The Pharma Innovation Journal 2017; 6(10): 412-415

www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277- 7695
ISSN (P): 2349-8242
NAAS Rating 2017: 5.03
TPI 2017; 6(10): 412-415

www.thepharmajournal.com
Received: 07-08-2017
Accepted: 08-09-2017

Ishwarya MV
Research Scholar, HITS,
Assistant Professor, CSE
Department Sri Sairam
Engineering College,
West Tambaram, Tamil Nadu,
India

Saptha Maaleekaa S
4th year CSE Department,
Sri Sairam Engineering College,
West Tambaram, Tamil Nadu,
India

Swetha G
4th year CSE Department,
Sri Sairam Engineering College,
West Tambaram, Tamil Nadu,
India

Anu Grahaa R
4th year CSE Department,
Sri Sairam Engineering College,
West Tambaram, Tamil Nadu,
India

Correspondence
Ishwarya MV
Research Scholar, HITS,
Assistant Professor, CSE
Department Sri Sairam
Engineering College,
West Tambaram, Tamil Nadu,
India

Privacy preservation of data using SG and SS models

Ishwarya MV, Saptha Maaleekaa S, Swetha G and Anu Grahaa R

Abstract
The paper aims at preserving the security of the military database when it is mined to judge about the soldier's performance without bias or to obtain statistical information. In the existing methods of privacy preservation, the database is sliced, bucketized and the tuples inside the bucket are shuffled. In addition to these processes some of the values are suppressed in order to prevent the sensitive information from being known to the miners. This masking technique decreases the efficiency of classification. These techniques also consume a lot of time. To overcome these problems, we introduce a technique in which generalization is done only to certain tuples of an attribute and then the table is sliced. In one of the sliced tables selective tuples are shuffled based on an algorithm. By selective generalization, classification can be done efficiently and by selective shuffling, less time is consumed. Thus the proposed technique ensures that the miner can mine efficiently with the table provided and at the same time privacy is preserved.

Keywords: Generalization, Slicing, Bucketisation, Shuffling, Data Mining, Quasi Identifier.

1. Introduction

Data mining is the process of obtaining useful information from a large set of available data. The database is given to a trusted third party so that the useful or statistical information is obtained from the available information. While providing the table to the miner, there is a high possibility that the identity of an individual may be revealed. The quasi identifiers are the group of attributes from which the identity of the person can be revealed. To avoid this leakage of information there are many techniques like generalization, slicing, suppression, shuffling and bucketisation each with its own pros and cons. Data mining is applied in various sectors like healthcare, finance, communication, military and so on**.** Here we present a novel technique for privacy preservation in military database, which can be given to the miner to judge the soldier's performance and at the same time we have ensured that the privacy of each soldier is preserved.

2. Existing Technology

2.1. Slicing:

Slicing is the process by which the table is partitioned in horizontal and vertical manner. Partitioning the attributes vertically by organizing the attributes which are highly dependent, together as columns and tuples are partitioned in horizontal way. In every bucket, values are shuffled in a haphazard way to cut the link. This technique has the advantage that it preserves the privacy to a great extent.

Table 1: Original table

Name	Age	Sex	Salary	No of wars attended	Location of service	Rating by peers(Out of 10)
AAA	24	M	15000	2	CC	7
BBB	36	M	20000	2	GG	5
CCC	27	F	25000	2	EE	9
DDD	32	M	20000	2	GG	6
AAA	27	M	30000	2	CC	10
EEE	24	M	15000	1	CC	8
CCC	30	M	20000	1	GG	8
FFF	24	F	20000	1	BB	7
GGG	36	M	20000	3	GG	7
HHH	27	F	25000	2	EE	9
DDD	24	F	20000	3	BB	7
KKK	29	M	35000	1	CC	8
SSS	33	M	30000	2	CC	8

Table 2: Vertically sliced table

{Age, Sex, Salary}	{No of wars attended, Location of service, Rating by peers (out of 10)}
{24,M,15000}	{2,CC,7}
{36,M,20000}	{2,GG,5}
{27,F,25000}	{2,EE,9}
{32,M,20000}	{2,GG,6}
{27,M,30000}	{2,CC,10}
{24,M,15000}	{1,CC,8}
{30,M,20000}	{1,GG,8}
{24,F,20000}	{1,BB,7}
{36,M,20000}	{3,GG,7}
{27,F,25000}	{2,EE,9}
{24,F,20000}	{3,BB,7}
{29,M,35000}	{1,CC,8}
{33,M,30000}	{2,CC,8}

