ZIBELINE INTERNATIONAL™ PUBLISHING

ISSN: 2590-4043
CODEN: AEMCDV

Acta Electronica Malaysia (AEM)

DOI: http://doi.org/10.26480/aem.01.2023.01.03

RESEARCH ARTICLE

AN EVALUATION OF DATA ANONYMIZATION METHODS FOR DATA PUBLISHING

Monika Singh

Faculty of information technology, Gopal Narayan Singh University.
****Corresponding Author Email: singhmoni@gmail.com***

ARTICLE DETAILS

Article History:

Received 07 December 2022
Revised 11 January 2023
Accepted 14 February 2023
Available online 23 February 2023

ABSTRACT

Data anonymization approaches have been the focus of research recently for several types of structured data, including tabular, graph, and item set data. In this article, we provide a succinct yet thorough assessment of a number of anonymization approaches, including generalisation and bucketization, which have been created for publishing microdata while protecting privacy. Recent research has demonstrated that generalisation results in significant information loss, particularly for high-dimensional data. Bucketization, however, does not stop membership disclosure. While slicing both prevents membership disclosure and preserves the data's superior utility than generalisation. The practical methods that can be employed to provide improved data utility and handle high-dimensional data are the main emphasis of this research.

KEYWORDS

Data anonymization, Privacy preservation, Data publishing, Data security

1. Introduction

Data mining, also known as Knowledge Discovery Data (KDD), is the process of reviewing data from various angles and condensing it into helpful knowledge. Many businesses with a strong consumer emphasis, such as retail, financial, communication, and marketing enterprises, use data mining nowadays. A strong new technique that has the potential to greatly assist businesses in focusing on the most crucial data in their data warehouses is the extraction of hidden predictive information from massive datasets. Knowledge discovery from databases uses a variety of methods and techniques, such as classification, clustering, regression, artificial intelligence, neural networks, association rules, decision trees, genetic algorithms, nearest neighbour method, etc. Data mining has become widely employed in recent years in a variety of scientific and engineering fields, including bioinformatics, genetics, medicine, education, and electrical power engineering. Data mining is all about gaining knowledge, and it has been claimed that knowledge is power. The ability to make strategic decisions, which will ultimately lead to the success of a business or organisation, depends on the accumulation of pertin

1.1 Anonymization of Data

In recent years, data anonymization techniques for privacy-preserving data posting have drawn a lot of interest. Information on a person, a family, or an organisation can be found in detailed data, also referred to as microdata. Generalization and bucketization are the most used anonymization methods (Tiangcheng et al., 2012). Each record contains a number of attributes that fall under the following categories: 1) The attributes that can be used to uniquely identify people are identifiers like Name or Social Security Number. 2) Some characteristics may be sensitive characteristics (SAs), such as disease and salary, and 3) Some characteristics may be quasi-identifiers (QIs), such as zip code, age, and sex, whose values may be used to identify a particular person.

Data that the recipient of the information cannot use to identify the patient. Any information that, when combined with other data kept by the receiver or divulged to them, could be used to identify the patient must be redacted, including the patient's name, address, and complete postcode. Only if the recipients of the data lack access to the "key" to trace the patient's identity may unique numbers be included. Technology, such as but not limited to preimage resistant hashes (such as one-way hashes) and encryption methods in which the decryption key has been lost, that transforms clear text data into an unreadable and irreversible form. Even when linked together with pointer or pedigree values that point users to the original system, record, and value (such as when supporting selective revelation), data is still regarded as anonymized. This also holds true when anonymized records can be linked together with, matched against, or combined with other anonymized records. While lowering the risk of unintentional disclosure, data anonymization allows the transfer of information across boundaries, such as between two departments within an agency or between two agencies, and in some environments, does so in a way that enables evaluation and analytics after anonymization.

2. Background

The following step is where the two methods diverge (Tiangcheng et al., 2012). To prevent tuples in the same bucket from being discriminated against by their QI values, generalization converts the QI-values in each bucket into "less specific but semantically consistent" values.

QI standards. In bucketization, the SA values in each bucket are randomly permuted to separate the SAs from the QIs.

A number of buckets with permuted sensitive attribute values make up the anonymized data.

The remainder of the essay is structured as follows: Background of the two primary privacy-preserving paradigms is described in Section II. The several methods of data anonymization for data publishing with privacy protection are described in Section III. Comparing the slicing technique with generalisation and bucketization is covered in Section IV of the outline. This essay is concluded in Section V.

There are two basic paradigms for protecting privacy that have been established: l-diversity and k-anonymity, which both prevent the linkage of a specific record with a sensitive attribute value (Tiangcheng et al., 2012; Martin et al., 2007).

