International Journal For Technological Research In Engineering
Volume 5, Issue 12, August-2018 *ISSN (Online): 2347 - 4718*

NEW APPROACH FOR SECURE ENCRYPTED FILE SHARING USING MODIFIED K-MEANS BASED CLUSTERING

Tanuja Sharma[1], Shanti Prakash Gehlot[2]
[1]M.Tech Scholar, [2]Assistant Professor, Sobhasaria Group of Institutions, Sikar

Abstract: In the present situation The Security is most or of at most significance when discussing record moving in systems. In the paper, the work has plot another innovative calculation to securely trade the data over system. The k-implies clustering calculation, exhibited by MacQueen in 1967 is a broadly utilized course of action to deal with the clustering issue. It portrays a given game-plan of n-data focuses in m-dimensional space into k-bunches whose fixations are gotten by the centroids. The issue with the security thought has been assessed, and that is the data is appropriated among various parties and the scattered data is to be ensured.

In this research work , made hurls or parts of record using the K-Means Clustering Algorithm intends to parcel n observations into K bunches in which each discernment has a place with the group with the nearest mean, filling in as a model of a group and the individual part is mixed using the key which is shared among sender and beneficiary. Further, the bunched records have been encoded by utilizing AES encryption calculation with the introduction of private key thought subtly shared between the included social affairs which gives an unrivaled security state. The articulation "clustering" is used as a piece of a couple of research systems to portray procedures for get-together of unlabeled data. These society have unmistakable phrasings and suppositions for the parts of the clustering system and the setting in which clustering is used. .

Keywords : Data Mining ,Clusters ,Clustering

I. INTRODUCTION

Data Mining suggests expelling or "mining" gaining from a ton of data. Data mining is useful for evacuating charming learning or productive and non-apparent data from database. It is moreover profitable for essential authority, question dealing with and data organization. In business world, data is growing progressively reliably along these lines affiliation's database is duplicating every year so there is a need to manage those data by applying the best possible estimation with the territory details. As data mining deals with the extraction of delicate data thusly it also requires to keep up the mystery and the insurance of the data. [1]

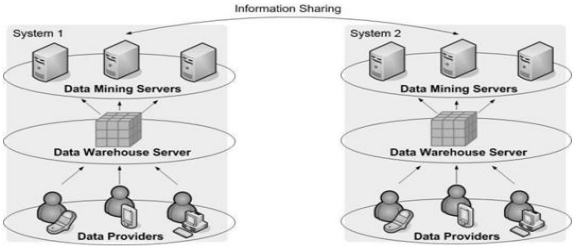

Fig 1. Data Mining Concept

Privacy Clustering[1,2] is a methodology to apply security to the encompassed bunch keeping at the highest point of the need list a definitive target to offer surety to the data proprietors that their data is being exchanged safely to the accompanying end. The vital reason for security saving is to ensure question regards 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 outline P to the dataset to accomplish protection. Clustering [26] is a strategy for get-together data objects into incomprehensible bunches so the data in a comparative group is close, yet having a spot with different bunch separate. A bunch is a get-together of data in a way that the articles with for all intents and purposes indistinguishable properties are amassed into comparative groups and request with novel properties are set into various bunches. The enthusiasm for managing the sharp growing data and taking in beneficial data from data, which makes clustering systems extensively related in various applications, for instance, counterfeit care, science, customer relationship association, data weight, data mining, data recovery, picture arranging, machine getting the hang of, publicizing, pharmaceutical, outline request, cerebrum science, estimations and whatnot. Bunch examination is a mechanical social affair that is used to watch the recorders of group and to base on a particular group for propel examination. Clustering is an unsupervised learning and does not rely upon predefined classes. Clustering strategy estimates the uniqueness between things by evaluating the segment between each match of articles. These measures join the Euclidean, Manhattan and Minkowski division. [1].

II. RELATED WORK

P.Deivanai, et. al, 2011 , Data Mining is a method of finding noteworthy data or data from the data assignment center. Unmistakable Security Protecting Data Mining estimations are made to safeguard confirmation and cover delicate data ought to be saved. In this paper, they proposed another mean protection saving data mining. In any case, they adjusted the records of the data set utilizing a novel "CTree" system and inconvenience the principal quality. By at that point, they encoded the delicate characteristics utilizing ASCII Code and unprecedented characters. . Accordingly they understood the calculation and endeavored on a downsized scale data of patient record and consequently temperamental data was exasperates effectively which will never uncover anybody's personality. In like manner, intriguing data can be recreated from abraded data, making ease of use of data.

