Fuzzy Systems and Data Mining VI
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA200721

Federated Learning in Big Data Application and Sharing

Yang Jing[1], Zhang Quan, Liu Kunpeng, Jin Peng, Zhao Guoyi
Customer Service Center of the State Grid Corporation of China

Abstract: In recent years, electricity big data has extensive applications in the grid companies across the provinces. However, certain problems are encountered including, the inability to generate an ideal model using the isolated data possessed by each company, and the priority concerns for data privacy and safety during big data application and sharing. In this pursuit, the present research envisaged the application of federated learning to protect the local data, and to build a uniform model for different companies affiliated to the State Grid. Federated learning can serve as an essential means for realizing the grid-wide promotion of the achievements of big data applications, while ensuring the data safety.

Keywords: Federated learning, big data application, blockchain, data sharing

1. Introduction

Electricity big data is known to possess widespread applications at companies under the State Grid in recent years, including the evaluation of electricity charge collection risk[1]-[5], electricity theft identification[6]-[8], overload of distribution transformer[9]-[10], and electric vehicle layout optimization[11]-[12]. For each grid company across the country, the model is trained, based on the already possessed business data. Currently, different companies in different provinces work in an isolated manner. In the era of big data, the urgent issues faced include the means to break data barriers and to perform collaborative modeling across the companies. Collaborative modeling across the companies can help to overcome the problems like low data quality and small sample size at each grid company, and also promote the grid-wide big data applications.

Federated machine learning provides a framework for machine learning, which facilitates the data use and development of models in machine learning across the institutions under the premise of users' privacy and data safety protection, and compliance to laws and regulations. Federated learning was initially proposed by Google[13] in 2016 and started to attract a lot of interest in the academic and industrial circles by 2019. At present, federated learning has become a hot technology in the field of artificial intelligence research and application. At present, two major difficulties need to be overcome in big data applications: (1) Data-related problems: In actual applications, the problems of limited data volume and low data quality are common, especially in the specialized fields. For example, electricity theft analysis, severe overload in the distribution transformer, customer complaint forecast, and big customer

[1] Corresponding Author, Yang Jing, Customer Service Center of the State Grid Corporation of China, Tianjin, China ,E-mail: zjuyangjing@126.com.

loss. In these fields, the annotated data is hardly enough to support model training. (2) Problems of privacy protection: Worldwide consensus has been reached as to data privacy and safety, and concerted efforts have been made to protect the data privacy and safety. In May 2018, the European Union (EU) published the General Data Protection Regulation (GDPR), which consists the regulations imposing the constraints on data acquisition, transmission, retention, and processing. Requirements on data privacy and safety protection have added to the difficulties in data acquisition, sharing, and exchange and brought an unprecedented challenge for the realization of many artificial intelligence technologies and applications.

Literature [14]-[18] shows the research and practice of confidential data protection during the model training process. In this study, we propose a uniform federate learning platform, which can be used for the sharing of a uniform model without data exchange across the companies under the State Grid. A particular emphasis was laid upon the application of the horizontal federated learning across the companies under the State Grid.

The advantages and features of federated learning used across the companies under the State Grid via a uniform platform are as follows: (1) The model is trained separately by using local data of each company, which satisfies the requirements for users' privacy and data safety protection; (2) The model algorithm is trained by grid-wide data with iterative optimization, thereby solving the problems of limited sample volume and difficult annotation; (3) All participants enjoy an equal status and are engaged in fair cooperation. In this way, the encrypted exchange of information and model parameters can be achieved, while safeguarding the independence of participants. Moreover, all participants could grow synchronously.

2. Federated learning

Federated learning defines the machine learning architecture, under which collaboration among the parties possessing different data is achieved through the design of a virtual model, without incurring the need for data exchange. The virtual model is an optimal model for pooling the data from different parties, where the local objectives are served in different regions according to the model. With federated learning, it is essential that the modeling results should be infinitely close to those of the conventional model. That is, the data from different parties are pooled together for the modeling. Under the federated mechanism where all participants enjoy equality of identity and status, it is possible to build a shared data strategy. As there is no data transfer, there will be no privacy leak or violation of the data norm. Therefore, the requirements for data privacy protection and legal compliance gets satisfied.

