xIlkogretim Online - Elementary Education Online,Year; Vol 20 (Issue 5): pp.2402-2408
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.05.262

A Novel Anonymity Algorithm for Privacy Preservation

Dr. Dhaval Jadhav, Assistant Professor, Vidyabharti Trust College of MCA, Bardoli, Dhaval.jadhav@vbtmca.ac.in
Dr. Ronak Panchal, Assistant Professor, Vidyabharti Trust College of MCA, Bardoli, dr.ronak.k.panchal@gmail.com

Abstract: Nowadays, data mining techniques play a major role in obtaining useful data from large amounts of data. Released information may include personal or sensitive information. It is necessary to protect this sensitive information from unauthorized access. This is why, privacy protection becomes an important part of data mining. Over the years, various privacy protection strategies have evolved to protect personal information. After all, Anonymization is one of the most important ways to maintain privacy. A variety of common anonymous methods are used to maintain privacy, however these methods have some flaws in maintaining personal privacy. The Framework for Efficient Anonymous Algorithm is therefore proposed here. Initially, the proposed algorithm aims to identify critical and sensitive data characteristics using the Principal Component Analysis based Attribute Selection Algorithm. The algorithm estimates Eigen values and Eigen vectors. Over time, the process of encryption was performed by introducing the Novel Based Anonymity Algorithm (NBA). Finally anonymous information is available that prevents unauthorized access to personal information. To evaluate the performance of the proposed algorithm many factors such as data usage, privacy levels and computer costs are compared to existing systems. From the experimental analysis, the effectiveness of the proposed system proves its superiority compared to other strategies.

Index Terms—Data Mining, Privacy Preservation, K-Anonymity.

I. INTRODUCTION

The people are aware of the privacy of their personal data and are reluctant to share their most sensitive information. Data Mining Privacy (PPDM) is the process of keeping personal data confidential or sensitive information without the use of data loss.

In day-to-day life, advances in information technology play a major role in leading to greater data retention. Extracting information from these large repositories requires a better way to make better decisions. The discovery of these information resources is done through a data mining process [1] .It is one of the main processes for retrieving information from a database. This data contains general sensitive information about certain individuals such as financial and health information which is disclosed mainly to many groups such as, users, owners, collectors and miners. This availability of large amounts of data has the potential to read more personal details.

For this purpose the concept of privacy preservation has been introduced which is considered a significant concern in data mining [2]. It is referred to as privacy protection for sensitive or personal information without sacrificing data usage. Due to privacy issues, users are often reluctant to share their sensitive information. In recent days, confidentiality has become increasingly common as the ability to store data has increased. The main purpose of this data mining privacy is to extract relevant information from a large amount of data and data considered during the protection period.

II. PROPOSED METHOD

This section describes the operational process for the NEAA privacy protection program. The flow of the projected system is shown in Fig. 1. Initially the installation data is available in a database where personal

information about individual users is stored. This input data is also processed by continuing with other processes. Thereafter the used information is stored in a database. From pre-processed data, attributes are selected based on the main object algorithm for object analysis. Then critical and sensitive qualities are acquired based on selected traits. The anonymity of each attribute is determined using an algorithm of the novel based on the unknown. Eventually unknown details are available as a result.

Initially the input data is available in a database containing user information. The process of sound removal and special letter removal is done in this preparatory step. The database may also contain various types of data such as letters, series, numeric values, etc. This unstructured data is converted into a structured format using a forwarding process.

