq 2$), respectively $\left\{1, 2, \ldots, \left|A_j\right|\right\}$. The distance between any two data points in $Dx_i = \{x_{i1}, x_{i2}, \ldots, x_{id}\}$ and $x_j = \{x_{j1}, x_{j2}, \ldots, x_{jd}\}$ is defined as the combination of Euclidean distance and Hamming distance as following:

$$Dis(x_i, x_j) = \sum_{r=1}^{p} (x_{ir} - x_{jr})^2 + \sum_{r=p+1}^{d} (1 - (x_{ir}, x_{jr})).$$

B. OUR PROPOSED ALGORITHM

The algorithm proposed in this paper is designed to address the privacy preservation issue of clustering for mixed dataset.

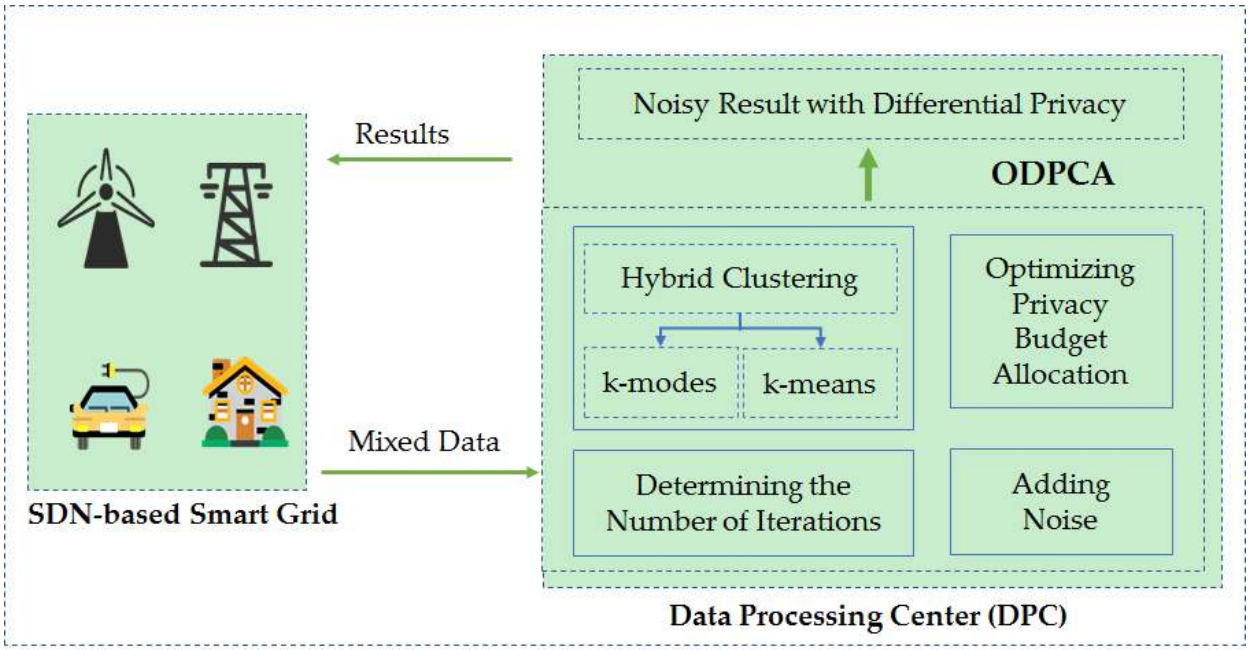

FIGURE 2. An overview of ODPCA algorithm in SDN-based smart grid.

We combine the K-means algorithm with K-modes algorithm to cluster mixed dataset and differential privacy and two mechanisms are applied in our proposed algorithm to ensure that the change of the centroids and the number of records of each cluster does not reveal private information when the data is changed.

The basic idea of proposed algorithm is to use k-means algorithm and k-modes algorithm to cluster numerical data and categorical data respectively. And in each iteration, privacy preserving is achieved by adding noise when updating the cluster center. Our proposed algorithm is outlined in **Algorithm 1.**

Algorithm 1 DP K-Means-and-Modes Algorithm

Input: D: datasets with n data points, each data has d attributes. K: number of clusters. T: number of iterations. ε: privacy budget.

