Certainty upon Empirical Distributions

Joan Garriga

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

Abstract. We address the problem of assessing the information conveyed by a finite discrete probability distribution, within the context of knowledge discovery. Our approach is based on two main axiomatic intuitions: (i) the minimum information is given in the case of a uniform distribution, and (ii) knowledge is akin to a notion of richness, related to the dimension of the distribution. From this perspective, 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. This includes a native value for the uncertainty related to unseen events. Our contributions are then faced up with entropy based measures. Finally, by implementing our measure in a decision tree induction algorithm, we show an empirical validation of the behavior of our measure with respect to entropy. Our conclusion is that the contributions of our measure are significant, and should lead to more robust models.

Key words: knowledge discovery, measures of information, entropy

1 Introduction

Many data mining tasks for knowledge discovery rely on the use of the so called *information measures*. Such measures are intended in order to select an optimal model (statistical model selection, graphical modeling), an optimal set of rules (classification rule mining), an optimal split at each node of a tree (induction of decision trees), an optimal discretization of a continuous variable, or whatever. In any case, all of them are particular forms of expressing knowledge learned from data, which in our 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 co-ocurrences of values, observed in the sample. Therefore, such measures intend to assess the amount of information conveyed by any finite discrete probability distribution observed in the data.

Among others, Shannon's entropy [11] is the most widely known measure of uncertainty associated to a probability distribution. Some attractive properties

hold for this measure that make it uniquely characterized [2]. A nice correspondence can be established between these properties and, what is commonly accepted as, a plausible axiomatic definition of knowledge [1]. This is the reason of its success, and the basis of a comprehensive later work.

Although entropy is strongly rooted within the information theory community, entropy's characterization does not properly attain to cover some aspects of knowledge. It is well known, for instance, that when applied to the induction of decision trees, entropy shows a certain bias for attributes with greater cardinality. Also, it yields some undesired results, when the attributes to be used, and/or the classes to be learned, have highly imbalanced frequencies. Thus, further generalizations have been developped, like Rényi's entropies of type α [10], and Daróczy's entropies of type β [3]. Furthermore, other solutions have been suggested in the form of combined entropies, (entropic gain [8], the u coefficient of Theil [13], the gain-ratio [9], the Kvalseth coefficient [6], or more recently the off-centered entropies [7]).

Alternatively, we propose a new measure of certainty [1], which is derived from a slightly different axiomatization of knowledge, that takes into account these aspects. As we will show, when applied to practical problems such as induction of decision trees, this perspective leads to some different results.

In section 2, we describe the aim and major contributions of our approach. In section 3, we briefly review the groundwork of the proposed measure. In section 4, we focus on entropy based measures, and how they relate with our contributions. In section 5, we give some experimental results. Finally, in section 6, we comment the results and summarize our conclusions.

2 Contributions

In a few words, we introduce a measure of *Certainty* upon empirical probability distributions. With this measure, we aim at overcoming two important aspects of knowledge that, in our opinion, are not yet properly covered.

2.1 The Cardinality Scaling of Knowledge

Let's figure the problem of modeling the price of a house from a set of features (city, square meters, garage, ...). We can deal with it as a regression problem, ok. But we also could try to discretize the price variable, and treat it as a classification problem. In this case, it is clear that, (overfitting issues apart), the higher the cardinality of the class variable in our final model, the more the information it expresses. This is a direct illustration of our intuition that knowledge is akin to a notion of richness, related to the dimension of the distribution. But we would like to depict a deeper idea of this concept.

[1] 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.

So, let's figure the following horse race betting example. We have a sample of races, in which two horses, (Tomcat and Apache), show equal number of victories. Apparently, there exist no compelling reasons to lean over either of them in future races. Now, let's suppose that we have an equal size second sample of races, in which Tomcat is running against two competitors, and we observe the same 0.5 winning risk for Tomcat. In this case, it could make sense to bet. Going further, if more competitors are involved in the race, and the sample keeps yielding the same 0.5 risk for Tomcat, the chances of its individual competitors are bound to decrease, and we are increasingly confident on Tomcat. Indeed, we are always bound to loose 0.5 of our bets in the long run, but it is clear that our epistemic state is quite different, and we are increasingly compelled to bet on those more populated races, though Tomcat's chances remain the same.

