Indian Journal of Science and Technology, *Vol 9(37), DOI: 10.17485/ijst/2016/v9i37/94290, October 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Anonymization in PPDM based on Data Distributions and Attribute Relations

Jitendra Kumar Jaiswal*[1], Rita Samikannu[1] and Ilango Paramasivam[2]

[1]School of Advanced Sciences, VIT University, Vellore - 632014, Tamil Nadu, India; jitendra.kjaiswal@vit.ac.in, ritasamikannu@vit.ac.in

[2]School of Computer Science and Engineering, VIT University, Vellore - 632014, Tamil Nadu, India; pilango@vit.ac.in

Abstract

Objectives: Privacy Preserving Data Mining techniques deal with the secure data publication or communication without revealing the private and sensitive information about any individual. Anonymization technique has been considered as one of the most effective techniques since it can provide better tradeoff between data utility and privacy preservation. **Methods/Statistical Analysis:** Existing anonymization techniques works on individual attributes and their cardinalities and they do not consider the relations among different attributes of the data. In this paper we have considered auxiliary information and entropy and mutual information to calculate distribution of entities in an attribute and relations among different attributes respectively. Based on these calculations we shall be analyzing the best generalization level for data anonymization. **Findings:** An adverse user can analyze the data with numerous possible perspectives viz. auxiliary information, theoretical and manual data analysis and try to exploit the data vulnerability, so improved data privacy can be achieved if we could also see with the adversary eyes. **Applications/Improvements:** Different other techniques can be applied to find distribution and relations on the basis of data background and its area of application.

Keywords: Auxiliary Information, Data Anonymization, Entropy, Mutual Information, Privacy Preserving Data Mining (PPDM)

1. Introduction

Privacy Preserving Data Mining enables the data storage and publication in such a way that private and sensitive information should remain preserved about any individual. Data are published or shared for scientific research purposes, health, medicine and diseases characteristics observations, statistical and economic analysis, forecasting and substantially many premises. In research areas, it can promote improved and appropriate research methodologies. A privacy concerned data set may contain direct identifiable attributes, partial identifiable attributes or quasi-quantifiers, sensitive data and general information about individuals or organizations. Direct identifiable information can straight away find out an individual from a data set and it may comprise name, mobile number, social security number, PAN number, voter ID, etc. Partial identifiable information represents a group of individuals and it may comprise gender, age, pin/zip codes, etc. Sensitive data can introduce a risk of discrimination, impairment or unwanted attention to an individual. Data are either openly published or shared or a description of the data that is, metadata, can be published without making the data itself openly accessible with conditions around access to the data, but the data sensitivity should always be considered.

We have considered a standard data repository namely UCI machine learning repository for machine learning databases and followed Adult data set for our calculation works[1].

1.1 Associated Works and Further Requirement of Data Analysis

Data generalization or suppression should be executed in such a way that the data remain utilizable and least

*Author for correspondence

exploitable. Measure of data distortion[2], calculation of information loss[3] and information utility[4], estimation of uncertainty[5], query accuracy calculation[6], etc. can perpetrate information gain substantially. Further approaches like local recoding methods[2,5] and multidimensional application[7] have been applied to increase the data anonymization in the table.

If a data table is already analyzed precisely with entities distributions and relations among different attribute sets, many vulnerable points can be estimated prior to the data publication. As much precisely the analysis of the data table is performed, the better the application of anonymization techniques can be applied with calculation of degree of suppression and data utility. In this paper we have worked on the selection of the set of attributes from a table and estimate the vulnerability of table with the application of regression and probability distribution techniques and prepared a subset of attributes on which further anonymization can be applied and achieve an improved privacy preservation along with the consideration of tradeoff factor between data suppression and data utilization.

When data generalization is applied, the domain consistency should also be considered at different level of generalization since data overlapping may also take place[2,8]. For classification utilities[9], suggested bottom up approaches which can deal with categorical data only. Further they proposed top-down approach called TDS (Top-Down Specification) method which deals only with single dimensional attributes[10]. For both categorical and numerical data they improved TDS to TDR (Top-Down Refinement) method with and without generalization taxonomy trees[11]. They have also proposed kACTUS for multidimensional suppression using decision tree algorithm C4.5. An effective algorithm called Mondrian for multidimensional generalization[7], which was further improved to InfoGain Mondrian[6] for classification utilities.

