Data & Knowledge Engineering 70 (2011) 1030–1045

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Information based data anonymization for classification utility

Jiuyong Li [a,*], Jixue Liu [a], Muzammil Baig [a], Raymond Chi-Wing Wong [b]

[a] School of Computer & Information Science, University of South Australia, Australia
[b] Department of Computer Science & Engineering, Hong Kong University of Science and Technology, Hong Kong

ARTICLE INFO

Article history:
Received 27 September 2010
Received in revised form 10 April 2011
Accepted 5 July 2011
Available online 22 July 2011

Keywords:
Privacy
Anonymization
k-anonymity
Classification
Mutual information
Kullback–Leibler divergence

ABSTRACT

Anonymization is a practical approach to protect privacy in data. The major objective of privacy preserving data publishing is to protect private information in data whereas data is still useful for some intended applications, such as building classification models. In this paper, we argue that data generalization in anonymization should be determined by the classification capability of data rather than the privacy requirement. We make use of mutual information for measuring classification capability for generalization, and propose two *k*-anonymity algorithms to produce anonymized tables for building accurate classification models. The algorithms generalize attributes to maximize the classification capability, and then suppress values by a privacy requirement *k* (IACk) or distributional constraints (IACc). Experimental results show that algorithm IACk supports more accurate classification models and is faster than a benchmark utility-aware data anonymization algorithm.

1. Introduction

1.1. Data privacy and anonymization

Privacy preservation has become a major issue in many data mining applications. Various organizations, such as hospitals, medical administrations and insurance companies, have collected a large amount of data over years. However, these organizations are reluctant to publish the data because of privacy concern. It is necessary to ensure privacy protection when data is published.

Anonymization is a major technique for protecting privacy in data publishing. For example, *k*-anonymity [19] protects data privacy by ensuring that the probability for identifying an individual in a published data set is at most $1/k$. A common way to achieve *k*-anonymity is to generalize values within the person identifiable attributes in a table, called *quasi-identifier*. For example, if the following information, "gender = male, age = 45, postcode = 5011", is too specific in a data set, e.g. fewer than *k* men of age 45 live in the suburb of postcode 5011. These people are potentially identifiable. If the record is generalized as "gender = male, age = 45–55, postcode = 5009–5012", more than *k* people will have the same person identifiable information in the data and hence their privacy is better preserved. The higher the privacy protection requirement is, the more the generalization will be. The most generalized form of a record is the "*, *, *". The replacement of values with "*"s is called *suppression*. This is equivalent to that nothing is published for identity-related attributes.

Much research work has been conducted to enhance the protection level of the *k*-anonymity model, such as, *l*-diversity [15], (α,k)-anonymity [23], and *t*-closeness [13]. These models impose further protective requirements on the published data. They block attribute inference channels from identity-related attributes to sensitive values in the data. These models provide strong privacy protection, but are not good for some data mining tasks, such as classification, because associations between some attributes and classes have been purposely hidden. Models preventing attribute inference, such as, *l*-diversity and *t*-closeness, impose an upper bound for classification accuracies [14].

* Corresponding author at: Mawson Lakes, SA 5095, Australia. Tel.: +61 8 83023898; fax: +61 8 83023381.
E-mail address: jiuyong.li@unisa.edu.au (J. Li).

0169-023X/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2011.07.001

The goal for publishing a data set is to make it useful rather than to lock it in owner's safe case. Building classification models is a major utility. For example, the hospital data is released to public for modeling causes of diseases. Normally, it is not an obligation for a data owner to build models but it is an obligation for a data owner to keep data privacy when the data is released. In many circumstances, *k* anonymity provides sufficient protection. For example, every released medical record has been authorized by a patient, and there is no privacy concern in the data itself. However *k*-anonymization is necessary for preventing the medical data set from being linked to other patient sensitive information such as DNA sequences.

Most research on data utility has focused on value precision. The purpose is to minimize value generalizations in an anonymized table. Criteria such as distortion [12], uncertainty [24], query accuracy [11], information loss [20], and information utility [25] capture this information directly. The smaller modifications are made to a data set, the better the anonymization is. To increase value precision of anonymized tables, many anonymization techniques, such as, multidimensional [10] and local recoding [12,24] methods, have been proposed. These methods reduce uncertainty or distortions of the anonymized data.

1.2. Related work and motivations

When the anonymized data is for building classification models, the utility requirement is quite different. Generalization is not a problem for building many interpretable classification models, such as decision trees and naive Bayes models. However, domain consistency is important and many precision based anonymization methods are not applicable [12,24]. For example, values generalized to overlapping intervals, such as (10–14), (11–12), and (13–17), are not good for classification model building. Generalized values mixed with different levels of an attribute taxonomy hierarchy, such as, 11th, 12th and senior secondary (which is a generalization of 11th and 12th) of Education attribute, are not good for building classification models either. Other methods are required for data anonymization for classification utility.

Many previous studies aiming at classification utility have been done. Iyengar [6] has firstly proposed an optimization approach to minimize class impurity in data generalization. The optimization has been shown impractical for medium and large data sets. Wang et al. have proposed a bottom up anonymization method for classification utility [21], which only handles categorical values. An improved method, called TDS (Top–Down Specialization method), from the same research group has been proposed [4]. TDS makes use of the single dimensional generalization approach. It is efficient and keeps good classification capability in the anonymized data. A further improvement of TDS is called TDR [5] (Top–Down Refinement). TDR improves functionalities of TDS greatly. It handles both categorical and numerical values with and without generalization taxonomy trees. It also handles data with multiple quasi-identifiers. Recently, Kisilevich et al. [7] have proposed a multi-dimensional suppression approach, called kACTUS, for classification-aware anonymization. kACTUS makes use of a decision tree, i.e. C4.5 [17], as a base for deciding multi-dimensional regions to be suppressed. The pioneering work in multi-dimensional generalization has been proposed by Lefevre et al. [10], called Mondrian. Mondrian has then been extended to InfoGain Mondrian for various utilities including classification [11]. InfoGain Mondrian has been shown to achieve better classification accuracy than the TDS, and is a benchmark algorithm for classification based anonymization. Other privacy classification work has been done on the publishing classification models without violating the *k*-anonymity constraint. Friedman et al. have proposed a method for building *k*-anonymous decision trees [3]. Sharkey et al. [18] have also proposed a method for publishing decision trees along with a pseudo data set generated by the tree model. The release of a model lacks great flexibilities to users in comparison to the release of data. Firstly, there are many different types of classification models. A data owner won't know which model that a user is interested in. Secondly, for the same type of models, many adjustable parameters will lead to different models. For example, some users are interested in the specificity and some are interested in the precision. Their required models are quite different. Therefore, in this paper, we consider data publishing instead of model publishing.

Let us look at three most recent and closely related methods in data publishing for classification utility: InfoGain Mondrian [11], TDR [5], and kACTUS [7]. Interestingly, they produce the same 10-anonymous table for Table 1(a) following very different paths. InfoGain Mondrian is a multi-dimensional generalization method, and it partitions the data space into a number of disjoined (hyper) rectangular regions by attributes in the quasi-identifier. The smallest partitioned regions (not optimized because the optimal solution is intractable), each of which contains at least *k* data points, are used for attribute value generalization. In this example, attribute Gender is partitioned along male and female. Any partition in attribute Age will lead to a region which has data points fewer than 10. Therefore, the Age attribute is kept at the top level "*".. TDR starts with a table with all values suppressed. TDR then tests attributes Gender and Age to find out which will lead to better tradeoff between information gain and anonymity loss. Attribute Gender wins and attribute Gender is refined, and as a result values of males and females are shown. Note that classes (problems) have been well separated by attribute Gender. Values in Age are kept suppressed because the release of values in Age does not improve classification performance. kACTUS firstly builds a decision tree on Gender and Age attributes. The decision tree contains one node with the test that "whether gender = male or not". Both outcome branches contain 11 data points each and hence they comply with 10-anonymity. Attribute Age has not been referenced in the decision tree and hence all values are suppressed.

A disadvantage of the anonymized table in Table 1(b) is that it suppresses too many values when the data set has satisfied the privacy requirement. It is true that no other anonymization method is able to improve the classification accuracy in this 10-anonymous table. However, when other attributes are taken into account, it will be useful to have relationships between attributes Age and Blood Pressure. In other large data sets, such relationships potentially help classification. Normally the quasi-identifier is only a part of all the attributes, and we should not assume that a classification model is built on the quasi-identifier only. The