Output: K noisy centroids $\left\{\hat{o}_1^{(T)}, \hat{o}_2^{(T)}, \ldots, \hat{o}_K^{(T)}\right\}$

1: Randomly select K points in the dataset D as the initial centroid $\left\{o_1^{(0)}, o_2^{(0)}, \ldots, o_K^{(0)}\right\}$;

2: for $t=1 \rightarrow T$ do

3: for $k=1 \rightarrow K$ do

4: for $j=1 \rightarrow p$ do

5: $\hat{o}_{kj}^{(t)} = \left(S_{kj}^{(t)} + Lap(\varepsilon_{kj})\right) / \left(C_{kj}^{(t)} + Lap(\varepsilon_{kj})\right)$

6: for $j = p+1 \rightarrow d$ do

7: for $r = 1 \rightarrow |A_j|$ *do*

8: $n_{kjr}^{(t)} = count_{kj}^{(t)}(r) + Geom(\alpha)$

9: $\hat{o}_{kj}^{(t)} = \arg\max_r n_{kjr}^{(t)}$

10: return $\left\{\hat{o}_1^{(T)}, \hat{o}_2^{(T)}, \ldots, \hat{o}_K^{(T)}\right\}$

C. ADDING NOISE

In each iteration of clustering process, privacy preserving is achieved by adding noise when updating the cluster center. For numerical data, we use Laplace mechanism [12] to add noise in the process of calculating the centroid. Laplace mechanism computes the result of function f on the dataset D by adding to $f(D)$ a random noise, as shown in the following equation:

$$A_f(D) = f(D) + Lap(\frac{GS_f}{\varepsilon}),$$

where

$$\Pr[Lap(\beta) = x] = \frac{1}{2\beta} e^{|x|/\beta}.$$

GS_f is the global sensitivity of function f, which is defined as following formula:

$$GS_f = \max \left\| f(D) - f(D') \right\|_1,$$

where D and D' differ in at most one data record. Noises are added to C_k, the number of points included in k-th cluster, and S_{kj}, the sum of the coordinates of the data points in the j-th dimension, respectively, and get noisy C'_k and S'_{kj}. Then, a new centroid is calculated by

$$\hat{o}_{kj} = S'_{kj}/C'_k, \quad j = 1, 2, \ldots, p, \ k = 1, 2, \ldots, K.$$

For categorical data, we use modes of attribute to update centroids in the clustering and we use Geometric Mechanism to add noise to the counts of all attribute values and take the value whose count is maximum for each attribute. Geometric Mechanism is a discrete variant of the Laplace mechanism with integral output range Z and it computes the result of

function f on the dataset D as $f(D)+z$, z is a random integral noise generated from a geometric distribution $Geom(\alpha)$:

$$P[Z=z]=\frac{1-\alpha}{1+\alpha}\alpha^{|z|},\quad z\in Z.$$

We can obtain some properties about geometric distribution as **proposition 1** shows.

Proposition 1: Suppose that $Z\sim Geom(\alpha)$, then we have

$$\begin{cases}E[Z]=0\\ Var[Z]=\dfrac{2\alpha}{(1-\alpha)^2}\end{cases}\tag{3}$$

Proof:

First, we can easily get that

$$E[Z]=\sum_{z=-\infty}^{z=\infty}\frac{1-\alpha}{1+\alpha}\cdot z\cdot\alpha^{|z|}=0.$$

Then,

$$\begin{aligned}Var[Z]&=E[Z^2]-E[Z]\\&=E[Z^2]\\&=\sum_{z=-\infty}^{z=\infty}\frac{1-\alpha}{1+\alpha}\cdot z^2\cdot\alpha^{|z|}\\&=2\cdot\frac{1-\alpha}{1+\alpha}\cdot\left(\sum_{z=0}^{z=\infty}z^2\cdot\alpha^z\right)\end{aligned}$$

and

$$\begin{aligned}\sum_{z=0}^{z=\infty}z^2\cdot\alpha^z&=\sum_{z=0}^{z=\infty}z(z-1)\cdot\alpha^z+\sum_{z=0}^{z=\infty}z\cdot\alpha^z\\&=\alpha^2\sum_{z=0}^{z=\infty}z(z-1)\cdot\alpha^{z-2}+\alpha\sum_{z=0}^{z=\infty}z\cdot\alpha^{z-1}\\&=\alpha^2\left(\sum_{z=0}^{z=\infty}\alpha^z\right)''+\alpha\left(\sum_{z=0}^{z=\infty}\alpha^z\right)'\\&=\alpha^2\left(\frac{1}{1-\alpha}\right)''+\alpha\left(\frac{1}{1-\alpha}\right)'\\&=\frac{\alpha(\alpha+1)}{(1-\alpha)^3}\end{aligned}$$

