TWO PARTY HIERARICHAL CLUSTERING OVER HORIZONTALLY PARTITIONED DATA SET

Priya Kumari[1] and Seema Maitrey[2]

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

ABSTRACT

Data mining is a task in which data is extracted from the large database to make itin an understandable form or structure so that it can be used for further use. In this paper we present an approach by which the concept of hierarchal clustering applied over the horizontally partitioned data set. We also explain the desired algorithm like hierarichal clustering, algorithms for finding the minimum closest cluster. In this paper wealso explain the two party computations. Privacy of any data is the most important thing in these days hence we present an approach by which we can apply privacy preservation over the two party which are distributing their data horizontally. We also explain about the hierarichal clustering which we are going to apply in our present method.

KEYWORD

Two party computations, Partitioning, clustering, k-means algorithm, Hierarichal clustering.

1. INTRODUCTION

Data mining is a very current research area in these days only because of its ability to extract the data from a large data set very efficiently. Data mining is a field in which the main aim is to extract or mine knowledge from a large amount of data [1]. In data mining generally the processing is done over the large volume of data that is stored in a database and search for pattern and relationships inside the data. Privacy is also the main point of focus in these days in between the researchers all of us have some data that we don't want to share with anyone hence whenever the situation arises that we wants to secure our data from others then we use have some approaches like association rule mining, classification and clustering. In this paper we are going to use clustering approach.

Clustering of data is a method by which the similar kind of cluster are grouped together and one by one each attribute comes to any cluster in the end of the approach. When we are dealing with some sensitive information then the privacy issue is a major concern because if any of the information is leak or compromise then that may result to effect or harm to individual or financial losses to any well stablished organisation. Clustering is used widely in many real time areas like in financial affairs, marketing, medical, chemistry, insurance, machine cleaning, data mining etc [2,3].

2. PRELIMINARIES

In this section we are going to present some preliminaries. We first introduce the partitioning approach by which we partition the data set. Then we explain about how to party works and after then the main approach or algorithms which we are going to implement to fulfil our approach.

2.1 Two party computation

Two party computationsare an approach in which mainly two party involve in computation. These two parties have their own data set in an equal amount but they don't expose their data to its corresponding party. In this way they a form an distance matrix and share their distance matrix to each other not the original data set by merging the two distance matrix of each party we come to a single matrix by which we can easily get the solution of our queries because these distance matrices compute the smallest distance of each cluster with the help of cluster center so any user get their result in a very reasonable time [4]. Two party computation query model which consist of mainly four entities these are following:-

1. Randomizer.
2. Computing engine.
3. Query front end engine.
4. Individual Database.

In this the randomizer and computing engine are comes in primary engine. The query front end engine which is responsible for receiving all of the queries from the users and then it forward these queries to the randomizer. An encoded query which is normally contains the type of query to its computing engine. The computing engine coordinates that query to individual database for computing the result of query.In our approach we apply hierarichal clustering over two party computations which are using horizontally partitioned data.

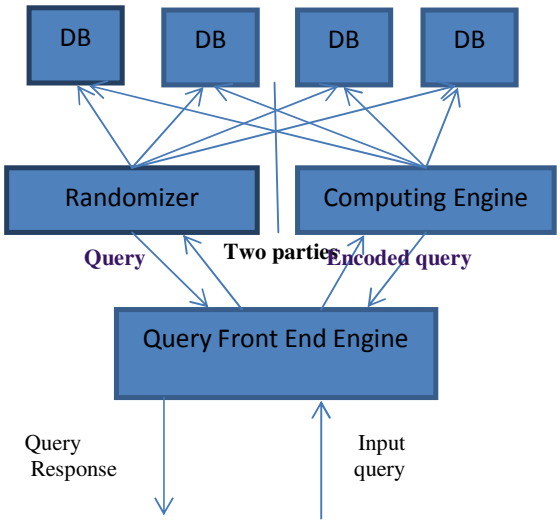


Figure1. Two Party computations Model

2.2 Partitioning Of Database

Partitioning of data base is the process in which we partition the data set in horizontal, vertical and arbitrary. In the partitioning we apply several techniques to full fill the task of preserving the data if we apply the perfect partitioning then we can easily make our approach good. First is horizontally partitioning this means that the partitioning apply on the data set in which we are deal with the data of a complete row we don't have to worry about the column information a complete row information comes in result.

