5. Fuzzy Set-Based Modeling of Case-Based Inference I

A close connection between fuzzy set-based (approximate reasoning) methods and the inference principle underlying similarity-based (case-based) reasoning has been pointed out recently [99, 407]. Besides, some attempts at combining case-based reasoning (or, more generally, analogical reasoning) and methods from fuzzy set theory have already been made [408], including the use of fuzzy sets for supporting the computation of similarities of situations in analogical reasoning [144], the formalization of aspects of analogical reasoning by means of similarity relations between fuzzy sets [48], the use of fuzzy set theory in case indexing and retrieval [209, 214], the case-based learning of fuzzy concepts from fuzzy examples [295], the use of fuzzy predicates in the derivation of similarities [40], and the integration of case-based and rule-based reasoning [138]. See [45, 49] for a more general framework of analogical reasoning.

This chapter continues this promising line of research. It is argued that fuzzy rules in conjunction with associated inference procedures provide a convenient framework for modeling the CBI hypothesis and for supporting the task of case-based inference as outlined in Section 2.4.

The remaining part of the chapter is organized as follows: Even though we assume the reader to be familiar with basics of fuzzy set theory, we recall the most important concepts from possibility theory in Section 5.1. The basic CBI framework we proceed from and the key idea of fuzzy rule-based modeling of the CBI hypothesis are introduced in Section 5.2. Diverse types of extensions of the basic model will then be discussed in Sections 5.3 and 5.4. Section 5.5 presents some experimental studies in the field of classification. The idea of calibrating a CBI model by combining qualitative modeling techniques with data-driven optimization methods is addressed in Section 5.6. Finally, some connections between the approach introduced in this chapter and related approaches in the field of fuzzy set theory are discussed in Section 5.7.

5.1 Background on possibility theory

In this section, we recall some basic concepts from possibility theory, as far as required for the current chapter. 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 [416]. However, as concerns the use in most applications, and in this book in particular, possibility theory is considered as a means for representing uncertain knowledge, that is, 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" [251, 127]. 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 [122]. 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 *certainty* and *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 [131] and [104].

5.1.1 Possibility distributions as generalized constraints

A key idea of possibility theory as originally introduced by Zadeh [416] 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 $\Pi_{\mathcal{K}}(A) \leq p$, 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*:[1] 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 $\Omega \longrightarrow P$ mapping 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:

$$\Pi_{\mathcal{K}}(A \cup B) = \max\left\{\Pi_{\mathcal{K}}(A),\, \Pi_{\mathcal{K}}(B)\right\}. \tag{5.1}$$

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:

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

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

$$\mathcal{N}_{\mathcal{K}}(A) = 1 - \Pi_{\mathcal{K}}(\Omega \setminus A) \tag{5.2}$$

for all $A \subseteq \Omega$:[2] An event A is necessary in so far as its complement (logical negation) is not possible.

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

[2] If the possibility scale P is not the unit interval $[0, 1]$, the mapping $1 - (\cdot)$ on the right-hand side of (5.2) is 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 [416] 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 $\pi_{\mathcal{K}}(\omega \,|\, F) = \mu(F \,|\, \omega)$: 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 simply 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 ($\mu_F(10) = 0$).

5.1.2 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 $\delta(\omega)$,[3] 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 [108]. As examples do express positive evidence, they do never change the distribution $\pi \equiv 1$. This distribution would only be changed if

[3] In [393], this type of distribution is called σ-distribution.

we *knew* from some other information source, e.g., that V can only take values $v \geq 6$, in which case $\pi(v) = 1$ for $v \geq 6$ and $\pi(v) = 0$ for $v \leq 5$.

The difference between modeling positive and negative evidence becomes especially clear when it comes to expressing complete ignorance. As already mentioned above, this 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 $\delta_\phi(\omega) \geq d$. 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

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

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 $\pi' \leq \pi$. 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 [121], 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.[4] For the measure Δ, the characteristic property (5.1) becomes

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

5.2 Fuzzy rule-based modeling of the CBI hypothesis

Rule-based modeling plays an important role in fuzzy systems research and will also turn out to be useful in the context of case-based inference. Fuzzy rules provide a local, rough and soft specification of the relation between variables X

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

and Y ranging on domains D_X and D_Y, respectively [124]. They are generally expressed in the form "if X is A then Y is B," where A and B are fuzzy sets associated with symbolic labels and modeled by means of membership functions on D_X resp. D_Y.[5]

There are several aspects which motivate the use of fuzzy rules in connection with case-based reasoning [100, 205]. Firstly, the CBI hypothesis itself corresponds to an *if-then* rule: "If two inputs are similar, then the associated outcomes are similar as well." Secondly, the notion of *similarity*, which lies at the heart of case-based reasoning, is also strongly related to the theory of fuzzy sets. Indeed, one of the main interpretations of the membership function of a fuzzy set is that of a similarity relation, i.e., degrees of membership can be thought of as degrees of similarity [126]. Thirdly, linked with the framework of possibility theory, fuzzy sets provide a tool for the modeling and processing of *uncertainty*. In connection with the *heuristic* character of CBR, this aspect seems to be of special importance. As already mentioned in Chapter 1, the CBI principle should not be understood as a deterministic rule. Within the context of fuzzy rules considered in this chapter, it will rather be interpreted in the following sense: "If two inputs are similar, it is *possible* that the associated outcomes are similar as well."

At a formal level, fuzzy rules can be modeled as possibility distributions constrained by a combination of the membership functions which define the antecedent and consequent part of the rule, where the concrete form of the constraint depends on the interpretation of the rule [124]. This way, they relate the concepts of *similarity* and *uncertainty*, thus providing the basis for methods of uncertain similarity-based inference. This is the main reason for their convenience as formal models of the CBI hypothesis

5.2.1 Possibility rules

The aforementioned interpretation of the CBI hypothesis is nicely captured by means of a so-called *possibility rule*, a special type of conjunction-based fuzzy rule. A possibility rule involving fuzzy sets A and B, subsequently symbolized by $A \rightarrow B$, corresponds to the statement that "the more X is A, the more *possibly* B is a range for Y." More precisely, it can be interpreted as a collection of rules "if $X = x$, it is possible at least to the degree $A(x)$ that B is a range for Y." The intended meaning of this kind of *possibility-qualifying* rule is captured by the following constraint which guarantees a certain lower bound to the possibility $\delta(x, y)$ that the tuple (x, y) is an admissible instantiation of the variables (X, Y):

$$\delta(x, y) \geq \min\{\, A(x),\, B(y) \,\}. \tag{5.3}$$

As suggested by the rule-based modeling of the relation between X and Y, these variables often play the role of an input and an output, respectively, and one

[5] We shall usually use the same notation for a label, the name of an associated fuzzy set, and the membership function of this set. Thus, $A(x)$ is the degree of membership of the element x in the fuzzy set A.

is interested in possible values of Y while X is assumed to be given. By letting $\delta(y \,|\, x) \stackrel{\text{df}}{=} \delta(x, y)$, the constraint (5.3) can also be considered as a lower bound to a *conditional* possibility distribution. That is, given the value $X = x$, the possibility that $Y = y$ is lower-bounded by $\delta(x, y)$ according to (5.3). Observe that *nothing* is said about Y in the case where $A(x) = 0$ since we then obtain the trivial constraint $\pi(y \,|\, x) \geq 0$. Besides, it should be noticed that the lower bound-interpretation is also consistent with conditional distributions $\delta(\cdot \,|\, x)$ which are not normalized, i.e., for which $\sup_y \delta(y \,|\, x) < 1$ (cf. Section 5.1).

5.2.2 Modeling the CBI hypothesis

The basic framework we shall proceed from in this chapter is a special type of generalized non-deterministic CBI setup (see Definition 2.7 and Remark 2.8 in Section 2.4.2). As in Chapters 3 and 4, a case c is a tuple $\langle s, r\rangle \in \mathcal{C} = \mathcal{S} \times \mathcal{R}$ consisting of an input $s \in \mathcal{S}$ and an associated output $r \in \mathcal{R}$. However, we do no longer assume that an input determines a unique outcome, i.e., cases $c = \langle s, r\rangle$ and $c' = \langle s', r'\rangle$ such that $s = s'$ but $r \neq r'$ might be encountered. In fact, the assumption of a functional relation $\varphi : \mathcal{S} \longrightarrow \mathcal{R}$ mapping inputs to unique outcomes would be too restrictive for the type of applications we have in mind in connection with the possibilistic approach. Rather, φ is now defined as a relation

$$\varphi \subseteq \mathcal{S} \times \mathcal{R} \tag{5.4}$$

and corresponds to a set of potential observations, i.e., existing (but perhaps not yet encountered) cases. As before, we assume data to be given in the form of a memory

$$\mathcal{M} = \big\{\langle s_1, r_1\rangle, \langle s_2, r_2\rangle, \ldots, \langle s_n, r_n\rangle\big\}$$

of observed cases. As an aside, note that $\mathcal{M}$ was formally treated as a sequence rather than a set in Chapters 3 and 4. This is not necessary within the possibilistic framework of this section. Moreover, we can abandon the assumption that $\mathcal{S}$ and $\mathcal{R}$ are countable.

As before, our focus is on case-based inference: Given a new input $s_0 \in \mathcal{S}$, the task is to predict the outcome $r_0 \in \mathcal{R}$ associated with s_0. This actually comes down to predicting the set $\{r \in \mathcal{R} \,|\, \langle s_0, r\rangle \in \varphi\}$ of potential outcomes, since we do no longer assume uniqueness. To this end, we shall derive a quantification of the *possibility* that $r_0 = r$, i.e., $\langle s_0, r\rangle \in \varphi$, for each outcome $r \in \mathcal{R}$. As will be seen in the remainder of this chapter, this kind of prediction makes the formulation of rather general types of queries possible, especially if s_0 is allowed to be incompletely specified.

The basic idea of the approach discussed in this chapter is to use a possibility rule as defined above in order to formalize the CBI hypothesis. In fact, interpreting the variables X and Y as degrees of similarity between two inputs and two outputs, respectively, and A and B as fuzzy sets of "large similarity degrees" (with strictly

increasing membership functions) amounts to expressing the following version of the CBI hypothesis: "The more similar two inputs are, the more *possible* it is that the corresponding outcomes are similar" [99]. In the same way as the probabilistic model of Chapter 4, this formalization takes the heuristic nature of the CBI hypothesis into account. In fact, it does not impose a deterministic constraint, but only concludes on the *possibility* of the outcomes to be similar.

In the sense of the above principle, an observed case $\langle s_1, r_1 \rangle \in \mathcal{M}$ is taken as a piece of evidence which qualifies similar (hypothetical) cases $\langle s, r \rangle$ as being possible. According to (5.3) it induces lower bounds[6]

$$\delta(s, r) \geq \min \left\{ \sigma_{\mathcal{S}}(s, s_1), \sigma_{\mathcal{R}}(r, r_1) \right\} \tag{5.5}$$

to the possibility that $\langle s, r \rangle \in \varphi$. This can be interpreted as a similarity-based *extrapolation* of case-based information: The observation $\langle s_1, r_1 \rangle$ is considered as a typical case or, say, prototype, which is extrapolated in accordance with the CBI hypothesis. The more similar $\langle s, r \rangle$ and $\langle s_1, r_1 \rangle$ are in the sense of the (joint) similarity measure

$$\sigma_{\mathcal{C}} : \big(\langle s, r \rangle, \langle s', r' \rangle \big) \mapsto \min \left\{ \sigma_{\mathcal{S}}(s, s'), \sigma_{\mathcal{R}}(r, r') \right\}, \tag{5.6}$$

the more plausible becomes the (hypothetical) case $\langle s, r \rangle$ and, hence, the larger is the (lower) possibility bound (5.5). In other words, a high degree of possibility is assigned to a hypothetical case as soon as the *existence* of a very similar case is guaranteed (by observation).

Applying (5.5) to all cases in the memory $\mathcal{M}$ we obtain the possibility distribution $\delta_{\mathcal{C}}$ defined by

$$\delta_{\mathcal{C}}(s, r) = \max_{1 \leq \imath \leq n} \min \left\{ \sigma_{\mathcal{S}}(s, s_\imath), \sigma_{\mathcal{R}}(r, r_\imath) \right\} \tag{5.7}$$

for all $c = \langle s, r \rangle \in \mathcal{S} \times \mathcal{R}$. This distribution can be interpreted as a possibilistic approximation of the relation φ in (5.4). It is of provisional nature and actually represents lower bounds to possibility degrees (the equality in (5.7) is justified by the principle of *maximal informativeness*, see Section 5.1.2). In fact, the degree of possibility assigned to a case c may increase when gathering further evidence by observing new sample cases, as reflected by the application of the maximum operator in (5.7).

Observe that similarity degrees (on the right-hand side) are turned into possibility degrees (on the left-hand side) by virtue of the functional relation (5.7). In fact, the latter reveals at a formal level that – according to our formalization – similarity is in direct correspondence with possibility: From the similarity of a case $\langle s, r \rangle$ to an observed case, (5.7) concludes on the possibility of this case itself.

The distribution (5.7) can be taken as a point of departure for various inference tasks. In particular, given a new input s_0, a prediction of the associated outcome r_0 is obtained in the form of the conditional distribution δ_{s_0} defined by

[6] Without loss of generality, we assume the membership functions of the fuzzy sets of "large similarity degrees" to be given by the identical function $\mathrm{id} : x \mapsto x$ on $[0, 1]$.

$$\delta_{s_0}(r) = \delta(r \mid s_0) \stackrel{\text{df}}{=} \max_{1 \leq \imath \leq n} \min \left\{ \sigma_{\mathcal{S}}(s_0, s_\imath),\, \sigma_{\mathcal{R}}(r, r_\imath) \right\}, \tag{5.8}$$

for all $r \in \mathcal{R}$, where $\delta_{s_0}(r)$ denotes the (estimated) *possibility* of the output r, i.e., the possibility that r corresponds to the true outcome r_0.

Example 5.1. The (real-world) Automobile Database[7] contains 205 cars, each of which is characterized by 26 attributes. Thus, let a case correspond to a car which is characterized by means of an attribute–value representation including properties, such as its horsepower and fuel-type. For the sake of simplicity, we shall consider only some of the attributes available, i.e., the memory $\mathcal{M}$ is actually a projection of the complete database. One of the attributes, namely the price of a car, has been chosen as the outcome associated with a case. The latter is hence a tuple $\langle s, r \rangle$, where the input $s = (a_1, \ldots, a_L)$ is a vector of attribute values describing a car, and r is the associated price. The similarity between two cars s and s' is defined as a combination of the similarities between the respective attribute values $a_\jmath$ and $a'_\jmath$ $(1 \leq \jmath \leq L)$.

To illustrate, suppose a car to be characterized by only one attribute, namely its horsepower. Thus, the CBI hypothesis should simply be understood in the sense that "cars with similar horsepowers (possibly) have similar prices." Let $\sigma_{\mathcal{S}}(s, s') = \sigma_{hp}(s, s') = \max\{1 - |s - s'|/100, 0\}$. Likewise, let the similarity between two outcomes (= prices) be given by $\sigma_{\mathcal{R}}(r, r') = \max\{1 - |r - r'|/10000, 0\}$. Fig. 5.1 shows the prediction (5.8) for $s_0 = 100$. This prediction corresponds to the "more or less" possible range of prices for the class of cars whose horsepower is 100. As can be seen, the evidence contained in the memory $\mathcal{M}$ of cases strongly supports prices between \$10,000 and \$17,000. At the same time, however, it does not completely rule out prices which are slightly lower or higher. □

The possibility distribution $\boldsymbol{\delta_{s_0}}$. According to (5.8), r is regarded as a possible output if there is a case $\langle s_\imath, r_\imath \rangle$ such that both, $s_\imath$ is close to s_0 and $r_\imath$ is close to r. Or, if we define the *joint similarity* between the case $\langle s_\imath, r_\imath \rangle$ and the (hypothetical) case $\langle s_0, r \rangle$ according to (5.6), this can be expressed by saying that the case $\langle s_0, r \rangle$ is regarded as possible if the existence of a similar case $\langle s_\imath, r_\imath \rangle$ is confirmed by observation. In other words, a similar case provides evidence for the existence of $\langle s_0, r \rangle$ in the sense of *possibility qualification*.[8]

Following the notational convention of Section 5.1, possibility degrees $\delta_{s_0}(r)$ 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 5.1.2. More specifically, the distribution $\delta_{s_0} : \mathcal{R} \longrightarrow [0, 1]$ is thought of as a *lower* rather than an upper bound. Particularly, $\delta_{s_0}(r) = 0$ must

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

[8] The idea of possibility qualification, already mentioned in Section 5.1, is usually considered in connection with natural language propositions [328, 417]. Here, possibility qualification is casuistic rather than linguistic.

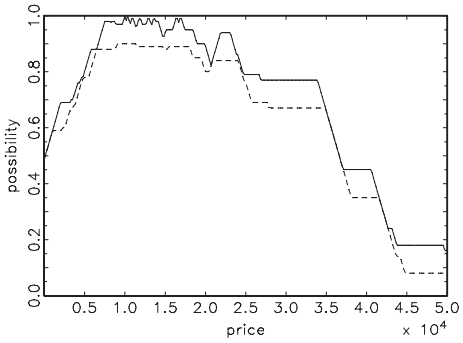


Fig. 5.1. Prediction (5.8) of the price of a car with horsepower $s_0 = 100$ (solid line) and prediction (5.32) for $90 \leq s \leq 110$.

not be equated with the impossibility of $r_0 = r$ but merely means that no evidence supporting the outcome r is available so far! In fact, δ_{s_0} is of provisional nature, and the degree of possibility assigned to an outcome r may increase when gathering further evidence by observing new cases, as reflected by the application of the maximum operator in (5.8). These remarks also make clear that the distribution δ_{s_0} is not necessarily normalized (in the sense that $\sup_r \delta_{s_0}(r) = 1$). In this connection, also note that there is not necessarily a unique actual world in the sense of the possible worlds semantics [51]. Since s_0 is not assumed to have a unique output, δ_{s_0} rather provides information about the set $\{r \in \mathcal{R} \,|\, \langle s_0, r\rangle \in \varphi\}$ of potential outcomes. Thus, the state of "complete knowledge" corresponds to the distribution δ_{s_0} with $\delta_{s_0}(r) = 1$ if $\langle s_0, r\rangle \in \varphi$ and $\delta_{s_0}(r) = 0$ otherwise.

In a classification context, where the outcomes r are class labels (i.e., $\mathcal{R}$ is a finite number of classes), the set of all inputs $s \in \mathcal{S}$ with the same output is sometime referred to as a *concept*. When being applied to all $s \in \mathcal{S}$, (5.8) yields "fuzzy" concept descriptions, that is possibilistic approximations of the concepts C_r $(r \in \mathcal{R})$:

$$C_r^{est} = \{(s, \delta_s(r)) \,|\, s \in \mathcal{S}\}, \tag{5.9}$$

where $\delta_s(r)$ is the degree of membership of $s \in \mathcal{S}$ in the fuzzy concept C_r^{est}, i.e., $C_r^{est}(s) = \delta_s(r)$. Note that these fuzzy concepts can overlap in the sense that $\min\{C_r^{est}(s), C_{r'}^{est}(s)\} > 0$ for $r \neq r'$ and $s \in \mathcal{S}$ (s has a positive degree of membership in two concepts C_r^{est} and $C_{r'}^{est}$, $r \neq r'$).[9]

The similarity measures $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$. Let us make some remarks on the similarity measures $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$. As mentioned previously, according to (5.8), the

[9] In practice, fuzzy and/or overlapping concepts seem to be the rule rather than the exception [3].

similarity of cases is in direct correspondence with the *possibility* assigned to an outcome. Roughly speaking, the principle expressed by (the fuzzy rule underlying) equation (5.8) gives rise to turn similarity into possibilistic support. Consequently, $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$ 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 NN estimation (cf. Section 2.2.1). Particularly, $\sigma_{\mathcal{S}}(s_0, s_\imath) = 0$ means that the $\imath$-th case is not considered as a relevant piece of information since it is not sufficiently similar to s_0. For computation, irrelevant cases in (5.8) can clearly be left out of account. Thus, it is enough to consider cases in a certain region around s_0. As opposed to the kNN approach, it is the size of this region rather than the number of neighboring cases which is fixed.

As in previous chapters, we assume $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$ to be reflexive and symmetric, whereas no special kind of transitivity is required.[10] In fact, the application of the maximum operator in (5.8) 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

$$\{\text{completely different}, \ldots, \text{very similar}, \text{identical}\}. \tag{5.10}$$

Correspondingly, degrees of possibility are interpreted in a qualitative way [251, 127]. That is, $\delta_{s_0}(r) < \delta_{s_0}(r')$ only means that outcome r is less supported than outcome r'; apart from that, the difference between the possibility degrees has no meaning.

Needless to say, a scale such as (5.10) is more convenient if cases 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 (cf. Section 2.2.1), 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 (5.8) from the point of view of observed cases, this estimation principle defines a (possibilistic) *extrapolation* of each case $\langle s_\imath, r_\imath \rangle$. In the original NN approach, which does not involve a distance measure on $\mathcal{R}$, a case $\langle s_\imath, r_\imath \rangle \in \mathcal{M}$ can only support the output $r_\imath$. This corresponds to the special case where $\sigma_{\mathcal{R}}$ in (5.8) is given by

[10] Let us mention again that relations satisfying reflexivity and symmetry are often called *proximity relations* in the fuzzy set literature, where similarity relations are defined as transitive proximity relations [100]. Anyway, we shall use the term similarity relation (similarity measure) henceforth without assuming transitivity.

$$\sigma_{\mathcal{R}}(r, r') = \begin{cases} 1 & \text{if} \quad r = r' \\ 0 & \text{if} \quad r \neq r' \end{cases}, \tag{5.11}$$

which is reasonable if $\mathcal{R}$ is a nominal scale, as, e.g., in concept learning.

By allowing for graded distances between outcomes, the possibilistic approach provides for a case $\langle s_\imath, r_\imath \rangle$ to support similar outcomes as well. This type of extended extrapolation is reasonable if $\mathcal{R}$ is a cardinal or at least ordinal scale. In fact, it should be observed that (5.8) 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 (= ouput) 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 [318, 289]. Indeed, (5.8) can be considered as a possibilistic counterpart of kernel-based density estimation. Let us furthermore mention that the consideration of graded distances between outputs is also related to the idea of class-dependent misclassification costs [290, 364].

5.3 Generalized possibilistic prediction

The possibility distribution δ_{s_0}, which specifies the fuzzy set of well-supported outputs, is a disjunctive combination of the individual support functions

$$\delta^\imath_{s_0} : r \mapsto \min\left\{\sigma_{\mathcal{S}}(s_0, s_\imath),\, \sigma_{\mathcal{R}}(r, r_\imath)\right\}. \tag{5.12}$$

In fact, the max-operator in (5.8) is special t(riangular)-conorm and serves as a generalized logical or-operator: $r_0 = r$ is regarded as possible if $\langle s_0, r \rangle$ is similar to $\langle s_1, r_1 \rangle$ OR to $\langle s_2, r_2 \rangle$ OR ... OR to $\langle s_n, r_n \rangle$.

Now, fuzzy set theory offers t-conorms other than max and, hence, (5.8) could be generalized as follows:

$$\begin{aligned} \delta_{s_0}(r) &\stackrel{\text{df}}{=} \delta^1_{s_0}(r) \oplus \delta^2_{s_0}(r) \oplus \ldots \oplus \delta^n_{s_0}(r) \\ &= \bigoplus_{1 \leq \imath \leq n} \min\left\{\sigma_{\mathcal{S}}(s_0, s_\imath),\, \sigma_{\mathcal{R}}(r, r_\imath)\right\} \\ &= 1 - \bigotimes_{1 \leq \imath \leq n} \max\left\{1 - \sigma_{\mathcal{S}}(s_0, s_\imath),\, 1 - \sigma_{\mathcal{R}}(r, r_\imath)\right\} \end{aligned}$$

for all $r \in \mathcal{R}$, where $\otimes$ and $\oplus$ are a t-norm and a related t-conorm, respectively. Recall that a t-norm is a binary operator $\otimes : [0,1]^2 \longrightarrow [0,1]$ which is commutative, associative, monotone increasing in both arguments and which satisfies the boundary conditions $x \otimes 0 = 0$ and $x \otimes 1 = x$ [227]. 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 $\otimes_L : (\alpha, \beta) \mapsto \max\{0, \alpha + \beta - 1\}$ the related t-conorm of which is the bounded sum $\oplus_L : (\alpha, \beta) \mapsto \min\{1, \alpha + \beta\}$.

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 $\langle s_0, r\rangle$ and $\langle s_\imath, r_\imath\rangle$ are similar if both, s_0 is similar to $s_\imath$ *and* r is similar to $r_\imath$. Consequently, this operator might be replaced by a t-norm, too. By doing so, (5.12) and (5.8) become

$$\delta^\imath_{s_0} : r \mapsto \sigma_{\mathcal{S}}(s_0, s_\imath) \otimes \sigma_{\mathcal{R}}(r, r_\imath) \tag{5.13}$$

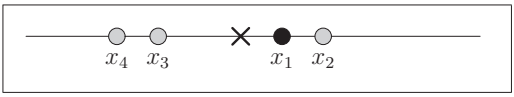
and

$$\delta_{s_0}(r) \stackrel{\text{df}}{=} \bigoplus_{1 \le \imath \le n} \sigma_{\mathcal{S}}(s_0, s_\imath) \otimes \sigma_{\mathcal{R}}(r, r_\imath), \tag{5.14}$$

respectively. Note, however, that a (fuzzy) logic-based derivation of the joint similarity is not compulsory. Particularly, the t-norm $\otimes$ in (5.14) 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 $\sigma_{\mathcal{S}}(s_0, s_\imath)$ and $\sigma_{\mathcal{R}}(r, r_\imath)$ by means of an operator which is not a t-norm. In that case, however, the "logical" interpretation of (5.14) is lost.

5.3.1 Control of compensation and accumulation of support

By choosing an appropriate t-conorm $\oplus$ in (5.14) one can control the accumulation of individual degrees of evidential support, especially the extent of compensation. To illustrate, consider the following classification scenario (with labels DARK and LIGHT), where $\sigma_{\mathcal{S}}(s_0, s_1) = 3/4$, $\sigma_{\mathcal{S}}(s_0, s_2) = \sigma_{\mathcal{S}}(s_0, s_3) = 1/2$, and $\sigma_{\mathcal{S}}(s_0, s_4) = 1/4$:



Should one prefer DARK or LIGHT as a classification of the new input (indicated by the cross)? The use of the max-operator as a t-conorm yields $\delta_{s_0}(\text{DARK}) = 3/4$ and $\delta_{s_0}(\text{LIGHT}) = 1/2$ 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 $\delta_{s_0}(\text{DARK}) = 3/4$ and $\delta_{s_0}(\text{LIGHT}) = 13/16$, 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 an output is already large, one is not surprised to see another (close) input having the same output. A small support increment then reflects the low information content related to the new observation [203].

5.3.2 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 $n \in \mathfrak{N}$ 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 $g : [0,1] \longrightarrow [0,\infty]$ such that $g(1) = 0$ and

$$\alpha \otimes \beta = g^{(-1)}(g(\alpha) + g(\beta)) \tag{5.15}$$

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

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

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 (5.15), one can establish an interesting connection between (5.14) and the weighted NN rule (cf. Section 2.2.1). To this end, let g be the additive generator of the t-norm[11] related to the t-conorm $\oplus$ used as an aggregation operator in (5.14). With $d_\imath = 1 - \sigma_{\mathcal{S}}(s_0, s_\imath) \otimes \sigma_{\mathcal{R}}(r, r_\imath)$ and $\omega_\imath = g(d_\imath)$, we can write (5.14) as

$$\delta_{s_0}(r) = 1 - g^{(-1)}(\omega_1 + \omega_2 + \ldots + \omega_n). \tag{5.16}$$

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 input. Then, (5.16) tells us that the possibility degree $\delta_{s_0}(r)$ is nothing else than a (monotone increasing) transformation of the sum of weights $\omega_\imath$. In other words, (5.14) 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 (2.8), however, the weight of a case according to (5.16) does not depend on other cases stored in the memory (cf. Section 5.3.5 below).

Consider the Lukasiewicz t-(co)norm as an example, for which we obtain $\omega_\imath = 1 - d_\imath = \sigma_{\mathcal{S}}(s_0, s_\imath) \otimes \sigma_{\mathcal{R}}(r, r_\imath)$ and

[11] This is not the t-norm used in (5.14) for defining a joint similarity measure.

$$\delta_{s_0}(r) = \min\{1, \omega_1 + \omega_2 + \ldots + \omega_n\}. \tag{5.17}$$

If, moreover, $\sigma_{\mathcal{R}}$ is given by (5.11), then $\delta_{s_0}(r)$ is nothing else than the bounded sum of the similarity degrees $\sigma_{\mathcal{S}}(s_\imath, s_0)$ between s_0 and the inputs $s_\imath$ with output $r_\imath = r$. Thus, (5.17) is basically equivalent to the global NN method, i.e., the weighted NN approach with $k = n$,[12] apart from the fact that it does not distinguish between outputs whose accumulated support exceeds 1 (this is an extreme type of saturation effect). For the probabilistic sum $\oplus_P$, the mapping between possibility degrees and the sum of weights is bijective:

$$\delta_{s_0}(r) = 1 - \exp\big(-(\omega_1 + \omega_2 + \ldots + \omega_n)\big).$$

In connection with the generalized model (5.14), the t-conorm $\oplus$ used for combining individual degrees of support defines another degree of freedom of the model. It is 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

$$\oplus_\rho : (\alpha, \beta) \mapsto \begin{cases} \max\{\alpha, \beta\} & \text{if } \rho = 0 \\ \alpha + \beta - \alpha\beta & \text{if } \rho = 1 \\ \min\{1, \alpha + \beta\} & \text{if } \rho = \infty \\ 1 - \ln_\rho \left(1 + \frac{(\rho^{1-\alpha}-1)(\rho^{1-\beta}-1)}{\rho-1}\right) & \text{otherwise} \end{cases} . \tag{5.18}$$

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.

5.3.3 Upper and lower possibility bounds

The possibility degree (5.14) represents the support (confirmation) of an output r gathered from similar cases according to the CBI hypothesis. Now, in the sense of this hypothesis, an observation $\langle s_\imath, r_\imath \rangle$ might not only confirm but also *disqualify* an output r. This happens if $s_\imath$ is close to s_0 but $r_\imath$ is not similar to r. A possibility distribution expressing degrees of *exclusion* rather than degrees of support and, hence, complementing (5.14) in a natural way is given by

$$\pi_{s_0} : r \mapsto \bigotimes_{1 \le \imath \le n} (1 - \sigma_{\mathcal{S}}(s_0, s_\imath)) \oplus \sigma_{\mathcal{R}}(r, r_\imath). \tag{5.19}$$

According to (5.19), an individual observation $\langle s_\imath, r_\imath \rangle$ induces a constraint on the outcome of s_0: An output r is disqualified by $\langle s_\imath, r_\imath \rangle$ if both, $\sigma_{\mathcal{S}}(s_0, s_\imath)$ is large and $\sigma_{\mathcal{R}}(r, r_\imath)$ is small. As opposed to this, $\langle s_\imath, r_\imath \rangle$ is completely ignored if

[12] The proper kNN rule cannot be emulated as in (2.10) since the weights $\omega_\imath$ depend on absolute distance (again, see Section 5.3.5 below).

$\sigma_{\mathcal{S}}(s_0, s_\imath) = 0$, in which case the individual support on the right-hand side of (5.19) is 1 ($\pi_{s_0} \equiv 1$ is an expression of complete ignorance: all upper possibility bounds are 1 since there is no reason to discredit any output). This approach is obviously in agreement with the constraint-based view of possibilistic reasoning (cf. Section 5.1.1). Moreover, the distribution (5.19) is again related to a special type of fuzzy rule [107].

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

$$\delta^*_{s_0}(r) = [\, \delta_{s_0}(r),\, \pi_{s_0}(r) \,]$$

with a lower bound $\delta_{s_0}(r)$ expressing a degree of confirmation, and an upper bound $\pi_{s_0}(r)$ expressing a degree of plausibility. The following cases show that the complementary distribution π_{s_0} can greatly improve the informational content of a possibilistic evaluation:[13]

- $\delta^*_{s_0}(r) = [0, 1]$: This is an expression of complete ignorance. Neither is r supported nor is it (partly) excluded by any observation. Thus, r is fully plausible though not confirmed at all.
- $\delta^*_{s_0}(r) = [0, 0]$: Clear evidence against r has been accumulated in the form of inputs similar to s_0 with outputs dissimilar to r.
- $\delta^*_{s_0}(r) \approx [1, 1]$: The output r is strongly supported through the observation of similar cases.

Notice that

$$\delta_{s_0}(r) > \pi_{s_0}(r) \tag{5.20}$$

indicates a kind of conflict [376] and is closely related to the problem of ambiguity in connection with the NN principle (cf. Section 2.2.1). In fact, (5.20) can occur if s_0 has close neighbors $s_\imath$ and $s_\jmath$ with quite dissimilar outputs $r_\imath$ and $r_\jmath$ (mathematically speaking, s_0 is a point of discontinuity). In this case, the evaluation of r is unsteady, and the support $\delta_{s_0}(r)$ should be taken with caution. The inequality in (5.20) might also trigger a revision process that aims at removing the conflict by means of a model adaptation.

5.3.4 Fuzzy logical evaluation

The values $\delta_{s_0}(r)$ in (5.14) can also be considered as membership degrees of a fuzzy set, namely the fuzzy set of "well-supported outputs". In fact, the possibility degree $\delta_{s_0}(r)$ can be seen as the truth degree, $\langle P(r) \rangle$, of the following (fuzzy) predicate $P(r)$: "There is an input close to s_0 with an output similar to r." $P(r)$ defines the property that qualifies r as a well-supported output.

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

Of course, one might easily think of alternative characterizations of well-supported outputs. 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 inputs, most of these inputs have the same output, and none of the moderately close inputs has a very different output."

