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

Privacy Preserving Decision Tree Classification Using RSA Encryption On Health Dataset

Ikmal Faizadh Roszaind[1], Maheyzah Md. Siraj*[2]

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

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

Abstract

Due to the advance in storage capacity, the capability to store a huge amount of data had been relatively increased exponentially, aid with the breakthrough of the information processing technology, large amount of databases can be processed in only a short amount of time compared to previous years. Data mining techniques, such as classification are often used on data to extract hidden information, which is why data mining also can be called as knowledge discovery. Unfortunately, there is a probability that during the event of data mining, the information can be exposes to several parties and this will lead to the breach of individual privacy. However, this problem can be solved by applying privacy preserving method during the event of data mining, which can be achieved through several techniques such as Perturbation, Randomization and Cryptographic. This paper discusses the cryptographic approach to privacy preserving data mining in providing confidentiality during the event of data mining. The objective of this research is to implement a cryptographic technique on dataset to make it encrypted for preserving privacy, using decision tree classification algorithm for data mining on health dataset and to evaluate and benchmark the proposed approach using quantification measurement in term of privacy level and classification accuracy to other data mining algorithm. The type of encryption that was used is RSA encryption, an asymmetric encryption. The result will then be measured by specific parameters measurements, which are the privacy level and mining quality (data loss).

Keywords: Data Mining, Privacy, Encryption, Decision Tree Classification

1.0 Introduction

Data can be view as raw information about an individual or an organization such as your medical information, credit card purchases, mails or emails, phone calls or even you social medias activities. This information is very valuable to certain organization, as it will benefit them in their field of business. The issues later that came with this facts is the privacy breach whether to an individual or an organization. Medical information, usage of credit cards and such are considered personal, which means only the users or the owners of those that can know about it. There is this term called data mining or knowledge discovery, which can be seen as a process of extracting data from large databases. This new knowledge then will be analyses including identifying patterns and relations between the knowledge and then, new

useful information can be concluded. These data sets typically contain sensitive individual information, which consequently get exposed to the other parties. In Privacy Preserving Data Mining (PPDM), we have to ensure that data privacy is maintained in the event of data mining.

There are different kinds of strategy that can be implemented to protect the data privacy during decision tree analysis of data mining process, and one of them is by cryptography. The attributes of the information are encrypted for privacy preservation. The encrypted data then is presented to the second party for decision tree analysis. The decision tree obtained on the original data and the obfuscated data are similar but by using this approach, the data proper is not revealed to the second party during the mining process and hence the privacy will be preserved.

Objectives of the project are : (i) to implement a cryptographic technique on dataset to make it encrypted for preserving privacy, (ii) to use a decision tree classification algorithm for data mining on health dataset, and (iii) to evaluate and benchmark the proposed approach using quantification measurement in term of privacy level and classification accuracy.

2.0 Methodology

This chapter will discuss the research methodology that will be used in researching the proposed method. This chapter will describe the steps that will be implement in the undergoing study that will be done in three phases. The main purpose of this research is to propose a method in which to apply a privacy preserving decision tree mining based on cryptography method on health data. The proposed method will then be compared to the current method available based on their efficiency and accuracy metrics. There are three phases that will be conducted in this research:

a) RSA Encryption As Privacy Preserving Method

First phase started with problem identification and will end with the analysis of the problem and a proposed solution. Initiation phase is made of two parts, the literature review and problem formulation. Literature review was completed by studying and analyzing the related work that are in the same research area. This include with understanding the models and algorithm of data mining, the distinctive properties that each one of the algorithm had and analyzing the approaches of the privacy preserving techniques. For this research, we focus on the cryptographic-based technique and decision tree based classification. The strengths, advantages and disadvantages of the PPDM algorithms over other approaches are also stated.

b) Implementation of Decision Tree Classification Algorithm

Phase II is about analysis, design and implementation. The output for this phase is a Java application, which is used to perform noise addition scheme mechanism with decision tree classification. This phase will include the analysis and the designing of the application and then it will be develop into usable java application. For the decision tree process, we will use the C4.5 algorithm proposed by J. Ross Quinlan, (1993). It will be used on the health data sets and will run on Weka software.

c) Evaluation and Benchmarking

The final phase in this research framework will be the evaluating and benchmarking the proposed PPDM algorithm. As mention is the literature review, we will evaluate the performance based on two things, the privacy level and mining quality (data loss). The

proposed algorithm, which is using Decision Tree Classification, will be benchmarked with other approaches, which are Random Substitution and K-anonymity. The data that will be used as target material for the develop application will be dataset from National Institute of Diabetes and Digestive and Kidney Disease in which the dataset are the Pima Indians Diabetes Database.

3.0 Result

There are many ways to benchmark PPDM algorithm, one of them is by measuring the privacy level. This can be done by calculating its hiding failure. Hiding failure were obtain from the percentage of restrictive pattern that are discovered from the sanitized database. It is measure as follow

$$HF = \frac{\#R_P(D')}{\#R_P(D)} \qquad (4.1)$$

$\#R_P(D)$ and $\#R_P(D')$ denote the number of restrictive patterns that can be identified from the original database *D* and the sanitized database *D'* respectively. Using the dataset that had been modified using the proposed method, the percentage for quatifying hiding failure was 10%. As we can see, the restrictive pattern discovered from the original database is 10% which is quite good. As Bertino *et al*, (2008) stated the ideal hiding failure is 0. The closer the percentage to 0, the better the data hiding failure parameter. This cryptographic based approach had shown that even it was not as good as under 5% for the hiding failure, it still did a good job in hiding sensitive information.

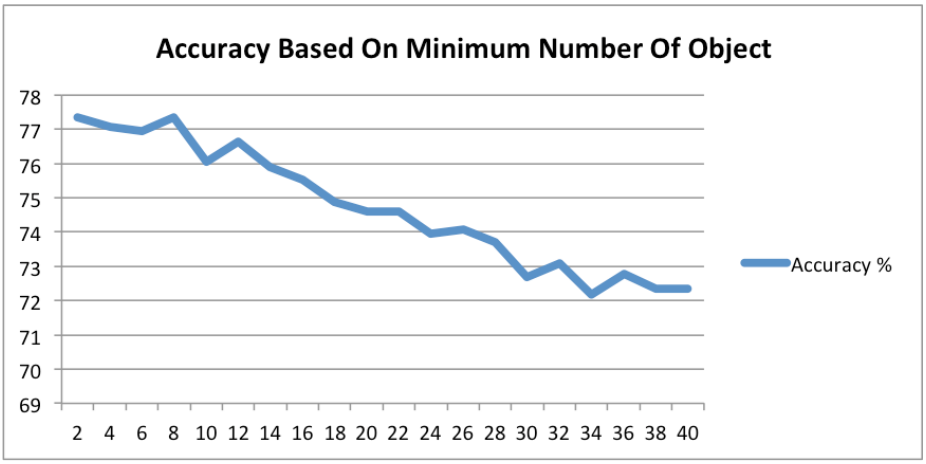

Figure 1 Accuracy Based On Minimun Number Of Object

Table 1: Comparison Of Accuracy on Data Mining Techniques

	Decision Tree Classification	**Logistic Regression**	**Support Vector Machine**
Accuracy	77.3438%	68.8802%	65.1042%

Table 2: Comparison of Mining Quality With Other Related Works.

Works	Accuracy
PPDM Based On K-Anonymity and Decision Tree Classification	82.4219%
PPDM Based On Random Substitution	82.1615%
PPDM Based On Cryptographic Method Using RSA Encryption	77.3438%

4.0 Discussion

The graph in Figure 1 above described the quality result of the data mining when proposed privacy preserving are applied. We can see that the accuracy of the result were somewhat constant with the highest accuracy was 77.3438% while the lowest was 72.3508% with the minimum number of object a set as manipulation variable. This show that proposed technique for privacy preserving technique did quite well in preserving the information but at the same time did not jeopardized the knowledge discovery process. As seen in Table 1 , we can see that Decision Tree Classification score the highest accuracy compared to other data mining techniques which are Logistic Regression and Support Vector Machine. This proved that the proposed technique is the most effective technique for conducting data mining purposes compared to the other two. Based on the comparison in Table 2, cryptography based approach score the lowest accuracy. This is because the individual itself data was modified, thus integrity of the data was changed and leads to the decrease of the accuracy during data mining. However, since cryptographic approach encrypted all data, we can conclude that the privacy level of this approach is relatively better compared to other works presented in the table above and more secure as the data was fully change from plain text to encrypted and cannot be read and understand unless the person has the key to decrypt it.

5.0 Conclusion

The major motive of this research is to discuss a privacy preservation technique that is meant to protect information during the data mining event. For this project the method used for privacy preserving is cryptographic method using RSA encryption. While cryptographic based techniques is well known to offer security privacy over sensitive information, this research project had proved to what extent this techniques can be used to preserved the privacy in the data mining process.

References

Rissman, J., Greely, H. T., & Wagner, A. D. (2010). Detecting individual memories through the neural decoding of memory states and past experience. Proceedings of the National Academy of Sciences, USA, 107, 9849-9854. doi:10.1073/pnas.1001028107

Aggarwal, C. C., & Philip, S. Y. (2008). A general survey of privacy-preserving data mining models and algorithms (pp. 11-52). Springer US.Bertino, E., Lin, D., & Jiang, W. (2008). A survey of quantification of privacy preserving data mining algorithms. In Privacy-preserving data mining (pp. 183-205). Springer US.

Domingo-Ferrer, J. (2008). A survey of inference control methods for privacy-preserving data mining. In Privacy-preserving data mining (pp. 53-80). Springer US. Han, J. (2002). How can data mining help bio-data analysis?. In BIOKDD (pp. 1-2).

Islam, M. Z. (2007). Privacy preservation in data mining through noise addition (Doctoral dissertation, University of Newcastle). Islam, M. Z., & Brankovic, L. (2003). Noise addition for protecting privacy in data mining. In Proceedings of The 6th Engineering Mathematics and Applications Conference (EMAC2003), Sydney (pp. 85-90).

Jans, M., Lybaert, N., & Vanhoof, K. (2007). Data mining for fraud detection: Toward an improvement on internal control systems? Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. Journal of healthcare information management, 19(2), 65. Mivule, K. (2013). Utilizing noise addition for data privacy, an overview. arXiv preprint arXiv:1309.3958.

Rao, K. S., & Rao, B. S. An Insight in to Privacy Preserving Data Mining Methods.

Qi, X., & Zong, M. (2012). An overview of privacy preserving data mining. Procedia Environmental Sciences, 12, 1341-1347.

Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier

.