International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 4, Issue 2, 2017, PP 14-19
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
DOI: http://dx.doi.org/10.20431/2349-4859.0402003
www.arcjournals.org

Privacy Preserving Data Mining Technique and Their Implementation

V. Prasanthi, Dr. V. Ananta Krishna

Department of Computer Science and Engineering, SriDevi Women Engineering College,
Vattinagulapally, Hyderabad, India

Abstract: *Privacy preservation in data mining has gained significant recognition because of the increased concerns to ensure privacy of sensitive information. It enables multiple parties to conduct collaborative data mining while preserving the privacy of their data. In this work, a cloud computing based protocol for privacy-preserving distributed K-means clustering over horizontally partitioned data, shared between N parties, is proposed. Clustering is one of the elementary algorithms used in the field of data mining. Traditional cryptographic methods use encryption techniques or secure multiparty computation (SMC) to ensure privacy of data. But privacy in these techniques is at the expense of additional communication cost, which limits their use in practical applications. Hence, to reduce these overheads, threshold cryptography is used in the proposed work as a privacy-preserving mechanism. The proposed scheme is faster as compared to the previous schemes and experimental results presented in this paper.*

Keywords: *Privacy Preserving, Data Mining, K-Means, Secure Multiparty Computation (SMC), Threshold cryptography.*

1. Introduction

Advances in computer networks and data acquisition techniques have enabled the collection and storage of huge amounts of data. This data is of no use until it is analyzed and converted into useful information, which is done in data mining. Data mining can help in predicting future trends and behaviors, which will aid the businesses to make knowledge-driven decisions. The first step in exploratory data mining is clustering the data according to some criterion. One of the most extensively used techniques for clustering in statistical data analysis is K-means clustering. Generally, multiple organizations collaborate to get more accurate patterns in their data. The sharing of private data can raise privacy concerns and so it must be kept as a secret from other organizations. As a solution to the privacy issues in distributed data-mining, privacy-preserving data-mining was introduced by Agarwal et al.[1] and Lindell and Pinkas [2]. Privacy-preserving distributed data-mining is the cooperative computation of data that is distributed among multiple parties without revealing any of their private data items. These schemes are primarily categorized into those using Data Perturbation or Randomization techniques and those using Cryptographic techniques.

The first category takes the data and replaces it from the same distribution sample or from the distribution itself, i.e., probability distribution approach, or by adding noise, i.e., value distortion approach. The latter category makes use of cryptographic protocols. The solution should not just be secure, i.e., it leaks, no additional useful information, but should also minimize the additional overheads in terms of communication and computation costs required to introduce privacy. The schemes using randomization techniques achieve partial privacy, but the advantage is that the communication costs are negligible. The schemes using cryptographic techniques,like Homomorphic Encryption and SMC, provide complete privacy but are computationally expensive. Hence, it is required to explore other techniques that have lesser overheads.

The work is focused on giving a privacy-preserving distributed K-means clustering algorithm which uses secret sharing paradigm to provide privacy. Methods using the secret sharing paradigm have very less communication overhead as compared to the traditional cryptographic techniques and hence are faster. A secret sharing scheme allows the splitting of a secret s into different pieces, which are called shares. These shares are distributed among a set of players or participants, denoted by P, such that only certain subsets of players can recover the secret using their respective shares.

These subsets are known as qualified or authorized subsets. The collection of such qualified subsets of participants is called the access structures, corresponding to the secret s. The scenario where N parties wish to collaborate for doing cluster analysis on their combined data is addressed. The clustering is done over the horizontally partitioned data, shared among N parties. The proposed solution is based on cloud computing and utilizes the services of R computational servers, where R > 2. In the proposed scheme, the collaborating users use Vandermonde's matrix based secret sharing to compute secret shares of their private data. They then send the shares securely to the servers for processing. The cloud of employed servers run the K-means algorithm over the secret shares.

B. Motivation and Contribution

The privacy-preserving distributed k-means clustering using secret sharing scheme proposed by Upmanyu et al.[3] has too much of storage overhead. The space required for shares at each of the R computational servers is equal to that of the data, i.e., total storage is R times the actual data.

This data expansion is important since it increases the cost overheads in terms of both storage and interaction among the servers. For applications such as data mining, it becomes even more critical as data mining deals with large volumes of data.

