Association Rules for Expressing Gradual Dependencies

Eyke Hüllermeier

Department of Mathematics and Computer Science
University of Marburg, Germany
eyke@mathematik.uni-marburg.de

Abstract. Data mining methods originally designed for binary attributes can generally be extended to quantitative attributes by partitioning the related numeric domains. This procedure, however, comes along with a loss of information and, hence, has several disadvantages. This paper shows that fuzzy partitions can overcome some of these disadvantages. Particularly, fuzzy partitions allow for the representation of association rules expressing a tendency, that is, a gradual dependence between attributes. This type of rule is introduced and investigated from a conceptual as well as a computational point of view. The evaluation and representation of a gradual association is based on linear regression analysis. Furthermore, a complementary type of association, expressing absolute deviations rather than tendencies, is discussed in this context.

1 Introduction

Data mining aims at extracting understandable pieces of knowledge from usually large sets of data stored in a database. It comes as no surprise that *rule-based* models play a prominent role in this field, as rules provide a simple and intelligible yet expressive means of knowledge representation. Among the related techniques that have been developed, so-called *association rules* (or associations for short) have gained considerable attraction [1]. An association rule is meant to represent dependencies between attributes in a databases. Typically, such a rule involves two sets A and B of binary attributes, also called features or items. The intended meaning of a rule symbolized as $A \rightharpoonup B$ is that a transaction (a data record stored in the database) that contains the set of items A is likely to contain the items B as well.

Generally, a database does not contain binary attributes only but also attributes with values ranging on (completely) ordered scales, e.g. cardinal or ordinal attributes. This has motivated a corresponding generalization of (binary) association rules [5]. The simplest approach, to be detailed in Section 2, is to partition the domain $\mathfrak{D}_X$ of a quantitative attribute X into intervals $A \subseteq \mathfrak{D}_X$ and to associate a new binary variable with each interval. This leads to interval-based association rules of the form $X \in A \rightharpoonup Y \in B$.

A slightly different type of association rule, particularly interesting in connection with quantitative attributes, has recently been considered in [2]. This

T. Elomaa et al. (Eds.): PKDD, LNAI 2431, pp. 200–211, 2002.

type of rule, which has a more statistical flavor, is of the following form:

$$X \in A \rightharpoonup mean(Y) = \bar{y}_A, \tag{1}$$

where X and Y are attributes and A is an interval. This rule says that the mean value of Y is $\bar{y}_A$ if the database is restricted to those transactions satisfying $X \in A$, an information which is clearly interesting if $\bar{y}_A$ deviates significantly from the overall (unconditional) mean $\bar{y}$. The basic idea underlying this approach can be summarized as follows: The (empirical) *distribution* of an attribute Y changes significantly when focusing on a certain subpopulation (a subset of the database). In this connection, a subpopulation is specified by the condition $X \in A$, and the change of the distribution is measured by the change of the mean. Clearly, the mean could be replaced by any other statistic of interest, for example the variance or the median, or even by the distribution of Y itself. See [3] for a closely related data mining method.

In this paper, we elaborate further on quantitative association rules. Especially, we propose a new type of rule which is able to express a kind of *tendency*, that is, a gradual dependence between attributes. In this connection, the idea of a "fuzzy" partition of a quantitative domain plays an important role. After having pointed out some difficulties caused by classical partitions (Section 2), this idea will be motivated in Section 3. In Section 4, two types of association rules will be introduced, namely the aforementioned rules expressing a tendency and a complementary type of rule expressing absolute deviations.

Notation. We proceed from a database $\mathcal{D}$, which is a collection of transactions (records) t. A transaction t assigns a value $t[A]$ to each attribute $A \in \mathcal{A}$, where $\mathcal{A}$ is an underlying set of attributes. We focus on cardinally scaled attributes X; the domain of an attribute X is denoted $\mathfrak{D}_X$. When discussing simple rules involving two (fixed) attributes X and Y, we consider the database $\mathcal{D}$ as a collection of data points $(x, y) = (t[X], t[Y])$, i.e. as a projection of the original database. This notation is generalized to rules involving more variables in a canonical manner. An association rule is written in the form $A \rightharpoonup B$, or sometimes $\{A \rightharpoonup B\}$, where A and B can be single items or sets of items.

