Privacy Preserving Clustering using Fully Homomorphic Encryption

Urvi Narang
M.Tech Student
Dept. of Computer Engineering
NIT, Surat
narang.urvi@gmail.com

Sankita Patel
Assistant Professor
Dept. of Computer Engineering
NIT, Surat
sankitapatel@gmail.com

Dr. D. C. Jinwala
Professor
Dept. of Computer Engineering
NIT, Surat
dcjinwala@gmail.com

ABSTRACT
In this paper, we focus on the privacy preserving scheme for distributed K-Means clustering. Various techniques have been suggested in the literature for privacy preserving distributed clustering which are either cryptography based or non-cryptography based. In the non-cryptography based techniques, there is a trade-off between privacy and accuracy. Whereas the cryptography based techniques provide higher level of privacy without loss of accuracy. However, existing cryptography based techniques are based on the Yao's Garbled circuit which incurs very high computational and communicational overheads and hence not scalable. In this paper, we attempt to reduce the computational and communication cost of existing techniques and propose a new cryptography based technique that is based on the Gentry's Fully homomorphic encryption scheme. We discuss the theoretical analysis of our proposed technique on a vertically partitioned dataset. Our technique can be applied to any data mining task that requires additive and multiplicative homomorphic operations as its basic building block.

Categories and Subject Descriptors
K.6.5 [**Management of Computing and Information System**]: Security and Protection.

General Terms
Security.

Keywords
Fully Homomorphic Encryption, Privacy Preservation, Clustering.

1. INTRODUCTION
With the advancement in the network, storage and data collection technology, organizations collect, store and share huge amounts of data about their customers, their transactions, financial records etc. This data forms a wealthy resource of information if intelligently processed. These days, data is usually distributed across various sources. Processing of this data calls for collecting the data from all the sources to a central location and then perform the mining operations on this wholesome data. But this jeopardizes the privacy aspect of the data which is not acceptable due to personal and legal concerns. Hence some measures are suggested in the literature which embeds privacy preserving mechanisms in the data mining tools.

The approaches suggested for Privacy Preserving Clustering can be classified either as Randomization based approaches (i.e., Non-Cryptographic) viz. Data perturbation, Fuzzy logic etc. or Cryptography based approaches viz. Vector Sums with SMC, Yao's Garbled Circuit.

In the non-cryptography based approaches, there is a trade-off between accuracy and privacy while in the Yao's approach; there is a big communication overhead because the circuit generated cannot be reused.

Table: 1 Literature Survey

Non-cryptography based Techniques	Data perturbation [3,4]	Trade-off between accuracy and privacy.
	Global Dissimilarity Matrix [6,7]	
	Fuzzy Logic [2,8,9]	
	Generative Models [37]	
Cryptography based Techniques	Vector Sums with SMC [5]	High Computational and communicational overheads.
	Yao's Garbled circuit [10,11,12]	

Hence there is a need to investigate an approach which provides: higher level of privacy; better accuracy; low communicational and computational overheads.

Moreover the cryptographic schemes suggested in literature are additively homomorphic schemes and cannot be applied to data mining tasks involving both additive and multiplicative homomorphic operations. Thus, there is a need to investigate a scheme which is both additively and multiplicatively homomorphic.

Therefore, in this work we propose a novel approach to Privacy Preserving Clustering using Fully Homomorphic Encryption scheme as it doesn't have the privacy-accuracy trade-off, and also, once the public key has been transferred it can be reused numerous times [14]. This reduces the computational and communicational costs to a great extent.

2. PROBLEM FORMULATION AND METHODOLOGY
Our problem setup involves a co-operative setup of vertically partitioned databases where each record has m attributes distributed among the m parties. Assuming the total number of data records to be n, these records need to be clustered into k clusters where k, m and n are input parameters.

Consider $\{x_1, x_2, \ldots x_m\}$ are the m attributes of each of the n objects, such that party i has attribute x_i of all the n objects. The centers for the k clusters are randomly picked up from these n objects and are known to all the parties.

Now the main concern is that, the data held by each of the parties is sensitive and needs to be kept private but at the same time the objects need to be clustered. Thus, we propose a scheme wherein the clustering operation will be performed on encrypted data thus ensuring the secrecy of the individuals' data and at the same time also achieving the clustering results. We use Gentry's Fully Homomorphic encryption scheme [36] for working on the encrypted data.

We use the K-means Clustering algorithm [1] as our base algorithm. In that we propose the distributed computation of the clusters in which all the parties collaborate in finding the closest centre for each object (since each party holds one attribute of all objects and can compute only partial distances).

2.1 Proposed Algorithm

The main aim of our algorithm is to efficiently and securely compute the distance between each object and all the cluster centers; and to assign the object to the cluster with minimum distance. The algorithm for finding the closest center for each object is as follows:

The PKI body generates the public and private keys. The Public key is distributed to all the parties while the Private key is sent to the Trusted Third Party (TTP). One party is randomly selected from all the parties known as Selected Party (SP) who securely computes the distances of the objects from the cluster centers. SP requests the parties to send encrypted partial distance (for the attribute held by the party). The parties send the encrypted partial distance X_{ij} to SP. X_{ij} is the partial distance computed for the i^{th} attribute of object X from the cluster j. SP then homomorphically adds the X_{ij}'s for each cluster j and compares the X_j of each cluster and sends the encrypted comparison result to the TTP.

