UTM Computing Proceedings
Innovations in Computing Technology and Applications
Volume: 1 | Year: 2016 | ISBN: 978-967-0194-81-3

Privacy Preserving Data Mining Based On Random Substitution and Decision Tree Classification

Mohamad Shafiq Aizuddin Ali Yasak[1], Maheyzah Md. Siraj*[2]

Department of Computer Science, Faculty of Computing,
Universiti Teknologi Malaysia,
81310 Johor Bahru, Johor, Malaysia

[1]nieshafiq@gmail.com, [2]maheyzah@utm.my

Abstract

In digitalization era, people frequently use the Internet to store and access their personal data and information in the large database system or data warehouse due to the advance in storage capacity and information processing. These data are able to extract for the analysis and interpretation purpose by using several data mining techniques such as classification. Unfortunately, some data that contain private and sensitive information about individual are exposed to the invasion by unauthorized parties during the process of data mining. This problem can be solved by adopting privacy preserving approaches such as data perturbation and cryptographic during the data mining process. This paper discusses the effectiveness of privacy preserving data mining based on random substitution and decision tree classification techniques in preserving better privacy and providing better mining quality. The objectives of this research is to develop a random substitution algorithm for preserving privacy on health dataset, implement a decision tree classification algorithm on preserved dataset for data mining and to evaluate and benchmark the proposed technique using quantification measurement in term of hiding failure on preserving dataset and classification accuracy on mining process. Random Substitution Perturbation algorithm was used as data perturbation technique and J48 algorithm as decision tree classification technique. The result will be measured and compared in term of privacy level and mining quality with other related works

Keywords: Privacy Preserving, Data Mining, Data Perturbation, Decision Tree Classification

1.0 Introduction

Data mining or the process of extracting knowledge or pattern from a large amount of data has recently emerged field that closely related in the world of database, statistics and artificial intelligence. In digitalization era, both organizations and individuals frequently use the Internet to store and access their personal data and information in the large database system or data warehouse. However, some data that contain private and sensitive information about individual are exposed to the invasion by unauthorized parties. Therefore, privacy preserving

data mining techniques are required in order to protect individual privacy while allowing data mining process.

Several methods of privacy preserving data mining have recently been explored base of four main approaches which are data perturbation, condensation approach, anonymization and cryptographic. Two approaches most widely adopted where data perturbation and cryptographic (Dowd *et al*., 2006). The cryptographic approach explored by Lindell and Pinkas, (2000), offers better privacy and accuracy but has poor performance. The data perturbation approach works on different data mining algorithms such as decision tree, association rules, clustering and etc., which offer better performance (Dowd *et al*., 2006).

The pioneer privacy preserving data mining decision tree method is introduced by (Agrawal and Srikant, 2006). Unfortunately, Kargupta *et al*., (2003) proved that the method was flawed and not suitable for a long period of time. This is because the perturbed data can be recovered by adversary easily. However, the framework introduced in Agrawal and Srikant, (2006) has its own privileges and useful features that should be maintained and rescued.

Therefore, the performance problem of existing PPDM techniques in preserving the sensitive data needs to be addressed in developing the better technique (Aggarwal *et al*., 2008). So, it will preserve the better privacy of sensitive data alongside it allows the data mining process.

Objectives of the projects are : (i**)** to develop a random substitution algorithm as data perturbation for preserving privacy in data, (ii) to implement a decision tree classification algorithm for data mining on health dataset, and (iii) to evaluate and benchmark the proposed PPDM using quantification measurement in term of hiding failure and data loss.

2.0 Methodology

This study is divided into three main phases which is the phase is about the studying the frameworks of the existing random substitution technique and developing the proposed algorithm. Second phase, the analysing existing decision tree classification and implementing the algorithm on health dataset. The last phase is about evaluating and benchmarking the proposed PPDM algorithm based on quantification measurement (Hiding Failure) and (Classification Accuracy). MATLAB and WEKA are used to run those algorithms.

a) Development of Random Substitution Algorithm

The first step in this phase is by studying the existing data perturbation privacy preserving framework from the previous works. The related information has been collected for analysis purpose including how the algorithm works, it matrices measurement and it advantages and disadvantages compared to other privacy preserving algorithms.The next step is the designing of Random Substitution Perturbation algorithm. The concept of this algorithm is to replace the value of each attribute by other value that has been chosen randomly from attribute domain using probabilistic model. MATLAB is used to run that algorithm.

b) Implementation of Decision Tree Classification Algorithm

The second phase is about the implementation of decision tree classification algorithm. The first step is done by studying the existing algorithm from previous works. Then, some relevant information is collected for implementing process.The C4.5 or J48 algorithm is implemented in this study. This algorithm works on the reconstructed dataset that generated by Matrix-based Reconstruction Matrix (MR) algorithm. The reconstructed dataset is to learn decision tree using existing decision tree mining algorithm besides it not violate the privacy

guarantee of original dataset. WEKA is used to run that algorithm.

c) Evaluation and Benchmarking

This phase is for evaluating and benchmarking the proposed algorithms. The Diabetes dataset is run under both algorithms to measure their performance.There are two matrices that will be used as quantification measurements which are privacy level and mining quality. The proposed PPDM techniques is benchmarked with Logistic Regression (LR) and Support Vector Machine (SVM) and with other related works.

3.0 Result

Table 1 below, shows the result of Hiding Failure on Random Substitution Perturbation (RSP) algorithm.

Table 1 Result of Hiding Failure on RSP algorithm

Test	Number of Restrictive Patterns Discovered from Sanitized Dataset	Number of Restrictive Patterns Discovered from Original Dataset	Hiding Failure (%)
1	106	1536	6.90
2	110	1536	7.16
3	101	1536	6.58
4	102	1536	6.64
5	32	1536	2.08

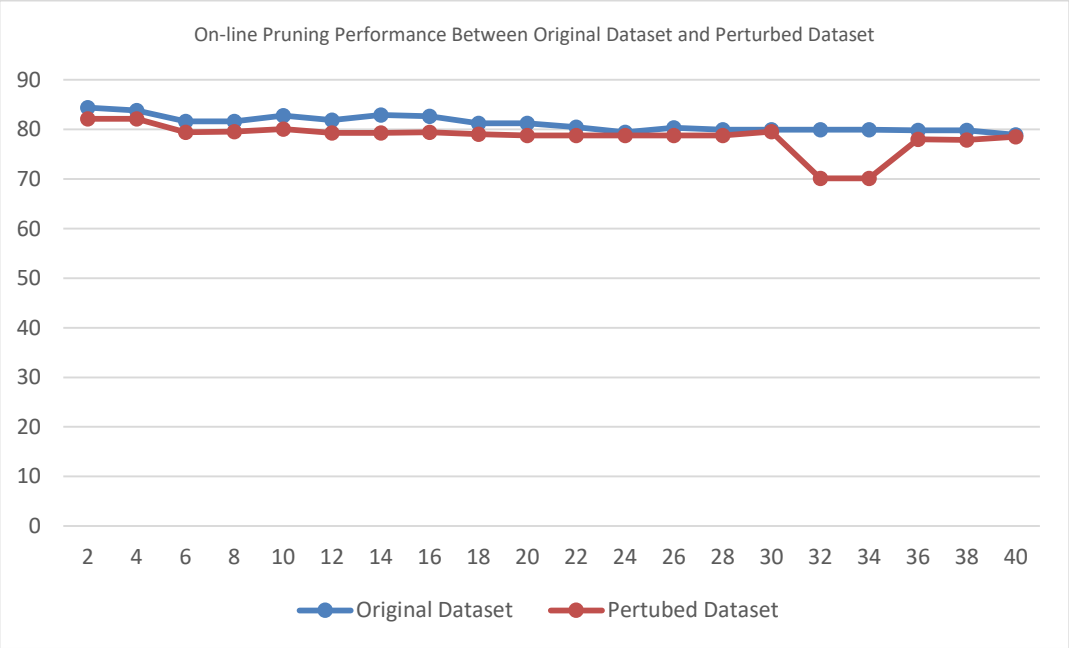

Figure 1 On-line Pruning Performance Between Original Dataset and Perturbed Dataset

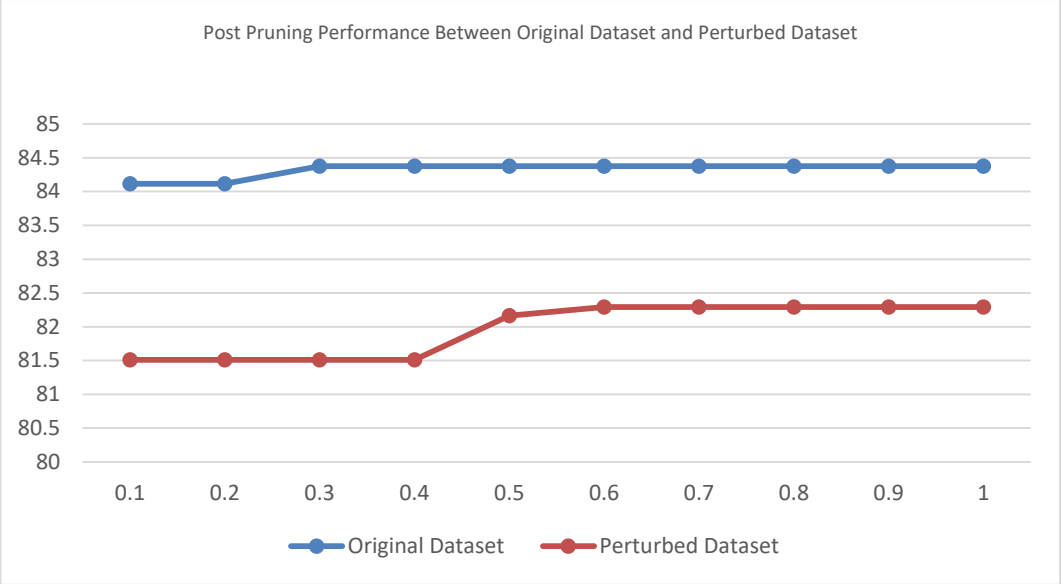

Figure 2 Post Pruning Performance Between Original Dataset and Perturbed Dataset

Figure 1 and 2 above, show the overall result of classification accuracy on J48 algorithm. While, Table 2 below shows comparison of proposed PPDM technique with other related works.

Table 2 Comparison of Mining Quality with Other Related Works

No.	Title	Accuracy (%)
1.	PPDM Based On Random Substitution	82.1615
2.	PPDM Based On K-Anonymity and Decision Tree Classification	82.4219
3.	PPDM Based On Cryptographic Based Using RSA Encryption On Health Dataset	77.3438

4.0 Discussion

In this study, RSP algorithm is successfully developed using MATLAB and J48 algorithm is implemented on WEKA. After performing randomized substitution perturbation, the result shows the number of restrictive patterns discovered from original data set is 1536. While the number of restrictive patterns discovered from sanitized data set is 106. So, the HF or the percentage of restrictive patterns that are discovered from sanitized data set is 6.90%. The randomization process has been performed for five times to get an average value of HF which is 5.87%. It can conclude that the performance of RSP algorithm is good since the percentage of it HF is close to 0%.

The classification accuracy result on perturbed dataset are quite close with the original datasets for both On-line Pruning and Post Pruning as shown in Figure 1 and 2 above. This is because, the perturbation process only happened to the instances in their respective columns and it does not modify the complexity of the dataset. There are some

matrices that affect the classification accuracy. In this study, there were no changes in the total number of nodes, number of leaves and tree depth which affect the classification accuracy. Besides that, the number of attributes that is perturbed also affect the classification accuracy obtained. In this study, only two attributes out of nine attributes are perturbed. So, it can conclude that data quality of proposed PPDM technique is good.

Finally, the result obtained is compared with other related works. This study is on the second place among other PPDM techniques where the different is only 0.2604% compared the first place as shown in Table 2 above.

5.0 Conclusion

Based on the result analysis on the conducted experiment, to get better performance in privacy preserving for data mining, several factors need to be considered. The analysis of the result obtained shown the proposed RSP algorithm provide better privacy level based on the Hiding Failure value obtained. Data quality for data mining process also shown better result. The accuracy of perturbed dataset from mining process quite close compared with the original dataset. It can be concluded that the proposed PPDM provide better and balance performance between privacy level and mining quality.

References

Agrawal, R., & Srikant, R. (2000, May). Privacy-preserving data mining. InACM Sigmod Record (Vol. 29, No. 2, pp. 439-450). ACM.

Aggarwal, C. C., & Philip, S. Y. (2004). A condensation approach to privacy preserving data mining. In Advances in Database Technology-EDBT 2004 (pp. 183-199). Springer Berlin Heidelberg.

Aggarwal, C. C., & Philip, S. Y. (2008). A general survey of privacy-preserving data mining models and algorithms (pp. 11-52). Springer US

Dowd, J., Xu, S., & Zhang, W. (2006). Privacy-preserving decision tree mining based on random substitutions. In Emerging Trends in Information and Communication Security (pp. 145-159). Springer Berlin Heidelberg.

Du, W., & Zhan, Z. (2002, December). Building decision tree classifier on private data. In Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14 (pp. 1-8). Australian Computer Society, Inc..

Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

Lindell, Y., & Pinkas, B. (2000, January). Privacy preserving data mining. InAdvances in Cryptology—CRYPTO 2000 (pp. 36-54). Springer Berlin Heidelberg.