The logical expression $P(\cdot)$ associated with such a specification can be used in place of the right-hand side in (5.14):

$$\delta_{s_0}(r) \stackrel{\text{df}}{=} \langle P(r) \rangle. \tag{5.21}$$

The decision rule related to (5.14) favors the outcome r_0^{est} 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 (5.21) can also mimic the original kNN rule: Consider the fuzzy proposition "r is supported by many of the k nearest neighbors of s_0", and let the fuzzy quantifier "many (out of k)" be modeled by the mapping $\imath \mapsto \imath/k$. Then, $\delta_{s_0}(r) = \imath/k$ iff $\imath$ among the k nearest neighbors have outcome r. In this case, possibility degrees (derived from fuzzy truth degrees) formally coincide with probability degrees.

5.3.5 Comparison of extrapolation principles

As already mentioned above, the possibilistic approach to CBI can also be considered as a kind of NN estimation. Thus, it seems interesting to have a closer look at this type of "possibilistic NN estimation" as an alternative to the *probabilistic* approach to estimation and decision making, which is in agreement with the original kNN rule (cf. Section 2.2.1).

Both the possibilistic and the probabilistic approach 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 output $r \in \mathcal{R}$. The second step, the NN *decision*, chooses one output on the basis of the NN estimation. Usually, the decision is given by the outcome with maximal support, and ties are broken by coin flipping. Still, in the case of a continuous (or at least ordinal) scale $\mathcal{R}$, 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

$$\nu(r \mid s_0, \mathcal{M}) = \alpha\left(\{\nu_{s_\imath}(r \mid s_0, \mathcal{M}) \mid \langle s_\imath, r_\imath\rangle \in \mathcal{M}\}\right) \tag{5.22}$$

and thus obtain the (maximal support) decision as

$$r_0^{est} = \arg\max_{r\in\mathcal{R}} \nu(r \mid s_0, \mathcal{M}). \tag{5.23}$$

In (5.22), $\nu_{s_\imath}(r \mid s_0, \mathcal{M})$ is the support of the hypothesis $r_0 = r$ provided by the case $\langle s_\imath, r_\imath\rangle$, and α is an aggregation function.

To reveal the original kNN rule and the probabilistic approach as special cases of (5.23), note that the probability distribution (2.6) is obtained by using the arithmetic sum as an aggregation function α and defining the support function as

$$\nu^p_{s_\imath}(r \mid s_0, \mathcal{M}) = \begin{cases} 1/k & \text{if } s_\imath \in \mathcal{N}_k(s_0) \text{ and } r = r_\imath \\ 0 & \text{otherwise} \end{cases}. \tag{5.24}$$

More generally, if $\mathcal{S}$ is a metric space, a support function can be defined as

$$\nu^p_{s_\imath}(r \mid s_0, \mathcal{M}) = \begin{cases} K_{d_k}(s_0 - s_\imath) & \text{if } r = r_\imath \\ 0 & \text{otherwise} \end{cases}, \tag{5.25}$$

where K is a kernel function. The index d_k denotes the distance between s_0 and its k-th nearest neighbor. It signifies that the kernel function is *scaled* so as to exclude exactly those inputs $s_\imath$ with $\Delta_{\mathcal{S}}(s_0, s_\imath) > d_k$. Proceeding from (5.25), and assuming that $\mathcal{R}$ is a finite set $\{\rho_1 \ldots \rho_m\}$, the probability distribution p_{s_0} is obtained by normalizing the supports

$$\nu^p(\rho_\jmath \mid s_0, \mathcal{M}) = \sum_{\langle s_\imath, r_\imath\rangle \in \mathcal{M}} \nu^p_{s_\imath}(\rho_\jmath \mid s_0, \mathcal{M}),$$

which yields

$$p_{s_0}(\rho_\jmath) = \frac{\nu^p(\rho_\jmath \mid s_0, \mathcal{M})}{\sum_{\imath=1}^m \nu^p(\rho_\imath \mid s_0, \mathcal{M})} \tag{5.26}$$

for all $\rho_\jmath \in \mathcal{R}$. 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 (5.14) is recovered by $\alpha = \oplus$ and

$$\nu^\delta_{s_\imath}(r \mid s_0, \mathcal{M}) = \sigma_{\mathcal{S}}(s_0, s_\imath) \otimes \sigma_{\mathcal{R}}(r, r_\imath). \tag{5.27}$$

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

Apart from that, however, a direct comparison is complicated by the similarity measure over outputs, $\sigma_{\mathcal{R}}$, which is used in (5.27) but not in (5.25). One possibility to handle this problem is to consider (5.27) only for the special case (5.11):

$$\nu^{\delta}_{s_\imath}(r \,|\, s_0, \mathcal{M}) = \begin{cases} \sigma_{\mathcal{S}}(s_0, s_\imath) & \text{if } r = r_\imath \\ 0 & \text{otherwise} \end{cases} . \tag{5.28}$$

Equation (5.28) reveals that the similarity measure $\sigma_{\mathcal{S}}$ now plays the same role as the kernel function K in (5.25).

Absolute versus relative support. An important difference between (5.25) and (5.28) is that an example $\langle s_\imath, r_\imath \rangle \in \mathcal{M}$ provides *relative* support of an output r in the probabilistic approach but *absolute* support in the possibilistic one. That is, $\nu^{\delta}_{s_\imath}(r \,|\, s_0, \mathcal{M})$ depends on the absolute similarity between s_0 and $s_\imath$ but is independent of further observations. In fact, we can actually write $\nu^{\delta}_{s_\imath}(r \,|\, s_0)$ in place of $\nu^{\delta}_{s_\imath}(r \,|\, s_0, \mathcal{M})$ since $\mathcal{M}$ does not appear on the right-hand side of (5.28): The support provided by observed examples $\langle s_\imath, r_\imath \rangle$ is bounded to nearby cases, decreases gradually with distance, and vanishes for completely dissimilar cases.

As opposed to this, the support $\nu^{p}_{s_\imath}(r \,|\, s_0, \mathcal{M})$ is relative and depends on the relation between the distance of $s_\imath$ to s_0 and the distances of other observations to s_0. This is reflected by the scaling of the kernel function in (5.25). On the one hand, this means that $\nu^{p}_{s_\imath}(r \,|\, s_0, \mathcal{M})$ can be large even though $s_\imath$ is quite distant from s_0. On the other hand, the extension of the memory $\mathcal{M}$ by another instance close enough to s_0 might exclude a quite similar observation $s_\imath$ from the neighborhood $\mathcal{N}_k(s_0)$. The corresponding re-scaling of the kernel function will then cancel the support provided by $\langle s_\imath, r_\imath \rangle$ so far. The induced thresholding effect appears especially radical (and might be questioned on such grounds) in connection with (5.24), where $\nu^{p}_{s_\imath}(r \,|\, s_0, \mathcal{M})$ is reduced from $1/k$ to 0, that is from full support to zero support.

The bounding of evidential support, as realized by the possibilistic approach, is often advisable. Consider a simple example: Let $\mathcal{S} = [0, 1]$ and

$$\varphi = \{(s, \mathbb{I}_{[1/2,1]}(s)) \,|\, s \in \mathcal{S}\}$$

and suppose inputs to be chosen at random according to a uniform distribution. Moreover, assume that a new input s_0 must be labeled, given a memory that consists of only a single observation $\langle s_1, r_1 \rangle$. Using the 1NN rule, the probability of a correct decision is obviously 1/2. Now, suppose that the NN rule is applied only if $|s_0 - s_1| \leq d$, whereas a decision is determined by flipping a coin otherwise (this is exactly the procedure that results from the possibilistic approach by defining $\sigma_{\mathcal{S}}$ in (5.8) by $\sigma_{\mathcal{S}}(s, s') = 1$ if $|s - s'| \leq d$ 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 cases are likely to provide misleading information in this example and, hence, the disregard of such cases is indeed advantageous. Loosely speaking, it is better to guess an output 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 (5.25) by

$$\nu^p_{s_\imath}(r \mid s_0, \mathcal{M}) = \begin{cases} K(s_0 - s_\imath) & \text{if } r = r_\imath \\ 0 & \text{otherwise} \end{cases}, \tag{5.29}$$

where the kernel function K is now fixed. That is, K is no longer scaled by the size of the neighborhood of s_0. This is exactly the estimation one derives by the reasoning in Section 2.2.1 if the generalized NN density estimation (2.14) is replaced by the simple kernel estimator:

$$\phi^{est}(s_0) = \frac{1}{n} \cdot \sum_{\imath=1}^{n} K(s_0 - s_\imath). \tag{5.30}$$

Here, the only problem occurs if $\nu^p(r \mid s_0, \mathcal{M}) = 0$ for all $r \in \mathcal{R}$. In this situation (of complete ignorance), a probability distribution cannot be derived by normalization.

Apart from that, (5.29) might indeed be preferred to (5.25) due to the reasons mentioned above. In fact, one should realize that one of the major reasons for using the NN density estimator (2.14) rather than the kernel estimator (5.30) is to guarantee the continuity of the density function ϕ^{est}. In the context of case-based inference or, say, instance-based learning this is not important, however, since one is not interested in estimating a complete density function but only a single value thereof. To the best of our knowledge, (5.25) and (5.29) have not been compared in a systematic way in IBL so far. Note that (5.29) 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 outcomes is often advised (see also our example in Section 5.3.7 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 $\mathcal{S} \times \mathcal{R}$. An estimation for the output r can then be derived by conditioning on s_0:

$$p_{s_0}(r) \propto \sum_{\langle s_\imath, r_\imath \rangle \in \mathcal{M}} \nu^p_{s_\imath}(r \mid s_0, \mathcal{M}) = \sum_{\langle s_\imath, r_\imath \rangle \in \mathcal{M}} K\left(s_0 - s_\imath,\, r - r_\imath\right).$$

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

Similarity versus frequency. The estimation principle underlying the probabilistic NN 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 (5.8) 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.[14] The generalized possibilistic approach based on (5.14) 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.[15] For example, consider the standard probabilistic setting where cases are chosen randomly and independently according to a fixed probability measure over $\mathcal{S} \times \mathcal{R}$. The possibility degree $\delta_{s_0}(r)$ will then converge to 1 with increasing sample size whenever $\langle s_0, r\rangle$ 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_r, $r \in \mathcal{R}$, whereas the probabilistic approach aims at estimating conditional probability distributions $p_{s_0} = \mathbb{P}(\cdot \,|\, s_0)$. Of course, this distinction is relevant only if the concepts are overlapping, that is, if the query s_0 does not have a unique outcome. Otherwise, a possibilistic and a probabilistic approach are equivalent in the sense that $s_0 \in C_r \Leftrightarrow \mathbb{P}(r \,|\, s_0) = 1$.

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 $C_1 \subseteq \mathcal{S}$, depends on the number of positive examples observed so far and hence contradicts an independence assumption (the probability of an output r, given the input s, is not independent of the data). In this case, a probabilistic estimation is clearly biased, whereas the possibility distribution (5.8) is not affected at all. Indeed, the information expressed by δ_{s_0} remains valid even if only negative examples $s_\imath \in C_0 = \mathcal{S} \setminus C_1$ have been presented so far: $\delta_{s_0}(1) = 0$ then simply means that no evidence for $s_0 \in C_1$ has been gathered as yet. Moreover, the value $\delta_{s_0}(0)$ reflects the available support for $s_0 \in C_0$. This support depends on the distance of s_0 to the observed negative examples. Note that $\delta_{s_0}(0) = 0$ is possible as well. In this case, no evidence is available at all, neither for nor against $s_0 \in C_1$. See Section 5.5.3 for a simulation experiment which concerns the aspect of robustness of NN estimation toward violations of the standard statistical assumptions.

[14] To a certain extent, this is related to the distinction between an *existential* and an *enumerative* analogy factor in models of analogical induction [281].

[15] Though such a relationship can be established, e.g., by interpreting possibility as upper probability [122] or fuzzy sets as coherent random sets [111].

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 $\Delta_{\mathcal{S}}(s_1, s_2) < \Delta_{\mathcal{S}}(s_1, s_3)$ whenever $r_1 = r_2 \neq r_3$. The output of an input 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 $k > 1$.

5.3.6 From predictions to 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. 2.1.

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 $\delta_{s_0}(r)$ remain rather small if no sufficiently similar cases are available. Particularly, the distribution $\delta_{s_0} \equiv 0$ is an expression of *complete ignorance* and reflects the absence of any relevant observation ($\sigma_{\mathcal{S}}(s_0, s_i) = 0$ for all s_i). A learning agent using this estimation "knows that it doesn't know" [359]. As opposed to this, a distribution such as, say, $\delta_{s_0} \equiv |\mathcal{R}|^{-1}$ (in the case of finite $\mathcal{R}$) indicates that some (small) evidence is available for each of the potential outcomes. These two situations cannot be distinguished in probability theory where they induce the same distribution $p_{s_0} \equiv |\mathcal{R}|^{-1}$ (if, as suggested by the principle of insufficient reason, complete ignorance is modeled by the uniform distribution).
- Possibility reflects absolute frequency: For example, suppose $\sigma_{\mathcal{S}}(s_0, s_i) = 1 - d > 0$ and $r_i = r'$ for all n inputs s_i stored in the memory. The probabilistic estimation (2.6) then yields the one-point distribution $p_{s_0}(r') = 1$ and $p_{s_0}(r) = 0$ for all $r \neq r'$. Thus, it suggests that $r_0 = r'$ is certain, even if n is rather small. With a compensating t-conorm such as the probabilistic sum $\oplus_P$, the extended estimation (5.14) yields $\delta_{s_0}(r') = 1 - d^n$ and $\delta_{s_0}(r) = 0$ for all $r \neq r'$. Thus, not only does the possibilistic support of the hypothesis $r_0 = r'$ reflect the distance but also the actual number of voting instances: $\delta_{s_0}(r')$ 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. 2.1, above) is present if there are several plausible outputs with similar

degrees of support, and ignorance (Fig. 2.1, below) is reflected by the fact that even the most supported output has a small degree of possibility. Thus, (5.14) can be taken as a point of departure for a decision making procedure that goes beyond the guessing of an outcome. For example, a possible line of action proceeding from (5.14) might be expressed by the following rules (involving thresholds $0 < d_{max} < d_{min} < 1$):

– If $\delta_{s_0}(r^*) \geq d_{min}$ for the most supported outcome r^* and $\delta_{s_0}(r) \leq d_{max}$ for all $r \neq r^*$, then let $r_0^{est} = r^*$.

– If $\delta_{s_0}(r^*) < d_{min}$, then gather further information.

– If $\delta_{s_0}(r^*) \geq \delta_{s_0}(r) \geq d_{min}$ for two outcomes $r^*, r \in \mathcal{R}$, then refuse a prediction.

The ECHOCARDIOGRAM DATABASE[16] 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 ($\delta_{s_0} \equiv 0$) is very different from telling him that his chance is 50% ($p_{s_0} \equiv 1/2$).

The discrepancy between a probabilistic and a possibilistic approach 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. For example, the method in [84], which derives a prediction in terms of a *belief function* (cf. Chapter 4), refers to the so-called transferable belief model [350] and, hence, turns the belief function (at the "credal" level) specifying the unknown outcome into a probability function (at the "pignistic" level) before making a decision. Thus, the support of individual outputs is expressed in terms of probability, and an NN estimation can be derived by taking one among the most probable outcomes, breaking ties at random.

Observe that, as a consequence of applying the maximum operator, a possibilistic NN decision derived from (5.8) coincides with the 1NN rule. The generalized version (5.14), 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 (5.29). Equation (5.16) shows that a possibility degree $\delta_{s_0}(r)$ is a monotone transformation of the sum of weights ω_i, and this relation is one-to-one if the pseudo-inverse $g^{(-1)}$ is actually the inverse g^{-1}. The similarity function $\sigma_{\mathcal{S}}$ can then be chosen

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

such that

$$\delta_{s_0}(r) \leq \delta_{s_0}(r') \;\Leftrightarrow\; p_{s_0}(r) \leq p_{s_0}(r').$$

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

To illustrate, consider the case where $\mathcal{S} = \mathfrak{R}^l$ and $\sigma_{\mathcal{R}}(r, r') = 1$ if $r = r'$ and 0 otherwise. Let K be a kernel function and define $\sigma_{\mathcal{S}}$ as $(x, y) \mapsto 1 - \exp(-K(x, y))$.[17] For the t-conorm $\oplus_P$, the weights in (5.16) are then given by $\omega_i = K(s_0 - s_i)$. Therefore,

$$\begin{aligned}\delta_{s_0}(r) &= 1 - \exp\left(-\sum_{\langle s_i, r_i\rangle \in \mathcal{M}\,:\, r_i = r} K(s_0 - s_i)\right) \\ &= 1 - \exp\left(-c \cdot p_{s_0}(s_i)\right),\end{aligned}$$

where $p_{s_0}(r)$ is the probability degree derived from (5.29) using the kernel function K and c is the normalization factor $c = \sum_{r' \in \mathcal{R}} p_{s_0}(r')$.

5.3.7 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 input is now a subject, and the output is given by the corresponding grade. We assume that grades are taken from the scale $\mathcal{R} = \{0, 1, \ldots, 10\}$, 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	–

Admittedly, it is not obvious how to define a reasonable similarity measure over the set of subjects. In fact, an ordinal measure – sufficient for the possibilistic approach (5.8) – appears much simpler than a cardinal one. Nevertheless, let us assume the following (cardinal) degrees of similarity:

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

[17] Formally, one might set $K(0) \stackrel{\text{df}}{=} \infty$ to ensure that $\sigma_{\mathcal{S}}$ is reflexive.