The proposed work aims at decreasing the storage as well as computation overheads. It also aims at improving the clustering results, i.e., quality of the clustered data or group cohesion and the number of iterations required to get the final clustering assignment.B**.**

2. Privacy Preserving Data Mining

2.1. Data-mining

Data-mining is the extraction of patterns and relationships by analyzing the data. This can help in predicting future trends and behaviors, which will aid the businesses to make knowledge-driven decisions. Generally multiple organizations come together to get more accurate relationships or patterns in their data.

These schemes are primarily categorized into those using Data Perturbation techniques and those using Secure Multiparty Computation (SMC). The first category takes the data and replaces it from the same distribution sample or from the distribution itself, i.e., probability distribution approach, or by adding noise, i.e., value distortion approach. These techniques achieve partial privacy, while the latter category makes use of cryptographic protocols to achieve complete privacy.

2.2. Clustering

The fitrst step in exploratory data mining is clustering the data according to some criterion. Exploratory data mining focuses on discovering new features in data. Clustering helps to find the "natural" grouping of unlabeled data. This partitioning leads to the formation of meaningful sub-groups or sub-classes that are called clusters. Clustering is used in many fields like machine learning, search engines, pattern recognition, image analysis, bio-informatics, anomaly detection, etc.

2.3. K-Means Clustering

One of the most extensively used techniques for statistical data analysis is k-means clustering. The basic algorithm of k-means clustering is

1. Pick the number of clusters required (k),
2. Assign initial cluster centers or seeds randomly,
3. Assign every data item to its nearest cluster center using some distance formula,
4. Calculate new cluster center as the mean of assigned items,
5. Repeat steps 3 and 4 until the termination criterion is met.

2.4. Upmanyu et al.'s scheme

The data that is to be clustered can be thought of as points in a D-dimensional Cartesian space. The data should be bounded, and should be scale invariant, i.e., scaling the axis won't affect the cluster assignment. To adopt this scheme, these two properties should be satisfied by the data. It is a three step protocol: 1) users compute the secret shares of their data using shatter function and securely send their secret shares over to the servers, 2) clustering is carried out over the shares by the servers, and 3) the users reconstruct the cluster assignment and the cluster centers using the merge function.

3. VADLNA.B[17] PRIVACY PRESERVING

The proposed work is the modification of the scheme proposed by Upmanyu et al.[3]. The (integer) data to be clustered is distributed among N users. The clustering is done over horizontally partitioned data.

Phase 1. Secret Share Generation Phase: In this phase, each of the N users computes the secret shares of his/her private data and then sends those shares over to the R servers. The secret sharing scheme based on Vandermonde's matrix is used for the computation of the shares. The order of Vandermonde's matrix used is DxR, and it is known only to the N users. At the end of phase one, the servers store the shares of users' secret data. As the servers only store the shares of the data, no information about the entities is

revealed. Let Xi be an entity made up of D attributes, then secret shares of data are computed as follows:

Algorithm 1 Share Generation

```
1: For each entity $X_l$ in the distributed database, l = 1 to L do
2: Compute $X_l^1$ = $[X_l]$1xD [V ]DxR mod p
3: Send $X_{lj}$ , j = 1 to R, to server j
4. End or
```

Phase 2. K-means Clustering phase: This phase deals with the K-means clustering algorithm on the aggregate of the shares available with the R servers.

Step 1. Seed Initialization: In this step, K entities, chosen from the database in a pseudo-random manner, are assigned as the K initial cluster centers. Each Ck is a R dimensional vector. The choice of seeds affect the final cluster assignment.

2. Seed Intialization

```
1. For each cluster $k, k$ =1 to $k$ do
2. Choose a random entity index $l$ , so that 1_$l$_ $L$
3. To initialize $C_k = X_l^{'}$ do
4. For each server $j, j = 1$ to $R$ do
5. Since $X_{lj}^{'}$ is the share corresponding to the entity $l$ with server $j$ ,initialize $C_{kj} = X_{lj}^{'}$, where $C_{kj}$ is the share of $C_k$ with server $j$ .
6. End for
End for
```

Step 2.1 Clustering: Cluster Assignment: In this step, each entity is assigned to the closest cluster center. The distance between each entity and each of the cluster centers is computed using Euclidean distance. It uses secure addition protocol for this. To find the minimum distance i.e., the closest cluster center, secure comparison protocol is used.

