Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 3, March 2014, pg.188 – 193

SURVEY ARTICLE

A Survey on Privacy Preservation in Data Publishing

V. Shyamala Susan[1], Dr. T. Christopher[2]

[1]Head, Department of Computer Science, A.P.C.Mahalaxmi College for Women, Thoothukudi, India

[2]Head & Assistant Professor, Department of Computer Science, Government Arts College, Udumalpet, India

[1] shyamalasusan@gmail.com; [2] chris.hodcs@gmail.com

*Abstract— **Privacy preservation is the most concentrated issue in data publishing, as the sensitive information should not be leaked. For this sake, several techniques such as generalization, bucketization are proposed, in order to deal with privacy preservation. However, generalization fails on high dimensional data because of dimensionality and it causes information loss due to uniform distribution. On the other hand, bucketization cannot achieve membership disclosure. All the above mentioned shortcomings are overcome by a technique named slicing. Slicing can handle high dimensional data too. It is proposed that slicing can be clubbed with the algorithm in order to increase the data utility and privacy.***

Keywords— Generalization, Bucketization, Slicing

I. INTRODUCTION

Data mining deals with the process of extracting useful data from the large databases and makes this into useful information. Data mining has got several dimensions such as text mining, web mining, data clustering, data classification and privacy preservation etc.

Privacy preservation in data publishing is the major topic of research in the field of data security. Data publication in privacy preservation provides methodologies for publishing useful information; simultaneously the privacy of the sensitive data has to be preserved.

Generally, there are two important phases for data publication. They are data collection and data publication. In the data collection phase, the data holder collects the data from the data owner.

The data publication phase deals with the release of collected data to a data miner or the public, denoted as the data recipient. The data recipient mines the published data.

Privacy preservation in data mining is sensitive, because the identity and other personal details should not be disclosed, while publishing data. On the other hand, a trade-off is observed with privacy preservation and data utility.

Several anonymization techniques are proposed to manage this trade-off. There are three approaches to achieve privacy preservation and they are perturbation, anonymization and cryptography.

1. Perturbation

This approach of privacy preservation can be used by the organizations when the data are about to be classified, so as to control the disclosure of sensitive data. There are several data perturbation methods such as

© 2014, IJCSMC All Rights Reserved

additive, multiplicative, matrix multiplicative, micro-aggregation, categorical, resampling, data swapping and shuffling, probability distribution and value distortion approaches are available to protect data [1].

2. Data Anonymization

Data anonymization reduces the identity disclosure. Data anonymization aims at removing the electronic track of the data, which can allow the intruder to gain access to the source of the data.

Organizations need to publish detailed data, commonly known as microdata, which consists of information about the entity which can be a person or an organization.

Each dataset maintains a table of the form D(Explicit identifier, Quasi identifier, sensitive attributes, non-sensitive attributes). An explicit identifier can explicitly identify the data owner and it can be the name, unique id etc.

Quasi identifier is a set of attributes that can identify the data owner but not in a full-fledged manner and it can be the gender, age etc. A sensitive attribute contains the person-specific information, which can be the salary or disease etc.

Non-sensitive attributes are the attributes that do not come under the mentioned categories. Data anonymization enables the data transfer in such a way that the risk of unintended disclosure is considerably reduced.

The two important privacy preserving approaches are k-anonymity and l-diversity. Out of these two, k-anonymity prevents the identification of individual records from the data and l-diversity prevents the association of an individual record with the sensitive value attribute. Sensitive attributes are revealed and it paves way for the background knowledge attack, in the k-anonymity.

K-anonymity

A k-anonymous database is a database in which the attributes are suppressed or generalized until each row is identical, with at least k-1 other rows. Thus, the k-anonymity prevents the definite database linkages.

Accurate data release is guaranteed by the k-anonymity. This concentrates on two techniques. They are generalization and suppression. K-anonymity guarantees that the information cannot be linked to groups with less than k individuals.

Thus, k-anonymity provides information disclosure whereas it cannot guarantee the attribute disclosure. There are three kinds of information disclosure and they are identity, attribute and membership disclosure.

Identity disclosure is when an individual is linked to a particular record in the published data. When the sensitive information about an individual is disclosed, then it is attribute disclosure. Membership disclosure is when the information from the dataset, about an individual is not disclosed [3].

