# Analysis of Privacy Preserving Clustering Approach over Horizontally Partitioned Data

Priya Kumari[1], Seema Maitrey[2]

[1]M.Tech (CSE) Student KIET Group of Institution, [2]Assistant Professor KIET Group of Institution

Priyayadhuvanshi23@gmail.com, seema.maitrey@gmail.com

**Abstract-** Data mining is the most current topic in research area. From a very long time we are working on this topic that is how we can secure our database. There are many problems which are associated with this topic like data missing, data lost, hence we explain some approaches like hierarichal clustering, homomorphic encryption, k-means clustering and SMC by these techniques database which is horizontally divided can be preserved by using clustering. We also focus on other approaches then clustering or the approaches which are used on vertically partitioned data. In this paper we focus on horizontally distributed data. We give a tabular literature review of some techniques in tabular form

**Keywords:** Agglomerative Clustering, Data Mining, privacy, Divisive Clustering, Hierarichal Clustering, Homomorphic Encryption, K-Means Clustering, SMC.

## 1. INTRODUCTION

Data mining is defined as the collection of a large amount of data set in which is extracted or mined knowledge from a large amount of dataset. It is a process in which we have to process a large amount of data set which is mainly called as database and then it search for some pattern or relationship within the stored data [1]. The main goal of datamining is to extract some information from a dataset and after that transform that into an understandable structure for future use.

Clustering is an good approach for privacy preserving data mining when we are going to deal with sensitive data or information now a day privacy issue is a major concern because a little loose of data or information result in financial or individual losses of any corporation or as well as in the some other area like science, insurance, medical machine learning [2].

Partitioning of data is mainly divided in three parts these are:

Vertically partitioned data, horizontally partitioned data, arbitrary partitioned data.

When we are going to perform any privacy preserving approach then first we have to see that on which type of data we want to apply that approach then we can easily made any approach the main problem of privacy preserving clustering is that in a data set there are n no of data objects which are distributed in a large partitioned dataset and these objects should be clustered in a k number of cluster which are based on their similarity the main focus on they are clustered without effecting the privacy of the database. But this approach is suitable for only when there are only one user when more than one user this approach are become less applicable so we are going to discuss some approach which are based on all three type of partitioning of data and also trying to focus on horizontal partitioning data approaches for more than two party [3].

In this paper we explain some efficient approaches which are existing and also discuss some benefits or limitations of privacy preserving approaches.

## 2. SOME BASIC PRIVACY PRISERVING TECHNIQUES

The main focus of data mining privacy preserving approach is to reduce data misuse. There are a number of methods that comes in going years only the researcher's focus on privacy and preserving of data mining. Some basic concepts that are basically used for privacy of data mining are given below:

**Random share**: A very popular method in privacy preserving data mining. Each will compute their intermediate values that are designated by uniformly distribute the random value where each of the party has taking one of its values. The actual intermediate value is the sum of their value. In this basically adds the noise to the every individual value hence as a result they cannot be aggregated and recovered of those randomised values shall give the final output and the final result is to be available for all other parties.

**Homomorphic Encryption**: The homomorphic encryption method mainly focuses on that people may jointly perform mining task which are based on private input they are providing. Hare we insure and secure by the transformation method that the encryption method gives them.

Homomorphic encryption schemes permit some certain computation over the encrypted value. Let (A, B, C, D) be the homomorphic encryption scheme where A is an algorithm which generate keys , B and C are the encryption and decryption function and D is message space. These have the following properties

(A, B, C) are symmetrically secure

$B(a_1) \times B(a_2)=B(a_1+a_2)$, for $a_1, a_2 \in D$.

**Secure scalar product protocol**: In this approach the scalar product and the vector product are of two different vector and computed securely that means that actual value of vector are not reviled only the result is reviled.

Let there is vector $X= (x_1,.....x_n)$, which is held by Alice , and $Y=(y_1, .....y_n)$, held by Bob. They need to securely compute scalar product $X \cdot Y = \sum i = 1 - n(xi * yi)$ hence in the end pf the protocol Alice only know $X \cdot Y$ not Y and this is also same for Bob.

**Secure Add and Compare**: Secure and compare is based on homomorphic encryption in this each party build a circuit there each have two input, then the first input of both party and also same as second input of both party are sum respectively after this return the result of these comparing value of the two sum [4].

Let there are two parties $p_1$, $p_2$ that have number $a_1$ and $b_1$ and $p_2$ has number $a_2$ and $b_2$ in this process it securely find that if $a_1+a_2<b_1+b_2$without effecting the following two pieces of information to the other party :

1. Numbers of the processer by each party.
2. Difference between $a_1+a_2$ and $b_1+b_2$.

## 3. PRIVACY PRESERVING CLUSTERING ALGORITHM

Clustering is a process in which the similar type of data are grouped in the same place and these similarity are based on their properties of the data but these group must be most dissimilar from the other groups these groups are called as cluster and the whole process is called as clustering [5]. In privacy preserving of data mining the clustering is an important approach in the database there are mainly three types of partitioning:

1. Vertically partitioning.
2. Horizontally partitioning.
3. Arbitrary partitioning.

### 3.1 Clustering over vertically partitioned data

Vertically partitioned data is that in which the data is collected from different sites collect information about the same set of entities and they collect different feature sets. In vertically partitioned data there are used in keeping medical records and also it is used in for keeping the record of call phone record of the same people which have medical record. As many clustering algorithm is exist in literature format here we give few of the details of some approaches used in vertically partitioned clustering.

In real time the vertically partitioning is used in various areas hence it is also used in various technical researches area for making the privacy strong like in hierarichal clustering we use vertically partitioned data.

Table 1
Privacy Techniques over Vertically Partitioned Data

| Authors | Year | Distribution models | Parties number n | Security tools | Main objective |
|---|---|---|---|---|---|
| Vaidya and Clifton[6] | 2003 | Vertically | n>2 | -Secure permutation<br>-Homomorphic encryption schemes<br>-Yao evaluation circuit | The first privacy preserving K-means algorithm based on secure multi-party computation. |
| Samet et.at [7] | 2007 | Vertically | n | -Secure multiparty addition<br>-Secure sum | A multi-party privacy preserving in K means algorithm. |
| Dogany et.al [8] | 2008 | Vertically | n>3 | -Additive secret sharing schemes | A new protocol based on additive secret sharing scheme instead of homomorphic encryption. |

Table 2
Privacy techniques over horizontally partitioned data

| Author | Year | Distribution mode | Parties number n | Security Tools | Main Objective |
|---|---|---|---|---|---|
| Jha et.al [9] | 2005 | Horizontal | 2 | -oblivious polynomial evaluation.<br>-homomorphic encryption schemes. | Comparing two protocol in term of computation and communication cost |
| Samet et.al [7] | 2007 | Horizontal | n | -secure multi-party addition.<br>-secure sum. | A multi-party privacy preserving in k means algorithm. |

### 3.2 Clustering over horizontally partitioned data

Horizontally partitioning is the important type partitioning in which the data are collected same set of information from different sites but about different entities. In horizontally partitioning data this is used in keeping credit card details of two different location credit union. Horizontal partitioning is the distribution of data on the bases of row division if we see a table of having some database then if we applying horizontal partitioning this means the database is divided by row the all of the main attributes are same but the record is different

### 3.3 Clustering over Arbitrary partitioned data

Clustering over arbitrary partitioned data is the most important type of partitioning in the clustering and this is n also important in the way that in this we have partitioning operation of both either it is vertical or horizontal both of the operation can be performed here hence it is defined as a method or approach in which randomly select that which approach is selected as the user need horizontal or vertical he can use by his convenience [6]. Here some of the approaches are given by using arbitrary partitioned data.

Table 3
Privacy techniques over Arbitrary partitioned data

| Author | Year | Distribution mode | Parties number n | Security Tool | Main Objective |
|---|---|---|---|---|---|
| Jagannathan and wright [10] | 2005 | Arbitrary | 2 | -Random shares.<br>-Secure scalar product.<br>-Yao evaluation circuit | A secure protocol which can be used on two data distribution horizontal and vertical. |
| Su et.al [11] | 2007 | Arbitrary | 2 | -Secure scalar product.<br>-Oblivious polynomial evaluation.<br>-Secure approximation technique. | Secure data standardisation and security improvement. |
| Bunn and ostrovsky [12] | 2007 | Arbitrary | 2 | -Paillier cryptosystems.<br>-Secure scalar product. | Resolving the problem of secure multi-party division and a new protocol for secure randomly selection of K first centres. |
| Sakuma and Kobayashi [13] | 2008 | Arbitrary | n | -Paillier cryptosystem.<br>-Random shares.<br>-Yao evaluation circuit. | Scalable and fault tolerant protocol. |