Algorithm3: Cluster Assignment

```
1. For each entity $X_l^{'}, l = 1$ to $L$ do
2. For each cluster $k, k = 1$ to $K$ do
```

```
3.  For each server $j, j = 1$ to $R$ do
4.  Calculate the distance between $X'_{lj}$ and $C_{kj}$ as
$D_{lkj} = (C_{kj} = X'_{lj})^2$
End for
```

Secure computation

```
1.  Compute $D_{lk} = \sum_{j=1}^{R} D_{lkj}$ using secure addition protocol
2.  End for
3.  Find the minimum $D_{lk}$ using secure comparison protocol
4.  Let the closest cluster be $C_k$ assign $X'_l$ to $k^{th}$ cluster
5.  End for
```

Step 2.2 Clustering: Update Cluster Centers: In this step, weighted mean of all the entities, in each cluster, is computed. The entity closest to the computed mean is chosen as the new cluster center. Then the reassignment of entities is done using the previous algorithm.

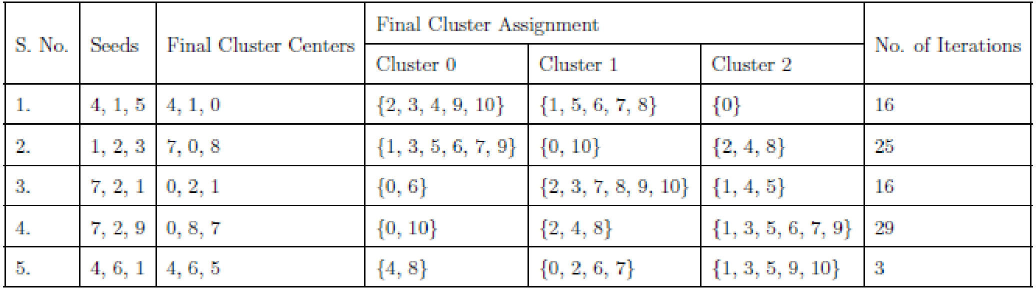
4. Experiment Results

In this experiment, we have taken sample data set int X[L][D]={{1,10,1,96,11,12,22,21,45,30},

{1,2,3,4,11,12,2,15,2,8},{45,67,45,34,11,3,9,12,1,13},{23,24,25,26,11,4,23,2,7,12},{12,23,34,45,66,33,5,2,1,5},{23,45,67,78,10,30,6,25,2,21},{12,34,57,78,45,55,7,30,1,9},{90,70,60,50,9,4,8,1,2,2},{23,45,67,89,23,34,9,19,1,4},{22,65,10,32,2,8,12,23,23,12},{2,34,11,10,1,3,23,24,24,21},{1,2,3,4,5,6,23,45,67,89},{10,20,30,40,50,60,35,67,89,34},{12,14,15,16,17,18,12,33,44,56},{98,65,43,32,21,33,23,4,56,67},{20,34,26,48,59,32,20,9,12,11},{23,43,21,45,54,67,68,52,3,50},{90,34,20,30,10,50,40,60,80,56},{67,56,67,78,89,90,80,70,58,48},{23,45,56,67,78,3,6,9,3,4}};

Implemented upmanyu and v.baby methods using Java and compared the results.The following table represents the existing method proposed by Upmanyu et.al [3] implemented with above data set.

