Measuring Variation Strength in Gradual Dependencies

C. Molina, J.M. Serrano
Department of Informatics
University of Jaén, Spain
e-mail {carlosmo, jschica}@ujaen.es

D. Sánchez[1], M.A. Vila
Department of Computer Science and A.I
University of Granada, Spain
e-mail {daniel, vila}@decsai.ugr.es

Abstract

In this paper we extend a previous definition of gradual dependence as a special kind of (crisp) association rule, in order to measure not only the existence of a tendency, but its strength. The new proposal is based on the idea of fuzzy association rule and the definition of variation strength in the degree of fulfilment of an imprecise property by different objects. We study the new semantics and properties of the resulting fuzzy gradual dependence, and we propose a way to adapt existing fuzzy association rule mining algorithms for the new task of mining such dependencies.

1 Introduction

Gradual dependencies are rules that represent a relation among the variation in the degree of fulfilment of imprecise properties by different objects [10]. This kind of rules express a *tendency*. Consider for instance a database containing data about weight and speed of a set of trucks, and consider the restrictions *high* related to weight and *slow* related to speed, represented by means of suitable fuzzy sets on the domains of the attributes. An example of gradual dependence is *the higher the weight, the lower the speed*, meaning that as the weight of a truck increases, its speed tends to decrease.

The variations in the membership degree considered in gradual dependencies can be of two types: *the more* and *the less*, meaning that the membership degree of the first object to the considered fuzzy set is greater or lower than the membership of the second one, respectively. Hence we can consider four types of gradual dependencies: *the more X is A, the more Y is B* (expressed as $(>, X, A) \rightarrow (>, Y, B)$), *the more X is A, the less Y is B* (expressed as $(>, X, A) \rightarrow (<, Y, B)$), and so on.

In [10], the evaluation and representation of these gradual associations is based on linear regression analysis. The starting point of this approach is the idea of *contingency diagram.* Given two attributes X and Y, fuzzy sets A and B defined on X and Y, respectively, and a database $\mathcal{D}$ containing pairs of values $(x, y) \in X \times Y$, a contingency diagram is a two-dimensional plot of points $(A(x), B(y))$ such that $A(x) > 0$. A gradual dependence, represented as a *tendency rule* $A \rightarrow^t B$, means that " ... an increase in $A(x)$ comes along with an increase in $B(y)$". The validity of the rule is assessed on the basis of the regression coefficients $[\alpha, \beta]$ of the line that approximates the points in the contingency diagram (α being the slope of the line) and the quality of the regression as given by the R^2 coefficient.

In [4] we introduced an alternative approach, in which a gradual dependence is a rule of the form $(*_1, X, A) \rightarrow (*_2, Y, B)$, with $*_1, *_2 \in \{<, >\}$. The dependence holds in $\mathcal{D}$ iff $\forall (x, y), (x', y') \in \mathcal{D}$, $A(x) *_1 A(x')$ implies $B(y) *_2 B(y')$. The discovery of such dependencies is based on mining for association rules in a suitable set of transactions obtained from the database. For that purpose we define items of the form $[>, X, A]$ and $[<, X, A]$, expressing the two possible tendencies of attribute X with respect to the restriction A, and one transaction associated to every pair of objects. An item of the form $[<, X, A]$ (resp. $[<, X, A]$) is in the transaction associated to the pair of objects (o, o') (with values x and x' of X respectively) iff $A(x) < A(x')$ (resp. $A(x) > A(x')$). This way, a gradual dependence $(*_1, X, A) \rightarrow (*_2, Y, B)$ in a database $\mathcal{D}$ corresponds to an association rule of the form $[*_1, X, A] \Rightarrow [*_2, Y, B]$ in the corresponding set of transactions (one for each pair of objects in $\mathcal{D}$). For example, *the higher the weight, the lower the speed* can be expressed by the association rule $[>, Weight, High] \Rightarrow [>, Speed, Low]$. Support and

[1] Corresponding author

accuracy of the rule are employed in order to measure the importance and accuracy of the gradual dependence.

The latter has the advantage that algorithms to discover gradual rules can be obtained by a simple modification of any (crisp) association rule discovery algorithm. However, the semantics of both approaches are different since in [10] the relation between the magnitude of variation in both variables is taken into account, whilst in [4] only the fulfilment of the variation is considered.

In order to illustrate the difference, let us come back to our first example. Let us suppose we have three trucks whose fulfilment of the restrictions *high weight*, *slow speed*, and *big size* is shown in table 1. Let us assess the two gradual dependencies *the higher the weight, the lower the speed* and *the higher the weight, the bigger the size* using the approaches in [10] and [4]. Using [4], both dependencies hold with total accuracy since every time a truck is heavier than another, it is slower and bigger. For this approach, both dependencies hold to the same degree. However, if we look at the contingency diagrams for both dependencies (figure 1), it can be seen that the slope of the regression line for the dependence *the higher the weight, the lower the speed* (the parameters of the regression line are approximately $[0.167, 0.167]$) is smaller than for the dependence *the higher the weight, the bigger the size* (approximately $[1.3, -0.67]$). In both cases, clearly, the regression line fits perfectly the points in the contingency diagrams, so the quality of the regression is $R^2 = 1$. Hence the second dependence is stronger than the first one.

