International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Gaussian Noise Multiplicative Privacy for Data Perturbation Under Multi Level Trust

[1]Ranjeet Kumar Rai, [2]Dr. Manish Varsney

Submitted: 27/05/2023 **Revised**: 18/07/2023 **Accepted**: 26/07/2023

Abstract: Data mining is the technique of exploring and analyzing huge blocks of information to find significant trends and patterns. Perturbation is a mechanism that has been introduced in the fields of celestial mechanics and mathematical physics. Each characteristic has a weight that represents how accurate and comprehensive it is. Database and data security administrators are forced to perform a difficult balancing act when it comes to granting employees access to organizational data. for this To use multiplicative data perturbation in conjunction with single level and multilayer trust the geometric type of multiplicative data perturbation will be carried out in this method, as well. When generating the perturbed copy, geometric perturbation involves the orthonormal matrix, translational matrix, and a random generated Gaussian noise vector, among other things. In the beginning, the orthonormal matrix will be used to perform the rotation perturbation, and then the translational matrix and Gaussian noise components will be added to it for the final perturbed copy. We can say that under single level trust, additive Gaussian data perturbation produces perturbed copies using uniform Gaussian noise. Regardless of their trust ratings, all data miners receive the same perturbed copy. Additive Gaussian data perturbation at multi level trust is studied for data miners at various trust levels. A common randomization technique that guarantees confidentiality and trustworthy data mining findings is data perturbation.

Keywords: *noise, Gaussian, data perturbation, , trust, multi level*

1. Introduction

Data mining (DM) is a method used in the corporate world to turn raw data into informative information. Using software that looks for patterns in vast volumes of data might help businesses understand their customers better [1]. This enables them to create marketing plans that are more successful, boost sales, and cut expenses. Successful data mining requires efficient data gathering, warehousing, and computer processing. The process of searching and analyzing vast blocks of data in order to find significant patterns and trends is known as data mining (DM) [2]. Applications for this technology include database marketing, credit risk management, fraud detection, spam email screening, and even user sentiment analysis. The data mining procedure can be divided into five parts. In the initial phase of data collecting, information is gathered and loaded into information warehouses. They either manage or keep the data after that on their own servers or in the cloud.

- **Perturbation:**

Perturbation is a mechanism that has been introduced in the fields of celestial mechanics and mathematical physics [3]. How accurate and complete it is. Every constraint involving this attribute is associated with a weight that represents the importance of the violation of that constraint [11][12]. The more faith a data miner has in someone, the less agitated they are when they are provided access to a copy of the data. In this case, a malicious data miner may obtain several altered copies of the same data through a number of different techniques, and he or she may combine these varied copies to collectively derive extra data that the data owner does not wish to make public. Avoiding such diversity attacks is the key challenge in offering MLT-PPDM services [4]. To properly mimic the distribution of the original data values, a novel reconstruction technique has been developed. By making use of these reconstructed distributions, classifiers can be created with accuracy that is on par with or superior to that of classifiers created with the original data. So, as compared to other methods, perturbation processes are the best at preserving privacy.

1.1 Perturbation Techniques for data protection

Organizations today collect massive amounts of data about their competitors, customers, and internal processes [13] [14]. Organizations are constantly struggling to fully utilize their data, and finding "unknown" knowledge within their enormous data vaults continues to be a highly desired objective. Giving employees access to organizational data requires database and data security managers to strike a delicate balance [5]. Access to big data repositories with individual records is extremely beneficial for sophisticated organizations that employ data mining and knowledge discovery methods (such as inductive learning algorithms, neural networks, etc.) to

[1]*Research Scholar, Department of Computer Science, MUIT Lucknow*
ranjeetrai2007@gmail.com
[2]*Prof. School of Engineering & Technology, MUIT , Lucknow*

find previously unidentified "patterns" in their data. The need to prevent specific "confidential" data items in an organizational database from being unlawfully disclosed by other parties is another significant issue that the database administrator must handle [6]. The range of this protection goes beyond more conventional data access problems (such as hackers and unauthorized access), and also involves concealing individual confidential record properties to prevent even authorized users from being able to identify particular records [15] [16].