In this way, different types of data are converted into numerical values using ASCII code. The value of each data can be processed up to 2 digits and this can be obtained as pre-processed data. Thereafter the used data can be stored in a database to continue with other processes. Thereafter the matrix for the covariance of the corresponding values in the data tested using,

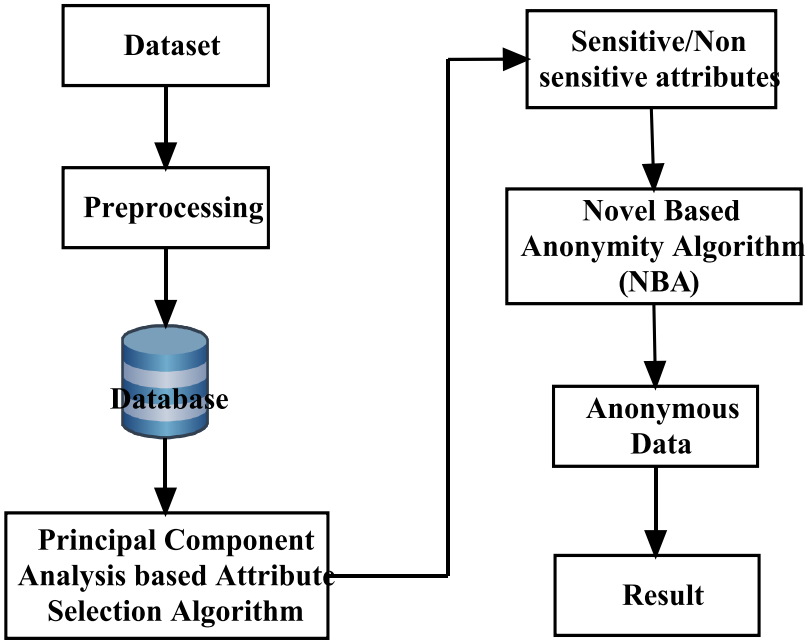
$$CVM_{X,Y} = {}^{X_i Y_i}/_{N} \qquad (1)$$

Where N = Number of scores in each set of data,
X_i = i^{th} raw score in the first set of scores,
Y_i = i^{th} raw score in the second set of scores,
$CVM_{X,Y}$ = Covariance of corresponding scores in the two sets of data

Then the Eigen values and Eigen vectors are calculated. From the Eigen vector and the original data the score of the attributes are obtained using the following equation,

$$SC = [ori_{dt}].[E_{Vec}] \qquad (2)$$

Where ori_{dt} = Original Data
E_{Vec} = Eigen Vector



Algorithm:

Input: Data Input

Output: Eigen Value (E_Val)

Processing data in advance.

Step 1: Delete Audio and delete special characters

Step 2: while queuing! = null

Step 3: words = line separated by ","

Step 4: alphabetically: alphabetically

Step 5: for i = 0 to length (alphabet)

Step 6: for i = 0 to length (String)

Step 7: if 48 <String <58

Step 8: sum = sum + str.character-48

Step 9: Finish if

Step 10: Finish

Step 11: while (total> 0)

Step 12: temp = sum% 10

Step 13: sum1 = sum1 + temp

Step 14: sum = sum / 10

Step 15: Finish in time

Step 16: Finish if

Step 17: Finish in time.

Step 18: So the data is then configured as numerical values.

Step 20: ito n, where n the number of columns.

Step 21: calculate the corresponding value covariance on two sets of data using equation (1)

Step 22: Calculate Eigen Values and Eigen Vectors

Step 23: [〖Cv〗 _mat]. [E_Vec] = [E_Val]. [E_Vec]

Where 〖Cv〗 _mat = Matrix of Covariance

E_Vec = Eigen Vector

E_Val = Eigen value

Step 24: Score SC SC = [〖ori〗 _dt]. [E_Vec]

When 〖ori〗 _dt = Original data

E_Vec = Eigen Vector

Step 25: Finish.

III. PERFORMANCE ANALYSIS

This section illustrates the effectiveness of the NEAA system. The performance of the proposed system is analyzed using the Mockaroo database [20]. Provides the opportunity to generate a limited amount of data records. It also helps to provide an opportunity to download the generated data as a .json, .xml, .sql file format. The effectiveness of the proposed plan is compared to existing strategies.

The type of anonymity for different types of data are tabulated in table 1. The data need not to be changed for both the alphabets and numbers in case on no anonymous type. For partial anonymous, the Hash code conversion is done for alphabetical data and the Eigen values are represented in a specific range for

numerical data. For full anonymous, the hash code conversion is carried out for both alphabetical and numerical data records.