In this paper we propose an extension to the approach in [4] that incorporates the magnitude of variation in the degree of fulfilment of the restrictions in both variables, with the objective of detecting the strength of the dependence in cases like the example above. The new approach is based on the concept of fuzzy association rule, and it is related to previous work about the discovery of fuzzy approximate dependencies [2].

The paper is organized as follows: in section 2 we briefly recall a previous approach to gradual dependencies and we extend it by considering membership variation and fuzzy association rules. In section 3 we introduce the particular case of fuzzy gradual dependencies generated by the approach to fuzzy association rules in [6]. Section 4 is devoted to mining issues and to show some experiments. Finally, section 5 contains our conclusions and future research.

Truck	High weight	Slow speed	Big size
t_1	0.2	0.2	0.2
t_2	0.5	0.25	0.6
t_3	0.8	0.3	1

Table 1: Membership degrees of *high weight*, *slow speed*, and *big size* for three trucks

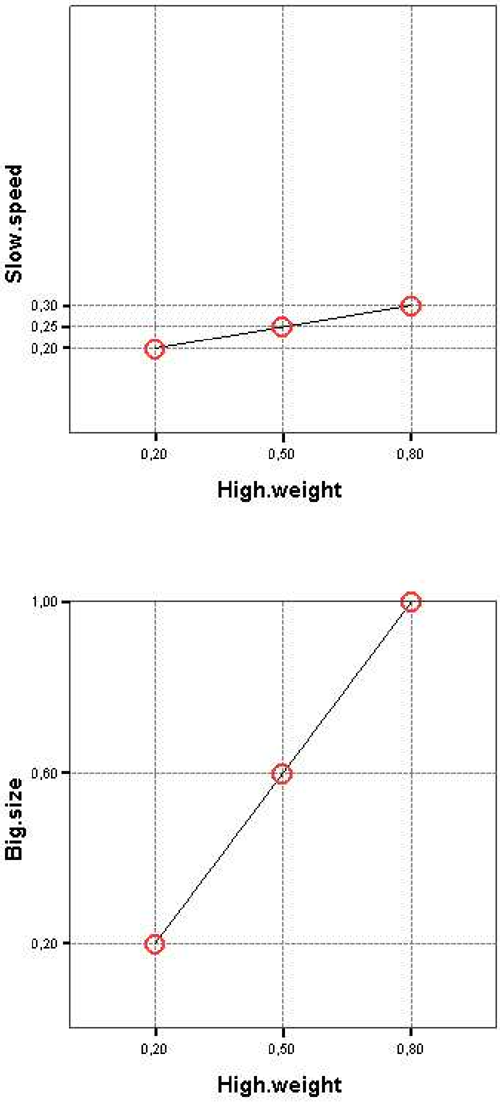

Figure 1: Contingency diagrams for the gradual dependencies *the higher the weight, the lower the speed* and *the higher the weight, the bigger the size* from the data in table 1

$$CF(I_1 \Rightarrow I_2) = \begin{cases} \frac{conf(I_1 \Rightarrow I_2) - supp(I_2)}{1 - supp(I_2)} & conf(I_1 \Rightarrow I_2) > supp(I_2) \\ \frac{conf(I_1 \Rightarrow I_2) - supp(I_2)}{supp(I_2)} & conf(I_1 \Rightarrow I_2) \leq supp(I_2) \end{cases} \tag{1}$$

2 Gradual dependencies with variation strength

In this section we extend our definition of gradual dependence [4] in order to incorporate variation strength in the assessment. First we briefly recall the definition in [4]. Then we define the concept of variation, and use it to extend the definition in [4] by using fuzzy association rules.

2.1 Our previous approach

In [4], a gradual dependence is defined as follows: let X and Y be two attributes, A and B fuzzy sets defined on the domains of X and Y, respectively, and a database $\mathcal{D}$ containing pairs of values $(x, y) \in X \times Y$. Let $*_1, *_2 \in \{<, >\}$. A gradual dependence of the form $(*_1, X, A) \rightarrow (*_2, Y, B)$ holds in $\mathcal{D}$ iff $\forall (x, y), (x', y') \in \mathcal{D}$, $A(x) *_1 A(x')$ implies $B(y) *_2 B(y')$.

This way, a gradual dependence is seen as a rule on a dataset consisting of pairs of objects of the original database. Hence, we use association rules in order to assess gradual dependencies in a database. As it is well known, given a set I of items and a bag T of transactions with $t \subseteq I$ $\forall t \in T$, an association rule is an expression of the form $I_1 \Rightarrow I_2$ with $I_1, I_2 \subset I$, $I_1 \cap I_2 = \emptyset$ [1]. This rule is said to hold in T iff every transaction that contains I_1 contains also I_2. The usual measures are *support* and *confidence*, the former being the number or percentage of transactions containing $I_1 \cup I_2$, and the latter being the percentage of transactions containing I_1 that contain I_2. Many other measures have been proposed, see for example [5, 12, 3, 9]. In this paper we shall employ Shortliffe and Buchanan's *certainty factors*, as proposed in [3]. Let $supp(I_j)$ be the support of the itemset I_j and let $supp(I_1 \Rightarrow I_2) = supp(I_1 \cup I_2)$ be the support of the rule. Let $conf(I_1 \Rightarrow I_2) = supp(I_1 \Rightarrow I_2)/supp(I_1)$ be the confidence. The certainty factor of the rule, $CF(I_1 \Rightarrow I_2)$, is defined in equation 1.

