www.ijiarec.com

ISSN:2348-2079

Volume-6 Issue-2

International Journal of Intellectual Advancements and Research in Engineering Computations

Secure cloud data storage as a service with minimum cost in cloud environment

S. Menaka, Mrs. N.Radha M.E.

M.E CSE, Department of CSE, Mahendra Engineering College

Assistant Professor, Department of CSE, Mahendra Engineering College

ABSTRACT

Privacy-preserving High order Possibility c-Means Algorithm that allows support for all users to conveniently access data over the cloud and control and detect the inside threat attack. Data owner is not able to control all over their data and security issues. The new security issues of Insider Threat Attack Various techniques are available to support user privacy and secure data sharing and detect of control the Insider Threat attack. An insider threat was the misuse of information through malicious intent, accidents or malware. The study also examined four best practices companies could follow to implement a secure strategy, such as business partnerships, prioritizing initiatives, controlling access, and implementing technology. This paper focus on various schemes to deal with secure data sharing such as Data sharing with forward security, secure data sharing for dynamic groups, Attribute based data sharing, encrypted data sharing and Shared Authority Based Privacy-preserving High-order Possibilistic c-Means Algorithm for access control of outsourced data. In this paper improve the could security issue's and inside threat attack.

Keywords: PPHOP, Data integrity, Authentication, Security

INTRODUCTION

Cloud computing as long as seemingly unrestricted virtualized resources to users because services across the whole Internet, while hiding platform and implementation de-tails. Today's cloud service providers offer both highly available storage and massively parallel computing resources at relatively

Low costs. Since cloud computing turn into prevalent, a rising quantity of data is being stored in the cloud and shared by users with specified privileges, which define the access rights of the stored data. One critical challenge of cloud storage services is the management of the ever-increasing volume of data.

Cloud storage service is used by many web applications, such as online social networks and web portals, to provide services to clients all over the world. In the web applications, data access delay and availability are critical, which affect cloud customers' incomes. Experiments at the Amazon portal [4] demonstrated that a small increase of 100ms in webpage load time significantly reduces user satisfaction, and degrades sales by one percent. For a request of data retrieval in the web presentation process, the typical latency budget inside a storage system is only 50-100ms [5]. In order to reduce data access latency, the data requested by clients needs to be handled by datacenters near the clients, which requires worldwide distribution of data replicas. Also, data replication between datacenters enhances data availability since it avoids a high risk of service failures due to datacenter failure, which may be caused by disasters or power shortages.

However, a single CSP may not have datacenters in all locations needed by a worldwide web application. Besides, using a single CSP may introduce a data storage vendor locking problem

Author for correspondence:
M.E CSE, Department of CSE, Mahendra Engineering College.

[6], in which a customer may not be free to switch to the optimal vendor due to prohibitively high switching costs. Storage providers charge clients for bandwidth (Transfer), data requests (Get/Put), and Storage. Thus, a client moving from one CSP to another pays for Transfer cost twice, in addition to the Storage cost. The clients are vulnerable to price hikes by vendors, and will not be able to freely move to new and better options. The quickly evolving cloud storage marketplace may leave a customer trapped with an obsolete provider later. The vendor lock-in problem can be addressed by allocating data to datacenters belonging to different CSPs. Building such redistributed cloud storage is faced with a challenge: how to allocate data to worldwide datacenters to satisfy application SLO (service level objective) requirements including both data retrieval latency and availability? The data allocation in this paper means the allocation of both data storage and Get requests to data centers.

To make data management scalable in cloud computing, reduplication [17] has been a well-known technique and has attracted more and more attention recently. Data reduplications is a dedicated data density technique for eliminating duplicate copies of repeat data in storage. The method is used to get better storage operation and can also be applied to network data transfers to reduce the number of bytes that must be sent. Instead of keeping multiple data copies with the same content, reduplications eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy

RELATED WORK

k-NN Classification over Semantically Secure Encrypted Relational Data

Recently, the cloud computing paradigm is revolutionizing the organizations' way of operating their data particularly in the way they store, access and process data. As an emerging computing paradigm, cloud computing attracts many organizations to consider seriously regarding cloud potential in terms of its cost-efficiency, flexibility, and offload of administrative overhead. Most often, organizations delegate their computational operations in addition to their data to the cloud. Despite tremendous advantages that the cloud offers, privacy and security issues in the cloud are preventing companies to utilize those advantages. When data are highly sensitive, the data need to be encrypted before outsourcing to the cloud. However, when data are encrypted, irrespective of the underlying encryption scheme, performing any data mining tasks becomes very challenging without ever decrypting the data. There are other privacy concerns, demonstrated by the following example.