Table 2 shows the critical values for the various indicators in the data records. Seven attributes are extracted here and criticalvalues of these attributes have been obtained. Depending on the critical values,the sensitivity level of a particular quality is determined.

Table 1 Anonymity Action Type

Content Types	***No Anonymous***	***Partial Anonymous***	***Full Anonymous***
Alphabets	No Changes	HashCode	Hashcode
Numbers	No Changes	MinValue< Number <MaxValue	Hashcode (Absolute)

Table 2. Sensitive Values of Attributes

Attributes	***Sensitive Value (x)***	***SQRT (x)***	***Level***
AT 1	147.97	12.16	L3
AT 2	36.8	6.06	L3
AT 3	29.27	5.41	L3
AT 4	21.66	4.65	L2
AT 5	20.28	4.5	L2
AT 6	18.47	4.29	L2
AT 7	12.9	3.59	L1

Figure 1 shows the loss of information with different symbols. Here the amount of data loss per attribute is measured and compared with the existing process. From the graph it is evident that the NEAA system has a slight reduction in information compared to that of the current system.

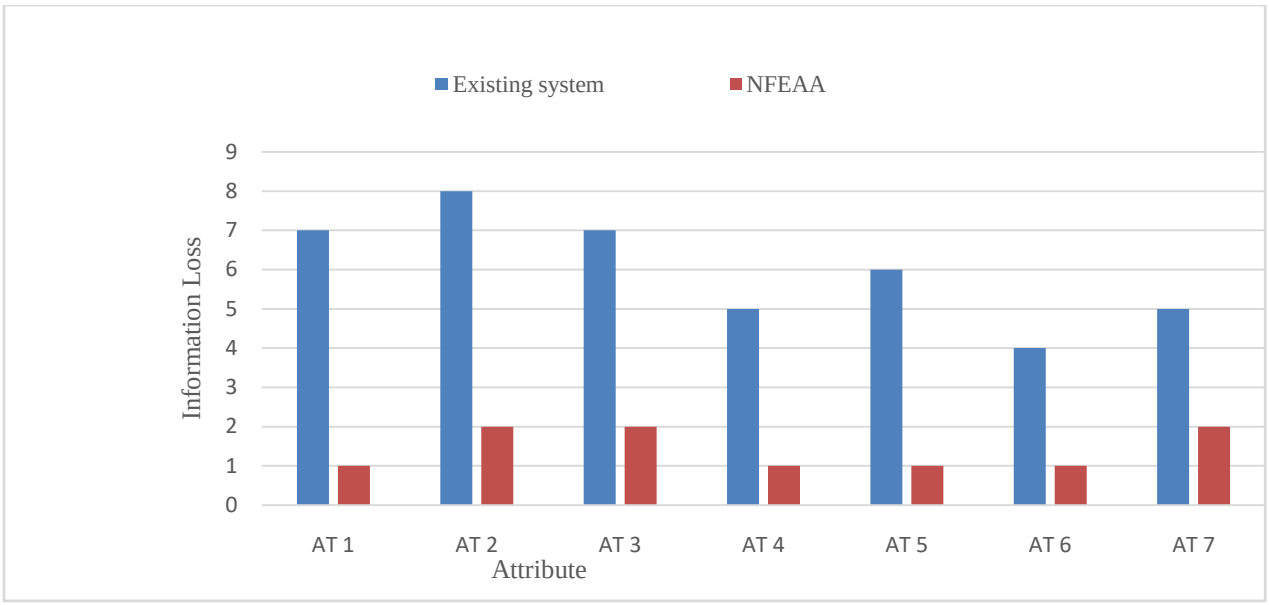


Fig. 1 Information Loss

Figure 2 describes the computational costs of the NEAA system are analyzed and compared with the various available systems [22]. Compared to the existing Chaudryet. al, the NEAA system has a minimum computational time of 35.6%. This shows that the proposed system provides better results than other existing programs.

