International Journal of Computer Science and Business Informatics

IJCSBI.ORG

A Survey on Privacy Preserving Data Mining Techniques

A. K. Ilavarasi
Assistant Professor, Department of Computer Science and Engineering,
Sona College of Technology,
Salem, India

B. Sathiyabhama
Head of the Department,
Department of Computer Science and Engineering,
Sona College of Technology,
Salem, India

S. Poorani
PG Scholar, Department of Computer Science and Engineering,
Sona College of Technology,
Salem, India.

ABSTRACT
Many kinds of anonymization techniques have been in the subject of research. This paper will present a detailed review of several anonymization techniques particularly in the area called "Privacy Preserved Data Mining". Recent experiments shown that some of the anonymization techniques like generalization, bucketization doesn't ensure the privacy preservation. And it is experimentally shown that slicing provides significant level of utility and also prevents membership disclosure. Thus, detailed analysis is done on the Post anonymization techniques and the necessity for privacy preservation is also reviewed in detail.

Keywords:
Anonymization, Privacy preservation, data mining, k-anonymity, l-diversity.

1. INTRODUCTION
Data mining is the process of analysing data from various perspectives and acquiring the useful information. Knowledge discovery is the ultimate goal of data mining. Nowadays, the data through the internet and other social media are plenty. Hence the privacy preservation deserves the serious attention. Privacy Preservation in Data Mining (PPDM) is a novel technique in data mining, where mining algorithms are incorporated. The significance of PPDM varies from different perspective because while publishing the data, the individual's identity and other details should not get disclosed. As well the information loss due to privacy preservation highly affects the data utility. PPDM, balance the trade-off between utility and privacy preservation by using various anonymization techniques.

2. TAXONOMY

In general, the personal identifications will be removed before publishing the data for mining purpose. Privacy preservation is a serious issue and it can be gained through different techniques. Figure 1 describes the taxonomy of Privacy preservation in data mining. The three main approaches of Privacy preservation are Perturbation, Anonymization and Cryptography.

2.1 Perturbation

The perturbation method for categorical data can be used by organizations to prevent or limit disclosure of confidential data for identifiable records when the data are provided to analysts for classification. Based on the needs of privacy protection the perturbation approach will ensure the statistical properties of the data. As the medical dataset has high probability of linking attack, perturbation can be effectively applied to such field.

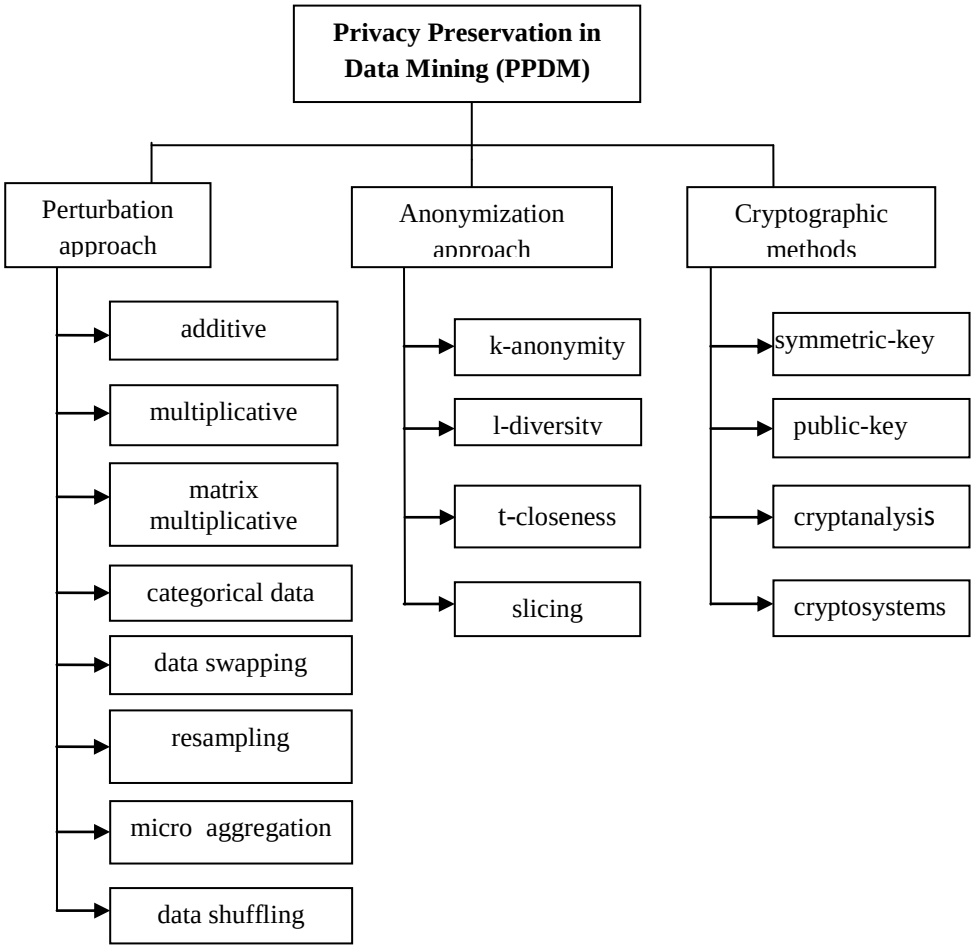

Figure 1. Taxonomy

There are different kinds of data perturbation methods available for data protection. The methods includes additive, multiplicative, matrix

multiplicative, micro aggregation, categorical, resampling, data swapping and data shuffling, probability distribution approach and the value distortion approach.

2.2 Data Anonymization

Anonymization reduces the risk of identity disclosure whereas the data remains still realistic. Micro data contains information about an individual, a household or an enterprise. Each such dataset will be having i) Personal identification like Name, Address or Social Security Number (SSN) which uniquely identifies an individual ii) Sensitive Attributes (SAs) like salary and disease iii) The values of Quasi Identifiers (QI) such as Gender, Age, Zip code will leads to identity disclosure when taken together. Two main Privacy Preserving approaches are k-anonymity and l-diversity.

k-anonymity prevents the identification of individual records in the data and l-diversity prevents the association of an individual record with the sensitive value attribute. k-anonymity has the limitations of revealing sensitive attributes and background knowledge attack. And it cannot be applied to high-dimensional data without complete loss of utility.

2.2.1 Generalization

Generalization is one of the conventional anonymization techniques. It was the widely used technique which replaces the QI values with "less-specific but semantically consistent value". Due to high dimensionality of the QI, the generalization would cause high information loss. Records in the equivalence class should be close to each other in order to avoid information loss. Another defect is the over generalization which makes the data useless. Effective analysis of attribute correlation also gets lost due to separate generalization of each attribute.From an l-diverse generalized table, an adversary can gain 1/l sensitive data of every individual.

Table 1. A 2 diverse generalize paper

age	Sex	Zipcode	Disease
[21,60]	M	[10001,6000]	Pneumonia
[21,60]	M	[10001,6000]	Dispepsia
[21,60]	M	[10001,6000]	Dispepsia
[21,60]	M	[10001,6000]	Pneumonia
[61,80]	F	[10001,6000]	Flu
[61,80]	F	[10001,6000]	Pneumonia
[61,80]	F	[10001,6000]	Dispepsia
[61,80]	F	[10001,6000]	Pneumonia

l-diversity makes the group of k different records that all share a particular quasi –identifier. A QI-group with m tuples is l-diverse, if each sensitive

value appears no more than m / l times in the QI-group. A table is l-diverse, if all of its QI-groups are l-diverse.

Table 2. Published voter's list

name	age	sex	Zip code	disease
John	23	M	10000	Pneumonia
Peter	35	M	13000	Flu
Martin	61	F	54000	Pneumonia

The defect of generalization is for the query like

SELECT COUNT (*) from Unknown-Micro data

WHERE Disease = 'pneumonia' AND Age in [0, 30]

AND Zip code in [10001, 20000]

Estimated answer for query A: $2 * p = 0.1$

Table 2.