George Mathew, et. al 2011 ,Guaranteeing grouping of

individual data and sparing assurance are essential when data is gathered from different associations for business fundamental initiative. They presented an estimation that develops data using estimations considering subject data from appropriated goals that satisfy demonstrated assurance criteria. The count keeps up entire commitment of data structures in the scattered data appeared differently in relation to the joined equivalent. Heterogeneous data mappings across finished regions can be obliged and breaking points can be set for overall minimum drenching for attributes to participate in the desire demonstrate building. Courses of action for thought and dismissal of non-far reaching properties among goals were introduced. Eventual outcomes of tests using data from helpful, propelled instruction, and social zones show the estimation of their figuring in controlled business endeavors, where shipping unrefined data outside gatekeeper association isn't reasonable.

AnitaParmar, et alI ,2011 In this the authorities found sensitive attribute and subsequently they supplanted known fragile characteristics with cloud characteristics ("?"). Finally the cleaned dataset is made from which unstable portrayal rules are not any more mined. In this way darken Qualities help in sparing security yet proliferation of remarkable data set is extremely troublesome.

Sara Mumtaz, et. al , 2011 , In this the investigators used the Data aggravation procedure which is in like manner called reliably adjusted twisting. They at first distorted one cell of 3-D square and after that twisting occurs fit as a fiddle. This assignment of mutilation method jams, and also outfits most extraordinary precision with achieve add up to request and high accessibility.

III. PROBLEM DEFINITION

Data mining oversees enormous database which can contain sensitive data. It requires data arranging which can uncover data or cases which may deal mystery and privacy responsibilities. Privacy defending data mining oversees disguising a person's sensitive character without giving up the accommodation of data. It has transformed into a basic locale of concern yet in the meantime this branch of investigation is in its beginning periods. People today have ended up being particularly mindful of the privacy interferences of their fragile data and are to a great degree reluctant to share their data. The standard idea in privacy ensuring data mining is twofold. In any case, sensitive unrefined data should be changed or trimmed out from the principal database, all together for the recipient of the data not to have the ability to exchange off privacy. Second, sensitive taking in which can be mined from a database by using data mining figurings should in like manner be denied. The essential focus in privacy securing data mining is to make estimations for changing the primary data by one means or another, with the goal that the private data and learning remain private even after the mining procedure. There are various strategies which have been gotten for privacy protecting data mining.

IV. PROPOSED WORK

A new modified k-means clustering is presented in this examination work which depends on the alphanumeric data and number of clusters. In this the execution of the algorithm is assessed based on the quantity of clusters and time parameters to contrast the proposed work and the current work.

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

- Splitting the File: Allows the sender to split the information into clusters such that it at the same time encodes the file utilizing AES encryption method.
- Joining the File: Allows the collector to join the file to get the first data utilizing a similar system. K-means is utilized as the base algorithm to make the correlation with the modified algorithm. The proposed work additionally makes the examination

Algorithm of the Proposed Work :
Step-1: Read the Excel .csv file containing the Sample data.
Step-2: Select the base file which forms the reason for clustering.
Step-3: Perform the Modified K-Means algorithm taking the alphanumeric field as the premise.
Step-4: Obtain the clusters for every algorithm.
Step-5: Split the primary data file based on the clusters and encode the files utilizing AES algorithm and the private key concept.
Step-6: Resultant scrambled files are then disregarded to collector.
Step-7: Receiver decodes the file utilizing a similar private key.

V. IMPLEMENTATION

The implementation is done in Visual Studio 2010 and the implementation is divided into the two phases :
5.1 Clustering : In this the various algorithms are compared with the proposed approach , in order to perform the comparative analysis in between the proposed and the existing approach.

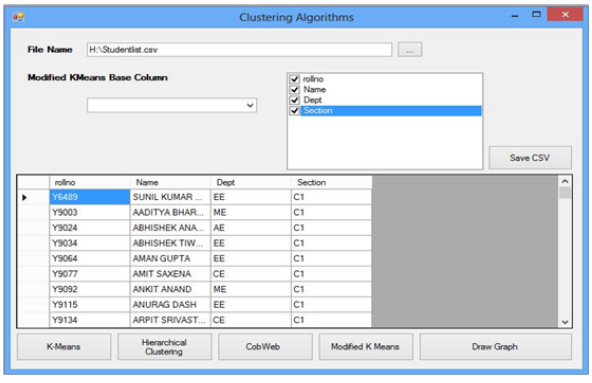

Fig 2. Clustering Section

5.2 Splitting and Joining : In this section , file is splited into the chunks on the basis of the clustering algorithm and also latter on joined to find the single resultant file.

