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} $\Rightarrow$ {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

D that contain both X and Y. Also, we define the support of itemset X as $supp(X) = \frac{|D_X|}{|D|}$; clearly $supp(X \Rightarrow Y) = supp(X \cup Y)$. A valid association rule is a rule with support and confidence greater than given thresholds. [1]

When a database also contains a quantitative attribute Q, it is possible to "binarize" Q by partitioning its range into p intervals and by replacing Q by new binary attributes $Q_1, \ldots Q_p$ such that $t[Q_k] = 1$ when the value of t for Q falls within the k^{th} interval, and 0 otherwise. We can then apply traditional mining algorithms to this transformed dataset [9]; these algorithms usually involve detecting all the *frequent itemsets*[1], and using them to construct valid association rules (e.g. Apriori algorithm [8]).

In [5, 6], authors distinguish between positive, negative and irrelevant examples of an association rule. A transaction t is called a *positive example* of $X \Rightarrow Y$, if $X \subseteq t$ and $Y \subseteq t$, a *negative example* if $X \subseteq t$ and $Y \not\subseteq t$ and an *irrelevant example* if $X \not\subseteq t$. It is clear that with this terminology, the support of $X \Rightarrow Y$ equals the relative fraction of database transactions that are positive examples to the rule. In [10], expressions of the form $X \Rightarrow Y$, $X \Rightarrow \neg Y$, $Y \Rightarrow \neg X$ and $\neg Y \Rightarrow \neg X$, where X and Y are itemsets, are introduced and called *negative association rules*. The understanding is that, e.g., each negative example of $X \Rightarrow Y$ is a positive example of $X \Rightarrow \neg Y$. However, this definition has an important drawback: a negative association rule {mobile} $\Rightarrow \neg$ {batteries, alarm clock} then means that customers who buy a mobile are unlikely to buy both batteries and alarm clocks. If a transaction t contains mobile and batteries, but no alarm clock, t is then a positive example to the rule because {mobile} $\subseteq t$ and {batteries,alarm clock} $\not\subseteq t$. More generally, if $Y \subseteq Y'$, then the support of $X \Rightarrow \neg Y'$ is not less than that of $X \Rightarrow \neg Y$, which (informally) means that for two rules with the same antecedent, the negative rule with longer consequent has larger support. This results in much more computations and uninteresting negative rules with long consequents.

In real life, a more desirable kind of knowledge may be {mobile} $\Rightarrow$ {$\neg$ alarm clock, batteries}, which means that customers buying mobiles are unlikely to buy alarm clocks *but* are likely to buy batteries. Therefore, we regard each item's complement as a new item in the database. That is, for the rule $X \Rightarrow Y$, X and Y are two disjoint itemsets of $I \cup I_c$, where $I = \{I_1, I_2, \ldots, I_m\}$ and $I_c = \{\neg I_1, \neg I_2, \ldots, \neg I_m\}$.

As rule quality measures, we complement[2] support and confidence with a so–called degree of implication (see e.g. [3, 5]). The latter measure interprets the arrow sign in $X \Rightarrow Y$ as an implication relationship, and is defined as

$$D_{imp}(X \Rightarrow Y) = \frac{|D_{X \to Y}|}{|D|} \tag{1}$$

[1] i.e., those meeting the support threshold.

[2] In [5] it was shown that under certain circumstances degree of implication may even replace confidence, but in principle the three measures can meaningfully co–exist. Degree of implication may be particularly relevant when considering incorporation of the mined association rules into a rule–based system (see e.g. [4]).

where $D_{X\to Y} = \{t \in D | X \not\subseteq t \text{ or } Y \subseteq t\}$. Clearly, this non–symmetrical measure computes the relative fraction of transactions that are not negative examples to the rule. A detailed investigation into this measure and its relationship to support and confidence was carried out in [5].

Because of the large size of the databases in real life applications, computations that require database scanning are by far the most time–consuming. It is therefore worthwhile to avoid them as much as possible. The following properties show that mining negative assocations, as well as using D_{imp}, do not require additional database scans.

Proposition 1. *No transaction simultaneously contains I_j and $\neg I_j$.*

During candidate frequent itemset generation, any itemset containing both an item and its complement can be pruned away immediately.

Proposition 2. $supp(X \Rightarrow \{I_k\}) + supp(X \Rightarrow \{\neg I_k\}) = supp(X)$.

Proposition 2 relates the support of a negative association rule to that of a corresponding positive rule. More generally, the following holds.

Proposition 3. *Let $X = \{J_1, \ldots, J_p, \neg J'_1, \ldots, \neg J'_q\}$ where $J_k, J'_l \in I$ and $X' = \{J_1, \ldots, J_p\}$. Then $supp(X)$ equals to*

$$\frac{|D_{X'}| - \sum_{i=1}^{q} |D_{X'\cup\{J'_i\}}| + \sum_{i=1}^{q}\sum_{j=i+1}^{q} |D_{X'\cup\{J'_i,J'_j\}}| + \ldots + (-1)^q |D_{X'\cup\{J'_1,\ldots,J'_q\}}|}{|D|}$$

Degree of implication can be derived from support, i.e. computing D_{imp} does not lead to additional database scans.

Proposition 4. *[3]* $D_{imp}(X \Rightarrow Y) = 1 - supp(X) + supp(X \cup Y)$

Finally, proposition 5 gives us a hint about how to choose meaningful threshold values in the definition of a valid association rule.

Proposition 5. $supp(X \Rightarrow Y) \leq conf(X \Rightarrow Y) \leq D_{imp}(X \Rightarrow Y)$

3 Positive and Negative Fuzzy Association Rules

In the framework of fuzzy association rules, transactions can be perceived as fuzzy sets in I, so $t[I_j] \in [0, 1]$; moreover, we assume $t[\neg I_j] = 1 - t[I_j]$. The idea is that a transaction can contain an item to a given extent. A standard approach to extend quality measures to fuzzy association rules is to replace set-theoretical operations by corresponding fuzzy set–theoretical operations. Specifically, we need extensions to the classical conjunction and implication. To this end, t-norms and implicators are used; some popular t-norms and implicators are listed in Table 1.

Table 1. Well-known t–norms and implicators (x, y in $[0, 1]$)

t-norm	implicator
$T_{\mathrm{M}}(x, y) = \min(x, y)$	$I_{\mathrm{M}}(x, y) = \max(1 - x, y)$
$T_{\mathrm{P}}(x, y) = xy$	$I_{\mathrm{P}}(x, y) = 1 - x + xy$
$T_{\mathrm{W}}(x, y) = \max(x + y - 1, 0)$	$I_{\mathrm{W}}(x, y) = \min(1 - x + y, 1)$

Support. Given a t-norm T, the degree to which a transaction t supports the itemset $X = \{x_1, \ldots, x_p\}$ is expressed by:

$$D_X(t) = T(t[x_1], t[x_2], \ldots, t[x_p]) \tag{2}$$

Support is defined, by means of the cardinality of a fuzzy set, as:

$$supp(X \Rightarrow Y) = \frac{\sum\limits_{t \in D} D_{X \cup Y}(t)}{|D|} = \frac{\sum\limits_{t \in D} T(D_X(t), D_Y(t))}{|D|} \tag{3}$$

Confidence.

$$conf(X \Rightarrow Y) = \frac{\sum\limits_{t \in D} D_{X \cup Y}(t)}{\sum\limits_{t \in D} D_X(t)} \tag{4}$$

Degree of Implication.

$$D_{imp}(X \Rightarrow Y) = \frac{\sum\limits_{t \in D} D_{X \to Y}(t)}{|D|} = \frac{\sum\limits_{t \in D} I(D_X(t), D_Y(t))}{|D|} \tag{5}$$

where I is an implicator. For a comparative study of the behaviour of various implicators w.r.t. fuzzy association rule mining we refer to [5].

