ELSEVIER

Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE @ DIRECT®

Artificial Intelligence 148 (2003) 335–383

Artificial
Intelligence

www.elsevier.com/locate/artint

Possibilistic instance-based learning

Eyke Hüllermeier

Department of Mathematics and Computer Science, University of Marburg, Marburg 35032, Germany

Received 16 July 2001; received in revised form 9 August 2002

Abstract

A method of instance-based learning is introduced which makes use of possibility theory and fuzzy sets. Particularly, a possibilistic version of the similarity-guided extrapolation principle underlying the instance-based learning paradigm is proposed. This version is compared to the commonly used probabilistic approach from a methodological point of view. Moreover, aspects of knowledge representation such as the modeling of uncertainty are discussed. Taking the possibilistic extrapolation principle as a point of departure, an instance-based learning procedure is outlined which includes the handling of incomplete information, methods for reducing storage requirements and the adaptation of the influence of stored cases according to their typicality. First theoretical and experimental results showing the efficiency of possibilistic instance-based learning are presented as well.

Keywords: Possibility theory; Fuzzy set theory; Machine learning; Instance-based learning; Nearest neighbor classification; Probability

1. Introduction

A major theme in machine learning concerns the problem of induction, that is the creation of general knowledge from particular examples or observed data. In this respect, *uncertainty* plays a fundamental role. To begin with, the data presented to learning algorithms is imprecise, incomplete or noisy most of the time, a problem that can badly mislead a learning procedure. But even if observations were perfect, the generalization beyond that data would still be afflicted with uncertainty. For example, observed data can generally be explained by more than one candidate theory, which means that one can never

E-mail address: eyke@mathematik.uni-marburg.de (E. Hüllermeier).

doi:10.1016/S0004-3702(03)00019-5

be sure of the truth of a particular theory. Consequently, inductive reasoning—by its very nature—is inseparably connected with uncertainty [13].

In fact, the insight that inductive inference can never produce ultimate truth can be traced back at least as far as Francis Bacon's epistemology. In his *Novum Organum*,[1] Bacon advocates a gradualist conception of inductive enquiry and proposes to set up *degrees of certainty*. Thus, from experience in the form of given data, one may at best conclude that a theory is *likely* to be true—not, however, that it is true with certainty. In machine learning and mathematical statistics, uncertainty of this type is generally handled by means of probabilistic methods. In Bayesian approaches, for example, an inference result is usually given in the form of a probability distribution over the space of candidate models, that is, each model (theory) is assigned a degree of probability.

In this paper, our interest concentrates on *possibility theory* [29] as an alternative calculus for modeling and processing uncertainty or, more generally, partial belief. By using possibility theory for handling uncertainty in learning procedures, inductive reasoning becomes *possibilistic* in the sense that certain generalizations are declared more or less plausible. In this paper, we shall employ possibility theory in the context of *instance-based learning* (IBL), a special approach to (supervised) machine learning. IBL relies on a kind of extrapolation principle[2] expressing a commonsense rule already suggested by David Hume:[3] "In reality, all arguments from experience are founded on the similarity, which we discover among natural objects, and by which we are induced to expect effects similar to those, which we have found to follow from such objects.... From causes, which appear *similar*, we expect similar effects. This is the sum of all our experimental conclusions." Thus, HUME suggests to extrapolate properties of one object to properties of similar ones. The idea of possibilistic induction, combined with this extrapolation principle, leads to the following inference pattern: The more similar two causes are, the more *plausible* it is that they have the same effects. Since possibility theory (in conjunction with fuzzy set theory) establishes a close connection between the concepts of similarity and uncertainty, it provides an excellent framework for translating this principle into a formal inference procedure.

This paper complements recent work on the use of possibility theory and fuzzy sets in instance-based reasoning [25–27]. The latter is more concerned with extending IBL by means of fuzzy set-based modeling techniques, whereas here the focus is on the learning process itself. More specifically, we introduce a method of possibilistic IBL, referred to as POSSIBL, which implements the above-mentioned inference pattern. Together, the two frameworks yield a powerful methodology of instance-based reasoning in which possibility theory and fuzzy set-based modeling are used, respectively, for representing gradation of uncertainty and evidential support and for complementing the data-driven inference procedure by means of domain-specific expert knowledge.

By way of background, Section 2 recalls some important ideas of possibility theory and Section 3 gives a brief review of instance-based learning and the NEAREST NEIGHBOR principle upon which it is based. Besides, the aspect of uncertainty in IBL is discussed

[1] Published in 1620.

[2] IBL does actually not realize induction proper, as will be discussed later.

[3] See, e.g., [45, p. 116].

in this section. In Section 4, a possibilistic extrapolation principle is introduced and compared to other principles commonly used in instance-based learning. Proceeding from this extrapolation principle, a method of possibilistic instance-based learning is developed in Section 5. Finally, Section 6 presents experimental studies. The paper concludes with a summary in Section 7.

2. Background on possibility theory

In this section, we recall some basic concepts from possibility theory, as far as required for the current paper. Possibility theory deals with "degrees of possibility". The term "possibility" is hence employed as a *graded* notion, much in the same way as the term "probability". At first sight, this might strike as odd since "possibility" is usually considered a two-valued concept in natural language (something is possible or not). Before turning to more technical aspects, let us therefore make some brief remarks on the semantics underlying the notion of "possibility" as used in possibility theory.

Just as the concept of probability, the notion of possibility can have different semantic meanings. To begin with, it can be used in the (physical) sense of a "degree of ease". One might say, for instance, that it is more possible for Hans to have two eggs for breakfast than eight eggs, simply because eating two eggs is more easy (feasible, practicable) than eating eight eggs [82]. However, as concerns the use in most applications, and in this paper in particular, possibility theory is considered as a means for representing uncertain knowledge, that means, for characterizing the epistemic state of an agent. For instance, given the information that Hans has eaten *many* eggs, one is clearly uncertain about the precise number. Still, three eggs appears somewhat more plausible (possible) than two eggs, since three is more compatible with the linguistic quantifier "many" than two.

It is important to note that a degree of possibility, as opposed to a degree of probability, is not necessarily a number. In fact, for many applications it is sufficient, and often even more suitable, to assume a qualitative (ordinal) scale with possibility degrees ranging from, e.g., "not at all" and "hardly" to "fairly" and "completely" [33,52]. Still, possibility degrees can also be measured on the cardinal scale $[0, 1]$, again with different semantic interpretations. For example, possibility theory can be related to probability theory, in which case a possibility degree can specify, e.g., an upper probability bound [31]. For convenience, possibility degrees are often coded by numbers from the unit interval even within the qualitative framework of possibility theory.

As a means of representing uncertain knowledge, possibility theory makes a distinction between the concepts of the *certainty* and the *plausibility* of an event. As opposed to probability theory, possibility theory does not claim that the confidence in an event is determined by the confidence in the complement of that event and, consequently, involves non-additive measures of uncertainty. Taking the existence of two quite opposite but complementary types of knowledge representation and information processing into account, two different versions of possibility theory will be outlined in the following. For a closer discussion refer to [34] and [24].

2.1. Possibility distributions as generalized constraints

A key idea of possibility theory as originally introduced by Zadeh [82] is to consider a piece of knowledge as a (generalized) constraint that excludes some "world states" (to some extent). Let Ω be a set of worlds conceivable by an agent, including the "true world" ω_0. With (incomplete) knowledge $\mathcal{K}$ about the true world one can then associate a possibility measure $\Pi_{\mathcal{K}}$ such that $\Pi_{\mathcal{K}}(A)$ measures the compatibility of $\mathcal{K}$ with the event (set of worlds) $A \subseteq \Omega$, i.e., with the proposition that $\omega_0 \in A$. Particularly, $\Pi_{\mathcal{K}}(A)$ becomes small if $\mathcal{K}$ excludes each world $\omega \in A$ and large if at least one of the worlds $\omega \in A$ is compatible with $\mathcal{K}$. More specifically, the finding that $\mathcal{A}$ is incompatible with $\mathcal{K}$ to some degree corresponds to a statement of the form $\Pi_{\mathcal{K}}(A) \leqslant p$, where p is a possibility degree taken from an underlying possibility scale P.

The basic informational principle underlying the possibilistic approach to knowledge representation and reasoning is stated as a *principle of minimal specificity*:[4] In order to avoid any unjustified conclusions, one should represent a piece of knowledge $\mathcal{K}$ by the *largest* possibility measure among those measures compatible with $\mathcal{K}$, which means that the inequality above is turned into an equality: $\Pi_{\mathcal{K}}(A) = p$. Particularly, complete ignorance should be modeled by the measure $\Pi \equiv 1$.

Knowledge $\mathcal{K}$ is usually expressed in terms of a *possibility distribution* $\pi_{\mathcal{K}}$, a mapping $\Omega \to P$ related to the associated measure $\Pi_{\mathcal{K}}$ through $\Pi_{\mathcal{K}}(A) = \sup_{\omega \in A} \pi_{\mathcal{K}}(\omega)$. Thus, $\pi_{\mathcal{K}}(\omega)$ is the degree to which world ω is compatible with $\mathcal{K}$.

Apart from the boundary conditions $\Pi_{\mathcal{K}}(\Omega) = 1$ (at least one world is fully possible) and $\Pi_{\mathcal{K}}(\emptyset) = 0$, the basic axiom underlying possibility theory after Zadeh involves the maximum-operator:

$$\Pi_{\mathcal{K}}(A \cup B) = \max\bigl\{\Pi_{\mathcal{K}}(A),\ \Pi_{\mathcal{K}}(B)\bigr\}. \tag{1}$$

In plain words, the possibility (or, more precisely, the upper possibility-bound) of the union of two events A and B is the maximum of the respective possibilities (possibility-bounds) of the individual events.

As constraints are naturally combined in a conjunctive way, the possibility measures associated with two pieces of knowledge, $\mathcal{K}_1$ and $\mathcal{K}_2$, are combined by using the minimum-operator:

$$\pi_{\mathcal{K}_1 \wedge \mathcal{K}_2}(A) = \min\bigl\{\pi_{\mathcal{K}_1}(A), \pi_{\mathcal{K}_2}(A)\bigr\}$$

for all $A \subseteq \Omega$. Note that $\pi_{\mathcal{K}_1 \wedge \mathcal{K}_2}(\Omega) < 1$ indicates that $\mathcal{K}_1$ and $\mathcal{K}_2$ are not fully compatible, i.e., that $\mathcal{K}_1 \wedge \mathcal{K}_2$ is contradictory to some extent.

The distinction between possibility and certainty of an event is reflected by the existence of a so-called *necessity measure* $\mathcal{N}_{\mathcal{K}}$ that is dual to the possibility measure $\Pi_{\mathcal{K}}$. More precisely, the relation between these two measures is given by $\mathcal{N}_{\mathcal{K}}(A) = 1 - \Pi_{\mathcal{K}}(\Omega \setminus A)$ for all $A \subseteq \Omega$:[5] An event A is necessary in so far as its complement (logical negation) is not possible.

[4] This principle plays a role quite comparable to the maximum entropy principle in probability theory.

[5] If the possibility scale P is not the unit interval $[0, 1]$, the mapping $1 - (\cdot)$ has to be replaced by an order-reversing mapping of P.

Worth mentioning is the close relationship between possibility theory and fuzzy sets. In fact, the idea of Zadeh [82] was to induce a possibility distribution from knowledge stated in the form of vague linguistic information and represented by a fuzzy set. Formally, he postulated that $\pi_{\mathcal{K}}(\omega) = \mu_F(\omega)$, where μ_F is the membership function of a fuzzy set F. To emphasize that ω plays different roles on the two sides of the equality, the latter might be written more explicitly as $\pi_{\mathcal{K}}(\omega \mid F) = \mu(F \mid \omega)$: Given the knowledge $\mathcal{K}$ that ω is an element of the fuzzy set F, the possibility that $\omega_0 = \omega$ is evaluated by the degree to which the fuzzy concept (modeled by) F is satisfied by ω. To illustrate, suppose that world states are just integer numbers. The uncertainty related to the vague statement that "ω_0 is a small integer" (ω_0 is an element of the fuzzy set F of small integers) might be translated into a possibility distribution that lets $\omega_0 = 1$ appear fully plausible ($\mu_F(1) = 1$), whereas, say, 5 is regarded as only more or less plausible ($\mu_F(5) = 1/2$) and 10 as impossible ($\mu_F(10) = 0$).

2.2. Possibility as evidential support

Possibility theory as outlined above provides the basis of a generalized approach to constraint propagation, where constraints are expressed in terms of possibility distributions (fuzzy sets) rather than ordinary sets (which correspond to the special case of $\{0, 1\}$-valued possibility measures). A constraint usually corresponds to a piece of knowledge that excludes certain alternatives as being impossible (to some extent). This "knowledge-driven" view of reasoning is complemented by a, say, "data-driven" view that leads to a different type of possibilistic calculus. According to this view, the statement that "ω is possible" is not intended to mean that ω is provisionally accepted in the sense of not being excluded by some constraining piece of information, but rather that ω is indeed supported or, say, confirmed by already observed facts (in the form of examples or data).

To distinguish the two meanings of a possibility degree, we shall denote a degree of *evidential support* or *confirmation* of ω by $\delta(\omega)$,[6] whereas $\pi(\omega)$ denotes a degree of compatibility.

To illustrate, suppose that the values a variable V can assume are a subset of $\mathcal{V} = \{1, 2, \ldots, 10\}$ and that we are interested in inferring which values are possible and which are not. In agreement with the example-based (data-oriented) view, we have $\delta(v) = 1$ as soon as the instantiation $V = v$ has indeed been observed and $\delta(v) = 0$ otherwise. The knowledge-driven approach can actually not exploit such examples, since an observation $V = v$ does not exclude the possibility that V can also assume any other value $v' \neq v$. As can be seen, the data-driven and the knowledge-driven approach are intended, respectively, for expressing *positive* and *negative* evidence. As examples do express positive evidence, they do never change the distribution $\pi \equiv 1$. This distribution would only be changed if we *knew* from some other information source, e.g., that V can only take values $v \geqslant 6$, in which case $\pi(v) = 1$ for $v \geqslant 6$ and $\pi(v) = 0$ for $v \leqslant 5$.

The distinction between modeling positive and negative evidence becomes especially clear when it comes to expressing complete ignorance. As already mentioned above, this

[6] In [75], this type of distribution is called σ-distribution.

situation is adequately captured by the possibility distribution $\pi \equiv 1$: If nothing is known, there is no reason to exclude any of the worlds ω, hence each of them remains completely possible. At the same time, complete ignorance is modeled by the distribution $\delta \equiv 0$. The latter does simply express that none of the worlds ω is actually supported by observed data.

Within the context of modeling evidential support, possibilistic reasoning accompanies a process of data accumulation. Each observed fact, ϕ, guarantees a certain degree of possibility of some world state ω, as expressed by an inequality of the form $\delta_\phi(\omega) \geqslant d$. The basic informational principle is now a principle of *maximal informativeness* that suggests adopting the smallest distribution among those compatible with the given data and, hence, to turn the above inequality into an equality. The accumulation of observations ϕ_1 and ϕ_2 is realized by deriving a distribution that is pointwise defined by

$$\delta_{\phi_1 \wedge \phi_2}(\omega) = \max\bigl\{\delta_{\phi_1}(\omega), \delta_{\phi_2}(\omega)\bigr\}.$$

As can be seen, adding new information has quite an opposite effect in connection with the two types of possibilistic reasoning: In connection with the knowledge-driven or constraint-based approach, a new constraint can only reduce possibility degrees, which means turning the current distribution π into a smaller distribution $\pi' \leqslant \pi$. In connection with the data-driven or example-based approach, new data can only increase (lower bounds to) degrees of possibility.

Closely related to the view of possibility as evidential support is a set-function that was introduced in [30], called measure of "guaranteed possibility": $\Delta(A)$ is the degree to which *all* worlds $\omega \in A$ are possible, whereas an event A is possible in the sense of the usual measure of "potential possibility", namely $\Pi(A)$ as discussed above, if at least one $\omega \in A$ is possible.[7] For the measure Δ, the characteristic property (1) becomes

$$\Delta(A \cup B) = \min\bigl\{\Delta(A), \Delta(B)\bigr\}.$$

3. Instance-based learning

In recent years, several variants of instance-based approaches to (supervised) machine learning have been devised, such as, e.g., memory-based learning [70], exemplar-based learning [64], or case-based reasoning [50]. Though emphasizing slightly different aspects, all of these approaches are founded on the concept of an *instance* or a *case* as a basis for knowledge representation and reasoning. A case (observation, example, ...) can be thought of as a single experience, such as a pattern (along with its classification) in pattern recognition or a problem (along with a solution) in case-based reasoning. To highlight the main characteristics of IBL it is useful to contrast it with *model-based* learning.[8]

Typically, IBL methods learn by simply storing (some of) the observed examples. They defer the processing of these inputs until a prediction (or some other type of query) is actually requested, a property which qualifies them as *lazy* learning methods [3].

[7] The latter semantics is clearly in line with the measure-theoretic approach underlying probability theory.

[8] Needless to say, there is no clear borderline between the two approaches. In fact, several learning techniques fall in-between (e.g., [22]) or combine concepts of both (e.g., [62]).

Predictions are then derived by combining the information provided by the stored examples in some way or other. After the query has been answered, the prediction itself and any intermediate results are discarded. As opposed to this, model-based or inductive approaches derive predictions in an indirect way: First, the observed data is used in order to induce a model, say, a decision tree or a regression function. Predictions are then obtained on the basis of this model (which can also serve other purposes such as explaining). As opposed to lazy learners, inductive methods are *eager* in the sense that they greedily compile their inputs into an intensional description (model) and then discard the inputs. In general, eager (model-based) algorithms have higher computational costs during the training phase than lazy (instance-based) methods where learning basically amounts to storing (selected) examples. On the other hand, lazy methods often have greater storage requirements, typically linear in the size of the data set, and higher computational costs when it comes to deriving a prediction.

Model-based learning is in line with parametric methods in (classical) statistics, whereas instance-based approaches to machine learning share important features with non-parametric statistics, such as, e.g., kernel smoothing techniques [74]. It deserves mentioning, however, that instance-based methods are not necessarily non-parametric [77]. Besides, the lazy learning paradigm is naturally related to what is called *transductive inference* in statistical learning theory [73]. Transductive inference is inference "from specific to specific". Thus, it stands for the problem of estimating some values of a function *directly*, given a set of empirical data. Instead of transductive inference we shall also employ the less pompous term "extrapolation" to denote this process: The known values of a function are extrapolated—in a locally restricted way—in order to estimate unknown values. This type of inference represents an alternative to the indirect (model-based) approach which estimates the complete functional relationship in a first step (induction) and evaluates this estimation at the points of interest afterwards (deduction).

3.1. Nearest Neighbor classification

The well-known NEAREST NEIGHBOR (NN) principle originated in the field of pattern recognition [16] and constitutes the core of the family of IBL algorithms. It provides a simple means to realize the aforementioned extrapolation of observed instances.

Consider the following setting that will be used throughout the paper: $\mathcal{X}$ denotes the instance space, where an instance corresponds to the description x of an object (usually in attribute-value form). $\mathcal{X}$ is endowed with a distance measure $\mathcal{D}_{\mathcal{X}}$.[9] $\mathcal{L}$ is a set of labels, and $\langle x, \lambda_x \rangle$ is called a labeled instance (or a case). In classification tasks, which are the focus of most IBL implementations, $\mathcal{L}$ is a finite (usually small) set $\{\lambda_1, \ldots, \lambda_m\}$ comprised of m classes. S denotes a sample that consists of n labeled instances $\langle x_\imath, \lambda_{x_\imath} \rangle$ $(1 \leqslant \imath \leqslant n)$. Finally, a new instance $x_0 \in \mathcal{X}$ is given, whose label λ_{x_0} is to be estimated.

In connection with the sample S, note that $\mathcal{X} \times \mathcal{L}$ corresponds to the set of *potential* observations. For each label $\lambda \in \mathcal{L}$, let $C_\lambda \subseteq \mathcal{X}$ denote the set of instances $x \in \mathcal{X}$ such

[9] $(\mathcal{X}, \mathcal{D}_{\mathcal{X}})$ is often supposed to be a metric space. From a practical point of view, it is usually enough to assume reflexivity and symmetry of $\mathcal{D}_{\mathcal{X}}$.

that $\langle x, \lambda\rangle$ can indeed be observed. C_λ is also referred to as a *concept*. For example, a bicycle belongs to the concept "two-wheelers" whereas a car does not. Formally, we can assume an underlying population $\mathcal{P}$ of entities such that each element $p \in \mathcal{P}$ is mapped to a labeled instance $\langle x(p), \lambda(p)\rangle$ in a unique way. Thus, x is an element of C_λ or, say, $\langle x, \lambda\rangle$ is an *existing* instance if there is at least one $p \in \mathcal{P}$ such that $\langle x, \lambda\rangle = \langle x(p), \lambda(p)\rangle$. Observe that the mapping $p \mapsto x(p)$ is not assumed to be injective (different elements of $\mathcal{P}$ might have the same description), which means that concepts can overlap ($C_\lambda \cap C_{\lambda'} \neq \emptyset$ for $\lambda \neq \lambda'$).

The NN principle prescribes to estimate the label of the yet unclassified point x_0 by the label of the closest sample point, i.e., the one which minimizes the distance to x_0. The k-Nearest Neighbor (kNN) approach is a slight generalization which takes the $k > 1$ nearest neighbors of a new sample point x_0 into account. That is, an estimation $\lambda_{x_0}^{\mathrm{est}}$ of λ_{x_0} is derived from the set $\mathcal{N}_k(x_0)$ of the k nearest neighbors of x_0, e.g., by means of the *majority vote* decision rule:

$$\lambda_{x_0}^{\mathrm{est}} = \arg\max_{\lambda \in \mathcal{L}} \operatorname{card}\bigl\{x \in \mathcal{N}_k(x_0) \mid \lambda_x = \lambda\bigr\}. \tag{2}$$

Not only can the NN principle be used for classification, it is also employable for realizing a (locally weighted) approximation of continuous-valued target functions. To this end, one reasonably computes the (weighted) mean of the k nearest neighbors of a new query point instead of returning the most common value.[10]

The inductive bias[11] underlying the NN principle corresponds to a *representativeness* or *closeness* assumption suggesting that similar (= closely located) instances have similar (or even the same) classification. This hypothesis, which gives rise to the similarity-guided extrapolation principle discussed in the introduction, is clearly of a heuristic nature. Still, theoretical properties of NN classification have been investigated thoroughly from a statistical perspective (e.g., [14]).[12] In fact, the origin of the NN approach can be found in work on non-parametric discriminatory analysis [38,39].

Besides, several conceptual modifications and extensions, such as distance weighting, which is discussed below, have been considered. Particularly, (editing) methods for selecting optimal training samples to be stored in the memory have been developed in order to improve classification performance [78] or to reduce computational complexity [41] or both. Other extensions aim at supporting the determination of adequate metrics and the optimal size of the neighborhood. Computational aspects have been addressed as well. For example, fast algorithms for finding nearest neighbors have been devised in order to improve computational efficiency [40,49,81].

[10] Shephard's interpolation method [67] can be considered as a special type of NN estimation.

[11] Roughly speaking, the inductive bias corresponds to the a priori assumptions on the identity of the model to be learned. Without a biased angle of view, observed data is actually meaningless and generalization beyond that data impossible [56].

[12] Needless to say, corresponding results can only be derived under certain statistical assumptions on the setting of the problem.

3.2. Uncertainty in NN classification

In statistical estimation theory, an estimated quantity is always endowed with a characterization of its reliability, usually in terms of a confidence measure and a confidence region. Alternatively, an estimation is given directly in the form of a probability distribution. As opposed to this, the NN principle in its basic form merely provides a point-estimation or, say, a decision rule, but not an estimation in a statistical sense. The neglecting of uncertainty makes this principle appear questionable in some situations [43]. To illustrate, Fig. 1 shows two classification problems. The new instance x_0 is represented by a cross, and dark and light circles correspond to instances of two different classes, respectively. In both cases, the kNN rule with $k = 5$ suggests DARK as a label for x_0. As can be seen, however, this classification is everything but reliable: In the above setting, the proportion of dark and light examples is almost balanced (apart from that, the closest points are light). This is a situation of *ambiguity*. The setting below illustrates a problem of *ignorance*: It is true that all neighbors are dark, but even the closest among them are actually quite distant.

A simple (yet drastic) step to handle this type of problem is to apply a reject option in the form of a distance or frequency threshold. That is, a classification or answer to a query is simply refused if the nearest neighbors are actually not close enough [15,36,72] or if the most frequent label among these neighbors is still not frequent enough [12,42].

A second possibility is to equal statistical methods (especially Bayesian ones) in deriving a probability distribution as an inference result. In fact, this is an obvious idea since NN techniques have originally been employed in the context of non-parametric density estimation [38,53]. Thus, a single decision can be replaced by an estimation in the form of a probability vector

$$\big(p_{x_0}(\lambda_1), \ldots, p_{x_0}(\lambda_m)\big), \tag{3}$$

where $p_{x_0}(\lambda_\imath) = \Pr(\lambda_\imath \mid x_0)$ is the probability that $\lambda_{x_0} = \lambda_\imath$, i.e., the conditional probability of the label $\lambda_\imath$ given the instance x_0. Taking the k nearest neighbors of x_0 as a point of departure, an intuitively reasonable approach is to specify the probability $p_{x_0}(\lambda_\imath)$ by the relative frequency of the label $\lambda_\imath$ among the labels of these neighbors: $p_{x_0}(\lambda_\imath) \doteq k_\imath / k$, where $k_\imath$ denotes the number of neighbors having label $\lambda_\imath$. In fact, this approach can also be justified theoretically, as will be shown in the following.

The NEAREST NEIGHBOR approach to *density estimation* (not to be confused with the one to classification) is closely related to kernel-based density estimation. An NN density estimator is a kernel estimator with variable kernel width [68]: The size of the

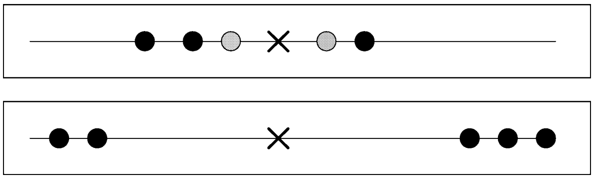


Fig. 1. Two situations of uncertainty in connection with the basic kNN rule, caused by the existence of more than one frequent class label among the nearest neighbors (top) and the absence of any close neighbor (bottom).

neighborhood of a point x_0 is adapted so as to include exactly k observations. Thus, consider a sample of n observations $x_1, \dots, x_n \in \mathbb{R}^l$ which are realizations of an l-dimensional random vector X with probability density $\phi : \mathbb{R}^l \to \mathbb{R}_{\geqslant 0}$. For $x_0 \in \mathbb{R}^l$ let v be the volume of the smallest sphere $V(x_0)$ around x_0 that contains k of these observations. The relation

$$\Pr\bigl(X \in V(x_0)\bigr) \approx \phi(x_0) \cdot v$$

(which holds true for small spheres) then suggests the following estimation of $\phi(x_0)$, the density at point x_0:

$$\phi^{\text{est}}(x_0) = \frac{k}{n \cdot v}. \tag{4}$$

Coming back to NN classification, consider a sample S that comprises $n = n_1 + \cdots + n_m$ observations, where $n_\imath$ denotes the number of tuples $\langle x, \lambda_x \rangle \in S$ such that $\lambda_x = \lambda_\imath$. Let x_0 be a new observation. Again, we choose an as small as possible hypersphere around x_0 which contains a set $\mathcal{N}_k(x_0)$ of k instances from S, where $k = k_1 + \cdots + k_m$ with $k_\imath = \text{card}\{x \in \mathcal{N}_k(x_0) \mid \lambda_x = \lambda_\imath\}$. The conditional probability density of x_0 (given the label) can now be estimated by

$$\phi^{\text{est}}(x_0 \mid \lambda_\imath) = \frac{k_\imath}{n_\imath \cdot v}, \tag{5}$$

where v denotes the volume of the hypersphere around x_0. Moreover, the unconditional density of x_0 and the prior probability of the label $\lambda_\imath$ can be estimated by

$$\phi^{\text{est}}(x_0) = \frac{k}{n \cdot v}, \qquad p^{\text{est}}(\lambda_\imath) = \frac{n_\imath}{n}, \tag{6}$$

respectively. For the probabilities in (3) one thus obtains

$$p_{x_0}(\lambda_\imath) = p^{\text{est}}(\lambda_\imath \mid x_0) = \frac{\phi^{\text{est}}(x_0 \mid \lambda_\imath) \cdot p^{\text{est}}(\lambda_\imath)}{\phi^{\text{est}}(x_0)} = \frac{k_\imath}{k}. \tag{7}$$

Remark 1. Note that the NN estimation of the conditional probability density (5) is actually given by

$$\phi^{\text{est}}(x_0 \mid \lambda_\imath) = \frac{k_\imath}{n_\imath \cdot v_\imath},$$

where $v_\imath$ is the volume of the smallest sphere around x_0 that contains all of the $k_\imath$ neighbors with label $\lambda_\imath$. Then, however, the probabilities

$$p_{x_0}(\lambda_\imath) = \frac{k_\imath \cdot v}{k \cdot v_\imath} \tag{8}$$

do not necessarily add up to 1. This problem is related to a general difficulty of NN density estimation. Namely, deriving (4) for all $x \in X$ leads to a non-normalized density function ϕ^{est} since each x requires a different hypersphere.[13]

[13] Apart from that, an NN density estimation may suffer from very heavy tails and an infinite integral.

Of course, (7) might be considered as a formal justification of the original kNN (decision) rule: The label estimated by the (majority vote) kNN rule is just the one of maximal (posterior) probability [18]. Still, one should be cautious with the distribution (7). Particularly, it is not clear how reliable the estimated probabilities $p_{x_0}(\lambda_\imath) = k_\imath / k$ actually are. It is possible to construct corresponding confidence intervals, but these are only asymptotically valid [68]. In fact, k is generally small and, hence, (7) not very reliable.[14] Improving the quality of predictions by simply increasing k obviously does not work since it also entails an enlarging of the hypersphere around x_0.[15]

3.3. Weighted NN rules

A straightforward modification of the kNN rule is to weight the influence of a neighboring sample point by its distance. This idea leads to replace (2) by

$$\lambda_{x_0}^{\text{est}} = \arg\max_{\lambda \in \mathcal{L}} \sum_{x \in \mathcal{N}_k(x_0):\ \lambda_x = \lambda} \omega(x \mid x_0, S), \tag{9}$$

where $\omega(x \mid x_0, S)$ is the weight of the neighbor x. There are different possibilities to define these weights. For example, let the neighbors $\mathcal{N}_k(x_0) = \{x_1, \ldots, x_k\}$ be arranged such that $d_\imath = \mathcal{D}_{\mathcal{X}}(x_\imath, x_0) \leqslant \mathcal{D}_{\mathcal{X}}(x_\jmath, x_0) = d_\jmath$ for $\imath \leqslant \jmath$. In [37], the weights are then determined as[16]

$$\omega(x_\imath \mid x_0, S) = \begin{cases} (d_k - d_\imath)/(d_k - d_1) & \text{if } d_k \neq d_1, \\ 1 & \text{if } d_k = d_1. \end{cases} \tag{10}$$

The weighting of neighbors appears reasonable from an intuitive point of view. For instance, a weighted kNN rule is likely to yield LIGHT rather than DARK as a classification in Fig. 1 (top). More general evidence for the usefulness of distance-weighting is provided in [54,58], at least in the practically relevant case of finite samples. In fact, in [5] it was shown that the *asymptotic performance* of the kNN rule is not improved by distance-weighting.

Note that the original kNN rule corresponds to the weighted rule with

$$\omega(x \mid x_0, S) = \begin{cases} 1 & \text{if } x \in \mathcal{N}_k(x_0), \\ 0 & \text{if } x \notin \mathcal{N}_k(x_0). \end{cases} \tag{11}$$

Thus, the NN rule can be expressed as a global principle involving the complete sample S of observations without loss of generality:

$$\lambda_{x_0}^{\text{est}} = \arg\max_{\lambda \in \mathcal{L}} \sum_{\langle x, \lambda_x \rangle \in S:\ \lambda_x = \lambda} \omega(x \mid x_0, S). \tag{12}$$

[14] An estimated probability is always a multiplicity of $1/k$. Particularly, $p_{x_0}(\lambda_\imath) \in \{0, 1\}$ in the special case $k = 1$, i.e., for the 1NN rule.

[15] Good estimations are obtained for *small* hyperspheres containing *many* points. Besides, asymptotic convergence generally assumes an adaptation of k as a function of n.

[16] See [54] for a modification that performed better in experimental studies; for other types of weight functions see, e.g., [79].

Interestingly enough, it is also possible to consider the probabilistic NN prediction (7) in the context of the weighted NN approach. Namely, (7) can be written as

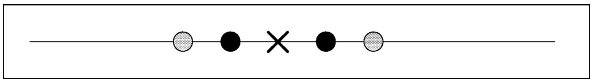
$$p_{x_0}(\lambda) = \sum_{\langle x,\lambda_x\rangle \in S:\ \lambda_x=\lambda} \omega(x \mid x_0, S), \tag{13}$$

with the weight function ω now being defined by

$$\omega(x \mid x_0, S) = \begin{cases} 1/k & \text{if } x \in \mathcal{N}_k(x_0), \\ 0 & \text{if } x \notin \mathcal{N}_k(x_0). \end{cases} \tag{14}$$

Again, (12) then amounts to choosing the label with maximal posterior probability.

Of course, in the following situation one would hardly advocate a uniform distribution suggesting that labels DARK and LIGHT have the same probability:



This example reveals a shortcoming of the weight function (14), namely the disregard of the *arrangement* of the neighbors. In fact, the derivation of the probabilistic NN estimation (7) disregards the actual distances and positions in the estimation of probability densities.[17] This, however, is only justified if the sphere containing the k nearest neighbors is indeed very small, which is usually not the case in practice. (Note that the label DARK is assigned a higher degree of probability than LIGHT according to (8), cf. Remark 1.)

In order to account for this problem, it is possible to combine the idea of weighting and probabilistic estimation. The use of the uniform weights (14) corresponds to the use of the (uniform) Parzen window in kernel-based density estimation [59]. By making use of a more general kernel function $K : \mathbb{R}^l \to \mathbb{R}_{\geqslant 0}$, a density function which is usually symmetric around 0, the NN density estimation (4) can be generalized as follows:

$$\phi^{\text{est}}(x_0) = \frac{1}{n} \cdot \sum_{\imath=1}^{n} K_{d_k}(x_0 - x_\imath), \tag{15}$$

where d_k is the distance between x_0 and its kth nearest neighbor and K_{d_k} is a re-scaling of a kernel function K (with $K(u) = 0$ for $|u| > 1$):

$$K_d : u \mapsto 1/d^l \cdot K(u/d).$$

The same reasoning as in Section 3.2 then suggests a weighted counterpart of (7):

$$p^{\text{est}}(\lambda \mid x_0) \propto \sum_{\langle x,\lambda_x\rangle \in S:\ \lambda_x=\lambda} K_{d_k}(x_0 - x). \tag{16}$$

[17] Taking positions into account becomes very tricky in instance spaces of higher dimension [86].

As can be seen, (16) is nothing else than an estimation derived from the weighted NN rule by means of normalization.[18] Thus, proceeding from weights such as (10), one simply defines a probability distribution p_{x_0} such that

$$p_{x_0}(\lambda) \propto \sum_{\langle x,\lambda_x\rangle\in S:\ \lambda_x=\lambda} \omega(x \mid x_0, S). \tag{17}$$

Related to this approach are extensions of NN classification which make use of fuzzy sets [6,8,46,47]. By weighting neighbors according to their distance, these methods compute a "fuzzy" classification

$$\lambda_{x_0}^{\mathrm{est}} = \big(u_{\lambda_1}(x_0), \ldots, u_{\lambda_m}(x_0)\big) \tag{18}$$

for a new instance x_0. That is, x_0 is not assigned a unique label in an unequivocal way. Rather, a degree of membership, $u_\lambda(x_0)$, is specified for each label λ. Consider as an example the fuzzy kNN algorithm proposed in [47]. The degree to which x_0 is assigned the label $\lambda_\imath$ (is classified into the $\imath$th class) is given by

$$u_{\lambda_\imath}(x_0) = \frac{\sum_{\jmath=1}^{k} u_{\imath\jmath}\,|x_0 - x_\jmath|^{-2/(m-1)}}{\sum_{\jmath=1}^{k} |x_0 - x_\jmath|^{-2/(m-1)}}, \tag{19}$$

where $u_{\imath\jmath} = u_{\lambda_\imath}(x_\jmath)$ is the membership degree of the instance $x_\jmath$ in the $\imath$th class. The possibility of assigning fuzzy membership degrees $u_{\imath\jmath}$ to labeled instances $x_\jmath$ is seen as a decisive feature. Turning the (non-fuzzy) label $\lambda_{x_\jmath}$ of an observed instance $x_\jmath$ into a fuzzy label allows one to adjust the influence of that instance if it is not considered prototypical of its class. The constant m in (19) determines the weighting of the distance between x_0 and its neighbors.

Clearly, (19) still has a probabilistic flavor since degrees of membership add up to 1.[19] However, the use of fuzzy labels makes it more general than (17). In fact, a fuzzy classification (18) can be written as

$$u_{\lambda_0}(x_0) \propto \sum_{\imath=1}^{n} u_{\lambda_0}(x_\imath) \cdot \omega(x_\imath \mid x_0, S).$$

Formally, the main difference between a probabilistic estimation and a fuzzy classification is hence the use of fuzzy labels in the latter approach: In the probabilistic case, an observed instance $\langle x, \lambda_x\rangle$ supports the label λ_x only. Depending on the "typicality" of the instance (it might concern a "boundary case" whose labeling was not unequivocal), it may also support labels $\lambda \neq \lambda_x$ in the case of fuzzy classification.

[18] Note, however, that (16) actually considers more than k instances if the kth nearest neighbor is not unique. See [58] for an alternative type of distance-weighting in kNN which unifies classification and density estimation.

[19] Formally, (19) might hence be interpreted as a probability distribution as well. It should be noted, however, that this interpretation might be criticized since the derivation of (19) does not assume an underlying probabilistic model.

3.4. IBL algorithms

Proceeding from the basic NN approach, a family of instance-based machine learning algorithms has been proposed in [2,4]. The simplest algorithm, known as IB1, mainly differs from the basic NN algorithm in that it normalizes the (numeric) attribute values of instances (which are characterized by means of an attribute–value representation) to guarantee that features are equally weighted, processes instances incrementally, and uses a simple method for tolerating missing attribute values. IB2 extends IB1 by using an editing strategy, i.e., it maintains a memory (case base) of selected cases called prototypes (falsely classified points are added as references). A further extension, IB3, aims at reducing the influence of noisy observations.[20] To this end, a classification record is maintained, which counts the correct and incorrect votes of the stored references. By weighting attribute values in the computation of the distance measure, IB4 and IB5 [2] take the relevance of features into account. The weights are adapted each time a new classification has been made.

To summarize, IBL algorithms (for concept learning) basically consist of three components [2]: A *similarity function* computes a numeric similarity between instances. A *classification function* decides on the membership of a newly presented instance in a concept, given the similarities between the new instance and the stored examples as well as the labels (and classification performance) of these examples. It yields a complete concept description when being applied to all (still unclassified) instances. After each classification task, a *concept description updater* derives a modified concept description by maintaining the memory of cases. The decision whether to retain or remove a case is based on records of the previous classification performance and the information provided by the new classification task.

As for the basic NN rule, some efforts have been made to improve the performance of IBL algorithms. Important points, some of which have already been mentioned above, include conceptual aspects such as the reduction of storage requirements by editing and prototype selection [55], the toleration of noise [4], the definition of similarity functions [80], and feature weighting or selection [77], as well as practical issues such as efficient techniques for indexing training examples [76]. Apart from classification, IBL techniques can also be employed for function approximation, that is to predict real-valued attributes [48,86].

4. Possibilistic extrapolation of cases

4.1. The basic estimation principle

The following type of possibilistic prediction was proposed in [23] and has been further developed in [25,27]:

$$\delta_{x_0}(\lambda_0) \doteq \max_{1 \leqslant \imath \leqslant n} \min\bigl\{\sigma_{\mathcal{X}}(x_0, x_\imath), \sigma_{\mathcal{L}}(\lambda_0, \lambda_\imath)\bigr\}, \tag{20}$$

[20] See also [78] for an early work along these lines.

for all $\lambda_0 \in \mathcal{L}$, where $\delta_{x_0}(\lambda_0)$ denotes the (estimated) *possibility* of the label λ_0, i.e. the possibility that $\lambda_{x_0} = \lambda_0$. Moreover, $\sigma_{\mathcal{X}}$ and $\sigma_{\mathcal{L}}$ are $[0, 1]$-valued similarity measures on $\mathcal{X}$ and $\mathcal{L}$, respectively.

4.1.1. The possibility distribution δ_{x_0}

According to (20), $\lambda_{x_0} = \lambda_0$ is regarded as possible if there is an instance $\langle x_\imath, \lambda_{x_\imath}\rangle$ such that both, $x_\imath$ is close to x_0 and $\lambda_{x_\imath}$ is close to λ_0. Or, if we define the *joint similarity* between the labeled instance $\langle x_\imath, \lambda_{x_\imath}\rangle$ and the (hypothetical) case $\langle x_0, \lambda_0\rangle$ to be the minimum of the similarities $\sigma_{\mathcal{X}}(x_0, x_\imath)$ and $\sigma_{\mathcal{L}}(\lambda_0, \lambda_{x_\imath})$, this can be expressed by saying that the case $\langle x_0, \lambda_0\rangle$ is regarded as possible if the existence of a similar case $\langle x_\imath, \lambda_{x_\imath}\rangle$ is confirmed by observation. In other words, a similar case provides evidence for the existence of $\langle x_0, \lambda_0\rangle$ in the sense of *possibility qualification*.[21]

Following the notational convention of Section 2, possibility degrees $\delta_{x_0}(\lambda_0)$ denote degrees of "guaranteed possibility". Thus, they are actually not considered as degrees of plausibility in the usual sense but rather as degrees of *confirmation* as introduced in Section 2.2. More specifically, the distribution $\delta_{x_0} : \mathcal{L} \to [0, 1]$ is thought of as a *lower* rather than an upper bound. Particularly, $\delta_{x_0}(\lambda_0) = 0$ must not be equated with the impossibility of $\lambda_{x_0} = \lambda_0$ but merely means that no evidence supporting the label λ_0 is available so far! In fact, δ_{x_0} is of provisional nature, and the degree of possibility assigned to a label λ_0 may increase when gathering further evidence by observing new examples, as reflected by the application of the maximum operator in (20).

This is completely in accordance with the use of possibility theory in connection with a special approach to fuzzy rule-based reasoning. Indeed, proceeding from the rule "The closer x to x_0, the more possible it is that λ_x is close to λ_{x_0}", the possibility distribution (20) has originally been derived as the inference result of a related approximate reasoning method [32]. The latter concerns an *example-based* approach to fuzzy rules where a single rule (case) is considered as a piece of data [84]. This contrasts with the *constraint-based* approach where a rule is modeled as an implication and several rules are combined conjunctively (a possibility distribution is then an *upper* bound, cf. Section 2.1).

It is natural to assume a possibility distribution $\pi : \Omega \to [0, 1]$ to be normalized (in the sense that $\sup_{\omega \in \Omega} \pi(\omega) = 1$) if $\pi(\omega)$ specifies the degree of plausibility that ω corresponds to the "true world" ω_0.[22] The above remarks make clear that this constraint does not make sense for δ_{x_0}. In this connection, it should also be noticed that there is not necessarily a unique actual world ω_0 in the sense of the possible worlds semantics [9]. Since x_0 is not assumed to have a unique label, δ_{x_0} rather provides information about the set $\{\lambda \in \mathcal{L} \mid x_0 \in C_\lambda\}$ of potential labels. Thus, the state of "complete knowledge" corresponds to the distribution δ_{x_0} with $\delta_{x_0}(\lambda) = 1$ if $x_0 \in C_\lambda$ and $\delta_{x_0}(\lambda) = 0$ otherwise.

[21] The idea of possibility qualification is usually considered in connection with natural language propositions [65,83]. Here, possibility qualification is casuistic rather than linguistic.

[22] Though generally accepted, this constraint is questioned by some authors. For example, a sub-normalized distribution might be allowed in order to express a kind of conflict.

When being applied to all $x \in \mathcal{X}$, (20) yields "fuzzy" concept descriptions, that is possibilistic approximations of the concepts C_λ ($\lambda \in \mathcal{L}$):

$$C_\lambda^{\mathrm{est}} = \left\{ \left(x, \delta_x(\lambda)\right) \mid x \in \mathcal{X} \right\}, \tag{21}$$

where $\delta_x(\lambda)$ is the degree of membership of $x \in \mathcal{X}$ in the fuzzy concept C_λ^{est}. Note that these fuzzy concepts can overlap in the sense that some x has a positive degree of membership in two concepts C_λ^{est} and $C_{\lambda'}^{\mathrm{est}}$, $\lambda \neq \lambda'$.[23]

4.1.2. The similarity measures $\sigma_{\mathcal{X}}$ and $\sigma_{\mathcal{L}}$

Let us make some remarks on the similarity measures $\sigma_{\mathcal{X}}$ and $\sigma_{\mathcal{L}}$. To begin with, notice that—according to (20)—the *similarity* of cases is in direct correspondence with the *possibility* assigned to a label. Roughly speaking, the principle expressed by (the fuzzy rule underlying) equation (20) gives rise to turn similarity into possibilistic support. Consequently, $\sigma_{\mathcal{X}}$ and $\sigma_{\mathcal{L}}$ are thought of as, say, support measures rather than similarity measures in the usual sense. They do actually serve the same purpose as the weight functions in Section 3.3. Particularly, $\sigma_{\mathcal{X}}(x_0, x_\imath) = 0$ means that the label $\lambda_{x_\imath}$ is not considered as a relevant piece of information since $x_\imath$ is not sufficiently similar to x_0. For computation, irrelevant cases in (20) can clearly be left out of account. Thus, it is enough to consider cases in a certain region around x_0. As opposed to the kNN approach, it is the size of this region rather than the number of neighboring cases which is fixed.

We assume $\sigma_{\mathcal{X}}$ and $\sigma_{\mathcal{L}}$ to be reflexive and symmetric, whereas no special kind of transitivity is required. In fact, the application of the maximum operator in (20) does even permit a purely *ordinal* approach. In this case, the range of the similarity measures is a finite subset $\mathcal{A} \subset [0,1]$ that encodes an ordinal scale such as

$$\{\text{completely different}, \ldots, \text{ very similar}, \text{identical}\}. \tag{22}$$

Correspondingly, degrees of possibility are interpreted in a qualitative way [33,52]. That is, $\delta_{x_0}(\lambda) < \delta_{x_0}(\lambda')$ only means that label λ is less supported than label λ'; apart from that, the difference between these values has no meaning.

Needless to say, a scale such as (22) is more convenient if instances are complex objects rather than points in a Euclidean space and if similarity (distance) between objects must be assessed by human experts (which is common practice in case-based reasoning). Note that an ordinal structure is also sufficient for the original kNN rule. In connection with distance-weighting, however, the structures of the involved measures become more important. In any case, one should be aware of the fact that a cardinal interpretation of similarity raises some crucial semantic questions if corresponding measures cannot be defined in a straightforward way. In the weighted kNN rule, for example, one patient that died from a certain medical treatment compensates for two patients that survived if the former is twice as similar to the current patient. But what exactly does "twice as similar" mean in this context?

Looking at (20) from the point of view of observed cases, this estimation principle defines a (possibilistic) *extrapolation* of each sample $\langle x, \lambda_x \rangle$. In the original NN approach,

[23] In practice, fuzzy and/or overlapping concepts seem to be the rule rather than the exception [1].

which does not involve a distance measure $\mathcal{D}_{\mathcal{L}}$ on $\mathcal{L}$, a case $\langle x_\imath, \lambda_{x_\imath} \rangle \in S$ can only support the label $\lambda_{x_\imath}$. This corresponds to the special case where $\sigma_{\mathcal{L}}$ in (20) is given by

$$\sigma_{\mathcal{L}}(\lambda, \lambda') = \begin{cases} 1 & \text{if } \lambda = \lambda', \\ 0 & \text{if } \lambda \neq \lambda', \end{cases} \tag{23}$$

which is reasonable if $\mathcal{L}$ is a nominal scale, as, e.g., in concept learning or pattern recognition (classification with $|\mathcal{L}| = 2$).

By allowing for graded distances between labels, the possibilistic approach provides for a case $\langle x_\imath, \lambda_{x_\imath} \rangle$ to support similar labels as well. This type of extended extrapolation is reasonable if $\mathcal{L}$ is a cardinal or at least ordinal scale. In fact, it should be observed that (20) applies to continuous scales in the same way as to discrete scales and thus unifies the performance tasks of classification and function approximation. For example, knowing that the price (= label) of a certain car is \$ 10,500, it is quite plausible that a similar car has exactly the same price, but it is plausible as well that it costs \$10,700. Interestingly enough, the same principle is employed in kernel-based estimation of probability density functions, where probabilistic support is allocated by kernel functions centered around observations [59,63]. Indeed, (20) can be considered as a possibilistic counterpart of kernel-based density estimation. Let us finally mention that the consideration of graded distances between labels is also related to the idea of class-dependent misclassification costs [60,71].

4.2. Generalized possibilistic estimation

The possibility distribution δ_{x_0}, which specifies the fuzzy set of well-supported labels, is a disjunctive combination of the individual support functions

$$\delta_{x_0}^{\imath} : \lambda_0 \mapsto \min\bigl\{\sigma_{\mathcal{X}}(x_0, x_\imath),\ \sigma_{\mathcal{L}}(\lambda_0, \lambda_{x_\imath})\bigr\}. \tag{24}$$

In fact, the max-operator in (20) is a so-called t(riangular)-conorm and serves as a generalized logical or-operator: $\lambda_{x_0} = \lambda_0$ is regarded as possible if $\langle x_0, \lambda_0 \rangle$ is similar to $\langle x_1, \lambda_{x_1} \rangle$ or to $\langle x_2, \lambda_{x_2} \rangle$ or ... or to $\langle x_n, \lambda_{x_n} \rangle$.

Now, fuzzy set theory offers t-conorms other than max and, hence, (20) can be generalized as follows:

$$\begin{aligned} \delta_{x_0}(\lambda_0) &\doteq \delta_{x_0}^1(\lambda_0) \oplus \delta_{x_0}^2(\lambda_0) \oplus \cdots \oplus \delta_{x_0}^n(\lambda_0) \\ &= \bigoplus_{1 \leqslant \imath \leqslant n} \min\bigl\{\sigma_{\mathcal{X}}(x_0, x_\imath),\ \sigma_{\mathcal{L}}(\lambda_0, \lambda_{x_\imath})\bigr\} \\ &= 1 - \bigotimes_{1 \leqslant \imath \leqslant n} \max\bigl\{1 - \sigma_{\mathcal{X}}(x_0, x_\imath),\ 1 - \sigma_{\mathcal{L}}(\lambda_0, \lambda_{x_\imath})\bigr\} \end{aligned}$$

for all $\lambda_0 \in \mathcal{L}$, where $\otimes$ and $\oplus$ are a t-norm and a related t-conorm, respectively. Recall that a t-norm is a binary operator $\otimes : [0,1]^2 \to [0,1]$ which is commutative, associative, monotone increasing in both arguments and which satisfies the boundary conditions $x \otimes 0 = 0$ and $x \otimes 1 = x$. An associated t-conorm is defined by the mapping $(\alpha, \beta) \mapsto 1 - (1 - \alpha) \otimes (1 - \beta)$. The t-norm associated with the t-conorm max is the min-operator. Other important operators are the product $\otimes_P : (\alpha, \beta) \mapsto \alpha\beta$ with related t-conorm $\oplus_P : (\alpha, \beta) \mapsto$

$\alpha + \beta - \alpha\beta$ and the Lukasiewicz t-norm $\otimes_L : (\alpha, \beta) \mapsto \max\{0, \alpha + \beta - 1\}$ with related t-conorm $\oplus_L : (\alpha, \beta) \mapsto \min\{1, \alpha + \beta\}$.

Observe that the minimum operator employed in the determination of the joint similarity between cases can be considered as a logical operator as well, namely as a fuzzy conjunction: Two cases $\langle x_0, \lambda_{x_0}\rangle$ and $\langle x_1, \lambda_{x_1}\rangle$ are similar if both, x_0 is similar to x_1 *and* λ_{x_0} is similar to λ_{x_1}. Consequently, this operator might be replaced by a t-norm, too. By doing so, (24) and (20) become

$$\delta^{\imath}_{x_0} : \lambda_0 \mapsto \sigma_{\mathcal{X}}(x_0, x_\imath) \otimes \sigma_{\mathcal{L}}(\lambda_0, \lambda_{x_\imath}) \tag{25}$$

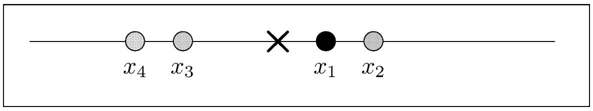
and

$$\delta_{x_0}(\lambda_0) \doteq \bigoplus_{1 \leqslant \imath \leqslant n} \sigma_{\mathcal{X}}(x_0, x_\imath) \otimes \sigma_{\mathcal{L}}(\lambda_0, \lambda_{x_\imath}), \tag{26}$$

respectively. Note, however, that a (fuzzy) logic-based derivation of the joint similarity is not compulsory. Particularly, the t-norm $\otimes$ in (26) need not necessarily be the one related to the t-conorm $\oplus$. For example, one might thoroughly take $\otimes = \min$ and $\oplus = \oplus_P$, or even combine the similarity degrees $\sigma_{\mathcal{X}}(x_0, x_\imath)$ and $\sigma_{\mathcal{L}}(\lambda_0, \lambda_{x_\imath})$ by means of an operator which is not a t-norm. In that case, however, the "logical" interpretation of (26) is lost.

4.2.1. Control of compensation and accumulation of support

By choosing an appropriate t-conorm $\oplus$ in (26) one can control the accumulation of individual degrees of evidential support, especially the extent of compensation. To illustrate, consider the following situation, where $\sigma_{\mathcal{X}}(x_0, x_1) = 3/4$, $\sigma_{\mathcal{X}}(x_0, x_2) = \sigma_{\mathcal{X}}(x_0, x_3) = 1/2$, and $\sigma_{\mathcal{X}}(x_0, x_4) = 1/4$:



Should one prefer DARK or LIGHT as a classification of the new point? The use of the max-operator as a t-conorm yields $\delta_{x_0}(\text{DARK}) = 3/4$ and $\delta_{x_0}(\text{LIGHT}) = 1/2$ and, hence, the decision DARK. The three moderately similar instances with label LIGHT do not compensate for the one very similar instance with label DARK. As opposed to this, the probabilistic sum $(\alpha, \beta) \mapsto \alpha + \beta - \alpha\beta$ brings about a compensation effect and entails $\delta_{x_0}(\text{DARK}) = 3/4$ and $\delta_{x_0}(\text{LIGHT}) = 13/16$, that is, a slightly larger possibility for LIGHT.

More generally, different t-conorms can model different accumulation modes, which typically entail a kind of saturation effect. In the case of the probabilistic sum $\oplus_P$, for example, an additional β-similar observation increases the current support α by $\beta(1 - \alpha)$. Thus, the larger the support already granted is, the smaller the absolute increase due to the new observation will be. This appears reasonable from an intuitive point of view: If the support of a label is already large, one is not surprised to see another (close) instance having the same label. A small support increment then reflects the low information content related to the new observation [44].

4.2.2. Possibilistic support and weighted NN estimation

A t-norm $\otimes$ is called Archimedian if the following holds: For all $x, y \in]0,1[$ there is a number $n \in \mathbb{N}$ such that $\otimes^{(n)}(x) < y$ (where $\otimes^{(n)}(x) = \otimes^{(n-1)}(x) \otimes x$ and $\otimes^{(1)}(x) = x$). It can be shown that $\otimes$ is a continuous Archimedian t-norm iff there is a continuous, strictly decreasing function $g : [0,1] \to [0,\infty]$ such that $g(1) = 0$ and

$$\alpha \otimes \beta = g^{(-1)}\bigl(g(\alpha) + g(\beta)\bigr) \tag{27}$$

for all $0 \leqslant \alpha, \beta \leqslant 1$, where the pseudo-inverse $g^{(-1)}$ is defined as

$$g^{(-1)} : x \mapsto \begin{cases} g^{-1}(x) & \text{if } 0 \leqslant x \leqslant g(0), \\ 0 & \text{if } g(0) < x. \end{cases}$$

The function g is called the *additive generator* of $\otimes$. For example, $x \mapsto 1 - x$ and $x \mapsto -\ln(x)$ are additive generators of the Lukasiewicz t-norm $\otimes_L$ and the product $\otimes_P$, respectively.

Based on the representation (27), one can establish an interesting connection between (26) and the weighted NN rule. To this end, let g be the additive generator of the t-norm[24] related to the t-conorm $\oplus$ used as an aggregation operator in (26). With $d_\imath = 1 - \sigma_{\mathcal{X}}(x_0, x_\imath) \otimes \sigma_{\mathcal{L}}(\lambda_0, \lambda_{x_\imath})$ and $\omega_\imath = g(d_\imath)$, we can write (26) as

$$\delta_{x_0}(\lambda_0) = 1 - g^{(-1)}(\omega_1 + \omega_2 + \cdots + \omega_n). \tag{28}$$

Since g is decreasing, it can be considered as a weight function that turns a distance $d_\imath$ into a weight $\omega_\imath$ associated with the $\imath$th instance. Then, (28) tells us that the possibility degree $\delta_{x_0}(\lambda_0)$ is nothing else than a (monotone increasing) transformation of the sum of weights $\omega_\imath$. In other words, (26) can be seen as a distance-weighted NN estimation, where the weight of a neighbor is determined as a function of its similarity to the new instance. As opposed to (9), however, the weight of a case according to (28) does not depend on other cases stored in memory (cf. Section 4.3.1 below).

Consider the Lukasiewicz t-(co)norm as an example, for which we obtain $\omega_\imath = 1 - d_\imath = \sigma_{\mathcal{X}}(x_0, x_\imath) \otimes \sigma_{\mathcal{L}}(\lambda_0, \lambda_{x_\imath})$ and

$$\delta_{x_0}(\lambda_0) = \min\{1, \omega_1 + \omega_2 + \cdots + \omega_n\}. \tag{29}$$

If, moreover, $\sigma_{\mathcal{L}}$ is given by (23), then $\delta_{x_0}(\lambda_0)$ is nothing else than the bounded sum of the similarity degrees $\sigma_{\mathcal{X}}(x_\imath, x_0)$ between x_0 and the instances $x_\imath$ with label $\lambda_{x_\imath} = \lambda_0$. Thus, (29) is basically equivalent to the global NN method, i.e. the weighted NN approach with $k = n$,[25] apart from the fact that it does not distinguish between labels whose accumulated support exceeds 1 (this is another type of saturation effect). For the probabilistic sum $\oplus_P$, the mapping between possibility degrees and the sum of weights is one-to-one:

$$\delta_{x_0}(\lambda_0) = 1 - \exp\bigl(-(\omega_1 + \omega_2 + \cdots + \omega_n)\bigr).$$

In connection with the generalized model (26), the t-conorm $\oplus$ used for combining individual degrees of support defines another degree of freedom of the model. It is

[24] This is not the t-norm used in (26) for defining a joint similarity measure.

[25] The proper kNN rule cannot be emulated as in (11) since the weights $\omega_\imath$ depend on absolute distance (again, see Section 4.3.1 below).

hence interesting to mention the existence of parameterized families of t-(co)norms which comprise commonly used operators as special cases. For example, the Frank-family is defined as

$$\oplus_\rho : (\alpha, \beta) \mapsto \begin{cases} \max(\alpha, \beta) & \text{if } \rho = 0, \\ \alpha + \beta - \alpha\beta & \text{if } \rho = 1, \\ \min\{1, \alpha + \beta\} & \text{if } \rho = \infty, \\ 1 - \ln_\rho\bigl(1 + \frac{(\rho^{1-\alpha}-1)(\rho^{1-\beta}-1)}{\rho-1}\bigr) & \text{otherwise.} \end{cases} \tag{30}$$

Proceeding from such a family of t-conorms, the degree of freedom of the model reduces to a single parameter, here ρ, which can be adapted in a simple way, e.g., by means of cross-validation techniques.

4.2.3. Upper and lower possibility bounds

The possibility degree (26) represents the support (confirmation) of a label λ_0 gathered from similar instances, according to the basic NN principle suggesting that similar instances have similar labels. Now, in the sense of this principle, an observation $\langle x_\imath, \lambda_{x_\imath} \rangle$ might not only confirm but also *disqualify* a label λ_0. This happens if $x_\imath$ is close to x_0 but $\lambda_{x_\imath}$ is not similar to λ_0. A possibility distribution expressing degrees of *exclusion* rather than degrees of support and, hence, complementing (26) in a natural way is given by

$$\pi_{x_0} : \lambda_0 \mapsto \bigotimes_{1 \leqslant \imath \leqslant n} \bigl(1 - \sigma_{\mathcal{X}}(x_0, x_\imath)\bigr) \oplus \sigma_{\mathcal{L}}(\lambda_0, \lambda_{x_\imath}). \tag{31}$$

According to (31), an individual observation $\langle x_\imath, \lambda_{x_\imath} \rangle$ induces a constraint on the label of x_0: A label λ_0 is disqualified by $\langle x_\imath, \lambda_{x_\imath} \rangle$ if both, $\sigma_{\mathcal{X}}(x_0, x_\imath)$ is large and $\sigma_{\mathcal{L}}(\lambda_0, \lambda_{x_\imath})$ is small. As opposed to this, $\langle x_\imath, \lambda_{x_\imath} \rangle$ is completely ignored if $\sigma_{\mathcal{X}}(x_0, x_\imath) = 0$, in which case the individual support on the right-hand side of (31) is 1 ($\pi_{x_0} \equiv 1$ is an expression of complete ignorance: all upper possibility bounds are 1 since there is no reason to discredit any label). This approach is obviously in agreement with the constraint-based view of possibilistic reasoning (cf. Section 2.1). Moreover, the distribution (31) is again related to a special type of fuzzy rule [26].

The possibility of a label λ_0 can now be characterized by means of an extended estimation, namely as a tuple

$$\delta^*_{x_0}(\lambda_0) = \bigl[\delta_{x_0}(\lambda_0), \, \pi_{x_0}(\lambda_0)\bigr]$$

with a lower bound $\delta_{x_0}(\lambda_0)$ expressing a degree of confirmation, and an upper bound $\pi_{x_0}(\lambda_0)$ expressing a degree of plausibility. The following cases show that the complementary distribution π_{x_0} can greatly improve the informational content of a possibilistic evaluation:[26]

- $\delta^*_{x_0}(\lambda_0) = [0, 1]$: This is an expression of complete ignorance. Neither is λ_0 supported nor is it (partly) excluded by any observation. Thus, λ_0 is fully plausible though not confirmed at all.

[26] Recall that positive and negative evidence cannot be distinguished in probability theory.

- $\delta^*_{x_0}(\lambda_0) = [0, 0]$: Clear evidence against λ_0 has been accumulated in the form of instances similar to x_0 with labels dissimilar to λ_0.
- $\delta^*_{x_0}(\lambda_0) \approx [1, 1]$: The label λ_0 is strongly supported through the observation of similar instances.

Notice that

$$\delta_{x_0}(\lambda_0) > \pi_{x_0}(\lambda_0) \tag{32}$$

indicates a kind of conflict and is closely related to the problem of ambiguity in connection with the NN principle (cf. Section 3.2). In fact, (32) can occur if x_0 has close neighbors $x_\imath$ and $x_\jmath$ with quite dissimilar labels $\lambda_{x_\imath}$ and $\lambda_{x_\jmath}$ (mathematically speaking, x_0 is a point of discontinuity). In this case, the evaluation of λ_0 is unsteady, and the support $\delta_{x_0}(\lambda_0)$ should be taken with caution. The inequality in (32) might also trigger a revision process that aims at removing the conflict by means of a model adaptation.

4.2.4. Fuzzy logical evaluation

The values $\delta_{x_0}(\lambda_0)$ in (26) can also be considered as membership degrees of a fuzzy set, namely the fuzzy set of "well-supported labels". In fact, the possibility degree $\delta_{x_0}(\lambda_0)$ can be seen as the truth degree, $\langle P(\lambda_0)\rangle$, of the following (fuzzy) predicate $P(\lambda_0)$: "There is an instance close to x_0 with a label similar to λ_0." $P(\lambda_0)$ defines the property that qualifies λ_0 as a well-supported label.

Of course, one might easily think of alternative characterizations of well-supported labels. Fuzzy set-based modeling techniques allow for translating such characterizations given in linguistic form into logical expressions. By using fuzzy logical connectives including t-norms, fuzzy quantifiers such as "a few" and fuzzy relations such as "closely located", one can specify sophisticated fuzzy decision principles that go beyond the simple NN rule. Example:

> "There are at least a few closely located instances, most of these instances have the same label, and none of the moderately close instances has a very different label."

The logical expression $P(\cdot)$ associated with such a specification can be used in place of the right-hand side in (26):

$$\delta_{x_0}(\lambda_0) \doteq \langle P(\lambda_0)\rangle. \tag{33}$$

The decision rule related to (26) favors the label $\lambda^{\mathrm{est}}_{x_0}$ that meets the requirements specified by $P(\cdot)$ best. This generalization appears especially interesting since it allows one to adapt the NN principle so as to take specific characteristics of the application into account.

Observe that (33) can also mimic the original kNN rule: Consider the fuzzy proposition "λ_0 is supported by many of the k nearest neighbors of x_0", and let the fuzzy quantifier "many (out of k)" be modeled by the mapping $\imath \mapsto \imath/k$. Then, $\delta_{x_0}(\lambda_0) = \imath/k$ iff $\imath$ among the k nearest neighbors have label λ_0. In this case, possibility degrees (derived from fuzzy truth degrees) formally coincide with probability degrees.

4.3. Comparison of extrapolation principles

So far, we have discussed two types of NN approaches to estimation and decision making: A probabilistic one, which is in agreement with the original kNN rule, and a possibilistic one introduced in this section. Both approaches can be considered as a two-step procedure. The first step derives a distribution that will subsequently be referred to as the NN *estimation*. This estimation defines a degree of support for each label $\lambda \in \mathcal{L}$. The second step, the NN *decision*, chooses one label on the basis of the NN estimation. Usually, the decision is given by the label with maximal support, and ties are broken by coin flipping. Still, in the case of a continuous (or at least ordinal) scale $\mathcal{L}$, a decision might also be obtained by some kind of averaging procedure.

In order to facilitate the comparison of the two approaches, we write degrees of evidential support in the general form

$$\nu(\lambda \mid x_0, S) = \alpha\big(\{\nu_x(\lambda \mid x_0, S) \mid \langle x, \lambda_x \rangle \in S\}\big) \tag{34}$$

and thus obtain the (maximal support) decision as

$$\lambda_{x_0}^{\text{est}} = \arg\max_{\lambda \in \mathcal{L}} \nu(\lambda \mid x_0, S). \tag{35}$$

In (34), $\nu_x(\lambda \mid x_0, S)$ is the support of the hypothesis $\lambda_{x_0} = \lambda$ provided by the labeled instance $\langle x, \lambda_x \rangle$, and α is an aggregation function.

To reveal the original kNN rule and the probabilistic approach as special cases of (35), note that the probability distribution (7) is obtained by using the arithmetic sum as an aggregation function α and defining the support function as

$$\nu_x^p(\lambda \mid x_0, S) = \begin{cases} 1/k & \text{if } x \in \mathcal{N}_k(x_0) \text{ and } \lambda = \lambda_x, \\ 0 & \text{otherwise.} \end{cases} \tag{36}$$

More generally, a support function can be defined as

$$\nu_x^p(\lambda \mid x_0, S) = \begin{cases} K_{d_k}(x_0 - x) & \text{if } \lambda = \lambda_x, \\ 0 & \text{otherwise,} \end{cases} \tag{37}$$

where K is a kernel function. The index d_k denotes the distance between x_0 and its kth nearest neighbor. It signifies that the kernel function is *scaled* so as to exclude exactly those instances $x_\imath$ with $\mathcal{D}_{\mathcal{X}}(x_0, x_\imath) > d_k$. Proceeding from (37), the probability distribution p_{x_0} is obtained by normalizing the supports

$$\nu^p(\lambda \mid x_0, S) = \sum_{\langle x, \lambda_x \rangle \in S} \nu_x^p(\lambda \mid x_0, S),$$

which yields

$$p_{x_0}(\lambda) = \frac{\nu^p(\lambda \mid x_0, S)}{\sum_{\jmath=1}^{m} \nu^p(\lambda_\jmath \mid x_0, S)} \tag{38}$$

for all $\lambda \in \mathcal{L}$. That is, the aggregation α is now the normalized rather than the simple arithmetic sum. Of course, since normalization does not change the mode of a distribution it has no effect on decision making and could hence be omitted from this point of view.

The possibilistic approach (26) is recovered by $\alpha = \oplus$ and

$$\nu_x^\delta(\lambda \mid x_0, S) = \sigma_{\mathcal{X}}(x_0, x) \otimes \sigma_{\mathcal{L}}(\lambda, \lambda_x). \tag{39}$$

As can be seen, the main difference between the probabilistic and the possibilistic approach concerns the definition of the individual support function ν_x and the aggregation of the corresponding degrees of support.

Apart from that, however, a direct comparison is complicated by the similarity measure over labels, $\sigma_{\mathcal{L}}$, which is used in (39) but not in (37). One possibility to handle this problem is to consider (39) only for the special case (23):

$$\nu_x^\delta(\lambda \mid x_0, S) = \begin{cases} \sigma_{\mathcal{X}}(x_0, x) & \text{if } \lambda = \lambda_x, \\ 0 & \text{otherwise.} \end{cases} \tag{40}$$

Eq. (40) reveals that the similarity measure $\sigma_{\mathcal{X}}$ now plays the same role as the kernel function K in (37).

4.3.1. Absolute versus relative support

An important difference between (37) and (40) is that an example $\langle x, \lambda_x \rangle \in S$ provides *relative* support of a label λ in the probabilistic approach but *absolute* support in the possibilistic one. That is, $\nu_x^\delta(\lambda \mid x_0, S)$ depends on the absolute similarity between x_0 and x but is independent of further observations. In fact, we can actually write $\nu_x^\delta(\lambda \mid x_0)$ in place of $\nu_x^\delta(\lambda \mid x_0, S)$ since S does not appear on the right-hand side of (40): The support provided by observed samples $\langle x, \lambda_x \rangle$ is bounded to nearby instances, decreases gradually with distance, and vanishes for completely dissimilar examples.

As opposed to this, the support $\nu_x^p(\lambda \mid x_0, S)$ is relative and depends on the relation between the distance of x to x_0 and the distances of other observations to x_0. This is reflected by the scaling of the kernel function in (37). On the one hand, this means that $\nu_x^p(\lambda \mid x_0, S)$ can be large even though x is quite distant from x_0. On the other hand, the extension of the sample S by another instance close enough to x_0 might exclude a quite similar observation x from the neighborhood $\mathcal{N}_k(x_0)$. The corresponding re-scaling of the kernel function will then cancel the support provided by $\langle x, \lambda_x \rangle$ so far. The induced thresholding effect appears especially radical (and might be questioned on such grounds) in connection with (36), where $\nu_x^p(\lambda \mid x_0, S)$ is reduced from $1/k$ to 0, that is from full support to no support at all.

The bounding of evidential support, as realized by the possibilistic approach, is often advisable. Consider a simple example: Let $\mathcal{X} = [0, 1]$ and $\lambda_x = \mathbb{I}_{[1/2,1]}(x)$[27] and suppose instances to be chosen at random according to a uniform distribution. Moreover, assume that a new instance x_0 must be labeled, given only one observation, x_1. Using the 1NN rule, the probability of a correct decision is obviously 1/2. Now, suppose that the NN rule is applied only if $|x_0 - x_1| \leqslant d$, whereas a decision is determined by flipping a coin otherwise (this is exactly the procedure that results from the possibilistic approach by defining $\sigma_{\mathcal{X}}$ in (20) by $\sigma_{\mathcal{X}}(x, x') = 1$ if $|x - x'| \leqslant d$ and 0 otherwise). A simple calculation shows that the probability of a correct decision is now $1/2 + d(1 - d)$. As can be seen, dissimilar instances

[27] $\mathbb{I}_A$ is the indicator function: $\mathbb{I}_A(x) = 1$ if $x \in A$ and 0 otherwise.

are likely to provide misleading information in this example and, hence, the disregard of such instances is indeed advantageous. Loosely speaking, it is better to guess a label at random than to rely on observations not similar enough.

Of course, the concept of absolute support is actually not reserved to the possibilistic approach but can be realized for the probabilistic method as well. To this end, one simply replaces (37) by

$$\nu_x^p(\lambda \mid x_0, S) = \begin{cases} K(x_0 - x) & \text{if } \lambda = \lambda_x, \\ 0 & \text{otherwise,} \end{cases} \tag{41}$$

where the kernel function K is now fixed. That is, K is no longer scaled by the size of the neighborhood of x_0. This is exactly the estimation one derives by the reasoning in Section 3.2 if the generalized NN density estimation (15) is replaced by the simple kernel estimator:

$$\phi^{\text{est}}(x_0) = \frac{1}{n} \cdot \sum_{\imath=1}^{n} K(x_0 - x_\imath). \tag{42}$$

Here, the only problem occurs if $\nu^p(\lambda \mid x_0, S) = 0$ for all $\lambda \in \mathcal{L}$. In this situation (of complete ignorance), a probability distribution cannot be derived by normalization.

Apart from that, (41) might indeed be preferred to (37) due to the reasons mentioned above. In fact, one should realize that one of the major reasons for using the NN density estimator (15) rather than the kernel estimator (42) is to guarantee the continuity of the density function ϕ^{est}. In the context of instance-based learning, however, this is not important since one is not interested in estimating a complete density function but only a single value thereof. To the best of our knowledge, (37) and (41) have not been compared in a systematic way in IBL so far. Note that (41) should actually be called a NEAR NEIGHBOR estimation since it involves the *near* rather than the *nearest* neighbors. The same remark applies to the possibilistic approach, of course.

Above, it has been argued that the consideration of graded degrees of similarity between labels is often advised (see also our example in Section 4.5 below). It should be mentioned, therefore, that the probabilistic approach might be extended in this direction as well. To this end, a *joint* probability density can be estimated based on a kernel function K, which is now defined over $\mathcal{X} \times \mathcal{L}$. An estimation for the label λ can then be derived by conditioning on x_0:

$$p_{x_0}(\lambda) \propto \sum_{\langle x, \lambda_x \rangle \in S} \nu_x^p(\lambda \mid x_0, S) = \sum_{\langle x, \lambda_x \rangle \in S} K(x_0 - x, \lambda - \lambda_x).$$

This is the most general form of a probabilistic estimation. Still, one should keep in mind that it requires $\mathcal{X} \times \mathcal{L}$ to have a certain mathematical structure, an assumption which is not always satisfied in applications (again, we refer to our example below).

Let us conclude this section with a final remark on related work in a different context. Interestingly enough, a distinction similar to ours between absolute and relative support has also been made in connection with cluster analysis. In *fuzzy* cluster analysis, a point may have a positive degree of membership in several classes. Still, in the classical approach [7] the membership degrees add up to 1 and must hence be interpreted as *relative* numbers. In [51], some difficulties caused by this constraint are discussed, and *possibilistic* clustering is advocated as an alternative. In this approach, a membership degree does indeed reflect the (absolute) compatibility of a point with the prototype of a cluster.

4.3.2. Similarity versus frequency

The estimation principle underlying the probabilistic approach combines the concepts of similarity (distance) and frequency: It applies a closeness assumption, typical of similarity-based reasoning, that suggests to focus on the most similar observations (or to weight observations by their distance). From the reduced set of supposedly most relevant instances, probabilities are then estimated by relative frequencies. This contrasts with the basic (max–min) possibilistic approach (20) which relies on similarity alone: The application of the maximum operator does not produce any compensation or reinforcement effect. Thus, possibility depicts the *existence* of supporting evidence, not its frequency.[28] The generalized possibilistic approach based on (26) allows for modes of compensation which combine both aspects. Especially, the operators mentioned above produce a kind of saturation effect, that is, a limited reinforcement effect: The increase of support due to the observation of a similar instance is a decreasing function of the support that is already available.

In this connection, it is important to realize the different nature of the concepts of possibility and probability. Particularly, it should be emphasized that the former is not interpreted in terms of the latter.[29] For example, consider the standard probabilistic setting where cases are chosen randomly and independently according to a fixed probability measure over $\mathcal{X} \times \mathcal{L}$. The possibility degree $\delta_{x_0}(\lambda_0)$ will then converge to 1 with increasing sample size whenever $\langle x_0, \lambda_0 \rangle$ has a non-zero probability of occurrence. In fact, the possibilistic approach is interested in the *existence* of a case, not in its probability. Roughly speaking, the major concern of this approach is the approximation of the concepts C_λ, whereas the probabilistic approach aims at estimating conditional probability distributions $p_{x_0} = \Pr(\cdot \mid x_0)$. Of course, this distinction is relevant only if the concepts are overlapping, that is, if the query x_0 does not have a unique label. Otherwise, a possibilistic and a probabilistic approach are equivalent in the sense that $x_0 \in C_\lambda \Leftrightarrow \Pr(\lambda \mid x_0) = 1$.

It is beyond question that the frequency of observations usually provides valuable information. Yet, the frequency-based approach does heavily rely on statistical assumptions concerning the generation of training (and test) data. Thus, it might be misleading if these assumptions are violated. Suppose, e.g., that the probability of observing a positive example, while learning a concept $C_1 \subseteq \mathcal{X}$, depends on the number of positive examples observed so far and hence contradicts an independence assumption (the probability of a label λ_x, given the instance x, is not independent of the data). In this case, a probabilistic estimation is clearly biased, whereas the possibility distribution (20) is not affected at all. Indeed, the information expressed by δ_{x_0} remains valid even if only negative examples $x_\imath \in C_0 = \mathcal{X} \setminus C_1$ have been presented so far: $\delta_{x_0}(1) = 0$ then simply means that no evidence for $x_0 \in C_1$ has been gathered as yet. Moreover, the value $\delta_{x_0}(0)$ reflects the

[28] To a certain extent, this is related to the distinction between an *existential* and an *enumerative* analogy factor in models of analogical induction [57].

[29] Though such a relationship can be established, e.g., by interpreting possibility as upper probability [31] or fuzzy sets as coherent random sets [28].

available support for $x_0 \in C_0$. This support depends on the distance of x_0 to the observed negative examples. Note that $\delta_{x_0}(0) = 0$ is possible as well. In this case, no evidence is available at all, neither for nor against $x_0 \in C_1$. See Section 6.3 for a simulation experiment which concerns the aspect of robustness of NN estimation toward violations of the standard statistical assumptions.

Apart from statistical assumptions, the structure of the application has an important influence. To illustrate, consider two classes in the form of two clusters such that the (known) diameter of both clusters is smaller than the distance between them, that is $\mathcal{D}_{\mathcal{X}}(x_1, x_2) < \mathcal{D}_{\mathcal{X}}(x_1, x_3)$ whenever $\lambda_{x_1} = \lambda_{x_2} \neq \lambda_{x_3}$. The label of an instance can then be determined with certainty as soon as the distance from its nearest neighbor is known. In other words, the 1NN rule which does not involve frequency information performs better than any kNN rule with $k > 1$.

4.4. NN estimations and NN decisions

In addition to the extrapolation principles let us compare the induced distributions, referred to as NN estimations, from a knowledge representational point of view, especially against the background of the two shortcomings of the NN rule illustrated in Fig. 1.

A crucial difference between a possibility distribution δ and a probability function p is that the latter obeys a normalization constraint that demands a total probability mass of 1, whereas no such constraint exists in possibility theory. Consequently, a possibility distribution is more expressive in some situations. Especially, the following points deserve mentioning:

- Possibility reflects ignorance: All possibility degrees $\delta_{x_0}(\lambda)$ remain rather small if no sufficiently similar instances are available. Particularly, the distribution $\delta_{x_0} \equiv 0$ is an expression of *complete ignorance* and reflects the absence of any relevant observation ($\sigma_{\mathcal{X}}(x_0, x_\imath) = 0$ for all $x_\imath$). A learning agent using this estimation "knows that it doesn't know" [70]. As opposed to this, a distribution such as, say, $\delta_{x_0} \equiv 1/m$ indicates that some (small) evidence is available for each of the m labels $\lambda_\imath$. These two situations cannot be distinguished in probability theory where they induce the same distribution $p_{x_0} \equiv 1/m$ (if, as suggested by the principle of insufficient reason, complete ignorance is modeled by the uniform distribution).
- Possibility reflects absolute frequency: For example, suppose $\sigma_{\mathcal{X}}(x_0, x_\imath) = 1 - d > 0$ and $\lambda_{x_\imath} = \lambda_1$ for all n instances $x_\imath$ stored in memory. The probabilistic estimation (7) then yields the one-point distribution $p_{x_0}(\lambda_1) = 1$ and $p_{x_0}(\lambda) = 0$ for all $\lambda \neq \lambda_1$. Thus, it suggests that $\lambda_{x_0} = \lambda_1$ is certain, even if n is rather small. With a compensating t-conorm such as the probabilistic sum $\oplus_P$, the extended estimation (26) yields $\delta_{x_0}(\lambda_1) = 1 - d^n$ and $\delta_{x_0}(\lambda) = 0$ for all $\lambda \neq \lambda_1$. Thus, not only does the possibilistic support of the hypothesis $\lambda_{x_0} = \lambda_1$ reflect the distance but also the actual number of voting instances: $\delta_{x_0}(\lambda_1)$ is an increasing function of n and approaches 1 for $n \to \infty$.

As can be seen, a probabilistic estimation can represent ambiguity, whereas the possibilistic approach captures both problems, ambiguity and ignorance: Ambiguity (Fig. 1, top) is present if there are several plausible labels with similar degrees of support, and ignorance (Fig. 1, bottom) is reflected by the fact that even the most supported label has a small degree of possibility. Thus, (26) can be taken as a point of departure for a decision making procedure that goes beyond the guessing of a label. For example, a possible line of action proceeding from (26) might be expressed by the following rules (involving thresholds $0 < d_{\max} < d_{\min} < 1$):

- If $\delta_{x_0}(\lambda^*) \geqslant d_{\min}$ for the most supported label λ^* and $\delta_{x_0}(\lambda) \leqslant d_{\max}$ for all $\lambda \neq \lambda^*$, then let $\lambda_{x_0}^{\mathrm{est}} = \lambda^*$.
- If $\delta_{x_0}(\lambda^*) < d_{\min}$, then gather further information.
- If $\delta_{x_0}(\lambda^*) \geqslant \delta_{x_0}(\lambda) \geqslant d_{\min}$ for two labels $\lambda^*, \lambda \in \mathcal{L}$, then refuse a prediction.

The ECHOCARDIOGRAM DATABASE[30] is a real-world example that is quite interesting in this respect. One problem that has been addressed by machine learning researchers in connection with this database is to predict from several attributes whether or not a patient who suffered from a heart attack will survive at least one year. Since data is rather sparse (132 instances and about 10 attributes), the possibilistic approach often yields estimations with low support for both alternatives, surviving and not surviving at least one year. This is clearly reasonable from a knowledge representational point of view and reveals an advantage of absolute over relative degrees of support. For example, telling a patient that your experience does not allow any statement concerning his prospect of survival ($\delta_{x_0} \equiv 0$) is very different from telling him that his chance is 1/2 ($p_{x_0} \equiv 1/2$).

Let us mention that a generalization of the kNN rule closely related to our approach has been developed in [19]. In this method, which is also motivated by the problems of ambiguity and ignorance in the original kNN rule, an estimation of the label λ_{x_0} is given in terms of a *belief function* [66] rather than a possibility distribution. See [27] for a comparison between the two approaches.

The discrepancy between a probabilistic and a possibilistic approach (or an approach based on belief functions) disappears to some extent if one is only interested in a final decision, that is if a decision must be made irrespective of the quality and quantity of the information at hand. The method in [19], for example, refers to the so-called transferable belief model [69] and, hence, turns the belief function (at the "credal" level) specifying the unknown label into a probability function (at the "pignistic" level) before making a decision. Thus, the support of individual labels is expressed in terms of probability, and an NN estimation can be derived by taking one among the most probable labels, breaking ties at random.

Observe that, as a consequence of applying the maximum operator, a possibilistic NN decision derived from (20) coincides with the 1NN rule. The generalized version (26), where several moderately similar examples can compensate for one very similar instance, comes closer to the original kNN rule. In fact, for certain special cases, the possibilistic approach is equivalent—from a decision making point of view—to the probabilistic approach based on the support function (41). Eq. (28) shows that a possibility degree $\delta_{x_0}(\lambda)$ is a monotone transformation of the sum of weights $\omega_\imath$, and this relation is one-to-one if

[30] Available at http://www.ics.uci.edu/~mlearn.

the pseudo-inverse $g^{(-1)}$ is actually the inverse g^{-1}. The similarity function $\sigma_{\mathcal{X}}$ can then be chosen such that

$$\delta_{x_0}(\lambda_\imath) \leqslant \delta_{x_0}(\lambda_\jmath) \quad \Leftrightarrow \quad p_{x_0}(\lambda_\imath) \leqslant p_{x_0}(\lambda_\jmath).$$

That is, labels which are better supported in a possibilistic sense are also more probable and vice versa.

To illustrate, consider the case where $\mathcal{X} = \mathbb{R}^l$ and $\sigma_{\mathcal{L}}(\lambda, \lambda') = 1$ if $\lambda = \lambda'$ and 0 otherwise. Let K be a kernel function and define $\sigma_{\mathcal{X}}$ as $(x, y) \mapsto 1 - \exp(-K(x, y))$.[31] For the t-conorm $\oplus_P$, the weights in (28) are then given by $\omega_\imath = K(x_0 - x_\imath)$. Therefore,

$$\delta_{x_0}(\lambda_\imath) = 1 - \exp\biggl(-\sum_{\langle x, \lambda_x\rangle \in S:\ \lambda_x = \lambda_\imath} K(x_0 - x)\biggr) = 1 - \exp\bigl(-c \cdot p_{x_0}(\lambda_\imath)\bigr),$$

where $p_{x_0}(\lambda_\imath)$ is the probability degree derived from (41) using the kernel function K and c is the normalization factor

$$c = \sum_{\lambda \in \mathcal{L}} p_{x_0}(\lambda).$$

4.5. *An illustrative example*

Here, we present a simple example for which the possibilistic approach might be considered superior to the probabilistic one. The task shall be to predict a student's grade in physics given some information on other grades of that student. Thus, an instance is now a subject, and the label is given by the corresponding grade. We assume that grades are taken from the scale $\mathcal{L} = \{0, 1, \ldots, 10\}$, where 10 is the best result. Moreover, we consider two scenarios S1 and S2:

Subject	S1	S2
Chemistry	–	10
French	–	3
Philosophy	–	3
Spanish	–	3
Sports	5	–

It is clearly not obvious how to define a reasonable similarity measure over the set of subjects. In fact, an ordinal measure—sufficient for the possibilistic approach (20)—appears much simpler than a cardinal one. Nevertheless, let us assume the following (cardinal) degrees of similarity:

$\sigma_{\mathcal{X}}$	Chem.	French	Phil.	Span.	Sports
Physics	3/4	1/3	1/3	1/3	0

[31] Formally, one might set $K(0) \doteq \infty$ to ensure that $\sigma_{\mathcal{X}}$ is reflexive.

Concerning the set of labels $\mathcal{L}$, graded degrees of similarity are clearly advised in this example. Let us define the similarity between two grades a and b to be

$$\sigma_{\mathcal{L}}(a,b) = \max\left\{1 - \tfrac{1}{5}|a-b|, 0\right\}.$$

Needless to say, our application does not define a statistical setup par excellence, which is a main reason why the probabilistic approach does hardly appear suitable. To begin with, a scenario as defined above cannot be considered as an independent sample (perhaps the information is censored if it comes from the student himself), not to mention the small number of observations. Moreover, a relative frequency interpretation does not make sense. Finally, the set $\mathcal{X}$ endowed with the similarity measure $\sigma_{\mathcal{X}}$ (as partly specified above) is likely to lack the mathematical (metric) structure that enables one to define a reasonable kernel function K (either on $\mathcal{X}$ or on $\mathcal{X} \times \mathcal{L}$). Consequently, the derivation of the kNN estimation in Section 3.2 is no longer valid. Clearly, nothing prevents us from still applying the formulae and simply interpreting the normalized degrees of additive support as degrees of probability. But one should keep in mind that this approach actually lacks a solid foundation.

The first scenario is a typical example of complete ignorance, for one does not have any relevant piece of information. It is true that the case base is not empty, but the grade in sports does not allow one to draw any conclusion on the grade in physics since these two subjects are very dissimilar. This is adequately reflected by the possibilistic estimation which yields $\delta_{x_0} = \delta_{\text{physics}} \equiv 0$. A probabilistic estimation with relative support is obviously not appropriate in this example. Since sports is the only neighbor one obtains a probability distribution that favors grade 5 for physics. Thus, it is clearly advised to use absolute rather than relative support. Then, however, a probability is actually not defined since the denominator in (38) is zero. One way out is to take the uniform distribution $p_{x_0} \equiv 1/11$ as a default estimation, but this raises the well-known question whether the latter is an adequate expression of complete ignorance (which is definitely denied by most scholars).

Scenario S2 reveals problems of weighting and aggregation. Undoubtedly, a weighted estimation should be preferred in this example. Still, the example shows that the definition and aggregation of weights can be tricky. What is the most likely grade? Particularly, is grade 3 for physics more likely than grade 10 or vice versa? The weighted kNN rule favors grade 3 since the three subjects which are moderately similar to physics compensate for the one (chemistry) which is very similar. Of course, this result might be judged critically. Especially, this example reveals a problem of interdependence which is not taken into account by means of a simple summation of weights. Namely, the two subjects Spanish and French are very similar by themselves. Thus, one might wonder whether the grade 3 should really count twice. In fact, one might prefer to consider the grades in French and Spanish as only one piece of evidence (suggesting that the student is not good at languages) instead of two pieces of distinct information. Formally, the problem is that the probabilistic approach makes an assumption of (conditional) independence which is no longer valid when taking *structural* assumptions about the application into account. Here, such assumptions correspond to the NN inductive bias, namely the hypothesis that similar instances have similar classifications. Given this hypothesis, the instances stored in the case base are no longer independent (grade 3 in French, in conjunction with this hypothesis, makes grade 3 in Spanish very likely).

The problem of interdependence cannot be taken into account as long as an estimation disregards the similarity between the instances stored in memory, as do all the estimations presented so far. Still, the aggregation operator $\oplus$ in the possibilistic approach provides a means for alleviating the problem. With $\oplus = \max$, for example, frequency does not count at all and one obtains $\delta_{x_0}(3) = 1/3 < 3/4 = \delta_{x_0}(10)$. The probabilistic sum $\oplus_P$ brings about a reinforcement effect but still yields $\delta_{x_0}(3) = 0.7 < 3/4 = \delta_{x_0}(10)$, a result that appears quite reasonable.

A second problem related to scenario S2 is that of ambiguity. Particularly, the probabilistic approach yields a bimodal distribution p_{x_0}, and the same is also true for most aggregation operators in the possibilistic approach. For example, (26) with $\oplus = \oplus_P$ (and $\otimes = \otimes_P$) yields $\delta_{x_0}(3) > \delta_{x_0}(7) < \delta_{x_0}(10)$. This result is not intuitive, for one might hardly judge an intermediate grade less possible than two extreme grades. To solve this problem, δ_{x_0} can be replaced by its convex hull

$$\lambda \mapsto \min\Bigl\{\max_{\lambda' \leqslant \lambda} \delta_{x_0}(\lambda'), \max_{\lambda' \geqslant \lambda} \delta_{x_0}(\lambda')\Bigr\}. \tag{43}$$

In our example, this leads to the following distribution:

λ	0	1	2	3	4	5	6	7	8	9	10
$\delta_{x_0}(\lambda)$	0	0.3	0.53	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.75

Of course, this prediction is still ambiguous in the sense that is supports several grades by means of high degrees of possibility. This is not a defect, however, but rather an adequate representation of the ambiguity which is indeed present in the situation associated with scenario S2.

The modification (43) of δ_{x_0} should not be considered ad-hoc. Rather, the convexity requirement can be thought of as a possibility-qualifying rule that complements the similarity-based justification of possibility degrees: The more possible two labels are, the more possible is any label in-between. This type of background knowledge and the associated constraints can be met more easily in the possibilistic approach than in the probabilistic one. In fact, the incorporation of background information is hardly compatible with non-parametric density estimation.

In summary, the example has shown the following advantages of the possibilistic approach: First, the interpretation of aggregated weights in terms of degrees of evidential support is often less critical than the interpretation in terms of degrees of probability. Second, a possibility distribution can represent ignorance. Third, the use of aggregation operators other than the arithmetic sum can be useful. Fourth, the possibilistic approach is more flexible and allows for incorporating constraints or background knowledge.

4.6. *Complexity issues*

Even though algorithmic aspects are beyond the scope of this paper, let us have a rough look at the computational complexity of our possibilistic approach to IBL. A straightforward implementation of the prediction (25) has a running time which is linear in the size $|S|$ of the sample and the number $|\mathcal{L}|$ of labels. In this respect, it is hence completely comparable to other instance-based learning methods.

In order to reduce the computational complexity, IBL approaches take advantage of the fact that a prediction is already determined by the nearest neighbors of the query instance. Thus, the consideration of each sample instance is actually not necessary, and efficiency can be gained by means of fast algorithms for finding nearest neighbors [40,49,81]. Such algorithms employ efficient similarity-based indexing techniques and corresponding data structures in order to find the relevant instances quickly.

The same idea can be applied in connection with our possibilistic approach. In fact, a possibility degree $\delta_{x_0}(\lambda)$ is completely determined by the neighborhood of the case $\langle x_0, \lambda\rangle$, that is the sample instances $\langle x, \lambda_x\rangle$ satisfying $\sigma_{\mathcal{X}}(x, x_0) > 0$ and $\sigma_{\mathcal{L}}(\lambda_x, \lambda) > 0$. As can be seen, apart from minor differences, the possibilistic method is quite comparable to other IBL methods from a complexity point of view. One such difference concerns the relevant sample instances. In the kNN approach, the number of relevant instances in always k, but the (degree of) relevance of an instance may change when modifying the case base. As opposed to this, the degree of relevance of a neighboring instance is fixed in the possibilistic approach, but the number of relevant instances can change.

Let us finally mention that efficiency can also be gained if the complete possibility distribution δ_{x_0} is not needed. In fact, quite often one will only be interested in those labels having a high degree of possibility. For example, one might be interested in a fixed number of maximally supported labels, or in those labels whose support exceeds a given possibility threshold. In such cases, the computation of $\delta_{x_0}(\lambda)$ can be omitted (or broken off) for certain labels λ.

5. Possibilistic instance-based learning

Proceeding from the NN estimation (26), we have developed a possibilistic method of instance-based learning, called POSSIBL. This section presents some extensions of the basic model which turn POSSIBL into a powerful and practically useful IBL algorithm.

5.1. Dealing with incomplete information

The problem of dealing with incomplete information such as missing attribute values in an important issue in machine learning [20,61]. For example, suppose that the specification of the new instance x_0 is incomplete, and let $X_0 \subseteq \mathcal{X}$ denote the instances compatible with the description of x_0. Moreover, recall the lower support-bound semantics of our possibilistic approach to IBL. The following generalization of (26) is in accordance with these semantics:

$$\delta_{x_0}(\lambda) \doteq \inf_{x \in X_0} \delta_x(\lambda) = \inf_{x \in X_0} \bigoplus_{1 \leqslant \imath \leqslant n} \sigma_{\mathcal{X}}(x, x_\imath) \otimes \sigma_{\mathcal{L}}(\lambda, \lambda_{x_\imath}). \tag{44}$$

Indeed, each potential candidate $x \in X_0$ gives rise to a lower bound according to (26), and without additional knowledge we can guarantee but the smallest of these bounds to be valid. This is in agreement with the idea of *guaranteed possibility* (cf. Section 2.2). The simplicity of handling incomplete information in a coherent (namely possibilistic) way is clearly a strong point of POSSIBL. Notice that the computation of the lower bound in (44) is in line with the handling of missing attribute values in IB1, where these values are assumed to be maximally different from the comparative value. Yet, the possibilistic solution appears more appealing since it avoids any default assumption. Indeed, inferring what is *possible* seems to be a reasonable way of dealing with missing attribute values and for handling incomplete and uncertain information in a coherent way.

More generally, imprecise knowledge about x_0 can be modeled in the form of a possibility distribution π on $\mathcal{X}$, where $\pi(x)$ corresponds to the degree of plausibility that $x_0 = x$. A graded modeling of this kind is useful, e.g., if some attributes are specified in a linguistic way. It suggests the following generalization of (44):

$$\delta_{x_0}(\lambda) \doteq \inf_{x \in \mathcal{X}} \big(\pi(x) \rightsquigarrow \delta_x(\lambda)\big), \tag{45}$$

where $\rightsquigarrow$ is a generalized implication operator that is reasonably chosen as the Gödel implication [35]:

$$\alpha \rightsquigarrow \beta \doteq \begin{cases} 1 & \text{if } \alpha \leqslant \beta, \\ \beta & \text{if } \alpha > \beta. \end{cases}$$

From a logical point of view, (45) specifies the extent to which *the label* λ *is supported by all plausible candidates for* x_0. Notice that the distributions δ_x and π in (44) have different semantics and express degrees of confirmation and plausibility, respectively (cf. Section 2). Particularly, π is assumed to be normalized, i.e., there is at least one instance x with $\pi(x) = 1$. One obviously recovers (44) from (45) for the special case where π is a $\{0, 1\}$-valued possibility distribution $\pi = \mathbb{I}_{X_0}$ and hence corresponds to a crisp subset $X_0 \subseteq \mathcal{X}$.

Similar generalizations can also be realized for coping with incompletely specified examples. Let the $\imath$th case in the memory be characterized by the set $X_\imath \times L_\imath \subseteq \mathcal{X} \times \mathcal{L}$. Then, (26) becomes

$$\delta_{x_0}(\lambda) \doteq \bigoplus_{1 \leqslant \imath \leqslant n} \inf_{\langle x, \lambda_x \rangle \in X_\imath \times L_\imath} \sigma_{\mathcal{X}}(x_0, x) \otimes \sigma_{\mathcal{L}}(\lambda, \lambda_x),$$

which is in accordance with (44). Moreover, we obtain

$$\delta_{x_0}(\lambda) \doteq \bigoplus_{1 \leqslant \imath \leqslant n} \inf_{\langle x, \lambda_x \rangle \in \mathcal{X} \times \mathcal{L}} \max\big\{\sigma_{\mathcal{X}}(x_0, x) \otimes \sigma_{\mathcal{L}}(\lambda, \lambda_x), 1 - \pi_\imath(x, \lambda_x)\big\}$$

if the $\imath$th case is characterized by means of a possibility distribution $\pi_\imath$ on $\mathcal{X} \times \mathcal{L}$ rather than by a crisp set $X_\imath \times L_\imath$. Observe that this expression can be combined with (45) in order to handle incomplete specifications of both, the sample cases and the new instance. Moreover, notice that the distribution δ_{x_0} will generally remain unaffected if an example is completely unspecified ($\pi_\imath \equiv 1$), which is clearly a reasonable property. See [27] for a more thorough discussion of handling incomplete information and for a more detailed derivation of the above extensions.

5.2. Discounting noisy and atypical instances

IBL is quite sensitive to noisy instances which should hence be discarded [2]. By noise one generally means incorrect attribute value information, concerning either the descriptive part x of a case or the label λ_x (or both). However, the problem of noise is also closely related to the "typicality" of a case. A typical instance is representative of its neighbors, whereas an exceptional (though not incorrect) instance has a label quite different from the labels of neighboring instances [85].

Recall that each case $\langle x_\imath, \lambda_{x_\imath}\rangle \in S$ is extrapolated by placing the support function or, say, "possibilistic kernel" (25) around the point $\langle x_\imath, \lambda_{x_\imath}\rangle \in \mathcal{X} \times \mathcal{L}$, just like a density (kernel) function is centered around each observation in kernel-based density estimation. Of course, the less representative (i.e., noisy or exceptional) an instance is of its neighborhood, the smaller the extent of extrapolation should be.

A simple learning mechanism that adapts the extent of extrapolation of stored cases can be realized by means of a slight generalization of the kernel function (25):

$$\delta^{\imath}_{x_0} : \lambda \mapsto m_\imath\big(\sigma_{\mathcal{X}}(x_0, x_\imath)\big) \otimes \sigma_{\mathcal{L}}(\lambda, \lambda_{x_\imath}). \tag{46}$$

Here, $m_\imath : [0,1] \to [0,1]$ is a monotone increasing modifier function with $m_\imath(1) = 1$. This function allows for discounting atypical cases. Roughly speaking, $m_\imath$ adapts the similarity between the instance $x_\imath$ and its neighbors. For example, $x_\imath$ is made completely dissimilar to all other instances by letting $(m_\imath | [0,1[) \equiv 0$. Replacing $\sigma_{\mathcal{X}}$ by the modified measure $m_\imath \circ \sigma_{\mathcal{X}}$ is closely related to the idea of local distance measures in NN algorithms.

Suppose that a new observation x_0 with label λ_{x_0} has been made, and consider a stored case $\langle x_\imath, \lambda_{x_\imath}\rangle$. Should this case be discounted in the light of the new observation? The fact that $\langle x_\imath, \lambda_{x_\imath}\rangle$ supports a label different from the observed label λ_{x_0} need not necessarily be a flaw. In fact, recall that $x_0 \in C_{\lambda_{x_0}}$ does not exclude that $x_0 \in C_\lambda$ for some $\lambda \neq \lambda_0$. In other words, neither the non-support of the observed nor the support of a different label can actually be punished. However, what can be punished is the disqualification of the label λ_{x_0} as expressed by the upper possibility model (31). Thus, it is reasonable to require that the degree of disqualification induced by $\langle x_\imath, \lambda_{x_\imath}\rangle$ is bounded:

$$1 - m_\imath\big(\sigma_{\mathcal{X}}(x_0, x_\imath)\big) \otimes \sigma_{\mathcal{L}}(\lambda_{x_0}, \lambda_{x_\imath}) \geqslant \beta, \tag{47}$$

where $\beta \gg 0$ is a constant.

The constraint (47) suggests an update scheme in which a stored case $\langle x_\imath, \lambda_{x_\imath}\rangle$ is (maybe) discounted every time a new observation $\langle x_0, \lambda_{x_0}\rangle$ is made: Let $\mathcal{F}$ denote a parameterized and completely ordered class of functions from which $m_\imath$ is chosen. An adaptation is then realized by

$$m_\imath \leftarrow \min\Big\{m_\imath, \sup\big\{f \in \mathcal{F} \mid 1 - f\big(\sigma_{\mathcal{X}}(x_0, x_\imath)\big) \otimes \sigma_{\mathcal{L}}(\lambda_{x_0}, \lambda_{x_\imath}) \geqslant \beta\big\}\Big\}. \tag{48}$$

The discounting of noisy and atypical instances through modifying possibilistic kernel functions appears natural and somewhat simpler than the method used in IB3 [2]. Firstly, possibilistic discounting is gradual, whereas an instance is either accepted or rejected (or is temporarily in-between) in IB3. Secondly, the question whether to discount an instance and to which extent is answered quite naturally in the possibilistic approach, where support is absolute and graded. In IB3, an instance is either punished or not, and the corresponding

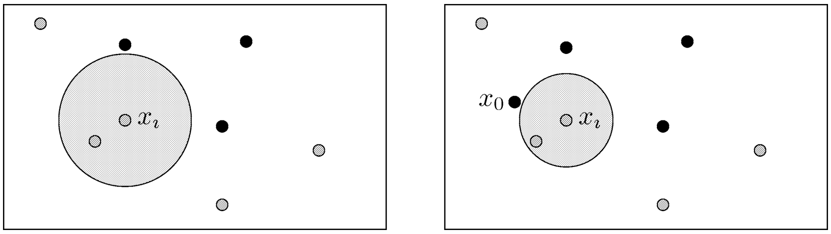


Fig. 2. Left: The large circle corresponds to the support function (possibilistic kernel) centered around $x_\imath$ and marks the extrapolation of label $\lambda_{x_\imath}$. Right: The support function is updated after observing a new instance which has a different label $\lambda_{x_0} \neq \lambda_{x_\imath}$ and hence must not be supported.

decision is based on a rule that appears reasonable but might still be considered ad-hoc ($x_\imath$ is discounted if $\mathcal{D}_{\mathcal{X}}(x_\imath, x_0)$ is smaller than or equal to the distance between x_0 and its closest *accepted* neighbor[32]).

The possibilistic adaptation scheme becomes rather simple for the special case $\mathcal{X} = \mathbb{R}^l$, $\mathcal{L} = \{0, 1\}$ and $m_\imath = \mathbb{I}_{]\gamma_\imath, 1]}$, where $0 \leqslant \gamma_\imath < 1$. If $\sigma_{\mathcal{X}}$ is a strictly decreasing function of Euclidean distance, then the support function (25) corresponds to a ball around $x_\imath$: $\delta^\imath_{x_0}(\lambda) = 1$ if $\lambda = \lambda_x$ and x_0 is located inside that ball and $\delta^\imath_{x_0}(\lambda) = 0$ otherwise. The parameter $\gamma_\imath$ is chosen as large as possible, but such that the support function does not cover any observed instance $x_\jmath$ with $\lambda_{x_\jmath} \neq \lambda_{x_\imath}$, that is $\gamma_\imath \leqslant |x_\imath - x_\jmath|$ holds true for all of those $x_\jmath$. Fig. 2 gives an illustration for $l = 2$.

This special case is a useful point of departure for investigating theoretical properties of POSSIBL. In [4], some convergence properties of IB1 have been shown for a special setup which makes statistical assumptions about the generation of training data and geometrical assumptions on a concept C_1 to be learned. For POSSIBL, one can prove similar properties under the same assumptions. More specifically, let $l = 2$, $\mathcal{X} = [0, 1] \times [0, 1]$ (the results can be generalized to any dimension $l > 2$ and any bounded region $\mathcal{X} \subseteq \mathbb{R}^l$) and consider a concept $C_1 \subseteq \mathcal{X}$. For the special case above, the POSSIBL approximation of C_1 is then given by

$$C_1^{\text{est}} = \bigcup_{\langle x_\imath, 1\rangle \in S} \mathfrak{B}_{\rho(x_\imath)}(x_\imath), \tag{49}$$

where $\mathfrak{B}_d(x_\imath) = \{x \in \mathcal{X} \mid |x - x_\imath| < d\}$ is the (open) d-ball around $x_\imath$ and

$$\rho(x_\imath) = \min\bigl\{|x_\jmath - x_\imath| \mid \langle x_\jmath, \lambda_{x_\jmath}\rangle \in S, \lambda_{x_\jmath} \neq \lambda_{x_\imath}\bigr\}. \tag{50}$$

Moreover, the approximation of $C_0 = \mathcal{X} \setminus C_1$ is given by

$$C_0^{\text{est}} = \bigcup_{\langle x_\imath, 0\rangle \in S} \mathfrak{B}_{\rho(x_\imath)}(x_\imath). \tag{51}$$

It is readily verified that $C_0^{\text{est}} \cap C_1^{\text{est}} = \emptyset$. However, $C_0^{\text{est}} \cup C_1^{\text{est}} = \mathcal{X}$ does not necessarily hold true. Thus, one may have $\delta_{x_0} \equiv 0$ for some instances $x_0 \in \mathcal{X}$ (which are then classified at random). Consequently, an approximation of concept C_1 should actually be

[32] Auxiliary rules are used if x_0 does not have an accepted neighbor.

represented by the tuple $(C_0^{\mathrm{est}}, C_1^{\mathrm{est}})$ which divides instances $x_0 \in \mathcal{X}$ into three groups: Those which (supposedly) belong to C_1 ($\delta_{x_0}(0) = 0, \delta_{x_0}(1) = 1$), those which do not ($\delta_{x_0}(0) = 1, \delta_{x_0}(1) = 0$), and those for which no evidence is available so far ($\delta_{x_0} \equiv 0$).

Now, a first desirable property is the convergence of the concept approximation, that is the convergence of C_0^{est} and C_1^{est} toward C_0 and C_1, respectively. In this context, however, the property of convergence itself has to be weakened since exact convergence cannot be achieved due to the fact that an NN classifier cannot guarantee the avoidance of wrong decisions at the boundary of a concept. Moreover, some assumptions on the generation of samples and on the geometry of the concept C_1 have to be made. Here, we make the same assumptions as in [4]: Instances are generated randomly and independently according to a fixed probability measure μ over $\mathcal{X}$. Furthermore, C_1 is a concept having a *nice* boundary, which is the union of a finite number of closed (hyper-)curves of finite size.

We employ the following notation: The ε-neighborhood of C_1 is the set

$$C_1^+(\varepsilon) \doteq \bigl\{x \in \mathcal{X} \mid \mathfrak{B}_\varepsilon(x) \cap C_1 \neq \emptyset\bigr\},$$

and the ε-core of C_1 is defined by

$$C_1^-(\varepsilon) \doteq \bigl\{x \in \mathcal{X} \mid \mathfrak{B}_\varepsilon(x) \subseteq C_1\bigr\}.$$

A set $A \subseteq \mathcal{X}$ is called an (ε, γ)-approximation of C_1 if there is a (measurable) set $N \subseteq \mathcal{X}$ with $\mu(N) \leqslant \gamma$ and such that

$$\bigl(C_1^-(\varepsilon) \setminus N\bigr) \subseteq (A \setminus N) \subseteq \bigl(C_1^+(\varepsilon) \setminus N\bigr).$$

Finally, let $C_{1,n}^{\mathrm{est}}$ and $C_{0,n}^{\mathrm{est}}$ denote, respectively, the possibilistic concept approximations (49) and (51) for $|S| = n$, i.e., after n observations have been made.

Lemma 2. *The equalities*

$$C_1^-(\varepsilon) = \mathcal{X} \setminus C_0^+(\varepsilon) \quad \textit{and} \quad C_0^-(\varepsilon) = \mathcal{X} \setminus C_1^+(\varepsilon)$$

hold true for all $0 < \varepsilon < 1$.

Proof. For $x \in C_1^-(\varepsilon)$ we have $\mathfrak{B}_\varepsilon(x) \subseteq C_1$, which means that $|x - x_1| < \varepsilon$ implies $x_1 \in C_1$. Consequently, there is no $x_0 \in C_0$ such that $|x - x_0| < \varepsilon$ and, hence, $x \notin C_0^+(\varepsilon)$. Now, suppose $x \in \mathcal{X} \setminus C_0^+(\varepsilon)$. Thus, there is no $x_0 \in C_0$ such that $|x - x_0| < \varepsilon$, which means that $|x - x_1| < \varepsilon$ implies $x_1 \in C_1$ and, hence, $x \in C_1^-(\varepsilon)$. The second equality is shown in the same way. □

Theorem 3. *Let $C_1 \subseteq \mathcal{X}$ and $0 < \varepsilon, \gamma, d < 1$. There is an integer n_0 such that the following holds true with probability at least $1 - d$: The possibilistic concept approximation $C_{1,n}^{\mathrm{est}}$ is a $(2\varepsilon, \gamma)$-approximation of C_1 and $C_{0,n}^{\mathrm{est}}$ is a $(2\varepsilon, \gamma)$-approximation of C_0 for all $n > n_0$.*

Proof. Let N denote the set of instances $x \in \mathcal{X}$ for which no sample $x_\imath \in S$ exists such that $|x - x_\imath| < \varepsilon$. In [4], the following lemma has been shown: $\mu(N) \leqslant \gamma$ holds true with probability $1 - d$ whenever

$$n > \lceil n_0 = \sqrt{2}/\varepsilon\rceil^2/\gamma^2 \cdot \ln\bigl(\lceil\sqrt{2}/\varepsilon\rceil^2/d\bigr). \tag{52}$$

Subsequently, we ignore the set N, that is we formally replace $\mathcal{X}$ by $\mathcal{X} \setminus N$, C_1 by $C_1 \setminus N$ and C_0 by $C_0 \setminus N$. Thus, the following holds true by definition: For each $x \in \mathcal{X}$ there is an instance $x_\imath \in S$ such that $|x - x_\imath| < \varepsilon$.

Now, consider any instance $x \in C_1^-(2\varepsilon)$. We have to show that $x \in C_{1,n}^{\mathrm{est}}$. Let $x_\imath \in S$ be an instance such that $|x - x_\imath| < \varepsilon$. For this instance we have $x_\imath \in \mathfrak{B}_\varepsilon(x) \subseteq C_1$, which means that $x_\imath$ belongs to C_1. Furthermore, $\mathfrak{B}_\varepsilon(x_\imath) \subseteq \mathfrak{B}_{2\varepsilon}(x) \subseteq C_1$ and, hence, $\rho(x_\imath) \geqslant \varepsilon$ for the value in (50). This implies that $x \in \mathfrak{B}_{\rho(x_\imath)}(x_\imath)$ and, therefore, $x \in C_{1,n}^{\mathrm{est}}$. Thus, we have shown that $C_1^-(2\varepsilon) \subseteq C_{1,n}^{\mathrm{est}}$.

Since the same arguments apply to C_0, the property $C_0^-(2\varepsilon) \subseteq C_{0,n}^{\mathrm{est}}$ can be shown in an analogous way. Thus, using Lemma 2,

$$C_{1,n}^{\mathrm{est}} \subseteq \mathcal{X} \setminus C_{0,n}^{\mathrm{est}} \subseteq \mathcal{X} \setminus C_0^-(2\varepsilon) = C_1^+(2\varepsilon).$$

Likewise, one shows that $C_{0,n}^{\mathrm{est}} \subseteq C_0^+(2\varepsilon)$. □

Roughly speaking, Theorem 3 guarantees that the 2ε-core of both, C_0 and C_1 is classified correctly (with high probability) if the sample S is large enough. In other words, classification errors can only occur in the boundary region. For being able to quantify the probability of an error, it is necessary to put restrictions on the size of that boundary region and on the probability distribution μ. Thus, let $\mathcal{C}$ denote the class of concepts $C_1 \subseteq \mathcal{X}$ that can be represented as the union of a finite set of regions bounded by closed curves with total length of at most L [4]. Moreover, let $\mathfrak{P}_\beta$ denote the class of probability distributions μ over $\mathcal{X}$ such that $\mu(A) \leqslant \mu_L(A) \cdot \beta$ for all Borel-subsets $A \subseteq \mathcal{X}$, where μ_L is the Lebesgue measure and $\beta > 0$.

Theorem 4. *The concept class $\mathcal{C}$ is polynomially learnable with respect to $\mathfrak{P}_\beta$ by means of the possibilistic concept approximation $(C_0^{\mathrm{est}}, C_1^{\mathrm{est}})$.*

Proof. If $C_1 \in \mathcal{C}$, then the size of the region $C_1^+(2\varepsilon) \setminus C_1^-(2\varepsilon)$ is bounded by $4\varepsilon L$. Consequently, the probability of that area is at most $\alpha = 4\varepsilon L\beta$. Since a classification error can only occur either in this region or in the set N as defined in Theorem 3 and the probability of N is at most γ, the probability of a classification error is bounded by $\alpha + \gamma$. Now, fix the parameters γ and ε as follows: $\gamma = e/2$, $\varepsilon = e/(8L\beta)$. By substituting these parameters into (52) one finds that the required sample size n is polynomial in $1/e$ and $1/d$. In summary, the following holds true for any $0 < e, d < 1$, $C_1 \in \mathcal{C}$, and $\mu \in \mathfrak{P}_\beta$: If more than $n(1/e, 1/d)$ examples are presented, where n is a polynomial function of $1/e$ and $1/d$, then, with probability $1 - d$, the possibilistic concept approximation has a classification error of at most e. This is precisely the claim of the theorem. □

5.3. From instances to rules

Selecting appropriate instances to be stored in memory and pruning the training set are important issues in IBL that have a strong influence on performance. Especially reducing the size of the memory is often necessary in order to maintain the efficiency of the system. The basic idea is to remove instances which are actually not necessary to achieve good concept descriptions. For example, imagine a concept having the form of a circle in some (two-dimensional) instance space. To classify inner points correctly by means of the kNN rule it might then be sufficient to store positive examples of that concept near the boundary.

In connection with POSSIBL, where support is absolute rather than relative, deleting instances from memory might produce "holes" in the concept description. An interesting alternative, which allows one to reduce the size of the memory and, at the same time, to fill "holes" in the concept description by interpolation, is based on the idea of merging instances and of generalizing cases into rules. This idea appears particularly reasonable since the possibilistic estimation principle is closely related to fuzzy rule-based reasoning. More precisely, each observation can be interpreted as a fuzzy rule, namely as an instance of a fuzzy meta-rule suggesting that similar instances have similar labels.

To illustrate the one-to-one correspondence between rules and cases in POSSIBL, let $\mathcal{X} = \mathbb{R}$, $\mathcal{L} = \{0, 1\}$ and suppose that two instances $x_1 = 4$ and $x_2 = 6$ with label 0 have been observed. The possibilistic kernels (25) induced by these cases are shown in Fig. 3. The first case is equivalent to the fuzzy rule "If x_0 is approximately 4 then $\lambda = 0$" if the fuzzy set "approximately 4" is modeled by the possibility distribution $\delta^1_{x_0}$ (the individual support function (25)). The rules associated with the two cases can be merged into one rule, say, "If x_0 is about 5 then $\lambda = 0$", where the fuzzy set "about 5" is modeled by the pointwise maximum, $\delta^1_{x_0} \vee \delta^2_{x_0}$, of $\delta^1_{x_0}$ and $\delta^2_{x_0}$ (Fig. 3, right).

The above procedure is closely related to several other techniques that have been proposed in connection with IBL. Viewing cases as maximally specific rules and the idea of generalizing cases into rules has been put forward in [21,22]. The method proposed in [64] generalizes cases by placing rectangles of different size around them. A new instance is then labeled by the nearest rectangle rather than by the nearest case. This is very similar to our approach, where rectangles are replaced by possibility distributions. Relations also exist with the idea of merging nearest neighbors of the same class, thereby generating new (pseudo-sample) prototypes [11]. In our example, the point 5 may be regarded as a pseudo-instance replacing 4 and 6 (and also endowed with a modified support function).

In the example in Fig. 3, the summarizing rule is exactly equivalent to the conjunction of the two individual rules. Of course, the merging procedure might also incorporate concepts of approximation and interpolation. For example, suppose $x_2 = 8$ rather than $x_2 = 6$. The replacement of $\delta^1_{x_0} \vee \delta^2_{x_0}$ by its convex hull $\delta : x \mapsto \max\{\delta^1_{x_0}(x), \delta^2_{x_0}(x), \mathbb{I}_{[5,7]}\}$ then goes beyond a simple combination since δ is larger than the pointwise maximum of $\delta^1_{x_0}$ and $\delta^2_{x_0}$ (e.g., $\delta^1_{x_0}(6) = \delta^2_{x_0}(6) = 0.5 < 1 = \delta(6)$). This kind of possibilistic induction can be reasonable and often allows for incorporating background knowledge. Particularly, replacing

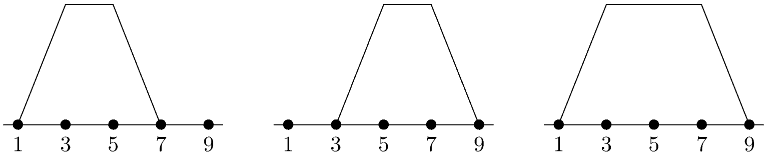


Fig. 3. Possibility distributions induced by two cases (left, middle) and the distribution associated with the summarizing fuzzy rule (right).

a possibilistic estimation δ_{x_0} by its convex hull is advised whenever a multimodal distribution does not make sense (as in our example in Section 4.5) or if the relation of observable cases (cf. Section 3.1) is even known to satisfy a convexity constraint of the form

$$x \in C_\lambda \cap C_{\lambda''} \quad \Rightarrow \quad x \in C_{\lambda'}$$

for all $\lambda < \lambda' < \lambda''$.

As can be seen, the extensions discussed here basically suggest a system that maintains an optimal rule base rather than an optimal case base, including the combination and adaptation of rules. These extensions are well-suited to the discounting of instances discussed in Section 5.2. Indeed, deriving one rule from several instances (or other rules) can be accomplished by replacing the latter by a pseudo-instance and defining an appropriate modifier function m for that pseudo-instance. Still, the extensions in this direction are premature and have not been implemented in POSSIBL yet.

6. Experimental studies

6.1. Preliminaries

This section presents some experimental studies providing evidence for POSSIBL's excellent performance in practice. We would like to emphasize, however, that it is not meant as an exhaustive comparative study covering several competing learning algorithms—and showing that POSSIBL is superior to all of its competitors. Apart from the fact that empirical studies are clearly of limited evidence,[33] one should realize that the primary motivation underlying POSSIBL is not another ε-improvement in classification accuracy but rather the enrichment of IBL by concepts of possibilistic reasoning (though the latter does clearly not exclude the former). Besides, one should keep the following points in mind. Firstly, POSSIBL has not been developed within a statistical framework. Thus, the type of problems for which POSSIBL is most suitable (see the example in Section 4.5) is perhaps not represented in the best way by standard (public) data sets commonly used for testing performance. Secondly, an important aspect of the possibilistic approach is the one of *knowledge representation*. But this aspect is neglected if—as in experimental studies—only the correctness of the final decision (classification accuracy) counts, not the estimated distribution. Thirdly, a comparison with other IBL algorithms might appear dubious since POSSIBL—in its most general form—is an *extension* of IBL and hence covers specific algorithms such as kNN as special cases.

Due to these difficulties, we have decided to apply a basic version of POSSIBL to several data sets from the UCI repository and to employ the kNN (resp. IB1) algorithm as a reference (we use kNN with $k = 1, 3, 5$ and the weighted 5NN rule with weight function (10)). Thus, we have refrained from tuning various degrees of freedom in order to optimize the performance of POSSIBL (an exception is only the experimental study presented in Section 6.4). Instead, we have applied the original max–min version (20), only extended

[33] It is well known that each algorithm has a selective superiority [10]. Thus, one will always find data sets for which a certain algorithm, at least after being tuned appropriately, performs better than others.

by the learning scheme presented in Section 5.2. The function $m_\imath$ in (46) was defined as $t \mapsto \exp(-\gamma_\imath(1-t))$, where $\gamma_\imath \geqslant 0$ is the discounting rate of the $\imath$th instance. The constant β in (47) was taken as 0.8.[34] In order to avoid difficulties due to the different handling of non-nominal class labels and the definition of similarity measures for non-numeric attributes, we have restricted ourselves to data sets for which all predictive attributes are numeric and for which the class label is defined on a nominal scale. The similarity $\sigma_{\mathcal{X}}$ is always defined as 1 minus the normalized Euclidean distance and the similarity $\sigma_{\mathcal{L}}$ is given by (23).

6.2. Classification accuracy

The experiments in this section were performed as follows: In a single simulation run, the data set is divided at random into a training set (the case base) and a test set, and the discounting rates $\gamma_\imath$ are adapted to the training set. A decision is then derived for each element of the test set by extrapolating the training set (but without adapting the discounting rates or expanding the case base any further), and the percentage of correct decisions is determined. Statistics are obtained by means of repeated simulation runs.

Results are summarized by means of statistics for the percentage of correct classifications (mean, standard deviation, minimum, maximum, 0.1-fractile, 0.9-fractile) (see Tables 1–5).

The experiments show that POSSIBL achieves comparatively good results and is always among the best algorithms. Thus, it can be said that a basic version of POSSIBL performs

Table 1
BALANCE SCALE DATABASE (625 observations, 4 predictive attributes, three classes, training set of size 300, 1,000 simulation runs)

Algorithm	mean	std.	min	max	0.1-frac.	0.9-frac.
POSSIBL	0.8776	0.0148	0.8215	0.9230	0.8584	0.8984
1NN	0.7837	0.0161	0.7323	0.8369	0.7630	0.8030
3NN	0.8117	0.0165	0.7630	0.8707	0.7907	0.8338
5NN	0.8492	0.0155	0.8030	0.8923	0.8307	0.8707
W5NN	0.7864	0.0164	0.7294	0.8428	0.7655	0.8067

Table 2
IRIS PLANT DATABASE (150 observations, 4 predictive attributes, three classes, training set of size 75, 10,000 simulation runs)

Algorithm	mean	std.	min	max	0.1-frac.	0.9-frac.
POSSIBL	0.9574	0.0204	0.8400	1.0000	0.9333	0.9733
1NN	0.9492	0.0196	0.8400	1.0000	0.9200	0.9733
3NN	0.9554	0.0175	0.8666	1.0000	0.9333	0.9733
5NN	0.9586	0.0181	0.8533	1.0000	0.9333	0.9866
W5NN	0.9561	0.0187	0.8400	1.0000	0.9333	0.9733

[34] Variations of this parameter had no significant influence.

Table 3
GLASS IDENTIFICATION DATABASE (214 observations, 9 predictive attributes, seven classes, training set of size 100, 10,000 simulation runs)

Algorithm	mean	std.	min	max	0.1-frac.	0.9-frac.
POSSIBL	0.6841	0.0419	0.5300	0.8400	0.6300	0.7400
1NN	0.6870	0.0410	0.5200	0.8200	0.6300	0.7400
3NN	0.6441	0.0421	0.4800	0.8100	0.5900	0.7000
5NN	0.6277	0.0412	0.4800	0.7800	0.5700	0.6800
w5NN	0.6777	0.0414	0.5000	0.8300	0.6200	0.7300

Table 4
PIMA INDIANS DIABETES DATABASE (768 observations, 8 predictive attributes, two classes, training set of size 380, 1,000 simulation runs)

Algorithm	mean	std.	min	max	0.1-frac.	0.9-frac.
POSSIBL	0.7096	0.0190	0.6421	0.7711	0.6868	0.7316
1NN	0.6707	0.0199	0.6132	0.7289	0.6447	0.6947
3NN	0.6999	0.0183	0.6447	0.7500	0.6763	0.7237
5NN	0.7190	0.0183	0.6553	0.7684	0.6947	0.7421
w5NN	0.6948	0.0188	0.6421	0.7474	0.6684	0.7184

Table 5
WINE RECOGNITION DATA (178 observations, 13 predictive attributes, three classes, training set of size 89, 1,000 simulation runs)

Algorithm	mean	std.	min	max	0.1-frac.	0.9-frac.
POSSIBL	0.7148	0.0409	0.5506	0.8652	0.6629	0.7640
1NN	0.7163	0.0408	0.5843	0.8652	0.6629	0.7640
3NN	0.6884	0.0407	0.5506	0.8315	0.6404	0.7416
5NN	0.6940	0.0392	0.5730	0.8090	0.6404	0.7416
w5NN	0.7031	0.0404	0.5730	0.8315	0.6517	0.7528

at least as well as the basic IBL (NN) algorithms. In other words, possibilistic IBL is in no way inferior to "standard" IBL as a basis for further improvements and sophisticated learning algorithms. This is exactly what we wanted to show.

Due to the special setting of our experimental studies, especially the choice of max as an aggregation operator and the use of a $\{0, 1\}$-valued similarity measure over $\mathcal{L}$, one might wonder how to explain the different performance of POSSIBL and the NN classifiers. In fact, in Section 4.4 it was argued that the possibilistic NN decision derived from (20) is actually equivalent to the 1NN rule when applying the maximum operator. It should hence be recalled that POSSIBL, as employed in the above experiments, involves an adaptation of the (absolute) possibilistic support that comes from stored cases, which in essence is responsible for the differences.

A very interesting finding is the following: In the above examples, classification performance of the kNN algorithm is generally an increasing or a decreasing function of k. POSSIBL, on the other hand, performs very well irrespective of the direction of that tendency, i.e., regardless of whether a smaller or a larger neighborhood should be called in. This can be taken as an indication of the robustness of the possibilistic approach.

6.3. Statistical assumptions and robustness

Let us elaborate a little more closely on the aspect of robustness. Above, it has been claimed that the possibilistic approach is more robust than other methods against violations of statistical assumptions of independence (see end of Section 4.3.2). This is clearly true for the possibilistic estimation δ_{x_0} the informational content of which remains meaningful even if data is not independent. Here, we would like to provide experimental evidence for the supposition that the possibilistic approach can indeed be advantageous from both, an estimation and a decision making point of view, if the sample is not fully representative of the population.

The experimental setup is determined as follows: The instance space is defined by $\mathcal{X} = \mathbb{R}$, the set of labels is $\mathcal{L} = \{-1, +1\}$, the class probabilities are $1/2$, the conditional probability density of x given λ_x is normal with standard deviation 1 and mean λ_x. In a single simulation run, a random sample of size $n = 20$ is generated, using class-probabilities of $1/2 - \alpha$ and $1/2 + \alpha$, respectively $(0 < \alpha \leqslant 1/2)$. Based on the resulting training set, which is not “fully representative” in the sense of [17], predictions are derived for 10 new instances. These instances, however, are generated with the true class-probabilities of $1/2$. For a fixed value α and a fixed prediction method, a misclassification rate $r(\alpha)$ is derived by averaging over 10,000 simulation runs.

Fig. 4 shows the misclassification rates for several methods. As was to be expected, $r(\cdot)$ is an increasing function of the sample bias α. The best results are of course obtained if the class-probabilities of the training set and the test set coincide, that is for $\alpha = 0$. The figure also reveals that the sensitivity of the kNN classifier increases with k. On the one hand, it is true that a larger k leads to better results for α close to 0. On the other hand, the performance decreases more quickly than for smaller k, and $k = 1$ is to be preferred for α close to $1/2$. This finding can also be grasped intuitively: The larger k, the more the kNN rule relies on frequency information, and the more it is affected if this information is misleading.

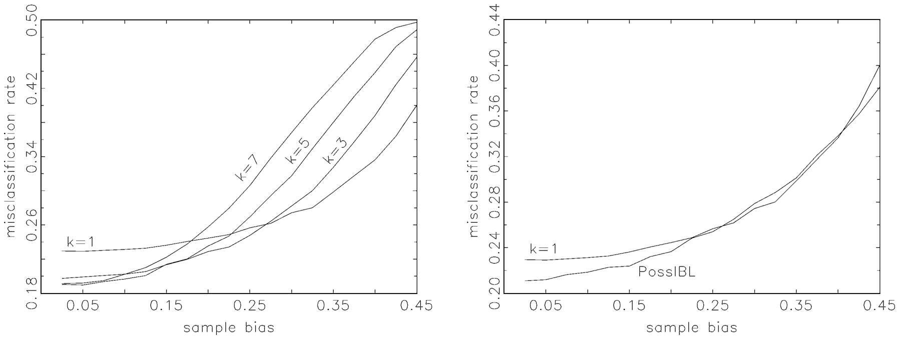


Fig. 4. Misclassification rates of kNN methods (left) and POSSIBL (right, in comparison with 1NN).

Apart from kNN methods, we have tested POSSIBL with $\oplus = \oplus_P$. The similarity measure $\sigma_{\mathcal{X}}$ was defined by the triangle $(x, y) \mapsto \max\{0, 1 - |x - y|/0.8\}$. Interestingly enough, this approach yields the most satisfactory results. For α close to 0 it is almost as good as the kNN rules with $k > 1$, and for α close to $1/2$ it equals the 1NN rule. Thus, the combination mode as realized by the probabilistic sum $(\alpha, \beta) \mapsto \alpha + \beta - \alpha\beta$ turns out to be reasonable under the conditions of this experiment. As already explained in Section 4.2, this operator produces a kind of saturation effect: It takes frequency information into account, but only to a limited extent (the larger the current support already is, the smaller the absolute increase due to a new observation). Thus, it is indeed in-between the 1NN rule and the kNN rules for $k > 1$. Intuitively, this explains our findings in the above experiment, especially that POSSIBL is more robust against the sample bias than kNN rules for $k > 1$.

6.4. *Variation of the aggregation operator*

An interesting question concerns the dependence of POSSIBL's performance on the specification of the aggregation operator $\oplus$ in (25). To get a first idea of this dependence, we have performed the same experiments as described in Section 6.2 above. Now, however, we have tested POSSIBL with different t-conorms.

More precisely, we have specified a t-conorm by means of the parameter ρ in (30), i.e., we have taken different aggregation operators from the Frank-family of t-conorms. POSSIBL was then applied to each data set with different operators $\oplus_\rho$. The simulation results are presented in Figs. 5–9. Each figure shows the average classification performance of POSSIBL (over 100 experiments) as a function of the parameter ρ. Please note the different scaling of the axes for the five data sets.

Confirming our previous considerations, the results show that in general different t-conorms are optimal for different applications. Still, POSSIBL's performance is quite robust toward the variation of the aggregation operator. That is, classification accuracy does not drop off too much when choosing a suboptimal operator.

A very interesting finding is the observation that the parameter $\rho = 0$ and, hence, the maximum operator is optimal if simultaneously the 1NN classifier performs well in comparison with other kNN classifiers. If this is not the case, as, e.g., for the BALANCE

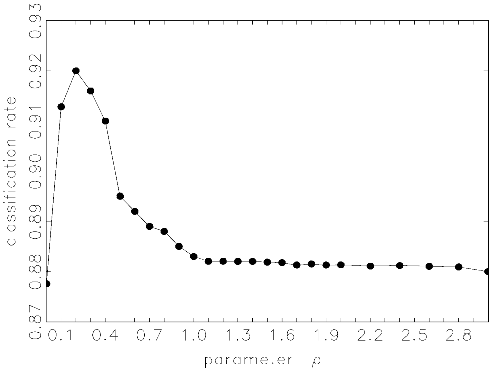


Fig. 5. Experimental results for the BALANCE SCALE data.

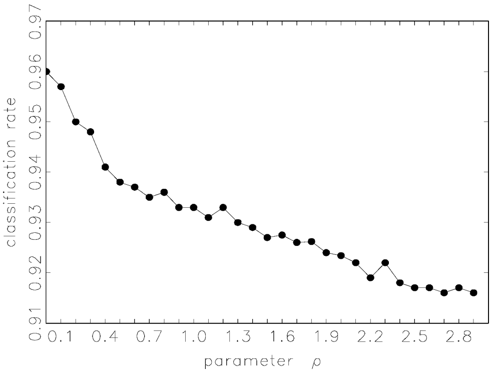


Fig. 6. Experimental results for the IRIS PLANT data.

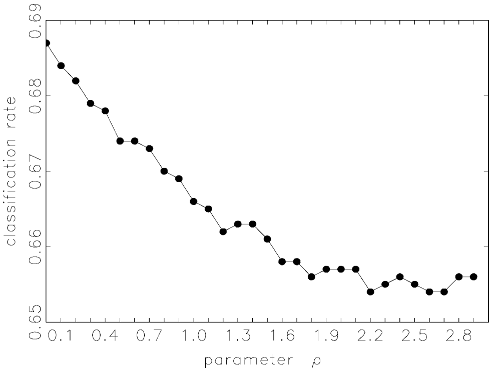


Fig. 7. Experimental results for the GLASS IDENTIFICATION data.

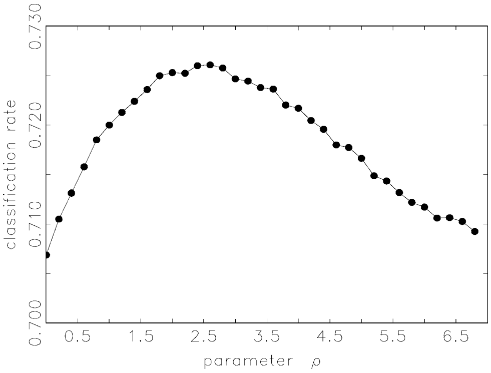


Fig. 8. Experimental results for the PIMA INDIAN DIABETES data.

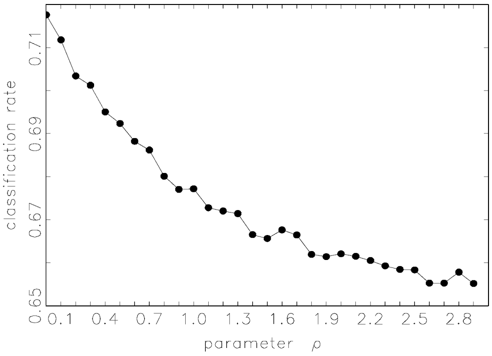


Fig. 9. Experimental results for the WINE RECOGNITION data.

SCALE and the PIMA INDIANS DIABETES data, parameters $\rho > 0$ achieve better results. This finding is not astonishing and can also be grasped intuitively. In fact, it was already mentioned that POSSIBL with $\oplus = \oplus_0 = \max$ is closely related to the 1NN classifier, as both methods do fully concentrate on the most relevant information. As opposed to this, aggregation operators $\oplus = \oplus_\rho$ with $\rho > 0$ combine the information from several neighbors in much the same way as do kNN classifiers with $k > 1$.

6.5. Representation of uncertainty

It was already mentioned that an important aspect of POSSIBL concerns the representation of uncertainty. The fact that POSSIBL can adequately represent the *ignorance* related to a decision problem is easily understood and does not call for empirical validation. To get a first idea of POSSIBL's ability to represent *ambiguity* we have derived approximations to two characteristic quantities, again using the experimental setup as described in Section 6.1.

Let D_1 denote the expected difference between the possibility degree of the predicted label $\lambda_{x_0}^{\mathrm{est}}$ and the possibility degree of the second best label, given that the prediction is correct:

$$D_1 \doteq \delta_{x_0}(\lambda_{x_0}) - \max_{\lambda \in \mathcal{L},\ \lambda \neq \lambda_{x_0}} \delta_{x_0}(\lambda).$$

Moreover, let D_0 denote the expected difference between the possibility degree of the predicted label $\lambda_{x_0}^{\mathrm{est}}$ and the possibility degree of the actually true label λ_{x_0}, given that $\lambda_{x_0} \neq \lambda_{x_0}^{\mathrm{est}}$:

$$D_0 \doteq \delta_{x_0}\big(\lambda_{x_0}^{\mathrm{est}}\big) - \delta_{x_0}(\lambda_{x_0}).$$

Ideally, D_0 is small and D_1 is large: Wrong decisions are accompanied by a large degree of uncertainty, as reflected by a comparatively large support of the actually correct label. As opposed to this, correct decisions appear reliable, as reflected by low possibility degrees assigned to all labels $\lambda \neq \lambda_{x_0}$.

Table 6

Database	D_0	D_1
BALANCE SCALE	0,094	0,529
IRIS PLANT	0,194	0,693
GLASS IDENTIFICATION	0,181	0,401
PIMA INDIANS DIABETES	0,211	0,492
WINE RECOGNITION	0,226	0,721

Table 6 shows approximations to the expected values D_0 and D_1, namely averages over 1,000 experiments. As can be seen, the reliability of a prediction is reflected very well by the possibilistic estimations.

7. Summary and future work

The idea underlying the method presented in this paper is to extend instance-based learning by concepts and techniques from possibility theory and fuzzy sets. Here, this idea has been realized in the form of a basic learning procedure called POSSIBL. Apart from discussing methodological aspects, the paper has started the investigation of theoretical properties of this approach (under standard statistical assumptions) and the validation of POSSIBL by means of experimental studies.

The application of possibility theory allows for realizing a graded version of the similarity-based extrapolation principle underlying IBL. Not only does this version appear very natural, it is also intuitively appealing. We have presented a detailed comparison of the possibilistic extrapolation principle and the commonly used approach which can be endowed with a probabilistic basis. Even though the two methods are based on quite different semantics, POSSIBL can formally be seen as an extension of the probabilistic approach. Indeed, it has been shown that the former—at least in its general form—can mimic the latter. Apart from that, the possibilistic approach has the following advantages:

- *Knowledge representation*: A possibilistic (instance-based) prediction is more expressive than a probabilistic one. Especially, the former is able to represent the *absolute* amount of evidential support as well as partial ignorance, a point which seems to be of major importance in IBL. Furthermore, the interpretation of aggregated degrees of individual support in terms of (guaranteed) possibility (degrees of confirmation) is generally less critical than the interpretation in terms of degrees of probability.
- *Scope for applications*: The possibilistic approach is more robust and extends the range of applications. Particularly, it makes no statistical assumptions about the generation of data and less mathematical assumptions about the structure of the underlying instance space. In fact, POSSIBL performs at least as well as standard NN techniques for typical (real-word) data sets. Beyond that, however, it can also be applied to data that violates certain statistical assumptions. Finally, the max–min version of POSSIBL can even be applied within a purely ordinal setting.
- *Support of extensions*: The possibilistic method is more flexible and supports several extensions of IBL. This includes the adaptation of aggregation modes in the

combination of individual degrees of support, the coherent handling of incomplete information, and the graded discounting of atypical cases. Moreover, it allows one to complement the similarity-based extrapolation principle by other inference procedures.

In the paper, we have outlined some extensions of the basic POSSIBL algorithm which deserve further investigation. This concerns particularly the ideas to automatically adapt a parameterized aggregation operator (Section 4.2.2) and to complement lower possibility bounds by means of upper bounds (Section 4.2), as well as the combination of instance-based and rule-based inference (Section 5.3). These extensions are important topics of ongoing research, which aims at realizing an efficient framework of *plausible instance-based learning* on the basis of possibility theory and fuzzy sets. In this regard, let us again mention the idea of supplementing IBL with fuzzy set-based modeling techniques. In fact, the methods in [27] allow for guiding and extending instance-based learning by means of domain knowledge and, thus, for combining knowledge and data in a flexible way. Parts of the possibilistic IBL framework have already been realized in connection with the PRETI project (Platform of Research and Experimentation in the Treatment of Information) maintained at the INSTITUT DE RECHERCHE EN INFORMATIQUE DE TOULOUSE.

Acknowledgements

The author wishes to thank three anonymous referees for useful comments which helped to improve the paper.

References

[1] D.W. Aha, Incremental, instance-based learning of independent and graded concept descriptions, in: Proc. 6th Internat. Workshop on Machine Learning, Ithaca, NY, Morgan Kaufmann, San Mateo, CA, 1989, pp. 387–391.

[2] D.W. Aha, Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms, Internat. J. Man-Machine Stud. 36 (1992) 267–287.

[3] D.W. Aha (Ed.), Lazy Learning, Kluwer Academic, Dordrecht, 1997.

[4] D.W. Aha, D. Kibler, M.K. Albert, Instance-based learning algorithms, Machine Learning 6 (1) (1991) 37–66.

[5] T. Bailey, A.K. Jain, A note on distance-weighted k-nearest neighbor rules, IEEE Trans. Systems Man Cybernet. 8 (4) (1978) 311–313.

[6] M. Béreau, B. Dubuisson, A fuzzy extended k-nearest neighbors rule, Fuzzy Sets and Systems 44 (1991) 17–32.

[7] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithm, Plenum Press, New York, 1981.

[8] J.C. Bezdek, K. Chuah, D. Leep, Generalized k-nearest neighbor rules, Fuzzy Sets and Systems 18 (1986) 237–256.

[9] R. Bradley, N. Swartz, Possible Worlds, Basil Blackwell, Oxford, UK, 1979.

[10] C.E. Brodley, Addressing the selective superiority problem: Automatic algorithm for model class selection, in: Proc. 10th Machine Learning Conference, 1993, pp. 17–24.

[11] C.L. Chang, Finding prototypes for nearest neighbor classifiers, IEEE Trans. Comput. 23 (11) (1974) 1179–1184.

[12] C.K. Chow, On optimum recognition error and reject tradeoff, IEEE Trans. Inform. Theory 16 (1970) 41–46.

[13] L.J. Cohen, An Introduction to the Philosophy of Induction and Probability, Clarendon Press, Oxford, 1989.

[14] T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, IEEE Trans. Inform. Theory 13 (1967) 21–27.

[15] B.V. Dasarathy, Nosing around the neighborhood: A new system structure and classification rule for recognition in partially exposed environments, IEEE Trans. Pattern Analysis and Machine Intelligence 2 (1) (1980) 67–71.

[16] B.V. Dasarathy (Ed.), Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques, IEEE Computer Society Press, Los Alamitos, CA, 1991.

[17] E.R. Davies, Training sets and a priori probabilities with the nearest neighbor method of pattern classification, Pattern Recognition Lett. 8 (1) (1988) 11–13.

[18] R. Lopez de Mantaras, E. Armengol, Machine learning from examples: Inductive and lazy methods, Data Knowledge Engrg. 25 (1998) 99–123.

[19] T. Denoeux, A k-nearest neighbor classification rule based on Dempster–Shafer Theory, IEEE Trans. Systems Man Cybernet. 25 (5) (1995) 804–813.

[20] J.K. Dixon, Pattern recognition with partly missing data, IEEE Trans. Systems Man Cybernet. 9 (10) (1979) 617–621.

[21] P. Domingos, Rule induction and instance-based learning: A unified approach, in: C.S. Mellish (Ed.), Proc. IJCAI-95, Montreal, Quebec, Morgan Kaufmann, San Mateo, CA, 1995, pp. 1226–1232.

[22] P. Domingos, Unifying instance-based and rule-based induction, Machine Learning 24 (1996) 141–168.

[23] D. Dubois, F. Esteva, P. Garcia, L. Godo, R. Lopez de Mantaras, H. Prade, Fuzzy set modelling in case-based reasoning, Internat. J. Intelligent Syst. 13 (1998) 345–373.

[24] D. Dubois, P. Hajek, H. Prade, Knowledge driven vs. data driven logics, J. Logic Language Inform. 9 (2000) 65–89.

[25] D. Dubois, E. Hüllermeier, H. Prade, Flexible control of case-based prediction in the framework of possibility theory, in: E. Blanzieri, L. Portinale (Eds.), Advances in Case-Based Reasoning, Proc. EWCBR-2000, 5th European Workshop on Case-Based Reasoning, Trento, Italy, Springer, Berlin, 2000, pp. 61–73.

[26] D. Dubois, E. Hüllermeier, H. Prade, Formalizing case-based inference using fuzzy rules, in: S.K. Pal, D.Y. So, T. Dillon (Eds.), Soft Computing in Case-Based Reasoning, Springer, Berlin, 2000, pp. 47–72.

[27] D. Dubois, E. Hüllermeier, H. Prade, Fuzzy set-based methods in instance-based reasoning, IEEE Trans. Fuzzy Systems 10 (3) (2002) 322–332.

[28] D. Dubois, H. Prade, Fuzzy sets and statistical data, European J. Oper. Res. 25 (1986) 345–356.

[29] D. Dubois, H. Prade, Possibility Theory, Plenum Press, New York, 1988.

[30] D. Dubois, H. Prade, Possibility theory as a basis for preference propagation in automated reasoning, in: Proc. 1st IEEE Internat. Conference on Fuzzy Systems (FUZZ-IEEE-92), San Diego, CA, 1992, pp. 821–832.

[31] D. Dubois, H. Prade, When upper probabilities are possibility measures, Fuzzy Sets and Systems 49 (1992) 65–74.

[32] D. Dubois, H. Prade, What are fuzzy rules and how to use them, Fuzzy Sets and Systems 84 (1996) 169–185.

[33] D. Dubois, H. Prade, Possibility theory: Qualitative and quantitative aspects, in: D.M. Gabbay, P. Smets (Eds.), Handbook of Defeasible Reasoning and Uncertainty Management Systems, Vol. 1, Kluwer Academic, Dordrecht, 1998, pp. 169–226.

[34] D. Dubois, H. Prade, P. Smets, Not impossible vs. guaranteed possible in fusion and revision, in: Proc. ESCQARU-2001, Toulouse, France, in: Lecture Notes in Comput. Sci, Vol. 2143, Springer, Berlin, 2001, pp. 522–531.

[35] D. Dubois, H. Prade, L. Ughetto, A new perspective on reasoning with fuzzy rules, in: N.R. Pal, M. Sugeno (Eds.), Advances in Soft Computing, Proc. AFSS International Conference on Fuzzy Systems, Calcutta, India, in: Lecture Notes in Artificial Intelligence, Vol. 2275, Springer, Berlin, 2002, pp. 1–11.

[36] B. Dubuisson, M. Masson, A statistical decision rule with incomplete knowledge about classes, Pattern Recognition 26 (1) (1993) 155–165.

[37] S.A. Dudani, The distance-weighted k-nearest-neighbor rule, IEEE Trans. Systems Man Cybernet. 6 (4) (1976) 325–327.

[38] E. Fix, J.L. Hodges, Discriminatory analysis: nonparametric discrimination: consistency principles, in: B.V. Dasarathy (Ed.), Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques, IEEE Computer Society Press, Los Alamitos, CA, 1991. Reprint of original work from 1951.

[39] E. Fix, J.L. Hodges, Discriminatory analysis: Nonparametric discrimination: Small sample performance, in: B.V. Dasarathy (Ed.), Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques, IEEE Computer Society Press, Los Alamitos, CA, 1991. Reprint of original work from 1952.

[40] J.H. Friedman, F. Baskett, L.J. Shustek, An algorithm for finding nearest neighbors, IEEE Trans. Comput. 24 (1975) 1000–1006.

[41] P.E. Hart, The condensed nearest neighbor rule, IEEE Trans. Inform. Theory 14 (1968) 515–516.

[42] M.E. Hellman, The nearest neighbor classification rule with a reject option, IEEE Trans. Systems Man Cybernet. 6 (1970) 179–185.

[43] E. Hüllermeier, Toward a probabilistic formalization of case-based inference, in: T. Dean (Ed.), Proc. IJCAI-99, Stockholm, Sweden, Morgan Kaufmann, San Mateo, CA, 1999, pp. 248–253.

[44] E. Hüllermeier, On the representation and combination of evidence in instance-based learning, in: Proc. ECAI-2002, 15th European Conference on Artificial Intelligence, Lyon, France, IOS Press, Amsterdam, 2002, pp. 360–364.

[45] D. Hume, An Enquiry concerning Human Understanding, Oxford University Press, New York, 1999.

[46] A. Józwik, A learning scheme for a fuzzy k-NN rule, Pattern Recognition Lett. 1 (1983) 287–289.

[47] J.M. Keller, M.R. Gray, J.A. Givens, A fuzzy k-nearest neighbor algorithm, IEEE Trans. Systems Man Cybernet. 15 (4) (1985) 580–585.

[48] D. Kibler, D.W. Aha, Instance-based prediction of real-valued attributes, Comput. Intelligence 5 (1989) 51–57.

[49] B.S. Kim, S.B. Park, A fast *k* nearest neighbor finding algorithm based on the ordered partition, IEEE Trans. Pattern Analysis and Machine Intelligence 8 (6) (1985) 761–766.

[50] J.L. Kolodner, Case-based Reasoning, Morgan Kaufmann, San Mateo, CA, 1993.

[51] R. Krishnapuram, J.M. Keller, A possibilistic approach to clustering, IEEE Trans. Fuzzy Systems 1 (2) (1993) 98–110.

[52] D.K. Lewis, Counterfactuals and comparative possibility, J. Philos. Logic 2 (1973).

[53] D.O. Loftsgaarden, C.P. Quesenberry, A nonparametric estimate of a multivariate density function, Ann. Math. Stat. 36 (1965) 1049–1051.

[54] J. Macleod, A. Lik, D. Titterington, A re-examination of the distance-weighted k-nearest neighbor classification rule, IEEE Trans. Systems Man Cybernet. 17 (4) (1987) 689–696.

[55] E. McKenna, B. Smyth, Competence-guided edition methods for lazy learning, in: Proc. 14th European Conference on Artificial Intelligence (ECAI-2000), Berlin, 2000, pp. 60–64.

[56] T.M. Mitchell, The need for biases in learning generalizations, Technical Report TR CBM-TR-117, Rutgers University, New Brunswick, NJ, 1980.

[57] I. Niiniluoto, Analogy and similarity in scientific reasoning, in: D.H. Helman (Ed.), Analogical Reasoning, Kluwer Academic, Dordrecht, 1988, pp. 271–298.

[58] G. Parthasarathy, B.N. Chatterji, A class of new KNN methods for low sample problems, IEEE Trans. Systems Man Cybernet. 20 (3) (1990) 715–718.

[59] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Statist. 33 (1962) 1065–1076.

[60] E.A. Patrick, F.P. Fischer, A generalized k-nearest neighbor rule, Inform. and Control 16 (2) (1970) 128–152.

[61] J.R. Quinlan, Unknown attribute values in induction, in: Proc. 6th International Workshop on Machine Learning, Morgan Kaufmann, San Mateo, CA, 1989, pp. 164–168.

[62] R. Quinlan, Combining instance-based and model-based learning, in: Proc. 10th International Conference of Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993, pp. 236–243.

[63] M. Rosenblatt, Remarks on some nonparametric estimates of a density function, Ann. Math. Statist. 27 (1956) 832–837.

[64] S. Salzberg, A nearest hyperrectangle learning method, Machine Learning 6 (1991) 251–276.

[65] E. Sanchez, On possibility qualification in natural languages, Inform. Sci. 15 (1978) 45–76.

[66] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, Princeton, NJ, 1976.

[67] D. Shepard, A two-dimensional interpolation function for irregularly spaced data, in: Proc. 23rd National Conference of the ACM, 1968, pp. 517–523.

[68] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, London, 1986.

[69] P. Smets, R. Kennes, The transferable belief model, Artificial Intelligence 66 (1994) 191–234.

[70] C. Stanfill, D. Waltz, Toward memory-based reasoning, Comm. ACM (1986) 1213–1228.

[71] M. Tan, Cost-sensitive learning of classification knowledge and its application to robotics, Machine Learning 13 (7) (1993) 7–34.

[72] I. Tomek, A generalization of the k-NN rule, IEEE Trans. Systems Man Cybernet. 6 (1976) 121–126.

[73] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[74] M.P. Wand, M.C. Jones, Kernel Smoothing, Chapman and Hall, London, 1995.

[75] J. Weisbrod, A new approach to fuzzy reasoning, Soft Comput. 2 (1998) 89–99.

[76] S. Wess, K.D. Althoff, G. Derwand, Using k-d trees to improve the retrieval step in case-based reasoning, in: S. Wess, K.D. Althoff, M.M. Richter (Eds.), Topics in Case-Based Reasoning, Springer, Berlin, 1994, pp. 167–181.

[77] D. Wettschereck, D.W. Aha, T. Mohri, A review and empirical comparison of feature weighting methods for a class of lazy learning algorithms, AI Rev. 11 (1997) 273–314.

[78] D.L. Wilson, Asymptotic properties of nearest neighbor rules using edited data, IEEE Trans. Systems Man Cybernet. 2 (3) (1972) 408–421.

[79] D.R. Wilson, Advances in instance-based learning algorithms, PhD Thesis, Department of Computer Science, Brigham Young University, Provo, UT, 1997.

[80] D.R. Wilson, T.R. Martinez, Improved heterogeneous distance functions, J. Artificial Intelligence Res. 6 (1997) 1–34.

[81] T.P. Yunck, A technique to identify nearest neighbors, IEEE Trans. Systems Man Cybernet. 6 (10) (1976) 678–683.

[82] L.A. Zadeh, Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets and Systems 1 (1978) 3–28.

[83] L.A. Zadeh, PRUF: A meaning representation language for natural language, Internat. J. Man-Machine Stud. 10 (1978) 395–460.

[84] L.A. Zadeh, Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets and Systems 90 (2) (1997) 111–127.

[85] J. Zhang, Selecting typical instances in instance-based learning, in: Proc. 9th International Conference on Machine Learning (ICML-92), Aberdeen, Scotland, 1992, pp. 470–479.

[86] J. Zhang, Y. Yim, J. Yang, Intelligent selection of instances for prediction in lazy learning algorithms, Artificial Intelligence Rev. 11 (1997) 175–191.