2 Problems with Binary Partitions

Most algorithms in data mining have been designed for binary variables, and methods capable of dealing with quantitative attributes are mostly extensions of these algorithms. A standard approach in this connection is to replace a quantitative attribute X with domain $\mathfrak{D}_X$ by a finite set of binary variables $X_{A_\imath}$ with domain $\{0, 1\}$, where the $A_\imath \subseteq \mathfrak{D}_X$, $1 \leq \imath \leq k$, are intervals such that $\bigcup_{\imath=1}^{k} A_\imath = \mathfrak{D}_X$. An attribute $X_{A_\imath}$ takes the value 1 if the related quantitative value x is covered by $A_\imath$ and 0 otherwise: $X_{A_\imath} = 1 \Leftrightarrow x \in A_\imath$. Algorithms for binary attributes can then be applied to the new (transformed) data set.

Needless to say, this type of binarization comes along with a loss of information, since the precise value x cannot be recovered from the values of the

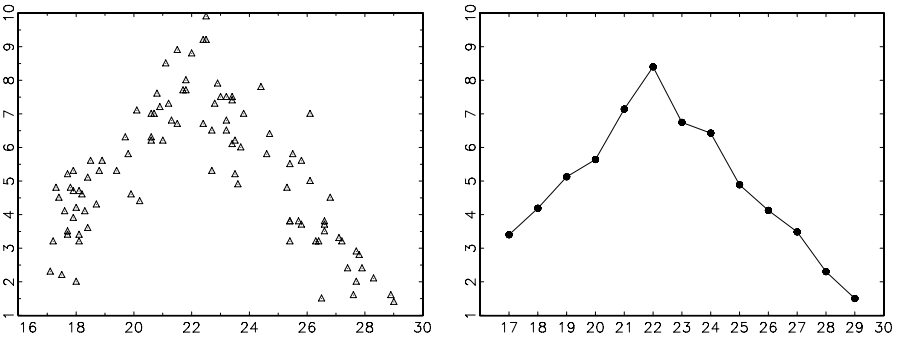

Fig. 1. Left: Observations (transactions) plotted as points in the instance space (x-axis: `weight`, y-axis: `perf`). Right: Mean performance for subclasses $D_w \doteq \{(x, y) \in D \,|\, w - 1/2 < x \leq w + 1/2\}$, $w = 17, 18, \ldots, 29$

binary variables $X_{A_1}, \ldots, X_{A_k}$. Likewise, a subset $A_\imath \times B_\jmath$ of the *instance space* $\mathfrak{D}_X \times \mathfrak{D}_Y$ becomes a "black box" when considering two variables X and Y. Information about the *distribution* of points $(x, y) \in A_\imath \times B_\jmath$ is hence completely lost, which makes it impossible to discover *local* interdependencies between X and Y. Consequently, interval-based rules might convey a misleading picture of the underlying data. These problems are further aggravated by the sharp boundaries between the intervals or, more generally, between the range of support and non-support of a binary feature.

To illustrate, consider an artificial data set comprised of 100 data points $(x_\imath, y_\imath)$. The related variables X and Y can be thought of as, say, the weight of a dog in kg (`weight`) and a certain physical performance (`perf`) measured on a scale ranging from 0 (bad) to 10 (excellent). The following table shows some of the data:

$\imath$	1	2	3	4	5	6	7	8	9	11 ...
$x_\imath$	22.4	25.8	22.4	27.7	27.7	25.7	20.2	18.1	18.1	17.4 ...
$y_\imath$	6.7	5.6	9.2	2.0	2.9	3.8	4.4	3.4	4.7	4.5 ...

Fig. 1 (left) shows the complete data as points in the instance space. On average, `perf` seems to increase with `weight` up to a value of about 22, and to decrease afterwards. This impression is confirmed by the right picture in the same figure, showing the mean values of `perf` separately for the subpopulations $D_w \doteq \{(x, y) \in D \,|\, w - 1/2 < w \leq w + 1/2\}$, $w = 17, 18, \ldots, 29$.

Now, a rule of the following kind would nicely characterize the above data: *The more normal the weight, the better the performance*. Note that this rule expresses a *tendency*, which cannot be accomplished by a (single) classical association rule. Moreover, the above rule involves a cognitive concept, namely "normal weight", which is here understood as "a weight close to 22 kg". Such concepts do often preexist in the head of a data miner, but are seldom adequately represented by an interval created in the course of a rule mining procedure.

In our example, a reasonable candidate for an interval-based association would be a rule of the form

$$\texttt{weight} \in [\,22-\gamma,\,22+\gamma\,] \;\rightharpoonup\; \texttt{perf} \in [\,7,\,10\,], \tag{2}$$

suggesting that a dog whose weight is close to 22 kg is likely to perform rather well. But how should γ be chosen? Note that (2), in conjunction with the rule

$$\texttt{weight} \notin [\,22-\gamma,\,22+\gamma\,] \;\rightharpoonup\; \texttt{perf} \in [\,0,\,7\,[\,,$$

induces a classification into low-performance and high-performance dogs. The borderline between these two groups is clearly arbitrary to some extent, and the classification will hardly reflect the true nature of the data.