Since ordinary sets are replaced by fuzzy sets, the properties mentioned in Section 2 need to be re–investigated. Proposition 1 does not generally remain valid because $T(x, 1 - x) = 0$ does not hold for every t-norm (it does hold for $T = T_W$), which means that item I_j and $\neg I_j$ can appear in an itemset simultaneously. To avoid meaningless rules like $\{I_j\} \Rightarrow \{\neg I_j\}$, we should explicitly include this restriction in the definition of a valid fuzzy association rule $X \Rightarrow Y$. For Proposition 2 to hold, $\sum\limits_{t \in D} D_{X \cup \{I_k\}}(t) + \sum\limits_{t \in D} D_{X \cup \{\neg I_k\}}(t) = \sum\limits_{t \in D} D_X(t)$ should hold. As was discussed in [6], for $T = T_P$ the proposition is valid. It can be verified that Proposition 3 is also valid for $T = T_P$, hence the optimization strategy to reduce the number of candidate itemsets can still be used. As discussed in [3], Proposition 4 is maintained for some t-norm/implicator combinations, in particular for (T_M, I_W), (T_P, I_P) and (T_W, I_M). Finally, Proposition 5 is valid as soon as Proposition 4 is valid.

4 Implementation and Discussion

To implement the fuzzy association rule mining procedure, we used a modified version of the Apriori algorithm. To guarantee that all simplifying properties from the previous section are valid, we chose $T = T_P$ and $I = I_P$. Note that these properties assure that the additional complexity caused by considering negative items and degree of implication, can be kept within very reasonable bounds, and the algorithm is definitely much more economical than straightforwardly applying Apriori, treating negative items as new database attributes. It is also very much preferable to the approach for mining negative association rules from [10] which involves the costly generation of infrequent as well as frequent itemsets.

Regarding the quality of the mined association rules, we observed that most of them are negative. This can be explained as follows: when for each transaction t and each collection $J_1, \ldots, J_p$ of $[0,1]$–valued positive attributes corresponding to a quantitative attribute Q, it holds that[3] $\sum_{i=1}^{p} t[J_i] = 1$, then at the same time $\sum_{i=1}^{p} t[\neg J_i] = p - 1$. In other words, the overall support associated with positive items will be 1, while that associated with negative items will be $p - 1$, which accounts for the dominance of the latter. Since typically p is between 3 and 5, the problem however manifests itself on a much smaller scale than in supermarket databases. To tackle it, we can e.g. use different thresholds for positive rules, and for rules that contain at least one negative item. However, this second threshold apparently should differ for every quantitative attribute since it depends on the number of fuzzy sets used in the partition. A more robust, and only slightly more time-consuming, approach is to impose additional filtering conditions and interestingness measures to prune away the least valuable negative patterns.

5 Conclusion

We introduced fuzzy negative association rules, and showed that their incorporation into mining algorithms does not cause additional database scans, making implementations efficient. Future work will focus on selecting adequate quality measures to dismiss uninteresting negative rules.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In: Proc. ACM–SIGMOD Int. Conf. on Management of Data (1993) 207–216

[3] Note that this is a very natural assumption, since it means that each transaction makes the same overall contribution to the support measure. It is automatically fulfilled for classical quantitative association rules.

2. Brin, S., Motwani, R., Silverstein, C.: Beyond Market Baskets: Generalizing Association Rules to Correlations. In: Proc. ACM SIGMOD on Management of Data (1997) 265–276
3. Chen, G.Q., Yan, P., Kerre, E.E.: Computationally Efficient Mining for Fuzzy Implication-Based Association Rules in Quantitative Databases. In: International Journal of General Systems (to appear)
4. Cornelis, C.: Two–sidedness in the Representation and Processing of Imprecise Information (in Dutch), Ph.D. thesis
5. De Cock, M., Cornelis, C., Kerre, E.E.: Elicitation of Fuzzy Association Rules from Positive and Negative Examples. Submitted.
6. Dubois, D, Hüllermeier, E., Prade, H.: A note on Quality Measures for Fuzzy Association Rules. In: LNAI, Vol. 2715 (2003) 346–353
7. Gyenesei, A.: A Fuzzy Approach for Mining Quantitative Association Rules. TUCS technical report 336, University of Turku, Finland (2000)
8. Srikant, R., Agrawal, R.: Fast Algorithms for Mining Association Rules. In: Proc. VLDB Conference (1994) 487–499
9. Srikant, R., Agrawal, R.: Mining Quantitative Association Rules in Large Relational Tables. In: Proc. ACM–SIGMOD Int. Conf. on Management of Data (1996) 1–12
10. Wu, X., Zhang, C., Zhang, S.: Mining Both Positive and Negative Association Rules. In: Proc. 19th Int. Conf. on Machine Learning (2002) 658–665