Treatment of the incomplete information in L-Fuzzy contexts*

Cristina Alcalde
Dp. Matemática Aplicada
Universidad del País Vasco
Plaza de Europa,1
20018 San Sebastián
mapalvac@sp.ehu.es

Ana Burusco
Dp. Automática y Computación
Universidad Pública de Navarra
Campus de Arrosadía
31006 Pamplona
burusco@si.unavarra.es

Ramón Fuentes-González
Dp. Automática y Computación
Universidad Pública de Navarra
Campus de Arrosadía
31006 Pamplona
rfuentes@si.unavarra.es

Abstract

The goal of this work is to find some tools that allow us to extract information from an interval-valued L-Fuzzy context in which there are unknown values.
Using the idea of frequent sets that appears in data mining we will give the definition of a frequent interval-valued L-fuzzy set and we will use this definition to decide when we can eliminate an object or an attribute of the context with some absent values.
Next, we will set up the implications between intervals those will allow us to replace the absent values and to limit, as much as possible, the lack of information in the context.
Finally, we will illustrate these results with an example.
Keywords: L-Fuzzy context, L-Fuzzy concept, frequent set, association rules, implications between attributes.

1 Introduction. Interval-valued L-Fuzzy context

The L-Fuzzy contexts have been studied [3, 7, 2] as a generalization of the formal concept theory of R.Wille [9].

In order to extract knowledge from a table with incomplete and ambiguous information, we will use the interval-valued L-Fuzzy context as an extension of the L-Fuzzy context given by an implication operator [3, 4] to the interval-valued case.

Definition 1 An *interval-valued L-fuzzy context* is a tuple $(\mathcal{J}[L], X, Y, R)$, with X and Y two finite sets (of *objects* and *attributes*) and R an interval-valued L-fuzzy relation between X and Y.

Definition 2 Let $\boldsymbol{I}$ be an interval-valued fuzzy implication defined on the lattice of intervals $(\mathcal{J}[L], \leq)$. Given $A \in \mathcal{J}[L]^X$ and $B \in \mathcal{J}[L]^Y$ two interval-valued L-fuzzy sets, the *derived sets* of A and B, denoted respectively by $A_1 \in \mathcal{J}[L]^Y$ and $B_2 \in \mathcal{J}[L]^X$, are defined as:

$$A_1(y) = \inf_{x \in X} \{\boldsymbol{I}(A(x), R(x, y))\}$$
$$B_2(x) = \inf_{y \in Y} \{\boldsymbol{I}(B(y), R(x, y))\}$$

(A_1 is the direct image of A under the relation R, and B_2 the inverse image of B under R.)

These operators are the extensions of the *derivation operator* $*$ of R.Wille [9] to our situation.

Definition 3 If M is a fixed point of the operator $\boldsymbol{\varphi}$ defined as $\boldsymbol{\varphi}(A) = (A_1)_2$, then (M, M_1) is an *interval-valued L-fuzzy concept* of the interval-valued L-fuzzy context $(\mathcal{J}[L], X, Y, R)$.

The set of interval-valued L-fuzzy concepts with the order relation $\preceq$ defined as $(A, B) \preceq (C, D) \Leftrightarrow A \leq C$ is a complete lattice.

2 Contexts with absent values

When we attempt to study a context where there are some unknown values in the relation $R(x, y)$,

*This work has been supported by the project TIC 2003-08693 of the Ministry of Science and Technology

we can treat those *absent values* in two different ways, using the cleaning and transformation techniques that appear in data mining [6]:

Elimination. We will eliminate the object or the attribute corresponding to an absent value when it is not a *frequent* object or attribute.

Replacement. We will choose a value to represent the lack of information and we will use the implications between attributes to try to predict the absent value.

3 Elimination of the absent values

In order to define the frequent objects and attributes, we are going to use the idea of *frequent sets* [1, 6] that appears in data mining.

Given a threshold that we will call minimum support, a set is said to be *frequent* if its support is greater or equal than this minimum support, understanding by support the quotient between the number of times that the elements of the set appear divided by the possible maximum number.

