Journal of Engineering and Applied Sciences 12 (8): 2138-2142, 2017
ISSN: 1816-949X

Exploring Utility of Extended Misusability Measure for Data Publications

[1]J. Pradeep Kumar, [2]A. Udaya Kumar and [3]T. Ravi
[1]Department of Computer Science Engineering, Aditya College of Engineering, Madanapalle, India
[2]Hindustan Institute of Technology and Science, Chennai, India
[3]Srinivasa Institute of Engineering and Technology, Chennai, India

Abstract: When data is published in the real world it is essential to ensure that privacy is not disclosed and the data is not misused. In our study earlier we proposed an extended misusability measure that helps in finding the probability of misuse of given dataset. The measure takes single or multiple publications as input and generates misusability score. This score determines the level of misusability possible with the given dataset. The misusability leads to possible disclosure of privacy. By computing misusability score, it is possible to anonymize sensitive attributes to achieve privacy preserving data publications and data mining as well. In this study, our aim is to demonstrate the real utility of our extended misusability measure. We proposed a framework with an underlying algorithm to sanitize data before publishing it or before it is subjected to mining. The proposed algorithm employs the measure and determines the need for sanitizing datasets. The algorithm in turn uses K-anonymity which one of the standard sanitization algorithms for preventing privacy attacks on the datasets. We built a prototype application that demonstrates the proof of concept. The empirical results revealed that our misusability measure has significant impact on the privacy preserving data publishing and privacy preserving data mining.

Key words: Anonymization, misusability measure, privacy preserving knowledge discovery, privacy preserving data publishing, significant

INTRODUCTION

Enterprises in the real world have been employing data mining techniques in order to obtain business intelligence to leverage business growth faster. Some organizations may also publish data for public use. In the former case, when data is outsourced to third party for mining, there is possibility of privacy attacks by malicious insiders. In the latter case also there is possibility of inference attacks to exploit privacy or sensitive details in the datasets. Therefore, it is essential to have a measure to know whether the data being published or given to third party is misusable. In other words, it is important to have a mis-usability measure so as to determine the level of sanitization of data. In our previous research (Kumar *et al*., 2016), we proposed an extended misusability measure which takes datasets as input and produce misusability score. The score is between 0.0 and 1.0. The more in score is the more in probability of misuse of data.

In this study our focus is on evaluating the utility of our extended misusability measure in order to ensure privacy preserving data publishing or privacy preserving data publishing. We proposed a framework and an algorithm to ensure this. We built a prototype application that demonstrates the proof of concept. The empirical results revealed that our misusability measure has significant impact on the privacy preserving data publishing and privacy preserving data mining.

Literature review: Matatov *et al*. (2010) explored feature set partitioning approach for PPDM. They tried to achieve the goal of data mining while preserving privacy to data. They proposed a new method known as Data Mining Privacy by Decomposition (DMPD). It is a genetic algorithm that that makes use of operators and fitness function for feature set partitioning and achieving PPDM. Turner *et al*. (2010) explored big data mining on genetic data using a hybrid approach. Brankovic and Castro (2011) focused on privacy issues while discovering knowledge through data mining process. They identified the following privacy issues:

- Disclosure of sensitive data
- Personal information used for secondary purposes
- Misusing data
- Gaining access to personal information at finer granularity
- Combination of patterns
- Identity disclosure from training sets
- Internal attackers
- Stereotypes

Corresponding Author: J. Pradeep Kumar, Department of Computer Science Engineering, Aditya College of Engineering, Madanapalle, India

Islam and Brankovic (2011) proposed a PPDM technique based on noise addition and clustering approaches. To this effect the proposed a framework and employed noise addition techniques such as perturbation of class attributes, perturbation of non-class numerical attributes, perturbation of categorical attributes and random categorical technique. Guo *et al.* (2012) explored knowledge discovery in the form of spatial patterns from mobility data. Spatial patterns can provide knowledge in both temporal and spatial domains.

