How to Estimate Amount of Useful Information, in Particular under Imprecise Probability

L. Longpré[1)], **O. Kosheleva**[2)], and **V. Kreinovich**[1)]

[1)]Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968, USA, {longpre, vladik}@utep.edu
[2)]Department of Teacher Education,University of Texas at El Paso, El Paso, TX 79968, USA, olgak@utep.edu

Keywords: *Amount of Information; Entropy; Utility; Imprecise Probabilities; Interval Uncertainty.*

Abstract

When we have a finite set of n alternatives, and we have no information about their probabilities, then a natural way to estimate the amount of uncertainty is to count the number of binary ("yes"-"no") questions needed to uniquely determine the alternative. In situations when we have several similar problems like that, the number of questions per problem is equal to $\log_2(n)$.

When we know the probabilities $p_1, \ldots, p_n$ of different alternatives, a reasonable idea is to consider the *average* number of binary questions needed to uniquely determine the alternative. This average number is equal to Shannon's entropy $-\sum\limits_{i=1}^{n} p_i \cdot \log_2(p_i)$.

In the continuous case, when the unknown(s) can take any of the infinitely many values from some interval, we need infinitely many binary questions to uniquely determine the exact value. To estimate uncertainty, it then makes sense to consider the average number of questions needed to determine each value with a given accuracy $\varepsilon > 0$. For small ε, this number is equal to $S - \log_2(\varepsilon)$, where $S = -\int \rho(x) \cdot \log_2(\rho(x))\, dx$ is the continuous version of Shannon's entropy – which can thus serve as a reasonable measure of uncertainty.

While these measures have many useful applications, in some cases, they are not adequate since they do not distinguish between useful information and information, which is of little importance to the user. For example, when we are trying to protect a salary database, a breach in which the hackers get access to the highest digit of the salary is very undesirable, while a breach in which they learn the smallest digit (i.e., the amount of cents) does not affect the users. In such cases, it is desirable to estimate the amount of *useful* information, i.e., information that affects the utility of different alternatives.

In this talk, we propose such a measure. Specifically, we propose to count the number of binary questions that are needed to determine each of the unknown variables with accuracy sufficient to determine the utility u with a given accuracy $\varepsilon > 0$. In this case, similar to the continuous version of Shannon's entropy, we also get an expression which does not depend on ε and which can therefore serve as a measure of the amount of useful information. We also discuss how to extend this measure to situations when we only have partial information about the probabilities – e.g., when we only know the intervals which contain the actual

(unknown) values of the corresponding quantities.