Interpreting GUHA Data Mining Logic in Paraconsistent Fuzzy Logic Framework

Esko Turunen,
TU Vienna

Intl. Center for Information and Uncertainty

29. 8. 2013

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Applying Boolean logic in data analysis and decision making causes anomalies: the law of the excluded middle is problematic, the use of classical quantifiers $\forall$ (for all) and $\exists$ (there exists) is clumsy and truth and falsehood need not to be each others complements.

To overcome these problems several non–classical logics were born. In various many–valued logics such as mathematical fuzzy logic the law of the excluded middle does not hold in general, in GUHA data mining logic there are several non–classical quantifiers e.g. 'in most cases', and in paraconsistent logic, besides true or false, a statement can be unknown or contradictory, too.

We show how GUHA logic is related to paraconsistent fuzzy logic.

GUHA - General Unary Hypotheses Automaton - introduced by Hájek in 1966 and still developing, is a method of automatic generation of hypotheses based on empirical data, thus a method of data mining. GUHA is a kind of automated exploratory data analysis: it generates systematically hypotheses supported by the data.

The GUHA method is based on well–defined first order monadic logic containing generalized quantifiers on finite models. A GUHA procedure generates statements on association between complex Boolean attributes.

A typical data matrix processed by GUHA has hundreds or thousands of rows and tens of columns. Exploratory analysis means that there is no single specific hypothesis that should be tested by our data; rather, the aim is to get orientation in the domain of investigation, analyze the behavior of chosen variables, interactions among them etc. – Let us see an example!

Contraception data description[B-Course] - Mozilla

http://b-course.hiit.fi/cmcexpl.html

home | library | feedback

Indonesian choice of contraceptive method

B-Course provides a public domain data sets called Contraceptive Method Choice so that one can try out B-Course without own data. Below you can find some details of our example data set.

Some information about the study

This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The samples are 1473 married women who were either not pregnant or do not know if they were at the time of interview. The data contains information about the current contraceptive method choice (no use, long-term methods, or short-term methods) of a woman together with her demographic and socio-economic characteristics.

(1) What implies contraception method?
(2) Are there 'above average' subgroups?

Variables

The data consists of the following ten variables:

10 columns, 1473 rows

1.	Wife's age:	numerical
2.	Wife's education:	1=low, 2, 3, 4=high
3.	Husband's education:	1=low, 2, 3, 4=high
4.	Number of children ever born:	numerical
5.	Wife's religion:	Non-Islam, Islam
6.	Wife's now working?	Yes, No
7.	Husband's occupation:	Categories 1, 2, 3, 4
8.	Standard-of-living index:	1=low, 2, 3, 4=high
9.	Media exposure:	Good, Not good
10.	Contraceptive method used:	No-use, Long-term and Short-term

Reference:
Lim, T.-S., Loh, W.-Y. & Shih, Y.-S. (1999). "A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms". *Machine Learning*. Forthcoming.

B-Course, version 2.0.0 — *CoSCo 2002*

The contemporary logical orthodoxy has it that, from contradictory premises, anything can be inferred. To be more precise, let $\models$ be a relation of logical consequence, defined either semantically or proof–theoretically. Call $\models$ explosive if it validates $\{\alpha, \neg\alpha\} \models \beta$ for every α and β (ex contradictione quodlibet). The contemporary orthodoxy, i.e., classical logic, is explosive, but also some non-classical logics such as intuitionist logic and most other standard logics are explosive. The major motivation behind paraconsistent logic is to challenge this orthodoxy. A logical consequence relation, $\models$, is said to be paraconsistent if it is not explosive. Thus, if $\models$ is paraconsistent, then even if we are in certain circumstances where the available information is inconsistent, the inference relation does not explode into triviality. Thus, paraconsistent logic accommodates inconsistency in a sensible manner that treats inconsistent information as informative.

In Belnap's first order paraconsistent logic (1977), four possible values associated with a formula Φ are true, false, contradictory and unknown:

• if there is evidence for Φ and no evidence against Φ, then Φ obtains the value true and

• if there is no evidence for Φ and evidence against Φ, then Φ obtains the value false.

• A value contradictory corresponds to a situation where there is simultaneously evidence for Φ and against Φ and, finally,

• α is labeled by value unknown if there is no evidence for Φ nor evidence against α.

More formally, the values are associated with ordered couples $T = \langle 1, 0 \rangle$, $F = \langle 0, 1 \rangle$, $K = \langle 1, 1 \rangle$ and $U = \langle 0, 0 \rangle$, respectively.

Perny, Tsoukias and Öztürk introduced a $[0,1]$–valued extension of Belnap's logic: the graded values are computed via

$$t(\Phi) = \min\{\alpha, 1-\beta\}, \tag{1}$$

$$k(\Phi) = \max\{\alpha+\beta-1, 0\}, \tag{2}$$

$$u(\Phi) = \max\{1-\alpha-\beta, 0\}, \tag{3}$$

$$f(\Phi) = \min\{1-\alpha, \beta\}, \tag{4}$$

where $\langle\alpha,\beta\rangle$, called evidence couple, is given; α and β is the degree of evidence of a statement Φ and against Φ, respectively. Moreover, the set of 2×2 evidence matrices of a form

$$\begin{bmatrix} f(\Phi) & k(\Phi) \\ u(\Phi) & t(\Phi) \end{bmatrix}$$

is denoted by $\mathcal{M}$. The values $f(\Phi), k(\Phi), u(\Phi)$ and $t(\Phi)$ are values on $[0,1]$ such that $f(\Phi)+k(\Phi)+u(\Phi)+t(\Phi)=1$. One of the most important features of paraconsistent logic is that truth and falsehood are not each others complements.

We observed 2007 that the operations in (1) – (4) are expressible in the Lukasiewicz structure, which is an example of an injective MV–algebra (not, in general, a Boolean algebra).
Lukasiewicz–Pavelka style fuzzy sentential logic is a complete logic (i.e. a–tautologies and a–provable formulae coincide). We prove that, having any set of injective MV-algebra L valued evidence couples $\langle \alpha, \beta \rangle$, the structure of the evidence matrices

$$\begin{bmatrix} \alpha^* \wedge \beta & \alpha \odot \beta \\ \alpha^* \odot \beta^* & \alpha \wedge \beta^* \end{bmatrix} \tag{5}$$

forms an injective MV–algebra, too. Here the operations $\odot, \wedge$ and * are the algebraic operations product, meet and complement, respectively, of the original injective MV–algebra L. In particular, on the real unit interval $a \odot b = \max\{0, a + b - 1\}$, $a \wedge b = \min\{a, b\}$, $a^* = 1 - a$ for all $a, b \in [0, 1]$. Moreover, in MV–algebras there is an additional operation $\oplus$, in the Lukasiewicz structure it is defined by $a \oplus b = \min\{1, a + b\}$, $a, b \in [0, 1]$.

Our result that continuous valued paraconsistent logic can be seen as a special case of Lukasiewicz–Pavelka style fuzzy logic has a consequence that a rich logical semantics and syntax is available. For example, all Lukasiewicz tautologies as well as Intuitionistic tautologies can be expressed in the framework of this logic. This follows by the fact that we have two sorts of logical connectives conjunction, disjunction, implication and negation interpreted either by the monoidal operations $\bigodot, \bigoplus, \longrightarrow, ^*$ or by the lattice operations $\wedge, \vee, \Rightarrow, ^\star$, respectively (however, neither * nor * is a lattice complementation). Besides, there are many other logical connectives available.

How is this paraconsistent fuzzy logic related to GUHA-logic?

Conceder, for example, the following fancied allergy matrix:

Child	Tomato	Apple	Orange	Cheese	Milk
Anna	1	1	0	1	1
Aina	1	1	1	0	0
Naima	1	1	1	1	1
Rauha	0	1	1	0	1
Kai	0	1	0	1	1
Kille	1	1	0	0	1
Lempi	0	1	1	1	1
Ville	1	0	0	0	0
Ulle	1	1	0	1	1
Dulle	1	0	1	0	0
Dof	1	0	1	0	1
Kinge	0	1	1	0	1
Laade	0	1	0	1	1
Koff	1	1	0	1	1
Olavi	0	1	1	1	1

Here ϕ could mean child is allergic to tomato and apple and ψ could mean child is allergic to milk.

A four–fold contingency table $\langle a, b, c, d \rangle$ related to these attributes is composed from numbers of objects in the data satisfying four different binary combinations of these attributes:

	ψ	$\neg\psi$
ϕ	a	b
$\neg\phi$	c	d

where

- a is the number of objects satisfying both ϕ and ψ,
- b is the number of objects satisfying ϕ but not ψ,
- c is the number of objects not satisfying ϕ but satisfying ψ,
- d is the number of objects not satisfying ϕ nor ψ,
- $m = a + b + c + d$.

Various relations between ϕ and ψ can be measured in the data by different four–fold table quantifiers, denoted by $\phi \sim \psi$, understood as functions with values on $[0, 1]$.

A statement connecting two attributes ϕ and ψ by basic double implicational quantifier is supported by the data if

$$a \geq n \text{ and } \frac{a}{a+b+c} \geq p,$$

where $n \in \mathcal{N}$ and $p \in [0, 1]$ are parameters given by user.

A fuzzy logic interpretation of this quantifier is the following

Given a data, the determining subset A is formed of cases that satisfy ϕ or ψ; there must be enough cases satisfying both of them. The data supports a relation 'ϕ implies ψ and ψ implies ϕ' if there are few cases in A not satisfying ψ or few cases not satisfying ϕ.

Our novel observation is that a value $\alpha = \frac{a}{m}$ can be seen as the degree of evidence that ϕ and ψ occur simultaneously, a value $\beta = \frac{b+c}{m}$ can be seen as the degree of evidence that ϕ and ψ do not occur simultaneously and a value $\frac{d}{m}$ the degree that ϕ and ψ do not occur at all – a kind of indifferent situation. Then

$$\alpha^* \wedge \beta = \beta,\ \alpha \odot \beta = 0,\ \alpha^* \odot \beta^* = \frac{d}{m},\ \alpha \wedge \beta^* = \alpha.$$

Therefore $\langle \frac{a}{m}, \frac{b+c}{m} \rangle$ can be seen as an evidence couple for a statement Φ: 'ϕ and ψ occur simultaneously'. The correspondent evidence matrix is then

$$\begin{bmatrix} f(\Phi) & k(\Phi) \\ u(\Phi) & t(\Phi) \end{bmatrix} = \begin{bmatrix} \frac{b+c}{m} & 0 \\ \frac{d}{m} & \frac{a}{m} \end{bmatrix}.$$

In practical data mining it happens that indifferent cases rule over interesting cases, i.e. value d in a four–fold contingency table is much bigger that values a, b, c. However, even in such cases it is useful to look for statements Φ such that the truth value of Φ is, say at least k times bigger than the falsehood of Φ, i.e. $\alpha \geq k\beta$, which is equivalent to $a \geq k(b+c)$. On the other hand such a statement Φ is stamped by label supported by the data if

$$\frac{a}{a+b+c} \geq p \text{ iff } a \geq \frac{p}{1-p}(b+c).$$

This means $k = \frac{p}{1-p}$, $p \neq 1$, or equivalently $p = \frac{k}{k+1}$. We have

Theorem

Given a data, all statements Φ such that the truth value of Φ is at least k times bigger than the falsehood of Φ in the sense of paraconsistent logic, can be found by using basic double implicational quantifier and setting $p = \frac{k}{k+1}$.

Examples

Consider the above data about children's allergies.
(a) Let ϕ stand for 'child is allergic to tomato and apple' and ψ stand for 'child is allergic to milk'.

Compute the corresponding contingency table, the evidence couple and the evidence matrix for a statement Φ: 'ϕ and ψ occur simultaneously'.

Solution. First write the corresponding table where the connective '&' is interpreted as a Boolean conjunction.

Child	Tomato & Apple	Milk
Anna	1	1
Aina	1	0
Naima	1	1
Rauha	0	1
Kai	0	1
Kille	1	1
Lempi	0	1
Ville	0	0
Ulle	1	1
Dulle	0	0
Dof	0	1
Kinge	0	1
Laade	0	1
Koff	1	1
Olavi	0	1

This leads to

	ψ	$\neg\psi$
ϕ	5	1
$\neg\phi$	7	2

Thus, the evidence couple is $\langle \frac{5}{15}, \frac{7+1}{15} \rangle$ and the correspondent evidence matrix is

$$\left[\begin{array}{cc} f(\Phi) & k(\Phi) \\ u(\Phi) & t(\Phi) \end{array} \right] = \left[\begin{array}{cc} \frac{8}{15} & 0 \\ \frac{2}{15} & \frac{5}{15} \end{array} \right]$$

Since $f(\Phi)$, the degree of falsehood of Φ, is larger that $t(\Phi)$, the degree of truth of Φ, we conclude that the given data does not support the statement that childen who are allergic to tomato and apple are simultaneously allergic to milk, too.

(b) Let ϕ stand for child is allergic to cheese and ψ stand for child is allergic to milk. Compute the corresponding contingency table, the evidence couple and the evidence matrix for the statement Φ: 'ϕ and ψ occur simultaneously'.

Solution. From the original data matrix we get the following contingency table

	Milk	¬Milk
Cheese	8	0
¬Cheese	4	3

Thus, the evidence couple is $\langle \frac{8}{15}, \frac{4+0}{15} \rangle$, and the correspondent evidence matrix is

$$\begin{bmatrix} f(\Phi) & k(\Phi) \\ u(\Phi) & t(\Phi) \end{bmatrix} = \begin{bmatrix} \frac{4}{15} & 0 \\ \frac{3}{15} & \frac{8}{15} \end{bmatrix}$$

We conclude: the truth of cheese allergy and milk allergy occur simultaneously is two times bigger than the paraconsistent falsehood and, thus, the data supports Φ.