K-anonymity is easily vulnerable to homogeneity and the background knowledge attack. The homogeneity attack occurs in the absence of diversity between the sensitive attributes for a certain block.

This may result in revealing sensitive information based on the non-sensitive information, which is known to the attacker. Background knowledge attack is the attack that reveals information about an individual, when the demographic information about the user is linked to the published data.

The main limitations of k-anonymity are it cannot hide the presence of an individual in the database, the sensitive attributes of individuals are revealed, it cannot safeguard against the attacks of type background knowledge, and degree of utility is completely reduced when it is applied over high dimensional data.

l-Diversity

The main theme of l-diversity is that the values of the sensitive attributes are needed to be represented well in each group. There are two broad categories of l-diversity techniques and they are generalization and permutation based.

Skewness attacks and the similarity attacks can easily affect the l-diversity. Skewness attack is the attack that cannot prevent attribute disclosure, when the overall distribution is skewed.

Similarity attack is the type of attack, through which the intruder can gain knowledge, when the sensitive information is distinct but semantically similar. It is very complicated to achieve l-diversity and also the privacy protection is not up to the mark.

2.1 Anonymization Techniques

There are two main data anonymization techniques namely generalization and bucketization. The main difference between these two is, the bucketization do not generalize the QI attributes.

2.1.1 Generalization

Generalization is the most commonly used anonymization technique that replaces Quasi-Identifier (QI) values with the less specific but semantically consistent values. After that, all QI values in a group will be generalized to the entire group extent in the QID space [4].

The generalization may lead to high information loss, due to the high dimensionality of QI. Records in the equalization class should be closer to each other, such that the information loss can be avoided.

Also, over generalization lands at the useless data. To analyze or mine data on the generalized table, the data analyst has to carry out the assumption of uniform distribution, that every value in a generalized interval is equally possible.

© 2014, IJCSMC All Rights Reserved

This reduces the utility of generalized data considerably. As each attribute is generalized separately, correlations between different attributes are lost.

2.1.2 Bucketization

Bucketization partitions the tuples in the table into buckets and then the sensitive attribute is separated from the non-sensitive attribute by randomly permuting the sensitive attribute values within each bucket. Now, the sanitized data contains the bucket with permuted sensitive values.

The data utilization is high in bucketization when compared to generalization. However, the limitations are bucketization did not prevent membership disclosure.

As the QI values are published in their original forms, the intruder can easily figure out the presence of an individual's record in the published data.

Bucketization requires a clear separation between QI and Sensitive Attribute (SA). In many datasets, the QI and SA cannot be easily differentiated. Finally, by separating the sensitive attributes from the QI attributes, bucketization violates the attribute correlations between the QI and the SA.

2.1.3 Slicing

Slicing is the new technique developed for privacy preservation, which has got many advantages over generalization and bucketization.

Slicing partitions the dataset both horizontally and vertically. Horizontal partitioning is done by grouping tuples into buckets. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes.

Each column consists of a subset of highly correlated attributes. At last, the values in each column are randomly permuted, in order to break the link between different columns, within each bucket.

The central theme of slicing is to break the association cross columns and to preserve the association within each column. This reduces the data dimensionality and better utility is rendered than generalization and bucketization. Slicing groups highly correlated attributes together, so as to preserve utility and also it preserves the correlation between such attributes. Slicing provides privacy by breaking the association between uncorrelated attributes, which are not frequent and so identifying.

The process of slicing ensures that for every tuple, there are multiple matching buckets. Initially, slicing partitions the attributes into columns. Each column contains a subset of attributes. Slicing partitions the tuples into buckets also.

Every bucket consists of a subset of tuples. This partitions the table horizontally. The values in each column are randomly permuted to break the linking between different columns.

Many algorithms such as generalization and bucketization were proposed for preserving privacy, but they end up in attribute disclosure. In order to overcome the issue, slicing is proposed and it consists of three phases namely attribute partitioning, column generalization and tuple partitioning.

2.1.3.1 Attribute Partitioning

Attribute partitioning algorithm partitions the attributes, so that the highly correlated attributes will be in the same column, which is favourable for both data utility and privacy.

In the perception of data utility, the highly correlated attributes are grouped and this preserves the correlation among the attributes.

When privacy is considered, the association of uncorrelated attributes present the risk of higher identification than the association of highly correlated attributes. This is because the association of uncorrelated attribute values is the least frequent and hence easily identifiable.