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

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

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 $\mathcal{S}$ endowed with the similarity measure $\sigma_{\mathcal{S}}$ (as partly specified above) is likely to lack a sufficiently strong mathematical (metric) structure, so that the derivation of the kNN estimation in Section 2.2.1 might no longer be 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 $\delta_{s_0} = \delta_{physics} \equiv 0$. 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 (5.26) is zero. One way out is to take the uniform distribution $p_{s_0} \equiv 1/11$ 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 [198]. Here, such assumptions correspond to the NN inductive bias, namely the CBI hypothesis that similar inputs have similar outputs. Given this hypothesis, the cases 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 the memory (cf. Section 4.5.3), 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 $\delta_{s_0}(3) = 1/3 < 3/4 = \delta_{s_0}(10)$. The probabilistic sum $\oplus_P$ brings about a reinforcement effect but still yields $\delta_{s_0}(3) = 0.7 < 3/4 = \delta_{s_0}(10)$, 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_{s_0}, and the same is also true for most aggregation operators in the possibilistic approach. For example, (5.14) with $\oplus = \oplus_P$ (and $\otimes = \otimes_P$) yields $\delta_{s_0}(3) > \delta_{s_0}(7) < \delta_{s_0}(10)$. This result is not intuitive, for one might hardly judge an intermediate grade less possible than two extreme grades. To solve this problem, δ_{s_0} can be replaced by its convex hull

$$r \mapsto \min \left\{ \max_{r' \leq r} \delta_{s_0}(r'), \max_{r' \geq r} \delta_{s_0}(r') \right\}. \tag{5.31}$$

In our example, this leads to the following distribution:

r	0	1	2	3	4	5	6	7	8	9	10
$\delta_{s_0}(r)$	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 (5.31) of δ_{s_0} should not be considered ad-hoc. Rather, the convexity requirement can be thought of as a possibility-qualifying rule that complements the case-based justification of possibility degrees: The more possible two outputs are, the more possible is any outcome 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: Firstly, 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. Secondly, a possibility distribution can represent ignorance. Thirdly, the use of aggregation operators other than the arithmetic sum can be useful. Fourthly, the possibilistic approach is more flexible and allows for incorporating constraints or background knowledge.

5.3.8 Complexity issues

A straightforward implementation of the prediction (5.13) has a running time which is linear in the size $|\mathcal{M}|$ of the memory and the number $|\mathcal{R}|$ of outcomes (resp. a discretization thereof). In this respect, it is hence completely comparable to other case-based learning methods.

In order to reduce the computational complexity, instance-based 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 [154, 411, 222]. 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 the possibilistic approach. In fact, a possibility degree $\delta_{s_0}(r)$ is completely determined by the neighborhood of the case $\langle s_0, r\rangle$, that is the sample instances $\langle s_\imath, r_\imath\rangle$ satisfying $\sigma_{\mathcal{S}}(s_\imath, s_0) > 0$ and $\sigma_{\mathcal{R}}(r_\imath, r) > 0$. As can be seen, apart from minor differences, the possibilistic method is quite comparable to other instance-based 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_{s_0} is not needed. In fact, quite often one will only be interested in those outcomes having a high degree of possibility. For example, one might be interested in a fixed number of maximally supported outcomes, or in those outcomes whose support exceeds a given possibility threshold. In such cases, the computation of $\delta_{s_0}(r)$ can be omitted (or broken off) for certain outputs r.

5.4 Extensions of the basic model

The previous section has introduced the main principles of the possibilistic approach to case-based inference (subsequently, for the sake of brevity, sometimes referred to as PoCBI). In this regard, the close connection to fuzzy rule-based

reasoning was especially emphasized. Besides, we highlighted the fact that possibilistic CBI can be considered as an alternative approach to NN estimation. This section presents some extensions of the basic model making PoCBI even more powerful and practically useful.

5.4.1 Dealing with incomplete information

The problem of dealing with incomplete information such as missing attribute values in an important issue in case-based reasoning and machine learning [88, 305]. For example, suppose that the specification of the new query s_0 is incomplete, and let $S_0 \subseteq \mathcal{S}$ denote the inputs compatible with the description of s_0. Moreover, recall the lower support-bound semantics of the possibilistic approach to CBI. The following generalization of (5.14) is in accordance with these semantics:

$$\delta_{s_0}(r) \stackrel{\text{df}}{=} \inf_{s \in S_0} \delta_s(r) = \tag{5.32}$$

$$= \inf_{s \in S_0} \bigoplus_{1 \leq \imath \leq n} \sigma_{\mathcal{S}}(s, s_\imath) \otimes \sigma_{\mathcal{R}}(r, r_\imath).$$

Indeed, each potential candidate $s \in S_0$ gives rise to a lower bound according to (5.14), 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 5.1.2). The simplicity of handling incomplete information in a coherent (namely possibilistic) way is clearly a strong point of possibilistic CBI. Notice that the computation of the lower bound in (5.32) is in line with the handling of missing attribute values in the IB1 algorithm (cf. Section 2.2.2), 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.

Example 5.2. Reconsider Example 5.1 and suppose that we are interested in, say, the price of a car whose horsepower is between 90 and 110. This amounts to predicting the outcome of an income s_0, in which the attributes are incompletely specified. Fig. 5.1 shows the prediction obtained for the max–min version of (5.32) for this example. □

More generally, imprecise knowledge about s_0 can be modeled in the form of a possibility distribution π on $\mathcal{S}$, where $\pi(s)$ corresponds to the degree of plausibility that $s_0 = s$. 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 (5.32):

$$\delta_{s_0}(r) \stackrel{\text{df}}{=} \inf_{s \in \mathcal{S}} \big(\pi(s) \rightsquigarrow \delta_s(r) \big), \tag{5.33}$$

where $\rightsquigarrow$ is a generalized implication operator that is reasonably chosen as the Gödel implication [134]:

$$\alpha \rightsquigarrow \beta \stackrel{\text{df}}{=} \begin{cases} 1 & \text{if} \quad \alpha \leq \beta \\ \beta & \text{if} \quad \alpha > \beta \end{cases} .$$

From a logical point of view, (5.33) specifies the extent to which *the output r is supported by all plausible candidates for s_0*. Notice that the distributions δ_s and π in (5.32) have different semantics and express degrees of confirmation and plausibility, respectively (cf. Section 5.1). Particularly, π is assumed to be normalized, i.e., there is at least one input s with $\pi(s) = 1$. One obviously recovers (5.32) from (5.33) for the special case where π is a $\{0,1\}$-valued possibility distribution $\pi = \mathbb{I}_{S_0}$ and hence corresponds to a crisp subset $S_0 \subseteq \mathcal{S}$.

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 $S_\imath \times R_\imath \subseteq \mathcal{S} \times \mathcal{R}$. Then, (5.14) becomes

$$\delta_{s_0}(r) \stackrel{\text{df}}{=} \bigoplus_{1 \leq \imath \leq n} \inf_{\langle s', r' \rangle \in S_\imath \times R_\imath} \sigma_{\mathcal{S}}(s_0, s') \otimes \sigma_{\mathcal{R}}(r, r'),$$

which is in accordance with (5.32). Moreover, we obtain

$$\delta_{s_0}(r) \stackrel{\text{df}}{=} \bigoplus_{1 \leq \imath \leq n} \inf_{\langle s', r' \rangle \in \mathcal{S} \times \mathcal{R}} \max \left\{ \sigma_{\mathcal{S}}(s_0, s') \otimes \sigma_{\mathcal{R}}(r, r'), 1 - \pi_\imath(s', r') \right\}$$

if the $\imath$-th case is characterized by means of a possibility distribution $\pi_\imath$ on $\mathcal{S} \times \mathcal{R}$ rather than by a crisp set $S_\imath \times R_\imath$. Note that this expression can be combined with (5.33) in order to handle incomplete specifications of both, the sample cases and the new query. Moreover, notice that the distribution δ_{s_0} will generally remain unaffected if an example is completely unspecified ($\pi_\imath \equiv 1$), which is clearly a reasonable property.

Interestingly enough, the above generalization does not only allow for dealing with incomplete (fuzzy) cases. It also suggests to lump together several (similar) cases stored in the memory. The idea, then, is to replace these cases by one "fuzzy case", the attributes of which are given by the disjunction of the attribute values of the individual cases. On the one hand, this procedure might improve efficiency, especially if the memory of cases is very large. On the other hand, some information might be lost when basing a prediction on one or several fuzzy cases: In fact, it is not difficult to show that the support $\delta_{s_0}(r)$ of a (hypothetical) case $\langle s_0, r \rangle$ derived from a set of observed cases can be larger (but not smaller) than the support obtained from the fuzzy case which combines the original observations. Nevertheless, the more similar the combined observations are, the better the approximation becomes. Of course, instead of replacing a set of cases by a fuzzy case, one might also think of simply selecting one of these cases which is prototypical of this set.[18]

[18] This is in line with the idea of generating prototypes by merging training samples – and thus reducing the size of the training set – which has been proposed in the context of NN classification [62].

5.4.2 Discounting noisy and atypical instances

Since case-based prediction and instance-based learning are quite sensitive to noisy instances, it is reasonable to discard those instances [5]. By noise one generally means incorrect attribute value information, concerning either the descriptive part s of a case or the outcome r (or both). However, the problem of noise is also closely related to the "typicality" of a case. A typical case is representative of its neighbors, whereas an exceptional (though not incorrect) case has an outcome quite different from the outputs of neighboring cases [419].

Recall that each case $\langle s_\iota, r_\iota \rangle \in \mathcal{M}$ is extrapolated by placing the support function or, say, "possibilistic kernel" (5.13) around the point $\langle s_\iota, r_\iota \rangle \in \mathcal{S} \times \mathcal{R}$, 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) a case 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 (5.13):

$$\delta^\iota_{s_0} : r \mapsto m_\iota \big(\sigma_\mathcal{S}(s_0, s_\iota)\big) \otimes \sigma_\mathcal{R}(r, r_\iota). \tag{5.34}$$

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

Suppose that a new observation s_0 with output r_0 has been made, and consider a stored case $\langle s_\iota, r_\iota \rangle$. Should this case be discounted in the light of the new observation? The fact that $\langle s_\iota, r_\iota \rangle$ supports an outcome different from the observed output r_0 need not necessarily be a flaw. In fact, recall that $s_0 \in C_{r_0}$ does not exclude that $s_0 \in C_r$ for some $r \neq r_0$. In other words, neither the non-support of the observed nor the support of a different outcome can actually be punished. However, what can be punished is the disqualification of the output r_0 as expressed by the upper possibility model (5.19). Thus, it is reasonable to require that the degree of disqualification induced by $\langle s_\iota, r_\iota \rangle$ is limited:

$$1 - m_\iota(\sigma_\mathcal{S}(s_0, s_\iota)) \otimes \sigma_\mathcal{R}(r_0, r_\iota) \;\geq\; \beta, \tag{5.35}$$

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

The constraint (5.35) suggests an update scheme in which a stored case $\langle s_\iota, r_\iota \rangle$ is (maybe) discounted every time a new observation $\langle s_0, r_0 \rangle$ is made: Let $\mathcal{F}$ denote a parameterized and completely ordered class of functions from which m_ι is chosen. An adaptation is then realized by

$$m_\iota \leftarrow \min\big\{m_\iota, \sup\{f \in \mathcal{F} \,|\, 1 - f(\sigma_\mathcal{S}(s_0, s_\iota)) \otimes \sigma_\mathcal{R}(r_0, r_\iota) \;\geq\; \beta\}\big\}. \tag{5.36}$$

The discounting of noisy and atypical instances through modifying possibilistic kernel functions appears natural and somewhat simpler than the method used in IB3 [5]. 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 decision is based on a rule that appears reasonable but might still be considered ad-hoc ($s_\imath$ is discounted if $\Delta_{\mathcal{S}}(s_\imath, s_0)$ is smaller than or equal to the distance between s_0 and its closest *accepted* neighbor[19]).

The possibilistic adaptation scheme becomes rather simple for the special case $\mathcal{S} = \mathfrak{R}^l$, $\mathcal{R} = \{0, 1\}$ and $m_\imath = \mathbb{I}_{]\gamma_\imath, 1]}$, where $0 \leq \gamma_\imath < 1$. If $\sigma_{\mathcal{S}}$ is a strictly decreasing function of Euclidean distance, then the support function (5.13) corresponds to a ball around $s_\imath$: $\delta^\imath_{s_0}(r) = 1$ if $r = r_\imath$ and s_0 is located inside that ball and $\delta^\imath_{s_0}(r) = 0$ otherwise. The parameter $\gamma_\imath$ is chosen as large as possible, but such that the support function does not cover any observed input $s_\jmath$ with $r_\jmath \neq r_\imath$, that is $\gamma_\imath \leq |s_\imath - s_\jmath|$ holds true for all of those $s_\jmath$. Fig. 5.2 gives an illustration for $l = 2$.

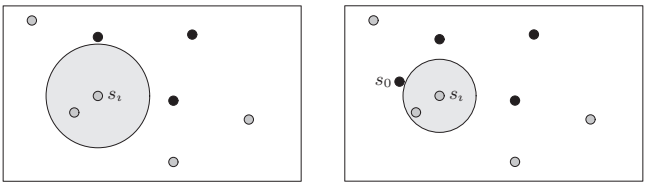


Fig. 5.2. Left: The large circle corresponds to the support function (possibilistic kernel) centered around $s_\imath$ and marks the extrapolation of outcome $r_\imath$. Right: The support function is updated after observing a new instance which has a different outcome $r_0 \neq r_\imath$ and hence must not be supported.

This special case, that we shall subsequently refer to as PossIBL, is a useful point of departure for investigating theoretical properties of the possibilistic approach in the context of concept learning. In [11], 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 $l = 2$, $\mathcal{S} = [0, 1] \times [0, 1]$ (the results can be generalized to any dimension $l > 2$ and any bounded region $\mathcal{S} \subset \mathfrak{R}^l$) and consider a concept $C_1 \subseteq \mathcal{S}$. For the special case above, the PossIBL approximation of C_1 is then given by

$$C_1^{est} = \bigcup_{\langle s_\imath, 1\rangle \in \mathcal{M}} \mathfrak{B}_{\rho(s_\imath)}(s_\imath), \tag{5.37}$$

[19] Auxiliary rules are used if s_0 does not have an accepted neighbor.

where $\mathfrak{B}_d(s_\imath) = \{s \in \mathcal{S} \,|\, |s - s_\imath| < d\}$ is the (open) d-ball around $s_\imath$ and

$$\rho(s_\imath) = \min\left\{|s_\jmath - s_\imath| \,|\, \langle s_\jmath, r_\jmath\rangle \in \mathcal{M},\, r_\jmath \neq r_\imath\right\}. \tag{5.38}$$

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

$$C_0^{est} = \bigcup_{\langle s_\imath, 0\rangle \in \mathcal{M}} \mathfrak{B}_{\rho(s_\imath)}(s_\imath). \tag{5.39}$$

It is readily verified that $C_0^{est} \cap C_1^{est} = \emptyset$. However, $C_0^{est} \cup C_1^{est} = \mathcal{S}$ does not necessarily hold true. Thus, one may have $\delta_{s_0} \equiv 0$ for some instances $s_0 \in \mathcal{S}$ (which are then classified at random). Consequently, an approximation of concept C_1 should actually be represented by the tuple (C_0^{est}, C_1^{est}) which divides instances $s_0 \in \mathcal{S}$ into three groups: Those which (supposedly) belong to C_1 ($\delta_{s_0}(0) = 0, \delta_{s_0}(1) = 1$), those which do not ($\delta_{s_0}(0) = 1, \delta_{s_0}(1) = 0$), and those for which no evidence is available so far ($\delta_{s_0} \equiv 0$).

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 [11]: Instances are generated randomly and independently according to a fixed probability measure μ over $\mathcal{S}$. 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

$$C_1^+(\varepsilon) \stackrel{\text{df}}{=} \{s \in \mathcal{S} \,|\, \mathfrak{B}_\varepsilon(s) \cap C_1 \neq \emptyset\},$$

and the ε-core of C_1 is defined by

$$C_1^-(\varepsilon) \stackrel{\text{df}}{=} \{s \in \mathcal{S} \,|\, \mathfrak{B}_\varepsilon(s) \subseteq C_1\}.$$

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

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

Finally, let $C_{1,n}^{est}$ and $C_{0,n}^{est}$ denote, respectively, the possibilistic concept approximations (5.37) and (5.39) for $|\mathcal{M}| = n$, i.e., after n observations have been made.

