Full-Exact Approach for
Frequent Itemset Hiding

Tolga Ayav, Department of Computer Engineering, Izmir Institute of Technology, Izmir, Turkey
Belgin Ergenc, Department of Computer Engineering, Izmir Institute of Technology, Izmir, Turkey

ABSTRACT

This paper proposes a novel exact approach that relies on integer programming for association rule hiding. A large panorama of solutions exists for the complex problem of itemset hiding: from practical heuristic approaches to more accurate exact approaches. Exact approaches provide better solutions while suffering from the lack of performance and existing exact approaches still augment their methods with heuristics to make the problem solvable. In this case, the solution may not be optimum. This work presents a full-exact method, without any need for heuristics. Extensive tests are conducted on 10 real datasets to analyze distance and information loss performances of the algorithm in comparison to a former similar algorithm. Since the approach provides the optimum solution to the problem, it should be considered as a reference method.

Keywords: Association Rule Hiding, Cost Model, Exact Approach, Itemset Hiding, Side Effect

1. INTRODUCTION

Data mining field that aims to find interesting patterns from huge amounts of data attracted the increasing interest of organizations, researchers and practitioners. However this growing use of data mining technology in different domains increased the concern for privacy leading to an active research area named privacy preserving data mining. From a general point of view, privacy issues related to the application of data mining can be classified into two main categories, namely data hiding and knowledge hiding. Data hiding aims to remove confidential or private information from data prior to its publication. Knowledge hiding, on the other hand is concerned with the sanitization of data leading to disclosure of confidential and private knowledge (Gkoulalas-Divanis and Verykios, 2010). The problem of knowledge hiding requires sanitizing the input database $D$ in such a way that a set of sensitive knowledge is hidden, while most of the information in $D$ is maintained.

Association rule mining or frequent itemset mining is a major data mining methodology, mostly known in the area of market basket analysis and aims to capture relationships present among items in a transactional database (Agrawal et al., 1993) (Agrawal and Srikant, 1994). Despite its benefit in modern business,
frequent itemset mining can also pose a threat to privacy and security in a database sharing environment when precautions are not taken in its implementation (Atallah et al., 1999) (Oliveira and Zaiane, 2002). Frequent itemset hiding is specialization of the generic knowledge hiding problem where the main requirement asks for lowering the support of sensitive itemsets in the input database D so that sanitized database can be produced. Secondary requirements are minimization of deleted items and loss of non-sensitive frequent itemsets (Bonchi and Ferrari, 2011).

Large body of research emerged in the field of itemset hiding, since the database owners are in need of sharing data with their competitors, for their mutual benefit without revealing strategic patterns in the form of sensitive itemsets. Due to combinatorial nature of the problem of itemset hiding, proposed sanitization methodologies span from simple, time and memory efficient heuristics (Oliveira and Zaiane, 2002) (Oliveira and Zaiane, 2003a) (Verykios et al., 2004) (Amiri, 2007) (Wu et al., 2007) (Keer and Singh, 2012) (Yildiz and Ergenc, 2012), border-based approaches (Sun and Yu, 2005) (Sun and Yu, 2007) (Moustakides and Verykios, 2008) and reconstruction based approaches (Mielikainen, 2003) (Guo, 2007) (Lin and Liu, 2007) (Boora et al., 2009) (Mohaisen et al., 2010) to exact hiding (Menon et al., 2005) (Gkoulalas-Divanis and Verykios, 2006) (Gkoulalas-Divanis and Verykios, 2008) (Gkoulalas-Divanis and Verykios, 2009b) algorithms that offer guarantees on the quality of the computed hiding solution at an increased computational complexity cost. Whatever the technique used in sanitization; different attributes are used in selecting the transaction, itemset in the transaction or the item in the itemset to modify. Sanitization techniques try to minimize distance and/or information loss while the number of sensitive itemsets, characteristics of the data sets or user defined support vary.

The motivation for this work is to find a full exact solution to the problem of itemset hiding. Although exact approaches claim to model the itemset hiding as optimization problem whose objective is to find the optimum solution, the ones proposed by Menon (Menon et al., 2005) and Divanis (Gkoulalas-Divanis and Verykios, 2008) cannot be considered as entirely exact since they rely not only on integer programming but also some heuristics. To the best of authors’ knowledge, this work is the first one that models the entire problem as integer programming so that an optimum solution can be achieved without requiring any heuristics. Evaluation tests are carried out to measure distance and information loss performance of the proposed approach in comparison to an existing algorithm.

Organization of the paper is as follows; Section 2 gives preliminary information related to the problem definition of itemset hiding, metrics for side effects and a motivating example that will be used to explain the algorithms in the following sections. Section 3 starts with the explanation of the full exact approach of itemset hiding problem. Section 4 gives detailed performance analysis of the proposed method on UCI benchmark datasets (Coenen, 2003) while changing the number of sensitive itemsets in comparison to a similar exact algorithm. Section 5 is dedicated to the survey of existing approaches. Finally, Section 6 covers the conclusion remarks.

2. PRELIMINARIES AND MOTIVATING EXAMPLE

In this section, formulation of the problem of itemset hiding and metrics used in minimizing the side effect during sanitization process will be described. In the last part, a motivating example is presented, that example will be used in the following section to better explain exact itemset hiding algorithms.

2.1. Problem Formulation

Let \( I = \{a_1, a_2, \ldots, a_m\} \) be a set of literals, called items. Any subset of \( I : I_k \subseteq I \) is called an itemset. Let \( D \) denote a transactional database, where each transaction \( T_i \) is a tuple \( \{i, I_i\} \).
Digital Rights Management and Corporate Hegemony: Avenues for Reform
Nikhil Moro (2013). *Ethical Data Mining Applications for Socio-Economic Development* (pp. 281-299).
www.igi-global.com/chapter/digital-rights-management-corporate-hegemony/76266?camid=4v1a

Optimizing ETL by a Two-Level Data Staging Method
www.igi-global.com/article/optimizing-.etl-by-a-two-level-data-staging-method/168485?camid=4v1a