2.1.3.2 Column Generalization

Column generalization is required for identity or membership disclosure protection. If a value of a column is unique in a column, a tuple with this unique column value can have only one matching bucket. This is not good for privacy protection, as in the case of generalization or bucketization, in which each tuple can belong to only one equivalence class or bucket.

2.1.3.3 Tuple partitioning

Two data structures are maintained in this algorithm. They are a queue of buckets Q and a set of sliced buckets SB. Initially, Q contains a bucket that includes all tuples and SB is empty.

The algorithm removes a bucket from Q, and then it splits the bucket into two buckets iteratively. If the sliced table after the split satisfies l-diversity, then the algorithm adds the two buckets at the end of the queue Q. Else, the bucket cannot be parted and the algorithm adds the bucket to the SB.

When the queue is empty, the sliced table is computed and the set of sliced buckets is SB.

3. Cryptographic Methods

Cryptography focuses on securing the information from adversaries. There are several dimensions of data security such as data confidentiality, integrity and authentication. Symmetric key cryptography, public key cryptography, cryptanalysis and cryptosystems are the widely used privacy preservation techniques.

4. Experimental Results

To evaluate the performance of the classification techniques several performance metrics are available.

The performance metrics considered are True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Detection Accuracy (DA) and Peak-to-Side lobe Ratio (PSR)

4.1 TRUE POSITIVE (TP):

True Positive (TP) is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$TP = \frac{Number\ of\ correctly\ classified\ data}{Total\ data} x100$$

True positive performance metrics calculates the number of data which classifies correctly

4.2 TRUE NEGATIVE (TN):

True Negative (TN) is defined as the proportion of negatives cases that were classified correctly, as calculated using the equation:

$$TN = \frac{Number\ of\ falsely\ classified\ data}{Total\ no.\ of\ data} x100$$

4.3 FALSE POSITIVE (FP):

False Positive (FP) is the proportion of negatives cases that were incorrectly classified as positive, as calculated using the equation:

$$FP = \frac{Number\ of\ correctly\ classified\ data}{Total\ no.\ of\ data} x100$$

4.4 FALSE NEGATIVE (*FN*):

False Negative (FN) is the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation:

$$FN = \frac{Number\ of\ falsely\ classified\ data}{Total\ no.\ of\ data} x100$$

4.5 CLASSIFICATION ACCURACY (DA):

Accuracy is a global measure providing the ratio of total well-classified images.

$$DA = \frac{TP + FP}{TP + FN + TN + FP} x100$$

The detection accuracy is calculated using the sum of true positive and true negative divided by the sum of true positive, false negative, true negative and false positive.

Table 1: Classification Accuracy

Technique & Data Set		TP	TN	FP	FN	ACC
K- Means	**OCC - 7**	**82**	**18**	**72**	**28**	**77.0**
	OCC -15	**89**	**11**	**88**	**12**	**88.0**
Weighted K - Means	**OCC - 7**	**75**	**25**	**73**	**27**	**74**
	OCC -15	**83**	**17**	**77**	**23**	**80**
Fuzzy C - Means	**OCC - 7**	**76**	**24**	**89**	**11**	**82.5**
	OCC -15	**85**	**15**	**91**	**09**	**87.6**
Ant Colony	**OCC - 7**	**69**	**31**	**64**	**36**	**66.5**
	OCC -15	**86**	**14**	**76**	**24**	**81.34**

© 2014, IJCSMC All Rights Reserved

Bee Colony	**OCC - 7**	**91**	**09**	**89**	**11**	**90.45**
	OCC -15	**96**	**4**	**88**	**12**	**95.36**

