Fuzzy Versus Quantitative Association Rules: A Fair Data-Driven Comparison

Hannes Verlinde, Martine De Cock, and Raymond Boute

***Abstract*—As opposed to quantitative association rule mining, fuzzy association rule mining is said to prevent the overestimation of boundary cases, as can be shown by small examples. Rule mining, however, becomes interesting in large databases, where the problem of boundary cases is less apparent and can be further suppressed by using sensible partitioning methods. A data-driven approach is used to investigate if there is a significant difference between quantitative and fuzzy association rules in large databases. The influence of the choice of a particular triangular norm in this respect is also examined.**

***Index Terms*—Data mining, fuzzy association rules, quantitative association rules, triangular norms.**

I. Introduction

The discovery of knowledge in databases, also called data mining, is a most promising and important research area. In data mining, association rules are often used to represent and identify dependencies between attributes in a database. The original idea dates back to the late 1970s [1], while its application for market basket analysis gained popularity at the beginning of the 1990s [2]. In the original context of association rule mining, data are represented by a table with binary values. The rows correspond to objects or transactions, while the columns correspond to attributes or items. The binary values denote whether a transaction contains a specific attribute. The purpose of association rule mining is to detect rules of the form $A \rightarrow B$, indicating that a transaction containing attribute A is likely to contain attribute B as well.

In most real-life applications, databases contain many other values besides 0 and 1. Very common, for instance, are quantitative attributes such as age or income, taking values from an ordered numerical scale, often a subset of the real numbers. One way of dealing with a quantitative attribute is to replace it by a few other attributes that form a crisp partition of the range of the original one. For instance, in a particular application, we might decide to replace age by the attributes young, middle-aged, and old corresponding to intervals [0,35[, [35,65[, and [65,100], respectively, while income can be replaced by low, medium, and high corresponding to the intervals [0,1000[, [1000,2000[, and [2000,10 000], respectively. The new attributes can be considered as binary ones (e.g., the value of young is 1 if the corresponding value of age belongs to [0,35[; otherwise it is 0), which reduces the problem to traditional association rule mining with binary values. The generated rules are now called quantitative association rules [3].

The starting point for fuzzy set theory [4] is that it is against intuition to model vague concepts such as young and high by crisp intervals. For why would a person be considered as young while he is younger than 35, and on his 35th birthday suddenly lose this status? In reality, the transition between being young and not being young is not abrupt but gradual. This is a very good argument for modeling vague concepts by fuzzy sets instead of crisp sets. Many researchers have already used this argument for the introduction of fuzzy association rules (see, e.g., [5]–[11]). In this process, a database containing quantitative attributes is replaced by one with values from [0,1] in a similar way as one does for quantitative association rules. It is said that, compared to quantitative association rules, fuzzy association rules correspond better to intuition and prevent overestimation of boundary cases (see, e.g., [7], [8], and [11]). The so-called sharp boundary problem states that binary algorithms either ignore or overemphasize the elements near the boundary of the intervals in the mining process. It is easy to construct toy examples to illustrate this point.

Association rule mining is, however, not developed to play around with small toy examples but to deal with large databases. Unfortunately, comparisons for large data sets of results obtained with a quantitative versus a fuzzy association rule mining algorithm are extremely hard to find in the literature. An additional problem is the partitioning of the range of attribute values in intervals (for quantitative association rule mining) and the construction of membership functions (for fuzzy association rule mining). The two extreme solutions to this problem are the expert-driven approach (an expert manually sets the interval boundaries and/or defines the membership functions) and the data-driven approach (they are generated automatically from the data table, see, e.g., [3] and [12]). In [11], results obtained with quantitative and fuzzy association rule mining are compared for two artificially created data sets. For the quantitative case, the attribute partitioning method proposed by Srikant and Agrawal [3] is applied. For the fuzzy case, however, the membership functions are constructed manually (but not given in the paper). This is an unfair footing for comparison because additional expert knowledge is injected into the fuzzy approach while the quantitative approach is fully automatical (data driven). It is clear that a fair comparison can only be done using either the same expert-driven approach or the same data-driven approach for both mining processes. In this paper, we experimentally investigate the latter.

II. Association Rule Mining

Recall that a fuzzy set A on a universe X is characterized by $X - [0, 1]$ mapping, also called the membership function of A. For x in X, $A(x)$ denotes the membership degree of x in the fuzzy set A. If the membership function only takes values in {0,1}, it coincides with the traditional set concept, which, in this context, is also called "crisp set."

Table I presents what could happen if we replace the quantitative attributes in a small database by either binary or fuzzy attributes.

Let X be the set of all transactions (in our example $|X| = 6$). We study rules of the form $A \rightarrow B$, where A and B are different attributes.[1] We use $A(x)$ to denote the value of attribute A for transaction x. In this way, A becomes either a crisp subset of X [$A(x) = 0$ corresponds to $x \notin A$ while $A(x) = 1$ means $x \in A$] or a fuzzy set in X [$A(x)$ denotes the membership degree of x in A]. The support and confidence of a (candidate) association rule $A \rightarrow B$ are defined as

$$\mathrm{supp}(A \rightarrow B) = \frac{|A \cap B|}{|X|}$$

$$\mathrm{conf}(A \rightarrow B) = \frac{|A \cap B|}{|A|}.$$

[1] For reasons explained in Section III-C, we only consider rules with one attribute in antecedent and consequent.

Manuscript received September 8, 2004; revised March 11, 2005. This work was supported by the bilateral China-Flanders cooperation project 174BO201. H. Verlinde is a Research Assistant of the Fund for Scientific Research–Flanders. M. De Cock is a Postdoctoral Fellow of the Fund for Scientific Research–Flanders. This paper was recommended by Associate Editor T. Sudkamp.

The authors are with Ghent University, B900 Ghent, Belgium.

Digital Object Identifier 10.1109/TSMCB.2005.860134