International Journal For Technological Research In Engineering
Volume 5, Issue 9, May-2018 *ISSN (Online): 2347 - 4718*

SECURE FILE TRANSFER VIA FILE SPLITING WITH USING MODIFIED K-MEANS ALGORITHM

Jyoti Yadav[1], Shailendra Soni[2]
[1]Research Scholar, [2]AP, [1,2]Computer Science Engg, St. Margaret Engg. College Neemrana

Abstract: In the current scenario The Security is most or of at most importance when talking about file transferring in networks. In the thesis, the work has design a new innovative algorithm to securely transfer the data over network. The k –means clustering algorithm, introduced by MacQueen in 1967 is a broadly utilized plan to solve the clustering problem. It classifies a given arrangement of n-information focuses in m-dimensional space into k-clusters whose focuses are gotten by the centroids. The issue with the privacy consideration has been examined, and that is the data is distributed among various gatherings and the disseminated information is to be safeguarded. In this thesis, created chucks or parts of file using the K-Means Clustering Algorithm aims to partition n observations into K clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of a cluster and the individual part is encrypted using the key which is shared between sender and receiver. Further, the bunched records have been encoded by utilizing AES encryption algorithm with the introduction of private key concept covertly shared between the involved parties which gives a superior security state. The term "clustering" is used in several research communities to describe methods for grouping of unlabeled data. These communities have different terminologies and assumptions for the components of the clustering process and the context in which clustering is used.

Key words: K-Means Clustering, Security, File Splitting

I. PRIVACY PRESERVING DATA MINING

Privacy Clustering[1,2] is a method to apply security to the framed cluster keeping in mind the end goal to give surety to the data proprietors that their data is being exchanged safely to the next end. The fundamental point of security saving is to ensure object values that are utilized for clustering examination. To accomplish this, every single individual should be ensured. The objective is to change D into D' i.e. moving D dataset into D' dataset by applying some example P to the dataset to accomplish protection. Clustering [26] is a technique of gathering data objects into unintelligible clusters so that the data in the same cluster is near, yet having a spot with different cluster contrast. A cluster is a social occasion of data in a way that the articles with comparable properties are assembled into comparative clusters and questions with unique properties are set into various clusters. The enthusiasm for sorting out the sharp extending data and taking in productive data from data, which makes clustering frameworks extensively associated in various applications, for instance, fake awareness, science, customer relationship organization, data weight, data mining, data recuperation, picture planning, machine learning, publicizing, Pharmaceutical, outline affirmation, cerebrum science, estimations and so forth. Cluster examination

is a mechanical assembly that is used to watch the scribes of cluster and to focus on a particular cluster for further examination. Clustering is an unsupervised learning and does not rely on upon predefined classes. Clustering method measures the uniqueness between things by measuring the partition between each pair of articles. These measures join the Euclidean, Manhattan and Minkowski division.

Privacy preserving saving data mining is the region of data mining that tries to defend the touchy data from spontaneous divulgence. Security safeguarding is fundamentally worried with ensuring against divulgence of individual data records. PPDM can be characterized by classifications. These are:

A. Data Distribution

The PPDM calculations can be initially partitioned Into two noteworthy classes,

1) Centralized data

In brought together database, data is put away in a solitary database.

2) Distributed data

In dispersed database, data is put away in various databases. Dispersed data situations can be further arranged into even and vertical data appropriations. Flat conveyances allude to the situations where diverse records of the same data traits are dwelled in better places. While in a vertical data dispersion, diverse properties of the same record of data are lived in better places.

B. Hiding Purposes

The PPDM calculations can be further characterized into two sorts:

1) Data Hiding Data: concealing alludes to the situations where the touchy data from unique database like personality, name, and address that can be connected, straightforwardly or in a roundabout way, to a distinct individual are covered up.

2) Rule Hiding: In standard concealing, the touchy learning (principle) got from unique database in the wake of applying data mining calculations is expelled. Dominant part of the PPDM calculations utilized data concealing systems.

C. Data Mining Tasks / Algorithms: The PPDM calculations are mostly utilized on the undertakings of order, affiliation control and clustering.

1) Association Rule: Association examination includes the disclosure of related standards, demonstrating trait esteem and conditions that happen as often as possible in a given arrangement of data.

