Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2007) 596–613
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set

Marie Plasse[a, b, *], Ndeye Niang[a], Gilbert Saporta[a], Alexandre Villeminot[b], Laurent Leblond[b]

[a] *CNAM, Laboratoire CEDRIC, 292 Rue St Martin Case 441, 75141 Paris Cedex 03, France*
[b] *PSA Peugeot Citroën, Zone Aéronautique Louis Bréguet, Route Militaire Nord, 78943 Vélizy Villacoublay, France*

Available online 1 March 2007

Abstract

A method to analyse links between binary attributes in a large sparse data set is proposed. Initially the variables are clustered to obtain homogeneous clusters of attributes. Association rules are then mined in each cluster. A graphical comparison of some rule relevancy indexes is presented. It is used to extract best rules depending on the application concerned. The proposed methodology is illustrated by an industrial application from the automotive industry with more than 80 000 vehicles each described by more than 3000 rare attributes.

Keywords: Association rules mining; Variable clustering; Large sparse matrix; Binary attributes; Rule relevancy index

0. Introduction

We consider the problem of discovering links between binary attributes in the case of large sparse matrices. Our sample data from the automotive industry consists of more than 80 000 vehicles each described by more than 3000 attributes. Each attribute is a binary variable equal to 1 if the vehicle has the attribute, 0 otherwise.

Our data can be considered as basket data and then a first idea is to mine association rules to find frequent co-occurrences of attributes. In our case, threshold configuration for support and confidence is particularly tricky. Minimum support has to be very low because vehicle attributes are extremely rare contrary to basket data. In addition, by a slight threshold variation, the number of rules increases rapidly.

To solve this problem we propose to cluster variables in order to build homogeneous groups of attributes and then mine association rules inside each of these groups. We have used several clustering methods and compared resulting partitions. The study shows that the combined use of association rules and classification methods is more relevant. Actually this approach brings about an important decrease in the number of rules produced. Furthermore, it appears that complex rules are always generated by the same grouped attributes identified through variable clustering.

* Corresponding author. CNAM, Laboratoire CEDRIC, 292 Rue St Martin Case 441, 75141 Paris Cedex 03, France. Tel.: +33 1 57 59 08 15; fax: +33 1 41 36 30 46.
E-mail address: marie.plasse@mpsa.com (M. Plasse).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.02.020

Even if we reduce number of rules, we still need to sort them from the most relevant to the less interesting. There are many indexes that measure statistical interest of association rules and the choice of one depends on the application. We have performed an empirical and graphical comparison to help to select the most appropriate.

After reviewing the basics of association rules mining in the first section, in the second we present an overview of variable clustering methods. Then in the third section we describe the combined use of these two methods. Finally, in the last section we compare some rule relevancy indexes. To illustrate our approach, each section contains a detailed example using industrial data.

1. Association rules mining

1.1. Algorithms to mine association rules

Association rules mining has been developed to analyse basket data in a marketing environment. Input data are composed of transactions: each transaction consists of items purchased by a consumer during a single visit. Output data are composed of rules. For example, a rule can be "90% of transactions that involve the purchase of bread and butter also include milk" (Agrawal et al., 1993). Even if this method has been introduced in the context of Market Business Analysis, it has many applications in other fields, like webmining or textmining. It can actually be used to search for frequent co-occurrences in every large data set.

A rule is an implication $A \rightarrow C$. The left part of the rule is called the antecedent and the right, the consequent. The sets A and C are disjointed as we cannot find the same item in both the antecedent and consequent. A rule makes sense thanks to its support $s = sup(A \rightarrow C) = P(A \cap C)$ and its confidence $c = conf(A \rightarrow C) = P(C/A)$.

The first efficient algorithm to mine association rules is *APriori* (Agrawal and Srikant, 1994). The first step of this algorithm is the research of frequent itemsets. The user gives a minimum threshold for the support and the algorithm searches all itemsets that appear with a support greater than this threshold. The second step is to build rules from itemsets found in the first step. The algorithm computes confidence of each rule and keeps only those where confidence is greater than a threshold defined by the user.

As we will see in the application, one of the main problems is to define support and confidence thresholds. *Apriori* is based on the property that every subset of a frequent itemset is also frequent. Candidate k-itemsets are generate in the kth read of the data set and their supports are computed in the $k+1$th read. If K is the largest size of frequent itemsets, the total number of reads is $K+1$. Other algorithms have been proposed to decrease the count of reads of the database and to improve computational efficiency. Among them, we can quote *Eclat* (Zaki, 2000), *Partition* (Savasere et al., 1995), *Sampling* (Toivonen, 1996), *DIC* (Brin et al., 1997a) or *FP-Growth* (Han et al., 2000). All of these algorithms furnish the same results as rules searching is deterministic. We have used *Apriori* and *Eclat* because they perform fastest on our sparse data.

1.2. Application of association rules mining to industrial data

To apply association rules mining, we consider vehicles as transactions and their binary attributes, as items.

1.2.1. A large sparse data set of industrial data

The sample data from the automotive industry consists of more than 80 000 vehicles each described by more than 3000 binary attributes. Simple graphics illustrate that we are dealing with extremely sparse data. Our binary matrix contains about 0.13% of "1". The most frequent attribute appears on 12% of vehicles but 97% of attributes appear on less than 1% of vehicles as shown in Fig. 1. In addition, a vehicle has an average of four attributes. A few vehicles have more than 10 attributes but most have between one and five (Fig. 2).

1.2.2. Too many association rules

With a minimum support of 500 vehicles and a minimum confidence of 50%, *the algorithms* produce 18 rules with a maximum size of three items (Table 1). As items are rare events, the minimum support threshold has to be reduced to identify less common links. We want to find rules with a minimum support of 100 vehicles. As Table 1 shows, the number and complexity of rules increase enormously.