6. Fuzzy Set-Based Modeling of Case-Based Inference II

In Chapter 5, it has already been shown that fuzzy rules can be modeled formally as possibility distributions constrained in terms of a combination of the membership functions which define, respectively, their antecedent and consequent part. This way, they relate the concepts of similarity and uncertainty, which is the main reason for their convenience as formal models of the CBI hypothesis. Work on fuzzy *if–then* rules has mainly concentrated on algebraic properties of (generalized) logical operators. However, going into the semantics of such rules, it turns out that different interpretations lead to different types of fuzzy rules, which can be associated with corresponding classes of implication operators [117].

The logical operator used for modeling the type of fuzzy rule that we have focused on in Chapter 5, a so-called possibility rule, is a conjunction (t-norm) rather than an implication. In fact, a possibility rule is considered, not as a logical implication in the strict sense, but rather as an example-oriented rule which encodes and extrapolates information derived from observations. In this context, a fuzzy rule "if X is A then Y is B" defines a case in the form of an ordered pair of data (A, B) which suggests the feasibility of further (similar) cases (or, more precisely, guarantees a certain degree of possibility of such cases).

As already pointed out, however, an alternative, *implication-based* type of fuzzy rule can be very useful in the context of CBI, both from a knowledge representation (Section 5.3.3) and a learning point of view (Section 5.6). In this chapter, we shall consider implication-based fuzzy rules in more detail. As will be seen, formalizing the CBI hypothesis in terms of implication-based rules involves a completely different approach to knowledge representation and inference. In fact, the use of implication-based fuzzy rules leads to a *constraint-based* approach which can be seen as a generalization of the constraint-based modeling of CBI in Chapter 3. That is, each rule associated with an observed case $\langle s_1, r_1 \rangle$ serves as a constraint: Given a new input s_0 similar to s_1, it rules out those outcomes which are not sufficiently similar to r_1. This way, an observation restricts the set of possible outputs resp. decreases the possibility of certain outcomes. Loosely speaking, a constraint-based (implication-based) fuzzy rule *excludes* outcomes which are *dissimilar* (while not saying anything about the similar ones), whereas an example-oriented (conjunction-based) rule *supports* outcomes which are *similar* (while reserving judgement concerning the ones which are dissimilar). The difference between the two approaches, which exactly corresponds to the distinc-

tion between certainty and plausibility resp. upper and lower possibility in Section 5.1, becomes also apparent from the way in which evidence from multiple cases is combined. In connection with implication-based rules, this evidence is aggregated by means of an intersection (resp. the application of a t-norm) which is a natural approach to combining constraints. As opposed to this, the disjunctive aggregation (resp. the application of a t-conorm) in the case of possibility rules corresponds to a data accumulation process.

The remaining part of the chapter is organized as follows: In Section 6.1 and Section 6.2, two basic models which make use of two types of implication-based fuzzy rules, namely *gradual rules* and *certainty rules*, are introduced. Section 6.3 considers case-based inference in the context of information fusion and provides a probabilistic interpretation which relates the gradual rule and the certainty rule model. The rating of cases based on the information they provide and the related idea of "exceptionality" of cases is considered in Section 6.4. Section 6.5 generalizes the previously introduced models by applying the CBI hypothesis in a locally resticted way.

Before going on, let us make a note on notation. As in Chapter 5, we shall denote by $\varphi \subseteq \mathcal{S} \times \mathcal{R}$ the set of potential observations, i.e., a case is always an element of the relation φ. Alternatively, we shall look at φ as a set-valued mapping $\varphi : \mathcal{S} \longrightarrow 2^{\mathcal{R}}$, i.e., we denote by $\varphi(s)$ the set $\varphi \cap (\{s\} \times \mathcal{R})$ of possible outcomes of the input s. We shall further abuse this notation and write $r = \varphi(s)$ instead of $(s, r) \in \varphi$ or $\{r\} = \varphi(s)$ if φ is an ordinary function. Again, we assume data to be given in the form of a (finite) 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 precedent cases. Let $\mathcal{M}^*$ denote the class of all finite memories $\mathcal{M} \subset \varphi$.

Finally, we restrict ourselves in this chapter to the qualitative version of possibility theory and, hence, to the operators min and max as t-norm and t-conorm, respectively. Thus, we assume that possibility (and hence similarity) is measured on an ordinal scale $\mathcal{L}$. (Though an exception is made in Section 6.3, where a possibilistic prediction is endowed with a probabilistic semantics.) We note, however, that all results can be transferred to the quantitative case in a more or less straightforward way.

6.1 Gradual inference rules

6.1.1 The basic model

Gradual rules [119] depict relations between variables X and Y which correspond to propositions of the form "the more X is A, the more Y is B," where A and B are fuzzy sets modeling certain symbolic labels. This can also be stated as "the

larger the degree of membership of X in the fuzzy set A, the larger the degree of membership of Y in B" or, even more precisely, as "the larger the degree of membership of X in the fuzzy set A, the larger the guaranteed lower bound to the degree of membership of Y in B." The intended semantics of such a rule can be expressed in terms of membership degrees by

$$A(X) \leq B(Y), \tag{6.1}$$

which is equivalent to the collection of constraints

$$\forall\, 0 < \alpha \leq 1 \,:\, X \in A_\alpha \Rightarrow Y \in B_\alpha,$$

where $A_\alpha = \{x \,|\, A(x) \geq \alpha\}$ denotes the α-cut of the fuzzy set A [119].

The constraint (6.1) induces a $\{0,1\}$-valued (conditional) possibility distribution $\pi_{Y|X}$, where $\pi_{Y|X}(y \,|\, x)$ denotes the possibility of $Y = y$ given that $X = x$:

$$\forall\, x \in D_X\, \forall\, y \in D_Y \,:\, \pi_{Y|X}(y \,|\, x) = A(x) \overset{\text{rg}}{\rightsquigarrow} B(y), \tag{6.2}$$

where $\overset{\text{rg}}{\rightsquigarrow}$ is the Rescher-Gaines implication ($\alpha \overset{\text{rg}}{\rightsquigarrow} \beta = 1$ if $\alpha \leq \beta$ and 0 otherwise) and D_X and D_Y are the domains of X and Y, respectively.

More generally, fuzzy gradual rules can be classified as *truth-qualifying rules*, the semantics of which are adequately modeled by means of so-called R(esiduated)-implications. An R-implication is derived from a t-norm $\otimes$ through residuation [118]:

$$\forall\, \alpha, \beta \in [0,1] \,:\, \alpha \rightsquigarrow \beta \overset{\text{df}}{=} \sup\{\, \gamma \,|\, \alpha \otimes \gamma \leq \beta \,\}. \tag{6.3}$$

An example is the implication operator $\rightsquigarrow$ defined as

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

Using this implication, the possibility of $Y = y$ is not restricted to the values 0 and 1 but may take any value in the interval $[0,1]$. Nevertheless, subsequently we will adhere to the model (6.2) which is referred to as a *pure gradual rule* in [46].

Within the context of our CBI framework, a gradual rule reads "the more similar two inputs are, the more similar are the associated outcomes" or, more precisely, "the more the similarity of inputs is in F, the more the similarity of outcomes is in G," with F and G being fuzzy sets of "large similarity degrees" (F and G are non-decreasing $\mathcal{L} \longrightarrow \mathcal{L}$ functions). In connection with (6.1) and an observed case $\langle s_1, r_1 \rangle$, this rule (completely) excludes the existence of other (hypothetical) cases $\langle s, r \rangle$ which would violate

$$F(\sigma_{\mathcal{S}}(s, s_1)) \leq G(\sigma_{\mathcal{R}}(r, r_1)). \tag{6.4}$$

Thus, given a new input s_0 and assuming $F = G = \text{id}$, (6.4) becomes

$$\forall \langle s, r\rangle \in \varphi \,:\, \sigma_{\mathcal{S}}(s, s_1) \leq \sigma_{\mathcal{R}}(r, r_1) \tag{6.5}$$

and, hence, leads to the restriction

$$r_0 \in \left\{ r \in \mathcal{R} \,|\, \sigma_{\mathcal{S}}(s_0, s_1) \leq \sigma_{\mathcal{R}}(r, r_1) \right\} \tag{6.6}$$

for the output r_0 associated with s_0. Since corresponding constraints are obtained for all cases of a memory $\mathcal{M}$, we finally derive the following prediction [99, 101]:

$$r_0 \in \widehat{\varphi}_{\mathcal{M}}(s_0) \stackrel{\text{df}}{=} \bigcap_{1 \leq \imath \leq n} \left\{ r \in \mathcal{R} \,|\, \sigma_{\mathcal{S}}(s_0, s_\imath) \leq \sigma_{\mathcal{R}}(r, r_\imath) \right\}. \tag{6.7}$$