*2) Classification :*Classification is the procedure of finding an arrangement of models (or capacities) that portray and recognize data classes

or ideas, with the end goal of having the capacity to utilize the model to foresee the class of items whose class name is obscure.

3) Clustering

Clustering Analysis concerns the issue of decaying or apportioning a data set (generally multivariate) into gatherings so that the focuses in one gathering are like each other and are as various as would be prudent from the focuses in different gatherings.

II. CLUSTERING

The unsupervised classification of examples, which incorporates perceptions, highlight vectors, or data things, into clusters is termed as clustering. A valuable stride in exploratory data investigation; the issue of clustering has engaged scientists in shifted controls and connections. In any case, clustering is mind boggling to decode and the distinction in sentiments and settings crosswise over groups has diminished the pace at which crucial nonexclusive philosophies and ideas are exchanged. The work endeavors to break down clustering examples and presents a brief outline of example clustering approaches from a viewpoint that takes a gander at example acknowledgment factually, alongside drawing on the critical ideas, hailed by clustering experts as essential. This work investigates the life structures of clustering, alongside methods, finding cross-cutting subjects and huge advances in the field. The work likewise depicts significant applications in view of clustering calculations, for example, picture division, data recovery and article acknowledgment. As it shows up from broad research, a legitimate cluster will contain designs that are comparative rather than an example having a place with another cluster. There exists an assortment of procedures for sorting out and speaking to data, gathering data components and measuring closeness (similitude) in data components, which frequently bring about a variety of clusters, both rich, yet confounding.

To better comprehend clustering, it is fundamental to see first the distinction between clustering (unsupervised order) and separate examination (regulated characterization). Directed characterization incorporates the procurement of pre-grouped examples that are named. The issue to be determined spins around the marking of an unlabeled example to its substantial cluster, and normally the designs, beforehand named are utilized for obtaining the depiction of classed, which are then used to name a fresh out of the box new example. In the example of clustering, the issue concerns the designation of unlabeled examples into important clusters. In a way these marks are connected to clusters likewise, however into classifications that are data driven, which means they are acquired exclusively from the current data. Considered valuable in exploratory example investigation, basic leadership, machine learning circumstances, gathering, data mining, picture division, design order and record recovery, clustering confronts issue because of absence of data. In numerous comparative issues because of absence of or being squeezed for earlier data, for example, factual models about the data, the basic leadership proficient must fall back on suspicions, a lesser number of which is viewed as attractive. In this manner, under these impediments the clustering strategy is particularly suitable in the investigation of between connections among the data focuses to make, regularly preparatory, appraisals of this structure. A phrasing utilized by a few examination groups to clarify the strategy for gathering unlabeled data, one encounters differing implications, connections, procedures and wordings for parts of 'clustering'. What's more, it is from here that the problem encompassing the extent of this overview, stems, since it would be an immense errand to create a really thorough review with the plenty of writing accessible for this field. For instance, the openness of the overview itself would be a test with the need to accommodate shifting presumptions and vocabularies identified with "clustering" from assorted groups.

III. PROBLEM STATEMENT

A. “Implementation of Privacy- Preserving Clustering in Data Mining”

Data mining manages huge database which can contain delicate data. It requires data planning which can reveal data or examples which may bargain secrecy and privacy commitments. Privacy safeguarding data mining manages concealing an individual's delicate character without relinquishing the convenience of data. It has turned into a critical region of concern yet at the same time this branch of exploration is in its early stages .Individuals today have turned out to be very much aware of the privacy interruptions of their delicate data and are extremely hesitant to share their data. The principle thought in privacy protecting data mining is twofold. To begin with, delicate crude data ought to be changed or trimmed out from the first database, all together for the beneficiary of the data not to have the capacity to trade off privacy. Second, delicate learning which can be mined from a database by utilizing data mining calculations ought to likewise be prohibited. The fundamental target in privacy protecting data mining is to create calculations for altering the first data somehow, so that the private data and learning stay private even after the mining procedure. There are numerous methodologies which have been received for privacy safeguarding data mining.

We can group them in light of the accompanying measurements:

- Data dispersion
- Data Change.
- Data Mining Calculation.
- Data or Principle Stowing Away.
- Privacy Conservation.