Age	**Sex**	**Zip code**	**Disease**
[21, 60]	M	[10001, 60000]	Pneumonia
[21, 60]	M	[10001, 60000]	Pneumonia

2.2.2 Bucketization

Anatomy is one of the new techniques for publishing the sensitive data. Anatomy protects privacy by releasing all the Quasi-Identifiers and sensitive values in two separate tables. This technique provides effective data analysis than the generalization. And it also achieves privacy-preserving publication by capturing the exact QI-distribution. The experimental results derive the highly accurate aggregate information with the average error below 10% when compared with that of generalization.

The QIT has the schema
$\left(A_1^{qi}, A_2^{qi}, ...A_d^{qi}, Group-ID\right)$

The ST has the schema

(Group-ID, A^s, Count)

Table 3. Quasi-entifier Table (QIT)

Age	Sex	Zipcode	Group-id

23	M	1100	1
27	M	13000	1
35	M	59000	1
59	M	12000	1
61	F	54000	2
65	F	25000	2
65	F	25000	2
70	F	30000	2

Table 4. Sensitive table (ST)

Group-ID	Disease	Count
1	Dyspepsia	2
1	Pneumonia	2
2	Bronchitis	1
2	Flu	2
2	Gastritis	1

David J.Martin et al. [3] describe the necessity of considering the attacker's background knowledge when discussing the privacy in data publishing. The polynomial time algorithm was proposed to measure the sensitive information disclosure in the worst-case. Thus the worst-case background knowledge helps to analyse the knowledge that the attacker possess. The background knowledge can be sanitized by two methods called bucketization and full-domain generalization.

Bucketization, partition the set of tuples T into buckets and each sensitive attribute will be randomly permuted within each bucket. The buckets will provide the sanitized data with permuted values. Bucketization has better utility than generalization but bucketization does not prevent membership disclosure. Bucketization does not have any clear separation between QIs and SAs.

Aggregate queries cannot be answered well with the presently available generalization based anonymization approaches. This problem is focused in [4], which provides a framework for accurate aggregate queries with permutation based anonymization. This would be more accurate than generalization based approach. Permutation based anonymization is carried out by data swapping techniques where privacy is achieved by exchanging the sensitive attributes and it provides high micro data utility.

Anonymization through permutation is carried out because of following reasons. The individual's identity can be recovered by three ways: (1) the link between the identifier and quasi-identifiers in the public database P; (2)

the link between the QIs in P and those in the deidentified micro data D; (3) the link between QIs and the sensitive value D. Breaking the associations of the above links will ensure privacy.

Domain generalization weakens only the second and third links. Instead of using domain generalization we can permute the association between the quasi-identifiers and the sensitive attributes. Even if an attacker can link an individual's identifier with tuple's QI he will not be able to know with certainty the exact value of the individual's sensitive attribute.

Table 5. 3-anonymity table after generalization- satisfies 3-diversity

		Quasi-identifiers			Sensitive
group-ID	tuple-ID	Age	zip code	gender	Salary
1	1	[31-40]	271*	*	$56,000
1	2	[31-40]	271*	*	$54,000
1	3	[31-40]	271*	*	$55,000
2	4	[41-50]	272*	*	$65,000
2	5	[41-50]	272*	*	$75,000
2	6	[41-50]	272*	*	$70,000
3	7	[51-60]	276*	*	$80,000
3	8	[51-60]	276*	*	$75,000
3	9	[51-60]	276*	*	$85,000

Table 6. 3-anonymous table after permutation

		Quasi-identifiers			Sensitive
group-ID	tuple-ID	Age	zip code	gender	Salary
1	1	40	27130	M	$54,000
1	2	38	27120	M	$55,000
1	3	35	27101	M	$56,000
2	4	41	27229	F	$65,000
2	5	43	27269	F	$70,000
2	6	47	27243	M	$75,000
3	7	52	27656	M	$75,000
3	8	53	27686	F	$80,000
3	9	58	27635	M	$85,000

Slicing discussed in [5], is one of the novel techniques which better preserves the data utility than generalization. Slicing also protects membership disclosure than bucketization. High dimensional data can be handled better by slicing based anonymization. Privacy is ensured by partitioning the attributes into columns and that breaks the association of

uncorrelated attributes. Data utility is preserved by preserving the highly correlated attributes.

