Instance-based prediction in the framework of possibility theory

Eyke Hullermeier, Didier Dubois, Henri Prade

Instance-Based Prediction in the Framework of Possibility Theory

Eyke Hüllermeier
Statistics and Decision Theory
University of Paderborn
Paderborn, Germany

Didier Dubois, Henri Prade
IRIT/CNRS
Université Paul Sabatier
Toulouse, France

Abstract

A possibilistic framework for instance-based prediction is presented which formalizes the generalization beyond experience by means of fuzzy rules. In comparison with related instance-based approaches such as the well-known NEAREST NEIGHBOR classifier, this method distinguishes itself by the following: First, by suggesting (guaranteed) degrees of possibility for competing outcomes rather than making precise predictions, it takes the uncertain character of similarity-based inference into account. Second, the possibilistic framework can easily be extended so as to cope with incompletely specified cases. Thirdly, the close connection between possibility theory and fuzzy sets suggests the extension of the basic model by means of fuzzy set-based (linguistic) modeling techniques. This paper especially highlights two of these aspects, namely the modeling of uncertainty and the handling of incomplete information.

Introduction

The term *instance-based reasoning* refers to a class of methods which make use of special techniques of (case-based) knowledge representation, (lazy) learning, and (similarity-based) inference. Well-known examples include the simple NEAREST NEIGHBOR classifier (Dasarathy 1991), instance-based learning algorithms (Aha *et al.* 1991), and case-based reasoning (Aamodt and Plaza 1994). Inference in these methods is generally realized by *extrapolating* the information provided by observed cases, based on some kind of *closeness* or *representativeness* assumption. Typically, the concept of *similarity* (or distance) plays a crucial role in the inference process. In case-based reasoning (CBR), for example, problem solving proceeds from the assumption that "similar problems have similar solutions." This type of closeness assumption will subsequently be referred to as the SBR hypothesis, where SBR stands for *similarity-based reasoning*.

In this paper, we formalize the SBR hypothesis and the related inference principle in the framework of fuzzy rules and fuzzy set-based approximate reasoning. Our approach can be seen as a possibilistic version of the NEAREST NEIGHBOR principle and thus provides a basis for corresponding extensions of instance-based learning and case-based reasoning. There are several motivations for combining SBR and fuzzy set-based modeling and reasoning techniques (Yager 1997), and especially the use of fuzzy rules in the context of SBR (Dubois *et al.* 1998). Particularly, the notion of *similarity*, which lies at the heart of SBR, is also strongly related to the theory of fuzzy sets since membership functions express proximity w.r.t. prototypical elements in the core of fuzzy sets. Moreover, possibility theory based on fuzzy sets provides a tool for modeling and processing *uncertainty*. In connection with SBR, this aspect seems to be of special importance if one realizes the *heuristic* character of this type of reasoning method (Hüllermeier 1999).

The basic framework we proceed from is stated in Section 2. In Section 3, we present a possibilistic model of similarity-based inference, referred to as PEC (Possibilistic Extrapolation of Cases). This model makes use of possibility rules, a special type of fuzzy rules, in order to formalize the SBR hypothesis.[1] In Section 4, our method is compared to the classical NEAREST NEIGHBOR principle and related approaches. In Section 5, we briefly outline some extensions of the basic model. Finally, Section 6 presents calibration techniques which allow one to adapt a model to the application at hand.

The Basic Framework

A *case* is a tuple $\langle s, r \rangle \in \mathcal{C} = \mathcal{S} \times \mathcal{R}$ consisting of a *situation* $s \in \mathcal{S}$ and an associated *result* or *outcome* $r \in \mathcal{R}$.[2] A case can be an arbitrarily complex object, not necessarily represented by a set of numeric attribute values. We do not assume that a situation determines a unique outcome, which would be too restrictive for certain applications. That is, cases $\langle s, r \rangle$ and $\langle s, r' \rangle$ might be encountered such that $r \neq r'$. Let $\varphi \subset \mathcal{S} \times \mathcal{R}$

[1] See (Dubois *et al.* 1998; Plaza *et al.* 1998) for a more logic-oriented formalization.

[2] We prefer these expressions for reasons of generality to the terms "problem" and "solution" which are commonly used in CBR.

denote the class of potential observations. Thus, a case is always an element of the relation φ.

Data is assumed to be given in the form of a memory

$$\mathcal{M} = \{\langle s_1, r_1\rangle, \langle s_2, r_2\rangle, \dots, \langle s_n, r_n\rangle\} \quad (1)$$

of precedent cases. The similarity of situations resp. results is specified by means of similarity relations $\sigma_{\mathcal{S}} : \mathcal{S} \times \mathcal{S} \to \mathcal{L}$, $\sigma_{\mathcal{R}} : \mathcal{R} \times \mathcal{R} \to \mathcal{L}$, where $\mathcal{L}$ is an *ordinal* scale whose top (bottom) element 1 (0) corresponds to complete (dis)similarity.

Our focus is on the performance task of prediction, namely the prediction of the *result* or *outcome* $r_0 \in \mathcal{R}$ associated with a new situation $s_0 \in \mathcal{S}$. To this end, we shall characterize the *possibility* of the candidates $r \in \mathcal{R}$ by means of a possibility distribution $\pi_{\mathcal{R}}$ on $\mathcal{R}$. Note that, seen from a machine learning point of view, the relation φ can be considered as a *concept* and, hence, instance-based prediction can be cast in the framework of *concept learning*. In fact, our point of departure is a *lower possibilistic approximation* of φ, that is, a possibility distribution $\pi_{\mathcal{C}}$, where $\pi_{\mathcal{C}}(s, r)$ is considered as a lower bound to the possibility that $\langle s, r\rangle \in \varphi$. For example, $\pi_{\mathcal{C}}(s, r) = 1$ means that $\langle s, r\rangle$ does definitely belong to φ, whereas $\pi_{\mathcal{C}}(s, r) = 0$ indicates that the membership of $\langle s, r\rangle$ cannot be guaranteed at all. More generally, a possibility degree $\pi_{\mathcal{C}}(s, r)$ reflects the extent of available evidence in favor of $\langle s, r\rangle \in \varphi$. Here, evidence means the observation of *similar* cases which belong to φ.

Possibilistic Extrapolation of Cases

Possibility Rules

Fuzzy rules provide a local, rough and soft specification of the relation between variables X and Y ranging on domains D_X and D_Y, respectively (Dubois and Prade 1996). 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.[3]

A *possibility rule* involving fuzzy sets A and B, subsequently symbolized by $A \rightsquigarrow B$, is a special type of fuzzy rule which corresponds to the statement that "the more X is A, the more *possible* 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 modeled by the following constraint which guarantees a certain *lower bound* to the possibility, $\pi(x, y)$, that (x, y) is an admissible instantiation of (X, Y):

$$\pi(x, y) \geq \min\{A(x), B(y)\}. \quad (2)$$

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 is interested in possible values of Y while X is assumed to be given. By letting $\pi(y \,|\, x) = \pi(x, y)$, the constraint (2) 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 $\pi(x, y)$ according to (2). 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 $\pi(\cdot \,|\, x)$ which are not normalized, i.e. for which $\sup_y \pi(y \,|\, x) < 1$.

Formalizing the SBR Hypothesis

A basic idea of the approach discussed in this paper is to use a possibility rule as defined above in order to formalize the SBR hypothesis. In fact, interpreting X and Y as degrees of similarity between two situations and two results, 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 SBR hypothesis: "The more similar two situations are, the more *possible* it is that the corresponding outcomes are similar" (Dubois *et al.* 1998). Note that this formalization takes the heuristic nature of the SBR 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 (2) it induces lower bounds[4]

$$\pi(s, r) \geq \min\left\{\sigma_{\mathcal{S}}(s, s_1), \sigma_{\mathcal{R}}(r, r_1)\right\} \quad (3)$$

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 extrapolated in accordance with the SBR hypothesis. The more similar $\langle s, r\rangle$ and $\langle s_1, r_1\rangle$ are in the sense of the (joint) similarity measure

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

the more plausible the (hypothetical) case $\langle s, r\rangle$ becomes and, hence, the larger is the (lower) possibility bound (3). 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 (3) to all cases in the memory $\mathcal{M}$ we obtain the possibility distribution $\pi_{\mathcal{C}}$ defined by

$$\pi_{\mathcal{C}}(s, r) = \max_{1 \leq i \leq n} \min\left\{\sigma_{\mathcal{S}}(s, s_i), \sigma_{\mathcal{R}}(r, r_i)\right\} \quad (4)$$

for all $c = \langle s, r\rangle \in \mathcal{S} \times \mathcal{R}$. This distribution can be interpreted as a possibilistic approximation of the concept

[3] We use the same notation for a label, the name of an associated fuzzy set, and the membership function of this set.

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

φ. It is of provisional nature and actually represents lower bounds to possibility degrees (the equality in (4) is justified by a principle of *maximal informativeness*). 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 (4).

Similarity-Based Inference

The distribution (4) can be taken as a point of departure for various inference tasks. For example, given a new situation s_0, a prediction of the associated outcome r_0 is obtained in the form of the conditional possibility distribution

$$\pi_{\mathcal{R}} : r \mapsto \pi_{\mathcal{R}}(r \mid s_0) = \pi_{\mathcal{C}}(s_0, r). \tag{5}$$

For illustrational purposes let a case $c = \langle s, r\rangle$ correspond to a car in the (real-world) AUTOMOBILE DATABASE,[5] where s is the horsepower and r the price of the car. In this small example, which only involves two attributes, the SBR hypothesis simply suggests that "cars with similar horsepower have similar prices." Let $\sigma_{\mathcal{S}} = f_{100}$ with

$$f_M : (x, x') \mapsto [\max\{1 - |x - x'|/M, 0\}]_{\mathcal{L}}, \tag{6}$$

where $\mathcal{L} = \{0, 1/10, \ldots, 1\}$ and $[x]_{\mathcal{L}} = \max\{\lambda \in \mathcal{L} \mid \lambda \leq x\}$. Moreover, let the similarity between two outcomes (= prices) be given by $\sigma_{\mathcal{R}} = f_{10000}$. Figure 1 shows the prediction (5) 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 represented by the 205 cases (cars) in the database 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.

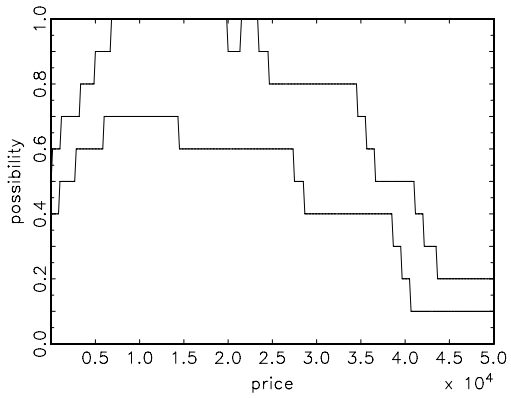

Figure 1: Prediction (5) of the price of a car with horsepower $s_0 = 100$ (upper curve) and prediction (7) for $60 \leq s_0 \leq 140$ (lower curve).

A straightforward computation of the distribution (5) has time-complexity $O(|\mathcal{M}| \cdot |\mathcal{R}|)$. Without going into detail, let us mention that a much more efficient computation is generally possible. First, note that the possibility degree $\pi_{\mathcal{R}}(r)$ is already determined by the case $\langle s', r'\rangle \in \mathcal{M}$ which is maximally similar to $\langle s_0, r\rangle$. Since only this case has to be found, searching the complete memory can be evaded by means of efficient similarity-based indexing methods and corresponding data structures. Second, one does not need to derive the complete distribution $\pi_{\mathcal{R}}$ if only those outcomes r with large possibility $\pi_{\mathcal{R}}(r)$ are of interest. Therefore, the memory $\mathcal{M}$ can be reduced to those cases $\langle s', r'\rangle$ for which s' is similar enough to s_0. Moreover, $\pi_{\mathcal{R}}(r)$ needs to be computed only for those results $r \in \mathcal{R}$ for which there is at least one case $\langle s', r'\rangle$ in the reduced memory such that r is similar to r'. In other words, one can concentrate on the neighborhoods of the results which have already been observed (in situations similar to s_0). This way, complexity might be reduced to almost $O(|\mathcal{M}|)$.

Dealing with Incomplete Information

Suppose that we are interested in, say, the price of a car whose horsepower is between 60 and 140. This amounts to predicting the outcome of a situation s_0 with incompletely specified attributes. The following generalization of (5) is in accordance with the semantics underlying our approach:

$$\pi_{\mathcal{R}}(r) = \pi_{\mathcal{R}}(r \mid \mathcal{S}_0) = \inf_{s \in \mathcal{S}_0} \pi_{\mathcal{C}}(s, r), \tag{7}$$

where $\mathcal{S}_0$ denotes the set of situations $s \in \mathcal{S}$ which are compatible with the characterization of s_0 (i.e. the interval $[60, 140]$ in our example). Indeed, each potential situation $s \in \mathcal{S}_0$ gives rise to a lower bound $\pi_{\mathcal{C}}(s, r)$ according to (4). Without additional knowledge, however, we can guarantee but the smallest of these bounds to be valid. This is in agreement with the idea of *guaranteed possibility* (of an event $X \subset \mathcal{X}$), which is formally derived from a possibility distribution π on $\mathcal{X}$ according to $\Delta(X) = \inf_{x \in X} \pi(x)$ (Dubois and Prade 1996). Figure 1 shows (7) for our example.

The prediction (7) can be generalized further by modeling imprecise knowledge about s_0 in the form of a possibility distribution π_0 on $\mathcal{S}$, where $\pi_0(s)$ corresponds to the degree of possibility that $s_0 = s$. A horsepower of 100, for instance, might appear somewhat more plausible than a horsepower of 80, even though the latter is not completely excluded. A graded modeling of $\mathcal{S}_0$ is useful, e.g., if some attributes are specified linguistically.

Observe that (7) can be interpreted as the possibility of the tuple $\langle s, r\rangle$ which is *guaranteed* by each *possible* situation $s \in \mathcal{S}_0$. Therefore – taking the possibility distribution π_0 which represents the imprecisely known situation s_0 into account – (7) can be generalized as follows:

$$\pi_{\mathcal{R}}(r) = \inf_{s \in \mathcal{S}} \max\{\pi_{\mathcal{C}}(s, r), 1 - \pi_0(s)\}. \tag{8}$$

One obviously recovers (7) from (8) by associating the set $\mathcal{S}_0$ in (7) with a related $\{0, 1\}$-valued possibility distribution, i.e. $\pi_0(s) = 1$ if $s \in \mathcal{S}_0$ and 0 otherwise.

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

(8) estimates the inclusion of the fuzzy set of situations compatible with $\mathcal{S}_0$ in the fuzzy set of situations which are possibly associated with the result r; it represents the *certainty* that a situation, fuzzily restricted by $\mathcal{S}_0$, is possibly associated with r. In the extreme case where $\mathcal{S}_0$ is completely unspecified ($\pi_0 \equiv 1$), (8) yields $\pi_{\mathcal{R}}(r) = \inf_{s \in \mathcal{S}} \pi_{\mathcal{C}}(s, r)$, that is, a fully uninformative result usually equal to 0. This is clearly in agreement with the idea that $\pi_{\mathcal{R}}$ is a lower bound.

A further generalization becomes necessary when allowing for incompletely specified sample cases. Let the ith case in the memory be characterized by the (crisp) set $\mathcal{C}_i = \mathcal{S}_i \times \mathcal{R}_i \subset \mathcal{C}$. Then, (4) becomes

$$\pi_{\mathcal{C}}(s, r) = \max_{1 \leq i \leq n} \inf_{c_i \in \mathcal{C}_i} \sigma_{\mathcal{C}}(\langle s, r \rangle, c_i),$$

which is in accordance with (4) and (7). Moreover, we obtain

$$\pi_{\mathcal{C}}(s, r) = \max_{1 \leq i \leq n} \inf_{c_i \in \mathcal{C}_i} \max\{\sigma_{\mathcal{C}}(\langle s, r \rangle, c_i), 1 - \pi_i(c_i)\}$$

if the ith case is characterized by means of a possibility distribution π_i rather than by a crisp set $\mathcal{C}_i$. Observe that this expression of $\pi_{\mathcal{C}}(s, r)$ can be inserted into (8) in order to handle incomplete specifications of both, the sample cases and the new situation.

Comparison with Related Methods

The Nearest Neighbor Principle

As already mentioned above, the possibilistic extrapolation of cases (PEC) is closely related to k-NEAREST NEIGHBOR (kNN) and instance-based learning (IBL) algorithms, which exploit the concept of similarity (distance) in order to predict the class (= outcome) associated with a new instance (= situation). The extrapolation principle as realized by kNN (and IBL) algorithms is best exemplified by the *majority vote* decision rule which derives an estimation $\widehat{c}_0$ of the class $c_0 \in C$ of a new sample point x_0, from the set X of the k nearest neighbors of x_0, according to

$$\widehat{c}_0 = \arg\max_{c \in C} \operatorname{card}\left(\{x \in X \mid \operatorname{class}(x) = c\}\right). \quad (9)$$

The principle of case extrapolation underlying PEC avoids two questionable properties of the basic kNN approach. This is mainly due to the fact that a prediction in the form of a possibility distribution is more expressive than a "point-estimation" such as (9), while being less ambitious since is only reflects lower bounds.

