Research on Homomorphic Encryption Clustering Algorithm for Smart Grid Privacy Preserving

Yun YANG[1, a], Xiao-Yong ZHANG[2, b], Zhu ZHU[2,c] and Juan LEI[2, d, *]

[1]State Grid Chongqing Electric Power Company, Chongqing, 400015, China

[2]State Grid Chongqing Electric Power Co. Electric Power Research Institute, Chongqing, 400014, China

[a]yy@cq.sgcc.com.cn, [b]qnzxy@cq.sgcc.com.cn, [c]zz@cq.sgcc.com.cn, [d] 61902966@qq.com

*Juan LEI

Keywords: the smart grid, privacy preserving, data mining.

Abstract: In order to solve the recent privacy issues in the process of the smart grid data mining, this paper analyzes the data mining techniques for the smart grid data and states the existing research methods of the privacy issues in the process of the smart grid data mining; Then this paper designs a homomorphic encryption clustering algorithm for smart grid privacy preserving. This algorithm uses homomorphic encryption to encrypt the clustering data in local stations sent to the global station. After processing the encrypted data, the global station sends the data back to the local stations for decryption. The experiment using the sample data of the database of the power quality detection network shows that we can classify the consumers by clustering their daily load curves. Finally, we demonstrate that this algorithm can protect the privacy of the smart grid data effectively by deducing the correctness and security of this algorithm.

Introduction

The smart grid is the integration of the custom power technology, the distributed power system and advanced information technique, and it is also a form of the grid intellectualization. It has arouses global common concerns, at the same time it becomes a significant developmental strategy of our country [1]. With the integration of information communication techniques and electricity power industry, the issue of the smart grid big data has been proposed, as a result of the scale，whose data produced by the grid system and the electricity market，increasing exponentially[2]. The sources of the smart grid data are divided in three parts: electricity generation side, electricity transmission and transformation side, and electricity demand side. That the data about customers using electric power are collected by power companies forms the marking large data-sets, related to the consumers. Discovery the knowledge from data-sets is an important researching direction of the construction of smart grid.

It will impact the intelligent process of our country grid system, if we neglect privacy leaking issues and the hidden risks [3]. According to the issues above, this paper selects k-means clustering algorithm to mine smart grid large data-sets, on the basis of studying the current smart grid data mining techniques. Furthermore, researching the data mining privacy preserving techniques, this paper presents a homomorphic encryption clustering algorithm that used for preserving the privacy of the smart grid data. That this algorithm can cluster the consumers is benefit to form reasonable electric power price mechanism, by using their daily load curves of electric power system.

Related Working

The knowledge discovery in Database is a complicated process containing many techniques. The famous process model is the multistage process model of large data-sets designed by Usama Fayyad. There are five steps in the KDD process: Data-Collection, Data-Preparation, Data-Storage and Management, Data-Mining, Knowledge-Exhibition. Recently, popular research topic of the behavior feature of electric consumers are classifying the electric consumers, market segment, electro-load forecasting, consumers responding under the electro-price promoting and so on [4].

Establishing and optimizing a system of time-of-use (TOU) tariff or categorized electricity tariff is benefit to promote the optimized allocation of electric power resources. As the electric power market come into a new stage in which guided by the market requirement, classifying the electric power consumers is one necessary measure to realize the optimal allocation of electric power resources [5]. This paper uses a consumers clustering algorithm based on features of load curves. This method extracts data that consumers daily load curves in a certain time from electric power data-sets, classifying them by seasons and holidays. And the load curves standing for different consumers are dealt into average daily load curves.

Compared with common information security issues, there are three issues in the field of privacy preserving of smart grid large data-sets: privacy preserving of the information of consumers, creditability of larger data-sets, and how to realize the access control of large data-sites. The architecture of privacy preserving techniques of smart grids large data-sets looks like Fig. 1[6].

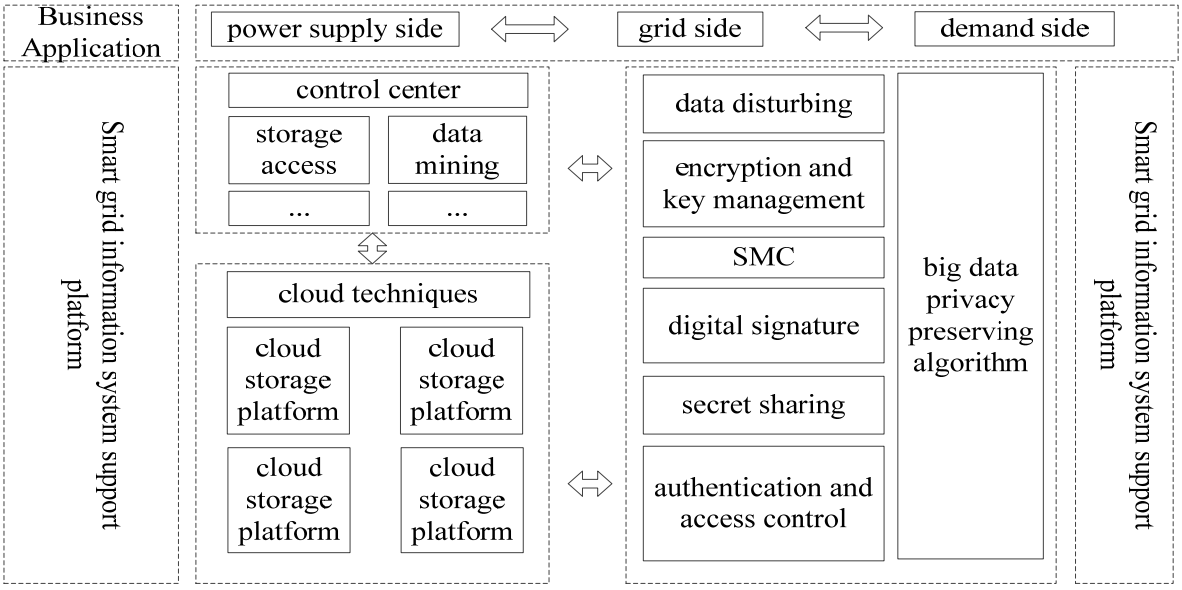

Fig. 1 Privacy preserving techniques of electric power large data-sets

As the data stored in the sites those are physical or logical isolation, these sites in the smart grid cloud platforms are autonomous units, and their data have private attributes. That every site needs to process the clustering model or clustering results together when mining data-sets, which leads to the issues of privacy leaking. The privacy leaking issues only take place in the process of numerical computation and result sharing. As a result, the following issues should be taken into account: (1) preserve privacy of each local sites, keeping the other participants from getting the data that do not bellowing themselves; (2) protect the data security, keeping the data from be stealing by the malicious attackers; (3) preserving the privacy in the process of data clustering, keeping the data privacy from leaking in the step of processing.

The researches of preserving privacy of smart grid data mining are not enough. The existing techniques are only used for clustering. In the paper [7], the method that constructing an optimal relationship tree, which can cover all neighbor nodes of one certain node and its shortest path can reach those not neighbor nodes before clustering step, can reduce data traffic effectively.

Design of Homomorphic Encryption Clustering Algorithm Used for Smart Grid Privacy Preserving

Homomorphic technology is to deal with encrypting plaintext and the encrypted results after being decrypted is the same with the results that come from the step of processing plaintext. The fully homomorphic encryption technology based on the ideal lattice, was proposed by Gray Gentry. The algorithm which has the operation for addition and multiplication- satisfying the Eq.1:

$$e(m) \otimes e(m') = e(m \oplus m') \quad (1)$$

The method can be called fully homomorphic encryption algorithm. Before transmitting, we use this technique to encrypt the data stored in the local stations to preserve privacy issues of the process of mining data in a global station. This paper uses the Homomorphic Encryption Scheme (HES) based on integer to encrypt consumers daily load curves of the smart grid system. First of all, take a number N satisfy this Eq.2:

$$N = P \times Q \tag{2}$$

(P and Q are all prime numbers). The X is plaintext to be encrypted, and the encrypting process is as follows Eq.3:

$$Y = E_p(X) = (X + P \times R) \bmod N \tag{3}$$

R is a random number satisfies uniform distribution in the range (1,Q).After the cipher-text Y being sent to the receiving end, we can get the plaintext by using the key p to decrypt Y. Decrypting method is as follows Eq.4:

$$X = E^{-1}(Y) = D_p(Y) = (Y) \bmod_p \tag{4}$$

The weaknesses steps of privacy preserving are cooperation calculations and intermediate results shared. This paper proposes a homomorphic encryption clustering algorithm used for smart grid privacy preserving. The basic idea is: (1) Local stations compute the clustering data of local stations, then they send this calculated results to the global station after homomorphic encryption; (2) Global station computes the results coming from local stations, in which decrypting the results and sending the results back to the global station; (3) According to the algorithm convergence conditions, global station determine whether to stop iterating. If stop iterating, outputs the clustering results, complete the clustering process; otherwise, continue iterating until convergence condition is satisfied. The implementation of this algorithm is as shown in Fig. 2:

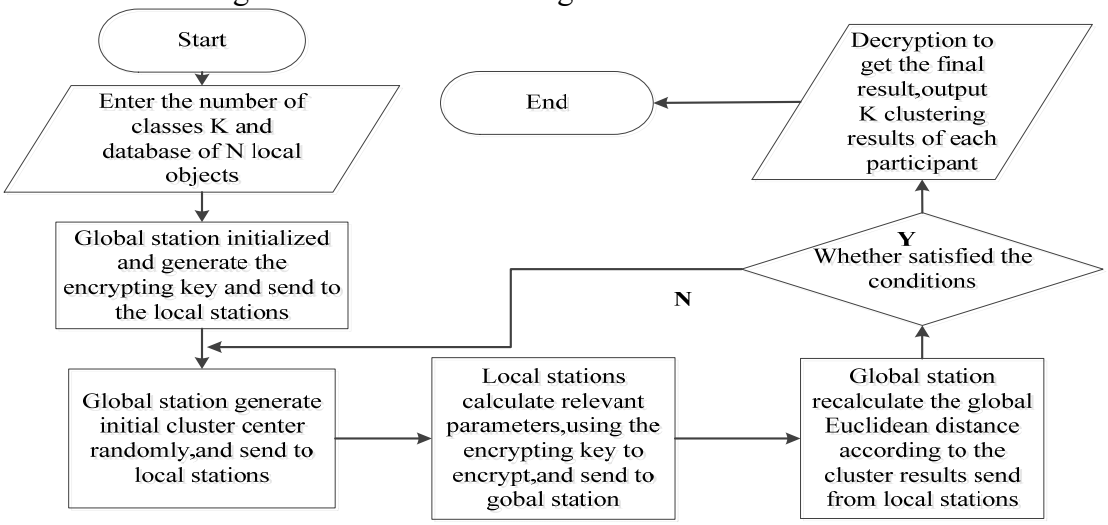

Fig. 2 The flow chart of k-means algorithm for privacy preserving

Experiment and Analysis

Preparing Work and Experiment

The daily load curves come from the sample, which we get from consumer's active power every 15mins in Hangzhou from power quality monitoring network database, and there are 15 kinds of different terminal load types. The user types and main electric power load as Table 1:

Table 1 User types and main electric power load

User type	The main electricity load	No
Colleges and Universities	lighting, air conditioning, laboratory equipment	1
Mobile base stations	communication device	2
staff dormitory	lighting, household appliances	3
Community	air conditioning, lighting, household appliances	4
Chemical Factory	grinding Equipment	5
Hotels	central air conditioning, lighting, elevators	6
Bank	central air conditioners, computers	7
Research Institution	electronic equipment, electric motors	8
Office Buildings	electronic equipment, air conditioning	9
Software Technology Park	computers, elevators, air conditioning	10
Pump Station	pump	11
Entertainment Center	elevator, air Conditioning	12
Powder plant	183m Mill	13
Concrete Factory	sand transport, recycling machine	14
Supermarket	fire-fighting equipment, freight elevators	15

We extract one-week data of each user from the database, and the daily load curves consist of 96 active sampling points. Tag consists of the user serial number and the time of week, in which the first two digits are load ID and next number stand for Monday to Sunday. For analysis, we extract curves from domain and frequency domain as the input and extract time domain feature extraction by using the idea of time-sharing computing, calculating a power value in different times of one day as characterization of the electric power load.

The cloud platform requires multiple parties taking part in, and requires homomorphic encryption and RSA public-key encryption technique. Local site of smart grid cloud computing platform uses homomorphic encryption and public key encryption of RSA to encrypt the result of clustering, sending to the global site of cloud computing platform. The center randomly generates k clustering center. Then the global site sends the results to the local site. Local sites receive the results come from the global site. Local sites computing results from decrypting result. Local sites encrypt them by RSA public key encryption, and then the results are sent to the center site. The process of iteration repeats until each cluster does not change.

Eventually, we get eight categories from 131 load curve in this experiment. Curve shapes in same category are similar. From consideration of periodicity of using electric power, 15 users can be divided into three categories: periodic users, workday users and stochastic users. The daily electric power load of periodic users, belong to the same cluster, are similar. The curves of workday users and weekend users are different. The users belong to the different clusters, including banks, institutions, office buildings and other users. The weekday load curves of stochastic users such as chemical factory, powder factory are divided into different clusters.

Analysis of the Validity and the Privacy

We use RSA public key encryption system and homomorphic encryption system to encrypt the computing results of sites to keep the data of the involved parties from leaking in semi-trusted environment. Since the RSA public key encryption system is only used for the encryption key and the process of encrypting with homomorphic encryption do not have impact on the result of clustering, the algorithm proposed by this paper can get accurate mining results.

The algorithm preserves the privacy in three levels: (1) The algorithm uses homomorphic encryption process encrypts the local clustering results, due to the random number R, so it can be used as a digital envelope to store clustering result of local sites. The global site gets encrypted data, so it cannot get any information of the local sites data; (2) Completing computing process of the local at the central site, the intermediate results are sent to the local site to be decrypt, and then they are sent back to the central site for the next operation. It can keep the global site from decrypting the privacy data from local sites effectively, preserving privacy of the smart grid data; (3) RSA public

key encryption system meets the security requirements, and so every participant can only get outputs and computing results of themselves. They cannot obtain any other data.

By a combination theorem, if a protocol and organizational process of its sub-protocol are semantic security, the protocol is security. All participants involved can ensure data security; the data clustered at the global site are homomorphic encrypted; what is more the plaintext does not appear at the transfer process e transfer process. Hence, the algorithm can preserve the privacy of the process of mining grid data.

Conclusions and Prospect

Based on the distributed environment, this paper proposes a privacy preserving clustering algorithm used in the process of mining large data-sets of smart grid. The proposed algorithm uses the RSA public key encryption system based on semantic security and the encryption system to protect the security of the participating parties. All parties involved use K-means algorithm to calculate the local clustering, and then the results will be encrypted. After that, the global site receives the local clustering results, and can complete the remaining mining work in the cloud. This algorithm can prevent all participants' privacy data leakage without increasing limited time complexity and decreasing the accuracy of mining results. This algorithm is used in sites that are horizontal distribution for mining data, and take the privacy of the intermediate results of communication processes in account. Because the clustering process is carried out in the cipher text, the public key encryption makes the intermediate results of the computation process can be protected, and the algorithm can get accurate clustering results. We set the smart grid data mining environment is semi trust, participants following the cooperative computation protocol are trusted, but data on the other side of the curious and malicious environment may contain malicious attackers. How to preserve the privacy in malicious environment is our next step research.

Acknowledgement

The authors acknowledge support from the Science Technology Project of State Grid Chongqing Electric Power Company (No. 2016Yudiankeji18).

References

[1] Yu. X.Y, and Qin. C, Status of Smart Grid basic concepts in Chinese. J. Sci. SCIENCE CHINA Information Sciences, 44(6), 694-701, (2014).

[2] Song. Y. Q, Zhou. G. L and Zhu. Y. L, Challenges and Status of Large Data Process in Smart Grid in Chinese, J. Power System Technology, 37(04):927-935, (2013).

[3] Lai. J. D, Based on the CSC Permanent Magnet Direct-Drive Wind Power System Coordination Control Method and Strategy Research in Chinese. (Doctoral dissertation, Hefei University of Technology), (2012).

[4] Shen. Y. L, Lu. Y and Chen. R. F, Research of Power User Behavior and Application Status Based on Big Data Techniques in Chinese. J. Sci. Automation of Electric Power Systems, 38(3), (2016).

[5] Chicco. G, Napoli. R, Postolache. P, and Scutariu. M, Customer Characterization Options for Improving the Tariff offer. IEEE Power Engineering Review, 18(1), 381-387.

[6] Wei. S.Q, Ren. H and Yang, Techniques of preserving privacy of smart grid data mining, Journal of Guangxi University (natural science edition) (03), 714-721, (2015).

[7] Li. F, Luo. B and Liu. P, Secure and privacy-preserving information aggregation for smart grids. International Journal of Security & Networks, 6(1), 28-39, (2011).