iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VIII **Month of publication:** August 2017

DOI: http://doi.org/10.22214/ijraset.2017.8226

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VIII, August 2017- Available at www.ijraset.com

A Proposed Technique for Privacy Preservation by Anonymization Method Accomplishing Concept of K-Means Clustering and DES

Priyanka Pachauri[1], Unmukh Datta[2]
[1]*Dept. of CSE/IT, MPCT College Gwalior, India*
[2]*Prof. Dept. of CSE/IT, MPCT College , Gwalior, India*

Abstract: ***Privacy-Preserving Data Mining also an essential branch of the data mining and an exciting topic in privacy preservation has gain particular attention in current years. Data mining has been substantially studied and useful into numerous fields which include the Internet of Things and the business growth. However, data mining tactics additionally take vicinity critical demanding situations due to enlarged sensitive statistics disclosure and the violation of privacy. This discussion describes the privacy concern that occurs due to data mining, particularly for the national security applications. We discuss PPDM by Anonymization Method in which we use K-means clustering in order to divide the given data and DES algorithm for encryption of data in order to prevent sensitive data from attacker.***

Keywords: privacy preservation, PPDM, K-meanS clustering, anonymization ,DES etc.

I. INTRODUCTION

Privacy is receiving extra attention partially as of counter-terrorism and the national security. At present we have heard so much approximately countrywide protection in media. This is mainly because people are now realizing that to handle terrorism, the government may need to collect information about individuals. There been much interest in the recent on applying the sa mining used to detect patterns which are unusual , terrorist activities and the fraudulent behavior. While all applications of data mining can give profit to the humans and save lives, there also negative side to this type of method , it could be a danger to the individuals privacy . This is due to data mining tools are present on the Web or, and even naive individuals can use these tools to mine information from stored data in various databases and files, and therefore violate the privateness of individuals. As we have stressed in papers to take out efficient data mining and mine necessary information for counter terrorism and national security, we gather all kinds of information about individuals.[1] However, this information could be a threat to individuals' privacy and civil liberties. This is causing major concern with different civil liberties unions. The aim is to carry out data mining and yet to the maintain privacy. This topic is known as privacy-preserving data mining .
In this paper we use anonymization technique along with K-means clustering and DES algorithm.

A. Anonymization

To guard identity of the individual when release of sensitive information is done , holders of data often being encrypt or eliminate explicit identifiers, like names and the unique security no: However, data which is unencrypted provides no assurance for anonymity. To preserve privacy, model of k-anonymity has been proposed by the Sweeney [2] which achieve k anonymity by means of generalization and the Suppression , K-anonymity, it is difficult for an imposter to decide the identity of the individuals in collection of data set containing personal information. For ex, the age of person might be generalized to a variety like youth, middle age and the adult with no specifying suitably, so as to decrease the threat of the identification. Suppression involves decrease the exactness of the applications and it don't liberate some information .By using this method it reduces the risk of detecting exact information.

B. K- Means Clustering Algorithm

K-mean is the greatest popular apportioning procedure of clustering. It turns out to be leading proposed by methods for Macqueen in 1967. K-mean is an unsupervised, non-deterministic, numerical, iterative technique for grouping. In k-mean each cluster is represented by using the suggest value of gadgets within the cluster.[3] Here we partition a hard and fast of n item into k cluster in order that intercluster similarity is low and intracluster similarity is excessive. Similarity is measured in time period of imply value of objects in a cluster. The set of rules consists of two separate levels.

1 st Phase pick k centroid randomly, where the value k is settled in advance.2 nd Phase each item in data set is identified with the closest centroid. Euclidean distance is used to measure the distance among every data object and cluster centroid.

Procedure

1) Arbitrarily pick out k data item from D dataset as preliminary cluster centriod.
2) Repeat
3) Assign every data item to the cluster to which object is most comparative construct absolutely with respect to the suggest estimation of the item in cluster;
4) Calculate the new suggest estimation of the data items for each cluster and refresh the mean esteem.
5) Until no trade.

C. DES(Data Encryption Standard)

Data Encryption Standard also called as (DES) algorithm has been all the rage secret key encryption algo and it is taken in use in lots of commercial and the financial applications. Though it was introduced in the year 1976, it has established resistant to all the forms of cryptanalysis. In addition, DES is a block cipher algorithm which means that it takes a fixed length of the message and encrypts it (encrypts the block), and returns a string in the same size. DES is the first encryption algorithm recommended by NIST (National Institute of Standards and Technology).