# 4. CLUSTERING TECHNIQUES OVER HORIZONTALLY PARTITIONED DATA

## 4.1 k-means clustering

The k-means clustering algorithm when it is applied over the horizontally partitioning of datasets then the distance computation does not violate itself its privacy it is because each party holds its all of the component of an entity. When it starts to compute the intermediate cluster centres then the problem arises and in this case there entities that belong to same cluster may come from numerous parties where they work for interest of protecting them. For computing this step we have to require knowledge of the number of entities of each of the cluster this number is the extra information that must not be exposed to any other party while they execute the protocol. The problem comes in random selection of K first centres of privacy in data distribution [14].

The privacy preserving protocol over the horizontally partitioned data should prevent the release of data and its additional information such as intermediate centres themselves.

Jha et. Al .give two protocols in k-means algorithm for the privacy of the data this K-means applied for only two parties. In the given scenario the entities details of the tom parties are kept confidential but the intermediate centre part are exposed to the both of party and both the party locally done the computation distance. Here the privacy and preserving are computed for each party to give the security here they use mainly two protocol the first one is OPE which is Oblivious Polynomial Evaluation that is given by Naor and Pinkas the another one that is the second protocol is DPE which is based on homomorphic encryption schemes.

In the given solution it is required that both of the party must be semi-honest to each other the main aim is experimentally assessment of two protocol. In these two schemes homomorphic encryption is more efficient then OPES if we take two perimeters that is communication cost and the computation.

Samet et al. they proposes an protocol in which the protocol is used to secure the entities of both of the parties for this it gives a method of division hence it exposé of number of entities of each party in the cluster. This protocol is also appropriate in multi-party computation environment but here also the intermediate centres are always revealed

### 4.2 Hierarichal clustering

Hierarichal clustering is one of the most important approaches in the privacy preserving clustering. In clustering we discuss many approaches in those approach k-means clustering is a popular approach but along with that many other approaches also exist each of the approach have its own area in which they are strong [15]. The another approach is the SMC secure multiparty computation this is also good if we are taking the data of two parties that means if two parties wants to share the data than they have two give some of the essential data and the other confidential data is hide from the others hence like this there are many approaches here we discus about the hierarichal approach [16].

In a hierarichal clustering approach we create an hierarichal decomposition of the data objects in this mainly we have two steps so after performing these two steps we complete the hierarichal clustering these two are given below.

It may be agglomerative and divisive by this we can know that how the hierarichal decomposition is formed:-

1. **Agglomerative:** It is a bottom up approach which starts with where each point forming the separate groups it work on merging the most closest objects and groups. It group all of the same property object in a single group and until it groups all of the object it does not terminate once it complete all of the objects into group then it execute the termination command.

2. **Divisive**: It is an top to down approach it starts form the cluster where all of the objects are in the same cluster in the each of the next iteration the big cluster is split into two other small cluster until it go the state where it have only one object in each cluster once it get that state where it found one cluster in each state then the termination condition is applied.

In hierarichal clustering there is only one fact that once a step is performed that means merge or split technique it cannot be go back to the previous state or we can say undone.

#### 4.2.1 Advantage of hierarichal clustering

- Hierarichal clustering have Embedded flexibility regarding the level of the granularity.
- It gives easily handling of any form of similarity or distance.
- It is applicable and consequent for any type of attribute.

#### 4.2.2 Disadvantage of hierarichal clustering

- It gives the vagueness of termination criteria.

- It is fact that most of the hierarchical algorithms do not revisit one constructed the intermediate clusters with the
- Purpose of their improvement.

