Research and application of carbon emission data sharing model based on privacy computing

Ying Shi¹, Lixing Wang¹, Jin Ao²*, Jie Song¹, Qian Li², Xuesen Zhang²

¹State Grid Jilin Electric Power Company Co., Ltd, Jilin, China
²State Grid Blockchain Technology (Beijing) Co., Ltd., Beijing, China

Abstract: In the context of the dual-carbon goal, improving the level of scientific and technological innovation of modern carbon emissions and achieving comprehensive green development are not only one of the trends and important reform directions of today's world development, but also the core content of achieving sustainable development. Taking the power industry as an important breakthrough, this paper uses blockchain technology and privacy computing technology to integrate data from the government, power grid, taxation and banks, and proposes a blockchain-based privacy computing algorithm for carbon emission scores to create an execution environment for trusted sharing of carbon emission data, expand channels for carbon emission data collection, and give full play to the value of big data on carbon emissions. Promote the realization of dual carbon goals as soon as possible.

1. Introduction

On September 22, 2020, at the General debate of the seventy-fifth session of the United Nations General Assembly, the General Secretary solemnly declared that "China will increase the intensity of its nationally determined contributions, adopt more powerful policies and measures[1-5], strive to peak carbon dioxide emissions before 2030, and strive to achieve carbon neutrality before 2060." To achieve carbon emission reduction, the energy industry as the main battlefield, the power industry as the main position, on December 28, 2020, the State Grid Co., Ltd. held a special meeting to study the "carbon peak, carbon neutral" action plan[6-8], to accelerate the clean and low-carbon energy transformation, promote green development put forward proposals for the State grid, formulate the State grid plan, and contribute to the power of the State grid. In early March, the State Grid Co., Ltd. issued the "carbon peak, carbon neutral" action plan[9-11], proposed to speed up the construction of a clean low-carbon, safe and efficient energy system, continue to promote carbon emission reduction, and clear the implementation path for the energy and power industry to implement carbon peak and carbon neutral[12-14].

With the long-term goal of "carbon neutrality and carbon peak" proposed, the dual carbon economy replaces the low carbon economy has become a hot spot, among which the establishment of a digital carbon economy model guided by the goal of "dual carbon" is an important means to reflect the changes of regional "carbon economy"[15-16]. Build a carbon emission data sharing and privacy calculation model for the government, judiciary, power plant, power grid, tax, users and other parties oriented to regional carbon emissions, so as to realize the "usable and invisible" carbon emission data, protect the data privacy of all parties, and realize the trusted sharing of carbon emission data.

Through the use of blockchain and privacy computing technology, this paper eliminates trust barriers in the encrypted environment, enhances the application scenario of urban carbon economy, improves the quality of urban data information and data sharing efficiency, and realizes the efficient management and operation of the city. Integrate data in the field of energy and government affairs, and create products such as macroeconomic dynamic and accurate prediction, quantitative analysis of industry production status, energy regulation, evaluation of investment promotion effect, and monitoring of new energy consumption, so as to provide auxiliary decision-making services for governments and enterprises.

2. Multidimensional data source analysis of user carbon emissions

In order to enhance the value of private data on blockchain and ensure data privacy security, this paper studies the blockchain-based multi-source carbon emission data analysis model to realize the privacy protection computing technology of carbon emission business. In order to achieve a more accurate and meaningful calculation scheme, this topic promotes the integration of multi-source carbon emission data resources through the integration of external data including government, power grid, taxation, banks, etc., and the power...
provided by the State grid. Since multi-source carbon emission data come from different departments and systems, there are problems such as different data structures, different names, data missing, conflicts and repeated redundancy, etc. Therefore, in the process of carbon emission data sharing system construction, the integration of multi-source carbon emission data needs specific processing.

"Multi-source" means that there are many sources and wide sources of information. Multi-source information system mainly refers to the comprehensive management information system of multi-source and massive privacy data generated in the daily operation and management of government, power grid, tax and bank. Because the power grid enterprise system is not interoperable, so the amount of information and information content are different, some power generation enterprises have mastered the data required for mass production, each department that has mastered the production data information has formed an independent information system, transportation enterprises and electricity sales enterprises are also the same. Therefore, in the whole power grid environment, the summary of carbon emission information of independent parties becomes the source of multi-source data of the system.