Giannotti *et al.* (2013) explored privacy preserving association rule mining from the data that has been outsourced. They proposed mining as a service concept and related architecture for Privacy Preserving Data Mining (PPDM). They proposed an attack model and counter measure for achieving PPDM. Waqar *et al.* (2013) proposed a technique known as dynamic reconstruction of metadata to preserve privacy of cloud users. Meta data refers to the data about data. Their technique focuses on the meta data construction to find misuse of data. Cooley and Smith (2013) focused on usability of IT. They proposed an approach for privacy preserving screen capture with respect to health IT. Perera focused on Internet of Things (IoT) which is the recent happening. With IoT they focused on privacy knowledge modelling. They describe how privacy modelling can be done in the context of IoT.

Dong *et al.* (2014) focused on a privacy preserving data sharing approach that is employed in cloud computing environment. Thus, they made the data sharing approach dependable and secure. They employed a technique known as Cipher text Policy Attribute Based Encryption (CP-ABE) in combination with other technique known as Identity Based Encryption (IBE). They proposed a system model and adversary model in order to evaluate their approach. Oksanen *et al.* (2015) explored privacy preservation in mobile sports data. They proposed a framework with many activities and at each level they introduced an awareness of privacy. Taneja and Singh (2015) focused on Electronic Medical Records (EMR) in health care domain for privacy preserving knowledge discovery. They used re-identification risk concept in order to strengthen their proposed work. Zhang *et al.* (2015) focused on feature learning on big data in privacy preserving fashion. They used a deep computation model for achieving this. They offloaded expensive operations to cloud while performing non resource intensive tasks in the local machine as part of privacy preserving deep computation.

Victor *et al.* (2016) made a good survey on the privacy models for big data. They focused on two aspects of privacy. They are privacy preserving data mining and privacy preserving data publishing. The first case is data is outsourced to third party which needs privacy while the second case is data is released to public use that needs privacy. In the privacy models they identified four kinds of attributes known as explicit identifiers that identify records uniquely, quasi identifiers that may be used to have inference attacks, sensitive attributes containing personal information and non-sensitive attributes that are not sensitive and can be directly disclosed to third parties. They also explored anonymization techniques such as generalization and suppression, anatomization and perturbation and perturbation.

Chen *et al.* (2016) made good survey of big data science and big data analytics. Their work mainly focuses on identifying research studys that are meant for big data science and big data analytics. Jiang *et al.* (2016) focused on e-Health clouds with three-factor authentication which is privacy preserving. The phases involved in their approach include initialization phase, registration phase and login phase. These phases are carried out with privacy preserving authentication mechanism in place. Li *et al.* (2016) focused on vertically partitioned databases for mining. They employed a technique to preserve privacy while performing data mining. In fact their focus was on privacy preserving association rule mining. Kawamoto (2016) focused on stream integration system with privacy preserved. As the sensor networks and WWS are producing data streams, they focused on building a model that caters to the needs of data streams with privacy preserving process. Yan *et al.* (2016) explored social networks for privacy preserving data mining. The techniques they proposed are distance grained differential privacy and item-grained differential privacy. They evaluated their techniques with collusion attacks to ensure that their technique preserves privacy. Dara and Muralidhara focused on collaborative intrusion detection and its related techniques while preserving privacy. They used the notion of advanced persistent threats and used global intelligence to achieve secure and privacy preserving intrusion detection.

MATERIALS AND METHODS

Proposed framework: We proposed a framework that is meant for evaluating the utility of our misuability measure. The extended misusability measure proposed by us in our previous research (Kumar *et al.*, 2016) is evaluated by using the framework. The purpose of the framework is to generalize the utility of the measure which helps in protecting data while publishing or while giving it to third parties for data mining purposes. As shown in Fig. 1, the framework takes dataset as input and computes misusability measure. More details on computing misusability measure are found in our study (Kumar *et al.*, 2016). However, some details of the measure are provided in this section. After computing the measure it is possible to determine whether the dataset has any sensitive fields and needs to be anonymized for ensuring non-disclosure

of privacy from the dataset. Based on the threshold for the misusability measure a sanitization technique named K-Anonymity is applied on the dataset in order to protect it from misusability.

Algorithm

Utility algorithm for extended misusability measure:

Initialize dataset D
Populate D with given dataset
Initialize misusability threshold mt
Populate mt with domain specific value provided by domain expert
Compute misusability score ms as explored by Kumar *et al.* (2016)
IF ms>mt THEN
Sanitize dataset using K-Anonymity
END IF

The proposed algorithm is meant for sanitizing datasets based on the misusablity score which is briefly explained in the ensuing sub section. Once misusability score is computed, the score is used to know whether sanitization is required based on domain expert input known as a threshold. Based on the threshold value it is determined whether K-Anonymity is needed or not.

Extended misusability measure (M-Score): This is the measure, we proposed in our previous study (Kumar *et al.*, 2016). It is used to determine whether a dataset is misusable and the probability of misusability. The equation that makes use of sensitivity function to compute raw record score is as follows:

$$RRS_i = \min\left(1, \sum_{S_j \in T} f\left(c, S_j[x_i]\right)\right)$$

Then, the Distinguishing Factor (DF) of the record is calculated. This will help as a measure to understand how the identity of a record is exposed by quasi identifiers. The computation of final record score is done as:

$$RS = \max_{0 \le i \le} \left(RS_i\right) = \max_{0 \le i \le} \left(\frac{RRS_i}{D_i}\right)$$

The result of record score is substituted in the formula that is meant for computing misusability as:

$$MScore = r^{1/x} \times RS = r^{1/x} \times \max_{0 \le i \le r} \left(\frac{RRS_i}{D_i}\right)$$

There is some problem with misusability measure as its value is unbounded. For this reason, it is essential to normalize, it that gives standard measure value. After normalization the misusability measure will be between 0.0 and 1.0.

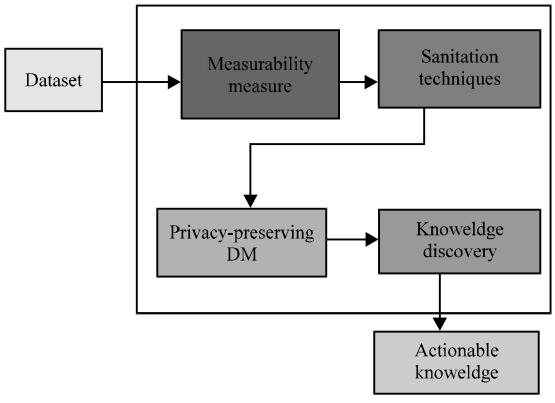

Fig. 1: Framework for application of extended M-Score

RESULTS AND DISCUSSION

Adult dataset is taken from UCI machine learning repository. The dataset contains anonymous census data of 32562 individuals from the 1990 US census. Table 1 shows the adult dataset after random sampling.

As shown in Table 1, the adult dataset chosen contains attributes like name, age, sex, zip code and disease. After applying the proposed algorithm using K-Anonymity the results is as given in Table 2. As shown in Table 2, the results of anonymization are presented.

Table 1: Adult dataset with random sampling

Names	Age	Sex	Zip code	Diseases
Bop	23	M	11000	Pneumonia
Ken	27	M	13000	Dyspepsia
Linda	65	F	25000	Gastritis
Alice	65	F	25000	Flu
Peter	35	M	59000	Dyspepsia
Sam	59	M	12000	Pneumonia
Jane	61	F	54000	Flu
Mandy	70	F	30000	Bronchitis
Jane	62	F	54000	Flu
Moore	79	F	30000	Bronchitis
Kjetil	30	M	12000	Flu
Stephen	54	F	13000	Bronchitis

Table 2: Result of anonymizing

Names	Age	Sex	Zip code	Diseases
*	23-27	M	11000-25000	Pneumonia
*	23-27	M	11000-25000	Dyspepsia
*	35-61	F	30000-59000	Gastritis
*	35-61	F	30000-59000	Flu
*	35-61	M	30000-59000	Dyspepsia
*	59-65	M	11000-25000	Pneumonia
*	59-65	F	11000-25000	Flu
*	59-65	F	11000-25000	Bronchitis
*	62-79	F	30000-59000	Flu
*	62-79	F	30000-59000	Bronchitis
*	62-79	M	30000-59000	Flu
*	62-79	F	30000-59000	Bronchitis

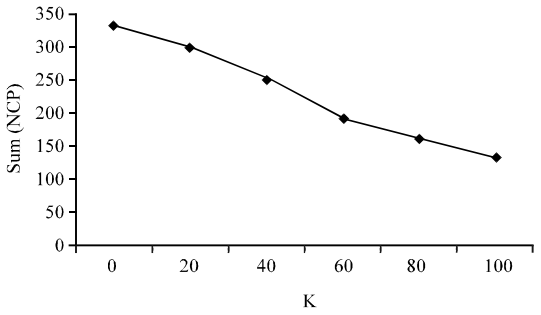

Fig. 2: Privacy level (K) versus Normalized Certainty Penalty (NCP)

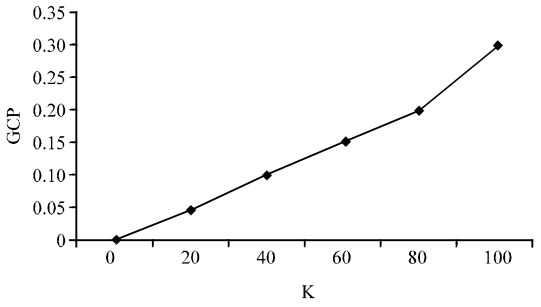

Fig. 3: Privacy level (K) versus Global Certainty Penalty (GCP)

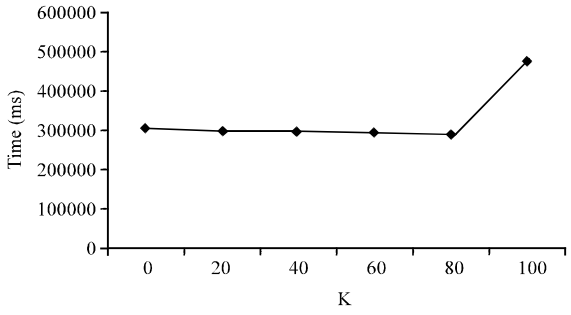

Fig. 4: Privacy level (K) versus time taken to run

Attributes with unique identifiers are suppressed to * while other attributes are anonymized using 2-anonymization. As can be seen in Fig. 2, the dynamics of NCP with respect to K value is presented. NCP decreases when K value increases.

As can be seen in Fig. 3, the dynamics of GCP with respect to K value is presented. GCP increases when K value increases. As can be seen in Fig. 4, it is evident that the time taken for processing data is presented. The time taken is same for most of the K values while the time taken dramatically increases when K value goes beyond 80.

CONCLUSION

Data privacy plays an important role when data is published for general public or it is given to any third party for discovering knowledge or business intelligence from it. Our research focused on proposing an extended misusability measure (Kumar *et al*., 2016) that helps in finding the probability of misusability of given dataset. The score obtained by the measure can help in understanding the possible misuse if the data is not anonymized. In this study, we proposed a framework that helps in utilizing the measure. In other words, the proposed framework evaluates the utility of extended misusability score to determine whether sanitization is required or not. We employed K-anonymity algorithm which is well known data mining algorithm for preserving privacy before publishing or outsourced for data mining purposes. We built a prototype application that demonstrates the proof of concept. The empirical results revealed that our misusability measure has significant impact on the privacy preserving data publishing and privacy preserving data mining.