Clearly, the extent to which the CBI hypothesis holds true depends on the respective application. Consequently, the formalization of this principle by means of the constraint (6.1) might be too strong, at least in connection with the underlying similarity relations $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$. That is, cases $\langle s, r\rangle, \langle s', r'\rangle$ might exist such that $\sigma_{\mathcal{S}}(s, s') > \sigma_{\mathcal{R}}(r, r')$, i.e., although the inputs are similar to a certain degree, the same does not hold for the associated outputs. This, however, contradicts (6.4). Thus, calling a prediction $\widehat{\varphi}_{\mathcal{M}}(s_0)$ *correct* (with respect to the case $\langle s_0, r_0\rangle$) if $r_0 \in \widehat{\varphi}_{\mathcal{M}}(s_0)$, the (general) correctness of the inference scheme (6.7) is not guaranteed in the sense that it might yield an incorrect prediction:

$$\exists\, \mathcal{M} \in \mathcal{M}^* \,\exists\, \langle s_0, r_0\rangle \in \varphi \,:\, r_0 \notin \widehat{\varphi}_{\mathcal{M}}(s_0).$$

That is, there are a memory $\mathcal{M}$ and a case $\langle s_0, r_0\rangle$ such that the set-valued prediction derived from $\mathcal{M}$ does not cover r_0. Note that the complete class φ of cases would have to be known in order to guarantee the correctness of (6.7) in the above sense. Needless to say, this condition is usually not satisfied.

6.1.2 Modification of gradual rules

Again, more flexibility can be introduced in the basic model (6.1) by means of a modifier, i.e., a non-decreasing function $m : \mathcal{L} \longrightarrow \mathcal{L}$. This leads to

$$\forall \langle s, r\rangle \in \varphi \,:\, m(\sigma_{\mathcal{S}}(s, s_1)) \leq \sigma_{\mathcal{R}}(r, r_1) \tag{6.8}$$

instead of (6.5). Moreover, (6.7) becomes

$$r_0 \in \widehat{\varphi}_{m,\mathcal{M}}(s_0) \stackrel{\text{df}}{=} \bigcap_{1 \leq \imath \leq n} \left\{ r \in \mathcal{R} \,|\, m(\sigma_{\mathcal{S}}(s_0, s_\imath)) \leq \sigma_{\mathcal{R}}(r, r_\imath) \right\}. \tag{6.9}$$

The application of the modifier m can be seen as "calibrating" the similarity scales underlying the set of inputs and the set of outputs such that (6.1) is always satisfied. As an extreme example of (6.8) consider the case where $m \equiv 0$, expressing the fact that the CBI hypothesis does not apply at all. In other words, the similarity of inputs (in the sense of $\sigma_{\mathcal{S}}$) does not justify any conclusions about

the similarity of outcomes (in the sense of $\sigma_{\mathcal{R}}$). Observe, however, that m can as well be utilized in order to strengthen (6.1). We might take, for instance, $m \equiv 1$ if all outcomes are always perfectly similar according to $\sigma_{\mathcal{R}}$! This type of modification of a gradual rule can be interpreted in the same way as the modification of a possibility rule (cf. Section 5.4.4).

We call a modifier *admissible* if it guarantees the correctness of the inference scheme (6.9), i.e.

$$\forall\, \mathcal{M} \in \mathcal{M}^* \,\forall\, \langle s_0, r_0 \rangle \in \varphi \,:\, r_0 \in \widehat{\varphi}_{m,\mathcal{M}}(s_0). \tag{6.10}$$

The modifier m defined by

$$m(x) = \sup \big\{ h(x') \,|\, x' \in D_{\mathcal{S}}, x' \leq x \big\} \tag{6.11}$$

for all $x \in D_{\mathcal{S}}$, where

$$h(x) = \inf_{\langle s,r\rangle, \langle s',r'\rangle \in \varphi : \sigma_{\mathcal{S}}(s,s') = x} \sigma_{\mathcal{R}}(r, r'),$$

is admissible. Moreover, it is maximally restrictive in the sense that

$$\forall\, \mathcal{M} \in \mathcal{M}^* \,\forall\, s_0 \in \mathcal{S} \,:\, \widehat{\varphi}_{m,\mathcal{M}}(s_0) \subseteq \widehat{\varphi}_{m',\mathcal{M}}(s_0)$$

holds true for each admissible (and non-decreasing) $m' : D_{\mathcal{S}} \longrightarrow \mathcal{L}$.[1] Taking the upper bound in (6.11) only guarantees that m is non-decreasing. In fact, (6.10) remains valid when replacing m by h, which obviously corresponds to the *similarity profile* as introduced in Section 3.1.[2] In other words, a modifier m defines a *strict similarity hypothesis* (see page 61) and thus obeys the "the more... the more..." assumption underlying the concept of a gradual rule: The modification by means of a non-decreasing function corresponds to the "stretching" and "squeezing" of the similarity scale underlying $\sigma_{\mathcal{S}}$. When interpreting $m \circ \sigma_{\mathcal{S}}$ as a new (adapted) similarity measure, $m \circ \sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{S}}$ are still *coherent* in the sense that

$$\sigma_{\mathcal{S}}(s_1, s_2) \leq \sigma_{\mathcal{S}}(s_3, s_4) \;\Rightarrow\; m(\sigma_{\mathcal{S}}(s_1, s_2)) \leq m(\sigma_{\mathcal{S}}(s_3, s_4)) \tag{6.12}$$

for all $s_1, s_2, s_3, s_4 \in \mathcal{S}$. As opposed to this, a non-increasing function h also puts the similarity degrees $x \in D_{\mathcal{S}}$ in a different order and, hence, violates (6.12).

Loosely speaking, (6.11) can be seen as a solution to the (optimization) problem of finding a modifier maximally restrictive among all the admissible ones. Estimating (6.11) from observed data (in the form of the memory $\mathcal{M}$) can be considered as a problem of *case-based learning*. Of course, a corresponding estimation will generally not allow for verifying the admissibility of a modifier in the sense of (6.10). In fact, (6.10) can be checked only for the *observed* cases, which means

[1] Here, we assume that $m'(x) \in \mathcal{L}$ for all $x \in D_{\mathcal{S}}$. More generally, a modifier is a $D_{\mathcal{S}} \longrightarrow [0,1]$ mapping.

[2] Recall, however, that φ as defined here is not necessarily a functional relation.

that the requirement of (global) admissibility has to be weakened. An obvious idea is to look for a maximally restrictive modifier m which is admissible, not necessarily for the complete relation φ, but at least for the memory $\mathcal{M}$. That is,

$$\forall \langle s, r\rangle \in \mathcal{M} \,:\, r \in \widehat{\varphi}_{m,\mathcal{M}}(s). \tag{6.13}$$

In addition to (6.13), it might appear natural to require

$$\forall\, s \in \mathcal{S} \,:\, \widehat{\varphi}_{m,\mathcal{M}}(s) \neq \emptyset. \tag{6.14}$$

That is, for each input s which might be encountered, the inference scheme (6.9) yields a non-empty (even if perhaps incorrect) prediction [100]. Needless to say, the additional requirement (6.14) makes the learning of a modifier more complex.[3] Note that the problem of learning the maximally restrictive modifier (6.13) can be approached by the algorithm proposed in Section 3.4 (cf. Remark 3.31).

Observe that $F = G = \mathrm{id}$ can be assumed for the fuzzy sets F and G in (6.4) without loss of generality (as long as G is strictly increasing). This becomes obvious from the constraint (6.8). Namely, $m(F(\sigma_{\mathcal{S}}(s, s')) \leq G(\sigma_{\mathcal{R}}(r, r'))$ is equivalent to $m'(\sigma_{\mathcal{S}}(s, s')) \leq \sigma_{\mathcal{R}}(r, r')$ with $m' = G^{-1} \circ m \circ F$.

Even though the approach (6.8) allows for the adaptation of the formal CBI model based on a gradual rule, this model remains rather restrictive. In fact, the above discussion has shown that the gradual rule model is closely related to the constraint-based approach of Chapter 3.[4] Consequently, it might lead to imprecise predictions for exactly the same reasons. Consider the following example, to which we shall return occasionally in subsequent sections.

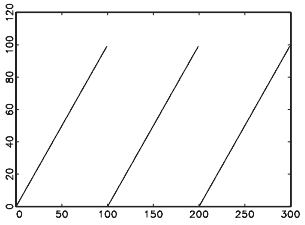

Fig. 6.1. Graph of the function $a \mapsto a \bmod 100$.

