Wireless Big Data for Differential Seclusion Preserving

K. C. Rajavenkatesswaran
(AP/IT)
Department of Information Technology
Nandha college of Technology
Erode, India

S. Dharini, P. Ranjitha, E. Vignesh, S. Vigneshkumar
(UG Scholar)
Department of Information Technology
Nandha college of Technology
Erode, India

***Abstract*— With the popularity of smart devices and the widespread use of machine learning methods, smart edges have become the mainstream of dealing with wireless big data. When smart edges use machine learning models some models may unintentionally store a small portion of the training data with sensitive records. To solve this privacy issue, in this paper, we proposed and implemented a machine learning strategy for smart edges using differential privacy. Our attention has been focused on privacy protection in training datasets in wireless big data scenario. Privacy protection adds Laplace mechanisms, and designed three different algorithms which satisfied differential privacy. This project introduces a privacy preserving approach that can be applied to decision tree learning, without connected loss of accuracy. It described an approach to the preservation that privacy of collected data samples in cases where information from the sample database has been partially lost. This approach converts the original sample data sets into a group of Non-Sensitive data sets, from which the original samples cannot be reconstructed without the entire group of unreal data sets. Meanwhile, an accurate analysis can be built directly from those unreal data sets. This can be applied directly to the data storage as soon as the first sample is collected. The Relevant Columns Values Swapping approach is compatible with other privacy preserving approaches, such as without cryptography, for extra protection.**

***Keywords:* Wireless Big Data, Smart Edges, Differential Privacy, Training Data Privacy, Machine Learning, Correlated Datasets, Laplacian Mechanism, TensorFlow**

I. INTRODUCTION

The smart edges have received extensive data processing, data analysis and data storage in wireless big data scenario . Smart edges bring enormous benefits in the aspect of analyzing and mining data, perceiving location information, such as localization and low latency . Some wireless big data, e.g., broadband download, online business, health sensing, etc, contains a lot of analysis and mining of effective information. However, it is inevitable to involve some privacy records when the smart edges use machine learning methods for data process- ing and prediction, such as license plates, tax information, and personal assets information. In recent years, it is often seen that hackers have exploited privacy holes in machine learning, and restored private sensitive training data from the model. Fredrikson et al. used the privacy leak of the computer vision classifier to expose the personal picture information from the training data. Thus, the problem of data privacy in machine learning is becoming more and more serious, especially for the privacy protection of training datasets. Once the data with sensitive information is maliciously attacked, it is most likely to be exploited by criminals. To guarantee the privacy protection of the training datasets, this paper first analyzes the possible privacy issues faced by the training datasets for smart edges in wireless big data scenario. There are a large number of personal edge nodes and business edge nodes in the smart edges, as shown in Fig. 1, and they own a certain ability to compute and process data. These nodes can make use of the edge cloud to make the network cooperate with the terminals, which can achieve the business localization processing, the service delay reduction, and the network efficiency promotion. As the most popular methods of data analysis, machine learning can efficiently analyze and mine valuable information hidden behind the wireless big data.

From the figure, we can see that smart edge nodes can analyze a variety of different wireless big data by using machine learning methods. For example, the analysts establish a machine learning model to analyze and predict the results of medical diagnosis, which is more conductive to the prediction and resolution of problems. Analysts can find out the common features of cancer patients by sorting out massive training datasets, thereby providing better help in diagnosing cancer. On the other hand, however, intruders may invade certain sensitive data individuals in datasets to achieve their ulterior motives. That is exactly the great challenge in privacy preserving at present: how to ensure that analysts are not likely to cause sensitive data leakage when analyzing data in machine learning model? Fortunately, differential privacy algorithm is a promising technology solution that can alleviate this tension. This approach allows analysts to perform benign aggregation.

