Closed-set-based Discovery of Representative Association Rules Revisited

José L Balcázar, Cristina Tîrnăucă

Departamento de Matemáticas, Estadística y Computación
Universidad de Cantabria, Santander, Spain
{joseluis.balcazar, cristina.tirnauca}@unican.es

Abstract. The output of an association rule miner is often huge in practice. This is why several concise lossless representations have been proposed, such as the "essential" or "representative" rules. We revisit the algorithm given by Kryszkiewicz (Int. Symp. Intelligent Data Analysis 2001, Springer-Verlag LNCS 2189, 350–359) for mining representative rules. We show that its output is sometimes incomplete, due to an oversight in its mathematical validation, and we propose an alternative complete generator that works within only slightly larger running times.

1 Introduction

Association rule mining is among the most popular conceptual tools in the field of Data Mining. We are interested in the process of discovering and representing regularities between sets of items in large scale transactional data. Syntactically, the association rule representation has the form of an implication, $X \to Y$; however, whereas in Logic such an expression is true if and only if Y holds whenever X does, an association rule is a partial implication, in the sense that it is enough if Y holds *most of the times* X does.

To endow association rules with a definite semantics, we need to make precise how this intuition of "most of the times" is formalized. There are many proposals for this formalization. One of the frequently used measures of intensity of this kind of partial implication is its *confidence*: the ratio between the number of transactions in which X and Y are seen together and the number of transactions that contain X. In most application cases, the search space is additionally restricted to association rules that meet a minimal *support* criterion, thus avoiding the generation of rules from items that appear very seldom together in the dataset (formal definitions of support and confidence are given in Section 2.1).

Many association rule miners exists, Apriori (see Agrawal et al. (1996)) being one of the most widely discussed and used. The major problem shared by all mining algorithms is that, in practice, even for reasonable support and confidence thresholds, the output is often huge. Therefore, several concise lossless representations of the whole set of association rules have been proposed. These representations are based on different notions of "redundancy". In one of these, a rule is redundant if it is possible to compute exactly its confidence and support from other information such as the confidences and supports of other *informative* rules (see