So, we can get

$$\begin{aligned}Var[Z]&=2\cdot\frac{1-\alpha}{1+\alpha}\cdot\left(\sum_{z=0}^{z=\infty}z^2\cdot\alpha^z\right)\\&=2\cdot\frac{1-\alpha}{1+\alpha}\cdot\frac{\alpha(1+\alpha)}{(1-\alpha)^3}\\&=\frac{2\alpha}{(1-\alpha)^2}.\end{aligned}$$

For categorical attribute A_j and centroid o_k, $j=p+1,\ldots,d$, $k=1,2,\ldots,K$, the noises are added to the counts of values $n_{kjr}=\text{count}_{kj}(r)+Geom(\alpha)$, $r=1,2,\ldots,|A_j|$. And then $\hat{o}_{kj}=\arg\max_r n_{kjr}$. To satisfy ε-differential privacy, we set $\alpha=e^{-\varepsilon}$ [19].

D. OPTIMIZING THE NUMBER OF ITERATION AND PRIVACY BUDGET ALLOCATION

An important issue of our algorithm is the allocation of privacy budgets. The choice of the number of iterations directly affects the allocation of the privacy budget. There are generally two ways to determine the number of iterations, which correspond to two different methods for the allocation of privacy budgets in each iteration of the clustering algorithm. One way is to fix the number of iterations. Another way is proposed in [38], in which the number of iterations is uncertain, each iteration consumes half of the remaining privacy budget. Based on the two ways of determining the number of iterations as described above, there are two main methods for the allocation of privacy budget. One is first to determine the number of iterations T, then the privacy budget for each iteration is ε/T; the other method is that the number of iterations is uncertain and the privacy budget for iteration t is $\varepsilon/2^{t+1}$. Considering that as the iterations proceed, the harm to the accuracy of results will increase with the privacy budget decreasing, we adopt the former method in which the number of iterations is fixed and the privacy budget to each iteration is equally allocated to improve the accuracy of the clustering results.

We introduce the following method of determining the number of iterations. Our method considers the impact of using differential privacy techniques on updating centroid. For numerical attributes, we analyze the mean square error (MSE) between noisy centroids and true centroids in one iteration proposed in [16]. And for categorical attributes, we consider the sum of variances caused by added noises. The sum of the MSE and variances is defined as the loss function of optimizing the number of iterations. The loss function of one centroid in one iteration is

$$Loss(\hat{o})=\mathrm{E}\left[\sum_{j=1}^{p}\left(\frac{S_j+\Delta S_j}{C+\Delta C}-\frac{S_j}{C}\right)^2\right]+\sum_{j=p+1}^{d}\sum_{r=1}^{|A_j|}Var(\Delta C_{jr})\tag{4}$$

We try to ensure that each iteration's Loss is no larger than the threshold, and allocate the minimal privacy budget to each iteration.

1) LOSS STUDY OF ONE ITERATION

We analyze the loss between noisy centroids and true centroids in one iteration.

Proposition 2: In one iteration of our proposed algorithm, the Loss is

$$Loss(\hat{o})\approx\frac{2K^2(1+\rho^2)}{N^2\varepsilon^2}+\frac{2(1-\varepsilon)|A|_q}{\varepsilon^2}(|A|_q=\sum_{j=p+1}^{d}|A_j|)\tag{5}$$

Proof: First, on the j-th numerical attribute,

$$MSE(\widehat{o}_j)=E\left[\left(\frac{S_j+\Delta S_j}{C+\Delta C}-\frac{S_j}{C}\right)^2\right]$$

$$= \frac{Var(\Delta S_j)}{C^2} + \frac{S_j^2 Var(\Delta C)}{C^4}.$$

We suppose that the privacy budget allocated to each attribute is equal, and on average $\rho = S_i/C$ and $C \approx N/K$. And noises added to S_j and C generated from same Laplace distribution. Hence, $MSE(\widehat{o_i})$ can be approximated as follows:

$$\sum_{j=1}^{p} MSE(\widehat{o_i}) \approx \frac{K^2 p}{N^2}\left(Var(\Delta S_i) + \rho^2 Var(\Delta C)\right)$$
$$= \frac{K^2 p(1+\rho^2)}{N^2\varepsilon^2}.$$