II. LITERATURE REVIEW

The main contributions of this paper are three folds: (i) the definition of the data collection and publication process, (ii) the privacy framework model and (iii) personalized anonymization approach. The experimental analysis is presented at the end; it shows this approach performs better over the distinct l-diversity measure, probabilistic l-diversity measure and k-anonymity with t-closeness measure [4].

The [5] proposes a singular, extra flexible generalization scheme. The experimental results of their study indicate that their approaches produce k – anonymization with less generalization compared to previous approaches. They conclude that a bottom-up approach for k – anonymization is preferable for small number of quasiidentifying attributes.

The k anonymity based method is illustrated in [6] is used to search for optimal feature set partitioning and [7] for cluster analysis. And [8] Proposes a data reconstruction approach to obtain k-anonymity safety in predictive data mining. In this method the probably identifying attributes are first mapped the usage of aggregation for numeric data and swapping for nominal data. A genetic set of rules technique is then implemented to the masked data to find a correct subset of it.

k-anonymity method is treated as the classical anonymization method and most of the studies are based on k-anonymity. The others are based on its improved methods like l-diversity, t-closeness, km -anonymization, (α,k) anonymity, p-sensitive k-anonymity, (k,e) anonymity, which are described in [9].

They provide a detailed survey of anonymization methods and also point out pitfalls in k anonymity. Previous works by Samarati and Sweeney [10,11] shows that the removal of the personally identifying information from data is insufficient for the data security, rather it is better to use k – anonymity method for publishing data. The quasi–identifier (QI), which is the combination of person specific identifiers are considered here for the process of anonymization. One of the common methods to achieve k –anonymity is to generalize identifiers (for example date of birth can be generalized to month of birth).

[12]There are various PPDM techniques such as anonymization, perturbation, randomization, condensation, cryptography. In this paper we have reviewed anonymization technique of PPDM such as k anonymity using generalization and suppression, p sensitive k anonymity, (α, k) anonymity, l diversity, m-invariance.

Sometimes the data should be publically published in its original form. Even though it is not encrypted and perturbed, some sort of precaution should be implemented before releasing the data in terms of anonymization [13]. This is a kind of generalization of some attributes which protects against identity disclosure. Anonymization can be obtained through techniques inclusive of generalization, suppression, data removal, permutation, swapping and so forth [13]

III. PROPOSED METHOD

The proposed algorithm is a try to present a new method for complex encrypting and decrypting data based totally on parallel programming in the sort of way that the new method can make use of multiple- core processor to acquire higher speed with better degree of protection.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VIII, August 2017- Available at www.ijraset.com

ALGORITHM

Partition of given dataset through the usage of K- means clustering
Summary ← partition
Dim ← choose dimension()
P ← frequency set(partition , dim)
Sv ← find median(P)
Lhs ← {t belongs to partition : t.dim <= sv}
Rhs ← {t belongs to partition: t.dim > sv}
Apply DES for encrypting the records.
Finally return union of lhs and rhs partition

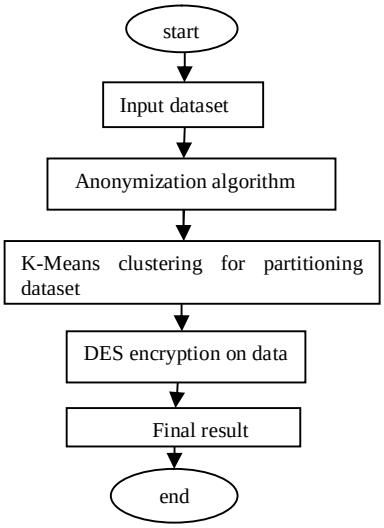


Fig. 1 Flow diagram Propose method

Procedure –

Step1 - consider the dataset for input.

Step2 - applies anonymization technique to that particular dataset.

Step3 - K-means clustering technique is used to partiton the data sets into clusters.

Step4 – DES encryption technique is used to suppress the data values.

Step5 – final result obtained by union of lhs and rhs values formed by anonymization technique.

IV. RESULT ANALYSIS

Table 1 describes the accuracy between Base method results and Propose method results.

