Similarity-Based Inference as Evidential Reasoning

Eyke Hüllermeier[1]

Abstract. We make use of a probabilistic model in order to formalize the basic assumption underlying case-based reasoning (CBR), suggesting that "similar problems have similar solutions." Taking this model as a point of departure, we propose a similarity-guided inference scheme in which case-based evidence is represented in the form of belief functions over the set of solutions, and in which the combination of evidence derived from individual cases is considered in the context of information fusion. Our approach is meant to support the overall process of problem solving by estimating the quality of potential solutions. Besides, it reveals that probabilistic methods and related techniques from the field of reasoning under uncertainty provide a convenient framework in which parts of the CBR methodology can be formalized. This framework seems particularly suitable since it allows for taking the heuristic and, hence, uncertain character of case-based problem solving into account.

1 INTRODUCTION

The "CBR hypothesis" which, loosely speaking, suggests that "similar problems have similar solutions," is a major assumption of case-based reasoning (CBR), and a guiding principle of the related problem solving methodology [14]. Recently, some attempts at formalizing this hypothesis in a systematic way and, thus, at making an important step toward a theoretical foundation of CBR have been made [8, 12, 15]. The probabilistic formalization advocated in [12] takes into account that the CBR hypothesis should be understood, not as a universally valid rule, but rather as a "rule of thumb" which only holds true in a general way. In fact, a probabilistic model, according to which similar problems are (at most) *likely* to have similar solutions, can be seen as a more faithful description of this assumption. Particularly, it still allows for "exceptions to the (CBR) rule."

In this paper, we put the probabilistic framework of [12] into a concrete (similarity-based) inference scheme. Our approach amounts to representing experience from a previously encountered case in the form of a belief function characterizing the solution to a new problem. The combination of individual pieces of evidence derived from several cases can thus be considered in the context of *information fusion.*

By putting emphasis on CBR as a *prediction* method [8, 9], the framework in [12] essentially concerns the REUSE process within the (informal) R^4 model of the so-called CBR cycle [1]. In order to point out the restriction to a certain aspect of CBR, we shall refer to the approach proposed in this paper as *similarity-based inference* (SBI). The latter is closely related to lazy learning algorithms [2], particularly those which are derivatives of the k-NEAREST NEIGHBOR (kNN) classifier [6]. Yet, there are also important methodological differences. Besides, SBI is not intended as a special performance task such as, e.g., classification or function learning.

The remaining part of the paper is organized as follows: The next section briefly reviews the probabilistic framework introduced in [12], and extends the related concept of a probabilistic similarity profile. In Section 3, we propose to look at cases as information sources, and to consider SBI as a problem of information fusion. Based on this interpretation, a method of processing case-based evidence is discussed in Section 4 and Section 5. The paper concludes with a brief summary and some remarks.

2 A PROBABILISTIC FRAMEWORK

The primitive concept of a *case* is thought of as a tuple consisting of a *situation* and a *result* or *outcome* associated with the situation.[2]

Definition 1 (SBI **setup)** *An* SBI *setup is a 6-tuple* $\Sigma = \langle(\mathcal{S}, \mu_{\mathcal{S}}), \mathcal{R}, \varphi, \sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, \mathcal{M}\rangle$*, where* $\mathcal{S}$ *is a finite set of situations endowed with a probability measure* $\mu_{\mathcal{S}}$ *(on* $2^{\mathcal{S}}$*),* $\mathcal{R}$ *is a set of results, and* $\varphi : \mathcal{S} \to \mathcal{R}$ *assigns results to situations. The functions* $\sigma_{\mathcal{S}} : \mathcal{S} \times \mathcal{S} \to [0, 1]$ *and* $\sigma_{\mathcal{R}} : \mathcal{R} \times \mathcal{R} \to [0, 1]$ *define (reflexive and symmetric) similarity measures over the set of situations and the set of results, respectively.* $\mathcal{M}$ *is a finite memory*

$$\mathcal{M} = \big(\langle s_1, r_1\rangle, \langle s_2, r_2\rangle, \ldots, \langle s_n, r_n\rangle\big) \tag{1}$$

of cases $c = \langle s, \varphi(s)\rangle \in \mathcal{S} \times \mathcal{R}$*. Let* $D_{\mathcal{S}} \doteq \{\sigma_{\mathcal{S}}(s, s') \mid s, s' \in \mathcal{S}\}$ *and* $D_{\mathcal{R}} \doteq \{\sigma_{\mathcal{R}}(\varphi(s), \varphi(s')) \mid s, s' \in \mathcal{S}\}$ *denote the sets of actually attained similarity degrees.*

The probability measure $\mu_{\mathcal{S}}$ in Definition 1 models the occurence of cases. Thus, it is assumed that situations are chosen repeatedly (and independently) according to $\mu_{\mathcal{S}}$.[3] This kind of statistical assumption (on the distribution of training sets) is typical of machine learning.

Even though we assume that situations determine outcomes, the derivation of results might involve a computationally complex process. In this connection, we understand *similarity-based inference* as a method supporting the overall process of problem solving by bringing plausible results of a new situation into focus. Thus, given an SBI *problem* $\langle\Sigma, s_0\rangle$ consisting of a setup Σ and a new situation $s_0 \in \mathcal{S}$, the task is to predict the result $r_0 = \varphi(s_0)$ associated with s_0. To this end, SBI performs according to the CBR *principle*: It exploits experience in the form of precedent cases to which it "applies" background knowledge in the form of the heuristic CBR hypothesis.

2.1 Probabilistic Similarity Profiles

Consider a problem $\langle\Sigma, s_0\rangle$ with a memory (1) of cases. According to the stochastic occurence of situations, the sequence $(s_1, \ldots, s_n, s_0)$

[1] Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier, France, email: eyke@irit.fr

[2] For reasons of generality, these expressions are preferred to the commonly used terms "problem" and "solution."

[3] $\mathcal{M}$ is hence a sequence of not necessarily different cases.

can be seen as the realization of a random sequence of situations $(S_1, \ldots, S_n, S_0)$ which is characterized by the probability measure

$$(\mu_{\mathcal{S}})^{n+1} \doteq \underbrace{\mu_{\mathcal{S}} \otimes \mu_{\mathcal{S}} \otimes \ldots \otimes \mu_{\mathcal{S}}}_{n+1 \text{ times}}. \quad (2)$$

This measure defines the (discrete) probability space $(\mathcal{S}^{n+1}, (\mu_{\mathcal{S}})^{n+1})$ underlying the SBI problem.

In accordance with the CBR hypothesis, SBI is particularly concerned with modelling the (similarity) relation between *pairs* of cases. Thus, we shall pay special attention to (2) with $n = 1$. The more general case $n > 1$ and the related problem of combining (uncertain) evidence obtained from different cases will be discussed in Section 5.

Consider a random tuple $(S, S') \in \mathcal{S} \times \mathcal{S}$ of situations. The random variable $Z = (X, Y)$, with $X = \sigma_{\mathcal{S}}(S, S')$ being the similarity of the situations, and $Y = \sigma_{\mathcal{R}}(\varphi(S), \varphi(S'))$ denoting the similarity of the associated outcomes, is then defined on the probability space $(\mathcal{S} \times \mathcal{S}, \mu_{\mathcal{S}} \otimes \mu_{\mathcal{S}})$ as the mapping

$$(s, s') \mapsto \left(\sigma_{\mathcal{S}}(s, s'), \sigma_{\mathcal{R}}(\varphi(s), \varphi(s'))\right).$$

Let $\mu_Z \doteq Z(\mu_{\mathcal{S}} \otimes \mu_{\mathcal{S}})$ be the induced probability distribution on $D_{\mathcal{S}} \times D_{\mathcal{R}}$ and define the marginal distributions μ_X on $D_{\mathcal{S}}$ and μ_Y on $D_{\mathcal{R}}$ in the same way. We write $\mu_{\mathcal{S}}(s)$ instead of $\mu_{\mathcal{S}}(\{s\})$ for $s \in \mathcal{S}$. We also use notations such as $(X = x)$ for events $X^{-1}(x)$ and $\mu_{Y|(X=x)} \doteq Y((\mu_{\mathcal{S}} \otimes \mu_{\mathcal{S}})(\cdot \mid X^{-1}(x)))$ to denote corresponding conditional probabilities.

Definition 2 (probabilistic similarity profile) *Consider an* SBI *setup* Σ *and let* $\mathcal{P}(D_{\mathcal{R}})$ *denote the class of probability measures over* $D_{\mathcal{R}}$*. The function*

$$H_\Sigma : D_{\mathcal{S}} \to \mathcal{P}(D_{\mathcal{R}}),\ x \mapsto \mu_{Y|(X=x)}$$

is called the probabilistic similarity profile (PSP) *of* Σ.

The PSP H_Σ is intended as a characterization of the *similarity structure* of the *system* under consideration, $(\mathcal{S}, \mathcal{R}, \varphi)$. For each degree of similarity $x \in D_{\mathcal{S}}$, it specifies the probability distribution of the similarity of results, i.e., of the random variable Y, given the similarity of two situations. It thus gives a precise meaning to the CBR assumption that "similar situations are *likely* to have similar outcomes." In fact, it clarifies the meaning of "likely" in terms of probability distributions and depicts its dependency on the similarity of situations.

