Comparing Probability Measures Using Possibility Theory: A Notion of Relative Peakedness ⋆

Didier Dubois [a] and Eyke Hüllermeier [b]

[a] *Institut de Recherche en Informatique de Toulouse, France*

[b] *Faculty of Computer Science, University of Magdeburg, Germany*

Abstract

Deciding whether one probability distribution is more informative (in the sense of representing a less indeterminate situation) than another one is typically done using well-established information measures such as, e.g., the Shannon entropy or other dispersion indices. In contrast, the relative specificity of possibility distributions is evaluated by means of fuzzy set inclusion. In this paper, we propose a technique for comparing probability distributions from the point of view of their relative dispersion without resorting to a numerical index. A natural partial ordering in terms of relative "peakedness" of probability functions is proposed which is closely related to order-1 stochastic dominance. There is also a close connection between this ordering on probability distributions and the standard specificity ordering on possibility distributions that can be derived by means of a known probability-possibility transformation. The paper proposes a direct proof showing that the (total) preordering on probability measures defined by probabilistic entropy refines the (partial) ordering defined by possibilistic specificity. This result, also valid for other dispersion indices, is discussed against the background of related work in statistics, mathematics (inequalities on convex functions), and the social sciences. Finally, an application of the possibilistic specificity ordering in the field of machine learning or, more specifically, the induction of decision forests is proposed.

Key words: probability distributions, possibility distributions, probability-possibility transformation, entropy, dispersion, stochastic dominance, specificity, machine learning, recursive partitioning, decision forest

⋆ A preliminary version of this paper appeared under the title "A notion of comparative probabilistic entropy based on the possibilistic specificity ordering", in LNAI 3571, Springer-Verlag, Berlin, pp. 848-859.

Email addresses: dubois@irit.fr (Didier Dubois), huellerm@iti.cs.uni-magdeburg.de (Eyke Hüllermeier).

1 Introduction

The principle of maximum entropy plays an important role in probability theory, especially in the case of incomplete probabilistic models (see e.g. Paris [25]). It is instrumental in selecting a probability distribution in agreement with the available constraints, preserving as much indeterminateness as possible. Moreover, entropy faithfully accounts for existing dependencies and only assumes independence where no justification to the contrary can be found [17,19]. There are axiomatic characterizations of the Shannon entropy function (Shore and Johnson [28]). Paris [25] has strongly advocated the selection of the maximum entropy probability as being a reasonable default choice under incomplete information. Entropy can also be viewed as one of the many dispersion indices that one can find in the literature (see Morales et al. [23]).

In possibility theory, "least commitment" information principles similar to entropy exist (e.g. Dubois et al. [10]): When a set of constraints delimits a family of possibility distributions, the least committed choice is the minimally specific distribution. The underlying idea is to consider any situation as being possible as long it is not explicitly ruled out by the constraints. This principle obviously suggests *maximizing* possibility degrees.

There also exists a natural *partial* information ordering between possibility distributions, called the *specificity relation*. This ordering is based on fuzzy set inclusion: If a possibility distribution $\pi : W \to [0, 1]$ is pointwisely dominated by another distribution $\pi' : W \to [0, 1]$, i.e., $\pi(w) \leq \pi'(w)$ for all $w \in W$, the former is said to be more specific than the latter (and strictly more specific if $\pi(w) < \pi'(w)$ for at least one $w \in W$). The natural measure of non-specificity in agreement with this partial ordering is the sum of the possibility degrees (also the scalar cardinality of the corresponding fuzzy set). [1]

Intuitively, there is a connection between ideas of probabilistic dispersion and possibilistic specificity: large dispersion and low specificity suggest distributions with wide supports. One may see some analogy between maximal entropy and minimal specificity principles, especially in the light of the Laplace indifference principle: In the possibilistic framework, the case of complete ignorance is adequately represented by the uniform distribution $\pi \equiv 1$ (all states w are completely possible). Likewise, if a unique probability distribution must be picked, the aforementioned indifference principle suggests selecting the uniform distribution $p \equiv |W|^{-1}$. For these distributions, the Shannon entropy and the additive possibilistic measure of non-specificity coincide with the Hartley entropy of a set (Higashi and Klir [15]), that is, the logarithm of the number of elements in the set. These authors use an additive index of possibilistic

[1] Of course, here we assume the domain W to be finite or at least countable. Otherwise, the sum must be replaced by an integral.

non-specificity that looks like Shannon entropy.

The temptation to relate specificity and entropy at a formal level is great. For instance, Klir [18] suggested equating numerical entropy and (additive) non-specificity indices for the purpose of transforming possibility distributions into probability distributions and conversely. This is debatable, however, because the entropy scale and the specificity scale are not commensurate. Maung [21] has tried to justify the principle of minimal specificity by adapting Paris' rationality axioms to the possibilistic setting.

Regarding the information-based comparison of distributions, there is an important difference between the probability and possibility settings. In the uncertainty literature, the comparison between probability distributions is often based on a type of entropy index without reference to an underlying intuitive *partial* ordering, which would be directly defined between probability distributions reflecting their relative informativeness. There are actually several entropy indices and dispersion indices (such as the Gini index) but the partial ordering that decides if a probability measure is more informative than another one is far less known. Yet, in information theory, authors such as Morales et al. [23], have pointed out that any well-behaved information measure is Schur-concave and satisfies a monotonicity condition with respect to a natural informativeness ordering between probability distributions. Also, there is an old paper by Birnbaum [1] suggesting such a qualitative comparison of probability functions on the real line in terms of what is called their *peakedness*, independently of the notion of entropy. It basically consists of checking the nestedness of confidence intervals of various confidence levels extracted from the probability distribution. Of course, the nestedness property of confidence intervals strongly suggests a similarity between the relative peakedness of probability distributions and the relative specificity of possibility distributions. On the other hand, the more peaked a probability distribution, the less spread out and, hence, the less indeterminate it is, and the lower its entropy should be.

The aim of this paper is to establish a connection between possibilistic specificity and a variant of the peakedness relation between probability distributions, known in mathematics, information theory, and the social sciences. This relation compares them in terms of dispersion and is refined by Shannon entropy, as well as many other information or dispersion indices. Checking the peakedness relation between two probability distributions comes down to comparing, in terms of specificity, possibility distributions whose cuts are optimal prediction intervals of the original probability distributions around their mode. These possibility distributions are in fact the most specific transforms from probability to possibility, already proposed by Dubois and Prade [4], and Delgado and Moral [3] in the eighties. The paper thus establishes a new link between possibility and probability theories. The proposed qualitative com-

parison test between probability distributions may arguably be considered as the natural information ordering between probability functions, something which is not always known in the uncertainty literature. We show that this type of ordering is akin to stochastic dominance. It also corresponds to a concept of majorization, studied in the early XXth century by Hardy and colleagues [14], for comparing vectors of positive numbers having the same sum (and hence cannot be compared component-wise), and furthermore used in the social sciences for the comparison of social welfare of societies of agents [24] .

The next section introduces a generalized notion of cumulative distribution, and describes the relative peakedness of probability functions in terms of stochastic dominance with respect to a particular choice of cumulative distribution. The relation between possibilistic specificity and probabilistic peakedness is shown, noticing that the chosen form of cumulative distribution corresponds to a well-known type of probability-possibility transformation. To make the paper self-contained, a direct proof, establishing the consistency between the possibilistic specificity ordering and the probabilistic entropy measure, is given in section 3. A discussion of related work in the statistical, mathematical, and social science literature is provided in section 4. It enables the obtained results to be generalized to a large class of dispersion indices. An application of the possibilistic specificity ordering in the field of machine learning or, more specifically, the induction of decision forests is proposed in section 5.

2 A notion of Comparative Dispersion: The Peakedness Ordering

When comparing probability distributions in terms of their informativeness (or dually in terms of dispersion), it is clear that the more peaked a distribution, the more informative, the less dispersed it is. Probability distributions on finite sets can be viewed as vectors the components of which sum to 1. Because of this property, it is difficult to compare probability distributions pointwisely. Therefore, many authors resort to information indices like Shannon entropy, or dispersion indices like the Gini index. The aim of this section is to propose a notion similar to stochastic dominance that captures the notion of relative peakedness of probability distributions, and to show its close relation to possibility theory, where the pointwise comparison of possibility vectors is the natural way to go when comparing distributions in terms of specificity.

Let $\Pr(\cdot)$ be a probability measure on the real line with density $p(\cdot)$. The *cumulative* distribution of $\Pr(\cdot)$ is denoted $F^p(\cdot)$ and defined by $F^p(x) = \Pr((-\infty, x])$. When comparing random variables X_1 and X_2 with cumulative distributions $F_1(\cdot)$ and $F_2(\cdot)$ respectively, it is usual (for instance in economy) to use the notion of *stochastic dominance*: X_1 stochastically dominates X_2 if and only if $F_1 \leq F_2$ (pointwisely). Stochastic dominance can be equivalently defined in terms of *survival functions* $S^p(x) = \Pr([x, +\infty)) : X_1$ stochastically dominates X_2 if and only if $S_1 \geq S_2$ (pointwisely). Strict dominance holds when $S_1 \geq S_2$ and $S_1(x) > S_2(x)$ for at least one x. Dominance thus defined is a natural approach to deciding whether one random variable is larger than another one, since when X_1 stochastically dominates X_2, the probability for X_1 being larger than any threshold x is always larger than the corresponding probability for X_2.

Interestingly, the notion of cumulative distribution is based on the existence of the natural ordering of numbers. Consider a probability distribution defined over a finite domain W of cardinality n. In this case, no obvious notion of cumulative distribution exists, unless W is endowed with a total preordering $\succeq$, that is, a reflexive, complete, and transitive relation:

Definition 1 *The $\succeq$-cumulative distribution of a probability distribution $p(\cdot)$ on a finite, totally preordered set $(W, \succeq)$ is the function $F^p_{\succeq} : W \to [0, 1]$ defined by $F^p_{\succeq}(w) = \Pr(\{u \in W : w \succeq u\})$.*

Consider another probability distribution $q(\cdot)$ on W. The corresponding stochastic $\succeq$-dominance relation between $p(\cdot)$ and $q(\cdot)$ can be defined by the pointwise inequality $F^p_{\succeq} \leq F^q_{\succeq}$. If the elements of W are numbered in such a way that $w_j \succeq w_i$ if and only if $i \leq j$, then $p(\cdot)$ can be viewed as a probability distribution on $\{1, 2 \ldots n\}$, and $F^p_{\succeq}$ coincides with a genuine survival function of $\Pr(\cdot)$ on $\{1, 2 \ldots n\}$. In other words, a generalized cumulative distribution can always be considered as a simple one, up to a reordering of elements. In the following, $p(w_i)$ is denoted p_i for short.

A probability distribution $p(\cdot)$ is more peaked than another one $q(\cdot)$ if the elements of W are more tightly clustered around the most frequent item(s) according to $p(\cdot)$ than around the most frequent item(s) according to $q(\cdot)$. Consider $\succeq$-cumulative distributions of $p(\cdot)$ and $q(\cdot)$, with respect to the orderings induced, respectively, by the probabilities p_i and q_i: $x_i \succeq^p x_j$ iff $p_i \leq p_j$ and $x_i \succeq^q x_j$ iff $q_i \leq q_j$. It is possible to use such generalized cumulative distributions to decide whether a probability distribution $p(\cdot)$ is more peaked than another one $q(\cdot)$. The idea is to define mappings from W to natural numbers $\{1, 2 \ldots |W|\}$ that correspond to the above suggested re-orderings of elements

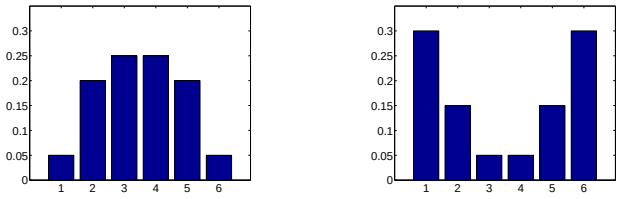

Fig. 1. The probability distribution on the left is (strictly) less peaked than the one on the right.

from the most probable to the least probable, and to use stochastic dominance on $\{1, 2 \ldots |W|\}$ to compare $p(\cdot)$ and $q(\cdot)$, the "largest" random variable on the integers corresponding to the most peaked one on W.

Let $a = O(p)$ be the *ordered* probability vector obtained from $p(\cdot)$ by rearranging the probability degrees p_i in a non-increasing order. That is,

$$a = (a_1 \ldots a_n) = (p_{\sigma(1)} \ldots p_{\sigma(n)}),$$

where σ is a permutation of $\{1 \ldots n\}$ such that $p_{\sigma(i)} \geq p_{\sigma(j)}$ for $i < j$. Likewise, we denote by $b = (b_1 \ldots b_n) = O(q)$ the ordered probability vector associated with $q(\cdot)$. Now, a and b can be viewed as probability distributions over the set $\{1, 2 \ldots n\}$.

Obviously, $F^p_{\succeq^p}(x_{\sigma(i)}) = \Pr(\{i \ldots n\}) = \sum_{k=i}^{n} a_k = S^a(i)$, in terms of survival functions:

Definition 2 *A probability distribution $p(\cdot)$ on W is said to be more peaked than a probability distribution $q(\cdot)$ in the wide sense if and only if $S^a(i) \leq S^b(i)$ for all $i = 1 \ldots n$, where $a = O(p)$, $b = O(q)$.*

The meaning of this definition is that if a random variable X_1 on W is more peaked than X_2, then for any integer i, the probability of picking a realization of X_1 not among the i most probable ones is less or equal to the probability of picking a realization of X_2 not among the i most probable ones. Hence, relative peakedness can be viewed as stochastic dominance in the appropriate space.

Example 3 *For the two probability distributions specified by the probability vectors*

$$p = (\ .05\ .20\ .25\ .25\ .20\ .05\ ),$$

$$q = (\ .30\ .15\ .05\ .05\ .15\ .30\ )$$

(see Fig. 1 for a graphical illustration) we obtain

$$S^a = (\,1.0\ .75\ .50\ .30\ .10\ .05\,),$$
$$S^b = (\,1.0\ .70\ .40\ .25\ .10\ .05\,).$$

Since $S^a \geq S^b$ *(and* $S^a(2) > S^b(2)$*),* $p(\cdot)$ *is (strictly) less peaked than* $q(\cdot)$*.*

2.2 Relative Peakedness and Possibilistic Specificity

A possibility distribution $\pi(\cdot)$ is a mapping from W to the unit interval such that $\pi(w) = 1$ for some $w \in W$. A possibility degree $\pi(w)$ expresses the absence of surprise about w being the actual state of the world. It generates a set function $\Pi(\cdot)$ called a possibility measure such that $\Pi(A) = \max_{w \in A} \pi(w)$. The degree of necessity (certainty) of an event A is computed from the degree of possibility of the complementary event A^c as $N(A) = 1 - \Pi(A^c)$.

In the following definition, we recall a basic notion from possibility theory (e.g. Dubois et al. [10]) already mentioned in the introduction.

Definition 4 *We say that a possibility distribution* $\pi(\cdot)$ *is* more specific *than a possibility distribution* $\rho(\cdot)$ *iff* $\pi \leq \rho$ *pointwisely. It is strictly more specific if* $\pi \leq \rho$ *and* $\pi(w) < \rho(w)$ *for at least one* $w \in W$*.*

Clearly, the more specific $\pi(\cdot)$, the more informative it is. If $\pi(w_i) = 1$ for some w_i and $\pi(w_j) = 0$ for all $j \neq i$, then $\pi(\cdot)$ is maximally specific (full knowledge); if $\pi(w_i) = 1$ for all i, then $\pi(\cdot)$ is minimally specific (complete ignorance).

A numerical degree of possibility can be viewed as an upper bound to a probability degree [8]. Namely, with every possibility distribution $\pi(\cdot)$ one can associate a non-empty family of probability measures dominated by the possibility measure:

$$\mathcal{P}(\pi) = \{\, \Pr(\cdot) \mid \Pr(A) \leq \Pi(A) \text{ for all } A \subseteq W \,\}.$$

On such a basis, it is possible to change representation from possibility to probability and conversely. Changing a probability distribution into a possibility distribution means losing information as the variability expressed by a probability measure is changed into incomplete knowledge or imprecision. Some principles for this transformation have been suggested in [9]. They come down to selecting a most specific element from the set of possibility measures dominating $\Pr(\cdot)$, that is,

$$\forall A \subseteq W\ :\ \Pi(A) \geq \Pr(A)$$

with $\Pi(A) = \max_{w \in A} \pi(w)$ and $\Pr(A) = \sum_{w \in A} p(w)$. A minimal consistency between the ordering induced by the probability distribution and the one of the possibility distribution, $\pi(w) > \pi(w')$ whenever $p(w) > p(w')$, is also required.

Let $\pi = T(p)$ be the possibility distribution derived from the probability distribution $p(\cdot)$ according to the following probability-possibility transformation suggested by Dubois and Prade [4]:

$$\pi_i = \sum_{j=i}^{n} a_j, \qquad i = 1 \ldots n. \tag{1}$$

where $a = O(p)$ and π_i is short for $\pi(w_{\sigma(i)})$. Obviously, $1 = \pi_1 \geq \ldots \geq \pi_n$. Moreover, the possibility measure $\Pi(\cdot)$ associated with $\pi(\cdot)$ dominates the corresponding probability measure $\Pr(\cdot)$.

It turns out that $T(p)$ is a maximally specific element of the family of possibility measures that dominate the probability function $\Pr(\cdot)$ induced by the distribution $p(\cdot)$; see Dubois and Prade [4], and Delgado and Moral [3]. Moreover, if the ordering induced by $p(\cdot)$ on W is linear (i.e., $a_i \neq a_j$ for all $i \neq j$) then $T(p)$ is the *unique* maximally specific possibility distribution which dominates $\Pr(\cdot)$ and respects the ordering induced by the probability assignment. When there are elements of equal probability, the uniqueness of the maximally specific dominating possibility distribution can be recovered if the ordering induced by $\pi(\cdot)$ on W is requested to be the same as the ordering induced by p (but then the equation defining $T(p)$ must be adjusted accordingly). The transformation T is hence called optimal.

It is patent that the possibility function $T(p)$ coincides with the survival function S^a with respect to the ordering induced by the probability values, as defined in the previous section.

In fact, any generalized cumulative (with respect to a weak order $\succeq$ on W) distribution $F^p_{\succeq}$ of a probability measure $\Pr(\cdot)$ with distribution $p(\cdot)$ on W can be viewed as a possibility distribution the associated measure of which dominates $\Pr(\cdot)$, i.e., $\max_{w \in A} F^p_{\succeq}(w) \geq \Pr(A), \forall A \subseteq W$. This property holds because a (generalized) cumulative distribution is constructed by computing the probabilities of events $\Pr(A)$ in a nested sequence defined by the ordering relation.

Probability-possibility transformations have been extended to the real line by Dubois et al. [9] (see also Dubois et al. [11]). Let $p(\cdot)$ be a unimodal continuous probability density with mode m. Suppose one tries to represent this information by means of an interval I. Intuitively, I must be narrow enough to be informative, and its probability must be high enough to let I be credible. It can be proved that the most narrow prediction interval I such that $\Pr(I) \geq \lambda$,

where λ is a fixed confidence level, is of the form $I_\lambda = \{\, x \,|\, p(x) \geq \theta \,\}$ for some threshold θ. Then, the most specific possibility transform (inducing the same ordering as $p(\cdot)$ on the real line) is $\pi = T(p)$ such that

$$\forall\, x \in R \,:\, \pi(x) = \pi(y) = 1 - \Pr([x, y]),$$

where $[x, y] = I_{p(x)}$. Clearly, $\pi(m) = 1$.

In this case, define an ordering relation $\geq_m$ on the real line such that $x \geq_m y$ if and only if $|m - x| \geq |y - m|$; then $\pi(x) = S_m(x)$ is the survival function of $p(\cdot)$ with respect to the ordering $\geq_m$.

As a result of this subsection, the peakedness relation for the comparison of probability functions can be described in terms of the relative specificity of their optimal probability transforms.

Definition 5 *Let $\pi = T(p)$ be the transformation (1) of an ordered probability vector a, i.e., $\pi_i = \sum_{j=i}^{n} a_j$. We say that a probability distribution $p(\cdot)$ on a finite set W is* more peaked *than a distribution $q(\cdot)$ on W iff $\pi_i \leq \rho_i$ for all $1 \leq i \leq n$, where $\pi = T(O(p))$ and $\rho = T(O(q))$. We say that $p(\cdot)$ is* strictly *more peaked than $q(\cdot)$ if it is more peaked and $\pi_i < \rho_i$ for at least one index $i \in \{1 \ldots n\}$.*

In the previous numerical example 1, $\pi = S^a$ and $\rho = S^b$, and $p(\cdot)$ is (strictly) less peaked than $q(\cdot)$ because $\pi(\cdot)$ is (strictly) less specific than $\rho(\cdot)$.

Subsequently, the peakedness relation is understood in the sense of this definition. The "less peaked than" relation is obviously invariant under permutations of the involved probability vectors. Therefore, we restrict our attention to ordered probability or possibility vectors in the next section.

3 From Peakedness to Dispersion Indices

The aim of this section is to prove that the peakedness relation, which is expressed in terms of possibilistic specificity, is consistent with the ordering of probability distributions induced by Shannon entropy and many other dispersion indices. As will be seen in the next section, this result is not completely new, and related results already exist in mathematics and some other fields outside the uncertainty community. However, to make the paper self-contained, we provide an explicit direct (and to the best of our knowledge novel) proof.

The most popular probabilistic information index is entropy.

Definition 6 *The* entropy *of a probability distribution $p(\cdot)$ is defined by*

$$E(p) \;=\; -\sum_{j=1}^{n} p_j \cdot \log p_j. \tag{2}$$

In the following, we consider a generalized form of entropy defined by

$$\Delta_\phi(p) = \sum_{j=1}^{n} \phi(p_j), \tag{3}$$

where the function $x \mapsto \phi(x)$ is strictly concave on $(0,1)$. (Note that, in particular, the function $x \mapsto -x\log(x)$ is strictly concave on $(0,1)$: its second derivative is given by $x \mapsto -1/x$). The family (3) covers many dispersions indices, for instance the Renyi family

$$\Delta_k(p) = \frac{\sum_{j=1}^{n}(p_j)^k - 1}{2^{1-k} - 1}$$

This is a concave function with $\lim_{k\to 1} \Delta_k(p) = E(p)$. The quadratic case ($k = 2$) is often considered. Besides, $\lim_{k\to 0} \Delta_k(p) = |\{i|p_i > 0\}| - 1$, the latter being the size of the support of $p(\cdot)$.

The main result of this paper claims that the ordering induced by the Δ_ϕ ordering (hence entropy in particular) refines the peakedness relation:

Theorem 7 *If a probability vector a is less peaked than a vector b, then $\Delta_\phi(a) \geq \Delta_\phi(b)$; if a is strictly less peaked than b, then $\Delta_\phi(a) > \Delta_\phi(b)$.*

Below, we shall prove this theorem in the following way: We construct a sequence of probability vectors $a^0, a^1 \ldots a^m$ such that $a^0 = a$, $a^m = b$ and a^{k+1} is more peaked than a^k. Moreover, this sequence will satisfy $\Delta_\phi(a^k) \geq \Delta_\phi(a^{k+1})$ (resp. $\Delta_\phi(a^k) > \Delta_\phi(a^{k+1})$) for all $1 \leq k \leq m-1$.

Remark 8 *Simple counterexamples can be constructed showing that an implication in the other direction, for instance that $E(a) \geq E(b)$ implies a to be less peaked than b, does not hold. In fact, such an implication cannot be expected since the entropy measure induces a total preorder on the class of probability measures, whereas the peakedness relation defines only a partial ordering. In other words, the former ordering is a proper refinement of the latter one.*

Other interesting indices fitting the framework are the Bayes probability of error, $e(p) = 1 - \max_i p_i$, and the Gini index

$$G(p) = \sum_{i,j=1}^{n} \min(p_i, p_j) - 1.$$

Interestingly, this information index is closely related to the following probability-possibility transform[5]

$$\hat{\pi}_i = \sum_{j=1}^{n} \min(p_i, p_j)$$

since $G(p) - 1$ is equal to its amount of imprecision $\sum_{j=1}^{n} \hat{\pi}_j$. It is easy to show that it can also be written in terms of the imprecision of the optimal probability-possibility transform (1) since $G(p) = 2\sum_{j=1}^{n} \pi_j$ as well. So, the above theorem trivially holds for the Gini index. It holds in the wide sense for $e(p)$. Other information indices can be found in [23].

3.2 Auxiliary Result

Let a and b denote two (ordered) probability vectors such that a is strictly less peaked than b. Starting with $a^0 = a$, a distribution a^{k+1} will be obtained from a distribution a^k by shifting a part of the probability mass a_j^k to a_i^k for appropriately defined indices $j > i$. More generally, a shifting operation $S(a, i, j, c)$ will transform an ordered vector $a = (a_1 \dots a_i \dots a_j \dots a_n)$ into the ordered vector

$$a^c = (a_1 \dots a_i + c \dots a_j - c \dots a_n).$$

Note that if $\pi = T(a)$ and $\pi^c = T(a^c)$ denote, respectively, the possibilistic transforms of a and a^c, then

$$\pi_k^c = \begin{cases} \pi_k & \text{if} \quad k \leq i \\ \pi_k & \text{if} \quad j < k \\ \pi_k - c & \text{if} \quad i < k \leq j \end{cases} \tag{4}$$

Thus, $\pi^c \leq \pi$ does obviously hold true, and a^c is strictly more peaked than a in the case where $c > 0$.

To guarantee a shifting operation $S(a, i, j, c)$ to be valid in the scope of turning a into b, the choice of c must satisfy the following conditions:

(i.) Proper ordering : $a_{i-1} \geq a_i + c$ and $a_j - c \geq a_{j+1}$
(ii.) Limited increase of specificity: $\pi^c \geq \rho = T(b)$

Recalling (4), the latter item means that

$$\pi_k^c \;=\; \sum_{i=k}^{n} a_i - c \;\geq\; \sum_{i=k}^{n} b_i \;=\; \rho_k$$

for all $i < k \leq j$. Define $d_k = a_k - b_k$. Since $\pi = T(a) \geq T(b) = \rho$ by assumption, we have $\sum_{m=k}^{n} d_m \geq 0$ for all $1 \leq k \leq n$. The condition $\pi^c \geq \rho$ can thus be written as

$$\forall i < k \leq j \;:\; c \leq \sum_{m=k}^{n} d_m.$$

To satisfy both (i.) and (ii.), we hence need

$$c \leq \min\left(\min_{i<k\leq j} \sum_{m=k}^{n} d_m,\; a_{i-1} - a_i,\; a_j - a_{j+1} \right). \tag{5}$$

Since $a \neq b$, there exists $j = \max\{k \,|\, a_k \neq b_k\}$. Of course, $a_j > b_j$ since $\pi \geq \rho$. By definition, we also have $d_j = a_j - b_j = \pi_j - \rho_j$. Since a and b are probability distributions, there must be some $i < j$ such that $b_i > a_i$. So, let

$$i \;=\; \max\left\{k \,|\, 1 < k \leq n,\, b_k > a_k \text{ and } a_{k-1} > a_k\right\} \tag{6}$$

if the set on the right-hand side is not empty (as will be assumed for the time being).

In order to simplify the upper bound on the number c, we first derive a lower bound on the quantity $\min_{i<k\leq j} \sum_{m=k}^{n} d_m$ that appears as the first argument on the right-hand side of (5).

Lemma 9 $\min_{i<k\leq j} \sum_{m=k}^{n} d_m \;\geq\; \min\{a_j - b_j, b_i - a_i\}$ *for all* $i < j$.

Proof: Define $D(k) = \sum_{m=k}^{n} d_m$ and $D = \min_{i<k\leq j} D(k)$. We consider two cases:
(a) $D = D(j)$. In this case, the lemma obviously holds, since $d_m = a_m - b_m = 0$ for $m > j$ and hence $D(j) = a_j - b_j$.
(b) $D < D(j)$. In this case, there must be an index k_0 with $i < k_0 < j$ and such that $D(k_0) < D(k_0 + 1)$. We claim that

$$D(i+1) < D(i+2) < \ldots < D(k_0). \tag{7}$$

In fact, since $D(k_0) < D(k_0 + 1)$ we have $a_{k_0} < b_{k_0}$. Thus, either $i = k_0 - 1$ (in which case (7) does trivially hold), or $a_{k_0-1} = a_{k_0}$ (since if $a_{k_0-1} > a_{k_0}$ and $a_{k_0} < b_{k_0}$, the index k_0 is a potential candidate for the choice of i). In the latter case, $a_{k_0-1} = a_{k_0} < b_{k_0} \leq b_{k_0-1}$ and therefore

$$D(k_0 - 1) = D(k_0) + (a_{k_0-1} - b_{k_0-1}) < D(k_0).$$

This argument can be repeated, showing that (7) does indeed hold. This in turn shows that $D = D(i+1)$. Moreover, we then have

$$D = D(i+1) = D(i) - (a_i - b_i) = D(i) + (b_i - a_i) \geq b_i - a_i$$

since $D(i) \geq 0$.
Overall, we get $D \geq a_j - b_j$ in case (a) and $D \geq b_i - a_i$ in case (b). Thus, the lemma does indeed hold. Q.E.D.

Now, if we let

$$c = \min\left(a_j - b_j,\, b_i - a_i,\, a_{i-1} - a_i \right) \tag{8}$$

then the above results and the fact that

$$a_j - a_{j+1} = (a_j - b_j) + (b_j - a_{j+1}) \geq (a_j - b_j) + (b_{j+1} - a_{j+1}) = (a_j - b_j)$$

guarantee that (5) is satisfied. Moreover, the constant c is strictly positive, since $a_{i-1} - a_i > 0$, $a_j - b_j > 0$, $b_i - a_i > 0$ by construction.

Let us now turn to the case where the right-hand side of (6) is empty.

Lemma 10 *Suppose that a is less peaked than b, and that the right-hand side on (6) is empty. Then $b_1 > a_1$.*

Proof: Suppose that a is less peaked than b. There is some $k < j$ such that $b_k > a_k$. Since the right-hand side on (6) is empty, it holds that $b_u > a_u$ implies $a_u = a_{u-1}$ for all $u < j$. Moreover, since $b_k > a_k$, this implies in turn $b_{k-1} \geq b_k > a_{k-1}$. The fact that $b_1 > a_1$ follows immediately by repeating this argument. Q.E.D.

Regarding the choice of c in the case of an empty right-hand side in (6), the only difference concerns the condition $a^c_{i-1} \geq a^c_i$ which simply becomes unnecessary. Hence, one can define

$$c = \min\left(a_j - b_j,\, b_1 - a_1 \right) \tag{9}$$

and apply the shifting operation $S(a, 1, j, c)$ in the same way as before.

3.3 Proof of the Main Result

Obviously, if the quantity c as defined in (8) (resp. (9)) is shifted from position j to position i (resp. position 1), then either $a^c_j = b_j$ or $a^c_i = b_i$ or $a^c_i = a_{i-1}$. In any case, at least one of the indices i or j will have a smaller value in the next iteration. Hence, the process of repeating the shifting operation, with i,

j, and c as specified above, is well-defined, admissible and turns a into b in a finite number of steps.

Given the above results, Theorem 7 follows immediately from the next lemma (recall that in each step of our iterative procedure, the constant c shifted from index j to index i is strictly positive):

Lemma 11 *Let* $\Delta_\phi(a) = \sum_{j=1}^{n} \phi(a_j)$ *where* ϕ *is strictly concave. Then* $\Delta_\phi(a) > \Delta_\phi(a^c)$ *for* $c > 0$.

Proof: It is easy to see that $\Delta_\phi(a) > \Delta_\phi(a^c)$ is equivalent to

$$\phi(a_i + c) + \phi(a_j - c) < \phi(a_j) + \phi(a_j).$$

Noting that $a_i > a_j$, this inequality holds because, by definition, the function $x \mapsto \phi(x)$ is strictly concave on $(0, 1)$. Q.E.D.

The above results show that the peakedness ordering proposed here underlies many probabilistic information indices, which turn out to be in agreement with possibilistic specificity. Theorem 7 is in particular valid for the standard Shannon entropy, and the logarithm $\log(\cdot)$ in (2) can be replaced by any monotone increasing function $F(\cdot)$ the second derivative $F''(\cdot)$ of which exists on $(0, 1)$ and satisfies $F''(x)/F'(x) > -2/x$ for all $0 < x < 1$ (where $F'(\cdot)$ denotes the first derivative). In fact, one might thus be tempted to require the property of coherence with the possibilistic specificity as a minimal prerequisite for any probabilistic measure of dispersion. That is, any index of dispersion D should satisfy the following axiom: For any probability assignments $p(\cdot)$ and $q(\cdot)$, define $\pi = T(O(p))$ and $\rho = T(O(q))$; if $\pi \geq \rho$ then $D(p) \geq D(q)$. Additional properties can then be required for selecting a particular dispersion index.

4 Related Work

The above results are in some sense not completely new. This section surveys three areas where closely related results or ideas can be found. First, we give a precise account of old mathematical results around a notion of *majorisation*, which is a generalization of peakedness to any vector of positive real values. Then, we note the presence of similar concerns in statistics, that originally inspired our work. Finally, we point out the application of majorisation in the social sciences.

The well-known book by Hardy, Littlewood, and Polya [14] [2] contains technical results that are equivalent to the main results of this paper. In section 2.18 of the book, the authors are interested in comparing vectors of values, the sum of components of which are equal (for instance probability assignments). Suppose a and b are two vectors of values arranged in decreasing order ($a_1 \geq a_2 \geq \cdots \geq a_n$), and whose sums of components are equal. They say that a is *majorised* by b if and only if $\sum_{i=1}^{j} a_i \leq \sum_{i=1}^{j} b_i, \forall j = 1 \ldots n$, which is equivalent to $\sum_{i=j}^{n} a_i \geq \sum_{i=j}^{n} b_i$ due to the equality $\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} b_i$. Thus, the majorisation of a by b exactly coincides with the fact that b is more peaked than a.

The question motivating the majorisation relation is that of comparing expressions called *symmetric means*, which consist of the average of $n!$ terms of the form $\prod_{j=1}^{n} u_i^{\alpha_i}$ ($u_i > 0, \alpha_i \geq 0$); the latter are obtained by the possible permutations of the coefficients u_i. As such a symmetric mean is stable under permutations of the $\alpha_i's$, comparing symmetric means, denoted $[\alpha]$, having different α exponents comes down to comparing the arranged vectors a. Hardy et al. prove that $[\alpha] \leq [\beta]$ as soon as a is majorised by b, the equality holding only when $[\alpha] = [\beta]$ or the coefficients u_i are equal. Interestingly, the result is proved using an elementary transfer notion of the form used above in section 3.2.

The authors then go on proving another result providing a necessary and sufficient condition for a to be majorised by b. Namely, they notice that this is equivalent to any component of a being a certain form of weighted average of the components of b. Namely, there exists a non-negative $n \times n$ weight matrix M such that the sum of elements in each row and each column is 1 (a so-called bistochastic matrix), and a is majorised by b if and only if $a = Mb$. A function $\Delta(p)$ mapping probability distributions to reals is said to be Schur-concave if $\Delta(Mb) \geq \Delta(b)$ for all bistochastic matrices M; see [23] for details and references.

In section 3.17 of the book, Hardy et al. prove a strong form of theorem 7, namely that $\sum_{j=1}^{n} \phi(a_j) \leq \sum_{j=1}^{n} \phi(b_j)$ holds for all continuous and convex functions ϕ if and only if a is majorised by b. To prove the result they show that the majorization relation can be induced by a suitable choice of the function ϕ, and the converse becomes obvious using the equivalent form $a = Mb$ since for convex functions the image of a weighted average of a set of values is less than the weighted average of the images of the values (Schur concavity).

Moreover, in the case when ϕ has positive second derivative everywhere, then

[2] A more modern text on majorisation is the one of Marshall, and Olkin [20].

$\sum_{j=1}^{n} \phi(a_j) = \sum_{j=1}^{n} \phi(b_j)$ only when the sets of coefficients in a and b are the same.

Our proof in the previous section is self-contained as it relates peakedness and dispersion indices in a direct way. The result of Hardy et al. indicates that the peakedness relation is the intersection of all total order relations induced by all dispersion indices of the form Δ_ϕ for a concave ϕ function.

4.2 Statistics

The term "peakedness" was coined by Birnbaum. In a paper in 1948 [1], he dealt with what he called the *quality* of a probability distribution, referring to its peakedness. Considering that the fourth moment of a distribution is not an appropriate measure of peakedness, he proposed a definition of the relative peakedness of distributions as follows:

Definition 12 *Let Y_1 and Y_2 be real random variables, associated with respective probability spaces $(\Omega_1, \mathcal{A}_1 \Pr_1)$, $(\Omega_2, \mathcal{A}_2 \Pr_2)$, and y_1 and y_2 real constants. Y_1 is said to be more peaked about y_1 than Y_2 about y_2 if and only if*

$$\Pr_1(\mid Y - y_1 \mid \geq t) \;\leq\; \Pr_2(\mid Z - z_1 \mid \geq t)$$

holds for all $t \geq 0$.

It is clear that the function

$$\pi_y(y_1 - t) = \pi_y(y_1 + t) = \Pr(\mid x - y_1 \mid \geq t) = 1 - \Pr([y_1 - t, y_1 + t])$$

is a possibility distribution, and easy to show that for any choice of y_1, its possibility measure dominates $\Pr(\cdot)$; see Dubois et al. [11]. In this paper, we adapted this definition in two ways: First, the results on the probability-possibility transforms clearly indicate that for unimodal densities, choosing y_1 as the mode of the distribution is reasonable. Moreover, Birnbaum [1] considers intervals whose common midpoint is y_1, yielding a symmetric possibility distribution even if the density is not symmetric by itself. Instead of intervals of the form $[y_1 - t, y_1 + t]$, we used intervals of the form $\{x \mid p(x) \geq \theta\}$, since they lead to a possibility distribution of the same shape as the probability density (and peakedness refers to the shape of this density anyway). The reason for this choice is that the width of intervals with a fixed confidence level is thus minimized. This change of definition enables peakedness to be defined for any referential set, not just the reals. Indeed, the set $\{x \mid p(x) \geq \theta\}$ makes sense in general, if measurability is ensured, while $[y_1 - t, y_1 + t]$ assumes the real line as an underlying domain. Here, we nevertheless restricted ourselves to the case of a finite referential set, because entropy indices are usually applied to such domains.

Now, for $\pi = T(a)$ it is clear that $\pi_i = 1 - \Pr(\{x \mid \Pr(\{x\}) \geq \theta\})$ if $a_{i-1} \geq \theta > a_i$, which recovers our variant of the original peakedness relation due to Birnbaum.

4.3 Social Sciences

Even though the proposed notion of relative informativeness, based on possibilistic specificity and Birnbaum peakedness, seems to be relatively unknown in the uncertainty literature, there is a subfield of the social sciences where the results obtained by Hardy et al. have apparently been exploited for some twenty years or so, in the study of social welfare orderings, and in particular, the modeling of social inequalities. [3]

We refer to the book by Moulin [24]. In this framework, W is a set of agents, whose welfare under some life conditions is measured by a utility function over W. The problem is to compare the quality of utility vectors $(u_1 \dots u_n)$ from the standpoint of social welfare. Under an egalitarian program of redistribution from the rich to the poor, the so-called Pigou-Dalton principle of transfer states that transferring some utility from one agent to an other one so as to reduce inequalities of utility values improves the social welfare of the population.

Formally, the transformation of a vector a into a vector a^c as in subsection 3.2 is known as a Pigou-Dalton transfer. The sequence of transformations we propose here is also used in this literature. Moreover, the role of entropy is played by so-called inequality indices. The counterpart to the possibility transform of a probability vector is called the Lorentz curve of the utility vector, and the counterpart of the peakedness ordering is called the Lorentz dominance relation.

One difference is that utility vectors do not sum to 1. But Lorentz dominance is precisely making sense for the comparison of utility vectors with equal sum. In this literature, dispersion indices are called inequality indices, and those of the form Δ_ϕ are called Atkinson indices.

Note that it would not be the first time that possibility-probability transformations find counterparts in the social sciences. For instance, a transformation from a belief function to a probability measure (obtained by generalizing the Laplace indifference principle) introduced in [4] and called *pignistic transformation* by Smets [29] is known in the social sciences as the Shapley value of cooperative games (see again Moulin [24]).

[3] The authors are grateful to Jérôme Lang for pointing out this connection.

5 An Application in Machine Learning

The entropy measure and related dispersion criteria are used in many research areas for diverse purposes. Since our results in previous sections have shown that the peakedness relation for probability distributions and, hence, the associated specificity ordering for possibility distributions is in agreement with entropy, the former could in principle be used as an alternative to the latter, at least if the potential incomparability between distributions is tolerated. In fact, recall that the entropy measure induces a total preorder on the set of probability distributions over a set $\mathcal{X}$, whereas peakedness only provides a partial order. On the other hand, while the latter seems to be a natural ordering in many applications, its refinement by means of the entropy measure is often done just because entropy is better known than other dispersion indices. To make this point concrete, the current section gives an example of the applicability of the peakedness relation as an alternative to the entropy measure in the field of machine learning.

5.1 Information Measures in Decision Tree Induction

A standard problem in supervised machine learning is to induce a classification function $\mathcal{X} \to \mathcal{Y}$ from a set of training examples $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \{y_1 \dots y_k\}$ is a finite set of elements called class labels. Instances x_i are typically characterized in terms of a feature vector of fixed length, i.e., the input space $\mathcal{X}$ is the Cartesian product of the domains of a fixed set of attributes (features); subsequently, we make the simplifying assumption that all these domains are finite.

The key idea of decision tree induction ([26]), by now one of the most popular machine learning methods, is to partition a set of training examples in a recursive manner, thereby producing a partitioning of the input space into decision regions that can be represented in terms of a tree structure. In the simplest case, partitioning is accomplished through (univariate) tests of the form $[F(x) = f_j]$, $j = 1 \dots m$, where F is a feature with domain $\{f_1 \dots f_m\}$ and $F(x)$ denotes the attribute value of the instance x. Each inner node of a decision tree is associated with a test of that kind and, hence, splits a subset of examples according to the value of the attribute F.

The generalization performance of a classification function in the form of a decision tree strongly depends on the selection of appropriate splitting attributes. Roughly speaking, all common learning algorithms seek to induce a "simple" tree, since the generalization performance of simple models is sup-

posedly superior to that of complicated models. [4] To make a selection at an inner node of a tree, each candidate attribute F is typically evaluated in terms of the *information gain*

$$E(Y) - E(Y \mid F), \tag{10}$$

where

$$E(Y) = -\sum_{i=1}^{k} p(y_i) \cdot \log p(y_i)$$
$$E(Y \mid F) = -\sum_{j=1}^{m} p(f_j) \sum_{i=1}^{k} p(y_i \mid f_j) \cdot \log p(y_i \mid f_j)$$

Here, $\{f_1 \dots f_m\}$ is the domain of attribute F, and $E(Y)$ is the Shannon entropy of the class distribution $(p(y_1) \dots p(y_k))$ in the current example set (i.e., $p(y_i)$ denotes the relative frequency of class label y_i among current examples). Moreover, $E(Y \mid F)$ is the conditional entropy of Y given F, namely a weighted average of the entropies of the class distributions in the subsets of examples that are produced by splitting according to the values of F. In probability theory, (10) is also known as the *mutual information*, i.e., the relative entropy between the joint distribution (of Y and F) and the product of the marginals. Despite this apparent theoretical justification, it is worth mentioning that selecting splitting attributes with maximal information gain is merely a heuristic approach which does not guarantee to produce a tree of minimal size. [5]

In the best case, an attribute splits a set of examples into "pure" subsets, i.e., subsets in which all examples do have the same class label; since a pure set of examples does not necessitate further splits, it defines a leaf of the decision tree that can reliably be labeled by the corresponding class. [6] As opposed to this, the worst situation is an example set with a uniform distribution over $\mathcal{Y}$, since this distribution does not suggest any particular classification. These two extreme situations are correctly captured by the entropy (information gain) measure. One might argue, however, that the interpolation between them, even if being based on the theoretically sound concept of mutual information, remains arbitrary to some extent. In fact, from a classification point of view, it is not obvious why a class distribution like $p = (0.5, 0.5, 0, 0)$ should be preferred to $q = (0.7, 0.1, 0.1, 0.1)$. More specifically, there is no obvious reason to expect the former to become "pure" (by further splitting) before the latter. And indeed, experimental studies [22] have shown that using entropy in (10) is neither superior nor inferior to using alternative information measures such as, e.g., the Gini index.

[4] This is the principle of Occam's razor.

[5] Besides, information gain in its basic form suffers from other problems such as, e.g., a systematic preference for attributes with many values.

[6] To prevent overfitting the data, splitting is usually stopped earlier.

In contrast, the peakedness relation can well be motivated from a classification point of view. Roughly speaking, if a distribution $p = (p_1 \dots p_k)$ over the class labels $\mathcal{Y}$ is more peaked than a distribution q, then classifying on the basis of p is easier or better than classifying on the basis of q. For example, since $p_1 \geq q_1$ (suppose that the distributions have already been reordered such that $p_1 \geq p_2 \geq \ldots \geq p_k$), the probability to guess the class label of a query instance $x_0 \in \mathcal{X}$ correctly is higher for p than for q. More generally, suppose that a prediction in terms of a *credible set*[7] of labels $C \subseteq \mathcal{Y}$ is desired. Such a prediction should reasonably consist of the $k' \leq k$ classes with highest probability, and since

$$\sum_{i=1}^{k'} p_i \geq \sum_{i=1}^{k'} q_i$$

for all $k' = 1 \dots k$, the credible sets derived from p have higher confidence than those derived from q, regardless of the size k'.

For the same reason, better performance is achieved in an alternative prediction scenario where, instead of estimating the class label of the query instance only once, this label must be guessed repeatedly until the true label is found [16]: the expected number of futile trials is then smaller for p than for q.

5.2 Lazy Decision Tree Learning

A lazy variant of decision tree learning has been introduced in [13]. This variant generates a separate classification tree for each query instance x_0. More specifically, it only generates one branch of the tree, namely the one which is needed to classify x_0. The test predicates along this branch are particularly tailored to the query: The splitting criterion (10) obviously seeks to maximize the information gain on *average*: $E(Y \mid F)$ is a weighted average of the form

$$\sum_{j=1}^{m} p(f_j) E(Y \mid f_j), \tag{11}$$

where the weights $p(f_j)$ are the (estimated) probabilities to encounter an instance x with $F(x) = f_j$. This strategy, however, is not reasonable if the instance to be classified is already known in advance. In other words, given that the attribute value $F(x_0)$ of the query is known, the entropy of the class distribution in those subsets of examples with a different value $f_j \neq F(x_0)$ is actually irrelevant. Correspondingly, instead of averaging the entropy over all these subsets, the lazy variant tries to maximize the information gain when going from the current set of examples to the subset of examples x_ℓ with

[7] This term is used in Bayesian statistics.

attribute value $F(x_\ell) = F(x_0)$:

$$E(Y) - E(Y \mid F(x_0)) = \sum_{i=1}^{k} p_F(y_i) \cdot \log p_F(y_i) - \sum_{i=1}^{k} p(y_i) \cdot \log p(y_i), \qquad (12)$$

where $p_F(y_i)$ is the probability (relative frequency) of the class y_i in the subset of examples with attribute value $F(x_0)$. [8]

The lazy variant of decision tree induction, as lazy learning methods in general, is of course more costly from a computational point of view, since a new model must be generated for each query instance. On the other hand, it often outperforms standard decision tree learning in terms of predictive performance. For details of the method as well as experimental results we refer to [13].

5.3 Ensembles of Decision Trees

A so-called *decision forest* is a special type of ensemble learning technique. Here, the key idea is to generate a whole set of models instead of only a single one. Viewing each of these models as a member of a committee, predictions are then made by means of majority voting: Given a new query, each model makes a vote in favor of a particular class, and the class with the maximal number of votes is predicted. [9] Under certain conditions, ensemble methods can reduce both the bias and the variance of predictions. Roughly speaking, if each individual model is sufficiently accurate and, at the same time, the ensemble is diverse enough, it is likely that incorrect predictions will "average out".

To generate a diverse ensemble of decision trees (from the same training data), different methods are conceivable. The key idea of *random forests* [2] is to modify deterministic decision tree induction as follows: At each inner node of the tree, the attribute with maximal information gain is selected, not among all potentially available attributes, but only among a randomly chosen candidate subset of fixed size K.

Interestingly enough, our specificity ordering suggests an alternative way to generate random forests: Instead of selecting a random subset of attributes first and choosing the best among these attributes afterwards, one could proceed the other way round: First the most promising candidates are selected,

[8] For technical reasons, the examples in the parent node are first re-weighted such that all classes are equi-probable; see [13] for details.

[9] Unsurprisingly, a large number of alternatives to and refinements of this simple aggregation procedure do exist.

namely those attributes that are optimal with respect to the specificity ordering, and then one among these candidates is chosen at random. As a potential advantage of this latter approach note that it does not assume the specification of the parameter K. Roughly speaking, instead of determining the size of the candidate set in a more or less arbitrary way, it is dynamically adapted in accordance with the ambiguity of the specificity ordering.

For such *alternative random forests* (ARF) we have implemented both a standard and a lazy variant. In the lazy version, given a query instance x_0 and a subset of training examples, the probability distribution $p_F(\cdot)$ in (12) is derived for each potential feature F. An attribute F becomes a candidate if its associated distribution is not dominated by any other attribute F', i.e., if there is no F' such that $p_{F'}(\cdot)$ is (strictly) more peaked than p_F. Finally, one among these candidate attributes is chosen at random, and the example set is split according to this attribute (viz. reduced to those examples having the same value as the query). Recursive partitioning thus produces a branch whose leaf node classifies the query x_0; the corresponding prediction is given by the majority of class labels in the leaf.[10] By repeating this process a certain number of times, an ensemble of decision branches is produced, and the overall classification is made by majority voting.

While the attribute selection in the lazy version only considers the peakedness of the distribution in one subset of examples, namely the one with the same attribute value as the query, the regular version (to induce standard trees) has to use a counterpart to the weighted average (11). A relatively straightforward solution is to associate with an attribute F the following distribution:

$$\sum_{j=1}^{m} p(f_j) \cdot p^j,$$

where $\{f_1 \dots f_m\}$ is the domain of F and $p^j = (p_1^j \dots p_k^j)$ is the distribution of the class labels in the subset of examples with attribute value f_j; more precisely, p^j is the distribution after reordering, i.e., $p_1^j \geq p_2^j \geq \dots \geq p_k^j$.

5.4 Experimental Results

The main purpose of the experimental studies was to compare the random forest (RF) method with the alternative (ARF) outlined above, both in the

[10] The recursive partitioning procedure stops if either all examples belong to the same class or if all attributes have already been used. As opposed to standard decision tree learning, the lazy variant does not need pruning strategies or premature stopping conditions in order to prevent overfitting.

case of regular ("eager") and lazy learning. Further, we compared the ensemble methods with the corresponding base learners, i.e., lazy and regular decision tree learning (LazyDT and DT). All methods have been implemented under the WEKA framework [31]. Since RF is already available, we only implemented the lazy variant LazyRF (the main difference again concerns the splitting measure, which in this case is (12)). As a decision tree learner we used the WEKA implementation of C4.5 [26]. ARF, LazyARF, and LazyDT were implemented from scratch.

Experimental studies were conducted using multiple benchmark datasets from the UCI repository. All numerical attributes have been discretized in advance using Fayyad & Irani's method [12]. For the ensemble methods we always generated 50 models. Table 1 shows the classification rates for the lazy methods, estimated by 10-fold cross validation (repeated 10 times), and Table 2 the corresponding results for the regular (non-lazy) approaches.

Interpretation of the results should be done with caution, since most differences in classification performance (two methods compared on a single dataset) are statistically not significant (at the 0.05 level of a simple t-test). Still, a closer examination of the results and a look at the simple win/loss statistics in Table 3 gives a relatively clear picture: The two ensemble methods are on a par and both outperform the corresponding base learner. With regard to the use of the specificity ordering in the context of decision tree learning, we consider this as a preliminary though very promising finding that motivates a closer examination and elaboration of this idea.

6 Conclusions and Perspectives

The main contribution of this paper is a reexamination and systematic exposition of a notion of relative information content that can decide if a probability distribution is more or less uncertain (or spread out) than another one (or whether the two distributions are not directly comparable). This ordering seems to be well-known in some scientific communities while being totally unknown in other ones. The surprising result is that the aforementioned comparison between probability distributions comes down to comparing two possibility distributions in the sense of their relative specificity, that is, in terms of fuzzy set inclusion! This test seems to be natural in the sense that it exactly captures the notion of relative peakedness of distributions, thus meeting our intuition. The fact that Shannon entropy as well as the Gini index (and many other ones, potentially) refine the peakedness relation corroborates this intuition. It sheds light on the meaning of these indices, by laying bare a common feature for them. The peakedness ordering offers a minimal robust foundation for probabilistic information indices. Finding an extension of these results to

dataset	LazyDT	LazyARF	LazyRF
autos (7,205,25)	76.28(9.61)	81.62(7.97)	79.21(8.53)
wisconsin-breast-cancer (100)	96.11(2.26)	96.04(2.29)	96.35(2.32)
bridges-version1 (6,107,12)	52.51(11.69)	54.77(13.38)	54.55(11.63)
horse-colic (2,368,22)	78.89(5.79)	81.85(6.00)	80.84(5.87)
dermatology (6,366,34)	89.84(4.55)	94.02(3.80)	92.71(3.76)
pima-diabetes (2,768,8)	73.44(4.49)	74.09(4.73)	73.53(4.76)
ecoli (8,336,7)	79.85(5.29)	79.68(5.07)	80.33(5.05)
Glass (7,214,9)	70.73(10.47)	72.13(10.07)	71.80(9.97)
haberman (2,306,3)	73.59(4.91)	73.00(4.90)	73.10(3.07)
cleveland-heart-diseas (5,303,13)	76.24(6.93)	79.71(6.71)	79.08(6.41)
hungarian-heart-diseas (5,294,13)	79.63(6.83)	80.14(6.87)	80.72(6.58)
hepatitis (2,155,19)	83.46(7.49)	83.62(8.49)	84.12(8.19)
iris (3,150,4)	94.00(5.88)	94.00(6.25)	94.00(5.72)
labor (2,57,16)	85.63(13.66)	83.83(15.79)	85.10(14.41)
liver-disorders (2,345,6)	56.85(4.20)	57.03(4.00)	57.34(4.64)
lymphography (4,148,18)	78.41(9.01)	82.88(8.31)	81.74(8.65)
tic-tac-toe (2,958,9)	84.01(3.66)	92.03(2.29)	92.15(2.69)
vote (2,435,16)	94.22(3.49)	94.96(3.04)	94.71(3.27)

Table 1
Results of the experimental studies for the lazy learners: Datasets (in brackets: number of classes, examples, attributes) and classification rates (in brackets: standard deviation).

continuous probability distributions, using differential entropy for instance, is an obvious next task.

Our discussion also shows that there is a degree of freedom in the choice of these indices, namely in the case of two distributions that cannot be compared by the peakedness relation but are ranked in opposite orders by, say, the entropy and the Gini index. This point needs further study, and mathematical insight from social sciences, where axiomatization results exist, might be useful in this regard. We note, however, that the situation is the same with the specificity relation in possibility theory where several non-specificity indices have been proposed (Higashi and Klir [15], Dubois and Prade [6], Yager [30], Ramer [27]) that disagree with each other. The same difficulty can be observed in the case of belief functions (Dubois and Prade [7]).

dataset	DT	ARF	RF
autos (7,205,25)	81.13(9.20)	83.50(8.08)	82.58(8.07)
wisconsin-breast-cancer (100)	94.82(2.70)	95.61(2.64)	95.72(2.38)
bridges-version1 (6,107,12)	41.95(4.62)	50.80(12.36)	48.02(10.83)
horse-colic (2,368,22)	78.32(6.36)	78.34(6.38)	81.49(5.60)
dermatology (6,366,34)	93.46(3.58)	96.53(3.06)	95.40(2.98)
pima-diabetes (2,768,8)	73.53(4.61)	74.01(4.65)	73.57(4.62)
ecoli (8,336,7)	79.86(5.03)	80.30(4.86)	80.12(5.28)
Glass (7,214,9)	71.29(10.92)	73.25(10.10)	72.50(10.37)
haberman (2,306,3)	73.59(4.91)	73.50(4.82)	73.36(3.43)
cleveland-heart-diseas (5,303,13)	76.33(7.16)	80.40(5.63)	78.55(6.18)
hungarian-heart-diseas (5,294,13)	78.94(6.93)	81.93(7.40)	80.28(6.89)
hepatitis (2,155,19)	80.17(8.83)	82.39(8.31)	81.86(9.38)
iris (3,150,4)	93.93(5.77)	93.47(5.84)	93.80(5.78)
labor (2,57,16)	83.97(14.61)	73.90(13.89)	84.67(14.28)
liver-disorders (2,345,6)	56.85(4.20)	57.37(3.83)	57.54(3.92)
lymphography (4,148,18)	72.71(9.61)	81.87(8.93)	77.19(9.03)
tic-tac-toe (2,958,9)	85.47(3.74)	90.63(2.89)	93.74(2.18)
vote (2,435,16)	95.05(3.23)	95.47(2.84)	95.74(2.91)

Table 2
Results of the experimental studies for the regular (non-lazy) learners: Datasets (in brackets: number of classes, examples, attributes) and classification rates (in brackets: standard deviation).

The notion of peakedness is easy to understand, but, compared to entropy and other numerical indices, it is quite weak, and its efficiency in probabilistic reasoning and decision making is still unclear. In his book [25], Jeff Paris advocates the use of conditional probability statements as a natural means for expressing knowledge and the maximal entropy principle as a natural tool for selecting a reasonable default probabilistic model of this knowledge. The above results suggest that the maximal entropy principle can be replaced by a weaker minimal peakedness principle in problems with incompletely specified probability distributions. Of course, the minimally peaked distribution in agreement with the constraints may fail to be unique, and the issue of choosing between them is an intriguing one. Anyway, the peakedness relation can be used in all problems where the information content of a distribution is relevant, for example in machine learning techniques à la decision tree induction, as suggested in the previous section. These issues constitute interesting topics

	LazyDT	LazyARF	LazyRF
LazyDT		3/1/14*	2/1/15*
LazyARF	14/1/3*		9/1/8
LazyRF	15/1/2*	8/1/9	

	DT	ARF	RF
DT		3/0/15*	0/2/16*
ARF	15/0/3*		11/0/7
RF	16/0/2*	7/0/11	

Table 3
Win/tie/loss statistics for the lazy learners (left) and the standard methods (right). A * indicates statistical significance at the 0.02 level of a Fisher sign test.

of future research.

Acknowledgements: The authors are grateful to Jürgen Beringer and Jérôme Lang for helpful comments. The paper also benefited from the suggestions of two anonymous reviewers.

References

[1] Birnbaum ZW. On random variables with comparable peakedness, *Annals of Mathematical Statistics*, 19, 1948, 76-81.

[2] Breiman L. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] Delgado M. and Moral S. On the concept of possibility-probability consistency, *Fuzzy Sets and Systems* , 21, 1987 311-318.

[4] Dubois D. and Prade H. On several representations of an uncertain body of evidence, In: *Fuzzy Information and Decision Processes*, M.M. Gupta, and E. Sanchez, Eds., North-Holland, Amsterdam, 1982, pp. 167-181.

[5] Dubois D. and Prade H. Unfair coins and necessity measures, *Fuzzy Sets and Systems*, 10(1), 15-20, 1983.

[6] Dubois D. and Prade H. A note on measures of specificity for fuzzy sets, *Int. J. of General Systems*, 10, 1985, 279-283.

[7] Dubois D. and Prade H. The principle of minimum specificity as a basis for evidential reasoning, In: *Uncertainty in Knowledge-Based Systems* (B. Bouchon, R.R. Yager, eds.), Springer Verlag, 1987, 75-84.

[8] Dubois D. and Prade H. When upper probabilities are possibility measures, *Fuzzy Sets and Systems* , 49,1992 65-74.

[9] Dubois D., Prade H. and Sandri S. On possibility/probability transformations. In: *Fuzzy Logic. State of the Art*, (R. Lowen, M. Roubens, eds.), Kluwer Acad. Publ., Dordrecht, 1993, 103-112.

[10] Dubois D., Nguyen HT. and Prade H. Possibility theory, probability and fuzzy sets: misunderstandings, bridges and gaps. In: *Fundamentals of Fuzzy Sets*, (Dubois, D. Prade,H., Eds.), Kluwer, Boston, Mass., The Handbooks of Fuzzy Sets Series, 343-438. 2000.

[11] Dubois D., Foulloy L., Mauris G. and Prade H. Possibility/probability transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliable Computing* 10, 273-297, 2004.

[12] Fayyad U. and Irani KB. Multi-interval discretization of continuous attributes as preprocessing for classification learning. In *Proceedings of the 13th international Joint Conference on Artificial Intelligence*, pages 1022–1027. Morgan Kaufmann, 1993.

[13] Friedman JH., Kohavi R. and Yun Y. Lazy decision trees. In *Proceedings* AAAI–96, pages 717–724, Menlo Park, California, 1996. Morgan Kaufmann.

[14] Hardy GH., Littlewood JE. and Polya, G. Inequalities, Cambridge University Press, Cambridge UK, 1952.

[15] Higashi and Klir G. Measures of uncertainty and information based on possibility distributions, *Int. J. General Systems*, 8, 43-58, 1982.

[16] Hüllermeier E. and Fürnkranz J. Learning label preferences: Ranking error versus position error. In Proceedings IDA05, *6th International Symposium on Intelligent Data Analysis*, number 3646 in LNCS, pages 180191, Madrid, 2005. Springer-Verlag.

[17] Hunter D. Causality and maximum entropy updating. *Int. J. Approx. Reasoning, 3(1): 379–406, 1989.*

[18] Klir G. A principle of uncertainty and information invariance, *Int. J. of General Systems*, 17, 1990, 249-275.

[19] Lukasiewicz T. Credal Networks under Maximum Entropy In C. Boutilier and M. Goldszmidt, editors, Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI 2000), pp. 363-370, Stanford, California, USA, July 2000. Morgan Kaufmann, 2000.

[20] Marshall A. and Olkin I. Inequalities: a Theory of Majorization and its Applications, Academic Press, New York, 1970.

[21] Maung I. Two characterizations of a minimum-information principle in possibilistic reasoning, *Int. J. of Approximate Reasoning*, 12, 133-156, 1995.

[22] Mingers J. *An Empirical Comparison of Selection Measures for Decision-Tree Induction.* Machine Learning, 3, 319–342, 1989.

[23] Morales D., Pardo L. and Vajda I. Uncertainty of discrete stochastic systems: general theory and statistical theory, IEEE Trans. on System, Man and Cybernetics, 26, 1–17, 1996.

[24] Moulin H. *Axioms of Cooperative Decision Making*. Cambridge University Press, Cambridge, MA, 1988.

[25] Paris J. *The Uncertain Reasoner's Companion*. Cambridge University Press, Cambridge, UK, 1994.

[26] Quinlan JR. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[27] Ramer A. Possibilistic information metrics and distances: Characterizations of structure, *Int. J. of General Systems*, 18, 1990, 1-10.

[28] Shore E. and Johnson RW. Axiomatic Derivation of the Principle of Maximum Entropy and The Principle of Minimum Cross-Entropy, *IEEE Trans. on Information Theory*, pp. 26–36, 1980.

[29] Smets P. Constructing the pignistic probability function in a context of uncertainty, *Uncertainty in Artificial Intelligence* 5 (Henrion M. et al., Eds.), North-Holland, Amsterdam, 29-39, 1990.

[30] Yager RR. On the specificity of a possibility distribution, *Fuzzy Sets and Systems*, 50, 1992, 279-292.

[31] Witten IH. and Frank E. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2 edition, 2005.