The aim of this example is just to show that knowledge about random events is conditioned, not only by the observed frequencies, but also by the dimension of the probability distribution under consideration, (or the cardinality of the set of possible outcomes). Therefore, we call this effect the *cardinality scaling of knowledge*, which we may regard as different levels of *quality* of knowledge. The roots of this effect must be seek, not in the decreasing chances of each competitor, but rather in the curse of dimensionality: given a finite fixed sample size, the higher the dimension of the sample space (the number of competitors), the lower the prior chance of observing a 0.5 risk for Tomcat, so its statistical significance varies.

Now let's think of it the other way round, as it happens to be in data mining processes, when, given a fixed sample, models of different complexity are taken under consideration: the statistical significance of the corresponding observed frequencies is different, and so must be the amount of knowledge they convey.

In fig.1 we show how entropy and certainty capture this aspect of knowledge. This depiction shows the pieces of information, that each elementary event, contributes to the total certainty/uncertainty of the random process, for a growing number of uniform competitors, (given by s), while Tomcat's chances remain the same. The bottom one is Tomcat's 0.5 risk contribution.

In essence, what this figure shows is the evolution of our epistemic state with respect to the outcome of the race, in relation with the growing number of runners. Both of them express the obvious increase of uncertainty with the number of competitors, but we can observe some differences:

1. the increasing rate of uncertainty is different, so we should assume that, at a global level, each approach expresses a different scaling of knowledge;
2. the bounding is also different, $\lim_{s\to\infty}(H_s) = \infty$, while $\lim_{s\to\infty}(K_s) = 0.25$;
3. the single event *Tomcat wins*, contributes a fixed (cardinality independent) amount to the total uncertainty, while its contribution to the total certainty is dependent on the number of runners (cardinality dependent);
4. as the number of runners increases, the competitors contribute together an increasing amount of uncertainty, while their joint contribution to the total certainty tends to zero.

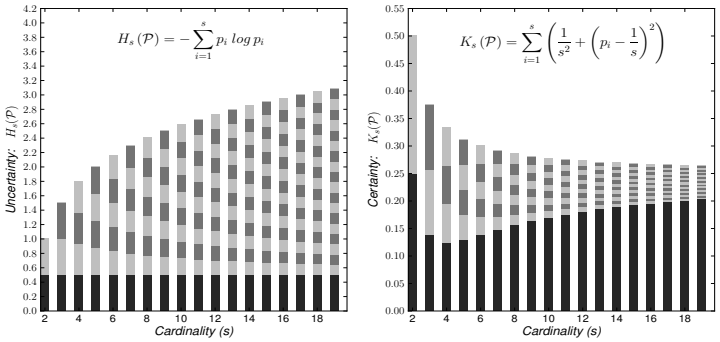

Fig. 1. Given $\mathcal{P}\left(0.5, \frac{0.5}{(s-1)}, ..., \frac{0.5}{(s-1)}\right)$, depiction of the elementary contributions to: (left) uncertainty as expressed by entropy; (right) certainty as expressed by equation 3, (promptly introduced in section 3). Please note the different scaling

Now, given an ideal, large enough, sample size, the question is: does our initial state of knowledge, (with two runners), evolve to an (almost) infinite uncertainty, as the number of runners grows (almost) indefinitely? This can be argued in many different ways, but our point of view is that the answer to this question is definitely no. Otherwise, this assumption would leave no room to our certainty that Tomcat is going to win half of the races.

Our reasoning comes from the certainty side. As the cardinality increases, our certainty decreases, and each competitor's contribution is less because their chances are lower. Up to here, this is correctly expressed by entropy. The difference in our measure is that, while being more uncertain about the outcome, the certainty part is increasingly due to Tomcat's chances. At the limit, we reach an ideal situation in which we just have the amount contributed by Tomcat. This is because the competitors are (tending to) infinite and uniform, two reinforcing reasons that explain a null contribution to the final certainty. At the same time, as each competitor's chances has almost vanished, Tomcat's victory seems to be amazingly guaranteed. But our certainty can not be one because Tomcat's chances are less than one, what leads to the maximum possible certainty, that is, just the value contributed by Tomcat in our initial state. We judge this as a more comprehensible description of our epistemic state, than a state of infinite uncertainty.

