A Novel Framework for Efficient Anonymity Algorithm

Dhaval A Jadhav,Ravikumar K
Rai University,Ahmedabad

Abstract— In recent days, the data mining techniques plays a significant role in finding useful information from large volume of data.The extracted data may contain some private information about the individuals. There may be a great chance of hacking the individual's personal information. Hence the preservation of privacy becomes an important aspect in data mining. Several privacy preserving techniques are developed for hiding the personal information about an individual. One of the major privacy preserving technique is the anonymization. Several traditional anonymity approaches are utilized for preserving the privacy of the individual. But still it has drawbacks in preserving the privacy about the personal information of an individual. Thus a Novel Framework for Efficient Anonymous Algorithm (NFEAA) is proposed in this work. Here the sensitive and non-sensitiveattributes of the data can be identified by using Principal Component Analysis based Attribute Selection Algorithm. In this algorithm, the Eigen values and Eigen vectors are estimated. Then the anonymization process is carried out by introducing a Novel Based Anonymity Algorithm (NBA). Finally the anonymous data is obtained which prevents the hacking of personal information about an individual. Here the success of privacy preservation can be determined by the performances such as data utility, privacy levels and computational cost. From the experimental analysis, the performance of the NFEAA system proves its superiority compared to the other techniques.

Keywords—Data Mining, Privacy Preservation, Principal Component Analysis, Anonymity.

I. INTRODUCTION

In day to day life, the advancements in information technology plays a major role which leads to the large volume of data storage.The extraction of information from these large repositories requires a proper mechanism for better decision making. The access of these information resources are done by the process of data mining[1].It is one among the core processes of discovering the knowledge from database. This data is comprised of typical sensitive information about the individual persons like financial and medical information which are mostly exposed to many parties like, users, owners, collectors and miners. This availability of large of volume of data has the ability to learn more information about an individual.

For this purpose the concept of privacy preservation has been introduced which is originated as the significant concern in data mining[2]. It is referred to as the privacy protection about sensitive or individual knowledge without sacrificing the utilization of the data. Because of the privacy intrusions, the users become unwilling to share their personal and sensitive information. In recent days, the privacy become more advanced as the ability of data storage gets increased. The major intention of this privacy preservation in data mining is to extract the appropriate information from large volume of data along with the thoughtful information during protection.

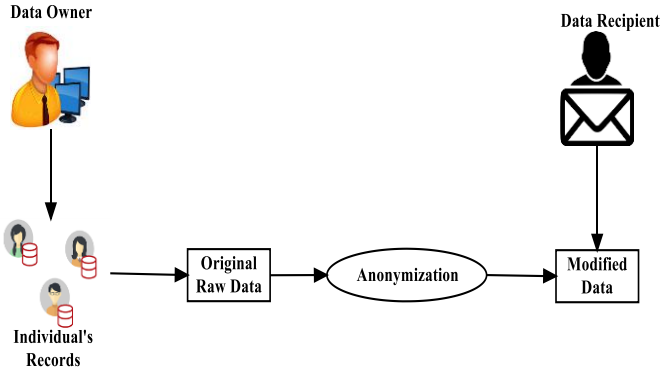

Fig. 1 Privacy Preservation

There are some different types of privacy preserving techniques such as heuristic based, cryptography base and reconstruction based techniques. Generally the fundamental form of data can be comprised of four kinds of attributes such as explicit identifiers, quasi identifiers, sensitive attributes and non-sensitive attributes[3].

- Attributes that consist of information which helps to identify the owner of the record by explicitly like name are referred to as Explicit Identifiers.
- Attributes that have the capability to identify the owner of a record which is integrated with the publicly available data are referred to as Quasi Identifier.
- Attributes which comprises the specific information about a sensitive person like salary, disease, etc. are referred to as Sensitive attributes.
- Attributes which does not creates any problem when the data is revealed to untrustworthy parties are referred to as Non-Sensitive attributes.