First, (9) does not reflect the absolute *distance* of the nearest neighbors (cf. Figure 2). In fact, the class of an instance can be extrapolated to instances which are hardly similar.[6] In order to avoid this effect, it has been proposed to apply a reject option (realized in the form of a distance threshold) according to which a classification is refused if the nearest neighbor is not near enough (Tomek 1976). In PEC, extrapolation of case-based information is *bounded* in the sense that results of situations are extrapolated only to similar situations. More precisely, the support of an outcome is graded according to the respective similarity.[7] Thus, the most plausible outcome, r, of a new situation s_0 might still be supported by a rather small possibility degree $\pi_{\mathcal{R}}(r)$ if the nearest neighbors of s_0 are actually quite dissimilar. Note that the extent of extrapolation is also graded in the *weighted* version of the kNN algorithm (Dudani 1976). There, however, the weight depends on the *relative* rather than the *absolute* distance of the respective instance.

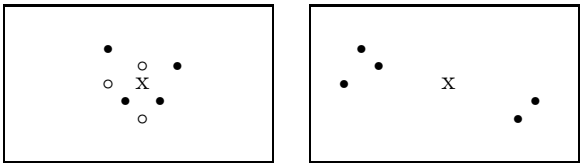

Figure 2: A simple kNN classification does not reflect the ambiguity (left) and the actual support (absolute distance of the neighbors, right) of a decision.

Second, IBL does principally realize a point-estimation when predicting a single class label. An estimation of this kind cannot represent the *ambiguity* caused by the existence of many different class labels among the nearest neighbors (which entails a large classification error, see again Figure 2). Again, the basic algorithm can be extended by a reject option in order to account for this problem (Hellman 1970). By providing a prediction in the form of a possibility distribution, i.e. a more general concept which can support different outcomes to different degrees, ambiguity can be reflected in a rather natural way in PEC. In fact, a prediction is ambiguous if there are several plausible outcomes having quite similar degrees of possibility.

The claim that IBL ignores the uncertainty related to a prediction deserves some further comments. In fact, this claim is actually not true if the kNN classifier is considered in the context of kernel-based density estimation (see the next section): Consider a set of data comprising $n = n_1 + \ldots + n_m$ points (= real-valued vectors) x, where n_i denotes the number of observations x such that $\operatorname{class}(x) = c_i$. Let x_0 be a new observation. We choose the smallest hypersphere around x_0 which contains a set X_0 of $k = k_1 + \ldots + k_m$ points, where $k_i = |\{x \in X_0 \mid \operatorname{class}(x) = c_i\}|$. The conditional probability densities of x_0 can now be estimated by $\varphi(x_0 \mid C = c_i) = k_i/(n_i \cdot v)$, where v denotes the volume of the hypersphere around x_0. Moreover, the unconditional density of x_0 and the prior probability of the class c_i can be estimated by $\varphi(x_0) = k/(n \cdot v)$ and

[6] This generally happens if only few observations have been made.

[7] As a nice side effect, the choice of k (size of neighborhood) thus becomes unnecessary.

$p(C = c_i) = n_i/n$, respectively. We thus obtain

$$p_i = p(C = c_i \mid x_0) = \frac{\varphi(x_0 \mid C = c_i) \cdot \widehat{p}(C = c_i)}{\varphi(x_0)} = \frac{k_i}{k}. \quad (10)$$

According to (10), the class estimated according to the (majority voting) kNN rule is just the one of maximal (posterior) probability. This result provides a formal justification of the rule. Besides, it is an obvious idea to not only consider the estimated class but the probability vector $(p_1, \ldots, p_m)$ obtained from (10) directly, which conveys an idea of the reliability of the prediction.

Still, it is not clear how reliable the estimated probabilities $p_i = k_i/k$ themselves are. It is possible to construct corresponding confidence intervals, but these are only asymptotically valid. In fact, the number k is generally small and, hence, (10) not very confident.[8] Improving the quality of predictions by simply increasing k does obviously not work, since it also entails an enlarging of the hypersphere around x_0.[9] One should also bear in mind that the density estimation based on the kNN principle suffers from further difficulties. For instance, deriving $\varphi(x)$ for all $x \in X$ leads to a non-normalized density function since each x requires a different hypersphere. Besides, it is not clear whether the above derivation can simply be generalized to the case where X is a non-metrical space.

The different extent to which extrapolation takes place in (the probabilistic version of) IBL and PEC is closely related to basic properties of the uncertainty frameworks underlying these approaches. IBL "enforces" a (probabilistic) prediction even if only few examples have been observed. Besides, the predicted probabilites (10) actually reflect the relative rather than the absolute similarity between a new situation and its neighbors. The prediction (10) remains unchanged, e.g. when doubling the distance between x_0 and all of its neighbors, which might then be rather faint. As opposed to this, PEC does not support outcomes before evidence in the form of similar cases is indeed available. In fact, a probabilistic approach is obliged to provide a probability distribution, and IBL takes the one that appears most faithful. As opposed to this, a possibilistic approach such as PEC can express *ignorance*: $\pi_{\mathcal{R}}(\cdot \mid s) \equiv 0$ simply means that no evidence about the outcome in the situation s is *as yet* available. In this connection, let us again emphasize the preliminary character of a possibilistic prediction which merely provides lower bounds.

[8] Note that the estimated probability according to (10) is 1 for the class of the nearest neighbor (and 0 for all other classes) in the special case $k = 1$, i.e. for the 1NN rule.

[9] Good estimations are obtained if both, the number of observations is *large* and the surrounding hypersphere is *small*.