Extending this idea to the concept theory, given a formal context (G, M, I), a frequent attribute is defined as:

Definition 4 (Stumme et al. [8]) Let $B \subseteq M$ and suppmin $\in [0, 1]$. The *support* of B in the context is the number (belonging to [0,1])

$$\text{supp}(B) = \frac{|B^*|}{|G|}$$

where B^* is the derived set of B. The set B is said to be a *frequent* attribute set if $\text{supp}(B) \geq$ suppmin.

We have adapted this definition to the case of L-Fuzzy and interval-valued L-Fuzzy contexts with the purpose of defining in which cases it is possible to eliminate an absent value.

From this point we will use $L = \mathcal{C} = \{0, 0.1 \ldots 0.9, 1\}$ and we will represent the absent values of R by elements belonging to $\mathcal{J}[\mathcal{C}]$.

Definition 5 [5] Let (L, X, Y, R) be an L-fuzzy context. Given $A \in L^X$, we define the *support of* A as the quotient (belonging to [0,1])

$$\text{supp}(A) = \frac{\sum_{y \in Y} A_1(y)}{|Y|}$$

In the same way, given $B \in L^Y$, we define the *support of* B as

$$\text{supp}(B) = \frac{\sum_{x \in X} B_2(x)}{|X|}$$

The L-Fuzzy set A (and similarly B) will be a *frequent* set when given a minimum support suppmin $\in [0, 1]$, the support of A will be greater or equal than suppmin.

Definition 6 Let $(\mathcal{J}[L], X, Y, R)$ be an interval-valued L-Fuzzy context. Given $A \in \mathcal{J}[L]^X$ an interval-valued L-Fuzzy set, we define the *support* of A as the interval (belonging to $\mathcal{J}$[0,1])

$$\text{supp}(A) = \left[\frac{\sum_{y \in Y} l_{A_1}(y)}{|Y|}, \frac{\sum_{y \in Y} u_{A_1}(y)}{|Y|} \right]$$

where the membership function of the set A_1 is given by $A_1(y) = [l_{A_1}(y), u_{A_1}(y)]$.

In the same way, we define the support of the interval-valued L-Fuzzy set $B \in \mathcal{J}[L]^Y$.

Definition 7 Fixed a value suppmin $\in [0, 1]$, that represents a minimum support and given an interval-valued L-Fuzzy set $A \in \mathcal{J}[L]^X$ (or similarly $B \in \mathcal{J}[L]^Y$), we will say that A is a *frequent* set if

$$[\text{suppmin}, \text{suppmin}] \leq \text{supp}(A)$$

(or similarly $\leq \text{supp}(B)$).

Definition 8 An object (or an attribute) of an interval-valued L-Fuzzy context is said to be *frequent* when the interval-valued L-Fuzzy set that represents it is a frequent set.

As we have said before, if one of the values of the relation $R(x, y)$ is unknown, we will eliminate the corresponding attribute or object if it is not frequent. However, to set up if an object or attribute is frequent, we need to know all the values of the relation.

In order to decide if we can eliminate the object or the attribute corresponding to an absent value, we will replace this absent value by $[1,1]$ and we will calculate the support with the new table, that will give us a value bigger or equal than the real support. We will eliminate the corresponding object or attribute if this obtained value is less than the fixed minimum support.

Example.

Let us suppose that we want to study the ingredients of a certain alimentary product for the main trademarks of this product. The information that we have is the approximate quantity of each ingredient that the manufacturer declares for 100 grammes of product.

Table 1: Quantity of the ingredient Ii in the trademark Mj

	$I1$	$I2$	$I3$	$I4$
$M1$	[20,25]	2	10	[10,15]
$M2$	20		5	15
$M3$	25	12	10	10
$M4$	22	10	[2,5]	
$M5$	20	5	[5,10]	5

Let us consider an interval-valued L-Fuzzy context $(\mathcal{J}[L], X, Y, R)$, with $L = \mathcal{C}$, where the set of objects, X, is the set of trademarks and the set of attributes and Y, the set of ingredients. The relation R is obtained transforming the table of data into a table of closed intervals and dividing each element by the greatest of the elements of the table, making the necessary rounds to obtain values that are members of L. The obtained interval-valued L-fuzzy relation is shown in the Table 2.

