Improving Privacy Preservation by Anonymization, Hierarchical Clustering & Des

Jay Prakash Maurya
Lakshmi Narain College of Technology, Bhopal, M.P, India,
jpeemaurya@gmail.com

Dr.Vivek Richhariya
Lakshmi Narain College of Technology, Bhopal, M.P, India,
vivekrich@gmail.com

Tripti Saxena
Lakshmi Narain College of Technology, Bhopal, M.P, India,
triptisaxena16@gmail.com

Puneet Nema
Lakshmi Narain College of Technology, Bhopal, M.P, India, puneetn@lnct.ac.in

Abstract: *A privacy policy is an arrangement of principles that unveils a portion the ways a gathering accumulates, oversees, uncovers and uses customer's data. Privacy preserving data mining an essential trademark in the advancement and assessment of calculations is the recognizable proof of suitable evaluation criteria and the development of related principles. This paper proposes Privacy preservation of sensitive knowledge using association rules for hiding sensitive information regulation. This paper focuses on preservation of information using data mining by anonymization method using hierarchical clustering for categorization of sensitive data and DES for encryption of categorized sensitive data. Though anonymizing huge data is a challenge for classic processes, PPDM is emerged for two critical desires: data analysis with a purpose to deliver better services and making sure the privacy rights of the data owners. The Proposed method works on critical issues of sensitive data using complex encrypting and decrypting algorithm on multiple-core processor to acquire higher speed with better degree of protection. Proposed method shows the accuracy ranges 90-95% and error rate 2.5-3.0 for input data set of medical sensitive data of patients of different age groups. Clustering algorithm help to focus critical age group those have sensitive diseases. The work can be helpful for embedded devices for securing sensitive information from smart health devices in which privacy is not maintained.*

Index terms: *Privacy preservation, PPDM, Anonymization, DES, Association rule, Encryption, Decryption*

I. INTRODUCTION & PREVIOUS WORK

PPDM aims to prevent the user's private information when transferred between multiple parties. Association rule hiding provides a solution to maintain privacy on transferred information. PPDM algorithm is evaluated on privacy level, data quality, complexity and ding failure. In last decade PPDM started (Nguyen X, et al; 2012) [1], and in last few year large number of improvements have been, in this research area. PPDM helps to find pattern from the mined data at same time secrecy of sensitive information data accuracy should be maintained. Major problem with data are trust, quality, and malicious data mining , intrusion detection systems that must be considered in the context of crucial online databases, thereby making this task more complex. The primary challenges for PPDM association rule hiding are excessive, expensive, unique data hiding and ought to be sufficient for terribly huge datasets. Below table -1 shows different algorithms comparison and conclusion that was cumulatively done by different authors in past years.

Table 1 Comparison Table of different algorithm

TITLE	ALGORITHM	PARAMETER	CONCLUSION
Anonymization of the Centralized and Distributed Social Networks is via Sequential Clustering	Anonymization algo and the SaNGreeA algorithm are used for the sequential clustering	Clustering coefficient, Diameter, the Average distance, the Effective diameter, the Epidemic threshold	The offered sequential clustering algo for anonymizing social networks. Those algo produce nonymizations via the clustering with better utility.
Data Mining for the Privacy Preserving Association Rules Based on the Improved MASK Algorithm.	Data Perturbation and the Query Restriction (DPQR)	Multi-parameters perturbation	The privacy- preserving degree and efficiency of time is completed. The DPQR is suitable for Boolean records.
K-Anonymity for Crowd sourcing Database	K-Anonymity algorithm	No. Of Tuples And Data spaces are used for measure the overall performance of the system.	The Outperforms standard K-Anonymity approaches on holding the adequacy of crowd sourcing.
Privacy Preserving Decision Tree Learning is Using Unrealized Data Sets	Tree learning Algo, decision tree generation are used.	Temperature Humidity, the Wind Play.	The decision tree algorithm is good with other security safeguarding methodologies, for example, cryptography, for additional protection.

