Journal's Title, Vol. x, 20xx, no. xx, xxx - xxx
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/

Confirmation Measures in a fuzzy-set-valued database: a focus on the neutrality scenario

Andrea Ellero

Venice School of Management
Ca' Foscari University of Venice

Paola Ferretti

Department of Economics
Ca' Foscari University of Venice

This article is distributed under the Creative Commons by-nc-nd Attribution License. Copyright © 20xx Hikari Ltd.

Abstract

Fuzzy Confirmation Measures (FCMs) assess the extent to which a fuzzy antecedent A supports a fuzzy conclusion B in a fuzzy association rule $A \Rightarrow B$. We investigate their use in fuzzy databases, where the record fields contain membership degrees associated with fuzzy attributes. In particular, we analise their properties in the neutrality case, associated to a value equal to zero of the FCM.

Mathematics Subject Classification: MSC2020: Primary 03E72; Secondary 68P15, 94D05, 68T37.

Keywords: Fuzzy Confirmation Measures, Bayesian Confirmation Measures, Fuzzy-set-database, neutrality

1 Introduction

Quantifying the degree to which an antecedent in a rule supports a conclusion has been the focus of a huge literature from various fields, including philosophy, statistics, and computer science. Modeling relationships in databases, including their extraction through data mining in the form $E \Rightarrow H$ linking evidence, or premise, E to the hypothesis, or conclusion, H, is essential for

informed decisions, data description or to uncover hidden patterns (see, e.g., [1], [5], [7]). In a database, a strong relationship is discovered when a large number of records possessing attribute E also possess attribute H, suggesting the presence of a regularity. Bayesian Confirmation Measures (BCMs) are commonly used to assess the relevance of an inductive rule, i.e., the extent to which a premise supports a conclusion. These measures take into account the prior probability $P(H)$, the posterior probability $P(H|E)$, and the probability of the antecedent $P(E)$ (see Section 2). Moving to a fuzzy-set framework makes it possible to consider database records that satisfy attributes only to a certain degree, thereby providing greater flexibility in describing the information contained in the data than in the crisp case. This can be achieved by using fuzzy association rules $A \Rightarrow B$, where A (we call it, again, the premise) and B (the conclusion) are fuzzy sets ([3], [4]). To construct measures for evaluating the relevance of a fuzzy association rule, the notion of probability used in the crisp case suitably generalised. We rely on the confidence of the rule, which provides the basis for defining Fuzzy Confirmation Measures (FCMs, see Section 2.2). In Section 3 we provide the computation of FCMs measures using very simple database examples that introduce the special case of neutrality scenario, further discussed in Section 4. Section 4 also compares the neutrality scenarios for FCMs and BCMs. Section 5 provides some conclusions.

2 Confirmation Measures

2.1 Bayesian Confirmation Measures

Bayesian Confirmation Measures (BCMs) are commonly used to assess the relevance of inductive rules. They compare the prior probability $P(H)$ with the posterior probability $P(H|E)$. Observing E confirms H whenever $P(H|E) > P(H)$, disconfirms it whenever $P(H|E) < P(H)$, and is neutral whenever $P(H|E) = P(H)$. Accordingly, a Bayesian Confirmation Measure $c(E, H)$ is a real-valued function such that:

$$\begin{cases} c(E,H) > 0, & \text{if } P(H|E) > P(H) \quad \text{(confirmation case)} \\ c(E,H) = 0, & \text{if } P(H|E) = P(H) \quad \text{(neutrality case)} \\ c(E,H) < 0, & \text{if } P(H|E) < P(H) \quad \text{(disconfirmation case)} \end{cases}$$

A couple of typical examples of BCMs are

$$d(E,H) = P(H|E) - P(H)$$

described by Carnap [2] and

$$K(E,H) = \frac{P(E|H) - P(E|\neg H)}{P(E|H) + P(E|\neg H)}$$

defined by Kemeny and Oppenheim [6].

2.2 Fuzzy Confirmation Measures in a fuzzy database

The idea of probability used in the crisp scenario must be substituted in order to develop measures that allow the assessment of the relevance of a fuzzy association rule, $A \Rightarrow B$ between fuzzy sets A and B.

To get more specific, let us review some fuzzy sets notations. Given a set of records $X = \{x_1, x_2, ..., x_n\}$, the database, consider two attributes of the records, say A and B, where each attribute is a fuzzy set defined by a membership function, $A, B : X \Rightarrow [0, 1]$, so that $A(x_j)$ and $B(x_j)$ measure the extent to which record x_j possesses each attribute. The standard negator $n(A) : X \Rightarrow [0, 1]$, defined as $n(A(x_j)) = 1 - A(x_j)$, is the fuzzy complement set $n(A)$ of A. When the membership functions can take only integer values (0 or 1) returns us to the crisp case.

We now focus on fuzzy association rules and on their evaluation. To this end, we adopt the approach proposed by Glass [4], in which conditional probability is replaced by fuzzy confidence:

$$conf(A \Rightarrow B) = \frac{\sum_{j=1}^{n} A(x_j) \otimes B(x_j)}{\sum_{j=1}^{n} A(x_j)}$$

where $\otimes$ denotes a t-norm used for fuzzy set intersection [1]. The definition of a BCM leads to Glass's [4] definition of a Fuzzy Confirmation Measure (FCM):

Definition 2.1. *A Fuzzy Confirmation Measure $fc(A, B)$ is a real-valued function such that:*

$$\begin{cases} fc(A,B) > 0, & \text{if } conf(A \Rightarrow B) > conf(n(A) \Rightarrow B) \quad \textit{(confirmation case)} \\ fc(A,B) = 0, & \text{if } conf(A \Rightarrow B) = conf(n(A) \Rightarrow B) \quad \textit{(neutrality case)} \\ fc(A,B) < 0, & \text{if } conf(A \Rightarrow B) < conf(n(A) \Rightarrow B) \quad \textit{(disconfirmation case)} \end{cases}$$

Each of the many existing BCMs naturally suggests a corresponding FCM, obtained by replacing the conditional probability $P(B|A)$ with the confidence $conf(A \Rightarrow B)$, and by interpreting $conf(T \Rightarrow A)$ (where T denotes a tautology) as the fuzzy counterpart of $P(A)$. For example, the BCM defined by Carnap as $d(E, H) = P(H|E) - P(H)$ gives rise to the FCM [2]:

$$d(A, B) = conf(A \Rightarrow B) - conf(T \Rightarrow B)$$

while

$$K(A, B) = \frac{conf(B \Rightarrow A) - conf(n(B) \Rightarrow A)}{conf(B \Rightarrow A) + conf(n(B) \Rightarrow A)}$$

is the fuzzy version of the BCM introduced by Kemeny and Oppenheim.

[1] The definition of confidence assumes a non-vanishing denominator.

[2] We will always use the same notation for both a BCM and its corresponding FCM.

3 Examples of FCMs in action

In this Section we give simple examples of using fuzzy confirmation measures in a database. We consider a database $X = \{x_1, x_2, x_3, x_4\}$ that contains four records, and assume that each record describes an individual in terms of two fuzzy attributes: wealth, denoted by B, and sympathy, denoted by A, as in the example reported in Table 1.

individual x_j	sympathy $A(x_j)$	wealth $B(x_j)$
x_1	0.9	0.8
x_2	0.7	0.6
x_3	0.2	0.3
x_4	0.1	0.2

Table 1: Tiny database example - confirmation case

In the following we shall use the product t-norm [4] [8], that is, the scalar product among vectors A and B. This choice allows to rewrite $conf(A \Rightarrow B)$ as

$$conf(A \Rightarrow B) = \frac{A \cdot B}{n\bar{A}}$$

where $A \cdot B$ is the inner product of (the images of) A and B and $\bar{A}$ the mean value of the coordinates of (the image of) A.

Using the product t-norm we have

$$conf(A \Rightarrow B) = \frac{\sum_{j=1}^{4} A(x_j)B(x_j)}{\sum_{j=1}^{4} A(x_j)} = \frac{0.72 + 0.42 + 0.06 + 0.02}{0.9 + 0.7 + 0.2 + 0.1} \approx 0.642$$

Similarly,

$$conf(T \Rightarrow B) = \frac{0.8 + 0.6 + 0.3 + 0.2}{4} = 0.475$$

it follows

$$d(A, B) = 0.642 - 0.475 = 0.167 \ ,$$

which reads that A positively influences B.

To compute $K(A, B)$ we need $conf(B \Rightarrow A)$ and $conf(\neg B \Rightarrow A)$. Since $B \cdot A = 1.22$ and $\sum B(x) = 1.9$, we obtain $conf(B \Rightarrow A) \approx 0.642$. Being $\neg B(x) = 1 - B(x)$ we have $\neg B \cdot A = 0.18 + 0.28 + 0.14 + 0.08 = 0.68$, $\sum \neg B(x) = 2.1$ and thus $conf(\neg B \Rightarrow A) = 0.68/2.1 \approx 0.324$. Therefore

$$K(A, B) \approx \frac{0.642 - 0.324}{0.642 + 0.324} \approx 0.330 \ .$$

Summarizing: $A \Rightarrow B$ is a fairly strong rule ($conf(A \Rightarrow B) \approx 0.642$) and knowing A increases the trust in B ($d(A, B) \approx 0.167 > 0$ and $K(A, B) \approx$

$0.330 > 0$). In other terms being sympathetic confirms having high wealth, in the fuzzy sense.

Since we want to focus on the neutral case, which corresponds to $conf(A \Rightarrow B) = conf(\neg A \Rightarrow B)$, we provide an example that illustrates this scenario. Aside from the trivial case where every record has the same value for attribute B, we can consider a situation like the one shown in Table 2.

.

x_j	sympathy $A(x_j)$	wealth $B(x_j)$
x_1	0.25	0.75
x_2	0.75	0.25
x_3	0.75	0.75
x_4	0.25	0.25

Table 2: Tiny database example - neutrality case

Actually, in this specific instance,

$$conf(A \Rightarrow B) = \frac{\sum_{j=1}^{4} A(x_j)B(x_j)}{\sum_{j=1}^{4} A(x_j)} = \frac{1}{2} = 0.5 \ , \quad conf(T \Rightarrow B) = \frac{2}{4} = 0.5 \ ,$$

so that $d(A, B) = 0$ (and, similarly, we have $K(A, B) = 0$), which indicates that A has no effect on B.

We examine the neutrality scenario in Section 4.

4 The neutrality scenario for FCMs

In the neutrality case, BCMs detect independence between two events E and H, whereas FCMs detect independence between the two fuzzy sets A and B. Clearly, the set of instances where the attributes exhibit neutrality is richer in the fuzzy case than in the crisp case. Let us first consider the FCMs case.

Proposition 1. *The neutrality condition for FCMs (that is $fc(A, B) = 0$) adopting the product t-norm can be rewritten as*

$$A \cdot B = n\bar{A}\bar{B}$$

where

$$\bar{A} = \frac{1}{n}\sum_{j=1}^{n} A(x_j), \qquad \bar{B} = \frac{1}{n}\sum_{j=1}^{n} B(x_j).$$

Proof. Let A and B be two fuzzy sets on $X = \{x_1, \ldots, x_n\}$, with

$$A(x_j), B(x_j) \in [0, 1], \qquad j = 1, \ldots, n.$$

The neutrality condition

$$conf(A \Rightarrow B) = conf(\neg A \Rightarrow B)$$

can be written as

$$\frac{\sum_{j=1}^{n} A(x_j)B(x_j)}{\sum_{j=1}^{n} A(x_j)} = \frac{\sum_{j=1}^{n}(1 - A(x_j))B(x_j)}{\sum_{j=1}^{n}(1 - A(x_j))}.$$

By simplifying, we obtain

$$n \sum_{j=1}^{n} A(x_j)B(x_j) = \left(\sum_{j=1}^{n} A(x_j)\right)\left(\sum_{j=1}^{n} B(x_j)\right).$$

Equivalently,

$$\sum_{j=1}^{n} A(x_j)B(x_j) = \frac{1}{n}\left(\sum_{j=1}^{n} A(x_j)\right)\left(\sum_{j=1}^{n} B(x_j)\right),$$

or

$$A \cdot B = n\bar{A}\bar{B}.$$

□

Observe that, setting $\mathbf{1} = (1, \ldots, 1)$ we can write

$$\sum_{j=1}^{n} A(x_j) = A \cdot \mathbf{1}, \qquad \sum_{j=1}^{n} B(x_j) = B \cdot \mathbf{1},$$

and the condition $A \cdot B = n\bar{A}\bar{B}$ can be rewritten as

$$A \cdot B = \frac{1}{n}(A \cdot \mathbf{1})(B \cdot \mathbf{1}) \ .$$

The last condition coincides with

$$(A - \bar{A}\mathbf{1}) \cdot (B - \bar{B}\mathbf{1}) = 0 \ ,$$

neutrality corresponds therefore to the case in which the centered membership-degree vectors

$$A - \bar{A}\mathbf{1} \quad \text{and} \quad B - \bar{B}\mathbf{1}$$

are orthogonal in $\mathbb{R}^n$. Equivalently, the membership-degree vectors A and B have zero empirical covariance, or, in statistical terminology, they are uncorrelated [8].

Let's see how the condition mentioned in Proposition 1 can be reexamined in relation to BCMs. BCMs can be represented using membership functions

A and B that can only take one of the integer values 0 and 1. The possibility to have neutrality are then rather restricted. For example, from $A \cdot B = n\bar{A}\bar{B}$ we have that $n\bar{A}\bar{B}$ must be an integer. Moreover $\sum_{j=1}^{n} A(x_j)B(x_j) \leq \min\{\sum_{j=1}^{n} A(x_j), \sum_{j=1}^{n} B(x_j)\}$. As one might expect, the orthogonality condition among vectors in R^n is thus subject to constraints that reduce the possibility of having neutrality.

Example 1. *Let us consider the special case $n = 3$. In the fuzzy setting, it is easy to provide nontrivial examples of neutrality, as with $A = (0, 0.5, 1)$ and $B = (0, 1, 0)$, geometrically they are vectors belonging to the unit cube. If, instead, as in the crisp case, $A, B \in \{0, 1\}^3$, i.e. , A and B are vertices of the unit cube, only vectors A and B with coordinates $(0, 0, 0)$ or $(1, 1, 1)$ satisfy the neutrality condition. In fact, assume that both A and B are neither equal to $(0, 0, 0)$ nor to $(1, 1, 1)$, it is necessary that $\sum_{j=1}^{3} A(x_j)B(x_j) \leq \min(\sum_{j=1}^{3} A(x_j), \sum_{j=1}^{3} B(x_j)) \leq 2$. Moreover*

$$3\bar{A}\bar{B} = 3 \cdot \frac{\sum_{j=1}^{3} A(x_j)}{3} \cdot \frac{\sum_{j=1}^{3} B(x_j)}{3}$$

must be an integer, so that $\sum_{j=1}^{3} A(x_j) \cdot \sum_{j=1}^{3} B(x_j)$ must be a multiple of 3, which is inconsistent with the above requirement.

To satisfy $A \cdot B = n\bar{A}\bar{B}$ of Proposition 1 with non trivial examples, the crisp case requires at least 4 records (e.g., consider $A = (1, 1, 0, 0)$ and $B = (1, 0, 1, 0)$).

5 Conclusions

Given a fuzzy rule $A \Rightarrow B$, a Fuzzy Confirmation Measure evaluates the extent to which the data support the rule. Under the product t-norm, a FCM takes value zero, corresponding to the neutrality case, when $A \cdot B = n\bar{A}\bar{B}$. This condition means that the centred membership-degree vectors are orthogonal, or equivalently, that the membership degrees of A and B have zero empirical covariance. In fuzzy databases, non-trivial neutrality cases can arise even in very small datasets, unlike in the binary, or crisp, setting.

Future research may investigate whether analogous characterisations can be obtained adopting other t-norms.

References

[1] Agrawal, R., Imieliński, T., Swami, A. (1993) Mining association rules between sets of items in large databases. In: Buneman P., Jajodia S.

(eds.) ACM SIGMOD international conference on Management of data. SIGMOD 1993. Rec. ACM, New York, NY, USA, 22(2):207–216.

[2] Carnap R. (1950) Logical Foundations of Probability. University of Chicago Press, Chicago.

[3] Celotto E., Ellero A., Ferretti P. (2019) Fuzzy Confirmation Measures (a)symmetry Properties. In: Torra V., Narukawa Y., Pasi G., Viviani M. (eds) Modeling Decisions for Artificial Intelligence. MDAI 2019. Lecture Notes in Computer Science, Springer, Cham, 11676:52–63.

[4] Glass D.H. (2008) Fuzzy confirmation measures. Fuzzy Sets Syst 159(4):475–490.

[5] Greco S., Słowiński R., Szczęch I. (2013) Finding meaningful Bayesian Confirmation Measures. Fundam Inform 127(1-4):161–176.

[6] Kemeny, J., Oppenheim, P. (1952) Degrees of factual support. Philos Sci 19:307–324.

[7] McGarry, K. (2005) A survey of interestingness measures for knowledge discovery. Knowl Eng Rev 20(1):39–61.

[8] Zadeh, L.A. (1968) Probability measures of Fuzzy events. J Math Anal Appl 23(2):421–427.