International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 9 35 - 39

Secure Algorithm for File Sharing Using Clustering Technique of K-Means Clustering

Rashmi Sharma
M.Tech Scholar
Dept. of Computer Science & Engineering
Kautilya Institute of Technology & Engineering, Jaipur
Email ID: sharmarashmi296@gmail.com

Mr. Satish Kumar Alaria
Assistant Professor,
Dept. of Computer Science & Engineering
Kautilya Institute of Technology & Engineering, Jaipur
Email ID: satish.alaria@gmail.com

Abstract: In the current scenario The Security is most or of at most importance when we are 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 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.

Keyword: *K-Means Clustering, Security, File Splitting etc.*

*****

I. INTRODUCTION

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.

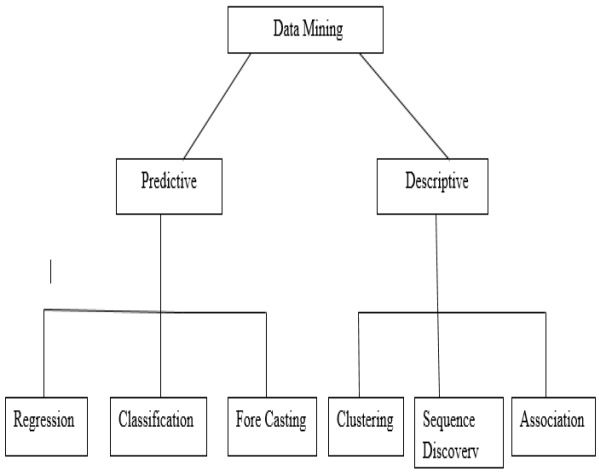


Fig.1. Data Mining Tasks/ Algorithm

II. CLASSIFICATION OF CLUSTERING ALGORITHMS

2.1 Hierarchical Clustering

Hierarchical clustering manufactures a cluster chain of importance or, as it were, a tree of clusters, otherwise called a Dendrogram. Each cluster hub contains youngster clusters and kin clusters. Such a methodology permits investigating data on various levels of granularity[4].
Various leveled clustering techniques are sorted into:

- Agglomerative (bottom-up)

- Divisive (top-down)

Hierarchical clustering (HC) calculations compose data into a hierarchical structure as indicated by the closeness network. The aftereffects of HC are typically portrayed by a parallel tree or dendrogram. The root hub of the dendrogram speaks to the entire data set and every leaf hub is viewed as a data object. The middle hubs, in this manner, depict the degree that the articles are proximal to each other; and the tallness of the dendrogram for the most part communicates the separation between every pair of items or clusters, or an article and a cluster. A definitive clustering results can be acquired by cutting the dendrogram at various levels. This representation gives extremely enlightening depictions and perception for the potential data clustering structures, particularly when genuine hierarchical relations exist in the data, similar to the data from transformative examination on various types of life forms. HC calculations are primarily delegated agglomerative strategies and divisive techniques.

2.2 Partitioning Clustering

Rather than hierarchical procedures, apportioned clustering strategies make a one level dividing of the data focuses. On the off chance that K is the wanted number of clusters, then apportioned methodologies commonly discover all K clusters on the double.

A partitional clustering calculation acquires a solitary parcel of the data in-stead of a clustering structure, for example, the dendrogram created by a hierarchical method. Partitional techniques have preferences in applications including vast data sets for which the development of a dendrogram is computationally restrictive. An issue going with the utilization of a partitional calculation is the decision of the quantity of craved yield clusters. The partitional procedures more often than not create clusters by enhancing a rule capacity characterized either locally (on a subset of the pat-terns) or universally (characterized over the majority of the examples). Combinatorial inquiry of the arrangement of conceivable marking for an ideal estimation of a foundation is unmistakably computationally restrictive. By and by, there-fore, the calculation is ordinarily run numerous times with various beginning states, and the best setup acquired from the greater part of the runs is utilized as the yield clustering.

K-Means Methods: K-Means[6] is the most celebrated separating procedure for clustering. It was firstly proposed by MacQueen in 1967. It is an unsupervised, non-deterministic, numerical, iterative procedure for clustering. In k-Means each cluster is distinguished by the mean estimation of articles in the cluster. In this work, Division of a plan of n thing into k cluster is being performed so that intercluster closeness is low and intracluster equivalence is high. Similarity is measured regarding mean estimation of things in a cluster. K-Means is broadly utilized due to its straightforwardness and the capacity to give speedy result. K-means clustering is a basic method to gathering things into k clusters. There are numerous routes in which k clusters may conceivably be framed. The nature of an arrangement of clusters can be measured utilizing the estimation of a target capacity which is taken to be the aggregate of the squares of the separations of every point from the centroid of the cluster to which it is allocated. It's required that estimation of this capacity to be as little as could be allowed. Next k focuses are chosen (by and large comparing to the area of k of the items). These are dealt with as the centroids of k clusters, or to be more exact as the centroids of k potential clusters, which at present have no individuals.