For categorical attributes, we can obtain that

$$Var(X) = 2\alpha/(1-\alpha)^2, \quad when\ X \sim Geom(\alpha).$$

Then, we can obtain

$$\sum_{j=p+1}^{d}\sum_{r=1}^{|A_j|} Var(\Delta C_{jr})$$
$$= \sum_{j=p+1}^{d} |A_j||Var(\Delta C_{jr})|$$
$$= \frac{2\alpha\,|A|_q}{(1-\alpha)^2} = \frac{2e^{-\varepsilon}\,|A|_q}{(1-e^{-\varepsilon})^2} \approx \frac{2(1-\varepsilon)\,|A|_q}{\varepsilon^2}.$$

From **Proposition 2** we can obtain *Loss* of all centroids is

$$Loss(\hat{O}) = \sum_{k=1}^{K} Loss(\hat{o})$$
$$\approx \frac{2K^3 p(1+\rho^2)}{N^2\varepsilon^2} + \frac{2(1-\varepsilon)K\,|A|_q}{\varepsilon^2}$$
$$= \frac{2K\left(K^2 p(1+\rho^2) + N^2(1-\varepsilon)\,|A|_q\right)}{N^2\varepsilon^2}. \quad (6)$$

2) DETERMINING THE NUMBER OF ITERATION

Based on our analysis above, then we calculate the minimal privacy budget ε^m allocated to each iteration. Suppose that ε^t is the privacy budget allocated to each iteration, then the privacy budget allocated to each attribute is $\varepsilon = \varepsilon_t/(d+1)$. We let the Loss of all centroids in one iteration should be no more than δ, and ε^m is calculated by

$$\frac{2K\left(K^2 p(1+\rho^2) + N^2(1-\varepsilon)\,|A|_q\right)}{N^2\varepsilon^2}$$
$$= \frac{2K\left(K^2 p(1+\rho^2)(1+d)^2 + N^2(1+d)(1+d-\varepsilon_t)\,|A|_q\right)}{N^2\varepsilon_t^2}$$
$$\leq \delta \quad (7)$$

Proposition 3: There exists a minimal ε^m value that satisfies Equation 1.

Proof: Equation 1 can be rewritten as

$$\begin{cases} a\varepsilon_t^2 + b\varepsilon_t + c \geq 0 \\ a = \delta N^2 \\ b = 2KN^2(1+d)|A|_q \\ c = -2KN^2(1+d)^2|A|_q - 2K^3(1+d)^2 p(1+\rho^2) \end{cases}$$

Regarding the unary quadratic equation of ε^t, it is only necessary to satisfy

$$\Delta = b^2 - 4ac \geq 0,$$

and this equation has a solution. And we can obtain that $a > 0, b > 0, c < 0$, and it is clearly that $\Delta = b^2 - 4ac > 0$. So, there are two values ε_1 and ε_2 make $a\varepsilon_t^2 + b\varepsilon_t + c = 0$. Moreover, we can get some properties based on the quadratic equation that $\varepsilon_1 \cdot \varepsilon_2 < 0$. We suppose that $\varepsilon_1 > 0$ and it is the minimal ε^m value from equation 1.

Then we can determine the number of iterations based on ε^m. We use the method proposed in [16]. For $\varepsilon \leq 2\varepsilon_m$, we set the number of iterations T to be 2, and the privacy budget allocated to each iteration is $\varepsilon/2$. For $\varepsilon > 2\varepsilon_m$, T is determined by the following equation:

$$T = \min\left\{7, \frac{\varepsilon}{\varepsilon_m}\right\}.$$

The privacy budget allocated to each iteration is ε/T.

VI. PRIVACY ANALYSIS OF PROPOSED ALGORITHM

Differential privacy has two characteristics: sequence combination and parallel combination, both play an important role in the allocation of privacy budget. If there are m random algorithms $A_1, A_2, \cdots, A_m$, and A_i $(1 \leq i \leq m)$ satisfies ε_i-differential privacy, then for the same dataset D, the sequence combination algorithm $\{A_1, A_2, \cdots, A_m\}$ satisfies ε-differential privacy, in which $\varepsilon = \sum_{i=1}^{m} \varepsilon_i$. If there is a random algorithm M and a dataset D, in which D is divided into disjoint subsets $\{D_1, D_2, \cdots, D_n\}$. If algorithm M satisfies ε-differential privacy, then the algorithm composed of the combination operation of M on $\{D_1, D_2, \cdots, D_n\}$ also satisfies ε-differential privacy.

