HiMod-Pert: Histogram Modification Based Perturbation Approach for Privacy Preserving Data Mining

Alpa Kavin Shah[1(✉)] and Ravi Gulati[2]

[1] MCA Department, Sarvajanik College of Engineering and Technology, Surat, India
alpa.shah@scet.ac.in
[2] Department of Computer Science, Veer Narmad South Gujarat University, Surat, India
rmgulati@vnsgu.ac.in

Abstract. Privacy Preserving Data Mining (PPDM) protects the disclosure of sensitive quasi-identifiers of dataset during mining by perturbing the data. This perturbed dataset is then used by trusted Third Party for effective derivation of association rules. Many PPDM algorithms destroy the original data to generate the mining results. It is essential that the perturbed data preserves the statistical inference of the sensitive attributes and minimize the information loss. Existing techniques based on Additive, Multiplicative and Geometric Transformations have minimal information loss, but suffer from reconstruction vulnerabilities. We propose Histogram Modification based method, viz. HiMod-Pert, for preserving the sensitive numeric attributes of perturbed dataset. Our method uses the difference in neighboring values to determine the perturbation factor. Experiments are performed to implement and test the applicability of the proposed technique. Evaluation using descriptive statistic metrics shows that the information loss is minimal.

Keywords: Privacy preserving data mining · Histogram Modification
Additive white Gaussian noise · Multiplicative perturbation
Geometric Data Perturbation

1 Introduction

Since last couple of decades, information collection over Internet is witnessing an exponential growth. More users have started providing their personal information in different Internet based activities like purchases/sales, auctions, entertainment, gaming, online surveys, to name a few. A person can now be easily and accurately linked based on his/her Internet activities, leading to a serious pose of privacy intrusion to the individuals. This vast pool of data has necessitated the need for efficient data mining protocols. Data mining which was limited and confined to narrower domain of Enterprises and Applications now encompasses Big Data and Cloud Computing.

Data collection has increased many-folds for research, trend analysis and more often collaborative mining results. It is vital that the information provided by the users should not breach their privacy. This concern has caught attention of researchers and is widely studied for improvements even today. PPDM algorithms tackle this issue by optimizing

Z. Patel and S. Gupta (Eds.): ICFITT 2017, LNICST 220, pp. 28–36, 2018.
https://doi.org/10.1007/978-3-319-73712-6_3

privacy and minimizing information loss. This segment of Data Mining guard individual's privacy in Data Mining applications while providing accurate results for mining. An efficient PPDM algorithm must maximize privacy and minimize information loss. Also, it is desirable that the computational cost of generating perturbed data should be minimal and feasible. This requirement augments need of techniques that can be easily implemented by the contributing party. PPDM methods must protect the privacy of data and prevent adversaries to derive correlation between the distributed data. Our study focuses on developing a robust method which will be implemented by contributing party before releasing the data to Third Party. The proposed method will preserve the privacy of sensitive attributes and minimize the information loss.

1.1 Our Contributions

We make following contributions with this research:

1. We propose a Histogram Modification based method for perturbing numeric attributes for achieving privacy.
2. The resultant perturbed data obtained from proposed method is evaluated for efficiency in mining using statistical metrics. Also, a comparison of the proposed technique with basic perturbation techniques is conferred to show the effectiveness of our *HiMod-Pert* method.

1.2 Organization of the Paper

The rest of the paper is organized as follows:

- Section 2 insights the literature survey;
- Section 3 details the proposed *HiMod*-Pert method;
- In Sect. 4, we present the experimental results and present a comparison with contemporary perturbation techniques using descriptive statistical metrics;
- Finally, in Sect. 5, we provide conclusions and propose a road map for future work.

2 Literature Survey

Perturbation methods discussed by authors Domingo-Ferrer et al. [20] and Herranz et al. [21] are widely used in PPDM because the computational cost is lower than Cryptographic and Secure Multiparty Computations based methods. The former also has an edge over, as these methods can be used either by the data owner or by Trusted Third Party. Statistical Databases (SDBs) worked by authors Adam and Wortmann [1], Duncan and Mukherjee [2] and Gopal et al. [3] provide summary statistical information without sacrificing individual's sensitive identifying attributes. Numerical sensitive attributes of an application after perturbation must preserve the descriptive statistics for accurate mining. Perturbation Methods change the original data in a way that the summary statistics of the perturbed data remains same as that of the original data. For data mining to be effective, the perturbed data must preserve the relationships amongst

the contributing attributes. Authors Liu et al. [4] have exemplified the applicability of perturbation techniques by distorting the original values with known distribution, a category of probability distribution based perturbation approach. Authors Bai Li and Sarkar [7] have described a tree-based perturbation method. In this method, the original dataset values are replaced with fixed set of values. This technique is a type of fixed-perturbation based technique where values belonging to same group are replaced with certain defined values.

Perturbation methods Clifton et al. [5] and Kargupta et al. [6] broadly fall into three basic categories viz. Additive, Multiplicative and Rotation based methods like Geometric Data Perturbation (GDP). In Additive based method, first introduced by Agrawal and Srikant [10], randomized noise from known distribution sample like Gaussian is added to original data. If x_i is the original data values, and ε is random noise from some distribution like Gaussian or Uniform, new perturbed value $x_i + \varepsilon$ will appear instead of x_i. Many reconstruction approaches worked by authors Agrawal and Aggarwal [11], Domingo-Ferrer et al. [12] and Kargupta et al. [13] ascertain the vulnerability in privacy breaches with the use of Additive methods.

Another category of perturbation is Multiplicative based approach in which the Euclidean Distance is preserved well between the perturbed data and the original dataset. If x_i are the original data values and R is rotation matrix, perturbed values are computed as $R * x_i$. Independent Component Analysis (ICA) suggested by Liu et al. [15] when applied to perturbed values generated by multiplicative methods, can approximate original values. Work done by authors Liu et al. [14, 15] and Giannella et al. [16] suggest that the Multiplicative based methods have high privacy breach probability. Geometric based perturbations proposed by Chen et al. [17] add a random translation to values perturbed by Gaussian distribution. Their work enhances the resilience of random perturbation against three types of inference attacks: Naïve Inference attacks, ICA-based attacks and Distance-Inference attacks.

Our motive to present this study is to overcome the vulnerability due to randomized approaches and possible data reconstruction from original data. Researchers have given special attention to this and have presented novel studies for gaining knowledge from perturbed data. The recovery approach also is dependent on relative noise. The randomized approach of adding/multiplying Gaussian or Uniform noise to the original data sets does not ensure quality of data recovery process. In a cloud based environment it is essential at times to verify the integrity of the perturbed data. Sang et al. [22] have proposed and experimented reconstruction based on Undetermined Independent Component Analysis (UICA) where attacker has full or zero background information about perturbation matrix. Their studies clearly reveal the vulnerability of perturbation methods based on random and orthogonal projections. The authors' prior work Shah and Gulati [24] has revealed that the Additive and GDP based perturbation preserves the statistical inference of the original dataset and multiplicative perturbation methods generate records with minimum information loss but does not preserve statistical inference.

The Histogram Modification method suggested by authors Ni et al. [8] and Tai et al. [9] is a type of Data Hiding mechanism that works on images. It is a branch of Steganography where sensitive information is embedded into an image, making hiding imperceptible to humans. The image at receiver's end can be restored and the secret

information can be retrieved. The Histogram Modification technique is not suitable when images have equal Histograms. In Histogram Modification Technique data hiding is performed based on the difference of adjoining pixel values. To successfully retrieve the secret message and image, receiver must be passed the various peak points and zero points.

3 The Proposed Method

Rather than directly perturbing the values based on noise, we propose Histogram Modification based method for perturbing values. The proposed method does not add noise to all sensitive attributes like generic Additive and Multiplicative methods. Our proposed method uses difference in the adjoining neighbor values of dataset to generate noise which will then be added to or subtracted from the sensitive values. We have used peak as a measure of average of the difference of the adjoining data values. This peak value along with difference between adjoining data value is used to compute the perturbation factor. This perturbation factor will be different for each value. Unlike randomized Gaussian noise, this perturbation factor is dependent on the integrity of the dataset

Algorithm *HiMod-Pert*: Given the sensitive numeric attribute of a dataset, this algorithm returns its corresponding privacy preserving perturbed value.

Input: Let S, S = {s_i, i=1, 2, 3….n}, be the numeric sensitive attribute of Dataset D needed to be privatized before making D public.

Output: P is the perturbed sensitive attribute, P = {p_i, i=1, 2, 3…n} generated after applying ***HiMod-Pert*** method.

Step 1: [Calculate the difference between adjoining attributes of S. If the number of attributes n is odd, for the last attribute, the mean is used to calculate the difference.]

$$\text{diff}_i = \begin{cases} |s_i - s_{i+1}|, & i \neq n \\ |s_i - \frac{1}{n}\sum_{i=1}^{n} s_i|, & i = n \text{ and } n \text{ is odd} \end{cases} \quad (1)$$

Step 2: [Determine Peak from the difference obtained in Step 1]

$$\text{Peak} = \frac{1}{n}\sum_{i=1}^{n} diff_i \quad (2)$$

Step 3: [Calculate the Perturbation Factor p_{factor} for each of the sensitive attributes s_i.]

$$P_{factor_i} = (\text{Peak} - \text{diff}_i)/\text{Peak} \text{ for i } = 1,2,3 \ldots \text{n} \quad (3)$$

Step 4: [Apply the Perturbation Factor to s_i calculated in Step 3]

$$P_i = \begin{cases} s_i, & if\ i = 1\ or\ diff_i \leq Peak, \\ s_i + P_{factor_i}, & if\ diff_i > Peak\ and\ s_i \geq s_{i-1} \\ s_i - P_{factor_i}, & if\ diff_i > Peak\ and\ s_i < s_{i-1} \end{cases} \quad (4)$$

Fig. 1. Algorithm for proposed *HiMod-Pert* method

values. The first value of the dataset is not perturbed. In last step, the perturbation factor is added or subtracted to the original values based on the adjoining values.

Our proposed work does not embed message bit as our aim is to perturb the values and not hide any data. It can be extended to embed a message bit for increased privacy. The message bit must be shared between the contributing parties before perturbation. The integrity will be compromised with actions like deleting a sensitive record, changing the values, subsequently adding significant information loss to the mining results. Having briefed up the basic logic of the proposed method, we will now outline the algorithm of *HiMod-Pert* method based on Histogram Modification for applicability in privacy preserving data mining. We have considered that the sensitive attribute is application specific and can be identified using Decision Tree. The algorithm can be iteratively applied to perturb all the sensitive numeric attributes of the dataset. Figure 1 on subsequent page details our proposed HiMod-Pert algorithm.

4 Experimental Evaluation

4.1 Setting Environment

Experimental setup was done in MATLAB tool. The privacy attributes (columns) of the test data are determined by using Decision Tree suggested by authors Matatov et al. [25] and Fung et al. [26]. The Decision Tree sorts the columns by importance which can then be chosen for perturbation. Selection of four different datasets based on sizes of small, medium and large were chosen to test the performance of the proposed method. We have considered two datasets viz. ADULT and BREAST-CANCER–W from UCI Repository [18]. Both datasets contain large number of records and they exhibit real-world scenario. We have perturbed the numeric attributes Age and ID of the ADULT and BREAST CANCER-W dataset respectively. Another dataset HALD is available inbuilt with MATLAB. We have used INGREDIENTS dataset array from it as it is a Statistical Database. Lastly, we have also used NBASalaries dataset available from [19]. The attribute Salary was considered confidential. Table 1 describes the datasets used for our experimentation purpose and details number of instances and attributes. To provide a comparative analysis with basic perturbations, we have also simulated functions for Additive Perturbation, Multiplicative Perturbation and Geometric Data Perturbation in MATLAB. These methods are used as a baseline for comparison against our *HiMod-Pert* method.

Table 1. Datasets

Dataset	Number of instances	Number of attributes
ADULT	32561	15
BREAST-CANCER Wisconsin	699	10
INGREDIENTS	13	4
NBASalaries	407	6

4.2 Experimental Results

We have implemented the *HiMod-Pert* method in MATLAB. To show the performance of the proposed method, descriptive statistical measures like Mean, Standard Deviation, Mean Square Error, Root of Mean Square, Mean Absolute Error and Euclidean Distance are taken into consideration. For effective comparison, Table 2 consolidates the results generated on original Dataset, Additive, Multiplicative, GDP and our proposed *HiMod-Pert* method for various statistical metrics.

Table 2. Results of descriptive statistical measures on original and perturbed dataset generated by Additive, Multiplicative, GDP and *HiMod-Pert* method

Perturbation Techniques	Mean	Standard Deviation	Mean Square Error	Root of Mean Square	Mean Absolute Error	Euclidean Distance
ADULT DATASET						
Original DS	38.58	13.64	–	–	38.5816	–
Additive	38.59	13.67	0.98	41.00	38.60	178.87
Multiplicative	0.26	41.03	3.35e+03	41.16	31.00	1.04e+04
GDP	39.30	13.64	1.75	42.17	39.90	238.54
HiMod-Pert	***38.5449***	***13.3428***	***0.3634***	***40.7889***	***38.5449***	***108.7724***
BREAST CANCER-W DATASET						
Original DS	1.07e+06	6.17e+05	–	–	1.07e+06	–
Additive	1.07e+06	6.17e+05	0.99	1.23e+06	1.07e+06	26.30
Multiplicative	3.45e+04	1.05e+06	2.57e+12	1.18e+06	8.23e+05	4.25e+07
GDP	1.07e+06	6.17e+05	1.057	1.24e+06	1.07e+06	27.13
HiMod-Pert	***1.0352e+06***	***1.7688e+05***	***2.1669***	***1.0353e+06***	***1.0352e+06***	***7.2115***
INGREDIENTS DATASET						
Original DS	48.1538	15.5609	–	–	48.1538	–
Additive	48.4089	15.3281	0.5395	50.5994	48.4089	2.6484
Multiplicative	31.1067	27.6588	732.2388	40.9120	31.1067	114.9051
GDP	48.3421	15.7801	0.56	51.3433	48.0952	3.4452
HiMod-Pert	***48.0404***	***15.3437***	***0.2274***	***50.2514***	***48.0404***	***1.7192***
NBASALARIES DATASET						
Original DS	4.4695e+06	4.6933e+06	–	–	4.4695e+06	–
Additive	4.4695e+06	4.6933e+06	0.8221	6.4768e+06	4.4695e+06	18.2919
Multiplicative	3.6857e+06	5.5919e+06	4.8124e+13	6.6916e+06	3.6857e+06	8.3264e+07
GDP	4.5673e+06	4.6924e+06	0.3412	6.4523+06	6.4523+06	15.2347
HiMod-Pert	***4.4695e+06***	***4.6933e+06***	***0.6380***	***6.4768e+06***	***4.4695e+06***	**16.1143**

4.3 Experimental Inferences

Statistical Measures are used to check the applicability of perturbation techniques for information loss and privacy breach. The use of probabilistic information loss discussed by Mateo-Sanz et al. [23] is used to evaluate the information loss of the perturbed data. Mean, Standard Deviation (SD), Mean Square Error (MSE), Mean Absolute Error (MAE) and Root Mean Square (RMS) are used to evaluate the information loss for the

perturbed dataset. These statistical measures are necessary to prove the information loss for perturbed sets but not sufficient to conclude the same.

Mean and Standard Deviation measures the univariate information loss after perturbation. The experiments show that the Mean and Standard Deviation of original dataset is very near to perturbed dataset generated by our proposed *HiMod-Pert* method. This ensures that the proposed method efficiently preserves the clusters of original dataset.

Mean Square Error is a measure of average of squares of deviation of the original values from the perturbed values. For perturbed values to accurately estimate the original values, mean square error should be near to 1. The proposed *HiMod-Pert* method will generate MSE near to 1. Unlike Multiplicative method, it is efficient in preserving this statistical metric. Mean Absolute Error forecast how close are the perturbed values to the original values. It measures the distance between values generated by perturbation methods and original unperturbed values. The values in all the four datasets in our experimentation show that the Mean Absolute Error is same as that of original.

Euclidean Distance is a measure of how the values in perturbed dataset are linked with the original values. Smaller Euclidean Distance suggests that the probability of linkage of perturbed values to the original values is high. The proposed method uses adjacent values for finding the perturbation factor. Hence the record linkage is high. Both Additive and GDP methods have Euclidean Distance measures very less, indicating high record linkage. Root Mean Square of an estimator is the measure of imperfection of the fit of the perturbed data to the original data. For our *HiMod-Pert*, the value of Root Mean Square is effectively retained. The result, shown in Table 2 suggests that descriptive statistics is preserved well by *HiMod-Pert* method.

5 Conclusions

HiMod-Pert - a method based on Histogram Modification for effectively preserving the privacy and optimally minimizing information loss is proposed. We have exploited the traditional method that is used in Image Steganography. The method uses the differences in neighbouring sensitive attributes to modify the original values. Unlike contemporary methods where the transformation is fixed or based on randomization, we have suggested use of conditional perturbation factor that will be computed for each privacy sensitive attribute. Our experiments show that the method is effective for balancing between information loss and disclosure risk. Future work encompasses in studying the impact of various attacks, variations caused due to compromise in integrity and optimizing the method to combat against attacks.

References

1. Adam, N.R., Wortmann, J.C.: Security-control methods for statistical databases: a comparative study. ACM Comput. Surv. **21**(4), 515–556 (1989)
2. Duncan, G.T., Mukherjee, S.: Optimal disclosure limitation strategy in statistical databases: deterring tracker attacks through additive noise. J. Am. Stat. Assoc. **95**(451), 720–729 (2000)

3. Gopal, R., Garfinkel, R., Goes, P.: Confidentiality via camouflage: the CVC approach to disclosure limitation when answering queries to databases. Oper. Res. **50**(3), 501–516 (2002)
4. Liu, L., Kantarcioglu, M., Thuraisingham, B.: The applicability of the perturbation based privacy preserving data mining for real-world data. Data Knowl. Eng. **65**, 5–21 (2007)
5. Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., Zhu, M.: Tools for privacy preserving distributed data mining. SIGKDD Explor. **4**(2), 38–44 (2002)
6. Kargupta, H., Datta, S., Wang, Q., Sivakumar, K.: Random data perturbation techniques and privacy preserving data mining. In: IEEE International Conference on Data Mining (2003)
7. Bai Li, X., Sarkar, S.: A tree-based data perturbation approach for privacy-preserving data mining. IEEE Trans. Knowl. Data Eng. **18**(9), 1278–1283 (2006)
8. Ni, Z., Shi, Y.Q., Ansari, N., Su, W.: Reversible data hiding. In: Proceedings of International Symposium on Circuits and Systems, Bangkok, Thailand, vol. 2, pp. 912–915, 25–28 May 2003
9. Tai, W., Yeh, C., Chang, C.: Reversible data hiding based on histogram modification of pixel differences. IEEE Trans. Circ. Syst. Video Technol. **19**(6), 906–910 (2009)
10. Agrawal, R., Srikant, R.: Privacy preserving data mining. In: Proceedings of ACM SIGMOD Conference on Management of Data, Dallas, Texas, pp. 439–450, May 2000
11. Agrawal, D., Aggarwal, C.C.: On the design and quantification of privacy preserving data mining algorithms. In: Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Santa Barbara, pp. 247–255 (2001)
12. Domingo-Ferrer, J., Sebé, F., Castellà-Roca, J.: On the security of noise addition for privacy in statistical databases. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 149–161. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-25955-8_12
13. Kargupta, H., Datta, S., Wang, Q., Sivakumar, K.: Random-data perturbation techniques and privacy preserving data mining. Knowl. Inf. Syst. **7**(4), 387–414 (2005). https://doi.org/10.1007/s10115-004-0173-6
14. Liu, K., Giannella, C., Kargupta, H.: An attacker's view of distance preserving maps for privacy preserving data mining. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) PKDD 2006. LNCS (LNAI), vol. 4213, pp. 297–308. Springer, Heidelberg (2006). https://doi.org/10.1007/11871637_30
15. Liu, K., Kargupta, H., Ryan, J.: Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Trans. Knowl. Data Eng. **18**(1), 92–106 (2006). https://doi.org/10.1109/TKDE.2006.14
16. Giannella, C., Liu, K., Kargupta, H.: Breaching Euclidean distance-preserving data perturbation using few known inputs. IEEE Trans. Knowl. Data Eng. **83**, 93–110 (2013). https://doi.org/10.1016/j.datak.2012.10.004
17. Chen, K., Sun, G., Liu, L.: Towards attack-resilient geometric data perturbation. In: Proceedings of the 2007 SIAM International Conference on Data Mining, Minneapolis, pp. 78–89 (2007)
18. Lichman, M.: UCI machine learning repository. School of Information and Computer Science, University of California, Irvine (2013). http://archive.ics.uci.edu/ml
19. https://github.com/Kjonge/DemoWorkbooks/blob/master/NBA%20salaries.xlsx
20. Domingo-Ferrer, J., Mateo-Sanz, J.M., Torra, V.: Comparing SDC methods for microdata on the basis of information loss and disclosure risk. In: Proceedings of the International Conference on New Techniques and Technologies for Statistics: Exchange of Technology and Knowhow, pp. 807–826 (2001)
21. Herranz, J., Matwin, S., Nin, J., Torra, V.: Classifying data from protected statistical datasets. Comput. Secur. **29**(8), 874–890 (2010). https://doi.org/10.1016/j.cose.2010.05.005

22. Sang, Y., Shen, H., Tian, H.: Effective reconstruction of data perturbed by random projections. IEEE Trans. Comput. **61**(1), 101–117 (2012)
23. Mateo-Sanz, J.M., Domingo-Ferrer, J., Sebé, F.: Probabilistic information loss measures in confidentiality protection of continuous microdata. Data Min. Knowl. Disc. **11**(2), 181–193 (2005). https://doi.org/10.1007/s10618-005-0011-9
24. Shah, A., Gulati, R: Evaluating applicability of perturbation techniques for privacy preserving data mining by descriptive statistics. In: Proceedings of 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, pp. 621–627, 21–24 September 2016
25. Matatov, N., Rokach, L., Maimon, O.: Privacy-preserving data mining: a feature set partitioning approach. Inf. Sci. **180**(14), 2696–2720 (2010). https://doi.org/10.1016/j.ins.2010.03.011
26. Fung, B.C.M., Wang, K., Yu, P.S.: Anonymizing classification data for privacy preservation. IEEE Trans. Knowl. Data Eng. **19**(5), 711–725 (2007)