A Survey on Algorithms in Game Theory in Big Data

D. Rasi and R. Mahaveerakannan

Abstract Conventional data processing application software for handling huge data or intricate data is eased nowadays due to the evolving technology specifically termed as big data. Three Vs, namely, volume, velocity and variety, play a significant role in big data which is a necessity for a particular technology as well as analytical approaches involved in its transformation into value. The major issue involved is that the steady growth of it and opportunities are not caught by the organization frequently as well as extraction of actionable data.

Keywords Big data · High-order PCM algorithm · Cloud

1 Introduction

Big data is the one that is greatly exploited for description of structured as well as unstructured data encompassed in an enormous manner. There are various benefits of big data such as management ease and procession or examination through traditional technologies in addition to tools like relational database, visualization, statistics of data, etc. The petabyte is the one mainly meant because of increasing big data size [1]:

$$1\text{Petabyte} = 1000\text{Terabytes}$$

D. Rasi (✉)
Department of Information Technology, Hindusthan College of Engineering and Technology, Coimbatore, Tamil Nadu, India

R. Mahaveerakannan
Department of Computer Science and Engineering, Saveethan School of Engineering, Thandalam, Chennai, Tamil Nadu, India

A. Ramu et al. (eds.), *International Conference on Computing, Communication, Electrical and Biomedical Systems*, EAI/Springer Innovations in Communication and Computing, https://doi.org/10.1007/978-3-030-86165-0_15

Big Data Characteristics

- *Volume:* The amount of data which is generated by particular online applications in the form of megabytes and gigabytes into petabytes is referred to as volume.
- *Variety:* The nature and the type of data. It has the capability to classify all the incoming data into various categories like structured, unstructured and semi-structured.
- *Velocity:* The speed of data involved in generating term velocity. It accepts the incoming flow of data and at the same time processes it fast to avoid creating bottlenecks.

Big Data Trustworthy

The structured big data is examined through a core platform, namely, Hadoop, which is regarded as a problem solver through utilizing certain convenient data analytical techniques. Hadoop is greatly utilized for scaling a particular server to a thousand of machines [2].

2 Related Work

Privacy-Preserving High-Order Possibilistic C-Means Algorithm for Big Data Clustering with Cloud Computing (PPHOPCM)

[3] utilized the possibilistic c-means (PCM) algorithm for clustering data mining and pattern recognition. Better outcomes are obtained for heterogeneous data involved in big data which is achieved through PCM and regarded as quite difficult. A high-order PCM algorithm (HOPCM) is presented to alleviate this issue which is considered as the main aim designed in tensor space. Privacy-preserving HOPCM algorithm is another vital technique adopted for data protection in cloud through BGV encryption. The PPHOPCM can effectively cluster a large data set by utilizing cloud computing deprived of private data disclosure.

This paper proposed PPHOPCM for big data clustering. It cannot be applied directly because merely small structured data sets might be designed:

1. It is also because HPCM algorithm is nothing but the extension version of PCM algorithm designed in tensor space, where tensor corresponds to a multidimensional array in mathematics utilized for heterogeneous data representation.
2. It also proposes PPHOPCM algorithm for private data protection in cloud through BGV encryption due to its greater efficiency.

Possibilistic C-Means Algorithm

It is regarded as a fuzzy clustering scheme which varies from classical approach in which data set is given by

Data set:

$$X = \{x_1, x_2, \ldots..x_n\}$$

Where:

PCM = c X n membership matrix:

$$U = \left\{u_{ij}\right\}$$

$$J_m\left(u,v\right) = \sum_{i=1}^{c}\sum_{j=1}^{n} U_{ij}{}^{m} \left\| x_j - V_i \right\|^2 + \sum_{i=1}^{c} n_i \sum_{j=1}^{c} n_i \sum_{j=1}^{n} \left(1 - u_{ij}\right)^m \tag{1}$$

Where:

$V = \{v_1, v_2, \ldots\ldots v_n\}$ set of clusters.
V_{ij} = Membership of x_i € v_i

