ABSTRACT

Association rule mining has been a highly active research field over the past decade. Extraction of frequency-related patterns has been applied to several domains. However, the way association rules are defined has limited people’s ability to obtain all the patterns of interest. In this chapter, the authors present an alternative approach that allows us to obtain new kinds of association rules that represent deviations from common behaviors. These new rules are called anomalous rules. To obtain such rules requires that we extract all the most frequent patterns together with certain extension patterns that may occur very infrequently. An approach that relies on anomalous rules has possible application in the areas of counterterrorism, fraud detection, pharmaceutical data analysis and network intrusion detection. They provide an adaption of measures of interest to our anomalous rule sets, and we propose an algorithm that can extract anomalous rules as well. Their experiments with benchmark and real-life datasets suggest that the set of anomalous rules is smaller than the set of association rules. Their work also provides evidence that our proposed approach can discover hidden patterns with good reliability.

INTRODUCTION

Traditionally, association rule (AR) mining is based on the definition of “frequent item-sets,” a concept that was originally derived using transactional data (Agrawal, Imielinski, & Swami, 1993). A significant drawback of the traditional approach to association rules is the large number of rules that are generated and processed. Even a small database often generates several thousand rules. In addition, potentially interesting “infrequent item-sets” are discarded a priori by the very definition of the association rules. These drawbacks have led to an increasing number of studies concerning interest measures (Geng & Hamilton, 2006; Tan, Kumar, & Srivastava, 2004), and they have prompted research into the connec-
tions between association rules and so-called “rare item-sets.”

The connection between association rules and rarities or anomalies is especially helpful in several specific knowledge domains in which the presence of unusual recorded data can prompt a decision to protect, counter-attack or to investigate further. Detection of anomalies is a specific case in the broader supercategory of hidden pattern detection. Anomalies are traditionally defined as deviations from a normal behavior (Denning, 1987). In network communications, this concept is particularly useful since it can be used to detect intrusions and to alert administrators to novel types of network attacks. In counter-terrorism, the detection of anomalies can help identify abnormal emails (or email content), unusual activities, etc. (Thuraisingham, 2004). However, a limitation of existing anomaly detection systems is the rate of false positives.

In this chapter, we propose a new set of anomalous association rules (AARs) that can reliably identify several kinds of hidden patterns, including those that may occur infrequently. A pattern that represents a common behavior consistent with the association rules, together with the anomalous patterns that are derived by association, can include non-frequent itemsets and can lead to reliable inferences about the data. Applications of anomalous rules extend from the areas previously highlighted to medical or agricultural domains, and also to general applications in which identifying rare patterns is considered important.

This chapter is organized as follows: we first offer a brief background on how to model deviations from a given common behavior, and we subsequently provide definitions of the anomalous rules and associated metrics that we used for our research. We also explore the effects of using various discretization methods. We present our algorithm for generating anomalous rules, and we illustrate the anomalous rule data mining process with simulated data.

Finally, we summarize the results of our experiments using real-world data. Our findings demonstrate the effectiveness of our chosen metrics and provide evidence that anomalous rules can be used to identify hidden patterns that may potentially reveal interesting associations.

**BACKGROUND**

Association rules (Agrawal et al., 1993) is a term that refers to patterns that relate sets of items. For a given set of \( k \) items, known as a \( k \)-itemset, an association rule has two components: an antecedent and a consequent, both of which may themselves involve \( k \)-itemsets. The association rule can be evaluated by two measures; support and confidence. Support quantifies the statistical significance of the pattern, and it is defined as the probability of a certain \( k \)-itemset being part of the dataset. The confidence, defined as the conditional probability between the pattern and the antecedent of the association rule, provides evidence of the relevance of a certain rule. An association rule is confirmed if its support and confidence values are greater than or equal to certain thresholds specified by the user.

Unlike the traditional frequent itemset framework that is used to obtain association rules, research into the detection of rare and hidden patterns must consider the specific definitions that pertain to each case. Several recent publications have explored the modeling of patterns that reflect deviations from a common behavior.

Exception rules (ERs) (Suzuki, 2002) represent changes in the consequent of an AR that are caused by the presence of a specific item. These changes are reinforced by a so-called reference rule, which treats both the item causing the exception and the new consequent. The reference rule states that the item causing the exception is not associated with the exception that it causes. In this context, the anomalous rules do not require the presence of
Related Content

Bayesian Belief Networks for Data Cleaning
[www.igi-global.com/chapter/bayesian-belief-networks-data-cleaning/26141?camid=4v1a](www.igi-global.com/chapter/bayesian-belief-networks-data-cleaning/26141?camid=4v1a)

Vector DNF for Datasets Classifications: Application to the Financial Timing Decision Problem
[www.igi-global.com/chapter/vector-dnf-datasets-classifications/26131?camid=4v1a](www.igi-global.com/chapter/vector-dnf-datasets-classifications/26131?camid=4v1a)

Advanced Data Mining and Visualization Techniques with Probabilistic Principal Surfaces: Applications to Astronomy and Genetics
[www.igi-global.com/chapter/advanced-data-mining-visualization-techniques/26144?camid=4v1a](www.igi-global.com/chapter/advanced-data-mining-visualization-techniques/26144?camid=4v1a)

SODA: A Service Oriented Data Acquisition Framework
[www.igi-global.com/chapter/soda-service-oriented-data-acquisition/63898?camid=4v1a](www.igi-global.com/chapter/soda-service-oriented-data-acquisition/63898?camid=4v1a)