Mathware & Soft Computing 15 (2008) 75-93

Mining Gradual Dependencies with Variation Strength

C. Molina[a], J.M. Serrano[a], D. Sánchez[1,b], M.A. Vila[b]
[a]Department of Informatics, University of Jaén, Spain
{*carlosmo, jschica*}*@ujaen.es*
[b]Department of Computer Science and A.I., University of Granada
C/ Periodista Daniel Saucedo Aranda s/n, 18071 Granada, Spain
{*daniel,vila*}*@decsai.ugr.es*

Abstract

In this paper we propose a definition of gradual dependence as a fuzzy association rule. Gradual dependencies represent tendencies in the variation of the degree of fulfilment of properties in a set of objects. We define the degree of variation of a certain imprecise property for a pair of objects as the difference between their membership degrees to the fuzzy set defining the property. When considering a transaction for every pair of objects and considering items representing positive and negative variations for each property of interest, fuzzy association rules become gradual dependencies and the accuracy and support of the former can be employed to assess the corresponding dependencies. 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.

Keywords: Gradual dependencies; gradual rules; approximate dependencies; association rules.

1 Introduction

Gradual dependencies represent tendencies in the variation of the degree of fulfilment of properties in a set of objects. Mining for gradual dependencies is interesting since they are one of the expressions that humans employ usually to describe their knowledge in a certain field.

The first approach to the definition and assessment of gradual dependencies was proposed in [13]. In this work, the evaluation and representation of gradual dependencies 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

[1]Corresponding author.

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 practice, 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 order to illustrate this, 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. Examples of gradual dependencies are *the higher the weight, the lower the speed*, meaning that as the weight of a truck increases, its speed tends to decrease, and *the higher the weight, the higher the speed*, meaning the opposite tendency.

An alternative, crisp approach to the definition of gradual dependence was introduced in [6]. In this approach 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 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 [13] the relation between the magnitude of variation in both variables is taken into account, whilst in [6] 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

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

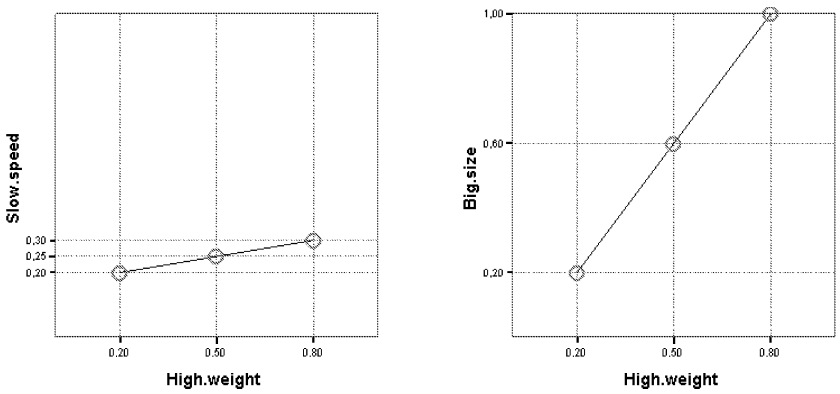

Figure 1: Contingency diagrams for the gradual dependencies *the higher the weight, the lower the speed* and *the higher the weight, the bigger the size* from the data in table 1

the higher the weight, the lower the speed and *the higher the weight, the bigger the size* using the approaches in [13] and [6]. Using [6], 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 [6] 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 [4]. We also explore the relationship to the approach in [13], both

from a theoretical and an experimental point of view.

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 [9]. Section 4 is devoted to mining issues and to show some experiments. Finally, section 5 contains our conclusions and future research.

2 Gradual dependencies with variation strength

In this section we extend our definition of gradual dependence [6] in order to incorporate variation strength in the assessment. First we briefly recall the definition in [6]. Then we define the concept of variation, and use it to extend the definition in [6] by using fuzzy association rules.