In[12] has also approached for the calculation of Correlation, Joint Entropy, Mutual Information along with Non Mutual Information and Time Delay Estimation for noisy environment. In[13] has applied a parallel data processing framework, namely MapReduce, for the problem of privacy preservation in the large scale data with minimum information loss of the Bottom-Up Generalization (BUG) approach. In[14] has applied Anonymization, Suppression, Generalization and Data Hiding to preserve the sensitive data from hospitals.

In the Section 2, we have summarized the basic definitions form the anonymization perspectives. We have analyzed the considered data set in Section 3 as per the data distribution basis and considered partial identifier attributes and sensitive attributes. In the Section 4, we have elaborated the effective application of anonymization techniques. In the Section 5, we have concluded our paper and proposed the future scope of this approach.

2. Privacy Preservation and Anonymization Techniques

When data is published or shared, direct identifiable attributes are removed in the very first step and rest is released with sensitive attributes and general information after the application of privacy preserving techniques. Data anonymization is one of the most effective techniques to preserve the sensitive information about an individual along with the consideration of data reusability factor. There are mainly three types of anonymization techniques: k-anonymization, l-diversity and t-closeness. Let us see some important definitions from the domain of PPDM.

- **Definition 1. Partial Identifiable Attributes.** Partial identifiable attributes or Partial Identifiers (PID) or Quasi Identifiers (QID) are identification factors which represent a group of people. For example, attributes set {age, sex, pin code} can represent a set of people of some age from a particular region.
- **Definition 2. Sensitive attributes.** Sensitive attributes are the information about individuals whose discloser can harm anyone in any perspectives. For example, serious diseases, account information, salary, work-class, etc.

  The determination of Partial Identifiers and sensitive attributes are decided by the domain experts.
- **Definition 3. k-Anonymization.** It is applied on partial identifiable attributes and it frames data attributes in such a way that the probability of finding out an individual from an attribute at most by 1/k or in other words we can say that it provides at least k-similar entities for an attribute set[15]. For example, a data set may contain Quasi-Identifier attributes as gender, age and pin code. Gender = 'male', age = 32, pin code = 632014, may represent all the male persons of aged 32 in the region 632014.

Anonymization is applied by the generalization of the data. This generalization is performed by the suppression and showing the records by the symbol '*' or by assigning some range values. This suppression can be performed by two approaches: Top-down approach and bottom-up approach. In the bottom-up approach the initial anonymization is carried out by suppressing all the Quasi-Identifier attributes as gender = '*', age = '*', and pin code = '*'. Further it can be proceeded as gender = '*', age = '3*' or 30-40 or a more wide range value and pin code = '6*****'. This generalization is performed on the basis of cardinality or number of entities in an attribute. The more the cardinality of an attribute set follows the lesser the suppression of data and vice versa. For example, if we are considering 10-anonymity and there are only five entities in the in the age group 30-40, we shall consider the wider range that may be 30-50 or 30-60. Top-down approach can initiate the execution of data from minimum level of suppression. Here the minimum level of suppression may be gender = '*', age = '32', and pin code = '632014'. It is obvious that k-anonymity may not be sufficient to deal with all the privacy concerns. So k-anonymization has been upgraded by other researchers with the techniques namely l-diversity and t-closeness which are applied on sensitive attributes.

- **Definition 4. l-diversity.** The application of l-diversity[16] technique concentrates on the diversification of sensitive entities in an equivalence class.
  A class considered for the application of anonymization techniques is called an equivalence class. So there are l-set of diverse entities in an equivalence class.
- **Definition 5. t-Closeness.** The concept of t-closeness[17] considers that there are at least t-closed entities in an equivalence class from the sensitive entities.

There are advanced works that have been done to improve k-anonymization and l-diversity such as k^m-anonymization[18], (α, k)-Anonymity[19], p-Sensitivity k-Anonymity[20], (k, e)-Anonymity[21], Distinctive l-diversity, Entropy l-diversity, Recursive (c, l)-diversity[16], etc. These techniques focus on the output of the data and check with some critical limits or number of attributes and entities of data which are being published. These techniques provide higher privacy concerns but there may be some natural relations among different attributes that can exploit the vulnerability of sensitive attributes. For example, increasing age can be related to obesity and increasing obesity can indicate the problem of high blood pressure, diabetes, heart diseases, etc. These points increase the probability of estimating vulnerable entities.