A PSP is a *global* model of a similarity structure in the sense that it applies in the same way to all (pairs of) cases. The stronger the similarity structure is developed, the more informative this model will be. In fact, the PSP provides a precise idea of the extent to which the CBR hypothesis holds true for the system under consideration.[4] However, quite often the CBR assumption will not be satisfied equally well for all parts of the *instance space* $\mathcal{S} \times \mathcal{R}$.[5] In such situations, the PSP might be misleading in the sense that it pretends too much precision for the "critical" regions and too little for those regions where the CBR assumption holds fairly.

One possibility of avoiding this problem is to partition the set $\mathcal{S}$ of situations and to derive respective local models. However, since φ is generally unknown, the definition of such 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 maintaining an individual similarity profile for each case in the memory. This approach is somehow comparable to the use of *local metrics* in kNN algorithms and instance-based learning, e.g., metrics which allow feature weights to vary as a function of the instance [18]. It leads us to introduce the concept of a *local similarity profile*.

Definition 3 (local similarity profile) *Consider a fixed situation* $s \in \mathcal{S}$*, and let* S *be distributed according to* $\mu_{\mathcal{S}}$*. Moreover, let* $X_s = \sigma_{\mathcal{S}}(s, S), Y_s = \sigma_{\mathcal{R}}(\varphi(s), \varphi(S))$*. The local probabilistic similarity profile associated with* s*, or* s-PSP*, is defined as*

$$H^s_\Sigma : D_{\mathcal{S}} \to \mathcal{P}(D_{\mathcal{R}}),\ x \mapsto \mu_{Y_s|(X_s=x)}.$$

A collection $H^{\mathcal{M}}_\Sigma = \{H^s_\Sigma \mid \langle s, \varphi(s)\rangle \in \mathcal{M}\}$ *of local profiles is called a local* $\mathcal{M}$-PSP.

One verifies that the (global) PSP (cf. Definition 2) is a (pointwise) weighted average of the local profiles associated with individual cases:

$$\forall\, x \in D_{\mathcal{S}} :\ H_\Sigma(x) \propto \sum_{s \in \mathcal{S}} \alpha(s, x) \cdot H^s_\Sigma(x),$$

where H_Σ denotes the PSP of a setup Σ, and H^s_Σ is the local PSP associated with $s \in \mathcal{S}$. Moreover, $\alpha(s, x) = \mu_{\mathcal{S}}(s) \cdot [X_s(\mu_{\mathcal{S}})](x)$ for all $s \in \mathcal{S}$, where $X_s : \mathcal{S} \to D_{\mathcal{S}}$ denotes the mapping $s' \mapsto \sigma_{\mathcal{S}}(s, s')$.

2.2 Probabilistic Similarity Hypotheses

Of course, knowledge about the (local) PSP of a certain setup Σ will generally be incomplete. This motivates the related concept of a *similarity hypothesis*, which is thought of as an estimation of a PSP. It can hence be seen as an expression of the CBR hypothesis at a formal level.

Definition 4 (similarity hypothesis) *A probabilistic similarity hypothesis is identified by a function* $H : D_{\mathcal{S}} \to \mathcal{P}(D_{\mathcal{R}})$*. A local* $\mathcal{M}$*-hypothesis is a collection* $H^{\mathcal{M}}$ *of hypotheses* $H^s : D_{\mathcal{S}} \to \mathcal{P}(D_{\mathcal{R}})$ *related to cases* $\langle s, \varphi(s)\rangle \in \mathcal{M}$.

Maintaining a local $\mathcal{M}$-hypothesis seems particularly reasonable if only few cases are stored in the memory. The estimation of local hypotheses H^s will then be practicable, even though it is computationally more expensive and does generally require more data than the (reliable) estimation of a global profile. Note that a global hypothesis H can reasonably serve as a prior estimation H^s when storing a new case $\langle s, r\rangle$ in the memory. Further observations can then be used for adapting this (local) hypothesis to the situation s.

A similarity hypothesis can originate from different sources. Firstly, it might express a quantification of the CBR assumption based on some (domain-specific) background knowledge. Secondly, it is a natural idea to consider the acquisition of hypotheses as a problem of *case-based learning*, i.e., to learn hypotheses from observed (pairs of) cases. This way, SBI combines *instance-based learning*, which essentially corresponds to the organization of a memory of cases, and *model-based learning*, namely the learning of similarity hypotheses. Within the probabilistic setting of this section, the learning of hypotheses comes down to estimating a class of probability distributions, and can hence be considered in the context of statistical inference. Interestingly enough, a Bayesian approach can be used

[4] This quantification might be carried even further by considering information measures for the probability distributions associated with a PSP.