[3] Verifying (6.14) is closely related to testing the *coherence* of a set of gradual rules [133].

[4] The approaches basically differ in the sense that the latter does not only allow for strict similarity hypotheses.

EXAMPLE 6.1. Let a CBI setup be defined as follows:

$$\begin{aligned}&\mathcal{S} = \mathcal{R} = \mathfrak{N}_0, \quad D_{\mathcal{S}} = D_{\mathcal{R}} = \{0,1\},\\ &\sigma_{\mathcal{S}}(a,b) = \sigma_{\mathcal{R}}(a,b) = 1 \Leftrightarrow |a-b| \leq 10,\\ &\varphi : \mathcal{S} \longrightarrow \mathcal{R}\,,\ a \mapsto a \bmod M.\end{aligned}$$

Thus, inputs and outputs correspond to natural numbers, and two inputs (outputs) are either completely similar or not similar at all. According to the definition of φ,

$$\begin{aligned}\varphi(a) = q \Leftrightarrow q \in \{0,1,\ldots,M-1\}&\\ \wedge\ \exists p \in \mathfrak{N}_0\ :\ a = pM + q.&\end{aligned}$$

Assuming M to be a rather large integer, we can hence say that $\varphi(s)$ and $\varphi(s')$ are "almost surely" similar whenever s and s' are similar. (See Fig. 6.1, where the graph of φ is illustrated for $M = 100$). Nevertheless, "exceptional" pairs of inputs s, s' for which $\sigma_{\mathcal{S}}(s,s') = 1$ and $\sigma_{\mathcal{R}}(\varphi(s),\varphi(s')) = 0$ still exist (e.g., $s = M-1$, $s' = M$). Thus, one has to take $m \equiv 0$ in order to guarantee the correctness of (6.9). Then, however, case-based inference via (6.7) becomes meaningless, since $\widehat{\varphi}_{m,\mathcal{M}}(s_0) = \mathcal{R} = \mathfrak{N}_0$ for all $s_0 \in \mathcal{S}$. □

This example suggests looking for generalized inference schemes which are less restrictive. In this chapter, we consider two possibilities of weakening the formalization of the CBI principle based on gradual rules. Firstly, we give up the requirement of its *global* validity, i.e., the fact that *one* modifier has to be determined such that (6.8) is satisfied for *all* (tuples of) cases. A related approach will be proposed in Section 6.5, where case-based inference will not be formalized by means of a single modifier, but by means of a set of ("locally valid") fuzzy rules. This idea is similar to the use of *local* similarity profiles in the constraint-based approach to CBI.

Secondly, (6.8) is obviously not very flexible in the sense that it does not allow for incorporating some tolerance toward exceptions into the inference process. In fact, the above example suggests looking for inference schemes which do not only distinguish between the possibility and impossibility of outcomes, but which are able to derive more expressive predictions using a *graded* notion of possibility. For this reason, we shall consider so-called *certainty rules* in Section 6.2 below. Replacing gradual rules by certainty rules is motivated in the same way as passing from constraint-based to probabilistic CBI as proposed in Chapter 4.

6.2 Certainty rules

A certainty rule corresponds to statements of the form "the more X is A, the more *certain* Y lies in B." More precisely, it can be interpreted as a collection of rules "if $X = x$, it is certain at least to the degree $A(x)$ that Y lies in B"

($x \in D_X$), which amounts to saying that the possibility of values outside B is bounded by $1 - A(x)$. This translates into the following constraint on the conditional possibility distribution $\pi_{Y|X}$ [124]:

$$\forall\, x \in D_X\,, y \in D_Y \,:\, \pi_{Y|X}(y \,|\, x) \leq \max\{1 - A(x), B(y)\}. \tag{6.15}$$

More generally, rules of this kind can be classified as *certainty-qualifying rules* [118]. The semantics of such rules is adequately captured by means of so-called S(trong)-implication operators. The latter is of the form $\alpha \rightsquigarrow \beta \stackrel{\text{df}}{=} n(\alpha) \oplus \beta$, where $n(\cdot)$ is a strong negation and $\oplus$ a t-conorm. A special case of an S-implication is the Kleene-Dienes implication in (6.15). Note that the mapping $x \mapsto 1 - x$ in (6.15) is actually thought of as the order-reversing mapping of the ordinal scale $\mathcal{L}$.

The upper bound (6.15) implies that the possibility of $Y = y$ is bounded by $1 - A(x)$ if $X = x$ and $B(y) = 0$, which means that y is outside of the support of B. Thus, the larger $A(x)$, the smaller the possibility that y lies outside of B. Within the framework of possibility theory, *certainty* is closely related to *impossibility*[5] and, hence, (6.15) indeed means that y lies in B with certainty $A(x)$.

Since a certainty rule is thought of as a constraint which holds true in general but still allows for exceptions (see e.g. [376]), it is more flexible than the approach based on gradual rules and seems to be particularly suitable as a formal model of CBI. In connection with the concept of a certainty rule, the CBI hypothesis can be understood as "the larger the similarity of two inputs is, the more *certain* it is that the similarity of corresponding outcomes is large," an interpretation which emphasizes the heuristic nature of this assumption.

Given a new input s_0, an observed case $\langle s_1, r_1 \rangle \in \mathcal{M}$ constrains the possibility of similarity degrees $y = \sigma_{\mathcal{R}}(r_0, r_1)$ according to the certainty rule model (6.15):

$$\pi(y \,|\, x) \leq \pi_{cert}(x, y) = \max\{1 - y, x\}, \tag{6.16}$$

where $x = \sigma_{\mathcal{S}}(s_0, s_1)$ is the similarity between s_0 and s_1. Since $r_0 = r$ implies $y = \sigma_{\mathcal{R}}(r, r_0)$, we thus obtain

$$\pi_{s_0}(r) = \pi(r \,|\, s_0) \leq \max\left\{1 - \sigma_{\mathcal{S}}(s_0, s_1), \sigma_{\mathcal{R}}(r, r_1)\right\} \tag{6.17}$$

for the possibility that $r \in \mathcal{R}$ corresponds to the unknown outcome r_0. The more similar the inputs s_0 and s_1 are, the more constrained the possibility of outcomes becomes according to (6.17). If, for instance, $\sigma_{\mathcal{S}}(s_0, s_1)$ is close to 1, the possibility bound $\pi(r \,|\, s_0)$ can only be large for outcomes which are very similar to r_1. If, however, $\sigma_{\mathcal{S}}(s_0, s_1)$ is very small, we also obtain a large possibility bound for outputs hardly similar to r_1. Particularly, (6.17) becomes trivial if $\sigma_{\mathcal{S}}(s_0, s_1) = 0$. The resulting possibility distribution $\pi \equiv 1$ reveals *complete ignorance*. That is,

[5] Formally, the certainty c of an event A and the possibility p of the complement of A are related according to $c = 1 - p$ (cf. Section 5.1).

the observed outcome r_1 says nothing about the unknown outcome r_0, because the corresponding inputs are not similar at all.

Since (6.17) applies to all cases of the memory, we obtain the possibility distribution

$$\begin{aligned} \pi_{s_0} : r \;\; &\mapsto \;\; \pi(r \,|\, s_0) \\ &\stackrel{\text{df}}{=} \min_{1 \le \imath \le n} \max \left\{ 1 - \sigma_{\mathcal{S}}(s_0, s_\imath), \sigma_{\mathcal{R}}(r, r_\imath) \right\}, \end{aligned} \tag{6.18}$$

which emerges from (6.15) under the application of the *minimal specificity principle*.[6] The constraint (6.18) can be generalized to

$$\begin{aligned} \pi_{s_0} : r \mapsto \pi(r \,|\, s_0) \\ = \min_{1 \le \imath \le n} m_2 \left(\max \left\{ 1 - m_1(\sigma_{\mathcal{S}}(s_0, s_\imath)), \sigma_{\mathcal{R}}(r, r_\imath) \right\} \right) \end{aligned} \tag{6.19}$$

by means of modifier functions $m_1, m_2 : \mathcal{L} \longrightarrow \mathcal{L}$. The associated certainty rule, denoted $m_1 \circ \sigma_{\mathcal{S}} \overset{m_2}{\rightsquigarrow} \sigma_{\mathcal{R}}$, corresponds to statements of the form "for m_1-similar inputs it is m_2-certain that the respective outputs are similar." As in the case of possibility rules, the modifier m_2 can be used for bounding the effect of a rule (cf. Section 5.4.4). Discounting a certainty rule can be realized, e.g., by means of a modifier $x \mapsto \max\{x, \lambda\}$, where the discounting factor λ guarantees a minimal degree of possibility.[7]