Table 2: Values of the relation R

R	$I1$	$I2$	$I3$	$I4$
$M1$	$[0.8, 1]$	$[0.1, 0.1]$	$[0.4, 0.4]$	$[0.4, 0.6]$
$M2$	$[0.8, 0.8]$		$[0.2, 0.2]$	$[0.6, 0.6]$
$M3$	[1,1]	[0,0]	$[0.4, 0.4]$	$[0.4, 0.4]$
$M4$	$[0.9, 0.9]$	$[0.2, 0.2]$	$[0.1, 0.2]$	
$M5$	$[0.8, 0.8]$	$[0, 0.2]$	$[0.2, 0.4]$	$[0.2, 0.2]$

In order to study if it is possible to eliminate the row or the column corresponding to the absent values, we are going to consider, for instance, the minimum support suppmin $= 0.5$ and we will eliminate the object or the attribute corresponding to an unknown value if it is not frequent, that is, if its support is not greater or equal than the minimum support. We will calculate the support of the objects $M2$ and $M4$ and the attributes $I2$ and $I4$ with the new relation $\widehat{R}$ in which we have replaced the absent values by $[1,1]$.

Table 3: Values of the new relation $\widehat{R}$

$\widehat{R}$	$I1$	$I2$	$I3$	$I4$
$M1$	$[0.8, 1]$	$[0.1, 0.1]$	$[0.4, 0.4]$	$[0.4, 0.6]$
$M2$	$[0.8, 0.8]$	$[1, 1]$	$[0.2, 0.2]$	$[0.6, 0.6]$
$M3$	[1,1]	[0,0]	$[0.4, 0.4]$	$[0.4, 0.4]$
$M4$	$[0.9, 0.9]$	$[0.2, 0.2]$	$[0.1, 0.2]$	$[1, 1]$
$M5$	$[0.8, 0.8]$	$[0, 0.2]$	$[0.2, 0.4]$	$[0.2, 0.2]$

The interval-valued L-fuzzy set associated to the object $M2$ is

$$M2 = \{M1/[0,0], M2/[1,1], M3/[0,0], M4/[0,0], M5/[0,0]\}$$

Using the Lukasiewicz R-implication (we can use another implication but the results are similar), we have obtained the derived set:

$$(M2)_1 = \{I1/[0.8, 0.8], I2/[1,1], I3/[0.2, 0.2], I4/[0.6, 0.6]\}$$

and then the support, $\text{supp}(M2) = \left[\frac{2.6}{4}, \frac{2.6}{4}\right]$.

With the fixed minimum support 0.5, $M2$ will be a *frequent* object.

Similarly, we can calculate the support of the interval-valued sets associated to the trademark $M4$ and to the ingredients $I2$ and $I4$, and we can say that with the chosen minimum support, the only set that is not frequent is the associated to $I2$ ($\text{supp}(I2) = \left[\frac{1.3}{5}, \frac{1.5}{5}\right]$), therefore that is the only element that we could eliminate.

We can eliminate more objects or attributes if we choose a higher value for the minimum support. For instance, if we chose the value suppmin $= 0.55$ in the previous case, we will eliminate $I4$ in addition to $I2$. However, we must take into account that the elimination of objects or attributes implies loss of information, an so, it is not advisable to work with very high minimum supports.

4 Replacement of the absent values

Once all the non frequent objects and attributes with absent values have been eliminated, our next task will be to choose the way to represent those unknown elements by means of a value that allows us to work with them.

We can consider the *lack of information* as the total imprecision, and so, we can represent it by the interval $[0,1]$.

The use of the interval $[0,1]$ to represent the unknown values is going to give us, in general, values with a great width of interval, because when the information that we have is not complete, then the knowledge that we can extract will have a certain degree of ambiguity. In any case, we will try to reduce to the maximum this ambiguity by means of the implications between attributes.

