Association Rule Hiding Methods

Vassilios S. Verykios
University of Thessaly, Greece

INTRODUCTION

The enormous expansion of data collection and storage facilities has created an unprecedented increase in the need for data analysis and processing power. *Data mining* has long been the catalyst for automated and sophisticated data analysis and interrogation. Recent advances in data mining and *knowledge discovery* have generated controversial impact in both scientific and technological arenas. On the one hand, data mining is capable of analyzing vast amounts of information within a minimum amount of time, an analysis that has exceeded the expectations of even the most imaginative scientists of the last decade. On the other hand, the excessive processing power of intelligent algorithms which is brought with this new research area puts at risk sensitive and confidential information that resides in large and distributed data stores.

Privacy and security risks arising from the use of data mining techniques have been first investigated in an early paper by O' Leary (1991). Clifton & Marks (1996) were the first to propose possible remedies to the protection of sensitive data and sensitive knowledge from the use of data mining. In particular, they suggested a variety of ways like the use of controlled access to the data, fuzzification of the data, elimination of unnecessary groupings in the data, data augmentation, as well as data auditing. A subsequent paper by Clifton (2000) made concrete early results in the area by demonstrating an interesting approach for privacy protection that relies on sampling. A main result of Clifton's paper was to show how to determine the right sample size of the public data (data to be disclosed to the public where sensitive information has been trimmed off), by estimating at the same time the error that is introduced from the sampling to the significance of the rules. Agrawal and Srikant (2000) were the first to establish a new research area, the *privacy preserving data mining*, which had as its goal to consider privacy and confidentiality issues originating in the mining of the data. The authors proposed an approach known as *data perturbation* that relies on disclosing a modified database with noisy data instead of the original database. The modified database could produce very similar patterns with those of the original database.

BACKGROUND

One of the main problems which have been investigated within the context of privacy preserving data mining is the so-called *association rule hiding*. Association rule hiding builds on the data mining area of *association rule mining* and studies the problem of hiding sensitive association rules from the data. The problem can be formulated as follows.

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of binary literals, called items. Let D be a transactional database, where each transaction T contains a set of items (also called an itemset) from I, such that $T \subseteq I$. A unique identifier TID (stands for transaction id) is associated with each transaction. We assume that the items in an itemset are sorted in lexicographic order. An *association rule* is an implication of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \varnothing$. We say that a rule $X \Rightarrow Y$ holds in the database D with *confidence* c if $|X \cup Y|/|X| \geq c$ (where $|X|$ is the cardinality of the set X) and *support* s if $|X \cup Y|/N \geq s$, where N is the number of transactions in D. An association rule mining algorithm proceeds by finding all itemsets that appear frequently enough in the database, so that they can be considered interesting, and by deriving from them all proper association rules that are strong (above a lower confidence level) enough. The association rule hiding problem aims at the prevention of a subset of the association rules from being disclosed during mining. We call these rules *sensitive*, and we argue that in order for a rule to become non-sensitive, its support and confidence must be brought below the minimum support and confidence threshold, so that it escapes mining at the corresponding levels of support and confidence. More formally we can state: Given a database D, a set R of rules mined from database D at a pre-specified threshold of support and confidence, and a subset R_h ($R_h \subset R$) of sensitive rules, the association

rule hiding refers to transforming the database D into a database D' of the same degree (same number of items) as D in such a way that only the rules in R-R_h can be mined from D' at either the pre-specified or even higher thresholds. We should note here that in the association rule hiding problem we consider the publishing of a modified database instead of the secure rules because we claim that a modified database will certainly have higher utility to the data holder compared to the set of secure rules. This claim relies on the fact that either a different data mining approach may be applied to the published data, or a different support and confidence threshold may be easily selected by the data miner, if the data itself is published.

It has been proved (Atallah, Bertino, Elmagarmid, Ibrahim, & Verykios, 1999) that the association rule hiding problem which is also referred to as the *database sanitization problem* is NP-hard. Towards the solution of this problem a number of heuristic and exact techniques have been introduced. In the following section we present a thorough analysis of some of the most interesting techniques which have been proposed for the solution of the association rule hiding problem.

MAIN FOCUS

In the following discussion we present three classes of state of the art techniques which have been proposed for the solution of the association rule hiding problem. The first class contains the *perturbation* approaches which rely on heuristics for modifying the database values so that the sensitive knowledge is hidden. The *use of unknowns* for the hiding of rules comprises the second class of techniques to be investigated in this expository study. The third class contains recent sophisticated approaches that provide a new perspective to the association rule hiding problem, as well as a special class of computationally expensive solutions, the *exact solutions*.

Perturbation Approaches

Atallah, Bertino, Elmagarmid, Ibrahim & Verykios (1999) were the first to propose a rigorous solution to the association rule hiding problem. Their approach was based on the idea of preventing disclosure of sensitive rules by decreasing the support of the itemsets generating the sensitive association rules. This reduced hiding approach is also known as frequent itemset hiding. The heuristic employed in their approach traverses the itemset lattice in the space of items from bottom to top in order to identify these items that need to turn from 1 to 0 so that the support of an itemset that corresponds to a sensitive rule becomes lower than the minimum support threshold. The algorithm sorts the sensitive itemsets based on their supports and then it proceeds by hiding all of the sensitive itemsets one by one. A major improvement over the first heuristic algorithm which was proposed in the previous work appeared in the work of Dasseni, Verykios, Elmagarmid & Bertino (2001). The authors extended the existing association rule hiding technique from using only the support of the generating frequent itemsets to using both the support of the generating frequent itemsets and the confidence of the association rules. In that respect, they proposed three new algorithms that exhibited interesting behavior with respect to the characteristics of the hiding process. Verykios, Elmagarmid, Bertino, Saygin & Dasseni (2004) along the same lines of the first work, presented five different algorithms based on various hiding strategies, and they performed an extensive evaluation of these algorithms with respect to different metrics like the execution time, the number of changes in the original data, the number of non-sensitive rules which were hidden (hiding side effects or false rules) and the number of "ghost" rules which were produced after the hiding. Oliveira & Zaiane (2002) extended existing work by focusing on algorithms that solely remove information so that they create a smaller impact in the database by not generating false or ghost rules. In their work they considered two classes of approaches: the pattern restriction based approaches that remove patterns completely from sensitive transactions, and the item restriction based approaches that selectively remove items from sensitive transactions. They also proposed various performance measures for quantifying the fraction of mining patterns which are preserved after sanitization.

Use of Unknowns

A completely different approach to the hiding of sensitive association rules was taken by employing the use of unknowns in the hiding process (Saygin, Verykios & Elmagarmid, 2002, Saygin, Verykios & Clifton, 2001). The goal of the algorithms that incorporate unknowns in the hiding process was to obscure a given