As described in section 5.3, the privacy of our proposed algorithm is achieved by adding Laplace noise and Geometric noise to the centroids in the process of iterations. Each iteration of our proposed algorithm is equivalent to the sequence combination of the random algorithm, the privacy budget of the entire algorithm is

$$\varepsilon = \sum_{t=1}^{T} \varepsilon_t.$$

In each iteration, $d+1$ noise will be added, including C, S_i of p numerical attributes and value count of q categorical attributes. Since one point is added or deleted to the data set, the maximum change of C and value count of categorical attributes is 1, the global sensitivity of them is 1. If numerical attributes of dataset are normalized to [0, 1], when adding or deleting a point from the dataset, the maximum change of each attribute S_i is 1. So, the global sensitivity of S_i is 1. According to equation 2, in each iteration, adding noise $Lap(1/\varepsilon)$ to C, adding noise $Lap(1/\varepsilon)$ to S_i, adding $Geom(e^{-\varepsilon})$ to count of categorical attributes can make the algorithm satisfy differential privacy, where $\varepsilon = \varepsilon_t/(d+1)$.

TABLE 1. Description of datasets.

Dataset	N	d	p	q	$\|A\|_q$	K
Heart	303	13	5	8	22	5
Adult	48842	9	6	3	26	5

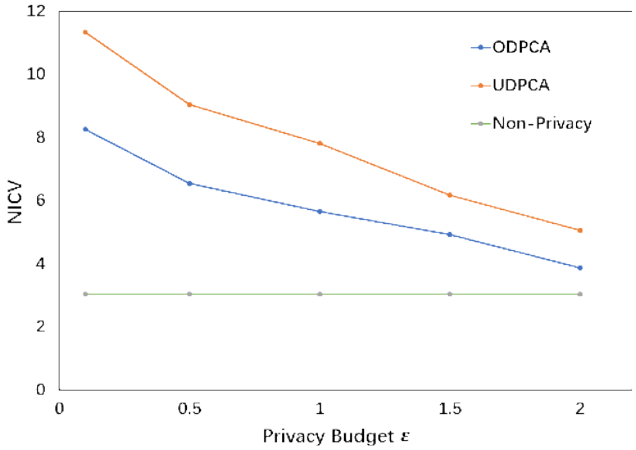

FIGURE 3. NICV of our proposed ODPCA, UDPCA and non-privacy on dataset adult with different privacy budget.

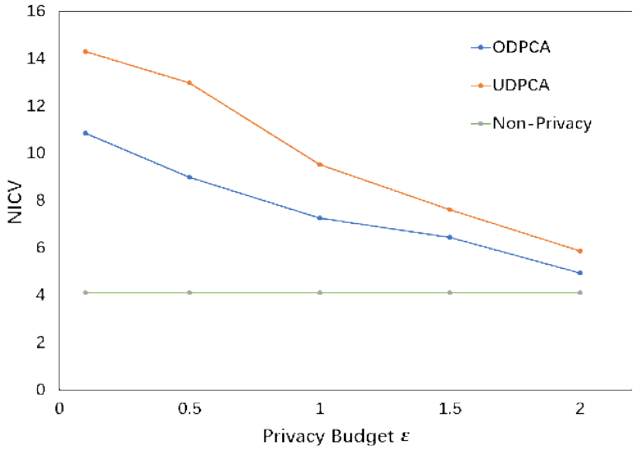

FIGURE 4. NICV of our proposed ODPCA, UDPCA and non-privacy on dataset heart with different privacy budget.

VII. PERFORMANCE EVALUATION

In this section, we conduct experiments to measure NICV of the algorithm for performance evaluation with two datasets. The experimental environment is set up as follows. CPU: Intel Core i7-6700 3.40GHz; RAM: 8GB; System: Windows 10. The clustering algorithm is developed in Python.

A. DESCRIPTION OF DATASETS

We experimented with two datasets Heart and Adult from the UCI Knowledge Discovery Archive database. Table 1 summarizes the two datasets. For the dataset Heart, the number of records is 303, the dimension of records is 13 and the number of clusters is 5. We choose 5 numerical attributes and 8 categorical attributes as the attributes in records. And there are 22 attribute values among all categorical attributes. For the dataset Adult, the number of records is 48842. We choose 6 numerical attributes and 3 categorical attributes as the attributes in records. And there are 26 attribute values among all categorical attributes. We set k=5 for this dataset according to variable ''race'' in the original dataset. And we set $\rho = 0.225$ [16].