4.1 Implications between attributes of an interval-valued L-Fuzzy context.

Some important tools for the extraction of knowledge are the *association rules.* We are going to be able to set up the implications between attributes by means of them.

The association rule idea was first presented by Agrawal et al. in [1] as an implication $B \Rightarrow C$, that is verified with a certain support and degree of confidence, where $B, C \subseteq M$ are two sets of attributes called the *antecedent* and the *consequent* respectively.

The *support* of a rule $B \Rightarrow C$ is the percentage of times that the rule predicts correctly, that is, the percentage of times that we have the attributes of B and those of C.

$$\mathrm{supp}(B \Rightarrow C) = \mathrm{supp}(B \cup C)$$

The *confidence* of a rule is the percentage of times that the rule is fulfilled when it is possible to be applied, that is, the percentage of times that containing the attributes of B also have those of C.

$$\mathrm{conf}(B \Rightarrow C) = \frac{\mathrm{supp}(B \cup C)}{\mathrm{supp}(B)}$$

Using the association rules and the definition of support of an L-Fuzzy set, we have defined in [5] the support and the degree of confidence of an implication between attributes as follows.

Definition 9 Let (L, X, Y, R) an L-fuzzy context and $B, C \subseteq L^Y$ two sets of attributes. The *support* of the implication $B \Rightarrow C$ is given by

$$\mathrm{supp}(B \Rightarrow C) = \mathrm{supp}(B \cup C) = \frac{\sum_{x \in X} (B \cup C)_2(x)}{|X|}$$

and represents the percentage of objects that share the attributes of B and C.

The *confidence* of the implication is

$$\mathrm{conf}(B \Rightarrow C) = \frac{\mathrm{supp}(B \cup C)}{\mathrm{supp}(B)} = \frac{\sum_{x \in X} (B \cup C)_2(x)}{\sum_{x \in X} B_2(x)}$$

and represents the percentage of objects that verify the implication, that is, the percentage of objects that having the attributes of B in a certain degree also have those of C in the same degree.

Using the definition of support of an interval-valued L-fuzzy set, we can extend the definition of support and confidence of an implication between attributes to the interval-valued case.

Definition 10 Let $(\mathcal{J}[L], X, Y, R)$ be an interval-valued L-Fuzzy context. Given the attribute sets $B, C \in \mathcal{J}[L]^Y$, we define the *support* of the implication $B \Rightarrow C$, $\mathrm{supp}(B \Rightarrow C)$, as the interval

$$\left[\frac{\sum_{x \in X} l_{(B \cup C)_2}(x)}{|X|}, \frac{\sum_{x \in X} u_{(B \cup C)_2}(x)}{|X|} \right]$$

where the membership function of the derived set B_2 is $B_2(x) = [l_{B_2}(x), u_{B_2}(x)]$

and the *confidence* of the implication, $\mathrm{conf}(B \Rightarrow C)$, is given by the interval

$$\left[\frac{\sum_{x\in X} l_{(B\cup C)_2}(x)}{\sum_{x\in X} u_{B_2}(x)}, \frac{\sum_{x\in X} u_{(B\cup C)_2}(x)}{\sum_{x\in X} l_{B_2}(x)} \wedge 1\right]$$

We are going to use the implications between attributes to try to estimate the value of the absent data.

Example.

Returning to the previous example, we have the relation given by the Table 4 where the two unknown values have been replaced by $[0,1]$.

Table 4: Values of the relation R

R	$I1$	$I2$	$I3$	$I4$
$M1$	$[0.8,1]$	$[0.1,0.1]$	$[0.4,0.4]$	$[0.4,0.6]$
$M2$	$[0.8,0.8]$	$[0,1]$	$[0.2,0.2]$	$[0.6,0.6]$
$M3$	$[1,1]$	$[0,0]$	$[0.4,0.4]$	$[0.4,0.4]$
$M4$	$[0.9,0.9]$	$[0.2,0.2]$	$[0.1,0.2]$	$[0,1]$
$M5$	$[0.8,0.8]$	$[0,0.2]$	$[0.2,0.4]$	$[0.2,0.2]$

We will study the implications between attributes in which they appear the attributes $I2$ and $I4$ with the purpose of finding those that are verified with a high level of confidence. For example:

• If we consider the implication $I_3 \Rightarrow I4$, we will obtain the following support and confidence:

$$\mathrm{supp}(I3 \Rightarrow I4) = \left[\frac{1.2}{5}, \frac{1.4}{5}\right] = [0.24, 0.28]$$

$$\mathrm{conf}(I3 \Rightarrow I4) = \left[\frac{1.2}{1.6}, 1\right] = [0.75, 1]$$

Between a 24 and a 28% of the trademarks use a great amount of the ingredients $I3$ and $I4$. The implication is verified with a high level of confidence, in a 75% or more of the marks in which the ingredient $I3$ appears, there is at least the same amount of the ingredient $I4$. That allows us to extend the rule to the mark $M4$ for which we do not know the quantity of the ingredient $I4$, and we will suppose that also in this case we have at least the same amount than that of the ingredient $I3$, that is, $R(M4, I4) \geq R(M4, I3) = [0.1, 0.2]$.

We obtain a condition that allows us to reduce the lack of information, and to represent that value by an interval with a smaller width, $R(M4, I4) = [0.1, 1]$.

• The implication $I4 \Rightarrow I1$ is verified with a support and a confidence:

$$\mathrm{supp}(I4 \Rightarrow I1) = \left[\frac{1.6}{5}, \frac{2.7}{5}\right] = [0.32, 0.54]$$

$$\mathrm{conf}(I4 \Rightarrow I1) = \left[\frac{1.6}{2.8}, 1\right] = [0.57, 1]$$

We can conclude that the percentage of marks in which the ingredient $I1$ appears at least in the same proportion than the ingredient $I4$, is more than a 57%. The implication is verified with a quite high level of confidence and then it is fulfilled also in the mark $M4$ with a high probability. We will suppose then that $R(M4, I4) \leq R(M4, I1) = [0.9, 0.9]$, and we will assign the value $R(M4, I4) = [0, 0.9]$.

We can obtain a smaller vagueness if we consider the two implications:

$$[0.1, 0.2] \leq R(M4, I4) \leq [0.9, 0.9]$$

and therefore, the vagueness will be smaller or equal than the one given by the value $R(M4, I4) = [0.1, 0.9]$.

In this way, we could analyze all the implications between attributes with any unknown value, with the purpose of obtaining relations among them that allow us to limit the degree of vagueness of each one of these values.

4.2 Implications in an L-fuzzy context with negation of attributes

In the previous section, we have used the association rules to set up the implications between the attributes that allow us to replace the absent values of an L-Fuzzy context in a suitable way.

Nevertheless, it is necessary to consider that these association rules do not work well when we study the negation of an attribute (for example, if we use the ingredient sugar, we do not use the ingredient saccharin). Even more, in some cases we can not find the important information by means of these rules [6].

We are going to study these situations to be able to analyze the implications between high values of an attribute and low values of another one.

In order to do it, from an original context $(\mathcal{J}[L], X, Y, R)$ we will construct $(\mathcal{J}[L], X, \widehat{Y}, R)$ where $\widehat{Y}$ will be formed by the values of Y and the negations of the attributes whose low values we want to analyze, so that $R(x_i, \bar{y}_j) = [1-b, 1-a]$ if $R(x_i, y_j) = [a, b]$.

From here, we can use the Definition 10 to study these implications.

Example.