The sensitive data or the identity of the record owners which are need to be hidden are done by utilizing an approach called anonymization. Also it assumes that the sensitive data must be retained for the purpose of analysis. The explicit identifiers must be removed but when the quasi identifiers are linked with the publicly available data, there exists a danger of privacy intrusion. This type of attacks are represented as linking attacks. One of the best example for this linking attack is the attributes like name, sex, DOB, race that are available in

public records. This can be utilized for inferring the identity of respective individual with the increased probability. The prediction of individual identity can be avoided by using k-anonymity model which helps to reduce the danger of privacy.

Several existing techniques are utilized for implementing the anonymization process [4]. Among these, k-anonymity is the most commonly utilized algorithm in current trends. But still they has the drawback of information loss during the data transformation. Also there exist two major limitations in the k-anonymity model. One is difficult to decide the attributes as available and not available in the external tables. The other thing is the adaptation of attack methods in real scenarios. Normally there is a common anonymity model for the queries that are given by the user. This leads to easily break the privacy. But in our work, different types of anonymization is carried out for different queries that are given by the same user. This makes the system hard to break the privacy of the data.

The main objectives of this research work are as follows,

- To preserve the sensitive data about the individual user, a novel anonymization technique is introduced.
- To select the sensitive attribute from the individual information of the user.

II. RELATED WORK

[5] developed privacy and anonymity models from the investigation of severaltypes of data and the inspirations involved by many groups and persons. Here the application and utility of particular risk reduction methods and tools were highlighted. The desire of the privacy and anonymity was also examined. The methods comprising legislation were proposed for ensuring the fulfillment of individual's privacy requirements. As a result it was inferred that there required an awareness and education programs for obtaining the better understanding of the issues by the technologists. [6] provided an overview on the Privacy Preserving Data Mining (PPDM) techniques depending upon the randomization, distortion, distribution, associative classification and k-anonymization.The main intention of PPDM was to integrate the existing data mining approaches for transforming the data into mask sensitive data. The key challenge was to transforming the data effectively and to recover the mining results from the transformed data. There was an urgent necessity for developing a robust, effective and scalable model for eliminating the existing issues like overhead in global mining computing, integrity of mining data, scalability, data utility and privacy preservation of growing data.

[7]recommended a confidentiality privacy which ensured the diversity of location by restricting the probabilistic analysis depending upon the adversary knowledge or the probability of user who were visiting a sensitive location. This anonymization approach worked on the map and distorted the sensitive portions of the trajectories. When the privacy parameters were fulfilled by the users, this approach had the capability for preserving the utilization. [8]utilized a novel clustering method for achieving k-anonymity through enhanced data distortion which assured reduced data loss. An additional restriction of less data loss was included during the process of clustering and it was not integrated with the conventional clustering techniques. A process of data release was supported by this approach in which the data would not be distorted more than they were required to attain k-anonymity. Also several suitable metrics were developed to measure the generalization quality and the new metrics were appropriate for both the categorical and numerical attributes. The results demonstrated that the proposed method offered less data loss compared to the existing techniques. Even though this approach has less execution but it was not fully optimized.

[9] introduced a new anonymization approach depending up on the k-anonymity via pattern based multidimensional suppression (kPBMS). The dimensionality of the data was reduced by utilizing feature selection in this approach. Then the attribute and record suppression were integrated for attaining k-anonymity. The proposed approach offered better accuracy but it missed the interacting features with less effect which became the major drawback. [10]suggested the idea of ego of data and analyzed the features of the ego of data in IOT. Here the two steps of data clustering was implemented like the spatial position of adjacent fuzzy clustering as the first step and the sampling time fuzzy clustering as the second step. By utilizing this way, the data with layout characteristics was classified as different equivalent classes which helped to obscure the particular position information of the data, eliminate the layout characteristics of tags and attain the anonymization protection. The efficiency of the protection of data could be improved by utilizing this suggested approach without reducing the anonymization quality and enhancing the data loss. But the main issues was that the data protection in IOT.

