To appear in IEEE Trans. on Knowledge and Data Engineering

Credible Case-Based Inference Using Similarity Profiles

Eyke Hüllermeier

Department of Mathematics and Computer Science

Philipps-Universität Marburg, Germany

eyke@informatik.uni-marburg.de

Abstract

In this paper, we propose a method for retrieving promising candidate solutions in case-based problem solving. Our method, referred to as *credible case-based inference*, makes use of so-called *similarity profiles* as a formal model of the key hypothesis underlying case-based reasoning (CBR), namely the assumption that similar problems have similar solutions. Proceeding from this formalization, it becomes possible to derive theoretical properties of the corresponding inference scheme in a rigorous way. In particular, it can be shown that, under mild technical conditions, a set of candidates covers the true solution with high probability. Thus, the approach supports an important subtask in case-based reasoning, namely to generate potential solutions for a new target problem, in a sound manner and, hence, contributes to the methodical foundations of CBR. Due to its generality, it can be employed for different types of performance tasks and can easily be integrated in existing CBR systems.

Keywords: case-based reasoning, prediction, instance-based learning.

ACM Taxonomy: heuristic methods [I.2.8.f], analogies [I.2.6.a], knowledge retrieval [I.2.13.h], decision support [I.2.1.c], probabilistic reasoning [I.2.3.l], machine learning [I.2.6.g].

1 Introduction

Longstanding research in artificial intelligence, knowledge engineering, and related fields has produced a number of paradigms for building intelligent and knowledge-based systems such as, e.g., rule-based reasoning, constraint processing, or probabilistic graphical models. Being one of these paradigms, case-based reasoning (CBR) has received a great deal of attention in recent years and has been used successfully in diverse application areas [7], ranging from web search [6] to legal reasoning [9]. CBR is inspired by human problem solving and has roots in cognitive psychology [29]. Its key idea is to tackle new problems by referring to similar problems that have already been solved in the past [23].

To illustrate this idea, consider the problem to give a talk on a certain topic. Instead of preparing new slides from scratch every time, one typically starts with having a look at the talks that one has already given in the past. Then, one takes the relevant slides from some of these talks, somehow combines and revises them, and finally comes up with a new set of slides. The key question in CBR concerns the automatization of this kind of problem solving.

A widely accepted framework for case-based reasoning is characterized by the so-called "CBR cycle". The latter reflects the main components necessary for realizing CBR, namely the retrieval and the intelligent use of stored cases, the update of experiences given in the form of cases, and the case base maintenance. The (informal) R^4 model of the CBR cycle describes the main steps of a single problem solving episode and consists of the following four phases [1]: Retrieve the case(s) from the case base which is (are) most similar to the target problem; Reuse the information provided by this (these) case(s) in order to generate a candidate solution for the new problem; Revise the proposed solution according to the special requirements of the new problem; Retain the new experience obtained in the current problem solving episode for future problem solving.

Especially difficult to automatize is the revision step. For example, reconsider the introductory example above: How to build a computer program that automatically prepares a new talk from old slides? This seems to be hardly possible, mainly because in this type of problem the adaptation steps are not well-defined and require a degree of understanding and creativity that goes far beyond the capabilities of current computers. If, one the other hand, a set of well-defined (formal) adaptation operators is available, these operators can in principle be used as search operators in a search space the states of which correspond to cases. As suggested in [8], CBR can then be cast as a search problem amenable to a computerized solution. We shall come back to this point at the end of the paper.

Even simpler than "systematic" or "combinatorial" adaptation of that kind is "null-adaptation", i.e., problems for which no adaptation is necessary at all. In particular, this includes *prediction problems*, such as classification and regression, that will be considered in more detail in section 4.

This paper contributes to the methodical foundations of CBR by developing a method for *case-based inference* (CBI) with some interesting theoretical properties. Here, case-based inference is considered as a part of CBR closely related to the retrieval step. More specifically, CBI tries to answer the following question: Given the current problem solving experience in the form of a case base (and background knowledge in the form of the underlying similar problems–similar solutions assumption), what solutions are likely to solve a new target problem?

The method that we shall propose will answer this questions in terms of a "credible solution set", that is, a set of candidate solutions that covers the correct solution with high probability. This approach, that we shall refer to as *credible case-based inference* (CCBI), will be introduced in its basic form in section 2. Practically motivated extensions of CCBI will then be presented in section 3. Section 4 is devoted to the application of CCBI to prediction problems.[1] Related work is briefly discussed in section 5. The paper concludes with a summary and some suggestions for future work in section 6.

[1] Applying the inference principle underlying our approach to instance-based regression has first been proposed in a paper presented at the 16th European Conference on Artificial Intelligence (ECAI), Valencia, Spain, 2004 [21].

2 Credible Case-Based Inference

In this section, we introduce the method of *credible case-based inference* (CCBI) for deriving credible solution sets in CBR. As a major tool, this method makes use of what we call a *similarity profile*, a function that establishes a connection between the similarity of problems and the similarity of solutions. This concept, as well as as the related concept of a *similarity hypothesis*, will first be discussed in sections 2.2–2.4. CCBI itself will then be introduced in section 2.5.

2.1 Case-Based Inference

Let $\mathcal{X}$ denote a problem space, that is, a set of potential problems in the application under consideration. More specifically, each $x \in \mathcal{X}$ is a formal representation of a problem. A problem space of such kind is a completely general concept; it includes special cases such as, e.g., *feature spaces* in supervised learning, where problems are instances characterized by a fixed number of attribute values, but also problems described by more complex structures like trees or graphs. Likewise, let $\mathcal{L}$ be a solution space, i.e., a set of potential solutions. For the sake of simplicity, we assume that each problem $x \in \mathcal{X}$ is associated with a unique (optimal) solution $\lambda_x \in \mathcal{L}$. We remark, however, that the approach presented in the remainder of the paper can be extended to the more general case where a problem can be solved by more than one solution and, hence, is associated with a subset of $\mathcal{L}$.

We assume the problem space $\mathcal{X}$ to be endowed with a similarity measure $\text{sim}_{\mathcal{X}}(\cdot)$; for each pair of problems $x, y \in \mathcal{X}$, $\text{sim}_{\mathcal{X}}(x, y)$ is a quantification of the similarity between x and y. Likewise, we assume a similarity measure $\text{sim}_{\mathcal{L}}(\cdot)$ to be defined on the solution space $\mathcal{L}$. For the sake of convenience, we assume that both measures are normalized to the range $[0, 1]$, where 1 means complete similarity and 0 complete dissimilarity. Moreover, we assume that the measures are reflexive and symmetric, i.e., $\text{sim}_{\mathcal{X}}(x, x) = 1$ and $\text{sim}_{\mathcal{X}}(x, y) = \text{sim}_{\mathcal{X}}(y, x)$ for all $x, y \in \mathcal{X}$. We like to emphasize, however, that we do not assume any kind of transitivity. In particular, we do not assume that $\mathcal{X}$ or $\mathcal{L}$ are metric spaces. This makes our approach widely applicable and more flexible than standard statistical inference methods, a point we shall return to later on.[2]

Finally, we assume a case base (memory) $\mathcal{M}$ to be given, that is, a collection of n cases of the form $\langle x_\imath, \lambda_{x_\imath} \rangle \in \mathcal{X} \times \mathcal{L}$, $1 \leq \imath \leq n$. This case base is a summary of the problem solving experience gathered so far. The task of case-based inference shall be to exploit that experience in order to predict the solution of a new target problem $x_0 \in \mathcal{X}$.

2.2 Similarity Profiles

CBR strongly relies of the assumption that similar problems have similar solutions. Even though this "CBR hypothesis" is in harmony with daily experience, it is a relatively vague heuristic, which in our opinion is one reason for the ad-hoc character of many CBR methods. Our point of departure is therefore a concretization of the CBR hypothesis in terms of a formal model. This will provide the basis of a sound inference procedure including assertions about the confidence of predictions.