Federated learning has three major components: data source, federated learning system, and users. The relationships between the three components is illustrated in Figure 1. In a federated learning system, data from all sources are pre-processed. The learning model is built collaboratively, and the output is the feedback to the users.

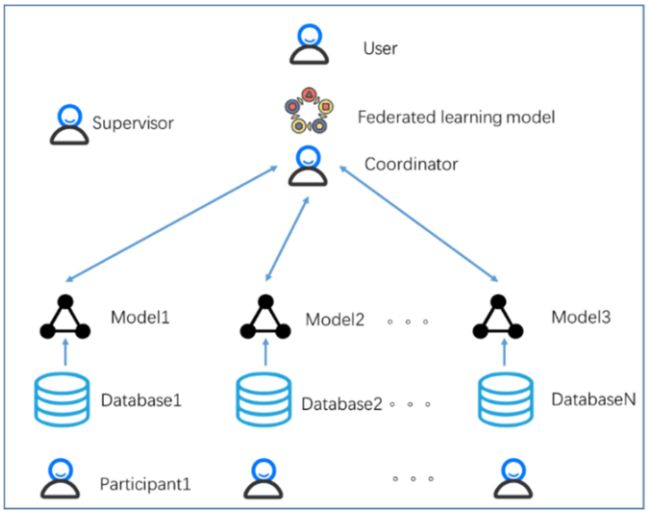

Figure. 1 Components of federated learning

The Customer Service Center of the State Grid Corporation of China has developed the supermarket for big data application sharing as a part of the exploration in federated learning applications. Based on the features of the State Grid Corporation of China, this supermarket offers a federated learning system along with the coordination and organization services. All companies under the State Grid Corporation of China are the participants and users of federated learning.

2.1 Vertical federated learning

This method employs two datasets that have many overlaps in users, but not in user features. The data are vertically partitioned (i.e., feature dimension), and the portions of data that share the same users but not the user features between the two datasets are selected and used for training. This method is known as vertical federated learning.

For example: Collaborative construction of an electricity charge overdue model is carried out between the grid company and the cooperative enterprise. The grid company possesses data Y, including the label data and overdue records. These data may be sufficient for building a good model. However, we may want to obtain more data, such as the label data and profile data of the cooperative party, in order to improve the performance and stability of the risk control model.

The problem with the conventional model is as follows: The cooperative enterprise cannot build the model independently due to the lack of data Y. It needs the grid company to bring data Y into the production environment of the cooperative enterprise for modeling. However, given the national laws on data protection and the strict regulations laid down by enterprises on their own data, the data X obtained from the grid company cannot be transmitted in full volume to the cooperative enterprise.

This problem can be addressed through vertical federated learning. As shown by the right side of the above figure, the data in both two parties share the same ID, though the features are different. The features lacking in one party can be made up by using the features from the other party.

Table 1 Vertical federated learning

Cooperative enterprise (bank)				Companies under the State Grid		
ID ID number	**X1 Account receivable age**	**X2 Monthly profit**	**X3 Grade**	**ID ID number**	**X4 Monthly electricity consumption**	**Y Electricity charge overdue**
u1	7	100	A1	u1	7	1
u2	8	200	A2	u2	8	1
u3	9	400	A1	u3	9	0
u4	10	500	A3	u4	10	0
u5	12	100	A4	u5	12	0

2.2 Horizontal federated learning

This method employs the two datasets that have many overlaps in user features, but not in users. The data are horizontally partitioned (i.e., user dimension), and the portions of data that share the same features, but not the same users between the two datasets are selected and used for training. This method is known as horizontal federated learning.

For example: The grid company in province A and the grid company in grid B collaboratively build an electricity theft model for the purpose of optimizing the electricity theft identification model. The need for collaborative modeling arises from the failure of the electricity theft models built separately by each party to meet the actual requirements for performance and stability. The federated learning mechanism can be utilized to make advantage of anti-electricity-theft samples from multiple parties to build a very expansive model without sample leak. Now with horizontal federated learning, the grid companies in both province A and B have (X, Y).