3. Analysis of Data Distributions and Attribute Relations

An improved degree of privacy preservation can be achieved with enhanced data analysis. In Table 1, we have observed data distribution in different attributes. Different version of anonymization can be applied on differently distributed data.

3.1 Initial Observations of Data Distribution

Let us observe a summary of different attribute from our considered data set in Table 1.

Our considered data set contains 32,561 entries and it is highly oriented towards Native-country (United–States – 89.6%), Race (White – 85.4) Work-class (Private – 69.7%) and Salary (~76% of population < =50 k). It is already anonymized up to some level of anonymization. Salary is completely anonymized into two instances only. Attributes "Race" and "Occupation" are anonymized with the instances "Other" or "Other-services" respectively.

3.2 Auxiliary Information

If an adversary tries to explore relationships among data attributes, he can find many inferences with the help of some auxiliary information. The prime source of auxiliary information is the Internet. We have collected some facts from Internet which can indicate some relationships among different attributes as follows:

As the age increase, the work experience increases and hence salary may also rise[22]. Same way higher education and job posts or work-class is proportional to higher salary[23,24]. Salary is highly affected by occupation, work experience, employee performance and motivation[25]. Relationships and marital status have remarkable communication between them[26].

There may be numerous relations among different attributes and on the basis of this fact we can also observe some relationships among different attributes of our considered data set which can help us in selection of attributes to apply further level of anonymization.

Table 1. Summary of different attributes of the total dataset

Age	**#inst. (%)**
[10,20)	2410 (7.4)
[20,30)	8162 (25.1)
[30,40)	8546 (26.2)
[40,50)	6983 (21.4)
[50,60)	4128 (12.7)
[60,70)	1792 (5.5)
[70,80)	441 (1.4)
[80,90)	99 (0.3)

Work-class	**#inst. (%)**
?	: 1836 (5.6)
Federal-gov	: 960 (2.9)
Local-gov	:2093 (6.4)
Never-Worked	:7 (0)
Private	: 22696 (69.7)
Self-emp-inc	:1116 (3.4)
Self-emp-not-inc	:2541 (7.8)
State-gov	:1298 (4)
Without-pay	:14 (0)

Native Country	**#inst. (%)**
United-States	:29170(89.6)
Mexico	: 643 (2.0)
?	: 583(1.8)
Philippines	: 198 (0.6)
Germany	: 137 (0.4)
Canada	: 121(0.4)
(Other)	: 1709(5)

Marital-Status	**#inst. (%)**
Divorced	:4443(13.6)
Married-AF-spouse	:23 (0.1)
Married-civ-pouse	:14976(46.0)
Married-spouse-absent	:418(1.3)
Never-married	:10683(32.8)
Separated	:1025 (3.1)
Widowed	:993 (3.0)

Occupation	**#inst. (%)**
?	: 1843 (5.7)
Adm-clerical	: 3770 (11.6)
Armed-Forces	: 9 (0)
Craft-repair	: 4099 (12.6)
Exec-managerial	: 4066 (12.5)
Farming-fishing	: 994 (3.1)
Handlers-cleaners	: 1370 (4.2)
Machine-op-inspct	: 2002 (6.1)
Other-service	: 3295 (10.1)
Priv-house-serv	: 149 (149)
Prof-specialty	: 4140 (12.7)
Protective-serv	: 649 (2.0)
Sales	: 3650 (11.2)
Tech-support	: 928 (2.8)
Transport-moving	: 1597 (4.9)

Education	**#inst. (%)**
10^{th}	: 933 (2.9)
11^{th}	: 1175 (3.6)
12^{th}	: 433 (1.3)
1^{st}-4^{th}	: 168 (0.5)
5^{th}-6^{th}	: 333 (1.0)
7^{th}-8^{th}	: 646 (2.0)
9^{th}	: 514 (1.6)
Assoc-acdm	: 1067 (3.3)
Assoc-voc	: 1382 (4.2)
Bachelors	: 5355 (16.4)
Doctorate	: 413 (1.3)
HS-grade	:10501(32.2)
Masters	: 1723(5.3)
Pre-school	: 51 (0.2)
Prof-School	: 576 (1.8)
Some-college	:7291(22.4)

Sex	**#inst. (%)**
Female	: 10771 (33.1)
Male	: 21790 (66.9)

Relationship	**#inst. (%)**
Husband	: 13193 (40.5)
Not-in-family	: 8305 (25.5)
Other-relative	: 981 (3.0)
Own-child	: 5068 (15.6)
Unmarried	: 3446 (10.6)
Wife	: 1568 (4.8)

Race	**#inst. (%)**
Amer-Indian-Eskimo	: 311 (1.0)
Asian-Pac-Islander	: 1039 (3.2)
Black	: 3124 (9.6)
Other	: 271 (0.8)
White	: 27816(85.4)

Salary #inst. (%)
<=50K : 24720 (75.9)
>50K : 7841 (24.1)

3.3 Attribute Selection to Apply Anonymization

Every attribute from a data set cannot be considered to apply anonymization approach. As per selection of partial identifiable attributes and sensitive attributes[27], we have considered partial identifier attributes and sensitive attributes for our experiment and shown in Table 2.

There are four partial identifier attributes and six sensitive attributes. A data set is published with partial identifiable attributes, sensitive attributes and general attributes. Any subset of sensitive attributes can be considered from the set of six attributes (2^n–1 subsets) for relation calculations. If we calculate degree of relation among different set of sensitive attributes on numeric basis, it will result with some values but there may not be always any relation among some attributes on real life basis.

Table 2. Attribute distributions of data

S. No.	Attributes	Types of Instances	Partial Identifier	Sensitive Attribute
	Age	73	Yes	No
	Work-class	9	No	Yes
	Education	16	No	Yes
	Marital-Status	7	No	Yes
	Occupation	15	No	Yes
	Relationship	6	No	Yes
	Sex	2	Yes	No
	Race	5	Yes	No
	Native-country	42	Yes	No
	Salary	2	No	Yes

4. Application of Anonymization Techniques

We have considered three prime anonymization techniques namely k-anonymization, l-diversity and t-closeness. The application of k-anonymization requires partial identifies or quasi-identifies whereas l-diversity and t-closeness are applied on sensitive attributes. We have followed decision trees to determine the level of anonymization.

4.1 Single Attribute Decision Tree

Tree based taxonomy hierarchy can assist us to classify data on the basis of single and multiple attributes both. Let us consider the data set $D = \{P, S, O\}$, where P is the partial identifier attributes set $P = \{P_1, P_2, ..., P_t\}$, S is the sensitive attributes set $S = \{S_1, S_2, ..., S_u\}$, and O is the set of other attributes irrelevant to anonymization. Partial identifier attributes are associated with sensitive attributes so there is a set of attribute taxonomy hierarchies $\{T_1, T_2, ..., T_t\}$ which consider the generalization level.

Single attribute decision tree are considered for every attributes separately. For single attribute a tree can be observed in Figure 1.

At the top most level, the attribute is most generalized and at the bottom values are in their original form. More levels can be generated by decreasing the intermediate range at different levels as per the availability of more instances. At the intermediate levels range values should be uniform otherwise overlapping case may arise. Every attribute taxonomy hierarchy may not follow the binary tree. Other attributes may follow the maximum number of children nodes as per the types of instances in that particular attribute[28].

4.2 Multiple Attributes Decision Tree

Multiple attributes decision trees are the conglomeration of single attribute decision trees at their best level of individual generalizations. Single attribute decision trees can be associated with proper taxonomy hierarchies[29] to get multiple attribute decision trees. Every attribute at its generalization level should be constructing generalization lattice with other attribute in its domain. We can get optimal normalized mutual information for an optimal node in a generalized lattice. Attribute selection at the different level of multiple attribute decision tree may depend on the number of branches of the different single attribute trees and the relationships among the different attributes. These values can be calculated with entropy, mutual information and normalized mutual information for different attributes.

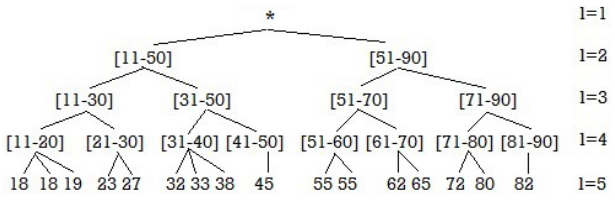

Figure 1. An example of single attribute tree.

4.3 Concerned Formulae for Determination Attributes for Data Anonymization

There are two aspects: The cardinality of different entities in an attribute set and mutual information of those entities with respect to different entities in different attributes. Cardinality can be calculated with entropy and mutual information can be found out with the help of mutual information gain formula.

Let C represent the classes of the data set and $\{c_1, c_2, ...c_i\}$ are the class labels. The entropy associated with c_i can be calculated with the following formula:

$$H(C) = \sum_{i=1}^{m} p\ (c_i) \times \log_2 \frac{1}{p\ (c_i)} = -\sum_{i=1}^{m} p\ (c_i) \times \log_2 p\ (c_i)$$

Where $p\ (c_i)$ are the probability distributions of the entities of class label c_i. Entropy is a non-negative value and it is zero when prediction of the random variable is certain, that is, there is only one type of entries for an attribute. Higher the entropy value relates to the lesser frequently occurrence of instances in the class. Entropy is basically average sum of the probability distribution of each entry in a class.

This formula provides uncertainty calculations of entities from an attribute without considering other entities from different attribute sets. Joint entropy for entities of two class labels (c_i, c_j) is given by:

$$H(c_i, c_j) = -\sum_{i,j=1,1}^{m,n} p\ (c_i, c_j) \times \log_2 p\ (c_i, c_j)$$

Conditional entropy is the measure of entropy of class label c_i when entropy of the class label c_j is already known.

$$H(c_i \mid c_j) = -\sum_{i,j=1,1}^{m,n} p\ (c_i \mid c_j) \times \log_2 p\ (c_i \mid c_j)$$

Mutual information between entities of two class labels is given by the following formula:

$$I(c_i;c_j) = -\sum_{i,j=1,1}^{m,n} p\ (c_i,c_j) \times \log \frac{p\ (c_i,c_j)}{p\ (c_i)p\ (c_j)}$$

Calculation of mutual information also results with non-negative values and higher the value of mutual information infer to the stronger relationship with other entity. Mutual information is reduction of conditional entropy of c_i with known c_j from entropy of c_i. So it can also be written as:

$$I(c_i;c_j) = H(c_i) - H(c_i \mid c_j)$$

And conditional mutual information for three attributes:

$$I(c_i;c_j \mid c_k) = H(c_i \mid c_k) - H(c_i \mid c_j,c_k) =$$

$$-\sum_{i,j,k=1,1,1}^{m,n,p} p\ (c_i,c_j,c_k) \times \log \frac{p\ (c_i,c_j \mid c_k)}{p\ (c_i \mid c_k)p\ (c_j \mid c_k)}$$

For multiple attributes relationships, calculation of mutual information for different subsets of variables thatare called normalized mutual information can be given as follows:

$$I_N(c_1,c_2,...c_{i-1};c_i) = \sum_i I(c_{i-1};c_i \mid c_{i-2},c_{i-3},....,c_1)$$

Let C' represents the new classes after generalization of considered attributes of the data set and $\{c_1',c_2',.....c_i'\}$ are the new class labels. So Kullback-Leibler divergence or relative entropy can be calculated as follows:

$$D(c_i \mid c_i') = \sum_{i=1}^{m} p\ (c_i) \times \log \frac{p\ (c_i)}{p\ (c_i')}$$

It is obvious that $D(c_i \mid c_i') \neq D(c_i' \mid c_i)$.

4.4 Determination of Best Generalization Levels for Data Anonymization

An anonymized table is produced with application of k-anonymization on quasi attributes and l-diversity and t-closeness applied on sensitive attributes. A power set of sensitive attributes can be considered for the calculations of relation among different attributes.

Attribute relations of Partial Identifiers with sensitive attributes should be considered for implementation of data anonymization. Sensitive attributes should be considered as per their minimum entropy first since minimum entropy refers to least entity distribution. On the basis of calculations of entropies of different attributes (Table 3), we have considered {Salary, Work-class} from sensitive attributes set since they are having minimum entropy and {Age, Sex} from quasi attributes to calculate the relations among different attributes set (Table 4).

Table 3. Entropy for each considered attribute

Attribute	Age	Sex	Race	Native Country	Work-class	Education	Marital-status	Occupation	Relationship	Salary
Entropy	1.7181	0.6347399	0.5536448	0.6541891	1.14229	2.031858	1.270989	2.437731	1.49333	0.5520113

Table 4. An example for different level of anonymization

Frequency	Partial Identifiers		Sensitive Attributes			Decision
	Age	Sex	Salary	Work-class	…	
1. Initial Anonymized Table with (k=12)						
5	[11-20]	Female	<=50k	Private	…	Yes
7	[11-20]	Male	<=50k	Private	…	No
2	[21-30]	Male	>50k	Local-gov	…	Yes
1	[21-30]	Male	>50k	Federal-gov	…	Yes
1	[51-60]	Female	<=50k	State-gov	…	No
7	[51-60]	Female	<=50	Private	…	Yes
5	[51-60]	Male	<=50k	Private	…	No

(*continued*)

2. Second level of Anonymization						
5	[11-30]	Female	<=50k	Private	…	Yes
7	[11-30]	Male	<=50k	Private	…	No
2	[11-30]	*	>50k	Local-gov	…	Yes
1	[11-30]	*	>50k	Federal-gov	…	Yes
1	[31-60]	*	<=50k	State-gov	…	No
7	[31-60]	Female	<=50	Private	…	Yes
5	[31-60]	Male	<=50k	Private	…	No
3. Third level of Anonymization						
5	[11-60]	Female	<=50k	Private	…	Yes
7	[11-60]	Male	<=50k	Private	…	No
2	[11-60]	*	>50k	Local-gov	…	Yes
1	[11-60]	*	>50k	Federal-gov	…	Yes
1	[11-60]	*	<=50k	State-gov	…	No
7	[11-60]	Female	<=50	Private	…	Yes
5	[11-60]	Male	<=50k	Private	…	No

Table 5. Information calculation for the three level of anonymization

Level	Entropy (Age)	Entropy (Salary)	Conditional Entropy (Salary| Age)	Mutual Information (Salary; Age)	Normalized Mutual Information
3	0.9586667	0.3404998	0	0.3404998	0.8134687
2	0.690594	0.3404998	0.2680727	0.07242705	0.5453959
1	0	0.3404998	0.3404998	0	0.4369676

On the basis of Table 4, we can get the information basis calculations in Table 5.

At the level 2, we are getting a better tradeoff for normalized mutual information. At the level 2, in Table 4, we have anonymized sex for less frequently occurring entities. At the level 2, if we again analyze the sex attribute, we can anonymize Male and Female with "*" for which frequencies are 2, 1 and 1. These information based calculations can be applied for other set of attributes based on data distributions[30] and an effective way of anonymization can be achieved with a better tradeoff.

5. Conclusion and Future Work

An adverse user can exploit the data vulnerabilities in spite of the existence of different Privacy Preserving Data Mining techniques on the basis of auxiliary information and manual observation of different attributes in the data table. We have analyzed the considered data set on the distribution basis and the relations among different attributes and applied the anonymization techniques. Every attribute and relations among them can be considered to the application of anonymization technique. In this paper, we have presented a way of selection of attributes to the application of data anonymization techniques that can provide us a better tradeoff. We have calculated the relationships among partial attributes and sensitive attributes with the help of entropy and mutual and conditional information gain calculations. In future we shall be applying other advanced methods like support vector machines, neural networks, etc. and analyzing relationships among different attributes and try to apply other privacy preserving techniques to get a better tradeoff between data utility and vulnerability concealment.

6. References

1. UCI Machine Learning Data Repository. Center for Machine Learning and Intelligent System. 2015. Available from: http://archive.ics.uci.edu/ml/datasets/Adult
2. Li J, Wong RCW, Fu AWC, Pei J. Anonymization by local recoding in data with attribute hierarchical taxonomies. IEEE Transactions on Knowledge and Data Engineering. 2008; 20(9):1181–94.

3. Wang H, Liu R. Privacy-preserving publishing microdata with full functional dependencies. Data and Knowledge Engineering. 2011; 70(3):249–68.
4. Zielinski MP, Olivier MS. On the use of economic price theory to find the optimum levels of privacy and information utility in non-perturbative microdata anonymization. Data and Knowledge Engineering. 2010; 69(5):399–423.
5. Xu J, Wang W, Pei J, Wang X, Shi B, Fu AWC. Utility-based anonymization using local recoding. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 06), ACM Press; USA. 2006. p. 785–90.
6. LeFevre K, DeWitt DJ, Ramakrishnan R. Workload-aware anonymization techniques for large-scale datasets. ACM Transactions on Database Systems. 2008; 33(3):17.1–47.
7. LeFevre K, DeWitt DJ, Ramakrishnan R. Mondrian multidimensional k-anonymity. 22nd International Conference on Data Engineering (ICDE 06); 2006. p. 25.
8. Li T, Li N. On the tradeoff between privacy and utility in data publishing. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 09), ACM; USA. 2009. p. 517–26.
9. Wang K, Yu PS, Chakraborty S. Bottom-up generalization: A data mining solution to privacy protection. 4th IEEE International Conference on Data Mining (ICDM04); DC. 2004. p. 249–56.
10. Fung BCM, Wang K, Yu PS. Anonymizing, classification data for privacy preservation. IEEE Transactions on Knowledge and Data Engineering. 2007; 19(5):711–25.
11. Kisilevich S, Rokach L, Elovici Y, Shapira B. Efficient multidimensional suppression for k-anonymity. IEEE Transaction on Knowledge and Data Engineering. 2010; 22(3):334–47.
12. Radhika S, Arumugam S. Improved non mutual information based multi-path time delay estimation. Indian Journal of Science and Technology. 2014 Aug; 7(8):1–6.
13. Irudayasamy A, Arockiam L. Parallel bottom-up generalization approach for data anonymization using map reduce for security of data in public cloud. Indian Journal of Science and Technology. 2015 Sep; 8(22):1–9.
14. Aruna Kumari D, Vineela Y, Mohan Krishna T, Sai Kumar B. Analyzing and performing Privacy Preserving Data Mining on medical databases. Indian Journal of Science and Technology. 2016 May; 9(17):1–9.
15. Sweeney L. k-anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge Based Systems. 2002; 10(5):557–70.
16. Machanavajjhala A, Johannes J, Gehrke G, Daniel D, Kifer K, Muthuramakrishnan M, Venkitasubramaniam V. l-diversity: Privacy beyond k-anonymity. ACM. 2007; 1(1):1–3.
17. Li N, Li T, Venkatasubramanian S. t-Closeness: Privacy beyond k-Anonymity and l-Diversity. 2007; p. 106–15.
18. Terrovitis M, Mamoulis N, Kalnis P. Privacy preserving anonymization of set valued data. VLDB; 2008. p. 115–25.
19. Wong RC, Li J, Fu AW. (α, k)-Anonymity: An enhanced k-anonymity model for Privacy-Preserving Data Publishing. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press; New York. 2006. p. 754–9.
20. Truta TM, Vinay B. Privacy protection: p-Sensitive k-anonymity property. Proceedings of the 22nd on Data Engineering Workshops, IEEE Computer Society; Washington. 2006. p. 1–10.
21. Zhang Q, Koundas N. Aggregate query answering on anonymized tables. Proc of ICDE; 2007 Apr. p. 116–25.
22. Age and earnings. 2015. Available from: http://www.statcan.gc.ca/pub/75-001-x/2009101/article/10779-eng.htm
23. The Connection between Education, income inequality and Unemployment. 2015. Available from: http://www.huffingtonpost.com/entry/the-connection-between-ed_b_1066401.html?section=india
24. Administrators in higher education salaries. 2015. Available from: https://www.higheredjobs.com/salary/salaryDisplay.cfm?SurveyID=30
25. Chaudhry MS, Sabir HM, Rafi N, Kalya MN. Exploring the relationship between salary satisfaction and job satisfaction: A comparison of public and private sector organizations. The Journal of Commerce. 2011; 3(4):1–14.
26. Chipperfield JG, Havens B. Gender differences in the relationship between marital status transitions and life satisfaction in later life. J Gerontol B Psychol Sci Soc Sci. 2001; 56(3):176–86.
27. Li J, Liu J, Baig M, Chi-Wong RCW. Information based data anonymization for classification utility. Data and Knowledge Engineering. 2011; 70(12):1030–45.
28. Frank E, Ian H. Witten. Selecting multiway splits in decision trees. New Zealand: Department of Computer Science, University of Waikato Hamilton; 2013. p. 1–18.
29. Tan PN, Steinbach M, Kumar V. Introduction to data mining. 2015. Available from: http://www-users.cs.umn.edu/~kumar/dmbook/index.php
30. Steuer R, Kurths J, Daub CO, Weise J, Selbig J. The mutual information: Detecting and evaluating dependencies between variables. Oxford Journals Bioinformatics. 2002; 18(2):231–40.