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 K-Anonymity and Decision Tree Classification

Abdul 'Azim Mohammad[1], Maheyzah Md. Siraj*[2]

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

[1]abdul.azim93@outlook.com, [2]maheyzah@utm.my

Abstract

Nowadays, there are an extensive amount of data stored in databases and other repositories. This is including the sensitive and confidential data which is needed to be protected and preserved the individual's sensitive identity without sacrificing the usability of data. These sensitive and private information are exposed to unauthorized personnel which lead to the information leaking and lead to the misuse of these data. Because of that, privacy preserving data mining was introduced to overcome these problems. Privacy preserving data mining has become popular as it the privacy level of the sensitive and confidential data must be protected from the unauthorized personnel. The purpose of this research is to develop an anonymization algorithm of privacy preserving technique with decision tree classification approach. This paper is mainly focuses on the k-anonymity techniques where it provides protection against identity disclosure. The k-anonymity technique used generalization and suppression method for achieving data anonymization. The privacy level and mining quality of the anonymized dataset then will be test using decision tree classification and then later compared with the other data mining technique which are logistic regression and support vector machine. The result of the research proves that the privacy level and data quality provides a better result compared to the other data mining technique.

Keywords: Privacy preserving, Data mining, Anonymization

1.0 Introduction

In this digitalization era, the technology expands at high rate from time to time. With technology development there are a lot of advantages and it helps our life easier. Existence of devices helps human works for daily basis and storing information. There are a lot of sensitive information stored within the device or cloud and this information need to be protected. Therefore, privacy of the data needs to be preserved in order to prevent the sensitive data from being misused by irresponsible people.

Data mining is crucial because its extract or mining knowledge from large amount of data. With the massive amount of data in database, it is important to develop an analysis for extracting data. There are a lot of possibilities where the stored data received threats.

Thus, privacy preserving data mining (PPDM) was developed in order to overcome with these threats.

Awareness of people makes security of the sensitive data need to be protected. Thus, PPDM is introduced as it one of the important research area that needs to be develop because it still in infancy stage. PPDM is a process of extraction of data from large database while it protects the sensitive data at the same time. PPDM hides individual sensitive data without sacrificing the usability of the data. There are several techniques of PPDM that have been developed by researchers and this numbers expected to grow to significant numbers as it is promising and being potential field.

Objectives of this project are : (1) to develop a k-anonymity algorithm as data anonymization for preserving privacy in data, (2) to implement decision tree classification algorithm for data mining on health dataset, and (3) to evaluate and benchmark the proposed approach using quantification measurements in terms of privacy level and mining quality.

2.0 Methodology

a) Phase 1: Privacy Preserving

The first phase is the dataset must first be anonymized. The anonymization process is done by using the ARX Anonymization Tool. By using this tool, the attribute selected for the anonymization process which is age, no of pregnant and class. The anonymization process uses generalization technique where the attribute of the selected dataset is generalized so that the anonymity of the data can be produced. The result collected after the anonymization process is the hiding failure after the data has been anonymized.

b) Phase 2: Data Mining

The second phase is the data mining process. In this research the data mining that used is decision tree classification. This data mining technique was introduced by Quinlan in his C4.5 book. The equation used is J48 equation. The data mining process is done by using WEKA tool. This tool is very suitable for the data mining process.

c) Phase 3: Privacy level & Mining Quality

The last phase is the quantification based on the privacy level and mining quality. The privacy level that used for quantification measurement is the hiding failure. The mining quality that is used for quantification is the accuracy. The accuracy of the anonymized dataset is measured by using based on the correctly classified instance.

3.0 Result

Process of anonymization of the dataset is the first phase of the operational framework in this research. In this phase, the pima Indian diabetes dataset is used as the data input and anonymized using k-anonymity technique. This technique is run using ARX, which is an efficient open source data anonymization framework that implemented in Java (ARX,2013).

In this phase, the quasi-identifier of the dataset must be recognized in order to choose what is the attributes that is suitable for anonymized. In this research, the attributes age,

number of pregnancy and class. In order to apply anonymization, each attribute K-anonymity technique is applied each attribute. The result is on the figure 5.1 below.

Table 1: The Anonymized Diabetes Dataset

preg	plas	pres	skin	insu	mass	pedi	age	class
[0,6]	148	72	35	0	33.6	0.627	[0,40}	[0,1]
[0,3]	85	66	29	0	26.6	0.351	[0,20]	[0,1]
[0,12]	183	64	0	0	23.3	0.672	[0,20]	[0,1]
[0,3]	89	66	23	94	28.1	0.167	[0,20]	[0,1]
[0,3]	137	40	35	168	43.1	2.288	[0,20]	[0,1]
[0,6]	116	74	0	0	25.6	0.201	[0,20]	[0,1]
[0,3]	78	50	32	88	31	0.248	[0,20]	[0,1]
[0,12]	115	0	0	0	35.3	0.134	[0,20]	[0,1]
[0,3]	197	70	45	543	30.5	0.158	[0,40}	[0,1]
[0,12]	125	96	0	0	0	0.232	[0,40}	[0,1]

The quantification of accuracy is the measure on how accurate the anonymized dataset compared to the original dataset. The higher the value of accuracy, the higher the proximity of the anonymized dataset to the original data. Accuracy of the anonymized dataset is really important as it show the value of usability of the dataset.