The same problem occurs in the approach in [2], where a rule would be specified as

$$\texttt{weight} \in [\,22-\gamma,\,22+\gamma\,] \;\rightharpoonup\; \mathit{mean}(\texttt{perf}) = f(\gamma). \tag{3}$$

Here, the (conditional) mean would be a decreasing function of γ, and the rule (3) would be considered as interesting only if $f(\gamma)$ is –in a statistical sense– *significantly* larger than the overall mean of `perf`. Again, there is no natural choice of the length of the interval. In [2], γ is basically determined by the confidence level of a t-test used for testing significance, which only removes but does not solve the problem.

Finally, let us mention two further problems of the interval-based approach. Firstly, sharp boundaries between intervals may lead to undesirable threshold effects, in much the same way as do *histograms* in statistics: A slight variation of the boundary points of the intervals can have a considerable effect on the histogram induced by a number of observations and may even lead to qualitative changes, that is changes of the shape of the histogram. Likewise, the variation of an interval can strongly influence the evaluation of a related association rule. Secondly, the interval-based approach becomes involved if the class of allowed intervals is not restricted in a proper way (for example in the form of fixed underlying partitions for the attributes). On the one hand, a rich class of intervals guarantees flexibility and representational power. On the other hand, one has to keep track of possible interactions between apparently interesting rules. For example, the antecedent and/or the consequent parts of two rules can overlap, which may cause problems of redundancy.

In summary, this section has pointed out the following difficulties of interval-based associations: Firstly, such rules are not able to express gradual dependencies between attributes. Secondly, some problems are caused by sharp boundaries: Their specification is often arbitrary, the evaluation of rules is sensitive toward the variation of boundary points, and rules are not very user-friendly due to a lack of readability and "cognitive relevance". Thirdly, additional complications and computational costs occur if interactions between interval-based rules are not excluded in advance by restricting the class of allowed intervals.

3 Fuzzy Partitions

The use of fuzzy sets in connection with association rules – as with data mining in general [8] – has recently been motivated by several authors (e.g. [6]). Among other aspects, many of the aforementioned problems can be avoided – or at least alleviated – by the use of *fuzzy* instead of crisp (non-fuzzy) partitions. A fuzzy subset of a set (domain) $\mathfrak{D}$ is identified by a so-called membership function, which is a generalization of the characteristic function $\mathbb{I}_A(\cdot)$ of an ordinary set $A \subseteq \mathfrak{D}$ [10]. For each element $x \in \mathfrak{D}$, this function specifies the degree of membership of x in the fuzzy set. Usually, membership degrees are taken from the unit interval $[0, 1]$, i.e. a membership function is a mapping $\mathfrak{D} \to [0, 1]$. We shall use the same notation for ordinary sets and fuzzy sets. Moreover, we shall not distinguish between a fuzzy set and its membership function, that is, $A(x)$ denotes the degree of membership of the element x in the fuzzy set A. Note that an ordinary set A can be considered as a "degenerate" fuzzy set with membership degrees $A(x) = \mathbb{I}_A(x) \in \{0, 1\}$.

Fuzzy sets formalize the idea of *graded membership*, which allows an element to belong "more or less" to a set. A fuzzy set can have "non-sharp" boundaries. Consider the above mentioned concept of "normal weights" as an example. Is it reasonable to say that 23.4 kg is a normal weight (for a dog in our example) but 23.5 kg is not? In fact, any sharp boundary will appear rather arbitrary. Modeling the concept "normal weight" as a fuzzy set A, it becomes possible to express, for example, that a weight of 22 kg is completely in accordance with this concept ($A(22) = 1$), 24 kg is a "more or less" normal weight ($A(24) = 0.5$, say), and 26 kg is clearly not normal ($A(26) = 0$).

As can be seen, fuzzy sets can provide a reasonable representation of linguistic expressions and cognitive concepts. This way, they act as an interface between a quantitative, numerical level and a qualitative level, where knowledge is expressed in terms of natural language. In data mining, fuzzy sets thus allow for expressing patterns found at the quantitative (database) level in a user-friendly way.

Concerning the class of fuzzy concepts underlying the rule mining process, we advocate a fixed partition for each attribute. Even though the assumption of a fixed partition is often regarded as critical, it appears particularly reasonable in the fuzzy case. Apart from a simplification of the rule mining procedure, a fixed partition specified by the user or data miner himself guarantees the interpretability of the rules. In fact, the user will generally have a concrete idea of terms such as "normal weight". Since it is the user who interprets the association rules, these rules should exactly reflect the meaning he has in mind and, hence, the user himself should characterize each linguistic expression in terms of a fuzzy set.

In this connection, it is worth mentioning that a given class of fuzzy concepts can be extended through the use of so-called (linguistic) modifiers. For example, applying the linguistic hedge (modifier) "almost" to the fuzzy concept "normal weight" – modeled by a fuzzy set A – yields the new concept "almost normal weight". Formally, this concept is represented by means of a suitable transfor-

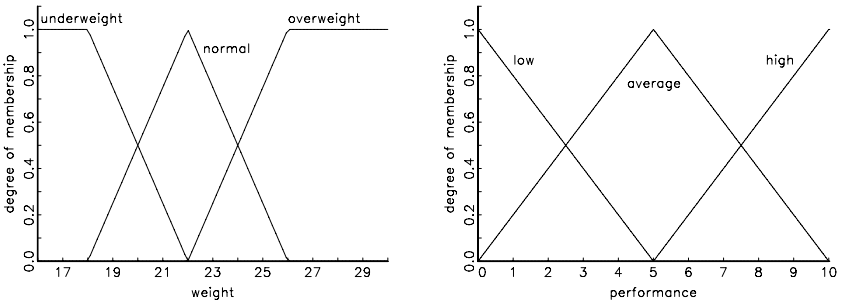

Fig. 2. Exemplary fuzzy partitions of the domains for weight and performance

mation of A. This way, a large number of interpretable fuzzy concepts can be built from a basic repertoire of fuzzy sets and modifier functions. However, we shall not elaborate any further on this aspect. Rather, we assume a fixed fuzzy partition for each attribute X. Formally, such a partition is defined as a class $\{A_1, \ldots, A_k\}$ of fuzzy sets $A_\imath : \mathfrak{D}_X \to [0,1]$ such that $\max_{1 \leq \imath \leq k} A_\imath(x) > 0$ for all $x \in \mathfrak{D}_X$. Fig. 2 shows fuzzy partitions of the domain of weights (using three fuzzy sets: underweight, normal, overweight) and the domain of performance (again with three fuzzy sets: low, average, high).

The discussion so far has shown that a (fixed) fuzzy partition can avoid some of the drawbacks related to classical partitions. Concerning the idea of association rules capable of expressing gradual dependencies between attributes, the following section will show that fuzzy partitions can also be beneficial in that respect.

4 Tendency and Deviation Rules

The basic quality measures for binary association rules $A \rightharpoonup B$ can be derived from the following contingency table:

	$B(y)=0$	$B(y)=1$	
$A(x)=0$	n_{00}	n_{01}	$n_{0\bullet}$
$A(x)=1$	n_{10}	n_{11}	$n_{1\bullet}$
	$n_{\bullet 0}$	$n_{\bullet 1}$	n

(4)

For example, the well-known support and confidence of a rule are given, respectively, by $\mathsf{supp}(A \rightharpoonup B) = n_{11}/n$ and $\mathsf{conf}(A \rightharpoonup B) = n_{11}/n_{1\bullet}$, where $n_{\imath\jmath}$ $(\imath, \jmath \in \{0,1\})$ is the number of tuples $(x,y) \in \mathcal{D}$ such that $A(x) = \imath$ and $B(y) = \jmath$.

In the fuzzy case, $A(x)$ and $B(y)$ can take any value in the unit interval. This suggests extending the above contingency table to a *contingency diagram* as shown in Fig. 3. A record $(x,y) \in \mathcal{D}$ gives rise to a point with coordinates (u,v) in this diagram, where $u = A(x)$ is the degree of membership of x in A (the abscissa) and $v = B(y)$ is the degree of membership of y in B (the ordinate). As

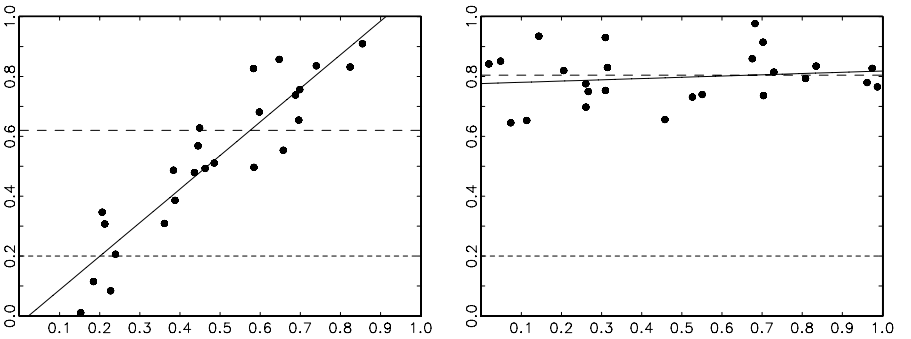

Fig. 3. Exemplary contingency diagrams for an underlying association $A \rightharpoonup B$. Each point is associated with a sample $(x, y) \in \mathcal{D}_A$: The abscissa corresponds to the membership of x in the fuzzy set A, the ordinate is the membership degree $B(y)$. The lines drawn by short and long dashes mark, respectively, the overall and conditional (given x is in A) mean value of $v = B(y)$. The third line is the regression line

