Dr. Dhaval Jadhav[1],
Dr. Ankit Patel[2],
Prof. Kajal Patel[3]

Preserving Privacy of Sensitive Data using Anonymization Technique

Journal of Electrical Systems

Abstract Today, data mining techniques play an important role in finding useful information from large amounts of data. The extracted data may contain some personal information about individuals. There is a high probability of hacking personal information. Therefore, protecting privacy becomes an important factor in data mining. Many privacy protection techniques have been developed to hide private information about people. An important process of protecting privacy is anonymization. Many anonymity methods are used to protect the privacy of individuals. However, it still has shortcomings in the personal protection of personal information. Therefore, a new approach for effective anonymization is proposed in this study. Here, the feature selection algorithm based on the main content analysis can be used to identify the negative features of the data.

In this algorithm, eigenvalues and eigenvectors are estimated. Anonymization is then accomplished by introducing the Novel Approach for Anonymization. Finally get anonymous data that protect personal data from hacking. The success of privacy protection here can be determined by performance such as electronic data, privacy level and cost accounting. From the experimental analysis, the performance of the proposed system shows its superiority over other systems.

Key Words: Privacy, Security, Anonymity, Data Mining

1. INTRODUCTION

Advances in information technology play an important role in daily life and result in a large amount of information. Extracting information from these large data centers requires a proper process to make better decisions. Access to this information is done through the data mining process [1]. It is one of the core thing of extracting knowledge from databases. These documents often contain important information about individuals such as financial and medical information that is disclosed to various parties such as users, owners, collectors and miners. The availability of this huge amount of data has the potential to learn more about people.

For this purpose, the concept of privacy, which has emerged as an important source of concern in data mining, has been put forward [2]. It is known as protecting the privacy of personal or personal information without compromising the information used. Due to privacy violations, users avoid sharing their personal and sensitive information. Privacy has become more important in recent days as information can be stored better.

The main purpose of data privacy protection is to extract only necessary data as well as desired data from large amounts of data during the protection period.

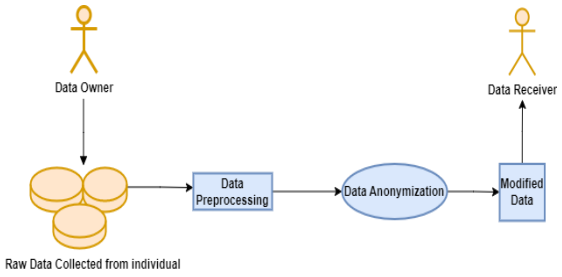

Fig. 1 Privacy Preservation

[1]*Associate Professor, VBTMCA, Umrakh*
dhaval.jadhav@vbtmca.ac.in
[2]*Associate Professor, VBTMCA, Umrakh*
ankit.patel@vbtmca.ac.in
[3]*Assistant Professor,VBTMCA,Umrakh*
kajal.patel@vbtmca.ac.in

There are several types of privacy protection mechanisms, such as heuristic-based, cryptography-based, and reconstruction methods. Mainly there are four kinds of attributes which represents the data: explicit identifiers, quasi identifiers, sensitive attributes and non-sensitive attributes [3].

Several existing mechanisms are used to implement anonymization [4]. Among them, k-anonymity is the most used algorithm in the current era. However, they still have the disadvantage of losing data during data conversion. The k-anonymity model also suffers from two major limitations. The first is that it is difficult to determine which attributes are present in an external tables. Another is the adaptation of the attack method to the real situation.

However, in our study, different types of anonymity are used for different questions given by the same user.

The main objectives of this research project are:

- ✓ New anonymization methods have been introduced to record sensitive information about individual users.
- ✓ Selection of sensitive attribute from given data.

2. RELATED WORK

[5] developed a model of privacy and anonymity by working with different types of information and inspiration from different groups and individuals. Application and utility specific mitigation and tools are discussed here. Also check the desire for privacy and anonymity. Approaches, including law, are suggested to ensure that the individual is subject to privacy. Therefore, it can be concluded that knowledge and training are needed to ensure a professional understanding of these issues.

[6] provides an overview of privacy-preserving data mining (PPDM) techniques based on randomization, perturbation, distribution, association distribution, and k-anonymization. The main purpose of PPDM is to combine existing data mining to transform data into masked data. The real challenge is to change data quality and recover mining profit from changed data. It is necessary to create a strong, effective and efficient system to eliminate existing problems such as the general burden of the world, computational mining, data mining integrity, scalability, data usage and data privacy protection.

[7] shows the privacy policy that ensures the diversity of the venue by limiting the surveillance results depending on the competitor's knowledge or the consequences of the user's access to the site.

This method of anonymity works on the map and covers sensitive areas of the road. The system can control usage when users are concerned about privacy. [8] used a new clustering method to achieve k-anonymity by improving data corruption, thus making the data smaller. Fewer information limitations are added in addition to the integration process that is not combined with the integration process. This approach supports a data publishing process in which data is not modified to the extent necessary to achieve k-anonymity.