Extensions of the NN Principle

A generalization of the k-NEAREST NEIGHBOR rule which is closely related to PEC and which is also motivated by the aforementioned problems has been proposed in (Denoeux 1995). In this method, each neighbor x_i specifies its "belief" about the class $c_0 \in C$ of a new pattern x_0 by means of a belief function (Shafer 1976) Bel_i resp. an associated mass distribution m_i such that

$$\mathsf{m}_i(\{c_i\}) = \alpha_i, \quad \mathsf{m}_i(C) = 1 - \alpha_i. \quad (11)$$

The degree of support of the hypothesis $c_0 = c_i$, expressed by the weight $0 < \alpha_i < 1$, is a decreasing function of the distance between x_0 and x_i. The evidence in the form of belief functions associated with the k neighbors is then aggregated by using DEMPSTER's rule of combination. Note that the belief structure (11) is consonant, which means that it can also be expressed in terms of a possibility distribution.

The main differences between (Denoeux 1995) and PEC are as follows:

(1) The combination of individual pieces of evidence is realized in different ways, namely by means of a max-aggregation in PEC and by means of DEMPSTER's rule in (Denoeux 1995). Note that the latter assumes the pieces of evidence to be *distinct* (Shafer 1976), which might not always be true in the context of classification.

(2) As in IBL, the method in (Denoeux 1995) 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 PEC.

(3) By focusing on classification as a performance task, the method in (Denoeux 1995) has been developed with a specific application in mind and can be seen as a purely data-driven approach. As will be seen below, PEC 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 SBR model by means of fuzzy set-based (linguistic) modeling techniques.

A further method which is closely related to PEC is the fuzzy kNN algorithm proposed in (Keller *et al.* 1985). In this paper, a new instance x_0 is not assigned to one class in an unequivocal way. Rather, a "fuzzy" classification is realized by specifying a degree of membership for each class. The degree to which x_0 is assigned to the ith class is given by

$$u_i(x_0) = \frac{\sum_{j=1}^{k} u_{ij} \left(1/|x_0 - x_j|^{2/(m-1)}\right)}{\sum_{j=1}^{k} 1/|x_0 - x_j|^{2/(m-1)}}, \quad (12)$$

where u_{ij} is the membership degree of the instance x_j in the ith class.[10] The constant m determines the weighting of the distance between x_0 and its neighbors. Note

[10] The possibility of assigning fuzzy memberships to la-

that (12) assumes a metrical structure and is not in accordance with an ordinal setting. Besides, the membership degrees in (12) sum up to 1 and, hence, represent *relative* rather than absolute evidence.

Kernel-Based Density Estimation

PEC can also be compared to kernel smoothing techniques from non-parametric statistics. Consider a set $\{x_1, \ldots, x_n\} \subset \mathbb{R}^m$ of realizations of a random variable X. A kernel-based estimation (KDE) of an underlying density function is then defined as

$$\varphi : x \mapsto \frac{1}{n} \sum_{i=1}^{n} \kappa_h (x - x_i) = \frac{1}{n} \sum_{i=1}^{n} \kappa \left(\frac{x - x_i}{h} \right), \quad (13)$$

where κ denotes the *kernel function*. Typical examples of κ include the PARZEN window and the normal kernel, the latter being defined as the density of the (multivariate) standard normal distribution. The so-called *kernel width* or *smoothing parameter*, h, has an important effect on the accuracy of the approximation (13). It plays a role somewhat similar to the bin-width of histograms.

The local generalization of observations as realized by (13), i.e. the allocation of probability mass by means of kernel functions, reflects the same line of thought as the SBR hypothesis. Namely, it is (implicitly) assumed that similar (= closely located) points have a similar probability of occurrence. Indeed, the striking similarity between (13) and (5) is revealed by writing the latter as

$$\pi : c \mapsto \max_{1 \leq i \leq n} \sigma_{\mathcal{C}}(c, c_i), \quad (14)$$

where $c_i = \langle s_i, r_i \rangle$ is the ith case in the memory. In fact, (14) can be seen as the possibilistic counterpart to (13): Instead of taking the average over a number of probability densities, the "possibilistic kernels" $\sigma_{\mathcal{C}}(\cdot, c_i)$ associated with the data are combined by means of the maximum operator. Thus, the sum and the product[11] in (13) are replaced by the maximum and the minimum in (14). Consequently, the generalization beyond observed data is completely grounded on the concept of *similarity* and does not fall back on the concept of *frequency*. Of course, the two types of extrapolation are not directly comparable, and their adequacy strongly depends on the application at hand. Here, let us only mention that a purely similarity-based extrapolation is an interesting alternative if frequency is not a reliable information source, e.g., if data is sparse or the assumption of independence is violated. See (Hüllermeier 2000) for a more thorough discussion along these lines.

beled instances is seen as an important feature of the algorithm. It allows, e.g., to increase (decrease) the influence of a labeled instance which is (not) considerded prototypical of a class.

[11] The kernel function κ can be thought of as the product of a set of one-dimensional density functions.

Controlling the Extrapolation

The possibilistic extrapolation of case-based information in the sense of (3) relies on the heuristic assumption underlying SBR. It should, therefore, take into account whether the related reasoning principle is actually valid. That is, the less the current application seems to meet the SBR hypothesis, the more cautious one should be when considering an observed case as evidence for the existence of similar cases. To this end, several extensions of the basic model can be realized which allow one to modulate the extent of similarity-based extrapolation.

Particularly, the basic model can be rendered more flexible by making use of so-called (linguistic) modifiers (Zadeh 1972) in (4), i.e. non-decreasing functions $m_1, m_2 : \mathcal{L} \to \mathcal{L}$. This leads to possibility rules $m_1 \circ A \overset{m_2}{\rightsquigarrow} B$ with associated distributions $\pi_{\mathcal{C}}$, where

$$\pi_{\mathcal{C}}(s, r) = \max_{1 \leq i \leq n} m_2 \left(\min \left\{ m_1(\sigma_{\mathcal{S}}(s, s_i)), \sigma_{\mathcal{R}}(r, r_i) \right\} \right).$$

Both modifiers 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 $\mathcal{L}$) m_1 and m_2 are, the stronger (in the sense of asserted possibility degrees) a case $\langle s_i, r_i \rangle$ is extrapolated. The modifier functions m_1 and m_2 provide a convenient way of incorporating domain-specific (expert) knowledge (expressed in a linguistic way), thereby combining data-driven and knowledge-based reasoning. Apart from expert knowledge, learning (calibration) methods can be used in order to adapt the inference principle to the application at hand.

