Mining Positive and Negative Fuzzy Association Rules*

Peng Yan[1], Guoqing Chen[1],
Chris Cornelis[2], Martine De Cock[2], and Etienne Kerre[2]

[1] School of Economics and Management, Tsinghua University,
Beijing 100084, China
{yanp,chengq}@em.tsinghua.edu.cn

[2] Fuzziness and Uncertainty Modelling Research Unit, Ghent University,
Krijgslaan 281 (S9), B–9000 Gent, Belgium
{chris.cornelis, martine.decock, etienne.kerre}@UGent.be
http://fuzzy.UGent.be

Abstract. While traditional algorithms concern positive associations between binary or quantitative attributes of databases, this paper focuses on mining both positive and negative *fuzzy* association rules. We show how, by a deliberate choice of fuzzy logic connectives, significantly increased expressivity is available at little extra cost. In particular, rule quality measures for negative rules can be computed without additional scans of the database.

Keywords: fuzzy association rules, positive and negative associations, quantitative attributes

1 Introduction and Motivation

Association rules [1], which provide a means of presenting dependency relations between attributes in databases, have become one of the most important fields in knowledge discovery. An association rule has the form $X \Rightarrow Y$, where X and Y are two separate sets of attributes (itemsets). An example of an association rule is {mobile, batteries} ⇒ {phone card}, which means that a customer who buys a mobile and batteries is likely to buy a phone card as well.

Since the attributes of real applications are not restricted to binary values but also quantitative ones like age and income exist, mining quantitative association rules is regarded meaningful and important. A straightforward approach to this problem is to partition attribute domains into intervals and to transform the quantitative values into binary ones, in order to apply the classical mining

* This work was partly supported by the National Natural Science Foundation of China (79925001/70231010), the MOE Funds for Doctoral Programs (20020003095), the Bilateral Scientific and Technological Cooperation Between China and Flanders (174B0201), and the Fund for Scientific Research Flanders.

M.Gh. Negoita et al. (Eds.): KES 2004, LNAI 3213, pp. 270–276, 2004.

algorithm [9]. To avoid abrupt transitions between intervals, *vagueness* has been widely introduced into the model of quantitative association rule mining because of its flexibility w.r.t. knowledge representation (see e.g. [3–7]). Indeed, a quantitative rule like "If the customers are between the ages of [30, 60], then they tend to buy electronics at a price of [$1000, $5000]", may lead to the so-called "boundary problem" [7]; e.g. a customer aged 29 with a purchase of $4000 is not accounted for in the model. On the other hand, "Middle-aged customers tend to buy expensive electronics" may be more flexible and would reflect this customer's buying behaviour. To deal with the sharp boundary problem, a number of fuzzy sets can be defined on the domain of each quantitative attribute, and the original dataset is transformed into an extended one with attribute values in the interval $[0, 1]$.

On another count, classical algorithms merely concern *positive* association rules, that is, only those itemsets appearing frequently together will be discovered. However, a negative rule such as $\{\neg$ high income $\} \Rightarrow \{\neg$ expensive electronics$\}$ is also useful because it expresses that people who are not rich generally do not buy expensive electronics. Although this kind of knowledge has been noted by several authors [2, 5, 10], we believe that the research on negative association rules has not received sufficient attention for the following reason: association rule mining first emerged in the domain of supermarkets, whose databases always contain thousands of goods (attributes) but each customer only buys few of them. In other words, most of the attribute values in a transaction are 0. If negative associations are also considered, a great deal of frequent negative patterns are generated, making algorithms unscalable and positive rules less noticed. In quantitative databases this problem is much less significant, because the fraction of attribute values equal to 0 is usually much smaller.

In this paper, in Section 2 we introduce positive and negative quantitative association rules in the classical (crisp) case. We show that, for the computation of the traditional rule quality measures of support and confidence, as well as the more logic–inspired degree of implication, the use of negative association rules does lead to additional database scans. Section 3 investigates the extension to a fuzzy framework, while Section 4 discusses important issues to be considered in a realistic application.

2 Positive and Negative Association Rules

Let $D = \{t_1, t_2, \ldots, t_n\}$ be a relational database of n tuples (or transactions) with a set of binary attributes (or items) $I = \{I_1, I_2, \ldots, I_m\}$; each transaction t in D can be considered as a subset of I, $t[I_j] = 1$ if $I_j \in t$, and $t[I_j] = 0$ if $I_j \notin t$ ($j = 1, 2, \ldots, m$). An association rule is of the form: $X \Rightarrow Y$, where X and Y are two disjoint non–empty subsets of I (itemsets). Support and confidence for rule $X \Rightarrow Y$ are defined as $supp(X \Rightarrow Y) = \frac{|D_{X \cup Y}|}{|D|}$ and $conf(X \Rightarrow Y) = \frac{|D_{X \cup Y}|}{|D_X|}$ respectively, where $|D|$ is the number of tuples in D, $|D_X|$ is the number of tuples in D that contain X and (hence) $|D_{X \cup Y}|$ is the number of tuples in