Quick Response Code	Access this article online	
	Website: www.actaelectronicamalaysia.com	**DOI:** 10.26480/mecj.01.2023.01.03

Cite The Article: Monika Singh (2023). An Evaluation of Data Anonymization Methods for Data Publishing. *Acta Electronica Malaysia, 7(1): 01-03.*

2.1 k-anonymity

When attributes are suppressed or generalised until every entry is identical to at least k-1 other rows, the database is said to be K-anonymous. Therefore, K-Anonymity prevents concrete database links. The accuracy of the data released is ensured by K-Anonymity. The K-anonymity proposal is particularly interested in two methods: generalisation and suppression (Ciriani et al., 2007). When sharing microdata, data owners frequently delete or encrypt explicit identifiers like names and social security numbers in order to safeguard respondents' identities. De-identifying information, however, does not ensure anonymity. Released data frequently includes additional information that can be linked to publicly accessible data to re-identify respondents and deduce information that was not intended for release. Examples of such additional data include birth date, sex, and ZIP code. K-anonymity, a recently proposed feature that encapsulates the protection of a microdata table with respect to potential re-identification of the respondents to which the data pertain, is one of the emerging concepts in microdata protection. Every tuple in the disclosed microdata table must be indistinguishably associated to at least k respondents in order to maintain k-anonymity. K-connection anonymity's to security measures that maintain the veracity of the data is among its intriguing features. In order to ensure privacy in data mining, the input (the data) was first perturbed before mining. The perturbation approach's flaw is that it doesn't have a formal foundation for demonstrating how much privacy is guaranteed. A second branch of privacy-preserving data mining was created at the same time, employing cryptographic methods. In light of this, it falls short of offering a comprehensive solution to the issue of privacy-preserving data mining. K-anonymity is one notion of privacy that has made significant progress in the public sphere and is currently accepted by both legislators and businesses (Sweeney, 2002). K-anonymity ensures that no information can be connected to groups of fewer than k people. Generalization for losses in k-anonymity

2.2 Substantial Amount of Data, Particularly for High-Dimensional Data.

The limitations of k-anonymity are that it cannot be used on high-dimensional data without completely losing its usefulness, it cannot be applied to high-dimensional data without revealing sensitive attributes of individuals, it cannot protect against attacks based on background knowledge, and it requires special methods if a dataset is anonymized and published more than once (Brickell and Shmatikov, 2008).

2.3 l- diversity

The term "l-diversity" comes next. Consider a collection of k unique records that are connected by a common quasi-identifier. This is advantageous since it prevents an attacker from using the quasi-identifier to locate the victim. However, what if the value they're interested in—for instance, the person's medical diagnosis—is shared by all the values in the group? "L-diversity" refers to the distribution of target values within a group (Ghinita et al., 2008). Currently, generalization and permutation-based approaches fall into two major types. A generalisation technique already in use would divide the data into discrete groups of transactions, each group including enough records with l-distinct, accurately represented sensitive elements.

2.4 Ii-Different Techniques for Anonymization

Generalization and bucketization are two often explored data anonymization methods. The fact that bucketization does not generalise the QI features is the primary distinction between the two anonymization methods.

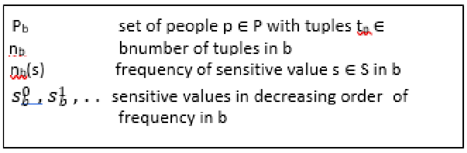
2.4.1 Generalization

One popular anonymous strategy is generalisation, which substitutes less-specific but semantically coherent values for quasi-identifier values. The full group extend in the QID space would thus be generalised to include all quasi-identifier values in a group (Ghinita et al., 2011). All data pertaining to a particular item in the current group is lost if at least two transactions in a group have different values in the same column (i.e., one contains the item and the other does not). All potential items in the log are covered by the QID utilised in this step. It is likely that any generalisation method would result in extremely substantial information loss, leaving the data unusable [8] due to the high dimensionality of the quasi-identifier, with the number of potential items in the range of thousands (Ghinita et al., 2008). Records in the same bucket must be close to each other for generalisation to be successful.

Generalising the records would prevent too much information from being lost. The majority of data points in high-dimensional data, however, are close to one another. No other distribution assumption can be justified, hence the data analyst must adopt the uniform distribution assumption that every value in a generalised interval or set is equally possible in order to execute data analysis or data mining operations on the generalised table. The data utility of the generalised data is dramatically decreased as a result. Additionally, correlations between several traits are lost as a result of each variable being generalised separately. The data analyst must make the assumption that any potential combination of attribute values is equally possible in order to investigate attribute correlations on the generalised table. This intrinsic generalisation issue makes it unable to analyse attribute correlations effectively.

2.4.2 Bucketization

The first method, which we refer to as bucketization, is dividing the tuples in T into buckets before separating the sensitive attributes from the non-sensitive ones by randomly permuting the sensitive attribute values within each bucket. The buckets with permuted sensitive values are what make up the sanitised data after that. Although all of our findings also apply for full-domain generalisation, in this paper we employ bucketization as the technique for creating the published data from the original table T (Martin et al., 2007). We now formalise our definition of bucketization. The tuples are divided into buckets (i.e., the table T is horizontally partitioned using some technique), and the column containing the S-values is subjected to an independent random permutation within each bucket. The resulting bucket set, designated by B, is then made public. For instance, the publisher might publish bucketization B if table T is the underlying table. Of course, the publisher can entirely conceal the identifying attribute (Name) and conceal some of the other non-sensitive attributes in order to increase privacy (Age, Sex, Zip). We use the following notation for a bucket b B.



P_b	set of people p ∈ P with tuples t_p ∈
n_b	bnumber of tuples in b
$n_b(s)$	frequency of sensitive value s ∈ S in b
$s_b^0, s_b^1, \ldots$	sensitive values in decreasing order of frequency in b

Although bucketization provides more useful data than generalisation, it has several drawbacks (Martin et al., 2007; Tiancheng et al., 2012). First off, membership disclosure is not prevented by bucketization. Bucketization releases the QI values in their original forms, making it possible for an enemy to determine whether or not a particular person has a record in the released data (He and Naughton, 2009). As demonstrated in, 87 percent of Americans can be uniquely recognised using just three characteristics (Birthdate, Sex, and Zipcode). A microdata typically includes many more features than just those three (for example, census data). This indicates that the bucketized table can be used to deduce the membership information for the majority of people.

Second, a distinct division between QIs and SAs is necessary for bucketization. The distinction between QIs and SAs might be difficult to make in many data sets. Third, bucketization breaks the attribute correlations between the QIs and the SAs by separating the sensitive attribute from the QI attributes.

The process of bucketization divides the table's tuples into buckets before separating the quasi-identifiers from the sensitive attribute by randomly permuting the values of the sensitive attribute in each bucket. A number of buckets with permuted sensitive attribute values make up the anonymized data. Bucketization has been utilised particularly for high-dimensional data anonymization (Lefevere et al., 2005; 2006). However, their strategy presupposes a distinct division between QIs and SAs. Additionally, membership information is made public because all QIs are provided with their exact values.

2.4.3 Slicing

In this research, we offer a unique data anonymization method called slicing to advance the current state of the art (Tiancheng et al., 2012). The data set is divided both vertically and horizontally by slicing. By organising qualities into columns based on how they relate to one another, vertical partitioning is accomplished. A subset of highly associated attributes are present in each column. The process of horizontal partitioning involves dividing tuples into buckets. To disrupt the connection between various columns, values in each bucket are then randomly permuted (or sorted). Slicing's fundamental goal is to destroy associations across columns while preserving associations inside each individual column. In comparison to generalisation and bucketization, this lowers the data's dimensionality while maintaining a higher level of utility.

Because it combines highly linked qualities together and maintains the relationships between them, slicing maintains utility. Because it destroys the linkages between uncorrelated qualities, which are rare and therefore identifying, slicing protects privacy. Be aware that when a data set contains both a QI and a single SA, bucketization must destroy their correlation whereas slicing can combine some QI attributes with the SA while maintaining attribute correlations with the sensitive attribute (Li et al., 2007). The main assumption that slicing protects privacy is that it makes sure that there are typically several matching bins for every tuple (Machanavajjhala et al., 2006; Bayardo and Agrawal, 2005). Attributes are first divided into columns via slicing. A subset of attributes are present in each column. dividing tuples into buckets while slicing. A subset of tuples are contained in each bucket. The table is divided horizontally as a result. To break the connection between several columns, values in each bucket are permuted randomly.

3. Discussion

The slicing strategy, which is superior to generalisation and bucketization for high dimension data sets, is compared in our discussion.

3.1 In Contrast to Generalisation

We want to be clear that our goal is not to do away with generalisation; there is no doubt that The fact that generalisation has received a lot of attention in the literature serves as evidence that it is an important approach. Instead, we want to offer a different privacy preservation solution that has its own benefits because it can keep more data features (Xiao and Tao, 2006; Xu et al., 2006). Indeed, anatomy doesn't always come out on top. It makes sense that anatomy would permit a higher breach probability than generalisation because it releases the QI-values directly. Nevertheless, as long as an adversary's background information does not exceed the amount permitted by the l-diversity model, such likelihood is always constrained by 1/l (Kifer and Gehrke, 2006). There are various generalisation recoding types. Local recoding is the type of recoding that safeguards the most data. In local recoding, tuples are first divided into buckets, and for each bucket, all of one attribute's values are then replaced with a generalised value. Since the same property value may be generalised differently depending on which bucket it appears in, such recoding is local. We now demonstrate that, if the same tuple partition is employed, slicing preserves more information than such a local recoding strategy. To do this, we demonstrate that slicing is superior to the next improvement of the local recoding method. One uses the multiset of exact values in each bucket rather than replacing more specific attribute values with a generalist value (Tiancheng et al., 2012). Two significant issues with generalisation are that it loses too much information due to the uniform-distribution assumption and fails on high-dimensional data due to the curse of dimensionality.

3.2 In Contrast to Bucketization

Following are some advantages of slicing over bucketization: First, membership disclosure can be avoided by using slicing to divide attributes into more than two columns (Tiangcheng et al., 2012; Martin et al., 2007; Xu et al., 2008). Our empirical analysis of an actual data set demonstrates that membership disclosure is not prevented by bucketization. Second, slicing can be utilised without a clear division between the sensitive attribute and the QI attributes, in contrast to bucketization, which demands it. Because there isn't a single external, publicly accessible database that can be used to establish which qualities the adversary already knows for a data collection like the census data, it is frequently difficult to distinguish between QIs and SAs (Dwork, 2008; Li and Li, 2009). Such data may benefit from slicing. The attribute correlations between the sensitive attribute and the QI attributes are also kept by enabling a column to contain both certain QI attributes and the sensitive attribute.

4. Conclusion

Handling high-dimensional data is a significant scientific issue. Generalization and bucketization are two common methods for data anonymization, according to the comparison above. These methods are intended for disseminating microdata while protecting privacy. Recent research has nevertheless demonstrated that generalisation results in significant information loss for high dimensional data. buckling up, then nonetheless, does not stop membership disclosure and does not apply to information where sensitive and quasi-identifying qualities are not clearly separated. On the other side, transaction databases can be made anonymous through slicing. Slices can be utilised for membership disclosure protection and preserve data utility better than generalisation. Slicing has the ability to handle high-dimensional data, which is an essential additional benefit.

References

Bayardo, R.J., and Agrawal, R., 2005. Data Privacy through Optimal k-Anonymization. in Proc. of ICDE, 2005, pp. 217–228.

Brickell, J., and Shmatikov, V., 2008. The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing. Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), Pp. 70- 78.

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

Dwork, C., 2008. Differential Privacy: A Survey of Results. Proc. Fifth Int'l Conf. Theory and Applications of Models of Computation (TAMC), Pp. 1-19.

Ghinita, G., Kalnis, P., Tao, Y., 2011. Anonymous Publication of Sensitive Transactional Data. in Proc. Of IEEE Transactions on Knowledge and Data Engineering, 23 (2), Pp. 161-174.

Ghinita, G., Tao, T., and Kalnis, P., 2008. On the Anonymization of Sparse High-Dimensional Data. Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), Pp. 715-724.

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

Kifer, D., and Gehrke, J., 2006. Injecting Utility into Anonymized Data Sets. Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), Pp. 217-228.

LeFevre, K., DeWitt, D.J., and Ramakrishnan, R., 2005. Incognito: Efficient Full-domain k-Anonymity. in Proc. of ACM SIGMOD, Pp. 49– 60.

LeFevre, K., DeWitt, D.J., and Ramakrishnan, R., 2006. Mondrian Multidimensional k-Anonymity. in Proc. of ICDE.

Li, N., Li, T., and Venkatasubramanian, S., 2007. t-Closeness: Privacy Beyond k-Anonymity and „-Diversity. Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), Pp. 106-115.

Li, T., and Li, N., 2009. 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.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M., 2006. l-diversity: Privacy beyond k- anonymity. In ICDE.

Martin, D., Kifer, D., Machanavajjhala, A., Gehrke, J., and Halpern, J., 2007. Worst-case background knowledge for privacy-preserving data publishing. In ICDE.

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

Sweeney, L., 2002. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 10 (5), Pp. 557–570.

Tiancheng, L., Ninghui, L., Jia, Z., and Ian M., 2012. Slicing: A New Approach for Privacy Preserving Data Publishing. Proc. Ieee Transactions On Knowledge And Data Engineering, 24 (3).

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

Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., and Fu, A.W.C., 2006. Utility- Based Anonymization Using Local Recoding. Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), Pp. 785-790.

Xu, Y., Wang, K., Fu, A.W.C., and Yu, P.S., 2008. Anonymizing Transaction Databases for Publication. Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), Pp. 767-775.