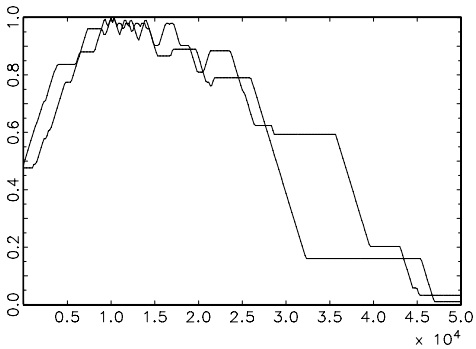

Figure 3: Prediction induced by two different rules (using the scale $\mathcal{L} = \{0, 1/100, \ldots, 1\}$). The overall prediction is the pointwise maximum.

Further extensions of the basic framework include the discounting of untypical cases, the restriction of inference rules to (fuzzily) bounded regions, and the modeling of the SBR hypothesis by combining several rules (Hüllermeier *et al.* 2000). Thus, a similarity-guided extrapolation principle might be expressed linguistically as follows: "Cars have similar prices if either the horsepower is *very similar* or both, the engine-size is *similar* and the peak-rpm is *approximately similar*." This

model corresponds to two rules and can be translated by making use of appropriate modifier functions and logical connectives. When making a prediction, each rule contributes a possibility distribution, and the overall prediction is obtained by taking the (pointwise) maximum (since a certain outcome can be supported by any of the two rules, see Fig. 3).

Interestingly enough, our framework is also well-suited for incorporating background knowledge which is expressed in the form of fuzzy rules or other types of constraints. As an example consider the following *convexity constraint* according to which intermediary prices of a car are not less possible than more extreme prices:

$$r' \leq r \leq r'' \;\Rightarrow\; \min\{\pi(s,r'),\pi(s,r'')\} \leq \pi(s,r)$$

for all $s \in \mathcal{S}$ and $r, r', r'' \in \mathcal{R}$. In order to satisfy this constraint it is necessary to replace the possibility distribution $\pi_{\mathcal{C}}$ by its convex hull

$$\pi_{\mathcal{C}}^{ch} : (s,r) \mapsto \min\left\{ \max_{r' \leq r} \pi_{\mathcal{C}}(s,r'), \max_{r \leq r''} \pi_{\mathcal{C}}(s,r'') \right\}.$$

Note that such constraints are generally more difficult to handle in related methods such as, say, kernel-based density estimation.

Calibration of SBR models

The methodological framework introduced so far provides a broad spectrum of modeling techniques, including the modification of similarity measures (via corresponding modifiers), the discounting of rules, and the discounting of individual cases. 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. Rather, the idea is that the expert specifies only the coarse structure of a model, i.e. the linguistic rules. The ultimate SBR model is then determined 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.

This section is meant to discuss this type of model calibration in more detail. More precisely, 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 $m_1 \circ \sigma_{\mathcal{S}} \rightsquigarrow \sigma_{\mathcal{R}}$. Each of these rules induces a related possibility distribution

$$(s,r) \mapsto \max_{1 \leq i \leq n} \max\left\{ m_1(\sigma_{\mathcal{S}}(s,s_i)), \sigma_{\mathcal{R}}(r,r_i) \right\}, \quad (15)$$

and the overall prediction $\pi_{\mathcal{C}}$ 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 parametrized functions. For instance, the modifier associated with the linguistic hedge "very" might be specified by the function $x \mapsto [x^{\alpha}]_{\mathcal{L}}$ with $\alpha > 1$. Likewise, the similarity of horsepowers, σ_{hp}, might be given by f_M in (6), where M plays the role of a parameter. All these parameters can be combined into one vector ϑ which determines the SBR 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 the smoothing parameter in kernel-based estimation of probability density functions.

In order to determine ϑ and, hence, a concrete SBR 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 e.g.

$$\sum_{c \in \mathcal{C}} \left(\pi_{\mathcal{C}}(c \,|\, \vartheta) - \pi_{\varphi}(c)\right)^2, \quad (16)$$

between the estimated distribution $\pi_{\mathcal{C}}(\cdot \,|\, \vartheta)$ and the (true) $\{0,1\}$-valued distribution π_{φ} defined by $\pi_{\varphi}(c) = 1 \Leftrightarrow c \in \varphi$. A similar procedure is used in kernel-based density estimation, where one way of determining the smoothing parameter h is to minimize the integrated squared error

$$\mathsf{ISE}(h) = \int \left(\varphi(x) - \varphi_h(x)\right)^2 \quad (17)$$

between the true density φ and the estimation φ_h. Of course, (16) cannot be derived since $\pi_{\varphi}(c)$ is not known for all $c \in \mathcal{C}$. In fact, the same problem occurs in (17), where the density φ is unknown. A possible way out is to apply a (leave-one-out) cross validation procedure which considers only the observed values, i.e. which approximates the integral by a weighted sum, and which replaces the density φ by a further estimation $\widehat{\varphi}$. This leads to the minimization of

$$\sum_{i=1}^{n} \left(\widehat{\varphi}_h(x_i) - \varphi_h(x_i)\right)^2, \quad (18)$$

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

The idea of such a cross validation can also be applied to (16). In this case, we do not even have to estimate the values $\pi_{\varphi}(c_i)$ since $\pi_{\varphi}(c_i) = 1$ holds true for each observation $c_i \in \mathcal{M}$. However, by restricting ourselves to the observed cases, the minimization problem becomes ill-posed. In fact, a trivial solution to the problem of minimizing

$$\sum_{c \in \mathcal{M}} \left(\pi_{\mathcal{C}}(c \,|\, \vartheta) - \pi_{\varphi}(c)\right)^2 \quad (19)$$

is given by $\pi_{\varphi} \equiv 1$. This simply means to choose the parameter ϑ such as to maximize the extrapolation of cases, a hardly convincing result.