Slicing partitions the dataset both horizontally and vertically. The objective of slicing is to break the association of poorly correlated attributes among columns but the association within each column will be preserved. Multiple matching buckets ensure privacy. Randomly permuting each values within each bucket will break the linking between different columns. The law of total probability calculates the probability of the sensitive value, $p\ (t,s)$

$$p(t,s) = \sum_B p(t,B)p(s|t,B) \quad (1)$$

Consider Tuple t which may have many matching buckets, in the whole data 'D' t's matching degree can be given as f (t) = $\sum_B$ f (t, B).

The probability that t is in bucket B is:

$$\mathrm{p}(\mathrm{t},\mathrm{B}) = \frac{f(t,B)}{f(t)} \quad (2)$$

l-diverse Slicing: A tuple t satisfies l-diversity iff for any sensitive value s,

$$p(t,s) \leq 1/1 \quad (3)$$

A Sliced table satisfies l-diversity iff every tuple in it satisfies l-diversity.

FACT: For any tuple t ϵ D, $\sum_s p\ (t,s) = 1$

PROOF:

$$\begin{aligned} \Sigma_S\, p(t,s) &= \Sigma_S \Sigma_B\, p(t,B)p(s|t,B) \\ &= \Sigma_B\, p(t,B)\Sigma_S\, p(s|t,B) \\ &= \Sigma_B\, p(t,B) \\ &= 1 \end{aligned}$$

Chi-square measure of correlation analysis is used as follows:

$$\varnothing^2(A1,A2) = \frac{1}{\min\{d1,d2\}-1} \Sigma_{i=1}^{d_1} \Sigma_{j=2}^{d_1} \frac{(f_{ij} - f_i \cdot f_j)^2}{f_i \cdot f_j} \quad (4)$$

Advantages of slicing over generalization:

[1] Generalization fails on high-dimensional data due to the curse of dimensionality.

[2] It also cause too much of information loss due to uniform-distribution.

Advantages of slicing over bucketization:

[1] Slicing prevents membership disclosure which bucketization fails to do.

[2] Better preservation of attribute correlation between sensitive attributes and the QI attributes.

Tiancheng Li and Ninghui Li, explains that there is no proper trade-off between privacy and utility. The trade off presents systematic methodology for measuring privacy loss and utility loss. [6] Also provides quantitative interpretations to the trade off which guides the data publishers to choose right privacy-utility trade-off.

COROLLORY: Privacy should be measured against the trivially-anonymized data whereas utility should be measured using the original data as the baseline.

Utility can be measured using "utility loss" instead of using "utility gain". Well achieved privacy-preserving method should result in zero privacy loss and zero utility loss.

Privacy loss can be measured using the JS divergence distance measure:

$$P_{loss}(t) = JS(Q, P(t)) = \frac{1}{2}\left[KL(Q, M) + KL(P(t), M\right]$$

where M = $\frac{1}{2}$ (Q+ P (t)) and KL (,) is the KL-divergence

$$KL(Q, P) = \Sigma_{iqi} \log \frac{q_i}{p_i}$$

The worst- case privacy loss is measured as the maximum privacy loss for all tuples in the data:

$$P_{loss} = \max_t P_{loss}(t) \quad (5)$$

Utility loss can be measured again using JS divergence:

$$Uloss(y) = JS(P_y, \overline{P_y}) \quad (6)$$

Because utility is an aggregate concept, utility loss is measured by averaging the utility loss U_{loss} (y) for all large population y. Maximum utility can be achieved when $U_{loss} = 0$.

$$Uloss = \frac{1}{|Y|} \Sigma y \in y\, Uloss(y) \quad (7)$$

where Y is the set of all large populations

In [7], the two k-anonymity algorithm was proposed which provides the most accurate classification models based on the mutual information obtained. It is also discussed that the data generalization should be based on classification capability of data rather than the privacy requirement to ensure the perfect anonymization.

Mutual information is used to measure which generalization level is best for the classification. The uncertainty associated with the set of class labels is described as:

$$H(C) = -\Sigma_{k=1}^{p} freq(C_k) \times \log_2 freq(C_k) \quad (8)$$

where H(C) indicates the classification uncertainty without using other attribute information.