II. LITERATURE SURVEY

(i) Embracing Big Data with Compressive Sensing: A Green Approach In Industrial Wireless Networks

Linghe Kong and Daqiang Zhang describe a new-generation industries heavily rely on big data to improve their efficiency. Such big data are commonly collected by smart nodes and transmitted to the cloud via wireless. Due to the limited size of smart node, the shortage of energy is always a critical issue, and the wireless data transmission is extremely a big power consumer. Aiming to reduce the energy consumption in wireless, this article introduces a potential breach from data redundancy. If redundant data are no longer collected, a large amount of wireless transmissions can be cancelled and their energy saved. Motivated by this breach, this article proposes a compressive-sensing-based collection framework to minimize the amount of collection while guaranteeing data quality. This framework is verified by experiments and extensive real-trace-driven simulations. In a big data collection system, smart nodes are usually distributed in the given area to sense data and transmit these data to the cloud via wireless communications. The cloud analyzes the collected data and provides customized service or production. Suppose there are a total of n nodes, and the period of monitoring time is evenly divided into t time slots. Every node collects data once per time slot at most. With the growth of the scale in industrial applications, the total collected data are very big. The big data can be represented as a large matrix X, where every element is the data collected by one node at one time slot. A matrix with no empty elements means that all data are collected, which indicates 100 percent data quality but costs nt wireless transmissions. On one hand, to reduce the number of transmissions, it is desired that only principal data are collected. Assume that the amount of principal data is r and r << nt. On the other hand, to guarantee the data quality, it is desired that the principal data are adequate to represent the whole big data, that is, the recovered matrix ^X is close to the complete X , where ^X is the matrix computed by compressive sensing using only principal data. From the above, we state the problem as follows: The green collection problem aims to minimize the amount of principal data r for energy saving and is constrained by ^X ≈ X for quality assurance.

Two main metrics are defined to measure the performance of green collection solutions:

1) **Energy Consumption Ratio a** : This ratio can be approximated as r / nt , that is, transmitting the amount of principal data over transmitting the total big data, in which we consider the consumption is equal for every transmission
2) **Data Error Ratio e:** The average error between recovered matrix and complete matrix, that is, e = || ^ X – X || / || X ||.

A green collection framework is proposed in this article to save energy in big-data-based smart industries. The core contribution of this framework is to reduce the number of transmissions by leveraging the compressive sensing theory. The evaluation results demonstrate that the proposed framework dramatically decreases the power consumption compared to existing approaches while the data quality is guaranteed.

(ii) Computing On Base Station Behavior Using ERLANG Measurement And Call Detail Record

Sihai Zhang and Dandan Yin et al describe the important aspects in this topic, including data set information, data analysis techniques, and two case studies. We categorize the data set in the telecommunication networks into two types, user-oriented and network-oriented, and discuss the potential application. Then, several important data analysis techniques are summarized and reviewed, from temporal and spatial analysis to data mining and statistical test.

Finally, we present two case studies, using Erlang measurement and call detail record, respectively, to understand the base station behavior. Interestingly, the night phenomenon of college students is revealed by comparing the base stations location and real-world map, and we conclude that it is not proper to model the voice call arrivals as Poisson process. Basically, in the telecommunication networks the data source can be categorized into user-oriented and network-oriented, corresponding to two fundamental components, mobile users/wireless devices who communicate with each other, and network devices which provide the wireless coverage, wireless transmission, positioning, data exchange and other functionalities. Everyone has her/his own living habits, like when to sleep, where to live, whom to play with, therefore these habits will definitely be reflected in her/his communication information, since mobile phones have become part and parcel of our life and an integral part for every individual. The practical communication activities of mobile users include voice calls, SMS, and other data traffic through all kinds of Internet Applications, coined here as user-oriented data. We note here that, despite of traditional mobile devices, Machine Type Communication (MTC) devices will also contribute even larger amount of data traffic in the coming future. Communication networks also produce giant amount of data when serving the mobile users, like spectrum measurements, device status report and etc, but this topic is beyond this paper's scope.