will be seen, the contingency diagram provides a useful point of departure for specifying an association between A and B. Note that (4) can indeed be seen as a special case of a contingency diagram: In the non-fuzzy (binary) case, all points are located in the four "corners" of this diagram.

4.1 Contingency Diagrams

In order to illustrate the concept of a contingency diagram, consider the exemplary diagrams shown in Fig. 3. The following information is provided: Each point in a diagram corresponds to a tuple $(A(x_\imath), B(y_\imath))$, where $(x_\imath, y_\imath) \in \mathcal{D}_A \doteq \{(x, y) \in \mathcal{D} \,|\, A(x) > 0\}$; the objects (x, y) with $A(x) = 0$ are ignored. The solid line is the regression line derived for the points $\mathcal{D}_A$, i.e. the linear approximation $u \mapsto \alpha\, u + \beta$ minimizing the sum of squared errors

$$\sum_{\imath=1}^{|\mathcal{D}_A|} \big(\, \alpha\, A(x_\imath) + \beta - B(y_\imath) \,\big)^2 . \tag{5}$$

The line drawn by short dashes marks the overall mean value of $v = B(y)$, that is the value $\bar{v} \doteq |\mathcal{D}|^{-1} \sum_{(x,y)\in\mathcal{D}} B(y)$. This is the average degree to which the objects in $\mathcal{D}$ have the property B. Finally, the line drawn by long dashes shows the conditional mean of $v = B(y)$, given that x is in A. Since A is a fuzzy set, this value is calculated as a weighted average:

$$\bar{v}_A \doteq \left(\sum_{(x,y)\in\mathcal{D}} A(x) \cdot B(y) \right) \cdot \left(\sum_{(x,y)\in\mathcal{D}} A(x) \right)^{-1} .$$

Now, consider the first diagram in more detail. As can be seen, there is a strong correlation between $A(x)$ and $B(y)$. In fact, the positive slope of the regression line suggests the following tendency: *The more x is in A, the more y is in B*. Moreover, the conditional mean $\bar{v}_A$ appears to be significantly larger than the overall mean $\bar{v}$. The basic idea of our approach, to be detailed below, is to derive a suitable (linguistic) representation of an association $A \rightharpoonup B$ on the basis of this information.

The regression line in the second diagram has a slope close to 0. Still, the conditional mean $\bar{v}_A$ is much larger than the overall mean $\bar{v}$, suggesting a rule of the following kind: *If x is in A, then y is more in B than usual*.

4.2 From Contingency Diagrams to Association Rules

Clearly, the information provided by the contingency diagram can be regarded as reliable only if the diagram contains enough points. First of all, we therefore apply the common support criterion: A rule $A \rightharpoonup B$ is taken into consideration only if $\mathsf{supp}(A \rightharpoonup B)$ exceeds a given threshold σ, where $\mathsf{supp}(A \rightharpoonup B) \doteq \sum_{(x,y)\in\mathcal{D}} A(x)$. Note that this definition of support differs from the usual definition of fuzzy support, which is $\mathsf{supp}(A \rightharpoonup B) \doteq \sum_{(x,y)\in\mathcal{D}} \min\{A(x), B(y)\}$.[1] In fact, it is modeled on the two types of association rules that will be introduced below: It corresponds to the (fuzzy) number of points considered when evaluating a rule of the first type and defines a lower bound to the number of involved points in second case.

Information from a Contingency Diagram. We proceed from the following information taken from the contingency diagram:

- The mean values $\bar{v}$, $\bar{v}_A$ (and the number of points $n_A \doteq |\mathcal{D}_A|$).
- The coefficients α, β of the regression line.
- A measure Q indicating the quality of the regression.