[5] In a game playing context, for example, the CBR principle hardly applies to certain "tactical" situations [16].

for combining the two aforementioned approaches: Taking a prior estimation as a point of departure, a similarity hypothesis is improved in the light of observed data. Since the problem of case-based learning is not addressed in this paper we shall subsequently assume the existence of a similarity hypothesis without scrutinizing its origin.

Now, consider an SBI problem $\langle \Sigma, s_0 \rangle$ and let H be a hypothesis related to the similarity profile H_Σ. Moreover, let $\langle s, r \rangle$ be a case from the memory $\mathcal{M}$. Knowing the similarity of situations, $x = \sigma_\mathcal{S}(s, s_0)$, the hypothesis H allows for characterizing the (unknown) similarity of outcomes, $\sigma_\mathcal{R}(r, r_0)$, by means of the random variable $Y \sim H(x)$. At this point, two characteristic properties of SBI become obvious. Firstly, SBI is *indirect* in the sense that predictions of outcomes are only obtained in a second step from predictions of similarity degrees, which are derived first. This necessitates the *transformation* of distributions on $D_\mathcal{R}$ into distributions on $\mathcal{R}$. Secondly, SBI is *local* in the sense that a PSP (just as the CBR hypothesis itself) supports the derivation of predictions from *single* cases. Given a memory of several cases, this calls for the *combination* of probabilistic evidence obtained from individual observations. The transformation and combination of evidence will be discussed in Section 4 and Section 5, respectively.

3 CASES AS INFORMATION SOURCES

The combination of probabilistic evidence in connection with SBI can be considered in a more general context, namely the *parallel combination of information sources*. The problem of combining concurrent pieces of (uncertain) evidence arises in many fields such as, e.g., robotics (sensor fusion) or knowledge-based systems (expert opinion pooling), and it has been dealt with in a probabilistic setting [11] as well as alternative uncertainty frameworks [5]. The combination of evidence derived from individual cases is perhaps best compared to that of expert opinion pooling: Each case corresponds to an expert, and the prediction of the unknown outcome associated with a case is interpreted as an expert statement. The task is to synthesize these statements.

A general framework for the parallel combination of information sources which seems suitable for our purpose has been introduced in [10]. A basic concept within this framework is that of an *imperfect specification*: Let Ω denote a (finite) 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 p_C is a probability measure over a (finite) set C of *specification contexts* and γ is a function $C \to 2^\Omega$.[6] The problem of combining evidence is then defined as generating one imperfect specification Γ from n imperfect specifications $\Gamma_1, \ldots, \Gamma_n$, issued by n different information sources.

From a semantical point of view, a specification context $c \in C$ can be seen as a physical or observation-related frame condition, and $\gamma(c)$ is the (most specific) characterization of ω_0 that can be provided by the information source in the context c. The value $p_C(c)$ can be interpreted as an (objective or subjective) probability of selecting c as a true context. An imperfect specification is thus able to model *imprecision* as well as *uncertainty*: The measure p_C accomplishes the consideration of (probabilistic) uncertainty. Moreover, the modelling of imprecision becomes possible due to the fact that γ is a *set-valued* function.

[6] Formally, an imperfect specification is nothing but a set-valued mapping on a probability space, a well-known concept in connection with random set approaches [7].

4 TRANSFORMATION OF EVIDENCE

According to the indirect approach realized by SBI, evidence concerning outcomes is derived in two stages, where the second step consists of translating evidence concerning similarity degrees, given in the form of probability measures, into evidence about results.

Consider a probability measure $\mu = H(\sigma_\mathcal{S}(s, s_0))$ over $D_\mathcal{R}$ which has been derived from a case $\langle s, r \rangle$, and which is taken as evidence concerning the similarity between r and the unknown outcome $r_0 = \varphi(s_0)$. When interpreting this case as an information source $\Gamma = (\gamma, p_C)$, the set of specification contexts is given by the set of possible degrees of similarity $y = \sigma_\mathcal{R}(r, r_0)$. That is,

$$\begin{aligned} C &= D_\mathcal{R}, \\ \gamma(c) &= \sigma_\mathcal{R}^{(-1)}(r, c), \\ p_C(c) &= \mu(c), \end{aligned} \tag{3}$$

where $\sigma_\mathcal{R}^{(-1)}(r, c) \doteq \{r' \in \mathcal{R} \mid \sigma_\mathcal{R}(r, r') = c\}$ for all $c \in C$. The set $\gamma(c)$ is obviously the most specific restriction of $\varphi(s_0)$ which can be derived from the case $\langle s, r \rangle$ in the context c, i.e., from the assumption that $\sigma_\mathcal{R}(\varphi(s), \varphi(s_0)) = c$ and the fact that $\varphi(s) = r$. Observe that we may have $\gamma(c) = \emptyset$ for some $c \in C$, which means that c cannot be a true context and that Γ is contradictory [10]. It is then necessary to replace Γ by a revised specification $\Gamma' = (\gamma', p_{C'})$. The latter is defined by

$$\begin{aligned} C' &= \{c \in C \mid \gamma(c) \neq \emptyset\}, \\ \gamma'(c') &= \gamma(c'), \\ p_{C'}(c') &= k \cdot p_C(c') \end{aligned}$$

for all $c' \in C'$, with k being the normalization factor, i.e.,

$$1/k = \sum_{c \in C \,:\, \gamma(c) \neq \emptyset} p_C(c). \tag{4}$$

Subsequently, the imperfect specification associated with a case $\langle s, r \rangle$ will always refer to the already revised specification.[7]

Observe that the imperfect specification Γ thus defined is closely related to the concept of a *mass distribution* in the belief function setting [17]: Let $\mathsf{m} : 2^\Omega \to [0, 1]$ be a mass distribution over a set Ω, i.e., $\mathsf{m}(\emptyset) = 0$ and $\sum_{A \subset \Omega} \mathsf{m}(A) = 1$. Moreover, let $\mathcal{A} = \{A_1, \ldots, A_m\} = \{A \subset \Omega \mid \mathsf{m}(A) > 0\}$ denote the (finite) set of *focal elements*. We can then associate an imperfect specification $\Gamma = (\gamma, p_C)$ with m:

$$\begin{aligned} C &= \{c_1, \ldots, c_m\}, \\ \gamma(c_k) &= A_k, \\ p_C(c_k) &= \mathsf{m}(A_k) \end{aligned}$$

for all $1 \leq k \leq m$. The other way round, each imperfect specification $\Gamma = (\gamma, p_C)$ induces an (information-compressed[8]) representation in the form of a mass distribution m, where

$$\mathsf{m}(A) = \sum_{c \in C \,:\, \gamma(c) = A} p_C(c) \tag{5}$$

for all $A \subset \Omega$ and $\mathsf{m}(A) > 0$ for a finite number of focal elements.

By making use of the relation between the mass function (5) and the imperfect specification (γ, p_C) associated with a case $\langle s, \varphi(s) \rangle$,

[7] We disregard cases for which (4) is not well-defined.

[8] A mass function does not define a unique imperfect specification.

the evidence about the outcome $\varphi(s_0)$ derived from $\langle s, \varphi(s)\rangle$ can be represented in the form of a belief function Bel and an associated plausibility function Pl over $\mathcal{R}$, where

$$\mathsf{Bel}(A) = \sum_{B \subset A} \mathsf{m}(B), \qquad \mathsf{Pl}(A) = \sum_{B \cap A \neq \emptyset} \mathsf{m}(B)$$

for all $A \subset \mathcal{R}$. $\mathsf{Bel}(A)$ and $\mathsf{Pl}(A)$ define degrees of belief and plausibility that $\varphi(s_0)$ is an element of A, respectively. These values can also be interpreted as lower and upper probabilities. Since the imperfect specification and, hence, the mass distribution associated with $\langle s, \varphi(s)\rangle$ is derived from the outcome $\varphi(s)$ and the probability measure $H(\sigma_{\mathcal{S}}(s, s_0))$, the above belief function corresponds to a transformation $\sigma_{\mathcal{R}}^{(-1)}$ which is now a mapping $\mathcal{R} \times \mathcal{P}(D_{\mathcal{R}}) \to \mathcal{F}(\mathcal{R})$, where $\mathcal{F}(\mathcal{R})$ denotes, say, the class of normalized uncertainty measures over $\mathcal{R}$:

$$\mathsf{Bel} = \mathsf{Bel}(H, s_0) = \sigma_{\mathcal{R}}^{(-1)}\left(\varphi(s), H(\sigma_{\mathcal{S}}(s, s_0))\right)$$

This transformation defines a generalization of $\sigma_{\mathcal{R}}^{(-1)}$ in (3).