This paper concentrates on the data analysis issue in the telecommunication networks, covering the data set categorization, commonly used data analysis techniques and two general case studies using Erlang measurement and CDR. In addition, our work presents several interesting findings about base station behavior. To be specific, we accomplish the spatial-temporal analysis to the GSM base station traffic of a city in Southern China, and comprehensively investigate the traffic pattern and spatial correlation, which brings some insight for future possible work. K-means method is adopted to help understand different patterns of base stations and finds that, different from common people, college students' special activity pattern, coined here as 'Night Burst', has been revealed through traffic of base stations near university campuses, based on spatial-temporal analysis. In spatial correlation part, we give our thinking and possible reason through the study of local Moran's I in the spatial dimension and finds that local Moran's I can be a tool to discover 'abnormal' stations in a region.

As to the CDR data, we have analyzed the characteristics of the call arrivals based on real call detail records of large-scale GSM base stations in Beijing over 30 days using MAVAR and chi-square test. First, the preliminary observation reported in this paper shows that the call arrival patterns vary over time and the location of stations. Second, the number of call arrivals in a minute has been found uncorrelated in short-range but time-correlation exist in long-range because of the violent fluctuation of the call arrivals in the long term (24 hours). Third, the call arrivals can be modeled as Poisson process in most cases, but the characteristic of call arrivals changes over time and space, so it is improper to model the call arrivals in one hour as Poisson distribution. The proposed work is a first step for such data analysis in mobile communication networks, and our observations have potential applications, such as cellular optimization, resource planning and etc.

(iii) Model Inversion Attacks That Exploit Confidence Information And Basic Countermeasures

Matt Fredrikson and Somesh Jha et al [3] describe a machine-learning (ML) algorithms are increasingly utilized in privacy-sensitive applications such as predicting lifestyle Choices, making medical diagnoses, and facial recognition. In a model inversion attack, recently introduced in a case study of linear classifiers in personalized medicine by Fredrikson, adversarial access to an ML model is abused to learn sensitive genomic information about individuals. Whether model inversion attacks apply to settings outside theirs, however, is unknown. Author develops a new class of model inversion attack that exploits confidence values revealed along with predictions. Our new attacks are applicable in a variety of settings, and we explore two in depth: decision trees for lifestyle surveys as used on machine-learning-as-a-service systems and neural networks for facial recognition. In both cases confidence values are revealed to those with the ability to make prediction queries to models. We experimentally show attacks that are able to estimate whether a respondent in a lifestyle survey admitted to cheating on their significant other and, in the other context, show how to recover recognizable images of people's faces given only their name and access to the ML model. We also initiate experimental exploration of natural countermeasures, investigating a privacy-aware decision tree training algorithm that is a simple variant of CART learning, as well as revealing only rounded confidence values. The lesson that emerges is that one can avoid these kinds of MI attacks with negligible degradation to utility.

It may be possible in some cases to use numeric approximations for the gradient function in place of the explicit gradient computation used above. This would allow a black-box adversary for both types of attack on facial recognition models. We implemented this approach using scipy's numeric gradient approximation, and found that it worked well for Softmax models—the reconstructed images look identical to those produced using explicit gradients. Predictably, however, performance suffers. While Softmax only takes about a minute to complete on average, MLP and DAE models take significantly longer. Each numeric gradient approximation requires on the order of 2 d black-box calls to the cost function, each of which takes approximately 70 milliseconds to complete. At this rate, a single MLP or DAE experiment would take 50–80 days to complete. Finding ways to optimize the attack using approximate gradients is interesting future work.

In proposed system demonstrated how the confidence information returned by many machine learning ML classifiers enables new model inversion attacks that could lead to unexpected privacy is- sues. By evaluating our model inversion algorithms over decision trees published on a ML-as-a-service marketplace, we showed that they can be used to infer sensitive responses given by survey respondents with no false positives. Using a large-scale study on Mechanical Turk, showed they can also be used to extract images from facial recognition models that a large majority of skilled humans are able to consistently re-identify. The proposed system explored some simple approaches that can be used to build effective counter measures to our attacks, initiating an experimental evaluation of defensive strategies. Although these approaches do not constitute full-fledged private learning algorithms, they illustrate trends that can be used to guide future work towards more complete algorithms. Our future efforts will follow this path, as continue to work towards new systems that are able to benefit from advances in machine learning without introducing vulnerabilities that lead to model inversion attacks.