Subsequently, the issue of privacy conservation in clustering can be expressed as takes after: Let D is a database and C be a set of clusters produced from D. The objective is to change D into D' so that the change T when connected to D must save the privacy of individual records, so that the discharged database D' hides the estimations of secret qualities, for example, compensation, ailment analysis, FICO score, and others.

IV. OBJECTIVES OF THE STUDY

1) To comprehend the working and execution of privacy saving method utilized as a part of clustering keeping in mind the end goal to handle vast databases.
2) To actualize privacy saving method for adjusting the first data somehow, so that the private data and learning stay private even after the mining procedure.

3) To recommend the conceivable enhanced arrangement keeping in mind the end goal to effectively find profitable, non-clear data from vast databases.

V. METHODOLOGY

In the proposed concept, clustering based security framework has been implemented. The activity of clustering task is done in the following steps:

 Feature Selection: It means how many patterns are available that is how many clustering algorithms are available in the list so that we can choose the best one.

 Pattern Definition: It defines the properties of individual pattern. For example; in k-means Euclidean distance is used to find dissimilarity between two patterns.

 Grouping: Grouping means clustering, making clusters in a way such that similar data objects are placed in same cluster and dissimilar data objects are placed in different cluster.

 Information Abstraction: Now, the useful information can be easily occupied from the above step that is Clustering. Now, the desired information can be extracted in an arranged manner.

Therefore, a new modified k-means clustering is introduced in this research work which is based on the alphanumeric data and number of clusters. In this the performance of the algorithm is evaluated on the basis of the number of clusters and time parameters to compare the proposed work with the existing work.

On the basis of number of clusters, two tasks are performed

1) Splitting the File: Allows the sender to split the information into clusters in such a way that it simultaneously encrypts the file using AES encryption technique.

2) Joining the File: Allows the receiver to join the file to get the original data using the same technique. K-means is used as the base algorithm to make the comparison with the modified algorithm. The proposed work also makes the comparison of two more algorithms i.e. Hierarchical and cob web. For each algorithm comparison is made on the basis of same number of clusters and time parameters.

. Algorithm for Proposed algorithm

Step 1: Read the data.
Step 2: Select the numbers of clusters (maximum 26 for words and 10 for numeric values).
Step 3: Set initial cluster randomly.
Step 4: Put object (data) to closet cluster.
Step 5:Recalculate the new cluster create clusters based on smallest distance.
Step 6: Split the main data file on the basis of the clusters.
Step 7: Now encryption process is apply on the data using AES algorithm.
Step 8:Resultant encrypted files are then passed over to receiver.
Step 9: Join the data.
Step 10: Receiver decrypts the data by AES algorithm.

VI. SIMULATION RESULTS

Main Screen

In the main screen we are provided with the various options in which can section the option according the operation which we want to perform.

The main screen contains the main menu which contains the following options:

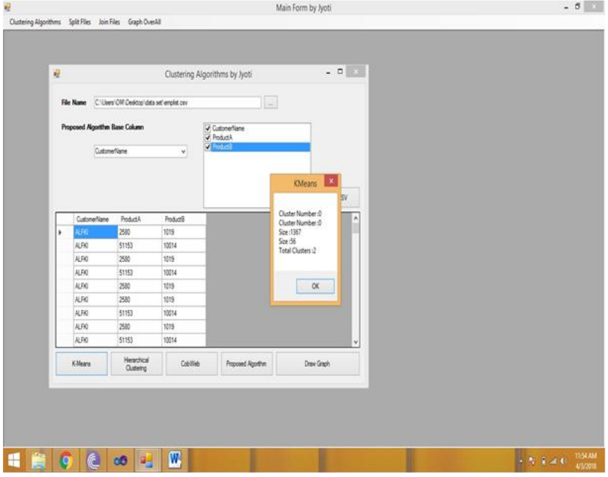
- Clustering Algorithms
- Split Files
- Join Files
- Graph Overall

Clustering Algorithm: This menu option will present the dialog for examining various clustering algorithms.

Split Files: It is used to split the main file using the various clustering algorithms.

Join Files: It is used to join the splited file in order to obtain again the single file.

Graph Overall: It is used for showing the comparison graph.



File Name C:\Users\OM\Desktop\data set\emplist.csv

Proposed Algorithm Base Column

CustomerName

☑ CustomerName
☑ ProductA
☑ ProductB

CustomerName	ProductA	ProductB
ALFKI	2580	1019
ALFKI	51153	10014
ALFKI	2580	1019
ALFKI	51153	10014
ALFKI	2580	1019
ALFKI	2580	1019
ALFKI	51153	10014
ALFKI	2580	1019
ALFKI	51153	10014

Cluster Number :0
Cluster Number :0
Size :1367
Size :56
Total Clusters :2

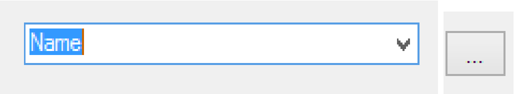
Fig.1.Form for Creating the Clusters Using Various Clustering Algorithms and Proposed algorithm.

This form is mainly used for the comparison basis and it will compare the algorithms on the basis of the number of clusters.

In this form we have taken the following algorithms,

1. K-Means
2. Hierarchical
3. CobWeb
4. Proposed algorithm

This process in the form, using the browse button we will select the CSV file containing the dataset.

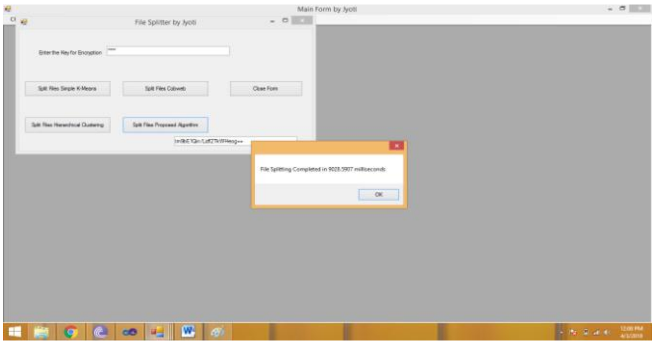


The combox automatically get populated with the columns in the .csv file and we will select the column containing the text values so that the .csv file contents will get sorted according the value of the selected field.

Click on the Save CSV button of save the .csv file containing the sorted data with name data.csv.

And this file will then used to form clusters in the proposed algorithm algorithm. Now the buttons of K-Means, Cobweb and Hierarchical will form the clusters using the WekaApi and then using the draw graph we can graphically compare the result on the basis of the number of clusters.

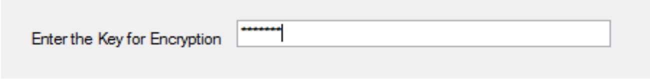
Therefore, in this form, base file is browsed from the specified location. Now "Proposed algorithm Base Column" is to be chosen to get the alphabetically organized data and then click on the "Save CSV" button to form the CSV file .After this, click on the "K-means, Hierarchical, Cob Web and Proposed algorithm" to get the desired clusters. Lastly click on graph button to get the graphical representation of number of clusters.


Main Form by Jyoti
File Splitter by Jyoti
Enter the Key for Encryption
Split Files Simple K-Means
Split Files Cobweb
Close Form
Split Files Hierarchical Clustering
Split Files Proposed Algorithm
OK

File Splitting Completed in 9028.5907 milliseconds

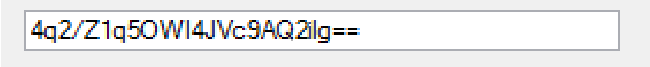
Fig.2 Form for Splitting the Files On The Basis Of Clusters Forms Using the Various Clustering Algorithms and Proposed algorithm.

In this form we will split the dataset file into clusters or chunks files. For this purpose the process which is adopted is , that we have to first write the key which is used for the encryption purpose, the enter the key text box ,



Enter the Key for Encryption *******

It will automatically generated into the corresponding hash map using the MD5 algorithm for the encryption purpose. And will be shown the label placed below in the form.