The mutual information is biased towards attributes of many values. Such bias should be avoided and this can be achieved by normalising the mutual information. I_N denotes normalized mutual information.

$$I_N(A_i(l);C) = \frac{I_N(A_i(l);C)}{H(A_{i(l)})} \quad (9)$$

The normalised mutual information of all possible generalization levels should be compared. The one with the highest normalized mutual information is the best for the classification. The algorithm maximises the classification capability by generalization. Suppression is done by privacy requirements K [IACk] or distributional constraints [IACc]. The proposed method IACk supports anonymization with better classification model than by utility-aware method.

In [8] both global recoding and local recoding are discussed. The global recoding method maps the domain of the QI attributes to generalized or changed values. Global recoding does not achieves effective anonymization in case of discernability and query answering accuracy. Local recoding covers both numerical and categorical data which global recoding fails to do. Here two algorithms namely the bottom-up algorithm and the top-down greedy search methods are used to perform local recoding. The bottom-up algorithm reduces the weighted certainty penalty which reflects the utility of anonymized data. The top down approach partition the table iteratively by using the binary partitioning. The number of groups that are smaller than k is much less than the worst case. Thus, the top down method is comparatively faster than the bottom-up method.

Raymond Chi-Wing Wong et al., describes the minimality attack as the knowledge of the mechanism or the algorithm of anonymization for data publication which will lead to privacy breech. The mechanism which tries

to minimize the information loss and such an attempt leads to minimality attack. The minimality attack deals with m-confidentiality which can prevent the attacks with less information loss.

The main objective of privacy preservation is to limit the probability of the linkage from any individual to any sensitive value set s in the sensitive attribute. The probability or credibility can be defined as:

Let T^* be the published table which is generated from T. Consider an individual o ϵ O and a sensitive value set s in the sensitive attribute. Credibility (o, s, K_{ad}) is the probability that the adversary can infer from T^* and background knowledge K_{ad} that o is associated with s.

A table T is said to satisfy m-confidentiality if, for any individual o and any sensitive value set s, Credibility (o, s, K_{ad}) does not exceed 1/m.

Then the information loss of a tuple t^* in T^* is given by,

$$Dist\left(T, T^*\right) = \frac{\sum t^* \in T^* IL\left(t^*\right)}{\left|T^*\right|} \quad (10)$$

The technique in [10] states that the quality of anonymized data can be better measured with the purpose for which the data been used. This can be done with the series of techniques like queries, classification and regression models which provide the high quality data. Hence large-scale datasets can be anonymized only based on their measure of usage. Two techniques called scalable decision tree and sampling are developed which allows anonymization algorithm to be applied to large datasets.

2.3 Cryptographic Methods

Cryptography is the technique which focuses mainly on securing the information from the third parties. Information security has various aspects like data confidentiality, authentication and data integrity. Cryptographic methods like symmetric-key cryptography, public-key cryptography, cryptanalysis and cryptosystems are widely used privacy preservation methods.

3. CONCLUSION

The detailed survey on various anonymization methods are carried out. Every anonymization techniques have their own significance. Generalization causes too much of information loss and bucketization fails in privacy preservation due to identity disclosure. Slicing performs better than generalization, bucketization and many other anonymization methods. Slicing provides high dimensional data by partitioning highly correlated attributes into columns and further breaks the association of uncorrelated attributes. Thus slicing in combination with correlation analysis has the high data utility and ensures privacy in PPDM.

REFERENCES

[1] Raymond Chi-Wing Wong et al., *"(alpha, k) Anonymity: An Enhanced k- anonymity Model for Privacy Preserving Data Publishing",* ACM, 2006.

[2] Xiaokui Xiao, Yufei Tao, *"Anatomy: Simple and Effective Privacy Preservation",* ACM, 2006.

[3] David J.Martin et al., *"Worst–Case Background Knowledge for Privacy-Preserving Data Publishing",* National Science Foundation under Grants.

[4] Q. Zhang et al., *"Aggregate Query Answering on Anonymized Tables",* In ICDE, 2007.

[5] Tiancheng Li et al., "Slicing: *A New Approach to Privacy Preserving Data Publishing",* ACM.

