Engineering Applications of Artificial Intelligence 26 (2013) 347–356

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

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

Correlation coefficient of linguistic variables and its applications to quantifying relations in imprecise management data

Shing-Chung Ngan *

Department of Systems Engineering and Engineering Management, City University of Hong Kong, Kowloon, Hong Kong, China

A R T I C L E I N F O

Article history:
Received 19 December 2011
Received in revised form
11 July 2012
Accepted 3 September 2012
Available online 12 October 2012

Keywords:
Correlation coefficient
Linguistic variables
Computing with words
Management data
Applied probability

A B S T R A C T

We frequently use the standard correlation coefficient to quantify linear relation between two given variables of interest in crisp industrial data. On the other hand, in many real world applications involving the opinions of experts, the domain of a variable of interest, e.g. the rating of the innovativeness of a new product idea, is oftentimes composed of subjective linguistic concepts such as *very poor*, *poor*, *average*, *good* and *excellent*. In this article, we extend the standard correlation coefficient to the subjective, linguistic setting, so as to quantify relations in imprecise industrial and management data. Unlike the correlation measures for fuzzy variables proposed in the literature, the present approach allows one to develop a correlation coefficient for linguistic variables that can account for and reflect the conditional dependence assumptions underlying a given data set. We apply the proposed method to quantify the degree of correlation between technology and management achievements of 15 large-scale machinery firms in Taiwan. It is shown that the flexibility of the present framework in allowing for the incorporation of appropriate conditional dependence assumptions to derive a correlation measure for linguistic variables can be essential in approximate reasoning applications.

1. Introduction

When we deal with crisp industrial data, correlation coefficient between two variables of interest (say the efficiency of removing certain chemical compounds versus the amount of catalysts available) is commonly used to quantify the strength of the linear relationship between the two variables. In many real world applications, however, an observation on an entity is oftentimes described in terms of vague and imprecise concepts (e.g., see Alaqtash et al., 2011, Chau, 2005, Cheng and Chau, 2001, Cheng et al., 2002, Lee and Li, 2011, Singh and Benyoucef, 2011, Zouggari and Benyoucef, 2012). Formally, such an observation can be viewed as a fuzzy linguistic variable whose domain is a collection of pre-specified fuzzy linguistic concepts (Herrera and Martinez, 2000a,2000b; Zimmermann, 2001). For example, in evaluating the innovativeness of a new product idea, a fuzzy linguistic variable can be used to represent the rating given by an expert, and the domain of the variable would consist of the fuzzy linguistic concepts *very poor*, *poor*, *average*, *good* and *excellent*. It has been of considerable interest to extend the standard correlation coefficient to the fuzzy setting, so that the correlation between two fuzzy linguistic variables, such as the innovativeness of a product idea and the perceived organizational capacity to realize that product idea, can be quantified (Bustince and Burillo, 1995; Chiang and Lin, 1999; Gerstenkorn and Manko, 1991; Hong and Hwang, 1995; Wang and Li, 1999; Yu, 1993).

As Liu and Kao (2002) pointed out, the works of Bustince and Burillo (1995), Gerstenkorn and Manko (1991), Hong and Hwang (1995), Wang and Li (1999) and Yu (1993) suffered from the deficiency that the correlation coefficient lies in the domain [0,1], as opposed to the conventional range of $[-1,1]$. On the other hand, while the correlation coefficient proposed in Chiang and Lin (1999) is within $[-1,1]$, it is a crisp number and thus the sense of fuzziness is lost. Hence, based on Zadeh′s extension principle, Liu and Kao (2002) developed a new correlation coefficient, which is a fuzzy number when at least one observation is fuzzy. In their method, a mathematical programming approach based on a pair of non-linear programs is used to compute the α-cuts of their fuzzy correlation coefficient. Subsequently, Hong (2006) refined the computational aspect of Liu and Kao′s method such that their fuzzy correlation coefficient can be determined without the aid of mathematical programming. However, the approaches of Liu and Kao (2002) and Hong (2006) remain problematic in that the membership function of the resultant fuzzy correlation coefficient can exceed the range $[-1,1]$, as it was shown in an example in which the method was applied to a ′Fats and Oil′ data set in Hong′s 2006 paper. Furthermore, while some formulations of fuzzy correlation coefficient (e.g., the centroid method of Hung and Wu, 2002) do result in the coefficient lying within the $[-1,1]$

* Tel.: +852 3442 8400; fax: +852 3442 0172.
E-mail address: scngan@cityu.edu.hk

0952-1976/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.engappai.2012.09.009

range, it is not clear how the postulates in those methods translate to what assumptions are actually imposed on the data. In fact, correlation measures for fuzzy variables proposed in the literature commonly do not possess mechanisms to accommodate the possibly different types of conditional dependence relations underlying given data sets (Hsu and Wu, 2010; Hung and Wu, 2007; Leu et al., 2008; Mitchell, 2004, 2006; Park et al., 2009a; Park et al., 2009b; Szmidt and Kacprzyk, 2010; Ye, 2010).

More generally, in working strictly within the framework of Zadeh's extension principle, theory development for fuzzy sets is highly non-trivial. For instance, substantial efforts are still being expended by various researchers in developing methodologies for foundational tasks such as ranking fuzzy numbers and defining arithmetic and logical operations on fuzzy numbers, resulting in a multitude of differing approaches (Abbasbandy and Hajjari, 2009; Chakraborty and Chakraborty, 2006; Chen and Sanguansat, 2011; Chu and Lin, 2009; Chu and Tsao, 2002; Hong, 2007; Kechagias and Papadopoulos, 2007; Kumar et al., 2011; Liu et al., 2008; Matarazzoa and Munda, 2001; Oussalah, 2002; Voxman, 1998). Extending the concept of correlation coefficient to the fuzzy setting is no exception. On the other hand, we have shown that by interpreting the membership function of a linguistic concept based on a probabilistic framework, and by abandoning Zadeh's extension principle in favor of relying on probabilistic arguments, much of the technical difficulties in developing theory involving linguistic variables can be bypassed (Ngan, 2011a, submitted for publication-a), with the resulting probabilistic linguistic computing (PLC) framework enabling a uniform approach for theory development. As this article will show, the PLC framework allows one to develop a correlation measure for linguistic variables that can account for and reflect the conditional dependence assumptions underlying a given data set, in addition to the advantages that the measure does not produce the undesirable artifact having results outside of the range $[-1,1]$, and that the measure is developed methodologically under the well-established axiomatic basis of probability.

A principal application of the proposed method is the analysis of engineering management phenomena, as data from this domain are not uncommon to be of vague and imprecise nature (Park et al., 2009a; Xu, 2012; Yang et al., 2009). In this article, we will apply the method to compute the correlation coefficient on a real industrial data set to quantify the strength of relation between technology and management achievements for 15 large-scale machinery firms in Taiwan, as part of an engineering management study.

This article is organized as follows: the whole of Section 2 will be devoted to reviewing the probabilistic interpretation of a linguistic set and the sampling distribution of such a set (previously described in Ngan, 2011a, submitted for publication-a). Then, the method of deriving correlation measures of a given pair of linguistic variables, which is a distinct development from (Ngan, 2011a, submitted for publication-a) and is built upon the PLC framework delineated in Section 2, will be described in Section 3—the reader is invited to see (Ngan, 2011a, submitted for publication-a) for the applications of the PLC framework to other areas. In Section 4, we will apply the proposed method to an engineering management study concerning large-scale machinery firms in Taiwan. Finally, discussion and suggestions for further development will be given in Section 5.

2. Basic methodologies for linguistic sets

2.1. A linguistic set

The characteristic function of a classical set assigns a value of either 0 or 1 to a given object, indicating whether the object is a member of that set. In contrast, a linguistic set, like the fuzzy set (Zimmermann, 2001), is endowed with a membership function which assigns a value in the range [0,1] to a given object, denoting the grade of membership of that object to the linguistic set. An example of a linguistic set is a so-called triangular linguistic number H, characterized by the membership function

$$f_H(x) = \begin{cases} (x-p)/(q-p) & \text{when } p \le x \le q \\ (x-r)/(q-r) & \text{when } q \le x \le r \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where x is a real number (see Fig. 1(a)). Following Ngan (2011a), for concreteness, let us regard H as representing a Person A's linguistic concept of an average-height individual, with x being the height of the individual under consideration. (Throughout this article, we will use the terms *linguistic set* and *linguistic concept* interchangeably.) Then, that individual with his associated height x will be considered by Person A as having "average height" to a degree determined by the membership function: an individual with height $x=q$ is definitely a member of the concept H according to Person A (as $f_H(q)=1.0$), whereas an individual with height shorter than p is definitely not (as $f_H(x)=0$ for any $x \le p$). In general, the domain of the membership function of a linguistic set can take on $\Re^n$ or other sets. In the special case when the domain is in $\Re$, the corresponding linguistic set is termed a linguistic number. Referring to Fig. 1a, q is a "crisp" number, whereas H is a linguistic number whose value is "around" q. (Note that while triangular membership functions have been frequently used in the fuzzy linguistic computing literature (see Martınez and Herrera, 2012 and the references therein), the subsequent formulation described in this article does not restrict membership functions to be of any particular shape, as we will see in Section 4, where the method is applied to engineering management data).

Thus far, the definition of a linguistic set is identical to that of a fuzzy set. A key difference between the two is that in the rest of this article, theory development for linguistic sets and linguistic variables will be based on probabilistic arguments, in place of Zadeh's extension principle. We therefore introduce the terminology *linguistic set* to distinguish it from *fuzzy set*.

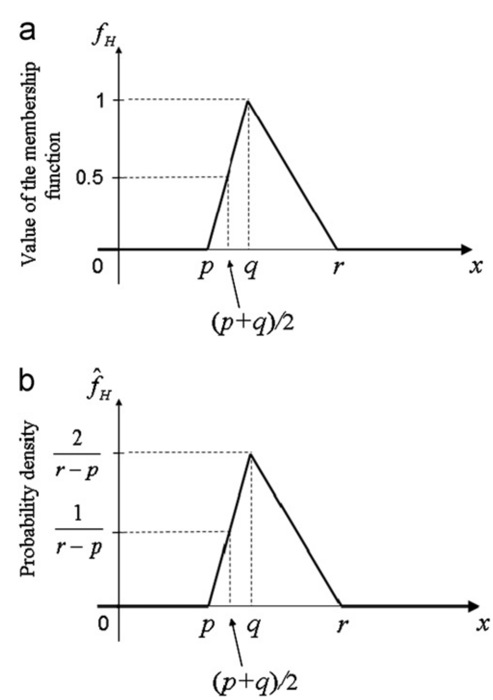

Fig. 1. (a) A triangular linguistic number H representing person A's linguistic concept of an "average height individual". (b) The sampling distribution corresponding to the linguistic number H. The area under the sampling distribution curve is equal to 1.