By eliminating Eq. 1, we get

$$U_{ij} = 1 / \left(1 + \left(d^2{}_{ij} / n_i\right)^{1/(m-1)}\right) \tag{2}$$

$$V_i = \frac{\sum_{j=1}^{n} Ux}{\sum_{j=1}^{n} U} \tag{3}$$

Where:

d = Distance between jth object of x_j and ith cluster.
V_i, η_i = scale parameters

$$\eta_i = \sum_{j=1}^{n} u_{ij}^m \times d_{ij}^2 / \sum_{j=1}^{n} u_{ij}^m \tag{4}$$

Distributed High-Order Possibilistic C-Means Algorithm Based on MapReduce

The distributed high-order possibilistic c-means (DHOPCM) algorithm is a notion on the basis of MapReduction and has been considered as an improved version of HOPCM pertaining to efficiency enhancement yielding proficient cloud computing programme model for massive data computing [4]. HOPCM exploits membership matrix as well as clustering centres for calculation purpose. Now, Map function is utilized for membership matrix computation along with Reduce function for clustering computations:

So:

Clustering centres V_i with X_i object:

$$V = \{v_1, v_2, \ldots\ldots v_i\}$$

The sub-matrix is

$$U = \{u_1, u_2, \ldots\ldots u_p\}$$

Data set X:

$$X = \{x_1, x_2, \ldots\ldots x_p\}$$

In MapReduce, we can establish each sub-matrix with their corresponding subsets to one computing node with the parameters (α_i, β_i) that are defined to calculate t computing node as follows:

$$\alpha^{(t)} = \sum_{k=(t-1)}^{tm/p} u$$

In Reduce phrase, the $\alpha_i^{(t)}$ $\beta_i^{(t)}$ where $t = 1,2,\ldots\ldots,P$

Then:

$$V_i = \qquad\qquad ( )$$

Reduce function is mainly deployed for clustering centre computation and dispatched to other computing nodes with another MapReduce function unit of convergence:

$$\text{The time complexity of DHOPCM is specified as O}\left(\text{tnc} / p\right) + \text{O}\left(\text{commu}\right)$$

Where:

P = number of computers engaged for accomplishing HOPCM.
O(commu) = communication overhead.

Hence communication quickly decreases time involved in clustering process estimation explicitly in centralized cloud computing platforms since ignoring can be done easily. Finally, computational complexity of DHOPCM is O(tnc/p).

Privacy-preserving high-order possibilistic c-means algorithm based on BGV
DHOPCM scheme is based on the MapReduce concept which rapidly increases clustering big data efficiency associated with cloud services. Generally, the private data gets affected due to disclosure as soon as it is processed in DHOPCM on cloud. Privacy-preserving HOPCM (PPHOPCM) scheme is introduced for private data production depending on BGV operations. This concept cannot be processed by cloud services, but it avoids private data disclosure. The PPHOPCM requires BGV operations for private data set security [5].

BGV Secure Operations
BGV is regarded as a completely homographic encryption technique. This procedure is used for the selection of a μ-bit modulus q bit and parameters as follows:

Dimensions(n) = $n(\lambda,\mu)$
Degree(d) = $d(\lambda,\mu)$
Distribution(X) = $X(\lambda,\mu)$

$$N = \left[(2n+1)\log 9 \right]$$

This format is mainly exploited for cipher text dimension reduction along with noise [6].

BGV technique comprises four secure operations, i.e. encryption, decryption, secure addition and secure multiplication, which are used for proposed PPHOPCM scheme implementation as follows:

Encryption
It encrypts a plaintext m R as a cipher text:

$$c \leftarrow m + A^{T}{}_{r} \in R_{q}^{\,n+1}.$$

Decryption
It decrypts a cipher text C to its plaintext:

$m \leftarrow (C, \mathrm{Sj} > \mathrm{mod}\ q) \mathrm{mod}\ 2)$ which uses its corresponding secret key Sj

Secure Addition
It adds two cipher texts like C1 and C2 to their sum as C4 on cloud C3 C1 + C2 mod qj and $C_4 \leftarrow (C_3, T(S_j\, S_j - 1) q_j, q_{j\text{-}1})$.