Lemma 5.3. The equalities

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

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

Proof. For $s \in C_1^-(\varepsilon)$ we have $\mathfrak{B}_\varepsilon(s) \subseteq C_1$, which means that $|s-s_1| < \varepsilon$ implies $s_1 \in C_1$. Consequently, there is no $s_0 \in C_0$ such that $|s - s_0| < \varepsilon$ and, hence, $s \notin C_0^+(\varepsilon)$. Now, suppose $s \in \mathcal{S} \setminus C_0^+(\varepsilon)$. Thus, there is no $s_0 \in C_0$ such that $|s-s_0| < \varepsilon$, which means that $|s-s_1| < \varepsilon$ implies $s_1 \in C_1$ and, hence, $s \in C_1^-(\varepsilon)$. The second equality is shown in the same way. □

Theorem 5.4. Let $C_1 \subseteq \mathcal{S}$ 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 $C_{1,n}^{est}$ is a $(2\varepsilon, \gamma)$-approximation of C_1 and $C_{0,n}^{est}$ is a $(2\varepsilon, \gamma)$-approximation of C_0 for all $n > n_0$. □

Proof. Let N denote the set of instances $s \in \mathcal{S}$ for which no $s_\imath \in \mathcal{M}^\downarrow$ exists such that $|s - s_\imath| < \varepsilon$. In [11], the following lemma has been shown: $\mu(N) \leq \gamma$ holds true with probability $1-d$ whenever

$$n > \lceil n_0 = \sqrt{2}/\varepsilon \rceil^2/\gamma^2 \cdot \ln\left(\lceil \sqrt{2}/\varepsilon \rceil^2/d\right). \tag{5.40}$$

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

Now, consider any instance $s \in C_1^-(2\varepsilon)$. We have to show that $s \in C_{1,n}^{est}$. Let $s_\imath \in \mathcal{M}^\downarrow$ be an instance such that $|s - s_\imath| < \varepsilon$. For this instance we have $s_\imath \in \mathfrak{B}_\varepsilon(s) \subseteq C_1$, which means that $s_\imath$ belongs to C_1. Furthermore, $\mathfrak{B}_\varepsilon(s_\imath) \subseteq \mathfrak{B}_{2\varepsilon}(s) \subseteq C_1$ and, hence, $\rho(s_\imath) \geq \varepsilon$ for the value in (5.38). This implies that $s \in \mathfrak{B}_{\rho(s_\imath)}(s_\imath)$ and, therefore, $s \in C_{1,n}^{est}$. Thus, we have shown that $C_1^-(2\varepsilon) \subseteq C_{1,n}^{est}$.

Since the same arguments apply to C_0, the property $C_0^-(2\varepsilon) \subseteq C_{0,n}^{est}$ can be shown in an analogous way. Thus, using Lemma 5.3,

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

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

Roughly speaking, Theorem 5.4 guarantees that the 2ε-core of both, C_0 and C_1 is classified correctly (with high probability) if the memory $\mathcal{M}$ 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 $C_1 \subseteq \mathcal{S}$ that can be represented as the union of a finite set of regions bounded by closed curves with total length of at most L [11]. Moreover, let $\mathfrak{P}_\beta$ denote the class of probability distributions μ over $\mathcal{S}$ such that $\mu(A) \leq \mu_L(A) \cdot \beta$ for all Borel-subsets $A \subseteq \mathcal{S}$, where μ_L is the Lebesgue measure and $\beta > 0$.

Theorem 5.5. The concept class $\mathcal{C}$ is polynomially learnable with respect to $\mathfrak{P}_\beta$ by means of the possibilistic concept approximation (C_0^{est}, C_1^{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 5.4 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 (5.40) 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. □

5.4.3 From instances to rules

As already mentioned in previous chapters, selecting appropriate cases to be stored in the memory $\mathcal{M}$ is an important issue in case-based reasoning and instance-based learning that has 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 cases which are actually not necessary to achieve good predictive performance. For example, consider the problem of concept learning and 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 cases from the 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 cases and of generalizing cases into rules. This idea appears particularly reasonable in light of the close relation between PoCBI and 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 inputs (possibly) have similar outputs.

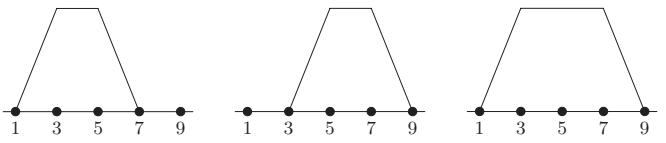


Fig. 5.3. Possibility distributions induced by two cases (left, middle) and the distribution associated with the summarizing fuzzy rule (right).

To illustrate this idea of a one-to-one correspondence between rules and cases, let $\mathcal{S} = \mathfrak{R}$, $\mathcal{R} = \{0, 1\}$ and suppose that two inputs $s_1 = 4$ and $s_2 = 6$ with $r_1 = r_2 = 0$ have been observed. The possibilistic kernels (5.13) induced by these cases are shown in Fig. 5.3. The first case is equivalent to the fuzzy rule "If s_0 is approximately 4 then $r = 0$" if the fuzzy set "approximately 4" is modeled by the possibility distribution $\delta^1_{s_0}$ (the individual support function (5.13)). The rules associated with the two cases can be merged into one rule, say, "If s_0 is about 5 then $r = 0$", where the fuzzy set "about 5" is modeled by the pointwise maximum, $\delta^1_{s_0} \vee \delta^2_{s_0}$, of $\delta^1_{s_0}$ and $\delta^2_{s_0}$ (Fig. 5.3, 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 [89, 90]. The method proposed in [327] 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 output (class label in classification), thereby generating new (pseudo-sample) prototypes [62].[20] 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. 5.3, the summarizing rule is exactly equivalent to the conjunction of the two individual rules. Of course, by weakening the requirement of equivalence, the merging procedure might also incorporate concepts of approximation and interpolation. For example, suppose $s_2 = 8$ rather than $s_2 = 6$. The replacement of $\delta^1_{s_0} \vee \delta^2_{s_0}$ by its convex hull $\delta : s \mapsto \max\{\delta^1_{s_0}(s), \delta^2_{s_0}(s), \mathbb{I}_{[5,7]}\}$ then goes beyond a simple combination since δ is larger than the pointwise maximum of $\delta^1_{s_0}$ and $\delta^2_{s_0}$ (e.g. $\delta^1_{s_0}(6) = \delta^2_{s_0}(6) = 0.5 < 1 = \delta(6)$). This kind of possibilistic induction can be reasonable and often allows for incorporating background knowledge. Particularly, replacing a possibilistic estimation δ_{s_0} by its convex hull is advised whenever a multimodal distribution does not make sense (as in our example in Section 5.3.7) or if the relation of observable cases (cf. page 22) is even known to satisfy a convexity constraint of the form

$$s \in C_r \cap C_{r''} \Rightarrow s \in C_{r'} \tag{5.41}$$

for all $r < r' < r''$.

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 cases discussed in Section 5.4.2. Indeed, deriving one rule from several cases (or other rules) can be accomplished by replacing the latter by a pseudo-case and defining an appropriate modifier function m for that pseudo-instance.

[20] Compare also with the idea of "fuzzy cases" discussed at the end of Section 5.4.1.

5.4.4 Modified possibility rules

The basic model of possibilistic CBI introduced in Section 5.2 can be rendered more flexible by making use of (linguistic) modifiers [413] in (5.7), i.e., non-decreasing functions $m_1, m_2 : [0,1] \longrightarrow [0,1]$. This leads to possibility rules $m_1 \circ A \overset{m_2}{\rightarrow} B$ with associated distributions

$$\delta_{s_0}(r) = \max_{1 \leq \imath \leq n} m_2 \left(\min \left\{ m_1(\sigma_{\mathcal{S}}(s_0, s_\imath)), \sigma_{\mathcal{R}}(r, r_\imath) \right\} \right), \tag{5.42}$$

or, when using generalized logical operators as suggested in Section 5.3,

$$\delta_{s_0}(r) = \bigoplus_{1 \leq \imath \leq n} m_2 \left(m_1(\sigma_{\mathcal{S}}(s_0, s_\imath)) \otimes \sigma_{\mathcal{R}}(r, r_\imath) \right).$$

Both modifiers in (5.42) control the extent to which a sample case is extrapolated, i.e., the extent to which other (hypothetical) cases are supported by an observation. The larger (in the sense of the partial order of functions on $[0,1]$) m_1 and m_2 are, the stronger (in the sense of asserted possibility degrees) a case $\langle s_\imath, r_\imath \rangle$ is extrapolated.

The modification (5.42) can be interpreted in different ways. Let us first consider the function m_1. In connection with the linguistic modeling of fuzzy concepts, modifiers such as $x \mapsto x^2$ or $x \mapsto \sqrt{x}$ are utilized for depicting the effect of linguistic hedges such as "very" or "almost" [413]. Applying the modifier m_1 defined by the mapping $x \mapsto x^2$ might thus be seen as replacing the original hypothesis that "similar inputs (possibly) induce similar outcomes" by the weaker assumption that only "*very* similar situations (possibly) induce similar outcomes." Thus, one interpretation of (5.42) is that of adapting the CBI hypothesis and, hence, the inference mechanism (but of maintaining the similarity measures): "The more two inputs are m_1-similar in the sense of $\sigma_{\mathcal{S}}$, the more possible it is that the respective results are (at least) similar in the sense of $\sigma_{\mathcal{R}}$."

According to a second interpretation the similarity measure $\sigma_{\mathcal{S}}$ is replaced by the measure $\sigma'_{\mathcal{S}} = m_1 \circ \sigma_{\mathcal{S}}$ in such a way that the CBI hypothesis applies in its original form:[21] "The more two inputs are similar in the sense of $\sigma'_{\mathcal{S}}$, the more possible it is that the respective results are (at least) similar in the sense of $\sigma_{\mathcal{R}}$." Roughly speaking, not the hypothesis is adapted to similarity, but similarity to the hypothesis. The extreme example $m_1 = \mathbb{I}_{\{1\}}$, indicating that the CBI hypothesis is not satisfied at all, again reveals that a similarity measure which is reasonable in the sense of inducing an appropriate extrapolation of observations does not necessarily appear natural. Indeed, interpreting $\sigma'_{\mathcal{S}} = m_1 \circ \sigma_{\mathcal{S}}$ as an improved measure suggests that inputs are not comparable at all.

The modifier m_2 does not act on a similarity measure but on the possibility-qualifying part of a rule. It can be thought of as modifying the possibility distribution

[21] One has to be careful with this interpretation, since modified measures do not necessarily inherit all (mathematical) properties of the original relations.

$$(s, r) \mapsto \max_{1 \leq \imath \leq n} \min\{m_1(\sigma_{\mathcal{S}}(s, s_\imath)), \sigma_{\mathcal{R}}(r, r_\imath)\} \tag{5.43}$$

associated with the possibility rule $m_1 \circ A \rightharpoonup B$. In fact, it allows for modeling rules of the form "for m_1-similar inputs it is m_2-possible that the respective results are similar," where "m_2-possible" stands for expressions like "more or less possible." Linguistic hedges such as "more or less" basically bring about a discounting of the distribution (5.43) and, hence, of the rule $m_1 \circ A \rightharpoonup B$.

Discounting a possibility distribution δ can be accomplished in different ways. A simple approach which is also applicable within the framework of qualitative possibility theory (where similarity and possibility are measured on ordinal scales) is to modify δ into $\min\{1 - \lambda, \delta\}$ [120]. The constant λ plays the role of a discounting factor and defines an upper bound to the support that can be provided by an underlying (possibility) rule. Indeed, δ remains unchanged if $\lambda = 0$. As opposed to this, the original support expressed by δ is completely annulled if the discounting is maximal ($\lambda = 1$). By taking m_2 as the mapping $x \mapsto \min\{1-\lambda, x\}$, the distribution (5.42) becomes

$$\delta_{\mathcal{C}} : (s, r) \mapsto \max_{1 \leq \imath \leq n} \min\left\{1 - \lambda, \min\left\{m_1(\sigma_{\mathcal{S}}(s, s_\imath)), \sigma_{\mathcal{R}}(r, r_\imath)\right\}\right\}. \tag{5.44}$$

Note that the similarity measure $\sigma_{\mathcal{R}}$ is not modified directly. Thus, it somehow determines the granularity of the extrapolation and, hence, the possibilistic approximation (5.44).

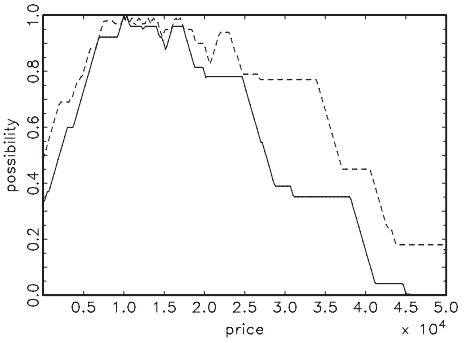


Fig. 5.4. Prediction (5.8) of the price of a car based on the original hypothesis (dashed line) and its modified version (5.44).

Example 5.6. Reconsider Example 5.1 with the hypothesis that "it is completely possible that cars with *very* similar horsepower have similar prices." Applying the modifier $m_1 : x \mapsto x^2$ to the similarity relation σ_{hp} and modeling the

(non-)effect of "completely" by $\lambda = 0$, the prediction δ_{s_0} based on (5.44) yields the possibility distribution shown in Fig. 5.4. Compared to the prediction (5.8), the degree of possibility is smaller for most of the prices $r \in \mathcal{R}$. This is caused by the fact that the CBI hypothesis is now modeled in a more cautious way. □

5.4.5 Combination of several rules

Rather than making use of a single possibility rule, the CBI hypothesis can be expressed by means of a combination (conjunction) of several rules. Suppose m such rules to be specified. Denoting by $\delta^k_{s_0}$ the possibility distribution (5.8) induced by the k-th rule ($1 \leq k \leq m$), the overall prediction is then given by

$$\delta_{s_0}(r) = \delta^1_{s_0}(r) \vee \delta^2_{s_0}(r) \vee \ldots \vee \delta^m_{s_0}(r). \tag{5.45}$$

The *disjunctive* combination in (5.45) shows that an outcome can be supported by any observed case in connection with any rule. Notice that each rule might involve different similarity relations, or different modifications of basic relations. Within our framework, it seems particularly interesting to compose new measures from a set of elementary relations (associated with individual attributes) by means of fuzzy set-based modeling techniques.

Suppose, as in the Example 5.1, that an attribute–value representation is used in order to characterize cases. That is, let inputs correspond to vectors $s = (a_1, \ldots, a_L) \in \mathcal{S} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_L$, where $\mathcal{A}_\jmath$ denotes the domain of the $\jmath$-th attribute. Moreover, let $\sigma_\jmath$ be an elementary similarity relation defined over $\mathcal{A}_\jmath$. By making use of logical connectives, the antecedent part of a possibility rule can then be composed of these elementary measures or modified versions thereof. Restricting ourselves to the logical connective $\wedge$, we obtain rules of the form

$$m_{11}(\sigma_1(a_1, a'_1)) \wedge \ldots \wedge m_{1L}(\sigma_L(a_L, a'_L)) \stackrel{m_2}{\rightarrow} \sigma_{\mathcal{R}}(r, r'). \tag{5.46}$$

Such rules can also be expressed as $\sigma'_{\mathcal{S}} \stackrel{m_2}{\rightarrow} \sigma_{\mathcal{R}}$, where

$$\sigma'_{\mathcal{S}}(s, s') = \bigotimes_{1 \leq \jmath \leq L} m_{1\jmath}(\sigma_\jmath(a_\jmath, a'_\jmath)), \tag{5.47}$$

provided that the elementary similarity relations in (5.47) are commensurate.

Of course, the antecedent part in (5.46) can be generalized such that only some of the attributes are used, i.e., each rule can concern different attributes. Leaving the $\jmath$-th attribute out of account can be interpreted in two ways. Firstly, this attribute might be irrelevant for the similarity of inputs, which is adequately reflected by $m_{1\jmath} \equiv 1$. Secondly, the rule might be interpreted as expressing a ceteris paribus condition, i.e., it might be assumed implicitly that $a_\jmath = a'_\jmath$. In this case, $m_{1\jmath}$ should be defined as $m_{1\jmath}(1) = 1$ and $m_{1\jmath}(x) = 0$ for $0 \leq x < 1$.[22] For

[22] Besides, $\sigma_\jmath$ should be separating.

example, when saying that two cars with similar horsepower have similar prices, it might be taken for granted that both cars have the same type of aspiration (standard or turbo).

Suppose that m possibility rules have been defined by using the same modifier m_2. Moreover, let $\sigma_{\mathcal{S}}^k$ $(1 \leq k \leq m)$ denote the (aggregated) measure (5.47) associated with the antecedent part of the k-th rule. Thus, the rules specify different conditions (in the form of conjunctions of similarity relations between attributes) which allow for drawing the same conclusion. The m individual rules are then equivalent to one (aggregated) rule of the form $\sigma_{\mathcal{S}} \overset{m_2}{\rightarrow} \sigma_{\mathcal{R}}$, where

$$\sigma_{\mathcal{S}}(s, s') = \bigoplus_{1 \leq k \leq m} \sigma_{\mathcal{S}}^k(s, s').$$

That is, the antecedent part of the aggregated rule corresponds to the disjunction of the antecedent parts of the individual rules.

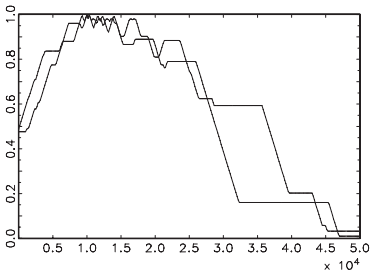


Fig. 5.5. Prediction (5.42) of the price of a car with horsepower 100, engine-size 110 and peak-rpm 5500, induced by two different rules.

Example 5.7. Reconsider Example 5.1 and let the following rules be given: (1) Cars with *very similar* horsepower possibly have similar prices. (2) Cars with *similar* engine-size and *approximately similar* peak-rpm (revolutions per minute) possibly have similar prices. Making use of the similarity measures $\sigma_{eng}(x, x') = \max\{1 - |x - x'|/100, 0\}$ and $\sigma_{rpm}(x, x') = \max\{1 - |x - x'|/1000, 0\}$, respectively, and modeling the effect of the linguistic hedge "approximately" by means of $x \mapsto \sqrt{x}$, the two rules yield the two predictions shown in Fig. 5.5. The overall prediction associated with the conjunction of the rules (i.e., the disjunction of the two premises) corresponds to the pointwise maximum of these distributions. □

Of course, different rules (5.46) will generally use different modifiers m_2. They should then be consistent in the sense that a strengthening of the antecedent

part of a rule does not entail a reduction of extrapolation. Thus, consider two rules (5.46) modeled by means of modifiers $m^1_{1\jmath}, m^1_2$ and $m^2_{1\jmath}, m^2_2$ $(1 \leq \jmath \leq L)$, respectively. The first rule is obviously redundant with respect to the second one if

$$\forall\, 1 \leq \jmath \leq L : m^1_{1\jmath} \leq m^2_{1\jmath} \quad \text{and} \quad m^1_2 \leq m^2_2.$$

In fact, we then have $\delta^1_{s_0} \leq \delta^2_{s_0}$ for the possibility distributions induced by these two rules in connection with any observed case.

Consider the following rules as an example: (1) For cars with *similar* horsepower it is *completely* possible that the associated prices are similar. (2) For cars with *very similar* horsepower it is *more or less* possible that the associated prices are similar. This example reveals that redundancy always emerges in connection with somewhat conflicting rules (a stronger condition entails a weaker conclusion). Therefore, redundant rules should be avoided.

5.4.6 Locally restricted extrapolation

So far, the possibility rules which define a model of the CBI hypothesis have been used *globally* in the sense that they apply to all cases of the input-output space $\mathcal{S} \times \mathcal{R}$. Needless to say, the CBI hypothesis does not necessarily apply equally well to all parts of this space. That is to say, the degree of extrapolation of a case $\langle s, r \rangle$ that can be justified by the CBI hypothesis might depend on the region to which it belongs.

In the Automobile Database database (cf. Example 5.1), for instance, the variance of the price is smaller for cars with aspiration "turbo" than for cars with aspiration "standard" (even though the average price is higher for the former). Thus, the hypothesis that similar cars possibly have similar prices seems to apply better to turbo than to standard cars. Likewise, a statistical analysis suggests that the variation of the price is an increasing function of the size of cars. Again, the smaller a car is, the better the CBI hypothesis seems to apply (at least if the similarity of two lengths x, x' is a function of $|x - x'|$). Consequently, the extrapolation of case-based information should be larger for small cars than for large cars.

In order to adapt the formalization of the CBI hypothesis one might think of defining different rules for different regions of the input space. Restricting the application of a rule to a certain (fuzzy) range of this space can be accomplished by means of a fuzzy partition $\mathcal{F}$ of $\mathcal{S}$. The condition part of a rule then appears in the form

$$F(s) \wedge F(s') \wedge m_1(\sigma_{\mathcal{S}}(s, s')), \tag{5.48}$$

where the fuzzy set $F \in \mathcal{F}$ is identified by its membership function $F : \mathcal{S} \longrightarrow [0, 1]$. The antecedent (5.48) can be associated with an extended possibility rule "the more both inputs are in F *and* the more similar they are, the more possible it is that the related outcomes are similar." This way, one might express, for

instance, that "it is completely possible that small cars of similar size have similar prices" and "it is more or less possible that large cars of similar size have similar prices." The fuzzy set F in (5.48) is then given by the set of small cars and large cars, respectively. Note that the attribute "aspiration" defines a crisp rather than a fuzzy partition.

On the basis of (5.48), the inference scheme (5.42) becomes

$$\delta_{s_0}(r) = \max_{1 \leq \imath \leq n} \min \big\{ F(s_0), F(s_\imath), \; m_2 \left(\min \left\{ m_1(\sigma_{\mathcal{S}}(s_0, s_\imath)), \sigma_{\mathcal{R}}(r, r_\imath) \right\} \right) \big\} . \tag{5.49}$$

Note that $\delta_{s_0} \equiv 0$ as soon as $F(s) = 0$, thus expressing that a rule has no effect outside its region of applicability. Besides, it is worth mentioning that (5.49) is closely related to ideas of discounting as discussed in previous sections. This becomes especially apparent when writing (5.49) in the form

$$\delta_{s_0}(r) = \max_{1 \leq \imath \leq n} m_{2\imath}(x_\imath), \tag{5.50}$$

with $x_\imath = \min\{m_1(\sigma_{\mathcal{S}}(s_0, s_\imath)), \sigma_{\mathcal{R}}(r, r_\imath)\}$ and $m_{2\imath} : x \mapsto \min\{F(s_0), F(s_\imath), m_2(x)\}$. In fact, (5.50) shows that the original support provided by the cases is discounted by means of the modifiers $m_{2\imath}$. As opposed to (5.44), however, this is not realized by using a constant factor λ. Rather, the discounting of a rule now depends on the inputs s and $s_\imath$ to which it is applied.

5.4.7 Incorporation of background knowledge

Our fuzzy set-based framework is also well-suited for incorporating background knowledge of more general nature (i.e., not necessarily related to similarity). This becomes especially apparent if such knowledge is also expressed in terms of fuzzy rules. For instance, an expert might be willing to agree that "a price of slightly more than \$40,000 for a car with horsepower of approximately 200 is completely possible." This can be formalized as a possibility rule $A \rightharpoonup B$, where A and B model the fuzzy sets of "approximately 200" and "slightly more than \$40,000." Such a rule can simply be added to the rule base induced by the memory of cases (cf. Section 5.4.3), thereby supplementing the "empirical" evidence which comes from observed cases.

A special type of (rule-based) background knowledge can be obtained by specifying "fictitious cases". One might specify, for instance, a fictitious car by means of some attribute values (which can be uncertain or vague) and then ask an expert for a typical (or possible) price. The fictitious observation thus defined can principally be treated in the same way as an observed one. This type of reasoning provides a convenient way of filling up sparse memories. It is also interesting from a knowledge acquisition point of view. Indeed, from a user (expert) perspective

it might appear less difficult to give some specific examples (e.g., by estimating prices of hypothetical cars) than to specify universally valid rules.

Apart from fuzzy rules, more general types of constraints can be used for expressing background knowledge. A nice example is the convexity constraint (5.41) according to which intermediary predictions are not less possible than more extreme ones. In order to satisfy such a constraint, a possibility distribution δ_{s_0} can simply be replaced by its convex hull (see (5.31) in Section 5.3.7).

5.5 Experimental studies

5.5.1 Preliminaries

This section presents some experimental studies providing evidence for the excellent practical performance of the possibilistic approach to case-based inference. More specifically, we shall focus on simple classification problems and investigate the PossIBL algorithm as introduced in Section 5.4.2. As in previous chapters, however, we would like to emphasize that our experiments are not meant as an exhaustive comparative study covering several competing learning algorithms – and showing that PossIBL is superior to all of its competitors. In fact, one should realize that the primary motivation underlying PossIBL (or, more generally, PoCBI) is not another ε-improvement in classification accuracy but rather the enrichment of instance-based learning (case-based reasoning) 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 5.3.7) 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, regarding other IBL algorithms, a comparison 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 reasons, we have decided to apply a basic version of PossIBL to several data sets from the UCI repository[23] 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 (2.9)). 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 5.5.4). Instead, we have applied

[23] `http://www.ics.uci.edu/~mlearn`.

the learning scheme from Section 5.4.2 with the original max–min version (5.8). The function $m_\imath$ in (5.34) was defined as $t \mapsto \exp(-\gamma_\imath\,(1-t))$, where $\gamma_\imath \geq 0$ is the discounting rate of the $\imath$-th case. The constant β in (5.35) was taken as 0.8.[24] 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 $\sigma_{\mathcal{S}}$ is always defined as 1 minus the normalized Euclidean distance and the similarity $\sigma_{\mathcal{R}}$ is given by (5.11).

5.5.2 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 memory $\mathcal{M}$) 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 memory any further), and the percentage of correct decisions is determined. Statistics are obtained by means of repeated 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 5.1. Results for the 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	0.9574	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

Table 5.2. Results for the Iris Plant Database (150 observations, 4 predictive attributes, three classes, training set of size 75, 10,000 simulation runs).

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

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 5.3. Results for the 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	0.7096	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 5.4. Results for the 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	0.7148	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

Table 5.5. Results for the Wine Recognition Data (178 observations, 13 predictive attributes, three classes, training set of size 89, 1,000 simulation runs).

Results are summarized in Tables 5.5.2–5.5.2 by means of statistics for the percentage of correct classifications (mean, standard deviation, minimum, maximum, 0.1–fractile, 0.9–fractile). The experiments show that PossIBL achieves comparatively good results and is always among the best algorithms. Thus, it is valid to conclude that even a very basic version of PossIBL performs 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.

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 $\mathcal{R}$, one might wonder how to explain the different performance of PossIBL and the NN classifiers. In fact, in Section 5.3.6 it was argued that the possibilistic NN decision derived from (5.8) 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.

5.5.3 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 page 185). This is clearly true for the possibilistic estimation δ_{s_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 as follows: The instance space is defined by $\mathcal{S} = \mathfrak{R}$, the set of class labels is $\mathcal{R} = \{-1, +1\}$, the class probabilities are $1/2$, the conditional probability density of the input s given the outcome r is normal with standard deviation 1 and mean r. 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 ($0 < \alpha \leq 1/2$). Based on the resulting training set, which is not "fully representative" in the sense of [78], 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 $f(\alpha)$ is derived by averaging over 10,000 simulation runs.

Fig. 5.6 shows the misclassification rates for several methods. As was to be expected, $f(\cdot)$ 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.

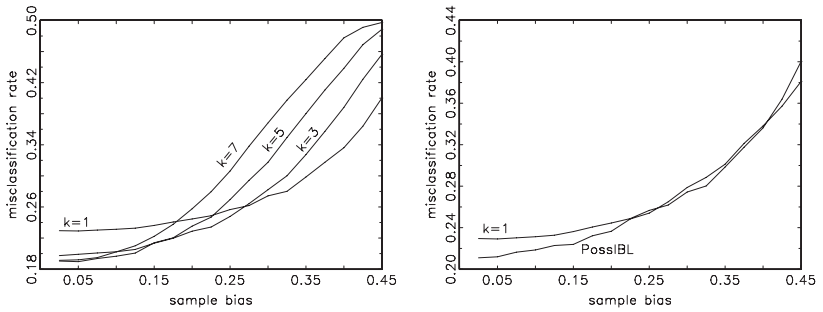


Fig. 5.6. Misclassification rates of kNN methods (left) and PossIBL (right, in comparison with 1NN).

Apart from kNN methods, we have tested PossIBL with $\oplus = \oplus_P$. The similarity measure $\sigma_{\mathcal{S}}$ 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 5.3, 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 $k > 1$.

Needless to say, what we considered here is only a particular setup in which PossIBL appears to be superior to standard kNN with regard to robustness. As robustness is a very multi-faceted aspect, one should not overlook that our results are preliminary and of limited significance.

5.5.4 Variation of the aggregation operator

An interesting question concerns the dependence of PossIBL's performance on the specification of the aggregation operator $\oplus$ in (5.13). To get a first idea of this dependence, we have performed the same experiments as described in

Section 5.5.2 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 (5.18), 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 results are presented in Appendix E. 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 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 $k > 1$.

5.5.5 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 5.5.1.

Let D_1 denote the expected difference (margin) between the possibility degree of the predicted label r_0^{est} and the possibility degree of the second best label, given that the prediction is correct:

$$D_1 \stackrel{\text{df}}{=} \delta_{s_0}(r_0) - \max_{r \in \mathcal{R}, r \neq r_0} \delta_{s_0}(r). \tag{5.51}$$

Moreover, let D_0 denote the expected difference between the possibility degree of the predicted label r_0^{est} and the possibility degree of the actually true label r_0, given that $r_0 \neq r_0^{est}$:

$$D_0 \stackrel{\text{df}}{=} \delta_{s_0}\left(r_{s_0}^{est}\right) - \delta_{s_0}\left(r_{s_0}\right). \tag{5.52}$$

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 $r \neq r_0$.

Table 5.5.5 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.

Dataset	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 5.6. Statistics (5.51) and (5.52) for PossIBL.

5.6 Calibration of CBI models

The methodological framework introduced in previous sections provides a broad spectrum of techniques for building a CBI model. Needless to say, it would be unrealistic to expect a human expert using these (linguistic) modeling techniques to come up with precise mathematical formalizations of related fuzzy concepts. Instead, a more reasonable approach is to let the expert specify the coarse structure of a model, in our case the fuzzy rules modeling the CBI hypothesis, and to determine the ultimate model in a second step by adapting the expert model to the observed data. This is to some extent comparable, say, to graphical modeling techniques such as Bayesian networks, where the user specifies the structure of the network (i.e., the qualitative part of the model), and the (conditional) probability distributions (i.e., the quantitative part) is learned from data.

In Section 5.4.2, we have already presented a learning scheme for adapting a possibilistic model to the application at hand, albeit for a very particular case (namely PossIBL, our possibilistic variant of IBL). This section is meant to discuss model calibration in more general terms, including the determination of similarity measures and modifier functions. More specifically, we consider the problem of determining modifiers m_1 and similarity measures $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$ in a set of rules of the form $m_1 \circ \sigma_{\mathcal{S}} \rightharpoonup \sigma_{\mathcal{R}}$. Each of these rules induces a related possibility distribution (5.7) or, when using aggregation operators other than max and min, the generalized version

$$(s, r) \mapsto \bigoplus_{1 \leq \imath \leq n} m_1(\sigma_{\mathcal{S}}(s, s_\imath)) \otimes \sigma_{\mathcal{R}}(r, r_\imath). \tag{5.53}$$

The overall distribution $\delta_{\mathcal{C}} : \mathcal{S} \times \mathcal{R} \longrightarrow [0,1]$, considered as a lower approximation of the relation φ in (5.4), is given by the union (pointwise maximum) of these distributions.

The basic idea is to proceed from similarity measures and modifiers which are specified in the form of parameterized functions. For instance, the modifier associated with the linguistic hedge "very" might be specified by the function $x \mapsto x^{\alpha}$ with $\alpha > 1$. Likewise, the similarity of horsepowers, σ_{hp}, might be given by the function

$$(x, x') \mapsto \max\left\{1 - \frac{|x - x'|}{M}, 0\right\}, \tag{5.54}$$

where M plays the role of a parameter (cf. Example 5.1). All these parameters can be combined into one vector θ which determines the CBI model and, hence, has a strong influence on the generalization beyond (via extrapolation of) observed cases. In this sense, it plays a role somewhat similar to, e.g., the smoothing parameter in kernel-based estimation of probability density functions.

In order to determine θ and, hence, a concrete CBI model from the memory $\mathcal{M}$ of observed cases, a kind of optimization criterion is needed. A reasonable idea is to minimize some distance, such as

$$\int_{\mathcal{C}} \left(\delta_{\mathcal{C}}(c \,|\, \theta) - \delta_{\varphi}(c)\right)^2 \, dc, \tag{5.55}$$

between the estimated distribution $\delta_{\mathcal{C}}(\cdot \,|\, \theta)$ and the (true) $\{0,1\}$-valued distribution δ_{φ} defined by $\delta_{\varphi}(c) = 1 \Leftrightarrow c \in \varphi$.

This is quite comparable with the determination of the *kernel width* or *smoothing parameter* h in kernel-based density estimation, where an underlying density function ϕ is estimated by

$$\phi_h : x \mapsto \frac{1}{n} \sum_{i=1}^{n} \kappa_h \left(x - x_i\right) = \frac{1}{n} \sum_{i=1}^{n} \kappa\left(\frac{x - x_i}{h}\right), \tag{5.56}$$

with κ being the *kernel function*.[25] The smoothing parameter h has an important effect on the accuracy of the approximation (5.56). It plays a role somewhat similar to the bin-width of histograms. One way of determining this parameter is to minimize the integrated squared error

$$\mathsf{ISE}(h) = \int \left(\phi(x) - \phi_h(x)\right)^2 \, dx \tag{5.57}$$

between the true density ϕ and the estimation ϕ_h.