Let $\Gamma = (\gamma, p_C)$ be the imperfect specification induced by a case $\langle s, \varphi(s)\rangle$. The application of a *generalized insufficient reason principle* [19] makes it possible to characterize $\varphi(s_0)$ by means of a probability measure $\Pr$ over $\mathcal{R}$. The latter is defined by

$$\Pr(A) \doteq \sum_{c \in C\,:\,\gamma(c) \cap A \neq \emptyset} p_C(c) \cdot \frac{|A \cap \gamma(c)|}{|\gamma(c)|} \tag{6}$$

for all $A \subset \mathcal{R}$, where $|X|$ denotes the cardinality of the set X. This measure is also called *betting function*, a term referring to the use of (6) in the context of decision making [19].

5 COMBINATION OF EVIDENCE

After having discussed the transformation of probabilistic evidence, let us now turn to the problem of combining evidence from several cases. That is, suppose we are given n imperfect specifications of the unknown outcome $\varphi(s_0)$, which have been derived from a memory $\mathcal{M}$ of n cases $\langle s_1, \varphi(s_1)\rangle, \ldots, \langle s_n, \varphi(s_n)\rangle$ in connection with a probabilistic similarity hypothesis H. The task shall be to aggregate these pieces of evidence.[9]

Suppose the similarity between $\varphi(s_0)$ and $\varphi(s_k)$ to be given by y_k, i.e.,

$$\forall\, 1 \leq k \leq n \,:\, \sigma_{\mathcal{R}}(\varphi(s_0), \varphi(s_k)) = y_k. \tag{7}$$

We can then derive the prediction $\varphi(s_0) \in \widehat{\varphi}_{y,\mathcal{M}}(s_0)$, where

$$\widehat{\varphi}_{y,\mathcal{M}}(s_0) \doteq \bigcap_{1 \leq k \leq n} \sigma_{\mathcal{R}}^{(-1)}(\varphi(s_k), y_k) \tag{8}$$

and $y = (y_1, \ldots, y_n)$. This corresponds to a *conjunctive combination* of the individual predictions $\sigma_{\mathcal{R}}^{(-1)}(\varphi(s_k), y_k)$. Within our probabilistic setting, the vector y of similarity degrees is actually a random variable $Y = (Y_1, \ldots, Y_n)$, and the related prediction (8) can hence be seen as a random set $\widehat{\varphi}_{Y,\mathcal{M}}(s_0)$. This idea comes down to considering the n cases as one information source, inducing the imperfect specification $\Gamma = (\gamma, p_C)$, where

$$\begin{aligned} C &= (D_{\mathcal{R}})^n, \\ p_C &= \mu(c), \\ \gamma(c) &= \textstyle\bigcap_{1 \leq k \leq n} \sigma_{\mathcal{R}}^{(-1)}(\varphi(s_k), c_k) \end{aligned} \tag{9}$$

for all $c = (c_1, \ldots, c_n) \in C$. The measure μ corresponds to the joint probability over $(D_{\mathcal{R}})^n$ characterizing the occurence of similarity vectors y, i.e., $\mu(y)$ is the probability of the event (7).

Treating n cases as one information source in the sense of (9) is an obvious way of combining evidence. What makes things difficult, however, is the fact that the joint probability measure μ over $(D_{\mathcal{R}})^n$ and, hence, the probability p_C in (9) are generally not known. It is also not possible to derive μ from the information provided by a PSP, which informs about the (conditional) distributions of *individual* similarity degrees: A PSP specifies the (unknown) similarity y_k between $\varphi(s_0)$ and $\varphi(s_k)$ by means of a probability measure $Y_k \sim \mu_{Y|(X=\sigma_{\mathcal{S}}(s_0,s_k))}$, given the similarity of the respective situations. The random variables Y_k $(1 \leq k \leq n)$, however, are not stochastically independent. Needless to say, an extended probabilistic model providing the required information will generally be intractable due to the huge number of joint measures it would have to specify.

If knowledge about the dependency structure is incomplete, the most reasonable way of combining evidence is to define the aggregated imperfect specification as the *convex combination* of the individual imperfect specifications.[10] Let $\Gamma_k = (\gamma_k, p_{C_k})$ denote the imperfect specification associated with the case $\langle s_k, \varphi(s_k)\rangle$, where

$$\begin{aligned} C_k &= \{k\} \times D_{\mathcal{R}}, \\ \gamma_k(c) &= \sigma_{\mathcal{R}}^{(-1)}(\varphi(s_k), y_k), \\ p_{C_k}(c) &= \mu_{Y|(X=\sigma_{\mathcal{S}}(s_0,s_k))}(y_k) \end{aligned}$$

for all $c = (k, y_k) \in C_k$. Thus, an element $c = (k, y_k)$ specifies the context in which the kth case is considered, and the similarity between the corresponding outcome $\varphi(s_k)$ and the unknown outcome r_0 is given by y_k. The convex combination $\Gamma = (\gamma, p_C)$ of $\Gamma_1, \ldots, \Gamma_n$ is then defined by $C = C_1 \cup \ldots \cup C_n$ and

$$\begin{aligned} \gamma(c) &= \gamma_k(c), \\ p_C(c) &= \alpha_k \cdot p_{C_k}(c) \end{aligned} \tag{10}$$

for all $c \in C_k$, where $\alpha_k \geq 0$ $(1 \leq k \leq n)$ and $\alpha_1 + \ldots + \alpha_n = 1$.