This Technique Has Two Advantages as Follows
It is a fully homographic encryption scheme supported by an arbitrary number of addition and multiplication operations simultaneously [7]. BGV technique yields a greater efficient output compared to other encryption schemes utilized for private data encryption of the large data sets.

Privacy-Preserving High-Order Possibilistic C-Means Algorithm on Cloud
[8] The grouping of X into c clusters is achieved with the help of secure HOPCM algorithm on cloud deprived of disclosure of the private data that is obtained through heterogeneous data set $X = \{\}$ with PHOPCM. The initialization of membership matrix $U = \{Uij\}$ with clustering centres $V = \{V\}$ is attained by PPHOPCM for encryption of membership matrix (μ) object in clustering client [9–11]. Also, updating of matrix and clustering centres is also performed by operating clustering (V) with other parameters on the cloud server by means of PPHOPCM algorithm. It is decrypted, while re-encryption gets updated based on the cloud for iteration. This process repetition is done till the convergence is met. The PPHOPCM can

effectively private data preservation. But data protection is not feasible due to clustering of heterogeneous data sets which are encrypted simultaneously in addition the complete clustering process that is executed on the cipher texts.

Game Theory-Based Correlated Privacy-Preserving Analysis in Big Data [GTCPA]

The greatest challenge involved in big data is privacy preservation. Privacy-preserving data publication (PPDP) is regarded as an extensive application of big data and an significant research field [12]. The trade-off amid privacy as well as utility of the single and independent data set is yet another challenge to be concentrated. This research work concentrates on the investigation of differential privacy parameter selections in correlated network data sets and maximizing each data set utility. There are several other challenges necessitated to be addressed which are as follows:

- Description of correlated relationship amid different data sets pertaining to privacy
- Design of reasonable extent about the utility of a sanitized data set
- Evaluation of data owners' value of privacy

Contributions

- Game model construction of multiple players for releasing their own data sets which is sanitized through anonymization mechanisms as well as measuring the differential privacy relationship of correlated data sets, the utility of sanitized data and the value of privacy.
- On the basis of game model, game analysis is performed.
- Anarchy price is utilized for efficiency assessment of the pure Nash equilibrium.

Game Theory

Game theory is being greatly utilized in data privacy game to analyse users' behaviour [13]. It can analyse competitive situations in a structured way. This theory understands the strategic situations. The basic principle for game theory is to find out an optimal solution.

Differential Privacy

Differential privacy generally refers to a standard meant for privacy definition in addition to rigorous mathematic definition which pertains to sensitivity of a query function. Two standard mechanisms, namely, Laplace mechanism and exponential mechanism, are generally used for attaining differential privacy. Multiple queries are considered together for privacy guarantee degradation for mitigating the privacy composition issues in differential privacy. Sequential composition is how the group of the privacy mechanisms gives differential privacy in isolation. While handling the multiple correlated data sets, privacy mechanism affording differential privacy over a data set might not hold same privacy guarantee. It is necessitated for assessment of relationship amid correlated data sets pertaining to privacy for computation of the real privacy level of a data set. Generally splitting is done for relationship of records about some user in different data sets as follows [14, 15]:

- *Direct Relationship*: This is merely meant for strict definition of relationship of two fully same records. For instance, a user concurrently defers to his tourist information to Facebook and Twitter. As a consequence, two different data sets have one same record about some user.
- *Indirect Relationship*: The direct relationship is more intricate and divided into two different records about some user or his correlated users, for instance, information streams of some user's activity, e.g., GPS record differential privacy and social networks records.