(iv) Clustering Of Electricity Consumption Behavior Dynamics toward Big Data Applications

Yi Wang And Qixin Chen describe a competitive retail market, large volumes of smart meter data provide opportunities for load serving entities to enhance their knowledge of customers' electricity consumption behaviors via load profiling. Instead of focusing on the shape of the load curves, this paper proposes a novel approach for clustering of electricity consumption behavior dynamics, where "dynamics" refer to transitions and relations between consumption behaviors, or rather consumption levels, in adjacent periods. First, for each individual customer, symbolic aggregate approximation is performed to reduce the scale of the data set, and time-based Markov model is applied to model the dynamic of electricity consumption, transforming the large data set of load curves to several state transition matrixes. Second, a clustering technique by fast search and find of density peaks (CFSFDP) is primarily carried out to obtain the typical dynamics of consumption behavior, with the difference between any two consumption patterns measured by the Kullback–Liebler distance, and to classify the customers into several clusters. To tackle the challenges of big data, the CFSFDP technique is integrated into a divide-and- conquers approach toward big data applications. A numerical case verifies the effectiveness of the proposed models and approaches.

Countries around the world have set aggressive goals for the restructuring of monopolistic power system towards liberalized markets especially on the demand side. In a competitive retail market, load serving entities (LSEs) will be developed in great numbers. Having a better understanding of electricity consumption patterns and realizing personalized power managements are effective ways to enhance the competitiveness of LSEs. Meanwhile, smart grids have been revolutionizing the electrical generation and consumption through a two-way flow of power and information. As an important information source from the demand side, advanced metering infrastructure (AMI), has gained increasing popularity worldwide; AMI allows LSEs to obtain electricity consumption data at high frequency, e.g., minutes to hours. Large volumes of electricity consumption data reveal information of customers that can potentially be used by LSEs to manage their generation and demand resources efficiently and provide personalized service. In this paper, a novel approach for the clustering of electricity consumption behavior dynamics toward large data sets has been proposed. Different from traditional load profiling from a static prospective, SAX and time-based Markov model are utilized to model the electricity consumption dynamic characteristics of each customer. A density-based clustering technique, CFSFDP, is performed to discover the typical dynamics of electricity consumption and segment customers into different groups. Finally, a time domain analysis and entropy evaluation is conducted on the result of the dynamic clustering to identify the demand response potential of each group's customers. The challenges of massive high-dimensional electricity consumption data are addressed in three ways. First, SAX can reduce and discredited the numerical consumption data to ease the cost of data communication and storage. Second, Markov model are modeled to transform long-term data to several transition matrixes. Third, a distributed clustering algorithm is then proposed for distributed big data sets. Limited by the data sets, the influence of external factors like temperature, day type, and economy on the electricity consumption is not considered in depth in this paper. Future works will focus on feature extraction and data mining techniques combining electricity consumption with external factors.

(v) Optimal Noise Adding Mechanisms for Approximate Differential Privacy

Quan Geng and Pramod Viswanath study the (nearly) optimal mechanisms in (E, δ) -differential privacy for integer-valued query functions and vector-valued (histogram-like) query functions under a utility-maximization/cost-minimization framework. Within the classes of mechanisms oblivious of the database and the queries beyond the global sensitivity, we characterize the tradeoff between and δ in utility and privacy analysis for histogram-like query functions, and show that the (, δ) -differential privacy is a framework not much more general than the (, 0) -differential privacy and (0 ,δ) -differential privacy in the context of 1 and 2 cost functions, i.e., minimum expected noise magnitude and noise power. In the same context of 1 and 2 cost functions, we show the near-optimality of uniform noise mechanism and discrete Laplacian mechanism in the high privacy regime (as (, δ) → (0 , 0)). We conclude that in (, δ) -differential privacy, the optimal noise magnitude and the noise power are (min ((1 /), (1 /δ))) and (min ((1 / 2), (1 /δ 2))) , respectively, in the high privacy regime. Differential privacy is a framework to quantify to what extent individual privacy in a statistical database is preserved while releasing useful statistical information about the database.

The basic idea of differential privacy is that the presence of any individual data in the database should not affect the final released statistical information significantly, and thus it can give strong privacy guarantees against an adversary with arbitrary auxiliary information. For more background and motivation of differential privacy, we refer the readers to the survey. The standard approach to preserve differential privacy for real-valued query function is to perturb the query output by adding random noise with Laplacian distribution. Recently, Geng and Viswanath show that under a general utility-maximization framework, for single real-valued query unction, the optimal -differentially private mechanism is the staircase mechanism, which adds noise with staircase distribution to the query output. The optimality of the staircase mechanism is extended to the multidimensional setting for histogram-like functions, where the sensitivity of the query functions is defined using the 1 metric. A relaxed notion of privacy, (E, δ) -differential privacy, was introduced by Dwork, and the standard approach to preserving (E, δ) -differential privacy is to add Gaussian noise to the query output. In this work, we study the (nearly) optimal mechanisms in (E, δ) -differential privacy for integer-valued query functions and vector-valued (histogram-like) query functions under a utility-maximization/cost-minimization framework, and characterize the tradeoff between and δ in utility and privacy analysis. Optimality in this work is defined with respect to the class of the mechanisms which are oblivious of the database and the properties of query functions except the global sensitivity. We refer the readers to the end of Section I.A for more details about the setting considered in this work.

III EXISTING SYSTEM

The existing system is a perturbation and randomization-based approach that protects centralized sample data sets utilized for decision tree data mining. Privacy preservation is applied to sanitize the samples prior to their release to third parties in order to mitigate the threat of their inadvertent disclosure or theft. In contrast to other sanitization methods, the approach does not affect the accuracy of data mining results. The decision tree can be built directly from the sanitized data sets, such that the originals do not need to be reconstructed.

Moreover, this approach can be applied at any time during the data collection process so that privacy protection can be in effect even while samples are still being collected. But the following assumptions are made in the existing system: first, as is the norm in data collection processes, a sufficiently large number of sample data sets have been collected to achieve significant data mining results covering the whole research target. Second, the number of data sets leaked to potential attackers constitutes a small portion of the entire sample database. Third, identity attributes (e.g., social insurance number) are not considered for the data mining process because such attributes are not meaningful for decision making. Fourth, all data collected are discretized; continuous values can be represented via ranged-value attributes for decision tree data mining.

(i) Drawbacks of Existing System

The existing system has following disadvantages,

- If the unrealized data sets with varying attributes are given to multiple parties then privacy preserving could not be achieved.
- Privacy preservation via data set complementation fails if all training data sets are leaked especially when given to multiple parties.
- Data may be leaked or stolen anytime during the storing process.

- Third parties may inadvertently disclose samples to malicious parties.
- Original samples cannot be reconstructed without the entire group of unreal data sets.
- Decision tree based data preserving is not possible in current technology.
- Usage of data set gives, efficient and real time results here are complicated.
- Multiparty privacy-preserving shared mining is very tedious.
- Privacy is not preserved in transferred dataset and original dataset.
- Privacy preserving mining benefits to multiple information providers are not flexible.

IV PROBLEM DEFINITION

A large body of research has been devoted to the protection of sensitive information when samples are given to third parties for processing or computing. It is in the curiosity of research to disseminate samples to a large audience of researchers, without making strong assumptions concern their trustworthiness. Even if information collectors ensure that information are released only to third parties with non-malicious intent (or if a privacy preserving approach can be applied before the data are released), there is always a chance that the information collectors may inadvertently disclose samples to malicious parties or that the samples are actively stolen from the collectors.

Samples may be leaked or stolen anytime during the storing process or while residing in storage. This paper focuses on anticipating such attacks from third parties for the whole lifetime of the samples. Contemporary research in privacy preserving data mining mainly falls into one of two categories: perturbation and randomization-based approaches, and secure multiparty computation (SMC)-based approaches. SMC approaches use cryptographic tools for collaborative data mining computation by multiple parties. Samples are distributed among completely different parties and they take part in the information computation and communication process. SMC analysis focuses on protocol development for protecting privacy among the involved parties or computation efficiency; however, centralized processing of samples and storage privacy is out of the scope of SMC.

V PROPOSED SYSTEM

The proposed system is a perturbation and randomization-based approach that protects centralized sample data sets utilized for decision tree data mining. Privacy preservation is applied to sanitize the samples prior to their release to third parties in order to mitigate the threat of their inadvertent disclosure or theft. In contrast to other sanitization methods, the approach does not affect the accuracy of data mining results. The decision tree can be built directly from the sanitized data sets, such that the originals do not need to be reconstructed. Moreover, this approach can be applied at any time during the data collection process so that privacy protection can be in effect even while samples are still being collected. But the following assumptions are made in the existing system: first, as is the norm in data collection processes, a sufficiently large number of sample data sets has been collected to achieve significant data mining results covering the whole research target. Second, the number of data sets leaked to potential attackers constitutes a small portion of the entire sample database. Third, identity attributes (e.g., social insurance number) are not considered for the data mining process because such attributes are not meaningful for decision making. Fourth, all data collected are discretized; continuous values can be represented via ranged-value attributes for decision tree data mining. In addition with all existing system approaches, the geometric perturbation is also implemented in the proposed system. In the proposed system, the privacy is preserved even if the data is spread across multi parties. Suppose a bank issues the account holders some of the attributes to more than one insurance agency. Then from the attributes of the table along with the records given to one insurance agency, other agency could not guess or identify the facts regarding the account holders. Likewise, if two agencies give their data set (retrieved from the bank) to other parties, they must not identify the facts by combing both data sets.

(i) Advantages of Proposed System

- From the two given data sets, the original facts cannot be guessed.
- Privacy is preserved even if the data is spread across multi parties
- Consider multiple service providers collaboratively providing the privacy preserving mining service to multiple data providers.
- Multiparty privacy-preserving shared mining is not tedious.
- Decision tree based data preserving is possible in current technology

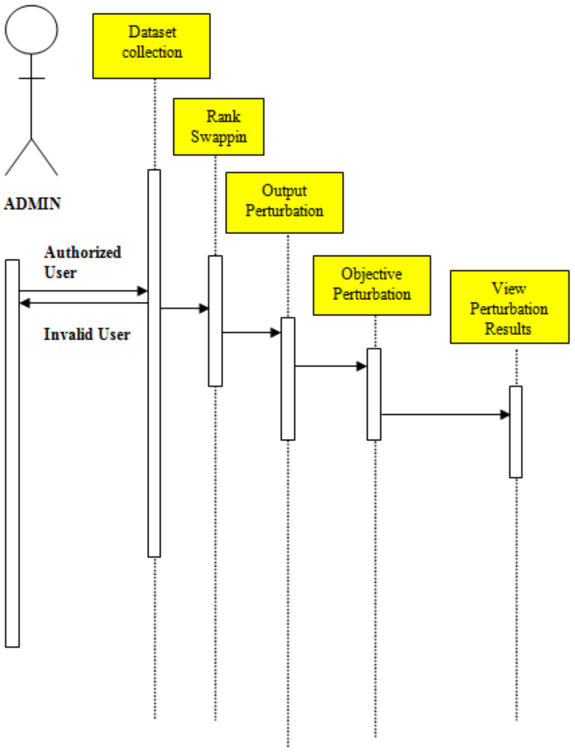
VI RELATED WORK

(i) Rank Swapping

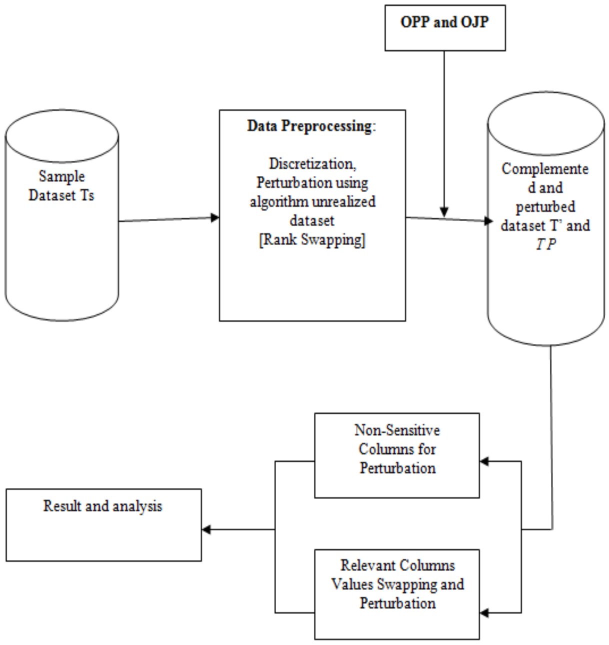
In this module, from the dataset, all the records are taken. Then all the columns are sorted such as second column values are sorted, then third column values and so on up to last column. Then each pair of rows is taken first. Then from second column to last column, all the column values are interchanged. If the row count is odd, then previous row of last row and last row values are interchanged. This perturbation is made to alter the values with no loss in data, i.e., no modification in data values.

(ii) Output Perturbation Algorithm

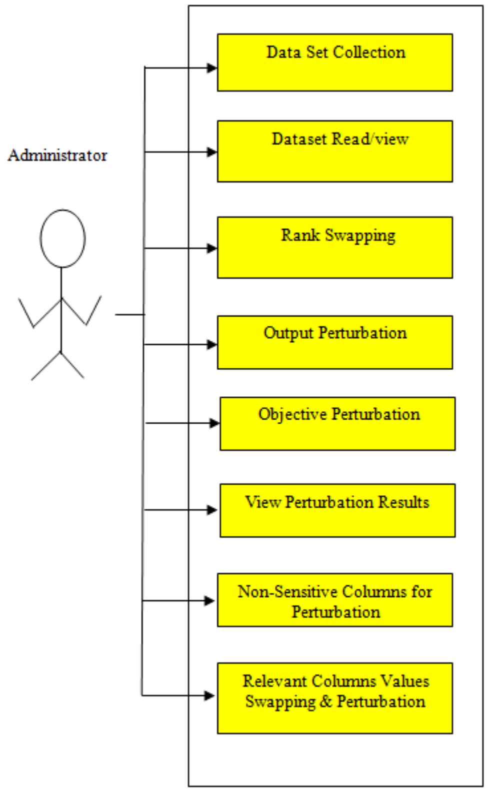
The datasets are taken such that D = {(x1,y1),(x2,y2),...,(xn,yn)} and D0 = {(x1,y1),(x2,y2),...,(x0n,y0 n)}, and suppose that the outputs of the algorithm are W(D) and W(D0), respectively. For the objective function W(D) = argminK(u,D), the output of the algorithm is W(D) + q, where q is a random Laplace noise.
In order to measure the quality of the prediction function u: X → Y in the training datasets, the empirical risk minimization method is employed to minimize the K(u,D) with a nonnegative loss function s : Y ×Y → R.



(a) Sequence diagram


OPP and OJP
Sample Dataset Ts
Data Preprocessing:
Discretization, Perturbation using algorithm unrealized dataset [Rank Swapping]
Complemented and perturbed dataset T' and TP
Non-Sensitive Columns for Perturbation
Relevant Columns Values Swapping and Perturbation
Result and analysis

(b) System flow diagram


Administrator
Data Set Collection
Dataset Read/view
Rank Swapping
Output Perturbation
Objective Perturbation
View Perturbation Results
Non-Sensitive Columns for Perturbation
Relevant Columns Values Swapping & Perturbation

(c) Use Case diagram

$$K(u,D) = \frac{1}{n}\sum_{i=1}^{n} s(u(x_i),y_i) + \lambda Z(u)$$

where Z(u) represents the smoothness of a function, u is a linear prediction function, and λ is an adjustable parameter. Here equilidiean distance is found out such that distance from origin (0,0) to (x1,y1), (x2,y2) up to (xn,yn) and summed out to prepare K(u,D). Like that five values are prepared and minimum value index is taken. That laplacian noise is used to perturbate the dataset.

(iii) Objective Perturbation Algorithm

Like the previous module, the datasets are taken such that D = {(x1,y1),(x2,y2),...,(xn,yn)} and D0 = {(x1,y1),(x2,y2),...,(x0n,y0 n)}, and suppose that the outputs of the algorithm are W(D) and W(D0), respectively. For the objective function W (D) = argminK (u,D). Here the laplacian noise is added to K (u,D)s and the minimum value laplacian noise is selected. That perturbated records are given as new dataset.

(iv) Identification Of Non-Sensitive Columns For Perturbation

In this module, the column with more similar values is identified as well as with very less similar values is also identified. Those column values are not perturbated and given with original values in the perturbated (other columns perturbated) dataset. This is carried out to eliminate the burden of perturbations less important columns.

(v) Relevant Columns Value Swapping And Perturbation

In this module, the columns with more similar values (but not exact) are identified. Those column values are swapped in the same row. This is carried out to non-loss perturbation of the dataset.

VII CONCLUSION

In this paper, we propose a machine learning approach with differential privacy for preserving training datasets privacy, and apply this machine learning approach to smart edges in wireless big data scenario. We first design two different algorithms OPP and OJP to satisfy differential privacy by adding Laplacian mechanism. In addition, we consider the privacy issues of correlated datasets, and prove the differential privacy preserving of correlated datasets via theoretical analysis. Last but not least, we establish the experiments on the Tensor flow, and evaluate our methods on four different datasets. We compare our OPP and OJP algorithms with two benchmark protocols, i.e., SGD, PATE-G. The experiment results show that the proposed methods can achieve high quality privacy preserving and accuracy assurance. If Z(.) is strongly convex and second derivable, the loss function s(.) satisfies the convexity of all sample data A and is derivable, then the algorithm OJP satisfies $\varepsilon_{p\,–\,\text{differential privacy}}$

VIII REFERENCES

1. X. Ding, Y. Tian, and Y. Yu, “A real-time big data gathering algorithm based on indoor wireless sensor networks for risk analysis of industrial operations,” IEEE Transactions on Industrial Informatics, vol. 12, no. 3, pp. 1232-1242, 2016.

2. L. Kong, D. Zhang, and Z. He, “Embracing big data with compressive sensing: a green approach in industrial wireless networks,” IEEE Communications Magazine, vol. 54, no. 10, pp. 53-59, 2016.

3. F. Xu, Y. Lin, and J. Huang, “Big data driven mobile traffic understanding and forecasting: a time series approach,” IEEE Transactions on Services Computing, vol. 9, no. 5, pp. 796-805, 2016.

4. S. H. Zhang, D. D. Yin, and Y. Q. Zhang, “Computing on base station behavior using erlang measurement and call detail record,” IEEE Transactions on Emerging Topics in Computing, vol. 3, no. 3, pp. 444- 453, 2015.

5. M. Fredrikson, S. Jha, and T. Ristenpart, “Model inversion attacks that exploit confidence information and basic countermeasures,” In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015, pp. 1322-1333.

6. O. Denas and J. Taylor, “Deep modeling of gene expression regulation in an erythropoiesis model,” In Representation Learning, ICML Workshop, 2013.

7. H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, “NFV: State of the art, challenges, and implementation in next generation mobile networks (vEPC),” IEEE Network, vol. 28, no. 6, pp. 18-26, 2014.

8. Y. Wang, Q. Chen, and C. Kang,“Clustering of electricity consumption behavior dynamics toward big data applications,” IEEE Transactions on Smart Grid, vol. 7, no. 5, pp. 2437-2447, 2016.

9. Q. Liu, C. C. Tan, J. Wu, and G. Wang, “Towards differential query services in cost-efficient clouds,” IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 6, pp. 1648-1658, 2014.

10. Y. Hong, J. Vaidya, H. Lu, P. Karras, and S. Goel, “Collaborative search log sanitization: Toward differential privacy and boosted utility,” IEEE Transactions on Dependable and Secure Computing, vol. 12, no. 5, pp. 504-518, 2015.