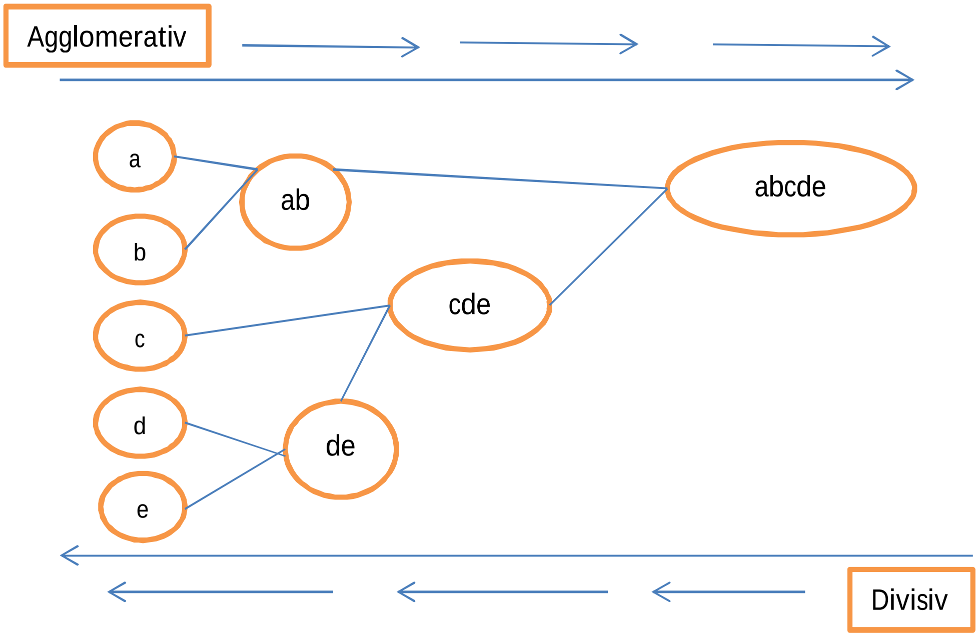

Fig 1: Hierarichal clustering over data items a, b, c, d, e.

### 4.3 Secure Multiparty Computation

SMC is introduced by Yao in 1982. The main concern of every privacy preserving technique is that how we make our data safe from the unauthorized user's secure multi-party computation is one of the best approaches in privacy preserving of database. In horizontally partitioned database as we know that the data is partitioned by the row not from the column hence the main attributes are same but the number of identities are different from each other there is no similarity between the distributed tables.

SMC is one of the approach in which there are mainly two types of methods by following them we can achieve the goal of privacy and preserving in this the first method is semi-honest in this the two party or we can say the person how involved in the process of computation follow the several protocols correctly but at the same time also try to infer the information of the other parties or persons from the data which they see during the execution of the protocols.

The second one is a malevolent model in this the malevolent parties may can do anything to infer secret information in this they can anytime abort the protocols and also send superior message, spoof message and collude with the other parties.

### 4.4 Homomorphic Encryption

In homomorphic encryption is used mainly to protect the data from the attackers how wants to steel the data which is confidential for the organisation or any personal user. In homomorphic encryption is used in horizontally partitioned database in a way that it apply some operations over the plain text and then the plane text convert into cipher text and again we have to follow some computation only after that we can get the encrypted data result. The encrypted result is same as the result which is come after performing over the plane text.

Raju R.,et al. obtain the multiple parties can conduct the data mining without effecting the data privacy this type of privacy and preserving add and multiply and exchange the technology an approach that add and multiply protocol which are based on homomorphic Encryption techniques which is defined to exchange the data by keeping its data private.

Jianming Zhu introduce a new approach for privacy preserving collaboration data mining in a distributive environment which is mainly based on homomorphic encryption and ElGamal encryption scheme. This approach is mainly used for K-nearest neighbour search it is good approach for prevent attacker from the database.

We have the idea that combined advantage of both the RSA public key cryptography and the homomorphic encryption scheme and the algorithms which are used for Paillier cryptosystem for computing global support which are used for horizontal partitioning of the database.

## 5. PROPOSED WORK

In this paper we discuss some very important privacy preserving clustering approaches for horizontally partitioned data there is a lot of work which can be done in this field of privacy preserving if database. In this paper we only analyse all of the possible methods over horizontally partitioned data. We are going to implement a method in our next paper that how we can use the hierarichal clustering for privacy preserving of two parties over the horizontal partitioned dataset. Horizontal partitioning has its own advantages like in banks and credit card details we use horizontal partitioning hence we propose an approach by which we can easily secure the dataset which is distributed between two parties and also horizontally partitioned we apply hierarichal clustering over that data so that they can securely distribute their data to each other. In our next paper we implement this approach

## 6. CONCLUSION

In this paper we have to come to a point that privacy preserving of data sets is very essential hence after analysing all of the privacy preserving clustering algorithm for horizontally partitioned data we come to know that each of the given approach like K-mean, homomorphic encryption, secure multiparty computation and the hierarichal clustering methods all are good for special purpose where k-means is a very important role for making the clusters where the data having same property are grouped in a cluster hence the cluster are formed up to limit all of the data is grouped in any cluster, hence this is an important approach on the other hand hierarichal clustering is an approach which is used in the recent technology [23]. Hierarichal clustering gives a new way for privacy of the database its two main methods give us to way for performing privacy computation. Hence this paper is helpful for analysing the clustering approach for horizontally distributed data set [24].

## 7. FUTURE WORK

In future we can apply hierarichal clustering over more than two parties in this paper we discuss all approach which are using one or two party not more than that so in future we can apply these method by further extension used them for multiple parties come to participate.

## 8. References

[1] Mohamed Ouda, Sameh Saliem (2015). Privacy preserving data mining in homogeneous collaborative clustering, the international general of information technology, vol-12 no-6 page no-604-612.

[2] Hirofumi Miyajima, Noritakd shigei (2017). New privacy protection clustering methods for secure multiparty computation, sciedu press artificial intelligence research,vol-6 no-1 page no-27-36.

[3] I.brankovice and V.Estivill-castro (1999). privacy issues in knowledge discovery and data mining, in proc.Austral inst comput.ethics conf, page no-89-99.

[4] G. Jagannathan, K. Pillaipakkamnatt, R. N. Wright and D. Umano, “Communication-Efficient Privacy Preserving Clustering,” Transactions on Data Privacy 3,Vol. 3, No. 1, 2010, pp.1-25.

[5] Sacca D., and Wiederhold G(1985). Database Partitionin in a Cluster of Processors. ACM TODS,Vol 10,No 1.

[6] Vaidya J.,(2008), A survey of privacy preserving methods across vertically partitioned data, ACM page no-206-215.

[7] Samet S., et al (2007), Privacy preserving K-means clustering in multi-party environment, International conference on security and cryptography page no-523-531.

[8] Doganay M., et al. (2008), Distributed privacy preserving clustering with additive secret sharing, International workshop on privacy preserving and anonymity in information society France page no-3-11.

[9] Jha S, Kruger L., (2005), privacy preserving sclustering, European symposium on research in computer security page no-397-417.

[10] Jagannathan G and Wright R,(2005),Privacy preserving distributed k-means clustering over arbitrarily distributed data, 11 ACM conference USA, page no-593-599.

[11] Su C.,et al, (2007), Privacy preserving k-means clustering via secure approximation, 21st International conference on advanced information networking and application workshop Oantario, Canada page no-385-391.

[12] Bunn P. and Ostrovsky R., Secure two party k-Means clustering, ACM, USA, page no-486-497.

[13] Sakuma S., et al,(2008), Large scale k-means clustering with user-centric privacy preservation, Knowledge discovery and data mining Berlin, page no-320-322.

[14] A. Fahad et al.,(2014), A survey of clustering algorithms for big data : taxonomy and empirical analysis, IEEE Trans vol-2 no-3 page no-267-279

[15] I.De and A. tripathy,(2104), a secure two party hierarchal clustering approach for vertically partitioned data set with accuracy measure , 2nd international symp. Vol-34 no-3 page no-153-162.

[16] G.Karpis and E-H.Han (1999), A hierarichal clustering algorithm using dynamic modelling , IEEE computer, vol-32 no-8 page no-75.

[17] Miyajima H, Shigei N (2016), A proposal of back propogation learning for secure multi-party computation methods, international multi-conference of engineers and computer scientests , pageno-381-386.

[18] J. Wang, J.Zhan (2008) , towards real time performance of data value hiding for frequent data update, IEEE International conference ,page no- 606-611.

[19] S.Xu and J.Han (2007), Single value decomposition based data distortion strategy for privacy protection knowledge and information system, vol-10 no-3 page no-348-361.

[20] Mohamed Ouda and Sameh Salem (2015), Privacy preserving data mining in homogeneous collaborative clustering, the international Arab journal of Information Technology vol-12 no-6 page no-604-612.

[21] Guang Li, Yadong Wang (2012), A privacy preserving classification method based on singular value decomposition, The International arab journal of Information Technology, vol-9 no-6.

[22] Amara M and said A (2011), Elliptic curve cryptography and its application, 7th international workshop on system, signal processing and application, Tipaza, Algeria page no-247-250.

[23] R.Akhtar and R.J.Choudhary (2013), privacy preserving two party k-means clustering in malicious model, IEEE 37th workshop, page no-121-126.