In this connection, it deserves mentioning that a possibilistic prediction $\pi_{\mathcal{C}}$ can principally not be "falsified": 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 $\pi_{\mathcal{C}}(c \,|\, \vartheta)$ is not necessarily a defect of the model, i.e. it does not necessarily indicate a poor choice of the parameter ϑ. (Recall that predicted possibility degrees are only lower bounds, and that low degrees are quite natural if the memory $\mathcal{M}$ does not contain many cases similar to c. Particularly, the lower possibility bound obtained from the remaining cases is 0 for completely "isolated" cases which are as yet not supported by any similar observation.) 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.

Interestingly enough, the (partial) exclusion of cases according to the SBR principle can be realized by means of a different type of extrapolation principle induced by a different sort of fuzzy rule: A *certainty rule* $A \rightsquigarrow B$ corresponds to the statement that "the more x is in A, the more *certain* it is that y is in B" and gives rise to a constraint of the form

$$\pi(x, y) \leq \max\{1 - A(x), B(y)\}.$$

Based on this interpretation, an SBR rule $m_1 \circ \sigma_{\mathcal{S}} \rightsquigarrow \sigma_{\mathcal{R}}$ entails the distribution

$$(s, r) \mapsto \min_{1 \leq i \leq n} \max\left\{1 - m_1(\sigma_{\mathcal{S}}(s, s_i)), \sigma_{\mathcal{R}}(r, r_i)\right\}. \tag{20}$$

This distribution, which actually represents upper bounds, defines the counterpart to (15). The overall prediction $\overline{\pi}_{\mathcal{C}}$, associated with a set of rules, is defined by the intersection (pointwise minimum) of the distributions (20). 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 situations are similar but the outcomes are rather different.

Example 1 *Consider as an example a case* $(100, 15000)$*, i.e. a car with horsepower 100 and price \$15,000, in connection with the similar horsepower–similar price principle. According to the possibility rule model (15), this case (partly)* supports *the case* $(110, 16000)$ *which has a similar horsepower and a similar price. According to the certainty rule model (20), 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.[12] Note that the prediction $\pi_{\mathcal{C}}$ derived from (15) and the prediction $\overline{\pi}_{\mathcal{C}}$ obtained from (20) might be *conflicting* in the sense that $\overline{\pi}_{\mathcal{C}}(c) < \pi_{\mathcal{C}}(c)$ for some 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 some other observation $c_2 \in \mathcal{M}$ (according to the certainty rule model). A situation of this kind indicates a defect of the underlying SBR model (the lower possibility bound is larger than the upper bound). Note that a conflict 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 $\pi_{\mathcal{C}}(c)$ and $\overline{\pi}_{\mathcal{C}}(c)$ will be close to 0 and 1, respectively.

Example 2 *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:*

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

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

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* SBR *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

[12] 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 (Ughetto *et al.* 1999; Weisbrod 1998).

(15) resp. (20) 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 *correct support* (of observed cases) and a *consistency* principle:

- Observed cases should be supported correctly, i.e. 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

$$\mathsf{supp}_\vartheta(c) = \pi_{\mathcal{C}}(c \,|\, \vartheta), \tag{21}$$

where $\pi_{\mathcal{C}}(\cdot \,|\, \vartheta)$ is derived from $\mathcal{M} \setminus \{c\}$ according to (15) 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

$$\mathsf{conf}_\vartheta(c) = \max\{0, \pi_{\mathcal{C}}(c \,|\, \vartheta) - \overline{\pi}_{\mathcal{C}}(c \,|\, \vartheta)\}, \tag{22}$$

where $\overline{\pi}_{\mathcal{C}}(c \,|\, \vartheta)$ is the distribution obtained from the certainty rule model (20). Note that the subtraction in (22) is actually not in line with a purely ordinal interpretation of possibility. Based on an adequate definition of the possibility scale, however, it will generally serve well enough as a rough specification of the discrepancy between the two predictions. Still, one might also think of using a purely qualitative measure of conflict:

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

The derivation of (21) and (22) for all cases in the memory yields n degrees of support and conflict, respectively. The overall support induced by the parameter ϑ, $\mathsf{supp}(\vartheta)$, can then be obtained by aggregating these values:

$$\mathsf{supp}(\vartheta) = A(\{\mathsf{supp}_\vartheta(c) \,|\, c \in \mathcal{M}\}) \tag{23}$$

with A being an aggregation function. A measure $\mathsf{conf}(\vartheta)$ 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

$$\mathsf{supp}(\vartheta) - \alpha \cdot \mathsf{conf}(\vartheta) \tag{24}$$

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

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 (23). Yet, an aggregation which is more in accordance with a qualitative setting is the SUGENO integral

$$\sup_{\alpha \geq 0} \min\{\alpha, \mu(F_\alpha)\}, \tag{25}$$

where $F_\alpha = \{c \in \mathcal{M} \,|\, \mathsf{supp}_\vartheta(c) \geq \alpha\}$ for $0 \leq \alpha \leq 1$. The measure μ in (25) can be taken as the counting measure, i.e. $\mu(A) = |A|/|\mathcal{M}|$ for all $A \subset \mathcal{M}$.

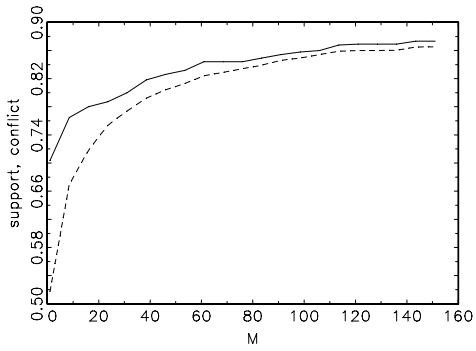

Figure 4: Support (solid line) and conflict as a function of the parameter M which defines the similarity measure for the attribute horsepower.