To begin, suppose that the CBR hypothesis has the following concrete meaning:

$$\forall\, x, y \in \mathcal{X} \; : \; \mathrm{sim}_{\mathcal{X}}(x, y) \; \leq \; \mathrm{sim}_{\mathcal{L}}\left(\lambda_x, \lambda_y\right) \tag{1}$$

Roughly speaking, (1) is a similarity constraint demanding that solutions are always at least as similar as problems. On the basis of this constraint, one can reason as follows: Consider a case $\langle x_1, \lambda_{x_1} \rangle$ from the case base $\mathcal{M}$ and let x_1 be α_1-similar to the target problem x_0, i.e. $\mathrm{sim}_{\mathcal{X}}(x_1, x_0) = \alpha_1$. According to (1), the unknown solution λ_{x_0} must then be an element of the α_1-neighborhood of λ_{x_1}, i.e., of the set

$$\mathcal{N}_{\alpha_1}(\lambda_{x_1}) \overset{\mathrm{df}}{=} \{\lambda \in \mathcal{L} \,|\, \mathrm{sim}_{\mathcal{L}}(\lambda, \lambda_{x_1}) \geq \alpha_1\}.$$

Likewise, if we have another case $\langle x_2, \lambda_{x_2} \rangle$ such that $\mathrm{sim}_{\mathcal{X}}(x_2, x_0) = \alpha_2$, the solution λ_{x_0} is also an element of the α_2-neighborhood of λ_{x_2} and, hence, of the intersection $\mathcal{N}_{\alpha_1}(\lambda_{x_1}) \cap \mathcal{N}_{\alpha_2}(\lambda_{x_2})$; see Fig. 1 for an illustration. Repeating the same argument for all

[2] In principle, we could even give up the symmetry assumption, though we retain it for ease of exposition. We note, however, that non-symmetric measures might indeed be of interest in CBR [30].

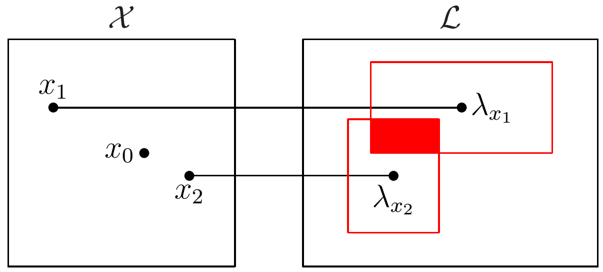

Figure 1: The known solutions of two problems restrict the solution of the target problem, which must be an element of the shaded region according to constraint (1).

cases in the case base, one finally derives the following restriction for the solution λ_{x_0}:

$$\lambda_{x_0} \in \bigcap_{\imath=1}^{n} \mathcal{N}_{\mathrm{sim}_{\mathcal{X}}(x_\imath, x_0)}(\lambda_{x_\imath}) \tag{2}$$

Needless to say, for a concrete application the similarity constraint (1) will usually not be satisfied, which in turn invalidates the above line of reasoning. Let us therefore consider a relaxation of this constraint:

$$\forall\, x, y \in \mathcal{X} \, : \, \zeta\left(\, \mathrm{sim}_{\mathcal{X}}(x,y)\,\right) \, \leq \mathrm{sim}_{\mathcal{L}}\left(\lambda_x, \lambda_y\right), \tag{3}$$

where $\zeta(\cdot)$ is an appropriate function $[0,1] \rightarrow [0,1]$. This function assigns to each similarity degree between two problems, α, the largest similarity degree $\beta = \zeta(\alpha)$ such that the following property holds: The solutions of two α-similar problems are guaranteed to be at least β-similar. We call $\zeta(\cdot)$ a *similarity profile*. More formally, a similarity profile is defined as follows: For all $\alpha \in [0,1]$,

$$\zeta(\alpha) \overset{\mathrm{df}}{=} \inf_{x,y \in \mathcal{X}, \mathrm{sim}_{\mathcal{X}}(x,y) = \alpha} \mathrm{sim}_{\mathcal{L}}\left(\lambda_x, \lambda_y\right). \tag{4}$$

We note that, by using $\zeta(\cdot)$ as defined in (4), the constraint (3) is satisfied *by definition*. In the worst case, $\zeta(\alpha) = 0$ for all α, which means that the similarity between two problems does not allow one to draw any conclusions about the similarity between the corresponding solutions. Fortunately, this situation rarely occurs in practice, where the CBR hypothesis is mostly satisifed at least to some extent. In fact, the similarity profile conveys a precise idea of the degree to which an application actually meets the CBR hypothesis. Roughly speaking, the "larger" $\zeta(\cdot)$ is, the more this hypothesis holds true.[3]