Table 2 Horizontal federated learning

Company in province A				
ID ID number	**X1 Daily electricity consumption of user**	**X2 Voltage**	**X3 Current**	**Suspected electricity theft**
u1	20	220	10	0
u2	30	220	20	0
u3	30	220	20	0
u4	40	220	20	0
u5	50	000	10	1
Company in province B				
ID ID number	**X1 Daily electricity consumption of user**	**X2 Voltage**	**X3 Current**	**Suspected electricity theft**
u6	20	220	0	1
u7	30	220	0	1
u8	30	220	0	1
u9	40	220	20	0
u10	50	000	10	1

3. Building a uniform grid-wide model through horizontal federated learning

The grid companies in different provinces share the same business-related need for the evaluation of electricity charge collection risk, electricity theft identification, overload of distribution transformer, work and production resumption analysis, and electric vehicle layout optimization. The model is trained separately by each company based on their own business data. At present, such work is done in an isolated manner. However, the State Grid Corporation of China faces an urgent need to build an effective uniform model for nationwide popularization by promoting a collaboration among the companies.

3.1 A brief introduction to the business scenarios

For the business scenarios of electricity theft and electricity charge collection risk, the following problems exist when the grid companies undertake the modeling work separately:

(1) Electricity theft is a small probability event, and the sample size of electricity theft is small. After eliminating the poor quality, the sample size gets even smaller.

(2) There are currently two pathways for the exploration of data sharing and data application sharing. The first is to pool the data from different parties. The second is to promote the wider use of mature model developed by a company in a certain province. However, there may be a problem of the data leak, and a general model is hardly possible due to the differentiation of businesses across the companies.

By forming an alliance for federated learning, different companies reach the federated protocol for multiparty cooperation. Without the need for sharing data across the parties, the federated big data ecosystem can be built to achieve the collaborative updates and optimization of the machine learning model. Thus, federated learning is very important in the optimization and integration of big data resources.

3.2 Architecture of horizontal federated learning

A typical architecture of horizontal federated learning is presented in the Figure 2. In this system, k participants sharing the same data structure are engaged in the collaborative learning of the machine learning model through parameters or cloud server. A typical hypothesis is that the participants are honest, while the server is honest and curious. Therefore, no participant is allowed to leak information to the server.

The training process of this system usually consists of the following four steps:

Step 1: The participants calculate the training gradients locally and mask the gradients through encryption, differential privacy or secret sharing technology. The masked results are sent to the server.

Step 2: The server implements secure pooling, without knowing about the information of the participants.

Step 3: The server sends the pooled results to the participants.

Step 4: The participants update their respective models using the deciphered gradients.

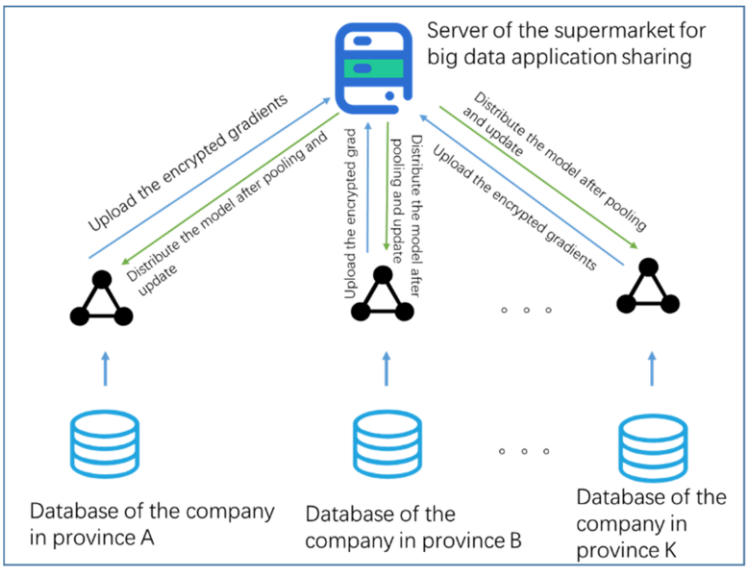

Figure. 2 Architecture of horizontal federated learning

The above steps are iterated until the convergence of the loss function, leading to the end of the entire training process. The above-described structure is independent of the type of machine learning algorithm used (eg., logistic regression or DNN). All participants share the final model parameters.

Safety analysis: If the gradient pooling is done by SMC or homomorphic encryption, it is usually indicated that the above-mentioned structure can protect a data leak from the influence of the semi-honest server. However, this structure can be vulnerable to the attack from an another safe mode. That is, the generative adversarial network (GAN) obtained by training in the collaborative learning process by the malicious participants.

