Knowledge-Based Systems 43 (2013) 82–94

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

Anonymizing classification data using rough set theory

Mingquan Ye [a,c,*], Xindong Wu [a,b], Xuegang Hu [a], Donghui Hu [a]

[a] Department of Computer Science, Hefei University of Technology, Hefei 230009, PR China
[b] Department of Computer Science, University of Vermont, Burlington, VT 05405, USA
[c] Department of Computer Science, Wannan Medical College, Wuhu 241002, PR China

ARTICLE INFO

Article history:
Received 16 April 2012
Received in revised form 23 December 2012
Accepted 8 January 2013
Available online 26 January 2013

Keywords:
k-Anonymity
Rough sets
Multi-level granulation
Attribute value taxonomy
Privacy preserving data mining

ABSTRACT

Identity disclosure is one of the most serious privacy concerns in many data mining applications. A well-known privacy model for protecting identity disclosure is *k*-anonymity. The main goal of anonymizing classification data is to protect individual privacy while maintaining the utility of the data in building classification models. In this paper, we present an approach based on rough sets for measuring the data quality and guiding the process of anonymization operations. First, we make use of the attribute reduction theory of rough sets and introduce the conditional entropy to measure the classification data quality of anonymized datasets. Then, we extend conditional entropy under single-level granulation to hierarchical conditional entropy under multi-level granulation, and study its properties by dynamically coarsening and refining attribute values. Guided by these properties, we develop an efficient search metric and present a novel algorithm for achieving *k*-anonymity, Hierarchical Conditional Entropy-based Top-Down Refinement (HCE-TDR), which combines rough set theory and attribute value taxonomies. Theoretical analysis and experiments on real world datasets show that our algorithm is efficient and improves data utility.

1. Introduction

Identity disclosure is one of the most serious privacy concerns in many data mining applications. Some organizations, such as hospitals and insurance companies, have collected a large amount of *microdata*, which refers to data published in its raw, non-aggregated form. The microdata can provide tremendous opportunities for knowledge-based decision making. However, these organizations are reluctant to publish the data because of privacy threats. One important type of privacy attack is the re-identification of individuals by joining data from multiple public tables; such an attack is called a *linking attack*. For example, according to [27], more than 85% of the population of the United States can be uniquely identified using their gender, zipcode, and date of birth. The minimal set of attributes in a table is called the *quasi-identifier* (QI), which can be joined with external information to re-identify individual records.

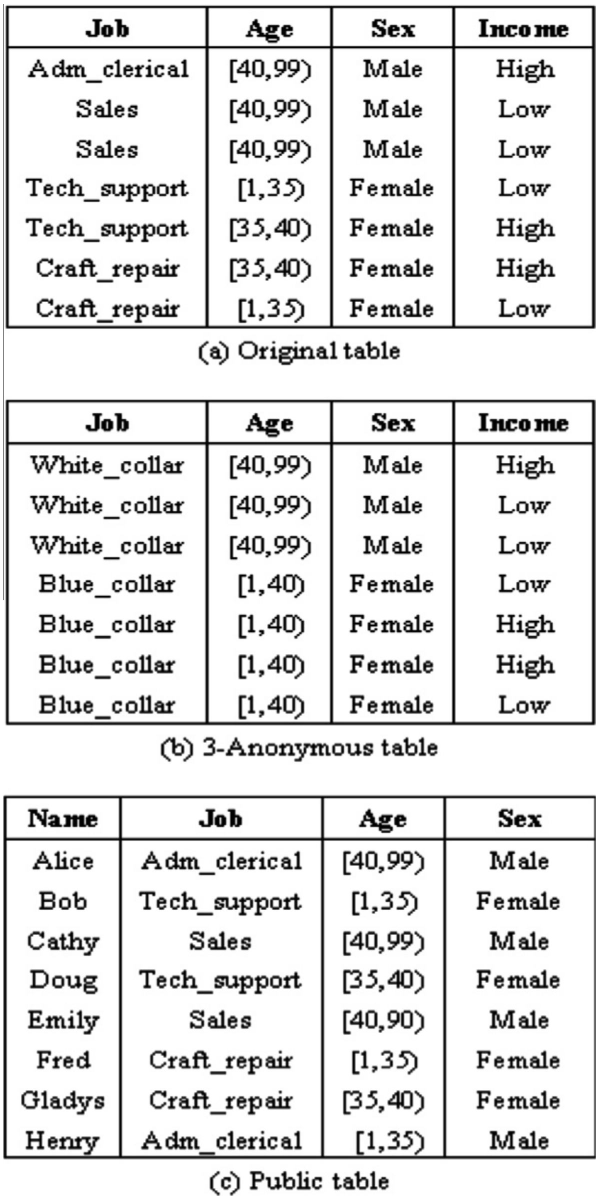
To prevent linking attacks through QI, *k-anonymity* was proposed [27]. A table satisfies k-anonymity if each record in the table is indistinguishable from at least $(k-1)$ other records with respect to certain QI attributes and such a table is called a *k-anonymous* table. Consequently, the probability of identifying an individual from a specific record through QI is at most $1/k$. This ensures that individuals cannot be uniquely identified by linking attacks. For example, Fig. 1 illustrates how k-anonymization hinders linking attacks. The joining of the original table in Fig. 1a with the public data in Fig. 1c would reveal that Alice's income is high and Bob's is low. Fig. 1b shows a 3-anonymous table that generalizes QI = {Job, Age, Sex} from the original table using the attribute value taxonomies in Fig. 2. The 3-anonymous table has two distinct groups on QI, "White_collar, [40,99), Male" and "Blue_collar, [1,40), Female". Because each group contains at least 3 records, the table is 3-anonymous. If we link the records in Fig. 1b to the records in Fig. 1c through the QI, each record is linked to either no record or at least 3 records in Fig. 1c. Therefore, the outcome of joining the 3-anonymous table with the public data is ambiguous.

Data in their original form often contain sensitive information about individuals. However, the data typically do not satisfy the k-anonymity requirement, and publishing such data would violate individual privacy. A task of the utmost importance is to modify the data so that the modified data remain practically useful for data mining while individual privacy is preserved. For example, drug companies and researchers may be interested in patient records for drug development. Data mining harnesses a large amount of patient data available for extracting knowledge crucial to the progress of drug research. Such additional uses of data are

* Corresponding author at: Department of Computer Science, Hefei University of Technology, Hefei 230009, PR China.

E-mail addresses: ymq@wnmc.edu.cn (M. Ye), xwu@cs.uvm.edu (X. Wu), jsjxhuxg@hfut.edu.cn (X. Hu), hudh@hfut.edu.cn (D. Hu).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.01.007



Job	Age	Sex	Income
Adm_clerical	[40,99)	Male	High
Sales	[40,99)	Male	Low
Sales	[40,99)	Male	Low
Tech_support	[1,35)	Female	Low
Tech_support	[35,40)	Female	High
Craft_repair	[35,40)	Female	High
Craft_repair	[1,35)	Female	Low

(a) Original table

Job	Age	Sex	Income
White_collar	[40,99)	Male	High
White_collar	[40,99)	Male	Low
White_collar	[40,99)	Male	Low
Blue_collar	[1,40)	Female	Low
Blue_collar	[1,40)	Female	High
Blue_collar	[1,40)	Female	High
Blue_collar	[1,40)	Female	Low

(b) 3-Anonymous table

Name	Job	Age	Sex
Alice	Adm_clerical	[40,99)	Male
Bob	Tech_support	[1,35)	Female
Cathy	Sales	[40,99)	Male
Doug	Tech_support	[35,40)	Female
Emily	Sales	[40,90)	Male
Fred	Craft_repair	[1,35)	Female
Gladys	Craft_repair	[35,40)	Female
Henry	Adm_clerical	[1,35)	Male

(c) Public table

Fig. 1. Examples for hindering linking attacks.

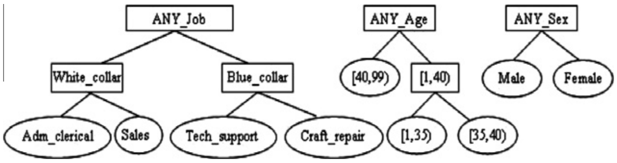


Fig. 2. Attribute value taxonomies for {Job, Age, Sex}.

important and should certainly be supported. However, privacy-sensitive information related to individual patients should be protected as well. To address the conflicting requirements of assuring privacy while supporting legitimate uses, the original data should be modified by applying some anonymization methods while ensuring that the anonymized data can be effectively used for data mining.

Generalization and suppression are popular anonymization methods. In generalization, quasi-identifier values are replaced with values that are less specific but are semantically consistent according to a given attribute value taxonomies. For example, in Fig. 2, the parent node, White_collar, is more general than its child nodes, Adm_clerical and Sales. The root node, ANY_Job, represents the most general value in Job. If the following information, "Job = Sales, Age = [35,40), Sex = Male", is too specific in a table, e.g., fewer than k men of age [35,40) work for sales, the probability for identifying these people to a specific record through {Job, Age, Sex} is greater than $1/k$. In this case, with the help of additional information, there is a chance that an attacker could uniquely identify these individual records from the data table. The larger the value of k results in greater generalization and better protection privacy. If the record is generalized as "Job = White_collar, Age = [1,40), Sex = male", more than k people will have the same person-identifiable information in the data, and therefore their privacy is better preserved. The most generalized form of a record is "ANY_Job, ANY_Age, ANY_Sex". When values are generalized to the highest level, this generalization is called suppression.

Meanwhile, information loss is an unfortunate consequence of anonymization. To make the anonymous data as useful as possible, the information loss must be minimized. The information metric for measuring the data usefulness can be categorized as a data metric and a search metric. A data metric measures the data quality in the entire anonymous table with respect to the data quality in the raw table. The problem of finding the optimal k-anonymous table using generalization has been proven to be NP-hard [1]. Therefore, heuristic algorithms are needed. A search metric is used to guide each step of the anonymization operations to identify an anonymous table with the maximum information or minimum distortion.

When the anonymous data are used to build classification models, protecting individual privacy in the data while ensuring that the data remain useful for building classification models is a challenge. Some data metrics, such as the *minimal distortion* [27] and the *discernibility metric* [1], have been considered for achieving k-anonymity. However, these data metrics do not consider any particular data mining task. As a result, the anonymous tables of these data metrics might not be suitable for every classification algorithm. Much research has been conducted to evaluate the data quality of anonymous tables for classification [12,16,30]. These efforts have not considered a search metric. For classification tasks, a more relevant approach is to search for a useful anonymization operation according to certain heuristics. An anonymization operation is ranked high if it preserves useful classification information. A search metric could be adopted to guide each step of the anonymization operations to identify an anonymous table using various anonymization algorithms, such as a greedy algorithm or a hill climbing optimization algorithm. Therefore, because the anonymous table identified by a search metric is eventually evaluated by a data metric, the two types of metrics usually share the same principle of measuring data quality. Some past research [8,31] has proposed a search metric based on the tradeoff principle between information gain and anonymity loss. However, this information gain metric is defined only for a single attribute in a single equivalence class and may not retain useful classification information.

In rough set theory, attribute reduction seeks to find a minimum subset of condition attributes that has the same classification ability as the set of all condition attributes with respect to the decision attributes [20,24,25]. The classification ability of all condition attributes with respect to the decision attributes can be measured by conditional entropy [29]. An anonymous table does not have to be close to the original table at the value level, but a classification model built on the anonymous table should be as good as a classification model built on the original table. For classification tasks, the data quality of a table can be considered as the classification ability. Therefore, we apply the conditional entropy to measure the classification ability of an anonymous table.

In this paper, we aim at releasing a k-anonymous table for modeling classification of the form $S(P \cup Q, D)$, where $P \cup Q$ is a finite set