Unfortunately, (5.57) cannot be derived since the true density ϕ is unknown, and the same remark of course also applies to (5.55), where $\pi_{\varphi}(c)$ is not known

[25] Typical examples of κ include the PARZEN window $u \mapsto \mathbb{I}_{[-1/2,1/2]^m}$ [289] and the normal kernel, the latter being defined as the density of the (multivariate) standard normal distribution.

for all $c \in \mathcal{C}$. A possible way out is to replace the true approximation error by an empirical error, namely the error for the observed cases. This can be done by means of a (leave-one-out) cross validation procedure which, in the case of kernel-based density estimation, approximates the integral by a weighted sum and replaces the density ϕ by a further estimation $\widehat{\phi}$ [185]. This leads to the minimization of

$$\sum_{\imath=1}^{n} \left(\widehat{\phi}_h(x_\imath) - \phi_h(x_\imath) \right)^2, \tag{5.58}$$

where $\widehat{\phi}_h(x_\imath)$ denotes the estimated (cross validation) density for the $\imath$-th observation $x_\imath$. Again, this value is obtained by means of a kernel-based estimation (using h as a smoothing parameter). As opposed to the derivation of $\phi_h(x_\imath)$, however, this estimation leaves the point $x_\imath$ itself out of account, i.e., it uses only the observations $\{x_1, \ldots, x_{\imath-1}, x_{\imath+1}, \ldots, x_n\}$.

The same idea can also be applied to (5.55). In this case, we do not even have to estimate the values $\delta_\varphi(c_\imath)$ since $\delta_\varphi(c_\imath) = 1$ holds true for each observation $c_\imath \in \mathcal{M}$. However, by restricting ourselves to the observed cases, the minimization problem becomes ill-posed. In fact, a trivial solution to the problem of minimizing

$$\sum_{c \in \mathcal{M}} \left(\delta_{\mathcal{C}}(c \,|\, \theta) - \delta_\varphi(c) \right)^2 \tag{5.59}$$

is given by $\delta_{\mathcal{C}}(\cdot \,|\, \theta) \equiv 1$. This simply means to choose the parameter θ such as to maximize the extrapolation of cases, a hardly convincing result.

In this connection, recall the problem that a possibilistic prediction $\delta_{\mathcal{C}}$ can principally not be "falsified" (cf. Section 5.4.2): The *non-support* of an actually *observed* case can be justified by the fact that no cases have (as yet) been observed which are similar enough. Thus, a small value $\delta_{\mathcal{C}}(c \,|\, \theta)$ is not necessarily a defect of the model, i.e., it does not necessarily indicate a poor choice of the parameter θ. (Predicted possibility degrees are only lower bounds, and low degrees are quite natural if the memory $\mathcal{M}$ does not contain many cases similar to c!) Moreover, it is hardly possible to object to the *support* of a yet *unobserved* case since it would require knowledge about the non-existence of that case (which is of course not available). As can be seen, the model based on possibility rules only indicates which cases are (provably) *possible*. It does not, however, point to those cases which appear *impossible*. In other words, the possibilistic model merely expresses the *support* but not the *exclusion* of cases. This contrasts with a probabilistic approach, where an event cannot be supported without (partly) excluding its complement at the same time.

Fortunately, as already pointed out in Sections 5.3.3 and 5.4.2, the (partial) exclusion of cases according to the CBI principle can be realized by means of a complementary type of extrapolation principle induced by a different sort of fuzzy rule, called certainty rule. The latter entails the distribution

$$(s,r) \mapsto \bigotimes_{1\le\imath\le n} (1-\sigma_{\mathcal{S}}(s,s_\imath)) \oplus \sigma_{\mathcal{R}}(r,r_\imath) \tag{5.60}$$

which actually represents upper bounds and thus defines the counterpart to (5.53). The overall prediction $\pi_{\mathcal{C}}$, associated with a set of rules of that type, is defined by the intersection (pointwise minimum) of the distributions (5.60). As can be seen, a certainty rule reduces the possibility of hypothetical cases which are somehow in conflict with observed cases, in the sense that the inputs are similar but the outcomes are rather different.

Example 5.8. Reconsider Example 5.1 with a case $(100, 15000)$, i.e., a car with horsepower 100 and price \$15,000. In connection with the similar horsepower–similar price hypothesis and the possibility rule model (5.53), this case (partly) *supports* the case $(110, 16000)$ which has a similar horsepower and a similar price. According to the certainty rule model (5.60), it (partly) *excludes* the case $(110, 5000)$ which has a similar horsepower but a rather different price. Observe that the possibility rule model will generally say little about the case $(110, 5000)$, as expressed by a small lower possibility bound. Likewise, the certainty rule model has not much to say about the car $(110, 16000)$ to which it assigns a large upper bound. □

In connection with the determination of optimal similarity measures and modifiers, the two models can complement each other in a reasonable way.[26] As already pointed out in Section 5.3.3, the prediction $\delta_{\mathcal{C}}$ derived from (5.53) and the prediction $\pi_{\mathcal{C}}$ obtained from (5.60) might be *conflicting* in the sense that $\pi_{\mathcal{C}}(c) < \delta_{\mathcal{C}}(c)$ for a case c. This can happen if c is supported by some observation $c_1 \in \mathcal{M}$ (according to the possibility rule model) and, at the same time, excluded by another observation $c_2 \in \mathcal{M}$ (according to the certainty rule model). A situation of this kind indicates a defect of the underlying CBI model (the lower possibility bound is larger than the upper bound). It occurs if a case c is similar to both, c_1 and c_2 (in the sense of the similarity measure $\sigma_{\mathcal{S}}$), and if c_1 indicates a result which is quite different (in the sense of $\sigma_{\mathcal{R}}$) from the one suggested by c_2. Besides, it should be noticed that a more or less isolated case c does not involve any conflict, since $\delta_{\mathcal{C}}(c)$ and $\pi_{\mathcal{C}}(c)$ will be close to 0 and 1, respectively.

Example 5.9. Suppose, for instance, that we have observed the cars $c_1 = (50, 5000)$, $c_2 = (100, 15000)$, and $c_3 = (75, 7000)$ and that we only distinguish between similar and dissimilar horsepowers resp. prices:

$$\sigma_{\mathcal{S}}(x,y) = \begin{cases} 1 & \text{if} \quad |x-y| \le \Delta \\ 0 & \text{if} \quad |x-y| > \Delta \end{cases},$$

$$\sigma_{\mathcal{R}}(x,y) = \begin{cases} 1 & \text{if} \quad |x-y| \le 5000 \\ 0 & \text{if} \quad |x-y| > 5000 \end{cases}.$$

[26] The joint use of lower and upper possibility bounds (derived, respectively, from possibility and certainty rules) has also been advocated in the context of approximate reasoning [376, 393].

For $\Delta = 30$, c_1 qualifies the case c_3 as being (completely) possible. However, since $\sigma_{\mathcal{S}}(75, 100) = 1$ as well, c_3 is disqualified by c_2 at the same time. This suggests to choose a smaller value for Δ, since otherwise the similar horsepower–similar price rule becomes invalid. More generally, a memory of n cases $\langle s_i, r_i \rangle$ calls for

$$\Delta \leq \min_{1 \leq i,j \leq n,\, \sigma_{\mathcal{R}}(r_i, r_j) = 1} |s_i - s_j|$$

in order to satisfy this rule. As can be seen, the stronger the variability in the horsepower–price relation is, the more restrictive the similarity between horsepowers has to be defined. In the more general case where similarity measures are not $\{0, 1\}$-valued, a conflict might appear in a less obvious way, and the degree to which the CBI hypothesis is satisfied can vary gradually. □

The above example reveals the following effect: The more similar the cases are made (through the definition of corresponding similarity measures and modifiers), the stronger is the degree of support resp. exclusion induced by a set of observations according to (5.53) resp. (5.60) and, hence, the larger the conflict becomes. Here, we take advantage of this effect in order to define meaningful modifier functions and measures of similarity. In fact, a reasonable optimization criterion is to find a tradeoff between a principle of *appropriate support* (of observed cases) and a *consistency* principle:

– Observed cases should be supported as much as possible by the other cases in the memory (e.g., in connection with a leave-one-out cross-validation).
– The conflict between the support and exclusion of these cases should be as small as possible.

Formally, we define the support attached to a case $c \in \mathcal{M}$ by

$$\mathsf{supp}_\theta(c) \stackrel{\text{df}}{=} \delta_{\mathcal{C}}(c \,|\, \theta), \tag{5.61}$$

where $\delta_{\mathcal{C}}(\cdot \,|\, \theta)$ is derived from $\mathcal{M} \setminus \{c\}$ according to (5.53) and $m_1, \sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}$ are determined by the parameter vector θ. Moreover, the conflict associated with the case c can be defined as

$$\mathsf{conf}_\theta(c) \stackrel{\text{df}}{=} \max\{0, \delta_{\mathcal{C}}(c \,|\, \theta) - \pi_{\mathcal{C}}(c \,|\, \theta)\}, \tag{5.62}$$

where $\pi_{\mathcal{C}}(c\,|\,\theta)$ is the distribution obtained from the certainty rule model (5.60). Note that, in the case where possibility is interpreted as an ordinal concept, one might think of replacing the subtraction in (5.62) by a purely qualitative measure of conflict:

$$\mathsf{conf}_\theta(c) = \begin{cases} 1 & \text{if} \quad \pi_{\mathcal{C}}(c\,|\,\theta) < \delta_{\mathcal{C}}(c\,|\,\theta) \\ 0 & \text{if} \quad \pi_{\mathcal{C}}(c\,|\,\theta) \geq \delta_{\mathcal{C}}(c\,|\,\theta) \end{cases} .$$

The derivation of (5.61) and (5.62) for all cases in the memory yields n degrees of support and conflict, respectively. The overall support induced by the parameter θ, $\mathsf{supp}(\theta)$, can then be obtained by aggregating these values:

$$\mathsf{supp}(\theta) = A(\{\mathsf{supp}_\theta(c)\,|\,c \in \mathcal{M}\}) \tag{5.63}$$

with A being an aggregation function. A measure $\mathsf{conf}(\theta)$ of conflict can be defined analogously. Finally, an optimal parameter θ is derived as a function of the support and the conflict thus defined, e.g., by maximizing

$$\mathsf{supp}(\theta) - \alpha \cdot \mathsf{conf}(\theta) \tag{5.64}$$

for some tradeoff parameter $\alpha \geq 0$ or by maximizing $\mathsf{supp}(\theta)$ under the condition that $\mathsf{conf}(\theta) \leq \alpha$.

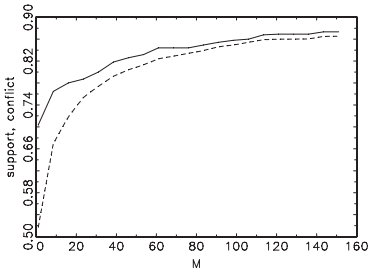


Fig. 5.7. Support (solid line) and conflict as a function of the parameter M which defines the similarity measure for the attribute horsepower.

In order to combine the degrees of support (conflict) associated with individual cases, one might use a simple average as an aggregation function A in (5.63). Alternatively, an aggregation which is more in accordance with a qualitative setting is the Sugeno integral

$$\int^{su} \mathsf{supp}_\theta \, d\mu = \sup_{\alpha \geq 0} \min\{\alpha, \mu(F_\alpha)\}, \tag{5.65}$$

where $F_\alpha = \{c \in \mathcal{M}\,|\,\mathsf{supp}_\theta(c) \geq \alpha\}$ for $0 \leq \alpha \leq 1$. The measure μ in (5.65) can be taken as the counting measure, i.e., $\mu(A) = |A|/|\mathcal{M}|$ for all $A \subseteq \mathcal{M}$.

EXAMPLE 5.10. Consider as a simple example the choice of the parameter M in (5.54) which defines the similarity measure σ_{hp} in connection with the similar horsepower–similar price hypothesis (using the same function with $M = 3000$ for the similarity $\sigma_{\mathcal{R}}$). Fig. 5.7 shows $\mathsf{supp}(M)$ and $\mathsf{conf}(M)$, defined according to (5.61), (5.62), and the aggregation (5.65) as a function of M. The choice of $\alpha = 3/4$ in (5.64) suggests $M = 76$ as an optimal parameter and leads to the prediction shown in Fig. 5.8. □

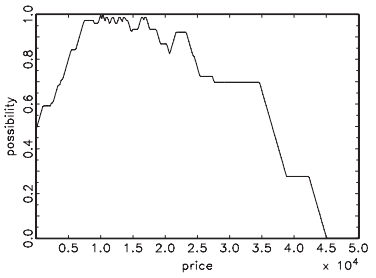


Fig. 5.8. Prediction of the price of a car with horsepower 100, where σ_{hp} is given by (5.54) with $M = 76$.

REMARK 5.11. The calibration method outlined above can be seen as a generalization of related probabilistic approaches. In the latter case, the support and the exclusion of a value always add up to 1. Therefore, a conflict cannot occur, and only the principle of correct support remains relevant. Note that this principle reduces to a principle of *maximal* support in the possibilistic model, as can be gathered from (5.59). In the probabilistic case, the correct support corresponds to the true probability, as expressed by (5.58). □

Let us finally mention that some standard estimation and optimization problems have to be solved in connection with a concrete application. This concerns, for example, the question whether all parameters can be identified by the optimization criterion. Besides, it should be noted that the method of finding an optimal CBI model outlined in this section amounts to solving a nonlinear optimization problem. It might hence be considered critical from the viewpoint of computational complexity, especially since a new parameter has to be derived each time the memory changes. One should realize, therefore, that a parameter estimation is usually not a time-critical problem since it can be solved "off-line." Note that the current optimal parameter can serve as a good initial value when using iterative improvement methods. In fact, a small variation of the memory, such as the adding of a new case, will generally change the optimal parameter but slightly.

5.7 Relations to other fields

This section is meant to explore relationships between the possibilistic approach to CBI outlined in previous sections (PoCBI) and some related methods. One can look at PoCBI from different directions. From the viewpoint of statistics and data analysis, it is formally somewhat similar to non-parametric (kernel-based) density estimation. However, as was already discussed in Section 5.3.5, it differs in using possibility theory and similarity instead of probability theory and frequency as major concepts. The use of fuzzy sets and possibility distributions instead of (in addition to) probability distributions is just the characteristic property that PoCBI shares with fuzzy data analysis, the fuzzy set-based counterpart (extension) to classical data analysis. Some relevant aspects of corresponding methods will be discussed in Section 5.7.1.

PoCBI combines rule-based and instance-based reasoning techniques: A memory of cases induces a set of rules and allows CBI to be realized as rule-based reasoning. Besides, Section 5.4.7 has shown that both techniques can be used in a complementary way. The combination of case-based and rule-based reasoning (as well as other hybrid approaches to machine learning) has recently received considerable attention, and it has already led to several interesting approaches [14, 61, 89, 174, 175, 246]. A combined approach is particularly advocated by the complementary merits of the two techniques, namely the suitability for representing general (background) knowledge of a domain in rule induction and specific knowledge in the form of observed cases in CBR. An obvious idea, for instance, is to use a complementary representation in which those cases are stored in the memory which are exceptions to a set of otherwise valid (default) rules. There are, however, other possibilities of combining rule induction and case-based reasoning, some of which have been realized in the Patdex system [18]. PoCBI can be considered from both directions. Since relationships between PoCBI and instance-based learning have already been discussed in Section 5.3.5, this section shall touch on some aspects in connection with more common approaches to fuzzy set-based approximate (rule-based) reasoning.

5.7.1 Fuzzy and possibilistic data analysis

The term *fuzzy data analysis* can have different meanings, depending on whether the adjective "fuzzy" refers to the observed *data* itself or to the *methods* used for analyzing the data. That is, a main differentiation must be made between the analysis of somehow uncertain or vague data (e.g., by means of generalized statistical methods [241]) and the use of fuzzy or possibilistic methods for processing data that has been observed precisely (e.g., fuzzy clustering of crisp data [32]). Fuzzy data analysis can also comprise both aspects, of course. It is then concerned with using fuzzy or possibilistic methods for supporting the analysis of vague data [22].

In connection with fuzzy data analysis it is important to distinguish between different types of incomplete knowledge, notably *uncertainty* and *imprecision*. Traditional statistical methods take the first phenomenon into consideration: The generation of data is modeled as a stochastic process, thus leading to random (but still precise) observations. The analysis of fuzzy data does not only consider uncertainty in the generation but also in the observation of data, i.e., it assumes observations to be afflicted with imprecision. In fact, the latter type of uncertainty, which must not be confused with randomness, is often present in practice. Firstly, the observed object itself can be vague in the sense that it might not be possible to identify or demarcate it exactly. Secondly, the measuring instrument or the underlying scale might not allow for identifying the (principally well-defined) object precisely. A standard example is the (linguistic) "value" of a number (which is exact as such) on a scale of linguistic expressions.

Subsequently, we shall briefly discuss some aspects of PoCBI in the context of different approaches to fuzzy data analysis. Qualitative data analysis generally aims at discovering some kind of structure or patterns in the data and, hence, is in line with desriptive statistics, exploratory data analysis, as well as much of current research in the emerging field of data mining and knowledge discovery [183]. Corresponding methods, such as (fuzzy) cluster analysis, mainly focus on single properties of the objects under study and are mainly interested in comparing the data. As in PoCBI, the concept of similarity thus plays a major role in such methods. Besides, PoCBI also helps in getting a more precise idea of the data. To this end, however, it already generalizes beyond the given observations (against the background of further knowledge), whereas qualitative methods consider these observations alone. Seen from this perspective, PoCBI might be considered as an extended form of *exploratory* or *descriptive* data analysis.

