xIlkogretim Online - Elementary Education Online,2021; Vol 20 (Issue 5): pp.2451-2457
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.05.267

A Novel Hybrid Approach of Suppression and Randomization for Privacy Preserving Data Mining

Vibhor Sharma, HSST,Swami Rama Himalayan University, Dehradun
Dheresh Soni, HSST,Swami Rama Himalayan University, Dehradun
Deepak Srivastava, HSST,Swami Rama Himalayan University, Dehradun
Dr. Pramod Kumar, Krishna Engineering College, Ghaziabad, India

Abstract- In the era of technology advancement, knowledge extraction from large amount of data is very much important task. The process of data mining is applied to get the useful information from the data stored in a centralized server for important decision making process of multiple organizations. When multiple organizations collect the data for mutual gain, it gets vulnerable to individual's private data. Different approaches such generalization, perturbation, cryptography and randomization are used for taking care of the confidentiality of any individual's private data. Each of these methods has their own pros and cons like in anonymization, huge loss of information can take place. Data that is used in the process of data mining contain many attributes which hold confidential data of an individual and many attributes can reveal the private information of an individual, if those are associated with each other. These attributes are called quasi identifiers (QID). Individually these attributes don't breach the security but in a combined way these may be vulnerable towards the security of private data. Thus, there is the requirement of an approach to overcome the problem of disclosing of private data through quasi identifiers. Our proposed method of combining the suppression and randomization presents the solution to this problem. The method conserves the data privacy with the zero information loss in the process of regaining the actual values. The proposed work is carried out by making a local centralized server and outcomes are matched up with anonymization process to obtain the better results.

Keywords Prime- Anonymization, Computation time, Privacy preservation, Randomization, Suppression.

I. INTRODUCTION

The term data mining is very well known procedure through which we get useful information from a large set of raw data [1]. Most of the organizations have adopted this technology for mutual gain in the process of multi-part computation. During the process of data mining, the basic requirement that is arisen nowadays is to secure the individual's private data; data related the production of an organization and information of customers and employees [2]. It has become a real challenge for researchers to come up with a solution to privacy preserving data mining. Most of the datasets hold the data in 2D form i.e. rows (records) and colons (attributes). The fields can be of different types. Attributes such as name, address and contact are called explicit identifier that shows the uniqueness of an individual in terms of disclosing its identity. Attributes such as pin code, date of birth and sex are called quasi identifier (QID) that can reveal the identity of an individual person if used in an association. Attributes such as salary, ailment and status of ailment are called sensitive attributes because these kinds of attributes contain the confidential information. These types of attributes should never be concealed due to their important roles in the discovery of useful information. Attributes that contain non-confidential information are called non-sensitive attributes. These attributes cannot be left out in cold due to their vulnerability of revealing an individual's identity in case of association. They can also behave like QID [3].

Privacy Preservation Methods

There are many privacy preservation methods have been introduced so far to overcome the problem of privacy breaching of confidential data during the process of data mining. Some of them are following [4]:

Anonymization

This process is applied to remove a person's private information from the given dataset to conserve the confidentiality of data. It preserves the data of some confidential fields and looks actual and real. The main drawback is that the data that has been anonymized cannot be regained in actual form. It can happen but with a huge information loss.

Perturbation
This approach is applied to data by adding some kind of noise to the rows of particular fields that preserves the actual values of confidential information and that information is not disclosed during the process of data mining. The drawback with perturbation is that the yielding process is not very much accurate.

Randomization
In randomization before data mining, a shuffling process takes place in vertical way i.e. the position of records are swapped and their mean doesn't change in such a way that it conceals the private information by shuffling the records. Alternatively, It just hides the actual identity by shuffling.

Cryptography
Using the method of cryptography, confidential information is encrypted with the help of cryptographic algorithms such as symmetric key and asymmetric key cryptography algorithms. Confidential records are encrypted in such a way that cannot be decrypted during the process of data mining.

These approaches are very much useful when it comes to protect the privacy of an individual record during the process of data mining when extraction of knowledge takes place. The trust of an individual should be maintained i.e. their private information should not be disclosed when the data is being analyzed to extract the useful information from the centralized location. But each approach has its own advantages and disadvantages. Here we are discussing the main drawback of anonymization process that is obtained by quasi identifiers. Linked attack takes place due to theses identifiers in an associated way. Because of these identifiers, confidential information of an individual can be theft by attackers. Thus, it becomes important to hide the actual data of an individual in quasi identifiers before the implementation of data mining process to rectify the problem of privacy breaching.

Other than the introduction part, this paper represents different sections where literature review is presented in section-2. The problem related to Anonymization process is discussed in section-3. Proposed work is explained in section-4. Experimental results are represented in section-5. Conclusion and future work of the proposed work is given in section-6.

II. RELATED WORK

Various researchers have shown an interest to overcome the problem of privacy preserving data mining so that any kind of confidential information of an individual could not be disclosed. For that purpose many approaches have been proposed in the mentioned domain of privacy protection of data. There are several methods such as K-anonymity [5] and I-diversity [6]. However, these approaches are very well known for improving the anonymity against frauds and attackers. These approaches have been discussed in [7] with their drawbacks in the process of giving protection to confidential data. In the last decade, lots of work has been done towards the solution of mentioned problem where hybrid approaches were used to presser the privacy of data. In [8], authors attained the way for privacy conservation of data during the process of data mining by combining randomization with approximate computation. A new approach in the form of randomized response algorithm was introduced in [9] to overcome the problem. Erlingsson et al. in [10] presented an approach to preserve the privacy locally with the help of filtration called bloom used for vector coding and impose a novel approach of randomization. In [11], authors represented a way to reduce the communication cost of the method presented in [10]. However, these hybrid approaches just increase the scope of randomization methods of privacy preservation and not represent any revolution on these techniques. In case of multiparty computation, the main goal is to provide privacy of confidential data in distributed manner where multiple participating sites are involved for their mutual gain from the process of data mining. In [12], authors multi-data ranking protocol for achieving the privacy in multi-party computation where detailed information is given related to different patterns used with performance and complexity comparisons. Arora et al. in [13] represented the study of comparison of different anonymization approaches and reached on a conclusion that loss of information is increased if the number of fields is increased in the data. T-Closeness [14] gives better results than K-anonymity [5] and I-diversity [6] but it also suffers from information loss.

III. ANONYMIZATION

Anonymization approach is used to hide the individual's identity due to assumption of the retention of confidential data for the process of data analysis. The hiding process is performed by the generalization of different attributes in dataset. But due to the process of generalization, information loss takes place when the process of actual data extraction from anonymized data is occurred. There are many attributes in dataset, using which an individual's identity can be revealed. These attributes are like sex, pin code, date of birth etc. Using these attributes, privacy can be disclosed if these data is added to the publicly available data [12]. K-anonymity can restrict this kind of problem.

The attacked over the network can intrude in dataset A using the neighborhood (other participating site/ imposter). Where imposter = (i) that can be attached over the network i.e. i ɣ Z(A). An imposter has m occurrences in A' which is the anonymized dataset of A. Imposter can be classified by 1/m credibility in A'.

Age	Gender	Zip Code
34	M	B1667
45	F	B1675
34	F	B1931
45	M	B1925
70	F	B1931
70	M	B1931
66	M	B0931

2-Anonymity

Age	Gender	Zip Code
<50	*	B16**
<50	*	B16**
<50	*	B19**
<50	*	B19**
>50	*	B19**
>50	*	B19**
>50	*	B09**

Figure1: 2-Anonymity (Process of Anonymization) [16]

Figure 1 show the process of anonymization where Zip Code's last two numbers are hidden from everybody. When data is published, linked attack can take place if an attacker could identify the records which are published publically as shown in Figure 2. Data can be published in a form of the table which contains all the attributes mentioned above.

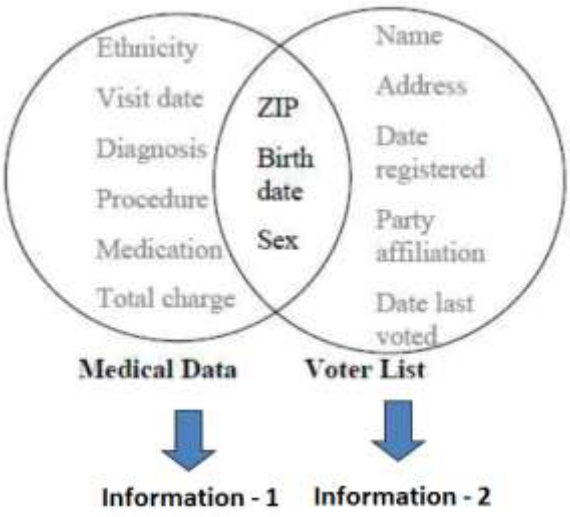

Figure 2: Linked Attack [17]

Through this kind of attack any fraudulent activity can take place by an attacker. The combination of information-1 and information-2 as shown in Figure 2 can reveal the confidential information about an individual and termed as linked attack. To avoid the linked attack, actual values of attributes in QID should be encrypted or hidden in such a way that attacker could not get the confidential information by the process of linking such kinds of attributes.

IV. PROPOSED WORK

Our proposed work is executed in terms of securing the identity of an individual's record. QID i.e. quasi identifier can play devastating role towards the task of providing security to individual's data in a data set. Anonymized data represents the security of confidential information but there can be huge information loss during the conversion process of data from anonymized level to actual level i.e. actual data cannot be regenerated again. Thus, the proposed work gives the solution by presenting a hybrid approach for preserving the privacy of data based on suppression and randomization. The approach aims to protect the data against linked attack using QID.

Input: Dataset Containing Original Values (A)

Output: Suppressed and Randomized Dataset (P)

Consideration: N datasets [A_1 ($r \times c_1$), A_2 ($r \times c_2$)..... An ($r \times c_n$)], where r is one same row values and c_1 is one value of different colon

Step-1: Remove the external identifiers.

Step-2: Select two numerical values randomly.

Step-3: Field named "Sex" is suppressed by both the selected random numerical values.

Step-4: Read the data "A" represented in form of a matrix R x C. Where R is the number of records and C is the desired number of colons.

Step-5: Generate a random matrix (RM) i.e. RM=random(R, C, SEED), where a discrete uniform distribution range is in the close interval [C, SEED], SEED is random initial value.

Step-6: **for each** dataset Ai

Step-7: Set $m = \sqrt{C} \times \sigma_r$ (where σ_r is single value of row)

Step-8: Set $c = 1/m$ ()

Step-9: Set $Q = c \times A_i \times RM$

Step-10: Set $P = \begin{bmatrix} P \\ Q \end{bmatrix}$

Step-11:**end for**

Step-12:return P

V. EXPERIMENTAL RESULTS

In this section, a set of experiments were held to evaluate the performance of proposed work.

Experiment environment

Implementation of our proposed work was done using Apache Tomcat 8.0 (local server) that was installed in Windows 10 operating system to make a virtual environment of a server that was used in a centralized manner. Minimum 4GB RAM was required and JAVA SE 14 version was installed in machine.

Experimental datasets

We considered hepatitis dataset [15] that is taken from the UCI Machine Learning Repository, a database from where many data sets are fetched for use in the domain of data mining. In this dataset, many attributes hold the confidentiality and some don't. The mentioned dataset contains 76 fields. All published work only used a subset of 14 attributes.

Information Loss

Following is the formula to obtain the loss of information in each attribute of data set in the process of obtaining actual values from randomized values:

$$\text{Loss of information} = (\text{Actual Values} - \text{New Values})^2 / (\text{Actual Values} + \text{New Values})$$

Table-2 represents the values of information loss in the process of anonymization, randomization and the proposed hybrid approach that is 0. The proposed approach represents no information loss. Figure-3 represents the comparison in the form of line charts.

Number of Patients	Anonymization	Randomization	Hybrid
1000	82.53	1.17	0.0
2000	81.92	1.93	0.0
3000	82.42	2.85	0.0
4000	84.04	1.39	0.0
5000	84.13	2.41	0.0
6000	84.24	1.45	0.0

Table-1: Values of information loss

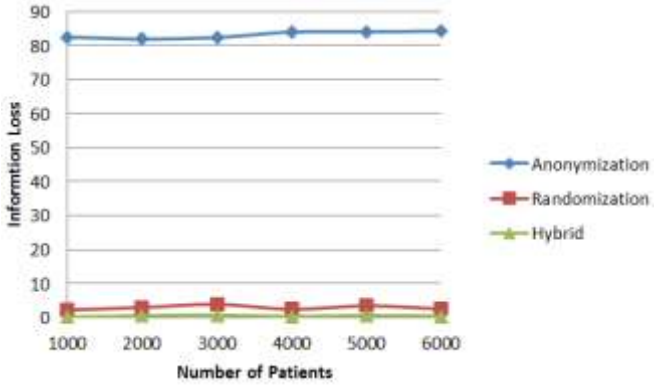

Figure-3: Comparison of information loss

Computation Time

Computation time is displayed in unit of milliseconds. Table-2 shows the computation time (in milliseconds) for anonymization, randomization and the proposed hybrid approach. Our proposed approach achieved the better results in less computation time as compared to anonymization and randomization. Figure-4 represents the comparison using line charts.

Number of Patients	Anonymization	Randomization	Hybrid
1000	285	245	221
2000	347	315	298
3000	432	402	378
4000	861	815	785
5000	1057	988	912
6000	1687	1546	1499

Table-2: Values of computation time (milliseconds)

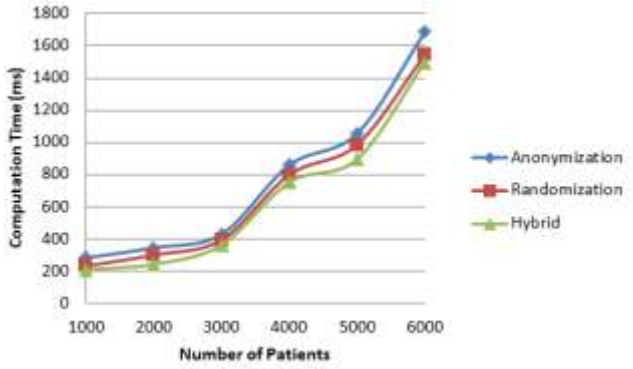

Figure-4: Comparison of computation time

Level of privacy preservation

The level of privacy preserved is computed in form of conserved number of characters of attributes' record of data set. Table-3 represents the number of characters conserved in anonymization, randomization and hybrid approach where our proposed approach achieves the maximum number of characters conserved. Figure-5 shows that comparison using line chart.

Number of Patients	Anonymization	Randomization	Hybrid
1000	6989	7221	11276
2000	35987	49879	85678
3000	72798	73254	174679
4000	175878	183454	509878
5000	245676	445135	735674
6000	321743	576598	913786

Table-3: Values of characters conserved

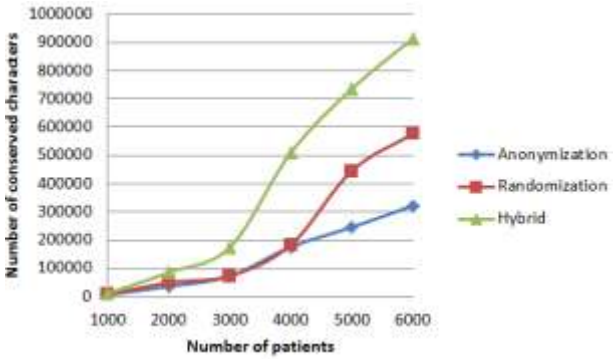

Figure-5: Comparison of characters conserved for privacy preservation

VI. CONCLUSION AND FUTURE WORK

In the paper, we proposed a hybrid way to preserve the privacy of data using suppression and randomization in the environment of local server where data is stored in a centralized manner. The experimental analysis represents the solution to hold the privacy of confidential attributes in such a way of conserving the individual's privacy durin the process of data mining. As anonymization process fails in the process of recovering the actual values from anonymized values, our proposed approach gives the positive solution to regain the actual values without any kind of information loss. Attributs having Boolean values can also be protected using our proposed algorithm. Hybrid approach achieves the goal in less computation time as compare to anonymization and randomization approaches which resolves the issue of information loss during the process of regaining actual data. In future, other techniques of privacy preservation can also be combined to gain the better results in less computation time and error rate.

REFERENCES

[1] M. B. Malik, M. A. Ghazi and R. Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", in proceedings of Third International Conference on Computer and Communication Technology, IEEE 2012, ISBN:978-1-4673-3149-4.

[2] K. Alotaibi, V. J. Rayward-Smith, W. Wang and Beatriz de la Iglesia, "Non-linear Dimensionality Reduction for Privacy Preserving Data Classification" in proceedings of 2012, ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, IEEE 2012, ISBN:978-1-4673-5638-1.

[3] Zhu, J. (2009, August). A new scheme to privacy-preserving collaborative data mining. In Information Assurance and Security, 2009. IAS'09. Fifth International Conference on (Vol. 1, pp. 468-471). IEEE.

[4] M. Suriyapriya, A. Joicy, Attribute Based Encryption with Privacy Preserving In Clouds. International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 Issue: 2

[5] Sweeney, Latanya. (2002) "k-anonymity: A model for protecting privacy." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10 (05): 557-570.
[6] Machanavajjhala, Ashwin Kumar Venkatanaga, Kifer, Daniel, Gehrke, Johannes, and Venkitasubramaniam Muthuramakrishnan. (2007) "L-diversity: Privacy beyond k-anonymity." ACM Transactions on Knowledge Discovery from Data (TKDD) 1 (1): 3-es
[7] Zhao, Jingwen, Chen, Yunfang, and Zhang, Wei. (2019) "Differential privacy preservation in deep learning: Challenges, opportunities and solutions." IEEE Access 7: 48901-48911.
[8] D.L. Quoc, M. Beck, P. Bhatotia, R. Chen, C. Fetzer, T. Strufe, Privacypreserving stream analytics: The marriage of randomized response andapproximate computing, Comput. Res. Repos. (1) (2017) 1–23.
[9] H. Cao, S. Liu, Z. Guan, L. Wu, H. Deng, X. Du, An efficient privacy-preserving algorithm based on randomized response in iot-based smartgrid, 2018, CoRR, abs/1804.02781.
[10] U. Erlingsson, V. Pihur, A. Korolova, RAPPOR: Randomized aggregatable privacy-preserving ordinal response, in: Proceedings of ACM SIGSAC'14,2014, pp. 1054–1067.
[11] R. Bassily, A.D. Smith, Local, private, efficient protocols for succincthistograms, in: Proceedings of ACM STOC'15, 2015, pp. 127–135.
[12] Tamanna Kachwala and Sweta Parmar, "An Approach for Preserving Privacy in Data Mining," International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)2014 pp. 370-373, ISSN: 2277 128X.
[13] Arora, D. K., Bansal, D., & Sofat, S. Comparative Analysis of Anonymization Techniques. In International Journal of Electronic and Electrical Engineering.ISSN 0974-2174 Volume 7, Number 8 (2014), pp. 773-778.
[14] Li, Ninghui, Li, Tiancheng, Venkatasubramanian, Suresh. (2007) "t-closeness: Privacy beyond k-anonymity and l-diversity." 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, IEEE: 106-115.
[15] https://archive.ics.uci.edu/ml/datasets/Hepatitis
[16] Zhang, S., Li, X., Tan, Z., Peng, T. and Wang, G., 2019. A caching and spatial K-anonymity driven privacy enhancement scheme in continuous location-based services. Future Generation Computer Systems, 94, pp.40-50.
[17] Sweeney, L 2002, 'Achieving k-anonymity privacy protection using generalization and suppression', International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 571-588.