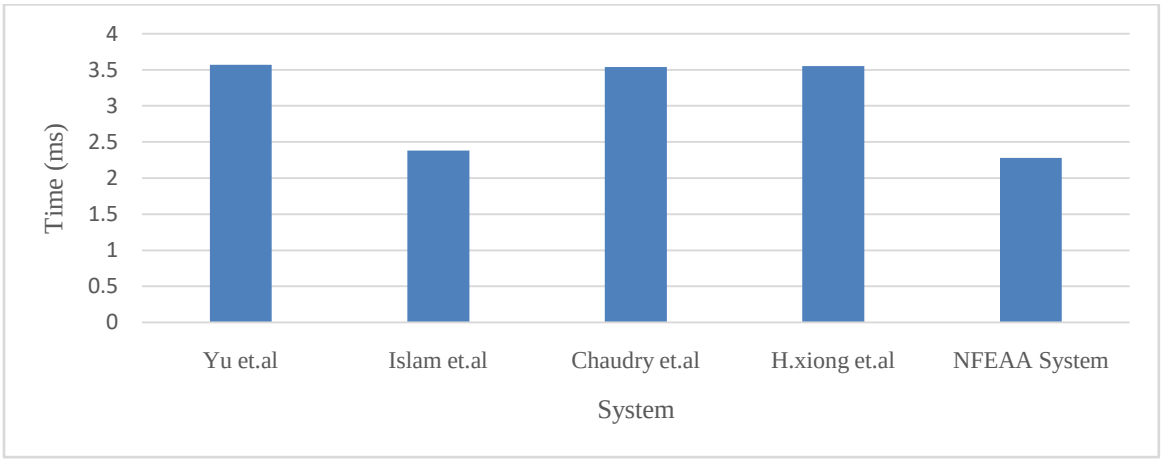


Fig. 2 Computational cost

Table 3. Comparison of Privacy models

Graph 2- Comparison of Privacy Models				
Model	***Run Time***	***Balance Point***	***Data Utility***	***Data Accuracy***
k anonymity	Low	Increases	Decreases	Medium
ℓ - diversity	High	Increases	Increases	High
ℓ - diversity applied k anonymity externaldatamodel	Very High	Increases	Increases	High

NEAA System	Low	Increases	Decreases	Very High

Table 3. shows the comparative analysis of the various privacy models. From the table it is discussed that the NEAA system provides better results in measures such as performance time, measurement point, data usage and accuracy. The results prove that the proposed system is better than other existing models.

IV. CONCLUSION AND FUTURE WORK

The main purpose of this work is to develop an effective anonymous algorithm for maintaining the privacy of personal information by individuals. Confidentiality often plays a major role in data mining techniques. In the current trend, several anonymous algorithms are designed for privacy that ends up digging data. But there are limits to privacy protection. The novel framework of the Efficient Anonymous Algorithm (NEAA) is therefore proposed for this work. Initially raw data about individual information is processed. Thereafter the used information is stored in a database. From this sensitive and sensitive information is determined using the Principal Component Analysis (PCA) based Attribute selection algorithm. The process of concealing words is done by introducing an novel algorithm based on Anonymous (NBA). Finally anonymous information can be obtained as a result. The effectiveness of the NEAA system can be verified by experimental analysis. The results concluded that the proposed framework provides better performance compared to existing systems.

REFERENCES

[1] A. Patil and S. Patil, "A review on data mining based cloud computing," *International Journal of Research in Science and Engineering,* vol. 1, pp. 1-14, 2014.

[2] H. Vaghashia and A. Ganatra, "A survey: privacy preservation techniques in data mining," *International Journal of Computer Applications,* vol. 119, 2015.

[3] K. Pasierb, T. Kajdanowicz, and P. Kazienko, "Privacy-preserving data mining, sharing and publishing," *arXiv preprint arXiv:1304.1877,* 2013.

[4] N. Victor, D. Lopez, and J. H. Abawajy, "Privacy models for big data: a survey," *International Journal of Big Data Intelligence,* vol. 3, pp. 61-75, 2016.

[5] C. W. Axelrod, "Ensuring online data privacy and controlling anonymity," in *Emerging Technologies for a Smarter World (CEWIT), 2015 12th International Conference & Expo on*, 2015, pp. 1-6.

[6] Y. A. A. S. Aldeen, M. Salleh, and M. A. Razzaque, "A comprehensive review on privacy preserving data mining," *SpringerPlus,* vol. 4, p. 694, 2015.

[7] S. B. Avaghade and S. S. Patil, "Privacy preserving for spatio-temporal data publishing ensuring location diversity using K-anonymity technique," in *Computer, Communication and Control (IC4), 2015 International Conference on*, 2015, pp. 1-6.

[8] M. I. Pramanik, R. Y. Lau, and W. Zhang, "K-anonymity through the enhanced clustering method," in *e-Business Engineering (ICEBE), 2016 IEEE 13th International Conference on*, 2016, pp. 85-91.

[9] A. Aristodimou, A. Antoniades, and C. S. Pattichis, "Privacy preserving data publishing of categorical data through k-anonymity and feature selection," *Healthcare technology letters,* vol. 3, pp. 16-21, 2016.

[10] M. Xie, M. Huang, Y. Bai, and Z. Hu, "The anonymization protection algorithm based on fuzzy clustering for the ego of data in the Internet of Things," *Journal of Electrical and Computer Engineering,* vol. 2017, 2017.

[11] S. Banerjee, V. Odelu, A. K. Das, S. Chattopadhyay, N. Kumar, Y. Park*, et al.*, "Design of an Anonymity-Preserving Group Formation Based Authentication Protocol in Global Mobility Networks," *IEEE Access,* vol. 6, pp. 20673-20693, 2018.

[12] L. Zheng, H. Yue, Z. Li, X. Pan, M. Wu, and F. Yang, "K-anonymity Location Privacy Algorithm based on Clustering," *IEEE Access,* 2017.

[13] Y. Gao, T. Luo, J. Li, and C. Wang, "Research on K Anonymity Algorithm based on Association Analysis of Data Utility."

[14] P. S. Rao and S. Satyanarayana, "Privacy preserving data publishing based on sensitivity in context of Big Data using Hive," *Journal of Big Data,* vol. 5, p. 20, 2018.
[15] Y. Wang, Z. Cai, Z. Chi, X. Tong, and L. Li, "A differentially k-anonymity-based location privacy-preserving for mobile crowdsourcing systems," *Procedia Computer Science,* vol. 129, pp. 28-34, 2018.
[16] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, "Enhancing data utility in differential privacy via microaggregation-based $$ k $$ k-anonymity," *The VLDB Journal—The International Journal on Very Large Data Bases,* vol. 23, pp. 771-794, 2014.
[17] X. Zhang, C. Liu, S. Nepal, S. Pandey, and J. Chen, "A privacy leakage upper bound constraint-based approach for cost-effective privacy preserving of intermediate data sets in cloud," *IEEE Transactions on Parallel and Distributed Systems,* vol. 24, pp. 1192-1202, 2013.
[18] V. Rajalakshmi and G. A. Mala, "Anonymization by data relocation using sub-clustering for privacy preserving data mining," *Indian Journal of Science and Technology,* vol. 7, pp. 975-980, 2014.
[19] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information security in big data: privacy and data mining," *IEEE Access,* vol. 2, pp. 1149-1176, 2014.
[20] "https://mockaroo.com/ ".
[21] G. B. Demisse, T. Tadesse, and Y. Bayissa, "Data Mining Attribute Selection Approach for Drought Modeling: A Case Study for Greater Horn of Africa," *arXiv preprint arXiv:1708.05072,* 2017.
[22] H. Xiong, J. Tao, and C. Yuan, "Enabling telecare medical information systems with strong authentication and anonymity," *IEEE Access,* vol. 5, pp. 5648-5661, 2017.
[23] P. M. V. Kumar and M. Karthikeyan, "l-diversity on k-anonymity with External Database for improving Privacy Preserving Data Publishing," *International Journal*