FCMRSA: Privacy Preservation with Clustering and Cryptographic Technique in Data Mining

Neha Garg  
M.Tech Research Scholar  
Department of Computer Science and Engineering  
Laxmi Devi Institute of Engineering & Technology  
Alwar, India

Mr. Hemant Pareek  
Assistant Professor of Computer Science and Engineering  
Laxmi Devi Institute of Engineering & Technology  
Alwar, India

Abstract— Data Mining can be one of the technologies used to extract knowledge from massive collection of data. On the other hand, being published, the sensitive information about individuals may be disclosed which create ethical or privacy issues. Due to privacy issues many individuals are reluctant to share their data to the public which leads to data unavailability. Thus, privacy should be an important concern in the field of Data Mining. Privacy Preserving Data Mining (PPDM) is becoming a popular research area to address various privacy issues. Clustering is a technique which makes cluster of useful objects which have resemble characteristics. Here, fuzzy c-means clustering algorithm is used. By the proposed work, privacy preservation of the data increased and it can be shown with the help of the results. RSA provide the result in minimum time which show that purpose produce result faster than existing approaches.

Keywords—Privacy Preserving Data Mining, Anonymization, Clustering, Fuzzy c-means clustering, RSA, encryption, decryption.

I. INTRODUCTION

Privacy Preserving Data mining techniques deal with the secure data publication or communication without revealing the private and sensitive information about any individual. Anonymization procedure has been considered as a standout amongst the best strategies since it can give the better tradeoff between information utility and privacy preservation [1].

Despite the fact that data mining is possibly helpful, numerous data holders are hesitant to give their information to data mining for the dread of abusing singular privacy. In recent years, study has been made to ensure that the sensitive information of individuals cannot be identified easily. The data holder in PPDM wants to randomize the data so that data mining outcome can be recuperated from the randomized data. To do as such, the data holder have to comprehend the data mining errands and algorithms included [2].

PPDM is utilized for the extraction of valuable knowledge from the huge measure of data, while ensuring the delicate data concurrently.

II. ANONYMIZATION IN PPDM

Anonymization alludes to a methodology where identity or/and sensitive data about record proprietors are to be covered up. Anonymization strategy goes for influencing the individual record to be indistinct among gathering records by using procedures of generalization and suppression. Distinctive properties in a data set may assume diverse jobs in either encouraging identification or facilitating sensitive information release. There are three fundamental sorts of Attributes:

(i) Explicit Identifier is an arrangement of attributes containing data that recognizes a record proprietor expressly, for example, name, SS number and so on.
(ii) Quasi Identifier is an arrangement of attributes that could possibly recognize a record proprietor when joined with freely accessible data.
(iii) Sensitive Attributes is an arrangement of attributes that contains sensitive person particular data, for example, infection, pay and so forth [3].
III. CLASSIFICATION OF CLUSTERING

Clustering [4] is a standout amongst the most fundamental methodology for managing the monstrous measure of research articles from an archive. Clustering algorithms have been utilized to bunch comparable articles into one bunch.

a) Hierarchical algorithms

Hierarchical clustering is a strategy for group investigation which tries to manufacture a hierarchy of groups. It is the availability based grouping algorithms. The hierarchical algorithms construct bunches step by step. Hierarchical clustering usually drop into two kinds: In hierarchical clustering, in single step, the data containing all objects to n clusters each containing a single is not divided into a specific bunch. It takes a progression of segments, which may keep running from a solitary group containing all items to n bunches each containing a solitary object. Hierarchical Clustering is subdivided into agglomerative strategies, which continue by an arrangement of combinations of the n objects into groups, and disruptive techniques, which discrete n objects successively into finer groupings.