2.2 Uncertainty about Unseen Events

The former example is an illustration of what we refer as the *algebra of knowledge*, that is, the way we measure the pieces of knowledge contributed by each one of the elementary events of the sample space, and the way these pieces should be combined, in order to yield a global measure of the information conveyed by their observed distribution. Obviously, this must include all that part of the

sample space that is not observed in the sample, what we call the *unseen events*. This is again, a well known consequence of the curse of dimensionality.

Decision trees, are a clear illustration of this issue. Given a fixed sample, the more expanded the tree, the higher the chances of getting to empty leaves. The set of empty leaves represents the unobserved part of the input space. Whenever the sample is representative enough of the domain, the existence of empty leaves could make sense: for instance, in a tree involving features as the age and profession of people, it would make sense to find the 80-years-old-sportsman leave empty, (though it may indeed occur in the real world). Unfortunately, real world cases are not so clear, and even very large samples, become not enough representative of the domain, as soon as the tree is just a few levels depth.

This lack of representation gives rise to an important amount of uncertainty. Therefore, it is a must to take empty leaves (unseen events) under consideration, when learning decision trees, (or more generally speaking, any kind of models), from data.

With regard to such basic concept of generalizing from the training data, entropy yields a puzzling result: the uncertainty contributed by an unseen event is zero!, as expressed by the weird mathematical artifact $H(0) = 0\, log\, \infty = 0$. This result is somewhat clashing, and according to this, we are lead to believe that some issues seem to be sneaking through entropy's axiomatization of knowledge. The evidence is that, at this point, one has to rely on *ad-hoc* regularization or smoothing procedures, in order to estimate a complete probability distribution from the frequencies observed in the sample.

Conversely, we start up with a slightly different axiomatic approach to knowledge. From there, we derive a measure of certainty which achieves both: it takes into account the cardinality scaling of knowledge, and it yields a natively smoothed piece of certainty about each single event, being it observed or not in the sample. 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 and offers a quite comprehensible insight of knowledge.

3 A Measure of Certainty

In the following, we denote by $\mathcal{P} = (p_1, p_2, \ldots, p_s)$, a 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. 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.

Our starting point is to measure the deviation of any such distribution, with respect to uniformity. Uniformity means equiprobability, which is 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 expression of such deviation is:

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

Now, let's introduce our simple axiomatic requests: (i) the minimum knowledge we can have is that given in case of uniformity, and (ii) knowledge is akin to a notion of richness, related with the cardinality of the distribution. From the combination of both we may derive that, in the worst case, knowledge should not be zero. A more consequent alternative is to consider that the minimum is 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,

$$K_s\left(\mathcal{P}\right) = U_s + \sum_{j=1}^{s} \left(p_j - U_s\right)^2 = \sum_{j=1}^{s} p_j^2 \tag{1}$$

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: 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,

$$K_{s+1}\left(\mathcal{Q}\right) = K_s\left(\mathcal{P}\right) - p_1^2\;\left(1 - K_2\left(\mathcal{T}\right)\right) \tag{2}$$

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.

3.1 Disjoint Dependent Events

One may think that there is nothing new in eq.1: it just looks like a simple translation of the euclidean distance to the uniform distribution, leading us to the well known *indexes of diversity/concentration*, long ago defined by [4], [5] or [12], among others. Furthermore, eq.1 is apparently independent of cardinality, and explicitly expresses that any unseen event has a null contribution to the total certainty.

The point comes with the algebra of knowledge that underlies this expression. In fact, for each elementary event we have that,

$$p_j^2 = \left(U_s + \left(p_j - U_s\right)\right)^2 = U_s^2 + \left(p_j - U_s\right)^2 + 2\,U_s\left(p_j - U_s\right)$$

Therefore, we see that the term U_s, (the certainty offset), is equally distributed among all possible outcomes of the distribution, yielding a term U_s^2, and the amount 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.