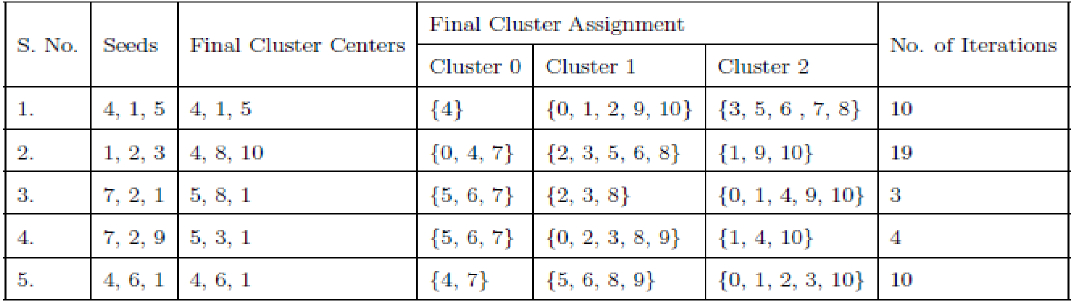
Table 1: Upmanyu et al.s' Scheme



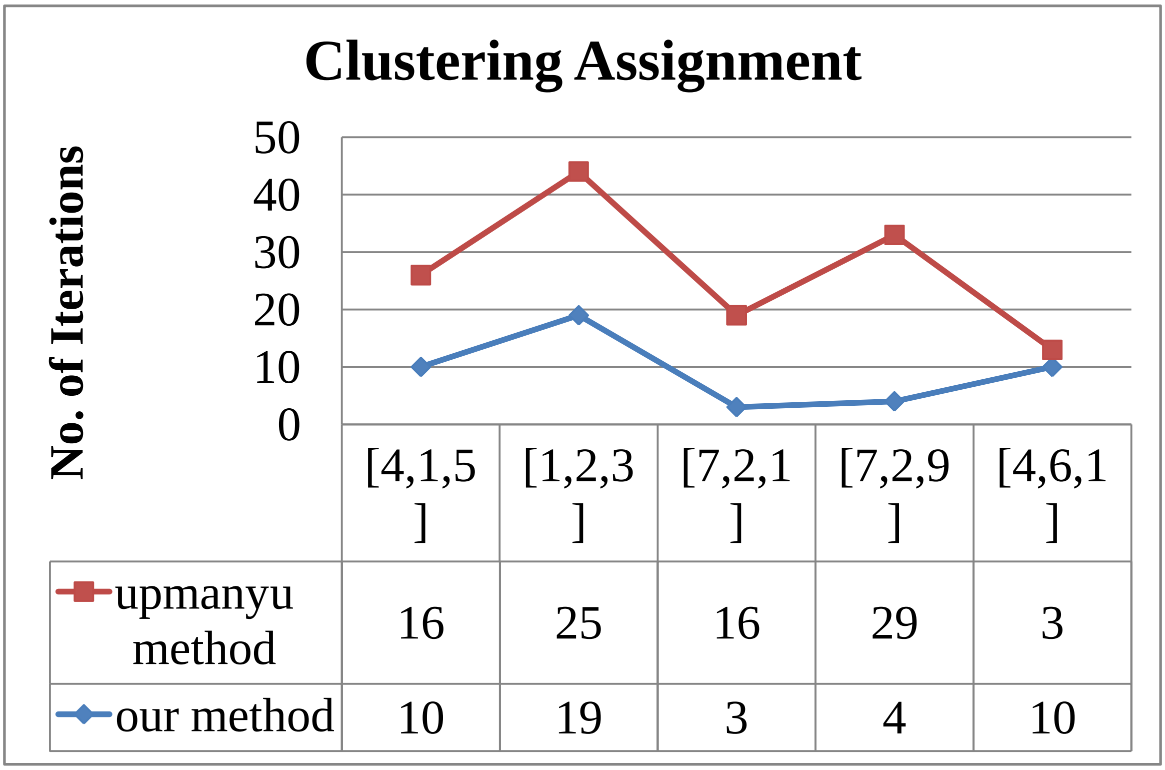
S. No.	Seeds	Final Cluster Centers	Final Cluster Assignment			No. of Iterations
			Cluster 0	Cluster 1	Cluster 2	
1.	4, 1, 5	4, 1, 0	{2, 3, 4, 9, 10}	{1, 5, 6, 7, 8}	{0}	16
2.	1, 2, 3	7, 0, 8	{1, 3, 5, 6, 7, 9}	{0, 10}	{2, 4, 8}	25
3.	7, 2, 1	0, 2, 1	{0, 6}	{2, 3, 7, 8, 9, 10}	{1, 4, 5}	16
4.	7, 2, 9	0, 8, 7	{0, 10}	{2, 4, 8}	{1, 3, 5, 6, 7, 9}	29
5.	4, 6, 1	4, 6, 5	{4, 8}	{0, 2, 6, 7}	{1, 3, 5, 9, 10}	3

The following table represents the our proposed method implemented with above data set.

