K-Medoids Computation Model Framework for Security in Distributed Architecture Networks

Lalitha Ariyapalli
Department of CSE
Sri Sivani College of Engineering
Srikakulam, AP, India

Rajendra Kumar Ganiya
Department of CSE
Sri Sivani College of Engineering
Srikakulam, AP, India

P Suresh Kumar
Department of CSE
Sri Sivani College of Engineering
Srikakulam, AP, India

`

ABSTRACT
In this paper, a method was proposed to maintain the networks with low cost for more processing of data. It contains simple framework to maintain the data in refine method and for secure data transfer. And also more data maintenance with clustering capability with secure mode.

Keywords
TDEA, DES

1. INTRODUCTION
In many of the researches in the recent years, the privacy preserving data clustering based on the k means algorithm and applying the method of secure multi-tasking communication with different datasets was discussed. There are two types of security models that is semi-honest adversary model and malicious adversary model. Semi honest model is very simple and more answered in privacy applications. It follows more rules and regulations using proper input and it is free to execute of the set of rules to confirm the security.

In existing clustering algorithms, the clustering is based on the minimize of the objective method which was mostly used and studied on datasets. In another research the computable function can be computed securely. It is in polynomial size Boolean circuit with given datasets and divides between the parties. But it is not enough to provide security for the data transfer. There are some other methods such as the holomorphic schemes. This is the hidden value result which is operated on hidden values.

Unsupervised learning deals with designing classifiers from a set of unlabeled samples. First cluster or group unlabeled samples into sets of samples that are “similar” to each other is one of the common approach for unsupervised learning. Once the clusters have been constructed, we can design classifiers for each cluster using decision-tree learning [3, 4]). Clusters can be used to identify features that will be useful for classification. We had discussed the problem of privacy-preserving algorithms for clustering in detail in this paper.

There are several applications of clustering [5]. Privacy-preserving clustering algorithm is a possible candidate for any application of clustering where privacy more concerns. For example, network traffic is collected at two ISPs, without revealing the individual traffic data the two ISPs want to cluster the joint network traffic. The present algorithm was used to attain joint clusters while concerning the privacy of the network traffic at the two ISPs.

All the more as of late, the information distortion methodology has been connected to boolean affiliation rules [6], [7]. Once again, the thought is to change information values such that reproduction of the qualities for any individual exchange is troublesome, however the tenets scholarly on the misshaped information are still substantial. One attractive component of this work is an adaptable meaning of protection; e.g., the capacity to accurately figure an estimation of "1" from the misshaped information can be considered a more prominent danger to protection than effectively learning a '0'.

The information distortion methodology addresses an alternate issue from our work. The values must be kept private from whoever is undertaking the mining. We rather expect that a few gatherings are permitted to see some of the information, simply that nobody is permitted to see all the information. Consequently, we have the capacity to get precise, as opposed to estimated, results.

The issue of secretly registering affiliation rules in vertically divided disseminated information has additionally been tended to [10]. The vertically divided issue happens when each exchange is part over various locales, with every site having an alternate arrangement of characteristics for the whole arrangement of exchanges. With flat parceling every site has an arrangement of complete exchanges. In social terms, with flat partitioning the connection to be mined is the relations' union at the locales. In vertical parceling, the relations at the individual locales must be joined to persuade the connection to be mined. The adjustment in the way the information is conveyed makes this a vastly different issue from the one we address here, bringing about an altogether different arrangement.

2. RELATED WORK
All in all, there are two methodologies for outlining security protecting machine learning algorithms. The primary methodology is to utilize changes to annoy the information set some time recently the algorithm is connected. This methodology for planning security saving grouping algorithms is taken by a few analysts [6]. A second way to deal with planning protection safeguarding algorithms is to utilize algorithms from the safe multiparty calculation writing. The benefit of this methodology over the irritation methodology is that formal certifications of protection can be given for these algorithms.