[6] Tiancheng Li and Ninghui Li, *"On the Trade off between Privacy and Utility in Data Publishing",* ACM 2009.

[7] Jiyoung Li et al.,*"Information Based Data Anonymization for classification utility",* Elsevier, 2011.

[8] Jian Xu et al.,*"Utility-Based Anonymization Using Local recoding",* ACM 2006.

[9] Raymond Chi-Wing Wong et al., *"Minimality Attack in Privacy Preserving Data Publishing",* ACM, 2007.

[10] Kristen Le Fevre and Raghu Ramakrishnan, *"Workload-Aware Anonymization Techniques for Large-Scale Datasets",* ACM, 2008.

[11] Rakesh Agarwal et al., *"Privacy-preserving Data Mining",* ACM SIGMOD Conference on Management of Data, 2000.

[12] G. Sai Chaitanya Kumar et al., *"Suppression of Multidimensional Data Using K-anonymity",* International Journal of Computer Science and Communication Networks, Vol 2(4), 501-505.

[13] Ali Inan, Murat Kantarcioglu, Elisa Bertino, *"Using Anonymized Data for Classification",* AFOSR.

[14] Benjamin C.M. Fung et al., *"Anonymizing Classification Data for Privacy Preservation"*, IEEE.

[15] Patrick Sharkey et al., *"Privacy-Preserving Data Mining through Knowledge Model Sharing",* NSF.

[16] Aris Gkoulalas-Divanis, Grigorios Loukides *"PCTA: Privacy-constrained Clustering-based Transaction Data Anonymization",* ACM 2011.

[17] L.Sweeney,*"Guarenteeing Anonymity When Sharing Medical Data, the Datafly System"* Journal of Informatics Association, pages 51-55, 1997.

[18] Neha V. Mogre, Girish Agarawal, Pragati Patil, *"A Review On Data Anonymization Technique For Data Publishing",* IJERT, 2012.

[19] L.Kaufman and P. Rousueeuw, *"Finding Groups in Data: An Introduction to Cluster Analysis"*, John Wiley & Sons, 1990.

[20] G.Ghinita et al.,*"On the Anonymization of sparse high-dimensional data*", In ICDE, pages 205-216, 2005.

[21] K.LeFevre et al., *"Mondrian multidimensional k-anonymity"*, In ICDE, page 25, 2006.

[22] Y. Xu, K. Wang, A.W.C.Fu, and P. S. Yu, *"Anonymizing transaction database for publication",* In KDD, pages 767-775,2008.

[23] H.Wang, R.Liu, *"Privacy-preserving publishing micro data with full functional dependencies",* Data & Knowledge Engineering, 2011.

[24] L.Sweeney, *"k-anonymity: A model for protecting privacy",* International Journal on uncertainty, Fuzziness and knowledge based systems, 2002.

[25] K.Wang and B.Fung, *"Anonymizing sequential releases",* In SIGKDD, 2006.

[26] X.Xiao and Y.Tao,*"Personalized Privacy Preservation",* In Sigmoid, 2006.

[27] G. Agarwal et al., *"Anonymizing tables",* In ICDT, pages 246-258, 2005.

[28] Agarwal et al., *"A framework for high-accuracy privacy-preserving mining",* IEEE, 2005.

[29] Srikant et al., *"Limiting privacy breaching in Privacy Preserving Data Mining",* ACM, 2003.

[30] V.S.Iyengar *"Transforming data to satisfy privacy constraints",* In KDD, 2002.

[31] J. Li, R.Wong, A. Fu, and J.Pei," *Acheiving anonymity by clustering in attribute hierarchical structures"*, In DaWak, pages 405-416, 2006.

[32] B.C.M Fung, K.Wang, and P.S.Yu,*"Top-down Specialization for Information and Privacy Preservation"*, In ICDE, 2005.

[33] K.Wang et al., *"Bottom-up Generalization: A data mining solution to privacy protection"*, In ICDM, 2004.

[34] P.Samarati, *"Protecting Respondents' Identities in Microdata Release"*, IEEE, 2001.

[35] K.Levere et al., *"Incognito: Efficient Full-Domain k-anonymity",* ACM, 2005.