Efficient Trustworthiness Management for Malicious User Detection in Big Data Collection (ETMMUC)

Data collection refers to aggregation of information which plays a vital role in big data, and there exists no guarantee for the data that the users offer. There is no option for the collector to validate the authenticity of every piece of information and the trustworthiness of users participated in the collection which is also considered the most significant. Besides user actions influences on trustworthiness have to be also investigated. Malicious users are also thereby prevented from raising their trustworthiness also given that false information might mislead the final outcomes, a security queue to record users' historical trust information, so that malicious users can be detected with high accuracy. Trustworthiness encompasses two main parts: familiarity and similarity. The computational complexity is alleviated through division of all the participated users into small groups on the basis of similarity, also assessing the trustworthiness of each group separately. The grouping strategy makes the trustworthiness aiding in representation of the trust level of the whole group.

System Model

System model is presented encompassing the theoretical basis for computation of trustworthiness of users and threat model.

Social relationship factor is regarded as a vital part for feature study which encompasses user's social activities pertaining to trustworthiness computation. Their amicability level is determined through the number of interactions amid users; hence it is beneficial for extemporizing this sort of relationship as well as the trustworthiness value delivered through a friend.

Threat Model It is presumed that few users are malicious as well as attacks are launched for compromising data collection. In a sort of attack, malicious users raise their trustworthiness through collaborating with others for sending and receiving an enormous amount of messages. Malicious users affording false information in the collection are considered as another sort of attack. If the data is considered as authentic by the collector, the final outcome will be misinformed and also distant from the real one. The collector has to concentrate on both sorts of attacks for data collection obstruction.

System Design

User Grouping

In data collection, the trustworthiness of users has evaluated through dividing the initial trustworthiness into two segments, i.e. (i) on the basis of familiarity and similarity, trustworthiness will be increased, and (ii) in accordance with the user activities over the data transference, initial trustworthiness will be adjusted.

Trustworthiness Calculation

1. *Initialization phase:* In this phase, users' initial trustworthiness has been attained, and the social relationship between the groups and others has been taken into account, in particular the groups known to be trusted by the collector. Let group G*i* be the friend of a trusted group *Gt* and trustworthiness of group G*i*, which has represented as *Tr*. Then, the estimation has performed through associated transaction amount and similarity. In addition, the trustworthiness of the group, those who are not directly associated with the trusted group, has estimated.
2. *Adjustment phase:* Here, users' trustworthiness has been adjusted, if activities seem to be compromised and malicious. In data transmission, trustworthiness has been considered for adjustment according to the performance of the groups. Consequently, it might be reduced by the abnormal performance of the groups (such as long delay in the data transmission process, massive packet repetition, etc.), during which the trustworthiness of the groups has been adjusted using the adjustment strategy. In the data collection, the trustworthiness of the contributed groups may gradually get revised, since the adversaries have the chance to convince the groups with authentic users.

Initial Trustworthiness

The estimation of users' initial trustworthiness has been carried out, and the trustworthiness of familiarity and similarity has been evaluated.

Familiarity Trustworthiness

Familiarity has been considered as the conveying factor of trustworthiness. For the attainment of familiarity amid G*l*, *a* and G*T*, the groups have included those who possess direct transaction with G*T* and the ones with indirect transactions. The familiarity between group G*I* and G*T* has indicated by the lower level of G*I* than the remaining groups, and it signifies its superior trustworthiness. The estimation of trustworthiness is increased by familiarity.

Malicious User Resistance

In order to conclude the reliability of the user-provided data, the data collector has facilitated by the trustworthiness estimation. Throughout the process, all the users are not malicious, if they are marked by lesser trustworthiness. In that case, the collector takes place to figure out.

Security Queue

In order to identify the abnormal trustworthiness, constructing the security queue has considered to be an efficient way. During the process, on the basis of familiarity and similarity of the users, the groups' initial trustworthiness has evaluated in

data collection. A security queue has been developed to register the historical data of the groups, and the abnormal trustworthiness has been identified, concerning the identification of malicious users. Alongside the constant size, the construction of the security queue has been processed, intending to register the trustworthiness information of group G*I*. The statement has registered the general information regarding G*I*s.

Abnormal Transaction Resistance

In the groups, post-identification of abnormal trustworthiness, the undesirable impacts created by the malicious users should be controlled, for which two scenarios have been conferred: (i) The malicious users are collectively started transmitting enormous messages within them, concerning the increment of their trustworthiness, during which the augmentation of trustworthiness needs to be controlled, which have driven by abnormal activities. (ii) The false data has intentionally been transmitted by malicious users for creating the outcome of the inauthentic collection in which the malicious users have been cumulated, during which the removal of inauthentic data needs to be identified. The sudden augmentation of trustworthiness has been effectively restricted by making the modifications on the abnormal transaction, through which the authentic trustworthiness estimation has safeguarded. In such scenarios, the performance of some users might have been wrongly marked as abnormal that leads them to reduced trustworthiness, which solely occurs during the estimation of trustworthiness while modifying it as per the user performance. In that case, the users can approach the collector by sending the request, since then the respective user's data that has been considered as abnormal will not impact their trustworthiness for the time being. Concurrently, the scenario of the concerned users will be individually considered and elaborately scrutinized by the collector to ensure their reliability. In accordance with the conclusion made by the collector, the respective users' trustworthiness will be increased or may be diminished. Besides, when the reliability of the users was unable to be identified by the collector, and the data has been claimed by the user as normal, the trustworthiness has mutually been marked by an average value, subsequently designated for additional observation.

3 Performance Evaluation

The assessment on the proposed trustworthiness management has been carried out and approached by contrasting it with the inclusion of data transmission trustworthiness, transmission delay impact on trustworthiness and handling of the malicious activities of users.

Advantages

- The high-order possibilistic c-means algorithm outcomes are clustering accuracy for big data, particularly a heterogeneous data.
- The clusters of big data from cloud services can be capably utilized by PPHOPCM, in which the disclosure of private data has been omitted.

Table 1 Comparison chart

Parameter	PPHOPCM (1)	GTCPA (2)	ETMMUC (3)
Objective	The comprehensive trustworthy data collection approach sensor cloud system to extend the data processing ability of WSNs	For defining the association within various data sets concerning privacy	For improving the recognition of malicious users and their unusual behaviour through security queue
Proposed model	PPHOPCM uses multiple mobile sinks to upload the data from wireless sensor networks to cloud	For multi-players' game model for releasing their data sets, where each has been sanitized through anonymization mechanism, we measure the differential privacy correlation of data sets	Malicious users enhance their trustworthiness by colliding with others through transmitting and receiving a huge quantity of messages
Merits	It evaluates the trustworthiness of sensors and mobile networks to conduct extensive simulations to evaluate the performance	Pure Nash equilibrium efficiency has been assessed through the price of anarchy	It is used to prevent group trustworthiness from abnormally growing by malicious users from various groups colliding collectively
Demerits	In cloud-based system, the quality of sensor cloud system is increased, and it causes physical layer attacks like node capture, DoS, integrity, etc.	Identifying the highly critical data publisher who greatly influences his/her neighbours' privacy level	It is a psychology-aware method that could be integrated to the estimation of trustworthiness in order to elevate its level

- This paper proposed the algorithm of HOPCM, during which the objective function of higher-order tensor space has been optimized for the clustering of heterogeneous data.
- The clustering efficiency has been enhanced by designing HOPCM algorithm based on MapReduce to utilize the cloud server.

Disadvantages

Due to the proficiency of psychology, it could be integrated to the estimation of trustworthiness in order to enhance it. There is a possible lack of efficient identification of malicious detection when receiving false information from users. To surpass this issue, designing a few approaches has been necessitated in Table 1. The clustering of big data with heterogeneous data becomes a little difficult to make it efficient which is caused by the following reason:

- To procure the optimal outcomes, the features from various modulate have linearly been concatenated, and the multifaceted association within heterogeneous data sets has been disregarded.

These are eligible to be applied with small data sets. However, in heterogeneous data, it shows its inability to efficiently process the clustering of large data sets due to its extensive time complexity.

4 Conclusion

In this paper, a higher-order privacy level from a few data set that relies on the privacy parameters of its own and the neighbours has been proposed. At the time when the privacy parameters have been considered by their respective data publisher to make the best use of the utility, the trade-off problem is getting transformed into a game problem. In order to assess the impact of the remaining data sets over the real privacy level of a single data set, the correlated differential privacy has been referred. Subsequently, the game model has been developed for multi-players, in which the differential privacy sanitizes the data published by each player. In addition, the adequate criterion of the pure Nash equilibrium's presence and the distinctiveness has been revealed, and its efficiency has been determined through the price of anarchy. Eventually, the comprehensive trails have been carried out to demonstrate the exactness of the proposed game analysis. Further study will be emphasized on the identification of the highly critical data publisher who greatly influences his/her neighbours' privacy level on the basis of game analysis, in the future. Following that, a few significant trials can be done, concerning the enhancement of utility in the game.

References

1. Wu, X., Zhu, X., Wu, G.Q., Ding, W.: Data mining with Big Data. IEEE Trans. Knowl. Data Eng. **26**(1), 97–107 (2014)
2. Ermis, B., Acar, E., Cemgil, A.T.: Link prediction in heterogeneous data via generalized coupled tensor factorization. Data Min. Knowl. Disc. **29**(1), 203–236 (2015)
3. Zhang, Q., Yang, L.T., Chen, Z.: Deep computation model for unsupervised feature learning on Big Data. IEEE Trans. Serv. Comput. **9**(1), 161–171 (2016)
4. Soni, N., Ganatra, A.: MOiD (multiple objects incremental DBSCAN)-a paradigm shift in incremental DBSCAN. Int. J. Comput. Sci. Inf. Secur. **14**(4), 316–346 (2016)
5. Xie, Z., Wang, S., Chung, F.L.: An enhanced possibilistic c-means clustering algorithm EPCM. Soft. Comput. **12**(6), 593–611 (2008)
6. Kamver, S., Schlosser, M., Molina, H.: Eigenrep reputation managementin p2p networks, ISWC 04 workshop on trust, security, and reputation on the Semantic Web, 2003
7. Song, S., Hwang, K., Zhou, R., Kwok, Y.: Trusted p2p transaction with fuzzy reputation aggregation. IEEE Int. Comput. **9**, 24–34 (2005)
8. Devikanniga, D.: Diagnosis of osteoporosis using intelligence of optimized extreme learning machine with improved artificial algae algorithm. Int. J. Intell. Netw. **1**, 43–51 (2020)
9. Bao, F., Chen, I.: Trust management for the internet of things and its application to service composition, IEEE conference on world of wireless, mobile and multimedia networks, pp. 1–6, 2012

10. Tran, T., Rahman, M., Bhuiyan, M., Kubota, A., Kiyomoto, S., Omote, K.: Optimizing share size in efficient and robust secret sharing scheme for big data. IEEE Trans. Big Data. **7**, 703 (2017)
11. Yu, S.: Big privacy: challenges and opportunities of privacy study in the age of big data. IEEE Access. **4**, 2751–2763 (2016)
12. Zakerzadeh, H., Aggarwal, C.C., Barker, K.: Privacy-preserving big data publishing, Proceedings of international conference on scientific and statistical database management, 26:1–26:11, 2015
13. Zhang, X., Leckie, C., Dou, W., Chen, J., Ramamohanarao, K., Salcic, Z.: Scalable local-recoding anonymization using locality sensitive hashing for big data privacy preservation, Proceedings of ACM international on conference on information and knowledge management, pp. 1793–1802, 2016
14. Fung, B.C.M.: Privacy-preserving data publishing: a survey of recent developments. ACM Comput. Surv. **42**(4), 14:1–14:53 (2010)
15. Srinivasa Rao, D., Berlin Hency, V.: Performance evaluation of congestion aware transmission opportunity scheduling scheme for 802.11 wireless LANs. Int. J. Intell. Netw. **2**, 34–41 (2021)