No. of records	**100**	**200**	**300**	**400**	**500**
Accuracy in base results	84.00	83.50	84.00	84.00	83.80
Accuracy in proposed results	96.00	98.00	99.00	98.75	98.20

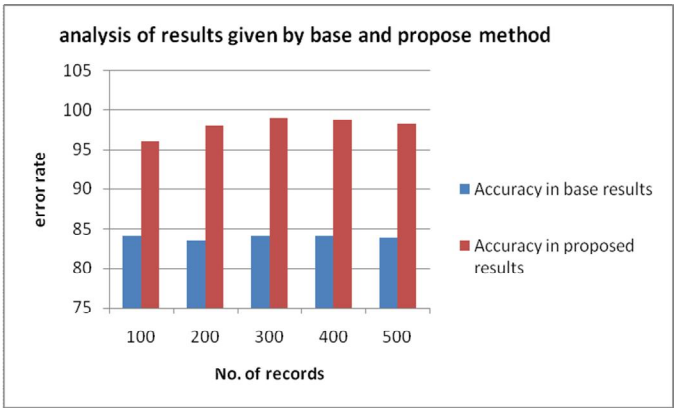


Fig.2 Above graph represents the analysis of accuracy in base and propose method results and concluded that propose method is best for preserving privacy.

Table II describes the error rate between Base method results and Propose method results.

No. of records	**100**	**200**	**300**	**400**	**500**
Accuracy in base results	84.00	83.50	84.00	84.00	83.80
Accuracy in proposed results	96.00	98.00	99.00	98.75	98.20

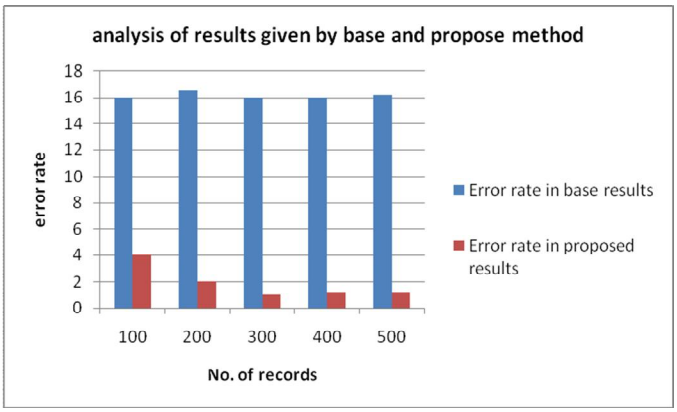


Fig.3 Above graph represents the analysis of error rate in base and propose method results and concluded that propose method is best for preserving privacy as the error rate in propose method is less than the error rate in base method.

V. CONCLUSION

Each privacy preserving technique has its own importance. Substantial efforts have been accomplished to address these needs. Data encryption and anonymization are widely adopted ways regarding privacy breach. However, encryption is not suitable for data that are processed and shared. Anonymizing huge data and dealing with anonymized data sets are nonetheless challenges for classic anonymization processes. Privacy- preserving data mining is emerged for to 2 critical desires: data analysis with a purpose to deliver better services and making sure the protection privileges of the data owners.

The results of our proposed work shows that by doing k-means clustering and encypting the data using DES method we can achieve more preservation of privacy.

REFERENCES

[1] Bhavani Thuraisingham," Privacy-Preserving Data Mining: Developments and Directions", IDEA GROUP PUBLISHING, Journal of Database Management, 16(1), 75-87, Jan-March 2005 77.

[2] Pingshui WANG," Survey on Privacy Preserving Data Mining", International Journal of Digital Content Technology and its Applications, Volume 4, Number 9, December 2010.

[3] Jyoti Yadav, Monika Sharma," A Review of K-mean Algorithm", International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 7- July 2013, ISSN: 2231-5381.

[4] M. Prakash G. Singarave "An approach for prevention of privacy breach and information leakage in sensitive data mining"2015 IEEE.

[5] Tiancheng Li, Ninghui Li, "Towards Optimal k-anonymization", Data & Knowledge Engineering, 2008 Elsevier.

[6] Nissim Matatov, Lior Rokach, Oded Maimon, "Privacy-preserving data mining: A feature set partitioning approach", Information Sciences 180 (2010) 2696–2720.

[7] Benjamin C. M. Fung, Ke Wang, Lingyu Wang, Patrick C.K. Hung, "Privacy-preserving data publishing for cluster analysis" , Data & Knowledge Engineering 68 (2009) 552–575.

[8] Dan Zhu, Xiao-Bai Li, Shuning Wu, "Identity disclosure protection: A data reconstruction approach for privacypreserving data mining", Decision Support Systems 48 (2009) 133–140.

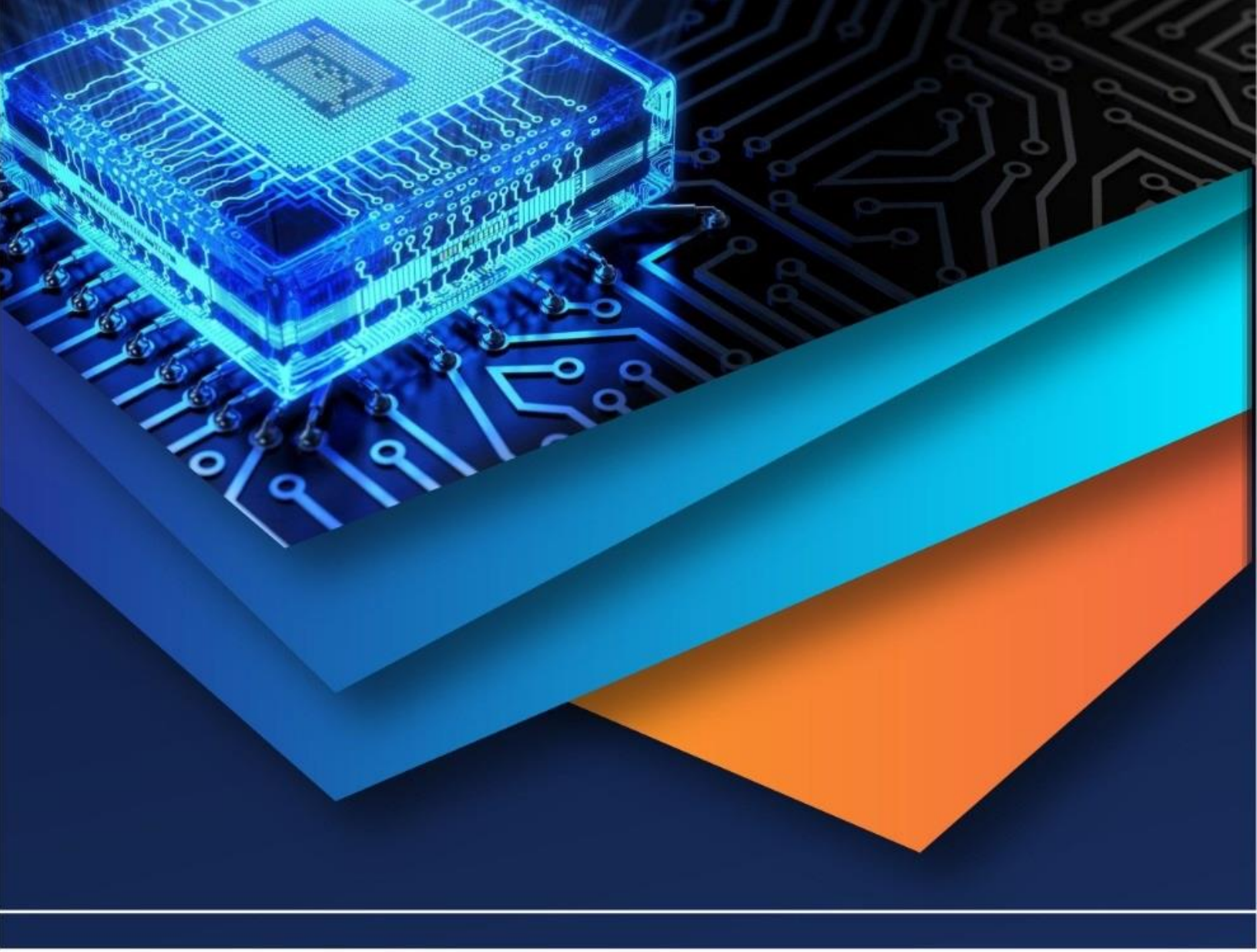
[9] Yan Zhao, Ming Du, Jiajin Le, Yongcheng Luo, "A Survey on Privacy Preserving Approaches in Data Publishing", First International Workshop on Database Technology and Applications, 2009.

[10] Samarati P, "Protecting respondent's privacy in Microdata release", IEEE Transactions on Knowledge and Data Engineering, 13:1010–1027.

[11] Sweeney L, "k-anonymity: A model for protecting Privacy", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557–570.

[12] Kiran Israni, Shalu Chopra," Survey on Anonymization Technique for Privacy Preserving Data Mining (PPDM)", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 11, November 2016, ISSN(Online): 2320-9801.

[13] Asmaa H.Rashid and Prof.dr. Abd-Fatth Hegazy, "Protect Privacy of Medical Informatics using K-Anonymization Model", IEEE Explore.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY