Acta Cybernetica **15** (2001) 305–320.

A Fuzzy Approach for Mining Quantitative Association Rules

Attila Gyenesei *

Abstract

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. Many algorithms have been proposed to find association rules in databases with quantitative attributes. The algorithms usually discretize the attribute domains into sharp intervals, and then apply simpler algorithms developed for boolean attributes. An example of a quantitative association rule might be "10% of married people between age 50 and 70 have at least 2 cars". Recently, fuzzy sets were suggested to represent intervals with non-sharp boundaries. Using the fuzzy concept, the above example could be rephrased e.g. "10% of married old people have several cars". However, if the fuzzy sets are not well chosen, anomalies may occur. In this paper we tackle this problem by introducing an additional fuzzy normalization process. Then we present the definition of quantitative association rules based on fuzzy set theory and propose a new algorithm for mining fuzzy association rules. The algorithm uses generalized definitions for interest measures. Experimental results show the efficiency of the algorithm for large databases.

1 Introduction

The goal of data mining is to extract higher-level information from an abundance of raw data. Mining association rules is one of the important research problems in data mining [11]. The problem of mining boolean association rules over basket data was introduced in [1]. Given a set of transactions where each transaction is a set of items, an association rule is an expression of the form $X \Rightarrow Y$, 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 transactions contain all of these items". Here 40% is called the confidence of the rule, and 5% the support of the rule. The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints. There are many known algorithms for mining boolean association rules (see [2], [4], [5], [10] and [13] for just a few examples).

*Turku Centre for Computer Science (TUCS), University of Turku, Department of Computer Science, Lemminkäisenkatu 14, FIN-20520 Turku, Finland, e-mail: gyenesei@cs.utu.fi

In practice the information in many, if not most, databases is not limited to categorical attributes (e.g. zip code, make of car), but also contains much quantitative data (e.g. age, income). The problem of mining quantitative association rules was introduced and an algorithm proposed in [12]. The algorithm involves discretizing the domains of quantitative attributes into intervals in order to reduce each domain into a categorical one. An example of such an association might be "10% of married people between 50 and 70 have at least 2 cars".

Without a priori knowledge, however, determining the right intervals can be a tricky and difficult task due to the "catch-22" situation, as called in [12], because of the effects of *small support* and *small confidence*. Moreover, these intervals may not be concise and meaningful enough for human experts to easily obtain nontrivial knowledge from those rules discovered.

Instead of using sharp intervals, fuzzy sets were suggested in [9] to represent intervals with non-sharp boundaries. The obtained rules are called fuzzy association rules. If meaningful linguistic terms are assigned to fuzzy sets, the fuzzy association rule is more understandable. The above example could be rephrased e.g. "10% of married old people have several cars". An algorithm for mining fuzzy association rules was proposed in [8], but the problem is that an expert must provide the required fuzzy sets of the quantitative attributes and their corresponding membership functions. It is unrealistic to assume that experts can always provide the best fuzzy sets for fuzzy association rule mining. Moreover, if the fuzzy sets are not well chosen, anomalies may occur. In this paper we will tackle this problem by introducing an additional fuzzy normalization process.

The rest of this paper is organized as follows. In the next section, we present a brief description of how existing algorithms can be used for the mining of quantitative association rules and how fuzzy techniques can be applied to the data mining process. Then we will introduce a fuzzy normalization process in Section 3. In the same section, we give the definitions of fuzzy association rules and interest measures. In Section 4 we propose a new algorithm for fuzzy quantitative association rules. In Section 5 the experimental results are reported, followed by a brief conclusion in Section 6.

2 Problem Description

Several efficient algorithms for mining boolean association rules have been presented. Boolean attributes can be considered a special case of categorical attributes [4] and it is relatively straightforward to generalize the boolean algorithms for categorical attributes. For quantitative attributes, however, the situation is not so simple. We either have to somehow transform the quantitative association rules problem into boolean one or to find new algorithms. Here we shall, in fact, apply both alternatives.

2.1 Mapping Quantitative Attributes to Boolean Ones

If the quantitative association rules problem can be mapped to the boolean association rules problem, any algorithm for finding boolean association rules can be used to find quantitative association rules. This mapping can be performed as follows [12]. Suppose that we have a database shown in Table 1.

RID	Age	Income	Status	RID	Age	Income	Status
1	19	1400	Unmarried	11	26	2000	Married
2	22	1600	Unmarried	12	31	2400	Married
3	31	2400	Unmarried	13	19	1400	Unmarried
4	18	1400	Married	14	27	2200	Married
5	23	1600	Married	15	31	2600	Married
6	30	2800	Married	16	15	1000	Unmarried
7	17	1200	Unmarried	17	24	1800	Unmarried
8	25	2000	Married	18	38	2600	Married
9	31	2200	Married	19	17	1200	Unmarried
10	19	1400	Unmarried	20	39	2400	Married

Table 1: An example database

Let the relational table contain a boolean field for each attribute value/interval for each quantitative attribute. Then the value of any such boolean field, which corresponds to $\langle attribute, v \rangle$, would be "1" if the *attribute* had v in the original record, and "0" otherwise. Table 2 shows this mapping for the example database given in Table 1. Age is partitioned into three intervals: 11..20, 21..30 and 31..40. For income, two intervals have been defined. The categorical attribute, Status, is represented by two boolean attributes: "Unmarried" and "Married".

RID	Age (11..20)	Age (21..30)	Age (31..40)	Income (1000..1800)	Income (2000..2800)	Status Unmarried	Status Married
1	1	0	0	1	0	1	0
2	0	1	0	1	0	1	0
3	0	0	1	0	1	1	0
4	1	0	0	1	0	0	1
5	0	1	0	1	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
19	1	0	0	1	0	1	0
20	0	0	1	0	1	0	1

Table 2: Mapping to boolean association rules problem

For example, Record 5, which had $\langle Age : 23 \rangle$ now has "Age : 11..20" equal to "0", "Age : 21..30" equal to "1", and Age : 31..40" equal to "0", etc.

2.2 Mapping Problems

Unfortunately, the mapping approach leads to two problems [12]:

- *Small support*: if an interval is too small, a rule containing this interval may not have the minimum support; either very few rules are generated or rules are nearly as specific as the data itself.
- *Small confidence*: if an interval is too large, a rule containing this interval in the antecedent may not have the minimum confidence; many rules with little information are generated.

An example of the problem of small support (also called "sharp boundary problem" in [9]) is shown in Figure 1, suppose $[11, 20]$, $[21, 30]$ and $[31, 40]$ are three intervals created on the quantitative attribute Age, with 35%, 35% and 30% supports. If the minimum support threshold is a bit greater than 35%, then none of the intervals has sufficient support. However, there are high frequencies at 19 and 31, so a small extension of the interval $[21, 30]$ would make it frequent.

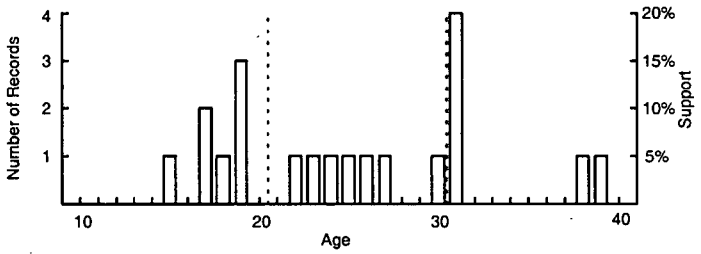

Figure 1: Example of small support problem

Of course, there is no restriction that the intervals should be disjoint. By letting them overlap, the sharp boundary problem can be overcome [12], see Figure 2.

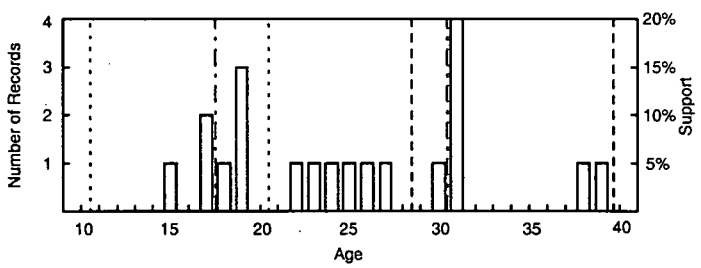

Figure 2: Overlapping adjacent intervals

Combining/overlapping adjacent intervals avoids the small support problem, and the number of intervals may be increased to avoid the small confidence problem. Unfortunately, this approach introduces a new problem [12]:

- *Many rules*: Consider an interval satisfying the minimum support. Then any range containing this interval will also satisfy the minimum support. Thus, the number of rules increases, and not all of them are interesting.

2.3 Fuzzy Approach

Instead of using sharp intervals, fuzzy sets were suggested in ([3], [9]) to represent intervals with non-sharp boundaries, as shown in Figure 3. Using fuzzy sets, an element can belong to a set with set membership value in [0, 1]. The lower histogram in Figure 3 shows membership values chosen for the middle fuzzy set.

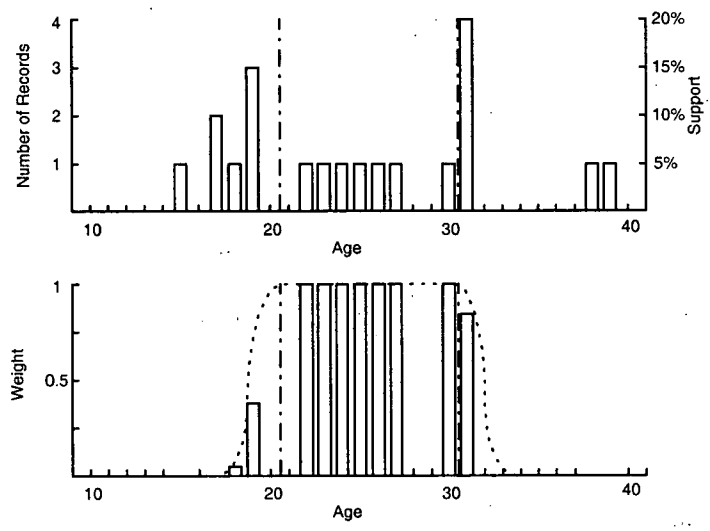

Figure 3: Fuzzy set

However, if the fuzzy sets are not well chosen, some anomalies occur. In Figure 1, the three intervals will be replaced by three fuzzy sets. Suppose the value 30 has membership degree of 0.9 in the second set and 0.3 in the third set. Then it will contribute 0.9 to the support of the second fuzzy set and 0.3 to the third one. However, this means that the value 30 will be more important than other values since the sum of its contributions to different fuzzy sets has become greater than 1. In the following section we will tackle this problem by introducing an additional fuzzy normalization process.

3 Inclusion of Fuzzyness in Associaton Rules

Our starting point is that the fuzzy sets and their membership functions are given. In [6] we gave a clustering algorithm for their automatic generation, but here we do not make any assumptions of the source of fuzzy sets. In this section we will first

introduce a fuzzy normalization process, to derive unbiased membership functions for the given fuzzy sets. Then we will give the generalized definition of a fuzzy association rule and related interest measures.

3.1 Fuzzy Normalization Process

Let $I = \{i_1, i_2, \ldots, i_n\}$ be the complete set of items where each i_j $(1 \leq j \leq n)$ denotes a categorical or quantitative (fuzzy) attribute. Further denote by $F(i_j) = \{\langle i_j, l\rangle \mid l = 1, \ldots, N(i_j)\}$ the set of fuzzy sets (or non-fuzzy categories), related to item i_j, where $N(i_j)$ represents the number of fuzzy sets (or number of categories). The membership function of $\langle i_j, l\rangle$ is denoted by $m_{\langle i_j,l\rangle}(v)$. If i_j is categorical, $m_{\langle i_j,l\rangle}(v) = 0$ or $m_{\langle i_j,l\rangle}(v) = 1$. If i_j is fuzzy, $0 \leq m_{\langle i_j,l\rangle}(v) \leq 1$. Thus, categories are special fuzzy sets, and can be handled similarly.

Let $t = \{t.i_1, t.i_2, \ldots, t.i_n\}$ be a transaction, where $t.i_j$, $(1 \leq j \leq n)$ represents the value of the j^{th} item. Value $t.i_j$ can be mapped to

$$\{(l, m_{\langle i_j,l\rangle}(t.i_j)) \mid \text{ for all } l, 1 \leq l \leq N(i_j)\}.$$

We define that $F(i_j)$ is a *'fuzzy partition'* if $\sum_{l=1}^{N(i_j)} m_{\langle i_j,l\rangle}(v) = 1$ for each v in domain i_j where i_j is fuzzy. This is a natural generalization to the non-fuzzy partitioning of a set into disjoint intervals covering the whole range. In practice, the sum may not always be equal to 1. We therefore define a normalization process as follows:

$$m'_{\langle i_j,l\rangle}(t.i_j) = \frac{m_{\langle i_j,l\rangle}(t.i_j)}{\sum_{l=1}^{N(i_j)} m_{\langle i_j,l\rangle}(t.i_j)}$$

Example. Suppose $I = \{status, age\}$ where *status* is a categorical attribute with the domain of {married, unmarried} and *age* is a quantitative attribute with three fuzzy sets {young, middle, old}. Note that it is possible to define other fuzzy set groups for this attribute. t = {unmarried, 25} will be mapped to {{(married, 0),(unmarried, 1)}, {(young, 0.2), (middle, 0.9), (old, 0.1)}}.

$(Status, married)$	$(Status, unmarried)$	$(Age, young)$	$(Age, middle)$	(Age, old)
0	1	0.2	0.9	0.1

Table 3: Without fuzzy normalization

Without normalization (Table 3), transaction t would increase the support of itemset $\{Status = unmarried, Age = young\}$ by 0.2, the support of itemset $\{Status = unmarried, Age = middle\}$ by 0.9, and the support of itemset $\{Status = unmarried, Age = old\}$ by 0.1. That is to say, this transaction will be counted $0.2 + 0.9 + 0.1 = 1.2$ times for the item *Age*. However, it is unreasonable for one transaction to contribute more than others, if the corresponding discrete sets are disjoint.

$(Status, married)$	$(Status, unmarried)$	$(Age, young)$	$(Age, middle)$	(Age, old)
0	1	0.167	0.75	0.083

Table 4: After fuzzy normalization

In contrast (Table 4), the normalization process will further transform the transaction t into {{(married, 0),(unmarried, 1)}, {(young, 0.167), (middle, 0.75), (old, 0.083)}}, for a total contribution of 1.0 for the item *Age*.

It should be noticed that normalization is not always order-preserving, with respect to the values of membership functions. It might even produce functions which are not concave. For example, suppose that we have a quatitative attribute *age* and three transactions $t_1 = \{25\}$, $t_2 = \{26\}$ and $t_3 = \{27\}$. Table 5 shows the original and normalized mappings of the transactions into membership values.

Transaction	*Age*	Original memberships $\{young, middle, old\}$	Normalized memberships $\{young, middle, old\}$
t_1	25	{0.20, 0.90, 0.10}	{0.167, 0.750, 0.083}
t_2	26	{0.20, 0.91, 0.11}	{0.164, 0.746, 0.090}
t_3	27	{0.18, 0.92, 0.12}	{0.148, 0.754, 0.098}

Table 5: Example of normalization anomaly

Notice the anomaly for the 'middle' fuzzy set: normalization changes the order of membership values. A sufficient (but not necessary) condition for concavity is that $\sum_{l=1}^{N(i_j)} m_{\langle i_j,l\rangle}(v)$ is constant for all v in domain i_j. In [6] we tackled this problem and showed how to create a fuzzy partition directly, without normalization.

3.2 Fuzzy Association Rule

After having obtained the fuzzy partitions and their corresponding membership functions for each fuzzy set of every quantitative attribute, a new transformed (fuzzy) database D^T is generated from the original database. Given a database $D^T = \{t_1, t_2, \ldots, t_n\}$ with attributes I and the fuzzy sets $F(i_j)$ associated with attributes i_j in I, we use the following form for a fuzzy association rule [9]:

$$\textbf{If } X = \{x_1, x_2, \ldots, x_p\} \textbf{ is } A = \{a_1, a_2, \ldots, a_p\}$$
$$\textbf{then } Y = \{y_1, y_2, \ldots, y_q\} \textbf{ is } B = \{b_1, b_2, \ldots, b_q\},$$

where $a_i \in F(x_i)$, $i = 1, \ldots, p$, and $b_j \in F(y_j)$, $j = 1, \ldots, q$. X and Y are ordered subsets of I and they are disjoint i.e. they share no common attributes. A and B contain the fuzzy sets associated with the corresponding attributes in

X and Y. As in the binary association rule, "X is A" is called the *antecedent* of the rule while "Y is B" is called the *consequent* of the rule. We also denote $Z = X \cup Y = \{z_1, \ldots, z_{p+q}\}$ and $C = A \cup B = \{c_1, \ldots, c_{p+q}\}$.

3.3 Fuzzy Itemset Measures - Support and Confidence

Let $D^T = \{t_1, t_2, \ldots, t_n\}$ be a database, where n denotes the total number of records ('transactions'). Let (Z, C) be an attribute-fuzzy set pair, where Z is an ordered set of attributes z_j and C is a corresponding set of fuzzy sets c_j. (From now on, we prefer to use the word "itemset" instead of "attribute-fuzzy set pair" for (Z, C) elements). If a fuzzy association rule $(X, A) \rightarrow (Y, B)$ is interesting, it should have enough fuzzy support $FS_{(Z,C)}$ and a high fuzzy confidence value $FC_{((X,A),(Y,B))}$, where $Z = X \cup Y$, $C = A \cup B$.

The fuzzy support value is calculated by multiplying the membership grade of each (z_j, c_j), summing them, then dividing the sum by the number of records [9]. We prefer the product operator as the fuzzy AND, instead of the normal minimum, because it better distinguishes high- and low-support transactions.

$$FS_{(Z,C)} = \frac{\sum_{i=1}^{n} \prod_{j=1}^{m} (t^i[(z_j, c_j)])}{n},$$

where m is the number of items in itemset (Z, C).

The fuzzy confidence value is calculated as follows:

$$FC_{((X,A),(Y,B))} = \frac{FS_{(Z,C)}}{FS_{(X,A)}}.$$

Both of the above formulas are direct generalizations of the corresponding formulas for the non-fuzzy case [1].

...	$(Age, middle)$	...	$(Income, low)$	...
	0.7		0.5	
	0.2		0.3	
...	0.5	...	0.2	...
	0.3		0.4	
	0.6		0.2	
	0.8		0.4	

Table 6: Part of a database containing fuzzy membership values

The following example illustrates the calculation of the fuzzy support and fuzzy confidence values. Let $Z = \{Age, Income\}$, $C = \{middle, low\}$ and a part of database shown in Table 6. The fuzzy support and confidence of (Z, C) are given by:

$$FS_{(Z,C)} = \frac{0.35+0.06+0.1+0.12+0.12+0.32}{6} = 0.178$$
$$FC_{((X,A),(Y,B))} = \frac{0.35+0.06+0.1+0.12+0.12+0.32}{0.7+0.2+0.5+0.3+0.6+0.8} = 0.345$$

3.4 Fuzzy Covariance and Correlation Values

Covariance is one of the simplest measures of dependence, based on the co-occurrence of the antecedent (X, A) and consequent (Y, B). If they co-occur clearly more often than what can be expected in an independent case, then the rule $(X, A) \rightarrow (Y, B)$ is potentially interesting. Piatetsky-Shapiro called this measure a *rule-interest* function [11]. We extend it to the fuzzy case, and define the covariance measure as:

$$FCov_{((X,A),(Y,B))} = FS_{(Z,C)} - FS_{(X,A)} \cdot FS_{(Y,B)}.$$

Covariance has generally the drawback that it does not take distributions into consideration. Therefore, in statistics, it is more common to use so called correlation measure, where this drawback has been eliminated. Again, we have to generalize the non-fuzzy formula to the fuzzy case, and obtain:

$$FCorr_{((X,A),(Y,B))} = \frac{FCov_{((X,A),(Y,B))}}{\sqrt{Var_{(X,A)} \cdot Var_{(Y,B)}}},$$

where

$$Var_{(X,A)} = FS_{(X,A)^2} - \left(FS_{(X,A)}\right)^2,$$
$$FS_{(X,A)^2} = \frac{\sum_{i=1}^{n} \left(\prod_{j=1}^{m} t^i\left[(x_j, a_j)\right]\right)^2}{n},$$

similarly for (Y, B).

These definitions are extensions of the basic formulas of variance and covariance. The value of the fuzzy correlation ranges from -1 to 1. Only a positive value tells that the antecedent and consequent are related. The higher the value is, the more related they are.

We use the information in Table 6 to illustrate the calculation of the fuzzy correlation value of a rule. Given the rule, "If *Age* is *middle* then *Salary* is *low*", the fuzzy covariance and correlation values of the rule are as follows:

$$FCov_{((X,A),(Y,B))} = 0.178 - 0.516 \cdot 0.333 = 0.006$$
$$FCorr_{((X,A),(Y,B))} = \frac{0.006}{\sqrt{0.045 \cdot 0.012}} = 0.258.$$

We defined the fuzzy extension of correlation measure, because it is an alternative to confidence, when measuring the dependence between the antecedent and consequent of a rule. We defined some other alternative measures of interestingness in [7]. In Section 5, we shall show results for both confidence and correlation.

4 Algorithm for Mining Fuzzy Quantitative Association Rules

An efficient algorithm for mining quantitative association rules has been proposed in [12]. However, a new algorithm is needed to solve the mining of fuzzy quantitative association rules. The problem of discovering all fuzzy quantitative association rules can be decomposed into two subproblems:

1. Find all itemsets that have fuzzy support $\left(FS_{(X,A)}\right)$ above the user specified minimum support (see Section 3.3). These itemsets are called *frequent itemsets.*

2. Use the frequent itemsets to generate the desired rules. The general idea is that if, say, X, Y, and $X \cup Y$ are frequent itemsets, then we can determine if the rule $X \Rightarrow Y$ holds by computing $FC_{((X,A),(Y,B))}$ (see Section 3.3). If this value is larger than the user specified minimum confidence value, then the rule will be interesting. We can also use the fuzzy correlation value $\left(FCorr_{((X,A),(Y,B))}\right)$ for this problem (see Section 3.4).

An algorithm for mining quantitative association rules has the following inputs an outputs.
Inputs: A database D, three threshold values *minsup*, *minconf* and *mincorr*.
Output: A list of interesting rules.
Notations:

D	the database
D^T	the transformed database
F_k	set of *frequent k-itemsets* (have k items)
C_k	set of *candidate k-itemsets* (have k items)
I	complete item set
minsup	support threshold
minconf	confidence threshold
mincorr	correlation threshold

Algorithm:
Main Algorithm $(minsup, minconf, mincorr, D)$
1 $F = \emptyset$;
2 $I = \text{Search}(D)$;
3 $(C_1, D^T) = \text{Transform}(D, I)$;
4 $k = 1$;
5 $F_k = \text{Checking}(C_k, D^T, minsup)$;
6 while $(|C_k| \neq 0)$ do
7 begin
8 $k = k + 1$;
9 if $k == 2$ then
10 $C_k = \text{Join1}(F_{k-1})$
11 else $C_k = \text{Join2}(F_{k-1})$;
12 $C_k = \text{Prune}(C_k)$;
13 $F_k = \text{Checking}(C_k, D^T, minsup)$;
14 $F = F \cup F_k$;
15 end
16 $Rules(F, minconf, mincorr)$;

The subroutines are outlined as follows:

1. **Search**(D): The subroutine accepts the database, finds out and returns the complete item set $I = \{i_1, i_2, \ldots, i_n\}$. For example, $I = \{Age, Income, Status\}$ for the database given in Table 1.

2. **Transform**(D, I): This step generates a new transformed (fuzzy) database D^T from the original database by user specified fuzzy sets. At the same time, the *candidate 1-itemsets* C_1 will be generated from the transformed database. (C_i is a set of sets of $(item, fuzzy\ set)$ pairs.) For example, $C_1 = \{\{(Age, young)\}, \{(Age, middle)\}, \{(Age, old)\}, \{(Income, low)\}, \{(Income, medium)\}, \{(Income, high)\}, \{(Status, unmarried)\}, \{(Status, married)\}\}$ is the complete set of *candidate 1-itemsets.*

3. **Checking**$(C_k, D^T, minsup)$: In this subroutine, the transformed (fuzzy) database is scanned and the fuzzy support $(FS_{(X,A)})$ of each candidate in C_k is calculated. A *k-itemset* in C_k is deleted if its fuzzy support is less than minsup. The remaining candidate itemsets will be kept in C_k. At the same time, the frequent itemsets F_k will be generated from C_k.

4. **Join1**(F_{k-1}): This Join step generates C_2 from F_1 as follows:
insert into C_2
select $\{(X, A), (Y, B)\}$
from $(X, A), (Y, B)$ in F_1
where $X \neq Y$

For example, after this Join step C_2 will be $C_2 = \{\{(Age, young),\ (Income, high)\}, \{(Age, middle), (Income, low)\}, \cdots\}$, but $C_2 \neq \{\cdots, \{(Age, young), (Age, middle)\}, \cdots\}$.

5. **Join2**(F_{k-1}): This Join step generates C_k from F_{k-1} as in [1]. For example, if we have $\{\{(Age, young), (Income, low)\}, \{(Age, young), (Balance, low)\}\}$ in F_{k-1}, $\{\{(Age, young),\ (Income, low), (Balance, low)\}\}$ will be generated in C_k.

6. **Prune**(C_k): During the prune step, an itemset S in C_k will be pruned if a subset of S does not exist in C_{k-1}.

7. **Rules**(F): Find the rules from the *frequent itemsets* F.

   For example, if $(Age, young)$ and $(Income, low)$ are frequent itemsets, then we get the $(Age, young) \Rightarrow (Income, low)$ rule, if its fuzzy confidence value (and fuzzy correlation value) is larger than the user specified minimum value.

5 Experimental Results

In this section, we will examine the accuracy and efficiency of our approach by experimenting with a real-life dataset. We applied our approach to a database called FAM95. This database contains data for the 63756 families that were interviewed in the March 1995 Current Population Survey (CPS), conducted by the Bureau of the Census for the Bureau of Labor Statistics. The data had 23 attributes: 7 quantitative and 16 categorical.

5.1 Interest Measures

In this experiment, we use six quantitative attributes to illustrate how the fuzzy concept gives more interesting rules than the discrete. The quantitative attributes were age of head in years ("head" is the reference person in a family), number of persons, children in family, education level of head, head's personal income and family income. Each quantitative attribute has three intervals/fuzzy sets. We choose the intervals by applying the well-known quantile-based partitioning, so that each interval gets the same number of attribute values.

Figure 4 shows the number of frequent itemsets for different minimum support. As expected, the number of frequent itemsets decreases as the minimum support increases from 10% to 45%. Fuzzy1 denotes the fuzzy method without normalization and Fuzzy2 denotes the fuzzy method with normalization. We can see that the fuzzy method with normalization gives fewer frequent itemsets than the fuzzy method without normalization. This method and the discrete interval method give similar numbers of frequent itemsets.

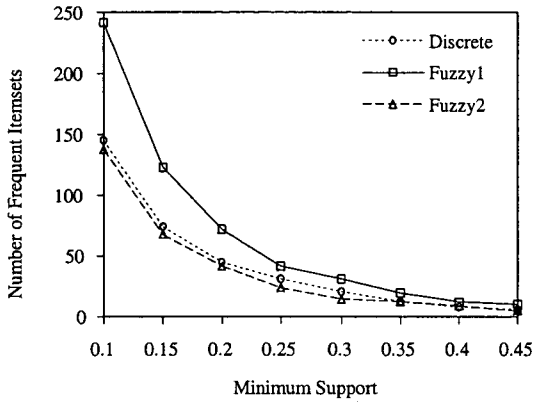

Figure 4: Number of frequent itemsets

Figures 5 and 6 show the number of interesting rules for different minimum confidence and correlation values. In both cases the minimum support was set to 20%. The results are quite similar to those of Figure 4.

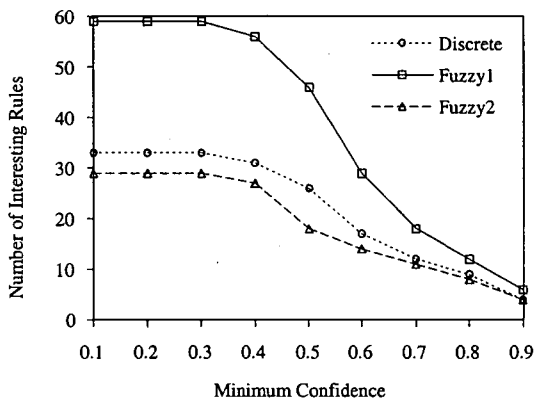

Figure 5: Effect of minimum confidence

We can see that the fuzzy method without normalization using confidence to calculate interest measure gives the highest number of expected interesting rules. However, in the correlation case the fuzzy method with normalization gives more rules than the others if the minimum correlation value was 0.6.

In the following we show some interesting rules. The minimum support was set

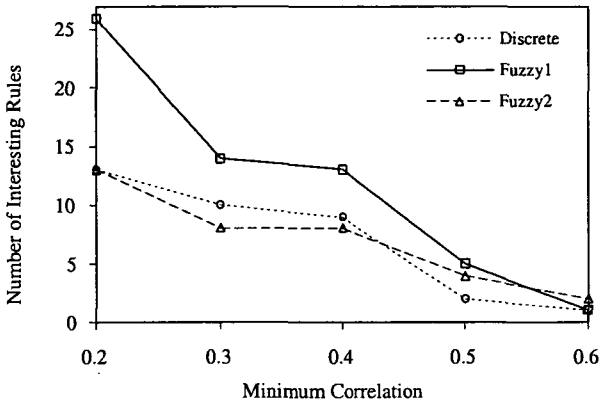

Figure 6: Effect of minimum correlation

to 20%, minimum confidence to 50%, and minimum correlation to 0.5.

IF *IncHead* is *low* THEN *IncFam* is *low*
IF *IncHead* is *medium* THEN *IncFam* is *medium*
IF *FamPers* is *low* AND *IncHead* is *low* THEN *IncFam* is *low*
IF *NumKids* is *low* AND *IncHead* is *low* THEN *IncFam* is *low*
IF *FamPers* is *low* AND *NumKids* is *low* AND *IncHead* is *low* THEN *IncFam* is *low*

5.2 Scale-Up Experiment

In this experiment, we will give the results on the performance of the algorithm using the confidence and correlation interest measures. The running time for the algorithm can be split into two parts:

- *Candidate generation.* The time for this is independent of the number of records.
- *Counting support, confidence and correlation.* The time for this is directly proportional to the number of records. When the number of records is large, this time will dominate the total time.

Thus we would expect the algorithm to have near-linear scaleup. This is confirmed by Figure 7, which shows the execution time as we increase the number of input records from 10000 to 64000. Note that we use five quantitative attributes in the database and each attribute has three fuzzy sets. We have set the user specified parameters such that both methods will give the same number of rules. The graph shows that the methods scale quite linearly for this dataset.

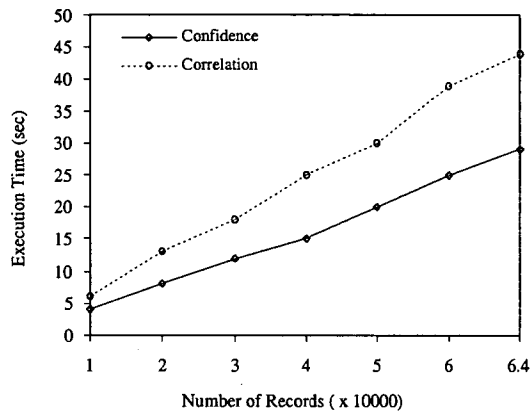

Figure 7: Scale-up: number of records

6 Conclusion

In this paper, we showed a new algorithm for mining fuzzy association rules, which were introduced in earlier papers. We assign each attribute with several fuzzy sets which characterize the quantitative attribute. Fuzzy sets provide a smooth transition between member and non-member of a set. We gave three different definitions for interest measure: fuzzy support, fuzzy confidence, and fuzzy correlation.

We showed two different methods of mining fuzzy quantitative association rules: without normalization, and with normalization. The unnormalized method gives the highest number of interesting rules. The normalized fuzzy method gives about the same number of rules as the discrete. However, either result set might not be included in the other.

We proposed a new algorithm for mining such quantitative association rules. Our experiments on a real-life dataset indicate that the algorithm scales linearly with the number of records. They also showed that the confidence interest measure gives better performance than the correlation interest measure.

Acknowledgment

I wish to thank Jukka Teuhola for his insightful comments and suggestions. Also thanks to the anonymous referee for promoting the clarity of the paper.

References

[1] Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. Proc. of ACM SIGMOD (1993) 207–216

[2] Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. Proc. of the 20th VLDB Conference (1994) 487–499

[3] Chan, Keith C.C., Au, Wai-Ho: Mining Fuzzy Association Rules. Proc. of CIKM Conference, LasVegas, Nevada, USA (1997) 209–215

[4] Fayyad, U.M., Uthurusamy, R.: Efficient algorithms for discovering association rules. AAAI Workshop on KDD, Seattle, Washington, (1994) 181–192

[5] Gyenesei A. Data mining approach for solving decision support problems of warehouse networks. POLVAX (in Hungarian), (1999) 69–93.

[6] Gyenesei, A.: Determining Fuzzy Sets for Quantitative Attributes in Data Mining Problems. Proc. of Advances in Fuzzy Systems and Evol. Comp. (2001) 48–53

[7] Gyenesei A. Interestingness Measures for Fuzzy Association Rules. In Proc. of the 5th European Conference on PKDD (accepted) (2001)

[8] Hong, T-P., Kuo, C-S, Chi, S-C.: Mining association rules from quantitative data. Intelligent Data Analysis 3 (5) (1999) 363–376

[9] Kuok, C.M., Fu, A., Wong, M.H.: Fuzzy association rules in databases. In ACM SIGMOD Record 27(1),(1998) 41–46

[10] Park, J.S., Chen, M-S., Yu, P.S.: An effective hash-based algorithm for mining association rules. Proceedings of ACM SIGMOD, (1995) 175-186.

[11] Piatetsky-Shapiro, G., Frawley, W.J.: Knowledge Discovery in Databases. Chapter 13. AAAI Press/The MIT Press, Menlo Park, California (1991)

[12] Srikant, R., Agrawal, R.: Mining quantitative association rules in large relation tables. Proc. of ACM SIGMOD (1996) 1–12

[13] Srikant, R., Agrawal, R.: Fast algorithms for mining association rules. Proceedings of the 20th VLDB Conference (1994) 487–499.