EXTRACTION OF DATA FROM BIGDATA USING MAP REDUCE

Map Reducing Methodology

A.Keerthivasan(BE-CSE)
M.Santhosh kumar(BE-CSE)
V.Vigneshwaren(BE-CSE)

Gnanamani College of Technology

Namakkal(po),India

Santhoshkumar1997be@gmail.com

vishthekingmaker@gmail.com

P.Kamarajapandiyan(Assitant Professor)
Gnanamani College of Technology
Namakkal(po),India

ABSTRACT

One important technique of fuzzy clustering in data mining and pattern recognition isthe Possibilistic c-means algorithm (PCM), has been widely used in image analysis and knowledge discovery. However, it is difficult for PCM to produce a good result for clustering big data, especially for heterogeneous data, since it is initially designed for only small structured dataset. To tackle this problem, the paper proposes a high-order PCM algorithm (HOPCM) for big data clustering by optimizing the objective function in the tensor space. Further, we design a distributed HOPCM method based on MapReduce for very large amounts of heterogeneous data. Finally, we devise a privacy-preserving HOPCM algorithm (PPHOPCM) to protect the private data on cloud by applying the BGV encryption scheme to HOPCM, In PPHOPCM, the functions for updating the membership matrix and clustering centers are approximated as polynomial functions to support the secure computing of the BGV scheme. Experimental results indicate that PPHOPCM can effectively cluster a large number of heterogeneous data using cloud computing without disclosure of private data.

Index Terms—big data clustering, cloud computing, privacy preserving, possibilistic c-means, tensor space.

1.Introduction

Big data is data set that are so voluminous and complex that traditional data processing application software are inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, updating and information privacy. There are three dimensions to big data known as Volume, Variety and Velocity.

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 ofobjects, 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 famoussocial websites, collects about 500 terabytes (TB) data every day [4]. These features of big data bring a challenging issue to clustering technologieseach object is typically a distribution over all the groups in the fuzzy clustering.

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.

2.1Possibilistic 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

2.2Big Data 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 [22] proposed two methods based on the vague information and the Fusion ART to learn the visual-textual correlations by measuring the image-text similarities.

3. HIGH-ORDER POSSIBILISTIC C-MEANS ALGORITHM BASED ON TENSOR REPRESENTATION MODEL

In this part, we present the HOPCM scheme for heterogeneous data clustering based on the tensor data representation model. The tensor data model represents each object by using a tensor [28]. For example, a colorful image can be represented as a 3-order tensor $R^{Iw} \times ^{Ih} \times ^{Ic}$, where I_w,I_h, and I_c denote width, height and color space, representatively. Specially, an image with 560 × 480 in the RGB color space can be represented by $R^{560} \times ^{480} \times ^{3}$. Furthermore, a piece of video with MPEG-4 format can be represented as a 4-order tensor $R^{Iw} \times ^{Ih} \times ^{Ic} \times ^{If}$ with I_f denoting the frames. The tensor model can represent any heterogeneous data object. More importantly, it can capture the complex correlations over the multiple modalities of each heterogeneous data object. The tensor-based representation models have been successfully used in big data analysis and mining in past few years [3, 11, 26]. Therefore, HOPCM extends the conventional possibilistic c-means algorithm using the tensor data representationmodel.

4.Algorithms

Algorithm 1: Secure Computation of the Function for Updatingu_{ij} on Cloud.

```
Input:$C(r), C(s), C(t) and C(d^2_{(T)ij})$
Output: $C(u_{ij})$
Using secure multiplication to compute: ;
$C_1 = C(s) \times C(d^2_{(T)ij} - \alpha)$;
Using secure multiplication to compute: ;
$C_2 = C(t) \times C(d^2_{(T)ij} - \alpha) \times C(d^2_{(T)ij} - \alpha)$;
Using secure addition to compute:
$C(u_{ij}) = C(r) + C_1 + C_2$;
    return $C(u_{ij})$;
```

Algorithm 2: The Privacy-preserving High-order Posssibilistic c-Means Algorithm.

```
Input: $X = \{X_1, X_2, ..., X_N\}$, $c$, $m$, $maxiter$ Output: $U = \{u_{ij}\}, V = \{v_i\}$
    Client: ;
    Randomly initialize the parameters ;
    Encrypt $X$, $m$, and $\eta$
Upload the ciphertexts to the cloud ;
```