2.3 Cobweb Clustering

- Cobweb is an incremental slope climbing methodology with bidirectional administrators - not backtrack, but rather could return in principle.
- Starts unfilled. Makes a full idea progressive system (order tree) with every leaf speaking to a solitary case/object. You can pick how somewhere down in the tree progression you need to go for the particular application within reach.
- Objects portrayed as ostensible property estimation sets.
- Each made hub is a probabilistic idea (a class) which stores likelihood of being coordinated (tally/complete), and for every quality, likelihood of being on, P(a=v|C), just numbers need be put away.
- Arcs in tree are just connections - hubs store data over all characteristics (dissimilar to ID3, and so on.)

III. RELATED WORK

“Cryptographic Technique- Privacy preserving data mining” (Year 2000) by Y.Lindell, B.Pinkas This paper was expected to exhibit fundamental thoughts from an expansive assortment of cryptographic exploration on secure circulated calculation, and their applications to data mining. The fundamental parameter that influences the plausibility of executing a protected convention taking into account the non specific development is the span of the best combinatorial circuit that figures the capacity that is assessed. There is additionally a legitimate toolset for calculations of cryptography. This methodology was particularly hard proportional when more than a couple gatherings were involved. [7]

“K- Anonymity- A Model for Protecting Privacy” (year 2002) by L. Sweeney This paper presented fundamental security models termed invalid guide, k-mapand wrong-delineate give insurance by guaranteeing that discharged data guide to no, k or mistaken substances, individually. They decided what number of people each discharged tuple really coordinates requires consolidating the discharged data with remotely accessible data and investigating other conceivable assaults. Making such a determination specifically can be a to a great degree troublesome assignment for the data holder who discharges information. [8]

"Privacy preserving association rule mining in vertically partitioned data" (year 2002) by J. Vaidya and C. Clifton In this the scientists did the conveyance of data vertically into sections. The objective of this work was to diminish correspondence cost. The thought was to discover critical data focuses or designs locally and utilize these to process the worldwide examples. The methodology of securing protection of appropriated sources was initially tended to for the development of choice trees. This work nearly took after the safe multiparty calculation approach, accomplishing "flawless" protection, i.e., nothing is found out that couldn't be reasoned from one's own data and the subsequent tree. The key knowledge was to exchange off calculation and correspondence cost for exactness, enhancing productivity over the bland secure multiparty calculation method. [9]

"Data Perturbation and features selection in preserving privacy" (year 2003)by HillolKargupta, SouptikDatta, Qi Wang and KrishnamoorthySivakumar Security is turning into an inexorably essential issue in many data mining applications. This paper scrutinized the utility of the arbitrary quality contortion strategy in protection preservation.The specialists in this paper noticed that irregular items (especially randommatrices) have "unsurprising" structures in the spectral domain and it builds up an arbitrary network based otherworldly filtering technique to recover unique data from the data set distorted by including irregular qualities. This paper displayed the oretical establishment of this sifting technique and extensive experimental results to show that by and large random data twisting protect almost no data security. The paper additionally pointed out the conceivable streets for the development of new protection safeguarding data mining systems like exploiting multiplicative and hued commotion for saving privacy in data mining applications.[10]

"A Condensation Approach to privacy preserving Data Mining" (year 2004) by CharuC.Aggarwal, Philip S. Yu This methodology works with pseudo-data as opposed to with alterations of unique data, this aides in preferable conservation of protection over strategies which essentially utilize adjustments of the first data. The utilization of pseudo-data no more requires the overhaul of data mining calculations, since they have the same arrangement as the first data.[11]

"L-Diversity privacy beyond K-Anonymity" (year 2006) by A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam Distributed data about people without uncovering touchy data about them is a critical issue. In a k-anonymized dataset, every record is vague from at any rate k −1 different records regarding certain "recognizing" properties. In this paper they appeared with two straightforward assaults that a k-anonymized dataset has some inconspicuous, however extreme protection issues. Initially, they demonstrated that an assailant can find the estimations of delicate properties when there is little differing qualities in those touchy characteristics. Second, assailants regularly have foundation learning, and they demonstrated that k-obscurity does not ensure security against aggressors utilizing foundation information. They gave a subtle element examination of these two assaults and they proposed a novel and effective protection definition called ℓ-differing qualities. Notwithstanding fabricating a formal establishment for ℓ-differing qualities, they appeared in a test assessment that ℓ-differences is down to earth and can be actualized efficiently. [12]

"Efficient Multi-Dimensional Suppression for K-Anonymity"(year 2010) by SlavaKisilevich, LiorRokach, Yuval Elovici, BrachaShapira Privacy preserving data mining manages concealing an individual's touchy character without yielding the ease of use of data. It has turned into a vital territory of concern yet at the same time this branch of exploration is in its early stages. The real territory of concern is that non-delicate data even may convey touchy data, including individual information, facts or examples. In this paper, they have concentrated all these condition of craftsmanship strategies. They made a plain examination of work done by various creators. Later on they dealt with a half breed of these strategies to save the protection of delicate data. [13]

