iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** VII **Month of publication:** July 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

ISSN: 2321-9653

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AND ENGINEERING TECHNOLOGY (IJRASET)

Performance of Multiplicative Data Perturbation for Privacy Preserving Data Mining

Bhupendra Kumar Pandya, Umesh Kumar Singh, Keerti Dixit, Kamal Bunkar

Institute of Computer Science

Vikram University, Ujjain

Abstract: Data mining is a well-known technique for automatically and intelligently extracting information or knowledge from a large amount of data, however, it can also disclosure sensitive information about individuals compromising the individual's right to privacy. Therefore, privacy preserving data mining has becoming an increasingly important field of research. Privacy preserving data mining is a novel research direction in data mining. In recent years, with the rapid development in Internet, data storage and data processing technologies, privacy preserving data mining has been drawn increasing attention. The goal of privacy preserving data mining is to develop data mining methods without increasing the risk of misuse of the data used to generate those methods. The topic of privacy preserving data mining has been extensively studied by the data mining community in recent years. A number of effective methods for privacy preserving data mining have been proposed. Most methods use some form of transformation on the original data in order to perform the privacy preservation

Key words: - privacy preservation, multiplicative data perturbation

INTRODUCTION

A Statistical database (SDB) is a database system that allows its users to retrieve aggregate statistics (e.g., sample mean and variance) for a subset of the entities represented in the database and prevents the collection of information on specific individuals. In the statistics community, there has been extensive research on the problem of securing SDBs against disclosure of confidential information. This is generally referred to as statistical disclosure control. Statistical disclosure control approaches suggested in the literature are classified into four general groups: conceptual, query restriction, output perturbation and data perturbation [1]. Conceptual approach provides a framework for better understanding and investigating the security problem of statistical database at the conceptual data model level. It does not provide a specific implementation procedure. The Query Restriction approach offers protection by either restricting the size of query set or controlling the overlap

among successive queries. The Output Perturbation approach perturbs the answer to user queries while leaving the data in the database unchanged. The Data Perturbation approach introduces noise into the database and transforms it into another version. This Chapter primarily focuses on the data perturbation approaches.

Adding random noise to the private database is one common data perturbation approach. In this case, a random noise term is generated from a prescribed distribution, and the perturbed value takes the form: $y_{ij} = x_{ij} + r_{ij}$, where x_{ij} is the i^{th} attribute of the j^{th} private data record, and r_{ij} is the corresponding random noise. In the statistics community, this approach was primarily used to provide summary statistical information (e.g., sum, mean, variance, etc.) without disclosing individual's confidential data. In the privacy preserving data mining area, this approach was considered [2,3] in for building decision tree classifiers from private data. Recently, many researchers have pointed out that additive noise can be easily filtered out in many cases that may lead to compromising the privacy [4,5]. Given the large body of existing signal-processing literature on filtered random additive noise, the utility of random additive noise for privacy-preserving data mining is not quite clear.

The Possible drawback of additive noise makes one wonder about the possibility of using multiplicative noise (i.e., $y_{ij} = x_{ij} * r_{ij}$) for protecting the privacy of the data. Two basic forms of multiplicative noise have been well studied in the statistics community [6]. One multiplies each data element by a random number that has a truncated Gaussian distribution with mean one and small variance. The other takes a logarithmic transformation of the data first, adds multivariate Gaussian noise, then takes the exponential function exp (.) of the noise-added data. As noted in the former perturbation scheme was once used by the Energy Information Administration in the U.S. Department of Energy to mask the heating and cooling degree days, denoted by x_{ij}. A random noise r_{ij} is generated from a Gaussian distribution with mean 1 and variance 0.0225. The random noise is further truncated such that the resulting number r_{ij} satisfies $0.01 \leq |r_{ij}-1| \leq 0.6$. The perturbed data $x_{ij}r_{ij}$ were released.

This research paper gives a brief review and Analysis of perturbation scheme I.