Remark 6.2. The modifier $x \mapsto \max\{x, \lambda\}$ corresponds to a special case of the discounting operation $x \mapsto (1 - \lambda) \otimes x + \lambda$ [402]. It is obtained by taking the generalized conjunction $\otimes$ as $(\alpha, \beta) \mapsto \max\{0, \alpha + \beta - 1\}$. The modifier $x \mapsto \min\{x, 1 - \lambda\}$, used as a discounting operation in the possibilistic framework of Chapter 5, emerges under the same conjunction from $x \mapsto (1 - \lambda) - (1 - \lambda) \otimes (1 - x)$. □

According to the gradual rule model, an observed case $\langle s_1, r_1 \rangle$ rules out the existence of other (hypothetical) cases completely, namely those which do not obey (6.8). Particularly, the set

$$\left\{ r \in \mathcal{R} \,|\, m(\sigma_{\mathcal{S}}(s_0, s_1)) \le \sigma_{\mathcal{R}}(r, r_1) \right\}$$

of outcomes regarded as possible for the input s_0 excludes outputs which are not similar enough, namely those outcomes $r \in \mathcal{R}$ with $\sigma_{\mathcal{R}}(r, r_1) < m(\sigma_{\mathcal{S}}(s_0, s_1))$. As opposed to this, a certainty rule (6.17) only gradually restricts the possibility of a case $\langle s, r \rangle$:

[6] According to this principle, each element of the domain of a possibility distribution is assigned the largest possibility in agreement with the given constraints. The principle is already discussed under the name *principle of maximal possibility* in [415] and has been introduced as an information-theoretic principle in [113].

[7] This contrasts with the discounting of possibility rules, where the application of the min-operator instead of the max-operator yields an upper rather than a lower possibility bound.

$$\pi(s,r) \leq \pi_{\mathcal{C}}(s,r) = \max\left\{1 - \sigma_{\mathcal{S}}(s,s_1), \sigma_{\mathcal{R}}(r,r_1)\right\}. \tag{6.20}$$

Thus, it does generally not exclude other cases completely. In fact, the possibility of a case $\langle s,r\rangle$ is 0 only if both, s is perfectly similar to s_1 and r is completely different from r_1. Given a new input s_0, we hence obtain $\pi_{s_0}(r) > 0$ as soon as $\sigma_{\mathcal{S}}(s_0,s_1) < 1$ or $\sigma_{\mathcal{R}}(r,r_1) > 0$. It is exactly this property which allows for the modeling of exceptional inputs and which seems advantageous in connection with the adaptation of CBI models.

Example 6.3. To illustrate this, let us reconsider Example 6.1. The fact that we have to take $m \equiv 0$ in connection with the gradual rule model means that a case $\langle s,r\rangle$ no longer constrains the possibility of outcomes associated with a new input s_0. Now, suppose that we define m_1 by $m_1(0) = 0$ and $m_1(1) = 1 - \varepsilon$ (and that we take $m_2 = \mathrm{id}$) in the certainty rule approach (6.19), where $0 < \varepsilon \ll 1$. Given a case $\langle s_1, r_1\rangle$ and a new input s_0 similar to s_1, we obtain

$$\pi_{s_0}(r) = \begin{cases} 1 & \text{if} \quad \sigma_{\mathcal{R}}(r,r_1) = 1 \\ \varepsilon & \text{if} \quad \sigma_{\mathcal{R}}(r,r_1) = 0 \end{cases}. \tag{6.21}$$

Thus, outcomes which are similar to r_1 are regarded as completely possible, but a positive (even if small) degree of possibility is also assigned to outcomes r which are not similar to r_1. This takes the existence of exceptional pairs of inputs into account. □

As pointed out in [99], a certainty rule (6.17) fails to modulate the width of the neighborhood around an observed outcome r_1 in terms of the similarity between s_0 and s_1, which a gradual rule would do. As expressed by (6.17), it only attaches a level of uncertainty (which depends on $\sigma_{\mathcal{S}}(s_0,s_1)$) to the fuzzy set $r \mapsto \sigma_{\mathcal{R}}(r,r_1)$ of outcomes close to r_1. A way of remedying this problem would be to use implication operators such as

$$\alpha \rightsquigarrow \beta = \begin{cases} 1 & \text{if} \quad \alpha \leq \beta \\ 1-\alpha & \text{if} \quad \alpha > \beta \end{cases} \tag{6.22}$$

or

$$\alpha \rightsquigarrow \beta = \begin{cases} 1 & \text{if} \quad \alpha \leq \beta \\ \max\{1-\alpha,\beta\} & \text{if} \quad \alpha > \beta \end{cases} \tag{6.23}$$

in place of $\max\{1-\alpha,\beta\}$ in (6.15).[8] Implications of that kind can be obtained from an R-implication $\rightarrow$ by contraposition, i.e., $\alpha \rightsquigarrow \beta = (1-\beta) \rightarrow (1-\alpha)$.

We then obtain the (generalized) model

$$\pi_{s_0} : r \mapsto \pi(r \mid s_0) = \min_{1 \leq \imath \leq n} m_2\left(m_1(\sigma_{\mathcal{S}}(s_0,s_\imath)) \rightsquigarrow \sigma_{\mathcal{R}}(r,r_\imath)\right). \tag{6.24}$$

[8] (6.23) is the R-implication and, at the same time, the S-implication related to a t-norm called the nilpotent minimum. Given a strong negation n, the latter is defined as $x \otimes y = \min\{x,y\}$ if $y > n(x)$ and $x \otimes y = 0$ otherwise [150].

This approach avoids the following effect which occurs under the application of the constraint (6.17): If the inputs s_0 and s_1 are similar enough, the bound of $\pi(r \mid s_0)$ in (6.17) only reflects the similarity between r and r_1. This, however, means that we generally have $\pi(r \mid s_0) < 1$ even for outcomes r which are rather similar to r_1. In fact, (6.17) reduces the possibility of $r_0 = r$ even if $\sigma_{\mathcal{S}}(s_0, s_1) \leq \sigma_{\mathcal{R}}(r, r_1)$. In this situation it appears to be more restrictive than a gradual rule. Observe that (6.22) to some degree combines the effect of gradual and certainty rules since $r_0 \in \sigma_{\mathcal{R}}(r_\imath, \cdot)_\alpha$ with certainty $\alpha = m_1(\sigma_{\mathcal{S}}(s_0, s_\imath))$ for all $1 \leq \imath \leq n$ (if $m_2 = \mathrm{id}$). Now, however, the certainty level and the level of the cut of the similarity relation $\sigma_{\mathcal{R}}(r_\imath, \cdot)$ are directly related (through m_1).

6.3 Cases as information sources

As in Section 4.5, we shall now look at cases as individual information sources and consider case-based inference as the parallel combination of such information sources. A corresponding (probabilistic) framework allows for a semantic interpretation of the prediction $\pi_{s_0} = \pi(\cdot \mid s_0)$ derived from a (modified) certainty rule. This interpretation gives a concrete meaning to a degree of possibility $\pi(r \mid s_0)$ and might hence be helpful in connection with the acquisition of modifiers (which act on possibility distributions). At the same time, it establishes a connection between the approaches presented in Section 6.1 and Section 6.2, showing that the latter can be seen as a generalization of the former (from a probabilistic point of view). Again, let us mention that we give up the ordinal interpretation of the underlying possibility scale in this section.

6.3.1 A probabilistic model

When making use of the CBI hypothesis formalized by means of a fuzzy rule, each observed case provides some evidence concerning the unknown outcome r_0. Given a memory $\mathcal{M}$ of n cases, the individual pieces of evidence have to be combined into a global constraint. Seen from this perspective, each case serves as an information source, and one task arising in connection with CBI is the parallel combination of these information sources. In Section 6.1, for instance, the evidence derived from an individual case $\langle s_1, r_1 \rangle$ is given in the form of a set $\mathcal{N}_{m(\sigma_{\mathcal{S}}(s,s_0))}(r)$ of possible candidates, where

$$\mathcal{N}_\alpha(r_1) \stackrel{\mathrm{df}}{=} \big\{ r \in \mathcal{R} \mid \alpha \leq \sigma_{\mathcal{R}}(r, r_1) \big\}$$

denotes the α-*neighborhood* of the outcome r_1. Moreover, the (conjunctive) combination of evidence is realized by means of the intersection (6.9).

Recall the framework of the parallel combination of information sources which has been outlined in Section 4.5: Let Ω denote a set of alternatives, consisting of all

possible states of an object under consideration and let $\omega_0 \in \Omega$ be the actual (but unknown) state. An imperfect specification of ω_0 is a tuple $\Gamma = (\gamma, p_C)$, where C is a (finite) set of *specification contexts*, γ is a mapping $\gamma : C \longrightarrow 2^{\Omega}$, and p_C is a probability measure over C. The problem of combining evidence is defined as generating an imperfect specification Γ of ω_0 which performs a synthesis among the n imperfect specifications $\Gamma_1, \ldots, \Gamma_n$ issued by different information sources.

In Section 6.1, the evidence derived from an individual case $\langle s_1, r_1 \rangle$, namely the set $\mathcal{N}_{m(x)}(r_1)$ with $m(x) = m(\sigma_{\mathcal{S}}(s_0, s_1))$ being the lower similarity bound (6.11), corresponds to a particular imperfect specification $\Gamma = (\gamma, p_{C_x})$:

$$\begin{array}{lcl} C_x & = & D_{\mathcal{R}}, \\ \gamma(c) & = & \mathcal{N}_c(r_1), \\ p_{C_x}(c) & = & \left\{ \begin{array}{lll} 1 & \text{if} & c = m(x) \\ 0 & \text{if} & c \neq m(x) \end{array} \right. . \end{array} \tag{6.25}$$

A context c is hence thought of as the lower similarity bound $m(x) \in D_{\mathcal{R}}$ associated with the similarity degree $x \in D_{\mathcal{S}}$. Observe that the information source $\langle s_1, r_1 \rangle$ is *correct* in the sense that the prediction $\gamma(c) = \mathcal{N}_c(r_1)$ contains the object $\omega_0 = r_0$ under the assumption that the context c is true (and the modifier m is admissible). It is also of *maximum specificity* since $\mathcal{N}_c(r_1)$ is the most specific characterization of r_0 that can be inferred by $\langle s_1, r_1 \rangle$ in this context.

The one-point distribution p_{C_x} in (6.25) suggests the lower similarity bound to be known precisely. In general, however, knowledge about $m(x)$ will be incomplete. Let us therefore assume p_{C_x} to be defined in a more general way, such that $p_{C_x}(c)$, the probability that $m(x) = c$, can take values between 0 and 1. Since $m(x) = c$ means that c defines the (largest) lower similarity bound, it implies $\sigma_{\mathcal{R}}(r_0, r_1) \in [c, 1]$. That is, the true similarity between r_1 and the unknown outcome r_0 is at least c. For $y \in D_{\mathcal{R}}$, the probability that $\sigma_{\mathcal{R}}(r_0, r_1) = y$ is hence bounded as follows:

$$\mathbb{P}(y) \leq \sum_{c \in D_{\mathcal{R}} : c \leq y} p_{C_x}(c).$$

When interpreting a possibility distribution π on $D_{\mathcal{R}}$ as an encoding of upper degrees of probability[9] – by virtue of the correspondence $\pi(y \,|\, x) = \mathbb{P}(y)$ – it is possible to trace the possibility distribution

$$\pi_{cert} : y \mapsto \pi_{cert}(y \,|\, x) = m(x) \rightsquigarrow y \tag{6.26}$$

derived from a (modified) certainty rule[10] back to a probabilistic specification of the similarity bound $m(x)$. Consider as an example (6.26) for the implication operator (6.22):

$$\pi_{cert}(y \,|\, x) = \left\{ \begin{array}{lll} 1 & \text{if} & m(x) \leq y \\ 1 - m(x) & \text{if} & m(x) > y \end{array} \right. . \tag{6.27}$$

[9] Here, we clearly give up the ordinal interpretation of the possibility scale.

[10] For the sake of simplicity, we restrict ourselves to certainty rules with one modifier in this section.

For $m(x) > 0$, (6.27) corresponds to the probability p_{C_x} defined by

$$p_{C_x}(c) = \begin{cases} 1 - m(x) & \text{if} \quad c = 0 \\ m(x) & \text{if} \quad c = m(x) \\ 0 & \text{if} \quad c \notin \{0, m(x)\} \end{cases} . \tag{6.28}$$

This model can be interpreted as follows: The lower similarity bound is estimated by $m(x)$, but this estimation is only correct with a certain probability. Particularly, (6.28) assigns a positive probability to the value 0, i.e., it does not exclude the existence of outcomes which are not similar at all (and hence entail $m(x) = 0$). Associating $m(x)$ with the interval $[m(x), 1]$, we might also interpret this model as a kind of confidence interval for a similarity degree $y = \sigma_{\mathcal{R}}(r_0, r_1)$, supplemented with a corresponding level of confidence.

Since $m(x) = c$ also implies

$$r_0 \in \left\{ r \in \mathcal{R} \,|\, \sigma_{\mathcal{R}}(r, r_1) \geq c \right\},$$

the possibility distribution

$$\pi_{s_0}(r) = m(\sigma_{\mathcal{S}}(s_0, s_1)) \rightsquigarrow \sigma_{\mathcal{R}}(r, r_1), \tag{6.29}$$

which is induced by an observed case $\langle s_1, r_1 \rangle$ in connection with a certainty rule, can be interpreted in the same way as the corresponding distribution (6.26). That is, the value $\pi_{s_0}(r)$ can be interpreted as an upper bound to the probability that $r_0 = r$.

The probability (6.28) reveals a special property of the uncertain prediction derived from the rule (6.27). Namely, the certainty level associated with the estimation of a similarity bound is in direct correspondence with the similarity degree itself. That is, the larger the estimation of the similarity bound $m(x)$ is, the larger will be the level of confidence attached to the confidence interval $[m(x), 1]$.[11]

6.3.2 Combination of information sources

So far, we have considered only one piece of evidence, derived from a single case $\langle s_1, r_1 \rangle$, and the imperfect specification related to the corresponding similarity bound $m(x)$, where $x = \sigma_{\mathcal{S}}(s_0, s_1)$. In general, the memory $\mathcal{M}$ contains several cases, and uncertainty concerning the complete modifier (6.11) has to be specified. Thus, let us define the set of specification contexts as $C = D_{\mathcal{R}}^{D_{\mathcal{S}}}$. Each context $c \in C$ corresponds to a function $c : D_{\mathcal{S}} \longrightarrow D_{\mathcal{R}}$ and, hence, specifies a lower similarity bound $c(x)$ for all $x \in D_{\mathcal{S}}$. Moreover, suppose a certainty rule with modifier m to be given and let p_C be defined on C in such a way that the marginal distributions correspond to the distributions p_{C_x} ($x \in D_{\mathcal{S}}$) induced by this rule.

[11] Needless to say, this property is not always appropriate.

The different information sources associated with cases in the memory now share a common set C of specification contexts. Let $\Gamma_\imath = (\gamma_\imath, p_C)$ $(1 \leq \imath \leq n)$ denote the imperfect specification associated with the $\imath$-th case $\langle s_\imath, r_\imath \rangle$. The mapping $\gamma_\imath$ is then given by

$$\gamma_\imath(c) = \mathcal{N}_{c(\sigma_\mathcal{S}(s_\imath, s_0))}(r_\imath)$$

for all $c \in C$. Making use of all cases and assuming the specification context $c \in C$ to be true, we can derive the prediction $r_0 \in \widehat{\varphi}_{c,\mathcal{M}}(s_0)$, where

$$\widehat{\varphi}_{c,\mathcal{M}}(s_0) = \bigcap_{1 \leq \imath \leq n} \left\{ r \in \mathcal{R} \,|\, c(\sigma_\mathcal{S}(s_0, s_\imath)) \leq \sigma_\mathcal{R}(r, r_\imath) \right\}. \tag{6.30}$$

This is in accordance with the gradual rule model that considers only one modifier and, hence, provides the corresponding set-valued prediction (6.30). In fact, (6.30) reveals that each context $c \in C$ corresponds to some modified gradual rule. In other words, a certainty rule can be interpreted as a "random" gradual rule, i.e., a class of (modified) gradual rules with associated probabilities. This relation between gradual and certainty rules is further explored in Appendix B.

When considering the modifier m as a random variable, the prediction of r_0 according to (6.30) becomes a random set, where $\widehat{\varphi}_{c,\mathcal{M}}(s_0)$ occurs with probability $p_C(c)$.[12] The probability that a certain output $r \in \mathcal{R}$ is an element of this set is given by

$$\mathbb{P}(r \in \widehat{\varphi}_{c,\mathcal{M}}(s_0)) = \sum_{c\,:\, r \in \widehat{\varphi}_{c,\mathcal{M}}(s_0)} p_C(c) \tag{6.31}$$

and defines an upper bound to the probability that $r_0 = r$. In connection with the idea of a randomized gradual rule model, (6.31) corresponds to the probability of selecting a (modified) gradual rule c which does not exclude the (hypothetical) case $\langle s_0, r \rangle$, i.e., for which (6.30) holds.

