A Differentially Private Big Data Nonparametric Bayesian Clustering Algorithm in Smart Grid

Zhitao Guan, *Member, IEEE*, Zefang Lv, Xianwen Sun, Longfei Wu, *Member, IEEE*,
Jun Wu, *Member, IEEE*, Xiaojiang Du, *Fellow, IEEE*, and Mohsen Guizani, *Fellow, IEEE*

***Abstract*—Smart systems, including smart grid (SG) and Internet of Things (IoT), have been playing a critical role in addressing contemporary issues. Taking full advantage of the big data generated by the smart grid can enhance the system stability and reliability, increase asset utilization, and offer better customer experience. To better support the data-driven smart grid, the machine learning technologies such as cluster analysis can be applied to process the massive data generated in smart grid. However, the process of cluster analysis may cause the disclosure of personal private information. In this paper, to achieve privacy-preserving cluster analysis in smart grid, we propose IDPC, a Differentially Private Clustering algorithm based on the Infinite Gaussian mixture model (IGMM). IDPC uses a combination of nonparametric Bayesian method and differential privacy. The nonparametric Bayesian method allows certain parameters to change along with the data and it is usually adopted in a clustering algorithm without a fixed number of clusters. The Laplace mechanism is used in data releasing process to make IDPC differentially private. We present how to make the nonparametric Bayesian clustering algorithm differentially private by adding Laplace noise. By security analysis and performance evaluation, IDPC is proved to be privacy-preserving as well as efficient.**

***Index Terms*—Differential privacy, Nonparametric Bayesian Method, Clustering, Big data, Smart grid.**

I. Introduction

SMART grid, a sensor-embedded smart electricity system allowing two-way communication between the utility and customers, is transformed from the traditional grids to achieve reliable, safe, economical, efficient, and environmentally friendly use of the grid. With the proliferation of intelligent devices, the data collected from the smart grid has also grown exponentially [1], [2]. The analytics of these data is the key to intelligently control of the production and distribution of electricity. As a promising solution, the advanced machine learning/deep learning (ML/DL) techniques can be applied to smart grid for data analysis, such as user electricity behavior analysis, power equipment monitoring and user classification [3]–[5].

However, the vast amount of data generated, processed and exchanged in smart grid is usually security-critical and privacy-sensitive, such as user information, power transmission and distribution data, hence has become the target of various attacks [6]–[8]. Although these data are of great value, the processing and analysis of involved sensitive information may cause leakage of users' privacy. Therefore, it is critical to guarantee privacy-preserving data analysis in smart grid. Differential privacy is proposed in [9] ensuring the privacy of all individuals in a dataset. One commonly adopted technique to achieve differential privacy is to add random noises, which obey a distribution that satisfies specific conditions, during data analysis or when publishing analysis results [10], [11].

Clustering is an important method in unsupervised learning. The main idea of clustering is to divide the dataset into several clusters according to the similarity between the data points, so that the similarities of data points in the same cluster are as high as possible, and the similarities between data points in different clusters are as low as possible. The algorithms we use for clustering, such as k-means and Gaussian Mixture modeling, need to specify the number of clusters in advance. However, in practice, due to the lack of previous knowledge of the datasets, we cannot accurately determine the number of clusters. In addition, for many real-world datasets, the number of clusters is uncertain. The nonparametric Bayesian method refers to a class of techniques that allows certain parameters to change with the data and it can be used to perform clustering without a fixed number of clusters. There have been some works on how to apply nonparametric Bayesian method in a clustering task [12], [13]. However, there has been no research about how to make it privacy preserving.

In our proposed algorithm, we combine the nonparametric Bayesian method and differential privacy to achieve privacy-preserving clustering with uncertain number of clusters. Instead of blindly set the number of clusters, we allow it to grow as more data are observed. Meanwhile, we make the nonparametric Bayesian clustering algorithm differentially private through some mechanisms.

Manuscript received September 10, 2019; revised February 13, 2020; accepted March 22, 2020. Date of publication April 6, 2020; date of current version December 30, 2020. The work was supported in part by Beijing Natural Science Foundation under Grant 4182060 and in part by the National Natural Science Foundation of China under Grant 61972148. Recommended for acceptance by Dr. Tarik Taleb. (*Corresponding author: Zhitao Guan.*)

Zhitao Guan, Zefang Lv, and Xianwen Sun are with the School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China (e-mail: guan@ncepu.edu.cn; 505098614@qq.com; 1186383599@qq.com).

Longfei Wu is with the Department of Mathematics and Computer Science, Fayetteville State University, Fayetteville, NC 28301 USA (e-mail: lwu@uncfsu.edu).

Jun Wu is with the Shanghai Jiaotong University, Shanghai 200240, China (e-mail: junwuhn@sjtu.edu.cn).

Xiaojiang Du is with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122 USA (e-mail: dxj@ieee.org).

Mohsen Guizani is with the Department of Computer Science and Engineering, Qatar University, Doha 2713, Qatar (e-mail: mguizani@ieee.org).

Digital Object Identifier 10.1109/TNSE.2020.2985096

2327-4697 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: ULAKBIM UASL - SELCUK UNIVERSITESI. Downloaded on December 07,2023 at 07:03:16 UTC from IEEE Xplore. Restrictions apply.

The contributions of our paper are mainly as follows:

1. We propose IDPC, an IGMM-based differentially private clustering algorithm for smart grid. IDPC combines the nonparametric Bayesian method with differential privacy to cluster data with non-fixed number of clusters in a private manner.
2. We propose a method to ensure that the algorithm satisfies differential privacy through some mechanisms. Specifically, we present how to make the nonparametric Bayesian clustering algorithm differentially private and show the detailed mechanism of adding noises.
3. We theoretically analyze the security of proposed algorithm and prove the efficiency through numerical experiments on two datasets. The experimental results demonstrate that proposed IDPC can achieve a tradeoff between privacy and utility.

The rest of this paper is organized as follows. In section II, we present the related work. Section III gives some preliminaries. In section IV, our model and its design goals are stated. Section V shows details of our proposed IDPC. In section VI, we show the security analysis. In Section VII, we evaluate the performance of IDPC. In Section VIII, the paper is concluded.

II. Related Work

With the rapid development of sensing and control technologies and the wide deployment of sensors, massive data such as power generation/consumption information and equipment information is produced in the smart grid [14]. Exploiting big data technologies to analyze these data to understand the state of the system can provide a basis for improving the stability and efficiency of the grid [15], [16]. Cluster analysis, a vital technology for big data analytics [17], [18], has wide applications in smart grid, such as user behavior analysis, power management, and equipment fault detection [19]–[21]. Reference [19] applied cluster analysis to residential smart meter data to discover behavior groups, which classified customers based on their demand and variability. Reference [20] used nodes and links to represent the buses and power transmission lines, abstracted the power transmission system into a network, and then mines the internal structure through hierarchical spectral clustering.

However, some sensitive information may be involved when applying advanced ML/DL techniques for data analysis, such as individual information and electricity usage data. Security and privacy issues in smart grids have always been a key research area [6]. There are some existing solutions aiming at preserving the private information of smart grid users [22]–[24]. Reference [22] proposed a practical privacy-preserving data aggregation scheme without using a trusted third party, in which user's personal data is masked within a virtual aggregation area while ensuring that the impact on the aggregation result is negligible.

There have been many works on data privacy protection for data analysis, including k-anonymity [25], l-diversity [26], differential privacy [9], and so on. To tackle with the problem that an attacker can steal data privacy by correlating background knowledge and clustering results, reference [27] applied differential privacy to clustering algorithm and proposed a k-means clustering method DP k-means (Differential Private k-means) which can cope with any background knowledge. This method realizes privacy preservation by adding appropriate random noise to the intermediate variables such as the sum of clustering records and the number of records in the iterative process of k-means algorithm. There have been many researches on how to improve the accuracy of differentially private clustering algorithms, from the perspective of privacy budget allocation and improving the clustering algorithms [28], [29]. Several papers (e.g. [30], [31]) have studied related security issues.