B. ACCURACY OF PROPOSED ALGORITHM

In this subsection, we compare the NICV of our proposed algorithms ODPCA with another differentially private cluster algorithm UDPCA, in which the number of iterations is uncertain, and cluster algorithm without privacy preserving through experimental testing. And in order to observe the impact of privacy budget on the availability of clustering result, we test with several different total privacy budget including 0.1, 0.5, 1, 1.5, 2. Results with two mixed datasets are shown in Fig.3 and Fig.4. It can be seen from the experimental results that the NICV values of our algorithm are smaller than the other one, proving that our algorithm outperforms the other algorithm. In addition, as the privacy budget increases, the NICV values of these two differentially privacy clustering algorithms gradually approach the NICV value of non-privacy algorithm. That is, as the degree of privacy preservation decreases, the accuracy of the clustering results increases.

VIII. CONCLUSION

To enable privacy-preserving cluster analysis in SDN-based smart grid, this paper proposed an optimizing and differentially private clustering algorithm for mixed data in smart grid. In our proposed algorithm, we combine the differentially private k-means algorithm and k-modes algorithm to cluster mixed data in a privacy preserving manner and design a mechanism to make the algorithm satisfy differential privacy. And we optimize the allocation of privacy budgets to improve the accuracy of clustering results. Specifically, we analyze the loss function that considers both numerical and categorical attributes between true centroids and noisy centroids, and the number of iterations of clustering algorithm is set to a fixed value determined by the total privacy budget and the minimal privacy budget allocated to each iteration calculated by setting loss function to a threshold. Finally, we prove that our proposed algorithm satisfies differential privacy and

experiment with two datasets to illustrate performance of our proposed ODPCA by comparing the Normalized Intra-Cluster Variance (NICV) produced by our algorithm in different privacy budget values. In the future, we will further improve the accuracy of differentially private clustering algorithm for mixed datasets.

REFERENCES

[1] J. Wu, M. Dong, K. Ota, J. Li, and Z. Guan, "Big data analysis-based secure cluster management for optimized control plane in software-defined networks," *IEEE Trans. Netw. Service Manag.*, vol. 15, no. 1, pp. 27–38, Mar. 2018.

[2] D. Kaur, G. S. Aujla, N. Kumar, A. Y. Zomaya, C. Perera, and R. Ranjan, "Tensor-based big data management scheme for dimensionality reduction problem in smart grid systems: SDN perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 10, pp. 1985–1998, Feb. 2018.

[3] T. Wang, G. Zhang, A. Liu, M. Z. A. Bhuiyan, and Q. Jin, "A secure IoT service architecture with an efficient balance dynamics based on cloud and edge computing," *IEEE Internet Things J.*, to be published.

[4] S. Ali, K. Wu, K. Weston, and D. Marinakis, "A machine learning approach to meter placement for power quality estimation in smart grid," *IEEE Trans. Smart Grid*, vol. 7, no. 3, pp. 1552–1561, Jul. 2016.

[5] E. Mocanu, P. H. Nguyen, W. L. Kling, and M. Gibescu, "Unsupervised energy prediction in a smart grid context using reinforcement cross-building transfer learning," *Energy Buildings*, vol. 116, pp. 646–655, Mar. 2016.

[6] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int. Conf. Theory Appl. Models Comput.*, 2008, pp. 1–19.

[7] Z. Guan, Y. Zhang, L. Zhu, L. Wu, and S. Yu, "EFFECT: An efficient flexible privacy-preserving data aggregation scheme with authentication in smart grid," *Sci. China Inf. Sci.*, vol. 62, no. 3, pp. 1–14, Mar. 2019.

[8] X. Zhang, L. Zhu, X. Wang, C. Zhang, H. Zhu, and Y.-A. Tan, "A packet-reordering covert channel over VoLTE voice and video traffics," *J. Netw. Comput. Appl.*, vol. 126, pp. 29–38, Jan. 2019.

[9] Z. Guan *et al.*, "Privacy-preserving and efficient aggregation based on blockchain for power grid communications in smart communities," *IEEE Commun. Mag.*, vol. 56, no. 7, pp. 82–88, Jul. 2018.

[10] Q. Zhang *et al.*, "A hierarchical group key agreement protocol using orientable attributes for cloud computing," *Inf. Sci.*, vol. 480, nos. 55–69, Apr. 2019.

[11] J. Wu, M. Dong, K. Ota, J. Li, and Z. Guan, "FCSS: Fog-computing-based content-aware filtering for security services in information-centric social networks," *IEEE Trans. Emerg. Topics Comput.*, to be published.

[12] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Conf. Theory Cryptogr.*, 2006, pp. 265–284.

[13] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical privacy: The SULQ framework," in *Proc. 20th ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst.*, 2005, pp. 128–138.

