Available online at www.sciencedirect.com
SCIENCE DIRECT®
Fuzzy Sets and Systems 149 (2005) 73–85
FUZZY sets and systems
www.elsevier.com/locate/fss

Elicitation of fuzzy association rules from positive and negative examples

M. De Cock*, C. Cornelis, E.E. Kerre

Fuzziness and Uncertainty Modelling Research Unit, Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 (S9), 9000 Gent, Belgium

Available online 19 August 2004

Abstract

The aim of this paper is to provide a crystal clear insight into the true semantics of the measures of support and confidence that are used to assess rule quality in fuzzy association rule mining. To achieve this, we rely on two important pillars: the identification of transactions in a database as positive or negative examples of a given association between attributes, and the correspondence between measures of support and confidence on one hand, and measures of compatibility and inclusion on the other hand. In this way we remove the "mystery" from recently suggested quality measures for fuzzy association rules.

Keywords: Fuzzy association rules; Support; Confidence; Positive and negative examples; Compatibility; Inclusion

1. Introduction

Association rules [1] provide a convenient and effective way to identify and represent certain dependencies between attributes in a database. Originally, association rules emerged in the domain of shops and customers; the basic idea is to identify frequent itemsets in market baskets, i.e., groups of products frequently bought together, so storekeepers may use this information to decide on what to put on sale, how to place merchandize on shelves to maximize a cross-selling effect, how to advertise, etc. Evidently, the application of association rules is not limited to marketing problems: in fact they can shed light on a wide

* Corresponding author.
E-mail addresses: Martine.DeCock@UGent.be (M. De Cock), Chris.Cornelis@UGent.be (C. Cornelis), Etienne.Kerre@UGent.be (E.E. Kerre).
URL: http://fuzzy.UGent.be.

0165-0114/$ - see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.fss.2004.07.010

range of knowledge discovery and decision making problems. Given the massive data archives maintained by most firms nowadays, it comes as no surprise that easy-to-handle and easy-to-grasp mechanisms like association rules have risen to great popularity.

Association rule mining is traditionally performed on a data table with binary attributes. Conceptually, a record x in the data table represents a customer transaction, whereas the attributes represent items that may be either purchased in that transaction, or not. Therefore, for each attribute A, $A(x)$ is either 1 or 0 indicating whether or not item A was bought in transaction x. An association rule is an expression of the form $A \Rightarrow B$ in which A and B are attributes, such as *cheese* $\Rightarrow$ *bread*. The meaning is that when A is bought in a transaction, B is likely to be bought as well. In an extended approach, the antecedent and the consequent of an association rule are sets of attributes. Considering this more general definition, however, would complicate the notation without providing additional benefit for the issues we want to deal with in this paper. Furthermore, since mining algorithms tend to generate too many rules, there is a trend to focus on simple association rules, i.e., those containing only one attribute in the consequent, and use them as building blocks to construct more general rules if required [8,9].

Association rules can be rated by a number of quality measures (for a recent, comprehensive overview of what is available, we refer to [24]), among which *support* and *confidence* stand out as the two essential ones. Support measures the statistical significance of a candidate rule $A \Rightarrow B$ as the fraction of transactions in which both A and B were bought. Confidence assesses the strength of a rule as the fraction of transactions containing A that contain B as well. The basic problem of mining association rules is then to generate all association rules $A \Rightarrow B$ that have support and confidence greater than user-specified thresholds.

In most real life applications, databases contain many other attribute values besides 0 and 1. Very common for instance are quantitative attributes such as *age* or *income*, taking values from a partially ordered, numerical scale, often a subset of the real numbers. One way of dealing with a quantitative attribute like *cost* is to replace it by a few other attributes that form a crisp partition of the range of the original one, such as $low = [0, 100[$, $medium = [100, 300[$ and $high = [300, +\infty[$. Now we can consider these new attributes as binary ones that have value 1 if the *cost* attribute equals a value within their range, and 0 otherwise. In this way, the problem is reduced to the mining procedure described above (the generated rules are now called quantitative association rules [23]). From an intuitive viewpoint, it makes more sense, however, to draw values from the interval $[0, 1]$ (instead of just $\{0, 1\}$), to allow records to exhibit a given attribute to a certain extent only. In this way binary attributes are replaced by fuzzy ones. The corresponding mining process yields fuzzy (quantitative) association rules (see, e.g., [4–7,9,11,13,15–17]).

In the traditional approach to association rule mining algorithms (including quantitative and fuzzy association rule mining), one merely thinks in terms of positive examples: especially when determining the degree of support, only the number of transactions in favour of the rule is accounted for. As we argued in [11], the remaining transactions can still be partitioned into those that actually violate the rule, and those which do not carry any relevant information. In other words, “not being a positive example" of a rule is not the same as “being a negative example". Realizing this provides deeper insight into the semantics of the quality measures as we will show in this paper.

On another count, it is sometimes also useful to detect negative associations (denoted $A \Rightarrow co\, B$), whose intended meaning is that transactions containing A are unlikely to contain B as well. As a somewhat frivolous example, we might quote *lucky-in-love* $\Rightarrow$ *co (lucky-in-games)*. Such patterns have received quite some attention lately (see, e.g., [6,21,26,28]); we will show that they can be embedded elegantly into our framework of positive and negative examples.