Perturbation Scheme: Let x_i be the ith attribute of a private database. Let x_i be the private value for the i^{th} attribute of the jth record in the database, i = 1, . . . , n, j=1, . . . , m. Let r_{ij} denote the random noise corresponding to xij. The perturbed data y_{ij} is

$$y_{ij} = x_{ij} * r_{ij} ,$$

where r_{ij} is independent and identically chosen from a Gaussian distribution with mean 1 (usually $\mu_i = 1$) and variance σi2. In other words, all r_{ij}'s for a given I follow the same distribution. In practice, the probability density of noise r (ignoring the subscript) is usually doubly truncated as follows:

$$F(r) = \frac{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(r-\mu)^2\right)}{\frac{1}{\sqrt{2\pi}\sigma}\int_A^B \exp\left(-\frac{1}{2\sigma^2}(r-\mu)^2\right)dr}$$

$$for\ A < r < B$$

$$= \frac{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(r-\mu)^2\right)}{\Phi\left(\frac{B-\mu}{\sigma}\right)-\Phi\left(\frac{A-\mu}{\sigma}\right)}$$

Where A and B are the lower and upper truncation bounds and Ø(A) stands for the cumulative probability up to A. The above equation can be further simplified as

$$KZ\left(\frac{r-\mu}{\sigma}\right),$$

Where K = $\frac{1}{\Phi\left(\frac{B-\mu}{\sigma}\right)-\Phi\left(\frac{A-\mu}{\sigma}\right)}$, and Z(x) = $\frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{1}{2}x^2\right)$.

Statistical Properties of the Perturbed Data:

It has been proved [6] that the mean and variance of the original data attributes can be estimated from the mean and variance of the perturbed data.

Mean of x_i

$E(x_i) = \frac{E(y_i)}{\mu_i + K\left[Z\left(\frac{A-\mu_i}{\sigma_i}\right) - Z\left(\frac{B-\mu_i}{\sigma_i}\right)\right]}$.

Because the data owner will release μI, σ_i, A and B, the data receiver can compute the expected value of x_i.

Variance of x_i:

$Var(x_i) = E(x_i^2) - (E(x_i))^2$,

Where $E(x_i)$ can be easily calculated from the above equation and $(E(x_i))^2$ can be computed from the follow equation:

$$Var[y_i] = E(x_i^2)E(r_i^2) - (E(x_i)E(r_i))^2$$

$$= E(x_i^2)\left\{\sigma_i^2 + \mu_i^2 + \sigma_i^2 K\left[\frac{A-\mu_i}{\sigma_i} Z\left(\frac{A-\mu_i}{\sigma_i}\right) - \frac{B-\mu_i}{\sigma_i} Z\left(\frac{B-\mu_i}{\sigma_i}\right)\right] + 2\sigma_i\mu_i K\left[Z\left(\frac{A-\mu_i}{\sigma_i}\right) - Z\left(\frac{B-\mu_i}{\sigma_i}\right)\right]\right\} -$$

$$(E(x_i))^2\left\{\mu_i^2 + \sigma_i^2 K^2\left[Z\left(\frac{A-\mu_i}{\sigma_i}\right) - Z\left(\frac{B-\mu_i}{\sigma_i}\right)\right]^2 + 2\sigma_i\mu_i K\left[Z\left(\frac{A-\mu_i}{\sigma_i}\right) - Z\left(\frac{B-\mu_i}{\sigma_i}\right)\right]\right\}.$$

Although the original attribute's mean and variance can be estimated from the perturbed data, the inner product and Euclidean distance among the data records are not necessarily preserved after perturbation. The following Results depict this situation.

Analysis of perturbation schemes with experimental result using matlab

Data to be used:-

In this study we have Students result database of Vikram University, Ujjain. We have randomly selected 7 rows of the data with only 7 attributes(Marks of Foundation, Marks of Mathematics, Marks of Physics, Marks of Computer Science, Marks of Physics Practical, Marks of Computer Science Practical and Marks of Job Oriented Project).

Perturbation Scheme I:-

In this scheme, a random number is generated from a normal distribution with mean 1 and variance .0225. The random number is truncated such that the resulting number e_j satisfies $.01 \le |e_j - 1| \le .6$.

Table 1: Original Data

Foundation	Maths	Physics	Com. Sc.	Phy. Prac.	Com. Sc. Prac	Project
56	73	38	42	39	42	42
49	47	22	36	37	42	39
55	57	40	33	39	42	40
60	50	34	53	37	41	38
50	37	11	25	38	41	38
48	61	31	36	40	43	41
61	64	40	40	39	42	39

ISSN: 2321-9653

Table 2: Random noise matrix.

1.012	1.007	1.016	0.972	1.006	1.032	1.007
1.041	1.080	0.995	1.016	0.982	1.007	1.007
0.949	1.062	0.997	1.036	1.019	0.983	0.980
1.019	0.969	1.033	1.011	0.974	1.030	0.999
1.007	1.068	1.031	1.023	0.975	0.961	0.996
0.970	1.016	1.031	1.016	0.981	0.99	1.014
0.990	0.998	1.015	0.993	0.933	0.994	1.024

Table 3: Dataset after perturbation.

56.677	73.562	38.611	40.858	39.257	43.359	42.301
51.021	50.784	21.898	36.580	36.344	42.307	39.274
52.204	60.551	39.888	34.210	39.779	41.286	39.221
61.163	48.481	35.139	53.583	36.045	42.264	37.974
50.358	39.526	11.348	25.582	37.086	39.421	37.859
46.587	61.995	31.988	36.588	39.271	42.901	41.579
60.404	63.909	40.604	39.726	36.416	41.771	39.959

Mean of original and perturbed data

Original data	54.14	55.57	30.85	37.85	38.42	41.85	39.57
Perturbed data	54.05	56.97	31.35	38.16	37.74	41.90	39.7

As seen in the table, the estimates of the means from the perturbed data are all close to those from original data.

Euclidean distance of original data

32.093	18.601	26.514	48.662	17.204	11.090	21.771
23.769	18.601	17.088	27.874	22.781	36.551	12.569
11.618	39.786	24.186	20.199	33.436	43.806	16.763

Euclidean distance of perturbed data

29.448	15.785	29.213	47.063	17.275	11.332	20.9493
23.965	19.244	16.167	24.917	25.303	36.701	10.4932
11.024	39.448	26.788	21.558	33.120	41.973	17.0505

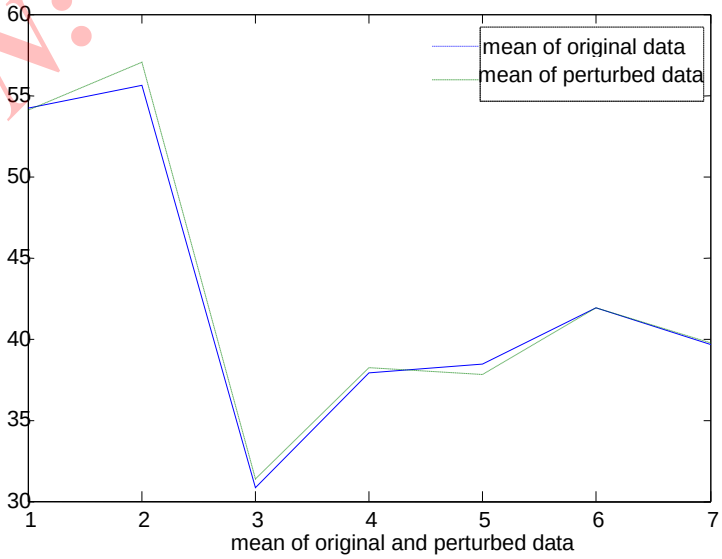


Figure 1

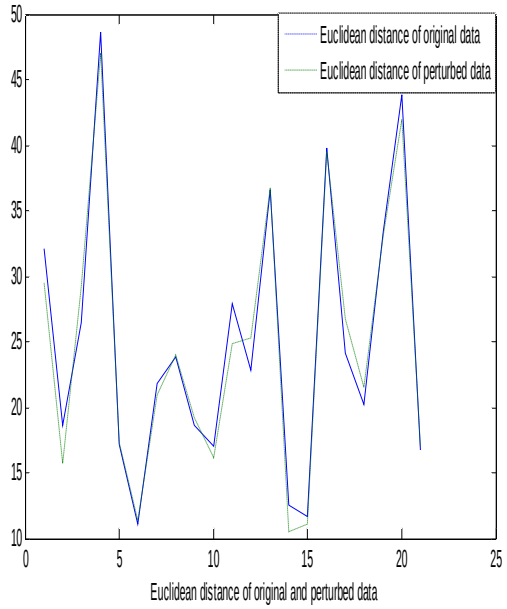


Figure 2

We have taken the original data which is result set of students. With this data we have generated a noise matrix with the help of Gaussian distribution with mean 1 and smaller variance. To generate the random noise we have used the normrnd() function of Matlab and this resultant noise data set is multiplied with the original data set to form the perturb data. We have evaluated mean of original and perturbed data with mean () fuction of Matlab. As seen in the graph 1, the estimates of the means from the perturbed data are all close to those from original data. We use pdist() function of Matlab to compute the Euclidian distance of original data set and the perturbed data.

We have plotted the graph 2 which shows the comparison between Euclidean Distances of original data and perturbed data after applying Perturbation Scheme I.

The above graph shows that although the original attribute mean can be estimated from the perturbed data, but the Euclidean Distance among the data records are not necessarily preserved after perturbation.

CONCLUSION

This research paper reviews first traditional multiplicative data perturbation techniques that have been studied in statistics community. The effectiveness of multiplicative data perturbation techniques for privacy preserving data mining have been analyzed and also the security of multiplicative data perturbation schemes after applying logarithmic transformation have been examined. These perturbations are primarily used to mask the private data while allowing summary statistics (e.g., sum, mean, variance and covariance) of the original data to be estimated.

On the surface, multiplicative perturbation seems to change the data more than additive perturbation. However, by taking logarithms on the perturbed data, the multiplicative perturbation turns into an additive perturbation.

For perturbation scheme I, the logarithmic transformation of yij gives us ln xij + ln rij , where the noise term ln rij is chosen independent and identically from some distribution.

The objective of these perturbation schemes is to mask the private data while allowing summary statistics to be estimated. However, problems in data mining are somewhat different. Data mining techniques, such as clustering, classification, prediction and association rule mining, are essentially relying on more sophisticated relationships among data records or data attributes,

but not simple summary statistics. The traditional multiplicative perturbations distort each data element independently, therefore Euclidean distance and inner product among data records are usually not preserved, and the perturbed data cannot be used for many data mining applications.

These perturbation schemes are equivalent to additive perturbation after the logarithmic transformation. Due to the large volume of research in deriving private information from the additive noise perturbed data, the security of these perturbation schemes is questionable.

REFERENCES

[1] N. R. Adam and J. C. Worthmann, "Security-control methods for statistical databases: a comparative study," ACM Computing Surveys (CSUR), vol. 21, 145-146 no. 4, pp. 515–556, 1989.

[2] R. Agrawal and R. Srikant, "Privacy preserving data mining," in Proceedings of the ACM SIGMOD Conference on Management of Data, Dallas, TX, May 2000, pp. 439–450.

[3] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in Proceedings of the twentieth ACM SIGMOD- IGACT-SIGART symposium on Principles of Database Systems, Santa Barbara, CA, 2001, pp. 247–255.

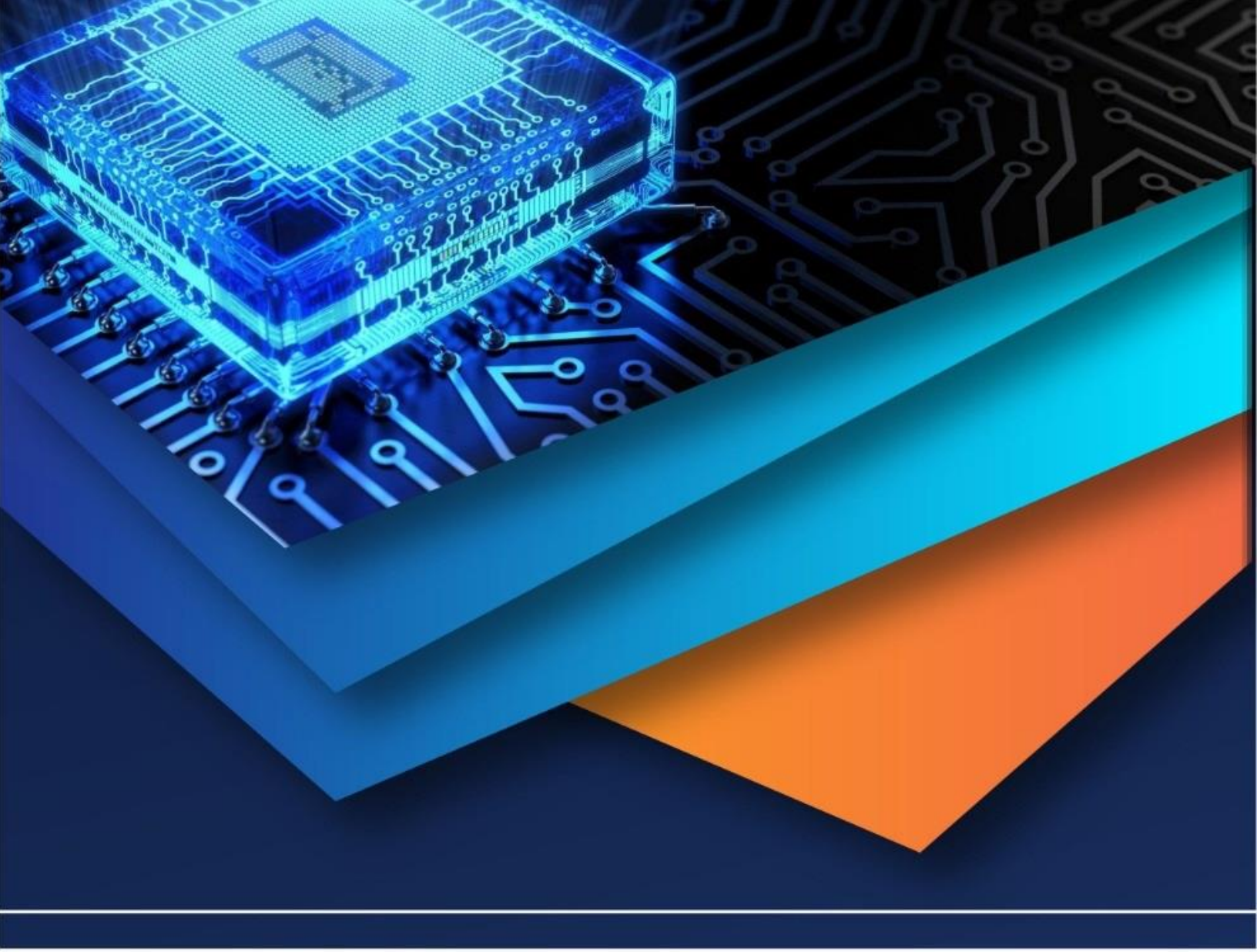
[4] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in Proceedings of the IEEE International Conference on Data Mining, Melbourne, FL, November 2003.

[5] S. Guo and X. Wu, "On the use of spectral filtering for privacy preserving data mining," in Proceedings of the 21st ACM Symposium on Applied Computing, Dijon, France, April 2006, pp. 622–626.

[6] J. J. Kim and W. E. Winkler, "Multiplicative noise for masking continuous data," Statistical Research Division, U.S. Bureau of the Census, Washington D.C., Tech. Rep. Statistics #2003-01, April 2003.

[7] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data," in Proceedings of the 2005 ACM SIGMOD Conference, Baltimroe, MD, June 2005, pp. 37–48.

[8] K. Muralidhar, R. Parsa, and R. Sarathy, "A general additive data perturbation
method for database security," Management Science, vol. 45, no. 10, pp. 1399–1415, 1999.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY