Some dependencies among attributes given by fuzzy confirmation measures

Jiří Kupka[1] **Iva Tomanová**[2]

[1,2]Institute for Research and Applications of Fuzzy Modeling, University of Ostrava.

Abstract

This paper is devoted to fuzzy association analysis. Our motivation to this work arose from mining of linguistic associations, however found relations among mined associations are valid in general. Three the most commonly used support and confidence measures are considered and various relations among found and previously known associations given by confirmation measures are studied. Good understanding of such relationships is essential for creating more efficient algorithms, for subsequent work with found associations as well as for cooperation with the consumer of the data mining process.

Keywords: Data mining, linguistic association, confirmation measures, support measure, confidence measure, association rule.

1. Introduction

Our paper is a contribution to *data mining* - a collection of methods that are used for searching for unknown and potentially valid knowledge from large scale datasets. One of first data mining methods was the GUHA method introduced e.g. in [9] (for a comprehensive survey we refer to [10], [16] and references therein).

In a recent paper [14] a theoretical background and also mining of associations that are expressible in natural language (i.e. associations of the form *"IF the area of the base of a cylinder is big AND the height of this cylinder is also big THEN the volume of this cylinder is big."*) via GUHA method was presented and very recently improved in [13]. For completeness, an analogous task is currently studied also in "crisp" association analysis (e.g., see [8]).

One of the main advantages of this approach is *interpretability* of mined rules - i.e., found associations can be easily interpretable to experts working with a chosen dataset. Our motivation was to proceed further in this way and, e.g., automatically to use certain expert knowledge (that can be linguistically and also mathematically expressed in the same way as mined associations) or to allow to cooperate directly with the end user in the data mining process etc. Except for this, reasonable description of the inner structure of a given dataset (resp. of the set of found associations) is required for creating efficient mining algorithms, and also for reasonable subsequent work with the set of found associations (e.g., see [14] and [13]) as well as for potential creating reasonable (e.g., sound, complete etc.) system of associations (e.g. [3]) etc. Similar tasks were already studied in "crisp" methods - GUHA method can serve as an example (see [10]), however, except for [7], there is no survey paper devoted to this task provided fuzzy confirmation measures are taken into consideration.

In this work three the most commonly used fuzzy confirmation measures are considered - see [5] and [7] for justification of their existence. For each pair of considered confirmation measures (i.e., support and confidence measure) we study several properties (look at the beginning of Section 4). Six of them are motivated by so-called Armstrong axioms that, among other things, can be used for database design (see e.g. [2]) and are also valid in fuzzy attribute logic developed e.g. in [3]. This logic can be applied to data sets similar to ours and, under some additional assumptions, establishes a complete and sound system of associations. Thus, it was a natural question under what conditions we can obtain similar relations in ordinary fuzzy association analysis. Further properties we have decided to study are motivated by properties that can be used in current methods of association analysis - for example, in GUHA method ([10]) or in known Apriori algorithm ([1] and also [4]).

We explain that some properties remain valid when we use our fuzzy confirmation measures. But we have also obtained some negative results and we demonstrated that the situation can be improved if some additional (expert) knowledge is applied to our properties. We would like to stress that our results are provided separately either for support or confidence measure if necessary. For details we refer to the end of Section 4 and [13].

This contribution contains a short survey of some results from [12] and some additional ones. The paper consists of six parts and is organized as follows - in Section 2 some basic notions are introduced. Chosen properties are described and obtained results are provided in Section 4. As finally, an illustrative example (in Section 5) and concluding comments (in Section 6) follow.

2. Preliminaries

Within this paper we use several common notions from fuzzy mathematics (see e.g. [15]) - hence we do not precisely define basic notions like *fuzzy set, t-norm, t-conorm, copula, residuated implication etc.* and their basic properties.

Let $\mathbb{R}$ denote the real line. We work with a data set $\mathcal{D}$ that is given by a real-valued two-dimensional table. The rows o_i of $\mathcal{D}$ correspond to *objects* (the set of objects is denoted $\mathcal{D}_o$) and columns X_i, $i = 1, 2, \ldots, k$, of $\mathcal{D}$ correspond to attributes. For each attribute X_i its context $[a_i, b_i]$ is chosen as a nondegenerated closed interval $[a_i, b_i] \subseteq \mathbb{R}$. Then relevant evaluative linguistic expressions represented by fuzzy sets can be defined ([14]). *Evaluative linguistic expressions* allow us to use several natural language expressions - for instance, a triplet of *atomic linguistic expressions* formed by expressions: *Big* (*Bi*), *Medium* (*Me*), *Small* (*Sm*).

Additionally, atomic linguistic expressions can be composed with various *linguistic hedges* (introduced by L. A. Zadeh): *Extremely* (*Ex*), *Significantly* (*Si*), *Very* (*Ve*), *More or Less*(*ML*), *Roughly* (*Ro*), *Quite Roughly* (*QR*), *Very Roughly* (*VR*).

There are several ways how to choose fuzzy sets representing evaluative linguistic expressions mentioned above. We can refer to an original mathematical model of linguistic expressions (Model I) elaborated in [14] or to another mathematical model (Model II) established in [13]. The latter mathematical model is based on fuzzy coverings (resp. partitions) and it was justified that sometimes it is necessary to use other forms of linguistic expressions - namely, so-called *specifying linguistic expressions* of the form "*Si Sm* `but not` *Ex Sm*".

Regardless of which mathematical model is chosen, we simply use the fact that linguistic expressions are represented by fuzzy sets. Then we can look for dependencies between given (usually disjoint) sets of attributes $\{\bar{X}_i\}_{i=1}^{p}$, $\{Y_j\}_{j=1}^{q} \subseteq \{X_i\}_{i=1}^{k}$ and we consider associations of the form

$$E(\{\bar{X}_i\}_{i=1}^{p}) \Rightarrow F(\{Y_j\}_{j=1}^{q}) \qquad (1)$$

where the left and right side of (1) is called *antecedent* and *succedent*, respectively. For simplicity we will write only $E \Rightarrow F$ instead of (1). In (1), F, resp. E, can represent various conjunctions `AND`'s (given by t-norms) and disjunctions `OR`'s (given by t-conorms) of respective fuzzy sets and $\Rightarrow$ represents a relationship between E and F given by chosen confirmation measures. Confirmation measures are specific mathematical expressions that confirm the validity of (1) in chosen data set - for instance, see support and confidence measures (4), (6), (7) and (8) below.

3. Association analysis from fuzzy viewpoint

There are several ways how to choose confirmation measures that determine linguistic associations (1). One of the best known methods for searching linguistic associations is so-called GUHA method ([10]). Its confirmation measures (called quantifiers) are computed from relevant four-fold tables. To construct such tables crisp partitions (induced by relevant fuzzy sets) of contexts of considered attributes are required ([10], [14] etc.). However, there are also other possibilities due to which we can work directly with fuzzy sets carrying linguistic labels.

For instance, in [5], the problem of choosing reasonable fuzzy confirmation measures was studied systematically and choices of various confirmation measures were justified especially in connection with a certain and very natural partition of the row set $\mathcal{D}_o$. The partition required therein is given by the condition

$$S_+(o_i) + S_-(o_i) + S_\pm(o_i) = 1, \text{ for any } o_i \in \mathcal{D}_o, \qquad (2)$$

where $S_+(o_i), S_-(o_i), S_\pm(o_i)$ denotes a *positive*, *negative* and *irrelevant* evaluation, respectively, of each row $o_i \in \mathcal{D}_o$ of a given rule (1). Such a partition of $\mathcal{D}_o$ can be of the form

$$\begin{aligned} S_+(o_i) &:= E(o_i) \otimes F(o_i), \\ S_-(o_i) &:= 1 - (E(o_i) \to F(o_i)), \\ S_\pm(o_i) &:= 1 - E(o_i) \end{aligned} \qquad (3)$$

where a t-norm $\otimes$ is so-called *copula* (e.g., Łukasiewicz t-norm is the smallest copula) and an implication operator $\to$ is given by $a \to b = (1 - a) + (a \otimes b)$. Under these assumptions, the partition (3) guarantees (2) for any possible rule (1). Additionally, the authors of [5] justified how any such partition induces meaningful fuzzy confirmation (i.e., support and confidence) measures.

Consequently, for the partition (3), the following (*t-norm-based*) support measure of (1) is suggested in the data set $\mathcal{D}$

$$supp_t(E \Rightarrow F) := \sum_{o_i \in \mathcal{D}_o} E(o_i) \otimes F(o_i). \qquad (4)$$

Remark 1 *However, the problem in [5] can be further specified. For instance, when we require the implication operator $\to$ to be a self implication (i.e., $a \to a = 1$), the authors obtained that $\otimes$ has to be the minimum t-norm ($\otimes(x, y) = \min\{x, y\}$), or, when we require $\to$ to be the strong implication (i.e., $a \to b = n(a) \otimes b$ for a strong negation $n(a)$) the solution is given only by the product t-norm ($\otimes(x, y) = xy$). For completeness we note that the partition is again given by (3).*

The authors of [11] discussed also *gradual* (resp. *certainty*) rules. Such rules are of the form "The

more the property E is true, the more the conclusion F is true". In that case, another definition of partition of $\mathcal{D}_o$ for $E \Rightarrow F$ was considered

$$\begin{aligned} S_+(o_i) &:= E(o_i) \otimes (E(o_i) \to F(o_i)), \\ S_-(o_i) &:= E(o_i) \otimes (1 - (E(o_i) \to F(o_i))), \\ S_\pm(o_i) &:= 1 - E(o_i). \end{aligned} \quad (5)$$

For this partition, the condition (2) is satisfied only for the product t-norm. Consequently, the following (*implication-based*) support measure is taken for (1)

$$supp_c(E \Rightarrow F) := \sum_{o_i \in \mathcal{D}_o} E(o_i) \cdot (E(o_i) \to F(o_i)), \quad (6)$$

where $\to$ represents any *generalized implication*. More precisely, an *implication operator* $\mathcal{I} : I \times I \to I$ for $I = [0, 1]$ is a generalization of the material implication if it satisfies, for $x, y, x_0, y_0 \in I$,

(I1) $\mathcal{I}(x, y) \leq \mathcal{I}(x_0, y)$ for $x_0 \leq x$,
(I2) $\mathcal{I}(x, y) \leq \mathcal{I}(x, y_0)$ for $y \leq y_0$,
(I3) $\mathcal{I}(1, y) = y$, and
(I4) $\mathcal{I}(0, 0) = 1$.

In order to keep preceding notation we put $\to := \mathcal{I}$. Sometimes, $x \to y$ denotes also a *product implication* that is equal to 1 for $x \leq y$ or to y/x otherwise.

Finally, if $\otimes$ is a continuous t-norm and $\to$ is derived from that t-norm through residuation, the (*minimum-based*) support measure is obtained:

$$supp_m(E \Rightarrow F) := \sum_{o_i \in \mathcal{D}_o} \min\{E(o_i), F(o_i)\}. \quad (7)$$

For support measures (4), (6) and (7), relevant *confidence* measures are defined by

$$conf_j(E \Rightarrow F) := \frac{supp_j(E \Rightarrow F)}{\sum_{o_i \in \mathcal{D}_o} E(o_i)} \quad (8)$$

for $j \in \{t, c, m\}$. Note that (8) cannot be strictly greater than 1 for any association $E \Rightarrow F$.

As usually, when a support and its relevant confidence measures are fixed, then we say that a given rule $E \Rightarrow F$ is *valid* if its support and confidence degrees are greater than or equal to preliminary given *support* and *confidence thresholds*.

Further, for given rules $E_1 \Rightarrow F_1$ and $E_2 \Rightarrow F_2$, $E_1 \Rightarrow F_1 \vdash_s E_2 \Rightarrow F_2$ denotes the fact that $supp(E_1 \Rightarrow F_1) \leq supp(E_2 \Rightarrow F_2)$. Similarly, $E_1 \Rightarrow F_1 \vdash_c E_2 \Rightarrow F_2$ denotes the fact that $conf(E_1 \Rightarrow F_1) \leq conf(E_2 \Rightarrow F_2)$ and finally $E_1 \Rightarrow F_1 \vdash E_2 \Rightarrow F_2$ means that $E_1 \Rightarrow F_1 \vdash_j E_2 \Rightarrow F_2$ for $j = \{s, c\}$. Analogous notation we can also use for sets of associations - the expression

$$A \Rightarrow B,\ C \Rightarrow D \vdash E \Rightarrow F$$

means that $E \Rightarrow F$ is valid, i.e., it has higher support and confidence degrees than either $A \Rightarrow B$ or $C \Rightarrow D$ etc.

Let us remark that we can also use associations of the form $C \Rightarrow C$. A considering such associations is very natural, their confidence degree has to be equal to 1. Consequently, the validity of this association implies that the linguistic expression represented by C has a sufficiently large support.

3.1. Using additional knowledge

Let us introduce a set $\mathcal{E}$ of associations (i.e., the set of *expert knowledge*) that can be provided to the data mining process. We would like to emphasize that linguistic and mathematical representation is the same for associations from $\mathcal{E}$ as well as for associations we want to mine in a given data-set.

Note that we need not to specify the inner structure of such expert associations (i.e., associations from $\mathcal{E}$). For a given unknown association $E \Rightarrow F$ we would like to test, associations from $\mathcal{E}$ (notation $A \Rightarrow^* B$) can describe information between the antecedent and succedent part of $E \Rightarrow F$ as well as between attributes contained either in the antecedent or succedent part of $E \Rightarrow F$, respectively. We would like to stress that it makes sense to deal only with confidence measures of associations from $\mathcal{E}$.

We would like to emphasize that, within this paper, associations from $\mathcal{E}$ are assumed to be *fully* valid in the dataset $\mathcal{D}$, i.e. we assume $conf_i(A \Rightarrow^* B) = 1$. According to the choice of a confidence measure, we can obtain some additional information. For instance, if $A(o_i) \leq B(o_i)$ holds for each $o_i \in \mathcal{D}_o$ then $conf_i(A \Rightarrow^* B) = 1$ is satisfied for t-norm-based confidence measures (and hence for the minimum-based one). When an implication-based confidence measure is considered, we can obtain the same condition provided $\to$ is a residuated implication of some t-norm. But if $\to$ is a generalized implication then only $B(o_i) = 1$ for any $o_i \in \mathcal{D}_o$ can be assumed.

4. Properties

In this section we study the following properties:
P1 $(A \texttt{ OR } B) \Rightarrow A$,
P2 $A \Rightarrow B,\ (B \texttt{ OR } C) \Rightarrow D \vdash (A \texttt{ OR } C) \Rightarrow D$,
P3 $A \Rightarrow B \vdash (C \texttt{ AND } A) \Rightarrow (C \texttt{ AND } B)$,
P4 $(A \Rightarrow B),\ (A \Rightarrow C) \vdash (A \Rightarrow (B \texttt{ OR } C))$,
P5 $A \Rightarrow (B \texttt{ OR } C) \vdash A \Rightarrow B$,
P6 $A \Rightarrow B,\ B \Rightarrow C \vdash A \Rightarrow C$,
P7 $A \Rightarrow B,\ C \Rightarrow D \vdash (A \texttt{ OR } C) \Rightarrow (B \texttt{ OR } D)$,
P8 $(A \texttt{ AND } B) \Rightarrow (C \texttt{ AND } D) \vdash (A \texttt{ AND } B \texttt{ AND } D) \Rightarrow C$.

In [2] the authors study axioms and inference rules used in database design. It should be also mentioned that the same rules are valid also in fuzzy attribute logic elaborated e.g. in [3]. These axioms and inference rules are described by Properties P1-P6.

The last two properties - Properties P7 and P8 are motivated by analogous properties that are used e.g. in GUHA method ([10]) or in the classic Apriori algorithm (see [1] and references therein). So, let us continue in our study property by property.

4.1. Properties P1 and P2

As regards Property P1, it has been explained in [13] that this property need not be satisfied in general. Thus let us study

$$A \Rightarrow B,\ (B \ \texttt{OR}\ C) \Rightarrow D \vdash (A \ \texttt{OR}\ C) \Rightarrow D.$$

We explained in [13] that, for t-norm-based measures,

$$A \Rightarrow B,\ (B \ \texttt{OR}\ C) \Rightarrow D \not\vdash_s (A \ \texttt{OR}\ C) \Rightarrow D$$

and also

$$A \Rightarrow B,\ (B \ \texttt{OR}\ C) \Rightarrow D \not\vdash_c (A \ \texttt{OR}\ C) \Rightarrow D.$$

The same negative expressions we obtain also for the minimum-based and implication-based confirmation measures - see the next two simple examples.

Example 1 *Consider a dataset with three rows and fuzzy sets A, B, C, D with values:*

	A	B	C	D
o_1	0.8	1	0.4	1
o_2	0.7	0.8	0.4	0.1
o_3	0.6	0.7	0.8	0.4

Then we immediately obtain $supp_m(A \Rightarrow B) = 2.1 > supp_m((B \,\texttt{OR}\, C) \Rightarrow D) = 1.5 > supp_m((A \,\texttt{OR}\, C) \Rightarrow D) = 1.3$. Moreover, $conf_m(A \Rightarrow B) = 1 > conf_m((B \,\texttt{OR}\, C) \Rightarrow D) = 0.57 > conf_m((A \,\texttt{OR}\, C) \Rightarrow D) = 0.56$. Additionally, this example contradicts

$$A \Rightarrow^* B,\ (B \ \texttt{OR}\ C) \Rightarrow D \vdash (A \ \texttt{OR}\ C) \Rightarrow D.$$

Example 2 *Consider a dataset with three rows and fuzzy sets A, B, C, D with values:*

	A	B	C	D
o_1	0.2	1	0.8	0.2
o_2	0.1	0.8	1	0.1
o_3	0.7	0.8	0.8	0.2

Then, for $x \to y := \max\{1 - x, y\}$, we get $supp_c(A \Rightarrow B) = 0.85 > supp_c((B \,\texttt{OR}\, C) \Rightarrow D) = 0.49 > supp_c((A \,\texttt{OR}\, C) \Rightarrow D) = 0.46$. Moreover, we obtain $conf_c(A \Rightarrow B) = 0.85 > conf_c(B \,\texttt{OR}\, C) \Rightarrow D) = 0.17 > conf_c((A \,\texttt{OR}\, C) \Rightarrow D) = 0.16$, i.e., it contradicts also

$$A \Rightarrow^* B,\ (B \ \texttt{OR}\ C) \Rightarrow D \vdash (A \ \texttt{OR}\ C) \Rightarrow D.$$

Analogously it could be demonstrated that using $A \Rightarrow^* B$ need not be justified for general t-norm based confirmation measures.

4.2. Property P3

It is shown in [13] that the validity of

$$A \Rightarrow B \vdash_c (C \ \texttt{AND}\ A) \Rightarrow (C \ \texttt{AND}\ B)$$

is satisfied neither for the minimum-based support measure nor for the confidence one. To ensure the validity of this rule it is necessary to require more - for instance, for the minimum-based confirmation measure, we have

$$A \Rightarrow^* B \vdash_c (C \ \texttt{AND}\ A) \Rightarrow (C \ \texttt{AND}\ B).$$

In the rest of this subsection we deal with an exact modification of Property P3 that can hold for support measures (7), (4) and (6). However, it can be shown that it does not hold for confidence measures (Example 3).

Lemma 1 *Let us consider the t-norm-based support measure given by* (4)*. Then*

$$(A \ \texttt{AND}\ C) \Rightarrow (B \ \texttt{AND}\ C) \vdash_s A \Rightarrow B.$$

Proof. We consider the t-norm-based support measure (4). We can obtain $supp_t((A \ \texttt{AND}\ C) \Rightarrow (B \ \texttt{AND}\ C)) = \sum_{o_i \in D_o} A(o_i) \otimes C(o_i) \otimes B(o_i) \otimes C(o_i)$ and it is smaller than $supp_t(A \Rightarrow B) = \sum_{o_i \in D_o} A(o_i) \otimes B(o_i)$ and therefore $supp_t((A \ \texttt{AND}\ C) \Rightarrow (B \ \texttt{AND}\ C)) \leq supp_t(A \Rightarrow B)$. □

From this lemma the next corollary easily follows.

Corollary 1 *Let us consider the minimum-based support measure given by* (7)*. Then*

$$(A \ \texttt{AND}\ C) \Rightarrow (B \ \texttt{AND}\ C) \vdash_s A \Rightarrow B.$$

Lemma 2 *Let us consider the implication-based support measure given by* (6) *and the product implication. Then*

$$(A \ \texttt{AND}\ C) \Rightarrow (B \ \texttt{AND}\ C) \vdash_s A \Rightarrow B.$$

Proof. We consider the implication-based support measure (6) with the product implication $\to$. Put $D_1 := \{o_i \in \mathcal{D}_o \mid A(o_i) \leq B(o_i)\}$ and $D_2 := \mathcal{D}_o \setminus D_1$ and consider two expressions:

$$A(o_i) \cdot (A(o_i) \to B(o_i)) \tag{9}$$

and

$$(A(o_i)C(o_i)) \cdot (A(o_i)C(o_i) \to B(o_i)C(o_i)). \tag{10}$$

For any $o_i \in D_1$ we easily obtain

$$A(o_i)C(o_i) = (10) \leq (9) = A(o_i). \tag{11}$$

Consequently, since $supp_c$ of $(A \ \texttt{AND}\ C) \Rightarrow (B \ \texttt{AND}\ C)$ and $A \Rightarrow B$ is counted as the sum of (10) and (9), respectively, we get that (11) implies

$$(A \ \texttt{AND}\ C) \Rightarrow (B \ \texttt{AND}\ C) \vdash_s A \Rightarrow B \tag{12}$$

on the set D_1.

Analogously, we can use an analogous argument for the set D_2, since we clearly obtain

$$B(o_i)C(o_i) = (10) \leq (9) = B(o_i)$$

for any $o_i \in D_2$. □

In the following example it is shown that this rule can not be proved for confidence measures.

Example 3 *We consider a dataset consisting of one object with the following values of fuzzy sets A, B, C: $A(o_1) = 0.9,\ B(o_1) = 0.5,\ C(o_1) = 0.1$. Then $conf_m((A\ \texttt{AND}\ C) \Rightarrow (B\ \texttt{AND}\ C)) = 1$ and it is greater than $conf_m(A \Rightarrow B) = 5/9$.*

Or when we consider a dataset consisting of two objects with values of fuzzy sets A, B, C: $A(o_1) = 0.9,\ B(o_1) = 0.5,\ C(o_1) = 0.1$ $A(o_2) = 0.6,\ B(o_2) = 0.8,\ C(o_2) = 0.2$. Then $conf_c((A\ \texttt{AND}\ C) \Rightarrow (B\ \texttt{AND}\ C)) = 0.8$ and it is greater than $conf_c(A \Rightarrow B) = 0.73$.

Remark 2 *Here and in subsequent counterexamples we consider the product implication for implication-based confirmation measures - i.e., the residuated implication induced by the product t-norm.*

4.3. Property P4

Let us study the rule

$$A \Rightarrow B,\ A \Rightarrow C \vdash A \Rightarrow (B\ \texttt{OR}\ C).$$

As we can see from the following lemma, the validity of this property is straightforward.

Lemma 3 (P4) *([12])Let us consider confirmation measures given by* (4), (6), (7) *and* (8). *Then*

$$A \Rightarrow B,\ A \Rightarrow C \vdash A \Rightarrow (B\ \texttt{OR}\ C).$$

Remark 3 *Clerly, since Property P4 is valid in general, it can be used also in connection with the expert knowledge (i.e, associations from $\mathcal{E}$) we consider in our task. Thus, e.g.,*

$$A \Rightarrow^* B,\ A \Rightarrow C \vdash A \Rightarrow (B\ \texttt{OR}\ C).$$

4.4. Property P5

In this subsection we focus on the property

$$A \Rightarrow (B\ \texttt{OR}\ C) \vdash A \Rightarrow B.$$

In [12] some examples demonstrate that this property does not hold in general for various confirmation measures.

According to examples from [12] we can claim, for all confirmation measures considered in this paper, that

$$A \Rightarrow (B\ \texttt{OR}\ C) \not\vdash_s A \Rightarrow B.$$

and

$$A \Rightarrow (B\ \texttt{OR}\ C) \not\vdash_c A \Rightarrow B.$$

However, it can be seen from the subsequent lemma we can specify some additional assumptions in order to ensure the validity of Property P5.

Lemma 4 (P5) *Let us consider the minimum-based confirmation measures* (7) *and* (8). *Then*

$$A \Rightarrow (B\ \texttt{OR}\ C), C \Rightarrow^* B \vdash A \Rightarrow B.$$

Proof. By our assumptions we have, for support and confidence thresholds r and γ,

$$supp_m(A \Rightarrow B\ \texttt{OR}\ C) := \sum_{o_i \in \mathcal{D}_o} \min\{A(o_i), \max\{B(o_i), C(o_i)\}\} \geq r \quad (13)$$

and

$$conf_m(A \Rightarrow B\ \texttt{OR}\ C) := \frac{\sum_{o_i \in \mathcal{D}_o} \min\{A(o_i), \max\{B(o_i), C(o_i)\}\}}{\sum_{o_i \in \mathcal{D}_o} A(o_i)} \geq \gamma. \quad (14)$$

We want to prove

$$supp_m(A \Rightarrow B) := \sum_{o_i \in \mathcal{D}_o} \min\{A(o_i), B(o_i)\} \geq r \quad (15)$$

and

$$conf_m(A \Rightarrow B) := \frac{\sum_{o_i \in \mathcal{D}_o} \min\{A(o_i), B(o_i)\}}{\sum_{o_i \in \mathcal{D}_o} A(o_i)} \geq \gamma. \quad (16)$$

According to Subsection 3.1, $C \Rightarrow^* B$ implies that $C(o_i) \leq B(o_i)$ for any $o_i \in \mathcal{D}_o$. Having this in mind, it is clear that (13) is equal to (15) and also (14) is equal to (16). □

The following counterexample shows that the lemma above can not be constructed for t-norm based and implication-based confirmation measures.

Example 4 *Consider dataset with one object and fuzzy sets A, B, C with values - $A(o_1) = 0.5,\ B(o_1) = 0.1,\ C(o_1) = 0.9$.*

Then $supp_t(A \Rightarrow (B\ \texttt{OR}\ C)) = 0.455$, resp. $conf_t(A \Rightarrow (B\ \texttt{OR}\ C)) = 0.91$, is strictly greater than $supp_t(A \Rightarrow B) = 0.05$, resp. $conf_t(A \Rightarrow B) = 0.01$.

Further, $supp_c(A \Rightarrow (B\ \texttt{OR}\ C)) = 0.5$, resp. $conf_c(A \Rightarrow (B\ \texttt{OR}\ C)) = 1$, is strictly greater than $supp_c(A \Rightarrow B) = 0.1$, resp. $conf_c(A \Rightarrow B) = 0.2$.

4.5. Property P6

In this subsection we consider the property

$$A \Rightarrow B,\ B \Rightarrow C \vdash A \Rightarrow C. \quad (17)$$

The authors of [12] demonstrated in counterexamples that this property is not valid in general in the set of mined associations, thus

$$A \Rightarrow B,\ B \Rightarrow C \not\vdash_s A \Rightarrow C$$

and

$$A \Rightarrow B,\ B \Rightarrow C \not\vdash_c A \Rightarrow C \quad (18)$$

for all support measures (4), (6) and (7), respectively. Additionally, there are the examples demonstrating that requiring some additional assumptions (for example $A \Rightarrow A,\ B \Rightarrow B,\ C \Rightarrow C$) need not lead to the validity of Property P6.

For completeness, there exists a lemma claiming that by using some expert knowledge we can reasonably use Property P6.

Lemma 5 *([12]) Let us consider confirmation measures given by* (8) *and* (4), (6), (7). *Then*

$$A \Rightarrow B,\ B \Rightarrow^* C \vdash A \Rightarrow C.$$

Remark 4 *As an easy corollary of Lemma 5 we can see that an ordinary transitivity ($A \Rightarrow^* B,\ B \Rightarrow^* C \vdash A \Rightarrow^* C$) is preserved in the set $\mathcal{E}$. On the other side, the property ($A \Rightarrow^* B,\ B \Rightarrow C \vdash A \Rightarrow C$) need not be valid in general.*

4.6. Properties P7

We can return to the original motivation ([5]) of establishing confirmation measures (8), (4), (6) and (7). And we can speak about a *positive*, *negative* and *irrelevant* part of the rule $E \Rightarrow F$ (notation $S_+(E \Rightarrow F)$, $S_-(E \Rightarrow F)$ and $S_\pm(E \Rightarrow F)$), respectively. Note that each $S_i(E \Rightarrow F)$, $i \in \{+, -, \pm\}$, can be seen as a fuzzy set on $\mathcal{D}_o$. In [5] confirmation measures (8), (4), (6) and (7) were established in order to satisfy

$$supp(E \Rightarrow F) = \sum_{o_i \in \mathcal{D}_o} S_+(E \Rightarrow F)(o_i)$$

and

$$conf(E \Rightarrow F) = \frac{\sum_{o_i \in \mathcal{D}_o} S_+(E \Rightarrow F)(o_i)}{\sum_{o_i \in \mathcal{D}_o} (S_+(E \Rightarrow F)(o_i) + S_-(E \Rightarrow F)(o_i))}$$

for given partitions satisfying (2).

It is easy to see from the last two expressions that having two valid associations $E_1 \Rightarrow F_1$, $E_2 \Rightarrow F_2$ with "disjoint" positive parts ensures the validity of $(E_1 \,\texttt{OR}\, E_2) \Rightarrow (F_1 \,\texttt{OR}\, F_2)$ whenever the linguistic OR is represented by pointwise maximum.

Therefore, we can work with a rule

$$A \Rightarrow B,\ C \Rightarrow D \vdash (A \,\texttt{OR}\, C) \Rightarrow (B \,\texttt{OR}\, D)$$

for fuzzy sets A, C with disjoint supports. Generally, the following results for P7 and its special case ($C \Rightarrow C$) can be provided.

Lemma 6 (P7) *([12]) Let us consider the minimum-based support measure given by* (7). *Then*

$$A \Rightarrow B,\ C \Rightarrow D \vdash_s (A \,\texttt{OR}\, C) \Rightarrow (B \,\texttt{OR}\, D),$$

Corollary 2 *([12]) For minimum-based support measure we have*

$$A \Rightarrow B,\ C \Rightarrow C \vdash_s (C \,\texttt{OR}\, A) \Rightarrow (C \,\texttt{OR}\, B).$$

Lemma 7 (P7) *([12]) Let us consider the t-norm-based support measure given by* (4). *Then*

$$A \Rightarrow B,\ C \Rightarrow D \vdash_s (A \,\texttt{OR}\, C) \Rightarrow (B \,\texttt{OR}\, D),$$

Analogously we obtain the subsequent corollary.

Corollary 3 *([12]) For any t-norm-based support measure we have*

$$A \Rightarrow B,\ C \Rightarrow C \vdash_s (C \,\texttt{OR}\, A) \Rightarrow (C \,\texttt{OR}\, B).$$

Lemma 8 (P7) *Let us consider the implication-based support measure given by* (6). *Then*

$$A \Rightarrow B,\ C \Rightarrow D \vdash_s (A \,\texttt{OR}\, C) \Rightarrow (B \,\texttt{OR}\, D) \quad (19)$$

Proof. Let us consider the implication-based support measure (6) with the product implication $\rightarrow$ and the following decomposition of $\mathcal{D}_o$ into $\mathcal{D}'_1, \mathcal{D}''_1, D''_2, D'_{21}$ and D'_{22} – $\mathcal{D}_1 := \{o_i \in \mathcal{D}_o \,|\, C(o_i) \leq D(o_i)\}$, $\mathcal{D}_2 := \mathcal{D}_o \setminus D_1$ and $\mathcal{D}'_1 := \{o_i \in \mathcal{D}_1 \,|\, A(o_i) \oplus C(o_i) \leq B(o_i) \oplus D(o_i)\}$, $\mathcal{D}''_1 := \mathcal{D}_1 \setminus \mathcal{D}'_1$, $\mathcal{D}'_2 := \{o_i \in \mathcal{D}_2 \,|\, A(o_i) \oplus C(o_i) \leq B(o_i) \oplus D(o_i)\}$ and $\mathcal{D}''_2 := \mathcal{D}_2 \setminus \mathcal{D}'_2$ and finally $\mathcal{D}'_{21} := \{o_i \in \mathcal{D}'_2 \,|\, A(o_i) \leq B(o_i)\}$ and $\mathcal{D}'_{22} := D'_2 \setminus \mathcal{D}'_{21}$. Finally, by $\oplus$ we denote a t-conorm of the product t-norm.

Let us study expressions

$$A(o_i) \cdot (A(o_i) \rightarrow B(o_i)), \quad (20)$$

$$C(o_i) \cdot (C(o_i) \rightarrow D(o_i)), \quad (21)$$

$$(A(o_i) \oplus C(o_i)) \cdot (A(o_i) \oplus C(o_i) \rightarrow B(o_i) \oplus D(o_i)) \quad (22)$$

Then (22) $= A(o_i) \oplus C(o_i)$ (or $B(o_i) \oplus D(o_i)$) on set $\mathcal{D}'_1$ (or $\mathcal{D}''_1$). In both cases we have

$$(21) \leq (22) \quad (23)$$

since (21) $= C(o_i)$ for any $o_i \in \mathcal{D}_1$. Similarly, (23) holds also on $\mathcal{D}''_2$ because, for $o_i \in \mathcal{D}''_2$

$$(21) = D(o_i) \leq B(o_i) \oplus D(o_i) < A(o_i) \oplus C(o_i) = (22).$$

Analogously, for any $o_i \in D'_{21}$

$$(20) = A(o_i) \leq A(o_i) \oplus C(o_i) = (22),$$

and for any $o_i \in D'_{22}$

$$(20) = B(o_i) < A(o_i) \leq A(o_i) \oplus C(o_i) = (22).$$

Consequently, (20) $\leq$ (22) holds for any $o_i \in \mathcal{D}'_2$. Since (23) holds for any $o_i \in (\mathcal{D}_o \setminus \mathcal{D}'_2) = \mathcal{D}_1 \cup \mathcal{D}''_2$, we obtain (19) directly from the definition of (6). □

From Lemma 8 we get the following corollary on the special case of Property P7.

Corollary 4 *([12]) For the implication-based support measure with the product implication we have*

$$A \Rightarrow B,\ C \Rightarrow C \vdash_s (C \,\texttt{OR}\, A) \Rightarrow (C \,\texttt{OR}\, B).$$

As regards the confidence measures, the following example demonstrates that Property P7 need not be proved for the minimum-based confidence measure. But it can be proven for the special case of Property P7 in the next lemma.

Example 5 *Consider minimum-based confidence measure and take a dataset consisting of three rows. Let fuzzy sets* A, B, C, D *be defined by* $A(o_1) = 0.9$, $B(o_1) = C(o_1) = D(o_1) = 0.1$, $A(o_2) = B(o_2) = C(o_2) = D(o_2) = 0.9$, $A(o_3) = B(o_3) = D(o_3) = 0.1$ *and* $C(o_3) = 0.9$.

Then $conf_m(A \Rightarrow B) = conf_m(C \Rightarrow D) = 11/19$ *and this expression is greater then* $conf_m(A \texttt{ OR } C \Rightarrow B \texttt{ OR } D) = 11/27$.

Lemma 9 (P7) *([12]) Let us consider the minimum-based confidence measure given by* (8). *Then*

$$A \Rightarrow B,\ C \Rightarrow C \vdash_c (C \texttt{ OR } A) \Rightarrow (C \texttt{ OR } B).$$

4.7. Property P8

The last property we consider is the condition:

$$(A \texttt{ AND } B) \Rightarrow (C \texttt{ AND } D) \vdash (A \texttt{ AND } B \texttt{ AND } D) \Rightarrow C.$$

It can be easily proven that this property can be valid for t-norm-based confirmation measures, and hence also for the minimum-based ones.

Lemma 10 (P8) *([12]) Let us consider the* t*-norm-based confirmation measures given by* (4)*and* (8). *Then*

$$(A \texttt{ AND } B) \Rightarrow (C \texttt{ AND } D) \vdash (A \texttt{ AND } B \texttt{ AND } D) \Rightarrow C.$$

Corollary 5 (P8) *([12]) Let us consider the minimum-based confirmation measures given by* (7)*and* (8). *Then*

$$(A \texttt{ AND } B) \Rightarrow (C \texttt{ AND } D) \vdash (A \texttt{ AND } B \texttt{ AND } D) \Rightarrow C.$$

Remark 5 *Since Property P8 is valid in general, it would be superfluous to study how to use associations from* $\mathcal{E}$.

However, for implication-based confirmation measures, the negative answer can be obtained - see the next example.

Example 6 *([12]) Consider data with attributes represented by fuzzy sets* A, B, C, D *having values* $A(o_i) = B(o_i) = 0.1$ *and* $C(o_i) = D(o_i) = 0.2$.

Then, for the product t-norm and its residuated implication, we obtain $supp_c((A \texttt{ AND } B) \Rightarrow (C \texttt{ AND } D)) = 0.01$. *This is greater than* $supp_c((A \texttt{ AND } B \texttt{ AND } D) \Rightarrow C) = 0.002$.

4.8. Summary

In this subsection we sketch obtained results for particular confirmation measures.

For minimum-based confirmation measures we have demonstrated that some rules (P1, P2, P3, P5, P6, P7) are not valid in general. However, when we can modify some of them (P3) or specify some conditions (P7) or expert knowledge (P4 and P6) in order to guarantee the validity of the considered rule. Finally, P4 and P8 are always valid.

For t-norm-based confirmation measures we have got that Properties P1, P2, P3, P5, P6 and P7 are not valid in general. Similarly as above, we can specify some conditions (for P7) or some expert knowledge (for P4 and P6) in order to get their validity. And as above, P4 and P8 are valid.

Finally we consider implication-based confirmation measures. For such measures, Properties P1, P2, P3, P5, P6 and P7 cannot be used in general. On the other side, P4 and P8 are always valid and for other rules some additional knowledge (for P4 and P6) or assumptions (for P7) can guarantee their validity.

5. Example

At the end of this paper we devise a simple example demonstrating how the mentioned results can be used in the data mining process. We use a dataset entitled NO2 downloaded from the web page: http://lib.stat.cmu.edu/modules.php. And for mining for associations we used the program LAMWin32 developed in our institute.

Our tools are the following (for details see [14])

- a model of evaluative linguistic expressions (more precisely, Model I),
- the implicational quantifier with parameters $r \geq 0.005$ and $\gamma \geq 0.2$.

The next two tables show some of found linguistic associations. The first two columns of first table represent associations of the form "$Hour \Rightarrow Temp$" and "$Temp \Rightarrow Y\_NO2$", an so on.

IF *Hour* is	THEN *Temp* is	IF *Temp* is	THEN *Y_NO2* is
$ML\ Me$	$ML\ Me.$	$ML\ Me$	$ML\ Me.$
Me	$ML\ Sm.$	$ML\ Sm$	$ML\ Me.$
$Ve\ Sm$	$ML\ Sm.$	$ML\ Sm$	$ML\ Me.$

IF *Hour* is	THEN *Y_NO2* is
$ML\ Me$	$ML\ Me.$
Me	$ML\ Me.$
$Ve\ Sm$	$ML\ Me.$

These tables demonstrate that, e.g., Property P6 is suitable for simplification of the data mining process. We can simplify the data mining process provided we have a suitable set $\mathcal{E}$ possessing associations from the right side of the first table. Then it would be sufficient to mine only for associations from the left side of that table. For example, in the first rows we can see the associations
"IF *Hour* is $ML\ Me$ THEN *Temp* is $ML\ Me$."
"IF *Temp* is $ML\ Me$ THEN *Y_NO2* is $ML\ Me$."
Then we immediately obtain another association
"IF *Hour* is $ML\ Me$ THEN *Y_NO2* is $ML\ Me$".

6. Conclusions, future work

In this paper we studied relations among attributes (see the beginning of Section 4) of a given data set that are given by common fuzzy confirmation measures. Within this contribution (and also [13]) we provided a short survey of found relations and counterexamples (see Subsection 4.8).

The most promising and detailed results can be obtained for the minimum-based confirmation measure. This jointly with other facts (e.g., together with less computational complexity of such confirmation measures) gives another argument for its use. According to our experience, the remaining confirmation measures are rather complex to use, especially when we use several attributes in considered expressions.

We realize that our results are rather negative and that there still are some open tasks devoted even to Properties P1-P8 considered in this contribution. On the other side, this contribution is, in fact, our first step in this area. We intend to extend our research, e.g., to study relations given by other properties of confirmation measures, for instance, by those admitting various dependencies as it was suggested in [7], to create novel algorithms for mining of linguistic associations using the knowledge we have discovered, or to specify formally how the user of the data mining process can cooperate with this process by its linguistically expressed knowledge.

This contribution partially contains results (without proofs) from [12] where the same problem was studied. But, since some tasks from [13] remained open therein, we have answered them within this paper - see Examples 1,2,3,4,5, Lemmas 1,2,4,8 and Corollary 1.

7. Acknowledgement

This work was supported by Research center 1M0572 "Data - Algorithms - Decision making" of the Ministry of Education of the Czech Republic.

References

[1] R. Agrawal and R. Srikant, Fast algorithms for mining association rules, Morgan Kaufmann Newport Beach, *In Proceedings of the* 20^{th} *International Conference on Very Large Databases*, (CA: AAAI Press 1994), pages 487–499, Santiago (Chile), 1994.

[2] W.W. Armstrong, Dependency structures of database relationships, *Proceedings of IFIP 1974* (IFIP 1974), pages 580–583, August 5-10, Amsterdam (North- Holland), 1974.

[3] R. Bělohlávek and V. Vychodil, Fuzzy attribute logic over complete residuated lattices, *Journal of Experimental & Theoretical Artificial Intelligence*, 18:471–480, 2006.

[4] M. Delgado, N. Marín, D. Sánchez and M.-A. Vila: Fuzzy Association Rules: General Model and Applications, *IEEE Transactions on Fuzzy Systems*, 11:214-225, 2003.

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

[6] D. Dubois and H. Prade, What are fuzzy rules and how to use them, *Fuzzy Sets and Systems*, 84:169–186, 1996.

[7] D. H. Glass, Fuzzy confirmation measures, *Fuzzy Sets and Systems*, 159:475 – 490, 2008.

[8] F. Guillet, H. J. Hamilton, *Quality measures in Data Mining*, Springer, Vol. 43 Springer 2007, ISBN 978-3-540-44911-9.

[9] P. Hájek: The question of a general concept of the guha method, *Kybernetika*, 505–515, 1968.

[10] P. Hájek and T. Havránek, *Mechanizing Hypothesis Formation, Mathematical Foundations for a General Theory*, Springer-Verlag, Berlin/Heidelberg/New York, 1978.

[11] E. Hüllermeier, Implication based fuzzy association rules, Springer-Verlag, *Proceedings of the* 5^{th} *European conference on principles and practice of knowledge discovery in databases* (PKDD 2001), pages 241–252, September 3-5, Freiburg (Germany), 2001.

[12] J. Kupka and I. Tomanová, Dependencies among attributes given by fuzzy confirmation measures, *Expert Systems with Applications*, submitted.

[13] J. Kupka and I. Tomanová, Some Extensions of Mining of Linguistic Associations, *NEURAL NETW WORLD*, 20:27–44, 2010.

[14] V. Novák, I. Perfilieva, A. Dvořák, Q. Chen, Q. and P. Yan, Mining pure linguistic associations from numerical data, *International Journal of Approximate Reasoning*, 4–22, 2008.

[15] V. Novák, I. Perfilieva and J. Močkoř, *Mathematical principles of fuzzy logic*, Kluwer Academic Publishers, Dordrecht, 1999.

[16] J. Rauch, Logic of Association Rules, *Applied Intelligence*, 22:9 – 28, 2005.