The certainty factor yields a value in $[-1, 1]$ and measures how our belief that I_2 is in a transaction changes when we are told that I_1 is in that transaction. Positive values indicate our belief increases, negative values mean our belief decreases, and 0 means no change. Certainty factors have better properties than confidence, and help to solve some of its inconveniences. In particular, it helps to reduce the number of rules obtained by eliminating those rules that correspond in fact to statistical independence or negative dependence (up to 80 % in some of our experiments). This is shown, among other properties of certainty factors as accuracy measures for association rules, in [3]. Finally, let us remark that the calculation of the certainty factor in the final step of any association rule mining algorithm is straightforward and does not modify the time complexity of the algorithm, since support of the consequent and support and confidence of the rule are all available in this step.

We employ association rules in order to mine for gradual dependencies as follows: let $GI^{\mathcal{D}} = \{[>, X, A], [<, X, A], [>, Y, B], [<, Y, B]\}$ be a set of items and $GT^{\mathcal{D}}$ be a set of transactions containing items from $GI^{\mathcal{D}}$. $GT^{\mathcal{D}}$ is obtained from $\mathcal{D}$ as follows: $\forall$ $o = (x, y), o' = (x', y') \in \mathcal{D}$ there is one transaction $gt_{oo'} \in GT^{\mathcal{D}}$ such that $[*, X, A] \in gt_{oo'}$ iff $A(x) * A(x')$ and $[*, Y, B] \in gt_{oo'}$ iff $B(y) * B(y')$, with $* \in \{<, >\}$. Let us remark that $GT^{\mathcal{D}}$ is a *crisp* set of transactions. Then, the gradual dependence $(*_1, X, A) \rightarrow (*_2, Y, B)$ holds in $\mathcal{D}$ iff the (crisp) association rule $[*_1, X, A] \Rightarrow [*_2, Y, B]$ holds in $GT^{\mathcal{D}}$. The support and confidence of the association rule $[*_1, X, A] \Rightarrow [*_2, Y, B]$ can be employed to assess the gradual dependence $(*_1, X, A) \rightarrow (*_2, Y, B)$. We usually employ support and certainty factor.

Let us remark that with this approach, the support of an item of the form $[*, X, A]$ is

$$supp([*, X, A]) = \frac{\left|\{gt_{oo'} \in GT^{\mathcal{D}} \mid A(x) * A(x')\}\right|}{|GT^{\mathcal{D}}|} \tag{2}$$

and hence the support of a dependence $(*_1, X, A) \rightarrow (*_2, Y, B)$ is the support of the itemset $\{[*_1, X, A], [*_2, Y, B]\}$, as equation 3 shows.

Some important and intuitive properties of this approach are the following: let c be an operator in $\{>, <\}$ such that $c(>) =<$ and $c(<) =>$. Then $supp(\{[*_1, X_1, A_1], \ldots, [*_k, X_k, A_k]\}) = supp(\{[c(*_1), X_1, A_1], \ldots, [c(*_k), X_k, A_k]\})$ (in particular $supp([*, X, A]) = supp([c(*), X, A])$). As a consequence, $supp((*_1, X, A) \rightarrow (*_2, Y, B)) = supp((c(*_1), X, A) \rightarrow (c(*_2), Y, B))$, and the same happens with confidence and certainty factor.

$$supp((*_1, X, A) \to (*_2, Y, B)) = \frac{\left|\left\{gt_{oo'} \in GT^{\mathcal{D}} \mid A(x) *_1 A(x') \wedge B(y) *_2 B(y')\right\}\right|}{|GT^{\mathcal{D}}|} \quad (3)$$

2.2 Membership variation

In the previous approach, only the fact that the membership degree is greater (or lesser) is taken into account. This way, the membership of the item $[*, X, A]$ in a transaction $gt_{oo'} \in GT^{\mathcal{D}}$ corresponding to a pair $o = (x, y), o' = (x', y') \in \mathcal{D}$ is defined as in equation 4.

$$gt_{oo'}([*, X, A]) = \begin{cases} 1 & A(x) * A(x') \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

With this definition, $A(x) = 0$ and $A(x') = 0.1$ yield the same result than $A(x) = 0$ and $A(x') = 1$. However, as we saw in the introduction, this can lead to obtain the same accuracy for dependencies that are intuitively different.

In order to avoid this problem, we propose to replace equation 4 by another expression that provides a degree in $[0, 1]$. We call this a *variation degree.* This way, $gt_{oo'}([*, X, A]) \in [0, 1]$.