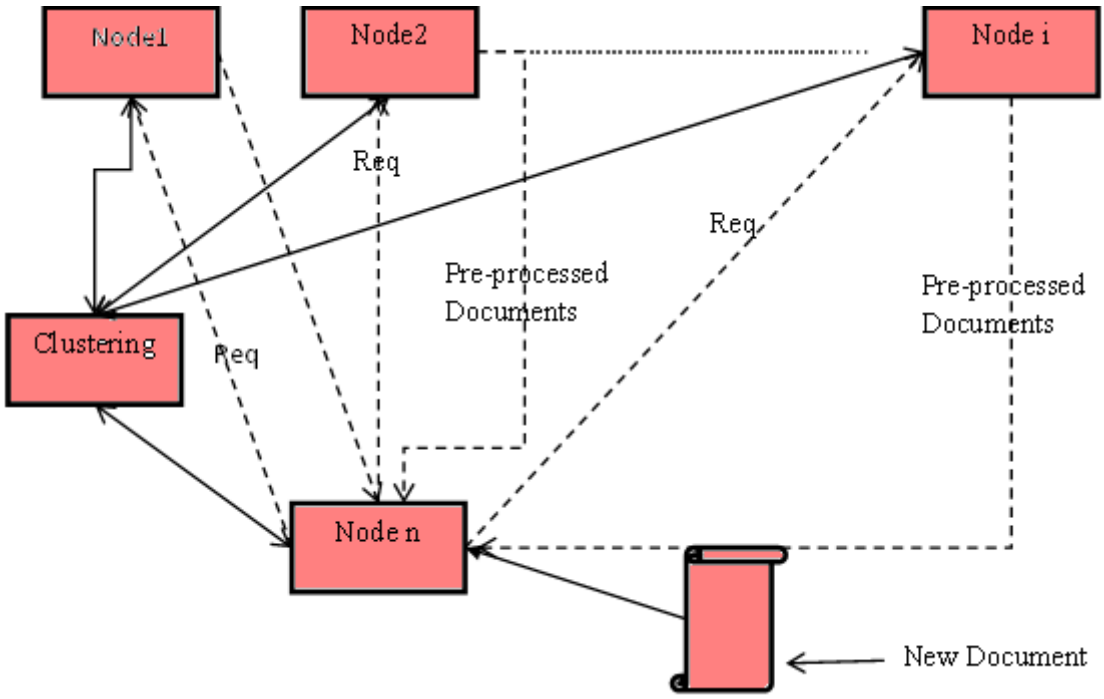
Our system takes after the fundamental methodology sketched out but that values are gone between the nearby information mining destinations as opposed to a brought together combiner. The two stages are finding hopeful item sets (those that are continuous on one or more destinations), and figuring out which of the competitor item sets meet the worldwide bolster/certainty limits.

Past work in security safeguarding information mining has tended to two issues. In one, the point is protecting client security by twisting the information values [4]. The thought is that the contorted information does not uncover private data,

and in this way is "sheltered" to use for mining. The key result is that the misshaped information, and data on the dispersion of the irregular information used to mutilate the information, can be utilized to create a rough guess to the first information dispersion, without uncovering the first information values. The conveyance is utilized to enhance mining results over mining theis shaped formation straightforwardly, principally through determination of split focuses to "canister" nonstop information. Later refinement of this methodology fixed the limits on what private data is revealed, by demonstrating that the capacity to remake the dispersion can be utilized to fix assessments of unique qualities in light of the misshaped information [7].

3. PROPOSED SYSTEM

In this paper, we are proposing an effective and secure data mining method with K Medoids and cryptographic approach for clustering the similar type of information. Initially the data points need to be shared the information which is at the individual data holders or players.



Here, we are stressing on mining approach not on cryptographic technique. For secure transmission of data, various cryptographic algorithms and key exchange protocols are available. The above diagram shows the architecture of proposed work. The individual peers at data holders are initially preprocess the raw data by eliminating the unnecessary features from datasets. The feature set in terms of term frequency and inverse document frequencies are computed after preprocessing the datasets. The file relevance matrix was used to reduce the time complexity while clustering the datasets. We are using the cosine similarity method which is the most widely used similarity measurement i.e. we try to find the key size of DES to guard beside brute force attacks, without needing a entirely new block cipher algorithm.

In the above table D(d1,d2….dn) represents set of documents at data holder or player and their respective cosine similarities, while computing the similarity between the centroids and documents, the time complexity is reduced.