The protocol developed by WeBank is used as a reference. The homomorphic encryption technology is used in this protocol, and the central parameter server has no way to know the parameters and models after pooling (this requirement can be loosened sometimes). In this way, the privacy of the participants is protected to the maximal extent. Moreover, the central parameter server is usually not involved in the training. Its only role is the pooling and distribution of the encrypted parameters.

3.3 Homomorphic encryption

Homomorphic encryption is a special encryption scheme, which allows the third parties to operate the encrypted data without a prior decryption, and the result of decryption is consistent with that of the plain text. Hence, the homomorphic encryption is an effective technology to protect the privacy of the users. The training process employs the homomorphic encryption to ensure that the intermediate training parameters have the same training effect with or without encryption. The data owner does not upload the model and data, hence, the other participants are unable to use the intermediate parameters to infer the content of source data.

Considering a logistic regression model as an example [19]:

$$(x_i, y_i) \quad , \quad i = 1,2,3, \ldots, n$$

where, $x_i \in R^m$, $y_i \in \{1,0\}$. (1)

Logistic maximizes the likelihood estimator as shown in Equation (2).

$$\prod_{i=1}^{n} P_r(y_i / x_i) = \prod_{i=1}^{n} \frac{1}{1 + \exp(-y_i(1, x_i)^T \beta)} \tag{2}$$

where, $\beta \in R^{m+1}$,Starting from an initial β_0,the gradient descent method at each step t updates the regression parameters using Equation (3).

$$\beta^{(t+1)} \leftarrow \beta^{(t)} + \frac{\alpha_t}{n} \sum_{i=1}^{n} \sigma(-z_i^T \beta^{(t)}) \cdot z_i \tag{3}$$

where, α_t is learning rate at step t,

and $z_i = y_i \cdot (1, x_i) \quad , \quad i = 1,2,3, \ldots, n$

The HEAAN scheme proposed by Cheon [20] [21] is adopted in this case. The HEAAN scheme supports the approximate arithmetic of the encrypted messages, so that the size of the parameters do not increase too much. Futhermore, it reduces a certain precision and greatly improves the efficiency. The Heaan scheme supports the key generation, encryption, decryption, addition and multiplication. At the same time, the scheme supports message packaging. If the encryption function is H, then the encryption gradient of each provincial company is shown by Equation (4).

$$G(u, t) = H(\frac{\alpha_t}{n} \sum_{i=1}^{n} \sigma(-z_i^T \beta^{(t)})) \tag{4}$$

where, G(u,t) is the Step t iterative encryption gradient for u-Th provincial company.

3.4 Technological realization

FATE, an open source program first maintained by WeBank, provides a secure computation platform to support the federated learning algorithms. FATE realizes the secure computation protocol based on homomorphic encryption and multiparty computation. It supports the secure computation in the federated learning framework and machine learning, including the typical machine learning algorithms as logistic regression and gradient boosted regression trees, and also in deep learning, transfer learning and other frontier algorithms. FATE not only provides a framework, but also runs some classical algorithms, including the linear regression, gradient boosted regression trees and other classifiers. It has been sufficiently verified in practice that FATE can be readily applied to the industrial field. If the developers are not willing to construct the federated learning model from scratch, they can borrow from the mature ones or make certain modifications on this basis.

The architecture of FATE[22] is illustrated in Figure 3.

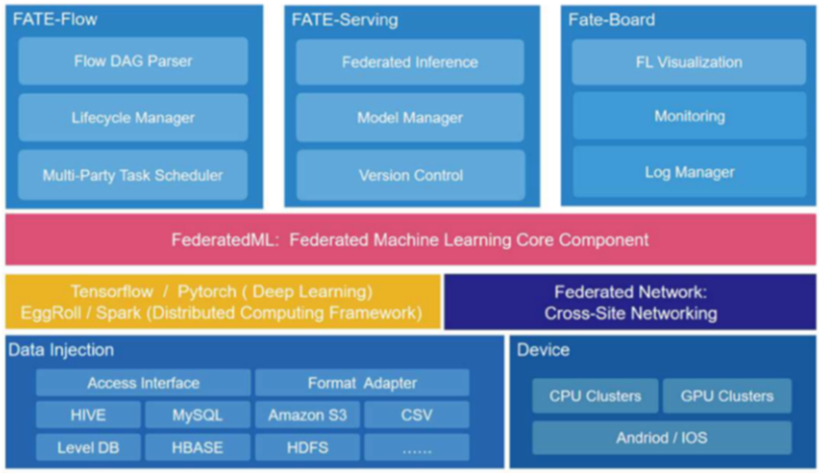

Figure. 3 Architecture of horizontal federated learning

Furthermore, FATE offers a set of friendly, cross-domain mutual information management protocols, which solves the difficulty of information security audit in federated learning. A simple, easy-to-use open source tool platform enables a multiparty data collaboration while protecting users' privacy and data safety and conforming to government regulations.

3.5 Applications of federated learning in the supermarket for big data application sharing

Federated learning is not only a technical standard, but also a business model. The Customer Service Center of the State Grid Corporation of China sets up a special space for federated learning in the supermarket for big data sharing, which can serve as a mutual benefit and collaboration platform for different companies under the State Grid. Stage 1. Model training: The sensitive data can be shared through the training model to the party needing these data. That is, the party needing the data, uploads the model to the sharing platform, from which the data provider downloads the model and trains it with its own data. After the training is done, the updated model is uploaded to the sharing platform. In this way, only the trained model is made available for the party needing the data, which allows for data sharing without the leak of sensitive data.

This method is further illustrated through the example of electricity theft identification. To this goal, the data on the users' daily electricity consumption, three-phase voltage, three-phase current, distribution transformer number, and user profiles are usually needed. The target variable is suspected electricity theft, to which the value of 1 or 0 is assigned.

The training process of the horizontal federated learning model can be summarized as follows: Initially, the initiator of federated learning for electricity theft identification uploads the original model to the supermarket for big data application sharing as the initial sharing model. Each participant downloads this model and independently calculates the gradient according to their respective data. Initially, the gradient is encrypted and sent to the server of the big data sharing supermarket. The federated

model of the big data sharing supermarket (server) implements weighted averaging of the models uploaded to the cloud without accessing the data from any client terminal. Thus a new sharing model is obtained. Later, the computation results are returned to each participant. Finally, the client terminals update their respective models using the deciphered gradients.

Let the encryption function of HEAAN scheme be H.Equation (5) can be deduced as:

$$G(u,t) = H(\frac{\alpha_t}{n}\sum_{i=1}^{n}\sigma(-z_i^T\beta^{(t)})) \tag{5}$$

In step t, the gradient values of provincial companies are G(u1,t),G(u2,t),……, G(um,t), which are uploaded to the server of federal learning management system. The server carries out weighted average as given by Equation (6).

$$\overline{G_t} = \sum_{i=1}^{n} p_i G(u_i,t) \tag{6}$$

The server is unable to decrypt G and also the model parameters. It only knows the homomorphic encryption domain, but it is unable to decrypt without the private key. The client downloads the parameters, decrypts it with its own private key, and then updates its own model.

Operational logic:

Step 1: t=1, Provincial companies get the provincial model gradient based on local data

$$g(u_i,t) = \frac{\alpha_t}{n}\sum_{i=1}^{n}\sigma(-z_i^T\beta^{(t)}) \tag{7}$$

Step 2: Gradient encryption

$$G(u_i,t) = H(g(u_i,t) \tag{8}$$

Step 3: Upload to the server of federal learning management system

Step 4:The server carries out weighted average

$$\overline{G_t} = \sum_{i=1}^{n} p_i G(u_i,t) \tag{9}$$

Step 5: The server sends it to the provincial company, provincial company get Decryption gradient,

$$g(u_i,t+1) = D(\overline{G_t}) \tag{10}$$

Step 6: Proceed to the next iteration

$$t = t + 1 \tag{11}$$

Step 7: When the overall loss function is less than the threshold value, the iteration is stopped