2.1 Our previous approach

In [6], 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 [7, 20, 5, 12]. In this paper we shall employ Shortliffe and Buchanan's *certainty factors*, as proposed in [5]. 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.

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

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 [5]. 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

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

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

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

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.

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.

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

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:

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

where

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

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 [6], 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 $(*_1, X, A) \rightarrow (*_2, Y, B)$ 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 [6] 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 [15, 9, 21, 11]).

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* $(*_1, X, A) \rightarrow (*_2, Y, B)$ *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 $(*_1, X, A) \rightarrow (*_2, Y, B)$.

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 [9] to obtain a particular definition of fuzzy gradual dependence.

3.1 Our approach to fuzzy association rules

In [9], 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 [10] 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 [9]. 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 $(*_1, X, A) \rightarrow (*_2, Y, B)$ in $\mathcal{D}$ is a fuzzy association rule $[*_1, X, A] \Rightarrow [*_2, Y, B]$ in $\tilde{G}T^{\mathcal{D}}$ that holds with support and confidence given by equations 12 and 13, where $\tilde{\Gamma}_{[*_1,X,A]}$ is a fuzzy subset of transactions such that $\tilde{\Gamma}_{[*_1,X,A]}(gt_{oo'}) = gt_{oo'}(\lfloor *_1, X, A \rfloor)$ (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.

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

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

$$conf(\lfloor *_1, X, A \rfloor \Rightarrow \lfloor *_2, Y, B \rfloor) =$$

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

The following properties from the approach in [6] keep holding:

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

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,

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

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 it holds that* $supp(\{[*_1, X_1, A_1], \ldots, [*_k, X_k, A_k]\}) = supp(\{[c(*_1), X_1, A_1], \ldots, [c(*_k), X_k, A_k]\})$

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 [13]:

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 $gt_{oo'}([*_1, X, A]) \leq gt_{oo'}([*_2, Y, B])\ \forall gt_{oo'} GT^{\mathcal{D}}$. 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* $A \rightharpoonup^t B[\alpha, \beta]$ *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 -1. Hence, we have covered all the possibilities and $|\alpha| \geq 1$. □

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 [13] (values are approximate)

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 a certain extent by considering a fixed number k of equidistributed levels (degrees) in the definition of the fuzzy sets. In [6] 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^2 and k^p.

We have developed algorithms based on similar principles for the discovery of fuzzy association rules [9] and fuzzy approximate dependencies [4]. Taking ideas from all these algorithms, we propose algorithm 1 in order to obtain the support of itemsets for mining fuzzy gradual dependencies. Its complexity is $O(n+pk^(p+2))$. Again, this solution is good to the extent that $pk^{p+2} < n^2$.

4.2 Experimental results

The following section is devoted to the description of the data sources employed in our experimentation as well as the discussion of results.

4.2.1 Soil database

We applied our techniques on a database containing soil information from the South and Southeast of the Iberian Peninsula under Mediterranean climate: Sierra Nevada, Sierra of Gádor and Southeast (involving part of the provinces of Murcia and Almería). Data was extracted from two Ph.D. Thesis and five cartographic sheets from LUCDEME, scale 1:100000 ([16], [8], [3], [17], [2], [18] and [19]). This database includes, among others, numeric attributes describing soil features as average temperature, raining and altitude, PH, and percentages of clay, sand, and chemical components in soil. The underlying numeric domain was fuzzified into a set of linguistic labels, $\{High, Medium, Low\}$, according to expert criteria.

Both crisp and fuzzy gradual dependencies were computed. Table 3 shows some of the most interesting rules involving only one attribute as antecedent and as consequent, according to a certainty factor threshold ($CF > 0.60$), fixed by the experts. For each rule, support (in %) and CF are shown, in both cases, crisp and fuzzy. For this particular case, both crisp and fuzzy results seem to be similar, nonetheless, support for fuzzy gradual dependencies use to be higher than in the crisp case, allowing a higher number of rules to be extracted.

