Decision Support Systems 48 (2009) 133–140

Contents lists available at ScienceDirect

Decision Support Systems

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

Identity disclosure protection: A data reconstruction approach for privacy-preserving data mining ☆

Dan Zhu [a,*], Xiao-Bai Li [b], Shuning Wu [c]

[a] Department of Logistics, Operations and MIS, Iowa State University, Ames, IA 50011, USA
[b] College of Management, University of Massachusetts Lowell, Lowell, MA 01854, USA
[c] ISO Innovative Analytics, 388 Market Street #750, San Francisco, CA 94111, USA

ARTICLE INFO

Article history:
Received 1 July 2008
Received in revised form 15 March 2009
Accepted 7 July 2009
Available online 15 July 2009

Keywords:
Privacy
Identity disclosure
k-Anonymity
Data mining
Genetic algorithm

ABSTRACT

Identity disclosure is one of the most serious privacy concerns in today's information age. A well-known method for protecting identity disclosure is *k*-anonymity. A dataset provides *k*-anonymity protection if the information for each individual in the dataset cannot be distinguished from at least $k-1$ individuals whose information also appears in the dataset. There is a flaw in *k*-anonymity that would still allow an intruder to discern the confidential information of individuals in the anonymized data. To overcome this problem, we propose a data reconstruction approach to achieve *k*-anonymity protection in predictive data mining. In this approach, the potentially identifying attributes are first masked using aggregation (for numeric data) and swapping (for nominal data). A genetic algorithm technique is then applied to the masked data to find a good subset of it. This subset is then replicated to form the released dataset that satisfies the *k*-anonymity constraint.

1. Introduction

Data-mining technologies have enabled organizations to extract useful knowledge from the data in order to better understand and serve their customers, and to gain competitive advantages [6,21,26]. While successful business applications of data mining are encouraging, there are increasing concerns about invasions to the privacy of personal information. A survey by Time/CNN [16] revealed that 93% of respondents believed companies selling personal data should be required to gain permission from the individuals whose information is being shared. In another study [9], more than 70% of participants responded negatively to questions related to the secondary use of private information. Concern about privacy threats has caused data quality and integrity to deteriorate. According to [34], 82% of online users have refused to give personal information and 34% have lied when asked about their personal habits and preferences.

This study deals with the conflict between privacy and data mining in organizational decision support. Organizations that use their customers' records in data-mining activities are obligated to take actions to protect the identities of the individuals involved. It has been demonstrated that personal identities cannot be adequately protected by simply removing identity attributes from released data. There has been extensive research in the area of statistical databases (SDBs) on how to protect individuals' sensitive data when providing summary statistical information. The privacy issue arises in SDBs when summary statistics are derived on very few individuals' data. In this case, releasing the summary statistics may result in disclosing confidential data. The methods for preventing such disclosure can be broadly classified into two categories: (i) query restriction, which prohibits queries that would reveal confidential data, and (ii) data perturbation, which alters individual data in a way such that the summary statistics remain approximately the same. In general, both methods have been extensively investigated and employed [1]. Problems in data mining are somewhat different from those in SDBs. A data-mining task, such as classification or numeric prediction, requires working on individual records contained in a dataset. As a result, query restriction is no longer applicable and data perturbation or anonymization becomes the primary approach for privacy protection in data mining. Further, predictive data mining essentially relies on discovering relationships between data attributes. Preserving such relationships may not be consistent with preserving summary statistics. Researchers in the data-mining community have proposed various methods to resolve the conflict between data mining and privacy protection [4,7,14,22,23]. For example, a method for building a decision tree classifier from perturbed data is proposed in [3]. A framework for mining association rules from transaction data that have been randomized is presented in [11]. A set of algorithms for hiding sensitive rules is proposed in [36]. Techniques for preserving privacy in distributed data mining are discussed in [8].

☆ A preliminary version of this work was presented at the 28th International Conference on Information Systems (ICIS), Montreal, Canada, December 2007.

* Corresponding author.
E-mail addresses: dzhu@iastate.edu (D. Zhu), xiaobai_li@uml.edu (X.-B. Li), swu@iso.com (S. Wu).

0167-9236/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2009.07.003

A well-known method for privacy protection, called *k*-anonymity, was proposed in [31,33]. The basic idea is to anonymize the data such that each individual cannot be distinguished from a group of other individuals in the data. The method has gained increasing popularity in privacy-preserving data mining. However, the *k*-anonymity approach would, in some circumstances, still allow a data intruder to disclose the individual confidential information in the *k*-anonymized data. To overcome this problem, we propose a data reconstruction approach to achieve *k*-anonymity protection in predictive data mining. In this approach, the potentially identifying attributes are first masked using aggregation (for numeric data) and swapping (for nominal data), without considering the *k*-anonymity constraint. A genetic algorithm technique is then applied to the masked data to find a good subset of it. This subset is then replicated to form the released dataset that satisfies the *k*-anonymity constraint. An experimental study is conducted to show the effectiveness of the proposed method.

2. Identity and confidentiality disclosure problem

A common practice for protecting identity disclosure is to remove identity related attributes from released data. Sweeney [33] demonstrated that this is not adequate in protecting personal identities. In fact, the author showed that 87% of the population in the United States can be uniquely identified using three demographic attributes: gender, date of birth, and 5-digit zip code. These attributes are normally not considered identity attributes. However, since they can potentially be used to uniquely identify a record, they are collectively called *quasi-identifier* (QI). The *k*-anonymity technique was proposed to address related identity disclosure problems [31,33]. A dataset provides *k-anonymity* protection if the values of the QI attributes for any individual matches those of at least $k-1$ other individuals in the same dataset. The anonymity is achieved by generalization and suppression of the QI values. With *k*-anonymity, individual identities are better protected. However, as indicated in [20], it is still likely for an intruder to disclose the confidential information of individuals in the *k*-anonymized data. The following example demonstrates the problem.

Table 1
An illustrative example.

ID	Age	Marital status	Blood pressure	Blood type	Test result
(a) Original patient data					
1	23	Never married	75/120	O	Negative
2	25	Never married	66/113	O	Positive
3	27	Never married	74/115	A	Negative
4	28	Never married	77/128	AB	Negative
5	32	Married	72/125	B	Negative
6	33	Married	93/147	O	Negative
7	35	Married	75/124	AB	Positive
8	37	Divorced	95/142	O	Negative
9	40	Widow(er)	88/146	A	Positive
10	43	Divorced	110/155	O	Positive
11	45	Married	90/140	O	Positive
12	45	Divorced	104/145	B	Positive
(b) k-anonymized patient data (k = 4)					
1	20–29	Never married	75/120	O	Negative
2	20–29	Never married	66/113	O	Positive
3	20–29	Never married	74/115	A	Negative
4	20–29	Never married	77/128	AB	Negative
5	30–39	Married	72/125	B	Negative
6	30–39	Married	93/147	O	Negative
7	30–39	Married	75/124	AB	Positive
8	30–39	Married	95/142	O	Negative
9	40–49	Married	88/146	A	Positive
10	40–49	Married	110/155	O	Positive
11	40–49	Married	90/140	O	Positive
12	40–49	Married	104/145	B	Positive

Table 1(a) shows a complete list of 12 patients administered at a hospital in a year for a sensitive disease. The test result is confidential. To protect privacy, the identity related attributes, such as name and address, were removed from the dataset. Knowing they were protected in this way, the patients authorized the hospital to share the data with related professionals and organizations for medical research purposes. However, it can be seen that the value combination of the Age and Marital Status attributes for each record is unique in the dataset. Therefore, it would not be difficult for an intruder to find the test results of a patient if he had some knowledge about the patient's age and martial status (quasi-identifier). Assume, for example, a medical school student, Allen, acquired this dataset from the hospital. If he knew that a 35-year-old, married classmate took this test at the hospital during the year, he can effectively identify his classmate as patient #7, who had a positive test result. Suppose Allen knew that one of his friends, aged 45 and divorced, was also in the list. Then he could also easily find that his friend was patient #12, who also had a positive test result.

The *k*-anonymity technique can help protect against such identity disclosure problems. Table 1(b) shows the anonymized dataset released by the hospital. The generalization method was applied to the original data where the Age values were grouped into three intervals and Marital Status values were combined into two groups (with Married representing three original categories: Married, Divorced and Widow). From this dataset, Allen can no longer identify his classmate (#7) or the classmate's test result. As far as his other friend (#12) is concerned, however, Allen is still able to access confidential information. Although he cannot identify which record is his friend's, he still knows that his friend has a positive test result, since all of the four records in the group containing his friend's record have the same test result.

The example above demonstrates that it is still quite possible for a data intruder to disclose the confidential information of an individual in the *k*-anonymized data. *k*-anonymity protects identity disclosure by generalizing different but similar QI attribute values into the same value. The new values produced by the generalization operation are still correct with respect to the generalized categories. Since confidential values (e.g., test result) remain unchanged in *k*-anonymity, individuals in a group are subject to high disclosure risk if their confidential values in the group are the same. To overcome this problem, Machanavajjhala et al. [20] proposed a new privacy principal, called *l*-diversity, which requires, in addition to *k*-anonymity, that the confidential attribute should include at least *l* "well-represented" values in the anonymized data. This additional constraint can sometimes be hard to satisfy and usually causes much larger group sizes.

Another drawback with the *k*-anonymity approach, in terms of data utility, is that it significantly changes the univariate statistical properties of the QI attributes, which are very important in statistical and data warehousing applications. This problem is due to the use of generalization and suppression methods: generalization creates new nominal values instead of keeping the original nominal values in the data, while suppression results in skewed distributions (partial suppression) or no univariate information at all (full suppression) for the QI attributes. This loss of univariate information also exists in other *k*-anonymity based techniques such as *l*-diversity. The problem becomes worse when the technique is geared towards specific data-mining algorithms, such as that proposed in [13].

The data reconstruction approach we propose addresses both the privacy protection and information loss problems mentioned above. Our proposed method masks the QI attributes by aggregating numeric values and swapping nominal values. Aggregation and swapping operations differ from generalization in that the aggregated and swapped values are "faked" values (as opposed to "correct" values produced by generalization). As a result, the proposed approach provides a better protection against identity disclosure. In addition, aggregation preserves approximately some important numeric univariate statistics (e.g., mean), while swapping preserves the univariate frequency distributions of nominal attributes.