[14] X. Du, M. Guizani, Y. Xiao, and H.-H. Chen, "Transactions papers a routing-driven elliptic curve cryptography based key management scheme for heterogeneous sensor networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1223–1229, Mar. 2009.

[15] Y. Tan *et al.*, "A root privilege management scheme with revocable authorization for Android devices," *J. Netw. Comput. Appl.*, vol. 107, no. 4, pp. 69–82, Apr. 2018.

[16] D. Su, J. Cao, N. Li, E. Bertino, and H. Jin, "Differentially private K-means clustering," in *Proc. ACM Conf. Data Appl. Secur. Privacy*, 2016, pp. 26–37.

[17] P. Mohan, A. Thakurta, E. Shi, D. Song, and D. Culler, "GUPT: Privacy preserving data analysis made easy," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2012, pp. 349–360.

[18] J. Zhang, X. Xiao, Y. Yang, Z. Zhang, and M. Winslett, "PrivGene: Differentially private model fitting using genetic algorithms," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2013, pp. 665–676.

[19] H. H. Nguyen, "Privacy-preserving mechanisms for K-modes clustering," *Comput. Secur.*, vol. 78, pp. 60–75, Sep. 2018.

[20] B. Claessens, P. Vrancx, and F. Ruelens, "Convolutional neural networks for automatic state-time feature extraction in reinforcement learning applied to residential load control," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3259–3269, Jul. 2018.

[21] S. Haben, C. Singleton, and P. Grindrod, "Analysis and clustering of residential customers energy behavioral demand using smart meter data," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 136–144, Jan. 2016.

[22] T. Wang *et al.*, "Data collection from WSNs to the cloud based on mobile fog elements," *Future Gener. Comput. Syst.*, to be published.

[23] C. Liang, Y.-A. Tan, X. Zhang, X. Wang, J. Zheng, and Q. Zhang, "Building packet length covert channel over mobile VoIP traffics," *J. Netw. Comput. Appl.*, vol. 118, pp. 144–153, Sep. 2018.

[24] X. Du, Y. Xiao, M. Guizani, and H.-H. Chen, "An effective key management scheme for heterogeneous sensor networks," *Ad Hoc Netw.*, vol. 5, no. 1, pp. 24–34, Jan. 2007.

[25] Z. Guan *et al.*, "APPA: An anonymous and privacy preserving data aggregation scheme for fog-enhanced IoT," *J. Netw. Comput. Appl.*, vol. 125, pp. 82–92, Jan. 2019.

[26] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, Oct. 2002.

[27] X. Du and H.-H. Chen, "Security in wireless sensor networks," *IEEE Wireless Commun.*, vol. 15, no. 4, pp. 60–66, Aug. 2008.

[28] C Liang, X Wang, X Zhang, Y. Zhang, K. Sharif, and Y.-A. Tan, "A payload-dependent packet rearranging covert channel for mobile VoIP traffic," *Inf. Sci.*, vol. 465, pp. 162–173, Oct. 2018.

[29] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: Privacy beyond K-anonymity," in *Proc. IEEE Int. Conf. Data Eng.*, Apr. 2006, p. 24.

[30] Z. Guan, J. Li, L. Wu, Y. Zhang, J. Wu, and X. Du, "Achieving efficient and secure data acquisition for cloud-supported Internet of Things in smart grid," *IEEE Internet Things J.*, vol. 4, no. 6, pp. 1934–1944, Dec. 2017.

[31] M. Abadi *et al.*, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.

[32] A. Ghosh, T. Roughgarden, and M. Sundararajan, "Universally utility-maximizing privacy mechanisms," in *Proc. ACM Symp. Theory Comput.*, 2009, pp. 351–360.

[33] Y. Xiao, V. K. Rayi, B. Sun, X. Du, F. Hu, and M. Galloway, "A survey of key management schemes in wireless sensor networks," *J. Comput. Commun.*, vol. 30, nos. 11–12, pp. 2314–2341, Sep. 2007.

[34] F. Mcsherry and K. Talwar, "Mechanism design via differential privacy," in *Proc. 48th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Nov. 2007, pp. 94–103.

[35] Z. Guan, Z. Lv, X. Du, L. Wu, and M. Guizani, "Achieving data utility-privacy tradeoff in Internet of medical things: A machine learning approach," *Future Gener. Comput. Syst.*, vol. 98, pp. 60–68, Sep. 2019.

[36] J. Ren, J. Xiong, Z. Yao, R. Ma, and M. Lin, "DPLK-means: A novel differential privacy K-means mechanism," in *Proc. IEEE 2nd Int. Conf. Data Sci. Cyberspace (DSC)*, Jun. 2017, pp. 133–139.

[37] F. D. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2009, pp. 19–30.

[38] C. Dwork, "A firm foundation for private data analysis," *Commun. ACM*, vol. 54, no. 1, pp. 86–95, Jan. 2011.

[39] D. Lam, M. Wei, and D. Wunsch, "Clustering data of mixed categorical and numerical type with unsupervised feature learning," *IEEE Access*, vol. 3, pp. 1605–1613, 2015.

[40] A. Pathak and N. R. Pal, "Clustering of mixed data by integrating fuzzy, probabilistic, and collaborative clustering framework," *Int. J. Fuzzy Syst.*, vol. 18, no. 3, pp. 339–348, 2016.

[41] J.-Y. Chen and H.-H. He, "A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data," *Inf. Sci.*, vol. 345, pp. 271–293, Jun. 2017.

[42] Z. Huang, "Extensions to the K-means algorithm for clustering large data sets with categorical values," *Data Mining Knowl. Discovery*, vol. 2, no. 3, pp. 283–304, Sep. 1998.

[43] Y Li, J Hu, Z Wu, C. Liu, F. Peng, and Y. Zhang, "Research on QoS service composition based on coevolutionary genetic algorithm," *Soft Comput.*, vol. 22, no. 23, pp. 7865–7874, 2018.

ZEFANG LV received the B.Eng. degree from Shandong University, in 2016. She is currently pursuing the master's degree with the School of Mathematics and Physics, North China Electric Power University. Her current research interests include secure machine learning and data privacy.

LIRONG WANG is currently pursuing the bachelor's degree with the School of Control and Computer Engineering, North China Electric Power University. Her current research interests include cyber security and secure machine learning.

ZHITAO GUAN (M'13) received the B.Eng. degree and Ph.D. degree in computer application from the Beijing Institute of Technology, China, in 2002 and 2008, respectively. He is currently an Associate Professor with the School of Control and Computer Engineering, North China Electric Power University. His current research interests include smart grid security, wireless security, and cloud security. He has authored over 50 peer-reviewed journal and conference papers in these areas.

JUN WU received the Ph.D. degree in information and telecommunication studies (GITS) from Waseda University, Japan. He was a Postdoctoral Researcher with the Research Institute for Secure Systems (RISEC), National Institute of Advanced Industrial Science and Technology (AIST), Japan, from 2011 to 2012. He was a Researcher with the Global Information and Telecommunication Institute (GITI), Waseda University, Japan, from 2011 to 2013. He is currently an Associate Professor of electronic information and electrical engineering with Shanghai Jiao Tong University, China. His research interests include the advanced computation and communications techniques of smart sensors, wireless communication systems, industrial control systems, wireless sensor networks, smart grids, and more. He has hosted and participated in several research projects for the National Natural Science Foundation of China, the National 863 Plan and 973 Plan projects, and so on. He is a member of the IEEE. He has been a Guest Editor of the IEEE Sensors Journal and a TPC Member of several international conferences, including WINCON 2011 and GLOBECOM 2015.

XIAOJIANG DU (M'04–SM'09) received the B.S. and M.S. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1996 and 1998, respectively, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland College Park, in 2002 and 2003, respectively. He was an Assistant Professor with the Department of Computer Science, North Dakota State University, from 2004 to 2009. He is currently a Professor with the Department of Computer and Information Sciences, Temple University. His research interests include security, wireless networks, computer networks, and systems. He has published over 200 journal and conference papers in these areas. He is a Life Member of ACM. He received the Excellence in Research Award from the Department of Computer Science, North Dakota State University, in 2009.

HONGTAO ZHAO is currently an Associate Professor with the School of Mathematics and Physics, North China Electric Power University. His current research interests include statistics and machine learning.

MOHSEN GUIZANI (S'85–M'89–SM'99–F'09) received the B.S. and M.S. degrees in electrical engineering and the M.S. and Ph.D. degrees in computer engineering from Syracuse University, in 1984, 1986, 1987, and 1990, respectively. He was the Associate Vice President of Qatar University, the Chair of the Computer Science Department, Western Michigan University, the Chair of the Computer Science Department, University of West Florida, and the Director of graduate studies at the University of Missouri-Columbia. He is currently a Professor with the Department of Computer Science and Engineering, Qatar University. He has authored or coauthored nine books and more than publications in refereed journals and conferences. His research interests include wireless communications and mobile computing, vehicular communications, smart grid, cloud computing, and security.