The TVpercent criteria to eliminate uninformative models among Association Rules

Le critère VT100 pour éliminer les modèles insuffisants parmi les règles d’association

Alain MORINEAU
MODULAD, Paris (France) – alain.morineau@modulad.fr

Ricco RAKOTOMALALA
Université Lumière, Lyon (France) - ricco.rakotomalala@univ-lyon2.fr

Keywords: Association rules, TVpercent criteria, test value, model comparison.

An association rule is written \{A=>C\} where A and C are sets of items. In a market basket analysis for example, items are products purchased by customers. The rule \{A=>C\} says that, when transactions in the database contain the items A, we probably observe also the items C. Such rules are elementary models based on binary observations: the simpler the model, the greater the number of compatible models. The classical context of association rule analysis may be viewed as a problem of comparing, ordering and choosing the best models among a huge set of potential models that are pulled from the database.

Association rule mining algorithms potentially extract a huge set of rules which have not the same relevance, and among which there are many redundant rules. Automating the process of culling the most interesting rules is an area of active research. The main purpose of this paper is to propose a criterion in order to rank and eliminate irrelevant rules.

In a first part, we introduce the Test-Value or TV criteria as a measure of how far we are from the independence hypothesis H0 which characterizes uninteresting models (Morineau, 1984 ; Lebart et al., 1995). This criterion is a straight transformation of the p-value associated with the test of H0. As such, it shares the property of any classical statistical index in data mining context: for the same constant phenomenon, the value inopportune increases with the number of observations.

In a recent paper, we proposed a standardized version of the Test-Value named TVpercent criterion (Morineau-Rakotomalala, 2006). Applications to association rules proved it to be an interesting criterion to eliminate statistically uninteresting rules without being influenced by the number of occurrences.

In this paper we show how this criterion can be used to solve the problem of eliminating uninformative models among association rules. Let A and X be itemsets of antecedents and C the consequent itemset. We are interested in comparing two rules that both pass the minimum thresholds of support (how often the rules apply), confidence (percentage of times the rules are correct) and TVpercent criteria (how far the rules from independence):

R1 : \{A=>C\}  
R2 : \{A and X=>C\}.
Rule R2 is more precise than R1 but confidence of R2 is smaller than confidence of R1. Does the precision supplied by R2 statistically counterbalance the loss in confidence? We answer the question the following way. In the case of independence, the null distribution of the number of occurrences (the support of the rule) is the hypergeometric distribution whose parameters are the total number of transactions, the frequency of the antecedent and the frequency of the consequent.

To compare R1 and R2 and evaluate whether X is a statistically irrelevant factor for the rule, we consider only the \( n(A) \) transactions with A as the whole population of transactions. Among them, there is a known number \( n(A,C) \) of transactions with the consequent C. We know the frequency \( n(A,X) \) of transactions containing A and X. Among them, the number \( n(A,X,C) \) of transactions containing A,X and C has an hypergeometric null distribution whose three parameters are the three preceding frequencies.

Using this null distribution, it is easy to compute the \( VT\text{percent} \) criterion to evaluate whether R2 supersedes R1 or R1 is sufficient.

We conclude this part of the paper with a complete strategy to eliminate uninformative rules. This may be exemplified here: suppose R1: \( \{A=>C\} \); R2: \( \{B=>C\} \) et R12: \( \{A,B=>C\} \). We compute the \( TV\text{percent} \) criterion which evaluates the partial contributions of A and B in rule R12 compared to R2 and respectively R1. If both \( TV\text{percent} \) pass the minimum threshold, we can add rule R12 to the base of interesting rules; if not we discard R12 as uninformative compared to R1 and R2.

The last part of the paper gives numerical illustrations using real databases of transactions which demonstrate how many rules can be eliminated with this statistical criterion.

Références