Table 2: Proposed Scheme



S. No.	Seeds	Final Cluster Centers	Final Cluster Assignment			No. of Iterations
			Cluster 0	Cluster 1	Cluster 2	
1.	4, 1, 5	4, 1, 5	{4}	{0, 1, 2, 9, 10}	{3, 5, 6 , 7, 8}	10
2.	1, 2, 3	4, 8, 10	{0, 4, 7}	{2, 3, 5, 6, 8}	{1, 9, 10}	19
3.	7, 2, 1	5, 8, 1	{5, 6, 7}	{2, 3, 8}	{0, 1, 4, 9, 10}	3
4.	7, 2, 9	5, 3, 1	{5, 6, 7}	{0, 2, 3, 8, 9}	{1, 4, 10}	4
5.	4, 6, 1	4, 6, 1	{4, 7}	{5, 6, 8, 9}	{0, 1, 2, 3, 10}	10



	[4,1,5]	[1,2,3]	[7,2,1]	[7,2,9]	[4,6,1]
■ upmanyu method	16	25	16	29	3
◆ our method	10	19	3	4	10

FIG2. *COMPARSION OF EXISTING METHOD*

5. CONLCUSION

A cloud computing based solution for privacy-preserving distributed K-means clustering is proposed which makes use secret sharing paradigm to do clustering of horizontally partitioned data among N users. In the proposed scheme, the storage required is very less as compared to the scheme proposed in [3]. Also the communication overheads while computing the distances or the closest cluster is very less in the proposed scheme. From the results, we can conclude that the number of iterations required, to get the final cluster assignment, is less in the proposed scheme.

REFERENCES

[1] R. Agrawal, R. Srikant, Privacy-preserving data mining, in: ACM Sigmod Record, Vol. 29, ACM, 2000, pp. 439–450.

[2] Y. Lindell, B. Pinkas, Privacy preserving data mining, in: Annual International Cryptology Conference, Springer, 2000, pp. 36–54.

[3] M. Upmanyu, A. M. Namboodiri, K. Srinathan, C. Jawahar, Efficient privacy preserving k-means clustering, in: Pacific-Asia Workshop on Intelligence and Security Informatics, Springer, 2010, pp. 154–166.

[4] R. Cramer, I. Damgard, J. B. Nielsen, Secure multiparty computation and secret sharing-an information theoretic approach, Book draft.

[5] Z. Lin, J. W. Jaromczyk, An efficient secure comparison protocol, in: Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on, IEEE, 2012, pp. 30–35.

[6] I. Damg°ard, M. Geisler, M. Krøigaard, Efficient and secure comparison for on-line auctions, in: Australasian Conference on Information Security and Privacy, Springer, 2007, pp. 416–430.

[7] S. Patel, V. Patel, D. Jinwala, Privacy preserving distributed k-means clustering in malicious model using zero knowledge proof, in: International Conference on Distributed Computing and Internet Technology, Springer, 2013, pp. 420–431.

[8] A. Shamir, How to share a secret, Communications of the ACM 22 (11) (1979) 612–613.

[9] G. Blakley, Safeguarding cryptographic keys, Proc. AFIPS 1979 NCC 48 (1979) 313–317.

[10] A. Beimel, Secret-sharing schemes: a survey, in: International Conference on Coding and Cryptology, Springer, 2011, pp. 11–46.

[11] S. Yakoubov, V. Gadepally, N. Schear, E. Shen, A. Yerukhimovich, A survey of cryptographic approaches to securing big-data analytics in the cloud, in: High Performance Extreme Computing Conference (HPEC), 2014 IEEE, IEEE, 2014, pp. 1–6.

[12] Y. Lindell, B. Pinkas, Secure multiparty computation for privacypreserving data mining, Journal of Privacy and Confidentiality 1 (1) (2009).

[13] Lichun Li, Rongxing Lu, Kim-Kwang Raymond Choo, Anwitaman Datta, and Jun Shao. 2016. Privacy-preserving-outsourced association rule mining on vertically partitioned databases. IEEE Transactions on Information Forensics and Security 11, 8, 1847-1861.

[14] Beaver, D. & Haber, S. (1992). Cryptographic Protocols Provably Secure Against Dynamic Adversaries. Advances in Cryptology, EUROCRYPT’92. Balatonfüred, , 307-323.

[15] Canetti, R. (2000). Security and Composition of Multiparty Cryptographic Protocols. Journal of Cryptology, 13 (1), 143-202.

[16] Chaum, D., Crepeau, C. & Damgard, I. (1988). Multi-party Unconditionally Secure Protocols. The 20th Annual ACM Symposium on Theory of Computing, STOC’88, pp.11-19.

[17] Vadlana Baby, N. Subhash Chandra, Distributed threshold k-means clustering for privacy preserving data mining, 03 November 2016.