A number of appropriate measures have also been developed to measure performance, and the new measures apply to both categorical and numerical measures. The results showed that the proposed method provides less information compared to the state-of-the-art technology. While this method is less work, it is not perfect.

[9] introduced a new anonymization method based on the k-anonymity of the model-based multivariate contentious system (kPBMS). Reduce the dimensionality of your data by using feature selection in this approach.

The attributes and data are then combined to achieve k-anonymization. The plan provided better accuracy, but missed less interactive interaction, which turned out to be a disadvantage. [10] proposed the concept of personal data and analyzed the characteristics of personal data in the Internet of Things. Two steps of data clustering are recognized here, i.e. the spatial position of the surrounding fuzzy set is taken as the first step and the time sampling fuzzy set is followed by the second step. In this way, documents with layout features are divided into different equal units, helping to hide documents' private information, remove text features, and realize anonymity protection.

The effectiveness of data protection can be increased using this plan, without reducing the quality of anonymity and increasing data loss. But the main problem is data protection in IoT.

3. PROPOSED METHOD

3.1 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.

At first, data entry is obtained from data containing users' personal information. In this preliminary step, noise removal and special symbols removal task is done. In addition, the file contains symbols, sequences, numbers, etc. This unstructured data is converted into a structured format during this preprocessing stage.

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,

In this stage, data variables are converted to numbers using ASCII codes. It is then analyzed using the covariance matrix of the corresponding scores in the data,

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

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,

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

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.

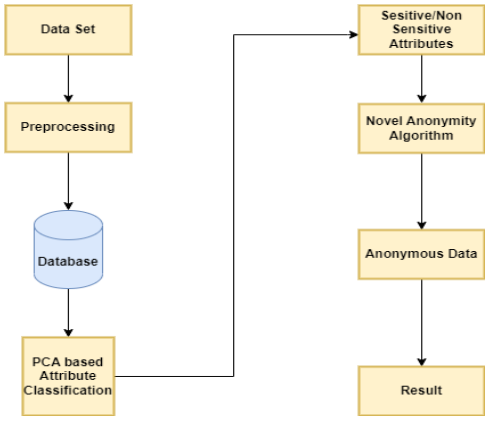

Fig. 2 Work flow of the Novel system

3.2 Attribute selection

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 assignn_1, n_2 as an assumption for categorizing the values and set a boundary value for the selection of attributes. Based on the sensitive values, the attributes are classified by using the equation,

By taking the preprocessed data as input, the features are selected according to the principal component analysis. It is a file size reduction tool that helps to download large numbers of converted files into small chunks containing most of the data. Here the eigenvalues of the data can be used as input. This eigenvalue is then divided into three groups. Let's set n_1, n_2 as assumptions for the distribution of the results and the limit value for the selection of features.

According to the sensitivity value, the behavior is separated from the equation,

$$(S_V) = \begin{cases} L_A & if\ (0 < E_{Val} < n_1) \\ S_A & if\ (n_1 < E_{Val} < n_2) \\ H_A & if\ (n_2 < E_{Val} < n_n) \end{cases} \quad (3)$$

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

Where

n_1, n_2 = Mid Value of Eigen valueE_{Val},

S_V = Sensitivity,

L_A = Less Sensitive Attribute,

S_A = Sensitive Attribute,

H_A = High Sensitive Attribute,

3.3 Proposed Novel Anonymity Algorithm

Let's consider attribute classification as a strategy for obtaining anonymous data. In this way, attributes are selected according to their classification. If the attribute type is less sensitive (L_A), the data will not be transformed. If the attribute type is sensitive (S_A), then it has to be partially converted. In this pass, Caesar cipher conversion is performed when the character is numeric. 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

$$A_{full} = \begin{cases} C_C & \text{if Numeric} \\ H_C & \text{Otherwise} \end{cases} \quad (4)$$

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,

$$D_S = \sum_i^n A_T \in (A_{no} \,||\, A_{pt} || A_{full}) \quad (5)$$

Where

D_S = Anonymised Data Set

A_{no}= No anonymised Attribute

A_{pt}= Partially anonymised Attribute

A_{full}= Fully anonymised Attribute

The proposed algorithm is defined as follows:

Novel Anonymity Algorithm

Input: *Classification of Attributes (C_A)*

Output: *Anonymised Data Set (D_S)*

Step 1: Classification of Sensitive Attribute (S_V)

Step 2: For No – Anonymous Type (A_{no})

If $A_T \in L_A$

Here, Attribute type (A_T) has no Conversion, So its remains same

$A_{no} = A_T$

End If

Step 3: For Partial Anonymous Type (A_{pt})

If $A_T \in S_A$

Here, Attribute type (A_T) has partial Conversion

End If

Step 4: For Fully Anonymous Types (A_{full})

If $A_T \in H_A$

A_T Converted to Fully Conversion using equation (4)

End If

Step 5: Anonymised Data Set (D_S) using equation (5)

4. PERFORMANCE ANALYSIS

This section demonstrates the performance of the proposed Anonymity 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.

This section describes the analysis of the proposed Anonymity System. The performance of the method was evaluated using the Mockaroo dataset [20]. Comparisons with the Existing Systems are shown here.

Table 1 describes the sensitivity levels. Here, a number of eigenvalues are estimated to determine the sensitivity level.

Levels	***Types***	***Range***	***Action***
L1	Low Sensitive	0<Eval<6	No Anonymous
L2	Sensitive	6<Eval<7	Partial Anonymous
L3	High Sensitive	Eval>7	Fully anonymous

Table 1 Sensitivity Levels

Table 2 describes the level of sensitivity for various attributes used in proposed system.

Attributes	*Sensitive Value (x)*	*SQRT (x)*	*Level*
AT 1	160	12.65	L3
AT 2	53	7.28	L3
AT 3	67	8.18	L3
AT 4	40	6.32	L2
AT 5	44	6.63	L2
AT 6	41	6.40	L2
AT 7	13.3	3.65	L1

Table 2 Sensitivity Levels of Attributes

Fig. 3 represents the comparison of information loss of the date set attributes used in proposed system with existing system.

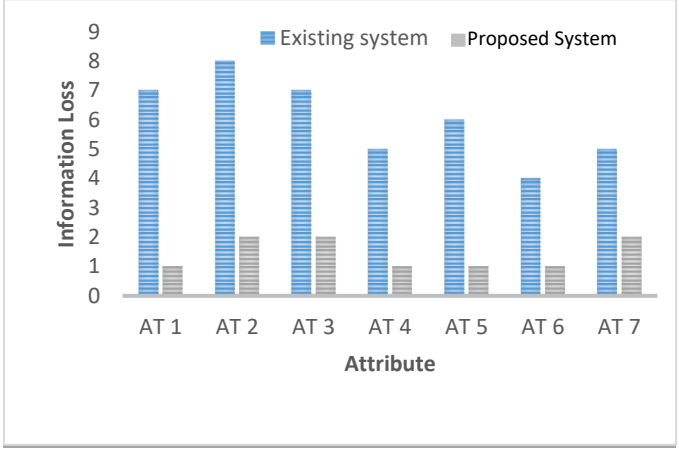

Fig. 3 Information Loss

Fig. 4 defines the computational cost comparisons for the proposed system is analyzed with the numerous existing systems [22]. It depicts that the proposed system has 35.6% less computational time as compared to Chaudry et. al system.

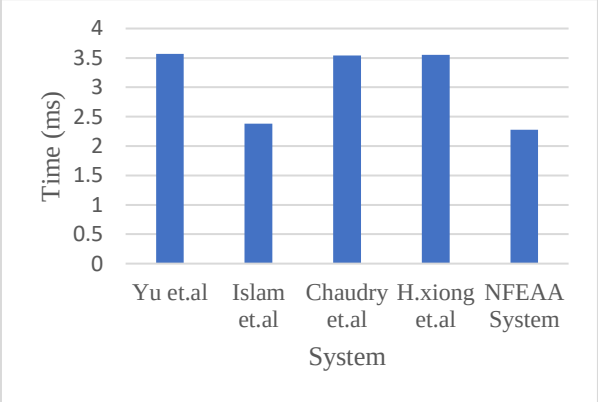

Fig. 4 Computational cost

The comparisons of privacy level for different attributes of the proposed system are estimated and shown in Fig. 5. From the results, it is verified that the proposed system gives higher privacy for various attributes over different existing systems.

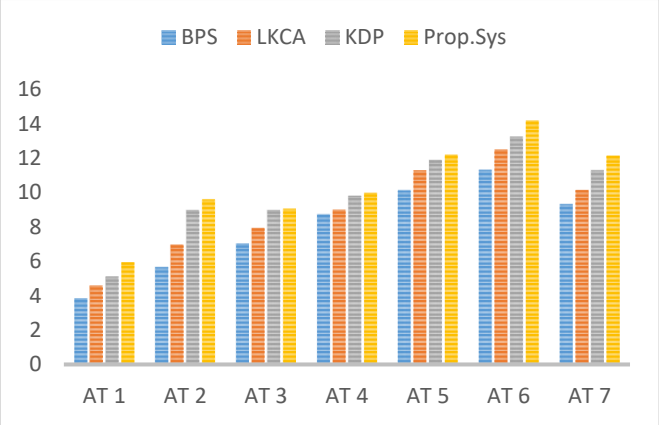

Fig. 5 Privacy Levels

Table 3. Illustrates the comparative analysis of various privacy models. From the table it is discussed that the proposed 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.

5. Conclusion and Future Work

The main purpose of this study is to propose an efficient anonymization algorithm for storing personal information. In general, privacy protection plays an important role in the data mining process. In the current model, some anonymous algorithms are designed to protect privacy in data mining. However, it has limited privacy scores. Therefore, a new framework for effective anonymization is proposed in this study. Raw information about personal data comes first. This pre-processed data is then stored in the database. From there, sensitive and non-sensitive data are determined using a behavior selection algorithm based on Principal Component Analysis. Finally, anonymous information may be obtained as a result.The effectiveness of the new system can be verified by experimental analysis. The results determined that the proposed system performs better than 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.