"A Hybrid Data Anonymization integrated with Suppression for Preserving Privacy in mining multi party data"(year 2011) by P.Deivanai, J. JesuVedhaNayahi and V.Kavitha Data Mining is a procedure of finding valuable data or information from the data distribution center. Different Security Protecting Data Mining calculations are created to safeguard protection and shroud delicate data ought to be saved. In this paper, they proposed another calculation for protection saving data mining. To begin with, they adjusted the records of the data set utilizing a novel "CTree" system and bother the essential quality. At that point, they encoded the delicate characteristics utilizing ASCII Code and extraordinary characters. . Along these lines they actualized the calculation and tried on a miniaturized scale data of patient record and consequently touchy data was bothered in an effective way which will never uncover anybody's personality. Likewise, unique data can be recreated from irritated data, making ease of use of data. [14].

IV. PROBLEM DEFINITION

The past part examined about the privacy issues while performing clustering in data mining. Hence, there is a need to protect privacy by utilizing some suitable algorithms there is a need to limit unapproved clients to counteract interruption, fakes, privacy ruptures and to ensure that individuals need not to stress over their delicate data which they would prefer not to share .Giving security to touchy data is the significant need in data mining as any unapproved access to proprietors data will bring up issues on the trust capacity of data extraction. Hence, this work will concentrate on executing the privacy of data in Data Mining.

V. PROPOSED SOLUTION

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

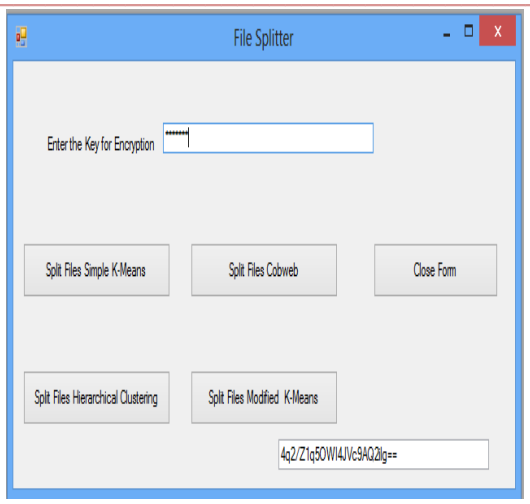
- Splitting the file
- Joining the file

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.

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

VI. PROPOSED ALGORITHM

Step 1: Read the Excel .csv file containing the Sample data.
Step 2: Select the base file which forms the basis for clustering.
Step 3: Perform the Modified K-Means algorithm taking the alphanumeric field as the basis
Step 4: Obtain the clusters for each algorithm.
Step 5: Split the main data file on the basis of the clusters and encrypt the files using AES algorithm and the private key concept.
Step 6: Resultant encrypted files are then passed over to receiver.
Step 7: Receiver decrypts the file using the same private key.



VII. CONCLUSION

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.

VIII. REFERENCES

[1] Rui Li, Denise de Vries, John Roddick, ”Bands Of 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, "A Privacy-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.
[21] E. G. Komishani and M. Abadi, "A Generalization-Based Approach for Personalized Privacy Preservation in Trajectory Data Publishing", in proceedings of 6'th International Symposium on Telecommunications (IST'2012), IEEE 2012.
[22] T. Jahan, G.Narsimha and C.V Guru Rao, "Data Perturbation and Features Selection in Preserving Privacy" in proceedings of 978-1-4673-1989-8/12, IEEE 2012.
[23] D.Karthikeswarant, V.M.Sudha, V.M.Suresh and A.J. Sultan, "A Pattern based framework for privacy preservation through Association rule Mining" in proceedings of International Conference On Advances In Engineering, Science And Management (ICAESM - 2012), IEEE 2012.
[24] M. N. Kumbhar and R. Kharat, "Privacy Preserving Mining of Association Rules on horizontally and Vertically Partitioned Data: A Review Paper", in proceedings of 978-1-4673-5116-4/12/$31.00_c, IEEE 2012.
[25] Zhengli Huang, Wenliang Du and Biao Chen, "Deriving Private Information from Randomized Data", in proceedings of 2012 Fourth International Conference on Computational Intelligence and Communication Networks, IEEE 2012.
[26] A.K Jain , M. K. Murty, P.J Flynn," Data Clustering: A Review", in proceedings of 16th International Conference on Intelligence in Next Generation Networks, IEEE 2012.
[27] Michael Beye, ZekeriyaErkin, Reginald L. Lagendijk, "Efficient Privacy Preserving K-means Clustering In a Three-Party Setting", 2011 IEEE.
[28] Teng-Kai Yu, D.T. Lee, Shih-Ming Chang,"Multi-Party k-Means Clustering with Privacy Consideration", IEEE DOI 10.1109/ISPA.2010.8
[29] Deepak S. Turaga, Michail Vlachos, Olivier Verscheure, "On K-Means Cluster Preservation using Quantization Schemes", 2009 IEEE
[30] Dongxi Li, Elisa Bertin ,Xun Yi, "Privacy of Outsourced k-Means Clustering", ASIA CCS'14