There are different possibilities to obtain the degree $gt_{oo'}([*, X, A])$. In this paper we propose to employ that of equation 5:

$$gt_{oo'}([*, X, A]) = v_*(A(x), A(x')) \quad (5)$$

where

$$v_*(a, b) = \begin{cases} |a - b| & a * b \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

As an example, let $o = (0, y)$, $o' = (0.1, y')$, and $o'' = (1, y'')$. Then, $gt_{oo'}([<, X, A]) = 0.1$, $gt_{oo''}([<, X, A]) = 1$, $gt_{o'o''}([<, X, A]) = 0.9$, $gt_{o'o}([<, X, A]) = gt_{o''o}([<, X, A]) = gt_{o''o'}([<, X, A]) = 0$.

The following proposition holds:

Proposition 2.1 *Equation 5 verifies*

1. $gt_{oo'}([*, X, A]) \in [0, 1]$
2. *Suppouse* $A(x) * A(x')$ *and* $A(x) * A(x'')$. *Then* $|A(x) - A(x')| > |A(x) - A(x'')|$ *implies* $gt_{oo'}([*, X, A]) > gt_{oo''}([*, X, A])$
3. $gt_{oo'}([*, X, A]) = gt_{o'o}([c(*), X, A])$

Proof: Trivial. □

We consider that the properties in proposition 2.1 must be verified by the variation degree, despite the way it is calculated.

2.3 A new approach to gradual dependencies

Taking variation degrees into account, we propose a new definition of gradual dependence as a modification of our definition in [4], as follows:

Definition 2.1 *Let X and Y be two attributes, A and B fuzzy sets defined on the domains of X and Y, respectively, and a database $\mathcal{D}$ containing pairs of values $(x, y) \in X \times Y$. Let $*_1, *_2 \in \{<, >\}$. A gradual dependence of the form $(*_1, X, A) \to (*_2, Y, B)$ holds in $\mathcal{D}$ iff $\forall o, o' \in \mathcal{D}$ with $o = (x, y)$ and $o' = (x', y')$, $v_{*_1}(A(x), A(x'))$ implies $v_{*_2}(B(y), B(y'))$.*

where v_* is that of equation 6. Let us remark that the implication that appears in this definition is a *fuzzy implication.* This has two main consequences: first, there are in fact different definitions of gradual dependence, depending on the implication considered. Second, a gradual dependence holds to a certain degree. Hence, we are working in fact with *fuzzy gradual dependencies.*

Now, we can extend our interpretation of gradual dependencies as association rules in [4] in order to consider the variation degree of items. A natural way to extend our first approach is to consider fuzzy association rules. There are many different approaches to the definition and assessment of fuzzy association rules. In general, the different extensions take as starting point, in one way or another, a generalization of transactions to fuzzy transactions as fuzzy subsets of items. The main difference between the different existing approaches is the way they assess the rules (see among others [11, 6, 13, 8]).

Using fuzzy association rules is natural in our case since each item has a membership degree to each transaction, so we have in fact a set of *fuzzy transactions*, i.e., fuzzy subsets of items. However, let us remark that since there is no a single definition of fuzzy gradual dependence, the approach employed for mining the fuzzy rules will define, in practice, a particular type of fuzzy gradual dependence.

Let $GI^{\mathcal{D}} = \{[>, X, A], [<, X, A], [>, Y, B], [<, Y, B]\}$ be a set of items and $\tilde{GT}^{\mathcal{D}}$ be a set of fuzzy transactions containing items from $GI^{\mathcal{D}}$. $\tilde{GT}^{\mathcal{D}}$ is obtained from $\mathcal{D}$ as follows: $\forall\ o = (x, y), o' = (x', y') \in \mathcal{D}$ there is one fuzzy transaction $gt_{oo'} \in \tilde{GT}^{\mathcal{D}}$ such that $gt_{oo'}([*, X, A]) = v_*(A(x), A(x'))$ and $gt_{oo'}([*, Y, B]) = v_*(B(y), B(y'))$, with $* \in \{<, >\}$.

Since a fuzzy association rule defines a special kind of fuzzy implication between the degrees of antecedent and consequent, we can conclude the following:

Proposition 2.2 *A fuzzy association rule $[*_1, X, A] \Rightarrow [*_2, Y, B]$ in $\tilde{GT}^{\mathcal{D}}$ defines a fuzzy gradual dependence $(*_1, X, A) \to (*_2, Y, B)$ in $\mathcal{D}$.*

i.e., fuzzy association rules in $\tilde{GT}^{\mathcal{D}}$ define some particular types of fuzzy gradual dependencies in $\mathcal{D}$.

Following proposition 2.2, the support and confidence (or other accuracy measures) of the fuzzy association rule $[*_1, X, A] \Rightarrow [*_2, Y, B]$ can be employed to assess a particular type of fuzzy gradual dependence $(*_1, X, A) \to (*_2, Y, B)$.

3 A particular definition of fuzzy gradual dependence

As we have seen, there are many possible ways to define fuzzy gradual dependencies, in particular starting from an specific approach to fuzzy association rules. In this paper we shall employ the approach to fuzzy association rules introduced in [6] to obtain a particular definition of fuzzy gradual dependence.

3.1 Our approach to fuzzy association rules

In [6], fuzzy association rules are defined and assessed as follows: let $I = \{i_1, \ldots, i_m\}$ be a set of items and $\tilde{T}$ be a set of fuzzy transactions, where each fuzzy transaction is a fuzzy subset of I. For every fuzzy transaction $\tilde{\tau} \in \tilde{T}$ we note $\tilde{\tau}(i_k)$ the membership degree of i_k in $\tilde{\tau}$. For an itemset I_0 we note $\tilde{\tau}(I_0) = \min_{i_k \in I_0} \tilde{\tau}(i_k)$ the degree to which I_0 is in a transaction $\tilde{\tau}$. A fuzzy association rule is an implication of the form $I_1 \Rightarrow I_2$ such that $I_1, I_2 \subset I$ and $I_1 \cap I_2 = \emptyset$. Notice that this is the same definition of a crisp association rule since, from the structural point of view, there is no difference. The difference is that for fuzzy rules the starting point is a set of fuzzy transactions, and the problem is how to assess the support and accuracy. Strictly speaking, what we call fuzzy association rules are association rules assessed on fuzzy transactions.

We call *representation* of the item i_k, noted $\tilde{\Gamma}_{i_k}$, to the (fuzzy) set of transactions where i_k appears, defined as in equation 7. This representation can be extended to itemsets as in equation 8.

$$\tilde{\Gamma}_{i_k}(\tilde{\tau}) = \tilde{\tau}(i_k) \tag{7}$$

$$\tilde{\Gamma}_{I_0}(\tilde{\tau}) = \min_{i_k \in I_0} \tilde{\Gamma}_{i_k}(\tilde{\tau}) = \min_{i_k \in I_0} \tilde{\tau}(i_k) = \tilde{\tau}(I_0) \tag{8}$$

In order to measure the interest and accuracy of a fuzzy association rule, we employ a semantic approach based on the evaluation of quantified sentences, using the fuzzy quantifier $Q_M(x) = x$, as follows:

- The support of an itemset I_0 is the evaluation of the quantified sentence Q_M of $\tilde{T}$ are $\tilde{\Gamma}_{I_0}$.
- The support of the fuzzy association rule $I_1 \Rightarrow I_2$ in $\tilde{T}$, $Supp(I_1 \Rightarrow I_2)$, is the evaluation of the quantified sentence Q_M of T are $\tilde{\Gamma}_{I_1 \cup I_2} = Q$ of T are $(\tilde{\Gamma}_{I_1} \cap \tilde{\Gamma}_{I_2})$.
- The confidence of the fuzzy association rule $I_1 \Rightarrow I_2$ in $\tilde{T}$, $Conf(I_1 \Rightarrow I_2)$, is the evaluation of the quantified sentence Q of $\tilde{\Gamma}_{I_1}$ are $\tilde{\Gamma}_{I_2}$.
- The certainty factor is obtained from support and confidence using equation 1.

We evaluate a quantified sentence of the form Q of F are G by means of method GD, defined in [7] as

$$GD_Q(G/F) = \sum_{\alpha_i \in \Lambda(G/F)} (\alpha_i - \alpha_{i+1}) Q\left(\frac{|(G \cap F)_{\alpha_i}|}{|F_{\alpha_i}|}\right) \tag{9}$$

where $\triangle(G/F) = \Lambda(G \cap F) \cup \Lambda(F)$, $\Lambda(F)$ being the level set of F, and $\Lambda(G/F) = \{\alpha_1, ..., \alpha_p\}$ with $\alpha_i > \alpha_{i+1}$ for every $i \in \{1, ..., p-1\}$, and considering $\alpha_{p+1} = 0$. The set F is assumed to be normalized. If not, F is normalized and the same normalization factor is applied to $G \cap F$.

It is possible to employ different fuzzy quantifiers, provided they verify certain properties [6]. We employ the quantifier Q_M since the resulting approach is a generalization of the ordinary association rule assessment framework in the crisp case (i.e., if the set of transactions is crisp, the measures described above yield the ordinary measures for support, confidence, and certainty factor). This is true only for Q_M. Other important properties defining the semantics of this proposal are those of equations 10 and 11.

$$Conf(I_1 \Rightarrow I_2) = 1 \text{ iff } \tilde{\tau}(I_1) \leq \tilde{\tau}(I_2) \quad \forall \tilde{\tau} \in \tilde{T} \tag{10}$$

$$CF(I_1 \Rightarrow I_2) = 1 \text{ iff } Conf(I_1 \Rightarrow I_2) = 1 \tag{11}$$

3.2 Fuzzy gradual dependence

Following the approach in the previous section, and using the quantifier $Q_M(x) = x$, a fuzzy gradual dependence $(*_1, X, A) \to (*_2, Y, B)$ in $\mathcal{D}$ is a fuzzy association rule $[*_1, X, A] \Rightarrow [*_2, Y, B]$ in $\tilde{GT}^{\mathcal{D}}$ that holds

$$supp([*_1, X, A] \Rightarrow [*_2, Y, B]) = \sum_{\alpha_i \in \Lambda\left(\left(\tilde{\Gamma}_{[*_1,X,A]} \cap \tilde{\Gamma}_{[*_2,Y,B]}\right)/\tilde{GT}^{\mathcal{D}}\right)} (\alpha_i - \alpha_{i+1}) \left(\frac{\left| \left(\tilde{\Gamma}_{[*_1,X,A]} \cap \tilde{\Gamma}_{[*_2,Y,B]}\right)_{\alpha_i} \right|}{|\tilde{GT}^{\mathcal{D}}_{\alpha_i}|} \right) \quad (12)$$

$$conf([*_1, X, A] \Rightarrow [*_2, Y, B]) = \sum_{\alpha_i \in \Lambda\left(\tilde{\Gamma}_{[*_1,X,A]}/\tilde{\Gamma}_{[*_2,Y,B]}\right)} (\alpha_i - \alpha_{i+1}) \left(\frac{\left| \left(\tilde{\Gamma}_{[*_1,X,A]} \cap \tilde{\Gamma}_{[*_2,Y,B]}\right)_{\alpha_i} \right|}{\left| \left(\tilde{\Gamma}_{[*_1,X,A]}\right)_{\alpha_i} \right|} \right) \quad (13)$$

with support and confidence given by equations 12 and 13, where $\tilde{\Gamma}_{[*_1,X,A]}$ is a fuzzy subset of transactions such that $\tilde{\Gamma}_{[*_1,X,A]}(gt_{oo'}) = gt_{oo'}([*_1, X, A])$ (similar for $\tilde{\Gamma}_{[*_2,Y,B]}$) and the α_i correspond to the union of the level sets of the fuzzy sets involved, arranged in decreasing order (in equation 13, $\tilde{\Gamma}_{[*_1,X,A]}$ must be normalized, otherwise we should normalize it first and apply the same factor to the intersection $\tilde{\Gamma}_{[*_1,X,A]} \cap \tilde{\Gamma}_{[*_2,Y,B]}$). The certainty factor is obtained as in equation 1.

The following properties from the approach in [4] keep holding:

Proposition 3.1 *Let c be an operator in $\{>, <\}$ such that $c(>) =<$ and $c(<) =>$. Then, $supp([*, X, A]) = supp([c(*), X, A])$.*

Proof: By proposition 2.1, $gt_{oo'}([*, X, A]) = gt_{o'o}([c(*), X, A])$ for every pair $o = (x, y)$, $o' = (x', y')$. Hence,

$$\Lambda\left(\tilde{\Gamma}_{[*,X,A]}/\tilde{GT}^{\mathcal{D}}\right) = \Lambda\left(\tilde{\Gamma}_{[c(*),X,A]}/\tilde{GT}^{\mathcal{D}}\right)$$

and $\forall \alpha_i$

$$\left|\left(\tilde{\Gamma}_{[*,X,A]}\right)_{\alpha_i}\right| = \left|\left(\tilde{\Gamma}_{[c(*),X,A]}\right)_{\alpha_i}\right|$$

so $supp([*, X, A]) = supp([c(*), X, A])$. □

Proposition 3.2 *The generalization to itemsets hold as well, so $supp(\{[*_1, X_1, A_1], \ldots, [*_k, X_k, A_k]\}) = supp(\{[c(*_1), X_1, A_1], \ldots, [c(*_k), X_k, A_k]\})$*

Proof: Same as proposition 3.1. □

Corollary 3.1 *It follows that:*

$supp([*_1, X, A] \Rightarrow [*_2, Y, B]) = supp([c(*_1), X, A] \Rightarrow [c(*_2), Y, B])$,

$conf([*_1, X, A] \Rightarrow [*_2, Y, B]) = conf([c(*_1), X, A] \Rightarrow [c(*_2), Y, B])$,

$CF([*_1, X, A] \Rightarrow [*_2, Y, B]) = CF([c(*_1), X, A] \Rightarrow [c(*_2), Y, B])$.

This last corollary implies that in order to assess all the possible gradual dependencies involving only items of the form $[*_1, X, A]$ and $[*_2, Y, B]$ it is enough to measure support and accuracy for $[<, X, A] \Rightarrow [<, Y, B]$ and $[<, X, A] \Rightarrow [>, Y, B]$.

The following propositions allow us to provide an interpretation of the semantics of our fuzzy gradual dependence and some relation to the approach in [10]:

Proposition 3.3 *$conf([*_1, X, A] \Rightarrow [*_2, Y, B]) = 1$ iff $v_{*_1}(A(x), A(x')) \leq v_{*_2}(B(y), B(y'))$ $\forall o, o' \in \mathcal{D}$*

Proof: $v_{*_1}(A(x), A(x')) \leq v_{*_2}(B(y), B(y'))$ $\forall o, o' \in \mathcal{D}$ iff $gt_{oo'}([*_1, X, A]) \leq gt_{oo'}([*_2, Y, B])$ $\forall gt_{oo'}\tilde{GT}^{\mathcal{D}}$. By equation 10, this is true iff $conf([*_1, X, A] \Rightarrow [*_2, Y, B]) = 1$. □

Proposition 3.4 *$CF([*_1, X, A] \Rightarrow [*_2, Y, B]) = 1$ iff $v_{*_1}(A(x), A(x')) \leq v_{*_2}(B(y), B(y'))$ $\forall o, o' \in \mathcal{D}$*

Proof: Immediate by proposition 3.3 and equation 11. □

Proposition 3.5 *If $CF([*_1, X, A] \Rightarrow [*_2, Y, B]) = 1$ then $A \rightarrow^t B[\alpha, \beta]$ holds with $|\alpha| \geq 1$.*

Proof: Let us consider first the dependence $[<, X, A] \Rightarrow [<, Y, B]$. If $CF([<, X, A] \Rightarrow [<, Y, B]) = 1$ then by proposition 3.4, $v_<(A(x), A(x')) \leq v_<(B(y), B(y'))$ $\forall o, o' \in \mathcal{D}$. As a consequence, $A(x) < A(x')$ implies $A(x') - A(x) \leq B(y') - B(y)$, i.e., $(B(y') - B(y))/(A(x') - A(x)) \geq 1$. Therefore, the slope of all the lines linking pairs of points in the contingency diagram is greater or equal than 1 (no points with membership 0 are considered in the diagram). Hence, the slope of the regression line for all the points is greater or equal than 1.

The proof is similar for the rule $[<, X, A] \Rightarrow [>, Y, B]$, but yielding $(B(y') - B(y))/(A(x') - A(x)) \leq -1$ and hence a slope for the regression line less or equal than

Rule	supp	conf	CF	α	β	R^2
$(>, Weight, High) \rightarrow (>, Speed, Low)$	0.33	0,1	0,06	0.167	0.167	1
$(>, Weight, High) \rightarrow (>, Size, Big)$	0.2	1	1	1.3	-0.67	1

Table 2: Assessment of the gradual dependencies of the example in the introduction using our new approach and that in [10] (values are approximate)

-1. Hence, we have covered all the possibilities and $|\alpha| \geq 1$. □

It is easy to show that the reciprocal of proposition 3.5 holds when $R^2 = 1$, but cannot be guaranteed otherwise.

In order to illustrate these results, let us come back to the example in the introduction. The assessment of the rules (approximate values) is shown in table 2. As expected, the new approach takes into account the variation membership and, instead of yielding two dependencies with confidence and certainty factor equal to one, only in the second case this happens. In fact, the first one has a very low accuracy. Let us remark also that for the second dependence, confidence and certainty factor are one and, at the same time, the slope of the corresponding regression line is greater than 1 (as expected since $R^2 = 1$).

4 Mining gradual dependencies

4.1 Algorithm

In general, the problem we face is that of mining gradual dependencies as association rules in a database $\mathcal{D}$ containing a description of a set of objects in terms of a set of attributes $\{X_1, \ldots, X_m\}$. For each attribute X_i we have a set of n_i fuzzy restrictions defined by fuzzy sets $\{A_{i1}, \ldots, A_{in_i}\}$. We consider a set of items $GI^{\mathcal{D}} = \{[*, X_i, A_{ij}]\}$ with $* \in \{<, >\}$, $i \in \{1, \ldots, m\}$, and $j \in \{1, \ldots, n_i\}$. We shall also consider a bag of fuzzy transactions $\tilde{GT}^{\mathcal{D}}$ containing items of $GI^{\mathcal{D}}$, and obtained from $\mathcal{D}$ as explained in previous sections. Finally, we impose an usual restriction on the rules: no pair of items appearing in the left or right part of a rule can share the same attribute.

A first approach to solve the problem of mining gradual dependencies would be simply to build the set $\tilde{GT}^{\mathcal{D}}$ of transactions and to apply any of the existing algorithms for mining fuzzy association rules. As it is well known, most of the existing algorithms work in two steps: the first one (the most computationally expensive) is to discover the frequent itemsets, i.e., those with support above a minimum user-defined threshold. In the second one, and starting from the frequent itemsets, those rules with enough accuracy are obtained.

The complexity of the second step is not modified as it depends on the number of frequent itemsets, and is not affected by the calculation of the certainty factor. However, the main inconvenience of this approach in our problem is the complexity of discovering the frequent itemsets with respect to the number of objects: while finding frequent itemsets in $\mathcal{D}$ has a complexity $O(n)$ in the number of objects (multiplied by another factors related to number of items and other, depending on the algorithm), finding frequent itemsets in $\tilde{GT}^{\mathcal{D}}$ has a complexity $O(n^2)$.

This problem can be solved to an extent by considering a fixed number of equidistributed levels (degrees) in the definition of the fuzzy sets. In [4] we proposed a solution for the approach presented in that paper (using crisp association rules). With that solution, the complexity of finding the support of itemsets of size p is $n + k^p$. The extent to which this solution is good depends on the relation between n and k^p.