2. Literature Review

Sun, X., Xu, R., Wu, L. et al (2021) [7] Edge computing's advantages of low latency several applications for data mining The use of computationally expensive data mining techniques is not possible with edge computing due to a lack of computing resources. Participants typically work together in an edge-cloud environment to train machine learning models that result in more accurate prediction outcomes. However, due to privacy issues, data owners can be reluctant to submit their own data. To handle such dissimilar goals, we concentrate on a computationally friendly tree-based distributed data mining technique with differential privacy. The foundation of our solution is a distributed ensemble technique. After being injected with the complicated noise, each person creates a good tradeoff between computation and data distribution accuracy, and shares it with the other participants. Then, in an adaptive ensemble approach, additional players are provided with the useful information provided by the decision models. Both the theoretical analysis and the trials show that our plan offers a useful data mining technique that can produce highly accurate predictions while upholding stringent data privacy.

Swapnil et. al (2017) [8] because to the Social media and Internet, the amount of data available has increased in recent years, as has its accessibility and availability. The DM process is used to search this massive data set and identify previously unknown relevant patterns and forecasts. Data mining is the process of connecting unrelated data in a meaningful way, analyzing it, and presenting the results in the form of valuable data patterns. DM has the potential to compromise sensitive and confidential data. Individual privacy is put at risk if some of the data leaks and identifies a person whose personal information was used in the data mining process. A number of privacy-preserving data mining (PPDM) strategies and techniques are available that are designed to secure sensitive data while yet yielding reliable datamining results. Incorporated into PPDM techniques and processes are many ways to data protection in the data mining process. The techniques employed in this paper were a bibliometric analysis and a systematic review of the literature.

Gunawan Dedi (2020) [9]Databases are increasingly used to collect and store data from a number of sources. Before the database owner does data analysis, such as by using data mining techniques on the databases, the gathering of data has little impact. In terms of quality, accuracy, and precision, data mining techniques and algorithms are significantly improving the information extraction process as they are now being developed. The data owner should take precautions to protect privacy because it is possible for some accessible data mining methods to reveal sensitive information about data subjects in databases. As a result, the field of data mining research is becoming more and more interested in privacy-preserving data mining (PPDM). Examining the detrimental effects of data mining technologies that result from the invasion of individuals' and organizations' privacy is the main objective of PPDM. Even though a sanitized database will have similar data utility to the original, it also guarantees that data miners won't be able to extract any personally identifying information from it. In this study, we provide a comprehensive overview of current PPDM tactics by classifying them according to the features of each strategy using taxonomy approaches.

Sulekh, V. Jane Varamani (2018) [10] Numerous fields, including the Internet of Things (IoT), the medical sector, and commercial development, have researched and used data mining extensively. However, these data mining methods face significant obstacles because of privacy concerns and greater sharing of sensitive information. A subset of data mining called privacy-preserving data mining (PPDM) works to prevent unauthorized or unlawful disclosure of people's personal information. private concerns violate someone's right to private and degrade the study participant. Along with harming one's social and financial standing, it would also bring about social disgrace, shame, and dishonor. In recent years, a number of data mining techniques have been developed that combine privacy-preserving techniques to conceal delicate item sets or patterns. Which privacy-preserving method offers the best protection for sensitive data is a crucial consideration in this situation. The performance of the algorithm and the outcome must both be evaluated after privacy-preserving techniques have been used. In this study, we investigate alternative noise-based privacy preservation techniques.

3. Objectives

- To investigate Protecting Data through ‘Perturbation’ Techniques.
- To evaluate Gaussian noise data perturbation under multi level trust.

4. Research Methodology

The first piece of research suggested makes use of Gaussian noise to perturb sensitive data in both single level and multilayer trust situations. In the first instance, additive data perturbation will be used to perturb the data by using Gaussian noise. Under a single degree of trust, Gaussian noise will be introduced into the sensitive data, and the resulting perturbed copy will be delivered evenly to all data miners, regardless of their trust levels. Different perturbed copies will be generated depending on the trust levels of the data miners, which will be achieved by multilayer trust. When the data miner will operating at a lower trust level, the amount of noise introduced will disproportionately more than when the data miner will operating at a higher trust level.

To use multiplicative data perturbation in conjunction with single level and multilayer trust the geometric type of multiplicative data perturbation will be carried out in this method, as well. When generating the perturbed copy, The mean value is µ and the variance is σ^2. Scalars or vectors can be used for the mean value and variance. The length of the variance vector must be the same as the length of the first seed vector. The covariance matrix in this situation is a diagonal matrix whose diagonal members are drawn from the variance vector. The output Gaussian random variables are uncorrelated because the off-diagonal elements are zero. geometric perturbation involves the orthonormal matrix, translational matrix, and a random generated Gaussian noise vector, among other things. In the beginning, the orthonormal matrix will be used to perform the rotation perturbation, and then the translational matrix and Gaussian noise components will be added to it for the final perturbed copy.

5. Result and Discussion

5.1 Gaussian noise for perturbation of data

Gaussian noise is a statistical noise with a probability density function that is comparable to that of the normal distribution in statistics. Gaussian distribution is another name for normal distribution. Equation gives the probability density function of Gaussian noise (1)

$$gf(x) = \left(\frac{1}{\sqrt{2\pi\sigma}}\right) e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)} \quad (1)$$

5.2 Privacy Preservation Estimation

The original data is reconstructed using Linear Least Square Error (LLSE) based estimation. The higher the error rate in the original data reconstruction, the more privacy will be retained. The noise component σZi^2 is supposed to have varying values for Gaussian noise data perturbation under multi level trust. The classifier accuracy for Gaussian data perturbation at multi level trust is shown in Table 1, with the values of all trust levels averaged. The findings for all three classifier under multi-level trust is depicts in Figure 1.

Table 1 Classifier accuracy for Gaussian data perturbation at Multi Level Trust

	Bank Dataset			**Credit Card Dataset**		
Classifier Accuracy	**Decision Tree**	**Naïve Bayes**	**kNN**	**Decision Tree**	**Naïve Bayes**	**kNN**
Original Data	90.7	89.2	84.9	77.7	54.2	75.3
Gaussian Additive	90.0	88.8	75.0	75.2	51.9	56.0
Gaussian Multiplicative	85.0	84.7	84.2	72.3	50.3	71.0

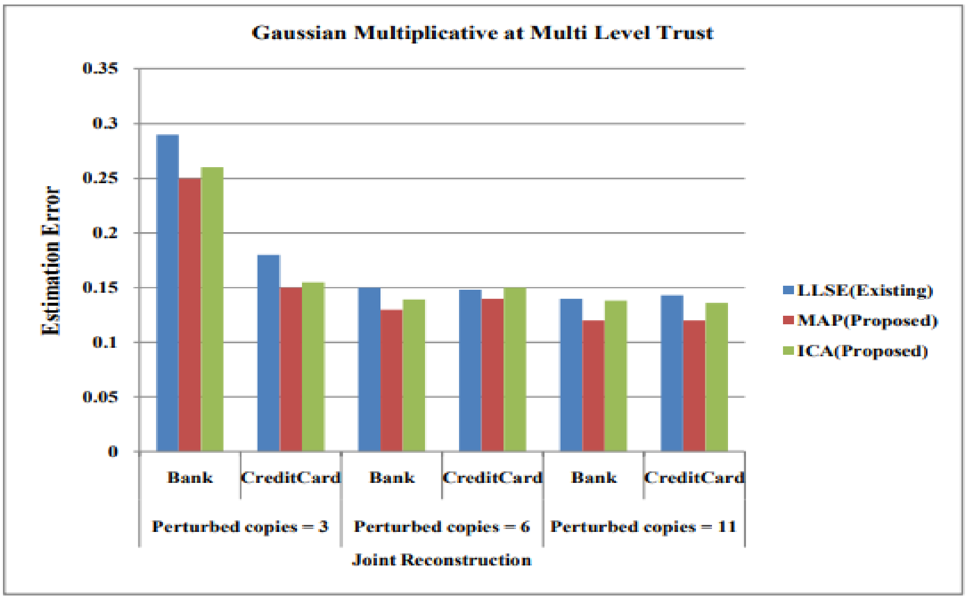

Fig 1: Gaussian Multiplicative Privacy measure under Multi Level Trust

Figure 1 shows the privacy accuracy for the Gaussian multiplicative approach based on the normalized estimation error. The number of perturbed copies represents the number of data miners with varying levels of trust. Copies=3 denotes the reconstruction of the original data from three perturbed data sets (correspondingly for copies 6 and 11). The normalized estimation error occurs when attempting to recreate the original data from perturbed data. If the estimation error is considerable, it suggests that the original data was not rebuilt precisely. In all noise filtering systems, it is obvious that the multiplicative form of data perturbation results in a higher error rate when reconstructing the original data, resulting in higher privacy.

Another Credit card dataset is used to test the approach. When the estimation error is high, the original data is reconstructed incorrectly. The Gaussian Multiplicative perturbation has a greater estimation error in both datasets. This demonstrates that this strategy outperforms the Gaussian additive method in terms of privacy accuracy. Perturbed copies for M levels are generated using a noise component σZi^2 taken from a random Gaussian distribution with multi level trust. The malevolent data miners are supposed to be aware of the noise distribution, mean, and covariance of the original and perturbed data. The results clearly reveal that joint estimation is growing for the Gaussian multiplicative technique, demonstrating that the multiplicative technique achieves the privacy goal more effectively. The estimation error does not change much when the number of perturbed copies available to malicious data miners rises, and it remains steady for varied available copies. In comparison to Gaussian multiplicative perturbation, this exhibits an increased privacy level.

6. Conclusion

We can say that under single level trust, additive Gaussian data perturbation produces perturbed copies using uniform Gaussian noise. Regardless of their trust ratings, all data miners receive the same perturbed copy. Additive Gaussian data perturbation at multi level trust is studied for data miners at various trust levels. Data perturbation is a popular randomization approach that ensures both accurate data mining results and privacy. The additive and multiplicative types of data perturbation have been used in previous research. The increasing amount of error rate depending on various sorts of attacks is used to quantify privacy. Data mining techniques that automatically translate data into knowledge may yield confidential information about a specific user, putting the user's right to privacy at risk. The results clearly reveal that joint estimation is growing for the Gaussian multiplicative technique, demonstrating that the multiplicative technique achieves the privacy goal more effectively.

References

[1] Mall, P. K., Narayan, V., Srivastava, S., Sabarwal, M., Kumar, V., Awasthi, S., & Tyagi, L. (2023). Rank Based Two Stage Semi-Supervised Deep Learning Model for X-Ray Images Classification: AN APPROACH TOWARD TAGGING UNLABELED MEDICAL DATASET. Journal of Scientific & Industrial Research (JSIR), 82(08), 818-830.

[2] Mall, P. K., Singh, P. K., Srivastav, S., Narayan, V., Paprzycki, M., Jaworska, T., & Ganzha, M. (2023). A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. *Healthcare Analytics*, 100216.

[3] THANGA REVATHI S (2017)" DATA PRIVACY PRESERVATION USING DATA PERTURBATION TECHNIQUES" International Journal of Soft Computing and Artificial Intelligence, ISSN: 2321-404X, International Journal of Soft Computing and Artificial Intelligence, ISSN: 2321-404X,

[4] S.Srijayanthi (2017)" A Comprehensive Survey on Privacy Preserving Big Data Mining" International Journal of Computer Applications Technology and Research Volume 6–Issue 2, 79-86, 2017, ISSN:-2319–8656

[5] Awasthi, S., Srivastava, A. P., Srivastava, S., & Narayan, V. (2019, April). A Comparative Study of Various CAPTCHA Methods for Securing Web Pages. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)* (pp. 217-223). IEEE.

[6] Ravi, A.T. & Chitra, S.. (2015). Privacy Preserving Data Mining. Research Journal of Applied Sciences, Engineering and Technology. 9. 616-621. 10.19026/rjaset.9.1445.

[7] Sun, X., Xu, R., Wu, L. et al. A differentially private distributed data mining scheme with high efficiency for edge computing. J Cloud Comp 10, 7 (2021). https://doi.org/10.1186/s13677-020-00225-3

[8] Dedi Gunawan (2020)" Classification of Privacy Preserving Data Mining Algorithms: A Review" Jurnal Elektronika dan Telekomunikasi (JET), Vol. 20, No. 2, December 2020, pp. 36-46

[9] V. Jane Varamani Sulekh (2018)" NOISE BASED PRIVACY PRESERVING DATAMINING TECHNIQUES" International Journal of Computer Engineering and Applications, Volume XII, Issue IV, April 18, www.ijcea.com ISSN 2321-3469

[10] Mr. Swapnil Kadam (2015)" Preserving Data Mining through Data Perturbation" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 11

[11] Shan, Jinzhao & Lin, Ying & Zhu, Xiaoke. (2020). A New Range Noise Perturbation Method based on Privacy Preserving Data Mining. 131-136. 10.1109/ICAIIS49377.2020.9194850.

[12] Luo, Zhifeng & Wen, Congmin. (2014). A chaos-based multiplicative perturbation scheme for privacy preserving data mining. Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS. 941-944. 10.1109/ICSESS.2014.6933720.

[13] Narayan, V., Awasthi, S., Fatima, N., Faiz, M., Bordoloi, D., Sandhu, R., & Srivastava, S. (2023, May). Severity of Lumpy Disease detection based on Deep Learning Technique. In *2023 International Conference on Disruptive Technologies (ICDT)* (pp. 507-512). IEEE.

[14] Fares, Tamer & Khalil, Awad & Mohamed, Bensaada. (2008). Privacy Preservation in Data Mining using Additive Noise.. 21st International Conference on Computer Applications in Industry and Engineering, CAINE 2008. 50-55.

[15] Narayan, V., Faiz, M., Mall, P. K., & Srivastava, S. (2023). A Comprehensive Review of Various Approach for Medical Image Segmentation and Disease Prediction. *Wireless Personal Communications*, 1-30.

[16] Awasthi, S., Srivastava, P. K., Kumar, N., Ojha, R. P., Pandey, P. S., Singh, R., ... & Bakare, Y. B. (2023). An epidemic model for the investigation of multi-malware attack in wireless sensor network. IET Communications.

[17] Smit, S., Popova, E., Milić, M., Costa, A., & Martínez, L. Machine Learning-based Predictive Maintenance for Industrial Systems. Kuwait Journal of Machine Learning, 1(3). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/139

[18] Sanapala, A. ., Lakshmi, B. J. ., Kundra, K. S. R. ., & Madhuri, K. B. . (2023). Air Pollution Detection and Control System Using ML Techniques. International Journal on Recent and Innovation Trends in Computing and Communication, 11(4), 219–225. https://doi.org/10.17762/ijritcc.v11i4.6442

[19] Anupong, W., Yi-Chia, L., Jagdish, M., Kumar, R., Selvam, P. D., Saravanakumar, R., & Dhabliya, D. (2022). Hybrid distributed energy sources providing climate security to the agriculture environment and enhancing the yield. Sustainable Energy Technologies and Assessments, 52 doi:10.1016/j.seta.2022.102142