We want to analyze, in the previous example, if the existence of the ingredient $I1$ implies the absence of the ingredient $I2$. To do it, we consider $\overline{I2}$ that represents the non existence of the ingredient $I2$ and we obtain the relation extended to the new set of attributes given by

Table 5: Values of the extended relation R

R	$I1$	$I2$	$I3$	$I4$	$\overline{I2}$
$M1$	$[0.8, 1]$	$[0.1, 0.1]$	$[0.4, 0.4]$	$[0.4, 0.6]$	$[0.9, 0.9]$
$M2$	$[0.8, 0.8]$	$[0, 1]$	$[0.2, 0.2]$	$[0.6, 0.6]$	$[0, 1]$
$M3$	$[1,1]$	$[0,0]$	$[0.4, 0.4]$	$[0.4, 0.4]$	$[1, 1]$
$M4$	$[0.9, 0.9]$	$[0.2, 0.2]$	$[0.1, 0.2]$	$[0, 1]$	$[0.8, 0.8]$
$M5$	$[0.8, 0.8]$	$[0, 0.2]$	$[0.2, 0.4]$	$[0.2, 0.2]$	$[0.8, 1]$

We calculate the support and the confidence of the implication $I1 \Rightarrow \overline{I2}$ applying the formulas previously defined:

$$\text{supp}(I1 \Rightarrow \overline{I2}) = \left[\frac{3.4}{5}, \frac{4.3}{5}\right] = [0.68, 0.86]$$

$$\text{conf}(I1 \Rightarrow \overline{I2}) = \left[\frac{3.4}{4.5}, 1\right] = [0.75, 1]$$

In this case, the implication is fulfilled with a very high support and level of confidence, this will allow us to decrease the vagueness of the values of the table which appear in this implication. A mark that has the ingredient $I1$ in a great amount will have the ingredient $\overline{I2}$ at least in the same amount, consequently,

$$R(M2, \overline{I2}) \geq R(M2, I1) = [0.8, 0.8]$$

We can replace the value $R(M2, \overline{I2}) = [0, 1]$ by another one with a smaller vagueness, $R(M2, \overline{I2}) = [0.8, 1]$.

So, we will have that the mark $M2$ has the ingredient $\overline{I2}$ at least with a membership grade 0.8, that is, it contains the ingredient $I2$ at the most with a membership grade 0.2. We can replace then $R(M2, I2) = [0, 0.2]$ with which we will eliminate, also in this case, a great part of the vagueness.

5 Future work

In future works, we will try to improve the process of elimination of the absent values by means of the techniques of fuzzy clustering and, after proceeding to the construction of the L-Fuzzy concepts lattice, we will analyze the influence that those absent values have in the obtained results.

References

[1] R.Agrawal, T.Imielinski, A.Swami. Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data.* Washington, (1993), pp.207-216.

[2] R.Bělohlávek, *Fuzzy Galois connections and fuzzy concept lattices: from binary relations to conceptual structures*, in: Novak V., Perfileva I. (eds.): Discovering the World with Fuzzy Logic, Physica-Verlag 2000, 462-494.

[3] A.Burusco, R.Fuentes-González. The Study of the L-Fuzzy Concept Lattice. *Mathware and Computing.* 1 No.3 (1994), pp.209-218.

[4] A.Burusco, R.Fuentes-González. Concept lattices defined from implication operators. *Fuzzy Sets and Systems.* 114 No.1 (2000), pp.431-436.

[5] A.Burusco, R.Fuentes-González, C.Alcalde. Utilización de reglas de asociación en la teoría de conceptos L-Fuzzy. *ESTYLF 2004.* Jaén (2004), pp.303-308.

[6] J.Hernández, M.J.Ramírez, C.Ferri. *Introducción a la minería de datos.* Pearson Prentice Hall, Madrid, (2004).

[7] S.Pollandt, *Fuzzy Begriffe: Formale Begriffsanalyse unscharfer Daten*, Springer, 1997.

[8] G.Stumme, R.Taouil, Y.Bastide, N.Pasquier, L.Lakhal. Computing iceberg concept lattices with TITANIC. *Data & Knowledge Engineering.* 42 (2002), pp.189-222.

[9] R.Wille, Restructuring lattice theory: an approach based on hierarchies of concepts,in: Rival I.(ed.),Ordered Sets, Reidel, (Dordrecht-Boston, 1982) 445-470.