Privacy Preserving Using Distributed K-means Clustering for Arbitrarily Partitioned Data

[1]Neha B. Jinwala, [2]Gordhan B. Jethava
[1]PG Student, [2]Assistant Professor
Parul Institute of Engineering and Technology, Limda
[1]neha.jinwala@gmail.com, [2]g.jethava@gmail.com

Abstract* - Advances in computer networking and database technologies have enabled the collection and storage of large quantities of data, also the freedom and transparency of information flow on the Internet has heightened concerns of privacy. Nowadays the scenario of one centralized database that maintains all the data is difficult to achieve due to different reasons including physical, geographical restrictions and size of the data itself. The data is normally maintained by more than one organization, each of which aims at keeping its information stored in the databases privately, thus, privacy-preserving techniques and protocols are designed to perform data mining on distributed data when privacy is highly concerned. Cluster analysis is a frequently used data mining task which aims at decomposing or partitioning a usually multivariate data set into groups such that the data objects in one group are most similar to each other. Distributed data mining is concerned with the computation of data that is distributed among multiple participants. Privacy preserving distributed data mining allows the cooperative computation of data without parties revealing their individual data. The work focuses on arbitrarily partitioned data which is a generalization of horizontally partitioned data and vertically partitioned data along with Shamir's Secret Sharing Schemes which was designed with the goal of achieving complete privacy for secure computation and communication between different parties. At the end of the work one can conclude that one achieves privacy with minimum or no leakage of the data thus satisfying the security constraint.

***Keywords* - Centralized K-means, Distributed K-means, Privacy preserving clustering, Shamir's Secret Sharing.**

I.INTRODUCTION

Due to availability of high speed internet, many organizations are able to gather large amount of data of their clients. Many a times this information gathered is sensitive in nature. Nowadays it is easy to get user specific information which is linked to other datasets. With this, various analysis and processing techniques have developed through which organizations and many government firms are publishing micro data (i.e. Data that contain separated information about individuals) for various purposes like data mining purposes, studying disease outbreak or economic patterns. These databases sometimes contain sensitive information about individuals keepingtheir privacy at risk.

Data mining is a practice of examining large databases in order to look for hidden patterns in a group of data that can be used to predict future behavior. True data mining techniques doesn't just change the presentation, but actually discovers previously unknown relationships among the data. So basically data mining is "*Knowledge Discovery in databases*". For example center of disease Control (CDC) may want data mining operations to identify trends and patterns in disease outbreaks such as understanding and predicting the progression of flu. Insurance companies have these data but they are not willing disclose it.

The amount of data collected could be "mined" for knowledge that could improve the performance of the organization. Well known data mining tasks include clustering, prediction, association rule mining. While much data mining occurs on data within the organization, it is common to use data from multiple sources in order to provide more precise and useful knowledge. However, the privacy and secrecy considerations would more likely prohibit organizations from willing sharing their data with each other. For example, an organization gathers all the details of its employees for record basis like employees address, bank details for transferring of salaries, cellphone numbers etc. However, if an adversary has an access to this data, he/she can easily discover the identities of individuals which could be misused putting at risk hundreds of thousands of workers, without even them being aware of it.

For data mining sufficient amount of accurate data is must, these include sharing of privacy sensitive data for analysis. The primary task in data mining is development of models about aggregated data. Can we develop accurate models without access to individual precise information in data records? Let's take instance of Inter-enterprise data mining: two different organizations org1 (Internet marketing company) and org2 (online retail company) both have different attributes for a common set of individuals. These organizations share their data for clustering to find the optimal customer targets so as to maximize return on investments. How can org1 and org2 learn about their clusters using each other's data without learning anything about the attribute values of each other? (This problem is called privacy preserving clustering over vertical partitioned data).