While qualitative methods focus on individual properties of an object, *quantitative* analysis is rather concerned with finding (invariant) *relations* between different features, e.g., by estimating (fuzzy) functional relationships (as supervised methods in machine learning).

Example 5.12. As a simple example of a quantitative method consider the fitting of a (parameterized) fuzzy set-valued mapping $F_\theta : \mathfrak{R} \longrightarrow \mathfrak{F}(\mathfrak{R})$ to a set of (fuzzy) observations $(x_k, Y_k) \in \mathfrak{R} \times \mathfrak{F}(\mathfrak{R})$ $(1 \leq k \leq n)$. This can be accomplished, e.g., by choosing the (fuzzy) parameter vector θ such that

$$\sum_{k=1}^{n} \|Y_k - F_\theta(x_k)\|$$

is minimized, where $\|\cdot\|$ is a (metric) distance measure on $\mathfrak{F}(\mathfrak{R})$, the class of fuzzy subsets of $\mathfrak{R}$ [86]. A further possibility is to minimize the spread of F_θ while somehow covering the data, e.g., while satisfying $Y_k \subseteq F_\theta(x_k)$ for all $1 \leq k \leq n$. The latter type of fuzzy regression analysis amounts to solving a linear programming problem if F_θ has a certain linear structure. □

Fuzzy methods like the one in Example 5.12 can be interpreted in different ways. Firstly, they can be seen as a generalized approximation (resp. interpolation) method, where scalar observations and functions are replaced by fuzzy set-valued observations and mappings, respectively. Such methods should basically be understood as describing the *given data*, as opposed to inductive statistical methods which draw conclusions about some underlying process which generates the data. For instance, the parameter θ in Example 5.12 is chosen such that F_θ fits the data optimally (e.g., in the sense of minimizing the sum of squared errors). It should not be interpreted, however, as an estimation of some true (but unknown) parameter which identifies a data-generating process. Consequently, fuzzy methods of such kind cannot fall back on a related model in order to make predictions. Rather, they have to rely on the same kind of assumptions as CBI, namely that the observations are to some degree representative and that similar outputs are generated by similar inputs [21].[27] It should be observed, however, that the extent of extrapolation (or interpolation) of outputs is principally not bounded, e.g., when fitting a fuzzy mapping to a set of observations and using that mapping for making predictions [87]. Seen from this perspective, corresponding methods seem to lack a solid basis for generalizing beyond observed data.

The use of fuzzy sets for modeling imprecision in the observation of (actually exact) data gives rise to a second interpretation which is related to possibility theory: A fuzzy set A attaches uncertainty to a crisp object (namely its core) and a degree of membership $A(x)$ is considered as the possibility of x being the true (only incorrectly observed) object. This interpretation has motivated the introduction of *possibilistic variables* as a counterpart to random variables. The related idea of a *possibilistic* generation of data leads to parameter estimation methods which parallel the maximum likelihood estimator in statistics (by using the minimum operator instead of the product) [22]. Corresponding methods thus fall into line with model-based approaches in mathematical statistics. Since each observation induces a possibility distribution $\pi = A$, this type of modeling is closely related to PoCBI. Still, the underlying semantics is very different. In the first case, *indistinguishability* is taken as a necessary evil, and $A(x)$ quantifies the possibility that the real object, x_0, is *actually given* by x. In the second case, *similarity* is exploited as a useful concept for pointing to the existence of other objects, and $\pi(x)$ is considered as the plausibility of encountering x (while knowing the current object x_0). As a further difference let us mention that the ensemble of fuzzy observations (the possibilistic data set) marks the *input* in possibilistic data analysis. It is further processed by means of generalized methods, such as possibilistic linear regression [367] or possibilistic cluster analysis [191]. In PoCBI, the union of possibility distributions principally corresponds to the output, whereas the input is given in the form of precise cases.

A third interpretation of fuzzy methods establishes a close connection between fuzzy sets (fuzzy data) and probability theory and makes use of concepts such

[27] Indeed, this assumption is implicitely made when fitting a *continuous* (fuzzy) mapping.

as like probabilistic sets [189], fuzzy random variables [241] or random fuzzy sets [303]. This approach calls for generalizations of classical statistical methods. It also leads to possibilistic reasoning methods which can be seen as a kind of approximate probabilistic inference. Let us mention the learning of possibilistic networks from data which is based on a probabilistic interpretation of possibility degrees (in terms of random sets) as an example [42, 43]. Possibilistic networks emerge from probabilistic networks (including Bayesian networks [292] and Markov networks [245]) by using possibility distributions instead of probability measures. This allows one to take uncertainty as well as imprecision into account [41]. Apart from the probabilistic semantics, they can hence be seen as the possibilistic counterpart to probabilistic networks in much the same way as POCBI can be considered as the possibilistic counterpart to kernel-based density estimation.

Graphical modeling by means of network structures is an example of a model-based approach which is capable of combining knowledge and data in various ways, a property which is often emphasized as a major benefit [187]. Typically, an expert specifies the structure of a network, i.e., the qualitative part, while the associated (conditional) probability or possibility distributions are learned from data. Compared with the use of rules (which define the qualitative part of the model in POCBI), knowledge hence appears in the form of (in)dependence relations between variables represented by means of a directed (acyclic) graph. Besides, the (conditional) probabilities or possibilities, i.e., the quantitative part of a network, correspond to the similarity measures and modifier functions in POCBI, which can be adapted to observed data by means of corresponding learning method (cf. Section 5.4.2).

In summary, POCBI has characteristics in common with both, qualitative and quantitative data analysis. It is close to qualitative approaches in making use of similarity as a basic concept and in supporting the description of data. Still, it is also concerned with generalizing and making predictions, a property it shares with possibilistic approximation or parameter estimation. As opposed to POCBI, however, such methods are mostly model-based. Besides, the meaning of a possibility distribution in POCBI greatly differs from the interpretation in the methods outlined in this section, the latter using such distributions for modeling uncertain or vague data, parameters or predictions.

5.7.2 Fuzzy set-based approximate reasoning

Fuzzy rule-based modeling and related approximate reasoning techniques are among the most popular applications of fuzzy set theory. Fuzzy rules have been used extensively for the linguistic modeling of functional relationships. The main idea of fuzzy control, for instance, is to simulate a human expert by constructing a control function from a set of linguistically specified *if-then* rules. In this context, a rule "if X is A then Y is B" represents (vague) partial knowledge about the graph of an underlying (control) function and is usually not considered as

a logical implication. Rather, it defines an (ordered) pair of (fuzzy) data (A, B) and should be understood in the sense of a possibility-qualifying rule. The union of fuzzy relations $A \times B$ associated with a number of rules defines a *fuzzy graph* [418]. It is thought of as a vague approximation of the underlying (control) function in much the same way as δ_{s_0} is interpreted as a (lower) approximation of the relation φ of cases.

Seen from this perspective, PoCBI is close to the interpretation of fuzzy rules originally outlined by Zadeh [414] and put into practice by Mamdani [259, 258]. Still, a major difference deserves mentioning: A human expert specifying points of the graph of a function is assumed to have knowledge about *absolute* values of that function. By providing similarity-based rules in PoCBI, he rather gives a description of how these values vary when changing the argument of the function. For example, an expert might know very little about prices of cars of a certain manufacturer. Still, his (case-based) experience might tell him that (at least in general) cars with similar horsepower and similar engine-size have similar prices. Then, learning about the price of one (typical) car of a certain manufacturer, he will also have an idea of the price of a similar car (produced by the same manufacturer).

Mathematically speaking, PoCBI assumes that a human expert can somehow specify, not a function itself, but the variation or derivative of the function. This knowledge can then be used for extrapolating observed data in the form of concrete values. By instantiating observed cases, PoCBI thus transforms a set of similarity-based rules into a (larger) set of ordinary fuzzy rules. In other words, an ordinary rule base is derived from a set of similarity-based rules in connection with a set of observations. Needless to say, this type of case-based derivation of a rule base might be interesting not only for CBR itself but also for other domains. In fuzzy control, for instance, it might reasonably complement other techniques for learning fuzzy rules (e.g. [2, 386]). In this sense, PoCBI can be seen from two perspectives. Firstly, as a method which makes use of fuzzy set-based modeling techniques in order to specify a CBR model, i.e., as an application of fuzzy set (possibility) theory in case-based reasoning. Secondly, as a method which allows one to transform case-based information into a fuzzy rule base, i.e., as an application of CBR techniques in (rule-based) approximate reasoning.

Of course, if the expert is also able to specify some values of a function it seems reasonable to combine PoCBI and the approach to approximate reasoning used in fuzzy control, an idea which has already been discussed in Section 5.4.7. Besides, it should be mentioned that a rule base thus obtained can be "tuned" in different ways. For instance, in order to reduce the size of the case base it will often be reasonable to merge several rules which originate from similar cases, i.e., to derive one general rule from a number of more specific rules (see, e.g., [406] and Section 5.4.3).

5.8 Summary and remarks

Summary

– In this chapter, we have outlined a possibilistic approach to case-based inference. The basic principle of this approach, referred to as PoCBI, is a kind of similarity-guided, possibilistic extrapolation of observed cases. According to this principle, which relies on the CBI hypothesis and which has been formalized within the framework of fuzzy rules, an already encountered case is taken as evidence for the existence of similar cases. This evidence is expressed in terms of degrees of possibility assigned to hypothetical cases and thus defines a possibilistic approximation of an underlying (but only partially observed) set of potential cases.

– A distinctive feature of PoCBI is the ability to combine knowledge and data in a flexible way. Even though it can be considered as a case-based method in the first place, (expert) knowledge still plays an essential role. Firstly, such knowledge is used for controlling the "possibilistic extrapolation" of sample cases, i.e., the local generalization beyond observed examples. Secondly, general background knowledge can supplement case-based information when it comes to making predictions. A prediction in the form of a possibility distribution may thus result from the combination of several ingredients, namely the observed cases, the (heuristic) "CBR knowledge" which dictates how to extrapolate the data, and background knowledge which supplements or modifies the extrapolation.

– One of the basic ideas of our approach is that of exploiting the merits of linguistic modeling techniques in the context of CBR. It does not mean, however, that a human expert is expected to come up with an optimal model from the start. Rather, it might be sufficient if he specifies a broad structure in a first step, including, e.g., the selection and combination of important attributes which appear together in a rule. A corresponding rule base can then be calibrated afterwards by means of the adaptation technique proposed in Section 5.6.

– From a learning point of view, the possibilistic approach has much in common with non-parametric statistical inference (kernel-based density estimation) and instance-based learning. In fact, the application of possibility theory allows for realizing a graded version of the similarity-based extrapolation principle underlying IBL which appears to be very natural and 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, the possibilistic variant (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 seems to have some advantages:

From a *knowledge representation* point of view, 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.

Regarding the *applicability*, the possibilistic approach is more robust and may thus extend 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, it was shown that 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. Also worth mentioning is that the max–min version of PossIBL can even be applied within a purely ordinal setting.

Finally, 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.

– PoCBI is also related to possibilistic data analysis. In this regard, it was found that it combines aspects of qualitative (descriptive, exploratory) and quantitative (inductive) methods and that it can be seen as a kind of extended exploratory data analysis. The comparison between PoCBI and fuzzy set-based approximate reasoning has shown that PoCBI applies fuzzy rules at a higher level. In connection with observed data, a set of such rules induces or, say, instantiates an "ordinary" (fuzzy) rule base. Thus, case-based and rule-based reasoning techniques can complement each other in a reasonable way.

Remarks

– The type of possibilistic prediction realized by PoCBI can be used in various ways, e.g., as in this chapter for classification or function approximation. Besides, it can be embedded into more complex reasoning procedures. In the context of case-based reasoning, for example, PoCBI can support the overall process of problem solving by bringing a set of potential solutions into focus: By providing estimations $\delta_{s_0}(r)$ of the possibility that r is the solution (= outcome) of the new problem (= input) s_0, or that r can at least be adapted in a suitable way, PoCBI allows one to focus on the most promising candidates and, hence, to improve the efficiency of case-based problem solving. Likewise, a prediction

in the form of a possibility distribution can provide useful information in the context of decision making (cf. Chapter 7).

– A case is often characterized by a set of attributes, and a similarity relation is given for each of these attributes (cf. Section 2.3.3 and Example 5.1). In this connection, it deserves mentioning that the derivation of a global similarity by means of an aggregation of individual similarity relations presupposes the individual measures to be *commensurate*: Given two measures σ_1 and σ_2 ranging on (numeric) scales $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively, the objects x and y are as similar as u and v iff the equality $\sigma_1(x, y) = \sigma_2(u, v)$ holds. This remark is particularly important in connection with ordinal scales (which might even have different cardinalities). At a formal level, commensurability can also be achieved by mapping similarity degrees from different (heterogeneous) scales into one common scale $\mathcal{L}$ before aggregation takes place.

– We have stressed the aspect that a possibilistic CBI model is essentially derived from the knowledge of an expert, and that data is only used for calibrating the model. Of course, other approaches which partly rely on user advice in model building exist as well, but often the user plays a less significant role or intervenes in a more indirect way. In the memory-based reasoning methodology presented in [217], for instance, the user can specify causal dependencies between variables by (partially) determining the structure of a probabilistic network. This network (eventually in a corrected form) is then used for deriving a similarity-measure which in turn controls the retrieval of cases (and, hence, the labeling of new cases in a classification task).

– Note that the possibilistic approximation of the relation φ in (5.4) will in general not converge toward (the $\{0, 1\}$-valued possibility distribution associated with) φ with an increasing sample size. Rather, some hypothetical cases similar to observed cases will always be supported with a positive degree of similarity even though they do actually not exist. This problem could be alleviated by controlling the extent of extrapolation as a function of the sample size.[28] This is comparable to a corresponding adaptation of the smoothing parameter in kernel-based density estimation. Notice, however, that an adaptation of this kind is already realized by the calibration of a CBI model (cf. Section 5.6), albeit in a more implicit way. Besides, it should be mentioned that an asymptotic influence of similarity might indeed be reasonable. It makes sense, e.g., if the sample is not representative and some cases are not accessible to observation [281].

– The generalization of the kNEAREST NEIGHBOR algorithm which has been proposed in [84] is also closely related to the possibilistic approach of this chapter. As already explained in Section 4.9, this approach specifies the unknown class c_0 of a new pattern x_0 in terms of a belief function. This belief function is

[28] The opinion that the influence of similarity should decrease if the sample size increases was already held by CARNAP in connection with the inductive logic-based modeling of analogical reasoning [60].

obtained by combining the individual belief functions induced by the neighbors of x_0, where the $\imath$-th neighbor $x_\imath$ specifies c_0 by means of a mass distribution $\mathsf{m}_\imath$ such that

$$\mathsf{m}_\imath(\{c_\imath\}) = \alpha_\imath, \quad \mathsf{m}_\imath(C) = 1 - \alpha_\imath. \tag{5.66}$$

Note that the belief structure (5.66) is consonant, which means that it can also be expressed in terms of a possibility distribution.

The main differences between [84] and PoCBI are as follows: Firstly, the combination of individual pieces of evidence is realized in different ways, namely by means of a $\oplus$-aggregation in PoCBI and by means of Dempster's rule in [84]. Note that the latter assumes the pieces of evidence to be distinct [349] which, as argued in Chapter 4, might not always be true in the context of classification.

Secondly, as in IBL, the method in [84] does not consider a similarity structure over the set of outcomes (classes). In fact, an instance only supports the class to which it belongs. As opposed to this, a case also supports *similar* outcomes in PoCBI.

Thirdly, by focusing on classification as a performance task, the method in [84] has been developed with a specific application in mind and can be seen as a purely data-driven approach. As has been seen in previous sections, PoCBI supports the combination of data and domain-specific (expert) knowledge in the more general context of case-based reasoning. This becomes possible through the close connection between possibility theory and the theory of fuzzy sets. In particular, this connection allows one to adapt a possibilistic CBI model by means of fuzzy set-based (linguistic) modeling techniques.

– When comparing the extrapolation principle of the possibilistic and the probabilistic NN principle (Section 5.3.5) we have emphasized the difference between absolute and relative support of a case. A similar distinction has also been made in the context of clustering. In fuzzy clustering, a point is not assigned to one class in an unequivocal way; rather, it may have a positive degree of membership in several classes. Still, in the classical approach the membership degrees are forced to sum to 1 [32]. Consequently, these membership degrees must be interpreted as *relative* numbers. This constraint (which has a probabilistic flavor) is relaxed in *possibilistic* clustering [240], where a membership degree does indeed reflect the (absolute) compatibility of a point with the prototype of a cluster.

– In the qualitative (max-min) version of PoCBI, the evidential support of a hypothetical case c basically corresponds to the maximal similarity between c and an observed case. Interestingly enough, the same value also plays an important role in a probabilistic model of analogical induction proposed in [281]. This value, which corresponds to the possibility degree (5.7) in our approach, is

called *analogy factor*.[29] In [281], however, this factor is not directly considered as a measure of evidence. Rather, it is used for modeling the influence of experience from similar situations when it comes to *updating* a degree of probability (of occurrence) associated with c.

[29] More precisely, it is qualified as an *existential* analogy factor. An *enumerative* factor which depends on the similarity of c, not only to the nearest neighbor, but to *all* observed cases is considered as an alternative.