REFERENCES

Brankovic, L. and V.E. Castro, 2011. Privacy issues in knowledge discovery and data mining. Privacy Issues Knowl., 1: 45-56.

Chen, Y., H. Chen, A. Gorkhali, Y. Lu and Y. Ma *et al*., 2016. Big data analytics and big data science: A survey. J. Manage. Anal., 3: 1-42.

Cooley, J. and S. Smith, 2013. Privacy-preserving screen capture: Towards closing the loop for health IT usability. J. Biomed. Inf., 46: 721-733.

Dong, X., J. Yu, Y. Luo, Y. Chen and G. Xue *et al*., 2014. Achieving an effective, scalable and privacy-preserving data sharing service in cloud computing. Comput. Secur., 42: 151-164.

Giannotti, F., L.V. Lakshmanan, A. Monreale, D. Pedreschi and H. Wang, 2013. Privacy-preserving mining of association rules from outsourced transaction databases. IEEE. Syst. J., 7: 385-395.

Guo, D., X. Zhu, H. Jin, P. Gao and C. Andris, 2012. Discovering spatial patterns in origin-destination mobility data. Trans. GIS., 16: 411-429.

Islam, M.Z. and L. Brankovic, 2011. Privacy preserving data mining: A noise addition framework using a novel clustering technique. Knowl. Based Syst., 24: 1214-1223.

Jiang, Q., M.K. Khan, X. Lu, J. Ma and D. He, 2016. A privacy preserving three-factor authentication protocol for E-Health clouds. J. Supercomputing, 72: 3826-3849.

Kawamoto, J., 2016. An implementation of privacy preserving stream integration system. Proceedings of the 2016 International Conference on Information Networking (ICOIN), January 13-15, 2016, IEEE, Fukuoka, Japan, ISBN:978-1-5090-1724-9, pp: 57-62.

Kumar, J.P., A.U. Kumar and T. Ravi, 2016. Beyond M-score for detection of misusability by malicious insiders. Proceedings of the 3rd International Conference on Computing for Sustainable Global Development (INDIACom), March 16-18, 2016, IEEE, Chennai, India, ISBN:978-9-3805-4421-2, pp: 155-159.

Li, L., R. Lu, K.K.R. Choo, A. Datta and J. Shao, 2016. Privacy-preserving-outsourced association rule mining on vertically partitioned databases. IEEE. Trans. Inf. Forensics Secur., 11: 1847-1861.

Matatov, N., L. Rokach and O. Maimon, 2010. Privacy-preserving data mining: A feature set partitioning approach. Inf. Sci., 180: 2696-2720.

Oksanen, J., C. Bergman, J. Sainio and J. Westerholm, 2015. Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data. J. Transp. Geogr., 48: 135-144.

Taneja, H. and A.K. Singh, 2015. Preserving privacy of patients based on Re-identification risk. Procedia Comput.Sci., 70: 448-454.

Turner, S.D., S.M. Dudek and M.D. Ritchie, 2010. Athena: A knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait Loci. Bio. Data Min., 3: 1-18.

Victor, N., D. Lopez and J.H. Abawajy, 2016. Privacy models for big data: A survey. Intl. J. Big Data Intell., 3: 61-75.

Waqar, A., A. Raza, H. Abbas and M.K. Khan, 2013. A framework for preservation of cloud users data privacy using dynamic reconstruction of metadata. J. Network Comput. Appl., 36: 235-248.

Yan, S., S. Pan, Y. Zhao and W.T. Zhu, 2016. Towards privacy-preserving data mining in online social networks: Distance-grained and item-grained differential privacy. Proceedings of the Australasian Conference on Information Security and Privacy, July 4-6, 2016, Springer, New York, USA., pp: 141-157.

Zhang, Q., L.T. Yang and Z. Chen, 2016. Privacy preserving deep computation model on cloud for big data feature learning. IEEE. Trans. Comput., 65: 1351-1362.