Additionally, table 4 shows those fuzzy rules with a higher certainty factor, involving two items in the antecedent. In both tables, 3 and 4, for summarizing reasons, operator $*$ stands for operator $>$ (resp. $<$), as well as operator $C(*)$ stands for $<$ (resp. $>$).

Some interesting relations can be seen in table 3. Let us pay attention, for example, to the relation *'The more the average temperature is Low, the more the average altitude is High'* (rule #2) that holds with a CF over 0.9, and looks clearly reasonable.

Also, a clear opposite relation between average temperature and average rainfall can be seen in rules #1, #7 and #14, as the lower the former the higher the latter (and viceversa).

Algorithm 1 Algorithm for measuring the support of itemsets for mining fuzzy gradual dependencies

$V(K \times K \times ... \times K) = 0$
$i \leftarrow 1$
while $i \leq p$ **do**
  $V(\lfloor A_1(x_i) \times K \rfloor, \lfloor A_2(x_i) \times K \rfloor, ..., \lfloor A_p(x_i) \times K \rfloor) + +$
end while
$\widetilde{Supp} \leftarrow \emptyset$
$i \leftarrow k$
while $i > 0$ **do**
  $elem \leftarrow 0$
  $\overrightarrow{j} \leftarrow 0$
  $d \leftarrow 1$
  while $d \geq 0$ **do**
    $j(d) + +$
    if $j(d) > K$ **then**
      $j(d) \leftarrow 0$
      $d \leftarrow d - 1$
    else if $d = K$ **then**
      $elem \leftarrow elem V(\overrightarrow{j}) \times V(j(1) + i, ..., j(N) + i)$
      $n \leftarrow i$
      while $n \leq K$ **do**
        $l \leftarrow 1$
        while $l \leq p$ **do**
          $\overrightarrow{m} \leftarrow (j(1) + i, ..., j(p) + i)$
          $h \leftarrow 1$
          while $h > 0$ **do**
            $m(h) + +$
            if $m(h) > K$ **then**
              $m(h) \leftarrow j(h) - 1$
              $h \leftarrow h - 1$
              if $h = l$ **then**
                $h \leftarrow h - 1$
              end if
            else if $h = p$ **then**
              $elem \leftarrow elem + V(\overrightarrow{j}) \times V(\overrightarrow{m})$
            else
              $h \leftarrow h + 1$
              if $h = l$ **then**
                $h \leftarrow h + 1$
              end if
            end if
          end while
        end while
        $n \leftarrow n + 1$
      end while
    else
      $d \leftarrow d + 1$
    end if
  end while
  $\widetilde{Supp} \leftarrow \widetilde{Supp} \cup \{\frac{i}{k}/elem\}$
  $i \leftarrow i - 1$
end while
return $GD_Q(\widetilde{Supp}/\widetilde{GT}^D)$

	Crisp		Fuzzy	
Rule	**s(%)**	**CF**	**s(%)**	**CF**
[, AvgTemp, 'Low'] → [*, AvgRainfall, 'High']*	6.023	0.974	8.727	0.907
[, AvgTemp, 'Low'] → [*, AvgAltitude, 'High']*	5.796	0.931	8.754	0.904
[, PH, 'Low'] → [*, AvgRainfall, 'High']*	2.332	0.796	4.403	0.773
[, AvgRainfall, 'High'] → [*, AvgAltitude, 'High']*	7.860	0.874	9.025	0.765
[, AvgAltitude, 'Low'] → [*, AvgRainfall, 'Low']*	7.073	0.748	18.394	0.754
[, PH, 'Low'] → [*, AvgTemp, 'Low']*	2.021	0.686	4.303	0.749
[, AvgRainfall, 'High'] → [*, AvgTemp, 'Low']*	6.023	0.660	8.727	0.741
[, PH, 'Low'] → [*, AvgAltitude, 'High']*	2.199	0.736	4.118	0.709
[, FE, 'High'] → [*, % Sand, 'Low']*	4.602	0.593	7.776	0.701
[, AvgTemp, 'High'] → [*, AvgRainfall, 'Low']*	5.344	0.624	18.336	0.693
[, AvgRainfall, 'Med.'] → [*, AvgTemp, 'Med.']*	1.014	0.420	10.222	0.683
[, AvgAltitude, 'Med.'] → [*, AvgTemp, 'Med.']*	1.825	0.529	9.454	0.682
[, AvgRainfall, 'Low'] → [*, AvgAltitude, 'Low']*	7.073	0.539	18.394	0.659
[, AvgRainfall, 'Low'] → [*, AvgTemp, 'High']*	5.344	0.390	18.336	0.656
[, % Clay, 'High'] → [*, % Sand, 'Low']*	3.216	0.573	4.848	0.640
[, % Sand, 'High'] → [*, % Clay, 'Low']*	7.790	0.540	12.537	0.613
[, AvgRainfall, 'High'] → [C(*), PH, 'High']*	6.333	0.629	7.858	0.601