III. PROPOSED METHOD

This section describes the working procedure of the NFEAA system for privacy preservation. The flow of the projectedsystem is shown in Fig. 1. Initially the input data is obtained from the dataset in which the personal information about the individual users are stored. This input data is further preprocessed for proceeding further processes. Then these preprocessed data are stored in database. From the preprocessed data, the attributes are selected based on the Principal Component Analysis algorithm. Then the sensitive and non-sensitive attributes are obtained based on the selected attributes. The anonymity for each attribute is determined by using a novel based anonymity algorithm. Finally the anonymous data is obtained as a result.

A. Preprocessing

Initially the input data is obtained from the dataset which is comprised of personal details of the users. The process of noise removal and special characters removal are done in this preprocessing step. Also the dataset may consists of different types of data such as characters, string, numerical values, etc. These unstructured data are converted into a structured format by using preprocessing technique.

In this approach, the different types of data are converted into numerical values using the ASCII code. The

values for each data can be process up to 2 digits and this can be obtained as a preprocessed data. Then this preprocessed data can be stored in a database for proceeding further processes. After this the covariance matrix for the corresponding scores in the data are evaluated using,

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

Where N = Number of scores in each set of data,

Xi = i^{th} raw score in the first set of scores,

Yi = 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,

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

Where ori_{dt} = Original Data

E_{Vec} = Eigen Vector

From this, the privacy score for the input data can be obtained and based on the Eigen values, the sensitive and non-sensitive attributes are classified. The algorithm used for obtaining the privacy score is described as follows.

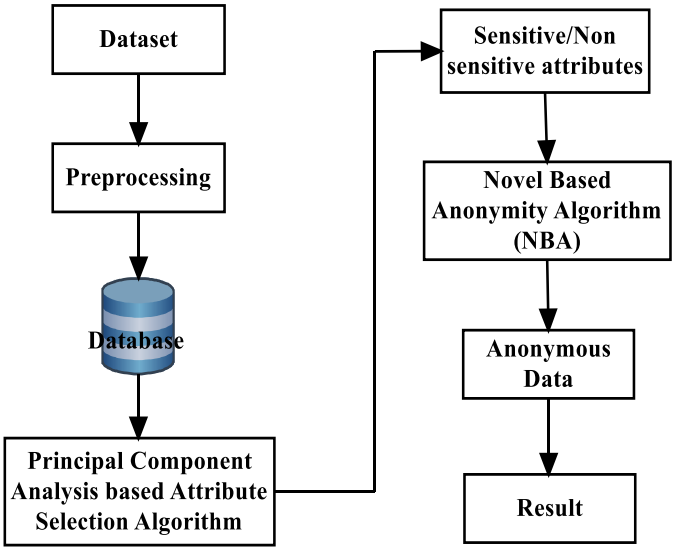

Fig. 1 Work flow of the NFEAAsystem

After the privacy score for the data is obtained using the following algorithm.

Algorithm I: Privacy score

Input: *Input Data Set*

Output: *Eigen Value (E_{Val})*

#Pre-Processing the data.

Step 1: Noise Removal and Removing special characters

Step 2: while line! = null

Step 3: alphabets = line split by ","

Step 4: for alphabet: alphabets

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: End if

Step 10: End for

Step 11: while (sum>0)

Step 12: temp = sum%10

Step 13: sum1=sum1+temp

Step 14: sum = sum/10

Step 15: End while

Step 16: End if

Step 17: End while.

Step 18: Hence the data was normalized as Numeric Values.

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

Step 21:calculate the covariance of corresponding scores in the 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} *= Covariance Matrix*

E_{Vec} *= Eigen Vector*

E_{Val} *= Eigen Value*

Step 24: Attribute Scoring $SC = [ori_{dt}].[E_{Vec}]$

Where ori_{dt} *= Original Data*

E_{Vec} *= Eigen Vector*

Step 25: End for.