To analyze carbon emission data access, big data is the basis. It is necessary to collect and access a large amount of information of enterprises as much as possible, establish enterprise user files and electricity consumption information files, and record users' monthly electricity, electricity, nuclear quantity, actual capacity, etc. Record the user's payment (owed) fee information, including the total number of monthly payments, the number of monthly payment of liquidated damages, the monthly electricity receivable, the amount of unpaid fees, the amount of electricity received in advance, the number of monthly reminder fees, the date of actual payment of electricity, the date of electricity issue, the payment method and other data. It is also necessary to access electricity theft and other default information, including the number of electricity theft, the amount of electricity theft, the number of illegal power consumption safety accidents, the number of electricity default, and the default electricity cost. Carbon emission data is closely related to the user's electricity consumption data, and a comprehensive multidimensional data source for carbon emission users is needed.

External data access: External information about users collected from multiple sources, including carbon emission measurement standards from government agencies, carbon emission policies, and carbon emission measurement from energy big data centers. Financial institutions' users' operating property status, audit details, raw material purchases, product benefits and other data, judicial environmental penalties, environmental data, tax records, tax evasion and other data.

3. User carbon emission scoring algorithm based on machine learning

This paper mainly quantifies the user's carbon emission score from these four perspectives.

(1) Consumer power consumption behavior index.
It focuses on the business characteristics of the user's electricity consumption, the user's monthly electricity consumption, electricity charge, nuclear quantity, actual capacity, actual payment data, and electricity theft data, so as to launch the user's carbon emission baseline data.

(2) Government electricity consumption behavior index.
It mainly includes the carbon emission data recorded by the user in the provincial energy big data center, as well as the industry carbon emission factor data collected by the National Bureau of Statistics, as auxiliary data of the user's electricity consumption, as well as the environmental penalties and environmental data recorded by the government judicial system on the user.

(3) User development index.
User development indicators analyze the growth and volatility of electricity consumption and reflect the operating status of enterprises. The power and capacity change dimension reflects the development trend of enterprise operation, the power volatility dimension reflects the stability of production and operation, and the industry comparison dimension reflects the actual value of electricity consumption and the tax situation of users.

(4) Bank economic data.
User operating property status, audit details, raw material purchase, product output, production plan and other data, the industry user dimension measures the overall development level of the industry.

Current carbon emissions assessments rely mainly on the basic information of experienced professionals to estimate the above four aspects. With the growth of the data age and the number of businesses, traditional manual audit methods are no longer suitable. A single model has its own advantages and disadvantages in computing speed and forecasting effect. Combining different models can give full play to their advantages and improve the generalization ability of the model. In this paper, a personal carbon emission risk assessment model integrating GBDT and LR algorithms is proposed. GBDT is used to transform data features, and then input data into LR for classification training. The advantages of the two algorithms are fully utilized, and the prediction accuracy and stability of the model are effectively improved.

Logistic Regression (also known as log-probability regression) is a generalized linear regression analysis model, which is widely used in evaluation and rating because of its simple and effective algorithm.

Suppose $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})^T$ (T stands for transpose). among $x_{ik}(k = 1, 2, \ldots, n)$ for the government, power grid, taxation, banks and other departments to evaluate the corresponding aspects of the k enterprise, carbon emission assessment is to find a classification function. $P(y_{i1} = 1) = f(x_{i1}, x_{i2}, \ldots, x_{in})$. Where $P(y_{i1} = 1)$ is the probability that an enterprise is classified into a category with bad
carbon emission, and $f$ is the classification function.

The evaluation steps of the Logistic carbon emission assessment method are as follows.

Based on previous statistics, $(X_i, Y_i), i = 1,2,\ldots,n$ is obtained. Where $X_i$ is the relevant data of the $i$ enterprise, $Y_i$ indicating the carbon emission of the enterprise ($Y_i=1$, indicating good carbon emission of the enterprise; $Y_i=0$, indicating bad carbon emission of the enterprise).

Establish a carbon emission assessment model, $P(Y_i=1) = p_i = \frac{1}{1+e^{-(R_{Y_i}+\sum_{k=1}^{K} p_k Y_k)}}$, Where $\beta_i, i = 0,1,\ldots,n$ is the model parameter. Estimation of the model. Solve the parameters in the model $\beta_i, i = 0,1,\ldots,n$. The most common way to solve this is to maximize the log-likelihood function: $\prod_{i=1}^{n} \prod_{j=1}^{d} \left[ \beta_i \sum_{x_{ij}} x_{ij} y_{ij} - \sum_{x_{ij}} \log(1 + e^{\beta_i \sum_{x_{ij}} x_{ij} y_{ij}}) \right]$, obtain $\beta_i$, substitute into $i=1,2,\ldots,n$. The corresponding carbon emission assessment model is obtained. $p_i = \frac{1}{1+e^{-(R_{Y_i}+\sum_{k=1}^{K} p_k Y_k)}}$. For an enterprise to be evaluated, its corresponding value of $x$ is substituted to obtain the corresponding value of $p_i$. The judgment rule is: if only $p_i >= c$ is used, the enterprise belongs to the class with good carbon emission. Otherwise, the enterprise belongs to the category of poor carbon emission.

