I J C T A, 9(28) 2016, pp. 357-361
© International Science Press

ISOLATION AND CONSERVATION PRACTICE IN DATA MINING

Ishwarya M.V* K. Ramesh kumar G Sruthi*** and P.S. Kshama*****

Abstract: This paper focuses on the methods to improve data utility to handle high dimensional data by using imbricating slicing method. Large databases are needed to collect and store data for large organizations. There exist several attacks in the records and attributes in a table. Privacy preservation is the important technique used to preserve sensitive information in a database. The main aim is to make the database more private and send it to the other party without any intruder involving in it. The existing techniques had disadvantages that are overcome by our proposed technique. We have combined the imbricating slicing and partial shuffling for more security and data utility with reduced time complexity. The proposed system is efficient and reliable because the database is partially shuffled and sliced into two parts. Whereas, slicing has more advantage over the latter as it preserves data utility.

Keywords: Data mining, Data anonymization, Privacy Preservation, overlapping slicing.

1. INTRODUCTION

Mining of the data is defined as the investigating and scrutinizing of data in various aspects and encapsulating it as a meaningful information. In Today's world, knowledge discovery is utilized by many organizations in different areas like retail marketing, financial, healthcare, communication, and global marketing. Retrieval of hidden foretelling information from bulky information is a new idea with great potential in the data warehouses and big data. It is a process which gets proceeded after data collection [3] where a data holder would give the collected data to a miner who will thereby manage to perform mining on the published data.

2. MATERIALS AND METHODS

2.1 Privacy Preservation

People are aware of the thefts to their data and are recommending the need to avoid disclosure (2).This decreases the efficient results of mining. In the most basic form of privacy-preserving of data, the data holder has specific identifiers of their own.(3)They are Explicit Identifier, Quasi Identifier, Sensitive Attributes, non-Sensitive Attributes, where Explicit Identifier are attributes that are record owners. Quasi Identifier is an attribute that is helpful in determining the owners which are not individually a primary key. Sensitive Attributes are in the sensitive information to a particular person where one must avoid disclosure and Non-Sensitive Attributes includes all attributes that are not considered above.

[Table 1] illustrates the original table in which the data is needed to be interpreted to avoid disclosure. (9).

* Research Scholar, HITS, Assistant Professor, CSE Dept, Sri SaiRam Engineering College, WestTambaram,TamilNadu,India Email: ishwarya.cse@sairam.edu.in

** Associate Professor, IT Dept, HITS, Tamil Nadu, India Email: krkumar@hindustanuniv.ac.in

*** B.E Computer science and Engineering, Sri Sairam Engineering College, West Tambaram, Tamil Nadu, India

Table 1

Full Name	*Sex*	*Age*	*Pin Code*
Alex	M	32	431278
Balu	M	36	431288
Caryl	F	30	431267
David	M	31	431333
Elec	F	35	431444

2.2 Data anonymization

Data anonymization is a type of technique in which the information is sanitized and the privacy is protected. It is the process of encrypting the sensitive information of the owner. (4) It is the technology that converts data into non readable form.The anonymization techniques generally classified are generalization and bucketization.

2.3 Generalization

To prevent the attack based on leakage of data due to the group of members having same set and same attribute and background knowledge attack, generalization technique is used.

In generalization technique the explicit identifiers are removed and the Quasi Identifiers are suppressed to a range of values. In this way, publishing of records are more secure and it preserves the sensitive information of the individuals. (5) To make generalization more effective the records are made close to each other in each bucket. They are permutated in such a way that the data are combined based on their correlation. (10) In high dimensional data, the uncorrelated data have some distances between them, because of that the correlation between them is lost. The disadvantage of this method is even huge amount of data loss which may lead to the guessing of the values without high dimensionality.

Table 2.
Illustrates the Table that is Generalized using this Technique.

Explicit attribute	*Quazi attribute*		*Sensitive attribute*
	Gender	Age	431***
Alex	M	30-36	431***
Balu	M	30-36	431***
Caryl	F	30-36	431***
David	M	30-36	431***
Elec	F	30-36	431***

2.4 Bucketization

In bucketization technique, the sensitive attributes are normally permuted randomly to fill in each of the buckets. The tuples are thus partitioned horizontally within each bucket. The identifying attributes can be masked which are considered as a primary attribute and the other sensitive attributes can be partially masked. Bucketization does not forestall the privacy from being opened. It uses the identifiers in original forms without any disclosure. The separation between the sensitive identifier and Quazi identifier is necessary to break the attribute correlation by making the peeping of third party difficult. The buckets are identified based on their B-ID. The disadvantage of this method is that the identification of Quazi and sensitive attributes is difficult in many areas and it also breaks the entire correlation between them.

Table 3.
Illustrates the Bucketized Table Using This Method.

Full Name	*Gender*	*Age*	*Pin code*	*BID*
Alex	M	32	431278	1
Balu	M	36	431288	2
Caryl	F	30	431267	2
David	M	31	431333	3
Elec	F	35	431444	3

3. RESULTS AND DISCUSSIONS

3.1 Slicing

The anonymization technique called slicing is the most efficient privacy preservation technique as compared to existing methods. The overlapping slicing technique from [3] is improvised for better efficiency.

3.2 Existing method

The procedure for slicing the attributes is by dividing the highly correlated records and it is Vertical Slicing. In horizontal partitioning, the grouping of tuples into buckets is implemented. The tuples inside the buckets are randomly sorted within the buckets to break the connection between the correlated columns. Slicing protects privacy because of sorting randomly where the set of records contain less risk. It does not separate the sensitive attributes and the Quasi attributes, it groups them in a random manner to avoid the correlation between them. This technique is followed by partial shuffling of the records [1].

Table 4.
Illustrates the Existing Overlapping Slicing Technique

Age, Gender	Pin Code, Gender
(22, M)	(14589, M)
(22, F)	(14586, F)
(33, M)	(14587, M)
(54, M)	(14588, M)
(56, M)	(145587, M)
(60, F)	(14577, F)
(60, M)	(14788, M)
(64, M)	(14566, M)

The [Table 4] has the following tuple where (33,M) can have 4 possible ways to correlate with. Therefore, a set of 4 tuple can have 16 possible ways to correlate with the other tuple.

3.3 Disadvantage of existing slicing technique:

Slicing technique might be considered to be the best because of the privacy risks with other techniques but their risks can lead to loss of security. The main problem of this type of risks is the background knowledge (6) which can be used to easily guess the records of a individual. Attribute partitioning divided the attributes in such a way that the divided columns are in correlated manner. Therefore, the privacy is not maintained. The uncorrelated attributes provide more risks as compared to correlated ones. The uncorrelated tuples are divided to store into the buckets randomly to protect the disclosure but the background knowledge is the main disadvantage in the earlier method even though if attribute partitioning holds good..

4. PROPOSED ALGORITHM- IMBRICATING SLICING TECHNIQUE

In the proposed method, the tables of the database are sliced into two halves. The proposed slicing process is executing in the O (1) irrespective of the size of the database. The sliced upper half of the table and the sliced lower right of the table is shuffled separately. Only the black records (touched records) are getting shuffled without touching the white records (untouched records) in both the parts of the table. The limitation in the existing shuffling process is that it may increase the computation cost so the advanced technique is introduced.

Table 5. Illustrates the imbricating Slicing Table.

Age, Gender	*Pincode, Disease*
(22, M)	(14589, cancer) White Record
(22, F)	(14586, flu) Black Record
(33, M)	(14587, Polycycstic) Black Record
(45,M)	(14588, flu) White Record
(56, M)	(14558, cancer) Black Record
(60, F)	(14577, bronchitis) Black Record
(60, M)	(14788,bronchitis) White Record
{64,M)	(14566, flu) Black Record

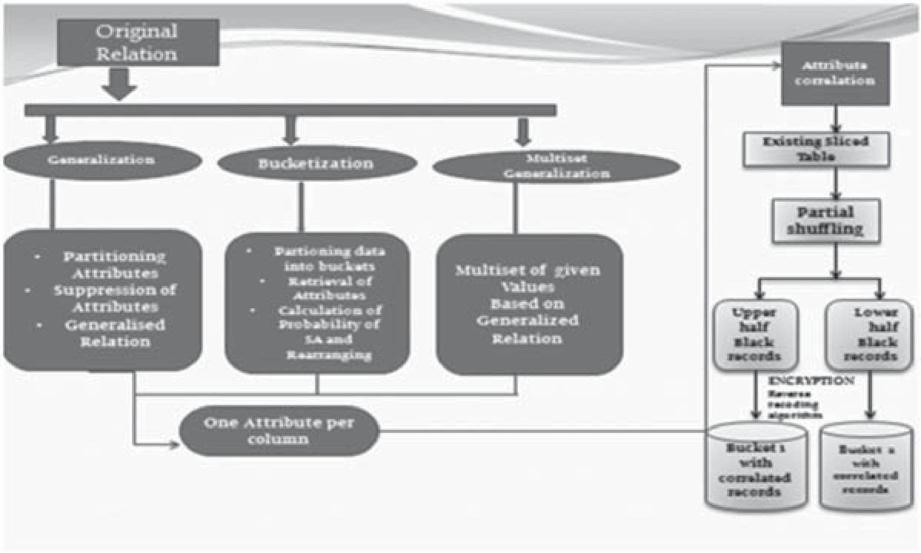
Original Relation
Generalization
Bucketization
Multiset Generalization
• Partitioning Attributes
• Suppression of Attributes
• Generalised Relation
• Partioning data into buckets
• Retrieval of Attributes
• Calculation of Probability of SA and Rearranging
Multiset of given Values Based on Generalized Relation
One Attribute per column
Attribute correlation
Existing Sliced Table
Partial shuffling
Upper half Black records
Lower half Black records
ENCRYPTION
Reverse encoding algorithm
Bucket 1 with correlated records
Bucket 2 with correlated records

Figure 1. Illustrates the Architecture of the Proposed System- Partial Shuffles with imbricating slicing.

The shuffling technique takes the complexity of O(logn). The two main reasons for not shuffling the white records are 1: It is not necessary to shuffle the untouched records. 2: As the records are untouched there is no way to access them illegally with their pattern.

The existing table having unique attribute can easily identify the matching record even if it is shuffled. In the above table, polycystic is the disease caused in women which can affect the pregnancy and fertility. Considering the older technique, as the disease can only affect women it's easy for the intruder who has a background knowledge to easily find out, if there is only one women in the tuple and this may affect privacy.The proposed system overcomes this advantage by even slicing and shuffling the fields that are to be shared with the other parties without the intruder looking into the table.

5. CONCLUSION

The concept of imbricating slicing overcomes the drawbacks of overlapping slicing thus by improving the privacy of data. The partial shuffling and imbricating slicing is the best privacy enhancement technique to improve the efficiency. As the slicing concept can even handle high dimensional data it is better to use the advanced slicing techniques. The database with partial shuffles reduces the cost of time and space by introducing only the black records to be shuffled.

References

1. Xuhua Ding, Yanjiang Yang, and Robert H. Deng ,” Database Access Pattern Protection Without Full-Shuffles**”,** IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 6, NO. 1, MARCH 2011.
2. Avinash Kumar Singh,Narayan P Keer,AnandMotwani,”A Review Of Privacy Preservation Technique”,NRI Institute of Research And Technology ,Bhopal,India.International Journal Of Computer Applications(0075-8887),Volume 90-No.3,March 2014.
3. Amar Paul Singh ,Mr.DhanshriParihar“A Review of Privacy Preserving Data Publishing Technique“,School of CSE Bahra University Shimla Hills, India. Research Article June 2013 Amar Paul Singh Page 32 International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-2, Issue-6)
4. 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
5. KeWang(Simon Fraser University), Philip S. Yu(IBM T. J. Watson Research Center), SouravChakraborty(Simon Fraser University),” Bottom-Up Generalization: A Data Mining Solution to Privacy Protection “.
6. D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J.Y. Halpern, “Worst-Case Background Knowledge forPrivacy-Preserving Data Publishing,” Proc. IEEE 23rd Int’l Conf. Data Eng. (ICDE), pp. 126-135, 2007.
7. Y. He and J. Naughton, “Anonymization of Set-Valued Data via Top-Down, Local Generalization,” Proc.Int Conf.Very Large DataBases(VLDB),pp-934-939,2009.
8. Priti V. Bhagat and 2Rohit Singhal, “An Overview of sectional Shuffle for Database Access Pattern Protection Using Reverse Encryption Algorithm”, 1M.Tech student, Department of Computer Science & Engineering, I.E.T. Alwar, Rajasthan, India ,2Astt. Professor, Department of Computer Science & Engineering, I.E.T. Alwar, Rajasthan, India. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, No 5, May 2013.
9. X. Xiao and Y. Tao, “Anatomy: Simple and Effective Privacy Preservation,” Proc. Int’l Conf. Very Large Data Bases (VLDB), pp. 139-150, 2006.
10. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Incognito: Efficient Full-domain k- Anony Anonymity,”inProc of ACMSIGMOD, 2005, pp. 49–60.