Recently, more importance has been placed on preserving the privacy of user-data aggregations (a case in point is the whistleblower Edward Snowden who has brought forth how privacy of end users is being compromised routinely which has led to a worldwide debate on security of user information), e.g., databases of personal information. Despite the risks involved, aggregating and analyzing data collections is enormously useful, in sighting trends and providing insights user preferences and requirements. It is from this balance between privacy and utility that the area of *privacy preserving data-mining* emerged [1].

II.RELATED WORK

In this section, a summarization of the pioneering work on privacy preserving data mining which shaped the research in this field can be deeply studied.

An overview of privacy preserving data mining focusing on distributed data sources can be studied in [9]. Firstly, all the databases that are gathered for mining are huge for which scalable techniques for privacy preserving data mining are needed. Comparison between different encryption schemes and secret sharing is done in which secret sharing schemes turns out to be one step ahead. It is also concluded that techniques which minimize the amount of computation and data transfer are needed in distributed environments [9]. Also, the new data types such as the spatio-temporal data collected by location-based services and other mobile service providers pose new types of threats to privacy for which existing techniques may not be adequate.

The related work in [6] uses group-based pseudo-data generation m order to preserve anonymity. In k-anonymity technique it requires that each record in an anonimized table to be indistinguishable with at least k-1 other records within the dataset, with respect to a set of quasi-identifier attributes. It protects against identity disclosure, but does not provide sufficient protection against attribute disclosure. Next, they discuss about the perturbation approach, which works under the need that the data server is not allowed to learn or recover precise records. The distribution of each data dimension *is* reconstructed independently. A brief discussion about the cryptographic techniques is done. Further discussion in the paper is about the condensation approach, which constructs constrained clusters in the data set, and then generates pseudo-data statistics of these clusters.

Few more techniques such as the heuristics-based techniques, cryptographic-based techniques and reconstruction-based techniques have been discussed in brief in [8].

Heuristics-based technique modifies the selected values using the perturbation-based method, blocking method or aggregation method. Cryptographic-based techniques conduct a secure computation based on multiparty private inputs, where each party only knows its private input and the finalized results. Reconstruction-based techniques randomize the original distribution of the data using the perturbation method and reconstruct the distribution from perturbed data. They have presented an efficient privacy preserving data mining platform which endows the mining project with higher privacy protection, higher scalability and low information loss.

The problem of protecting the underlying attribute values when sharing the data for clustering has been addressed in [12]. Main focus of this paper is on privacy preserving clustering over centralized data. A novel spatial data transformation method called Rotation- based transformation (RBT). The major features of the data transformation are that it is independent of any clustering algorithm, it has sound mathematical foundation, it is efficient and accurate and lastly it does not rely on intractability hypothesis from algebra.

Techniques from secure multiparty computation [2] form one approach to privacy preserving data mining. A problem of trust based privacy preserving clustering of participants has been addressed using the Secure Clustered Multiparty Computation (SCMC) approach.

Yao's general protocol for circuit evaluation [9] can be used to solve any two party privacy preserving distributed data mining problem. But it is purely based on theory, and in reality data mining usually involves millions or billions of data items. Privacy preserving solutions have been presented with respect to horizontally partitioned data as well as vertically partitioned databases. An introduction to arbitrarily partitioned data has been given by Jagannathan and Wright [10] which is a generalization of both horizontally and vertically partitioned data. In this partitioned data different attributes for different items can be owned by different party. Protocols developed for this model can be applied to both horizontally and vertically partitioned data, as well as to data anywhere in between. A privacy preserving k means clustering algorithm has been proposed in the work.

Furthermore, an efficient algorithm for privacy preserving distributed k-means clustering using Shamir's secret sharing scheme has been proposed in the works of [4]. The approach collaboratively computes cluster means and hence avoiding the presence of a trusted third party.

III.PROPOSED WORK

Overview

Since a lot of data is being collected over the distributed network and privacy has become a main concern for the users. The organizations are sharing a lot of data with the researchers for the analyses purpose which is compromising the privacy of individual users. The researchers analyze sensitive data of individual user's which is now making them hesitant to share their private data on the network. For this purpose many researchers have proposed different techniques such as Yao had proposed circuit evaluation technique. This technique consisted of scrambled circuit if encryption of all bit values on all possible wires. But this approach proved to be expensive since it required complicated encryptions.

Goldwasser and Micalli developed the homomorphic cryptosystems which proved to be practically impossible to do the encryption and decryption due to its probabilistic message expansion.

The cryptosystem was further improved by Benaloh which allowed the encryption of larger block sizes at a time, but still it was not suitable as the decryption was based on exhaustive search. Paillier developed a cryptosystem which provides fast encryption and decryption algorithms which is reasonable if worked for larger plaintexts.

Later secret sharing was proposed by Shamir in 1979. Shamir developed a system which has one party having a secret which it has to distribute among n parties in such a way that none of the n parties alone can recover the secret.

Work Flow

Assume that a data set D consists of n instances with m attributes. The data set is to be vertically partitioned into p parties, so that each party contains its own database with m_i attributes, where i is the specified number of attributes.

Now all the parties want to find the final clusters over total partitioned data but trying to protect their own privacy. So, every party learns final k clusters and nothing else.

The algorithm runs in a divide, conquer and merge strategy. So, the first step is each party will compute k number of clusters on their own data sets. The second step consists of each party which will compute the difference between each record and each of the k cluster centers. So, n × k matrix will be computed by each party. Until this stage there will be no privacy issues as all the computations are done by the parties on their own databases. In the third stage each party sends their distance matrices to the other parties, along with the k cluster centers in a randomized form.

Next each party computes all possible combinations of cluster centers from the total ak clusters, where p is the number of parties and k is the number of clusters. So, finally k^p cluster centers will be formed. Now each party will have information about these k^p clusters and they will compute the distance between each point and the k^p cluster centers. The minimum closest cluster will be chosen for n data points and finally will be merged into k clusters.

The merging is done by choosing a best pair of clusters C_i and C_j and then clusters are replaced by $(C_i \cup C_j)$. The best pair is chosen which has the least error. Let C_1 and C_2 be two clusters being considered for a merge. Let C.weight denote the number of object s in cluster C. the error $C_1 \cup C_2$ is

$$\text{Error}_w (C_1 \cup C_2) = \frac{(C_1.weight * C_2.weight * dist^2(C_1,C_2))}{(C_1.weight + C_2.weight)}$$

Where $dist(C_1,C_2)$ is the distance between the centers of C_1 and C_2.

IV.EXPERIMENTAL RESULTS

The parameters that the proposed work considers are

- Communication and computational complexity: As the parties need to communicate with each other in order to compute the closest cluster for each point
- Privacy.

Pseudo Code

Algorithm 1 Shamir's Secret Sharing

Input: D: Secret value, N: Set of Parties to distribute shares, K: Number of shares required to reconstruct the secret.

Output:

```
Phase1: Generating and sending secret shares
    1. Select a random polynomial $q(x) = a_{k-1}x_{k-1} + \dots + a_1x_1 + a_0$ where $a_{k-1} \neq 0$ and $a_0 = D$.
    2. Choose n publicly known distinct random values $x_1, x_2, \dots, x_n$ such that $x_i \neq 0$
    3. Compute the shares of each node $p_i$, where share(i) = q(x)
    4. For i = 1 to n do
    5. Send share i to node $P_i$
    6. End for
Phase2: Reconstruction
Require: Every party is given a point (a pair of input to the polynomial and output)
    7. Given subset of these pairs, find the coefficients of the polynomial using interpolation
    8. The secret is the constant term.
```

Algorithm 2 Privacy Preserving k-clustering

Input*: P Party's k cluster centers, P Party's distance matrix.*

Output*: Assignment of cluster number to objects.*

```
    1. Each party compute k-cluster center (C1,C2,…,Ck) from $m_{i.}$
    2. Each compute the distance matrices $M_{party}$.
    3. Each party randomly share cluster centers using Shamir's Secret sharing algorithm and distance matrices with each other.
    4. Each party form all possible cluster centers from the existing cluster's information i.e. $k^p$ clusters will be
    5. formed.
    6. Closest cluster.
    7. Find minimum value on each row of X matrix to find closest cluster for each instance.
    8. Place n instances to appropriate closest clusters.
    9. Merge $k^p$ clusters to form final k clusters.
```

Algorithm 2: Closest-cluster

Input: Distance matrix of each party (n × k).

Output: Closest cluster assignment for n instances, a matrix X ($n \times k^p$) that holds the distance between each pair of n points and k^p cluster centers.

```
for p=1 to n
    l=0
    for q=1 to k
        for r=1 to k
            l=l+1
            $X_{pl} = a_{pq} + b_{pr}$
        End for
    End for
Return X
```

RESULTS

The protocol proposed by Doganay et.al [7] assumes non colluding party which is impractical in real world because if we violate the assumption then they can reveal the secret of a particular party. On the contrary the proposed protocol does not reveal the secret value of a party even if all the remaining parties exchange their shares since each party execute Shamir's secret sharing algorithm. The combination of shares of other parties cannot reveal the secret, since all n shares are needed to reveal the secret.

Theoretical Analysis

Privacy

In the proposed approach, the secret value s_i of party P_i cannot be revealed even if all the remaining parties exchange their shares, since, each party P_i executes Shamir's Secret Sharing algorithm with a random polynomial of degree n-1. The values of that polynomial at n different points are needed in order to compute the coefficients of the corresponding polynomial. Each party P_i computes the value of its polynomial at n points as shares, keeping one of these shares for itself and sends the remaining to n-1 shares to other parties.

Further, no party learns anything more than its prescribed output, because every party shares its local cluster means as the secret; for which it chooses different polynomial randomly. It is not possible for a party to determine the secret values of other parties, since the individual polynomial coefficient is not known to other parties. Disclosure of intermediate cluster means during the program execution is prevented as intermediate cluster means are calculated at each site and there is no need to communicate them.

Correctness

Each party is guaranteed that the output that it receives is correct. Assuming the party P_i has private vector A_i. They have to perform addition of all shares to get the secret value. The secret value is the constant term of the sum polynomial. The linear equations need to be solved, not knowing there are n unknown coefficients and n equations.

Computation cost

Computation cost depends on the initial clusters and the no. of iterations required for finding final clusters. Computational complexity for computing the distance by each party is O (nk). The computational complexity for closest cluster procedure is O (nk^2). Since it runs n times for each instance and for each instance it takes O (k^2) time. Total computational complexity is O (nk^2).

Communication cost

For communication complexity, each party sends k shares to each other. Assuming that it takes c bits to represent each share then the total communication complexity is O (kc). To send the distance matrix O (nk) time is taken. Also, the cost of secret sharing is 2P (P-1), where P is the number of party. Thus total communication cost is O (nk) + 2P (P-1).

Experimental Results

Table 1 Test Case for Glass Dataset

Sr No.	Scheme Name	Accuracy (%)	Communication Cost (Bytes)	Computation Cost (ms)	No of Iterations
1	Local K-Means	-	240	1.39989E+12	20
2	Distributed K-Means	100	528	25476	20
3	Privacy Preserving Distributed K-Means	100	624	125327	20

Table 2 Test Case for Forest Covertype Dataset

Sr No.	Scheme Name	Accuracy (%)	Communication Cost (Bytes)	Computation Cost (ms)	No of Iterations
1	Local K-Means	-	1320	1.39989E+12	20
2	Distributed K-Means	99	2628	798330	20
3	Privacy Preserving Distributed K-Means	99	2784	1432529	20

V.CONCLUSION AND FUTURE WORK

From the literature survey one can conclude that though there are many schemes proposed for partitioning the data arbitrarily, none of the scheme is able to achieve complete privacy, as well as reduce the communication and computational cost. Hence, the proposed algorithm collaboratively computes cluster means and hence avoid trusted third party.

Our algorithm supports arbitrarily partitioning in presence of semi honest adversary model applied on two different data sets (one being large and one being small). As a future work we intend to extend our algorithm in presence of malicious adversary model.

REFERENCES

[PAPERS]

[1] Privacy Preserving Data mining; D. Aruna Kumari, K. Rajasekhara Rao and M. Suman, Proceedings of the 48th Annual Convention of CSI-Volume II, Springer International Publishing Switzerland, 2014.

[2] Towards Secure Clustered Multi-Party Computation: A Privacy-Preserving Clustering Protocol; Sedigheh Abbasi, Stelvio Cimato and Ernesto Damiani; ICT-EurAsia 2013, LNCS 7804, pp. 447-452, 2013, IFIP International Federation for Information Processing 2013.

[3] A Survey of Clustering of Partitioned Data in a distributed network; S. Harippriya, Prof. T. Kalaikumaran and Dr. S. Karthik, International Journal of Science, Engineering and Technology Research (IJSETR), Volume 2, Issue 2, February 2013, ISSN-2278-7798.

[4] An Efficient Approach for Privacy Preserving Distributed k-Means Clustering based on Shamir's Secret Sharing Scheme; Sankita Patel, Sweta Garasia and Devesh Jinwala, IFIPTM 2012, IFIP AICT 374, pp. 129-141,2012.

[5] Privacy Preserving K-Means Clustering: A Survey Research; Fatima Meskine and Safia Nait Bahloul, The International Arab Joumal of lnfonnation Technology, Vol. 9, No.2, March 2012.

[6] A survey on privacy preserving data mining; Jian Wang, Yongcheng Luo, Yan Zhao and Jiajin Le; First International workshop on Database Technology and Application, IEEE 2009.

[7] Distributed Privacy Preserving Clustering with Additive Secret Sharing, Doganay M., Pederson T., Saygin Y., Savas E., and Levi A., in Proceedings of the International Workshop on Privacy and Anonymity in Information Society Table, Nates, France, pp. 3-11, 2008.

[8] An Efficient Privacy-Preserving Data Mining Platform; Ronggong Song, Larry Korba and George Yee; National Research Council Canada, 2008.

[9] Secret Sharing vs Encryption-based techniques for privacy preserving data mining; Thomas Pederson, Yucel Saygin and Erkay Savas; UNECE/ Eurostat Work Session on SDC, 2007.

[10] Privacy-Preserving Distributed k-Means Clustering over Arbitrarily Partitioned data; Geetha Jagannathan and Rebecca Wright; In Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2005.

[11] Privacy Preserving Clustering; Somesh Jha, Luis Kruger and Patrick McDaniel; ESORICS 2005, LNCS 3679, pp. 397-417, Springer, Heidelberg (2005).

[12] Achieving Privacy Preservation when Sharing data for Clustering; Stanley R.M. Oliveria and Osmar R. Zaiane, SDM 2004, LNCS3178, pp.67-82, 2004, Springer Heidelberg(2004).

[13] Distributed Data Mining: Algorithms, Systems and Applications; Byung-Hoon Park and Hillol Kargupta, 2002, pp. 341-358.

[14] Simjava: A Discrete Event Simulation Library for Java; Fred Howell and Ross McNab; Dept. of Comp. Sci. Univ. of Edinburgh, 1998.

[15] Simulation of Parallel and Distributed Systems: A Taxonomy and Survey of Tools; Anthony Sulistio, Chee Shin Yeo, and Rajkumar Buyya.

[LINKS]

[16] Data Mining Wiki Available: http://en.wikipedia.org/wiki/Data_mining.

[17] NetBeans Wiki Available: http://en.wikipedia.org/wiki/NetBeans.

[THESIS]

[18] Privacy Preserving Clustering in Data Mining; Jitendra Kumar and Binit Sinha; Department of Computer Science and Engineering, National Institute of Technology, Rourkela.