Pertinent Exploration of Privacy Preserving Perturbation Methods

Vijaya Pinjarkar, Amit Jain, Anand Bhaskar, Prateek Srivastava

Abstract*: Digital era generates a huge amount of data in many sectors like education, medical, banking, business, marketing, etc. which can be used for research motive, analysis, prediction of trends, statistics, etc. Data mining techniques are useful in finding patterns, trends, and knowledge from such huge data. The data holders are not ready to share data because there are chances of privacy leakage. Sharing of such data immensely helps researchers to obtain knowledge from it, especially medical data. Privacy preserving data mining is one way where researchers will get mine data for gaining knowledge without breaching the privacy. In the medical sector there is a branch called the mental health section, where high confidentiality of data is maintained and is needed. Owners are not ready to share data for research motives. Mental health is nowadays a topic that is most frequently discussed when it comes to research. PPDM allows sharing data with the researcher, where the privacy of data is maintained by using perturbation techniques giving relief to doctors (owner of data). The current paper experiments and analyses different perturbation methods to preserve privacy in data mining

Keywords—Mental health, Perturbation, Privacy preserving data mining

I. INTRODUCTION

Data Mining empowers an organization to extract meaningful patterns, knowledge, analysis, the prediction from an enormous pool of data. Current years have seen tremendous development in the methods of data collection. Digital era generates a huge amount of data in many sectors like education, medical, banking, business, and marketing, etc. which can be used for research ground, analysis and prediction of trends, statistics, etc. Different data mining techniques are useful in finding patterns, trends, and knowledge from such huge data. But the data owners are not ready to share their data because there are changes of privacy leakage. This data contains attribute information like age, account no, disease type, family history, etc.

Revised Manuscript Received on February 22, 2020.

***** Correspondence Author

Vijaya Pinjarkar, Ph.D. Scholar, department of Computer Science and Engineering, Sir Padampat Singhania University, Udaipur, Rajasthan, Assistant Professor K.J. Somaiya Institute of Engineering & Information Technology, Mumbai, India.
Email: vijaya.pinjarkar@spsu.ac.in

Amit Jain, Assistant Professor, Computer Science and Engineering, Sir Padampat Singhania University, Udaipur, Rajasthan, India.
Email: amit.jain@spsu.ac.in

Anand Bhaskar, Head & Assistant Professor, Electronics & Communication Engineering, Sir Padampat Singhania University, Udaipur, Rajasthan, India. Email: anand.bhaskar@spsu.ac.in

Prateek Srivastava, Assistant Professor, Computer Science and Engineering, Sir Padampat Singhania University, Udaipur, Rajasthan, India.
Email: prateek.srivastava@spsu.ac.in

Sharing of such data immensely helps researchers to obtain knowledge from it, especially medical data. Privacy preserving data mining is a way where researchers will get mine data for obtaining knowledge without breaching privacy. In the medical sector there is a branch called the mental health section, where high confidentiality of data is maintained and is needed. The Doctors (Owners) are not ready to share data for research motives. Mental health is nowadays a topic that is most frequently discussed when it comes to research. The increasing use of technology leads to a lifestyle of less physical activities. Also, the constant pressure in the industry will makes the employee vulnerable to mental disorder. This vulnerability arises due to peer pressure, anxiety attack, depression and many more reasons.

PPDM allows sharing data with the researcher, where the privacy of data is maintained giving relief to the data owner (the doctor).

Authors in [1] and [2] have introduced PPDM and have given different methods for implementing PPDM. It refers to the study of various ways to anonymize data for hiding an individual's identity by preventing disclosure of sensitive attributes.

Major healthcare organizations store patient's data in digital form using the electronic healthcare record (EHR) system [3]. The data mining techniques can be used to extract useful patterns/knowledge from the collected data.

Privacy becomes an important concern while publishing the data. Medical research majorly focuses on distributed data mining and privacy preserving. There are two types of partitions in distributed privacy preserving data mining a) Horizontally partition data and b) vertically partition data.

In horizontally partition data, all the participants have the same schema and the different number of transactions while in vertically partition database, all the participants have different schema and an equal number of transactions. PPDM understands different perturbation methods for data publishing. The greatest challenge for perturbation is balancing privacy protection and data quality.

Section I of this paper presents an introduction about PPDM and different types of PPDM. Section II discusses the related work and existing PPDM methods. Section III is about the proposed method and implementation of existing and proposed PPDM perturbation method. Section IV is concerned to the conclusion and future work.

II. RELATED WORK

Figure 1. indicates the classification of Privacy Preserving Data Mining algorithm.

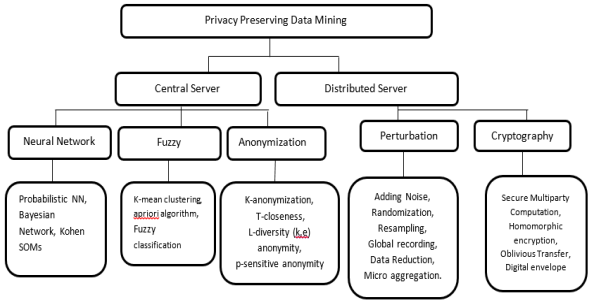

Figure1: Classification of Privacy Preserving Data Mining algorithm.

Various authors put their work for different methods like k-anonymity; and its disadvantages are discussed in [4]. The drawback of k-anonymity is that it does not take into consideration the distribution of the sensitive values which can lead to privacy breaches when the sensitive values are distributed in a skewed away, so [5] yakked about l-diversity method expands the k-anonymity model. The l-diversity problem of skewed sensitive values distribution by requiring that the distribution of the sensitive values in each equivalence class is solved in [6], which discussed the t-closeness method. Randomization is used in [7] where noise is added in the data. Resampling and Reduction are introduced and discussed in [2]. Generalization or Suppression of an attribute in the original data set is called Global recording described by [8]. For the protection of individual records, Micro-aggregation [4] is used where one can mine, distribute and publish individual records without providing any personal information that can be associated with specific individuals. In Randomization, noise is added in data. it can be done in two ways like additive randomization and multiplicative randomization. The following table shows the comparison of additive randomization and multiplicative randomization methods.

Table I: Types of Randomization Method

Randomization method	Description	Advantages	Disadvantages.
Additive Noise [2]	Data is randomized by adding noise with a known statistical distribution.	1.Performs independently for each captured value. 2.Preserves statistical properties after reconstruction of the original distribution. 3.Noise reduction techniques cannot be used accurately.	1.Limits data utility to the use of aggregate distribution. 2.Masking extreme values require great quantities of noise, severely degrading data utility.
Multiplicative Noise [9]	Data is randomized by multiplicative noise with a known statistical distribution.	1.The reconstruction of original individual values is difficult. 2.Perform independently for each captured value. 3.Perform statistical properties after reconstruction of the original distribution.	1.Limits data utility to the use of aggregate distribution. 2.Masking extreme values. 3.Require great quantities of noise, severely degrading data utility.

A. Exiting PPDM perturbation method.

Different PPDM perturbation techniques including noise, randomization, resampling, global recording, data reduction, micro aggregation etc. In this paper, the randomization perturbation method is discussed. The randomization method is a technique for privacy-preserving data mining in which noise is added to the data to mask the attribute values of records [10] with two types Additive and Multiplicative.

Additive Randomization adds noise to sensitive records of the dataset. The amount of noise added in additive perturbation is independent of the sensitive attributes. [1] Considered a set of data records denoted by $X = \{x_1 \ldots x_N\}$. For record $x_i \in X$, a noise component is added which is drawn from the probability distribution $f_Y(y)$. These noise components are drawn independently and are denoted $y_1 \ldots y_N$. Thus, the new set of distorted records are denoted by $x_1 + y_1 \ldots x_N + y_N$. We denote this new set of records by $z_1 \ldots z_N$.

Thus, if X be the random variable denoting the data distribution for the original record, Y is the random variable describing the noise distribution, and Z be the random variable denoting the final record, we have:

$$Z = X + Y \qquad (1)$$

$$X = Z - Y \qquad (2)$$

In multiplicative perturbation, noise ε is multiplied to sensitive attributes. The sensitive data is derived from a known distribution.

$$Y = X^* \varepsilon \qquad (3)$$

III.PROPOSED PPDM PERTURBATION METHOD.

We have proposed a perturbation method where we will try to overcome the

drawback of Additive randomization and multiplicative randomization mentioned in II.

In the present work we first have pre-processed data, then pre-processed data is encoded.

Encoding is the process of converting the data or a given sequence of characters, symbols, alphabets, etc., into a specified format, for the secured transmission of data.

On encoded data, the mathematical reckoning like adding RMS, sorting of data is performed to prevent privacy of data before sharing it to the third party. The output of is called Randomized data.

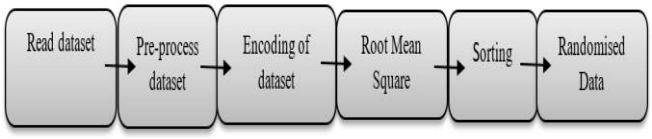

Figure2: Implementation Flow of Proposed method.

A. Implementation

All the experiments are performed on a Windows Operating System (Windows 10, 64bit) with 4 GB RAM.

For the implementation of perturbation methods, we have considered the mental health dataset and performed privacy preserving perturbation on it.

Dataset for mental health for the research work is taken from an online available dataset, provided by an OSMI (Open Sourcing Mental Illness) survey 2014. It mainly has information like age, gender, location of work, type of work, is he/she self-employed, total of 24 questions are included. This dataset has a total of 1258 records.

We consider the sensitive attribute whose privacy is critical in the area under consideration.

Research to provide summary of statistical information with minimal information loss and maximum data utility is the need of the hour.

Table II: Relation of Attribute, Attribute value and Encoding

Attribute	Attribute value	Encoding
age	18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 53, 54, 55, 56, 57, 58, 60, 61, 62, 65, 72	1, 2, 3, 4, 5, 6, 7, 8,9,10,11,12,13,14, 15,16,17,18,19,20,21, 22,23,24,25,26,27,28, 29,30,31,32,33,34,35, 36,37,38,39,40,41,42, 43,44,45
gender	'female', 'male', 'trans'	0, 1, 2
self_employed	'Yes','No'	1, 0
family_history	'Yes','No'	1, 0
treatment	'Yes','No'	1, 0
work_interfere	"Don't know", 'Never', 'Often', 'Rarely', 'Sometimes'	0, 1, 2, 3, 4
no_employees	'01-05', '100-500', '26-100', '500-1000', '06-25', 'More than 1000'	0, 1, 2, 3, 4, 5
remote_work	'No', 'Yes'	0,1
tech_company	'No', 'Yes'	0,1
benefits	"Don't know", 'No', 'Yes'	0, 1, 2
care_options	'No', 'Not sure', 'Yes'	0, 1, 2
wellness_program	"Don't know", 'No', 'Yes'	0, 1, 2
seek_help	"Don't know", 'No', 'Yes'	0, 1, 2
anonymity	"Don't know", 'No', 'Yes'	0, 1, 2
leave	"Don't know", 'Somewhat difficult', 'Somewhat easy', 'Very difficult', 'Very easy'	0, 1, 2, 3, 4
mental_health_consequence	'Maybe', 'No', 'Yes'	0, 1, 2
phys_health_consequence	'Maybe', 'No', 'Yes'	0, 1, 2
coworkers	'No', 'Some of them', 'Yes'	0, 1, 2
supervisor	'No', 'Some of them', 'Yes'	0, 1, 2
mental_health_interview	'Maybe', 'No', 'Yes'	0, 1, 2
phys_health_interview	'Maybe', 'No', 'Yes'	0, 1, 2
mental_vs_physical	"Don't know", 'No', 'Yes'	0, 1, 2
obs_consequence	'No', 'Yes'	0, 1
mental health disorder in the past	'Maybe', 'No', 'Yes'	0, 1, 2
have a mental health disorder	'Maybe', 'No', 'Yes'	0, 1, 2
diagnosed with a mental health condition by a medical professional	'No', 'Yes'	0, 1

Dataset is in .csv format which is first pre-processed by using "Jupyter Notebook".

Data cleaning is the process by which unwanted data is discarded to make it appropriate for further analysis. Incorrect format, errors while capturing, missing data acts as garbage data. Many of the attributes have empty values as input so default values will be assigned to it. Finally, data thoroughly checked to ensure that no vital part of it is missed out. The next step is encoding the data.

Encoding is the process of converting the data or a given sequence of characters, symbols, alphabets, etc., into a specified format, for the secured transmission of data. The root means square(RMS) of encoded data is calculated. RMS is the arithmetic mean of the squares of a set of numbers.

last but one step is sorting which means a rearrangement of a set of numbers. The output of sorting step is called randomized data. The final output is randomized data, which will conserve the privacy of data.

For the implementation of the Additive and Multiplicative perturbation method on an above dataset, we have perturbed "age" and "family history" as sensitive attributes.

B. Result and Discussion

Comparison of Additive Perturbation, multiplicative Perturbation and the Proposed Perturbation method

Table III: Experimental results of descriptive statistics calculation for Original dataset, Additive Perturbation, Multiplicative Perturbation and Proposed Perturbation method.

Details	Mean	Standard Deviation
Original dataset	14.042	7.157
Additive Perturbation	18.042	7.157
Multiplicative Perturbation	28.084	14.313
Proposed Perturbation Method	14.042	7.157

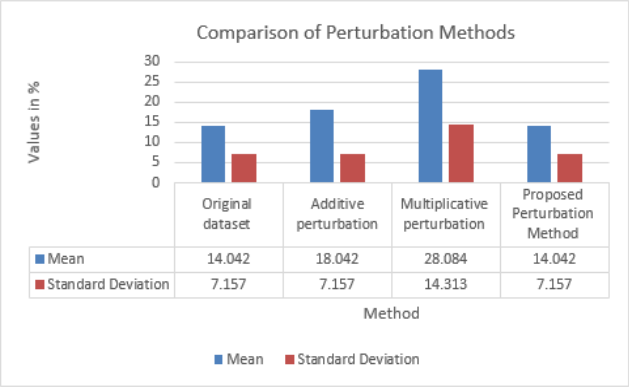

Figure3: Comparison of Additive Perturbation, Multiplicative Perturbation method and Proposed Perturbation Method

To signify minimal information loss, the mean of the perturbed dataset should be near to the mean of the original dataset.

From the table3 and figure3, mean of Additive perturbation, Multiplicative perturbation is greater than the mean of the Original dataset which means there is data loss when we implement the Additive perturbation, Multiplicative perturbation. The proposed method mean is closer to the mean of the original dataset, which implies that there is less data loss in the proposed method.

Standard Deviation is a statistical term used to measure the amount of variability or dispersion around an average. Technically, it is a measure of volatility. Dispersion is the difference between the actual and the average value. The larger this dispersion or variability is, the higher is the standard deviation. The standard deviation of Additive perturbation and proposed perturbation method refers closely to the standard deviation of the original dataset. Mean proposed method gives more data accuracy compared to Additive and multiplicative perturbation method.

IV. CONCLUSION AND FUTURE WORK

In privacy preserving randomization data mining perturbation shows an imperative role, where the privacy of data is maintained by adding noise in it. In the current paper, experimental work performed to compare PPDMP methods with proposed method. The aim of the proposed method is that owners could share data with researchers without sacrificing data confidentiality. Based on experimental work performed in the current paper, it can be inferred that the mean of the proposed perturbation method is close to mean of the original dataset. This signifies that there is no data loss while perturbation is performed on the dataset. The standard deviation of Additive perturbation and proposed perturbation method is referred closely to the standard deviation of the original dataset. The proposed method gives more data accuracy compared to the Additive and Multiplicative perturbation method. The experiments and the results described in this paper can be used to resolve the compromise between privacy and information loss. Randomized data can further be encrypted to achieve confidentiality of data. Privacy and security both can be accomplished which can be extended as future work of this paper.

REFERENCES

1. Aggarwal C.C., Yu P.S. An Introduction to Privacy-Preserving Data Mining. In: Aggarwal C.C., Yu P.S. (eds) Privacy-Preserving Data Mining. Advances in Database Systems, vol 34. Springer, Boston, MA, (2008)
2. Agrawal R., Srikant R., "Privacy-preserving data mining", ACM SIGMOD international conference on Management of data, 29 (2), pp 439-450 (2000).
3. Tang P.C., McDonald C.J. (2006) Electronic Health Record Systems. In: Shortliffe E.H., Cimino J.J. (eds) Biomedical Informatics. Health Informatics. Springer, New York, NY
4. Samarati P., Sweeney L., "Generalizing data to provide anonymity when disclosing information", PODS '98 Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of database systems, 809-812 (1998).
5. Machanavajjhala A., Kifer D., J. Gehrke, M. Venkita Subramaniam, " ℓ -diversity: Privacy beyond k-anonymity ", ACM Transactions on Knowledge Discovery from Data, 1(1), 809-812 (2007).
6. Li N , Li T, Venkatasubramanian S , "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, 106-115 (2007).
7. Xiao X., Tao Y., "Personalized privacy preservation", In Proceedings of ACM Conference on Management of Data (SIGMOD'06), 229-240, (2006).
8. Liew C. K., Choi U. J. & Liew C. J., "A Data Distortion by Probability Distribution", ACM Transactions on Database Systems,10(3), 395–411 (1985).
9. KimJ. J., WinklerW. E., "Multiplicative noise for masking continuous data", Technical Report Statistics 2003-01, Statistical Research Division, US Bureau of the Census, Washington D.C., pp1-18 (2003).
10. Agrawal D. Aggarwal C. C. On the Design and Quantification of Privacy-Preserving Data Mining Algorithms. ACM PODS Conference, 2002.
11. Li X. B.& Sarkar S., "A Tree-based Data Perturbation Approach for Privacy-Preserving Data Mining." IEEE Transactions on Knowledge and Data Engineering, 18, 1278–1283 (2006).

AUTHORS PROFILE

Ms. Vijaya Umesh Pinjarkar, is Ph.D. scholar in Sir Padampat Singhania University, Udaipur, India and an Assistant Professor in the department of Information Technology, K.J. Somaiya Institute of Engineering and Information Technology,Mumbai University. She has completed M.E. (Information Technology) from Mumbai University, India in 2013 and B.E. (Information Technology) from Sant Gadge Baba Amravati University, India in 2005. She has guided undergraduate projects Her area of research is data mining, cryptography, network security, machine learning. She has nearly 15 years of teaching experience. She has publications in International and national conferences and Journals.

Dr. Amit Jain, is presently working as Assistant Professor in Computer Science and Engineering Department, Sir Padampat Singhania University, Udaipur, India. He has completed his Doctoral Degree in Computer Engineering in year 2016. He is having teaching and administrative experience of 23 years. He has taught to Ph.D, post-graduate and graduate students of engineering. He has about 30 research publications in International Journals and Conferences. He holds the post of Associate Editor in many International Research Journal. He is reviewer in IEEE, Inderscience and Elsevier and Scopus Indexed Journals.

Dr. Anand A. Bhaskar, is presently serving as Head & Assistant Professor in the Department of Electronics & Communication Engineering from the Sir Padampat Singhania University, Udaipur, Rajasthan, India. He received Ph.D. in Electronics from the Saurashtra University, Gujarat in 2009. He has more than 30 research papers in international/national journals and conferences to his credit. His current research interest areas include Embedded Systems, Wireless Sensor Networks, and Energy Harvesting etc. He has organized various conferences and FDP in the field of his expertise and having 14 years of experience of teaching at undergraduate and postgraduate and doctorate level. He is proponent of utilizing technology to enhance teaching and proficient in clearly explaining complex topics to students with all level of scientific understanding.

Dr. Prateek Srivastava, received the B.Tech and M. Tech degree from Uttar Pradesh Technical University. He received PhD degree in Computer Science and Engineering from Sir Padampat Singhania University, Rajasthan. He had worked as an Assistant Professor in the Department of Computer Science and Engineering at Hindustan College of Science and Technology from 2005 to 2011.Presently he is associated with Sir Padampat Singhania University. His research interests include system modelling, refinement of distributed systems, verification and reasoning of critical properties using formal techniques.