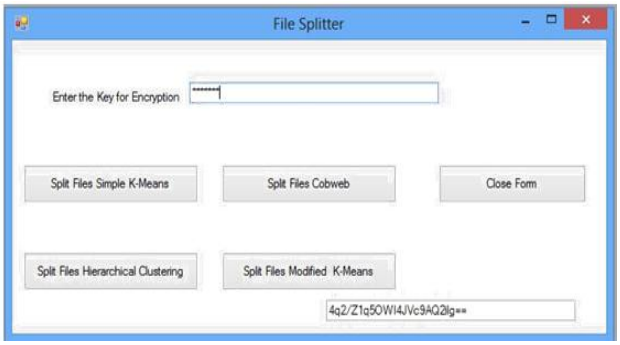

Fig 3. File Splitting

In second form , the 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.

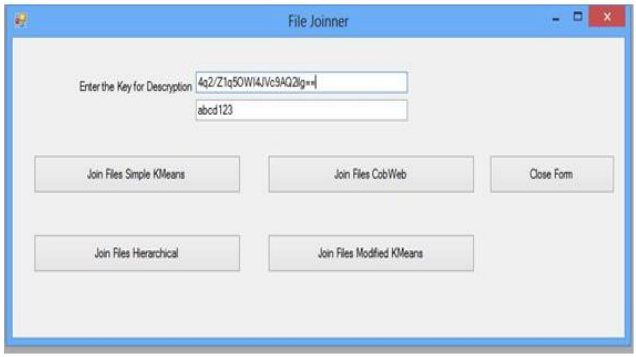

Fig 4. File Joining

VI. RESULT ANALYSIS

The result analysis involves the execution of the algorithm using the analysis of all algorithms on the dataset and then comparing the results.

Parameters	Kmeans	Hierarchical	CobWeb	Mod. KMeans
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

Table 1: Analysis for Data Set 1

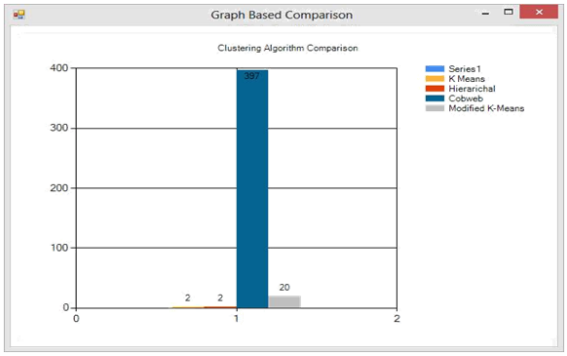

Fig 5. Graph According To Number of Clusters for Dataset 2

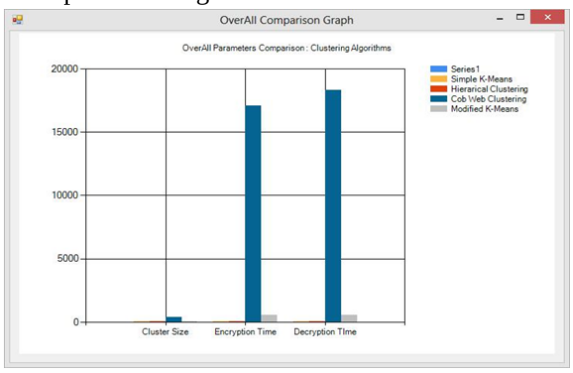

Fig 6. Graph for Overall Comparison between Clustering Algorithms Including the Modified K-Means for Dataset

Overall comparison is made on the basis of number of clusters, encryption time and decryption time. 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.

VII. CONCLUSION

The proposed count is capable from different points of view, similarly as number of bunches structures which are neither too less nor too simply more, with the goal that the data can be fairly appropriated besides beneficial to the extent the time objectives. The Modified K-Means estimation traces groups of dataset in an all together request as appeared by their properties. The estimation performs encryption and unscrambling procedure to offer privacy to the dataset. This guarantees proprietor that their data is safely exchanging over systems. In this way, this will enable clients to safely exchange their data and thusly have a managed game-plan of bunches to evacuate the required data.

REFERENCES

[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, SouptikDatta, Qi Wang and KrishnamoorthySivakumar," 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.