The TTP decrypts the comparison result and sends back to SP. This is an iterative process until the X_j's of all parties are compared to get the minimum X_j for each object. SP then assigns the object to the cluster with minimum X_j and announces it to all parties. Parties re-compute the new center C for each cluster based on the object assignment and compares if the old C and new C are same and sends back the results to SP. If they are same, SP declares the new C as the output else it restarts the whole process.

2.2 Cost Analysis

Assuming the number of iterations required for convergence to be *I*, the computation and communication complexities are defined in Table.2.

Table: 2 Computation and Communication Complexities

	Computation Costs	**Communication Costs**
Parties	$O(Iknm * Homo.Enc.Cost)$	Ikn messages sent to SP
SP	$O(Iknm * Homo.Enc.Cost + Iknm * Homo.Add\ Cost + Ik\ Homo.Comparisons)$	Ikn messages sent to TTP + In messages broadcast
TTP	$O(Ik)$ Decryptions	Ikn messages sent to SP

3. CONCLUSION

In this paper, we have discussed an approach for Privacy Preserving Clustering based on Fully Homomorphic Encryption [FHE] that achieves better accuracy as compared to the non-cryptographic schemes and low computation and communication cost as compared to the Yao's Garbled Circuit based techniques. This is due to the fact that in FHE, once the public key has been transferred it can be reused numerous times.

Our scheme can be applied to any data mining approach that requires additive and multiplicative homomorphism.

4. REFERENCES

[1] S.R.M. Oliveira and O.R. Zaiane, Privacy Preserving Clustering by Data Transformation, Proc. 18th Brazilian Symp. Databases, pp.304-318, Oct. 2003.

[2] Kalita M.,; Bhattacharyya D.K., and Dutta M., "Privacy Preserving Clustering-A Hybrid Approach," ADCOM 2008. 16th International Conference on , pp.123,130, 14-17 Dec. 2008.

[3] Ali I., Selim V.K., Yu C.S., Erkay S. and Ayca A.H., "Privacy preserving clustering on horizontally partitioned data," Data & Knowledge Engineering, vol. 63, no. 3, pp. 646-666, April 2007.

[4] Thomas B.P., Erkay S., Yucel S., and Selim V.K., "Efficient Privacy Preserving Distributed Clustering Based on Secret Sharing," in Emerging Technologies in Knowledge Discovery and Data Mining.: Springer, 2007, pp. 280-291.

[5] Merugu S., and Ghosh J., "Privacy-preserving distributed clustering using generative models," Data Mining, 2003. ICDM 2003. Third IEEE International Conference on , pp.211-218.

[6] G.Manikandan, V.Vaithiyanathan, and B. Karthikeyan, "A Fuzzy Based Approach For Privacy Preserving Clustering," Journal of Theoretical and Applied Information Technology, vol. 32, no. 2, pp. 118 - 122, October 2011.

[7] K.S. Rani, and M.N. Lakshmi, "A Privacy Preserving Clustering Method Based On Fuzzy Approach And Random Rotation Perturbation," International Journal Publications of Problems and Applications in Engineering Research, vol. 4, no. 1, pp. 174-177, 2013.

[8] K.S. Rani, and M.N. lakshmi, "Privacy Preserving Clustering Based on Fuzzy Data Transformation Methods," IJARCSSE, vol. 3, no. 8, pp. 1027-1033, August 2013.

[9] Yu T.K., Lee D.T., Chang S.M., and Zhan J., "Multi-Party k-Means Clustering with Privacy Consideration," in Parallel and Distributed Processing with Applications (ISPA), 2010 International Symposium on.: IEEE, 2010, pp. 200-207.

[10] Clifton C., and Vaidya J., "Privacy-preserving k-means clustering over vertically partitioned data," in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining.: ACM, 2003, pp. 206-215.

[11] Geetha J., Krishnan P., and R. N. Wright, "A New Privacy-Preserving Distributed k-Clustering Algorithm.," in *SIAM International Conference on Data Mining (SDM)*, 2006.

[12] Geetha J., Pillaipakkamnatt K., Wright R. N. ,Umano D., "Communication-Efficient Privacy-Preserving Clustering.," Transactions on Data Privacy, vol. 3, no. 1, pp. 1-25, 2010.

[13] Yao A.C., "How to generate and exchange secrets," in Foundations of Computer Science, 1986., 27th Annual Symposium on, Yao A. C., Ed.: IEEE, 1986, pp. 162-167.

[14] Mohammed G.K., Russell P., and Xun Y.I., "Fully homomorphic encryption based two-party association rule mining," Data & Knowledge Engineering , vol. 76-78, no. 0, pp. 1-15, 2012.

[15] M.V. Dijk, C. Gentry, S. Halevi, and V. Vaikuntanathan, "Fully homomorphic encryption over the integers," in Advances in Cryptology--EUROCRYPT 2010: Springer, 2010, pp. 24-43.

[16] Kamber,Pei ,Han, *Data Mining: Concepts and Techniques, 3rd Edition*. San Francisco: Morgan Kaufmann, July 2011