Table 3: (Fuzzy) Gradual Dependencies from Soil databases (Fuzzy CF > 0.60)

Rule	**supp.(%)**	**CF**
[[, AvgRainfall, 'Med.'], [*, AvgTemp, 'Low']] → [*, AvgAltitude, 'High']*	0.177	0.995
[[C(*), AvgRainfall, 'Low'], [C(*), Fe, 'Low']] → [C(*), AvgTemp, 'High']	0.421	0.995
[[C(*), AvgTemp, 'Low'], [, % Clay, 'High']] → [C(*), AvgRainfall, 'High']*	0.698	0.994
[[, AvgRainfall, 'Med.'], [C(*), AvgAltitude, 'High']] → [*, AvgTemp, 'Med.']*	1.487	0.993
[[C(*), AvgTemp, 'Low'], [C(*), Carbonates, 'High']] → [C(*), AvgRainfall, 'High']	0.129	0.993
[[, AvgTemp, 'Low'], [*, AvgAltitude, 'Med.']] → [*, AvgRainfall, 'High']*	0.145	0.992

Table 4: Fuzzy Gradual Dependencies involving two itemsets in the antecedent (soil data)

Finally, a similar opposite relation can be found between percentages of sand and clay (rules #16 and #17), a fact that experts can explain as it is easier to find a higher percentage of clay than of sand in wet soils, and viceversa, it is easier to have a higher percentage of sand than of clay in dry soils.

For the rules of table 3, we have computed the regression parameters α, β and R^2 according to [13]. The results are shown in table 5. However, we cannot compare the results from our approach with the one proposed in [13] because most of the gradual dependencies do not have a high enough value of R^2 to consider $[\alpha, \beta]$ representative (only 76 gradual dependencies have $R^2 > 0.5$). Lineal regression is not able to fit well to the data dispersion in this situation.

Keeping in mind this fact, if we only consider the gradual dependencies when $R^2 > 0$ (figure 2) we can appreciate that for CF values near zero, α values are near this point. Elsewhere in the range of alpha values are more scattered, although there is some tendency to increase along with CF when it presents values greater than 0.5. In the case of R^2 the behavior is very similar (figure 3).

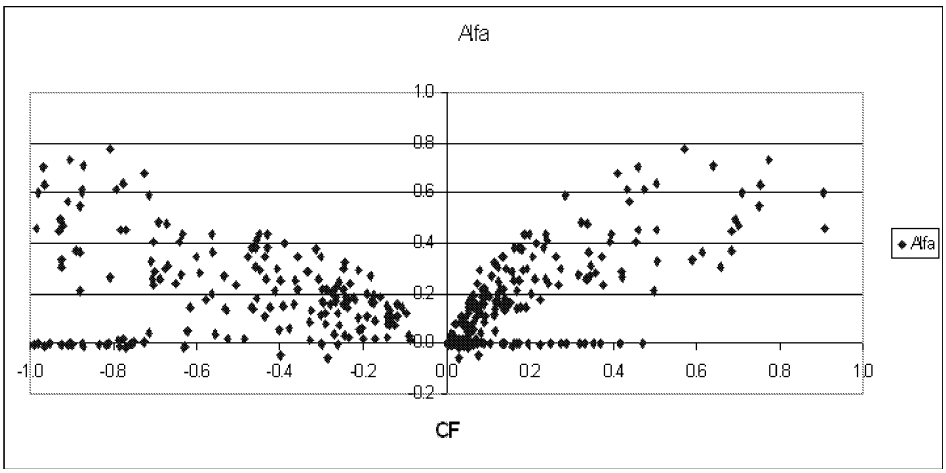

Figure 2: α values and associated CF when $R^2 > 0$

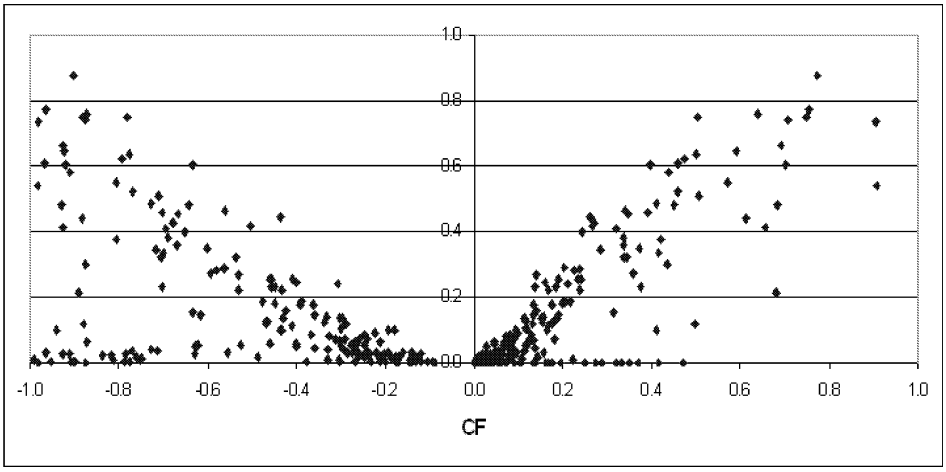

Figure 3: R^2 values and associated CF when $R^2 > 0$

Rule	α	β	R^2
[, AvgTemp, 'Low'] → [*, AvgRainfall, 'High']*	0.459	0.623	0.541
[, AvgTemp, 'Low'] → [*, AvgAltitude, 'High']*	0.601	0.477	0.736
[, PH, 'Low'] → [*, AvgRainfall, 'High']*	0.732	0.184	0.879
[, AvgRainfall, 'High'] → [*, AvgAltitude, 'High']*	-0.741	1.767	-3.408
[, AvgAltitude, 'Low'] → [*, AvgRainfall, 'Low']*	0.631	0.389	0.773
[, PH, 'Low'] → [*, AvgTemp, 'Low']*	0.549	0.365	0.752
[, AvgRainfall, 'High'] → [*, AvgTemp, 'Low']*	-0.678	1.610	-3.367
[, PH, 'Low'] → [*, AvgAltitude, 'High']*	0.602	0.364	0.739
[, FE, 'High'] → [*, % Sand, 'Low']*	0.469	0.539	0.606
[, AvgTemp, 'High'] → [*, AvgRainfall, 'Low']*	0.495	0.466	0.665
[, AvgRainfall, 'Med.'] → [*, AvgTemp, 'Med.']*	0.446	0.543	0.483
[, AvgAltitude, 'Med.'] → [*, AvgTemp, 'Med.']*	0.373	0.632	0.212
[, AvgRainfall, 'Low'] → [*, AvgAltitude, 'Low']*	-0.021	0.868	-0.391
[, AvgRainfall, 'Low'] → [*, AvgTemp, 'High']*	0.303	0.523	0.415
[, % Clay, 'High'] → [*, % Sand, 'Low']*	0.710	0.244	0.758
[, % Sand, 'High'] → [*, % Clay, 'Low']*	0.366	0.606	0.440
[, AvgRainfall, 'High'] → [C(*), PH, 'High']*	0.562	-0.601	-1.833

Table 5: Regression parameters following the approach in [13] for the Gradual Dependencies in table 3

4.2.2 STULONG database

On the other hand, we have employed as data source a medical database, available for these data mining purposes. The study (STULONG) was realized at the 2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital, U nemocnice 2, Prague 2 (head. Prof. M. Aschermann, MD, SDr, FESC), under the supervision of Prof. F. Boudík, MD, ScD, with collaboration of M. Tomečková, MD, PhD and Ass. Prof. J. Bultas, MD, PhD. The data were transferred to electronic form by the European Center of Medical Informatics, Statistics and Epidemiology of Charles University and Academy of Sciences (head. Prof. RNDr. J. Zvárová, DrSc). The data resource is on the web pages http://euromise.vse.cz/challenge2003. At present time the data analysis is supported by the grant of the Ministry of Education CR Nr LN 00B 107.

The STULONG database includes a twenty years lasting longitudinal study of the risk factors of the atherosclerosis in a population of 1417 middle aged men. It consists of several data files, but we have centered on only one of them, named *Entry*, which contains information obtained from entry examinations. The whole set of attributes describes Social characteristics, Physical activity, Smoking and Drinking habits, Physical examination, Biochemical examination, and Risk factors, among others.

Categorical attributes were unchanged, but, as in the previous case, we fuzzified numeric attributes, defining again a set of linguistic labels, $\{High, Medium, Low\}$. In order to accomplish this, we discretized the numeric domains into sets of equi-depth intervals (see [14] for further explanations on the procedure), which were later transformed into fuzzy sets, as trapezoidal distributions, relaxing the boundaries between them.

For this data source, we reduced the experimental results to fuzzy gradual

dependencies only. Tables 6 and 7 shows several gradual dependencies obtained from this dataset. For space saving purposes, we have included only those rules with $CF > 0.60$. Again, we note $*$ instead of $<$ (resp. $>$) and $C(*)$ instead of $>$ (resp. $<$).

We can observe some totally accurate ($CF = 1$) gradual dependencies, as between attributes *Intensity of Smoking* and *How Long have been Smoking*. In general, most of the rules shown in the table seem to be reasonable as, for example, those relating drinking habits with the quantity of beverage (wine, liquor).

Also, it seems to appear a direct relation between the different types of blood pressure (I and II, and systolic and diastolic), as the variation in one attribute is closely related to the variation, in the same direction, in the other attribute. Anyway, a further and deeper study, involving additional techniques, should be necessary to gain more information, as for this particular case, gradual dependencies act as an exploratory technique.

Rule	**supp.(%)**	**CF**
[, IntensOfSmoking, 'High'] → [*, HowLongSmoking, 'High']*	19.723	1.000
[, BloodPressIIDiast, 'High'] → [*, BloodPressIDiast, 'High']*	11.156	0.766
[, BloodPressISyst, 'Low'] → [*, BloodPressIISyst, 'Low']*	17.743	0.718
[, BloodPressIDiast, 'High'] → [*, BloodPressIIDiast, 'High']*	11.156	0.703
[, BloodPressISyst, 'High'] → [*, BloodPressIISyst, 'High']*	13.832	0.702
[, BloodPressIIDiast, 'Low'] → [*, BloodPressIDiast, 'Low']*	19.076	0.686
[, BloodPressIDiast, 'Low'] → [*, BloodPressIIDiast, 'Low']*	19.076	0.686
[, BloodPressIISyst, 'Low'] → [*, BloodPressISyst, 'Low']*	17.743	0.664
[, BloodPressIDiast, 'High'] → [*, BloodPressISyst, 'High']*	10.241	0.614

Table 6: Fuzzy Gradual Dependencies from STULONG data (CF > 0.60)

Rule	**supp.(%)**	**CF**
[[C(*), BloodPressISyst, 'Low'], [C(*), BloodPressIIDiast, 'Low']] → [C(*), BloodPressIISyst, 'Low']	11.771	0.877
[[, BloodPressIDiast, 'Low'], [*, BloodPressIISyst, 'Low']] → [*, BloodPressIIDiast, 'Low']*	12.232	0.847
[[, BloodPressISyst, 'Low'], [*, BloodPressIIDiast, 'Low']] → [*, BloodPressIDiast, 'Low']*	11.195	0.816
[[, BloodPressIDiast, 'Low'], [C(*), BloodPressIISyst, 'High']] → [*, BloodPressIIDiast, 'Low']*	10.991	0.806
[[C(*), BloodPressISyst, 'Low'], [C(*), BloodPressIDiast, 'Low']] → [C(*), BloodPressIISyst, 'Low']	11.437	0.800
[[, BloodPressIISyst, 'High'], [C(*), BloodPressIIDiast, 'Low']] → [C(*), BloodPressIDiast, 'Low']*	10.991	0.799
[[, BloodPressIDiast, 'Low'], [*, BloodPressIISyst, 'Low']] → [*, BloodPressISyst, 'Low']*	11.437	0.778
[[C(*), BloodPressISyst, 'Low'], [C(*), BloodPressIDiast, 'Low']] → [C(*), BloodPressIIDiast, 'Low']	11.195	0.773
[[, BloodPressIISyst, 'Low'], [*, BloodPressIIDiast, 'Low']] → [*, BloodPressIDiast, 'Low']*	12.232	0.745
[[, BloodPressIISyst, 'Low'], [*, BloodPressIIDiast, 'Low']] → [*, BloodPressISyst, 'Low']*	11.771	0.714

Table 7: Fuzzy Gradual Dependencies involving two itemsets in the antecedent (STULONG data)

The comparison with the approach in [13] yields results similar to those obtained for the soil database.

5 Conclusions

We have extended our definition of gradual dependence in [6] 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 [13] and [6].

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 [11]. 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.

Acknowledgements

This work has been supported by the Spanish Ministerio de Educación y Ciencia under the project grants TIN2006-07262 and TIN2006-15041-C04-01.

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] J. Alias. *Mapa de suelos de Mula. Mapa 1:100000 y memoria.* LUCDEME; MAPA-ICONA-University of Murcia, 1986.

[3] J. Alias. *Mapa de suelos de Cehegin. Mapa 1:100000 y memoria.* LUCDEME; MAPA-ICONA-University of Murcia, 1987.

[4] 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.

[5] 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.

[6] 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*, 15(5):559 – 570, 2007.

[7] 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.

[8] G. Delgado, R. Delgado, E. Gamiz, J. Párraga, M. Sánchez Marañon, J. Medina, and J.M. Martín-García. *Mapa de Suelos de Vera.* LUCDEME, ICONA-Universidad de Granada, 1991.

[9] 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.

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

[11] 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.

[12] 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.

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

[14] F. Hussain, H. Liu, C.L. Tan, and M. Dash. Discretization: An Enabling Technique. Technical Report, The National University of Singapore, June 1999.

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

[16] C. Oyonarte. *Estudio Edáfico de la Sierra de Gádor (Almería). Evaluación para usos forestales.* PhD thesis, University of Granada, 1990.

[17] A. Pérez Pujalte. *Mapa de suelos de Tabernas. Mapa 1:100000 y memoria.* LUCDEME; MAPA-ICONA-CSIC, 1987.

[18] A. Pérez Pujalte. *Mapa de suelos de Sorbas. Mapa 1:100000 y memoria.* LUCDEME; MAPA-ICONA-CSIC, 1989.

[19] M. Sánchez-Marañón. *Los suelos del Macizo de Sierra Nevada. Evaluación y capacidad de uso (in Spanish).* PhD thesis, University of Granada, 1992.

[20] 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.

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