Generalizing our above line of reasoning, the constraint (3) suggests the following counterpart to (2) for predicting the label λ_{x_0}:

$$\lambda_{x_0} \in C(x_0) \overset{\mathrm{df}}{=} \bigcap_{\imath=1}^{n} \mathcal{N}_{\zeta(\mathrm{sim}_{\mathcal{X}}(x_\imath, x_0))}(\lambda_{x_\imath}) \tag{5}$$

This inference scheme is obviously correct in the sense that $C(x_0)$ is guaranteed to cover λ_{x_0}, a property that follows immediately from the definition of the similarity profile $\zeta(\cdot)$. We call $C(x_0)$ a *credible solution set* and refer to the inference scheme itself as CCBI (Credible Case-Based Inference). As an interesting property of the prediction (5), note that it can also suggest solutions that have never been observed so far. Roughly speaking, while traditional CBR methods are typically restricted to retrieve solutions stored in the case base, our approach is also able to "interpolate" between the observed solutions.

2.3 Similarity Hypotheses

The application of the inference scheme (5) requires the similarity profile $\zeta(\cdot)$ to be known, a requirement that will usually not be fulfilled. This motivates the related concept of a *similarity hypothesis*, a function $h : [0,1] \rightarrow [0,1]$, which is thought of as an approximation of a similarity profile. We say that $h(\cdot)$ is *admissible* if $h(\cdot) \leq \zeta(\cdot)$, i.e., $h(\alpha) \leq \zeta(\alpha)$ for all $\alpha \in [0,1]$. A hypothesis $h(\cdot)$ is called *stronger* than a hypothesis $h'(\cdot)$ if $h'(\cdot) \leq h(\cdot)$ and $h(\cdot) \not\leq h'(\cdot)$.

It is obvious that using an admissible hypothesis $h(\cdot)$ in place of the true similarity profile $\zeta(\cdot)$ within the inference scheme (5) leads to predictions

$$C^{est}(x_0) \overset{\mathrm{df}}{=} \bigcap_{\imath=1}^{n} \mathcal{N}_{h(\mathrm{sim}_{\mathcal{X}}(x_\imath, x_0))}(\lambda_{x_\imath}) \tag{6}$$

[3] Of course, $[0,1] \rightarrow [0,1]$ mappings are only partially ordered: $\zeta(\cdot) \leq \zeta'(\cdot)$ if $\zeta(x) \leq \zeta'(x)$ for all $x \in [0,1]$.

that are correct in the sense that $\lambda_{x_0} \in C^{est}(x_0)$. Indeed, $h(\cdot) \leq \zeta(\cdot)$ implies

$$\mathcal{N}_{\zeta(\text{sim}_{\mathcal{X}}(x_\imath,x_0))}(\lambda_{x_\imath}) \;\subseteq\; \mathcal{N}_{h(\text{sim}_{\mathcal{X}}(x_\imath,x_0))}(\lambda_{x_\imath})$$

for all cases $\langle x_\imath, \lambda_{x_\imath}\rangle$ and, hence, $C(x_0) \subseteq C^{est}(x_0)$.

Yet, assuming the profile $\zeta(\cdot)$ to be unknown, one cannot guarantee the admissibility of a hypothesis $h(\cdot)$ and, hence, the correctness of (6). In other words, it might happen that $\lambda_{x_0} \notin C^{est}(x_0)$. In fact, we might even have $C^{est}(x_0) = \emptyset$ (in which case the prediction is definitely incorrect). Nevertheless, taking for granted that $h(\cdot)$ is indeed a good approximation of $\zeta(\cdot)$, it seems reasonable to derive $C^{est}(x_0)$ according to (6) as an approximation of $C(x_0)$, that is, to realize case-based inference as a kind of approximate reasoning. In fact, our results below will show that, by using suitable hypotheses, the probability of incorrect predictions can be bounded and becomes (arbitrarily) small for large case bases.

2.4 Learning Similarity Hypotheses

Our discussion so far has left open the question of how to specify a similarity hypothesis in an appropriate way. An obvious idea in this connection is to induce such a hypothesis from the observed cases. Before going into detail, note that the overall approach thus obtained can be seen as a combination of case-based and model-based learning. In fact, adapting the similarity hypothesis is a kind of model-based learning, since a similarity hypothesis is a model of the CBR hypothesis.

Given a hypothesis space $\mathcal{H}$, i.e., a class of functions $h : [0,1] \rightarrow [0,1]$, learning amounts to choosing one among these hypotheses on the basis of the data. But which of the hypotheses are interesting candidates? Of course, first of all a hypothesis $h(\cdot)$ should be consistent with the data given, that is, the similarity constraint should be satisfied for all cases in $\mathcal{M}$:

$$\forall\, \langle x, \lambda_x\rangle, \langle y, \lambda_y\rangle \in \mathcal{M} \;:\; \text{sim}_{\mathcal{X}}(x,y) = \alpha \;\Rightarrow\; \text{sim}_{\mathcal{L}}(\lambda_x, \lambda_y) \geq h(\alpha). \qquad (7)$$

Denote by $\mathcal{H}_C \subseteq \mathcal{H}$ the set of hypotheses that are consistent in this sense. Among two consistent hypotheses $h(\cdot)$ and $h'(\cdot)$, where $h(\cdot)$ is stronger than $h'(\cdot)$, we should prefer the former since it leads to more precise predictions. Thus, we call a hypothesis $h_*(\cdot)$ optimal if $h_* \in \mathcal{H}_C$ and if there is no hypothesis $h \in \mathcal{H}_C$ such that $h(\cdot)$ is stronger than $h_*(\cdot)$. The following observation is very simple to prove:

Observation 1 *Suppose the hypothesis space $\mathcal{H}$ to satisfy $h \equiv 0 \in \mathcal{H}$ and $(h, h' \in \mathcal{H}) \Rightarrow (h \vee h' \in \mathcal{H})$, where $h \vee h'$ is the pointwise maximum $x \mapsto \max\{h(x), h'(x)\}$. Then, a unique optimal hypothesis $h_* \in \mathcal{H}$ exists, and $\mathcal{H}_C = \{h \in \mathcal{H} \,|\, h \leq h_*\}$.*

Given the assumptions of this observation, learning a similarity hypothesis can be realized as a *candidate-elimination* algorithm [26], where $h_*(\cdot)$ is a compact representation of the *version space*, i.e., the subset $\mathcal{H}_C$ of hypotheses from $\mathcal{H}$ which are consistent with the training examples.

A very simple representation of hypotheses, that will nevertheless turn out to be very useful, is a step function

$$h : x \mapsto \sum_{k=1}^{m} \beta_k \cdot \mathbb{I}_{A_k}(x), \tag{8}$$

where $A_k = [\alpha_{k-1}, \alpha_k)$ for $1 \leq k \leq m-1$, $A_m = [\alpha_{m-1}, \alpha_m]$, and $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_m = 1$ defines a partition of $[0, 1]$. ($\mathbb{I}_A(\cdot)$ denotes the indicator function of set A.) For a fixed underlying partition, we denote the space of all step functions by $\mathcal{H}_{step}$.

The strongest hypothesis $h_* \in \mathcal{H}_{step}$ consistent with the cases in the case base $\mathcal{M}$ is characterized by the coefficients

$$\beta_k \overset{\mathrm{df}}{=} \min_{x_\imath, x_\jmath : \mathrm{sim}_{\mathcal{X}}(x_\imath, x_\jmath) \in A_k} \mathrm{sim}_{\mathcal{L}}(\lambda_{x_\imath}, \lambda_{x_\jmath}) \tag{9}$$

for $1 \leq k \leq m$, where $\min \emptyset = 1$ by definition. We call this hypothesis the *empirical similarity profile* and, since it is directly derived from the case base $\mathcal{M}$, denote it by $h_{\mathcal{M}}(\cdot)$; see Fig. 2 for an illustration.