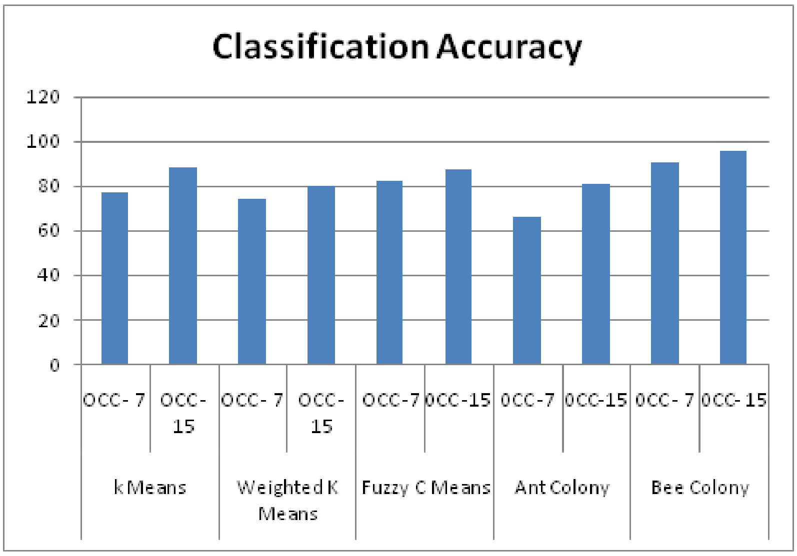

Fig 1: Classification Accuracy

Thus, bee colony proves its classification accuracy; in future this algorithm can be clubbed with another algorithm to arrive at better accuracy.

II. Conclusion

A detailed survey on various anonymization techniques is carried out. Every technique has got its own significance. Generalization leads to information loss and bucketization cannot assure privacy preservation because of identity disclosure.

Slicing provides high dimensional data by partitioning highly correlated attributes into columns and also it breaks the association of uncorrelated attributes. Thus, slicing in combination with correlation analysis ensures high data utility and ensures privacy of data.

References

[1] A. K. Ilavarasi, B. Sathiyabhama, "A Survey on Privacy Preserving Data Mining Techniques", International Journal of Computer Science and Business Informatics, Vol. 7, No. 1. NOVEMBER 2013

[2] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati On K-Anonymity. In Springer US, Advances in Information Security (2007).

[3] Inan.A,Kantarcioglu.M,and Bertino.e, "Using Anonymized Data for Classification," Proc. IEEE 25th Int"l Conf. Data Eng. (ICDE), pp. 429-440, 2009.

[4] Gabriel Ghinita, Member IEEE, Panos Kalnis, Yufei Tao," Anonymous Publication of Sensitive Transactional Data" in Proc. Of IEEE Transactions on Knowledge and Data Engineering February 2011 (vol. 23 no. 2) pp. 161-174.

[5] Gabriel Ghinita, Member IEEE, Panos Kalnis, Yufei Tao," Anonymous Publication of Sensitive Transactional Data" in Proc. Of IEEE Transactions on Knowledge and Data Engineering February 2011 (vol. 23 no. 2) pp. 161-174.

[6] D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J.Y. Halpern, "Worst-Case Background Knowledge for Privacy- Preserving Data Publishing," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 126-135, 2007.

[7] X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 139-150, 2006.

[8] Y. He and J. Naughton, "Anonymization of Set-Valued Data via Top-Down, Local Generalization," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 934-945, 2009.

[9] D. Kifer and J. Gehrke, “Injecting Utility into Anonymized Data Sets,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD), pp. 217-228, 2006.
[10] T. Li and N. Li, “On the Tradeoff between Privacy and Utility in Data Publishing,” Proc. ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp. 517-526, 2009.
[11] Y. Xu, K. Wang, A.W.-C. Fu, and P.S. Yu, “Anonymizing Transaction Databases for Publication,” Proc. ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp. 767-775, 2008.
[12] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A.W.-C. Fu, “Utility- Based Anonymization Using Local Recoding,” Proc. 12th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp. 785-790, 2006.
[213] C. Dwork, “Differential Privacy: A Survey of Results,” Proc. Fifth Int’l Conf. Theory and Applications of Models of Computation (TAMC), pp. 1-19, 2008.
[14] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey on recent developments. ACM Computing Surveys, 42(4), December 2010.

Authors Bibliography

V. Shyamala Susan, received her post graduate degree in MCA from Coimbatore Institute of Technology, Coimbatore and M.phil degree is earned from M.S University, Tirunelveli. At present, she is working as the Head, Department of Computer Science in A.P.C. Mahalaxmi College for women, Thoothukudi, India. She has presented many papers in national conferences. She has got eleven years of teaching experience and her area of interest is data mining.

Dr.T.Christopher, has earned M.Phil in both physics and computer science and received his doctorate in computer science. He owns M.ED., too. Presently, he is working as the Head and Assistant Professor, Department of Computer Science in Government Arts College, Udumalpet, India. His area of specialization is data mining and network security.