For the purpose of addressing the problem that it is difficult to determine the number of clusters in some datasets, the nonparametric Bayesian method has been applied in clustering algorithm [32], [33]. In nonparametric Bayesian clustering, we need not to specify the number of clusters to be a fixed number in advance and it can change along with the data. Reference [32] used the nonparametric Bayesian method to determine the quantity of transmitting devices in the primary user spectrum and identify the primary user emulation attacks. Nevertheless, existing nonparametric Bayesian clustering algorithms have not considered that the release of results may disclose private information in the dataset. In IDPC, we combine the nonparametric Bayesian method and differential privacy to achieve privacy-preserving nonparametric Bayesian clustering algorithm.

III. Preliminaries

We introduce the definition of differential privacy and Dirichlet process in this section.

A. Differential Privacy

Through some mechanisms to change the distribution of query results, so that an adversary cannot obtain the personal privacy information by comparing two queries results on neighboring datasets, and one can realize the differential privacy protection of the individual privacy information.

Definition 1 ε -differential privacy:

If for any two neighboring datasets *D* and *D'*, a mechanism *F* and any possible output $O \in Range(F)$, there exists

$$\Pr(F(D) \in O) \leq e^{\varepsilon} \cdot \Pr(F(D') \in O) \tag{1}$$

Then, *F* satisfies ε -differential privacy. The neighboring datasets D and D' refer to two datasets with one data point different at most. ε represents the privacy budget and it denotes the level of privacy guarantee. The smaller the privacy budget, the greater the degree of privacy preserving.

B. Dirichlet Process

We first review the definition of the beta distribution and the Dirichlet distribution. Beta distribution is a conjugate prior to

Authorized licensed use limited to: ULAKBIM UASL - SELCUK UNIVERSITESI. Downloaded on December 07,2023 at 07:03:16 UTC from IEEE Xplore. Restrictions apply.

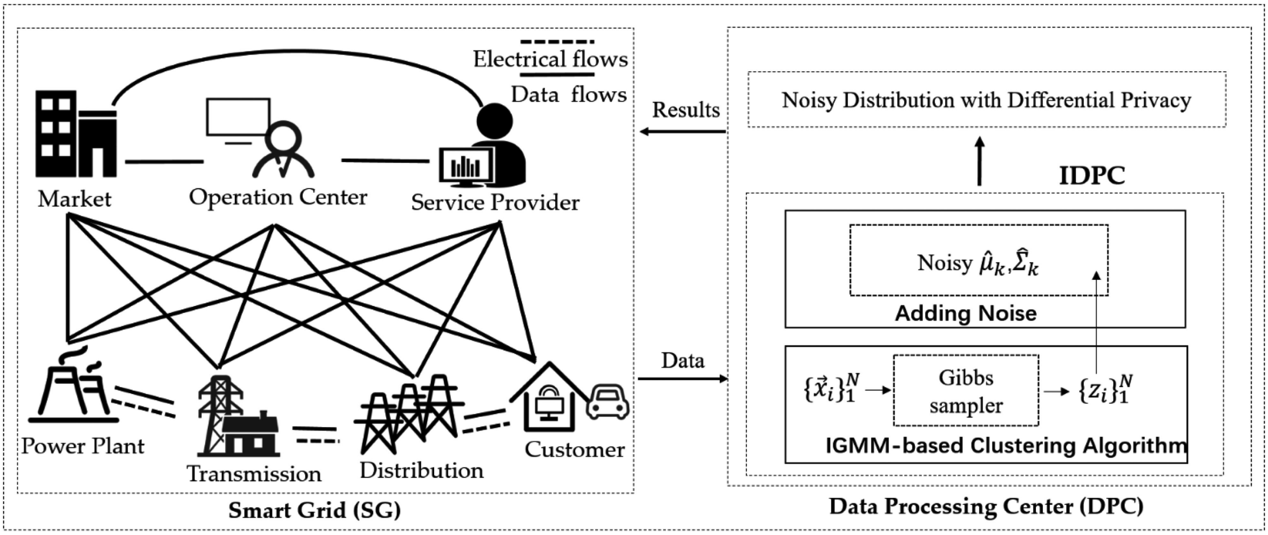


Fig. 1. Our Proposed IDPC Algorithm in Smart Grid.

the binomial distribution. Given $\alpha > 0, \beta > 0$, its probability density function is

$$f(x; \alpha, \beta) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, x \in [0,1] \\ 0, else \end{cases} \tag{2}$$

where

$$\frac{1}{B(\alpha, \beta)} = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{3}$$

$\Gamma(\cdot)$ is the Gamma function and

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt.$$

Dirichlet distribution [34] is defined as the distribution on the K-dimensional probability simplex $\{\omega_1, \omega_2, \ldots, \omega_K\}$, with $\sum_{i=1}^K \omega_i = 1$. Given $\alpha_1, \alpha_2, \ldots, \alpha_K > 0$, the probability density function of Dirichlet distribution is

$$f(\omega_1, \ldots, \omega_K; \alpha_1, \ldots, \alpha_K) = \begin{cases} \frac{\prod_{i=1}^K \omega_i^{\alpha_i - 1}}{B(\alpha)}, & \omega_i \in [0,1] \\ 0, & else \end{cases} \tag{4}$$

where

$$B(\alpha) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)} \tag{5}$$

Definition 2 **Dirichlet Process (DP)** [35]:

The random distribution G on sample space Θ is a Dirichlet process, if for every finite measurable partition $(A_1, A_2, \ldots, A_k)$ of Θ satisfies

$$(G(A_1), \ldots, G(A_k)) \sim Dir(\alpha H(A_1), \ldots, \alpha H(A_k)) \tag{6}$$

and G is written as

$$G \sim DP(\alpha, H) \tag{7}$$

H is a base distribution, which can be seen as the expectation of G, $E(G(A)) = H(A)$; α is the concentration parameter, which stands for the strength of prior.

IV. Models and Goals

A. An Overview of Our Model

Fig. 1 shows our system model. In this system, massive amount of data, such as the electricity usage data generated from smart home, is transferred to DPC (Data Processing Center). And the analysis results obtained in DPC will be returned to smart grid. The analysis results can then be applied to increase the efficiency of management in power systems and enhance the stability and reliability of smart grid.

Nevertheless, in the process of data analysis and result release, personal private information may be leaked. In IDPC, we achieve differentially private data analysis by adding random noises to the released analysis results to protect the sensitive information.

B. Design Goals

To solve the problem of privacy preservation while clustering the massive data generated in smart grid, we design our algorithm with the following two major goals:

1) Privacy preservation: considering a pair of neighboring datasets, regardless of how much background knowledge an adversary owns, he cannot obtain any specific information of an individual by accessing the statistical data of associated dataset or the released analytic results.
2) Accuracy: considering the tradeoff between the accuracy and the degree of privacy preservation, a balance must be achieved.

Authorized licensed use limited to: ULAKBIM UASL - SELCUK UNIVERSITESI. Downloaded on December 07,2023 at 07:03:16 UTC from IEEE Xplore. Restrictions apply.

C. Security Model

In our system, we assume that the DPC is trustful. However, an adversary has the capability to alter the datasets transmitted to DPC and has the access to the data analysis results released. Differential privacy can achieve a strong privacy guarantee by changing the distribution of query results on datasets through some mechanisms, such as adding noise to query results, so that an adversary cannot get specific individual information by comparing two query results on two neighboring datasets.

V. Description of Our Scheme

We will give the details our scheme in this section. We proposed a differentially private nonparametric Bayesian clustering algorithm based on infinite Gaussian mixture modeling to address the issues of uncertain number of clusters and privacy disclosure in the clustering process. The nonparametric Bayesian method for clustering and its combination with differential privacy will be explained below.

We assume that a d-dimensional dataset $D = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$ has N data points. Our proposed algorithm IDPC mainly consists of two parts. In the first part, we utilize the nonparametric Bayesian method to cluster the dataset, whose number of clusters is uncertain or unknown to us. The main goal of this part is to find the cluster label $\vec{z} = \{z_1, z_2, \ldots, z_N\}$. Specifically, $z_i = k$ demonstrates that x_i belongs to *kth* cluster, and two data points having the same cluster label belong to the same cluster. The second part is to utilize the differential privacy to address the issue of privacy disclosure during the releasing process of the clustering results.

A. Our Proposed Algorithm

Clustering is an important method in unsupervised learning. The algorithms we use to cluster, such as k-means and Gaussian Mixture modeling, need to specify the number of clusters to be a fixed number. However, in practice, due to the lack of experience and background knowledge of the datasets, we usually cannot determine the number of clusters accurately. Additionally, for many real-world datasets, the number of clusters is uncertain. Instead of fixing the number of clusters to be discovered, we allow it to grow as more data are observed. The nonparametric Bayesian method refers to a class of techniques that allows certain parameters to change with the data.

The proposed algorithm we designed is to address the privacy issue of the nonparametric Bayesian clustering algorithm. The main idea is to apply nonparametric Bayesian method to cluster dataset and then apply differential privacy to the released cluster results so that the results will not disclose individual information about the dataset. In the result releasing process, privacy preservation is realized by adding random noises to the parameters of released distributions. We use the nonparametric Bayesian method to solve cluster task and apply differential privacy to it. The detailed process of our proposed algorithm is outlined in **Algorithm 1**. The first part of the algorithm is to implement the clustering algorithm in dataset D and obtain cluster label for each data point. The second part is to estimate the parameters of each Gaussian distribution (for every cluster) given the data points involved and add noises to parameters to ensure the released distributions satisfy differential privacy. Then, the released noisy distribution estimated from dataset can be utilized to obtain information about the dataset and predict the cluster label for new observations without revealing the privacy of raw dataset. We will discuss these two parts in detail in the following subsections.

Algorithm 1: IGMM-based Differentially Private Clustering Algorithm

```
Input: d-dimensional dataset D: $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$. $\varepsilon$ : privacy budget.
Output: noisy Gaussian mixture distribution with K components
Assign all data points into one cluster, $K = 1$.
for $iter = 1 \rightarrow T$ do
  for $i = 1 \rightarrow N$ do
    Remove $\vec{x}_i$ from its current cluster.
    for $k = 1 \rightarrow K$ do
      Compute probability, $p_k$, of assigning $\vec{x}_i$ to an existing cluster k.
    end for
    Compute probability, $p_{K+1}$, of assigning $\vec{x}_i$ to an unpresented cluster.
    Sample $z_i$ according to $\{p_1, p_2, \ldots, p_{K+1}\}$.
    if $z_i > K$ then
      update $K = K+1$.
  end for
end for
for $k = 1 \rightarrow K$ do
  $O_k = \{\vec{x}_i | z_i = k, i = 1, \ldots, N\}$
  Compute $\omega_k = \frac{1}{N}\sum_{i=1}^{N} p_{ik}$, $p_{ik}$ stands for the posterior probability that $\vec{x}_i$ belongs to cluster k.
  for $j = 1 \rightarrow d$ do
    $\mu_{kj} = \frac{\sum_{x_l \in O_k} x_{lj} + noise}{|O_k| + noise}$
  end for
    $\Sigma_k = \frac{1}{|O_k|}\sum_{x_l \in O_k} (\vec{x}_l - \vec{\mu}_k)(\vec{x}_l - \vec{\mu}_k)^T$
end for
return noisy Gaussian mixture model
    $p(x) = \sum_{i=1}^{K} \omega_i \mathcal{N}(\vec{\mu}_i, \Sigma_i)$
```

The nonparametric Bayesian method mainly consists of two parts, generative model and inference model. The generative model is related to data generation, which gives a hypothesis about how the observations are generated and from which distribution. Inference model is used to estimate the associated parameters given the observations based on the generative model. Both of them will be discussed in the next two subsections.

B. Generative Model

In this subsection, we describe two generative model for nonparametric Bayesian clustering. In our proposed scheme, we assume that observations are generated from mixture Gaussian modeling. Next, we will describe the generative models for fixed number of clusters and non-fixed number of clusters based on Gaussian mixture modeling, respectively. The model for fixed number of clusters is an extension of the model for non-fixed number of clusters.

Authorized licensed use limited to: ULAKBIM UASL - SELCUK UNIVERSITESI. Downloaded on December 07,2023 at 07:03:16 UTC from IEEE Xplore. Restrictions apply.

1) Fixed Number of Clusters: We first start with the generative model for fixed number of clusters called finite Gaussian mixture model (FGMM) [36], then extend it to the generative model for non-fixed number of clusters named infinite Gaussian mixture model.

We assume that there are K mixture weights to model the dataset *D*, and the probability density function is

$$p(x) = \sum_{i=1}^{K} \omega_i \mathcal{N}(\vec{\mu}_i, \Sigma_i) \tag{8}$$

where ω_i is the weight of the *i*th component in the mixture model, with

$$\sum_{i=1}^{K} \omega_i = 1, 0 \le \omega_i \le 1 \tag{9}$$

N is the multivariate Gaussian distribution and its probability density function is given as

$$p(\vec{x}|\vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{2/d} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu}) \right) \tag{10}$$

The mixture weight ω_i represents the probability of which belongs to the *i*th cluster, i.e., $p(z_j = i) = \omega_i$. The parameters $\vec{\mu}_i$ and Σ_i are the mean vector and covariance matrix of the *i*th Gaussian mixture, respectively. In FGMM, the mixture weights are assumed to follow a $Dir(\vec{M})$ distribution and the parameters of mixture are assumed to follow the base distribution $\vec{H}$. Under the above assumptions, we can describe FGMM as follow. We first create the mixture weights $\vec{\omega} = \{\omega_1,\ \omega_2, \ldots, \omega_K\}$ following a $Dir(\vec{M})$ distribution. Given the weights, we can generate the cluster label for each data point following a multinomial distribution. Then we can know which cluster each data point belongs to, i.e., we can determine the distribution each data point follows. Given the distribution, we can then generate random samples.

The inverse Wishart distribution is chosen to be the prior distribution of the Gaussian distribution [37]. According to [38], we compute the conjugate prior distribution for mean vector $\vec{\mu}_i$ and covariance matrix Σ_i with Gaussian inverse Wishart (GIW) distribution as following

$$\Sigma_i \sim \mathrm{IW}_{\upsilon_0}(\Lambda_0^{-1}) \tag{11}$$

$$\vec{\mu}_i | \Sigma_i \sim \mathcal{N}(\vec{\mu}_0, \Sigma_i / \kappa_0) \tag{12}$$

where IW is the inverse Wishart distribution, $\Lambda_0^{-1}, \upsilon_0,\ \vec{\mu}_0, \kappa_0$ are the hyperparameters and integrated into $\vec{H}$. The hyperparameters $\vec{H}$ can be interpreted as following: $\vec{\mu}_0$ is the prior mean for $\vec{\mu}_i$, and κ_0 indicates confidence about it. Λ_0^{-1} is prior about Σ_i, and υ_0 represents the confidence about that. The probability density function of the inverse Wishart distribution is given as

$$p(\Sigma) = \frac{\left|\Lambda^{-1}\right|^{\upsilon/2} |\Sigma|^{-\frac{\upsilon+d+2}{2}} \exp\left[-\frac{tr(\Sigma_i^{-1}\Lambda^{-1})}{2} \right]}{2^{\frac{\upsilon d}{2}} \Gamma_d(\upsilon/2)} \tag{13}$$

where Λ is a $d \times d$ scale matrix, $\mathrm{tr}(\cdot)$ represents the sum of the diagonal elements of a matrix, and $\Gamma_d(\cdot)$ is the multivariate Gamma function. Then, the conjugate prior probability density function is written as

$$p(\vec{\mu}_i, \Sigma_i) = \mathrm{GIW}(\vec{\mu}_i, \Sigma_i | \Lambda_0^{-1}, \upsilon_0, \vec{\mu}_0, \kappa_0) \tag{14}$$

GIW is defined as

$$\begin{aligned} &\mathrm{GIW}(\vec{\mu}_i, \Sigma_i | \vec{H}) \\ &\sim \mathcal{N}(\vec{\mu}_0, \Sigma_i / \kappa_0) \cdot \mathrm{IW}_{\upsilon_0}(\Sigma_i | \Lambda_0^{-1}, \upsilon_0) \\ &= \frac{|\Sigma|^{-\frac{\upsilon_0+d+2}{2}} \exp\left[-\frac{\kappa_0}{2} (\vec{\mu}_i - \vec{\mu}_0)^2 \Sigma_i^{-1} - \frac{tr(\Sigma_i^{-1}\Lambda_0^{-1})}{2} \right]}{2^{\frac{\upsilon_0 d}{2}} \Gamma_d(\upsilon_0/2)(2\pi/\kappa_0)^{\frac{d}{2}} \left|\Lambda_0^{-1}\right|^{-\frac{\upsilon_0}{2}}} \end{aligned} \tag{15}$$

An important but difficult problem in FGMM is how to determine the number of clusters. In practice, due to the lack of experience and background knowledge of the datasets, we usually cannot determine the number of clusters accurately, and for many real-world datasets, the number of clusters is uncertain. Therefore, the FGMM cannot model these datasets. Instead, the IGMM will be adopted to solve this problem, as described in the next section.

2) Non-Fixed Number of Clusters: In the FGMM, the number of clusters is assumed to be a fixed number. However, in reality, we usually cannot obtain exact knowledge about the number of clusters. IGMM can be used to solve this problem by setting $K \to \infty$ in FGMM. In IGMM, we assume that the number of clusters is infinite but it is a finite number at a certain time. In the FGMM, we choose Dirichlet distribution as the prior of mixture weights. However, in the infinite case, we cannot obtain mixture weights directly by sampling from Dirichlet distribution. Instead, infinite mixture weights are sampled by another process named as the stick breaking construction [29] and it is defined as follow. Suppose that there is a stick with length 1 and we let $\beta_k {\sim} Beta(1, \alpha) for\ k = 1, 2, 3 \ldots,$ which are regarded as fractions for how much we take away from the remainder of the stick every time. Then the mixture weights $\{\omega_k\}_{k=1}^{\infty}$ can be calculated by the length we take away each time and this process can be written as follow:

$$\omega_1 = \beta_1, \omega_2 = (1 - \beta_1)\beta_2, \ldots, \omega_k = \beta_k \prod_{j=1}^{k-1} \left(1 - \beta_j\right), \ldots \tag{16}$$

Then we can obtain the infinite mixture weights with $\sum_{k=1}^{\infty} \omega_k = 1$.

Fig. 2 shows the graphical representation of the infinite Gaussian mixture modeling and illustrates that how the data points are generated. In IGMM, the number of clusters is

Authorized licensed use limited to: ULAKBIM UASL - SELCUK UNIVERSITESI. Downloaded on December 07,2023 at 07:03:16 UTC from IEEE Xplore. Restrictions apply.

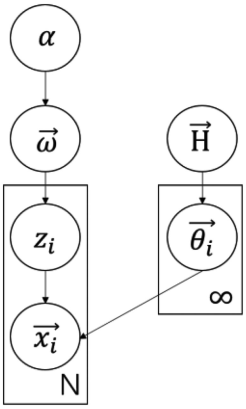


Fig. 2. Illustration of IGMM.

uncertain and assumed to be infinite. Each cluster can be described as a multivariate Gaussian distribution as in the FGMM, which can be denoted by parameters including the mean vector $\vec{\mu}$ and the covariance matrix Σ, which are integrated into $\vec{\theta}$ in Fig. 2. Different to the finite case, the mixture weights obtained from the hyperparameter α is calculated by the stick breaking construction defined above.

C. Inference Model

In this subsection, we describe how we use nonparametric Bayesian methods to solve the clustering problem, i.e., to find the cluster label, z_i, for each data point in *D*. Now, our goal is to find the parameter $\vec{z} = \{z_1, z_2, \ldots, z_N\}$ given the observations. Given the prior distributions of other parameters in the generative model, we want to find the joint distribution $\vec{z} = \{z_1, z_2, \ldots, z_N\}$ and then we can sample from this distribution to obtain the cluster labels. In our proposed algorithm, due to the difficulty of deriving the expression of posterior distribution, Gibbs sampler [39], an efficient method to generate samples from the univariate distribution, is adopted to obtain samples following the joint distribution approximately.

In the process of Gibbs sampler, we need to sample from the conditional distribution of z_i given the other cluster labels $\vec{z}_{-i}$ and $\vec{z}_{-i}= \{\mathrm{z}_1, \cdots \mathrm{z}_{i-1}, \mathrm{z}_{i+1}, \ldots, \mathrm{z}_N\}$. By applying the Bayes' rule, we can obtain the posterior distribution of z_i as following:

$$\begin{aligned} & P(z_i = k|\vec{z}_{-i}, \vec{x}, \alpha, \vec{H}) \\ & = P(z_i = k|\vec{z}_{-i}, \vec{x}_i, \vec{\theta}_k, \alpha, \vec{H}) \\ & \sim P(z_i = k|\vec{z}_{-i}, \alpha) P(\vec{x}_i|z_i = k, \vec{z}_{-i}, \vec{\theta}_k, \alpha, \vec{H}) \\ & \sim P(z_i = k|\vec{z}_{-i}, \alpha) P(\vec{x}_i|\vec{z}_{-i}, \vec{\theta}_k, \vec{H}) \end{aligned} \tag{17}$$

We omit the normalized factor in the formula above. We can easily find that $P(\vec{x}_i|\vec{z}_{-i}, \vec{\theta}_k, \vec{H})$ is the likelihood and obtain a Gaussian distribution according to our assumption. In order to determine the expression of posterior distribution of the cluster label z_i, we need to derive the expression of $P(z_i = k|\vec{z}_{-i}, \alpha)$ in the above formula.

From the generative model described in the previous subsection, we can know that the cluster label for each data point is generated from the mixture weights. So, we need to integrate $\vec{\omega}$ and give the prior distribution of cluster labels

$$p(\vec{z}|\alpha) = \int_{\omega} p(\vec{z}|\omega) p(\omega|\alpha) d\omega \tag{18}$$

where

$$p(\vec{z}|\vec{\omega}) = \prod_{k=1}^{K} \omega_k^{n_k} \tag{19}$$

n_k is the number of data points involved in the *kth* cluster. And

$$\begin{aligned} p(\vec{\omega}|\alpha) & \sim Dir(\alpha/K, \alpha/K, \ldots, \alpha/K) \\ & = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \omega_k^{\frac{\alpha}{K}-1} \end{aligned} \tag{20}$$

Hence, we have

$$\begin{aligned} p(\vec{z}|\alpha) & = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \int_{\vec{\omega}} \prod_{k=1}^{K} \omega_j^{n_k + \frac{\alpha}{K} - 1} \\ & = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha/K)}{\Gamma(\alpha/K)} \end{aligned} \tag{21}$$

In order to obtain the conditional prior for a single cluster label given the others, we keep all but a single cluster label fixed in the above and we can obtain

$$p(z_i = k|\vec{z}_{-i}, \alpha) = \frac{n_{-i,k} + \alpha/K}{N + \alpha - 1} \tag{22}$$

where $n_{-i,k}$ represents the number of data points in the *kth* cluster before $\vec{x}_i$ are observed. To satisfy the infinite case, we let $K \rightarrow \infty$ and the conditional prior distribution can be written as

$$p(z_i = k|\vec{z}_{-i}, \alpha) = \begin{cases} \frac{n_{-i,k}}{N+\alpha-1}, if n_{-i,k} > 0 \\ \frac{\alpha}{N+\alpha-1}, if n_{-i,k} = 0 \end{cases} \tag{23}$$

where $n_{-i,k}$=0 means that there is no data point assigned to the *kth* cluster.

As for another term $P(\vec{x}_i|\vec{z}_{-i}, \vec{\theta}_k, \vec{H})$ in formula (17), we also need to find two expressions as $p(z_i = k|\vec{z}_{-i}, \alpha)$. According to [38], due to our choice of conjugate prior, we can obtain the expression of $P(\vec{x}_i|\vec{z}_{-i}, \vec{\theta}_k, \vec{H})$ by the multivariate Student-t distribution. Therefore, we can obtain that

$$P(\vec{x}_i|\vec{z}_{-i}, \vec{\theta}_k, \vec{H}) \sim t_{\upsilon_n - d + 1}\left(\vec{\mu}_n, \frac{\Lambda_n(\kappa_n + 1)}{\kappa_n(\upsilon_n - d + 1)}\right) \tag{24}$$

where *t* represents the multivariate Student-t distribution and $\upsilon_n - d + 1$ denotes the number of degrees of freedom. The other parameters are defined as follows:

Authorized licensed use limited to: ULAKBIM UASL - SELCUK UNIVERSITESI. Downloaded on December 07,2023 at 07:03:16 UTC from IEEE Xplore. Restrictions apply.

$$\begin{aligned}\vec{\mu}_n &= \frac{\kappa_0}{\kappa_0+N}\vec{\mu}_0 + \frac{N}{\kappa_0+N}\bar{x}\\ \kappa_n &= \kappa_0 + N\\ \upsilon_n &= \upsilon_0 + N\\ \Lambda_n &= \Lambda_0 + \mathbf{S} + \frac{\kappa_0 n}{\kappa_0+N}(\bar{x}-\vec{\mu}_0)(\bar{x}-\vec{\mu}_0)^T \end{aligned} \tag{25}$$

where $\bar{x}$ is the mean of data points in *D*, *d* is the dimension of data point. $\vec{\mu}_l, \kappa_l, \upsilon_l, \Lambda_l$ are the updated hyperparameters after observing a new data point and *S* is defined as

$$\mathbf{S} = \sum_{i=1}^{N}(\vec{x}_i - \bar{x})(\vec{x}_i - \bar{x})^T \tag{26}$$

For the case that assigning a data point to an unpresented cluster, we can obtain the expression of $p(\vec{x}_i, \vec{H})$ as:

$$p(\vec{x}_i, \vec{H}) \sim t_{\upsilon_o - d+1}\left(\vec{\mu}_0, \frac{\Lambda_0(\kappa_0+1)}{\kappa_0(\upsilon_0-d+1)}\right) \tag{27}$$

And the probability density function of the multivariate Student-*t* distribution is given as:

$$\begin{aligned} t_v(\vec{x}|\vec{\mu},\Lambda) &= \frac{\Gamma((d+v)/2)}{\Gamma(v/2)}\frac{|\Lambda|^{1/2}}{(\pi\upsilon)^{d/2}}\\ &\quad\times\left[1+\frac{(\vec{x}-\vec{\mu})^2\Lambda^{-1}}{\upsilon}\right]^{-(d+v)/2} \end{aligned} \tag{28}$$

where υ is the number of degree of freedom, $\vec{\mu}$ is the mean vector, and Λ is a $d \times d$ scale matrix.

In conclusion, we obtain posterior distributions for two cases. The first case is assigning a data point to an existing cluster and the expression of distribution is written as:

$$\begin{aligned} &P(z_i = k|\vec{z}_{-i}, \vec{x}, \alpha, \vec{H})\\ &\sim \frac{n_{-i,k}+\alpha/K}{N+\alpha-1} t_{\upsilon_n - d+1}\left(\vec{\mu}_n, \frac{\Lambda_n(\kappa_n+1)}{\kappa_n(\upsilon_n-d+1)}\right) \end{aligned} \tag{29}$$

Another case is assigning a data point to an unpresented cluster, in which there is no data point assigned, and the expression of distribution is written as:

$$\begin{aligned} &P(z_i \neq j, \forall j \neq i, |\vec{z}_{-i}, \vec{x}, \alpha, \vec{H})\\ &\sim \frac{\alpha}{N+\alpha-1} t_{\upsilon_0 - d+1}\left(\vec{\mu}_0, \frac{\Lambda_0(\kappa_0+1)}{\kappa_0(\upsilon_0-d+1)}\right) \end{aligned} \tag{30}$$

Fig. 3 shows the process of using nonparametric Bayesian method to perform clustering algorithm.

D. Adding Noises

In this subsection, we will present how to make the nonparametric Bayesian clustering algorithm differentially private. The basic idea is to derive the sensitivity of the Gaussian distribution parameters and then to add Laplace noise, so that the released distribution is guaranteed to be differentially private. After the first part of our proposed algorithm, we will obtain cluster labels for each data point and we can estimate the parameters of the Gaussian distribution for every cluster given the data points involved. However, releasing the final distribution directly may cause privacy disclosure. A malicious analyst can mine individual information by analyzing the released clustering results. In our algorithm IDPC, privacy preservation is realized by adding noise when computing the parameters of the released Gaussian distributions. Given the data points sampled from a multivariate Gaussian distribution, we need to estimate the mean vector $\vec{\mu}$ and covariance matrix Σ. Their maximum likelihood estimates are written as

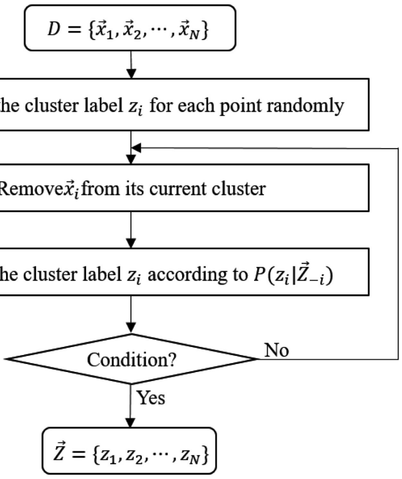

$D = \{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_N\}$
Initialize the cluster label z_i for each point randomly
Remove$\vec{x}_i$from its current cluster
Sample the cluster label z_i according to $P(z_i|\vec{Z}_{-i})$
Condition?
No
Yes
$\vec{Z} = \{z_1, z_2, \cdots, z_N\}$

Fig. 3. Graphical model representation of IGMM-based clustering algorithm.

$$\hat{\vec{\mu}} = \frac{1}{M}\sum_{i=1}^{M}\vec{y}_i = \bar{y} \tag{31}$$

$$\hat{\Sigma} = \frac{1}{M}\sum_{i=1}^{M}(\vec{y}_i - \bar{y})(\vec{y}_i - \bar{y})^T \tag{32}$$

where *M* is the number of data points sampled from this Gaussian distribution. Then, we add noises to these two parameters to prevent privacy leakage to malicious analysts. It can be seen that the mean vector is involved in the calculation of the covariance matrix. Therefore, we only add noise to the calculation process of the mean vector, and then use noisy mean vector to calculate the covariance matrix.

There are some mechanism to achieve differential privacy, such as the Laplace mechanism [40] and the Exponential mechanism. We choose Laplace mechanism to add noise. Laplace mechanism achieve differential privacy by adding a random noise generated from Laplace distribution to the result of *g* on the dataset *D*, and it can be written as:

$$A_g(D) = g(D) + Lap(GS_g/\varepsilon) \tag{33}$$

where

$$\Pr[Lap(\beta) = x] = \frac{1}{2\beta}e^{-\frac{|x|}{\beta}} \tag{34}$$

Authorized licensed use limited to: ULAKBIM UASL - SELCUK UNIVERSITESI. Downloaded on December 07,2023 at 07:03:16 UTC from IEEE Xplore. Restrictions apply.

GS_g is the global sensitivity of g, which is defined as follows:

$$GS_g = \max\|g(D) - g(D')\|_1 \tag{35}$$

where *D* and *D*´are neighboring datasets.

In the process of calculating the mean vector of cluster *k*, we need to count the number of data points belonging to this cluster, c_k, and the sum of each dimension *j*, s_{kj}. Then, the *j*th dimension of mean vector $\vec{\mu}_k$ is calculated by $\mu_{kj} = s_{kj}/c_k$. We will add noise to c_k and s_{kj} respectively to obtain c'_k and s'_{kj}. The noisy mean vector is obtained by $\mu'_{kj} = s'_{kj}/c'_k$.

Two important tasks of adding noise is to derive the global sensitivity of the target function and allocate the privacy budget. Next, we will derive the global sensitivity of the mean vector, which determines the distribution of noises added. Adding or deleting a point in dataset *D* will only affect one cluster of the final result. So, the maximum change of the number of data points c_k is 1, i.e., $GS(c_k) = 1$. We normalize the dataset *D* to $[0,1]^d$ in advance. Thus, the maximum change of each dimension s_{kj}, caused by adding or deleting a point, is 1, i.e., $GS(s_{kj}) = 1$.

According to the Laplace mechanism, to achieve differential privacy, we should add noise for each mean vector $\vec{\mu}_k$. Specifically, add noise $Lap(1/\varepsilon_0)$ to c_k and add noise $Lap(1/\varepsilon_j)$ to s_{kj}, where

$$\sum_{j=1}^{d} \varepsilon_{kj} + \varepsilon_{k0} = \varepsilon^k \tag{36}$$

ε_{k0} and ε_{kj} are the privacy budgets allocated to c_k and s_{kj}, and ε^k is the privacy budget allocated to $\vec{\mu}_k$. Since the range of data in *D* is [0,1], ε_{k0} and ε_{kj} satisfy $\varepsilon_{k0} : \varepsilon_{kj} = 1:1$. So, considering (36), we have

$$\varepsilon_{kj} = \varepsilon_{k0} = \varepsilon^k/(d+1) \tag{37}$$

The noise added to c_k and s_{kj} is $Lap((d+1)/\varepsilon^k)$. Considering that the clustering result of *D* is equal to dividing this dataset to *K* disjoint subsets, we just need to make sure that every calculating process of $\vec{\mu}_k$ satisfies ε -differential privacy, then our algorithm IDPC satisfies ε -differential privacy. Therefore, we can obtain that $\varepsilon^1 = \varepsilon^2 = \cdots = \varepsilon^K = \varepsilon$. Fig. 4 shows the main idea of ensuaring our algorithm differentially private and the mechanism we design.

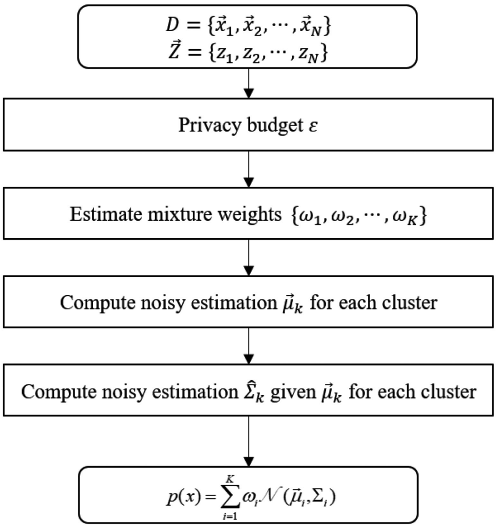

$D = \{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_N\}$
$\vec{Z} = \{z_1, z_2, \cdots, z_N\}$
Privacy budget ε
Estimate mixture weights $\{\omega_1, \omega_2, \cdots, \omega_K\}$
Compute noisy estimation $\vec{\mu}_k$ for each cluster
Compute noisy estimation $\hat{\Sigma}_k$ given $\vec{\mu}_k$ for each cluster
$p(x) = \sum_{i=1}^{K} \omega_i \mathcal{N}(\vec{\mu}_i, \Sigma_i)$

Fig. 4. Graphical model representation of adding noises.

VI. Security Analysis

In this section, we will present the security analysis of our proposed algorithm and demonstrate that this algorithm satisfies ε -differential privacy. To analyze the security of IDPC theoretically, we first give two characteristics of differential privacy, including the sequential composition and the parallel composition. We design our mechanism of privacy budget allocation according to these two properties. The sequential composition property can be interpreted as that *n* random algorithms are sequentially applied to the dataset *D*. Let $M_1, M_2, \ldots, M_n$ be *n* random algorithms, and M_i satisfies ε_i -differential privacy. Then, for dataset *D*, the composition algorithm $\{t_1 = M_1(D), t_2 = M_2(D, t_1), \ldots, t_n = M_n(D, t_1, \ldots, t_{n-1})\}$ satisfies ε -differential privacy, in which $\varepsilon = \sum_{i=1}^{n} \varepsilon_i$. The parallel composition property can be interpreted as that a random algorithm is applied to the subsect of dataset *D* respectively. Suppose a dataset *D* is divided into disjoint subsets $\{D_1, D_2, \ldots, D_n\}$ and a random algorithm *A* satisfies ε -differential privacy. Then the parallel operations of *M* on $\{D_1, D_2, \ldots, D_n\}$ satisfy ε -differential privacy.

We have discussed how to make nonparametric Bayesian clustering algorithm differentially private in subsection 5.4, and differential privacy is achieved by adding Laplace noises to the mean vectors of each Gaussian distribution in our proposed algorithm. The clustering result of *D* is equal to dividing this dataset into *K* disjoint subsets. So according to the parallel combination characteristic, if we let $\varepsilon^1 = \varepsilon^2 = \cdots = \varepsilon^K = \varepsilon$, then our proposed algorithm will satisfy ε -differential privacy. In the process of calculating mean vector $\vec{\mu}_k$, *d+1* noises will be added to c_k and s_{kj}. According to the sequence combination characteristic, we just need to make the total privacy budget allocated to cluster *k* equal to ε^k. Since the global sensitivity of c_k and s_{kj} are both 1 under our assumption that the dataset *D* is normalized to $[0,1]^d$, the noise added to c_k and s_{kj} are $Lap(1/\varepsilon_0)$ and $Lap(1/\varepsilon_j)$. We can easily obtain that if each dimension is normalized to [0, 1], the privacy budgets allocated to c_k and s_{kj} satisfy $\varepsilon_{k0} : \varepsilon_{kj} = 1:1$. Considering that

$$\sum_{j=1}^{d} \varepsilon_{kj} + \varepsilon_{k0} = \varepsilon^k \tag{38}$$

we have $\varepsilon_{kj} = \varepsilon_{k0} = \varepsilon^k/d + 1$.

VII. Performance Evaluation

To accomplish privacy-preserving cluster analysis in smart grid, we presented an IGMM-based differentially

Authorized licensed use limited to: ULAKBIM UASL - SELCUK UNIVERSITESI. Downloaded on December 07,2023 at 07:03:16 UTC from IEEE Xplore. Restrictions apply.

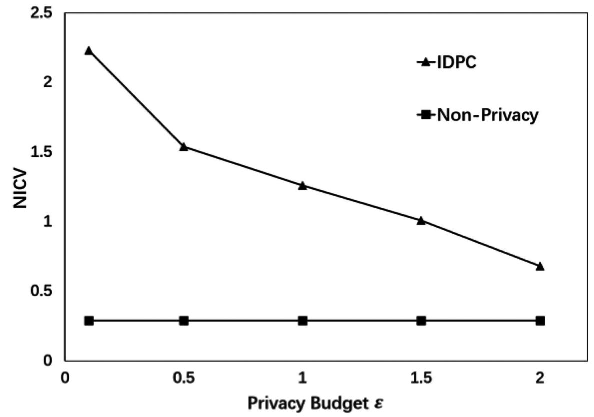


Fig. 5. NICV of our proposed IDPC and Non-privacy on dataset Blood with different privacy budgets.

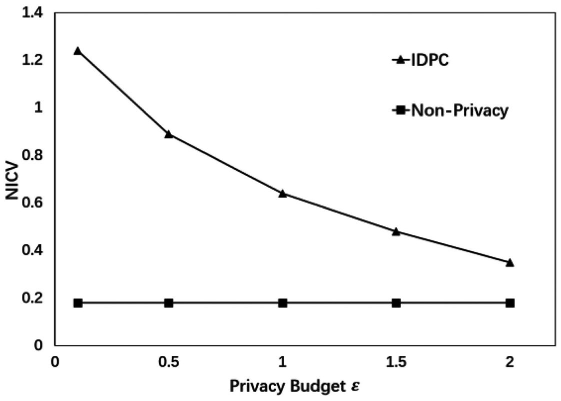


Fig. 6. NICV of our proposed IDPC and Non-privacy on dataset Adult with different privacy budgets.

private clustering (IDPC) algorithm in this paper, which takes a combination of the nonparametric Bayesian method and differential privacy. We use the nonparametric Bayesian method to perform clustering algorithm on dataset and then apply differential privacy to the released clustering results so that an adversary cannot obtain individual private information by analyzing clustering results. Specifically, privacy preservation is achieved by adding noises to the parameters of the released distributions. In this section, we implement numerical experiments to illustrate the efficiency of our algorithm by comparing the Normalized Intra-Cluster Variance (NICV) of IDPC and non-privacy clustering algorithm.

