Fuzzy Information and Data Analysis

Reinhard Viertl
Vienna University of Technology, Austria

INTRODUCTION

The results of data warehousing and data mining are depending essentially on the quality of data. Usually data are assumed to be numbers or vectors, but this is often not realistic. Especially the result of a measurement of a continuous quantity is always not a precise number, but more or less non-precise. This kind of uncertainty is also called fuzziness and should not be confused with errors. Data mining techniques have to take care of fuzziness in order to avoid unrealistic results.

BACKGROUND

In standard data warehousing and data analysis data are treated as numbers, vectors, words, or symbols. These data types do not take care of fuzziness of data and prior information. Whereas some methodology for fuzzy data analysis was developed, statistical data analysis is usually not taking care of fuzziness. Recently some methods for statistical analysis of non-precise data were published (Viertl, 1996, 2003).

Historically fuzzy sets were first introduced by K. Menger in 1951 (Menger, 1951). Later L. Zadeh made fuzzy models popular. For more information on fuzzy modeling compare (Dubois & Prade, 2000).

Most data analysis techniques are statistical techniques. Only in the last 20 years alternative methods using fuzzy models were developed. For a detailed discussion compare (Bandemer & Näther, 1992; Berthold & Hand, 2003).

MAIN THRUST

The main thrust of this chapter is to provide the quantitative description of fuzzy data, as well as generalized methods for the statistical analysis of fuzzy data.

Non-Precise Data

The result of one measurement of a continuous quantity is not a precise real number but more or less non-precise. For details see (Viertl, 2002). This kind of uncertainty can be best described by a so-called *fuzzy number*. A fuzzy number x^* is defined by a so-called *characterizing function* $\xi : I\!R \to [0,1]$ which obeys the following:

$$\exists x_0 \in I\!R : \xi(x_0) = 1 \qquad (1)$$

$\forall \delta \in (0,1]$ the so-called δ–cut $C_\delta[\xi(\cdot)]$ defined by $C_\delta[\xi(\cdot)] := \{x \in I\!R : \xi(x) \geq \delta\} = [a_\delta, b_\delta]$ is a finite closed interval. (2)

Examples of non-precise data are results on analogue measurement equipments as well as readings on digital instruments.

For continuous vector quantities real measurements are not precise vectors but also non-precise. This imprecision can result in a vector $(x_1^*, \cdots, x_k^*)$ of fuzzy numbers x_i^*, or more generally, in a so-called *k-dimensional fuzzy vector* $\underline{x}^*$. Using the notation $\underline{x} = (x_1, \cdots, x_k) \in I\!R^k$ a k-dimensional fuzzy vector is defined by its *vector-characterizing function* $\zeta : I\!R^k \to [0,1]$ obeying

$$\exists \underline{x}_0 \in I\!R^k : \zeta(\underline{x}_0) = 1 \qquad (1)$$

$\forall \delta \in (0,1]$ the δ–cut $C_\delta[\zeta(\cdot)]$ defined by $C_\delta[\zeta(\cdot)] := \{\underline{x} \in I\!R^k : \zeta(\underline{x}) \geq \delta\}$ is a compact simply connected subset of $I\!R^k$ (2)

Remark: A vector of fuzzy numbers is essentially different from a fuzzy vector. But it is possible to construct a fuzzy vector $\underline{x}^*$ from a vector $\xi_1(\cdot), \cdots, \xi_k(\cdot)$ of fuzzy numbers. The vector-characterizing function $\zeta(\cdot, \cdots, \cdot)$ of $\underline{x}^*$ can be obtained by

$$\zeta(x_1,\cdots,x_k)=\min\{\xi_1(x_1),\cdots,\xi_k(x_k)\} \quad \forall (x_1,\cdots,x_k)\in \mathbb{R}^k$$

Examples of *2*-dimensional fuzzy data are light points on radar screens.

Descriptive Statistics with Fuzzy Data

Analysis of variable data by forming histograms has to take care of fuzziness. This is possible based on the characterizing functions $\xi_i(\cdot)$ of the observations x_i^* for $i=1(1)n$. The height h_j^* over a class K_j, $j=1(1)k$ of the histogram is a fuzzy number whose characterizing function $\eta_j(\cdot)$ is obtained in the following way:

For each δ–level the δ–cut $C_\delta[\eta_j(\cdot)]=[\underline{h}_{n,\delta}(K_j),\overline{h}_{n,\delta}(K_j)]$ of $\eta_j(\cdot)$ is defined by

$$\overline{h}_{n,\delta}(K_j)=\frac{\#\{x_i^*: C[\xi_i(\cdot)]\cap K_j\neq\varnothing\}}{n}$$

$$\underline{h}_{n,\delta}(K_j)=\frac{\#\{x_i^*: C_\delta[\xi_i(\cdot)]\subseteq K_j\}}{n}.$$

By the representation lemma of fuzzy numbers hereby the characterizing functions $\eta_j(\cdot), j=1(1)k$ are determined:

$$\eta_j(x)=\max_{\delta\in(0,1]}\delta\cdot I_{C_\delta[\eta_j(\cdot)]}(x) \quad \forall x\in\mathbb{R}$$

The resulting generalized histogram is also called *fuzzy histogram*. For more details compare (Viertl & Hareter, 2004).

Fuzzy Probability Distributions

In standard data analysis probability densities are considered as limits of histograms. For fuzzy data limits of fuzzy histograms are fuzzy valued functions $f^*(\cdot)$ whose values $f^*(x)$ are fuzzy numbers. These fuzzy valued functions are normalized by generalizing classical integration with the help of so-called δ*–level curves* $\underline{f}_\delta(\cdot)$ and $\overline{f}_\delta(\cdot)$ of $f^*(\cdot)$, which are defined by the endpoints of the δ–cuts of $f^*(x)$ for all $x\in Def[f^*(\cdot)]$:

$$C_\delta[f^*(x)]=[\underline{f}_\delta(x),\overline{f}_\delta(x)] \quad \text{for all } x\in Def[f^*(\cdot)]$$

The generalized integral of a fuzzy valued function $f^*(\cdot)$ defined on M is a fuzzy number I^* denoted by

$$I^*=\int\!\!\!\!\!\!-_M f^*(x)dx,$$

which is defined via its δ–cuts

$$C_\delta[I^*]=\left[\int_M \underline{f}_\delta(x)dx,\int_M \overline{f}_\delta(x)dx\right] \quad \forall\,\delta\in(0,1]$$

in case of integrable δ–level curves $\underline{f}_\delta(\cdot)$ and $\overline{f}_\delta(\cdot)$.

Fuzzy probability densities on measurable spaces (M,A) are special fuzzy valued functions $f^*(\cdot)$ defined on M with integrable δ–level curves, for which

$$\int\!\!\!\!\!\!-_M f^*(x)dx=1_+^*,$$

where 1_+^* is a fuzzy number fulfilling

$$1\in C_1[1_+^*] \text{ and } C_\delta[1_+^*]\subseteq(0,\infty) \quad \forall\,\delta\in(0,1].$$

Based on fuzzy probability densities so-called *fuzzy probability distributions* P^* on A are defined in the following way:

Denoting the set of all classical probability densities $\varphi(\cdot)$ on M which are bounded by δ– level curves $\underline{f}_\delta(\cdot)$ and $\overline{f}_\delta(\cdot)$ by

$$S_\delta=\{\varphi(\cdot): \underline{f}_\delta(x)\le\varphi(x)\le\overline{f}_\delta(x) \quad \forall\, x\in M\}$$

the fuzzy associated probability $P^*(A)$ for all $A\in A$ is the fuzzy number whose δ– cuts $C_\delta[P^*(A)]=[\underline{P}_\delta(A),\overline{P}_\delta(A)]$ are defined by

$$\overline{P}_\delta(A)=\sup_{\varphi\in S_\delta}\int_A\varphi(x)dx$$

$$\underline{P}_\delta(A)=\inf_{\varphi\in S_\delta}\int_A\varphi(x)dx.$$

By this definition the probability of the extreme events $\varnothing$ and M are precise numbers, i.e.