In vertically partitioning the data is distributed in columns if any of the query come that means that the information of a complete row comes in result this is good for if we are requesting for the data of a single attribute.

The third partitioning approach is arbitrary partitioning that means that the horizontal and vertical both type of approaches comes in this. When the query is requested from the database at that time it is decided to apply either horizontal or vertical approach.

In all of the approaches the arbitrary approach is good but there is not much exploration in this field but this approach is applied by using k-means clustering.

Table 1. Record of Students

S.No	Name	Branch	Id
1	Ram	CSE	234
2	Rohan	CSE	235
3	Geeta	CSE	223
4	Pooja	ECE	342

Table 2. **V**ertically Partitioned

S.No	Name
1	Ram
2	Rohan
3	Geeta
4	Pooja

Table 3.Horizontally Partitioned

S.No	Name	Branch	Id
1	Ram	CSE	234
2	Rohan	CSE	235

In our approach we are going to use horizontally partitioned database hence we have the data which is organised by row [5]. If we are taking data by a single attribute we have the all information about a single attribute in a same time hence this is good approach for the banks and academics. For this partitioning we use hierarchical clustering and also we are using agglomerative approach with some encryption techniques.

3. CLUSTERING

Clustering is an approach which is best if we are dealing with some sensitive data or information. The privacy is the major issue because there are many chances that the information is leak. Hence clustering is the most appropriate approach for making the privacy strong[6] [7] [8].

Clustering is mainly said that if we are clustering some data then we have to find the data which is most similar in their properties hencethey are cluster in a single group. Each group is different from the other groups either in size, number of objects and their dimension and also they have different data types.

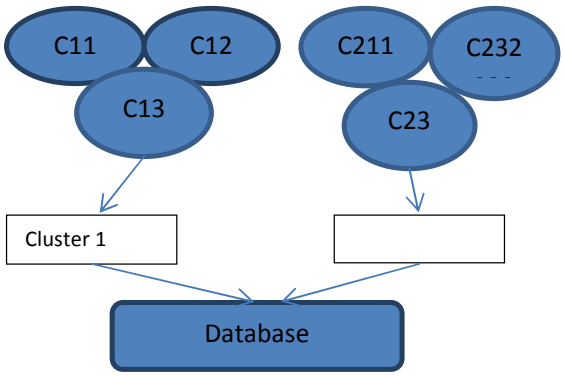


Figure 2. Clustering of Database

Clustering is a data mining technique which is unsupervised data analysis [9]. It offers advanced and more abstracted view to the dataset which is complex to handle if we are using simple techniques. There are many clustering based privacy techniques which are given by researchers. Types of clustering are following:-

1. Hierarchical clustering.
2. K-means clustering.
3. Density based clustering.
4. Self-organised maps EM clustering.

These are basic types of clustering algorithm which are mainly used in these days in many approaches.

4. HIERARCHICAL CLUSTERING

Hierarichal clustering is one of the clustering approaches [10]. In hierarchal clustering is mainly divided in two methods which are following:-

1. Agglomerative approach.
2. Divisive approach.

4.1 Agglomerative Approach

Agglomerative approach is one of the hierarichal clustering approach which is applied over the database. In agglomerative approach this make cluster of database from its bottom to its top

hence it is also called as bottom up approach. This is the most commonly used approach in the field of clustering of data sets.

