Expert Systems With Applications 67 (2017) 107–114

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# Regression analysis based on linguistic associations and perception-based logical deduction

Jiří Kupka, Pavel Rusnok*

*IT4Innovations - Division University of Ostrava - IRAFM, University of Ostrava, 30. dubna 22, 701 03 Ostrava 1, Czech Republic*

ARTICLE INFO

*Article history:*
Received 24 January 2016
Revised 17 August 2016
Accepted 18 August 2016
Available online 20 August 2016

*Keywords:*
Linguistic variables
Implicative fuzzy inference
Regression
IF-THEN rules
Fuzzy associations
Linguistic associations

ABSTRACT

We propose a new generalized model of linguistic variables based on fuzzy partition and its subpartitions. We use this new model for mining relationships between linguistic variables (linguistic associations) from a data set. These relationships can be interpreted as fuzzy IF-THEN rules in the implicative fuzzy inference engine, which is an extended version of the implicative inference called Perception-based Logical Deduction. We show that our extension leads to statistically significant improvements with respect to the previous model used with the help of original and successful Perception-based Logical Deduction. We perform the comparison with different measures of rule quality and five datasets. We can obtain improvements in prediction precision while retaining the interpretability of the models. We also compare our method with the classical machine learning methods and obtain a similar quality of precision, which is very encouraging because interpretability usually leads to worse precision.



## 1. Introduction

This work contributes to highly relevant topics in data analysis - namely, association analysis (e.g., Dubois, Hüllermeier, and Prade, 2006; Kim, Lee, Han, and Yongtae, 2016; Novák et al., 2008; Parkinson, Ward, and Somaraki, 2016 and Sahoo, Goswami, & Das, 2015) and regression (e.g., Ho, Lee, Feng, & Yen, 2012; Yang, Tsoka, Liu, & Papageorgiou, 2016) analyses. In both topics, we consider models allowing expert knowledge to be represented in natural language sentences (Novák, 2008). Usually, the prediction precision of mathematical models and their interpretability are opposed. The increase in the quality of one of the features decreases the other. In this article, we introduce a new mathematical model based on Perception-based Logical Deduction (see Novák, 2005; Novák & Perfilieva, 2004), which is an implicative fuzzy inference mechanism based on linguistics semantics that enables the users to create models described with expressions of natural language. Our mathematical model increases the accuracy of the inference mechanism used in regression analysis while maintaining the underlying linguistic semantics, which are crucial for human-computer interactions.

In this work, we introduce a general method that acts as a regression (e.g., predictive) method. Therefore, after some preprocessing (e.g., learning) steps, it can be used for the prediction of unknown numerical variables. The proposed method is based on a procedure that was successfully used for time series prediction (e.g., in Štěpnička, Dvořák, Pavliska, & Vavříčková, 2011; Štepnička, Cortez, Donate, & Štepničková, 2013). In these articles, a combination of two methods (the first one used for mining linguistic associations and creating, with their help, a linguistic description of the analyzed dataset, and the second one acting as the inference mechanism using this linguistic description of the dataset) was first used. However, the first mining method elaborated in Novák et al. (2008) had some drawbacks (Kupka & Tomanová, 2010) resulting mainly from the original model of evaluative linguistic expressions and from the transformation allowing to use the first data mining method GUHA (Hájek & Havránek, 1978). Therefore, the authors of Kupka and Tomanová (2010) proposed another model of linguistic expressions, which can be based on fuzzy partition and which (for some parameters) naturally extends the original model of linguistic expressions from Novák et al. (2008).

Below in Section 2, we elaborate on a new method based on fuzzy partitions of each variable. Compared with Štěpnička et al. (2011) and Štepnička et al. (2013), we have changed the model of fuzzy sets and the mining procedure, and we have also had to adapt the inference mechanism (originally called *Perception-based Logical Deduction*) because this inference method was built on the original model of linguistic expressions (see Novák et al., 2008 and references therein). Further, we demonstrate (Example 2) how our new model extends the original model of linguistic expressions and also how the deductive process generates predicted values from mined linguistic associations (Section 2.7).

* Corresponding author.
*E-mail addresses:* Jiri.Kupka@osu.cz (J. Kupka), Pavel.Rusnok@osu.cz (P. Rusnok).

http://dx.doi.org/10.1016/j.eswa.2016.08.053
0957-4174/

In the second part of this work (Section 3), we provide four tests. In the first one, we compare the proposed method with the original method from Štěpnička et al. (2011) and mostly obtain significant improvements in the prediction. In the second part, we demonstrate the influence of so-called lift on the final prediction and demonstrate that we again obtain some improvements. In the last but one part of our experiments, we demonstrate the use of a more detailed fuzzy partition in the proposed method. Finally, we also compare the proposed method to some standard approaches and show that we can obtain comparable results.

Here, we would like to stress two significant advantages of our approach. Our experiments confirmed that we not only improved the method that was recently successfully used for time series prediction (e.g., Štěpnička et al., 2011; Štepnička et al., 2013) but also proposed a method whose ability in prediction is fully comparable to standard non-fuzzy approaches. Additionally, as an important side effect of our method, we can obtain linguistic descriptions of the dataset and also of the deductive process. This may be useful in further human-computer interactions.

The second advantage is that we propose a really flexible and general model. For some parameters (e.g., for three subpartitions - see Example 2), we can obtain an extension of the method from Štěpnička et al. (2011). However, the freedom in choosing the number of subpartitions makes the proposed method truly flexible. We can usually lose the linguistic origin of mined associations; however, we should obtain better precision in prediction in the case of need. In this case, the procedure of mining associations and the deductive process itself remain the same. Consequently, this can be immediately used in several tasks; for instance, we can compare how much in terms of prediction we can gain if we avoid using linguistic descriptions of variables, or the method itself allows fuzzy and non-fuzzy partitions and their influence to be compared.

## 2. Methods, notation

In this section, we introduce our proposal. The proposed method is partially based on some older ideas and techniques, and these are introduced in this section as well. Namely, in Section 2.2, a generalization of the model of evaluative linguistic expressions from Novák et al. (2008) is introduced. This new model of linguistic expressions requires another definition of specificity ordering (Section 2.3), and perception-based logical deduction also must be adapted to it (Section 2.6). In the remaining subsection, we merely repeat some background necessary for our method.

### 2.1. Fuzzy mathematics

Let us briefly recall a few elementary notions from fuzzy mathematics. By $\mathbb{N}$, $I$ and $\mathbb{R}$, denote the set of natural numbers, the interval [0, 1] and the set of real numbers, respectively. A *fuzzy set* $A$ on a universum $[a,b]\subseteq\mathbb{R}$ is a map $A$: $[a, b] \to I$ (notation $A \subset\!\!\!\sim\, [a,b]$). The family of fuzzy sets on $[a, b]$ is denoted by $\mathcal{F}([a,b])$. Because a fuzzy set $A$ is defined as a map, all notions related to maps (such as continuity, uniform continuity, and upper semi-continuity etc.) may be considered for fuzzy sets as well. Further, for any $\alpha \in (0, 1]$, an *α-cut* $[A]_\alpha$ of $A$ is defined by

$$[A]_\alpha := \{x \in [a,b] \mid A(x) \geq \alpha\}.$$

A *support* $supp(A)$ of a given fuzzy set $A$ is usually defined as

$$supp(A) = \overline{\{x \in [a,b] \mid A(x) > 0\}}$$

where $\overline{\{\ldots\}}$ stands for a topological closure. A *kernel* of the fuzzy set $A$ is defined as a set of all points $x \in [a, b]$ for which $A(x) = 1$.

A *t-norm* $\otimes$ is a binary map $\otimes$: $I \times I \to I$ that is commutative, associative, and monotone in each argument, and $\otimes(1, \cdot)$ is the identity map. The most widely known examples of t-norms are the minimum, the product and Łukasiewicz t-norms $\otimes_M = \wedge, \otimes_P$ and $\otimes_L$. A *Łukasiewicz implication* $\to_L$ is defined as $a \to_L b = \min(1, 1 - a + b)$. Finally, a *negation* is a map $\neg$: $I \to I$ defined by $\neg(x) = 1 - x$ (Fig. 1).

A *fuzzy number* $A$ on $[a, b]$ is a function $A : [a,b] \to \mathbb{I}$ for which each $\alpha$-cut is a nonempty, closed (possibly degenerated) interval. Let $a = x_1 < x_2 < \cdots < x_p = b,\ n \geq 2$, be fixed *n*odes within $[a, b]$. Fuzzy sets $A_1, \ldots, A_p : [a,b] \to I$ establish a *fuzzy partition* of $[a, b]$ if they fulfill the following conditions (for simplicity, let $x_0 = x_1$ and $x_p = x_{p+1}$):

1. for every $k = 1, \ldots, p$, $supp(A_k) = [x_{k-1}, x_{k+1}]$;
2. for every $k = 1, \ldots, p$, $A_k$ is continuous on $[x_{k-1}, x_{k+1}]$;
3. for every $x \in [a, b]$, $\sum_{k=1}^{p} A_k(x) = 1$;
4. for every $k = 1, \ldots, p$, $A_k(x_k) = 1$.

Within this work, fuzzy numbers are called *convex fuzzy sets* as well.

### 2.2. Linguistic expressions

In this subsection, we introduce a general model based on fuzzy partitions of intervals. Within each fuzzy partition, we allow merging of two neighboring convex fuzzy sets, which results in another convex fuzzy set. Each such fuzzy set could represent its own linguistic expression. We also allow such merging only within some fuzzy subpartitions of the original fuzzy partition, which, see Example 2, can help us generalize the model of linguistic expressions elaborated in Novák et al. (2008). For $[a,b] \subseteq \mathbb{R}$, consider a finite sequence of points $\{a_i\}_{i=1}^{p},\ a_1 = a < a_2 < \ldots < a_{p-1} < a_p = b$. Then, for $i, j \in \{1, 2, \ldots, p\},\ i < j$, $P(i, j)$ denote a fuzzy partition on $[a_i, a_j]$. For a fuzzy partition $P(i, j) = \{\mathcal{A}_k\}_{k=i}^{j}$, one can consider fuzzy sets of the following form, for $k, l \in \{i, i+1, \ldots, j\},\ k \leq l$,

$$A(k,l)(x) = \begin{cases} \mathcal{A}_k(x) & x \leq a_k, \\ 1 & a_k \leq x \leq a_l, \\ \mathcal{A}_l(x) & a_l \leq x. \end{cases}$$

Note that $A(i, i)$ is allowed, and then $A(i, i) = \mathcal{A}_i$. Clearly, for each fuzzy partition $P(i, j)$, $i < j$, the *convex hull* $C(i, j)$ of the fuzzy partition $P(i, j)$ can be defined in the following way (Fig. 2):

$$C(i,j) = \{A(k,l) \mid i \leq k \leq l \leq j\}.$$

Our initial situation can be described as follows. For a given fuzzy partition $P(1, p)$ on $[a, b]$, one can choose finitely many *fuzzy subpartitions* $P(i_1, i_1 + p_1),\ P(i_2, i_2 + p_2),\ \ldots,\ P(i_r, i_r + p_r)$ (and related convex hulls $\mathcal{C}_1 := C(i_1, i_1 + p_1),\ \mathcal{C}_2 := C(i_2, i_2 + p_2),\ \ldots,\ \mathcal{C}_r := C(i_r, i_r + p_r)$), where $i_1 < i_2 < \ldots < i_r$ and $\sum p_i = p$. The system $\mathcal{C}^{[a,b]} = \bigcup_{i=1}^{r} \mathcal{C}_i$(resp. $\mathcal{C}^X$) of fuzzy sets contained in such convex hulls is called the *linguistic description of the interval* $[a, b]$ (resp. *of the variable* $X$). In the text below, the interval $[a, b]$ is also called the *context* of the variable $X$. In this way, we obtain a general model of fuzzy sets that may represent various linguistic descriptions of the variable $X = [a, b]$. Namely, our work follows the original model of evaluative linguistic expressions (resp. descriptions) elaborated, e.g., in Novák et al. (2008). Later in Kupka and Tomanová (2010) (see also Dvořák, Štěpnička, & Štěpničková, 2014), it has been shown that the original model of evaluative expressions might have some drawbacks if it is used in some data mining methods. Therefore in Kupka and Tomanová (2010), a new model of linguistic expressions based on fuzzy partitions was suggested and, in fact, is fully elaborated here. To make this paper more legible, we omit all linguistic background (Novák et al., 2008). However, we would like to stress here that fuzzy sets from $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_r$ on $X$ admit linguistic evaluation and form one of the possible extensions of the original model of evaluative linguistic expressions (see Example 2).