B. Attribute selection using Principal Component Analysis Algorithm

The preprocessed data is taken as an input, in which the attributes are selected based on the principal component analysis. It is a dimension reduction tool which helps to reduce a large volume of data variable to a small set that is comprised of most of the information. Here the Eigen value for the data can be taken as an input. Then this Eigen value is categorized in to three categories. Let us assign n_1,n_2 as an assumption for categorizing the valuesand set a boundary value for the selection of attributes. Based on the sensitive values, the attributes are classified by using the equation,

$$(C_A) = \begin{cases} L_S & if\ (0 < E_{Val} < n_1) \\ S_T & if\ (n_1 < E_{Val} < n_2) \\ H_S & if\ (n_2 < E_{Val} < n_n) \end{cases} \quad (3)$$

Then the type of anonymity can be fixed as partial or fully anonymous by using the following expressions,

$A_{Ty} \rightarrow A_T \in L_S = A_n$ (4)

$A_T \in S_T = A_p$ (5)

$A_T \in H_S = A_f$ (6)

Where A_{Ty} = Anonymity Type,

C_A = Classification of Attribute,

n_1,n_2 = Mid-Value of E_{Val},

L_S = Less Sensitive,

S_T = Sensitive,

H_S = High Sensitive,

A_n = No Anonymous,

A_p = Partial Anonymous,

A_f = Fully Anonymous

Algorithm II: Sensitive Attribute Selection

Input: *Eigen Value (E_{Val})*

Output: *Classificationof Attributes (C_A)*

Step 1: E_{Val} = 0 to n_n, where n_n is the end number.

Step 2: Split the E_{Val} into three categories.

Step 3:For that categorical value we assign n_1,n_2 as assumption.

Step 2: Set boundary values for attribute selection.

Step 3: Classify the Attributes (C_A) according to the sensitive values using equation (3)

Step 4: Fixing Anonymity type (A_{Ty}) using equations (4), (5) and (6).

C. Novel Based Anonymity Algorithm (NBA)

Let us consider the classification of attributes as an input for obtaining the anonymity of that data. In this approach the attributes are selected based on the classification of attributes. If the attribute type is less sensitive, then the there is no conversion process in the data and this can be represented as

$$A_n = A_{Ty} \quad (7)$$

If the attribute type is sensitive, then it has partial conversion. In this conversion, when the attribute is numerical then the Caesar cipher conversion is carried out. When the attribute is a character then the partial conversion is done which means the data can be represented in a particular range (i.e.,A_{Ty} Age consist 33 means its P_C value is "30<age<40"). This can represented as,

$$A_p = \begin{cases} C_C & \text{if } (A_{Ty} = \text{Numeric}) \\ P_C & \text{Otherwise} \end{cases} \quad (8)$$

Where C_C = Caesar Cipher with key value must positive or negative

P_C = Partial character

If the attribute type is high sensitive, then it has fully conversion. In this conversion, when the attribute type is numerical, then the Caesar cipher conversion is done with the key value of either negative or positive values. When the attribute type is character or string, then the hash code conversion is carried out. This can be represented as

$$A_f = \begin{cases} C_C & \text{if } (A_{Ty} = \text{Numeric}) \\ H_C & \text{Otherwise} \end{cases} \quad (9)$$

Where C_C = Caesar Cipher with key value either positive or negative

H_C = Hash Code for the attribute fields

By using this, the anonymity of the data can be obtained by,

$D_A = \sum_i^n A_{Ty} \in (A_n \,||\, A_p || A_f)$ (10)

The novel based anonymity algorithm is described as follows:

Algorithm III: Novel Based Anonymity Algorithm (NBA)

Input: *Classificationof Attributes (C_A)*

Output: *Anonymity Data (D_A)*

Step 1: Select all the attributes according to (C_A)

Step 2: For No – Anonymous Type (A_n)

If $A_{Ty} \in L_S$

Here, Attribute type (A_{Ty}) has no Conversion, So its remains same

$\rightarrow A_n = A_{Ty}$

End If

Step 3: For Partial Anonymous Type (A_p)

If $A_{Ty} \in S_T$

Here, Attribute type (A_{Ty}) has partial Conversion using equation (8)

End If

Step 4: For Fully Anonymous Types (A_f)

If $A_{Ty} \in H_S$

A_{Ty} Converted to Fully Conversion using equation (9)

End If

Step 5: Anonymity Data (D_A) using equation (10)

IV. PERFORMANCE ANALYSIS

This section demonstrates the performance of the NFEAA system.The performance of the proposed system is analyzed using Mockaroo dataset[20]. It offers the possibility to generate a limited number of data records. Also it helps to give the possibility for downloading the generated dataset as .json, .xml, .sql file format. The performance of the proposed system is compared with the existing techniques.