Logistic regression can easily update the model and absorb new data, which is easy for researchers to understand and operate. However, this model has limitations in its ability to adapt to data and scenarios, and it cannot handle a large number of multi-class features and variables, so it has weak adaptability when applied to the big data of carbon emission risk that changes at any time.

### 3.1. Gradient lifting iterative decision tree

Gradient Boosting Decision Tree (GBDT) is an ensemble learning method, which is an application of Boosting strategy combined with decision tree algorithm. GBDT builds the final predictive model by combining multiple learners, with multiple classifiers trained in turn and the outcome of each tree determined by the previous results. For a complex task, the judgment obtained by combining the judgments of multiple learners is better than the judgment made by any one of them alone.

If a log-likelihood loss function similar to logistic regression is adopted, the data representation of a training set sample is $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, the maximum number of iterations is $T$, the loss function is $L$, and the output is a strong learner $f(x)$. The expression of the loss function is: $L(y, f(x)) = \log(1 + \exp(-yf(x)))$, Where $y \in \{-1,1\}$. Then the negative gradient error at this time is: $r_t = -\frac{\partial L(y_t, f_t(x))}{\partial f_t(x)} = \frac{y_t}{1+\exp(y_tf_t(x))}$. Since the above formula is difficult to optimize, approximation is generally used instead: $c_t = \frac{\sum_{j=1}^{d} Y_{ij} r_t}{\sum_{j} r_t[1-|r_t|]}$. The final learner is: $f(x) = f_0(x) + \sum_{t=1}^{T} \sum_{j=1}^{d} Y_{ij} \beta_{tj} x_j$.

GBDT has strong expression ability, high prediction accuracy, and does not need to carry out complex feature engineering and feature transformation. At the same time, it can flexibly handle various types of data, including continuous and discrete values. The disadvantages of GBDT are also obvious. The credit risk data has a high dimension and some features have a strong sparsity. The use of GBDT will greatly increase the computational complexity.

### 3.2. Automated scoring model for privacy data

As a generalized linear model, LR algorithm is simple and interpretable, and can be used for massive data. However, LR algorithm has limited learning ability and high requirements on data features, which can easily lead to underfitting. Therefore, before classification training, effective feature engineering is needed to extract features from the original data to obtain better classification results. In 2014, Facebook proposed GBDT+LR combined model for CTR prediction, and Boosting Tree model uses its own feature combination ability for feature engineering. Boosting Tree model itself has the ability of feature screening and higher-order feature combination. Boosting Tree model can generate new discrete feature vectors for input of LR model by GBDT for feature screening and combination, and can get better prediction effect.

First, the training set constructs a series of decision trees through GBDT to form a strong learner. The path from the root node to the leaf node can be regarded as a feature combination of different features, and a leaf node corresponds to a discrete feature. Then, the feature processing is passed into the LR classifier for secondary training through one-hot coding. As shown in Figure 1, the training process of GBDT+LR fusion model is as follows:

The new discrete features constructed by GBDT are shown in the figure. Assume that $f_{m-1}$ and $f_{m}$ are two decision trees generated in the training process of GBDT algorithm, with 5 leaf nodes respectively, where the number 1 indicates that the results predicted by the training sample $x$ through the decision tree fall on the leaf node. Then for tree $f_{m-1}$, the predicted result can be represented by One-Hot coding as $[0,1,0,0,0]$. As shown in Figure 2, assuming that the number of iterations of GBDT algorithm is $x$ and all weak classifiers have a total of $y$ leaves, each of the $y$ pieces of original data will be converted into an $y$-dimensional sparse vector, in which $x$ elements are 1 and $x-y$ elements are 0, then a new training set with dimension $m \times x \times y$ will eventually be formed.

![Figure 1. GBDT+LR model training diagram](image-url)

1. Construct training set
2. Feature extraction
3. DT classifier training
4. Decision tree T1
5. Decision tree T2
6. Extract new features
7. DT classifier training
8. Decision tree T3
9. …
10. Decision tree Tn
11. …
12. Test set test
function edge node using its local data set. Therefore, the local loss average of the local loss function is minimized by minimizing the weighted nodes data sample.

Its main features are distributed and privacy protection. It developing and promising machine learning technology. The carbon business system is large and involves multiple parties, and federated learning technology is a rapidly growing technology. The joint use of blockchain technology and federated learning technology can effectively protect data privacy, so this topic uses same as blockchain technology, this technology can also be used as a sample of federated learning data.