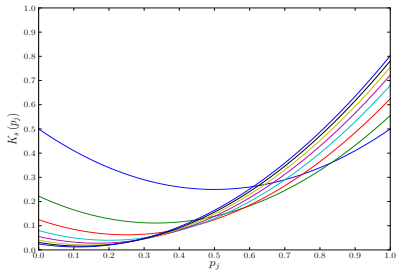

Fig. 2. Single event's Certainty for $s = \{2, 4, 8, 16, 32, 64, 128, \infty\}$

Consequently, if p_j is the observed probability of occurrence of event e_j, our knowledge about the outcome of e_j is the composition of two terms: a cardinality dependent offset, and a deviation with respect to the uniform distribution,

$$K_s\left(p_j\right) = U_s^2 + \left(p_j - U_s\right)^2 \tag{3}$$

While remaining consequent with eq.1, that is, $K_s\left(\mathcal{P}\right) = \sum_{j=1}^{s} K_s\left(p_j\right)$, this expression, shown in fig. 2, offers a quite different picture of the elementary contributions to global certainty:

- It is explicitly dependent on s. In the limit, where this measure would hardly apply, certainty meets (square) probabilities,

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

- The minimum value is coherently given at the point of equiprobability, where we have $K_s\left(U_s\right) = U_s^2$,
- It is continuous at zero, yielding a value greater than zero for any unseen event, that is, $K_s\left(0\right) = U_s^2 + U_s^2$. This value is the expression of a conservative attitude with respect to future coming examples, while, at the same time, it is an assertion of the certainties with which the rest of events have been observed.
- Being consequent with the previous, for any event with an observed probability of one, the measure yields a value lower than one, $K_s\left(1\right) = U_s^2 + C_s^2$
- In case of uniformity, or minimum information, we have,

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

- In the case of observing only one event, we have maximum certainty, (which does not mean absolute certainty),

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

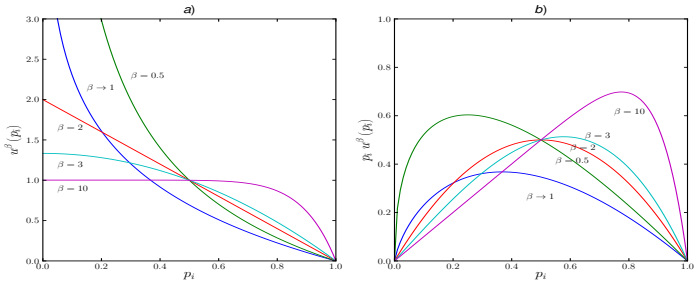

Fig. 3. a) Single event's uncertainty; b) Single event's weighted contribution.

Together with eq.2, this features synthesize the additive algebra of knowledge that is implicit by the measure of certainty. 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.

4 Entropy based Measures

Though initially not conceived as such, Shannon's entropy [11] is, by far, the most widely used measure of information,

$$H\left(\mathcal{P}\right) = H\left(p_1, p_2, \ldots, p_n\right) = \sum_{i}^{n} p_i log_2 \frac{1}{p_i} \tag{4}$$

Further generalizations of entropy have been defined, first by Rényi [10],

$$H_\alpha\left(\mathcal{P}\right) = H_\alpha\left(p_1, p_2, \ldots, p_n\right) = \frac{1}{1-\alpha} log_2 \left(\sum_{i=1}^{n} p_i^\alpha\right), \tag{5}$$

(with, $\alpha > 0$, and $\alpha \neq 1$; in the limiting case of $\alpha \to 1$, Rényi's entropy tends to Shannon's entropy), and later by Daróczy [3],

$$H^\beta\left(\mathcal{P}\right) = \sum_{i=1}^{n} p_i\, u^\beta\left(p_i\right) \text{ , where, } u^\beta\left(p_i\right) = \frac{2^{\beta-1}}{2^{\beta-1}-1}\left(1 - p_i^{\beta-1}\right)$$

what yields the so called entropies of type β, (with $\beta > 0$, and $\beta \neq 1$),

$$H^\beta\left(\mathcal{P}\right) = H^\beta\left(p_1, p_2, \ldots, p_n\right) = \frac{2^{\beta-1}}{2^{\beta-1}-1}\left[1 - \sum_{i=1}^{n} p_i^\beta\right] \tag{6}$$

(in the limiting case of $\beta \to 1$, Daróczy generalization tends to Shannon's entropy, and setting $\beta = 2$, yields the quadratic entropy, $H^2 = 2\left(1 - \sum_i^n p_i^2\right) = 2\sum_i^n p_i\left(1 - p_i\right)$, identical to the so called Gini index [4]).

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.4 is, indeed, a mean value, [10]. Daróczy generalization is even more explicit, by introducing the function of uncertainty $u^{\beta}\,(p_i)$ for a single event e_i.

The second contribution refers to the shape of the curve, determined by the β factor. As to our concern, this curve resolves a particular value of uncertainty for unseen events, (undetermined in Shannon's entropy), given by $u^{\beta}\,(0) = \frac{2^{\beta-1}}{2^{\beta-1}-1}$. In any case, the global uncertainty remains as a mean value, and therefore this particular value is meaningless by itself.

This is what we depict in fig.3: at the left (fig.3a), we plot the values of uncertainty $u^{\beta}\,(p_i)$ for different values of β, with special emphasis on Rényi's uncertainty, $u^{\beta\to 1} = -log\,(p_i)$, and on the quadratic entropy, $u^2 = 2\,(1-p_i)$; at the right (fig.3b), we plot the corresponding weighted contribution of a single event, showing how the contribution of unseen events, inevitably, vanishes.

As already stated, some reasons exist to believe that entropy does not cover a proper axiomatization of knowledge. If we do believe that unseen events are to be taken under consideration, we easily come to the conclusion that the algebra of knowledge does not fit well with the concept of a weighted mean measure. Furthermore, if we do believe that knowledge is akin to a notion of quality, we may yet find a reason to understand entropy's bias. It is argued that entropy is cardinality dependent, (being its maximum given by $log\,(s)$), but as long as there is no room for unseen events, this argument becomes worthless.

Conversely, we are giving a different answer to each one of these questions: certainty's algebra of knowledge is just additive (not a weighted mean), cardinality dependent, and yields not null values for unseen events. Despite of this, expansibility is implicit in this algebra. With respect to entropy, expansibility follows straightforward from the fact of being a weighted mean. Herein, the connotations of certainty's expansibility are much stronger.

5 Empirical Validation

We have run some experiments to empirically study the behavior of our measure, by implementing it in a decision tree algorithm, and comparing it with the landmark decision tree C4.5 algorithm [8].

The classical implementation of an ID3 algorithm has the drawback of expanding the tree until all leaf nodes are pure, or no more attributes are left to split on. Therefore, the C4.5 algorithm was developed, with two special enhancements: subtree replacement and subtree raising. Both of them are postpruning operations, at the cost of some accuracy on the training set. These operations are based on some weak statistical reasoning [15], and they involve some parameters. However, they seem to work well in practice, even with the default values suggested for the parameters.

Thus, our challenge is clear: if certainty expresses a proper cardinality scaling of knowledge, not only it should be able to choose the optimal attribute to split

on at each node, but also it should stop expanding the tree, whenever further splits are unnecessary. Therefore, the algorithm stands parameter free, and no postpruning operations are needed.

In order to carry out our evaluation, we have used the publicly available machine learning tool Weka [15], implementing our measure in the ID3 decision tree algorithm. Implementing certainty is straightforward: at each node we compare the class marginal distribution certainty, $K_s(class)$, with the class conditional distribution certainty, $K_s(class \mid att.)$, given each one of the pending attributes to split on. If there is no candidate yielding $K_s(class \mid att.) > K_s(class)$, the node is not expanded. If some exist, we choose the one with lower marginal distribution certainty, $K_r(att)$, (r refers to the attribute's cardinality), what means a more equilibrated marginal distribution. Thus, we split on the attribute which ensures a better coverage of all of its branches.

