On Mining Sensitive Rules to Identify Privacy Threats

Irene Díaz¹, Luis J. Rodríguez-Muñiz², and Luigi Troiano³

¹ Computer Science Department, University of Oviedo, Spain
sirene@uniovi.es
² Department of Statistics and O.R., University of Oviedo, Spain
luisj@uniovi.es
³ Department of Engineering, University of Sannio, Italy
troiano@unisannio.it

Abstract. Data mining techniques represent a useful tool to cope with privacy problems. In this work an association rule mining algorithm adapted to the privacy context is developed. The algorithm produces association rules with a certain structure (the premise set is a subset of the public features of a released table while the consequent is the feature to protect). These rules are then used to reveal and explain relationships from data affected by some kind of anonymization process and thus, to detect threats.

Keywords: disclosure control, association rules, data privacy, anonymity.

1 Introduction

The huge increase in digital data has led to emerge some concerns about data privacy, especially nowadays because WWW makes easy linking information about individuals obtained from different sources. It is easy to access to micro-data (data that are not summarized by some statistics) which are generally organized in tables whose attributes can (i) lead to identities, such as address, name, social security number, and (ii) release sensitive information, such as diseases and income, such those regarding census, medical issues, finance and others. In particular those attributes that are directly linked to identity are know as identifiers, whilst other attributes related at some extent to identity potentially able to identify an individual are known as quasi-identifiers.

In this environment data mining techniques are becoming extremely important to infer hidden information from data collections by providing patterns or existing models in data collections. The work presented in this paper tries to take advantage of association rules to study the possible threats in a given anonymous table. Firstly, an association rule based algorithm is developed to obtain a rule base where the antecedent set is a subset (possibly the whole set) of the quasi-identifier set while the consequent of each rule is the sensitive variable. After that, we will study the robustness of an anonymous table against attacks.
performed via this rule base. We will also study the conditions a rule must satisfy to be helpful in revealing information in a data set. It is worth to note that, differently form \( l \)-diversity and \( k \)-anonymity which are aimed at describing overall properties of released table, the proposed algorithm is aimed at describing single relationships between data, and thus to reveal leaks in the anonymization scheme.

The rest of the paper is organised as follows. Section 2 presents basic concepts in Data Privacy as well as the main metrics used to obtain an anonymous table. Section 3 shows some existing works in the context of data mining in privacy. Section 4 describes the RuleMiner algorithm. Finally, Section 5 shows an initial example of the procedure and Section 6 draws the conclusions and our future plans.

2 Privacy by Means of Metrics

Different metrics for measuring the level of privacy guaranteed by Statistical Disclosure Control (SDC) have been proposed over the time. They study the need for protection centred on limiting the ability to link released information to other external data. That limitation is controlled by identifying all attributes in the private information that could be used for linking with external information to uniquely identify individuals in the population. Such attribute set is named quasi-identifier.

Among them, in [15] and [17] is defined \( k \)-anonymity with respect to a quasi-identifier as the property that makes each record of a released table indistinguishable with at least \( k - 1 \) other records. Therefore, \( k \)-anonymity requires that each equivalence class contains at least \( k \) records. This property assures that if the released data satisfies \( k \)-anonymity with respect to a quasi-identifier, then the combination of the released data an the external sources on which the quasi-identifier is based, cannot link on the quasi-identifier or a subset of its attributes to match fewer than \( k \) records. Machanavajjhala et al. propose \( l \)-diversity [14] for providing privacy when the data publisher does not know what kind of information manages the attacker. Thus, \( l \)-diversity requires that the distribution of a sensitive attribute in each equivalence class has at least \( l \) values.

However \( l \)-diversity is limited in its assumption of adversarial knowledge. To avoid this lack, \( t \)-closeness [11] formalizes the idea of global background knowledge by requiring that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table. This effectively limits the amount of individual-specific information an observer can learn.

Other approaches try to prevent range disclosure, i.e., when an attacker is able to link an individual not to a specific sensitive value, but to a set of values that collectively disclose some sensible information [13]. In this sense, fuzzy set theory provides a natural framework to analyze data generalization and to identify threats to privacy (see [54]). Since generalization is about grouping elements in classes, and membership cannot be sharply defined, a class of elements can be regarded as fuzzy set. Privacy is preserved and disclosure protected, if the