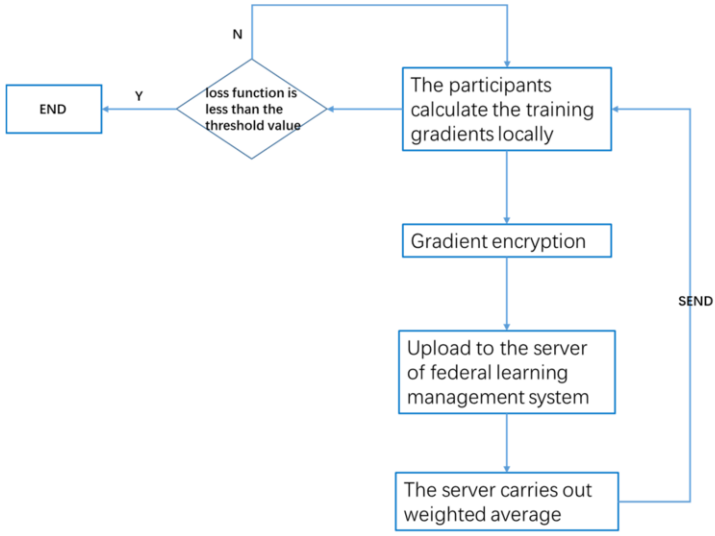

Figure. 4 Horizontal federated learning Operational logic

Stage 2. Application of a common model: The common model is shared through a remote technology. The data users can have a direct access to the big data application products by uploading the data or by processing the data with the middle platform. Application products are usually related to specific business scenarios, including the whole process of data processing and integration, computation, analysis and display. These products have a strong specificity for businesses, and can be shared via the remote technology. The party needing the data can use the application products according to the instruction manual. The data processing can be performed according to the use workflow for the application products.

The process of common model sharing can be summarized as follows: The cooperative party of the federated learning uploads the common model to the big data application platform, which can developed into the abnormal electricity consumption identification products for the use of grid companies in other provinces. This cooperative party will get credits as a reward, so as to facilitate the popularization of the achievements of big data applications.

Practical application: Five provincial companies participate in the federal study of electricity theft identification. The data format of provincial companies collects and processes electricity charge data, voltage data and current data according to the same standard, and marks electricity stealing situation, the data standards is shown in the Table 3.

Table 3 Data standards

Customers data				
Customer id	**Tariff data set**	**Voltage data set**	**Current data set**	**Suspected electricity theft**
id1	T1	V1	C1	1
id2	T2	V2	C2	1
id3	T3	V3	C3	1
id4	T4	V4	C4	0
……	……	……	……	……

The application effect is shown in the Table 4.

Table 4 Model effect

Model effect			
Model	**Sample size**	**Prediction accuracy (%)**	**Recall (%)**
Provincial companies independently identify the electricity theft (Logistic)	10000(avg)	30.4(avg)	50%(avg)
Provincial companies use the federal learning model (Logistic)	10000(avg) 50000(total)	70.5(avg)	60%(avg)

4. Conclusion and Future Work

Through the practical application of federal learning in the analysis of electricity theft in five provincial companies, it can be seen that the effect of federal learning is better than that of individual provincial companies.

We have discussed the multiparty collaboration and application of horizontal federated learning within the State Grid. The electricity businesses across the companies in different provinces have considerable overlaps horizontally, and the federated learning enables the trans-party model training. However, a cross-domain collaboration may be needed for the construction of the electricity big data application sharing ecosystem. For example, the collaboration between the banks and internet companies. Vertical federated learning is a good way to expand the knowledge of the attributes of electricity users in the financial and Internet fields. These features, if pooled in an encrypted state, can be used to promote the performance of the sharing model and in the construction of the electricity big data application ecosystem. In the future, the application of vertical Federation learning is the focus of our research and exploration.

References

[1] Li Jianbin, Wu Binbin, Zhu Yakui, Wang Yue, an Yagang, Zhao Shasha. Risk prediction and control of customer electricity tariff based on big data analysis [J]. Power Systems and Big Data ,2019, 22 (2):1-6.

[2] He Rong, Zhang Xiangdong, Qiu Lin, Chen libing, Zhou Qian, Zhang Yan. Electricity tariff risk prevention and control based on clustering and logistic regression [J]. Power Systems and Big Data ,2019,22(12): 42-49.

[3] Zhao Hong, Shen Jianzhong, Wang Jun, Zhang Cheng, Qu Qing. Risk prediction model of electricity charge recovery based on customer portrait and machine learning algorithm and its application [J]. Microcomputer Applications,2020, 36(2):93-96.