4q2/Z1q5OWI4JVc9AQ2ilg==

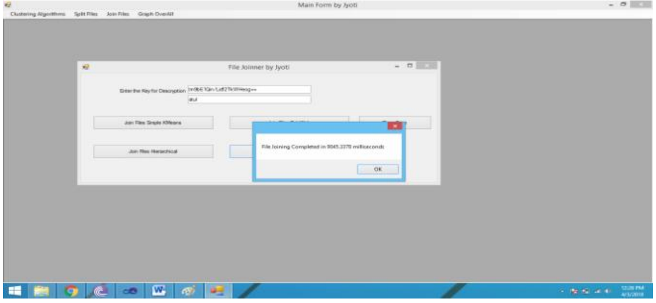
And the splitting in this form is performed on the behalf of the button on which the used clicked. There are four buttons in this form,

1. Split Files K-Means
2. Split Files CobWeb
3. Split Files Hierarchical
4. Split Files Proposed algorithm.

In any of the button we click the process which is followed is that the first the dataset file is splitted on the basis of the clusters calculated according to the algorithm selected and then these splitted files are encrypted in order to enhance the security.

.part is the extension which is given to the splitted files and .part_e is the extension which is given to the file resulted after the encryption.

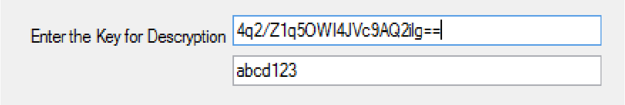
Therefore, in the second form, files are split on the basis of clusters formed. In this form, entering the private key is the first task to be performed. The key is in encrypted form to avoid any privacy attack. Now click on the "Split Files Simple K-means" button to split the files into clusters and in encrypted form. Do same for all the algorithms to get the encrypted clusters of the original data.



File Joining Completed in 9045.2370 milliseconds

Fig.3.Form for Joining The Files On The Basis Of Clusters Forms Using the Various Clustering Algorithms and Proposed algorithm

In this form we will join the dataset chunk file into main resultant file. For this purpose the process which is adopted is, that we have to first write the encrypted hash in the textbox and it will automatically results in the encryption key which is used for the decryption purpose



Enter the Key for Descryption 4q2/Z1q5OWI4JVc9AQ2ilg==
abcd123

And the joining in this form is performed on the behalf of the button on which the used clicked.

There are four buttons in this form,

1. Join Files K-Means
2. Join Files CobWeb
3. Join Files Hierarchical
4. Join Files Proposed algorithm.

In any of the button we click the process which is followed is first encrypted files are taken from the current directory with the extension _e and they are then decrypted using the key which is provided to the receiver end and after that the normal files will be obtained for the encrypted file and then the joining will be performed in order to get the initial or the original dataset file.

Therefore, in the last form, Joining of file task is to be performed on the basis of clusters formed. First task is to enter the key which was previously entered at the sender side to access the records. Next, click on join files simple K-means button to get the original file from the encrypted clusters.

VII. RESULT ANALYSIS

According to the results it is clear that the proposed Proposed algorithm results in the decent number of clusters which are well enough for the security and handling purpose.

The comparison results are shown in the following two tables for the two datasets.

Table 1 Comparison table of dataset1

Parameters	K-Means	Hierarchical	Cobweb	Proposed algorithm
No. of Clusters	2	2	334	21
Time taken for Encryption+ Splitting in milliseconds	61	59	12321	570
Time Taken for Decryption+ Joining in milliseconds	59	57	15167	567

Table 2 Comparison table of dataset 2

Parameters	Kmeans	Hierarchical	CobWeb	Proposed algorithm
Nos of Clusters	2	2	397	20
Time taken for Encryption+Splitting in milliseconds	68	58	17077	546
Time Taken for Decryption+ Joining in milliseconds	71	63	18296	547

These tables represent the result of an experimental study of privacy preserving clustering. The comparison of four algorithms is shown in the tables.

The four algorithms are:

1. K-means
2. Hierarchical
3. Cobweb
4. Proposed algorithm

Proposed algorithm is the proposed algorithm. The concept behind it is to provide better organized clusters in order to use data in an efficient manner.

In proposed algorithm 21 clusters (Dataset 1) are formed but in existing K-means only 2 cluster (Dataset 1) are formed. In proposed algorithm clusters are formed on the basis of alphabetically order thus giving much better result as compared to K-means which is forming 2 clusters and that too in some random manner.Hierarchical clustering is giving hierarchy of clusters but not in an organized manner.Cobweb is forming too many clusters which create memory issues.Therefore, Proposed algorithm is better as compared to the K-means, hierarchical, Cob Web and Proposed algorithm algorithm.

