Information Sciences 256 (2014) 138–161

Contents lists available at ScienceDirect

Information Sciences

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

Improving accuracy of classification models induced from anonymized datasets

Mark Last [a,*], Tamir Tassa [b], Alexandra Zhmudyak [a], Erez Shmueli [a]

[a] Department of Information Systems Engineering, Ben-Gurion University of the Negev, Israel
[b] Department of Mathematics and Computer Science, The Open University, Israel

ARTICLE INFO

Article history:
Received 14 February 2012
Received in revised form 6 May 2013
Accepted 28 July 2013
Available online 12 August 2013

Keywords:
Privacy preserving data publishing
Privacy preserving data mining
k-Anonymity
ℓ-Diversity
Non-homogeneous anonymization
Classification

ABSTRACT

The performance of classifiers and other data mining models can be significantly enhanced using the large repositories of digital data collected nowadays by public and private organizations. However, the original records stored in those repositories cannot be released to the data miners as they frequently contain sensitive information. The emerging field of Privacy Preserving Data Publishing (PPDP) deals with this important challenge. In this paper, we present NSVDist (Non-homogeneous generalization with Sensitive Value Distributions)—a new anonymization algorithm that, given minimal anonymity and diversity parameters along with an information loss measure, issues corresponding non-homogeneous anonymizations where the sensitive attribute is published as frequency distributions over the sensitive domain rather than in the usual form of exact sensitive values. In our experiments with eight datasets and four different classification algorithms, we show that classifiers induced from data generalized by NSVDist tend to be more accurate than classifiers induced using state-of-the-art anonymization algorithms.

1. Introduction

A vast amount of information of all types is collected daily about people by governments, corporations and individuals. As a result, there is an enormous quantity of privately-owned records that describe individuals' finances, interests, activities, and demographics. These records often include sensitive data and may violate the privacy of the users if published. This information is becoming a very important resource for many systems and corporations that may enhance and improve their services and performance by inducing novel and potentially useful data mining models. One common practice for releasing such confidential data without violating privacy is applying regulations, policies and guiding principles for the use of the data. Such regulations usually entail data distortion operations such as generalization or random perturbations. The challenge with this approach is that, on one hand, data leakage can still occur, and, on the other hand, the data and the resulting data mining models may become nearly useless after excessive distortion [7].

The emerging research field of Privacy Preserving Data Publishing (PPDP) is targeting this challenge [7]. It aims at developing techniques that enable publishing data while minimizing distortion for maintaining utility on one hand, and ensuring that privacy is preserved on the other hand. In this paper we present a new privacy-preserving data publishing method, which is shown to maintain the predictive utility of supervised classification algorithms that are trained on the published data. The predictive utility is measured by the classification accuracy of the induced classification models, when applied

* Corresponding author. Tel.: +972 8 6461397.
E-mail address: mlast@bgu.ac.il (M. Last).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.07.034

to new, previously unseen data. As we explain in the related work section (Section 2), we assume that the validation data can be kept in its original non-distorted form.

A closely related research area is Privacy Preserving Data Mining (PPDM) that was initiated in 2000 by [1]. PPDM algorithms aim at anonymizing data towards its release for specific data mining goals, so that the data utility is maximized, on one hand, and its privacy is preserved on the other hand. The developed PPDM algorithms are tailored to specific data mining tasks and algorithms. For example, if the data needs to be used for inducing a decision-tree classifier, the corresponding PPDM algorithm will aim at achieving anonymization while incurring a minimal loss of accuracy in the resulting classifier. In PPDP, on the other hand, the exact purposes of the data release are unknown and it is needed to anonymize the data using utility measures that are not targeted to a specific data mining algorithm.

It is customary to distinguish between four types of attributes in the database table that needs to be published (see [3]):

- Identifiers—attributes that uniquely identify an individual (e.g. `name`);
- Quasi-identifiers—publicly-accessible attributes that do not identify a person, but some combinations of their values might yield unique identification (e.g., `gender`, `age`, and `zipcode`);
- Sensitive information—attributes of private nature, such as medical or financial data (in this paper, we follow the common assumption of a single sensitive attribute, which is identical to the class attribute); and
- Other non-sensitive attributes that, on one hand, cannot be used for identification since they are unlikely to be accessible to the adversary, and, on the other hand, do not represent information of sensitive nature. (Those attributes can be ignored in our discussion.)

A common practice in PPDP and PPDM is to remove the identifiers and to generalize or suppress the quasi-identifiers in order to protect the sensitive data of individuals from being revealed. Generalization means that the original values of quasi-identifiers are replaced with less specific values, whereas in case of suppression no values are released at all. The sensitive data is usually retained unchanged.

In the past years, several models were suggested for maintaining privacy when disseminating data. Most approaches evolved from the basic model of k-anonymity [38]. In that model, the practice is to remove the identifiers and generalize the quasi-identifiers as described above, until each generalized record is indistinguishable from at least $k-1$ other generalized records, when projected on the quasi-identifiers. Consequently, an adversary who wishes to trace a record of a specific person in the anonymized table, will not be able to trace that person's record to subsets of less than k anonymized records.

As an example, consider the basic table in Table 1(a), having the quasi-identifiers `Age` and `Zipcode` and the sensitive attribute `Disease`. Table 1(b) is a corresponding 2-anonymization. An adversary who wishes to trace Eve's record in it may infer that it is one of the last two records, but they are equally likely, whence the probability of correct identification is $1/2$. Many algorithms were suggested in the literature for k-anonymization, e.g. [2,10,11,13,19,24,25,35,36,39].

The k-anonymity model on its own does not provide a sufficient level of privacy. Its main weakness is that it does not guarantee sufficient diversity in the sensitive attribute within each equivalence class (or block) of records that are indistinguishable by their generalized quasi-identifiers. Namely, even though it guarantees that every record in the anonymized table is indistinguishable from at least $k-1$ others, it is possible that the distribution of the sensitive values in those records discloses "too much" information. To mitigate this problem, Machanavajjhala et al. [28] proposed the security measure of ℓ-diversity. That measure requires that each block of indistinguishable records will have at least ℓ "well represented"

Table 1
A table and corresponding anonymizations.

Name	Age	Zipcode	Disease
(a) The original table			
Alice	30	10,055	Measles
Bob	21	10,055	Flu
Carol	21	10,023	Angina
David	55	10,165	Flu
Eve	47	10,224	Diabetes
(b) Homogeneous anonymization			
	21–30	100**	Measles
	21–30	100**	Flu
	21–30	100**	Angina
	47–55	10***	Flu
	47–55	10***	Diabetes
(c) Non-homogeneous anonymization			
	21–30	10,055	Measles
	21	100**	Flu
	21–30	100**	Angina
	47–55	10***	Flu
	47–55	10***	Diabetes