Mining Association Rules Using Frequent Closed Itemsets

Nicolas Pasquier
Université de Nice-Sophia Antipolis, France

INTRODUCTION

In the domain of knowledge discovery in databases and its computational part called data mining, many works addressed the problem of association rule extraction that aims at discovering relationships between sets of items (binary attributes). An example association rule fitting in the context of market basket data analysis is *cereal* $\wedge$ *milk* $\rightarrow$ *sugar* (support 10%, confidence 60%). This rule states that 60% of customers who buy cereals and sugar also buy milk, and that 10% of all customers buy all three items. When an association rule support and confidence exceed some user-defined thresholds, the rule is considered relevant to support decision making. Association rule extraction has proved useful to analyze large databases in a wide range of domains, such as marketing decision support; diagnosis and medical research support; telecommunication process improvement; Web site management and profiling; spatial, geographical, and statistical data analysis; and so forth.

The first phase of association rule extraction is the data selection from data sources and the generation of the data mining context that is a triplet D = (O, I, R), where O and I are finite sets of objects and items respectively, and R $\subseteq$ O $\times$ I is a binary relation. An item is most often an attribute value or an interval of attribute values. Each couple $(o, i) \in$ R denotes the fact that the object $o \in$ O is related to the item $i \in$ I. If an object o is in relation with all items of an *itemset* I (a set of items) we say that o contains I.

This phase helps to improve the extraction efficiency and enables the treatment of all kinds of data, often mixed in operational databases, with the same algorithm. Data-mining contexts are large relations that do not fit in main memory and must be stored in secondary memory. Consequently, each context scan is very time consuming.

Table 1. Example context

OID	Items
1	A C D
2	B C E
3	A B C E
4	B E
5	A B C E
6	B C E

BACKGROUND

The support of an itemset I is the proportion of objects containing I in the context. An itemset is frequent if its support is greater or equal to the minimal support threshold defined by the user. An association rule r is an implication with the form r: $I_1 \rightarrow I_2 - I_1$ where I_1 and I_2 are frequent itemsets such that $I_1 \subset I_2$. The confidence of r is the number of objects containing I_2 divided by the number of objects containing I_1. An association rule is generated if its support and confidence are at least equal to the minsupport and minconfidence thresholds. Association rules with 100% confidence are called *exact association rules*; others are called *approximate association rules*. The natural decomposition of the association rule-mining problem is:

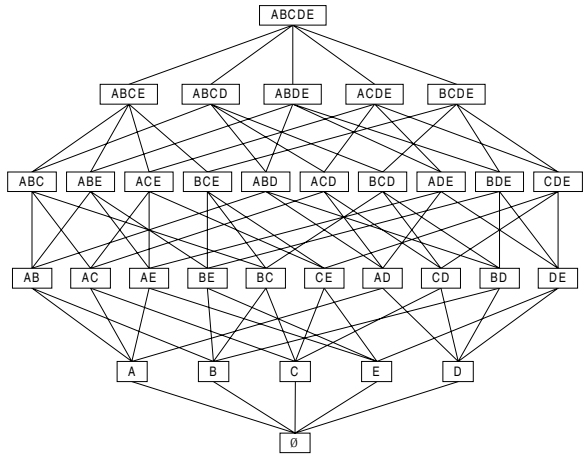
1. Extracting frequent itemsets and their support from the context.
2. Generating all valid association rules from frequent itemsets and their support.

The first phase is the most computationally expensive part of the process, since the number of potential frequent itemsets $2^{|I|}$ is exponential in the size of the set of items, and context scans are required. A trivial approach would consider all potential frequent itemsets at the same time, but this approach cannot be used for large databases where I is large. Then, the set of potential frequent itemsets that constitute a lattice called *itemset lattice* must be decomposed into several subsets considered one at a time.

Level-Wise Algorithms for Extracting Frequent Itemsets

These algorithms consider all itemsets of a given size (i.e., all itemsets of a level in the itemset lattice) at a time. They are based on the properties that all supersets of an infrequent itemset are infrequent and all subsets of a frequent itemset are frequent (Agrawal et al., 1995). Using this property, the candidate k-itemsets (itemsets of size k) of the k^{th} iteration are generated by joining two frequent (k-1)-itemsets discovered during the preceding

Figure 1. Itemset lattice



iteration, if their *k*-1 first items are identical. Then, one database scan is performed to count the supports of the candidates, and infrequent ones are pruned. This process is repeated until no new candidate can be generated.

This approach is used in the well known APRIORI and OCD algorithms. Both carry out a number of context scans equal to the size of the largest frequent itemsets. Several optimizations have been proposed to improve the efficiency by avoiding several context scans. The COFI* (El-Hajj & Zaïane, 2004) and FP-GROWTH (Han et al., 2004) algorithms use specific data structures for that, and the PASCAL algorithm (Bastide et al., 2000) uses a method called *pattern counting inference* to avoid counting all supports.

Algorithms for Extracting Maximal Frequent Itemsets

Maximal and minimal itemsets are defined according to the inclusion relation. Maximal frequent itemsets are frequent itemsets of which all supersets are infrequent. They form a border under which all itemsets are frequent; knowing all maximal frequent itemsets, we can deduce all frequent itemsets, but not their support. Then, the following approach for mining association rules was proposed:

1. Extracting maximal frequent itemsets and their supports from the context.
2. Deriving frequent itemsets from maximal frequent itemsets and counting their support in the context during one final scan.
3. Generating all valid association rules from frequent itemsets.

These algorithms perform an iterative search in the itemset lattice *advancing* during each iteration by one level from the bottom upwards, as in APRIORI, and by one or more levels from the top downwards. Compared to preceding algorithms, both the number of iterations and, thus, the number of context scans and the number of CPU operations carried out are reduced. The most well known algorithms based on this approach are PINCER-SEARCH (Lin & Kedem, 1998) and MAX-MINER (Bayardo, 1998).

Relevance of Extracted Association Rules

For many datasets, a huge number of association rules is extracted, even for high minsupport and minconfidence values. This problem is crucial with correlated data, for which several million association rules sometimes are extracted. Moreover, a majority of these rules bring the same information and, thus, are redundant. To illustrate this problem, nine rules extracted from the mushroom dataset (ftp://ftp.ics.uci.edu/pub/machine-learning-databases/mushroom/) are presented in the following. All have the same support (51%) and confidence (54%), and the item *free gills* in the antecedent:

1. free_gills → edible
2. free_gills → edible, partial_veil
3. free_gills → edible, white_veil
4. free_gills → edible, partial_veil, white_veil
5. free_gills, partial_veil → edible
6. free_gills, partial_veil → edible, white_veil
7. free_gills, white_veil → edible
8. free_gills, white_veil → edible, partial_veil
9. free_gills, partial_veil, white_veil → edible

The most relevant rule from the viewpoint of the user is rule 4, since all other rules can be deduced from this one, including support and confidence. This rule is a non-redundant association rule with minimal antecedent and maximal consequent, or minimal non-redundant rule, for short.

Association Rules Reduction Methods

Several approaches for reducing the number of rules and selecting the most relevant ones have been proposed.

The application of templates (Baralis & Psaila, 1997) or Boolean operators (Bayardo, Agrawal & Gunopulos, 2000) allows selecting rules according to the user's preferences.

When taxonomies of items exist, generalized association rules (Han & Fu, 1999) (i.e., rules between items of different levels of taxonomies) can be extracted. This produces fewer but more general associations. Other statistical measures, such as Pearson's correlation or c^2,