Graphs:

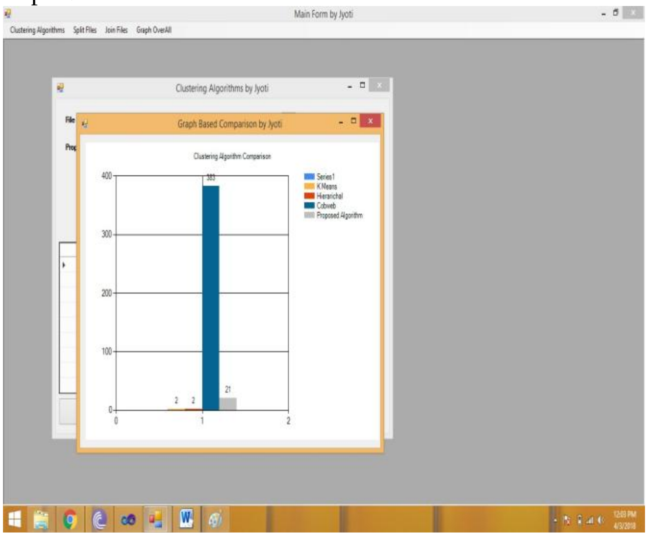


Fig 4 Graph According To Number of Clusters for Dataset 1

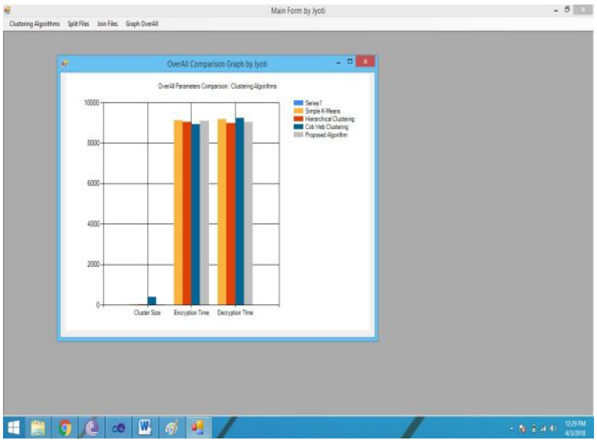


Fig 5 Graph for Overall comparison between clustering algorithms including the proposed algorithm for Dataset 1

The graph for overall comparison shows the graphical representation of all compared parameters i.e.:

Number of clusters

Encryption time (time taken for splitting)

Decryption time (time taken for joining)

According to these parameters, each algorithm has its number of clusters which are made on basis of clustering using Weka tool, Encryption time for encrypting splited files at sender side and Decryption time for decrypt files by joining at receiver side. These Algorithms have different number or clusters, different encryption time and different decryption time. On the basis of this result, comparison can be made. Proposed algorithm algorithm shows efficient results on the basis of alphabetically centralized concept.

In Graph 1, graphs according to the number of clusters are formed. There are two clusters formed for the K-means and hierarchical clustering. For Cobweb 334 clusters are formed which consumes too much memory. Lastly in case of proposed algorithm only 21 clusters are formed in an organized manner.

In Graph 3 same comparison is made on the basis of dataset 2. Number of clusters are made using Dataset 2 which also shows the same result as result on Dataset 1. All Algorithms have different number of clusters which are made on their concept.

In Graph 4 overall comparison is made on the basis of number of clusters, encryption time and decryption time. Graph 4 shows the final comparison between number of clusters, Encryption Time, Decryption Time for all algorithms. There is a File which can be splited into datasets. These splited files are further encrypted at sender side using encryption key provided by sender. This key is further used by receiver at receiver side for decrypt the file and get original file by joining.

VIII. CONCLUSION& FUTURE WORK

In a web associated universe of interpersonal organizations, the delicate individual data should be ensured. The world is confronting numerous privacy issues, so to beat this issue Adjusted K-Means calculation is being presented.

The proposed calculation is proficient from various perspectives, as far as number of clusters structures which are neither too less nor too all the more, so that the data can be equitably appropriated furthermore productive as far as the time imperatives. The Changed K-Means calculation frames clusters of dataset in an in order request as indicated by their properties. The calculation performs encryption and unscrambling procedures to give privacy to the dataset. This guarantees proprietor that their data is safely exchanging over systems. Along these lines, this will permit clients to safely exchange their data and in this way have a sorted out arrangement of clusters to extricate the required data.

According to the future extension, the proposed calculation can be further adjusted to proficiently secure the huge data and can improve the security by proposing some new encryption and decoding calculations for this reason in our future study and work.As per the future scope we can say that the proposed algorithm we will further modify to efficiently secure the big data and we will also try to enhance the security by proposed some new encryption and decryption algorithms for this purpose in our future study and work.

REFERENCE

[1] Rui Li, Denise de Vries, John Roddick,"BandsOf Privacy Preserving Objectives: Classification of PPDM Strategies", 2011 CRPIT.

[2] G. Jagannathan, K. Pillaipakkamnatt, and R.N. Wright, "A New Privacy-Preserving Distributed Clustering Algorithm," in Proceedingsof the Sixth SIAM International Conference on Data Mining, 2006.

[3] Sharaf Ansari, SailendraChetlur, SrikanthPrabhu, N. GopalakrishnaKini, GovardhanHegde, Yusuf Hyder, "An overview of clustering algorithms used in data mining", ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 12, December 2013.

[4] Yogita Rani and Dr. Harish Rohil, "A Study of Hierarchical Clustering Algorithm", International Journal of Information and Computation Technology.ISSN 0974-2239 Volume 3, Number 11 (2013)

[5] Neha B. Jinwala, Gordhan B. Jethava, "Privacy Preserving Using Distributed K-means Clustering for Arbitrarily Partitioned Data", 2014 IJEDR

[6] JyotiYadav, Monika Sharma,"A Review of K-mean Algorithm", International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 7- July 2013.

[7] Y. Lindell, B.Pinkas, "Privacy preserving data mining", in proceedings of Journal of Cryptology, 5(3), 2000.

[8] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy", in proceedings of Int'l Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 2002.

[9] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data", in The Eighth ACM

[10] Hillolkargupta, Souptik Datta, Qi Wang and Krishnamoorthy Sivakumar," Data Perturbation and features selection in Preserving Privacy", IEEE 2003.

[11] C. Aggarwal , P.S. Yu, "A condensation approach to privacy preserving data mining", in proceedings of International Conference on Extending Database Technology (EDBT),pp. 183–199, 2004. 746

[12] A.Machanavajjhala, J.Gehrke, D. Kifer and M. Venkitasubramaniam, "I-Diversity: Privacy Beyond k- Anonymity", Proc. Int'l Con! Data Eng. (ICDE), p. 24, 2006.

[13] SlavaKisilevich, LiorRokach, Yuval Elovici, BrachaShapira, "Efficient Multi-Dimensional Suppression for K-Anonymity", inproceedings of IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 3. (March 2010), pp. 334-347, IEEE. 2010.

[14] P.Deivanai, J. JesuVedhaNayahi and V.Kavitha," A Hybrid Data Anonymization integrated with Suppression for Preserving Privacy in mining multi party data" in proceedings ofInternational Conference on Recent Trends in Information Technology, IEEE 2011.

[15] G. Mathew, Z. Obradovic,"APrivacy-Preserving Framework for Distributed Clinical Decision Support", in proceedings of 978-1-61284-852-5/11/$26.00 ©2011 IEEE.

[16] A.Parmar, U. P. Rao, D. R. Patel, "Blocking based approach for classification Rule hiding to Preserve the Privacy in Database", in proceedings of International Symposium on Computer Science.

[17] S. Mumtaz, A. Rauf and S. Khusro, "A Distortion Based Technique for Preserving Privacy in OLAP Data Cube", inproceedings of 978-1-61284-941-6/11/$26.00, IEEE 2011.

[18] H.C. Huang, W.C. Fang, "Integrity Preservation and Privacy Protection for Medical Images with Histogram-Based Reversible Data Hiding", in proceedings of 978-1-4577-0422-2/11/$26.00_c, IEEE 2011.

[19] Jinfei Liu, Jun Luo and Joshua Zhexue Huang "Multiple Attributes with Different Sensitivity requirements", in proceedings of 11th IEEE International Conference on DataMining Workshops, IEEE 2011.

[20] K. Alotaibi, V. J. Rayward-Smith, W. Wang and Beatriz de la Iglesia, "Non-linear Dimensionality Reduction for Privacy- Preserving Data Classification" in proceedings of 2012ASE/IEEE International Conference on Social Computing and2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, IEEE 2012.