A Fuzzy Approach for Mining Interesting Quantitative Association Rules

Attila Gyenesei

During the last ten years, data mining, also known as knowledge discovery in databases, has established its position as a prominent and important research area. Mining association rules is one of the important research problems in data mining. The problem of mining boolean association rules over basket data was introduced in [Agrawal et al.,1993]. Given a set of transactions, where each transaction is a set of items, an association rule is an expression of the from $X \Rightarrow Y$, where X and Y are sets of items. An example of an association rule is: "40% of transactions that contain beer and potato chips also contain diapers; 5% of all transaction contain all of these items". Here 40% is called the confidence of the rule, and 5% the support of the rule.

The problem of mining quantitative association rules was introduced in [Srikant et al.,1996]. For illustration, a table with three non-key attributes is shown in Table 1. Age and NumCars are quantitative attributes, whereas Married is a categorical attributes.

RecordID	Age	Married	NumCars
5	23	No	1
8	25	Yes	1
14	27	No	0
18	38	Yes	2
19	39	Yes	2

(min. support = 40%, min. confidence = 50%)

Rules (Sample)	Support	Confidence
$\langle Age : 30..39 \rangle$ and $\langle Married : Yes \rangle \Rightarrow \langle NumCars : 2 \rangle$	40%	100%
$\langle NumCars : 0..1 \rangle \Rightarrow \langle Married : No \rangle$	40%	66.6%

Table 1: Example of Quantitative Association Rules

Several efficient algorithms for mining quantitative association rules have been published (see [Srikant et al.,1996] for an example). The algorithms find the association rules by partitioning the attribute domain, combining adjacent partitions, and then transforming the problem into a binary one. Although these quantitative association rule mining algorithms can solve some problems introduced by quantitative attributes, they introduce some other problems. The first problem is caused by the sharp boundary between intervals. For example, as shown in Figure 1, suppose $[11, 20]$, $[21, 30]$ and $[31, 40]$ are three intervals created on a quantitative attribute, and the minimum support is 40%.

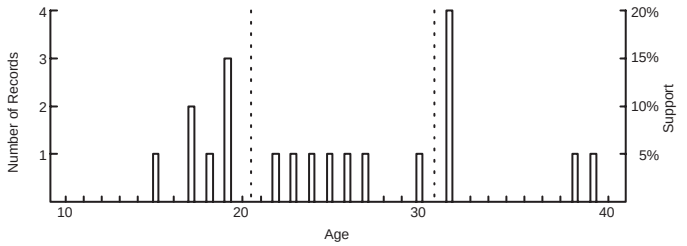

Figure 1: Example of Sharp Boundary Problem

In this case, none of these intervals will have enough support. However, the interval $[21, 30]$ should be interesting if we consider the values near both sides.

In this paper, we deal with mining fuzzy quantitative association rules of the following form:

$$\text{If } X = \{x_1, x_2, \ldots, x_p\} \text{ is } A = \{f_1, f_2, \ldots, f_p\} \text{ then } Y = \{y_1, y_2, \ldots, y_q\} \text{ is } B = \{g_1, g_2, \ldots, g_q\},$$

where

$$f_i \in \{\text{fuzzy sets related to attribute } x_i\},\ g_j \in \{\text{fuzzy sets related to attribute } y_j\}$$

and X, Y are itemsets, A and B contain the fuzzy sets associated with the corresponding attributes in X and Y. "X is A" is called the antecedent of the rule while "Y is B" is called the consequent of the rule. With fuzzy sets, a person may be both a member of "old" with 80% membership, and also a member of "middle-age" with 20% membership. An example of such an rule might be "If *Age* is *old* then *NumCars* is *small*".

In this paper we present a new algorithm for mining association rules based on fuzzy set theory. The algorithm uses new definitions for interesting measures. Finally, we describe the results of using approach on a real-life dataset.

References

[Agrawal et al.,1993] R. Agrawal, T. Imielinski and A. Swami (1993), *Mining association rules between sets of items in large databases*, ACM SIGMOD, Washington, DC, USA, pp. 207–216.

[Srikant et al.,1996] R. Srikant and R. Agrawal (1996), *Mining quantitative association rules in large relation tables*, Proceedings of ACM SIGMOD, pp. 1–12.