In our approach, we are implementing K Medoids algorithm with recentoird computation instead of single random selection at every iteration,

$Cos(d_m,d_n)= (d_m * d_n)/Math.sqrt(d_m * d_n)$

Where

	d1	d2	d3	d4	d5
d1	1.0	0.77	0.45	0.32	0.67
d2	0.48	1.0	0.9	0.47	0.55
d3	0.66	0.88	1.0	0.77	0.79
d4	0.89	0.67	0.67	1.0	0.89
d5	0.45	0.88	0.34	0.34	1.0

Fig2: Similarity Matrix

Algorithm:

We have a objects having b variables that will be classified into clusters. Let us define i-th variable of object.

The algorithm comprises of three steps with the condition, k < a X(i =1…a ,j=1…..b.)

Step 1: Here the selection of initial Medoids was performed.

1-1. Compute the distance between every pair of all objects by using Euclidean distance as a dissimilarity measure

1-2. After calculating at each object, sort them in ascending order. Select the objects having minimum value as initial group Medoids and assign each object to the nearest Medoids.

1-3. Calculate the sum of distance from all objects to their Medoids and also find the current optimal value,

Step 2: Here the new Medoids were found by replacing the current Medoids in each cluster by minimizing the total distance to other objects in its cluster.

Step 3: At first, the each object to the nearest new Medoids was assigned. Calculate the new optimal value and the sum of distance from all objects to their new Medoids. If the new optimal value is equal to the previous one, then stop the algorithm. Otherwise, go back to the Step 2.

Triple DES:

Triple DES is the Triple Data Encryption Algorithm (TDEA) block cipher. It applies the Data Encryption Standard (DES) cipher algorithm with three options,

The three keying options defined are:

- Keying option 1: All the three keys defined are independent.
- Keying option 2: Here, K1 and K2 are independent, and K3 = K1.
- Keying option 3: All the three keys defined are identical, i.e. K1 = K2 = K3.

Keying option 1 is the strongest, with 3 x 65 = 195 independent key bits.

Keying option 2 provides less security, with 2 x 55 = 110 key bits. This option is stronger than simply DES encrypting twice, e.g. with K1 and K2.

Keying option 3 is no better than DES, with only 65 key bits. Because of the first and second DES operations simply cancel out, the keying option 3 provides backward compatibility with DES. In general, Triple DES with three independent keys (keying option 1) has a key length of 195 bits (three 65-bit DES keys), but because of the meet-in-the-middle attack, the effective security it provides is only 110 bits. Here the Keying option 2 reduces the key size to 110 bits. However, this option is vulnerable to certain attacks like chosen-plaintext attacks or known-plaintext attacks.

4. CONCLUSIONS

The proposed method gives us the productive protection saving information grouping over disseminated systems. The nature of grouping instrument is improved with preprocessing, importance grid and centroid calculation in K-Medoids algorithm. The present cryptographic system describes the protected transmission of information between the information holders and information recoveries. We can improve the security by building up an effective key trade convention and cryptographic methods while transmission of information between information holders or players.

5. REFERENCES

[1] Privacy Preserving Decision Tree Learning Using Unrealized Data Sets Pui K. Fong and Jens H. Weber-Jahnke, Senior Member, IEEE Computer Society.

[2] Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang.

[3] Anonymization of Centralized and Distributed Social Networks by Sequential Clustering Tamir Tassa and Dror J. Cohen.

[4] Privacy Preserving Clustering S.Jha, L. Kruger, and P. McDaniel.

[5] Tools for Privacy Preserving Distributed Data Mining , Chris Clifton, Murat Kantarcioglu, Xiaodong Lin, Michael Y. Zhu.

[6] S.Datta, C. R. Giannella, and H. Kargupta, "Approximate distributed K-Means clustering over a peer-to-peer network," *IEEETKDE*, vol. 21, no. 10, pp. 1372–1388, 2009.

[7] K. M. Hammouda and M. S. Kamel, "Hierarchically distributedpeer-to-peer document clustering and cluster summarization,"*IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 5, pp. 681–698, 2009.

[8] M. Eisenhardt, W. Muller, and A. Henrich, "Classifying documents by distributed P2P clustering" in *INFORMATIK*, 2003.

[9] H.-C. Hsiao and C.-T.King, "Similarity discovery in structured P2P overlays," in *ICPP*, 2003.

[10] Privacy Preserving Clustering By Data Transformation Stanley R. M. Oliveira, Osmar R. Za¨ıane.