The imperfect specification $\Gamma = (\gamma, p_C)$ defined by

$$\gamma(c) = \widehat{\varphi}_{c,\mathcal{M}}(s_0)$$

for all $c \in C$ (and C, p_C as above) corresponds to the *conjunctive pooling* of the information sources $\Gamma_1, \ldots, \Gamma_n$. This kind of combination is justified by the fact that all information sources are correct with respect to all specification contexts $c \in C$. Within a possibilistic setting, conjunctive pooling comes down to deriving the intersection of possibility distributions. In fact, it is not difficult to show that (6.31) is bounded from above by the possibility distribution π_{s_0} derived from a certainty rule in connection with a number of cases. That is,

$$\mathbb{P}(r \in \widehat{\varphi}_{c,\mathcal{M}}(s_0)) \leq \pi_{s_0}(r) = \min\left\{ \pi^1_{s_0}(r), \ldots, \pi^n_{s_0}(r) \right\} \tag{6.32}$$

for all $r \in \mathcal{R}$, where $\pi^\imath_{s_0}$ denotes the possibility distribution derived from the $\imath$-th case according to (6.29). The interpretation of possibility degrees as upper

[12] Observe, however, that $c \neq c' \not\Rightarrow \widehat{\varphi}_{c,\mathcal{M}}(s_0) \neq \widehat{\varphi}_{c',\mathcal{M}}(s_0)$.

approximations of probabilities is hence in agreement with the application of the minimum operator in (6.19), i.e., with making use of this operator in order to combine the possibility distributions derived from individual cases.

Appendix B shows that the above probability distribution p_C, where $p_C(c)$ is the probability of the gradual rule associated with the context (= modifier) c, is unique under the assumption that the operator modeling the implication-based fuzzy rule satisfies a certain (non-)monotonicity condition. This might be considered as an interesting result, especially with regard to the combination of evidence in the probabilistic framework of Section 4.5.3. As pointed out there, the joint probability measure μ in (4.27) is generally not defined in a unique way.

According to the interpretation proposed in this section, the certainty rule approach can be seen as a generalization of the approach based on gradual rules, in the sense that the lower similarity bounds, which guarantee the correctness of the set-valued prediction of r_0, are no longer assumed to be precisely known. The incomplete knowledge concerning these bounds is characterized by means of a probability distribution. This allows for interpreting the case-based inference scheme in Section 6.2 as a kind of approximate probabilistic reasoning. More precisely, a prediction $\pi(\cdot \mid s_0)$ specifies possibility degrees $\pi(r \mid s_0)$ which can be seen as upper bounds to the probability that the unknown output r_0 is given by the outcome r.

6.4 Exceptionality and assessment of cases

Considering cases as individual information sources, as we have done in Section 6.3, suggests to rate their contribution to the prediction of outcomes. In fact, the assessment of information sources is supported by most frameworks for the combination of evidence. The basic idea, then, is to realize some kind of weighted aggregation procedure or to modify (discount) the information provided by a source according to its reliability.[13] In Section 4.6, this idea has already been discussed in the context of the probabilistic approach to CBI.

Recall that, given the same information in the form of a context $c \in C$, i.e., a modifier specifying lower similarity bounds, different cases provide different specifications of the unknown outcome r_0: Considering this modifier and the new input s_0, a case $\langle s, r \rangle$ provides a prediction of r_0 in the form of a possibility distribution which supports outcomes in the neighborhood of r. Such a specification might hence be misleading, e.g., if the outcome r is rather "untypical."

Example 6.4. Consider again Example 6.1 and suppose that $s_0 = M - 1$ and $s_1 = M + 1$. In accordance with the certainty rule model (6.21) of this example

[13] See e.g. [272] for various approaches to the discounting of expert opinions within a generalized probabilistic framework.

(cf. Section 6.2), the case $\langle s_1, r_1\rangle = \langle M+1, 1\rangle$ strongly supports the outcomes $\{0, \ldots, 11\}$ which are similar to $r_1 = 1$. It almost rules out all other outputs, including the true outcome $r_0 = M - 1$. Loosely speaking, the (otherwise useful) information about similarity relations, specified by the certainty rule, is "misinterpreted" by $\langle s_1, r_1\rangle$. Even though the advice to disqualify outcomes which are not similar to r will lead to good predictions for the majority of cases $\langle s, r\rangle$, it is hardly reasonable when taken up in connection with an "exceptional" pair of cases, such as $\langle s_0, r_0\rangle$ and $\langle s_1, r_1\rangle$. □

The above example makes clear that exceptionality is not necessarily a property of an individual input or case. Rather, the label of exceptionality applies to *pairs* of cases. In fact, $\langle s_1, r_1\rangle$ is exceptional only in connection with inputs $s = M - k$, where $1 \leq k \leq 9$, but it will lead to correct predictions for all other inputs. Moreover, the decision whether to call two cases exceptional will often not be as obvious as in our example, where only two degrees of similarity are distinguished. Making use of richer scales including intermediate degrees of similarity, exceptionality will become a gradual property.

Interestingly enough, the certainty rule framework suggests computing a degree of exceptionality in the following way:

$$\mathsf{ex}(\langle s, r\rangle, \langle s', r'\rangle) \stackrel{\text{df}}{=} 1 - \pi_{cert}(\sigma_{\mathcal{R}}(r, r') \,|\, \sigma_{\mathcal{S}}(s, s')). \tag{6.33}$$

That is, the exceptionality of the tuple of cases $\langle s, r\rangle$, $\langle s', r'\rangle$ is inversely related to the possibility of observing $\sigma_{\mathcal{R}}(r, r')$-similar outcomes for $\sigma_{\mathcal{S}}(s, s')$-similar inputs, as specified by the certainty rule model.[14] The more $\langle s, r\rangle$ and $\langle s', r'\rangle$ violate the certainty rule, the more exceptional they are in the sense of (6.33).

It is worth mentioning that (6.33) also makes sense in connection with the gradual rule model. Applying (6.33) to the possibility distribution (6.2) induced by a gradual rule, a tuple of cases is either completely exceptional or not exceptional at all. In fact, (6.33) may also be seen as a reasonable generalization of this rather obvious definition of exceptionality. This again reveals the difference between the gradual and the certainty rule model: The former is indeed not *tolerant* toward exceptions in the sense that each violation of the rule is "punished" by classifying the involved cases as *completely* exceptional ones. As opposed to this, exceptionality is a gradual property in the certainty rule model.

Even though a gradual or certainty rule can only be violated by *tuples* of cases and, hence, exceptionality should be considered as a property of pairs of cases, it seems intuitively clear in our example that the most unreliable information sources are those cases $\langle s, r\rangle$ with s close to integers kM ($k \in \mathfrak{N}_0$). The closer an input is to such a point, the more likely the case might be called exceptional. In fact, one possibility of regarding exceptionality as a property of an individual

[14] Again, note that $x \mapsto 1 - x$ in (6.33) actually represents the order-reversing mapping of a possibility scale.

case $\langle s, r\rangle$ is to consider the likelihood or possibility of $\langle s, r\rangle$ to be exceptional with respect to a new case $\langle s_0, r_0\rangle$. Thus, one might think of generalizing (6.33) as follows:

$$\mathsf{ex}_1(\langle s, r\rangle) \stackrel{\text{df}}{=} \sup_{\langle s', r'\rangle \in \varphi} \mathsf{ex}(\langle s, r\rangle, \langle s', r'\rangle). \tag{6.34}$$

Assigning a degree of exceptionality to a case in the sense of (6.34) can be interpreted as rating the reliability of this case. Of course, this degree of exceptionality depends on the formalization of the underlying rule. In other words, a case is exceptional not by itself but only with respect to a particular rule: Changing the rule by means of a modifier also changes the degree of exceptionality of the case. For instance, the modification of a gradual rule, as proposed in Section 6.1.2, can be interpreted as adapting the rule in such way that no exceptional cases exist at all. Likewise, no case is exceptional with respect to the certainty rule in its weakest form, as formalized by $m_1 \equiv 0$ in (6.19). In connection with the certainty rule model (6.21) of Example 6.1, we obtain

$$\mathsf{ex}_1(\langle s, r\rangle) = \begin{cases} 1-\varepsilon & \text{if} \quad \exists\, k \in \mathfrak{N}_0 \,:\, |s - kM| \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

for all $\langle s, r\rangle \in \varphi$.