Suppose an insurance company outsourced its encrypted customers database and relevant data mining tasks to a cloud. When an agent from the company wants to determine the risk level of a potential new customer, the agent can use a classification method to determine the risk level of the customer. First, the agent needs to generate a data record q for the customer containing certain personal information of the customer, e.g., credit score, age, marital status, etc. Then this record can be sent to the cloud, and the cloud will compute the class label for q. Nevertheless, since q contains sensitive information, to protect the customer's privacy, q should be encrypted before sending it to the cloud.

The above example shows that data mining over encrypted data (denoted by DMED) on a cloud also needs to protect a user's record when the record is a part of a data mining process. Moreover, cloud can also derive useful and sensitive information about the actual data items by observing the data access patterns even if the data are encrypted [2], [3]. Therefore, the privacy/security requirements of the DMED problem on a cloud are threefold: (1) confidentiality of the encrypted data, (2) confidentiality of a user's query record, and (3) hiding data access patterns.

Existing work on privacy-preserving data mining (PPDM) (either perturbation or secure multi-party computation (SMC) based approach) cannot solve the DMED problem. Perturbed data do not possess semantic security, so data perturbation techniques cannot be used to encrypt highly sensitive data. Also the perturbed data do not produce very accurate data mining results. Secure multi-party computation based approach assumes data are distributed and not encrypted at each participating party. In addition, many intermediate computations are performed based on non-encrypted data. As a result, in this paper, we proposed novel methods to effectively solve the

DMED problem assuming that the encrypted data are outsourced to a cloud. Specifically, we focus on the classification problem since it is one of the most common data mining tasks. Because each classification technique has their own advantage, to be concrete, this paper concentrates on executing the k- nearest neighbor classification method over encrypted data in the cloud computing environment.

LITERATURE REVIEW

Privacy-Preserving Data Mining

Arrayal and Srikant, Lindell and Pinkas were the first to introduce the notion of privacy-preserving under data mining applications. The existing PPDM techniques can broadly be classified into two categories: (i) data perturbation and (ii) data distribution. Agrawal and Srikant [10] proposed the first data perturbation technique to build a decision-tree classifier, and many other methods were proposed later. However, as mentioned earlier in Section 1, data perturbation techniques cannot be applicable for semantically secure encrypted data. Also, they do not produce accurate data mining results due to the addition of statistical noises to the data. On the other hand, Lindell and Pinkas proposed the first decision tree classifier under the two-party setting assuming the data were distributed between them. Since then much work has been published using SMC techniques. We claim that the PPkNNproblem cannot be solved using the data distribution techniques since the data in our case is encrypted and not distributed in plaintext among multiple parties. For the same reasons, we also do not consider secure k-NN methods in which the data are distributed between two parties.

Query Processing over Encrypted Data

Various techniques related to query processing over encrypted data have been proposed, however, we observe that PPkNN is a more complex problem than the execution of simple kNN queries over encrypted data. For one, the intermediate k-nearest neighbors in the classification process should not be disclosed to the cloud or any users. We emphasize that the recent method in reveals the k-nearest neighbors to the user. Second, even if we know the k-nearest neighbors, it is still very difficult to find the majority class label among these neighbors since they are encrypted at the first place to prevent the cloud from learning sensitive information.

Third, the existing work did no address the access pattern issue which is a crucial privacy requirement from the user's perspective. In our most recent work, we proposed a novel secure k-nearest neighbor query protocol over encrypted data that protects data confidentiality, user's query privacy, and hides data access patterns. However, as mentioned above, PPkNN is a more complex problem and it cannot be solved directly using the existing secure k-nearest neighbor techniques over encrypted data. Therefore, in this paper, we extend our previous work in and provide a new solution to the PPkNN classifier problem over encrypted data.

Limitation of Drawback

- ✓ Classification is one of the commonly used tasks in data mining applications.
- ✓ To the rise of various privacy issues, many theoretical and practical solutions to the classification
- ✓ Problem.
- ✓ We focus on solving the classification problem over encrypted data.
- ✓ The proposed protocol protects the confidentiality of data, privacy of user's input query, and hides the data access pattern.