Algorithm 3:The High-order Posssibilistic c-Means Algorithm.

```
Input: $X = \{X_1, X_2, ..., X_N\}$, $c$, $m$, $maxiter$
Output: $U = \{u_{ij}\}, V= \{v_i$
for$iteration = 1, 2, ..., maxiter$ d
    for$i = 1, 2, ..., c$ d
        $v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}$;
        $\eta_i = \frac{\sum_{j=1}^{n} u_{ij}^m \times d^2_{(T)ij}}{\sum_{j=1}^{n} u_{ij}^m}$;
    for$i = 1, 2, ..., c$ d
        for$j = 1, 2, ..., n$ d
            $u_{ij} = \frac{1}{(1+(d^2_{(T)ij} / \eta_i)^{1/(m-1)})}$;
```

5.PRIVACY-PRESERVING HIGH-ORDER POSSIBILISTIC C-MEANS ALGORITHM BASED ON BGV

To protect the private data, we devise a privacy-preserving HOPCM scheme (PPHOPCM) based on BGV in this section. The proposed scheme cannot only employ cloud servers to increase the clustering efficiency for large amounts of heterogeneous data, but also avoid the disclosure of the private data. BVG secure operations required for PPHOPCM are described first, followed by the details of the proposed scheme.

5.1BGV Secure Operations

BGV is a leveled fully homomorphic encryption technique. It uses a Setup procedure to select a μ-bit modulus q and the following parameters: the dimension $n = n(\lambda,\mu)$, the degree $d = d(\lambda,\mu)$, the distribution $\chi = \chi(\lambda,\mu)$, and $N = \lceil(2n + 1)\log q\rceil$.

Furthermore, a key Switching procedure and a modulus Switching procedure are implemented in the BGV scheme. The former is used to reduce the dimension of the ciphertext while the latter is aims to reduce the noise.

The BGV technique has four major secure operations, i.e., encryption, decryption, secure addition and secure multiplication, required for implementing our proposed PPHOPCM scheme, listed as follows [15].

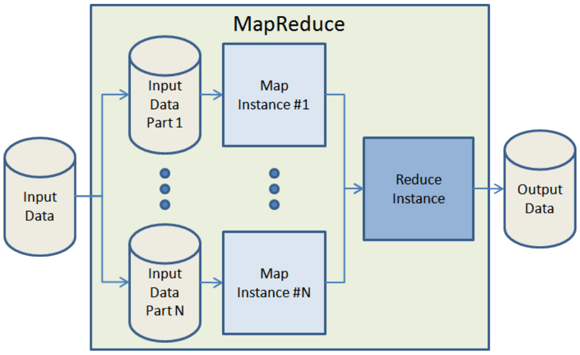
(1) *Encryption*: Encrypt a plaintext $m \in R_2$ as a ciphertext $c \leftarrow m + A^T r \in R_q^{n+1}$.

(2) *Decryption*: Decrpt a ciphertext c to its plaintext $m \leftarrow ((<c,s_j>\bmod q) \bmod 2)$ using the corresponding secret key s_j.

(3) *SecureAddition*: Add two ciphertexts, i.e., c_1 and c_2, to their sum c_4 on cloud by $c_3 \leftarrow c_1 + c_2 \bmod q_j$, and $c_4 \leftarrow$ Re$_{fresh(c_3, \tau(s_j' \rightarrow s_{j-1}), q_j, q_{j-1})}$.

(4) *SecureProduct*: Multiply two ciphertexts, i.e., c_1 and c_2, to their product c_4 on cloud by $c_3 \leftarrow c_1 \otimes c_2 \bmod q_j$, and

$c_4 \leftarrow \mathrm{Re}fresh(c_3, \tau(s_j' \rightarrow s_{j-1}), q_j, q_{j-1})$.



6.Complexity Analysis

Now, we estimate the computation complexity and the communication complexity of the PPHOPCM scheme. We use ADD and MUL to represent the time cost of one addition operation and one multiplication, representatively.

Computation Cost. Assume that the dataset $X = \{x_1,x_2,\ldots,x_k\}$, each represented by a T-order tensor

Communication Cost. Before performing PPHOPCM, the client× uploads×$k \prod^{T}_{t=1} I_t+c+1$ messages, with $(k\prod_{t=1}^{T} I_t+c+1)(n+1)\ \mu$ bits, to the cloud. And then, the client exchanges $c\times(k+1)$ messages, with $c\times(k+1)\times(n+1)\times\mu$ bits, with the cloud in each iteration.

7.EXPERIMENTS

To estimate the clustering accuracy and efficiency of our schemes, we perform the proposed algorithms on the cloud platform including 20 nodes, each with 3.2 GHz Core i7 CPU and 8GB memory. We first compare our HOPCM algorithm with HOPCM-15, wPCM and PCM in clustering accuracy on two representative big data sets, i.e., NUS-WIDE and SNAE2. And then, we evaluate the clustering efficiency of the PPHOPCM algorithm by comparison with HOPCM and DHOPCM. At last, we estimate the scalability of PPHOPCM and DHOPCM based on speedup.

7.1 Data Sets and Evaluation Criteria

Two representative big data sets, i.e., NUS-WIDE and SNAE2, are used to estimate the clustering accuracy and efficiency of our schemes. NUS-WIDE is downloaded from Flikr.com [32]. It consists more than 260,000 images which are grouped by 81 classes. All the images are annotated by some texts, constituting a heterogeneous dataset. To evaluate the robustness of our proposed schemes, we sampled 80, 000 representative images which can be averaged to 8 subsets, each grouped by 14 categories, from NUSWIDE. SNAE2, downloaded from Youtube, includes 1800 pieces of videos, grouped by four classes, i.e., sport, news, advertisement and entertainment. Each video consists 100 frames, represented by a 4-order tensor in our schemes.

7.2Performance Evaluation of HOPCM

The task of this experiment is to evaluate the clustering accuracy of HOPCM in terms of $E*$and *ARI* by comparison with three representative possibilistic clustering algorithms, i.e., HOPCM-15, wPCM, and PCM. HOPCM-15 is proposed by Zhang et al. [11] for heterogeneous data clustering while wPCM is a weighted PCM scheme. Different from our HOPCM scheme, HOPCM-15 learns features from heterogeneous data using improved auto-encoder model before clustering. For wPCM and PCM, we concatenate the attributes of each modality to form a single vector for clustering heterogeneous data.

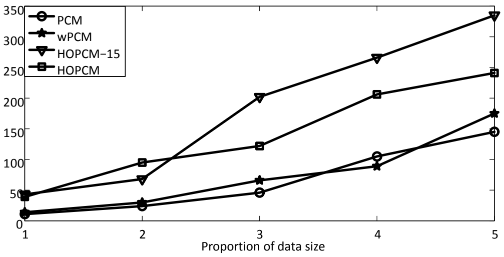


Fig. 2. Running time on the NUS-WIDE dataset

TABLE

Clustering result in terms of $E*$on NUS-WIDE to evaluate HOPCM

Algorithm/dataset	1	2	3	4	5	6	7	8	Overall
PCM	4.52	6.15	5.22	5.47	5.51	4.12	4.29	5.69	5.78
wPCM	4.21	3.96	4.11	5.23	3.85	4.18	4.64	5.25	4.83
HOPCM-15	1.98	2.57	2.91	2.63	2.12	2.91	3.29	2.08	2.93
HOPCM	2.06	2.13	2.26	3.08	2.03	2.67	2.25	2.18	2.72

8.CONCLUSION

In this paper, we proposed a high-order PCM scheme for heterogeneous data clustering. Furthermore, cloud servers are employed to improve the efficiency for big data clustering by designing a distributed HOPCM scheme depending on MapReduce. One property of the paper is to use the BGV technique to develop a privacy-preserving HOPCM algorithm for preserving privacy on cloud. Experimental results show

PPHOPCM can cluster big data by using the cloud computing technology without disclosing privacy.

Reference

1) Qingchen Zhang, Laurence T. Yang, Zhikui Chen, and Peng Li,"PPHOPCM: Privacy-preserving High-order Possibilistic c-Means Algorithm for Big Data Clustering with Cloud Computing" vol.1 , pp.2332-7790, 2016.

2) 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

3) 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.

4) 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.