Observe that the set of specification contexts in (10) is given by the *union* of the individual contexts, whereas it is defined as the *product* in (9). In fact, the convex combination (10) does not consider combined events (7) since the probabilities of these events are unknown. Rather, the incomplete specification Γ should be interpreted as follows: First, one of the n cases in the memory is chosen at random, where α_k is the probability of selecting the kth case. Then, the imperfect specification associated with the selected case is considered, and one of the contexts of this specification is chosen according to the corresponding probability measure.

An important question in connection with the combination of evidence concerns the determination of the weights α_k. An obvious possibility is to define them as normalized degrees of similarity, i.e., $\alpha_k = \sigma_{\mathcal{S}}(s_0, s_k) / \sum_{i=1}^{n} \sigma_{\mathcal{S}}(s_0, s_i)$, which is common practice in locally weighted approximation [4] and instance-based prediction of numeric values [13]. However, one might also think of more sophisticated approaches which take, say, the "typicality" or "reliability" of individual cases into account. The latter might be estimated, e.g., by comparing the predictions obtained from a case to actually observed outcomes in a sequence of inference problems.

Now, consider an SBI problem $\langle \Sigma, s_0\rangle$. Let $\mathsf{m}_k(H, s_0)$ and $\mathsf{Bel}_k(H, s_0)$ denote, respectively, the mass distribution and belief

[9] Of course, one might think of utilizing only a limited number of $k < n$ cases.

[10] Other types of aggregation such as, e.g., conjunctive or disjunctive pooling might be reasonable as well. This, however, presupposes special assumptions about the dependency structure [10].

function induced by the kth case $\langle s_k, \varphi(s_k)\rangle$ which corresponds to the kth specification Γ_k. That is,

$$\mathsf{Bel}_k(H, s_0) = \sigma_{\mathcal{R}}^{(-1)}\left(\varphi(s_k), H(\sigma_{\mathcal{S}}(s_0, s_k))\right),$$

with H being a hypothesis related to H_Σ. The corresponding function Bel associated with the convex combination (10) is then given by

$$\mathsf{Bel}(H, \mathcal{M}, s_0) = \sum_{k=1}^{n} \alpha_k \cdot \mathsf{Bel}_k(H, s_0). \tag{11}$$

In plain words, combining evidence at the instance level comes down to deriving the convex combination of the belief functions induced by individual cases where the weight of a case depends on characteristics such as, e.g., similarity, typicality, or precision. Observe that the global hypothesis H in (11) is replaced by the local hypotheses associated with the respective cases if SBI proceeds from a local $\mathcal{M}$-hypotheses $H^{\mathcal{M}}$:

$$\mathsf{Bel}(H^{\mathcal{M}}, \mathcal{M}, s_0) = \sum_{k=1}^{n} \alpha_k \cdot \mathsf{Bel}_k(H^{s_k}, s_0)$$

Given a setup Σ with memory $\mathcal{M}$, a prediction (11) can principally be derived for all situations in $\mathcal{S}$. This way, the similarity-based inference scheme can be generalized to a "belief function-valued" approximation of φ:[11]

$$\widehat{\varphi}_{H,\mathcal{M}} : \mathcal{S} \to \mathcal{F}(\mathcal{R}),\ s \mapsto \mathsf{Bel}(H, \mathcal{M}, s).$$

Of course, it is not necessary to derive a prediction (11) for those situations which have already been observed and are stored in $\mathcal{M}$, since the corresponding outcome can simply be retrieved from the memory. That is, $\widehat{\varphi}_{H,\mathcal{M}}$ should actually be defined as

$$\widehat{\varphi}_{H,\mathcal{M}}(s) = \begin{cases} \mathsf{Bel}_{\{\varphi(s)\}} & \text{if} \quad \langle s, \varphi(s)\rangle \in \mathcal{M} \\ \mathsf{Bel}(H, \mathcal{M}, s) & \text{if} \quad \langle s, \varphi(s)\rangle \notin \mathcal{M} \end{cases},$$

where $\mathsf{Bel}_{\{\varphi(s)\}}(A) = 1$ for $A \supset \varphi(s)$ and $\mathsf{Bel}_{\{\varphi(s)\}}(A) = 0$ otherwise.