Here, we take Q as the usual R^2 coefficient, defined as

$$R^2 \doteq 1 - \frac{\sum_{i=1}^{n_A} e_i^2}{\sum_{i=1}^{n_A} (v_i - \bar{v}_A)^2},$$

where $e_i = v_i - (\alpha\, u_i + \beta)$, $(u_i, v_i) = (A(x_i), B(y_i))$. Of course, R^2 can be replaced or complemented by other measures. In this connection, it should also be mentioned that $A(x_i)$ and $B(y_i)$ might be related in a monotone though *nonlinear* way.[2] In such a case, a linear regression might lead to poor quality measures. Even though we restrict ourselves to the linear case in this paper, the method could clearly be extended in the direction of more general regression functions. For example, a straightforward (and easy to implement) generalization is to fit a polynomial of degree 2 to the data.

[1] The minimum operator is sometimes replaced by other combination operators.
[2] The Durbin-Watson test statistic is a useful indicator in this respect.

Note that simple formulae exist for the coefficients α, β in (5), e.g.

$$\beta = \frac{n_A \sum_{i=1}^{n_A} u_i v_i - \sum_{i=1}^{n_A} u_i \sum_{i=1}^{n_A} v_i}{n_A \sum_{i=1}^{n_A} u_i^2 - (\sum_{i=1}^{n_A} u_i)^2}, \tag{6}$$

$$\alpha = \bar{v}_A - \beta\, \bar{u}_A. \tag{7}$$

Let us anticipate a possible criticism of this derivation of regression coefficients: If the *marginal points* of the form $(A(x), B(y)) = (u, 0)$ are regarded as *censored* observations,[3] simple linear regression techniques are actually not applicable and must be replaced by more sophisticated methods, such as Tobit regression models. Anyway, since our focus is on association rules rather than regression analysis, we shall not deepen this aspect further and rather proceed from (6–7) which yields at least good approximate results.

Generation of Rules. On the basis of the above information, two types of rules will be generated. The first type of rule, called *deviation rule* and denoted $A \rightarrow^d B$, expresses a (significant) deviation of the conditional mean. Suppose the points $(x, y) \in \mathcal{D}$ to be divided into two (fuzzy) samples: one for which $x \in A$ and one for which $x \notin A$. A point (x, y) belongs to the first sample, S_1, with degree $A(x)$ and to the second sample, S_2, with degree $1 - A(x)$. Let $\bar{v}_1 = \bar{v}_A$ and

$$\bar{v}_2 = \frac{\sum_{(x,y)\in\mathcal{D}} (1 - A(x))B(y)}{\sum_{(x,y)\in\mathcal{D}} 1 - A(x)}$$

denote, respectively, the average of the membership degrees $B(y)$ in S_1 and S_2. These averages can be considered as estimations of underlying parameters (expected values) ν_1 and ν_2. Moreover, $\delta = \bar{v}_1 - \bar{v}_2$ is a simple point estimation of the deviation $\nu_1 - \nu_2$. Now, suppose $A \rightarrow^d B$ to be considered as interesting if $\nu_1 - \nu_2 > \Delta$, where Δ is a user-defined threshold. How can the "interestingness" of $A \rightarrow^d B$ on the basis of $\delta = \bar{v}_1 - \bar{v}_2$ be decided? Statistically speaking, the question is whether $\delta = \bar{v}_1 - \bar{v}_2$ is *significantly* larger than Δ. An appropriate decision principle is provided by the t-test adapted to the fuzzy case [7]:[4]

$$T = \frac{\bar{v}_1 - \bar{v}_2 - \Delta}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}, \tag{8}$$

where $n_1 = |S_1| = \sum_{(x,y)\in\mathcal{D}} A(x)$, $n_2 = |S_2| = |\mathcal{D}| - n_1$, and s_1^2, s_2^2 denote, respectively, the variance of $B(y)$ for the two (fuzzy) samples. The deviation is considered to be significant at the .05 confidence level if $T > 1.645$.

Note that the denominator in (8) will generally be small (s_1^2, s_2^2 are upper-bounded by 1). In fact, it is not difficult to prove that $T > 1.645$ as soon as

$$\delta \;>\; \Delta^* \doteq \Delta + \frac{1.645}{\sqrt{|\mathcal{D}|\sigma(1-\sigma)}}, \tag{9}$$

[3] The membership of y in B cannot be negative.

[4] This test compares the difference between the mean values of two *fuzzy* populations.

where σ is the support threshold. The right-hand side in (9) can be seen as a modified threshold that includes a "confidence offset".

Once a deviation has been found to be significant, an adequate deviation rule can be defined on the basis of δ. This can be done by appending the corresponding averages to the rule, which is then of the form $\{A \rightharpoonup^d B\ [\bar{v}_1, \bar{v}_2]\}$. Another possibility is to present the rule in a linguistic form, paraphrasing the deviation δ by terms such as "slightly more", "more", or "much more". In our example above, one would find the rule $\{\,\texttt{normal} \rightharpoonup^d \texttt{high}\ [.40, .08]\,\}$, which could be translated as follows: *If the weight is normal, then the performance is much higher than usual.*

Note that we have only tested for *positive* deviations. Of course, one could also represent negative deviations, using terms such as "less" or "much less" associated with values $\delta < 0$. However, this does again cause problems of redundancy: If B_1 and B_2 are complementary concepts in the sense that $B_1(y)$ and $B_2(y)$ are negatively correlated (such as low and high performance), then the positive deviation for B_1 will come along with the negative deviation for B_2 and vice versa. As in classical association analysis, we shall henceforth concentrate on positive deviations.

A second type of rule, called tendency rule and denoted $A \rightharpoonup^t B$, represents a gradual dependence between the concepts A and B. More precisely, it indicates that an increase in $A(x)$ comes along with an increase in $B(y)$. The validity of such a rule is judged on the basis of the regression coefficients (6–7) and the quality measure Q. For example, a simple decision principle is to reject a rule iff Q falls below a given threshold or the slope of the regression line, α, is too small: $Q < Q_{min}$ or $\alpha < \alpha_{min}$. If a rule is accepted, it might be presented in the form $\{A \rightharpoonup^t B\,[\alpha, \beta]\}$. Alternatively, a linguistic representation is possible: *The more x is in A, the more y is in B.* Again, this representation can be refined in dependence on the specific values of α and β. In our example above, the rule $\{\,\texttt{normal} \rightharpoonup^t \texttt{high}\,[0.65, -0.05]\,\}$ would be supported ($R^2 = 0.77$): *The more normal the weight, the higher the performance.*

4.3 Rules with Compound Conditions

So far, we have only considered simple rules involving two attributes. However, the approach outlined above easily extends to rules with a compound antecedent: Consider a rule of the form $A_1, \ldots, A_m \rightharpoonup B$, where $A_\imath$ is an element of the fuzzy partition of $\mathfrak{D}_{X_\imath}$, the domain of attribute $X_\imath$ $(1 \leq \imath \leq m)$, and B is an element of the fuzzy partition of variable Y. The antecedent of this rule stands for a conjunction of the conditions $x_\imath \in A_\imath$.

In fuzzy set theory, the logical conjunction is modeled by means of a so-called *t-norm.* This is a binary operator $\otimes : [0,1] \times [0,1] \to [0,1]$ that is associative, commutative, non-decreasing in both arguments, and satisfies $\alpha \otimes 1 = 1 \otimes \alpha = \alpha$ for all $0 \leq \alpha \leq 1$. The most important t-norms are the minimum $(\alpha, \beta) \mapsto \min\{\alpha, \beta\}$, the product $(\alpha, \beta) \mapsto \alpha\beta$, and the Lukasievicz t-norm $(\alpha, \beta) \mapsto \max\{\alpha + \beta - 1, 0\}$.

In the special case of an association $A \rightharpoonup B$ involving two variables X and Y, a value $A(x)$ corresponds to the degree to which $x \in A$ is satisfied (the fuzzy

truth degree of the proposition $x \in A$). In the more general case, this value is given by the conjunction $A_1(x_1) \otimes A_2(x_2) \otimes \ldots \otimes A_m(x_m)$. Actually, this comes down to considering the attribute in the condition part as an m-dimensional variable $X = (X_1, \ldots, X_m)$. As before, a rule can then be written in the form $A \rightharpoonup B$, where the fuzzy set A is defined as

$$A : \mathfrak{D}_{X_1} \times \ldots \times \mathfrak{D}_{X_m} \to [0,1]\,,\ (x_1, \ldots, x_n) \mapsto A_1(x_1) \otimes \ldots \otimes A_m(x_m).$$

Again, one thus obtains a point $(u, v) = (A(x), B(y))$ for each transaction $(x, y) = (x_1, \ldots, x_m, y) \in \mathcal{D}$ and, hence, a contingency diagram as introduced above. In other words, a rule with a compound condition part can be evaluated in the same way as a rule with a simple antecedent.