Privacy-Preserving and Outsourced Multi-User k-Means Clustering

Clustering is one of the commonly used tasks in various data mining applications. Briefly, clustering is the unsupervised classification of data items (or feature vectors) into groups (or clusters) such that similar data items reside in the same group. It has immense importance in various fields, including information retrieval, machine learning, pattern recognition, image analysis, and text mining. Some real-life applications related to clustering include categorizing results returned by a search engine in response to a user's query, grouping persons into categories based on their DNA information, etc.

In general, if the data involved in clustering belongs to a single entity (hereafter referred to as a user), then it can be done in a trivial fashion. However, in some cases, multiple users, such as companies, governmental agencies, and health care

organizations, each holding a dataset, may want to collaboratively perform clustering task on their combined data and share the clustering results. Due to privacy concerns, users may not be willing to share their data with the other users and thus the distributed clustering task1 should be done in a privacy-preserving manner. This problem, referred to as privacy-preserving distributed clustering (PPDC), can be best explained by the following example: • Consider two health agencies (e.g., the U.S. CDC and the public health agency of Canada) each holding a dataset containing the disease patterns and clinical outcomes of their patients. Since both the agencies have their own data collecting methods, suppose that they want to cluster their combined datasets and identify interesting clusters that would enable directions for better disease control mechanisms.

METHODOLOGY

This paper proposes a privacy preserving high-order PCM scheme (PPHOP) for clustering. PCM is one important scheme of 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 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 heterogeneous data in 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 boost the effectiveness for cluster big data, we design a distributed HOPCM algorithm based on Map Reduce 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 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.

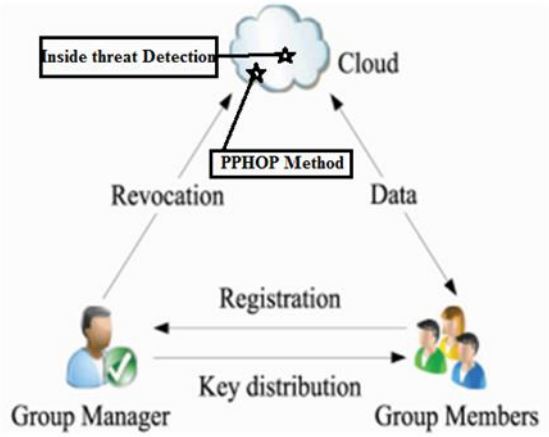

Fig.2 Architecture

RESULT AND DISCUSSION

In this section, we evaluate the performance of the high-order c-means algorithm (PPHOP) in terms of clustering accuracy and execution time by comparison with HOPCM and SHDC. SHDC is proposed. for heterogeneous data clustering. PPHOP is proposed in our previous work. Different from our HOFCM and HOPCM scheme, SHDC maps high-order features to low dimensional spaces by utilizing the Clique Expansion and it uses the k-

means algorithm to cluster the processed heterogeneous objects.

Since PPHOFCM obtains higher ARI values than PPHOP. This is because PPHOP often produces some coincident clusters. PPHOFCM obtains the lower values of ARI than PPHOP due to the approximation of the membership matrix updating function by using the Taylor function and the multivariable Taylor function, respectively. Specially, PPHOFCM reduces approximately 2.5% accuracy drops than PPHOP for the two datasets. The accuracy drops are so slight that it is allowed by massive heterogeneous data in Internet of Things. Although the approximation of the membership matrix updating function incurs slight accuracy drops, PPHOP can improve the clustering efficiency significantly by employing the cloud computing without the disclosure of the original heterogeneous data.

