Expert Systems with Applications 36 (2009) 9784–9792

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Genetic algorithm-based clustering approach for *k*-anonymization

Jun-Lin Lin *, Meng-Cheng Wei

Department of Information Management, Yuan Ze University, 135 Yuan-Tung Road, Chung-Li 320, Taiwan

ARTICLE INFO

Keywords:
k-Anonymity
Clustering
Genetic algorithm

ABSTRACT

k-Anonymity has been widely adopted as a model for protecting public released microdata from individual identification. This model requires that each record must be identical to at least $k-1$ other records in the anonymized dataset with respect to a set of privacy-related attributes. Although anonymizing the original dataset to satisfy the requirement of *k*-anonymity is easy, the anonymized dataset must preserve as much information as possible of the original dataset. Clustering techniques have recently been successfully adapted for *k*-anonymization. This work proposes a novel genetic algorithm-based clustering approach for *k*-anonymization. The proposed approach adopts various heuristics to select genes for crossover operations. Experimental results show that this approach can further reduce the information loss caused by traditional clustering-based *k*-anonymization techniques.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

Privacy protection is an important societal concern. Protection of public released microdata from individual identification becomes increasingly important as the public becomes increasingly concerned with privacy. Most privacy protection techniques work by randomizing (Agrawal & Srikant, 2000; Lin & Cheng, 2009) or generalizing Samarati, 2001 the original data, but can also degrade the quality of the data. Therefore, a dilemma exists between data quality and data privacy.

As a privacy-preserving approach, has been extensively studied in recent years, the *k*-anonymity model (Samarati, 2001; Sweeney, 2002) works by ensuring that each record of a table is identical to at least $k-1$ other records with respect to a set of privacy-related attributes, called *quasi-identifiers*, that could be used to identify individuals by linking these attributes to external datasets. For instance, consider the hospital data in Table 1, where the attributes *ZipCode*, *Gender* and *Age* are considered as quasi-identifiers. Table 2 shows a 3-anonymization version of Table 1, where anonymization is achieved via generalization at the *attribute* level (Ciriani, di Vimercati, Foresti, & Samarati, 2007), i.e., if two records contain the same value at a quasi-identifier before anonymization, then they are generalized to the same value at the quasi-identifier by the anonymization process. Table 3 shows another 3-anonymization version of Table 1, where anonymization is achieved via generalization at the *cell*-level (Ciriani et al., 2007), i.e., two cells with same value could be generalized to different values (e.g., value "75275" in the *ZipCode* column and value "*Male*" in the *Gender* column). Because anonymization via generalization at the cell-level generates data containing different generalization levels within a column (e.g., values "7527*" and "75275" in the *ZipCode* column and values "*Male*" and "*Person*" in the *Gender* column of Table 3), utilizing such data becomes more complex than utilizing the data generated via generalization at the attribute level. However, in terms of data quality, generalization at the cell-level causes less information loss than generalization at the attribute level, and therefore is adopted in this study.

Anonymization via generalization at the cell-level can proceed in two steps, partitioning and anonymizing. In the partitioning step, the dataset to be anonymized is partitioned into several groups such that each group contains at least *k* records. Then, in the anonymizing step, records in the same group are generalized such that their values at each quasi-identifier are identical. Minimizing the information loss incurred by the anonymizing step requires that the partitioning step places similar records (with respect to the quasi-identifiers) in the same group. In data mining, clustering is an effective means of partitioning records into clusters such that records within a cluster resemble each other, while records in different clusters are clearly distinct from each other. Hence, clustering techniques have been successfully adapted for *k*-anonymization (Byun, Kamra, Bertino, & Li, 2007; Chiu & Tsai, 2007; Lin & Wei, 2008; Lin, Wei, Li, & Hsieh, 2008; Loukides & Shao, 2007).

Genetic algorithms (GA) are well known for their global search capabilities. The constraint of *k*-anonymity means that traditional GA-based clustering techniques cannot be easily adapted for the *k*-anonymization problem without causing much overhead. Section 2.3 provides details. This work proposes a GA-based clustering approach for *k*-anonymization. This approach starts with a dataset that has been partitioned using a traditional clustering-based

* Corresponding author. Tel.: +886 3 4638800x2611.
E-mail addresses: jun@saturn.yzu.edu.tw (J.-L. Lin), mongcheng@gmail.com (M.-C. Wei).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.02.009

Table 1
Patient records of a hospital.

ZipCode	Gender	Age	Disease	Expense
75275	Male	22	Flu	100
75277	Male	23	Cancer	3000
75278	Male	24	HIV+	5000
75275	Male	33	Diabetes	2500
75275	Female	38	Diabetes	2800
75275	Female	36	Diabetes	2600

Table 2
Anonymization at attribute level.

ZipCode	Gender	Age	Disease	Expense
7527*	Person	[21–30]	Flu	100
7527*	Person	[21–30]	Cancer	3000
7527*	Person	[21–30]	HIV+	5000
7527*	Person	[31–40]	Diabetes	2500
7527*	Person	[31–40]	Diabetes	2800
7527*	Person	[31–40]	Diabetes	2600

Table 3
Anonymization at cell-level.

ZipCode	Gender	Age	Disease	Expense
7527*	Male	[21–25]	Flu	100
7527*	Male	[21–25]	Cancer	3000
7527*	Male	[21–25]	HIV+	5000
75275	Person	[31–40]	Diabetes	2500
75275	Person	[31–40]	Diabetes	2800
75275	Person	[31–40]	Diabetes	2600

k-anonymization technique, called the Hybrid method (Lin et al., 2008). The output of the Hybrid method is encoded as a population of chromosomes. Various heuristics are then adopted to select genes to undergo the crossover operations to reduce the information loss of the dataset. This approach is, to our knowledge, the first GA-based clustering method proposed for k-anonymization at the cell-level. Experimental results demonstrate that the proposed approach further reduces the information loss caused by the Hybrid method.

The rest of this paper is organized as follows. Section 2 reviews basic concepts on k-anonymization with a focus on clustering-based methods for k-anonymization and GA-based clustering techniques. Section 3 describes the proposed GA-based approach for k-anonymization. Section 4 presents a performance analysis of the proposed approach. Conclusions are finally drawn in Section 5, along with recommendations for further research.

2. Basic concepts

The k-anonymity model has attracted considerable attention in recent years. Many approaches have been proposed for k-anonymization and its variations. Please refer to Ciriani et al. (2007) for a survey of various k-anonymization approaches. This section first describes the concept of information loss, which measures the effectiveness of various k-anonymization approaches. Several recently proposed clustering methods for k-anonymization are then reviewed. Finally, GA-based clustering methods are briefly reviewed, and their possible adaption to the k-anonymization problem is discussed.

2.1. Information loss

Information loss refers to the amount of information lost due to k-anonymization. This work adopts the definition of information loss in Byun et al. (2007). Some notations are defined first to facilitate the discussion, and are used throughout this paper. Let $\mathscr{T}$ denote the dataset to be anonymized, which is described by m numeric quasi-identifiers $N_1,\ldots,N_m$ and q categorical quasi-identifiers $C_1,\ldots,C_q$. Let $M=\{1,2,\ldots,m\}$ and $Q=\{1,2,\ldots,q\}$ denote two index sets. Each categorical attribute $C_{i\in Q}$ is associated with a taxonomy tree T_{C_i}, which is used to generalize the values of this attribute.

Consider a set of records $\mathscr{P}\subseteq\mathscr{T}$. Let $\widehat{N}_i(\mathscr{P})$, $\check{N}_i(\mathscr{P})$ and $\overline{N}_i(\mathscr{P})$, respectively, denote the max, min and average values of the records in $\mathscr{P}$ with respect to the numeric attribute $N_{i\in M}$; let $C_i(\mathscr{P})$ denote the set of values of the records in $\mathscr{P}$ with respect to the categorical attribute $C_{i\in Q}$, and let $T_{C_i}(\mathscr{P})$ denote the maximal subtree of T_{C_i} rooted at the lowest common ancestor of the values in $C_i(\mathscr{P})$. Then, the *diversity* of $\mathscr{P}$, denoted by $D(\mathscr{P})$, is defined as:

$$D(\mathscr{P})=\sum_{i\in M}\frac{\widehat{N}_i(\mathscr{P})-\check{N}_i(\mathscr{P})}{\widehat{N}_i(\mathscr{T})-\check{N}_i(\mathscr{T})}+\sum_{i\in Q}\frac{H(T_{C_i}(\mathscr{P}))}{H(T_{C_i})} \tag{1}$$

where $H(T)$ represents the height of a tree T.

The *centroid* $\overline{\mathscr{P}}$ of $\mathscr{P}$ is a record whose value of attribute $N_{i\in M}$ equals $\overline{N}_i(\mathscr{P})$, and the value of attribute $C_{i\in Q}$ equals the root of the tree $T_{C_i}(\mathscr{P})$. Anonymizing the records in $\mathscr{P}$ means generalizing these records to the same values with respect to each quasi-identifier, and can be done in either of two ways. One method is simply replacing the value of each record at quasi-identifiers with the centroid of $\mathscr{P}$. The other method is to replace the value of a numeric attribute $N_{i\in M}$ by an interval $[\check{N}_i(\mathscr{P}),\widehat{N}_i(\mathscr{P})]$, and replace the value of a categorical attribute $C_{i\in Q}$ by the root of $T_{C_i}(\mathscr{P})$. These two methods differ only in terms of whether they generalize a numeric attribute to an interval or a mean. The amount of *information loss* incurred by anonymization on $\mathscr{P}$ is defined as:

$$L(\mathscr{P})=|\mathscr{P}|\times D(\mathscr{P}) \tag{2}$$

where $|\mathscr{P}|$ represents the number of records in $\mathscr{P}$.

Let $\mathbf{P}=\{\mathscr{P}_1,\ldots,\mathscr{P}_{|\mathbf{P}|}\}$ be a partitioning of $\mathscr{T}$, namely, $\bigcup_{i\in P}\mathscr{P}_i=\mathscr{T}, \mathscr{P}_i\neq\emptyset$, and $\mathscr{P}_{\hat{i}}\cap\mathscr{P}_{\check{i}}=\emptyset$ for any $\hat{i}\neq\check{i}$ where $i,\hat{i},\check{i}\in P=\{1,2,\ldots,|\mathbf{P}|\}$. The total information loss of $\mathbf{P}$ is the sum of the information loss of each $\mathscr{P}_{i\in P}$, as defined below:

$$TL(\mathbf{P})=\sum_{i=1}^{|\mathbf{P}|}L(\mathscr{P}_i) \tag{3}$$

To maximize the data quality of $\mathscr{T}$ after anonymization, an anonymization method should construct a partitioning $\mathbf{P}$ that minimized the total information loss of $\mathbf{P}$. Therefore, the k-anonymization problem can be formally defined as a constraint optimization problem as follows.

[**Problem Definition**] Given a dataset $\mathscr{T}$, find a partitioning $\mathbf{P}$ of $\mathscr{T}$ such that the total information loss $TL(\mathbf{P})$ is minimized subject to the constraint that $|\mathscr{P}|\geqslant k$ for any $\mathscr{P}\in\mathbf{P}$.

Solanas, Sebé, and Domingo-Ferrer (2008) proposed another method of calculating the total information loss for the case of $\mathscr{T}$ without any categorical quasi-identifier (i.e., $Q=\emptyset$). In this case, given a partitioning $\mathbf{P}=\{\mathscr{P}_1,\ldots,\mathscr{P}_{|\mathbf{P}|}\}$ of $\mathscr{T}$, the value at each numeric quasi-identifier $N_{i\in M}$ of each record in a partition $\mathscr{P}_j\in\mathbf{P}$ is generalized to the corresponding mean $\overline{N}_i(\mathscr{P}_j)$, and the total information loss of partitioning $\mathbf{P}$ is defined as

$$TL(\mathbf{P})=\frac{\sum_{j=1}^{|\mathbf{P}|}\sum_{x\in\mathscr{P}_j}(x-\overline{x}_j)^2}{\sum_{x\in\mathscr{T}}(x-\overline{x})^2} \tag{4}$$

where $\overline{x}_j$ and $\overline{x}$, respectively, denote the centroids of $\mathscr{P}_j$ and $\mathscr{T}$. The distance between a record x and a centroid ($\overline{x}_j$ or $\overline{x}$) is calculated using the Euclidean distance over all numeric quasi-identifiers. Since this definition of information loss is limited to datasets without any categorical quasi-identifier, it is not adopted in this current work.

Download English Version:

https://daneshyari.com/en/article/388797

Download Persian Version:

https://daneshyari.com/article/388797

Daneshyari.com