Chapter 7.14
Mining for Mutually Exclusive Items in Transaction Databases

George Tzanis
Aristotle University of Thessaloniki, Greece

Christos Berberidis
Aristotle University of Thessaloniki, Greece

ABSTRACT

Association rule mining is a popular task that involves the discovery of co-occurences of items in transaction databases. Several extensions of the traditional association rule mining model have been proposed so far; however, the problem of mining for mutually exclusive items has not been directly tackled yet. Such information could be useful in various cases (e.g., when the expression of a gene excludes the expression of another), or it can be used as a serious hint in order to reveal inherent taxonomical information. In this article, we address the problem of mining pairs of items, such that the presence of one excludes the other. First, we provide a concise review of the literature, then we define this problem, we propose a probability-based evaluation metric, and finally a mining algorithm that we test on transaction data.

INTRODUCTION

Association rules are expressions that describe a subset of a transaction database. When mining for such patterns, it is quite often that we come up with a large number of rules that appear to be too specific and not very interesting. A rule that relates two specific products in a market basket database is not very likely to be really strong compared to a rule that relates two groups or two families of products. Hierarchical relationships among items in a database can be used in order to aggregate the weak, lower-level rules into strong, higher-level rules, producing *hierarchical, multiple level,* or *generalized association rules*. However, such information is not always explicitly provided, although it might exist.

Mining for taxonomies is a really challenging task that, to the best of our knowledge, has not been approached yet. Taxonomies are conceptual

hierarchies, implemented by *is-a* relationships. The discovery of such relationships would involve the complete description and formulation of concepts that are more general or more specific than others. To learn taxonomies from data implies the automatic extraction of human concepts from the data, with the use of an algorithm. To our understanding, this is virtually impossible. However, we believe that when mining for various types of patterns, one can get serious hints about possible hierarchical relationships. Let us say, for instance, that a supermarket customer is vegetarian. Then it would be really rare for this customer to buy both veggie burgers *and* red meat. It seems that the two products *exclude* each other. When one of them is present, then the probability to also find the other one is very low. Motivated by that observation, we propose a method for mining for *mutually exclusive* items. Such information is also useful regardless of its use as a taxonomy clue. In this article, we define the problem of mining for mutually exclusive items. We propose a probability-based mutual exclusion metric and a mining algorithm that we test on transaction data.

The article is organized as follows. The next section presents the required background knowledge. This is followed by a short review of the relative literature. The next section contains the description of the proposed approach, definitions of terms and notions used, the proposed algorithm, a novel metric for measuring the mutual exclusion, and an illustrative example of our approach. Next we present our experiments, and then we discuss the presented approach. The final section contains our conclusions and our ideas for future research.

PRELIMINARIES

The association rules mining paradigm involves searching for co-occurrences of items in transaction databases. Such a co-occurrence may imply a relationship among the items it associates. These relationships can be further analyzed and may reveal temporal or causal relationships, behaviors, and so forth.

The formal statement of the problem of mining association rules can be found in Agrawal, Mannila, Srikant, Toivonen, and Verkamo (1996). Given a finite multiset of transactions D, the problem of mining association rules is to generate all association rules that have support and confidence at least equal to the user-specified minimum support threshold (*min_sup*) and minimum confidence threshold (*min_conf*), respectively.

The problem of discovering all the association rules can be decomposed into two subproblems (Agrawal, Imielinski, & Swami, 1993):

1. The discovery of all itemsets that have support at least equal to the user-specified *min_sup* threshold. These itemsets are called *large* or *frequent* itemsets.
2. The generation of all rules from the discovered frequent itemsets. For every frequent itemset F, all nonempty subsets of F are found. For every such subset S, a rule of the form $S \Rightarrow F\text{-}S$ is generated, if the confidence of the rule is at least equal to the user-specified *min_conf* threshold.

Another method to extract strong rules is the use of *concept hierarchies*, also called *taxonomies,* that exist in various application domains, such as market basket analysis. Taxonomy is a concept tree, where the edges represent is-a relationships from the child to the parent. An example of such a relationship is: "Cheddar is-a cheese is-a dairy product is-a food is-a product." When a taxonomy about a domain of application is available, a number of usually high-confidence rules that are too specific (having low support) can be merged, creating a rule that aggregates the support and, therefore, the information, in a higher abstraction level of the individual rules. In other words, "looser" associations at the lower levels of the taxonomy are summarized, producing "winner"