Published By:
International Journal of Engineering Research in Current Trends

Secure and Privacy Preserving Smartphone which is Based on Traffic Information Systems	KeyGen(n) algorithm	GSC which is (group signature center) Accuracy, Simulation, time stamp	A localization algorithm, which is suitable for the GPS location samples, and evaluated it through the realistic simulations.
On Design and Analysis of Privacy-Preserving the SVM Classifier	Data mining algo, Classification algorithm, kernal adatron algorithm and the data fly algorithm.	Taken a toll parameter, Kernal parameter are utilized to the quantify the execution of the framework.	PPSVC can accomplish comparable arrangement exactness to the first SVM classifier. By securing the delicate substance of support vectors.
Privacy-Preserving Gradient-Descent Methods	GA	Languages demonstrating smoothing parameters, weight parameters are utilized to gauge the execution of the framework	The secure constructing blocks are scalable and the proposed protocols allow us to decide a higher comfy protocol for the packages for every scenario.
A Data Mining Perspective in PPDM Systems	C5.0 data mining algorithm, Commutative RSA cryptographic algorithm	Area secured by roc, bend data set id, affectability, specificity-1	Overcomes the overheads arising because of key trade and key computation with the aid of adopting the cryptographic algorithm
Incentive Compatible Privacy-Preserving Data Analysis	Data analysis algorithms	Deterministically no cooperatively computable (DNCC).	Claim 5.1, the length of the last stride in a PPDA assignment is in DNCC, it is constantly conceivable to make the whole PPDA errand fulfilling the DNCC demonstrates.
Privacy and Quality Preserving Multimedia Data Aggregation for the Participatory Sensing Systems Outlier detection anomaly the detection algorithm, the secure hash algorithm	Outlier detection anomaly detection algorithm, secure hash algorithm.	Detection rate, data range. Indices, anomaly score	A general process for computing bounds on nonlinear privacy preserving data-mining (PPDM) approach with the applications to detection anomaly.

II. METHOD

This research study was conducted on research articles of time [2012-2019], a scope is founded that gives a relevance with privacy preservation of sensitive data with an approach of association rule hiding. The experiment was carried on data set healthcare and associated work assignments to peoples in different organization. The method of PPDM to extract data row and anonymization, on 5 different datasets are used to perform the execution of the proposed work and compare them with the existing work. One data file has been taken in which there are three attributes such as NS UID, NS Age and S disease. There are 15 records are used for different person with their age and disease related to them. Each dataset has different number of records and these are utilized to demonstrate the effectiveness of the proposed work.

1. Data file:

Sl_No.	NS_UID	NS_Age	S_Disease
1	55612	29	Cancer
2	55675	21	Flu
3	55627	25	Heart Disease
4	55646	43	Heart Disease
5	55672	48	Flu
6	55655	47	Cancer
7	55647	34	Heart Disease
8	55622	30	Flu
9	55634	36	Cancer
10	55685	55	Flu
11	55681	58	Flu
12	55694	72	Cancer
13	55698	65	Heart Disease
14	55688	59	Heart Disease
15	55690	65	Heart Disease

2. DATA100

S.No.	UID	Age	Sensitive Information
20	292175	65	Exec-managerial

40	265477	62	Prof-specialty
73	124744	79	Prof-specialty
80	51618	66	Other-service
98	124191	76	Exec-managerial

Query for Extracted input.

- (NS UID (preservation AND OF AND CONFIDENTIAL AND INFORMATION) AND NS UID (data and mining)) AND NS Age > 29
- (NS UID (rule and hiding AND approaches) AND NS UID(data AND association) AND NS UID(data and mining))

The extracted data was partitioned into clusters and analysis of clusters was done. Encryption was approached using DES having 16 rounds (not aligned), a mixer (no swapper), sixteen level L0-L15 for encryption and vice versa, K1-K16, 128 bit size. Initial and final permutation was invertible.

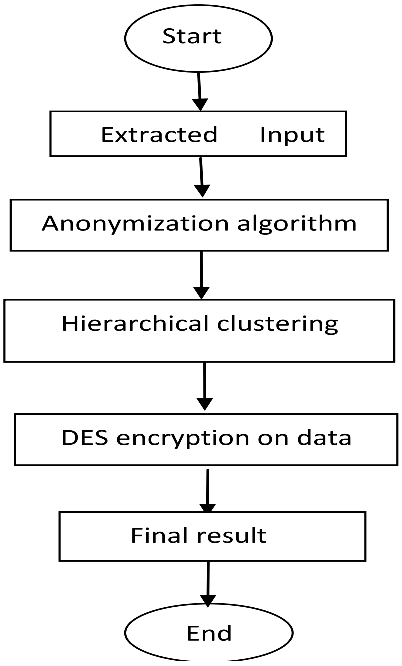
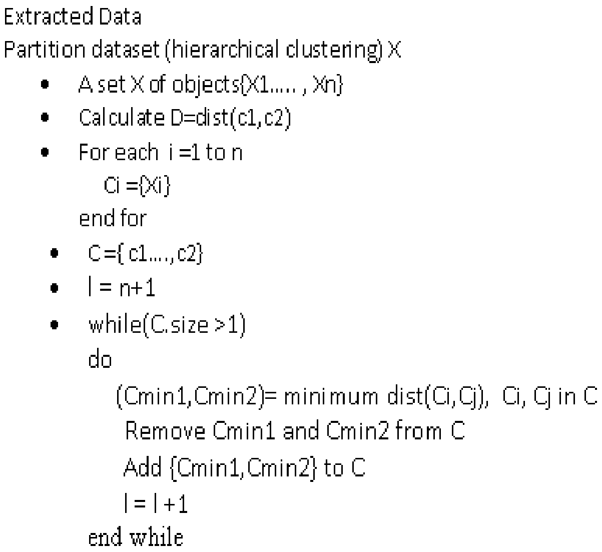


Figure 1. Flow Chart.

III. RESULT

There 5 different datasets are used to perform the execution of the proposed work and compare them with the existing work. One data file has been taken in which there are three attributes such as ***NS_UID, NS_Age and S_disease***. There records of different person with their age and disease related to them. Each dataset has different number of records and these are utilized to show the effective nature of the proposed work. The best model selection is based in how accurate a classier predicts the labels of unlabeled instances of data objects [Table 2]. Error rate is calculated for base and proposed work [Table 3]. Implementation of Proposed work on MATLAB replacing PPDM without anonymization with improved DES configuration as discussed in method.



```
Extracted Data
Partition dataset (hierarchical clustering) X
  • A set X of objects{X1..... , Xn}
  • Calculate D=dist(c1,c2)
  • For each i =1 to n
        Ci ={Xi}
    end for
  • C={ c1....,c2}
  • l = n+1
  • while(C.size >1)
    do
        (Cmin1,Cmin2)= minimum dist(Ci,Cj), Ci, Cj in C
        Remove Cmin1 and Cmin2 from C
        Add {Cmin1,Cmin2} to C
        l = l +1
    end while
```

Figure 2.Clustering Algorithm.

Apart from the implementation results on error rate and accuracy conclusion of various literature reading states description [Table 1] on different attribute direct towards implementation of future research in PPDM.

Table 2 Accuracy of classifier for Base and Proposed work

No. of records	100	200	300	400	500
Accuracy in base results	84.00	83.50	84.00	84.00	83.00
Accuracy in proposed Results	96.00	98.00	98.00	97.500	97.00

Table 3 Accuracy of classifier for Base and Proposed work

No. of records	100	200	300	400	500
Error rate in base results	16.00	16.50	16.00	16.00	16.20
Error rate in proposed results	4.00	2.00	2.00	2.50	3.00

IV. CONCLUSION

The main aim of this paper is to make the research community intimate the current state-of-art in relevant field and the future of previous techniques can get improved. The work done in this paper provides new issue for better understanding about the growth of this field related to PPDM. Authors have evaluated different association rule hiding algorithms on different parameters like efficiency, scalability, privacy level, hiding failure, and quality of data. PPDM is applicable to different Data Mining fields like classification, clustering, association rule hiding, etc. The aim to improve PPDM for association rule hiding the latest decade, works great in field of information science.

V. REFERENCES

[1] Nguyen XC, Le HB, Cao TA (2012). An enhanced scheme for privacy-preserving association rules mining on horizontally distributed databases. In: Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF) IEEE, pp: 1-4.

[2] Doganay MC, Pedersen TB, Saygin Y, Savaş E, Levi A (2008). Distributed privacy preserving k-means clustering with additive secret sharing. In: Proceedings of the 2008 international workshop on Privacy and anonymity in information society ACM, pp: 3-11.

[3] Moustakides G V and Verykios V S (2008). A maxmin approach for hiding frequent itemsets. Data and Knowledge Engineering 65(1):75–89.

[4] Adhvaryu R, Domadiya N (2012). An Improved EMHS Algorithm for Privacy Preserving in Association Rule Mining on Horizontally Partitioned Database. In: Security in Computing and Communications Springer Berlin Heidelberg, pp: 272-280.

[5] Aggarwal CC, Philip SY (2004). A condensation approach to privacy preserving data mining. In: Advances in Database Technology-EDBT Springer Berlin Heidelberg, pp. 183-199.

[6] Moustakides G V and Verykios V S (2006). A max–min approach for hiding frequent itemsets. In: Workshops Proceedings of the 6th IEEE International Conference on Data Mining (ICDM), pp: 502–506.

[7] Bogdanov D, Talviste R, Willemson J (2012). Deploying secure multi-party computation for financial data analysis. In: Financial Cryptography and Data Security Springer Berlin Heidelberg, pp:57-64.

[8] Dnyanesh P, Akhtar WS, Loknath S, TN R (2012). Perturbation Based Reliability And Maintaining Authentication In Data Mining. In: International Conference on Advances in Computer and Electrical Engineering, pp: 59-63.

[9] Li G, Wang Y (2012). A Privacy-Preserving Classification Method Based on Singular Value Decomposition. In: Int. Arab J. Inf. Technol.: 9(6):529-34.

[10] Li G, Xi M (2015). An Improved Algorithm for Privacy-preserving Data Mining Based on NMF. In: Journal of Information & Computational Science, 12(9), pp: 3423–3430.

[11] Domadiya NH and Rao UP (2013). Hiding sensitive association rules to maintain privacy and data quality in database. In: Advance Computing Conference, IEEE, pp: 1306-1310.

[12] Gaitán-Angulo M., Cubillos Díaz J., Viloria A., Lis-Gutiérrez JP., Rodríguez-Garnica P.A. (2018) Bibliometric Analysis of Social Innovation and Complexity (Databases Scopus and Dialnet 2007–2017). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer