UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

# Comparative Study on Perturbation Techniques in Privacy Preserving Data Mining

Desmond Ko Khang Siang, Siti Hajar Othman and Raja Zahilah Raja Mohd Radzi
Faculty of Computing,
Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Johor, Malaysia.
Email: desmondkks@hotmail.com, hajar@utm.my, zahilah@utm.my

Submitted: 12/01/2018. Revised edition: 26/03/2018. Accepted: 8/05/2018. Published online: 21/05/2018

***Abstract*—Data Mining is a computational process that able to identify patterns, trends and behaviour from large dataset. With this advantages, data mining has been applied in many fields such as finance, healthcare, retail and so on. However, information disclosure become one of an issue during data mining process. Therefore, privacy protection is needed during data mining process which known as Privacy Preserving Data Mining (PPDM). There are several techniques available in PPDM and each of the techniques has its' own benefits and drawbacks. In this research, perturbation technique is selected as privacy preserving technique. Perturbation technique is a method that alters the original data value before the application of data mining. In PPDM applications, perturbation technique able to provide a protection of data privacy but the accuracy of data should not be ignored too. In this research, three perturbation techniques are selected which are additive noise, data swapping and resample. For data mining techniques, two methods of classification are selected which are Naïve Bayes and Support Vector Machines (SVM). With the selection of these techniques, the experimental results are evaluated based on the hiding failure, accuracy and precision. For overall result, resample is selected as the best perturbation technique in naïve bayes and SVM classification for both glass and ionosphere datasets.**

**Keywords — Privacy Preserving Data Mining (PPDM), Perturbation, Accuracy, Hiding Failure**

## I. INTRODUCTION

In the era of modernization, the technology is evolving very quickly especially in the development of network, storage, data collection and transfer. With this technology, enormous of data will be sharing, sending or transferring around the internet. Nowadays, these technologies are very important in business operation because it able to provide many organizations to work effectively, efficiently, save time and increase an organization working quality.

In order to handle the huge amount of data in the database, data mining technique are introduced. Data mining is the process of sorting through huge dataset to identify the patterns and establish the relationships to solve problems by undergoing data analysis process [1]. From the definition of data mining, it consists of ability to uncover the hidden patterns and relationships of data so that a prediction of impact businesses can be made. Unfortunately, a normal process of data mining does not provide any security to protect that particular data. Therefore, Privacy Preserving Data Mining (PPDM) has become one of the popular trends in privacy and security for handling large amount of sensitive information. PPDM concept is often more complex and referred to "getting valid data mining results without learning the underlying data values" [2] and the algorithms of data mining will analyze the side-effects incur in data privacy. After analysis process, it comes to the main objective in PPDM which is develop algorithms for modifying the original data so that the private data and private knowledge will remain private even after the mining process [3]. Besides that, PPDM main considerations is two folds [3]. First is sensitive raw data such as identifiers, names, addresses and so on should be modified from the original database. This is because it able to prevent the recipient of the data to compromise or alter another person's privacy. Second is sensitive knowledge which should be excluded and can be mined from a database by using data mining algorithms. This is because a knowledge has the ability to compromise data privacy.

Perturbation approach is one of the PPDM technique that transform the data to another form. In perturbation technique, it alters the original data directly instead of reconstruct another data. This technique is chosen and selected by some of the data

owner due to they do not want to expose their privacy [4]. In PPDM applications, although perturbation technique provides a protection of privacy for data owners, but the accuracy of data should not be ignored. In order to know how efficient the PPDM technique used, an evaluation of data quality and privacy level is required.

## II. Related Work

PPDM is a combination techniques between privacy preserving techniques and data mining techniques. Based on the studies and related works, PPDM techniques can be classified into five major parts. The existing techniques or approaches in PPDM are Anonymization, Condensation, Cryptography, Randomization and Perturbation [5].

### *A. Anonymization Approach*

In anonymization technique, it is mainly focus on the concern of the hidden identity and sensitive data of a particular information and assume for analysis preservation. This technique aim is to make the record of a particular individual very similar among a large set of records by applying the generalization and suppression techniques [6]. It consists of three different types which are k-anonymization, l-diversity and t-closeness. Based on definition of k-anonymization, if a dataset T is separated into a partition and each group $G_i$ ($1 \leq i \leq$ p) of that partition (p) must at least contains of K-records and dataset T is either being generalized or suppressed, then the dataset T is considered as satisfied the k-anonymity [7]. Next is l-diversity which is based on the concept of inside the group diversity of sensitive values. Based on definition of l-diversity, if each group consists of one sharing combination of key attributes, it must be at least one "well-represented" values for each confidential attribute, then the dataset will be considered aa satisfy l-diversity [8]. Moreover, the other privacy rules designed by [9] are known as l-diversity. This technique invented and bring some advantages than k-anonymity. This is because k-anonymity dataset vulnerable to strong attacks due to lack of diversity in private attribute [9].

### *B. Condensation Approach*

Condensation approach was introduced by [10]. It is one of the PPDM techniques that uses a methodology that condenses the data into various groups of same size. Every group has a size at least k which allude to indistinguishability level of privacy persevering. The higher the indistinguishability level, it meant the higher the amount of privacy contain in that data.

### *C. Cryptography Approach*

Cryptography approach introduced by [11] and stated that this approach can provide security and safety for the sensitive attributes of dataset. Cryptography technique became popular in PPDM field after introducing due to two reasons [11]. Firstly, this approach provides several privacy models that including the methodologies for the purpose of proving and quantifying. Second reason is various kind of tools and mechanism provided and proven that cryptography is the fastest grown. The goal of cryptography technique is used to encrypt all the relevant data and share the data by using a dedicated algorithm which able to generate results and send to all recipients [12]. When there are multiple parties involve, an addition party will be included which is a third party or also known as "trusted party".

### *D. Randomization Approach*

Randomization technique was introduced by [13]. Based on this technique, it is also known as randomize response. In this approach, the data will be in disordered or confusing state so that the central place could not know the exact probabilities better than a pre-defined threshold, whether the data obtained from any resources that consists of real information or false information.

### *E. Perturbation Approach*

Perturbation approach is one of most popular and simple technique that used in PPDM. This technique is functioning by perturbing the original values in dataset with some synthetic data values during the process. After that, the result of perturbed data releases will be used for data analysis. In the outcome of the process, the perturbed data that produced in statistical information shown that it does not has large different compare to the original data. So, perturbation technique able to control the disclosure of statistical due to some characteristics of this technique which is simple features, efficient and capable to maintain the statistical information. With this technique, the attackers not able to launch any attack to obtain the sensitive information from the published data.

## III. Justification of Techniques and Evaluation

In this section, the selected techniques in perturbation approach for this project will be discussed. The techniques that are going to use in this project are normalization, additive noise, data swapping and resample.

### *A. Data Preprocessing and Perturbation Techniques*

In order to perform PPDM, data preprocessing which is normalization is performed to remove the noise of dataset. After that, follow by selected perturbation techniques for this research.

#### *1) Normalization*

Normalization is one of the technique that will implement in the beginning stage of PPDM which known as pre-processing stage. The main purpose of this technique is to helps to map the data in a certain range of scale which also improving the effectiveness in analysis of data. Besides that, the accuracy and efficiency for implementing the data mining algorithms can be increased too [14]. Normalization technique consists of three different ways which are Min-Max, Z-Score and Decimal Scaling normalization. However, the Min-Max normalization will be the one that focus on this project. In Min-Max normalization, it is performing on linear alteration on the original datasets. The values in the data will be normalized by setting a certain range which in the range of minimum 0.0 and maximum 1.0.

$$z = (B - A)\frac{x - \min(x)}{\max(x) - \min(x)} + A \quad (1)$$

$$z = (0.8) * (\frac{x - \min(x)}{\max(x) - \min(x)}) + 0.1 \quad (2)$$

where z = normalized data, x = ($x_1$, … , $x_n$), A = predefined of lower boundary, B = predefined of higher boundary, min (x) = minimum value of x, max (x) = maximum value of x [15].

*2) Additive Noise*

Additive noise in perturbation approach is a technique of addition of random noise into the actual data. The values of R can be a positive or negative values but in the range between 0 and 1. This is because normalization has defined the values of instances between 0 and 1. The equation of additive noise is shown as the following:

$$Y = X + R \quad (3)$$

Where X = original data, R = additive noise and Y = perturbed data.

*3) Data Swapping*

Data swapping is one of the technique in random perturbation approach. The main goal of data swapping is preserving the amount of information but the process will perturb the data values randomly in order to maintain the confidentiality of data. Initially, the data swapping will select the data by random, then it will continue by searching the swapping partner for each of the selected data according to their similarity of characteristics. If the characteristics are matched, then the data values will be swapped between the selected data and their swapping partners.

*4) Resampling*

Resample is a protection method that control the disclosure of numerical microdata. Resample is functioning by replacing the t samples with the n values from the original data. After that, the sample will be sorted and find the average of sample values. At the end of this technique, a resample of synthetic datasets will be produced with have the same distributional characteristics as the original microdata. A random subsample of dataset will be produced by applying the sampling either with or without the replacement of sample values.

*B. Selected Data Mining Techniques*

After implementation of privacy preserving techniques, data mining such as naïve bayes and support vector machines are implemented on perturbed dataset respectively.

*1) Naïve Bayes*

Naïve Bayes classifier is known as a family of simple probabilistic classifier. This technique applied the Bayes' theorem and assuming each of the features in a class are strong independent. The Naïve Bayes classifier usually can be completed based on the prior of probability and the likelihood to a class. It was introduced for the purpose of text categorization in the early of 1960s and it still remains popular nowadays [16]. This technique is highly scalable which is required a few number of parameters linear in the number of variables for a learning problem. Besides that, the maximum likelihood training used the evaluation of close form expression which applying the linear time to complete it. Moreover, Naïve Bayes is works faster and simpler in order to test the dataset and it also able to perform well for multi-class prediction. The formula of Naïve Bayes is shown in the following:

$$P(c \mid x) = \frac{P(x \mid c)\, P(c)}{P(x)} \quad (4)$$

where P (c | x) = the posterior probability of class/target and provided predictor/attribute,
P (c) = the prior probability of class,
P (x | c) = the likelihood which probability of predictor provided class,
P (x) = the prior probability of predictor

*2) Support Vector Machines (SVM)*

Support vector machine (SVM) is a one of the supervised machine learning algorithm that developed for classification problems where the datasets used will teach SVM regarding to the classes so that SVM able to classify any of new data. In the other word, SVM analyses the data and recognizes the pattern of datasets which is the role of classification. It functions is classified the data into various classes by searching for a line or hyperplane that divides the training dataset into classes. Since there will be many linear hyperplane, the margin maximization will be used. It is a method that maximize the distance between different classes that involved [17]. If the hyperplane that maximizes the distance of classes are identified, then the probability to generalize the hidden data will be increased.

*C. Selected Dataset from Weka 3.8*

Glass dataset is created by B. German which works in the Central Research Establishment at Home Office Forensic Science Service that located at Aldermaston. This dataset was donated to the public by Vina Spiehler in September 1987. This dataset was used for the study of classification for types of glass that appear in the crime investigation. This is because glass that left in the crime scene can be a relevant evidence. In this dataset, it consists of 10 attributes and 214 instances which all attributes are in numerical form.

Ionosphere dataset is one of the datasets that provided in the Weka software. Ionoshpere dataset is owned by the Space Physics Group of Applied Physics Laboratory in Johns Hopkins University. This dataset was donated to the public by Vince Sigillito in 1989. This dataset was used for investigating of usage of backprop and perceptron for the

training algorithm. +In this dataset, it consists of 35 attributes and 315 instances. For the attributes, all the 34 are predictor attributes which in continuous values, while 35$^{th}$ is in options which is good or bad.

### *D. Performance Measurements*

The quality of PPDM algorithms can be assessed through the evaluation of parameters such as performance, utilities of data, uncertainty level and the resistance of data mining [18]. Besides that, there are also several popular performance measurements in PPDM such as the level of protection in privacy, the time of execution, hiding failure and difference in percentage value. Unfortunately, several existing works have been proposed regarding to the evaluation of performance measurement but there is no specific metric that accepted by research community [18]. However, there are several existing papers shown the acceptable at least 68.2% accuracy for glass dataset and at least 86% accuracy for ionosphere dataset from Wlodzislaw Duch in 2010.

#### *1) Data Quality*

In the perspective of data mining, data quality is a measure of data consistency between data views in information system and the data in the real world. There are a few methods of measurement that have been used and applied by researchers to analyse the quality of data mining result. Classification is one of the data mining methods which can use the confusion matrix in order to calculate the accuracy and precision. The confusion matrix is shown in Table I:

TABLE I. CONFUSION MATRIX

| Current Classes | Predicted Classes | | |
|---|---|---|---|
| | | True Class | False Class |
| | True Class | True Positive (TP) | False Positive (FP) |
| | False Class | False Negative (FN) | True Negative (TN) |

With the confusion matrix, accuracy can be calculated. Accuracy is shown the effectiveness of a classifier used by calculating the ratio between the number of corrected classified cases and the sum of number of cases. Besides that, accuracy also considered closely related to information loss after the PPDM is implemented [18]. The higher the percentage of accuracy, the better the quality of data. The formula of accuracy is shown as the following:

$$\text{Accuracy} = \frac{TP\ Rate+TN\ Rate}{TP\ Rate+TN\ Rate+FP\ Rate+FN\ Rate} \quad (5)$$

Besides that, the precision of a dataset also can be done by making use of confusion matrix. Precision is the instances used that are classified into their exact classes. The formula of precision is shown as the following:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (6)$$

#### *2) Privacy Level*

The privacy level is a measurement of the privacy revelation of a dataset. This measurement can be completed by determined the common sensitive pattern of information between the original data and the perturbed data.

Moreover, the privacy level of published data can be determined by using the hiding failure measurement. This measurement usually will take place after the process of data sanitization on the original database. The formula of hiding failure (HF) is shown as the following:

$$HF=\frac{\#\text{Rp}(\text{D}')}{\#\text{Rp}(\text{D})} \quad (7)$$

Where $\#R_p$ (D') and $\#R_p$ (D) are the number of restrictive patterns that discover, while D' represent sanitized dataset and D represent original dataset.

## IV. EXPERIMENTAL RESULT AND DISCUSSION

### *A. Result of Privacy Level*

In this research, privacy level is used for the measurement the level of hiding failure for perturbed dataset after the implementation of selected perturbation techniques such additive noise, data swapping and resample. The privacy level can be estimated by using hiding failure which is the portion of sensitive information that are not hidden after implementing selected perturbation techniques. The lower the percentage of hiding failure, the better the privacy preserving technique is preserved the sensitive information within the datasets.

Table II that resample technique is the best among selected privacy preservation techniques after implementing in glass dataset. This is because resample technique showed and obtained the lowest percentage of hiding failure compare to additive noise and data swapping techniques. In glass dataset, there are 1926 instances in total and resample technique successfully preserved 1663 instances which is the highest value compare to additive noise and data swapping. While the rest of instances are the instances that fail to hide which still able to discover after privacy preserving techniques are implemented. Next is follow by data swapping technique which the percentage of hiding failure is 13.71%. There are 264 instances still can be discovered after privacy preserving technique is applied. The number of instances can be discovered for data swapping and resample only differ in one instances but it showed a clear difference in percentage. While additive noise has the highest percentage of hiding failure which is 14.84% and 286 instances can be discovered after the privacy preserving technique is implemented.

TABLE II. RESULT OF PRIVACY LEVEL FOR GLASS DATASET

| Perturbation Techniques | Restrictive Pattern Discovered in Sanitized Glass Dataset | Restrictive Pattern Discovered in Normalized Glass Dataset | Hiding Failure (%) |
|---|---|---|---|
| Additive Noise | 286 | 1926 | 14.84 |
| Data Swapping | 264 | 1926 | 13.71 |
| Resample | 263 | 1926 | 13.66 |

TABLE III: RESULT OF PRIVACY LEVEL FOR IONOSPHERE DATASET

| Perturbation Techniques | Restrictive Pattern Discovered in Sanitized Ionosphere Dataset | Restrictive Pattern Discovered in Normalized Ionosphere Dataset | Hiding Failure (%) |
|---|---|---|---|
| Additive Noise | 923 | 11934 | 7.73 |
| Data Swapping | 857 | 11934 | 7.18 |
| Resample | 822 | 11934 | 6.89 |

In Table III, the result of privacy level for ionosphere dataset showed that resample technique has the lowest percentage of hiding failure which is 6.89%. In ionosphere dataset, there are 11934 instances and resample technique successfully hide 11112 instances and the rest of 822 instances still able to be discovered although the privacy preserving technique is implemented. Next is follow by data swapping which has 7.18% of hiding failure after implementation on ionosphere dataset. However, data swapping is ranked as second place because the percentage of hiding failure is higher than resample. The number of instances still can be discovered is 857 instances and the number of instances discovered for resample is 822. This show that the number of instances discover for data swapping and resample only 35 instances in differ. Meanwhile, additive noise is ranked as last place because obtained the highest percentage of hiding failure which is 7.73% and 923 instances can be discovered after implementing on the dataset.

### *B. Result of Data Quality*

Table IV showed the result of accuracy and precision after PPDM techniques are implemented. The perturbation techniques such as additive noise, data swapping and resample are implemented first then follow by data mining methods. After data mining methods are implemented, data quality is evaluating based on two parameters which are accuracy and precision. In Naive Bayes, additive noise and data swapping obtained the same accuracy and precision result which is 90.654% and 90.700% for glass dataset. While resample obtained the lowest result which is 88.785% of accuracy and 88.600% of precision in perturbed glass dataset after Naïve Bayes is applied. By using glass dataset, the benchmark cannot be made since the result of additive noise and data swapping are the same, therefore another dataset which is ionosphere dataset is required. For ionosphere dataset, data swapping obtained the highest accuracy and precision compare to other two methods which is 83.476% of accuracy and 84.800% of precision. This can be concluded that data swapping is the best among selected perturbation techniques in Naive Bayes method.

Meanwhile, when SVM is implemented in three perturbation technique, the results are different compare to Naïve Bayes for both datasets. In glass dataset, resample has the highest result which is 92.056% of accuracy and 91.900% of precision. While data swapping is ranked as second place and follow by additive noise. For ionosphere dataset, additive noise and data swapping obtained the same percentage of accuracy and precision which is 90.3134% of accuracy and 91.200% of precision.

TABLE IV: RESULT OF PRIVACY LEVEL FOR IONOSPHERE DATASET

| Perturbation | Additive Noise | | Data Swapping | | Resample | |
|---|---|---|---|---|---|---|
| Data Mining | Naïve Bayes | | | | | |
| Datasets | Glass | Ionosphere | Glass | Ionosphere | Glass | Ionosphere |
| Accuracy (%) | 90.654 | 82.906 | 90.654 | 83.476 | 88.785 | 82.621 |
| Precision | 0.907 | 0.844 | 0.907 | 0.848 | 0.886 | 0.841 |
| | | | | | | |
| Data Mining | Support Vector Machines | | | | | |
| Datasets | Glass | Ionosphere | Glass | Ionosphere | Glass | Ionosphere |
| Accuracy (%) | 90.654 | 90.313 | 91.589 | 90.313 | 92.056 | 92.593 |
| Precision | 0.904 | 0.912 | 0.914 | 0.912 | 0.919 | 0.934 |

However, it does not affect the benchmarking since resample also obtained the highest accuracy and precision which is 92.593% and 93.400% respectively. Therefore, resample is selected as the best in benchmarking after the perturbed dataset implementing SVM methods.

## V. FUTURE WORKS

In this research, there are three privacy preserving techniques with two data mining algorithms applied on glass and ionosphere datasets. From the result obtained, the data

mining applied in both datasets obtained large difference of accuracy and precision. Therefore, a research of data mining methods can be studied, experimented and evaluated regarding which data mining methods more appropriate to which kinds of dataset.

Besides that, there are other privacy preserving techniques such as randomization, cryptography, condensation and anonymization approach that can be used for experimenting. Moreover, there are plenty of data mining models other than classification such as association rule, clustering and regression.

Moreover, some of the data owner prefer perturbation technique because it can help to prevent the disclosure of data privacy [19]. However, since perturbation technique direct distort the original data, it will be difficult for data owner if he requests for reversible the process of PPDM. Therefore, a reversible PPDM can be considered as one of the future works

## VI. CONCLUSION

In conclusion, resample technique is selected as most secured in privacy level which obtained the least of percentage in hiding failure for both dataset. In Naïve Bayes, data swapping shown as the most accurate and precise when Naïve Bayes is implementing in glass dataset and ionosphere dataset. However, resample is selected as most accurate and precise when SVM mining method is implementing in both dataset. For overall, resample is selected as the best perturbation among these three techniques.

## REFERENCES


[1] Margeret Rouse. (2008). Data Mining. Available at http://searchsqlserver.techtarget.com/definition/data-mining.

[2] Clifton, C., Kantarcioglu, M. & Vaidya, J. (2002). Defining Privacy for Data Mining. Proceeding of the National Science Foundation Workshop on Next Generation Data Mining, Baltimore, MD, USA, pp. 126-133.

[3] Vassilios, S. Verykios, Elisa Bertino. (2004). State of the Art in Privacy Preserving Data Mining. *Proceeding of Special Interest Group on Management of Data (SIGMO) Record*, 33(1), 50-57.

[4] Burcu, D. O., Murat, O., Mehmet, K. and Huseyin, P. (2015). A Survey: Deriving Private Information from Perturbed Data. *Artif Intell Rev*, 547-569.

[5] Sharma, M., Chaudhary, A., Mathuria, M., Chaudhary, S., and Kuma, S. (2014). An Efficient Approaches for Privacy Preserving in Data Mining.

[6] A. S. Shanthi and Dr. M. Karthikeyan. (2012). A Review on Privacy Preserving Data Mining. *IEEE International Conference on Computational Intelligence and Computing*.

[7] Jian-min, Han, Cen Ting-Ting and Yu Hui-Qun. (2008). An Improved V-MDAV Algorithm for l-diversity. *Information Processing (ISIP), 2008 International Symposiums.*

[8] M. E. Nergiz, C. Clifton and A. E. Nergiz. (2009). Multirelational k-anonymity. *Knowledge and Data Engineering, IEEE Transactions on*, 21(8), 1104-1117.

[9] A. Machanavajjhala, J. Gehrke, Kifer D and Venkitasubramaniam M. (2006). L-Diversity: Privacy Beyond k-Anonymity. *Proceeding of ICDE*.

[10] Aggrawal, C. C., and Philip, S. Y. (2004). A Condensation Approach to Privacy Preserving Data Mining. *In Advances in Database Technology EDBT 2004: 183-199*. Springer Berlin Heidelberg.

[11] Sven Laur, Helger L. and Taneli M. (2006). Cryptographically Private Support Vector Machines. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 618-624.

[12] Matwin, S. (2013). Privacy Preserving Data Mining Techniques: Survey and Challenges. In *Discrimination and Privacy in the Information Society*. Springer Berlin Heidelberg, 209-221.

[13] Agrawal R. and Srikant R. (2000). Privacy Preserving Data Mining. *SIGMOD '00 Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*: 439-450.

[14] Ryan Stephens and Ron Plew. (2002). The Database Normalization Process. *Sans Teach Yourself SQL in 24 Hours*. 3rd Edition.

[15] S. Gopal Krishna Patro and Kishore Kumar Sahu. (2015). Normalization: A Preprocessing Stage. Available from <http://www.arxiv.org> [26 November 2016].

[16] Russell, Stuart; Norvig, Peter. (2003). *Artificial Intelligence: A Modern Approach*. 2nd ed. Prentice Hall. ISBN 978-0137903955.

[17] G. Chen, Wang T. J., Gong L. Y. and Perfecto H. (2010). Multi-class Support Vector Machine Active Learning for Music Annotation. *International Journal of Innovative Computing, Information and Control*. 6(3A), March 2010.

[18] Elisa B., Dan L. and Wei J. (2008). A Survey of Quantification of Privacy Preserving Data Mining Algorithms. *Privacy-Preserving Data Mining*, 183-205

[19] Burcu D. O., Murat O., Mehmet K. and Huseyin P. (2015, Sep). A Survey: Deriving Private Information from Perturbed *Data. Artif Intell Rev*, 547-569.