Research on Cluster Mining Algorithms for Personal Privacy Protection in the Background of Big Data

HUANG Tianming
College of Arts and law, Wuhan Donghu University, P.R.China

Abstract. With the rapid development of cloud computing and mobile Internet technology, huge amounts of data are generated on the network at every moment. In the era of big data, most of the data are designed in distributed storage, stored on many independent sites. The typical data mining algorithm can not meet the requirement of privacy data security in the big data background. In the big data background, distributed data mining is faced with the need to solve data security, privacy disclosure and other technical problems. Therefore, this paper applies homomorphic encryption technology to typical K-means clustering mining algorithm, and designs a distributed privacy protection PP-kmeans algorithm based on big data. Experiments show that this algorithm can realize data privacy security protection. And more accurate big data clustering mining effect is achieved.

1. Introduction
Faced with the massive data generated by the network, the analysis and processing of big data has become the key to follow-up research. At this time, we need to use a variety of typical data mining technology. In the context of big data, data mining, which has attracted much attention, refers to the use of data mining technology to discover implicit and valuable information from big databases with NB level, more noise, diversified storage formats and distributed storage generated by the network [1]. Relevant information extracted from data mining may leak some sensitive information. Therefore, the premise of data mining is to ensure the privacy and security of users. In the era of big data, data is often stored in scattered sites. In the process of coordinated big data mining, data transmission and sharing of intermediate results of data mining are needed, which may lead to hidden dangers of data security and leaks of user privacy [2]. To solve this problem of data security protection, experts have proposed data mining technology based on privacy protection. This technology is to extract valuable information while protecting the privacy of user data, so as to effectively solve the problem of data leakage and privacy protection. With the rapid increase of data nowadays, protecting data privacy and security has become particularly urgent and important. Accordingly, distributed data mining can not only achieve the same effect as centralized data mining, but also protect the data privacy of distributed sites. Therefore, privacy-preserving distributed data mining (PPDDM) has become a research hotspot in the field of data mining in recent years [3]. PPDDM aims to design a multi-party collaborative data mining algorithm based on privacy protection technology. This algorithm can realize the orderly collaborative execution of data mining operations by multiple sites, and also can protect the private information of each site itself, taking into account the accuracy of data mining and the effectiveness of privacy protection. Based on this, this paper intends to apply homomorphic encryption technology to typical K-means clustering mining algorithm, designs and proposes a distributed privacy-preserving PP-kmeans algorithm based on big data. Under the condition of meeting the security and privacy-
preserving requirements of big data mining, Big data mining can be carried out, and more accurate
data mining results can be obtained [4].

2. Related Technology

2.1. Big data mining security technology
Big data mining is to extract and mine knowledge from massive, irregular and diverse data. When
each site carries big data mining tasks in collaboration, the first thing to consider is how to realize data
mining while preventing data privacy leakage at each site. Privacy-based data mining algorithms can
be divided into the following categories: privacy-based association rule data mining algorithm,
privacy-based classification data mining algorithm, privacy-based clustering data mining algorithm,
privacy-based sequential pattern data mining algorithm and so on. Secondly, we need to take targeted
constraints to strengthen the self-restraint management of each site, so as to ensure the smooth
progress of big data mining and at the same time to minimize the occurrence of data privacy leaks. In
recent years, a certain number of academic research results have emerged. Literature [5] explores the
efficiency of data mining algorithms based on privacy protection and privacy protection security of
data mining under the basic conditions of semi-honest model and malicious model. Literature [6] then
explores the sequential pattern data mining technology based on data security and privacy protection,
designs a data mining algorithm based on important sequence attribute hiding, and implements it.
Effective data mining privacy protection; literature [7] studies the design of privacy-preserving data
mining algorithm in distributed environment, which effectively solves the privacy leakage and data
security problems in data mining process.

2.2. Privacy data protection technology
Privacy data involved in the process of big data mining mainly includes personal-related private
information, such as personal basic information, work-related information, personal property and
medical record information, personal social dynamic information, etc. Privacy protection of big data
mining should not only protect the relevant privacy data of each site in the process of data mining, but
also not steal the privacy data of other sites. In addition, it can ensure the overall expected effect of
data mining. In conclusion, the research usually adopts privacy protection technology of data
encryption in typical data mining algorithms, and uses homomorphic encryption technology to encrypt
the original data. When data mining, the encrypted ciphertext is directly processed, this can not only
guarantee the security of private data, but also not affect the effect of data mining. Homomorphic
encryption technology does not decrypt the original data, but directly uses big data mining algorithm
to perform complex computing operations on the encrypted data, and can get the same results as the
original data operation before data encryption. The key research results related to this topic are as
follows: literature [8] explores and analyses the homomorphic encryption technology, and develops and
 designs an improved scheme based on the efficiency of homomorphic encryption algorithm, and
achieves good results. This paper considers that for addition and multiplication, if an encryption
algorithm can find its corresponding homomorphic operation, that is:

\[ E(d) \otimes E(d') = E(d \otimes d') \]  \hspace{1cm} (1)

In this way, it can be called homomorphic encryption algorithm.

3. Distributed K-means Clustering Mining Algorithms for Privacy Protection

3.1. Algorithmic Thought
In the distributed environment, data mining based on big data requires all sites to compute clustering
results jointly, which will lead to data security and privacy leakage. This paper takes the main site as
the center, calculates the clustering results independently from the site, and assists the main site to
jointly construct the operation pattern of big data mining. Data privacy may be violated in the process
of clustering calculation and result sharing and transmission. Therefore, it is necessary to protect the
data and privacy security of each site itself, avoid extracting the privacy data used in clustering mining of other sites, and also pay attention to data security in the process of clustering mining of big data, and cooperate in clustering analysis, calculation and seeking. The data security and data privacy protection of clustering analysis results in sharing and sending each other are obtained. Big data mining supported by distributed environment design has the characteristics of multiple sites participating, multiple sites communicating with each other, and interacting with the main site to transmit the final and intermediate results. Cooperative clustering analysis and calculation, sharing and transmission of intermediate clustering analysis results are the most easily leaked links of data privacy. Therefore, homomorphic encryption is used to encrypt the original data. To select data $A$, data $A$ must satisfy the following equation requirements:

$$A = D_1 \times D_2$$  \hspace{1cm} (2)

Data $D_1$ and $D_2$ are prime numbers.

If $Z$ is the plaintext that needs to be encrypted, the mathematical formula of the whole encryption process can be expressed as follows:

$$Z' = E_k(Z) = (Z + D_1 \times S) \mod A$$  \hspace{1cm} (3)

Among them, $S$ is a random data which satisfies uniform distribution in $(1, D_2)$ data set. After the encrypted cipher text $Z'$ is transmitted to the target site, the corresponding decryption operation is performed using the encryption key $k$, and the corresponding plaintext $X$ after decryption is obtained, that is:

$$X = E^{-1}(Z') = D_k(Z') = (Z') \mod k$$  \hspace{1cm} (4)

In this paper, homomorphic encryption technology is applied to typical K-means clustering mining algorithm, and a distributed privacy-preserving PP-kmeans algorithm based on big data is proposed. The idea of the algorithm can be summarized as follows.

1. Calculate the clustering analysis results of local sites from the site, calculate the distance from each data point to the initial clustering center sent by the main site, expand the initial classification processing, and then send the clustering analysis results to the main site after homomorphic encryption operation.

2. The main site will calculate the global Euclidean distance of the clustering data mining results sent from the site, recalculate the global clustering center points, and then send them to the slave site. After decryption, the clustering analysis data mining results will be obtained and sent back again.

3. According to the convergence condition of the global clustering analysis algorithm, the main site decides whether to stop the cyclic operation or not. If the condition is satisfied, the final clustering data mining results will be output after the decryption operation by the main site; otherwise, the cycle will continue until the exit condition of the loop is satisfied.

The design flow of PP-kmeans algorithm for distributed privacy protection based on big data is shown in Figure 1.
3.2. Algorithm description

Input: Data set $X_i$ for cluster analysis from each site, number of objects $x_i(i = 1, \ldots, n, n \geq 3)$ for each data set $X_i$, initial set of $k$ clustering centers

Output: $k$ categories

Step 1 The main site uses encryption algorithm to generate the encrypted key pairs $(a_i, b_i)$, and sends the values of $a_i$ and $k$ points to each slave site $C_i(i = 1, \ldots, n, n \geq 3)$, receiving the encrypted information from the main site respectively.

Step 2 Loop start

Step 2.1 According to the initial $k$ points of the main site, the Euclidean distance between the $C_i$ data object points and the $k$ points contained in the data set $X_i$ of the slave site is calculated, and each $C_i$ is added to the classification of the $k$ centers nearby according to the distance size.

Step 2.2 Calculate the location of the $k$ center of each slave site and the number of data sets $x_{ij}(i = 1, \ldots, n, j = 1 \ldots, k, n \geq 3)$. Encrypt data sets $e_{ij}$ and $x_{ij}$ by using full homomorphic encryption technology, get the encrypted clustering data mining results, namely $x_{ij}' = A_i(x_{ij})$ and $e_{ij}' = A_i(e_{ij})$, and send the encrypted results $x_{ij}'$ and $e_{ij}'$ back to the main site.

Step 2.3 The master station receives the encrypted clustering analysis result data $e_{ij}'$ and $x_{ij}'$ from each slave station, separately calculates the global $Y = \sum_{i=1, j=1}^{i=3, j=3} x_{ij}' \times e_{ij}'$ and $Z = \sum_{i=1, j=1}^{i=3, j=3} x_{ij}'$, and sends the global clustering analysis result and the new $k$ centers to the slave station after encrypting.

Step 2.4 The $Y'$ and $Z'$ of the encryption processing from the master site are received from the slave site respectively. After the corresponding decryption operation, the $Y = \sum_{i=1, j=1}^{i=3, j=3} e_{ij} \times x_{ij}$ and $Z = \sum_{i=1, j=1}^{i=3, j=3} x_{ij}$ is obtained. The clustering analysis results are encrypted by RSA public key encryption technology and sent back to the master site.

Step 2.5 The main site receives the encrypted data $Y', Z'$, performs the corresponding decryption operation, calculates the global Euclidean distance, and recalculates the joint data clustering center.
According to the results of the global clustering analysis, and calculates
\[ e_j(i = 1, \cdots, k) \]
according to the results of the global clustering analysis, and calculates
\[ \sum_{i=1}^{k} \sum_{j=1}^{n} b_{ij}(x_j, a_i). \]

Until each cluster center does not change, the cycle ends, and the final global cluster analysis results are output at the same time.

4. Analysis of Experimental Results

The experimental environment is as follows: the operating system is Windows 7, the CPU is 2.4 GHz, and the memory is 4 G. The experimental data are set by machine learning data set Adult, which contains 48800 pieces of data, each of which contains 14 attributes. In the experiment, 80% data records were randomly extracted to train the clustering analysis model, and the remaining 20% data records were used to test the accuracy of the clustering analysis model. In the distributed data storage mode, the classification accuracy of the PP-kmeans algorithm based on big data is 86.2%, which is 3.2% higher than that of the traditional K-means algorithm.

The analysis shows that the distributed privacy-preserving PP-kmeans algorithm designed in this paper increases the process of encrypting, decrypting and sending results to each other due to the clustering analysis results between the main site and the slave site, which makes the execution time of the distributed privacy-preserving PP-kmeans algorithm based on big data slightly longer than that of the traditional K-means algorithm, but increases encryption and decryption. The decryption operation is mainly linear operation with high efficiency and little increment of running time consumption. Therefore, the distributed privacy protection algorithm based on big data proposed in this paper can protect data security without affecting the actual effect of data mining.

5. Conclusion

In this paper, a distributed privacy protection algorithm based on big data is designed for distributed storage with horizontal partitioning. In this algorithm, homomorphic encryption technology is applied to typical K-means clustering mining algorithm to realize the data security and privacy protection function of each site itself and when sending data. Each site uses a typical K-means clustering mining algorithm to compute the clustering results of the slave sites, and encrypts the results of clustering analysis by using homomorphic encryption technology. The main site receives the results of clustering analysis after the slave sites are encrypted, and then calculates the global Euclidean distance, adjusts the global clustering center point continuously, and completes the global clustering data mining work. The efficiency analysis and experimental tests show that the algorithm can improve the accuracy of the final clustering analysis data mining results and effectively protect the data security of each site with less computing time.

References

[1] Li Tao, Zeng Chunqiu. Data Mining in the Age of Big Data—Viewing Big Data Mining from the Perspective of Application[J]. Big Data, 2015, 1(4): 57-80.
[2] Feng Dengguo, Zhang Min, Li Hao. Big Data Security and Privacy Protection[J]. Journal of Computer Science, 2014, 37(1): 246-258.
[3] MONA S, HIREN D J. Privacy preserving data mining techniques in a distributed environment[J]. International Journal of Computer Applications in Technology, 2014, 94(6): 21-27.
[4] Zhang Ni, Zhang Yunyong, Hu Kun. Big Data Security Technology and Application[M]. Beijing: People's Posts and Telecommunications Publishing House, 2014.
[5] KANTA R CIOGLU M, KA R DES O. Privacy preserving data mining applications in the malicious model[C]// Seventeenth IEEE International Conference on Data Mining Workshops. Omaha, NE, USA: IEEE, 2017: 717-722.
[6] Zhu Yuquan, Hu Tianhan, Chen Gen. Research on Privacy Protection in Sequential Pattern
Mining[J].Computer Applied Research,2018,26(7) :2489-2491.

[7] Jia Zhe. Research on Information Mining and Privacy Protection in Distributed Environment[D]. Beijing University of Posts and Telecommunications, 2017.

[8] Wu Xiaoyuan. Research on Information Mining and Privacy Protection in Distributed Environment[D]. Xi'an University of Electronic Science and Technology, 2017.