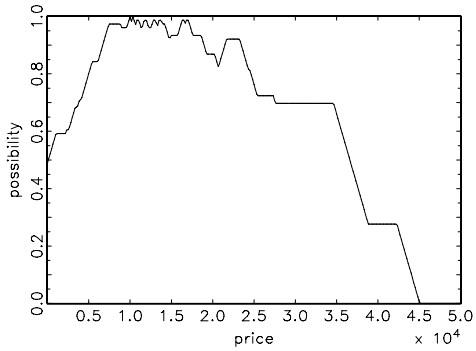

Figure 5: Prediction of the price of a car with horsepower 100, where $\sigma_{hp} = f_{76}$.

Example 3 *Consider as a simple example the choice of the parameter M which defines $\sigma_{hp} = f_M$ in connection with the similar horsepower–similar price hypothesis (using $\sigma_{\mathcal{R}} = f_{3000}$). Figure 4 shows $\mathsf{supp}(M)$ and $\mathsf{conf}(M)$, defined according to (21), (22), and the aggregation (25) as a function of M. The choice of $\alpha = 3/4$ in (24) suggests $M = 76$ as an optimal parameter and leads to the prediction shown in Figure 5.*

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 (19). In the probabilistic case, the correct support corresponds to the true probability, as expressed by (18).

Concluding Remarks

The basic inference principle underlying instance-based reasoning relies on a similarity-guided extrapolation of observed cases. We have formalized this principle in the framework of possibility theory and fuzzy sets: An already encountered case is taken as evidence for the existence of similar cases, and this evidence is expressed in terms of degrees of possibility assigned to hypothetical cases. This inference principle provides an interesting alternative to the classical NEAREST NEIGHBOR principle and extensions thereof. Here, we have emphasized that our method adequately represents the uncertainty involved in the heuristic SBR principle and also allows for dealing with incompletely specified observations.

PEC has been implemented in the framework of an experimental information system maintained at IRIT (PRETI, for Platform of Research and Experimentation in the Treatment of Information). This platform uses a database describing houses, specified in terms of about 30 attributes (binary, symbolic or numerical), which can be rented weekly for vacations. The system offers various facilities apart from PEC (e.g. flexible querying allowing for the expression of the user's preferences, querying by examples, ...) which help the user to retrieve houses of interest and to understand the situation of the market. Especially in the latter aspect, PEC is useful. In fact, apart from making predictions, it can also support certain types of data analysis. For instance, suppose the user suspects that some attributes, say, the number of people who can be accommodated or the distance to the sea, can influence the price. Then, using PEC, a possibility distribution for the price is built on the basis of the prices of the houses more or less similar to the type of house the user is looking for. It may happen that the range of more or less possible prices is quite large, perhaps due to the fact that the chosen attributes do not really influence the price, or that they do not sufficiently specify the type of house. In this latter case, other attributes may be added to the specification, e.g. the presence of a dish-washer. This might lead to a more focused range of prices. The feature selection problem, i.e. the automatic determination of decisive attributes, is an important topic in instance-based learning. The reconsideration of this problem in the framework of PEC might provide new insights and, hence, is an interesting point of further research.

References

A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1):39–59, 1994.

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

B.V. Dasarathy, editor. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Comp. Soc. Press, 1991.

T. Denoeux. A k-nearest neighbor classification rule based on Dempster-Shafer Theory. IEEE *Transactions on Systems, Man, and Cybernetics*, 25(5):804–813, 1995.

D. Dubois and H. Prade. What are fuzzy rules and how to use them. *Fuzzy Sets and Systems*, 84:169–185, 1996.

D. Dubois, F. Esteva, P. Garcia, L. Godo, R. Lopez de Mantaras, and H. Prade. Fuzzy set modelling in case-based reasoning. *International Journal of Intelligent Systems*, 13, 1998.

S.A. Dudani. The distance-weighted k-nearest-neighbor rule. IEEE *Transactions on Systems, Man, and Cybernetics*, SMC-6(4), 1976.

M.E. Hellman. The nearest neighbor classification rule with a reject option. IEEE *Transactions on Systems, Man, and Cybernetics*, SSC-6:179–185, 1970.

E. Hüllermeier, D. Dubois, and H. Prade. Knowledge-based extrapolation of cases: A possibilistic approach. In *Proceedings* IPMU–2000, pages 1575–1582, Madrid, July 2000.

E. Hüllermeier. Toward a probabilistic formalization of case-based inference. In *Proceedings* IJCAI–99, pages 248–253, 1999.

E. Hüllermeier. Similarity-based inference: Models and applications. Technical Report 00-28 R, IRIT, October 2000.

J.M. Keller, M.R. Gray, and J.A. Givens. A fuzzy k-nearest neighbor algorithm. IEEE *Transactions on Systems, Man, and Cybernetics*, SMC-15(4):580–585, 1985.

E. Plaza, F. Esteva, P. Garcia, L. Godo, and R. Lopez de Mantaras. A logical approach to case-based reasoning using fuzzy similarity relations. *Information Sciences*, 106, 1998.

G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

I. Tomek. A generalization of the k-NN rule. IEEE *Transactions on Systems, Man, and Cybernetics*, SMC-6:121–126, 1976.

L. Ughetto, D. Dubois, and H. Prade. Implicative and conjunctive fuzzy rules: A tool for reasoning from knowledge and examples. In *Proceedings* AAAI-99, Orlando, 1999.

J. Weisbrod. A new approach to fuzzy reasoning. *Soft Computing*, 2:89–99, 1998.

R.R. Yager. Case-based reasoning, fuzzy systems modelling and solution composition. In *Proc.* ICCBR-97, pages 633–643, 1997.

L.A. Zadeh. A fuzzy-set theoretic interpretation of linguistic hedges. *J. Cybernetics*, 2(3):4–32, 1972.