International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems

FUZZY (AND INTERVAL) TECHNIQUES IN THE AGE OF BIG DATA: AN OVERVIEW WITH APPLICATIONS TO ENVIRONMENTAL SCIENCE, GEOSCIENCES, ENGINEERING, AND MEDICINE

VLADIK KREINOVICH

Department of Computer Science, University of Texas at El Paso
El Paso, TX 79968, USA, vladik@utep.edu

RUJIRA OUNCHAROEN

Department of Mathematics, Faculty of Science
Chian Mai University, Thailand, rujira.o@cmu.ac.th

Received (received date)
Revised (revised date)

In some practical situations – e.g., when treating a new illness – we do not have enough data to make valid statistical conclusions. In such situations, it is necessary to use expert knowledge – and thus, it is beneficial to use fuzzy techniques that were specifically designed to process such knowledge. At first glance, it may seem that in situations when we have large amounts of data, the relative importance of expert knowledge should decrease. However, somewhat surprisingly, it turns out that expert knowledge is still very useful in the current age of big data. In this paper, we explain how exactly (and why) expert knowledge is useful, and we overview efficient methods for processing this knowledge. This overview is illustrated by examples from environmental science, geosciences, engineering (in particular, aircraft maintenance and underwater robots), and medicine.

Keywords: expert knowledge; fuzzy techniques; environmental science; geosciences; aircraft maintenance; underwater robots; medicine.

1. Introduction

Ideal case when we have enough data. Ideally, all our decisions should be supported by data: we should have enough data to decide which medicine is the most efficient against a given disease; we should have enough data to estimate, based on the previous records, where a tornado will go, etc.

Enough data means, in particular, that we have observed many similar situations in the past, situations when different decisions were made. By analyzing all these situations, for each possible decision, we can estimate the probabilities of different outcomes. We can then uses these probability estimates to select the decision which is the best for a given situations. Numerous *statistical* methods are known for estimating the corresponding probabilities and for making decisions in

such situations.

In practice, we often do not have enough data. In real life, we often do not have enough data to make a reliable decision. In such situations, we depend on *experts* who rely on their experience and on their intuition to make decisions: a medical expert decides how to treat a patient, a skilled pilot decides how best to control the plane in an emergency situation, etc.

Need for fuzzy techniques. In some sense, experts are human measuring instruments. For example, just like sensors can measure a patient's temperature, blood pressure, etc., a medical expert can supplement these measurements with a diagnosis. And just like measurements are imprecise – the measurement result is, in general, somewhat different from the actual (unknown) value – experts' opinions are imprecise. Because of this natural analogy, traditionally, statistical techniques were used to describe expert's uncertainty as well.

In many cases, such an approach works well, but in many other cases, we have a problem: in contrast to a measuring instrument that always returns numbers, experts often formulate their opinion by using imprecise ("fuzzy") words from natural language, like "small", "most probably". To capture the meaning of such words, L. Zadeh came up with the idea of *fuzzy logic*[40]; see also[18,25].

Fuzzy logic: a brief reminder of the main idea. The main idea behind fuzzy logic that, because of fuzziness of natural language, statements like "2.0 is small" are not absolutely true and not absolutely false, the expert has some degree of belief in this statement, and some degree of belief in its negation. In the computer, "true" is usually represented by 1 and "false" by 0. So, to describe intermediate degrees of belief, Zadeh proposed to use real numbers between 0 and 1.

This extension necessitates the extension of the usual "and"- and "or"-operations from the usual 2-valued set $\{0, 1\}$ to the whole interval $[0, 1]$. There are many possible extensions of this type[18,25]. Computationally the simplest are the extensions $f_{\&}(a, b) = \min(a, b)$ and $f_{\vee}(a, b) = \max(a, b)$; operations $f_{\&}(a, b) = a \cdot b$ and $f_{\vee}(a, b) = a + b - a \cdot b$ are also widely used.

At first glance, the role of experts should decrease in the age of big data. Experts' opinion is important when we have few data points. When we have few records of patients with given symptoms, the experience of an expert who encountered several such cases in his practice provides a valuable additional information.

By this logic, the importance of the expert's opinion can be gauged by the ratio of expert's prior situations to recorded fasts: the smaller this ratio, the less important the expert's opinion. In the current era of big data, when we record millions and billions of data points, expert's opinion should be, in most practical situations, of very small value.

Surprisingly, expert opinion is still important. Contrary to the above logic, practice shows that expert knowledge is still very important[35]. For example, in spite of the millions of new meteorological data point coming every day, human experts improve the accuracy of computer temperature forecasts by 10%[10,35] and precipitation forecasts by 25%[9,35]. Moreover, contrary to the expectations, while the amount of metrological data grows with years, these percentages do not change. Similarly, experts improve the accuracy of computer-based economic forecasts by 15%[23,35], and this ratio also does not change with time.

A possible explanation may be that while, e.g., a medical doctor has an experience with only a few patients, the doctor builds on the intuition and knowledge of the whole medical community, including doctors from the past. Thus, in effect, the doctor's opinions are (indirectly) based on a large amount of data.

Problem: what is the best way to handle expert knowledge? Since expert knowledge is important, it is therefore necessary to handle this knowledge. As we have mentioned earlier, there are different methods for handling this knowledge: there are more traditional statistical methods, there are fuzzy methods. Which methods should we use?

What we do in this paper. Our answer to the above question is that both techniques are needed. There are many examples when statistical methods work well. In this paper, we provide practical examples where fuzzy methods bring a definite advantage. There are many such examples; due to the size limitations, we mostly concentrate on examples which are most fundamental and most general. For the same reason, we mostly concentrate on applications that we know best – since we had our own experience with them. Of course, there are many other application areas where fuzzy techniques have an advantage; many of these examples are described in other papers from this special issue.

The paper is structured as follows. In a short next section, we provide a brief comparative overview of probabilistic and fuzzy techniques. In the following sections, on the example of different stages of data processing process, we show how fuzzy techniques can help.

2. Probabilistic and Fuzzy Techniques: A Brief Comparison

Probabilistic techniques: a brief overview. In the probabilistic approach to uncertainty, we estimate the probability of different possible values, and we use these probabilities to make decisions.

In principle, if we have a very large number of observations, we can estimate the probability by computing the corresponding frequencies: e.g., by processing the census data, we can find the probability of a family to have two children. However, in many cases, we do not have that much data. In such situations, two main techniques are applied to find the corresponding distributions.

In some cases, we know the type of the probability distribution – e.g., we know that the distribution is normal. In precise terms, this means that we know a finite-parametric family of distributions, and we are sure that the actual probability distribution belongs to this family. In this case, to get a full description of the distribution, we need to find the values of the corresponding parameters. The most commonly used way to find these values is to select "most probable" values. This is known as the *Maximum Likelihood* approach; see, e.g.,[34].

In other cases, we do not know the type of a distribution. In this case, if several different probability distributions are consistent with the data, it makes sense to avoid fake certainty and to select a consistent-with-data distribution with the largest possible uncertainty. This is known as the *Maximum Entropy* approach; see, e.g.,[17].

Fuzzy techniques: a brief overview. In fuzzy techniques, we start with the natural-language statements that describe the experts' knowledge. We then elicit, from the experts, the degrees describing different natural-language words, and we find the "and"- and "or"-operations that best describe the reasoning of these particular experts[18,25].

3. First Stage of Data Processing: Gauging Measurement Accuracy

Need to gauge accuracy. To properly process data, it is important to know the accuracy of different data values, i.e., the accuracy of different measurement results and expert estimates; see, e.g.,[27,28,30,32]. In many cases, this accuracy information is available, but in many other practical situations, we do not have this information. In such situations, it is necessary to extract this accuracy information from the data itself.

Extracting uncertainty from data: traditional approach. The usual way to gauge of the uncertainty of a measuring instrument is to compare the result $\widetilde{x}$ produced by this measuring instruments with the result $\widetilde{x}_s$ of measuring the same quantity x by a much more accurate ("standard") measuring instrument.

Since the "standard" measuring instrument is much more accurate than the instrument that we are trying to calibrate, we can safely ignore the inaccuracy of its measurements and take $\widetilde{x}_s$ as a good approximation to the actual value x. In this case, the difference $\widetilde{x} - \widetilde{x}_s$ between the measurement results can serve as a good approximation to the desired measurement accuracy $\Delta x = \widetilde{x} - x$.

Traditional approach cannot be applied for calibrating state-of-the-art measuring instruments. The above traditional approach works well for many measuring instruments. However, we cannot apply this approach for calibrating state-of-the-art instrument, because these instruments are the best we have. There are no other instruments which are much more accurate than these ones – and which

can therefore serve as standard measuring instruments for our calibration.

Such situations are ubiquitous; for example:

- in *environmental sciences*, we want to gauge the accuracy with which the Eddy covariance tower measure the Carbon and heat fluxes; see, e.g.,[3];
- in *geosciences*, we want to gauge how accurately seismic[8], gravity, and other techniques reconstruct the density at different depths and different locations.

How state-of-the-art measuring instruments are calibrated: case of normally distributed measurement errors. Calibration of state-of-the-art measuring instruments is possible if we make a usual assumption that the measurement errors are normally distributed with mean 0. Under this assumption, to fully describe the distribution of the measurement errors, it is sufficient to estimate the standard deviation σ of this distribution.

In many situations, we have several similar measuring instruments. For example, we can have two nearby towers, or we can bring additional sensors to the existing tower. In such a situation, instead of a single measurement result $\widetilde{x}$, we have two different results $\widetilde{x}^{(1)}$ and $\widetilde{x}^{(2)}$ of measuring the same quantity x. Here, by definition of the measurement error, $\widetilde{x}^{(1)} = x + \Delta x^{(1)}$ and $\widetilde{x}^{(2)} = x + \Delta x^{(2)}$ and therefore, $\widetilde{x}^{(1)} - \widetilde{x}^{(2)} = \Delta x^{(1)} - \Delta x^{(2)}$.

Each of the random variables $\Delta x^{(1)}$ and $\Delta x^{(2)}$ is normally distributed with mean 0 and (unknown) standard deviation σ (i.e., variance σ^2). Since the two measuring instruments are independent, the corresponding random variables $\Delta x^{(1)}$ and $\Delta x^{(2)}$ are also independent, and so, the variance of their difference is equal to the sum of their variances $\sigma^2 + \sigma^2 = 2\sigma^2$. Thus, the standard deviation σ' of this difference is equal to $\sqrt{2} \cdot \sigma$. We can estimate this standard deviation σ' based on the observed differences $\widetilde{x}^{(1)} - \widetilde{x}^{(2)}$ and therefore, we can estimate σ as $\dfrac{\sigma'}{\sqrt{2}}$.

Specificity of geophysical applications. In the geosciences applications, when we usually have only one seismic map, only one gravity map, etc. In such situations, the above approach does not work, so we need an alternative method.

In such situations, we have several measurement results $\widetilde{x}_k^{(i)}$ of the same (unknown) quantity x_k, with, in general, different standard deviations $\sigma^{(i)}$.

In such situations, the traditional statistical approach does not work. From the statistical viewpoint, a natural idea is to use the Maximum Likelihood method, i.e., to find the unknown values x_j and $\sigma^{(i)}$ which maximize the corresponding likelihood:

$$\prod_i \prod_k \frac{1}{\sqrt{2\pi} \cdot \sigma^{(i)}} \cdot \exp\left(-\frac{\left(\widetilde{x}_k^{(i)} - x_k\right)^2}{2\left(\sigma^{(i)}\right)^2}\right).$$

The problem with this approach is that the above likelihood attains its maximum value – infinity – when one of the values $\sigma^{(i)}$ is equal to 0, and $x_k = x_k^{(i)}$ for this i.

In other words, in this situation, the Maximum Likelihood approach implies that one of the measuring instruments is absolutely accurate. This is definitely not true: we know that the accuracies of different instruments are of the same order of magnitude – otherwise, we would not have observed the improvement when we add less accurate measurements.

Clearly, in this case, we need supplement the purely statistical techniques with expert knowledge.

Expert knowledge can help: idea. The fact that the measurement error is equal to $\sigma^{(i)}$ means that $\widetilde{x}_k^{(i)} \approx x_k$, with accuracy of order $\sigma^{(i)}$. In its turn, the fact that the accuracy is of order $\sigma^{(i)}$ means that the mean square difference $\frac{1}{n}\cdot\sum_{k=1}^{n}\left(\widetilde{x}_k^{(i)} - x_k\right)^2$ is approximately equal to $\left(\sigma^{(i)}\right)^2$ (where n is the total number of measured quantities x_k). The larger n, the more accurate is this approximate equality.

Similarly, the fact that the measurement errors corresponding to different measurements $i \neq j$ means that $\frac{1}{n}\cdot\sum_{k=1}^{n}\left(\widetilde{x}_k^{(i)} - x_k\right)\cdot\left(\widetilde{x}_k^{(j)} - x_k\right) \approx 0$.

We can eliminate x_k from these approximate formulas if we take into account that $\overline{x}_k^{(i)} - \overline{x}_k^{(j)} = \left(\overline{x}_k^{(i)} - x_k\right) - \left(\overline{x}_k^{(j)} - x_k\right)$ and thus,

$$\left(\overline{x}_k^{(i)} - \overline{x}_k^{(j)}\right)^2 = \left(\overline{x}_k^{(i)} - x_k\right)^2 + \left(\overline{x}_k^{(j)} - x_k\right)^2 - 2\left(\overline{x}_k^{(i)} - x_k\right)\cdot\left(\overline{x}_k^{(j)} - x_k\right).$$

By adding the expressions corresponding to different k and taking into account that we already know the corresponding sums, we conclude that

$$e_{ij} \stackrel{\text{def}}{=} \frac{1}{n}\cdot\sum_{k=1}^{n}\left(\overline{x}_k^{(i)} - \overline{x}_k^{(j)}\right)^2 \approx \left(\sigma^{(i)}\right)^2 + \left(\sigma^{(j)}\right)^2.$$

Resulting method of estimating accuracy. For every three measuring instruments, we this get three values e_{ij} for which:

$$e_{12} = \left(\sigma^{(1)}\right)^2 + \left(\sigma^{(2)}\right)^2; \quad e_{13} = \left(\sigma^{(1)}\right)^2 + \left(\sigma^{(3)}\right)^2;$$

$$e_{23} = \left(\sigma^{(2)}\right)^2 + \left(\sigma^{(3)}\right)^2.$$

Here, we have a system of three linear equations with three unknowns, from which we can uniquely determined all three desired variances $\left(\sigma^{(i)}\right)^2$:

$$\left(\sigma^{(1)}\right)^2 = \frac{e_{12} + e_{13} - e_{23}}{2}; \quad \left(\sigma^{(2)}\right)^2 = \frac{e_{12} + e_{23} - e_{13}}{2};$$

$$\left(\sigma^{(3)}\right)^2 = \frac{e_{13} + e_{23} - e_{12}}{2}.$$

Comment. Similar ideas can be used for distributions which are different from normal[32] and for estimating spatial resolution of different maps or images of the same area[26,33].

4. Second Stage: Data Processing Itself

Need for data processing. We are often interested in a physical quantity y that is difficult (or impossible) to measure or estimate directly: distance to a star, amount of oil in a well. A natural idea is to measure y *indirectly*: we find easier-to-measure (or easier-to-estimate) quantities $x_1, \ldots, x_n$ related to y by a known relation $y = f(x_1, \ldots, x_n)$, and then use the results $\widetilde{x}_i$ of measuring x_i to estimate $\widetilde{y}$ as $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$. This is known as *data processing*.

Estimating uncertainty of the results of data processing: a problem. Measurements and expert estimates are never 100% accurate. The actual value x_i of i-th auxiliary quantity can differ from its estimate $\widetilde{x}_i$; in other words, there are *approximation errors* $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$. Because of that, the result $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ of data processing is, in general, different from the actual value y: $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n) \neq f(x_1, \ldots, x_n) = y$. It is desirable to describe the inaccuracy $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$ of the result of data processing. For this, we must have information about the inaccuracy with which we know the values x_i.

Case when we know probabilities. In many practical situations, we also know the *probabilities* of different values Δx_i. It is usually assumed that Δx_i is normally distributed with 0 mean and known standard deviation. In this case, we can use, e.g., Monte-Carlo simulations to estimate the probabilities of different values of Δy.

Sometimes, we do not know the probabilities. In practice, we can determine the desired probabilities by *calibration*, i.e., by comparing the results $\widetilde{x}_i$ of our measuring instrument with the results $\widetilde{x}_i^{\,\text{st}}$ of measuring the same quantity by a "standard" (much more accurate) measuring instrument. However, there are two cases when calibration is not done: (1) cutting-edge measurements (e.g., in fundamental science), when our measuring instrument is the best we have, and (2) measurements on the shop floor, when calibration of measuring instrument is too expensive.

In both cases, the only information we have is the upper bound on the measurement error. In such cases, we have interval uncertainty about the actual values x_i; see, e.g.,[30].

Maximum Entropy (MaxEnt) approach: a brief reminder. Traditional engineering approach to uncertainty is to use probablistic techniques, based on probability density functions (pdf) $\rho(x)$ and cumulative distribution functions (cdf) $F(x) \stackrel{\text{def}}{=} P(X \leq x)$. As we have mentioned, in many practical applications, it is very difficult to come up with the probabilities. In such applications, many dif-

ferent probability distributions are consistent with the same observations. In such situations, a natural idea is to select one of these distributions – e.g., the one with the largest entropy $S \stackrel{\text{def}}{=} -\int \rho(x) \cdot \ln(\rho(x))\, dx$; see, e.g.,[17].

Often, the Maximum Entropy approach works. This approach often leads to reasonable results. For example, for the case of a single variable x, if all we know is that $x \in [\underline{x}, \overline{x}]$, then MaxEnt leads to a uniform distribution on $[\underline{x}, \overline{x}]$. For several variables, if we have no information about their dependence, MaxEnt implies that different variables are independently distributed.

Sometimes, MaxEnt does not work. Sometimes, the results of MaxEnt are misleading. As an example, let us consider the simplest algorithm $y = x_1 + \ldots + x_n$, with $\Delta x_i \in [-\Delta, \Delta]$. In this case, $\Delta y = \Delta x_1 + \ldots + \Delta x_n$. The worst case is when $\Delta_i = \Delta$ for all i, then $\Delta y = n \cdot \Delta$.

What will MaxEnt return here? If all Δx_i are uniformly distributed, then for large n, due to the Central Limit Theorem, Δy is approximately normal, with $\sigma = \Delta \cdot \dfrac{\sqrt{n}}{\sqrt{3}}$.

With confidence 99.9%, we can thus conclude that $|\Delta y| \le 3\sigma$; so, we get $\Delta \sim \sqrt{n}$, but, as we mentioned. it is possible that $\Delta = n \cdot \Delta \sim n$ which, for large n, is much larger than $\sqrt{n}$.

The conclusion from this example is that a traditional statistical approach can be very misleading, especially if we want guaranteed results – and we do want guaranteed results in high-risk application areas such as space exploration or nuclear engineering.

How fuzzy can help: Zadeh's extension principle and interval computations. In a situation when we do not know probabilities, instead of trying to get the probability distribution out of thin air, let us go back and describe what we know. What we know is that each estimate $\widetilde{x}_i$ is close to the actual (unknown) value x_i, with the difference $\Delta x_i = \widetilde{x}_i - x_i$ of order Δ_i.

To process this knowledge, according to the general fuzzy methodology, we fist need to elicit, from the experts, for each pair of values Δx_i and Δ_i, the degree $\mu(\Delta x_i, \Delta_i)$ to which the above statement is true.

The numerical values of Δx_i and Δ_i depend on what measuring unit we use. For example, if we replace meters with centimeters, all the values get multiplied by 100. In general, when we replace the original measuring unit with a new unit which is λ times smaller, we get $\Delta x_i' = \lambda \cdot \Delta x_i$ and $\Delta_i' = \lambda \cdot \Delta_i$. It is reasonable to require that the corresponding degree does not change if we simply change the measuring unit, i.e., that $\mu(\Delta x_i, \Delta_i) = \mu(\lambda \cdot \Delta x_i, \lambda \cdot \Delta_i)$ for all possible values Δx_i, Δ_i, and λ. In particular, if we take $\lambda = \dfrac{1}{\Delta_i}$, we conclude that $\mu(\Delta x_i, \Delta_i) = \mu_0\left(\dfrac{\Delta x_i}{\Delta_i}\right)$, where $\mu_0(x) \stackrel{\text{def}}{=} \mu(x, 1)$.

For some quantities like current, the choice of a sign is also arbitrary; in this sense, it makes sense to require that the degree does not change if we simply change the sign – which changes of Δx_i. This leads to $\mu_0(-x) = \mu_0(x)$.

Clearly, the function $\mu_0(x)$ should be decreasing when $x > 0$.

Now, we need to describe the degrees to which different values y are possible. A value y is possible if *for some* values $x_1, \ldots, x_n$ for which $f(x_1, \ldots, x_n) = y$, the value x_1 is possible, *and* the value x_2 is possible, $\ldots$, and the value x_n is possible. From the logical viewpoint, "for some" means that this property either is true for one tuple, *or* for another tuple, etc. If we use the simplest operations min for "and" and max for "or", we conclude that

$$\mu(y) = \max_{x_1,\ldots,x_n : f(x_1,\ldots,x_n)=y} \min(\mu_1(x_1), \ldots, \mu_n(x_n)),$$

where $\mu_i(x_i) = \mu_0\left(\dfrac{x_i - \widetilde{x}_i}{\Delta_i}\right)$. This formula is known as *Zadeh's extension principle*[18,25].

At first glance, this expressions sounds computationally complex, but it can be simplified if we use *α-cuts* $x_i(\alpha) \stackrel{\text{def}}{=} \{x_i : \mu_i(x_i) \geq \alpha\}$. Specifically, one can easily check that for a continuous function $f(x_1, \ldots, x_n)$ and for functions $\mu_i(x_i)$ which are different from 0 only on a bounded interval, we have $\mu(y) \geq \alpha$ if and only if there exist value $x_i \in x_i(\alpha)$ for which $y = f(x_1, \ldots, x_n)$. Thus, $y(\alpha) = f(x_1(\alpha), \ldots, x_n(\alpha))$, where the *range* in the right-hand side is defined as

$$f(X_1, \ldots, X_n) \stackrel{\text{def}}{=} \{f(x_1, \ldots, x_n) : x_i \in X_1 \,\&\, \ldots \,\&\, x_n \in X_n\}.$$

For the above membership functions, each α-cut $x_i(\alpha)$ is an interval $[\widetilde{x}_i - k(\alpha) \cdot \Delta_i, \widetilde{x}_i + k(\alpha) \cdot \Delta_i]$, where $k(\alpha)$ is the largest value x for which $\mu_0(x) \geq \alpha$. When the sets X_i are intervals, techniques for computing the ranges $f(X_1, \ldots, X_n)$ are known as *interval computations*; there are many efficient techniques for such computations, see, e.g.,[11,14,19,24].

Back to our example: fuzzy techniques have a clear advantage here. In the above example, the range of the function $y = x_1 + \ldots + x_n$ when $x_i \in [\widetilde{x}_i - \Delta, \widetilde{x}_i + \Delta]$ is easy to compute, since this function is increasing in each of its variables. Thus, its largest possible value is attained when each x_i attains its largest value $x_i = \widetilde{x}_i + \Delta$ and is thus equal to $\overline{y} = \widetilde{y} + n \cdot \Delta$. Similarly, the smallest possible value of y is equal to $\underline{y} = \widetilde{y} - n \cdot \Delta$, and possible values of y form an interval $[\widetilde{y} - n \cdot \Delta, \widetilde{y} + n \cdot \Delta]$.

In contrast with the misleading result of the probabilistic approach, this fuzzy-motivated result is in good accordance with the original problem.

First practical application: estimating probability of failure of a complex system. In many real-life applications (e.g., in *aircraft maintenance*), we need to estimate the probability of failure of a complex system (such as an aircraft as a whole or one of its subsystems).

Complex systems are usually built with redundancy allowing them to withstand the failure of a small number of components. It is reasonable to assume that we know the structure of the system. As a result, for each possible set of failed components, we can tell whether this set will lead to a system failure.

Usually, it is assumed that failures of different components are independent events. In this case, if for each component A, we know the probability $P(A)$, then we can use Monte-Carlo simulations to estimate the probability of the system's failure. In practice, however, these probabilities $P(A)$ come from experience and/or from expert estimates and are, therefore, also only known with uncertainty. For example, we may know a confidence interval $[\underline{P}(A), \overline{P}(A)]$ for this probability.

Based on this information, we need to estimate the probability P of the system's failure. In principle, we can use a statistical (MaxEnt) approach, thus assume that the values $P_i(A)$ are independent and uniformly distributed in the corresponding intervals – but that, as we have mentioned, can drastically underestimate the desired failure probability – which, for aircraft maintenance, could lead to a disaster. A safer method is to use fuzzy and interval techniques. The resulting estimates have indeed been successfully applied in aircraft maintenance[12,13,20] and in civil engineering[4].

Comment. Please note that for these particular applications, we have developed algorithms[4,20] which, for these applications, are more computationally efficient than general interval computation techniques[11,14,19,24].

Second practical application: to medicine. Neurological disorders – e.g., the effects of a stroke – affect human locomotion (such as walking). In most cases, the effect of a neurological disorder can be mitigated by applying an appropriate rehabilitation. For the rehabilitation to be effective, it is necessary to be able to correctly diagnose the problem, to assess its severity, and to monitor the effect of rehabilitation; see, e.g.,[5,21,29,38].

At present, this is mainly done subjectively, by experts who observe the patient. This is OK for the initial diagnosis, but for rehabilitation, a specialist can see a patient only so often, and it is definitely desirable to have a constant monitoring of how well rehabilitation works. For such a monitoring, we need to be able to automatically gauge how well the patient progresses based on an automatic observation (measurement) of the patients gait. Measuring the gait is indeed possible. For that, we can attach different sensors to the patient, e.g., inertial sensors that measure the absolute and relative location of different parts of the body during the motion, and electromyograph (EMG) sensors that measure the electric muscle activity during the motion.

We can then record the results $x(t)$ of each sensor during a gait cycle. Based on these observed signals, and on the signals corresponding to healthy patients, we need to gauge how severe is the original gait disorder and gauge whether the current rehabilitation procedure is helping – by comparing the measured gait signals $x(t)$ with the gait signal $x_0(t)$ corresponding to healthy people.

The severity of the disorder is determined by the differences $\Delta x(t) \stackrel{\text{def}}{=} x(t) - x_0(t)$: $S = S(\Delta x(t_1), \ldots, \Delta x(t_n))$, where t_i are moments of time at which measurements were made. The differences $\Delta x(t)$ are usually reasonably small, so we can expand the dependence S in Taylor series and ignore terms which are quadratic or of higher order in Δx_i. As a result, we get a linear dependence $S = \sum_{i=1}^{n} c_i \cdot \Delta x(t_i)$, where c_i are the corresponding partial derivatives.

The problem is that we do not know the exact values of c_i, we only known that they are all bounded by some constant M. Similarly to the above case, we can try both approaches for this problem.

First, we can use the probabilistic (MaxEnt) approach. In this approach, we thus assume that the values c_i are independently uniformly distributed on the interval $[-M, M]$. Under this assumption, similarly to the above simple example, the largest possible value of S (within a given confidence level) is proportional to the standard deviation of S, i.e., equal to $\text{const} \cdot \sqrt{\sum_{i=1}^{n} (\Delta x(t_i))^2} \approx \text{const} \cdot \sqrt{\int (\Delta x(t))^2 \, dt}$.

Alternatively, we can use the fuzzy approach, which, via interval computations, leads to $\text{const} \cdot \sum_{i=1}^{n} |\Delta x(t_i)| \approx \text{const} \cdot \int |\Delta x(t)| \, dt$; see, e.g.,[1].

Empirical data shows that the fuzzy approach is in much better accordance with the doctor's evaluations[1,2,31,39]. This can be explained by the fact that – similarly to airplane maintenance – we want to make sure that the patient performs correctly under all possible circumstances before declaring the therapy a success. Thus, we are interested in making sure that the patient performs well even under the worst-case scenario – while the probabilistic approach, by its very origin, checks only the average-case performance.

5. Possibility of Outliers

Localizing underwater robots: case study. In the above examples, we assume that every measurement result comes from measuring the corresponding quantity. However, there are practical situations when a significant proportion of sensor data are *outliers* that do not measure the intended quantity. Let us describe an example of such a situation: a problem of localization of a mobile underwater robot.

To locate the robot, stationary sonars placed at known locations periodically send a ping signal in all directions; they send signals one after another, so that signals from different sonars do not get mixed up. When the sonar's signal reaches the robot, this signal gets reflected, and part of the reflected signal gets back to the emitting sonar. The sonar then measures the signal's "travel time" t_i as the difference between the emission time and the time when the sonar received the reflected signal. During this travel time, the signal travelled to the robot and back. So, the overall path of the signal is double the distance d_i from the robot to the corresponding sensor i. Once we know the speed of sound v, we can multiply the

measured time interval t_i by this speed, divide by two, and get the distance $d_i = (v \cdot t_i)/2$ to the robot.

In the ideal case, when we know the exact distances, it is sufficient to known three distances d_i to find the exact location of the robot. In practice, because of the ever-present noise, we can only measure the distance d_i with some accuracy.

Usually, the manufacturer's specification for the sonar provide us with the upper bound Δ on the corresponding measurement error (provided, of course, that we are observing the reflection from the robot and not from some other object). Thus, once we know the estimated distance to the i-th sonar, i.e., the value $\widetilde{d}_i = (v \cdot t_i)/2$, then the actual (unknown) distance d_i can take any value from the interval $[\widetilde{d}_i - \Delta, \widetilde{d}_i + \Delta]$. If the signal indeed comes from the robot, then, for each sonar i, we would thus be able to conclude that the robot is located in the zone S_i formed by the two spheres centered around this sonar: the zone between the sphere corresponding to distance $\widetilde{d}_i - \Delta$ and the sphere corresponding to the distance $\widetilde{d}_i + \Delta$. If all the recorded values $\widetilde{d}_i$ corresponded to the robot, then we could find the set S of possible locations of the robot as the intersection of the sets S_i corresponding to all m sonars.

In real life, some measurements do come from reflections from other objects. In this case, some of the sets S_i reflect locations of these other objects. We should therefore take into consideration that some of the measurement are faulty.

In principle, we can use a probabilistic approach here: estimate the probability of each sensor to be an outlier, and then – in accordance with the Maximum Likelihood idea – dismiss measurement results which are most probable to be outliers. However, this approach sometimes leads to misleading results[6,7,36]. The reason for this is similar to previous examples: we assume that all probability distributions are uniform on the corresponding intervals, and, as a result, underestimate the inaccuracy of the computation results. This can potentially lead to a disaster: if we get a wrong impression of the robot's location, it may bump into obstacles and damage itself.

Instead, we used fuzzy and interval computations, and this leads to a much more reliable robot localization[6,7,22,36,37].

Comment. For this problem, in [6,7,22,36], we also developed special (faster) modifications of the general interval techniques from[14,15,16]. For this problem, reducing computation time is very important: when the fast-moving robot is close to the shore, we need to compute its location in real time, to avoid its bumping into numerous near-shore obstacles.

6. Conclusions

In some practical situations – e.g., when treating a new illness – we do not have enough data to make valid statistical conclusions. In such situations, it is necessary to use expert knowledge – and thus, it is beneficial to use fuzzy techniques that were specifically designed to process such knowledge.

At first glance, it may seem that in situations when we have large amounts of data, the relative importance of expert knowledge should decrease. However, somewhat surprisingly, it turns out that expert knowledge is still very useful in the current age of big data.

In this paper, on several practical examples, we show how fuzzy techniques can help on every stage of data processing: when we gauge the accuracy of measurement results, when we actually process the resulting data, and when we take into account that some sensor reading may be outliers. Case studies include examples from environmental science, geosciences, engineering (in particular, aircraft maintenance and underwater robots), and medicine.

Acknowledgements

This work is supported, in part, by Chiang Mai University, Thailand, and by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721.

References

1. M. Alaqtash, T. Sarkodie-Gyan, and V. Kreinovich, "Assessment of functional impairment in human locomotion: a fuzzy-motivated approach", *Proceedings of the 2012 Annual Conference of the North American Fuzzy Information Processing Society NAFIPS'2012*, Berkeley, California, August 6–8, 2012.
2. M. Alaqtash et al., "Application of wearable sensors for human gait analysis using fuzzy computational algorithm", *Engineering Applications of Artificial Intelligence* **24**(6) (2011) 1018–1025.
3. M. Aubinet, T. Vesala, and D. Papale (eds.), *Eddy Covariance – A Practical Guide to Measurement and Data Analysis* (Springer, Dordrecht, Hiedelberg, London, New York, 2012).
4. M. Beer, M. De Angelis, and V. Kreinovich, "Towards efficient ways of estimating failure probability of mechanical structures under interval uncertainty", *Proceedings of the American Society of Civil Engineers (ASCE) Second International Conference on Vulnerability and Risk Analysis and Management ICVRAM'2014 and Sixth International Symposium on Uncertainty Modelling and Analysis ISUMA'2014*, Liverpool, UK, July 13–16, 2014, pp. 320–329.
5. R. Begg, D.T.H. Lai, and M. Palaniswami, *Computational Intelligence in Biomedical Engineering* (CRC Press, Boca Raton, Florida, 2007).
6. Q. Brefort et al., "If we take into account that constraints are soft, then processing constraints becomes algorithmically solvable", *Proceedings of the IEEE Symposium on Computational Intelligence for Engineering Solutions CIES'2014*, Orlando, Florida, December 9–12, 2014, pp. 1–10.
7. Q. Brefort et al., "Towards fast and reliable localization of an underwater object: an interval approach", *Journal of Uncertain Systems* **9** (2015) to appear.
8. J.A. Hole, "Nonlinear high-resolution three-dimensional seismic travel time tomography", *Journal of Geophysical Research* **97** (1992) 6553–6562.
9. Hydro Meteorological Prediction Center, National Oceanic and Atmospheric Association, *HPC % Improvement to NCEP Models (1-Inch Day 1 QPF Forecast)*, http://www.hpc.ncep.noaa.gov/images/hpcvrf/linQPFImpann.gif

10. Hydro Meteorological Prediction Center, National Oceanic and Atmospheric Association, *HPC Pct Improvement vs MOS (Max Temp MAE: Stations Adjusted* ≥ 1 *F)*, http://www.hpc.ncep.noaa.gov/images/hpcvrf/max1.gif
11. Interval computations website http://www.cs.utep.edu/interval-comp
12. C. Jacob et al., "Estimating probability of failure of a complex system based on partial information about subsystems and components, with potential applications to aircraft maintenance", *Proceedings of the International Workshop on Soft Computing Applications and Knowledge Discovery SCAKD2011*, Moscow, Russia, June 25, 2011.
13. C. Jacob et al. "Uncertainty handling in quantitative BDD-based fault-tree analysis by interval computation". *Proceedings of the 5th International Conference on Scalable Uncertainty Management SUM'2011*, Dayton, Ohio, October 10–13, 2011, Springer Lecture Notes in Computer Science, Vol. 6929.
14. L. Jaulin et al., *Applied Interval Analysis* (Springer, London, 2001).
15. L. Jaulin and E. Walter, "Guaranteed robust nonlinear minimax estimation", *IEEE Transaction on Automatic Control* **47**(11) (2002) 1857–1864.
16. L. Jaulin, E. Walter and O. Didrit, "Guaranteed robust nonlinear parameter bounding", *Proceedings of Symposium on Modelling, Analysis and Simulation*, part of *IMACS Multiconference on Computational Engineering in Systems Applications CESA96*, Lille, France, July 9–12, 1996, Vol. 2, pp. 1156–1161.
17. E.T. Jaynes and G.L. Bretthorst, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, UK, 2003).
18. G.J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications* (Prentice Hall, Upper Saddle River, New Jersey, 1995).
19. V. Kreinovich, "Interval computations as an important part of granular computing: an introduction", In: W. Pedrycz, A. Skowron, and V. Kreinovich (eds.), *Handbook on Granular Computing* (Wiley, Chichester, UK, 2008), pp. 3-31.
20. V. Kreinovich et al., "Estimating probability of failure of a complex system based on inexact information about subsystems and components, with potential applications to aircraft maintenance", In: I. Batyrshin and G. Sidorov (eds.), *Proceedings of the 10th Mexican International Conference on Artificial Intelligence MICAI'2011*, Puebla, Mexico, November 26 – December 4, 2011, Springer Lecture Notes in Artificial Intelligence, Vol. 7905, pp. 70–81.
21. D.T.H. Lai, R.K. Begg, and M. Palaniswami, "Computational intelligence in gait research: a perspective on current applications and future challenges", *IEEE Transactions on Information Technology in Biomedicine* **13**(5) (2009) 687–702.
22. F. Le Bars et al., "Interval slam for underswater robots – a new experiment", *Proceedings of the 8th IFAC Symposium on Nonlinear Control Systems NOLCOS'2010*, Bologna, Italy, September 1–3, 2010.
23. S.K. McNees, "The role of judgment in macroeconomic forecasting accuracy", *International Journal of Forecasting* **6**(3) (1990) 287–299.
24. R.E. Moore, R.B. Kearfott, and M.J. Cloud, *Introduction to Interval Analysis* (SIAM Press, Philadelphia, Pennsylvania, 2009).
25. H.T. Nguyen and E.A. Walker, *A First Course in Fuzzy Logic* (Chapman and Hall/CRC, Boca Raton, Florida, 2006).
26. O. Ochoa, M. Ceberio, and V. Kreinovich, "How to describe spatial resolution: an approach similar to the Central Limit Theorem", *Applied Mathematical Sciences* **4**(63) (2010) 3153–3160.
27. O. Ochoa, A. Velasco, and C. Servin, "Towards model fusion in geophysics: how to estimate accuracy of different models", *Journal of Uncertain Systems* **7**(3) (2013) 190–197.

28. O. Ochoa et al., "Model fusion under probabilistic and interval uncertainty, with application to Earth Sciences", *International Journal of Reliability and Safety* **6**(1–3) (2012) 167–187.
29. J. Perry, *Gait Analysis: Normal and Pathological Function* (Slack Inc., Thorofare, New Jersey, 1992).
30. S. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice* (American Institute of Physics, New York, 2005).
31. T. Sarkodie-Gyan et al.,"Measurement of functional impairments in human locomotion using pattern analysis", *Measurement* **44** (2011) 181–191.
32. C. Servin at al., "How to gauge accuracy of measurements and of expert estimates: beyond normal distributions", *Proceedings of 3rd World Conference on Soft Computing*, San Antonio, December 15–18, 2013, pp. 339–346.
33. C. Servin, A. Velasco, and V. Kreinovich, "How to estimate relative spatial resolution of different maps or images of the same area?", *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics SMC'2014*, San Diego, California, October 5–8, 2014, pp. 3507–3511.
34. D.J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures* (Chapman & Hall/CRC, Boca Raton, Florida, 2011).
35. N. Silver, *The Signal and the Noise: Why Some Predictions Fail – but Some Don't* (Penguin Press, New York, 2012).
36. J. Sliwka et al., "Processing interval sensor data in the presence of outliers, with potential applications to localizing underwater robots", *Proceedings of the 2011 IEEE International Conference on Systems, Man, and Cybernetics SMC'2011*, Anchorage, Alaska, October 9–12, 2011, pp. 2330–2337.
37. J. Sliwka et al., "Using interval methods in the context of robust localization of underwater robots", *Proceedings of the 30th Annual Conference of the North American Fuzzy Information Processing Society NAFIPS11*, El Paso, Texas, March 18–20, 2011.
38. D.A. Winter, *Biomechanics and Motor Control of Human Movement* (Wiley-Interscience, Hoboken, New Jersey, 1990).
39. H. Yu et al., "Analysis of muscle activity during gait cycle using fuzzy rule-based reasoning", *Measurement* **43**(9) (2010) 1106–1114.
40. L.A. Zadeh, "Fuzzy sets", *Information and Control* **8** (1965) 338–353.