Table 1 describes the level of sensitivity. Here a certain range of Eigen values are assumed for determining the level of sensitivity. This tables shows that the Eigen values ranges from 0 to 4 is considered as L1, low sensitive. Thus no anonymity is carried out. For the Eigen values ranges from 4 to 5 are considered as L2, sensitive. This has partial anonymity. Similarly for the Eigen values that are greater than 5 are considered as L3, High sensitive. This has fully anonymity.

Table 1 Level of sensitivity

Levels	*Types*	*Range*	*Action*

L1	Low Sensitive	0<Eval<4	No Anonymous
L2	Sensitive	4<Eval<5	Partial Anonymous
L3	High Sensitive	Eval>5	Fully anonymous

The type of anonymity for different types of data are tabulated in table 2. 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 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 3 illustrates the sensitive values for different attributes in the data records. Here seven attributes are taken and the sensitive values for these attributes are obtained. Based on the sensitive values, the level of sensitivity for the particular attribute is determined.

Table 3. 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

Fig. 3 depicts the information loss for different attributes. Here the amount of information loss for each attribute is measured and compared with the existing technique. From the graph it is observed that the NFEAA system has less information loss when compared to that of the existing system.

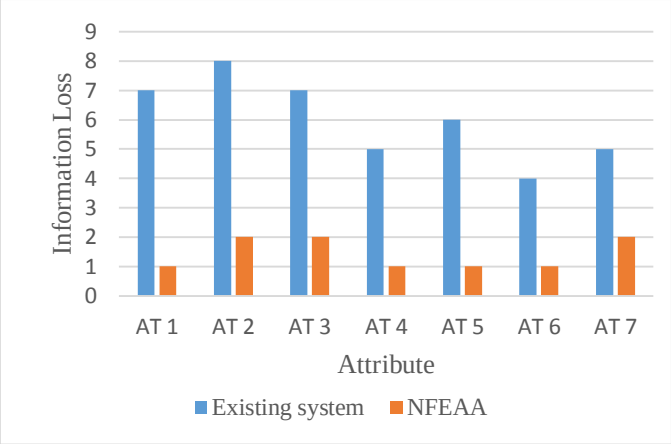

Fig. 3 Information Loss

The comparative analysis of Eigen values for different attributes are determined and shown in Fig. 4. Here the NFEAA system is compared with the existing technique [21]. From the results it is noted that the NFEAA technique offers better Eigen values than the existing technique.

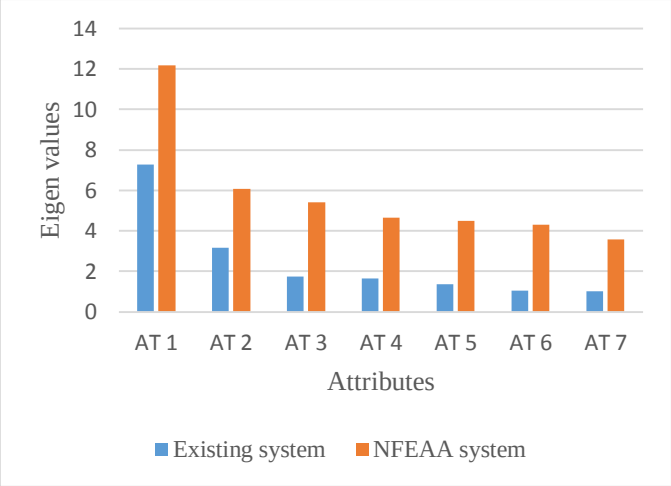

Fig. 4 Comparison of Eigen values

Fig. 5 describes the computational cost for the NFEAA system is analyzed and compared with the various existing systems [22]. When compared to the existing Chaudryet. al system, the NFEAA system has 35.6% less computational time. This shows that the proposed system offer better results than the other existing systems.