Based on the WEKA, the J48 decision tree algorithm have a few parameters that need to set in order to perform calculation of the accuracy of the data. The value of parameter that matter the most in the determination of the accuracy are minimum number of object and confidence factor. The minimum number of object is set from 2 until 40 and its increasing by 2 each time. The confidence factor level is set from 0.1 to 1.0 by increasing by 0.1 each time. The table 5.8 and 5.9 show the detail of the accuracy of the anonymized dataset.

Table 2: Accuracy of The DTC Data Mining Based on Minimum Number of Object

No.	confidenceFactor	minNumObj	Accuracy (%)
1	1	2	82.4219
2	1	4	80.3385
3	1	6	79.4271
4	1	8	78.5156
5	1	10	77.8646
6	1	12	78.6458
7	1	14	78.125
8	1	16	77.9948
9	1	18	77.3438
10	1	20	77.3438

Based on the HF of the dataset, the result shown that the anonymized datasets yield a slightly better result in hiding function as the percentage of HF which are around 86 to 87 percent. Although the hiding failure is quite high, but it still produces a little increase in term of hiding failure. Hence, it can be concluded that the privacy level of dataset can be increased by anonymizing the dataset by applying k-anonymity technique.

Table 3: Hiding Failure of the Anonymized Dataset

No. of k	Anonymized Database	Original Database	Hiding Failure (%)
10	85.60811	98.04067	87.31898
20	85.03044	98.04067	86.72976
30	86.00767	98.04067	87.72652
40	85.58478	98.04067	87.29518
50	86.04389	98.04067	87.76347

4.0 Discussion

In this section, the comparison with the other data mining technique must be done in order to find out what is the best result in terms of accuracy. The other mining technique such as Support Vector Machine(SVM) and Logistic Regression(LR) also must be test to the anonymized dataset in order to test its accuracy. Therefore, the table below show the comparison of the LR, SVM and DTC.

Table 4: Comparison of LR, SVM and DTC.

Data Mining Technique	Accuracy
DTC	82.4219%
LR	77.9948%
SVM	75.7813%

In this section, the comparison with the other PPDM work is discussed. The other PPDM works are Privacy Preserving Data Mining Based on Random Substitution and Decision Tree Classification (Shafiq, 2016) and Privacy Preserving Data Mining Based on Cryptographic Based Using RSA Encryption and Decision Tree Classification (Ikmal, 2016). The value that compared with the other works is the accuracy after the decision tree classification data mining. The table 5.10 show the result of the comparison.

Table 5: Comparison of Other Works.

Works	Accuracy
This research	82.4219%
Shafiq(2016)	82.1615%
Ikmal(2016)	77.3438%

5.0 Conclusion

Based on this research experiment, it can be concluded that the privacy preserving technique specifically k-anonymity technique does offer better result for PPDM approach. The privacy of data after applying anonymization technique also proven yield better result in term of low number of percentage of hidden failure. The potential of sensitive data

disclosure also can be reduced by applying anonymization technique. Data mining process which has been conducted in this research also produce a better result in term of mining quality. The mining quality after applying the PPDM technique also increase as the number of accuracy also increased.

References

Aggarwal, C. C. and Yu, P. S. (2008). A General Survey of Privacy-Preserving Data Mining Models and Algorithms. Privacy-Preserving Data Mining. (11-52). US: Springer US.

Liu, L., Kantarcioglu, M., & Thuraisingham, B. (2008). The applicability of the perturbation based privacy preserving data mining for real-world data. Data & Knowledge Engineering, 65(1), 5-21.

Bertino, E., Lin, D. and Jiang, W. (2008). A Survey of Quantification of Privacy Preserving Data Mining Algorithms. Privacy-Preserving Data Mining. (183-205). US: Springer US.

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

Roughan, M., & Zhang, Y. (2006, September). Privacy-preserving performance measurements. In Proceedings of the 2006 SIGCOMM workshop on Mining network data (pp. 329-334). ACM.

Taneja, S., Khanna, S., Tilwalia, S., & Ankita (2014). A Review on Privacy Preserving Data Mining: Techniques and Research Challenges. International Journal of Computer Science and Information Technologies (IJCSIT), Vol 5(2), 2310-2315.

Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., & Theodoridis, Y. (2004). State-of-the-art in privacy preserving data mining. ACM Sigmod Record, 33(1), 50-57.