[4] Wang Zongwei, Zhao Guoqi, Jin Peng, Yang Jing, Wang Hailong, Zhang Zuobin. Research on the risk identification model of electricity tariff based on customer credit system [J].Machine Design and Manufacturing Engineering ,2019 ,48(10): 99-104.

[5] Xie Ying, Wang Zheng, Zhao Yongliang. Research on risk identification model of customer electricity consumption behavior [J]. Zhejiang Electric Power, 2017,36(12):53-56

[6] Qiang Hao, Dai Qiaoyun, Wu Ke, Du Jian, Yin Xinbo, Chen Chen. Research on anti stealing technology of variable structure BP neural network based on big data [J].Journal Of Jiangsu University Of Technology,2019,25(2):10-14

[7] Li Zhi Peng, Hou Huiyong, Jiang Si fan, Wan can, Zheng Ruimin. Line loss calculation and power stealing analysis based on artificial neural network [J]. Southern Power System Technology, 2019,13(2):8-12

[8] Li Bo, Cao min, Zhu Yuanjing, Li Shilin, Zhang Linshan, Lin Cong, Wang Xianpei. A joint power stealing detection method based on network characteristics and user behavior analysis [J]. Engineering Journal of Wuhan Universit,2019, 52（12）:1121-1128

[9] Qiu xiaogeng, Dong Xiangyu, Zhang Peng, Liang Wei, Guo Jun, Bai Kaifeng, Liang Zhixian, Feng Chao. Analysis of distribution transformer overload warning based on big data [J]. POWER SYSTEMS AND BIG DATA 2018,21(10):38-42

[10] He Jianzhang, Wang Haibo, Ji Zhixiang, Meng Xiangjun, Zhang Tao. Heavy overload prediction of distribution transformer based on Stochastic Forest theory [J]. Power System Technology, 2018,21(10):38-42

[11] Xie Yuande, Zhang Lin, Deng Shali, he jiemeng. Study on optimization network layout of electric vehicle charging facilities [J]. Practice and understanding of Mathematics, 2020,50(10):168-176

[12] Sun Yue, Jiang Cheng, Wang Zhizhi, Tang chunsen. Optimal layout of dynamic wireless power supply system for electric vehicles based on PSGA [J]. Power system automation, 2019, 43（9）: 125-131

[13] McMahan H B ,Moore E , Ramage D , etal . Communication-efficient learning of deep networksfrom decentralized data[C]//Proceedings of the 20th
International Conference on Artificial Intelligence and Statistics (AISTATS) 2017. JMLR : W&CP volume 54.

[14] Jia Yanyan, Zhang Zhao, Feng Jian, Wang Chunkai Applications of federated learning models in the processing of confidential data [J]. Journal of China Academy of Electronics and Information Technology, 2020, 15 (1):43-49.

[15] Wang Chunkai, Feng Jian. A study on the applications of federated learning in the insurance industry [J]. Journal of Insurance Professional College (Bimonthly), 2020, 34 (1):13-17.

[16] Wang Yashen. A review of the technological development of federated learning oriented towards data sharing and exchange [J]. Unmanned Systems Technology, 2019, 2(6); 58-62.

[17] Pan Rusheng, Han Dongming et al. Visualization of federated learning: challenges and architecture [J]. Journal of Computer-Aided Design & Computer Graphics, 2020 (32):1-6.

[18] Liu Junxu, Meng Xiaofeng. A review of the privacy protection studies in machine learning [J]. Journal of Computer Research and Development, 2020, 57 (2): 346－362.

[19] CHENH, GILAD-BACHRACHR, HAN K,etal. Logisticregression overencrypted datafrom fully homomorphic encryption[J]. BMC Medical Genomics, 2018, 11(S4): 3–12. [DOI: 10.1186/s12920-018-0397-z]

[20] CHEON J H, KIM A, KIM M, et al. Homomorphic encryption for arithmetic of approximate numbers[C]. In:Advances in Cryptology—ASIACRYPT 2017, Part I. Springer Cham, 2017: 409–437. [DOI: 10.1007/978-3-319-70694-8_15]

[21] CHEON J H, HAN K, KIM A, et al. Bootstrapping for approximate homomorphic encryption[J]. IACR CryptologyePrint Archive, 2018: 2018/153. https://eprint.iacr.org/2018/153
[22] Artificial Intelligence Department of Weizhong bank, et al. Federal learning white paper[M].2020.4