PRIVACY PRESERVING DATA MINING USING ANONYMIZATION AND K-MEANS CLUSTERING ON LABOR DATASET

SAMAHAH SOLEHAH AHMAD ZAHARI

A project reportsubmitted in partial fulfilment of the
requirements for the award of the degree of
Master of Science (Information Security)

School of Computing
Faculty of Engineering
Universiti Teknologi Malaysia

JANUARY 2019

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. I wish to express my sincere appreciation to my main thesis supervisor, Dr. Maheyzah Md. Siraj, for encouragement, guidance, critics and friendship. I am also very thankful to my co-supervisor, Dr. Siti Hajar Othman for their guidance, advices and motivation. Without their continued support and interest, this thesis would not have been the same as presented here.

I am also indebted to Universiti Teknologi Malaysia (UTM) for funding my Master study. Librarians at UTM, and every staff in our faculty also deserve special thanks for their assistance in supplying the relevant materials.

My fellow postgraduate student should also be recognised for their support. My sincere appreciation also extends to all my colleagues and others who have helped me throughout this journey at various occasions. Their views and tips are very much useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family members for all emotional and financial supports.

ABSTRACT

Privacy Preserving Data Mining (PPDM) has recently become an important research area. There are some issues and problems related to PPDM have been identified. Information loss occur when the original of data are modified to keep the privacy of those data. Effects of PPDM also cause the level of data quality become lower. Aim of this research is to minimize information loss and increase the accuracy of mining result while maintaining the privacy level of data. A randomization approach based on anonymization and clustering algorithms are proposed in order to minimize the information loss and improve the accuracy of data clustering quality for PPDM results. Anonymization method is used in order to generalize and supress the data and limit the disclosure risk. Besides, the accuracy of data mining results could be increased by applying clustering using K-Means and EM algorithms. Labor dataset is used in this research and all instances are numerical value. WEKA tool is used to perform clustering algorithm on the labor dataset. Outcome for this research is the privacy level of dataset was increased while the information loss is minimized. The experimental results also show that the proposed method provides better result in privacy level of data mining.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

PPDM	-	Privacy Preserving Data Mining
CPU	-	Central Processing Unit
EM	-	Expectation Maximization
ME	-	Misclustering Error
UTM	-	Universiti Teknologi Malaysia
WEKA	-	Waikato Environment for Knowledge Analysis

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

Nowadays, technology has developed rapidly that it becomes really advanced in short amount of time. Technologies have become the heart of modern daily life, it acts as medium in helping people in their jobs. Every process nowadays done by using computational devices which process all the information and are stored digitally. Those information can be anything such as peoples' information, research progress, strategies, products' information, marketing information, and etc. which can be crucial to certain or several involved entities. Thus, they are stored with security provision in mind, which prevent them to be accessed by unauthorized person. Information and communication technology, high performance hardware components such as processor and storage capacity, rapid and efficient access to social networks, advanced database technologies, among other factors, have made it possible to generate high volume of data efficiently. Reduction in cost of hardware, computerization of many tasks and access to the technologies also contribute greatly to the efficient production of data.

Data, to be meaningful, need to be processed and analyzed to generate useful information out of it. The process of information generation from data is mostly applicable to small-scale data. As security itself become crucial with the advancement of technology, threats to them are also growing. It needs to be set in mind that providing security does not fully satisfied the safety of the information, some of the stored information itself needs to be shared or published for certain reasons. Those information/data then mined to achieve the purpose of their publication or sharing activities which results important information and knowledge. This process is called Data Mining. The act of data mining itself could lead an important knowledge or sensitive information to be stolen or breach with or without the owner knowing because the data source is out of the owner's control. Data mining is a domain that involves many methods from many areas such as statistics, database management, information retrieval, data warehousing, machine learning and pattern recognition (Patel and Donga, 2015). There are three basic data mining techniques, which are classification clustering and association rule.

Classification, as the name implies, involves assigning class value to an instance of database/dataset. Before the actual classification stage, learning stage should take place. At the learning stage, sample dataset, known as training set, with class labels for each instance is provided. Based on class labels of the training set, classifiers or models are generated. Next stage is the actual classification. Given an instance of dataset, whose class value is unknown, a classifier is employed to estimate appropriate class value to the instance. Classification is categorized as supervised learning because class labels are predefined in the training set (Patel and Donga, 2015). Clustering involves grouping of instances, based on similarities in their attribute values. Each group or cluster is given a description label. Unlike classification, clustering is categorized as unsupervised learning. This is because the clusters' labels are not predefined (Jain and Srivastava, 2013).

As the technology for generating data and mining it improves, threat to data security and privacy also increases, which creates data privacy concern. In the case of data mining, privacy preserving data mining PPDM, emerged as a general research area and measure to preserve privacy in data mining.

1.2 Problem Background

There are four main approaches used in privacy preserving data mining, which are data perturbation, condensation approach, anonymization, and cryptography. Each of the approaches has its own advantages and disadvantages. Such as data perturbation which protects against attribute disclosure and easy to implement but it provides less privacy and easier to breach (Kavitha and Vanathi, 2014). Condensation approach which provides better preservation but deemed as wasteful since the use of it no longer necessitates the redesign of data mining algorithm. Anonymization approach which is the most favorable approach that obfuscate data and protects against identity disclosure but not attribute disclosure. And also the cryptographic technique which offers a well-defined model for privacy but less on performance and difficult to scale.

Data perturbation is also called the randomization method in which it is used in the context of distorting the data (Aggarwal and Philip, 2008b), it can be done by using noise addition, multiplication or other more sophisticated approach (Traub et al., 1984; Adam and Worthmann, 1989). Perturbation method protects information from attribute disclosure, it is a simple approach. However, because its simplicity, this approach is easier be broken and provides less privacy (Kavitha and Vanathi, 2014). Whereas, Data anonymization is used to protect personal identities while releasing the truth information. It limit the disclosure risk to suitable level whereas maximizing data utility (Kavitha and Vanathi, 2014). In order to limit the risks of revealing the key attribute, Samarati (2001) and Sweeney (2002) introduce the k-anonymity model. It is a model in which stated that every record in an anonymized table to be indistinguishable with at least k other records in the dataset (Samarati, 2001; Sweeney, 2002). k-anonymity is the most technique used in data anonymization in Privacy Preserving Data Mining. It used the generalization and suppression for data anonymization while the information released are being truthful (Nayak and Devi, 2011). It gives the user anonymity for at least 1/k probability for an individual to be related to the real tuple.

The information given before address the problem in privacy level of k-anonymity method, k-Anonymity only provides protection against identity disclosure and not attribute disclosure whereas the perturbation methods provide opposite (Kavitha and Vanathi, 2014; Aggarwal and Philip, 2008b). A quasi-identifiers could also provide information because it uses a real values in range. Thus, a modified algorithm is proposed by perturbing the quasi-identifier attributes in k-anonymity algorithm in order to gives more protection of the information released. The perturbation method used is tree-based data perturbation method since it grouped the dataset into several subsets where it perturbed one attribute into its average value (Li and Sarkar, 2006). The grouping and resulting a same value is believes that it can help the anonymization process of k-anonymity since k-anonymization process grouped the dataset to minimal of k-1 similar value.

1.3 Problem Statement

There are some problems that rose during performing data mining process . Data mining process is a process to manage large database and pick the relevant information for some purposes. This process causes some information loss when transmitted through network medium. Some PPDM algorithms which have been introduced also reduce the data quality and accuracy of PPDM results. Labor dataset is important in organization since it contain all the account information of the company salary. Hence, the data privacy should be enhanced at all cost There are two main problems raised from large amount of dataset. Firstly, it is difficult to manage and easy to guess the correct value of dataset by intruders that lead to the privacy issues. Second is it is only some data needed to perform a task to select the only significant features and it decrease the accuracy level of PPDM results. The original values of data set are modified in order to preserve the data privacy which causes the damage on original values of data and decrease the accuracy of PPDM results.

1.4 Research Goal

The goal of this research is to minimize information loss and increase the mining result accuracy in term of privacy level and data clustering quality for privacy preserving data mining. In order to achieve this, enhanced PPDM approach was proposed by the anonymization of data with clustering algorithms where the dataset are firstly normalized then using against k-means and EM clustering algorithm.

1.5 Research Objective

In order to accomplish the aim of this project, a few objectives have been identified:

a) To apply anonymization technique for privacy preservation

b) To cluster the labor dataset using k-means and EM algorithm.

c) To evaluate the privacy preservation data mining (PPDM) using accuracy and data loss metrics.

1.6 Research Questions

In order to investigate these questions, the project considers the following questions to answer the above hypotheses:

a) How to minimize the information loss and increase the accuracy in PPDM?

b) How to normalize and generalize the dataset?

c) How to perform experiment of the proposed method and evaluate the performance?

1.7 Scope of Study

a) Scope is very important in any research and project as it can limit the area of research to a specific field and make sure that the research or project is not out from its early defined limitations. The scopes in this project are:

b) The dataset being used in the study are labor dataset which predict whether the income increase each year in 3 years duration.

c) The dataset is in term of numerical value.

d) The data mining software which is used for implementing the evaluation is Weka 3.8.3.

e) The process implementation and analysis are done using ARX Anonymization Tools and WEKA tool.

1.8 Significance of Study

The study is becoming significant due to the lack of privacy in data mining techniques. This research is to provide the optimal privacy preservation in data mining by minimizing the information loss during mining task and increasing the accuracy of mining result. This research was an enhanced PPDM approach which can minimize the information loss during mining task. The accuracy of data mining results can be increased while keeping the integrity of data. Hence, PPDM approach proposed can perform an optimal privacy preserving data mining for more sanitized dataset.

In this thesis, we have discussed several other ways of privacy preserve techniques. We discussed other existing methods to make data mining happen in a secure way. We have discussed these techniques in detail in Chapter 2 and their contribution in preserving privacy during data mining process is also included.

1.9 Structure of the Thesis

This project paper contains five chapters as follow:

Chapter two briefly explain about the databases and data mining, the nature of privacy preserving data mining, the existing methods in privacy preservation alongside with the weaknesses which reduce threats and create risks for the system. Then, the existing methods and solutions for the threats will be discussed and analyzed in order to find research gap where the significant of the study exist.

REFERENCES

Aldeen, Y., Salleh, M. & Razzaque, M. (2015). A comprehensive review on privacy preserving data mining. SpringerPlus, 4, 1-36.

Aggarwal, C. C. (2007, April). On randomization, public information and the curse of dimensionality. In Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on (pp. 136-145). IEEE.

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). An introduction to privacy-preserving data mining. In Privacy-Preserving Data Mining (pp. 1-9). Springer US.

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

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

Bhanumathi, S. and Sakthivel (2013). A New Model for Privacy Preserving Multiparty Collaborative Data Mining. 2013 International Conferences on Circuits, Power and Computing Technologies (ICCPCT-2013). IEEE, 845-850.

Blanton, M. (2011). Achieving Full Security in Privacy-Preserving Data Mining. 2011 IEEE International Conference on Privacy, Security, Risk and Trust, And IEEE International Conference on Social Computing. IEEE, 925-934.

Bora, S. P (2011). Data mining and ware housing. Electronics Computer Technology (ICECT), 2011 3rd International Conference on, 8-10 April 2011 2011. 1-5.

Chidambaranathan, S. (2014). A New Hybrid Algorithm for Privacy Preserving Data Mining. International Journal of Engineering Sciences & Research Technology. Vol. 3 (8), 147-156.

Coronel, C. & Morris, S. (2016). Database Systems: Design, Implementation, & Management, Cengage Learning.

Domingo-Ferrer, J., & Torra, V. (2001). A quantitative comparison of disclosure control methods for microdata. Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, 111-134.

Domingo-Ferrer, J., & Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control. Knowledge and Data Engineering, IEEE Transactions on, 14(1), 189-201.

Du, W. and Zhan, Z. (2003). Using Randomized Response Techniques for Privacy-Preserving Data Mining. Washington, DC, USA: Copyright 2003 ACM 1-58113-737-0/03/0008.

Grandison, T. and Evfimievski, A. (2009). Privacy-Preserving Data Mining. Retrieved on 12 October 2014 from http://www.tyronegrandison.org/uploads/1/8/8/1/18817082/ppdm_encyclopedia.pdf.

Jain, D., Khatri, P., Soni, R. & Chaurasia, B. K. (2012). Hiding sensitive association rules without altering the support of sensitive item(s). Advances in Computer Science and Information Technology. Networks and Communications. Springer.

Jain, N. & Srivastava, V. (2013). Data Mining techniques: A survey paper. IJRET: International Journal of Research in Engineering and Technology, 2, 2319-1163.

Jain, Y. K., Yadav, V. K. & Panday, G. S. (2011). An efficient association rule hiding algorithm for privacy preserving data mining. International Journal on Computer Science and Engineering, 3, 2792-2798.

Jyoti, D. R. (2015). A Review on Privacy Preserving Data Mining. International Journal of Emerging Trends in Science and Technology, 2.

Kavitha, M. R., & Vanathi, D. (2014). A Study of Privacy Preserving Data Mining Techniques. International Journal, 3(4).

Li, X. B., & Sarkar, S. (2006). A tree-based data perturbation approach for privacy-preserving data mining. Knowledge and Data Engineering, IEEE Transactions on, 18(9), 1278-1283.

Mandapati, S., Bhogapathi, R. B. and Chekka, R. B. (2013). A Hybrid Algorithm for Privacy Preserving in Data Mining. I.J.Computer Network and Information Security, 2013, 08, 47-53.

Matwin, S. (2013). Privacy-Preserving Data Mining Techniques: Survey and Challenges. In: CUSTERS, B., CALDERS, T., SCHERMER, B. & ZARSKY, T. (eds.) Discrimination and Privacy in the Information Society. Springer Berlin Heidelberg.

Mivule, K. (2013). Utilizing Noise Addition for Data Privacy, an Overview. Retrieved on 15 October 2014 from http://arxiv.org/ftp/arxiv/papers/1309/1309.3958.pdf.

Mynavathi, R., Sowmiya, N. and Vanitha, D. (2014). Survey of Various Techniques to Provide Multilevel Trust in Privacy Preserving Data Mining. International Journal of Innovative Research in Computer and Communication Engineering. Vol. 2, (Special Issue 1), 118-123.

Natarajan, R., Sugumar, D. R., Mahendran, M. & Anbazhagan, K. (2012). A survey on Privacy Preserving Data Mining. International Journal of Advanced Research in Computer and Communication Engineering, 1.

Nayak, G., & Devi, S. (2011). A survey on privacy preserving data mining: approaches and Techniques. International Journal of Engineering Science and Technology (IJEST), 3(3), 2117-2133.

Nithi and Maheyzah. (2015). K-Anonymity Data Obfuscation and Tree-Based Data Perturbation in Privacy Preserving Data Mining. Master of Computer Science (Information Security), Universiti Teknologi Malaysia.

Parmar, S., Gupta, M. P. & Sharma, M. P. (2015). A Comparative Study and Literature Survey on Privacy Preserving Data Mining Techniques.

Patel, K. & Donga, J. (2015). Practical Approaches: A Survey on Data Mining Practical Tools. Foundations, 2.

Patel, M., Richhariya, P. and Shrivastava, A. (2014). A Novel Approach for Data Mining Clustering Technique Using Neural Gas. 2014 Fourth International Conference on Advanced Computing & Communication Technologies.

Ple, R. and Stephens, R. (2003). The Database Normalization Process. Retrieved on 1 November 2014 from http://www.informit.com/articles/article.aspx?p=30646.

Prakash and Singaravel (2014). A Review on Approaches, Techniques and Research Challenges in Privacy Preserving Data Mining. Australian Journal of Basic and Applied Sciences, 8(10) July 2014, Pages: 251-259.

Reddy, O. H. & Singh, P. (2015). Preserving Privacy in Data Mining by Data Perturbation Technique. International Journal, 4.

Saranya, K., Premalatha, K. & Rajasekar, S. S (2015). A survey on privacy preserving data mining. Electronics and Communication Systems (ICECS), 2015 2nd International Conference on, 26-27 Feb. 2015. 1740-1744.

Sharma, M., Chaudhary, A., Mathuria, M., Chaudhary, S. and Kumar, S. (2014). An Efficient Approach for Privacy Preserving in Data Mining. IEEE 978-1-4799-3140-8.

Shorayha and Maheyzah. (2015). Integration of PSO and Clustering Algorithms for Privacy Preserving Data Mining. Master Master, Universiti Teknologi Malaysia.

Sukhdev, S. & Vasava, H. Privacy Preserving Data Mining With Classification And Encryption Methods.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

Vaidya, J., Clifton, C. W. & Zhu, Y. M. (2006). Privacy preserving data mining, Springer Science & Business Media.

Vaidya, J., Kantarcıoğlu, M. & Clifton, C. (2008). Privacy-preserving naive bayes classification. The VLDB Journal—The International Journal on Very Large Data Bases, 17, 879-898.

Vaidya, J., Shafiq, B., Basu, A. & Hong, Y (2013). Differentially private naive bayes classification. Web Intelligence (WI) and Intelligent Agent Technologies (IAT), IEEE/WIC/ACM International Joint Conferences on, 2013. IEEE, 571-576.

Waikato, T. U. O. (2016). Weka 3: Data Mining Software in Java [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/ [Accessed 07/04/2016].

Wikipeadia. (2015). Certified Quality Auditor (CQA) [Online]. Wikipeadia. Available: https://en.wikipedia.org/wiki/Certified_Quality_Auditor [Accessed 09/ 10/ 2016].

Xu, L., Jiang, C., Wang, J., Yuan, J. & Ren, Y. (2014). Information Security in Big Data: Privacy and Data Mining. Access, IEEE, 2, 1149-1176.

Yang, L., Wu, J., Peng, L. and Liu, F. (2014). Privacy-Preserving Data Mining Algorithm Based on Modified Particle Swarm Optimization. In D.S. Huang et al. (Eds): ICIC 2014, LNAI 8589, pp. 529-541. Springer International Publishing Switzerland

Yi, X. & Zhang, Y. (2013). Equally contributory privacy-preserving k-means clustering over vertically partitioned data. Information Systems, 38, 97-107.