The experimental setup includes some heterogeneous data sets from the UCI repository [14]. The ID3 algorithm does not deal with continuous or missing values. Therefore, examples with missing values have been discarded, and continuous values have been discretized to a reasonable number of equal width intervals. In all runs, we use a 10 fold cross validation method. Whenever independent train and test sets are available, we also perform an independent train/test classification. For the C4.5 algorithm we always use the default parameters.

The results are shown in table 1. We also show the results yielded by the entropic-gain ID3 original algorithm. Thus, it is easier to figure out the way how entropy tends to cover the sample space, and the posterior effect of the pruning phase of the C4.5 algorithm.[2]

6 Conclusions

Let's note the trade off between the complexity of the model, (column labeled *treeSize*), the dimension of the input space, (column *leaves*), the unobserved part of that space, (column *nullLvs*), and accuracy (column *%correct*). To better appreciate this tradeoff, the column labeled *%uncovered*, specifies the ratio of empty leaves with respect to the total number of leaves.[3]

Useless to denie it: the C4.5 algorithm yields somewhat better accuracies. Ok, this is just due to the greedy behavior of entropy, with respect to the conservative attitude of certainty. But, at a little cost in accuracy, we get dramatic reductions in the complexity of the models, and dramatic reductions in the unobserved part of the input space, along with significant reductions in computational cost. This means that, beyond the accuracies yielded by a ten fold cross validation, we can be much more confident on the behavior of certainty models with respect to future coming examples.

[2] Regarding to some of the data sets used, better accuracies have been reported using other methods. Please keep in mind, that the aim of the experiment is just to compare the behavior of certainty and entropy as measures of information.

[3] The tree size values refer to the basic model build upon the whole training set.

DataBase	setSize	attr.	Clssf.	tree	treeSize	nodes	leaves	nullLvs.	%uncovered	%correct
BreastCancer	683	10	10fld	ID3	211	21	190	95	50.00	91.65
			10fld	C4.5	61	6	55	14	25.45	93.41
			10fld	Crt.	51	5	46	4	8.70	95.46
SegmentChallenge	1500	20	10fld	ID3	390	44	346	193	55.78	93.92
			10fld	C4.5	213	23	190	102	53.68	94.93
			10fld	Crt.	171	27	144	48	33.33	91.73
OpticalDígits	5620	65	10fld	ID3	11493	676	10817	7582	70.09	44.11
			10fld	C4.5	4023	241	3782	2334	61.71	63.02
			10fld	Crt.	1769	104	1665	333	20.00	54.02
	1797		testSet	C4.5	3010	177	2833	1737	61.31	56.82
	1797		testSet	Crt.	1225	72	1153	198	17.17	54.26
penDigits	10992	17	10fld	ID3	5798	527	5271	2955	56.06	86.69
			10fld	C4.5	2366	215	2151	1068	49.65	89.16
			10fld	Crt.	1805	164	1641	342	20.84	86.85
	3498		testSet	C4.5	1915	174	1741	910	52.27	84.08
	3498		testSet	Crt.	1288	117	1171	227	19.39	81.76
letterRecognition	20000	17	10fld	ID3	30561	1910	28651	21832	76.20	73.53
			10fld	C4.5	13409	838	12571	9033	71.86	77.73
			10fld	Crt.	4657	291	4366	2060	47.18	71.65
Soybean	562	36	10fld	ID3	50	51	116	31	26.72	83.77
			10fld	C4.5	69	22	47	10	21.28	91.81
			10fld	Crt.	161	63	98	3	3.06	87.72
CarEvaluation	1728	7	10fld	ID3	408	112	296	0	0.00	89.35
			10fld	C4.5	182	51	131	0	0.00	92.36
			10fld	Crt.	213	58	155	0	0.00	94.21
			trainSet	C4.5	182	51	131	0	0.00	96.30
			trainSet	Crt.	213	58	155	0	0.00	96.30
Nursery	12960	9	10fld	ID3	1159	320	839	0	0.00	98.19
			10fld	C4.5	511	152	359	0	0.00	97.05
			10fld	Crt.	1031	274	757	0	0.00	96.37
			trainSet	C4.5	511	152	359	0	0.00	98.13
			trainSet	Crt.	1031	274	757	0	0.00	98.59

Table 1. Comparison of trees induced by entropy (ID3, C4.5) and certainty (Crt.).

In the OpticalDigits, PenDigits and LetterRecognition data bases, the features are vectors of integers ranging from 0 to 16. At this level of cardinalities, entropy begins to show some undesired behavior, and the differences between both measures become evident: C4.5 yields an extremely large uncovered part of the input space, and we should be cautious about relying on the accuracy figures given.

Soybean, CarEvaluation and Nursery, are exceptional cases in which certainty yields more complex models, with larger input spaces. The reason is that there exists an extra sample subspace that is sufficiently covered by the sample, and all this extra information can be efficiently exploited. For the Soybean case, though yielding a more complex model, the unobserved part of its input space is still much smaller. On the other hand, the CarEvaluation and Nursery examples are two special data bases, in which all attributes are perfectly balanced, and the sample space is completely covered, thus, no unseen events exist. Under such conditions, certainty tends to exploit all the information available and expands the tree over the whole sample space. Again, that extra information seems to be efficiently used and does not turn into excessive overfitting. Furthermore, the classification results over the training set, (also included in table 1), show that, once all the information about the sample space is available, the accuracies achieved are exactly the same.

In summary, certainty's algebra of knowledge seems to work well. The cardinality scaling of the measure, along with the uncertainty of unseen events, seems to guarantee a proper comparison of the information conveyed by distributions of different dimensions. In this case, it allows implementing a parameter free algorithm that stops the expansion of the tree at a reasonable level. This is a significant contribution, since pruning is the most computationally expensive part of tree induction. Thus, although we should not expect the best results in a validation phase, it looks as a promising tool, whenever the goal is to get some fast, robust and reliable knowledge.

References

1. Aczél, J.; Forte, B.; Ng, C.T. (1974) Why Shannon and Hartley entropies are natural. Adv. Appl. Probab. 1974, vol.6, pp.131-146.
2. Aczél, J.; Daróczy, Z. (1975) On Measures of Information and Their Characterizations. Academic Press: New York.
3. Daróczy, Z. (1970) Generalized information functions. In *Information and Control*, vol.16, pp.36-51.
4. Gini, C.W. (1912) Variability and Mutability, contribution to the study of statistical distributions and relations. In *Studi Economico-Giuricici della R. Universita de Cagliari*.
5. Herfindahl, O.C. Concentration in the U.S. Steel Industry. Unpublished doctoral dissertation, Columbia University, 1950.
6. Kvalseth, T.O. (1987) Entropy and correlation: some comments. In *IEEE transactions on Systems, Man and Cybernetics*, vol.17(3), pp.517-519.
7. Lenca, P.; Lallich, S.; Vaillant, B. (2010) Construction of an Off-Centered Entropy for the Supervised Learning of Imbalanced Classes: Some First Results. In *Communications in Statistics- Theory and Methods*, vol.39:3, pp. 493-507.
8. Quinlan, J.R. (1986) Induction of decision trees. In *Machine Learning*, vol.1(1), pp.81-106.
9. Quinlan, J.R. (1993) C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann.
10. Rényi, A. (1961) 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.
11. Shannon, C.E. (1948) A Mathematical Theory of Communication. *The Bell System Technical Journal*, vol.27, pp.379-423,623-656, July, October, 1948.
12. Simpson, E.H. (1949) Measurement of Diversity. *Nature*, vol.163, pp.688, Macmillan Publishers Ltd. April, 1949.
13. Theil, H. (1970) On the estimation of relationships involving qualitative variables. In *The American Journal of Sociology*, vol.76(1), pp.103-154.
14. Frank, A.; Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
15. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. (2009) The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.