Let us briefly hint at two properties of (6.34). Firstly, this definition of exceptionality is completely independent of any kind of *frequency*, i.e., the value $\mathsf{ex}_1(\langle s, r\rangle)$ should not be understood as a probability of $\langle s, r\rangle$ being exceptional with respect to some other case. Of course, defining exceptionality of an individual case by using an averaging operator in place of the supremum in (6.34) seems intuitively appealing and would clearly make sense within a probabilistic setting. Recall, for instance, the probabilistic interpretation of the certainty rule model proposed in Section 6.3. According to this interpretation, a certainty rule can be seen as a collection of (modified) gradual rules to each of which is attached a certain probability. Since a case is either exceptional or not with respect to a fixed gradual rule, it is an obvious idea to derive a corresponding probability of being exceptional with respect to a certainty rule.

Secondly, (6.34) is rather strict in the sense that it implies

$$\mathsf{ex}(\langle s, r\rangle, \langle s', r'\rangle) \leq \min\{\mathsf{ex}_1(\langle s, r\rangle), \mathsf{ex}_1(\langle s', r'\rangle)\} \tag{6.35}$$

for all cases $\langle s, r\rangle$ and $\langle s', r'\rangle$. In other words, having encountered an exceptional tuple of cases, *both* cases are considered to be exceptional. This principle can obviously be weakened by concluding on the exceptionality of *at least one* of the two cases. This leads to the constraints

$$\mathsf{ex}(\langle s, r\rangle, \langle s', r'\rangle) \leq \max\{\mathsf{ex}_1(\langle s, r\rangle), \mathsf{ex}_1(\langle s', r'\rangle)\} \tag{6.36}$$

for all $\langle s, r\rangle, \langle s', r'\rangle \in \mathcal{S}$. Indeed, (6.36) will often appear more reasonable than (6.35). For instance, modifying the mapping φ in Example (6.1) according to

$$\varphi(s) = \begin{cases} M & \text{if} \quad a \bmod M \neq 0 \\ 0 & \text{if} \quad a \bmod M = 0 \end{cases}$$

suggests to call the cases $\langle 0, 0\rangle, \langle M, 0\rangle, \langle 2M, 0\rangle \ldots$ exceptional and to consider all other cases to be (completely) normal. As opposed to this, (6.35) does not only qualify a case $\langle kM, 0\rangle$ itself as exceptional, but also all neighbored cases $\langle kM + a, M\rangle$ such that $1 \leq |a| \leq 10$.

A natural idea is to discount the information provided by a case based on its level of exceptionality. As already mentioned before, discounting a fuzzy restriction F over a domain D within the qualitative min-max framework amounts to modifying F into $\max\{\lambda, F\}$, where λ is a discounting factor [120]. Indeed, F remains unchanged if $\lambda = 0$. As opposed to this, the modified restriction becomes trivial (and corresponds to the complete referential D) if the discounting is maximal ($\lambda = 1$). This approach can be applied to the result of case-based inference by identifying discounting factors with degrees of exceptionality. It amounts to computing

$$\pi(r \mid s) = \min_{1 \leq \imath \leq n} \max \left\{ \mathsf{ex}_1(\langle s_\imath, r_\imath\rangle), m(\sigma_{\mathcal{S}}(s, s_\imath)) \rightsquigarrow \sigma_{\mathcal{R}}(r, r_\imath) \right\}. \tag{6.37}$$

If exceptionality is equivalent to complete exceptionality, as in the gradual rule model, (6.37) comes down to removing the exceptional cases from the memory. Apart from that, the usual inference process is realized. In other words, (6.37) then corresponds to the gradual rule approach ($\rightsquigarrow$ is the Rescher-Gaines implication) restricted to the normal cases. When using the certainty rule model in (6.37), i.e., when modeling $\rightsquigarrow$ by implication operators such as (6.22) or (6.23), the level of uncertainty of an individual prediction is increased in accordance with the degree of exceptionality of the corresponding case. The CBI hypothesis underlying the generalized approach might then be characterized as follows: "The larger the similarity between s and s_0 and the less exceptional the input s, the more certain our conclusion on the similarity between the associated outputs r and r_0."

Interestingly enough, the modifications outlined above suggest a further way of adaptation: Not the strength of the rule is adapted to the class φ of cases, but the influence of each case is modulated in accordance with its exceptionality relative to the (predefined) rule. In this connection, it also seems worth mentioning that assigning degrees of exceptionality to cases in such way that (6.36) is satisfied leads to an interesting problem from both, a mathematical as well as a semantical point of view. In addition to observed cases, one might think of using an (a priori) expert assessment of the exceptionality of cases (which then correspond to triples $\langle s, r, e\rangle$) in order to solve this problem, all the more since the minimization of some objective function subject to the constraints (6.36) might not guarantee a unique solution.

6.5 Local rules

The rule-based approaches to CBI outlined in previous sections are *local* in the sense that the information provided by different cases is processed and combined independently. They are, however, *global* in the sense that a (modified) fuzzy rule constitutes a constraint which is assumed to be globally valid. This becomes especially apparent in connection with the gradual rule approach, where an (admissible) modifier m specifies (conditional) lower bounds to the similarity of outcomes which hold true for all (pairs of) cases. It has already been pointed out in Section 6.1 that this requirement often entails rather imprecise predictions, caused by the fact that admissible modifiers might not be very restrictive.

Instead of looking for a global rule, which is valid up to some exceptions – as discussed in connection with the certainty rule model in previous sections – one might weaken the principle of a gradual rule by specifying rules which are somehow "locally" valid. In this section, we follow the idea of adapting a fuzzy rule to each case of the memory more directly rather than the one of associating instantiations of a global rule with all observed cases (and perhaps discounting these instantiations in the sense of Section 6.4). This approach is quite similar to the specification of *local* similarity profiles and hypotheses in connection with the constraint-based and probabilistic approaches to CBI discussed in previous chapters. It differs, however, from the solution proposed in connection with the possibility rule model (cf. Section 5.4.6), where local rules have not been defined for individual cases, but for different (fuzzy) regions of the space of inputs.

Let us again consider the gradual rule model. The problem that global validity might lead to (local) predictions which are unnecessarily imprecise is already certified by Example 6.1. In fact, the necessity of taking $m \equiv 0$ leads to the useless predictions $\widehat{\varphi}_{m,\mathcal{M}}(s_0) = \mathfrak{N}_0$. Loosely speaking, a CBI strategy is not applicable because the hypothesis of similar inputs having similar outcomes is not globally satisfied. Still, it seems desirable to make use of the observation that the mapping φ in the example is piecewise linear, i.e., that the CBI hypothesis is satisfied at least *locally*. One possibility of doing this is to partition the set $\mathcal{S}$ of inputs and to derive corresponding local models (cf. Section 5.4.6). In our example, the idea to partition $\mathcal{S}$ into sets of the form

$$\{kM, kM+1, \ldots, kM+(M-1)\} \qquad (k \in \mathfrak{N}_0)$$

suggests itself. However, since φ is generally unknown, the definition of a partition will not always be obvious, all the more if $\mathcal{S}$ is non-numerical.

Here, we consider a second possibility, namely that of associating an individual (local) rule with each case of the memory. Thus, the idea is to define rules of the form "the more similar an input is to s, the more similar the associated outcome is to r" for each case $\langle s, r\rangle$ in the memory. The validity of such a (gradual) rule is already guaranteed by the (non-decreasing) modifier

$$m_{\langle s,r\rangle}(x) = \sup\left\{h_{\langle s,r\rangle}(x') \,|\, x' \in D_{\mathcal{S}}, x' \leq x\right\}, \tag{6.38}$$

for all $x \in D_{\mathcal{S}}$, where

$$h_{\langle s,r\rangle}(x) = \inf_{\langle s',r'\rangle \in \varphi\,:\,\sigma_{\mathcal{S}}(s,s')=x} \sigma_{\mathcal{R}}(r,r'). \tag{6.39}$$

Since the infimum in (6.39) is taken over a smaller set of cases, (6.38) is obviously more restrictive than (6.11). Based on (6.38), the inference scheme (6.9) can be replaced by

$$r_0 \in \bigcap_{1\leq\imath\leq n} \left\{r \in \mathcal{R} \,|\, m_{\langle s_\imath,r_\imath\rangle}(\sigma_{\mathcal{S}}(s_0,s_\imath)) \leq \sigma_{\mathcal{R}}(r,r_\imath)\right\}. \tag{6.40}$$

In our example, the maximally constraining (admissible) modifier for a case $\langle s,r\rangle = \langle s,\varphi(s)\rangle$ is simply given by

$$m_{\langle s,r\rangle}(x) = \begin{cases} x & \text{if} \quad 10 \leq s \bmod M \leq M-9 \\ 0 & \text{otherwise} \end{cases}.$$

Based on a sufficiently large number of observations, the mapping φ can hence be approximated rather accurately. More precisely, the prediction (6.40) converges toward

$$\widehat{\varphi}(s_0) = \begin{cases} \{\varphi(s_0),\ldots,20\} & \text{if} \quad 0 \leq \varphi(s_0) < 20 \\ \{\varphi(s_0)\} & \text{if} \quad 20 \leq s_0\varphi(s_0) < M-20 \\ \{2M-12-\varphi(s_0),\ldots,M+9\} & \text{if} \quad M-20 \leq \varphi(s_0) < M \end{cases}$$

with an increasing number of observations.

Observe that a local rule can be taken as an indication of the (prediction) quality of a case $\langle s,r\rangle$ and can hence support the design of an optimal case base. The more restrictive a rule can be made by means of a modifier $m_{\langle s,r\rangle}$, the more it will contribute to precise predictions. As in our example, good local rules will generally be provided by "typical" cases, the outcomes of which are at least to some degree representative of similar inputs. In this sense, a modifier can also be seen as an assessment of a case (cf. Section 6.4). A modifier $m_{\langle s,r\rangle} < \mathrm{id}$, for instance, brings the discounting of a case about, whereas a modifier $m_{\langle s,r\rangle} > \mathrm{id}$ produces the opposite effect. Particularly, letting $m_{\langle s,r\rangle} \equiv 0$ comes down to leaving the corresponding case out of account, i.e., to remove it from the memory.

Let us mention that a (globally admissible) gradual rule can be seen as a collection of rules

$$\begin{aligned} \alpha(x):\ \sigma_{\mathcal{S}}(s_1,s_2) = x \ \Rightarrow \\ \forall\, r_1 \in \varphi(s_1)\, \forall\, r_2 \in \varphi(s_2)\ :\ \sigma_{\mathcal{R}}(r_1,r_2) \geq m(x), \end{aligned}$$

each of which is an aggregation of more specific (local) rules [115] associated with cases $\langle s, r\rangle \in \varphi$. More precisely, a rule $\alpha(x)$ can be seen as an approximation in the form of a disjunction

$$\alpha(x) = \bigvee_{\langle s,r\rangle \in \varphi} \alpha(\langle s, r\rangle, x) \tag{6.41}$$

of local rules

$$\alpha(\langle s, r\rangle, x) : \ (\langle s_1, r_1\rangle = \langle s, r\rangle) \wedge (\sigma_{\mathcal{S}}(s_1, s_2) = x) \Rightarrow \\ \forall\, r_2 \in \varphi(s_2) \ : \ \sigma_{\mathcal{R}}(r, r_2) \in [m_{\langle s,r\rangle}(x), 1]. \tag{6.42}$$

Since the disjunction in (6.41) is taken over all cases $\langle s, r\rangle \in \varphi$, the global rule $\alpha(x)$ depends on the similarity degree alone. Observe that (6.11) and (6.38) are related through

$$\forall\, x \in D_{\mathcal{S}} \ : \ m(x) = \inf_{\langle s,r\rangle \in \varphi} m_{\langle s,r\rangle}(x),$$

which shows that taking the disjunction of the consequent parts in (6.42) comes down to bounding similarity degrees from below and which again reveals the restrictive nature of the gradual rule model.

Interestingly enough, a certainty rule can be seen as a more general fusion of local rules (6.42), taking into account that some conclusions might be less plausible (or might occur less often) than others and, hence, may lead to a *weighted* union of conclusions instead of a disjunction.

Let us finally mention that the idea of adapting a rule-based formalization of the CBI hypothesis to individual cases applies to certainty rules in the same way as to gradual rules. Observe that local certainty rules can be seen as a combination of the two aforementioned generalizations of the gradual rule model. In fact, these rules are local and tolerant toward exceptions at the same time.

6.6 Summary and remarks

Summary

– The objective of this chapter was to elaborate in more detail on implication-based fuzzy rules as an alternative model of the inference process in case-based reasoning. It has been shown that this type of rule leads to an approach which deviates considerably from the possibility rule model discussed in Chapter 5. In fact, implication-based fuzzy rules realize a *constraint-based* approach in much the same way as the method proposed in Chapter 3: Already encountered cases are looked at as evidence for (partially) ruling out other (hypothetical) cases, not similar enough to the observed ones. As opposed to this, a possibility rule is a *conjunction-based* rule and gives rise to an *example-oriented* approach:

Observed cases are considered as pieces of data which provide evidence for the possibility of observing similar cases.

– We have distinguished between two types of implication-based rules. The first type (gradual rules) assumes a kind of closeness relation between the similarity of inputs and the similarity of outcomes which is not tolerant toward exceptions. Given a new input, the observed cases which constitute the memory are taken as evidence for either allowing or completely excluding certain outcomes. A second type of rules (certainty rules) only uses case-based information for deriving conclusions about the *possibility* of outcomes. They are more expressive and allow for the *partial* exclusion of outputs. Moreover, they can formalize situations in which the CBI hypothesis holds true "in general" up to some exceptions to the "similar inputs-similar outputs" rule.

– The use of modifier functions has been proposed for modulating the "strength" of fuzzy rules. This way, it becomes possible to adapt the formal model according to the extent to which the CBI hypothesis actually holds true for the respective application.

– The meaning of exceptionality of cases has been discussed in connection with the idea of discounting cases which might be seen as somewhat unreliable or misleading information sources. The discounting of cases, in conjunction with a modification of the basic inference scheme, presents a further possibility of model adaptation.

– Local rules have been introduced as a second direction of generalizing the basic model. There are different motivations for this step: In the gradual rule model, it is true that the instantiation of a (globally) admissible rule by different cases leads to correct predictions. However, inference results might be poor since this rule will often hardly be constraining. In the certainty rule model, the multiple instantiation of the same global rule leads to difficulties in connection with exceptional (still not discounted) cases. This might cause inconsistencies and an exaggerated exclusion of (rather possible) cases. We have also pointed out a close relation between local rules and the assessment of cases. In fact, the determination of a modifier for an individual case can be seen as a rating of the typicality or prediction quality of that case. Particularly, a modifier can make a local rule completely ineffective, which amounts to removing the corresponding case from the memory. Next to the idea of exceptionality with respect to a global rule, the concept of local rules thus presents a further possibility of rating and discounting cases.

Remarks

– In this chapter, we have refrained from discussing several issues which have already been considered in connection with the possibility rule model in Chapter 5. This concerns especially the extensions of the basic model, discussed in

Sections 5.3 and 5.4. These techniques can as well be applied to the CBI model which proceeds from implication-based fuzzy rules.

– The combination of possibility and certainty rules has already been proposed as a basis of the calibration method in Section 5.6. Besides, there are other motivations for using implication-based and conjunction-based rules jointly. In [205], for instance, it is argued that a combination of the two types of rules can greatly improve the informational contents of (possibilistic) case-based predictions. In fact, as already pointed out in Section 5.3.3, the degree $\delta_{s_0}(r)$ derived from a possibility rule can be seen as a degree of *confirmation* of the outcome r and actually defines a *lower* possibility bound. As opposed to this, the degree $\pi_{s_0}(r)$ obtained in connection with a certainty rule model reflects the degree to which past experience (in the form of the memory $\mathcal{M}$) *excludes* the output r and determines an *upper* degree of possibility. Recall the following extreme examples from Section 5.3.3:

  (a) $\delta_{s_0}(r) = 0$, $\pi_{s_0}(r) = 1$: A situation of complete ignorance. Neither is r supported nor (partly) excluded by any observation. Thus, r is fully plausible though not confirmed at all.

  (b) $\delta_{s_0}(r) = 0$, $\pi_{s_0}(r) = 0$: Clear evidence against r has been accumulated in the form of inputs similar to s_0 with outputs dissimilar to r.

  (c) $\delta_{s_0}(r) = 1$, $\pi_{s_0}(r) = 1$: The output r is strongly supported through the observation of similar cases.[15]

  The above cases emphasize the advantage of the combined approach. The example-based (possibility rule) model alone cannot distinguish between (a) and (b). It goes without saying, however, that it makes a great difference from an epistemic point of view whether a case is not supported simply because no similar cases have been observed or whether indeed some evidence against this case has been accumulated (through the certainty-rule model of the CBI principle). The constraint-based model cannot distinguish between the cases (a) and (c). Again, however, it might be important to know whether an outcome r seems completely possible for the input s_0 only because no input has been observed which is similar to s_0 or whether r is indeed supported by the observation of cases $\langle s, r\rangle$ such that s is similar to s_0 (which requires $\pi_{s_0}(r) > 0$).

[15] In fact, a possibility degree of 1 requires the observation of a *perfectly* similar case. If the similarity relations are separating, this means that $\langle s, r\rangle$ itself has already been encountered.