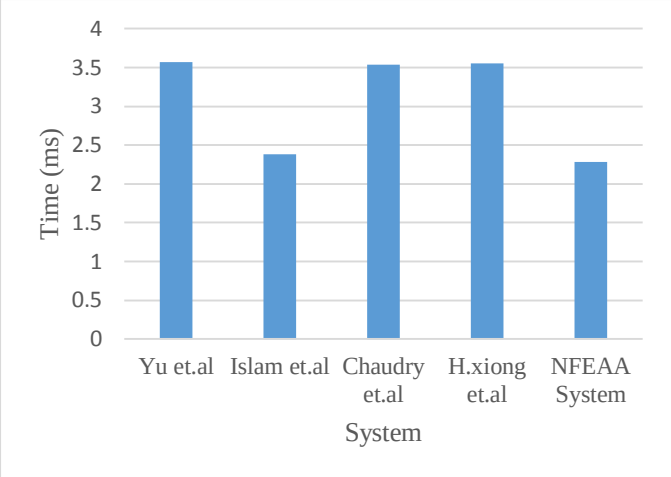

Fig. 5 Computational cost

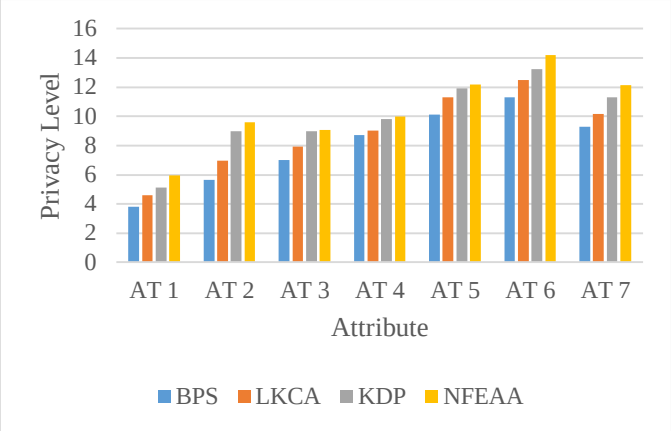

Fig. 6 Privacy Level

The privacy level of the NFEAA system is evaluated for different attributes and shown in Fig. 6. Also the analysis of privacy level for the NFEAA system is compared with several existing techniques [23]. In this graph, the NFEAA system offers 12.35% increased privacy model compared to the existing LKCA technique. From the results, it is observed that the NFEAA system offers higher privacy models for various attributes compared to that of the existing techniques.

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
NFEAA System	Low	Increases	Decreases	Very High

Table 3. illustrates the comparative analysis of various privacy models. From the table it is discussed that the NFEAA system offers better results in the measures like running time, balance point, data utility and accuracy. The results proves that the proposed system is superior to the other existing models.

V. CONCLUSION AND FUTURE WORK

The main intention of this work is to propose an efficient anonymous algorithm for preserving the privacy of the personal information about the individuals. Generally the privacy preservation plays a significant role in the data mining techniques. In current trends, several anonymous algorithm are developed for privacy preserving in data mining. But it has the limitation of privacy preservation. Hence a Novel Framework for Efficient Anonymous Algorithm (NFEAA) is proposed in this work. Initially the raw data about the individual's information are preprocessed. Then these preprocessed data are store in a database. From this the sensitive and non-sensitive data are determined by utilizing the Principal Component Analysis (PCA) based Attribute selection algorithm. The anonymization process is carried out by introducing a novel based Anonymity (NBA) algorithm. Finally the anonymous data can be obtained as a result. The performance of the NFEAA system can be validated by the experimental analysis. The results concluded that the proposed framework offers better performance compared to the 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 of Computer Applications*, vol. 54, 2012.