We have developed algorithms based on similar principles for the discovery of fuzzy association rules [6] and fuzzy approximate dependencies [2]. At this moment we are working in an algorithm whose complexity will be $n + k^{p+1}$ for itemsets of size p. We shall describe it in a future paper.

4.2 Experiments

In [4] we performed a small experiment on a real database with data about soils and weather in Granada (southeast of Spain). The data collected in a set of farms includes among others attributes about average temperature, raining and altitude, ph, and percentages of clay and sand in the soil. Fuzzy sets *High*, *Medium* and *Low* have been defined on the domain of each attribute. Our intention is not to discuss the dependencies obtained, but to show the variation between the crisp and fuzzy approaches in a real dataset.

Table 3 shows a set of gradual dependencies obtained that were found interesting by experts, with their support and certainty factor, and the variation in support and certainty factor when using fuzzy rules instead of crisp ones. As expected, in all the cases, support and certainty factor diminished. It is remarkable the case of dependence 5, that is detected using the crisp approach but not by the fuzzy one.

		Crisp		**Fuzzy**	
#	**Dependence**	**supp**	**CF**	**supp**	**CF**
1	$(>, ATEMP, High) \rightarrow (>, ALT, Low)$	0,14	0,93	0,015	0,53
2	$(>, ATEMP, Low) \rightarrow (>, ALT, High)$	0,13	0,93	0.087	0,9
3	$(>, ARAIN, Medium) \rightarrow (>, ATEMP, Medium)$	0,24	0,9	0,1	0,68
4	$(>, ARAIN, Low) \rightarrow (>, ALT, Low)$	0,3	0,85	0,18	0.65
5	$(>, ATEMP, Low) \rightarrow (>, ARAIN, Low)$	0,12	0,84	-	-
6	$(>, CLAY, High) \rightarrow (>, SAND, Low)$	0,09	0,84	0,04	0,64

Table 3: Some gradual dependencies obtained from a real database about soil characteristics and weather in olive cultivation farms, using crisp and fuzzy association rules.

5 Conclusions

We have extended our definition of gradual dependence in [4] in order to incorporate variation strength. For that purpose we have introduced the new notion of degree of variation associated to a pair of objets. We have provided a definition of gradual dependence on the basis of fuzzy association rules over a set of fuzzy transactions obtained from the original dataset by using the degree of variation. We have shown that the new approach is better in capturing the variation strength in gradual dependencies, and we have shown some properties that explain the semantics of the new approach, as well as some results that relate the new approach to the approaches in [10] and [4].

Several research avenues remain open. First, we want to investigate the semantics of fuzzy gradual dependencies obtained by using other approaches to fuzzy association rules, like the measures introduced in [8]. Second, we are working in an algorithm able to reduce the complexity of the mining process when employing existing algorithms for mining fuzzy association rules. Finally, we will apply our techniques to mine for fuzzy gradual dependencies in real databases.

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. Of the 1993 ACM SIGMOD Conference*, pages 207–216, 1993.

[2] F. Berzal, I. Blanco, D. Sánchez, J.M. Serrano, and M.A. Vila. A definition for fuzzy approximate dependencies. *Fuzzy Sets and Systems*, 149(1):105–129, 2005.

[3] F. Berzal, I. Blanco, D. Sánchez, and M.A. Vila. Measuring the accuracy and interest of association rules: A new framework. *Intelligent Data Analysis*, 6:221–235, 2002.

[4] F. Berzal, J.C. Cubero, D. Sánchez, J.M. Serrano, and M.A. Vila. An alternative approach to discover gradual dependencies. *Int. Journal of Uncertainty, Fuzziness and Knowledge-based Systems*. Submitted.

[5] S. Brin, R. Motwani, J.D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. *SIGMOD Record*, 26(2):255–264, 1997.

[6] M. Delgado, N. Marín, D. Sánchez, and M.A. Vila. Fuzzy association rules: General model and applications. *IEEE Transactions on Fuzzy Systems*, 11(2):214–225, 2003.

[7] M. Delgado, D. Sánchez, and M.A. Vila. Fuzzy cardinality based evaluation of quantified sentences. *International Journal of Approximate Reasoning*, 23:23–66, 2000.

[8] D. Dubois, E. Hüllermeier, and H. Prade. A systematic approach to the assessment of fuzzy association rules. *Data Mining and Knowledge Discovery*, 13(2):167–192, 2006.

[9] M. Hahsler and K. Hornik. New probabilistic interest measures for association rules. Technical report, Department of Statistics and Mathematics, Vienna University of Economics and Business Administration, 2006.

[10] E. Hüllermeier. Association rules for expressing gradual dependencies. In *Proceedings PKDD 2002 Lecture Notes in Computer Science 2431*, pages 200–211. 2002.

[11] Chan-Man Kuok, Ada Fu, and Man Hon Wong. Mining fuzzy association rules in databases. *SIGMOD Record*, 27(1):41–46, 1998.

[12] C. Silverstein, S. Brin, and R. Motwani. Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2:39–68, 1998.

[13] T. Sudkamp. Examples, counterexamples, and measuring fuzzy associations. *Fuzzy Sets and Systems*, 149(1):57–71, 2005.