b) K-Means algorithms

K-Means clustering is one of the partitioning algorithms which is widely used in the data mining. This clustering partitions n documents with regards to content data into k groups. It is a simple method to group a given article dataset through a specific no. of clusters (suppose k clusters) predetermined from the earlier. The principle thought is to characterize k centroids, one for each group. The following stage is to take each point belong toward a given article informational index and partner it to the closest centroid. At the point when no point is pending, the initial step is finished and an early gathering age is finished. Now, re-ascertain k new centroids coming about because of the past step. After figuring of k new centroids, another binding must be done between a similar article informational index focuses and the closest new centroid. A loop has been produced. Because of this loop one can see that the k centroids change their area well ordered until the point when no more changes are finished.

c) Fuzzy C-Means algorithms

Fuzzy C-Means (FCM) is a standout amongst the most famous fuzzy clustering algorithms. FCM is a strategy for grouping that allows membership levels and utilizes them to relegate information components to at least one clusters. It utilizes proportional distance to process fuzzy weights. Each component of the universe can belongs to any fuzzy set with a level of membership that shifts from 0 to 1. FCM presents the fuzziness for the belongingness of every object and can hold more data of the dataset. This calculation works by allocating membership to every data point of article dataset relating to each group center based on a distance between the cluster center and the data point. After every emphasis membership and cluster centers are modernized.

IV. RSA CRYPTOGRAPHIC ALGORITHM

The most noteworthy methodology of public key cryptography algorithm is RSA [5], which can oppose all the well-known passwords attacks up until this point. RSA algorithm, which is named after the creators, is the primary algorithm that can be utilized both for data encryption and digital signatures RSA algorithm's security relies upon the trouble of disintegration of vast numbers. In the algorithm, two huge prime numbers are utilized for building the public-key and the private-key. It is evaluated that the trouble of speculating the plaintext from signal key and the cipher text equivalents to that decay of the result of two extensive prime numbers. RSA algorithm has been used as possible authentication methods in ISAKMP / Oakley framework.

RSA cryptosystem utilizes the mode n, the littlest nonnegative finish the rest of the lines of the task, where n is the result of two distinct primes p and q. RSA is portrayed as following:

To start with, the creation process of keys is as per the following,

1) Randomly produces two primes P and Q of length K/2 bit
2) Calculate the public key public Key=P*Q; (public Key's length is k-bit)
3) Generate a random encryption key keyE, 2<=keyE<=Φ(n)-1, where GCD(keyE,Φ(n))=1;

This is the necessary and sufficient conditions for solvability of the decryption key keyE *keyD mod Φ(n)=1,Φ(n) is known as the Euler function of n, the value is Φ(n) = (P1)*(Q-1)

4) Calculate the decryption key, keyD=keyE-1 mod (n) , keyE-1 is inverse for the decryption key keyD. The formula of the original equation is keyE * keyD mod Φ(n)=1

Now, the public key, encryption key and decryption key are all created.
Then, the process of encryption of the plaintext and decryption of ciphertext is as follows:
1) Encryption: \( C = M \text{key}E \mod \text{publicKey} \); where \( M \) is plaintext, \( C \) is ciphertext.

2) Decryption: \( M = C \text{key}D \mod \text{publicKey} \); in which \( M \) plaintext, \( C \) is ciphertext.

V. LITERATURE SURVEY

Chunhui Piao et al. [2018] To deal with the issues of high delay, low unwavering quality and strong congestion in the administration cloud, a differential privacy structure for distributing legislative statistical data based on a government cloud-fog hybrid computing model is proposed. In light of the system, a data publishing algorithm utilizing the MaxDiff histogram is proposed, which is utilized to understand the capacity of client privacy given by fog processing. Hypothetical examination and trial results demonstrate that the proposed privacy protection data publishing algorithm can decrease query sensitivity and successfully enhance the viability of published data [6].

Lingjuan Lyu et al. [2018] displayed a 2-stage privacy-preserving collaborative fuzzy clustering plan. The randomization based plan annyois information in two phases: the primary, nonlinear stage foils Bayesian estimation assault and mitigates collusion attack, though the second, straight stage opposes autonomous part analysis attack. In the main stage, every member irrigates his/her information by going the information through a nonlinear capacity called rehashed Gompertz (RG); in the second stage, he/she at that point extends his/her annoyed information to a lower measurement in a (nearly) distance-preserving manner, utilizing a particular random projection (RP) matrix. The nonlinear RG work is intended to mitigate maximum a posteriori (MAP) estimation attacks, while random projection opposes independent component analysis (ICA) attacks and guarantees clustering accuracy [7].

Igor V. Anikin and Rinat M. Gazimov [2017] recommended the alteration of DBSCAN clustering algorithm for vertically divides big data. This modification provides privacy of the information which is transmitted via telecommunication channels between nodes of big data storage. This modification is founded on organization of secure channel which is based on homomorphic and asymmetric encryption. The first one is used for hiding the information and the second one is used for creation the encryption channel. Suggested privacy preserving DBSCAN clustering algorithm provides data confidentiality at all steps of DBSCAN algorithm [8].

Alabatelatif et al. [2017] presents a realistic system that exploits cloud assets to give a lightweight and adaptable privacy preserving inconsistency discovery benefit for sensor information. A lightweight Homomorphic Encryption conspire is utilized to guarantee information security and protection with any computational impediments defeat through an advantageous data processing a model that utilizes a solitary private server teaming up with an arrangement of open servers inside a cloud server centre. Virtual nodes actualized on open servers perform granular anomaly detection activities on encoded information. Far-reaching experimentation exhibits reliably high identification exactness with less overhead in a cloud-based anomaly detection display that is both lightweight and adaptable while guaranteeing data privacy [9].

J. Jesu Vedha Nayahi and V. Kavitha [2017] in this paper, an anonymization algorithm dependent on bunching and flexible to likeness attack and probabilistic derivation attack is proposed. The anonymized data is dispersed on the Hadoop Distributed File System. The strategy accomplishes a superior exchange off among privacy and utility. In our work, the information utility is estimated as far as precision and F-Measure concerning distinctive classifiers. Analyses demonstrate that the exactness, F-Measure and the execution time of the classification algorithms on the privacy-protected informational indexes framed by the proposed grouping calculations are superior to the current calculations [10].

Jiawei Yuan et al. [2016] anticipated a realistic privacy-preserving K-means clustering method that can be effectively redistributed to cloud servers. Our plan permits cloud servers to perform bunching specifically over encoded datasets while accomplishing practically identical computational multifaceted nature and exactness contrasted and clusterings over decoded ones. We likewise examine secure incorporation of Map-Reduce into our plan, which makes our plan to a great degree appropriate for distributed computing condition. Intensive security examination and numerical investigation do the execution of our plan as far as security and proficienciy. Experimental evaluation over a 5 million objects dataset further validates the practical performance of our scheme [11].

Vadlana Baby et al. [2016] in this paper they proposed an efficient distributed threshold privacy-preserving k-means clustering algorithm that use the code based threshold secret sharing as a privacy-preserving mechanism. Construction involves code based approach which allows the data to be divided into multiple shares and processed separately at different servers. Our protocol takes less number of iterations compare with existing protocols and it do not require any trust among the servers or users. We furnish analyze results with correlation and security examination of the proposed plan as well [12].

VI. PROPOSED METHODOLOGY

In the initial step, we get the data from the database in which the operations can be performed. The overall dataset has been fetched from the database and divides into group which is known as clusters. For the clustering procedure, fuzzy c-means clustering has been utilized. This algorithm works by doing out membership to every data point relating toward each bunch fixate on the premise of distance between the cluster center and the data point. Progressively the information is close to the cluster center more
is membership towards the specific cluster center. Plainly, the summation of participation of every datum point ought to be equivalent to one. After every cycle, membership and cluster centers are simplified utilizing objective function. Now the data has been encrypted using RSA Algorithm where it contains various operations for the key generation and then decryption has performed and gets the optimal results obtained.

Proposed Algorithm:

Step: 1 Start
Step: 2 Input dataset from the database
Step: 3 Apply Fuzzy C-Means clustering
   a. Choose ‘c’ cluster centers
   b. Compute fuzzy membership by
   \[ \mu_{ij} = 1 / \left( \sum_{k=1}^{d} \left( \frac{d_{ij}}{d_{ik}} \right)^{(2/m-1)} \right) \]
   c. Allocate coefficient haphazardly to every data points. Figure fuzzy centers by
   \[ v_j = \left( \sum_{i=1}^{n} \left( \mu_{ij} \right)^{m} x_i \right) / \left( \sum_{i=1}^{n} \left( \mu_{ij} \right)^{m} \right), \]
   \[ \forall j = 1, 2, ..., c \]
   d. Repeat (b) & (c) until the algorithm has converged (i.e. the minimum ‘J’ value achieved)
   e. Compute the centroid for each cluster
   \[ c_k = \frac{\sum_{x} W_k(x)^{m} x}{\sum_{x} W_k(x)^{m}} \]
   f. For every data point, calculate its coefficients of being in the clusters.
   g. Foremost intention of fuzzy c-means algorithm is to reduce:
   \[ J(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{c} \left( \mu_{ij} \right)^{m} \| x_i - v_j \|^2 \]
   Where,
   \[ \| x_i - v_j \| \] is the Euclidean distance b/w ith and jth cluster center.
   \[ \mu_{ij} \] is the fuzzy membership.
   \[ v_j \] is fuzzy center.
   \[ J \] is the objective function.

Step: 4 Apply RSA Algorithm over the output
   a. Randomly Generate 2-prime no.
   b. Calculate public key
   c. Randomly generate Encryption key
   d. Find Decryption key
   e. Perform Encryption
   f. Perform Decryption

Step: 5 Get optimal result
Step: 6 Exit

Fig. 3 Flowchart of Proposed Work

VII. RESULT ANALYSIS

The simulation of the proposed work has done with MATLAB 2018. Here, experiments have performed on number of records at different-different values. From this, we achieved encrypted and decrypted information. There are two graphs demonstrated below which show that the proposed technique has better accuracy and less error rate. Elapsed Time line graph also demonstrated that time has taken by proposed method is lesser than the existing method.

Table I: Comparison of Elapsed Time between existing and Propose Techniques

<table>
<thead>
<tr>
<th>No. of records</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Existing</td>
<td>23.28</td>
<td>42.44</td>
<td>63.77</td>
<td>81.85</td>
<td>109.10</td>
</tr>
<tr>
<td></td>
<td>866</td>
<td>951</td>
<td>27</td>
<td>91</td>
<td>15</td>
</tr>
<tr>
<td>Propose</td>
<td>3.448</td>
<td>7.029</td>
<td>10.028</td>
<td>12.960</td>
<td>15.516</td>
</tr>
<tr>
<td></td>
<td>806</td>
<td>58</td>
<td>13</td>
<td>48</td>
<td>91</td>
</tr>
</tbody>
</table>
Fig. 4: Select dataset

Fig. 5: Encrypted Information

Fig. 6: Decrypted Information

Fig. 7: Final Result

Fig. 8: Accuracy comparison graph

Fig. 9: Error rate among the base and propose approach
Conclusion

Data mining alludes to the way toward recovering knowledge by finding new and relative patterns from big datasets. Privacy-preserving data mining regard as the problem of running data mining algorithms on private information that should be uncovered even to the gathering running the algorithm. PPDM attempt to conquer this issue by securing the privacy of information without giving up the integrity of information. The goal of this paper is too discussed about the clustering with the introduction of clustering and RSA algorithm on privacy preserving technique which is helpful in mining large amount of data with reasonable efficiency and security.

References

[11] Jiawei Yuan, Yifan Tian, Student "Practical Privacy-Preserving MapReduce Based K-means Clustering over Large-scale Dataset" 2168-7161 (c) 2016 IEEE.