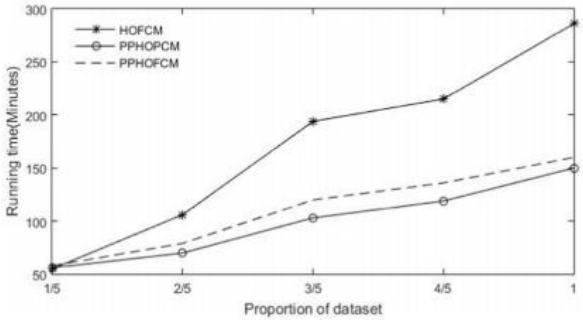

Fig .1 Running time

CONCLUSION

We proposed a high-order c-means algorithm to cluster heterogeneous data in IoT. Furthermore, we devise a privacy-preserving high-order technique to enhance the efficiency by employing the cloud computing. BGV is used to protect the private data when performing the high-order c-means on cloud. The idea of this paper is motivated by our previous work of PPHOP for multimedia data. However, there are at least two differences between PPHOP and PPHOFCM. First, PPHOP utilized tensor distance, instead of Euclidean metric, to capture the correlations of heterogeneous data. Second, PPHOFM utilizes the multivariable Taylor technique to transform the membership matrix updating function to a polynomial function to remove the division and exponentiation operations that are not supported while PPHOP uses the Taylor technique. Experiments implied that PPHOFCM outperforms PPHOP for clustering heterogeneous data in IoT.

REFERENCES

[1]. A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, M. Ayyash, Internet of things: a survey on enabling technologies, protocols, and applications, IEEE Commun. Surv. Tutor. 17 (4) (2015) 2347–2376.

[2]. Y. Sun, H. Song, A.J. Jara, R. Bie, Internet of things and big data analytics for smart and connected communities, IEEE Access 4, 2016, 766–773.

[3]. R. Zuech, T.M. Khoshgoftaar, R. Wald, Intrusion detection and big heterogeneous data: a survey, J. Big Data 2(1), 2015, 1–41.

[4]. X. Qiu, Y. Qiu, G. Feng, P. Li, A sparse fuzzy c-means algorithm based on sparse clustering framework, Neurocomputing 157, 2015, 290–295.

[5]. L. Meng, A.H. Tan, D. Xu, Semi-supervised heterogeneous fusion for multimedia data co-clustering, IEEE Trans. Knowl. Data Eng. 26(9), 2014, 2293–2306.

[6]. Y. Chen, L. Wang, M. Dong, Non-negative matrix factorization for semisupervised heterogeneous data coclustering, IEEE Trans. Knowl. Data Eng. 22(10), 2010, 1459–1474.

[7]. R. Bekkerman, M. Sahami, E. Learned-Miller, Combinatorial Markov random fields, in: Proceedings of the Seventeenth European Conference on Machine Learning (ECML), 2012, 30–41.
[8]. Z. Brakerski, C. Gentry, V. Vaikuntanathan, (Leveled) fully homomorphic encryption without bootstrapping, in: Proceedings of the Innovations in Theoretical Computer Science Conference, 2012, 309–325.
[9]. S. Chen, F. Wang, C. Zhang., Simultaneous heterogeneous data clustering based on higher order relationships, in: Proceedings of the IEEE International Conference on Data Mining, 2007, 387–392.
[10]. Q. Zhang, L.T. Yang, Z. Chen, F. Xia, A high-order possibilistic-means algorithm for clustering incomplete multimedia data, IEEE Syst. J. PP 99, 2015, 1–10.
[11]. B.A. Pimentel, R.M.C.R. de Souza, Multivariate fuzzy c-means algorithms with weighting, Neurocomputing 174, 2016, 946–965.
[12]. T.C. Havens, J.C. Bezdek, C. Leckie, L.O. Hall, M. Palaniswami, Fuzzy c-means algorithms for very large data, IEEE Trans. Fuzzy Syst. 20(6), 2012, 1130–1146.
[13]. N.R. Pal, K. Pal, J.M. Keller, J.C. Bezdek, A possibilistic fuzzy c-means clustering algorithm, IEEE Trans. Fuzzy Syst. 13(4), 2005, 517–530.
[14]. X. Y.Wang, J. Bu, A fast and robust image segmentation using FCM with spatial information, Digit. Signal Process. 20(4), 2010, 1173–1182.
[15]. Q. Zhang, L. T. Yang, and Z. Chen, "Privacy Preserving Deep Computation Model on Cloud for Big Data Feature Learning," IEEE Transactions on Computers, 65(5), 2016, 1351-1362.
[16]. N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A Possibilistic Fuzzy c-Means Clustering Algorithm, "IEEE Transactions on Fuzzy Systems, 13(4), 2005, 517-530.
[17]. M. Yang and C. Lai, "A Robust Automatic Merging Possibilistic Clustering Method," IEEE Transactions on Fuzzy Systems, 19(1), 2011, 26-41.