6 CONCLUDING REMARKS

We have proposed a general framework of similarity-based inference which combines principles and concepts from instance-based reasoning and reasoning under uncertainty: Observed cases are evaluated against the background of the heuristic CBR hypothesis, which is formalized by means of a probabilistic model. This way, the available case-based evidence concerning, say, the solution to a new problem is quantified in the form of a belief function over the set of candidates. SBI thus supports the overall process of (case-based) problem solving by providing a (preliminary) estimation of the suitability of potential solutions.

A probabilistic approach to SBI seems appealing from several perspectives. Firstly, it provides an adequate formalization of the CBR hypothesis, since it emphasizes the heuristic nature of this assumption. In fact, characterizing the *belief* in the unknown solution by means of an uncertainty measure seems more appropriate than simply giving a "point-estimation." Secondly, a probabilistic model has a clear semantic interpretation (whether subjective or objective) which seems advantageous from the viewpoint of knowledge representation. It also facilitates modelling and knowledge acquisition tasks. Thirdly, the probabilistic approach makes the powerful methodological framework of probabilistic reasoning and statistical inference accessible to CBR.

It has already been mentioned that our approach is related to (classification or estimation) methods based on the NN principle such as, e.g., instance-based learning. It can be seen as a generalization of such methods in the sense that individual predictions are given in the form of belief functions instead of, say, precise class labels. These predictions are synthesized by means of a linear combination instead of, e.g., majority voting. However, there are also important methodological differences. Without going into detail, let us only remark that most instance-based methods make use of the CBR hypothesis by more indirect means, somehow taking its validity for granted. As opposed to this, SBI fits an explicit model of the CBR assumption, namely a probabilistic similarity hypothesis, to the current application, thereby combining instance-based and model-based learning. Needless to say, that taking the validity of the CBR hypothesis into account and pointing out the credibility of proposed solutions seems indispensible for certain applications of CBR such as, e.g., experience-based reasoning in medicine.

[11] This function corresponds somehow to what is called an *extensional concept description* in instance-based learning [3].

REFERENCES

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

[2] *Lazy Learning*, ed., D.W. Aha, Kluwer Academic Publ., 1997.

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

[4] C.G. Atkeson, A.W. Moore, and S. Schaal, 'Locally weighted learning', *Artificial Intelligence Review*, **11**, 11–73, (1997).

[5] *Aggregation and Fusion of Imperfect Information*, ed., B. Bouchon-Meunier, Physica-Verlag, Heidelberg, 1998.

[6] *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, ed., B.V. Dasarathy, IEEE Computer Society Press, 1991.

[7] A.P. Dempster, 'Upper and lower probability induced by a random closed interval', *Annals of Math. Stat.*, **39**, 219–246, (1968).

[8] 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**, 345–373, (1998).

[9] B. Faltings, 'Probabilistic indexing for case-based prediction', in *Proc.* ICCBR-*97*, pp. 611–622. (1997).

[10] J. Gebhardt and R. Kruse, 'Parallel combination of information sources', in *Handbook of Defeasible Reasoning and Uncertainty Management Systems, Vol. 3*, eds., D.M. Gabbay and Ph. Smets, 393–439, Kluwer Academic Publishers, (1998).

[11] C. Genest and J.V. Zidek, 'Combining probability distributions: A critique and an annotated bibliography', *Statistical Science*, **1**(1), 114–148, (1986).

[12] E. Hüllermeier, 'Toward a probabilistic formalization of case-based inference', in *Proc.* IJCAI-99, pp. 248–253, (1999).

[13] D. Kibler and D.W. Aha, 'Instance-based prediction of real-valued attributes', *Computational Intelligence*, **5**, 51–57, (1989).

[14] J.L. Kolodner, *Case-based Reasoning*, Morgan Kaufmann, 1993.

[15] E. Plaza, F. Esteva, P. Garcia, L. Godo, and R. Lopez de Mantaras, 'A logical approach to case-based reasoning using fuzzy similarity relations', *Journal of Information Sciences*, **106**, 105–122, (1998).

[16] C. Reiser and H. Kaindl, 'Case-based reasoning for multi-step problems and its integration with heuristic search', in *Proceedings* EWCBR-94, pp. 113–125, (1994).

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

[18] R. Short and K. Fukunaga, 'The optimal distance measure for nearest neighbor classification', IEEE *Trans. Inf. Theory*, **27**, 622–627, (1981).

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