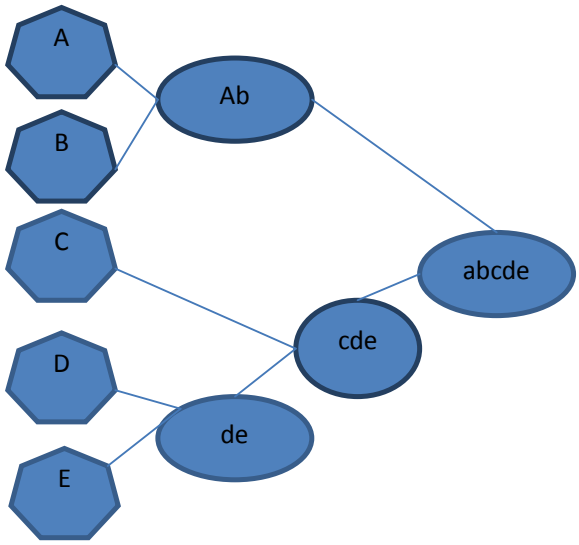


Figure 3. Agglomerative Approach

4.2 Divisive Approach

In divisive approach hierarichal clustering is applied over the database from top to bottom. Hence this approach is called as top to bottom approach.

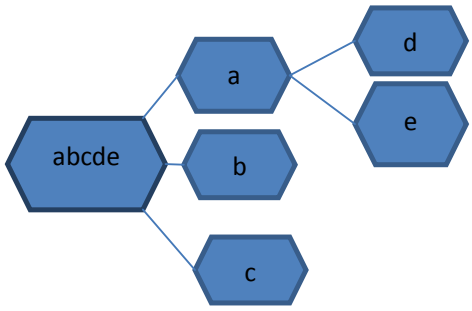


Figure 4. Divisive Approach

5. PRIVACY PRESERVING CLUSTERING ALGORITHM

5.1 K-Means clustering.

In our approach we are using k-means clustering algorithm for partitioning of datasets. The followings steps are followed:

- Let X={X1,X2,……Xn} are the data elements and v={v1, v2,…….vc} are the set of cluster center.
- First randomly select ‘c’ cluster center.
- Calculate the distance between each of data element and cluster center.
- Assign the minimum distance of each element from the entire cluster center.
- These steps are repeated until all of the elements come to a cluster.

5.2 Euclidian Distance Matrix

Euclidian distance matrix is an nxn matrix which represents the spacing of n points in any Euclidian space.

Let there are two party P1 and P2 these are distributing the database D. P1 have the distance matrix of first nxk elements and P2 have another nxk+1.

Now both of the parties have two set of data and the k-cluster center and distance matrix.
Before sharing the distance matrix we have to apply encryption over the data. In our approach we use two encryption algorithms which are following:

1. SHA1
2. MD5

These two algorithms are applied over the data of two parties. Each of the party uses a different encryption technique. Hence it is hard to understand for each of the party about the original dataset.

The main advantage of this approach is that one dataset is damage then that will not affect the other dataset.

5.3 Privacy Preserving Hierarichal agglomerative clustering.

Input: P1 have his cluster center and distance matrix and P2 have its own cluster center and distance

Matrix.

Output:

- Assign all of the elements to a cluster.
- P1 compute k-cluster center (c_1, c_2, $c_{3....}c_k$) from the first attributes.
- P2 compute in the same manner the left attributes (c_{k+1}, c_{k+2} ….c_{2k}).
- P1 and P2 compute their cluster center and distance matrix as M_{P1} and M_{P2} respectively.
- P1 and P2 randomly share their cluster center and distance matrix. We use permute share algorithm for sharing this information between them.
- P1 and P2 make the all possible cluster from the existing cluster informationi.e. k^2 cluster will be formed.
- Make a closest cluster from each party.
- Find the minimum value of each row of X matrix to find closest cluster for each instance that is if the i^{th} column have minimum value in j^{th} instance then that will become the closest i^{th} cluster.
- Place each instance to its appropriate closest cluster.
- Merge k^2 cluster to make the final k cluster.

5.4 Algorithm for closest cluster.

Input: Given distance matrix of P1 and distance matrix of P2.
Output: first assign the closest cluster for n instance. A matrix X(nxk^2) that holds the distance between each pair of n points and the k^2 cluster center.

- For a=1 to n
- 1=0
- For b= 1to k
- For c=1 to k
- 1=1+1
- Xa1= p_{ab} +q_{ac}
- end for
- end for
- end for
- Return X

6. EFFICIENCY AND PRIVACY ANALYSIS

P1 and P2 both compute k-cluster on their own data set and then they share it to each other after encrypting the value of data in distance matrix. The encryption is done through two algorithm which are SHA1 and MD5 which takes O(k) time for each party. In the next step computational complexity for computing the distance matrix by each party is calculated as O(nk). In the next step hierarichal k-clustering take $O(k^2)$ time for computation. The computational complexity for closest cluster is $O(nk^2)$. In the last step for each instance run time is $O(k^2)$.

So the total time complexity is $O(nk^2)$

.

Both of the party send or receive their k cluster independently but in an encrypted form. So the information of the parties does not public to other including the opposite party. They share only the distance matrix but this distance matrix is only the distance computed between the cluster center and instance. So the information is not leak. After merging the final k-cluster center is exposing to each other. Hence the privacy preserving by using hierarichal clustering algorithm over two parties using horizontally partitioned data is secure and does not leak any information.

7. EXPERIMENTAL RESULT

In the given approach we take a small database of 500 students. There records have weight and height. These records are distributed in four clusters by using k-means clustering algorithm. After this on two clusters we apply SHA-1 and on the other we apply MD-5 algorithm for encryption of the data or information that these cluster have. After encryption these are shared between these two party P1 and P2. The overall approach is explained briefly above.

Here we give the comparison of our approach with some other techniques which are similar in work but take more time in query processing. We take basically k-means clustering over horizontally partitioned data , hierarichal clustering of two parties over vertically partitioned data and k-means clustering over vertically partitioned data.

A brief description is given in figure that how the given approach is better than the existing methods.

RISHI\SQLEXPRESS.clusterr - dbo.cluster | SQLQuery2.sql - not connected*

name	height	weights	s_no
Rishabh	65	220	1
Ankit	73	160	2
Shivam	59	110	3
Tarun	61	120	4
Shubham	75	150	5
Sudhanshu	67	240	6
Vivek	68	230	7
Shivanshu	70	220	8
Saurabh	62	130	9
Prabhat	66	210	10
Pawan	77	190	11
Vishal	75	180	12
Unais	74	170	13
Mohit	70	210	14
Amit	61	110	15
Devesh	58	100	16
Shailendra	66	230	17
Ravi	59	120	18
Krishna	68	210	19
Shobhit	61	130	20
Sumit	71	215	21
Dhoni	72	210	22
Sachin	61	125	23
Feroz	75	175	24

Figure 5. Some of data set on which hierarichal clustering is applied

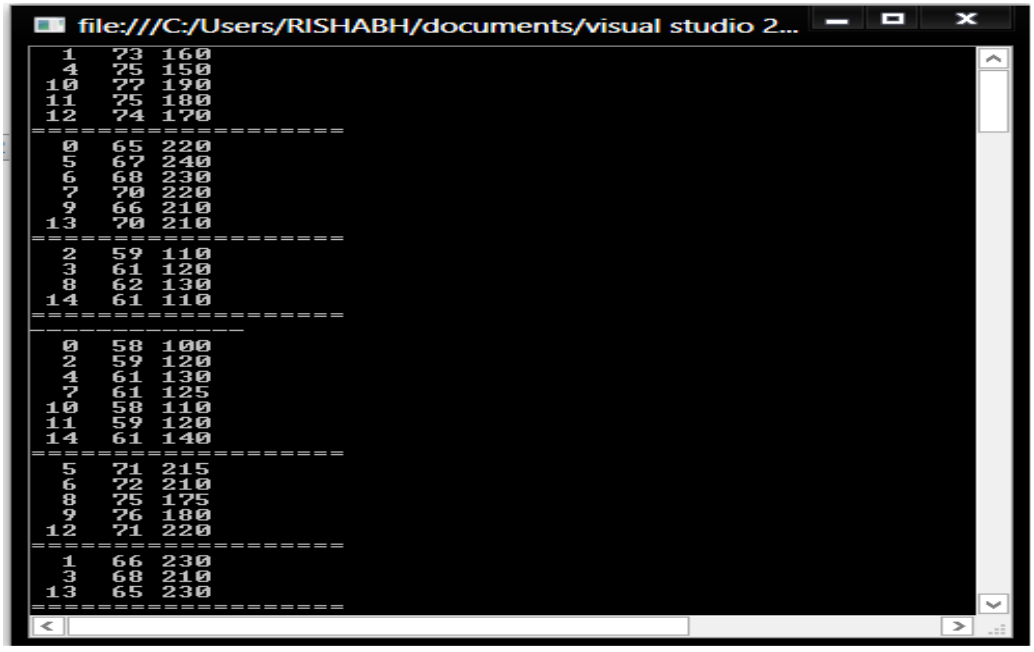
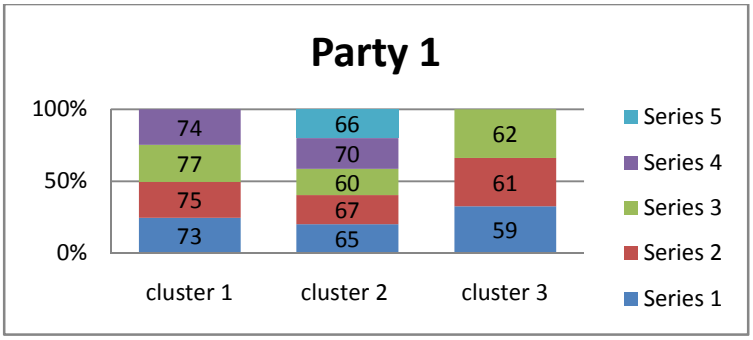
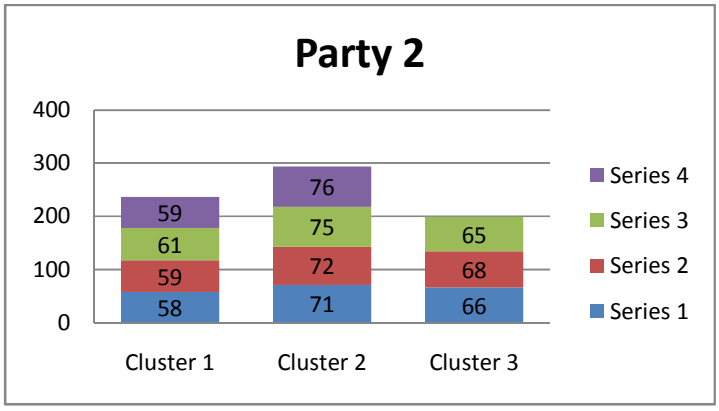


Figure 6: After clustering the database is divided in two party and 6 clusters



Graph 1: Cluster formed in Party1



Graph 2: Cluster formed in Party 2

The above graph shows the representation of data after the original database is distributed in between the two parties. Hence there are total six cluster are formed in our experiment. On these two different set of cluster we use two different encryption algorithms. Both the parties are unaware of the encryption technique of other. Hence the privacy of data is high.

s_no	height	weights
1	73	BE057D4CA44C10A0FC1DFCFFD99CCE1490291DC7
4	75	13682AC418603AA0966369D46BBF282F562ACF47
10	77	3A2DC677D8E85AC856541744E288D504882FEB36
11	75	EC7F1F65067126F3B2BD1037DE8A18D0DB2EC84B
12	74	717B2F3D8816830549097908C134E1729C516542

Figure 5.Encryption over party 1 using SHA-1

s_no	height	weights
0	58	F899139DF5E1059396431415E770C6DD
2	59	DA4FB5C6E93E74D3DF8527599FA62642
4	61	9B8619251A19057CFF70779273E95AA6
7	61	3DEF184AD8F4755FF269862EA77393DD
10	58	5F93F983524DEF3DCA464469D2CF9F3E
11	59	DA4FB5C6E93E74D3DF8527599FA62642
14	61	1385974ED5904A438616FF7BDB3F7439

Figure 6.Encryption over party 2 using MD5

Table 4. Comparison among different cluster approach

Types of cluster	k-means clustering over HPD	Hierarichal with HPD	k-means clustering over VPD	Hierarichal with VPD
No of database scanning	300	150	300	150
Running time (sec)	3.6304	1.342	3.6309	1.451

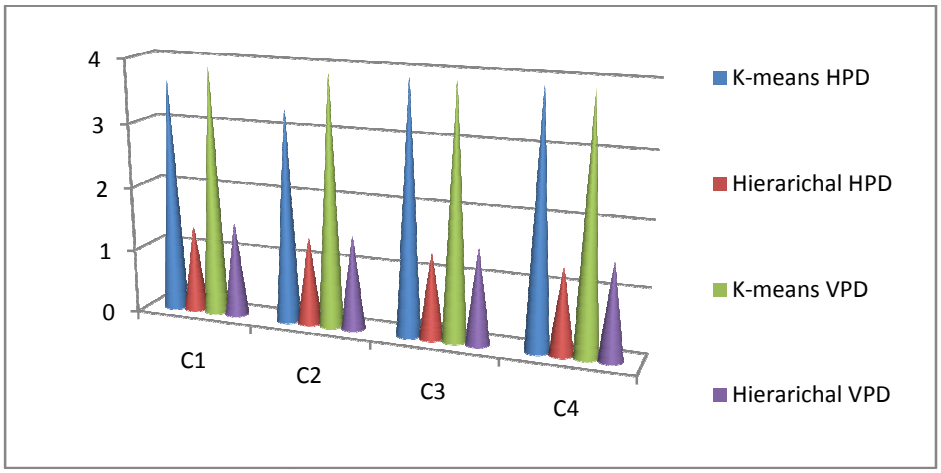


Figure 7.Execution time of each approach

8. CONCLUSION

In this paper we analyse the privacy preserving problems for horizontally partitioned data. There are various techniques used to solve the problems like adding noise or encryption data value. In this paper a hierarichal clustering approach for horizontally partitioned data for two parties is a novel approach to secure data.

9. FUTURE RESEARCH WORK

The future research work can be to find solution for hierarichal clustering for multiparty which can be apply over horizontal and vertically partitioned data. The hierarichal clustering can be further enhancing for arbitrary partitioned data.

REFERENCES

[1] J. W. Han and M. Kamber, “Data Mining: Concepts and Techniques,” 2 nd Edition, China Machine Press, Beijing, 2006.

[2] J. Vaidya and C. Clifton, “Privacy Preserving K-Means Clustering over Vertically Partitioned Data” Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington DC, USA, 2003, pp. 206-215.doi:10.1145/956750.956776

[3] T. K. Yu, D. T. Lee, Shih-Ming Chang and Justin Zhan, "Multi-Party k-Means Clustering with Privacy Consideration," International Symposium on Parallel and Dis-tribute Processing with Applications, IEEE Computer Society, 2010, pp. 200- 207.

[4] P. Bunn and R. Ostrovsky, "Secure Two-Party k-Means Clustering," In Proceedings of the 14th ACM Conference on Computer and Communications Security, 2007, pp. 486-497. doi:10.1145/1315245.1315306

[5] J. S. Vaidya, "Privacy Preserving Data Mining over Vertically Partitioned Data," Ph.D. Thesis, Purdue University, 2004, pp. 1-149.

[6] V. ESTIVILL-CASTRO, Why so many clustering algorithms: A position paper, SIGKDD Explorations Newsletter, 4 (2002), pp. 65–75.

[7] J. Vaidya and C. Clifton, "Privacy Preserving K-Means Clustering over Vertically Partitioned Data," Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington DC, USA, 2003, pp. 206-215. doi:10.1145/956750.956776

[8] T. K. Yu, D. T. Lee, Shih-Ming Chang and Justin Zhan, "Multi-Party k-Means Clustering with Privacy Consideration," International Symposium on Parallel and Distributed Processing with Applications, IEEE Computer Society, 2010, pp. 200- 207.

[9] G. Jagannathan and R. N. Wright, "Privacy Preserving Distributed k-Means Clustering over Arbitrarily Partitioned Data," Proceedings of the 11th ACM, SIGKDD International Conference on Knowledge Discovery and Data Mining, USA, 2005, pp. 1-7.

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