As shown in Table 1. As the display link of big data, data visualization makes the results at a glance with three-dimensional, easy-to-understand charts and graphs, which is convenient for data analysts and platform operators to understand the deep meaning of carbon emission data, so as to more effectively participate in the data analysis process and improve the data analysis results.

### 4. Carbon emission score privacy calculation algorithm based on blockchain

The carbon business system is large and involves multiple parties, and federated learning technology is a rapidly developing and promising machine learning technology. Its main features are distributed and privacy protection. It enables edge devices to collaboratively train shared global models published by task publishers without the need to upload local data. After each edge device obtains the global model, it trains the model locally with its own locally private data. The edge device then uploads new model parameters (i.e., updates the model locally). The same as blockchain technology, this technology can also effectively protect data privacy, so this topic uses blockchain technology and federated learning technology to comprehensively build privacy data protection technology to achieve effective protection of data privacy.

Each mobile device has a local training data set that can be used as a sample of federated learning data. Specifically, suppose that each edge node $n$ has a local data sample $s_n$. The total size of the data sample of $n$ edge nodes $\sum_{n=1}^{N} s_n = S$. Then, the goal of federated learning is to reduce the global loss function $l(\Phi)$, and the global loss function is minimized by minimizing the weighted average of the local loss function $l_n(\Phi)$ trained by each edge node using its local data set. Therefore, the local loss function $l_n(\Phi)$ and the global loss function $l(\Phi)$ are calculated as follows:

$$
\min_{\Phi} l(\Phi) = \frac{1}{S} \sum_{n=1}^{N} s_n f_n(\Phi)
$$

Where, $f_n(\Phi)$ is the loss function of sample data in the local data set of edge nodes.

In global iteration $t$, all edge nodes work together to obtain their average gradient $\Lambda_n$ by training their local data set using an optimization algorithm. General federated learning uses stochastic gradient descent (SGD) to iterate. SGD iteratively selects a set of training samples, calculates their gradient values with respect to $\Phi(t)$, and minimizes $l_n(\Phi)$ using the gradient step in the corresponding direction. Suppose, given the learning rate of edge node $n$ is $h_n$, then its local model is updated as:

$$
\Phi_{n}^{(t+1)} = \Phi^{(t)} - h_n \Lambda_n.
$$

Therefore, the task publisher updates all the local models uploaded from each edge node through weighted aggregation to update the shared global model $\Phi^{(t+1)}$, as follows:

$$
\Phi^{(t+1)} = \sum_{n=1}^{N} \frac{s_n}{S} \Phi_{n}^{(t+1)}.
$$

It can be seen that edge nodes with higher quality and reliable local data samples in federated learning can make local loss function $l_n(\Phi)$ and global loss function $l(\Phi)$ converge faster. As a result, high-quality, reliable edge nodes can significantly improve the accuracy and learning efficiency of the entire federated learning task, and can also reduce the time and energy consumption of federated learning training.

The federated learning module is the basis of the entire task and is responsible for the main body of data interaction. The main process of the algorithm is shown in the figure 3. The user carbon emission scoring algorithm based on machine learning proposed in the previous section is deployed by the power grid on the chain. The government, bank, power grid and tax are taken as four nodes to obtain the user carbon emission scoring algorithm based on machine learning on the chain. The model synthesis contract merges the model, and the merged model is re-deployed on the chain in the form of a smart contract to score carbon emissions of carbon emission users.

![Federal learning flow chart for carbon emission scoring](image.png)

**FIG. 3** Federal learning flow chart for carbon emission scoring

### 5. Conclusion

With the long-term goal of "carbon neutrality and carbon peak" proposed, the dual-carbon economy has become a hot spot to replace the low-carbon economy, and the zero-carbon economy has begun to emerge. The privacy calculation algorithm of carbon emission score based on blockchain proposed in this paper provides a trusted data basis for carbon emission data sharing, data security, data...
privacy protection and other functions. Explore the potential value of carbon emission data, standardize and standardize carbon emission data system. Expand the channels for carbon emission data collection, give full play to the value of carbon emission big data, and achieve the goal of reducing and promoting dual carbon as soon as possible.

**Acknowledgment**

This work was supported by the State Grid Jilin Electric Power Company Co., Ltd Science and Technology Project “Research on Urban Carbon Economy Model Based on Privacy Computing” (SGDSQK00QKJS2200079).

**References**

10. Sun H , Gao G . Research on the carbon emission regulation and optimal state of market structure: Based on the perspective of evolutionary game of different stages[J], 2022, 56(4):2351-2366. DOI:10.1051/ro/2022088.