As concerns the problem of redundancy and interaction between association rules, it is important to mention that *none* of the following properties hold:

$$\begin{aligned} A_1 \rightharpoonup B \ \wedge\ A_2 \rightharpoonup B \quad &\Rightarrow \quad A_1, A_2 \rightharpoonup B \\ A_1, A_2 \rightharpoonup B \quad &\Rightarrow \quad A_1 \rightharpoonup B \ \vee\ A_2 \rightharpoonup B \end{aligned}$$

Still, some kind of pruning is clearly advisable. Especially, this concerns the relation between a rule $A \rightharpoonup B$ and its specializations $A^+ \rightharpoonup B$ with $A \subsetneq A^+$. For example, given the deviation rule $A \rightharpoonup B\,[\bar{v}_1, \bar{v}_2]$, a rule $A^+ \rightharpoonup B\,[\bar{v}_1^+, \bar{v}_2^+]$ will not be interesting if $\bar{v}_1^+ \leq \bar{v}_1$. More generally, one might adopt a *minimum improvement constraint* in order to eliminate unnecessarily complex rules [4].

4.4 Rule Mining and Computational Aspects

How does one find interesting instances of the two types of association rules introduced in this section? As already mentioned above, the first step is to find the frequent itemsets. To this end, any of the existing procedures can be used, for example the Apriori algorithm (for quantitative attributes [9]). Note that an itemset is now a class of fuzzy sets $\{A_1, \ldots, A_m\}$, where $A_\imath$ is an element of the fuzzy partition of an attribute $X_\imath$ (and $X_\imath \neq X_\jmath$ for all $\imath \neq \jmath$). The frequent itemsets determine the condition parts of the *candidate rules*.

In order to evaluate a candidate rule $A \rightharpoonup^t B$ one needs to compute the regression coefficients α and β as well as the quality measure Q. A look at (6–7) reveals that α and β can be derived by a single scan of the database. Afterwards, the quality measure Q (which is here taken as R^2) can be obtained, which makes one further scan necessary.

The evaluation of a rule $A \rightharpoonup^d B$ comes down to computing the deviation δ as well as the test statistic (8). This requires two scans of the database, since the computation of the variances s_1^2 and s_2^2 in (8) assumes the mean values $\bar{v}_1$ and $\bar{v}_2$ to be known. Therefore, these values have to be derived first. Alternatively, an approximate evaluation can be obtained on the basis of (9), which requires only a single scan.

In summary, it can be seen that the rule mining procedure is quite efficient. Apart from the search for frequent itemsets, it merely requires two additional scans of the database.

5 Concluding Remarks

We have introduced two types of quantitative association rules, referred to as deviation rules and tendency rules. The former type of rule is basically a fuzzy counterpart to the approach in [2]. The latter type of rule is able to represent gradual dependencies between attributes. This becomes possible by the use of fuzzy partitions for the attributes' domains.

Let us conclude with some remarks. (1) We have applied our approach to several data sets from the UCI repository for which we obtained rather promising results. These experimental studies are not reported here due to limited space; the technical report [7] provides a more detailed exposition. (2) So far, our approach assumes fixed underlying partitions comprised of preexisting "cognitive concepts". On the one hand, this assumption appears especially reasonable in the fuzzy case. On the other hand, one cannot deny that the observation of data might also influence the formation of cognitive concepts. In our running example, for instance, a concept "ideal weight" (which coincides with our definition of "normal weight") might well be established on the basis of the data. Extending the approach so as to support the discovery of such cognitive concepts is an interesting challenge for future work.

References

1. R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the* ACM SIGMOD *Conference on Management of Data*, pages 207–216, Washington, D. C., 1993. 200
2. Y. Aumann and Y. Lindell. A statistical theory for quantitative association rules. In *Proc. 5th ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999. 200, 203, 211
3. S. D. Bay and M. J. Pazzani. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5:213–246, 2001. 201
4. R. J. Bayardo, R. Agrawal, and D. Gunopolos. Constraint-based rule mining in large, dense databases. *Data Mining and Knowledge Discovery*, 4:217–240, 2000. 210
5. T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Mining optimized association rules for numeric attributes. In *Proc. 15th ACM Symposium of Principles of Database Systems*, 1996. 200
6. E. Hüllermeier. Implication-based fuzzy association rules. In L. De Raedt and A. Siebes, editors, *Proceedings* PKDD–01*, 5th European Conference on Principles and Practice of Knowledge Discovery in Databases*, number 2168 in LNAI, pages 241–252, Freiburg, Germany, September 2001. Springer-Verlag. 204
7. E. Hüllermeier. Fuzzy association rules. 21. Workshop *Interdisziplinäre Methoden der Informatik*, Universität Dortmund, 2001. 208, 211
8. W. Pedrycz. Data mining and fuzzy modeling. In *Proc. of the Biennial Conference of the NAFIPS*, pages 263–267, Berkeley, CA, 1996. 204
9. R. Skrikant and R. Agrawal. Mining quantitative association rules in large relational tables. In *Proceedings of the* ACM SIGMOD *International Conference on Management of Data*, pages 1–12, 1996. 210
10. L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965. 204