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)

An overview of Multiplicative data perturbation for privacy preserving Data mining

Keerti Dixit, Bhupendra Pandya

Institute of Computer Science

Vikram University

Ujjain

Abstract—Privacy is an important issue when one wants to make use of data that involves individuals' sensitive information. Research on protecting the privacy of individuals and the confidentiality of data has received contributions from many fields, including computer science, statistics, economics, and social science. In this paper, we survey research work in privacy-preserving data Mining. This is an area that attempts to answer the problem of how an organization, such as a hospital, government agency, or insurance company, can release data to the public without violating the confidentiality of personal information. We focus on privacy criteria that provide formal safety guarantees, present algorithms that sanitize data to make it safe for release while preserving useful information, and discuss ways of analyzing the sanitized data. Many challenges still remain. This overview provides a summary of the current and traditional multiplicative data perturbation techniques for privacy preserving Data Mining.

Key words:- privacy preservation, multiplicative data perturbation

INTRODUCTION

Privacy preserving data mining in a broad sense has been an area of research since 1991 [1] both in the public and private [2] sector and has also been discussed at numerous workshops and international conferences [3]. Recent interest in the collection and monitoring of data using data mining technology for the purpose of security and business-related applications has raised serious concerns about privacy issues. Sometimes, individuals or organizational entities may not be willing to disclose the sensitive raw data; sometimes the knowledge and/or patterns detected by a data mining system may be used in a counter-productive manner that violates the privacy policy. The main objective of privacy preserving data mining is to develop algorithms for modifying the original data or modifying the computation protocols in some way, so that during and after the mining process, the private data and private knowledge remain private while other underlying data patterns or models can still be effectively identified.

LITERATURE ON PRIVACY PRESERVING DATA MINING

DATA HIDING:

The main objective of data hiding is to transform the data so that the private data remains private during and/or after data mining operations.

Data Perturbation:

Data perturbation techniques can be grouped into two main categories, which we call the value distortion technique and probability distribution technique. The value distortion technique perturbs data elements or attributes directly by either some other randomization procedures. On the other hand, the probability distribution technique considers the private database to be a sample from a given population that has a given probability distribution. In this case, the perturbation replaces the original database by another sample from the same [estimated] distribution or by the distribution itself.

Note that there has been expensive research in the area of statistical databases [SDB] on how to provide summary statistical information without disclosing individual's confidential data. The privacy issues arise when the summary statistics are derived from data of very few individuals. A popular disclosure control method is data perturbation, which alters individual data in a way such that the summary statistics remain approximately the same. However, problems in data mining become somewhat different from those in SDBs. 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 just simple summary statistics. This research work specifically focuses on data perturbation for privacy preserving data mining. In the following, we will primarily discuss different perturbation techniques in the data mining area. Some important perturbation approaches in SDBs are also covered for the sake of completeness.

ADDITIVE PERTURBATION

The work in proposed an additive data perturbation technique for building decision tree classifiers. In this technique, each client has a numerical attribute xi and the server [or data miner] wants to learn the distribution of these attributes o build a classification model. The clients randomize their attributes xii by adding random noise ri drawn independently from a known distribution such as a uniform distribution or Gaussian distribution. The server [or data miner] collects the values of xii + ri and reconstructs xi's distribution using a version of the Expectation-Maximization [EM] algorithm.

This algorithm probably converges to the maximum

likelihood estimate of the desire original distribution.

Kargupta et al [4,5,6], later questioned the use of random additive noise and pointed out that additive noise can be easily filtered out in many cases that will possibly compromise the privacy. To be more specific, they proposed a random matrix based Spectral Filtering [SF] technique to recover the original data from the perturbed data. Their empirical results have shown that the recovered data can be reasonably close to the original data. However, two important questions remain to be answered: (1) What are the theoretical lower bound and upper bound of the reconstruction error; and (2) What are the key factors that influence the accuracy of the data reconstruction. Guo and Wu [7] investigated the Spectral Filtering technique and derived an upper bound for the Frobenius norm of the reconstruction error using matrix perturbation theory. They also proposed a Singular Value Decomposition (SVD)-based reconstruction method and derive a lower bound for reconstruction error. They then proved the equivalence between the SF and SVD approach, and as a result, the lower bound of SVD approach can also be considered as the lower bound of the SF approach. Huang et al. [8] pointed out that the key factor that decides the accuracy of data reconstruction is the correlation among the data attributes. Their results have shown that when the correlations are high, the original data can be reconstructed more accurately, that is, more private information can be disclosed. They further proposed two data reconstruction methods based on data correlations: one used the principal Component Analysis [PCA], and the other used the Bayes Estimate [BE] technique, which in essence processing literature on filtering random additive noise, the utility of random additive noise for privacy preserving data.

MULTIPLICATIVE DATA PERTURBATION

Data perturbation refers to a data transformation process typically performed by the data owners before publishing their data. The goal of performing such data transformation is two-fold. On one hand, the data owners want to change the data in a certain way in order to disguise the sensitive information contained in the published datasets, and on the other hand, the data owners want the transformation to best preserve those domain-specific data properties that are critical for building meaningful data mining models, thus maintaining mining task specific data utility of the published datasets. Two basic forms of multiplicative noise have been well studied in the statistics community [9]. 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 xij. A random noise rij is generated from a Gaussian distribution with mean 1 and variance 0.0225. The random noise is further truncated such that the resulting number rij satisfies $0.01 \leq |rij-1| \leq 0.6$. The perturbed data xijrij were released.

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.

In summary these multiplicative perturbations have the following advantages and disadvantages.

The multiplicative perturbation is relative, that is, large values in the original data are perturbed more than smaller value.

In practice, the first perturbation scheme is good if the data disseminator only wants to make minor changes to the original data; the second scheme assures higher security than the first one but maintains the data utility in the log-scale.

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.

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 application.

In this paper, we will discuss multiplicative data perturbations. This category includes three types of particular perturbation techniques: Rotation Perturbation, Projection Perturbation, and Geometric Perturbation. Comparing to other multi-dimensional data perturbation methods, these perturbations exhibit unique properties for privacy preserving data classification and data clustering. They all preserve (or approximately preserve) distance or inner product, which are important to many classification and clustering models. As a result, the classification and clustering mining models based on the perturbed data through multiplicative data perturbation show similar accuracy to those based on the original data. The main challenge for multiplicative data perturbations thus is how to maximize the desired data privacy. In contrast many other data perturbation techniques focus on seeking for the better trade-off between the level of data utility and accuracy preserved and the level of data privacy guaranteed.

Definition of Multiplicative Perturbation

We will first describe the notations used in this chapter, and then describe three categories of multiplicative perturbations and their basic characteristics.

Notations

In privacy-preserving data mining, either a portion of or the entire data set will be perturbed and then exported. For example, in classification, the training data is exported and the testing data might be exported, too, while in clustering, the entire data for clustering is exported. Suppose that X is the exported dataset consisting of N data rows (records) and d columns (attributes, or dimensions). For presentation convenience, we use $X_{d\times N}$, $X = [x1 \ldots x_N]$, to denote the dataset, where a column x_i $(1 \leq i \leq N)$ is a data tuple, representing a vector in the real space R^d. In classification, each of such data tuples x_i also belongs to a predefined class, which is indicated by the class label attribute y_i. The class label can be nominal (or continuous for regression), and is public, i.e., privacy-insensitive. For clear presentation, we can also consider X is a sample dataset from the d-dimension random vector $X = [X_1, X_2, \ldots, X_d]^T$. As a convention, we use bold lower case to represent vectors, bold upper case to represent random variables, and upper case to represent matrices or datasets.

ROTATION PERTURBATION

This category does not cover traditional “rotations” only, but literally, it includes all orthonormal perturbations. A rotation perturbation is defined as following G(X):

$G(X) = RX$

The matrix $R_{d\times d}$ is an orthonormal matrix [10], which has following properties. Let R^T represent the transpose of R, r_{ij} represent the (i, j) element of R, and I be the identity matrix. The rows and columns of R are orthonormal, i.e., for any column j, $\sum_{i=1}^{d} r^2_{ij} = 1$, and for any two columns j, and k, $j \neq k$, $\sum_{i=1}^{d} r_{ij} r_{ik} = 0$. A similar property is held for rows. This definition infers that

ISSN: 2321-9653

$R^TR = RR^T = I$

It also implies that by changing the order of the rows or columns of an orthogonal matrix, the resulting matrix is still orthogonal. A random orthonormal matrix can be efficiently generated following the Haar distribution [11]. A key feature of rotation transformation is that it preserve the Euclidean distance of multi-dimensional points during the transformation. Let x^T represent the transpose of vector x, and $||x|| = x^Tx$ represent the length of a vector x. By the definition of rotation matrix, we have

$$||Rx|| = ||x||$$

Similarly, inner product is also invariant to rotation. Let $<x,y> = x^Ty$ represent the inner product of x and y. We have $<Rx,Ry> = x^TR^TRy = <x,y>$

In general, rotation also preserves the geometric shapes such as hyperplane and hyper curved surface in the multidimensional space [12]. We observed that since many classifiers look for geometric decision boundary, such as hyperplane and hyper surface, rotation transformation will preserve the most critical information for many classification models.

There are two ways to apply rotation perturbation. We can either apply it to the whole dataset X [13], or group columns to pairs and apply different rotation perturbations to different pairs of columns [14].

PROJECTION PERTURBATION

Projection perturbation refers to the technique of projecting a set of data points from a high-dimensional space to a randomly chosen lower-dimensional subspace. Let $P_{k\times d}$ be a projection matrix.

$$G(X) = PX$$

Why can it also be used for perturbation? The rationale is based on the Johnson-Lindenstrauss Lemma [15].

Theorem:

For any $0 < \varepsilon < 1$ and any integer n, let k be a positive integer such that $k \geq \frac{4\,In\,n}{\varepsilon^2/2-\varepsilon^3/3}$.

Then, for any set S of n data points in d dimensional space R^d, there is a map $f : R^d \rightarrow R^k$ such that, for all $x \in S$,

$$(1-\varepsilon)\,||x-x||^2 \leq ||f(x)-f(x)||^2 \leq (1+\varepsilon)||x-x||^2$$

where $||\cdot||$ denotes the vector 2-norm.

This lemma shows that any set of n points in d-dimensional Euclidean space could be embedded into a o (logn/ ε^2) - dimensional space, such that the pair-wise distance of any two points are maintained with small error. With large n (large dataset) and small ε (high accuracy in distance preservation), the ideal dimensionality might be large and may not be practical for the perturbation purpose. Furthermore, although this lemma implies that we can always find one good projection that approximately preserves distances for a particular dataset, the geometric decision boundary might still be distorted and thus the model accuracy is reduced. Due to the different distributions of dataset and particular properties of data mining models, it is challenging to develop an algorithm that can find random projections that preserves model accuracy well for any given dataset. In paper [16] a method is used to generate random projection matrix. The process can be briefly described as follows. Let P be the projection matrix. Each entry ri,j of P is independent and identically chosen from some distribution with mean zero and variance σ^2. A row-wise projection is defined as

$$G(X) = G(X) = \frac{1}{\sqrt{k\sigma}}\,PX$$

Let x and y be two points in the original space, and u and v be their projections.

The statistical properties of inner product under projection perturbation can be shown as follows.

$$E[u^tv - x^ty] = 0$$

and

$$Var[u^tv - x^ty] = \frac{1}{k}\left(\sum_i x_i^2\ \sum_i y_i^2 + (\sum_i x_i\ y_i)^2\right)$$

Since x and y are not normalized by rows, but by columns in practice, with large dimensionality d and relatively small k, the variance is substantial. Similarly, the conclusion can be extended to the distance relationship. Therefore, projection perturbation does not strictly guarantee the preservation of

ISSN: 2321-9653

distance/ inner product as rotation or geometric perturbation does, which may ignificantly downgrade the model accuracy.

Sketch-based Approach

Sketch-based approach is primarily proposed to perturb high-dimensional sparse data [17], such as the datasets in text mining and market basket mining.

A sketch of the original record x = (x_1, . . . , x_d) is defined by a r dimensional vector s = (s_1, . . . , s_r), r « d, where

$$S_j = \sum_{i=1}^{d} x_i r_{ij}$$

The random variable r_{ij} is drawn from {-1,+1} with a mean of 0, and is generated from a pseudo- random number generator [18], which produces 4-wise independent values for the variable r_{ij} . Note that the sketch based approach defers from projection perturbation with the following two features. First, the number of components for each sketch, i.e., r_{ij}, can vary across different records, and is carefully controlled so as to provide a uniform measure of privacy guarantee across different records. Second, for each record, r_{ij} is different − there is no fixed projection matrix across records.

The sketch based approach has a few statistical properties that enable approximate calculation of dot product of the original data records with their sketches. Let s and t with the same number of components r, be the sketches of the original records x and y, respectively. The expected dot product x and y is given by the following.

$$E[\langle x, y\rangle] = \langle s, t\rangle / r$$

and the variance of the above estimation is determined by the few non-zeros entries in the sparse original vectors

$$Var(\langle s, t\rangle / r) = \sum_{i=1}^{d} \sum_{l=1}^{d} x_i^2 y_i^2 - (\sum_{i=1}^{d} x_i y_i)^2 / r$$

On the other side, the original value x_k in the vector x can also be estimated by privacy attackers, the precision of which is determined by its variance

$(\sum_{i=1}^{d} x_i^2 - x_k^2)/r$, k = 1. . . d. The larger the variance is, the better the original value is protected. Therefore, by decreasing r the level of privacy guarantee is possibly increased. However, the precision of dot-product estimation (Eq. 7.1)

GEOMETRIC PERTURBATION

Geometric perturbation is an enhancement to rotation perturbation by incorporating additional components such as random translation perturbation and noise addition to the basic form of multiplicative perturbation Y = R×X. We show that by adding random translation perturbation and noise addition, Geometric perturbation exhibits more robustness in countering attacks than simple rotation based perturbation [19]. Let $t_{d\times 1}$ represent a random vector. We define a translation matrix as follows.

Definition: Ψ is a translation matrix if Ψ =[t, t, . . .,t]$_{d\times n}$, i.e., $\Psi_{d\times n} = t_{d\times 1} 1^T_{N*1}$ .where $1_{N\times 1}$ is the vector of N '1's. Let Δd×N be a random noise matrix, where each element is Independently and Identically Distributed (iid) variable ε_{ij} , e.g., a Gaussian noise N(0, σ^2). The definition of geometric perturbation is given by a function G(X),

$$G(X) = RX + \Psi + \Delta$$

Clearly, translation perturbation does not change distance, as for any pair of points x and y, $\|(x+t)-(y+t)\| = \|x-y\|$. Comparing with rotation perturbation, it protects the rotation center from attacks and adds additional difficulty to ICA-based attacks. However, translation perturbation does not preserve inner product.

In [9], it shows that by adding an appropriate level of noise", one can effectively prevent knowledgeable attackers from distance-based data reconstruction, since noise addition perturbs distances, which protects perturbation from distance-inference attacks. For example, the experiments in [19] shows that a Gaussian noise N(0, σ^2) is effective to counter the distance-inference attacks. Although noise addition prevents from fully preserving distance information,

a low intensity noise will not change class boundary or cluster membership much. In addition, the no se component is optional − if the data owner makes sure that the original data records are secure and no people except the data owne knows any record in the original dataset, the noise component can be removed from geometric perturbation.

ISSN: 2321-9653

CONCLUSION

We have reviewed the multiplicative perturbation method as an alternative method to privacy preserving data mining. The design of this category of perturbation algorithms is based on an important principle: by developing perturbation algorithms that can always preserve the mining task and model specific data utility, one can focus on finding a perturbation that can provide higher level of privacy guarantee. We described three representative multiplicative perturbation methods − rotation perturbation, projection perturbation, and geometric perturbation. All aim at preserving the distance relationship in the original data, thus achieving good data utility for a set of classification and clustering models. Another important advantage of using these multiplicative perturbation methods is the fact that we are not required to re-design the existing data mining algorithms in order to perform data mining over the perturbed data. One observation is that both the above mentioned techniques require much more samples (or background knowledge) to work effectively in the high dimensional case. Thus, random projection techniques should generally be used for the case of high dimensional data, and only a smaller number of projections should be retained in order to preserve privacy. Thus, as with the additive perturbation technique, the multiplicative technique is not completely secure from attacks. A key research direction is to use a combination of additive and multiplicative perturbation techniques in order to construct more robust privacy preservation.

REFERENCES

[1]http://www.cs.ualberta.ca/%7Eoliveira/psdm/pub_by_year.html or k. Liu at

http://www.cs.umbc.edu/~kunliul/research/privacy_review.html

[2] For example research carried out by IBM, see

http://www.almaden.ibm.com/software/disciplines/iis/

[3] See for example overview up to 2004 at

http://www.cs.ualberta.ca/%7Eoliveira/psdm/workshop.html

[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, November 2003.

[5] K.Liu, H.Kargupta, and J.Ryan, "Random projection-based multiplicative data perturbation for privacy preserving istributed data mining,"IEEE Transaction on knowledge and Data Engineering [TKDE], vol.18, no.1, January 2006.

[6] K.liu, C. Giannella, and H. Kargupta,"An attacker's view of distance preserving maps For privacy preserving distributed data mining," in proceeding of the10th European conference on principles and practice of knowledge Discovery in Databases [PKDD'06], Berlin, Germany, September 2006.

[7] 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.

[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.

[9] 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. 1963, ch. XII, pp. 213–217.

[10] SADUN, L. Applied Linear Algebra: the Decoupling Principle. Prentice Hall, 2001.

[11] STEWART, G. The efficient generation of random orthogonal matrices with an application to condition estimation. SIAM Journal on Numerical Analysis 17 (1980).

[12] CHEN, K., AND LIU, L. A random geometric perturbation approach to privacy-preserving data classification. Technical Report, College of Computing, Georgia Tech (2005).

[13] CHEN, K., AND LIU, L. A random rotation perturbation approach to privacy preserving data classification. Proc. of Intl. Conf. on Data Mining*(ICDM)* (2005).

[14] OLIVEIRA, S. R. M., AND ZA¨IANE, O. R. Privacy preservation when sharing data for clustering. In Proceedings of the International Workshop on Secure Data Management in a Connected World (Toronto, Canada, August 2004), pp. 67–82.

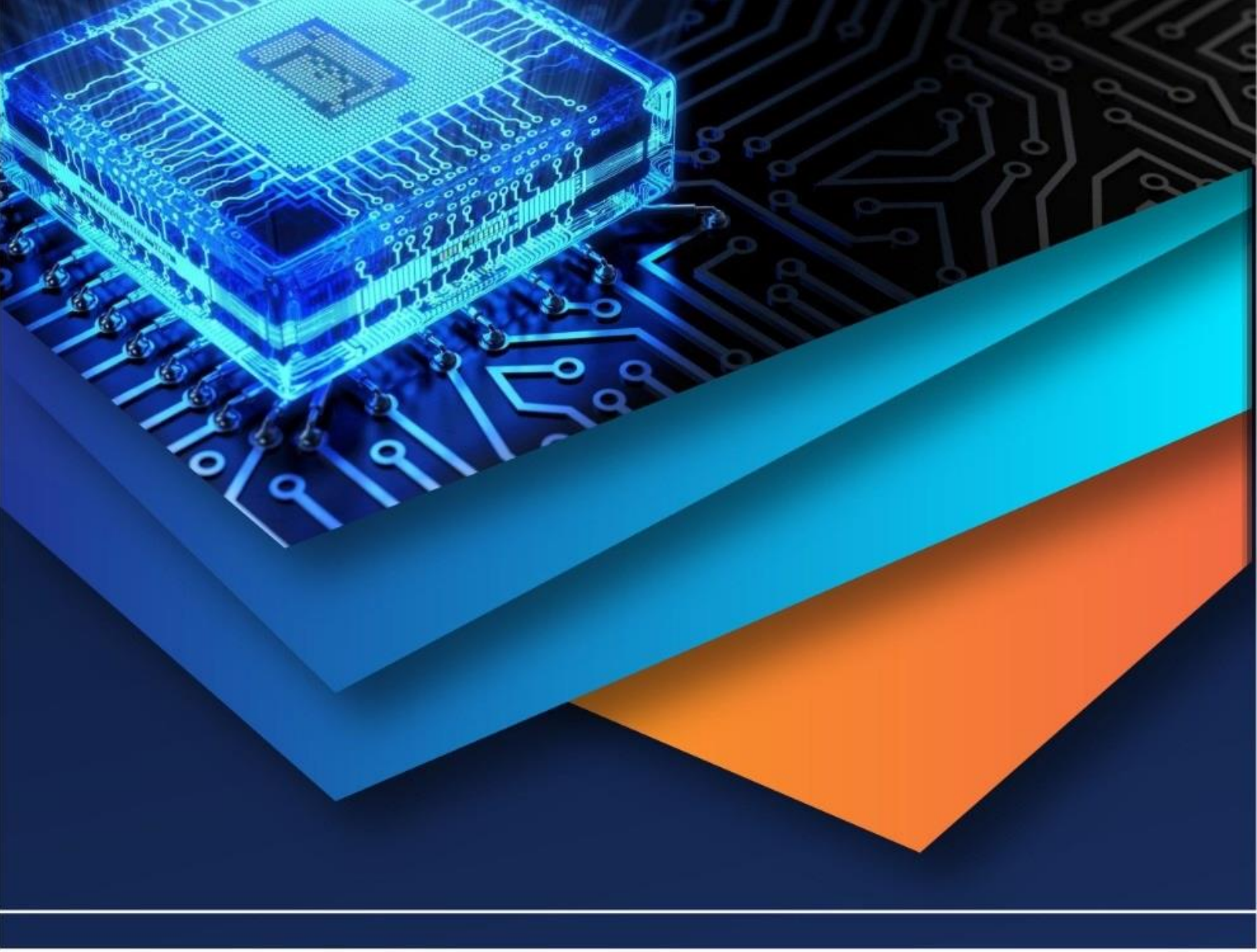
[15] JOHNSON, W. B., AND LINDENSTRAUSS, J. Extensions of lipshitz mapping into hilbert space. Contemporary Mathematics 26 (1984).

[16] LIU, K., KARGUPTA, H., AND RYAN, J. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Transactions on Knowledge and Data Engineering (TKDE) 18, 1 (January 2006), 92–106

[17] AGGARWAL, C. C., AND YU, P. S. On privacy-preservation of text and sparse binary data with sketches. *SIAM Data Mining Conference* (2007).

[18] ALON, N., MATIAS, Y., AND SZEGEDY, M. The space complexity of approximating the frequency moments. *Proc. of ACM PODS Conference* (1996).

[19] CHEN, K., AND LIU, L. Towards attack-resilient geometric data perturbation. SIAM Data Mining Conference (2007).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY