PPHOPCM: Privacy-preserving High-order Possibilistic c-Means Algorithm for Big Data Clustering with Cloud Computing

Shailaja B N
C Byregowda Institute of Technology
Department of Computer science and Engineering
Kolar, India
shylybn1995@gmail.com

Subhashini R
Prof. Department of Computer science and Engineering
C Byregowda Institute of Technology
Department of Computer science and Engineering
Kolar, India

Abstract- **As one essential approach of fuzzy clustering in statistics mining and pattern recognition, the possibilistic c-method set of rules (PCM) has been broadly used in photo evaluation and understanding discovery. However, it is tough for PCM to produce a good result for clustering large records, particularly for heterogenous information, considering that it is first of all designed for best small dependent dataset. To address this trouble, the paper proposes a excessive-order PCM algorithm (HOPCM) for huge records clustering by means of optimizing the goal function within the tensor area. Further, we design a allotted HOPCM approach based on MapReduce for extremely big amounts of heterogeneous statistics. Finally, we devise a privacy-maintaining HOPCM set of rules (PPHOPCM) to protect the non-public statistics on cloud by using making use of the BGV encryption scheme to HOPCM, In PPHOPCM, the capabilities for updating the club matrix and clustering centers are approximated as polynomial capabilities to guide the steady computing of the BGV scheme. Experimental effects suggest that PPHOPCM can efficiently cluster a big quantity of heterogeneous statistics the usage of cloud computing without disclosure of private statistics.**

INTRODUCTION

As personal computing technology and social websites, such as Facebook and Twitter, become increasingly popular, big data is in the explosive growth [1]. Big data are typically heterogeneous, i.e., each object in big data set is multi-modal [2]. Specially, big data sets include various interrelated kinds of objects, such as texts, images and audios, resulting in high heterogeneity in terms of structure form, involving structured data and unstructured data. Moreover, different types of objects carry different information while they are interrelated with each other [3]. For example, a piece of sport video with meta-information uses a large number of subsequent images to display the exercise process and uses some meta-information, such as annotation and surrounding texts, to show additional information which are not displayed in the video, for instance the names of athletes. Although the subsequent images pass on different information from the surrounding texts, they describe the same objects from different perspectives. Furthermore, big data are usually of huge amounts. For example, Facebook, the famous social websites, collects about 500 terabytes (TB) data every day [4]. These features of big data bring a challenging issue to clustering technologies. Clustering is designed to separate objects into several different groups according to special metrics, making the objects with similar features in the same group [5, 6]. Clustering techniques have been successfully applied to knowledge discovery and data engineering [7]. With the increasing popularity of big data, big data clustering is attracting much attention from data engineers and researchers.

THE PROPOSED SYSTEM

A high-order clustering algorithm for big data by using the tensor vector space to model the correlations over the multiple modalities. However, it is difficult for them to cluster big data effectively, especially heterogeneous data, due to the following two reasons.

First, they concatenate the features from different modalities linearly and ignore the complex correlations hidden in the heterogeneous data sets, so they are not able to produce desired results.

Second, they often have a high time complexity, making them only applicable to small data sets. Thus, they cannot cluster large amounts of heterogeneous data efficiently.

To tackle the above problems, this paper proposes a high-order PCM scheme (PPHOPCM) for big data clustering. PCM is one important scheme of fuzzy clustering. PCM can reflect the typicality of each object to different clusters effectively and it is able to avoid the corruption of noise in the clustering process. However, PCM cannot be applied to big data clustering directly since it is initially designed for the small structured dataset. Specially, it cannot capture the complex correlation over multiple modalities of the heterogeneous data object.

The paper proposes a high-order PCM algorithm by extending the conventional PCM algorithm in the tensor space. Tensor is called a multidimensional array in mathematics and it is widely used to represent heterogenous data in big data analysis and mining. In this paper, the proposed HOPCM algorithm represents each object by using a tensor to reveal the correlation over multiple modalities of the heterogeneous data object. To increase the efficiency for clustering big data, we design a distributed HOPCM algorithm based on MapReduce to employ cloud servers to perform the HOPCM algorithm. However, the private data tends to be in disclosure when performing HOPCM on cloud. Take the medical data which is a typical type of big data for example. A large amount of private information such as personal email address and diagnostic data is included in the medical records. The disclosure of the private information will threaten people's lives and property greatly. Therefore, to protect the private data on cloud, we propose a privacy preserving HOPCM scheme by using the BGV technique that is of high efficiency. Unfortunately, BGV does not support the division operations and square root operations that are the necessary computation in the functions for updating the membership matrix and clustering centers in the HOPCM algorithm although it is a fully homomorphic encryption scheme. To tackle this issue, we use the Taylor's theorem to transform these functions to polynomial functions to remove these operations. We conduct the experiments on the two representative big data sets, i.e., NUS-WIDE and SNAE2, to assess the clustering accuracy and efficiency of our algorithms by comparison with three representative possibilistic c-means clustering algorithms, namely HOPCM-15, wPCM and PCM. Results demonstrate that HOPCM outperforms other algorithms in clustering accuracy for big data, especially for heterogeneous data. Furthermore, PPHOPCM can use cloud servers cluster big data efficiently without disclosure of the private data. Therefore, our contributions are summarized as the following three aspects:

- The conventional PCM algorithm cannot cluster heterogeneous data. Aiming at this problem, the paper proposes a high-order PCM algorithm by optimizing the objective function in the high-order tensor sp contributions are summarized as the following three aspects:
- The conventional PCM algorithm cannot cluster heterogeneous data. Aiming at this problem, the paper proposes a high-order PCM algorithm by optimizing the objective function in the high-order tensor space for heterogeneous data clustering.
- To employ cloud servers to improve the clustering efficiency, we design a distributed high-order possibilistic algorithm based on MapReduce.

2.RELATED WORK

This section reviews the related work on the possibilistic c-means algorithm and heterogeneous data clustering methods. As the preliminary, the PCM algorithm is described first, followed by the heterogeneous data clustering methods.

2.1 Possibilistic c-Means Algorithm

The possibilistic c-means algorithm is one of fuzzy clustering schemes. Different from the traditional clustering schemes which assign each object into only one group fuzzy clustering schemes assign each object into multiple groups. Specially, the assignment of each object is typically a distribution over all the groups in the fuzzy clustering.

Given a data set X = {x1, · · · , xn},

PCM is defined as a c×n membership matrix

U = {uij}, with the following objective function:

Jm(U, V) = ∑c i=1 ∑n j=1 u m ij ||xj − vi ||2 + ∑c i=1 ηi ∑n j=1 (1 − uij) m_____________ (1)

where V = {v1, · · · , vc} represents the set of clustering centers, uij denotes the membership of xj belonging to vi . By minimizing Eq. (1), the membership matrix and the clustering centers can be updated by Eq.(2) and Eq.(3)

uij = 1 (1+(d 2 ij /ηi) 1/(m−1)) , ∀i, j______(2)

vi = ∑n j=1 um ∑ ijxj n j=1 um ij ________ (3)

where dij denotes the distance between the j-th object xj and the i-th clustering center vi , and ηi is a scale parameter which can be estimated by using ηi = ∑n j=1 um ij×d 2 ∑ ij n j=1 um ij _________(4)

Typically, the computational complexity of the traditional possibilistic c-means algorithm is dominated by calculating the distance between each object xj and every clustering center vi, which requires o(n × c) for each iteration. So, PCM has a ace for Furthermore, PPHOPCM can use cloud servers cluster big data efficiently without disclosure of the private data. Therefore, PCM is able to avoid the corruption of noise in the clustering big data sets. However, PCM is sensitive to initial parameters and usually produces a coincident clustering result. Aiming at this problem, FPCM and PFCM were proposed by combining PCM and FCM. Xie et al. [5] developed an enhanced PCM algorithm by grouping the data set into one main subset and one assistant subset to avoid the coincident result. In addition, PCM is not robust to the additional parameters. To tackle this problem, Yang et al. proposed an unsupervised PCM scheme to improve the robustness of the conventional PCM algorithm. To cluster non-spherical data sets, some kernel-based possibilistic clustering algorithms have been proposed by mapping the objects of the data set into high order data space [18]. Other PCM variants include weighted PCM algorithm and sample-weighted PFCM algorithm. Although these algorithms can improve the performance of the conventional PCM clustering, they are all limited in the structured data clustering. Therefore, the paper proposes a high-order PCM algorithm to cluster heterogeneous data

2.2 Bigdata clustering

Over the past few years, some algorithms have been proposed for big data clustering, especially for heterogeneous data sets. Early works focused on image-text co-clustering by information fusion [10]. Specially, many algorithms first extracted the image features and the text features separately, and then concatenated them into a single vector. [21] However, these methods are difficult to produce desired clustering results since they cannot capture the complex correlations over the bi-modalities of the objects by concatenating the features in linear way. To tackle this problem, Jiang and Tan proposed two methods based on the vague information and the Fusion ART to learn the visual-textual correlations by measuring the image-text similarities. Most of heterogeneous data clustering schemes are developed depending on graph theory. They usually transform the heterogeneous data clustering task into a graph partitioning problem. The most representative scheme of this type is the bipartite graph partition scheme proposed by Gao [8] for image-text clustering by interpreting the clustering task as a tripartite graph. Afterward, they extended this method for heterogeneous data clustering. The similar work is the isoperimetric co-clustering algorithm proposed by Rege et al. [23]

3.CRYPTANALYSIS OF A HOMOMORPHIC ENCRYPTION SCHEME

Homomorphic encryption permits making specific operations on personal statistics which remains encrypted. While packages including cloud computing require to have a sensible answer, the encryption scheme should be steady. In this article, we element and examine in-depth the homomorphic encryption scheme proposed through Zhou and Wornell in. From the analysis of the encryption scheme. The first attack permits to recover a mystery plaintext message broadcasted to more than one customers. The 2nd assault performs a designated ciphertext key recovery assault and it changed into implemented and established. The ultimate attack is a related chosen plaintext decryption assault

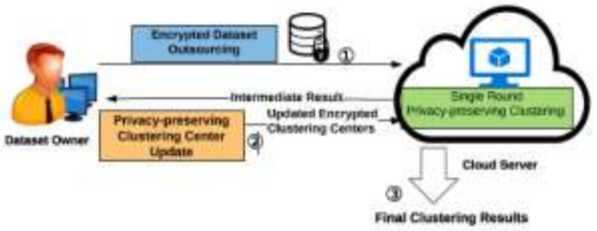
Encrypted Dataset Outsourcing
①
Dataset Owner
Intermediate Result
Privacy-preserving Clustering Center Update
②
Updated Encrypted Clustering Centers
Single Round Privacy-preserving Clustering
Cloud Server
③
Final Clustering Results

Fig. 1. System Architecture

Fig .1:System Architecture

In this work, the cloud server is considered as the cloud server will honestly follow the designed protocol but try to disclose content of the dataset as much as possible. This assumption is consistent with existing works on privacy-preserving outsourcing in cloud computing. Based on available information to the cloud server, we consider the following threat models in terms of the privacy protection of data.

Ciphertext Only Model: The cloud server only has access to all encrypted data objects in the dataset, all encrypted clustering centers, and all intermediate outputs generated by the cloud server. this stronger threat model, the cloud server has all information as in the Ciphertext Only Model. In addition, the cloud server may have some background information about the dataset (e.g., what is the topic of the dataset?), and get a small number of data objects in the dataset. We also consider the cloud server is not able to obtain the clustering centers from background information, since they are generated based on all data objects on the fly during the clustering process.

1.System Setup and Data Encryption;

2) Single Round MapReduce Based Privacy-preserving Clustering

3) Privacypreserving Clustering Center Update.

In Stage 1 fig 2, the owner first setups the system by selecting parameters for PPHOPCM and MapReduce. The owner then generates encryption keys for the system, and encrypts the dataset for clustering. In Stage 2, the cloud server performs a round of clustering and allocates encrypted objects to their closet clustering centers. After that, the cloud server returns a small amount of encrypted information back to the owner as the intermediate outputs.

In Stage 3, the owner updates clustering centers based on information from the cloud server and his/her own secret keys. These new centers are sent to the cloud server in encrypted format for the next round of clustering. Stage 2 and Stage 3 will be iteratively executed until the clustering result does not change any more or the predefined number of iterations is reached.

4.DESIGN AND IMPLEMENTATION OF MODEL

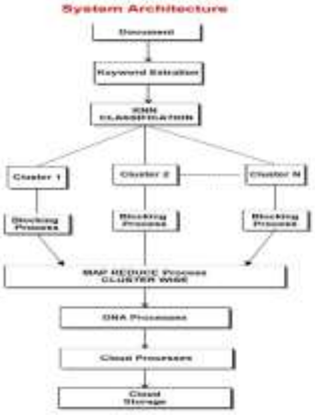

Fig 2: System design and implementation

Document choice here specifically comes within the first component, after selecting a report the principle crucial keywords has to be extracted from that file and after that keywords extraction from that file through KNN classification divides that statistics in to a unique clusters. Cluster evaluation or clustering is the venture of grouping a hard and fast of gadgets in this kind of manner that objects in the equal organization known as a cluster are more comparable in some feel to each aside from to those in different organizations. And then the blockading procedure of records will happens. Map reducing is an innovative generation by means of which we are able to reduce a large space in a larger records units. The MapReduce set of rules contains important duties, namely Map and Reduce. Map takes a set of records and converts it into every other set of statistics, wherein character elements are damaged down into tuple. After the mapreducing technique, Easy Accessibility and Usability. You do not need to be tech-savvy in order to keep your records on line.Recovery of Data is Easy. Every document or information is susceptible to some of failures and setback so ultimately we will save that information in a cloud.

Training Data

Select the cluster category from local system and stored in database based on selected cluster id.

Upload File

Select the file from local system and click to upload option.

Preprocessing

- In this selected file have to remove unnecessary words.
- Comparing file content with trained dataset. If it is matching with trained dataset then increasing the count of category code(Cluster id). Which category code having max count that file is belongs to that category(Cluster).
- User already selected file has to stored in that cluster.
- Select file from cluster table then selected files will get divided into small blocks (500 bytes each block). And each block content will get encryption by using DNA Algorithm (Encryption Key)

eg: packet size=500;
File Size=3000
=3000/500
=6 blocks

Generate hash tag for all block.Compare generated hash block with existing hash tag from database if hash tag matched in that case we will not upload that block into hadoop, we will increase number of instance of that block in database table.If hash tag not matched in that case we will add that block hash details in database and upload that block in hadoop.LBA - Logical Block Addressing technique is used to identify what are the blocks are present in a file.

Download File

Select the file in download list.Get the LBA based on file id Each encrypted blocks has to get decrypt by using DNA algorithm. User has to select a file to download. Using LBA has to find block numbers which are in selected file. Whether all the blocks required for the file is available, if all the blocks are available in Hadoop storage space and download blocks, while downloading itself all the encrypted blocks will get decrypt by using DNA algorithm (Decrypt Key) then merge the blocks and give it to the user.

5.RESULTS AND ANALYSIS

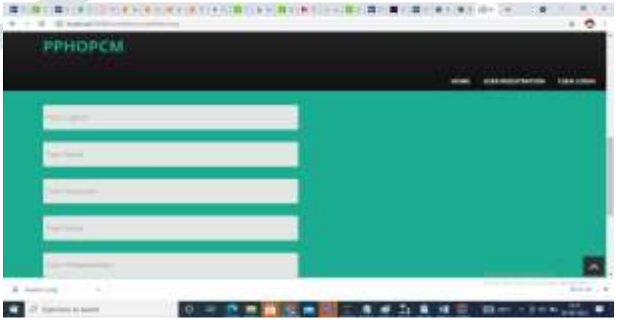

Fig.3: Registration process

User registration,right here is the person registration and enter is the User statistics like call password e-mail identification should be entered and output could be All data ought to be stored into the database and user registered correctly message ought to be seemed and real output as expected and finally end result is skip

Fig.4:login process

In this analysis the login process that is take a look at case for the person login here is the module check.And the input can be user enters his call and password. Expected Output is When user enters his proper username and password , consumer home web page must be redirected Actual Output and observation is bypass.

Fig.5:Upload file

Select the report from nearby device and click on to add option . Expected Output is After person logged in, consumer need to pick out the record and click upload ,accompanied by using document must get uploaded to cloud and correctly uploaded must be induced and actual output as anticipated

Fig.6:Download file

Select the file in download list. Get the LBA based on file identity. Each encrypted blocks has to get decrypt by using DNA algorithm.User has to select a file to download. Using LBA has to find block numbers which are in selected file.Whether all the blocks required for the file is available, if all the blocks are available in Hadoop storage space and download blocks, while downloading itself all the encrypted blocks will get decrypt by using DNA algorithm(Decrypt Key) then merge the blocks and give it to the user.

6.CONCLUSION AND FUTURE ENHANCEMENT

A. Conclusion

Here it proposed a excessive-order PCM scheme for heterogeneous data clustering. Furthermore, cloud servers are employed to improve the efficiency for huge information clustering by using designing a allotted HOPCM scheme depending on MapReduce. Experimental outcomes show PPHOPCM can cluster large facts with the aid of the usage of the cloud computing generation without disclosing privacy. The performance of PPHOPCM and DHOPCM can be similarly progressed whilst using greater cloud servers, making them extra suitable for large facts clustering, since they are of excessive scalability confirmed by way of the experimental effects.

B. Future Enhancements

In this work, the proposed schemes are preliminarily evaluated on two consultant heterogeneous datasets. In the future paintings, the proposed algorithms will be similarly verified on large actual datasets.

REFERENCES

[1] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, ”Data Mining with Big Data,” IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, Jan. 2014.

[2] B. Ermis, E. Acar, and A. T. Cemgil, ”Link Prediction in Heterogeneous Data via Generalized Coupled Tensor Factorization,” Data Mining and Knowledge Discovery, vol. 29, no. 1, pp. 203-236, 2015.

[3] Q. Zhang, L. T. Yang, and Z. Chen, ”Deep Computation Model for Unsupervised Feature Learning on Big Data,” IEEE Transactions on Services Computing, vol. 9, no. 1, pp. 161-171, Jan. 2016.

[4] N. Soni and A. Ganatra, ”MOiD (Multiple Objects Incremental DBSCAN) - A Paradigm Shift in Incremental DBSCAN,” International Journal of Computer Science and Information Security, vol. 14, no. 4, pp. 316-346, 2016.

[5] Z. Xie, S. Wang, and F. L. Chung, ”An Enhanced Possibilistic c-Means Clustering Algorithm EPCM,” Soft Computing, vol. 12, no. 6, pp. 593-611, 2008.

[6] Q. Zhang, C. Zhu, L. T. Yang, Z. Chen, L. Zhao, and P. Li, ”An Incremental CFS Algorithm for Clustering Large Data in Industrial Internet of Things,” IEEE Transactions on Industrial Informatics, 2015. DOI: 10.1109/TII.2017.2684807.

[7] X. Zhang, ”Convex Discriminative Multitask Clustering,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 1, pp. 28-40, Jan. 2015.

[8] B. Gao, T. Liu, T. Qin, X. Zheng, Q. Cheng, and W. Ma, ”Web Image Clustering by Consistent Utilization of Visual Features and Surrounding Texts,” in Proceedings of the 13th Annual ACM International Conference on Multimedia, 2005, 112-121.

[9] Y. Chen, L. Wang, and M. Dong, ”Non-Negative Matrix Factorization for Semisupervised Heterogeneous Data Coclustering,” IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1459-1474, Oct. 2010.

[10] L. Meng, A. Tan, and D. Xu, ”Semi-Supervised Heterogeneous Fusion for Multimedia Data Co-Clustering,” IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 9, pp. 2293-2306, Aug. 2014.