Towards Comprehensive Concept Description
Based on Association Rules

Petr Berka
University of Economics
W. Churchill Sq. 3, 130 67 Prague, Czech Republic
berka@vse.cz

Abstract. The paper presents two approaches to post-processing of association rules that are used for concept description. The first approach is based on the idea of meta-learning; a subsequent association rule mining step is applied to the results of ”standard” association rule mining. We thus obtain ”rules about rules” that in a condensed form represent the knowledge found using association rules generated in the first step. The second approach finds a ”core” part of the association rules that can be used to derive the confidence of every rule created in the first step. Again, the core part is substantially smaller than the set of all association rules. We experimentally evaluate the proposed methods on some benchmark data taken from the UCI repository. The system LISp-Miner has been used to carry out the experiments.

Keywords: concept description, association rules, meta-learning.

1 Introduction

Concept description is one of the typical data mining tasks. According to CRISP-DM methodology concept description ”aimes at understandable descriptions of concepts or classes” [9]. Concept description is thus similar to classification as there are predefined classes (given by values of the target attribute) we are interested in. But unlike to classification, the focus of concept description is on understandability not on classification accuracy. So association rules, decision rules or decision trees are preferred to model the concepts.

In our paper we will focus on concept description using association rules. Association rules have been proposed by R. Agrawal in the early 90th as a tool for so called market basket analysis [2]. An association rule has the form of an implication

\[ X \implies Y \]

where \(X\) and \(Y\) are sets of items and \(X \cap Y = \emptyset\). An association rule expresses that transactions containing items of set \(X\) tend to contain items of set \(Y\), so e.g. a rule \(\{A, B\} \implies \{C\}\) says, that customers who buy products \(A\) and \(B\) also often buy product \(C\). Such statements can be used to guide the placement of goods in a store, for cross-selling or to promote new products. This idea of association
rules can be applied to any data in the tabular, attribute-value form. So data describing values of attributes can be analyzed in order to find associations between conjunctions of attribute-value pairs (categories). Let us denote these conjunctions as \( \text{Ant} \) (antecedent) and \( \text{Suc} \) (succedent) and the association rule as

\[
\text{Ant} \implies \text{Suc}
\]

When using association rules for concept description, \( \text{Suc} \) will be a category of the target attribute.

The two basic characteristics of an association rule are support and confidence. Support is the estimate of the probability \( P(\text{Ant} \land \text{Suc}) \), (the frequency of \( \text{Ant} \land \text{Suc} \) is the absolute support), confidence is the estimate of the probability \( P(\text{Suc}|\text{Ant}) \).

In association rule discovery the task is to find all rules with support and confidence above the user-defined thresholds \( \minconf \) and \( \minsup \). There is a number of algorithms, that perform this task. The probably best-known algorithm \textit{apriori} proceeds in two steps. All frequent itemsets are found in the first step during breath-first search in the space of all frequent itemsets. Then, association rules with a confidence of at least \( \minconf \) are generated in the second step [2]. Another well known algorithm is \textit{FP-Growth}. This algorithm uses FP-tree to generate frequent itemsets. This way of representing the itemsets reduces the computational costs because (unlike apriori) it requires only two scans of the whole data [12]. The found frequent itemsets are then again splitted into antecedent and succedent to create a rule. The search space of all possible itemsets (or conjunctions of categories) can be very huge. For \( K \) items, there is \( 2^K \) itemsets, for \( K \) categorial attributes \( A_1, A_2, ... A_K \), having \( v_1, v_2, ... v_K \) distinct values, the number of all possible conjunctions is

\[
\prod_{i=1}^{K} (1 + v_i) - 1.
\]  

The main problem when using association rules for data mining is their interpretation. Usually we end up with a huge number of associations and each of them might be interesting for the domain expert or end-user. So some automatic support for the interpretation in the form of association rules post-processing would be of a great help. We present some ideas in this direction and show their experimental evaluation using \textit{LISp-Miner}, a data mining toolbox for mining different types of rules, that is under development at the University of Economics, Prague [16,18].

The rest of the paper is organized as follows. Section 2 reviews work related to the problem of post-processing of association rules, section 3 gives an overview of GUHA method and the 4FT rules, section 4 introduces the concept of association meta-rules and describes how they can be obtained using 4FT-Miner procedure, section 5 presents \textit{KEX}, another procedure of the \textit{LISp-Miner} system.