Table 3: Bucketized table

{Age, Sex, Salary}	{No of wars attended, Location of service, Rating by peers(out of 10)}
{24,M,15000}	{2,CC,7}
{24,M,15000}	{1,CC,8}
{32,M,20000}	{2,GG,6}
{36,M,20000}	{2,GG,5}
{30,M,20000}	{1,GG,8}
{24,F,20000}	{1,BB,7}
{36,M,20000}	{3,GG,7}
{24,F,20000}	{3,BB,7}
{27,F,25000}	{2,EE,9}
{27,F,25000}	{2,EE,9}
{29,M,30000}	{1,CC,8}
{33,M,30000}	{2,CC,8}
{29,M,35000}	{1,CC,8}

2.2. Bucketization
It is the process by which tuples in a table are put into different buckets and thereby separating the quasi-identifiers (QI) from the attributes which
are sensitive by shuffling the sensitive attribute values in each bucket. However the biggest disadvantage of bucketization is membership disclosure and it cannot be applied to the tables that does not have a clear distinguish between sensitive attributes and quasi identifiers.

2.3. Suppression
Suppression is one of the techniques for achieving k-anonymity. In suppression certain values are replaced by asterisk symbol either fully or partly. Though this technique preserves privacy it has poor data utility, that too when applied to data mining algorithms like classification.

Table 4: Table after suppression

{Age, Sex, Salary}	{No of wars attended, Location of service, Rating by peers (out of 10)}
{2*,M,15000}	{2,CC,7}
{2*,M,15000}	{1,CC,8}
{3*,M,20000}	{2,GG,6}
{3*,M,20000}	{2,GG,5}
{3*,M,20000}	{1,GG,8}
{2*,F,20000}	{1,BB,7}
{3*,M,20000}	{3,GG,7}
{2*,F,20000}	{3,BB,7}
{2*,F,25000}	{2,EE,9}
{2*,F,25000}	{2,EE,9}
{2*,M,30000}	{1,CC,8}
{33,M,30000}	{2,CC,8}
{2*,M,35000}	{2,CC,10}

3. Proposed Technique
3.1. Algorithm
Let PT be the private table containing attributes A1,…, An where A1 is the first attribute and An is the last attribute. Let Ai,….,Aj be the set of quasi identifiers of PT such that (Ai,….,Aj) ⊆ (A1,…,An).Let the total number of tuples in PT be denoted as r. Hence let t1,…,tr represent the tuples of PT. The algorithm is as follows:

1. Select the quasi identifier with the highest number of unique values say Am such that Am⊆ Ai,….,Aj.
2. Perform selective generalization on Am as described in points 2.1 to 2.2.

2.1 Let G1,…,Gn be groups such that tuples in each group have same value of Am. The tuples not in any group of G1,….,Gn are generalized.

2.2 For the tuples in G1,….,Gn we consider the remaining quasi identifiers of Ai,….,Aj. For each group in G1 to Gn repeat step 2.2.1. For c in 1 to n in 2.2.1:

2.2.1. For each tuple in Gc repeat steps 2.3.1.1 to 2.3.1.2.

2.2.1.1. For a tuple ensure that it has at least one more tuple in the same group which should have all the quasi identifier values (Ai,….,Aj) same as it. If so go to step 2.2.1. Else go to step 2.2.1.2.

2.2.1.2. Generalize the tuple.

3. For each generalized tuple in PT repeat step 3.1.

3.1. Select tuples which have unique quasi identifier set Ai,….,Aj.

4. Slice PT such that each sliced table contains highly correlated values. Let the sliced tables of PT be B1,….,Bk, such that k is the total number of sliced tables.
5. In the sliced tables select a table Bh in B1,….,Bk such that it has at least one quasi identifier.
6. Perform selective shuffling on the selected table Bh. This is done by shuffling the tuples selected in step 3.

3.2. Selective generalization (SG):
Based on the above algorithm we perform selective generalization to our table to show how it works. In the selected quasi identifier (say in our table age) to generalize we perform selective generalization. Firstly we try to identify the tuples that have the same age value. In the following table the same colored tuples have same age value.

Table 5: Selective generalization

Name	Age	Sex	Salary	No of wars attended	Location of service	Rating by peers(Out of 10)
AAA	24	M	15000	2	CC	7
BBB	36	M	20000	2	GG	5
CCC	27	F	25000	2	EE	9
DDD	32	M	20000	2	GG	6
AAA	27	M	30000	2	CC	10
EEE	24	M	15000	1	CC	8
CCC	30	M	20000	1	GG	8
FFF	24	F	20000	1	BB	7
GGG	36	M	20000	3	GG	7
HHH	27	F	25000	2	EE	9
DDD	24	F	20000	3	BB	7
KKK	29	M	35000	1	CC	8
SSS	33	M	30000	2	CC	8

Now the tuples in black color are unique tuples, each having unique age values. So, such tuples cannot be evicted from generalization. Considering grouped tuples we first check their remaining quasi identifiers (sex, salary, location of service).As per the proposed algorithm in a given group (same color) forevery tuple in a group ensure that it has at least one more tuple in the same group which should have all the quasi identifier values same as it. For example considering red group tuples we can see that the tuples AAA and EEE have same quasi identifier values (24, M, 15000, CC) and the tuples FFF and DDD have same quasi identifier values (24, F, 20000, BB), so we need not generalize it as it can't be identified because of its commonness in all quasi identifier values with at least one more tuple. Considering the yellow group tuples, tuples CCC and HHH have same quasi identifier values (27, F, 25000, EE), which need not be generalized but the tuple AAA having different quasi identifier values (27, M, 30000, CC) from CCC and HHH, need to be generalized. Considering the green group tuples since both of them have different values for the quasi identifier "location of service" we generalize them.

Table 6: Generalization

Name	Age	Sex	Salary	No of wars attended	Location of service	Rating by peers(Out of 10)
AAA	24	M	15000	2	CC	7
BBB	30-40	M	20000	2	GG	5
CCC	27	F	25000	2	EE	9
DDD	30-40	M	20000	2	GG	6
AAA	20-30	M	30000	2	CC	10
EEE	24	M	15000	1	CC	8
CCC	30-40	M	20000	1	GG	8
FFF	24	F	20000	1	BB	7
GGG	30-40	M	20000	3	GG	7
HHH	27	F	25000	2	EE	9
DDD	24	F	20000	3	BB	7
KKK	20-30	M	35000	1	CC	8
SSS	30-40	M	30000	2	CC	8

3.3 Slicing and Selective Generalization

In the above table after performing selective generalization, we can see that some generalized tuples still have unique quasi identifier set which is a threat to privacy. For example tuples like AAA (yellow group) and KKK both have age in the range 20-30, but they differ in the quasi identifier salary which makes them unique and hence identifiable. Similarly SSS also differs in both salary and location with the similar ranged tuples BBB and DDD. So before slicing we select such tuples as per the algorithm. After selection we slice the table using one of the existing slicing algorithms that has the best time efficiency. In the sliced tables we select any table as per our wish (with the constraint that it should have at least one quasi identifier) and shuffle the tuples that we selected before slicing process. By doing selective shuffling we have eliminated the possibility of privacy breech to certain records that the possibility of being identified (eg records like SSS, KKK) even after the generalization process. Moreover selective generalization consumes less time as compared to full generalization as no existing shuffling algorithm can guarantee a time efficiency of O (1) and hence the time efficiency of shuffling process depends on input size.

Table 7: Selection of tuples to be shuffled

Name	Age	Sex	Salary	No of wars attended	Location of service	Rating by peers(Out of 10)
AAA	24	M	15000	2	CC	7
BBB	30-40	M	20000	2	GG	5
CCC	27	F	25000	2	EE	9
DDD	30-40	M	20000	2	GG	6
*AAA	20-30	M	30000	2	CC	10
EEE	24	M	15000	1	CC	8
CCC	30-40	M	20000	1	GG	8
FFF	24	F	20000	1	BB	7

GGG	30-40	M	20000	3	GG	7
HHH	27	F	25000	2	EE	9
DDD	24	F	20000	3	BB	7
*KKK	20-30	M	35000	1	CC	8
*SSS	30-40	M	30000	2	CC	8

Tuples with asterisk are selected

Table 8: Sliced Tables

{Age, Sex, Salary}	{No of wars attended, Location of service, Rating by peers(out of 10)}
{24,M,15000}	{2,CC,7}
{30-40,M,20000}	{2,GG,5}
{27,F,25000}	{2,EE,9}
{30-40,M,20000}	{2,GG,6}
*{20-30,M,30000}	{2,CC,10}
{24,M,15000}	{1,CC,8}
{30-40,M,20000}	{1,GG,8}
{24,F,20000}	{1,BB,7}
{30-40,M,20000}	{3,GG,7}
{27,F,25000}	{2,EE,9}
{24,F,20000}	{3,BB,7}
*{20-30,M,35000}	{1,CC,8}
*{30-40,M,30000}	{2,CC,8}

Table 9: After selective shuffling

{Age, Sex, Salary}	{No of wars attended, Location of service, Rating by peers(out of 10)}
{24,M,15000}	{2,CC,7}
{30-40,M,20000}	{2,GG,5}
{27,F,25000}	{2,EE,9}
{30-40,M,20000}	{2,GG,6}
*{20-30,M,30000}	{2,CC,8}
{24,M,15000}	{1,CC,8}
{30-40,M,20000}	{1,GG,8}
{24,F,20000}	{1,BB,7}
{30-40,M,20000}	{3,GG,7}
{27,F,25000}	{2,EE,9}
{24,F,20000}	{3,BB,7}
*{20-30,M,35000}	{2,CC,10}
*{30-40,M,30000}	{1,CC,8}

The second sliced table is selectively shuffled.

4. Conclusion

By selective generalization the loss of information is reduced. Classification can also be done efficiently since only selected values are generalized. Since even after generalization some tuples are unique, shuffling is performed for those tuples to enhance the privacy. Time consumed for shuffling depends upon the number of records to be shuffled. Hence by selective shuffling we reduce the time required for shuffling by a considerable amount. Thus this technique guarantees both data utility and data privacy and can be applied to high dimensional data. This method can be further enhanced by devising a technique which can be applied to table consisting of a quasi-identifier in which all the values are unique.

References

1. Preet Chandan Kaur, Tushar Ghorpade, Vanita Mane. Analysis of Data Security by using Anonymization Techniques, 6th International Conference - Cloud System and Big Data Engineering (Confluence), 2016.
2. Samarati P, Sweeney L. Protecting privacy when disclosing information: k anonymity and its enforcement through generalization and suppression," In Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, 1998, 1-191998.
3. Sweeney L. k-anonymity: a model for protecting privacy," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems. 2002; 10(5):557570,
4. Xuhua Ding, Yanjiang Yang, Robert H. Deng, Database Access Pattern Protection Without Full-Shuffles, IEEE transactions on information forensics and security, 2011; 6(1).
5. KeWang. (Simon Fraser University), Philip S. Yu (IBM T. J. Watson Research Center), Sourav Chakraborty (Simon Fraser University), Bottom-Up Generalization: A Data Mining Solution to Privacy Protection.
6. He Y, Naughton J. Anonymization of Set-Valued Data via Top-Down, Local Generalization, Proc. Int Conf. Very Large Data Bases (VLDB) 2009, 934-939.
7. Friedman A, Wolff R, Schuster A. Providing k-Anonymity in Data Mining, Intl J Very Large DataBases, 2008; 17(4):789-804.
8. Bayardo RJ, Agrawal R. Data Privacy through Optimal k- Anonymization, in Proc. of ICDE, 2005, 217-228.
9. Neha V Mogre, Girish Agarwal, Pragati Patil. A Review on Data Anonymization Technique for Data Publishing" Proc. International Journal of Engineering Research & Technology (IJERT). 2012; 1(10) ISSN: 2278-0181.
10. Machanavajjhala J, Gehrke D, Kifer M. Venkitasubramaniam, l-Diversity: Privacy Beyond k-Anonymity." in Proc. of ICDE, 2006.