An Additive Rotational Perturbation Technique for Privacy Preserving Data Mining

Sangeetha Mariammal[a], S Dr.A.Kavithamani[b], and Srikaran Baradhwaj[c]

[a]Assistant Professor, CSE Department.
[b]Associate Professor, EEE Department, Coimbatore Institute of Technology, Coimbatore-641014.
[c]UG Student, CSE Department.

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April2021

Abstract: As the usage of internet and web applications emerges faster, security and privacy of the data is the most challenging issue which we are facing, leading to the possibility of being easily damaged. The privacy preservation techniques like condensation, randomization and tree structure etc., are having limitations, they are not able to maintain proper balance between the data utility and privacy and it may have the problem with privacy violations. This paper presents an Additive Rotation Perturbation approach for Privacy Preserving Data Mining (PPDM). In this proposed work, various dataset from UCI Machine Learning Repository was collected and it is protected with a New Additive Rotational Perturbation Technique under Privacy Preserving Data Mining. Experimental result shows that the proposed algorithm's strength is high for all the datasets and it is estimated using the DoV (Difference of Variance) method.

Keywords: data perturbation, privacy preserving data mining, additive rotation perturbation, difference of variance

1. Introduction

Data Mining has emerged by the nature of discovering useful information from large datasets, addresses many challenges including the privacy issues during data mining and it has been an active and an interesting research area.

Various data mining methods are consolidating security assurance systems, have been created dependent on various irritation draws near. Late examination in the region of PPDM has been dedicated a lot of exertion to decide the compromise among security and utility, the requirement for information disclosure, which is significant to improve dynamic cycles. PPDM assists with securing individual, exclusive or touchy data, to empower coordinated effort between various information proprietors and furthermore to agree to the authoritative strategies. This paper targets to executing the Perturbation methods for ensuring the protection of the client information. Perturbation techniques have been evolved as a solution to provide confidentiality on users' data by converting it into an incomprehensible form. Data Perturbation involves modification of data by adding a small noise or changing the structure of the data. Perturbation techniques are given in Figure 1.

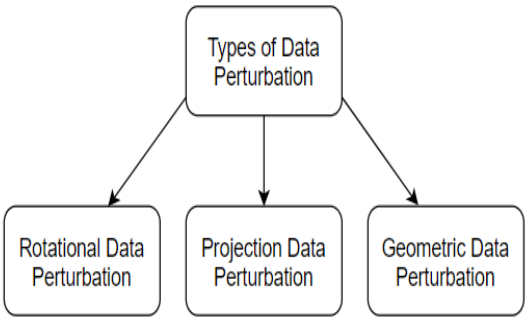

Fig 1. Types of Data Perturbation

1.1 Rotational Perturbation algorithm

In this strategy the estimation of the two credits in the grid are turned however the significance of the worth is ensured.

1.2 Projection Perturbation Algorithm

The data perturbation is done by moving the data value in high dimensional space to the lower dimensional space randomly.

1.3 Geometric Perturbation Algorithm

Hybrid strategy with the mix of revolution, interpretation and enhancing the given information esteem in the grid to give nature of information conservation for the most part for clusters are called geometric perturbation approach.

This paper summarizes the different chapters. Chapter II describes about literature survey of related works, Chapter III presented the proposed methodology, Chapter IV shows experimental discussion and result analysis. Chapter V concluded the proposed work.

2. Literature Survey

Bipul Roy (2014) the way to deal with PPDM which comprises of delivering the information as a synopsis that permits the assessment of specific classes of total inquiries while concealing the individual records. One might say, outline broadens randomization; however a rundown is frequently expected to be a lot more limited, in a perfect world of sub-straight size as for the first dataset. The thought returns to factual information bases, where two rundown strategies were contemplated and broadly applied inspecting and plain information portrayal. Inspecting compares to supplanting the private dataset with a little example of its records, frequently joined with concealment or bother of their qualities to forestall re-ID the techniques dependent on the information annoyance approach fall into two principle classifications known as likelihood circulation class and fixed information irritation class. In the likelihood appropriation class, the security control technique replaces the first information bases by another example from a similar circulation or by the dissemination itself. Then again, the fixed information bother techniques examined in the writing review has been grown solely for either mathematical information or downright information. These techniques as a rule necessitate that a devoted changed information base is made for optional use, and they have developed from a basic strategy for a solitary property to multi characteristic strategies.

Md Nadeem Ahmed and Mohd Hussain (2014) the approach identified with Web Services zeroing in on the different weaknesses and assaults and uncovering a portion of the current accessible shielding strategies which, being not versatile, are not adequate counter measure for those Web Services assaults. In this structure, at first, typical client response, practices and administration demand/reaction are caught and profiled. Operators which go about as sensors are then additionally used to distinguish the presumed things. Moreover understanding on these farfetched things is finished by utilizing affiliation rule-based, bunching and successive standard based methods along with fuzzy rationale. Record esteems and assault markers are then connected to these things to demonstrate the degree of earnestness or the high likelihood of them being genuine assaults. As clarified with the models that the lower the list esteem, the higher is the likelihood that the abnormality is a veritable assault and further preventive advances would then be able to be taken. Subsequent to advancing this structure as the primer thought, further innovative work towards conceptualizing it will be on-going later on.

Territories of interest ought to be equipped into two, first towards investigating the different information mining and fuzzy rationale calculations with the mean to advance the presentation of the system, and move towards making sure about the Semantic Web. By and by, there are as yet numerous investigates about security arrangements in Web-based applications. However, it can distinguish that past investigates don't have proposed a total case for illuminating the security issue and improving the exhibition issue. The ISPWAD of this paper coordinates Secure Web Application Project (SWAP) and Role Based Access Control (RBAC) ways to deal with give an absolute answer for disposing of security hazard and improving framework throughput in planning Web-based applications. The motivation behind ISPWAD approach is to fix the security hole and to improve the handling execution during the plan of Web-based applications. Additionally, it has been represented the strategy to actualize the safe Web-based applications with tuning execution.

Keke Chen (2016) has presented the random rotation based approach of multidimensional perturbation for privacy preserving scheme. Irregular pivot annoyance irritates numerous sections in a single change, which presents new difficulties in assessing the security ensure for multidimensional bother. Planning of a brought together protection metric dependent on esteem range standardization and multicolumn security structure model is finished. With this bound together security metric model one can ready to locate the neighborhood ideal turn bother as far as protection ensure. The bound together protection metric additionally empowers us to recognize and break down the versatility of the revolution irritation approach against the ICA-based information recreation assaults. Here the exploratory outcome shows that the mathematical revolution approach not just jam the exactness of the pivot invariant classifiers, yet in addition gives a lot higher security ensure, contrasted with the current multi-dimensional bother strategies.

P. Bertok et.al. (2018) have proposed a data stream perturbation algorithm (P^2RoCAl). It gives higher exactness, proficiency and assault versatility than comparable techniques. It was indicated that the runtime intricacy is represented by grouping when the quantity of properties is kept steady. The calculation shows the most pessimistic scenario runtime intricacy of O(n3) when the quantity of tuples is kept up as a steady. This makes it conceivable to work with persistently developing information streams and enormous information.

The P2RoCAl technique shows preferable order corrects over its competitors. P2RoCAl additionally shows higher versatility against the assaults, for example, credulous assessment, I/O assaults, and ICA assaults. This P2RoCAl strategy is a successful irritation technique for information streams and huge information. One possible use of P2RoCAl may be the accuracy wellbeing space where countless IoT gadgets are or will be utilized to screen an individual's body, exercises, and practices.

3. Proposed Methodology

3.1 Algorithm: Additive Rotation Perturbation

Input: CSV File with complete numerical data.
Output: CSV File with perturbed data after applying both rotation and addition modules.

Steps:

1. Read the data from the input csv file.
2. Find the data length of the input csv file.
3. Round off the data length to the nearest largest square number.
4. Store the input data in a list.
5. Append zeros to the list such that the length of the input data is the square number that was calculated in Step 3.
6. Convert the list into a square matrix of order n where n is calculated as the square root of the number calculated in Step 3.
7. The square matrix is rotated 90 degrees in the clockwise direction. The rotation is done from the top left of the matrix from 2x2 , 3x3 upto nxn.
8. The modified matrix is now converted into a list.
9. Add the second number with the first number, third number with the second and so on until the last digit is added to the second last one. The last digit is left untouched so that it acts as a reference while recovering the original data.
10. The resultant perturbed list is stored in a csv file.

The overall process flow of proposed algorithm is shown in Figure 2.This Proposed perturbation algorithm can work efficiently for all datasets. The time complexity of the proposed scheme is O(n2). There are two phases in this algorithm. They are rotation and addition phases.

Phase 1: This rotation phase deals with the rotation of the square matrix from 2*2 upto n*n matrix where n is the order of the square matrix.

Phase 2 : This addition phase deals with the addition of the successive numbers after converting the rotated matrix form into a list of value.

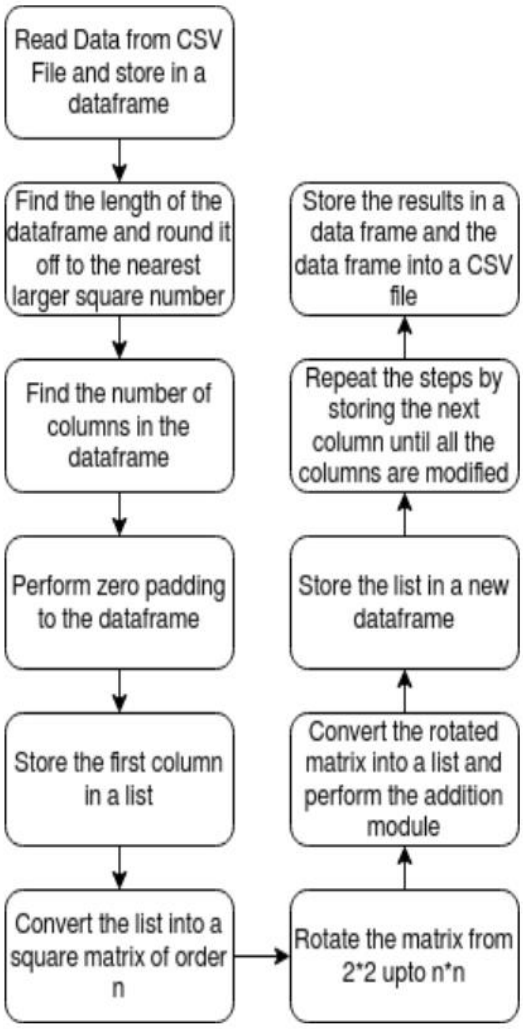

Fig 2. Overall Process Flow of Proposed Algorithm

4. Results and Discussion

The algorithm was implemented in python 3 and applied for various datasets having numeric data. A sample dataset with 4 columns and 9 rows is taken up as shown in Figure 3 and the perturbation results are obtained as shown in Figure 4.

	C1	C2	C3	C4
0	1	7	45	10
1	2	8	55	20
2	3	9	65	30
3	4	10	75	40
4	5	11	76	41
5	6	12	77	42
6	7	13	78	43
7	8	14	79	44
8	9	15	80	45

Fig 3. Sample Data Set

	0	1	2	3
0	12.0	24.0	154.0	84.0
1	9.0	21.0	151.0	81.0
2	12.0	24.0	154.0	84.0
3	10.0	22.0	134.0	64.0
4	3.0	15.0	100.0	30.0
5	10.0	22.0	125.0	55.0
6	15.0	27.0	157.0	87.0
7	9.0	21.0	142.0	72.0
8	3.0	9.0	65.0	30.0

Fig 4. Final Renovated Dataset

Table 1: Variance of the Original and Perturbed Sample Dataset

Sample Dataset	Variance of the original dataset	Variance of the perturbed dataset
Column 1	7.5	15.94445
Column 2	15	45.22223
Column 3	168.75	997.22223
Column 4	322.5	1504.91667

Table 1 shows the Difference of Variance (DoV) values after applying variance method between original and perturbed sample dataset. The Caller Feedback dataset contains 5041 rows and 4 columns of data. The variance of the original and perturbed Caller Feedback dataset is given in Table 2. The area dataset has 5476 rows of data and 10 columns. The variance of the original and modified area dataset is given below in Table 3.The Iris dataset contains 150 rows and 4 columns of data. The variance of the original and modified dataset is provided in the Table 4.

Table 2: Variance of the Original and Perturbed Caller Feedback Dataset

Caller Feedback Dataset	Variance of Original Dataset	Variance of Modified Dataset
Column 1	2083750	6374532.671
Column 2	2083752.372	6374538.036
Column 3	2083858.513	6374777.621
Column 4	2084874.349	6377068.663

Table 3: Variance of the Original and Perturbed Area Dataset

Area Dataset	Variance of the original dataset	Variance of the modified dataset
Column 1	359.50299	719.08548

Column 2	13520.58463	30351.77952
Column 3	23529977.38864	47124491.32973
Column 4	23530920.10812	47126475.98150
Column 5	23530920.13971	47126476.05265
Column 6	23530920.16883	47126476.11424
Column 7	23535692.07993	47136206.04397
Column 8	25149438.21896	50477108.36397
Column 9	28689496.65718	57888556.45340
Column 10	28717488.74509	57948591.63394

Table 4 Variance of the Original and Modified Iris Dataset

Iris Dataset	Variance of Original Dataset	Variance of Modified Dataset
Column 1	0.68570	9.04116
Column 2	0.87372	11.75630
Column 3	3.98688	21.73737
Column 4	4.56930	23.35522

The experimental results measures how far each number in the set the higher the DoV value, then privacy level is also high. There is no direct method for the recovery of the data and thus, even if the person who is trying to access the data gets an idea about the data, they won't be able to recover the data completely.

.5. Conclusion

The proposed additive rotational perturbation algorithm has been devised to work as predicted. Level of privacy is measured through variance and according to DoV values, it is observed that the protection and privacy of the dataset has increased after the perturbation algorithm is run over it. This proposed algorithm will be of great use for machine learning applications since this works mainly on numerical data. This algorithm helps in protecting data that is being shared or being stored. Also it is highly helpful for companies and organizations that share large amounts of data. So the proposed algorithm is an efficient and simple method to increase the privacy and security of the sensitive data. This algorithm can also be extended to non-numeric data by converting them to their ASCII codes; thereby one can get a numeric dataset.

References

1. Aruna Kumari .D, Y. Vineela, T. Mohan Krishna and B. Sai Kumar, (2016) *"Analyzing and Performing Privacy Preserving Data Mining on Medical Databases", Indian Journal of science and Technology, 9(17).*
2. Aradhyula, T.V., Bian, D., Reddy, A.B., Jeng, Y.R., Chavali, M., Sadiku, E.R. and Malkapuram, R., 2020. Compounding and the mechanical properties of catla fish scales reinforced-polypropylene composite–from biowaste to biomaterial. *Advanced Composite Materials*, *29*(2), pp.115-128.

3. Arunkarthikeyan K., Balamurugan K. & Rao P.M.V (2020) Studies on cryogenically treated WC-Co insert at different soaking conditions, Materials and Manufacturing Processes, 35:5, 545-555, DOI: 10.1080/10426914.2020.1726945
4. Babu, U.V., Mani, M.N., Krishna, M.R. and Tejaswini, M., 2018. Data Preprocessing for Modelling the audulteration detection in Gasoline with BIS. *Materials Today: Proceedings*, *5*(2), pp.4637-4645.
5. Bertok .P, D. Liu b, S. Camtepe b, I. Khalil a, (2018) *"Efficient data perturbation for privacy preserving and accurate data stream mining", 48.*
6. Bipul Roy, (2014) *"Performance analysis of clustering in privacy preserving data mining",International journal of computer applications and information security, 5, Issue II*.
7. Clifton, C. (2003) *Tutorial: Privacy-preserving data min-ing. Proc. of ACM SIGKDD Conference.*
8. Ezhilarasi, T.P., Kumar, N.S., Latchoumi, T.P. and Balayesu, N., 2021. A Secure Data Sharing Using IDSS CP-ABE in Cloud Storage. In Advances in Industrial Automation and Smart Manufacturing (pp. 1073-1085). Springer, Singapore.
9. Han .J and M. Kamber. (2007) *Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.*
10. Keke Chen Ling Liu, (2010) *A Random Rotation Perturbation Approach to Privacy Preserving Data Classification, Kno.e.sis Publication.*
11. Keke Chen , Gordon Sun , Ling Liu. (2007) *Towards Attack-Resilient Geometric Data Perturbation , Kno.e.sis Publications.*
12. Palaniswami .S and A Rajaram, (2010) *"The modified security scheme for data integrity in MANET," Inter. Jour. of Engg. Comp. Sci., 1(1): 1-6.*
13. Pujari .A .K. (2007) *Data Mining Techniques. Universities Press*.
14. Mahalle .V .S .Prof , Pankaj Jogi , Urvashi Ingale , Shubham Purankar , Samiksha Pinge. (2017) *"Data Privacy Preserving Using Perturbation Technique, Asian Journal of Convergence in Technology, 3, Issue 3.*
15. Vaidya J, and Clifton C, (2002) *Privacy preserving association rule mining in vertically partitioned data. Proc. of ACM SIGKDD Conference.*
16. Yarlagaddaa, J., Malkapuram, R. and Balamurugan, K., 2021. Machining Studies on Various Ply Orientations of Glass Fiber Composite. In Advances in Industrial Automation and Smart Manufacturing (pp. 753-769). Springer, Singapore.
17. Yarlagaddaa, J. and Malkapuram, R., 2020. Influence of carbon nanotubes/graphene nanoparticles on the mechanical and morphological properties of glass woven fabric epoxy composites. *INCAS Bulletin*, *12*(4), pp.209-218.