A. Description of Datasets

We did the numerical experiment on two datasets, including Blood and Adult, from the UCI Knowledge Discovery Archive database. There are 748 data points in dataset Blood and we choose 4 attributes as our experiment dataset. The dataset Adult consists of private information of individuals and has 48842 data points. We choose 5 continuous attributes as our experiment dataset.

Next, we discuss how we can set the hyperparameters in the nonparametric Bayesian clustering algorithm, including the concentration parameter α, and $\vec{\mu}_0$, κ_0, Λ_0^{-1}, υ_0. α stands for prior belief on the distribution. We can easily find that the probability of assigning a data point to an unpresented cluster depends on the value of α and the ratio of α and N. Therefore, we set $\alpha = [N/20]$ so that the probability of crea'ting a new cluster will not be too high or too small. $\vec{\mu}_0$ represents the initial mean vector, and κ_0 is associated with the dispersion of the clusters and our confidence of $\vec{\mu}_0$. Generally, we set $\vec{\mu}_0$ to be the mean vector of all data points and choose κ_0 to be 0.5. Another two hyperparameters, Λ_0^{-1}, υ_0, are associated with the covariance matrix, which are the parameters of Wishart distribution. υ_0 represents our confidence of Λ_0, which depends on the difference between Λ_0 and the updated covariance matrix. We set Λ_0 to be a diagonal matrix of 0.1 and υ_0 to be 10.

B. Accuracy

In this subsection, we evaluate our proposed IDPC by comparing the NICV of IDPC with the general nonparametric Bayesian clustering algorithm without differential privacy. We set different total privacy budgets in IDPC to illustrate the influence of the level of privacy guarantee on accuracy of clustering results. Figs. 5 and 6 are the analysis results on the two datasets. It can be seen from the experimental results that our proposed algorithm achieves privacy preservation while having an acceptable impact on the utility. As for the relationship between accuracy of proposed algorithm and privacy budgets, the accuracy of our algorithm approaches the general nonparametric Bayesian clustering algorithm as the privacy budget increases.

VIII. Conclusion

To accomplish privacy-preserving cluster analysis in smart grid, we presented an IGMM-based differentially private clustering (IDPC) algorithm in this paper, which takes a combination of the nonparametric Bayesian method and differential privacy. In IDPC, the nonparametric Bayesian method is applied in the clustering algorithm, to allow certain parameters to change with the. The Laplace mechanism is used in the data releasing process to make our proposed algorithm differentially private. We presented how to make the nonparametric Bayesian clustering algorithm differentially private by adding noises. Finally, we theoretically analyze the security of proposed algorithm and prove the efficiency through numerical experiments on two datasets. The experimental results demonstrate that our proposed algorithm can achieve a balance between the privacy and utility.

In future research, we will work on the privacy leakage issues during the process of the nonparametric Bayesian clustering and how to combine nonparametric Bayesian method with differential privacy to achieve more optimized nonparametric Bayesian clustering with both high accuracy and privacy preserving. In addition, we will continue to study the combination of other machine learning and deep learning algorithms with differential privacy and optimization of achieving tradeoff between privacy and utility.

Authorized licensed use limited to: ULAKBIM UASL - SELCUK UNIVERSITESI. Downloaded on December 07,2023 at 07:03:16 UTC from IEEE Xplore. Restrictions apply.

References

[1] D. Alahakoon and X. Yu, “Smart electricity meter data intelligence for future energy systems: A survey,” *IEEE Trans. Ind. Informat.*, vol. 12, no. 1, pp. 425–436, Feb. 2016.

[2] X. Du, M. Zhang, K. Nygard, S. Guizani, and H. H Chen, “Self-healing sensor networks with distributed decision making,” *Int. J. Sensor Netw.*, vol. 2, no. 5/6, pp. 289–298, 2007.

[3] H. Farhangi, “The path of the smart grid,” *IEEE Power Energy Mag.*, vol. 8, no. 1, pp. 18–28, Jan./Feb. 2010.

[4] T. Taleb, Y. Hadjadj-Aoul, and K. Samdanis, “Efficient solutions for enhancing data traffic management in 3GPP networks,” *IEEE Syst. J.*, vol. 9, no. 2, pp. 519–528, Jun. 2015.

[5] T. Taleb, I. Afolabi, and M. Bagaa, “Orchestrating 5G network slices to support industrial internet and to shape next-generation smart factories,” *IEEE Netw.*, vol. 33, no. 4, pp. 146–154, Jul. 2019.

[6] P. McDaniel and S. McLaughlin, “Security and privacy challenges in the smart grid,” *IEEE Secur. Privacy*, vol. 7, no. 3, pp. 75–77, May/Jun. 2009.

[7] X. Du, Y. Xiao, S. Ci, M. Guizani, and H. H. Chen, “A routing-driven key management scheme for heterogeneous sensor networks,” in *Proc. IEEE Int. Conf. Commun.*, Glasgow, Scotland, Jun. 2007, pp. 3407–3412.

[8] X. Du, M. Guizani, Y. Xiao, and H. H. Chen, “Defending DoS attacks on broadcast authentication in wireless sensor networks,” in *Proc. IEEE Int. Conf. Commun.*, Beijing, China, May 2008, pp. 1653–1657.

[9] C. Dwork, “Differential privacy: A survey of results,” in *Proc. Int. Conf. Theory Appl. MODELS Comput.*, 2008, pp. 1–19.

[10] H. Shin, S. Kim, J. Shin, and X. Xiao, “Privacy enhanced matrix factorization for recommendation with local differential privacy,” *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 9, pp. 1770–1782, Sep. 2018.

[11] X. Ren *et al.*, “Lopub: High-dimensional crowdsourced data publication with local differential privacy,” *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 9, pp. 2151–2166, Sep. 2018.

[12] H. Wallach, S. Jensen, L. Dicker, and K. Heller, “An alternative prior process for nonparametric bayesian clustering,” *J. Mach. Learn. Res.*, vol. 9, pp. 892–899, 2008.

[13] P. Wang, K. Laskey, C. Domeniconi, and M. Jordan, “Nonparametric bayesian co-clustering ensembles,” in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Spain, Barcelona, 2010, pp. 435–440.

[14] B. Ahmad, A. Vajda, and T. Taleb, “Impact of network function virtualization: A study based on real-life mobile network data,” in *Proc. IEEE Int. Wireless Commun. Mobile Comput. Conf.*, Paphos, Cyprus, Sep. 2016, pp. 541–546.

[15] P. Diamantoulakis, V. Kapinas, and G. Karagiannidis, “Big data analytics for dynamic energy management in smart grids,” *Big Data Res.*, vol. 2, no. 3, pp. 94–101, 2015.

[16] W. Yuan, P. Deng, T. Taleb, J. Wang, and C. Bi, “An unlicensed taxi identification model based on big data analysis,” *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 6, pp. 1703–1713, Jun. 2016.

[17] A. Shirkhorshidi, S. Aghabozorgi, T. Wah, and T. Herawan, “Big data clustering: A review,” in *Proc. Int. Conf. Comput. Science Appl.*, 2014, pp. 707–720.

[18] T. Taleb, D. E. Bensalem, and A. Laghrissi, “Smart service-oriented clustering for dynamic slice configuration,” in *Proc. IEEE Global Commun. Conf.*, Waikoloa, HI, USA, Dec. 2019, pp. 1–6.

[19] 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.

[20] R. Sánchez-García *et al.*, “Hierarchical spectral clustering of power grids,” *IEEE Trans. Power Syst.*, vol. 29, no. 5, pp. 2229–2237, Sep. 2014.

[21] E. Ortjohann, P. Wirasanti, M. Lingemann, W. Sinsukthavorn, S. Jaloudi, and D. Morton, “Multi-level hierarchical control strategy for smart grid using clustering concept,” in *Proc. Int. Conf. Clean Elect. Power*, 2011, pp. 648–653.

[22] Y. Liu, W. Guo, C. Fan, L. Chang, and C. Cheng, “A practical privacy-preserving data aggregation (3PDA) scheme for smart grid,” *IEEE Trans. Ind. Informat.*, vol. 15, no. 3, pp. 1767–1774, Mar. 2019.

[23] 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, 2019, Art. no. 032103.

[24] D. He, N. Kumar, S. Zeadally, A. Vinel, and L. Yang, “Efficient and privacy-preserving data aggregation scheme for smart grid against internal adversaries,” *IEEE Trans. Smart Grid*, vol. 8, no. 5, pp. 2411–2419, Sep. 2017.

[25] 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.

[26] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “l-diversity: Privacy beyond k-anonymity,” in *Proc. IEEE Int. Conf. Data Eng.*, 2006, p. 24.

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

[28] D. Su *et al.*, “Differentially private K-means clustering,” in *Proc. ACM Conf. Data Appl. Secur. Privacy*, 2016, pp. 26–37.

[29] N Hiep, “Privacy-preserving mechanisms for k-modes clustering,” *Comput. Secur.*, vol. 78, pp. 60–75, 2018.

[30] X. Du, M. Rozenblit, and M. Shayman, “Implementation and performance analysis of SNMP on a TLS/TCP base,” in *Proc. 7th IFIP/IEEE Int. Symp. Integr. Netw. Manage.*, Seattle, WA, USA, May 2001, pp. 453–466.

[31] X. Huang and X. Du, “Achieving big data privacy via hybrid cloud,” in *Proc. IEEE INFOCOM Workshops*, Toronto, ON, Canada, Apr. 2014, pp. 512–517.

[32] N. Nguyen, R. Zheng, and Z. Han, “On identifying primary user emulation attacks in cognitive radio systems using nonparametric bayesian classification,” *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1432–1445, Mar. 2012.

[33] P. Varela, J. Hong, T. Ohtsuki, and X. Qin, “IGMM-based co-localization of mobile users with ambient radio signals,” *IEEE Internet Things J.*, vol. 4, no. 2, pp. 308–319, Apr. 2017.

[34] Y. The, “Dirichlet process,” in *Proc. Encyclopedia Mach. Learn.*, 2010, pp. 280–287.

[35] Y. The, “Dirichlet processes: Tutorial and practical course,” Gatsby Computational Neuroscience Unit, University College London, 2007.

[36] C. Rasmussen, “The Infinite gaussian mixture model,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Apr. 2000, pp. 554–560.

[37] F. Wood, and M. Black, “A nonparametric bayesian alternative to spike sorting,” *J. Neuroscience Methods*, vol. 173, no. 1, pp. 1–12, Aug. 15, 2008.

[38] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*, 2nd ed. London, U.K.: Chapman & Hall/CRC, 2003.

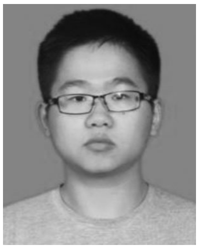
[39] P. Resnik, and E. Hardisty, “Gibbs sampling for the uninitiated,” UMIACS, College Park, MD, Tech. Rep., 2009.

[40] C. Dwork, F. Mcsherry, and K. Nissim, “Calibrating noise to sensitivity in private data analysis,” in *Proc. Conf. Theory Cryptography*, 2006, pp. 265–284.

Zhitao Guan (Member, IEEE) received the B.Eng. and Ph.D. degrees in computer application from the Beijing Institute of Technology, Beijing, 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, Beijing, China. He has authored more than 70 peer-reviewed journal and conference papers in these areas. His current research focuses on smart grid security, secure machine learning, and data privacy.

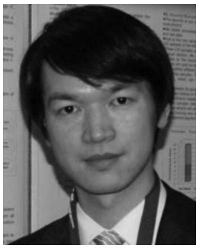
Zefang Lv received the B.Eng. degree from Shandong University, Jinan, China, in 2016. She is currently working toward the master degree with the School of Mathematics and Physics, North China Electric Power University, Beijing, China. Her current research focuses on secure machine learning and data privacy.

Authorized licensed use limited to: ULAKBIM UASL - SELCUK UNIVERSITESI. Downloaded on December 07,2023 at 07:03:16 UTC from IEEE Xplore. Restrictions apply.



Xianwen Sun received the B.Eng. degree from North China Electric Power University, Beijing, China, in 2018. He is currently working toward the master degree with the School of Control and Computer Engineering, North China Electric Power University. His current research focuses on secure machine learning and data privacy.

Longfei Wu (Member, IEEE) received the B.E. degree in telecommunication engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in July 2012. He received the Ph.D. degree in computer and information sciences from Temple University, Philadelphia, PA, USA, in July 2017. He is currently an Assistant Professor with the Department of Mathematics and Computer Science, Fayetteville State University, Fayetteville, NC, USA. His research interests are the security and privacy of networked systems and modern computing devices, including mobile devices, Internet-of-Things, implantable medical devices, and wireless networks.



Jun Wu (Member, IEEE) received the Ph.D. degree in information and telecommunication studies from Waseda University, Tokyo, Japan. He is an Associate Professor of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. He was a Postdoctoral Researcher for the Research Institute for Secure Systems, National Institute of Advanced Industrial Science and Technology, Japan, from 2011 to 2012. He worked as a Researcher for the Global Information and Telecommunication Institute, Waseda University, Japan, from 2011 to 2013. 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 been a Guest Editor for the IEEE Sensors Journal and a TPC Member of several international conferences.

Xiaojiang Du (Fellow, IEEE) received the B.S. and M.S. degrees in electrical engineering (Automation Department) from Tsinghua University, Beijing, China, in 1996 and 1998, respectively. He received the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, MD, USA, in 2002 and 2003, respectively. He is a tenured Full Professor and the Director of the Security and Networking (SAN) Lab in the Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. He has authored more than 400 journal and conference papers in these areas, as well as a book published by Springer. His research interests are security, wireless networks, and systems. He won the Best Paper Award at IEEE GLOBECOM 2014 and the Best Poster Runner-up Award at the ACM MobiHoc 2014. He serves on the editorial boards of three international journals. He was the lead Chair of the Communication and Information Security Symposium of the IEEE International Communication Conference 2015, and a Co-Chair of Mobile and Wireless Networks Track of IEEE Wireless Communications and Networking Conference 2015. He was a Technical Program Committee Member of several premier ACM/IEEE conferences. He is a Life Member of ACM.

Mohsen Guizani (Fellow, IEEE) received the B.S. (with distinction) and M.S. degrees in electrical engineering, the M.S. and Ph.D. degrees in computer engineering from Syracuse University, Syracuse, NY, USA, in 1984, 1986, 1987, and 1990, respectively. He is currently a Professor at the Computer Science and Engineering Department in Qatar University, Qatar. Previously, he served in different academic and administrative positions at the University of Idaho, Western Michigan University, University of West Florida, University of Missouri-Kansas City, University of Colorado-Boulder, and Syracuse University. His research interests include wireless communications and mobile computing, computer networks, mobile cloud computing, security, and smart grid. He is currently the Editor-in-Chief of the IEEE Network Magazine, serves on the editorial boards of several international technical journals and the Founder and Editor-in-Chief of Wireless Communications and Mobile Computing journal (Wiley). He is the author of nine books and more than 600 publications in refereed journals and conferences. He guest edited a number of special issues in IEEE journals and magazines. He also served as a member, Chair, and General Chair of a number of international conferences. Throughout his career, he received three teaching awards and four research awards. He also received the 2017 IEEE Communications Society WTC Recognition Award as well as the 2018 AdHoc Technical Committee Recognition Award for his contribution to outstanding research in wireless communications and Ad-Hoc Sensor networks. He was the Chair of the IEEE Communications Society Wireless Technical Committee and the Chair of the TAOS Technical Committee. He served as the IEEE Computer Society Distinguished Speaker and is currently the IEEE ComSoc Distinguished Lecturer. He is a Senior Member of ACM.

Authorized licensed use limited to: ULAKBIM UASL - SELCUK UNIVERSITESI. Downloaded on December 07,2023 at 07:03:16 UTC from IEEE Xplore. Restrictions apply.