A Native Measure of Certainty for Unseen Events

Joan Garriga

Dptmnt. de Llenguatges i Sistemes Informàtics,
Universitat Politècnica de Catalunya,
jgarriga@lsi.upc.edu
http://www.lsi.upc.edu/~jgarriga/

Abstract. Dealing with sparsity is still an open question in data mining. As soon as the dimension of the sample space becomes high, the number of unseen events or rare configurations in the sample, contribute a great amount of uncertainty. Existing methodologies offer partial solutions, often based on assumptions about certainly unknown prior distributions. In this work, we present an assumption free approach. We define a statistic that has a clear interpretation in terms of a *measure of certainty*, and we build up a plausible hypothesis, that offers a comprehensible insight of knowledge, with a consistent algebraic structure and a consistent set of properties, yielding a native value of uncertainty for unseen events. This hypothesis is summarized in a set of postulates that characterize such a measure. Also, we face up our measure with some close existing references, mainly, entropy based measures, in order to highlight the contributions of our approach. Finally, we show how this measure is implemented in a general context of statistical modeling.

Key words: data mining, sparsity, statistical modeling, measures of information, entropy

1 Introduction

Many data mining tasks for knowledge discovery, rely on the use of the so called predictive association measures, or more generally, measures of information. Such measures are intended in order to select an optimal model (statistical model selection, graphical modeling, classifiers), an optimal set of rules (association or classification rule mining), an optimal split at each node of a tree (induction of decision trees), or whatever. In any case, they are particular forms of expressing knowledge learned from data, which in this context, means the degree of certainty with respect to the outcome of a random variable. But, regardless to the final objective of the mining process, (let's suppose that no prior knowledge is available), knowledge is invariably and uniquely expressed by occurrences and concurrences of values, observed in the sample. Therefore, such measures intend to asses the amount of information conveyed by any (finite discrete) probability distribution estimated from data. This refers to marginal, as well as conditional, probability distributions.

The main difficulty herein, lies on the concept of knowledge itself. Knowledge is definitely an elusive concept. It is well known that, while learning models from data, one faces a subtle trade off between complexity and accuracy, which may lead to many different combinations of marginal and conditional distributions. Therefore, any assessment of the information conveyed by a probability distribution should be specially sensible to both concepts. [1]

Let's suppose that we are interested in the relation of an attribute X^c, (a sort of a class attribute, or more generally, a consequent set), with respect to a set of other attributes $\Pi = \{X^1, X^2, \ldots, X^m\}$, (a sort of joint explanation of the class, or more generally, a set of antecedents). The most direct way to figure this out, is an implication $\Pi \rightarrow X^c$ and its conditional probability distribution. But we want to emphasize, that we refer to the most general case, in which a set of features define together an input space, (for instance, a decision tree). Let's note, also, that we are not making an explicit distinction between supervised or unsupervised learning.

With respect to Π, we may decide to include or discard some of the features (feature subset selection), and we may have to decide how to partition the input space (discretization and clustering). Be what it may, the joint cardinality of the sample space under consideration, grows geometrically with the cardinality of the features, and can easily become very high. The problem here, is not only that of overfitting. Also, the explanation matrix will probably be very sparse, however large it is the sample, because there will certainly be particular joint configurations of the antecedents that are going to be very rare among the population, or even non-existent. [2]

We refer to these rare configurations, not present in the sample, as *unseen events*. With respect to them, one can adopt two different strategies: to discard them, by estimating a probability of zero, that is, consider that they will never happen, (*identical distribution* assumption), or not to discard them, in which case one should assume a certain amount of probability for them to occur.[3]

In summary, in many cases the sample may not convey all the information about the domain. Particularly, the higher the dimension of the input space, the sparser the input matrix, and this contributes an amount of uncertainty that can not be obviated,

The same argument applies to a validation set. Is it fair to assume that a validation set is representative enough of the whole distribution? Achieving a good rate of classification accuracy in a validation process is, indeed, an indi-

[1] Although some relation with MDL approaches ([7],[12]) may be perceived, we note that we do not specifically advocate in this direction, as it will be shown along this work.

[2] Let's note that, when tackling real world problems, the *iid* assumption does not always hold. In many cases, a real underlying distribution simply does not exist. Then, increasing the sample size will not always overcome this situation. Conversely, we may unnecessary burden, and probably bias, the sample, yielding hardly better results with important additional computational cost.

[3] Also some connections with Good-Turing estimation [6] or with bayesian approaches may be perceived here, and again we note that we are meaning none of this directions.

cation of the goodness of a model, but, up to which degree can we rely on its completeness? Hence, tunning for an extra 1% of accuracy, and/or using such rates as a ranking feature of particular algorithms or methodologies, may not make quite sense.

Among others, Shannon's entropy is the most widely known measure of uncertainty associated to a probability distribution. This measure is uniquely characterized and satisfies some attractive properties. A nice correspondence can be established between these properties and, what is commonly accepted as, a plausible axiomatic definition of knowledge. This is the reason of its success, and the basis of a comprehensive work. But, in relation to the framework above described, entropy's characterization does not attain to cover this aspect of uncertainty.

In this work, we build up a plausible alternative hypothesis, more suitable to deal with the arguments given above. We start up with some axiomatic intuitions about knowledge, from which a direct measure of certainty [4] is derived. This measure is characterized by an analog set of properties to those holding for entropy. But, in our case, the algebra of knowledge is more clearly stated, offering a quite comprehensible insight of knowledge, and contributing some important advantages.

2 Deviation from Minimum Information

In the following, we denote by $\mathcal{P} = (p_1, p_2, \ldots, p_s)$, a generalized finite discrete probability distribution, where $\mathcal{P}$ is a vector of observed frequencies over the set of disjoint dependent events $\Omega = \{e_1, e_2, \ldots, e_s\}$ observed in a sample. [5] Also, we denote by $C_s = (s-1)/s$ and $U_s = 1/s$, what we call, the *certainty* and *uncertainty* factors associated to the cardinality s of the distribution.

The basic idea of our approach is to measure the deviation of any such distribution, with respect to what is commonly called uniformity. Uniformity means equiprobability, which is obviously the most uninformative distribution about the outcome of a random variable. Thus, our interpretation follows straightforward: the larger the deviation, the greater the amount of knowledge expressed by that distribution.

The most direct expression of such deviation is:

$$\Delta(\mathcal{P}) = \sum_{j=1}^{s} (p_j - U_s)^2 \quad . \tag{1}$$

[4] Certainty and uncertainty are indeed quite the same thing: just different degrees of knowledge. But we want to emphasize this aspect, as opposed to entropy based measures, which are measures of uncertainty

[5] We refer to the extended concept of *generalized finite discrete probability distributions*, as expressed in [11]. Such extension allows to consider a simple sequence $p_1, p_2, \ldots, p_n$ of nonnegative numbers such that, $0 < \sum_i^n p_i \leq 1$. We denote, $W(\mathcal{P}) = \sum_i^n p_i$, as the weight of the distribution $\mathcal{P}$. Thus, the weight of an ordinary distribution is equal to 1. A distribution which has a weight less than 1 is called an *incomplete distribution*.

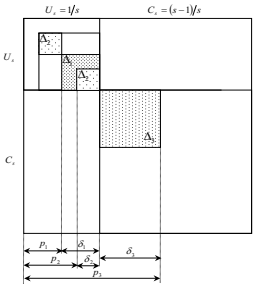


Fig. 1. Graphical depiction of knowledge conveyed by $\mathcal{P} = (p_1, p_2, p_3)$

This deviation is just a statistic, probably biased with respect to population constants. However, uniformity is a clearly defined reference point, independent of either sample space or sampling scheme under consideration. Therefore, stepping aside from any distributional assumptions about the sample space, and focusing exclusively on the information conveyed by the sample, deviation with respect to uniformity yields always a relative idea of the information conveyed by any estimated distribution. In other words, we intend to measure the information conveyed by a distribution, not by its proximity to any assumed underlying distribution, but rather from its deviation with respect to the worst imaginable case, a sort of *absolute* reference baseline. Note that, while proceeding this way, we are not concerned with any kind of validation process.

2.1 Geometric Interpretation

This deviation has a pretty illustrative geometric interpretation. Let's fix a square with an area equal to one and let's imagine that this area represents the absolute knowledge. Let's divide each side at the point corresponding to $1/s$, so that we get two portions, according to the certainty and uncertainty factors, as it is shown in fig.1. We refer to the crossing point as the point of minimum information.

For each p_j, we have a deviation $\delta_j = (p_j - U_s)$, and a square deviation $\Delta_j = (p_j - U_s)^2$. It is easily observed that square deviations are areas relative to the full square, so that we may regard them as a graphical representation of the amount of knowledge contributed by each event e_j.

It is straightforward that: (i) deviations sum up to zero, $\sum_{j=1}^{s} \delta_j = 0$, and (ii) square deviations sum up to $\sum_{j=1}^{s} \Delta_j = \sum_{j=1}^{s} (p_j)^2 - U_s$.

Ideally, the maximum certainty is given when only one particular event is observed in the sample, (let's say e_m, with $p_m = 1$). In this case, the square deviation is maximum, and is equal to $1 - U_s$. Thus, with respect to the absolute knowledge, we still have a lack of knowledge amounting U_s, and what we get is the shadowed areas shown in fig.2 for different values of s.

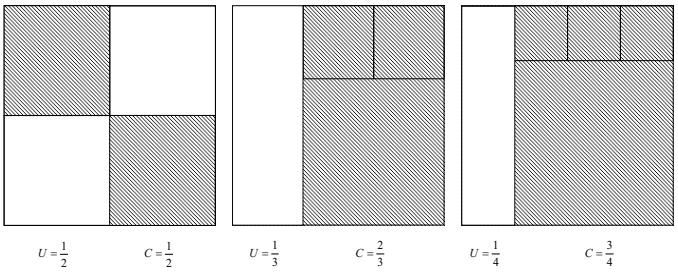


Fig. 2. Areas of certainty and uncertainty for $s = 2, 3, 4$

For each cardinality, the shadowed, and not shadowed areas, represent the relation between certainty and uncertainty with respect to the absolute knowledge given by the full square. Absolute knowledge, or absence of uncertainty, would only be achieved with an infinite cardinality, that is, a continuous feature.

Let's highlight the special consequence of taking measures of deviation with respect to minimum information: a notion of *richness*, or quality, not only quantity, of knowledge, is inherently related to the dimension of the distribution. This results in a very important feature of our measure, which we call the *cardinality scaling of knowledge*.

3 A Measure of Certainty

The minimum knowledge we can have is that given in case of uniformity, expressed by the uncertainty factor. That is, at the point of minimum information we have U_s, and it increases as the square deviations increase. The most direct expression of this idea is,

$$K(\mathcal{P}) = U_s + \sum_{j=1}^{s} \Delta_j = \sum_{j=1}^{s} p_j^2 \tag{2}$$

The geometric interpretation of this expression is shown in fig.3 a), in which the shadowed and not shadowed areas represent the relative measures of certainty and uncertainty associated to an example distribution.

It is straightforward to show that the following properties hold: (i) normalization, (ii) monotonicity (with respect to deviation), (iii) symmetry and (iv) expansibility.

And yet a fifth property holds, in relation to the composition of two successive random variables, as it is graphically shown in fig.3 b): given $\mathcal{P} = (p_1, p_2, \ldots, p_s)$ and $\mathcal{T} = (t, 1-t)$, and their composition $\mathcal{Q} = (t\,p_1, (1-t)\;p_1, p_2, \ldots, p_s)$, we have,

$$K(\mathcal{Q}) = K(\mathcal{P}) - p_1^2\;(1 - K(\mathcal{T})) \tag{3}$$

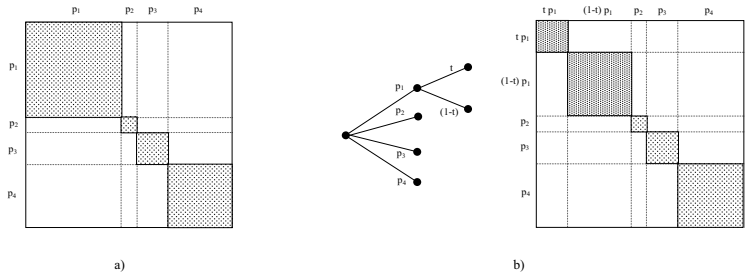


Fig. 3. a) Certainty measure for $\mathcal{P} = (p_1, p_2, p_3, p_4)$; b) Successive composition of $\mathcal{P} = (p_1, p_2, p_3, p_4)$ and $\mathcal{T} = (t, 1-t)$.

This looks quite natural: our knowledge about the final outcome of the successive composition of two distributions, is the certainty of the first distribution except for the additional uncertainty contributed by the second distribution. This relation synthesizes the additive algebra of knowledge that is implicit by our measure of certainty.

3.1 Disjoint Dependent Events

In terms of disjoint dependent events, the composition shown in fig.3 b), can also be expressed as,

$$K\left(t\,p_1, (1-t)\;p_1, p_2, \ldots, p_s\right) = p_1^2\,K\left(t, 1-t\right) + \sum_{j=2}^{s} p_j^2 \tag{4}$$

But we are implicitly assuming that a particular amount of uncertainty, the term U_s in eq.2, is inherently due to the dimension of the distribution. Thus, each p_j^2 is in fact,

$$p_j^2 = \left(U_{s+1} + \left(p_j - U_{s+1}\right)\right)^2 = \left(U_s + \left(p_j - U_s\right)\right)^2$$

Therefore, though apparently independent of cardinality, it is indeed a deviation with respect to minimum information: the term U_s is equally distributed among all possible outcomes of the distribution, yielding a term U_s^2, and the amount really contributed by each e_j is, $\left(p_j - U_s\right)^2 + 2\,U_s\left(p_j - U_s\right)$, from which the second term globally cancels out.

Consequently, if p_j is the observed probability of occurrence of event e_j, our knowledge about the outcome of e_j is given by,

$$K\left(p_j\right) = U_s^2 + \Delta_j^2 \tag{5}$$

which is consequent with eq.2, so that we have, $K\left(\mathcal{P}\right) = \sum_{j=1}^{s} K\left(p_j\right)$.

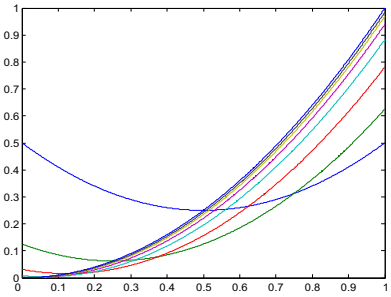


Fig. 4. $K\left(p_j\right)$ for $s = \{2, 4, 8, 16, 32, 64, 128, \infty\}$

Such an additive assumption contributes other remarkable features:

- The measure is explicitly dependent on s, thus expressing the notion of quality of knowledge as a function of the cardinality, as it is shown in fig.4. In the limit, where this measure would hardly apply, knowledge meets (square) probabilities,

$$\lim_{s\to\infty} K\left(p_j\right) = p_j^2 \quad .$$

- The minimum value is coherently given at the point of equiprobability, where,

$$K\left(U_s\right) = U_s^2 \quad .$$

- It is continuous for an observed probability of zero, so that, as long as we correctly estimate the dimension of the probability distribution, it yields a value greater than zero for any unseen event,

$$K\left(0\right) = U_s^2 + U_s^2 \quad .$$

- Conversely, for any event with an observed probability of one, as long as a meaningful cardinality is going to be higher than one, the measure yields a value lower than one,

$$K\left(1\right) = U_s^2 + C_s^2 \quad .$$

- In case of uniformity, or minimum information, we have,

$$K\left(\mathcal{P}\right) = \sum_{j=1}^{s} K\left(p_j\right) = U_s \quad .$$

- In case of maximum deviation, we have maximum knowledge, (which does not mean maximum certainty about the only seen event),

$$K\left(\mathcal{P}\right) = K\left(1\right) + \left(s-1\right) K\left(0\right) = 1 \quad .$$

- and finally, the relation of composition given by eq.3, can be alternatively postulated as,

$$K\left(t\, p_1, \left(1-t\right) p_1, p_2, \ldots, p_s\right) = p_1^2\, K\left(t, 1-t\right) + K\left(p_2, \ldots, p_s\right) \quad . \qquad (6)$$

3.2 Conditional Certainty

Let's denote $(\mathcal{Q}|\ e_i) = \{q_1|\ e_i, q_2|\ e_i, \ldots, q_s|\ e_i\}$ as the observed probability distribution of a random variable, given event e_i, drawn from probability distribution $\mathcal{P} = \{p_1, p_2, \ldots, p_r\}$. Also, let's denote $\Delta_{ji} = (q_j|\ e_i - U_s)^2$ as the square deviations of this conditional distribution. The information conveyed by $(\mathcal{Q}|\ e_i)$ is analogously given by,

$$K\left(\mathcal{Q}|\ e_i\right) = U_s + \sum_{j=1}^{s} \Delta_{ji} = \sum_{j=1}^{s} \left(q_j|\ e_i\right)^2 \tag{7}$$

As desirable, in case of independence, this expression yields, $K\left(\mathcal{Q}|\ e_i\right) = K\left(\mathcal{Q}\right)$, but moving away from that point in either direction, we have a clear expression of dependence in terms of $K\left(\mathcal{Q}|\ e_i\right)$:

- the minimum knowledge we can have is $K\left(\mathcal{Q}|\ e_i\right) = U_s$, which corresponds to the point of uniformity;
- at the point of independence we will have $K\left(\mathcal{Q}|\ e_i\right) = K\left(\mathcal{Q}\right) \geq U_s$;
- from that point on, $K\left(\mathcal{Q}|\ e_i\right) > K\left(\mathcal{Q}\right)$ and we may begin to consider a possible relation of dependence;
- and in case of absolute dependence, $K\left(\mathcal{Q}|\ e_i\right) = 1$.

That is, it would make sense to consider independence for the existing subset of distributions conveying less knowledge than that given in case of independence, an consider higher rates of dependence as long as knowledge is higher.

With respect to the whole distribution $\mathcal{P}$, it makes sense to consider a weighted mean expression of the amount of knowledge conveyed by the conditional distribution given each e_i, that is, [6]

$$K\left(\mathcal{Q}\,|\mathcal{P}\right) = \sum_{i=1}^{r} p_i\ K\left(\mathcal{Q}\,|e_i\right) = U_s + \sum_{i=1}^{r} p_i \sum_{j=1}^{s} \Delta_{ji} = \sum_{i=1}^{r} p_i \sum_{j=1}^{s} \left(q_j|\ e_i\right)^2 \tag{8}$$

In case of independence, eq.8, yields also, $K\left(\mathcal{Q}|\ \mathcal{P}\right) = K\left(\mathcal{Q}\right)$.

3.3 Joint Distributions

Let's denote $(\mathcal{P}, \mathcal{Q}) = \{p_{11}, \ldots, p_{1s}, \ldots, p_{r1}, \ldots, p_{rs}\}$ as the joint probability distribution of two random variables observed in the sample. Also, let's denote $U_{r,s} = U_r\, U_s$, and $\Delta_{ij} = (p_{ij} - U_{rs})^2$ as the square deviations of this joint distribution. The information conveyed by this distribution is,

$$K\left(\mathcal{P}, \mathcal{Q}\right) = U_{rs} + \sum_{i,j}^{r,s} \Delta_{ij} = \sum_{i=1,j=1}^{r,s} p_{ij}^2 \tag{9}$$

[6] Note that, being e_i an unseen event, $K\left(\mathcal{Q}\,|e_i\right) = U_s$. Thus p_i stands here as the basic form of $\theta_i \in \Theta$, a parametrical model, intrinsic to our approach, yielding assumption free priors for unseen events, the description of which does not fit in this paper.

In this case, we have the following expression of dependence in terms of $K(\mathcal{P},\mathcal{Q})$:

- The minimum knowledge we can have is $K(\mathcal{P},\mathcal{Q}) = U_{rs}$, which corresponds to uniformity.
- In case of independence, we have $\sum_{i,j}^{r,s} p_{ij}^2 = \sum_i^r p_i^2 \sum_j^s q_j^2$, thus,

$$K(\mathcal{P},\mathcal{Q}) = K(\mathcal{P})\ K(\mathcal{Q}) \geq U_{rs} \tag{10}$$

  We may refer to this relation as *multiplicativity* of knowledge (certainty), as an analog to the *additivity* of entropy (uncertainty), of two independent random variables.
- From that point on, we may begin to consider a relation of dependence, and following from $\sum_{i,j}^{r,s} p_{ij}^2 = \sum_i^r p_i^2 \sum_j^s (q_j|\ e_i)^2$, we get,

$$K(\mathcal{P},\mathcal{Q}) = \sum_{i=1}^{r} p_i^2\, K(\mathcal{Q}\,|e_i) = \sum_{i=1}^{r} p_i\, K_i(\mathcal{Q}\,|\mathcal{P}) \tag{11}$$

  So, what we have is a particular composition of the marginal and conditional knowledge, that may be regarded as a weighted mean expression of the contributions to $K(\mathcal{Q}\,|\mathcal{P})$ in eq.8 of each particular outcome of the antecedent, (what we denote in eq.11 as $K_i(\mathcal{Q}\,|\mathcal{P})$).
- In case of dependence given any event e_i, we have $\forall e_i\,,\ K(\mathcal{Q}|\ e_i) \geq K(\mathcal{Q})$, therefore,

$$K(\mathcal{P},\mathcal{Q}) = \sum_{i=1}^{r} p_i^2\, K(\mathcal{Q}\,|e_i) \geq \sum_{i=1}^{r} p_i^2\, K(\mathcal{Q}) = K(\mathcal{P})\ K(\mathcal{Q}) \tag{12}$$

  So, also an analog to the *subadditivity* of entropy holds, to which we may refer as *supermultiplicativity* of knowledge about two dependent random variables.
- Finally, in case of absolute dependence, we have $\forall e_i\,,\ K(\mathcal{Q}|\ e_i) = 1$, and consequently, $K(\mathcal{P},\mathcal{Q}) = K(\mathcal{P})$

3.4 Characterization of the Measure

We have shown how our measure of certainty applies to marginal, conditional, and joint distributions, as well as to each elementary event, and we have shown also the properties and relations that hold. Hereof, we can give a set of postulates, characterizing our measure:

1. *$K(\mathcal{P})$ is a symmetric function of the elements of $\mathcal{P}$.*
2. *For each element $\{p\} \in \mathcal{P}$, $K(\{p\})$ is a continuous function of p for $0 \leq p \leq 1$. Note that continuity of $K(\{p\})$ is supposed even for $p = 0$.*
3. *$K(\{1/s\}) = (1/s)^2$ is a minimum.*
4. *$K(t\,p_1, (1-t)\,p_1, p_2, \ldots, p_s) = p_1^2\, K(t,(1-t)) + K(p_2,\ldots,p_s)$ for any distribution $\mathcal{P}$ and for $0 \leq t \leq 1$*

Postulate number 4 is a synthesis of the algebraic structure of knowledge implicit by our measure. Knowledge is defined as the sum of the pieces contributed by disjoint dependent events, and as the (square) weighted sum of knowledge about combined events.

Expansibility is also implicit in this postulate. It is worth mentioning that, with respect to entropy, the connotations of this property are much stronger, because the piece of knowledge contributed by an unseen event is not zero, $K(0) = U_s^2$. Conversely, the piece of uncertainty contributed by an unseen event is the unnatural and weird mathematical artifact $H(0) = -0\,log(\infty) = 0$, so that expansibility follows straightforward.

At this point, it is obvious that an axiomatic definition of knowledge, that explicitly detaches, what is information, from what is probabilities, is underlying our approach. [7]. Some subtle differences exist among these two concepts, which are mainly expressed by the following two intuitions: (i) knowledge can never be zero, and (ii) knowledge is akin to a notion of quality related with the cardinality of the distribution.

On the basis of these differences, we build up a plausible hypothesis that offers a quite comprehensible insight of knowledge, with some notable features:

- it is based on the fact of moving the gravity center to the point of equiprobability, or minimum information;
- it allows for a consistent definition of an algebraic structure of knowledge, from distribution level to elementary event level;
- it yields a cardinality dependent measure of certainty, relating knowledge to a notion of quality;
- it yields a native measure of uncertainty associated to unseen events, as well as of certainty about a single observed event.

4 Close Existing References

An extensive literature exists concerning references which are close to our measure. They all have in common to be classified under taxonomic labels like *indexes of diversity*, or *measures of heterogeneity*. On the other hand, none of them seems to be conceptually close to our posing.

4.1 Diversity Indexes

Several formulations of indexes of diversity (or concentration) have appeared. From its origins, back to Gini's index [4] in 1912, to Yule's Characteristic K of stylistic diversity [17] (1944), Hirschman's index of trade concentration [9] (1945), Simpson's index of diversity [14] (1948), Herfindahl's coefficient of concentration [8] (1950), Greenberg's index of linguistic diversity [5] (1956), or Agresti's variability [2] (1990), among many others.

[7] We don't depict an axiomatic definition of knowledge in this paper, but it is clear that it should be slightly different from Kolmogorov's axiomatization of probability [10], in the context of finite sample spaces

Considering an infinite population such that each individual belongs to one of Z groups, and being $\pi_1, \ldots, \pi_Z$, $(\sum \pi = 1)$, the proportions of individuals in the various groups, the probability of choosing two individuals from group i is π_i^2. Then, $\lambda = \sum_i \pi_i^2$ can be interpreted as the probability that two individuals chosen at random and independently from the population will be found to belong to the same group. Therefore, λ is a measure of the concentration of the classification. This is the most common basis of this group of measures.

While this posing may look quite distant from ours, the formal similarity is clear. Let's note however, that *diversity* embrace two basic aspects of the population, which are exactly the same components that we combine in our measure of knowledge: (i) *richness*, the number of different groups in the population, as an analog to richness, or quality, of knowledge, and (ii) *evenness*, the uniformity in the number of individuals of the various groups, as a baseline reference.

What is most significant to us, is that: (i) any of them can be regarded from the point of view of knowledge about the outcome of the experiment at hand, (ii) the exact formal expression that we derive, is independent of sample size, (for instance, see [14]).

4.2 Entropy Based Measures

Though initially not conceived us such, Shannon's entropy is, by far, the most widely used diversity index. It was introduced in [13], within the framework of communication theory, and quickly reinterpreted by the information theory community, with the following formal definition:

Let $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$ be a finite discrete probability distribution, that is $\forall i\, , p_i \geq 0$, and $\sum_i^n p_i = 1$. The amount of uncertainty of $\mathcal{P}$, that is, the amount of uncertainty concerning an experiment, the possible results of which, have the probabilities $p_1, p_2, \ldots, p_n$, is called the entropy of the distribution, given by,

$$H(\mathcal{P}) = H(p_1, p_2, \ldots, p_n) = \sum_i^n p_i log_2 \frac{1}{p_i} \tag{13}$$

The simplest set of postulates which uniquely characterize the quantity given by eq. 13 is [1]:

1. *$H(\mathcal{P})$ is a symmetric function of the elements of $\mathcal{P}$.*
2. *$H(p, 1-p)$ is a continuous function of p for $0 \leq p \leq 1$.*
3. *$H(1/2, 1/2) = 1$.*
4. *$H(t\, p_1, (1-t)\, p_1, p_2, \ldots, p_n) = H(p_1, p_2, \ldots, p_n) + p_1 H(t, 1-t)$ for any distribution $\mathcal{P}$ and for $0 \leq t \leq 1$.*

The analogy of our measure's characterization, (sec. 3.4), is evident, though a main difference is expressed by postulate 4: while the entropy of combined events is a sum of the weighted uncertainty of the successive events, knowledge is defined such that, the final certainty, is the initial certainty, minus the (square) weighted uncertainty contributed by each successive event (eq. 3). The difference itself, is illustrative of the consistency of both posing.

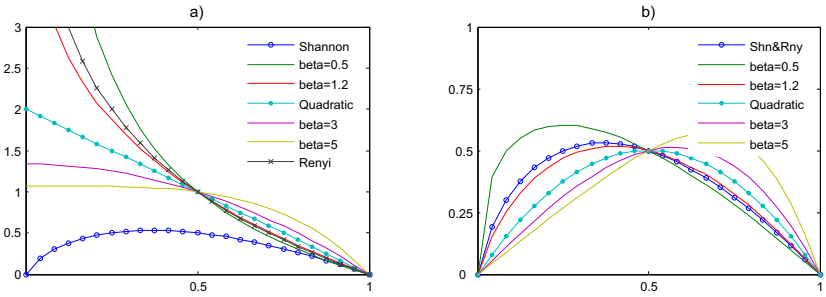


Fig. 5. a) Single event's uncertainty; b) Single event's contribution.

Shannon's entropy has been the basis for a comprehensive posterior work. As far as it is of our concern, let's just highlight two generalizations given for entropy measures:

– Rényi's extension of entropy to the set of generalized probability distributions, [11], defined by,

$$H_\alpha(\mathcal{P}) = H_\alpha(p_1, p_2, \ldots, p_n) = \frac{1}{1-\alpha} log_2 \left(\sum_{i=1}^{n} p_i^\alpha \right), \tag{14}$$

with, $\alpha > 0$, and $\alpha \neq 1$. In the limiting case of $\alpha \to 1$, Rényi's entropy tends to Shannon's entropy.

– Daróczy entropies of type β, [3], defined by, $H^\beta(\mathcal{P}) = \sum_{i=1}^{n} p_i\, u^\beta(p_i)$, being $u^\beta(p_i) = \frac{2^{\beta-1}}{2^{\beta-1}-1}\left(1 - p_i^{\beta-1}\right)$, what yields,

$$H^\beta(\mathcal{P}) = H^\beta(p_1, p_2, \ldots, p_n) = \frac{2^{\beta-1}}{2^{\beta-1}-1}\left[1 - \sum_{i=1}^{n} p_i^\beta\right] \tag{15}$$

with, ($\beta > 0$, and $\beta \neq 1$). In the limiting case of $\beta \to 1$, Daróczy generalization tends also to Shannon's entropy, and setting $\beta = 2$, yields the quadratic entropy, $H^2 = 2\left(1 - \sum_i^n p_i^2\right) = 2\sum_i^n p_i\,(1 - p_i)$, (identical to the so called Gini index).

The contribution of Rényi's extended notion of entropy is that the term $-log(p_i)$, in Shannon's expression, is interpreted as the entropy of the generalized distribution consisting of the single probability p_i, becoming thus evident that eq.13 is, indeed, a mean value, [11]. Daróczy generalization is even more explicit, by introducing the function of uncertainty $u^\beta(p_i)$ for a single event e_i. Together with postulate 4, (adequately adjusted in each case), the assignment of each event's contribution to the total uncertainty, states an implicit algebra of knowledge. This is depicted in fig.5 a), where we plot $u^\beta(p_i)$ for different values of β, with special emphasis on Rényi's uncertainty, $u^{\beta\to 1} = -log(p_i)$, and on the quadratic entropy, $u^2 = 2\,(1 - p_i)$. Also, $h(p_i) = -p_i\, log(p_i)$ is plotted as

the single event's uncertainty assumed from Shannon's original posing. In fig.5 b), the corresponding weighted contributions are plotted.

With respect to our algebra of knowledge (refer to sec.3.1 and fig.4), some significant differences arise:

- In our case, certainty and contribution, of each single event, is the same: knowledge is just the sum of the parts, not a weighted mean.
- At single event's level, uncertainty is not dependent on cardinality. This is, by all means, counterintuitive for us: the probability of observing a particular frequency of occurrence of an event, can not be the same, being this event one out of two, or one out of more.
- The value of uncertainty for an unseen event, $2^{\beta-1}/\left(2^{\beta-1}-1\right)$, is somewhat arbitrary and not easy to comprehend. On the other hand, it is zero when the observed probability of an event is 1, excessively confident to us, (having assumed a cardinality greater than one).
- Even assuming these values, there is no space left for unseen events in the algebra of entropy. Otherwise, one should assume the puzzling result that the contribution of an unseen event is zero, (while its uncertainty is greater than zero). Despite of this, expansibility is systematically ascribed to entropy as a property.
- Certainty is symmetric (at any level) with respect to its minimum, given at the point of minimum information, while contributions to entropy present eccentric maximums, (except for H^2), which are not easy to comprehend.
- Additionally, the question arises of what is the right value of $\alpha \backslash \beta$ for each task at hand, contributing some additional confusion.

Furthermore, though at distribution level, entropy is upper bounded by $2^{\beta-1}/\left(2^{\beta-1}-1\right)\left(1-n^{1-\beta}\right)$ ($log\, n$ for Shannon's and Rényi's entropies), we should be cautious about interpreting this as a dependence on cardinality. Just think about a sample in which we observe uniform frequencies for the $(n-1)$ outcomes of a random variable, while the remaining one is unseen. In such a case, we get a value of uncertainty of $log(n-1)$, when the right entropy should be higher, as far as we know nothing at all about the unseen event. For this same reason, we assign an assumption free value to this case, which, at least, is consequent with a baseline scaling of knowledge.

5 Structural Evidence Functions

In a general case, we will be interested in assessing the *utility* of considering a particular pattern of association between an antecedent set Π and a consequent set X^c, having observed the marginal distributions $P(\Pi)$ and $P(X^c)$, and the conditional distribution $P(X^c|\,\Pi)$, in the sample. We denote in the following, $K(X) \equiv K(P(X))$, for simplicity.

As $K(X^c)$ and $K(X^c|\,\Pi)$ are clashing sources of information about the consequent, in order to adeptly measure the quality of a pattern, we define the following functions of structural evidence:

– *Coherence*, a measure of the *reliability* of the pattern, given by the weighted mean conditional certainty,

$$C^{\Pi,c} = K\left(X^c \mid \Pi\right)$$

– *Presence*, a measure of the *representativity*, or bias, of the sample with respect to the consequent, given by the opposite of the marginal certainty,

$$B^c = U_s + \left(1 - K\left(X^c\right)\right)$$

Note that, $U_s \leq K\left(X^c \mid \Pi\right) \leq 1$ while $0 \leq \left(1 - K\left(X^c\right)\right) \leq C_s$. Inasmuch both measures must be equilibrated, we need to add the term U_s to $\left(1 - K\left(X^c\right)\right)$.

Thus, from the composition of the above two functions, we can define a sort of *certainty gain* as,

$$K_G\left(X^c \mid \Pi\right) = B^c\, C^{\Pi c}$$

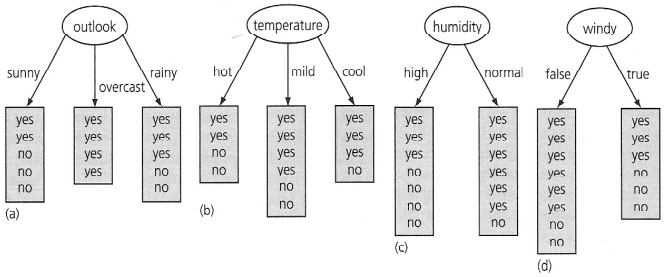
Note, that this expression is the analog to entropy based measures of *information gain*. But, our formalism allows to push this idea further, by including the *presence* of the antecedent, $B^{\Pi} = U_r + \left(1 - K\left(\Pi\right)\right)$, so as to take also into account the bias in the antecedent set. Thus, we get what we call, the *utility* of the pattern, given by,

$$U^{\Pi c} = B^{\Pi}\, B^c\, C^{\Pi c} \tag{16}$$

The composition of all amounts of information involved in a pattern, is a step forward in order to get a relative measure of the quality of the information given by alternative patterns under consideration. But a necessary condition to combine antecedent and consequent, with potentially different cardinalities, is that the measure must express a right cardinality scaling of knowledge. This is, of course, related with the fact of having a balanced measure of certainty for unseen events.

As a little illustration of our approach, we will refer to the *weather* example, extracted from [16]. This is a tiny dataset with 14 examples and attributes: *outlook* (sunny,overcast,rainy), *temperature* (hot,mild,cool), *humidity* (high,normal), *windy* (false,true), and the class *play* (yes,no). In order to select an attribute to place at the root node of a decision tree, we consider the tree stumps shown in fig. 6. Also we show a table, where we summarize the information gain for each attribute, based on Shannon's entropy, resulting from the following usual measures: a) entropic gain, $\left(h\left(Y\right) - h\left(Y \mid X\right)\right)$, b) u coefficient of Theil, $\left(h\left(Y\right) - h\left(Y \mid X\right)\right)/h\left(Y\right)$, c) gain-ratio, $\left(h\left(Y\right) - h\left(Y \mid X\right)\right)/h\left(X\right)$, d) Kvalseth coefficient, $2\left(h\left(Y\right) - h\left(Y \mid X\right)\right)/\left(h\left(X\right) + h\left(Y\right)\right)$. On the other hand, our measures of *presence*, *coherence*, *certainty gain* and *utility* are also given.

All measures yield the same ranking of attributes, being *outlook* the best choice. However, *humidity* is also a good choice because it has a highly selective right branch. Now, let's imagine that *humidity*, originally a continuous feature,



Attr.	s	1/s	event	n	Yes	No	h-i	H(Y/X)	hi	H(X)	Gain	GainRat	uTheil	Kvalseth	Prs.	C-I	Coh.	Cgain	Utl.
Play	2	0,5		14	9	5		**0,940**							0,959				
outlook	3	0,3	*sunny*	5	2	3	0,971	0,694	0,531	1,577	**0,247**	**0,156**	**0,262**	**0,196**	0,997	0,520	0,657	0,630	**0,655**
			overcast	4	4	0	0,000		0,516							1,000			
			rainy	5	3	2	0,971		0,531							0,520			
humidity	2	0,5	*high*	7	3	4	0,985	0,788	0,500	1,000	0,152	0,152	0,161	0,157	1,000	0,510	0,633	0,607	0,633
			normal	7	6	1	0,592		0,500							0,755			
windy	2	0,5	*false*	8	6	2	0,811	0,892	0,461	0,985	0,048	0,049	0,051	0,050	0,990	0,625	0,571	0,548	0,566
			true	6	3	3	1,000		0,524							0,500			
temperature	3	0,3	*hot*	4	2	2	1,000	0,911	0,516	1,557	0,029	0,019	0,031	0,023	0,986	0,500	0,560	0,537	0,552
			mild	6	4	2	0,918		0,524							0,556			
			cool	4	3	1	0,811		0,516							0,625			
humidity	3	0,3	*high*	7	3	4	0,493	0,493	0,500	1,296	**0,448**	**0,345**	**0,476**	**0,400**	0,895	0,510	0,755	0,724	0,675
(i)			*medium*	1	0	1	0,000		0,272							1,000			
			low	6	6	0	0,000		0,524							1,000			
humidity	3	0,3	*high*	7	3	4	0,985	0,635	0,500	1,432	**0,305**	**0,213**	**0,324**	**0,257**	0,935	0,510	0,684	0,656	0,639
(ii)			*medium*	2	1	1	1,000		0,401							0,500			
			low	5	5	0	0,000		0,531							1,000			

Fig. 6. Tree stumps for the weather data.

had been differently discretized, splitting the *normal* interval into *medium* and *low*, so that just the discordant example of that branch would have fallen in the new *medium* category. In such case, (shown in the table as *humidity(i)*), this attribute becomes highly informative, what is clearly indicated by all entropy based measures. *Utility* selects it as well, though it is not so confident about this new partition. The reason is that the *medium* branch is poorly represented, this is reflected in the attribute's *presence*, and the value of *utility* is heavily penalized. Following this line of reasoning, we present also the case $humidity(ii)$, in which the *medium* branch includes also an example with $play = yes$. In such case, though with lower values, entropy based measures still select this attribute as the best choice. Conversely, our measure of *utility* does not select it any more. To the previous reason, we must add now the minimum information contributed by this branch. In summary, this balance between presence and coherence, leads to a balance between the size of the tree and the number of examples at each leaf.

6 Conclusions

To put in a few words, this work can be summarized in the principle of making no assumptions about what is not known. Instead, we introduce a sort of bar measure, which always makes the same exact assumption about what is not known, in a cautious and balanced way. As such, it allows us to pick an optimal choice from a set of options, based on what we certainly can observe.

Our main contribution is to build up a plausible hypothesis from which a measure of certainty is derived, yielding a consistent algebraic structure and a consistent set of properties. With respect to entropy based measures, our approach contributes some important advantages, namely, a comprehensible insight of knowledge, with a native cardinality scaling, from which an assumption free measure of certainty for unseen events is derived. This hypothesis is summarized in a set of postulates that characterize such a measure of knowledge. The question remains, about whether this characterization is unique of this measure.

References

1. Aczél, J.; Forte, B.; Ng, C.T. Why Shannon and Hartley entropies are natural. Adv. Appl. Probab. 1974, 6, 131-146.
2. Agresti, A. Categorical Data Analysis. John Wiley and Sons, Inc. 24-25. 1990.
3. Daróczy, Z. Generalized information functions. In *Information and Control*, 16:36-51, 1970.
4. Gini, C.W. Variablity and Mutability, contribution to the study of statistical distributions and relations. In *Studi Economico-Giuricici della R. Universita de Cagliari*. 1912.
5. Greenberg, J.H. The measurement of linguistic diversity. *Language*, vol.32, n1, pp.109-115. January-March, 1956.
6. Gale, W.; Sampson, G. Good-Turing Smoothing without Tears. Journal of Quantitative Linguistics, 2(3):217–237, 1995.
7. Grünwald, P. A Tutorial Introduction to the Minimum Description Length Principle. In *Advances in Minimum Description Length: Theory and Applications*, MIT Press (2005).
8. Herfindahl, O.C. Concentration in the U.S. Steel Industry. Unpublished doctoral dissertation, Columbia University, 1950.
9. Hirschman, A.O. National Power and the Structure of Foreign Trade. Berkeley, 1945.
10. Kolmogorov, A.N. Foundations of the theory of probability. Edited by N.Morrison, Chelsea Publishing Company, New York, 1956.
11. Rényi, A. On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol.1, pp.547-561, University of California Press, 1961.
12. Rissanen, J. Modeling by the shortest data description. In *Automatica*, 14, 465-471, 1978.
13. Shannon, C.E. A Mathematical Theory of Communication. *The Bell System Technical Journal*, vol.27, pp.379-423,623-656, July, October, 1948.
14. Simpson, E.H. Measurement of Diversity. *Nature*, vol.163, pp.688, Macmillan Publishers Ltd. April, 1949.
15. Wehenkel, L. On uncertainty measures used for decision tree induction. In *International Congress on Information Processing and management of Uncertainty in Knowledge Based Systems*. Granada, Spain, pp. 413-418, 1996.
16. Witten, I.H.; Frank, E. Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco, 2000.
17. Yule, G.U The Statistical Study of Literary Vocabulary. *Cambridge University Press*, 1944.