Now, suppose that the case base $\mathcal{M}$, in which n cases are stored, is to be extended by

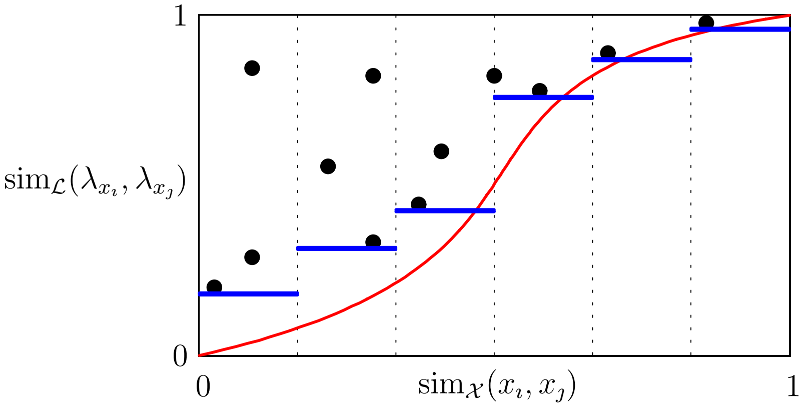

Figure 2: Each pair of cases $\langle x_\imath, \lambda_{x_\imath} \rangle$ and $\langle x_\jmath, \lambda_{x_\jmath} \rangle$ contributes a point (α, β) in the "similarity space", where $\alpha = \mathrm{sim}_{\mathcal{X}}(x_\imath, x_\jmath)$ and $\beta = \mathrm{sim}_{\mathcal{L}}(\lambda_{x_\imath}, \lambda_{x_\jmath})$. By definition, these points are located above the similarity profile, which is here shown by the solid (red) line. The empirical similarity profile is given by the step function indicated by the solid (blue) horizontal lines.

a newly observed case $\langle x_{n+1}, \lambda_{x_{n+1}} \rangle$. Updating the empirical similarity profile $h_{\mathcal{M}}(\cdot)$ can then be accomplished by passing the iteration

$$\beta_{\kappa(x_{n+1}, x_\jmath)} \ \leftarrow \ \min \left\{ \beta_{\kappa(x_{n+1}, x_\jmath)}, \mathrm{sim}_{\mathcal{L}}(\lambda_{x_{n+1}}, \lambda_{x_\jmath}) \right\} \qquad (10)$$

for $1 \leq \jmath \leq n = |\mathcal{M}|$. The index $1 \leq \kappa(x, y) \leq m$ is defined for problems $x, y \in \mathcal{X}$ by $\kappa(x, y) = k \Leftrightarrow \mathrm{sim}_{\mathcal{X}}(x, y) \in A_k$. As can be seen, the time complexity of updating the empirical profile is linear in the size of the case base.

2.5 Credible Case-Based Inference

The updating scheme (10) suggests a CBR process in which prediction and learning are repeated alternately in the style of incremental supervised learning, as shown in Algorithm 1: At each point of time, we dispose of a case base $\mathcal{M}$ with an associated empirical similarity profile $h_{\mathcal{M}}(\cdot)$. Having to predict the solution of a new problem x_0, an estimation $C^{est}(x_0)$ is derived from $\mathcal{M}$ and $h_{\mathcal{M}}(\cdot)$ according to (6). The system exploits $C^{est}(x_0)$ in order to support the (external) problem solving process (procedure `solve` in Algorithm 1). If, in the course of problem solving, the correct solution λ_{x_0} has become available, $\langle x_0, \lambda_{x_0} \rangle$ is added as the $(n+1)$-th case $\langle x_{n+1}, \lambda_{x_{n+1}} \rangle$ to the case base and the empirical profile $h_{\mathcal{M}}(\cdot)$ is updated according to (10).

Algorithm 1 CCBI

```
Input: a sequence of query inputs
Output: a sequence of credible estimation for outputs
initialize the original case base $\mathcal{M}$ (perhaps empty)
repeat
    take next query $x_0$
    derive estimation $C^{est}(x_0)$ from $\mathcal{M}$ and $h_{\mathcal{M}}(\cdot)$ according to (6)
    $\lambda_{x_0} \leftarrow \texttt{solve}(x_0, C^{est}(x_0))$
    $\mathcal{M} \leftarrow \mathcal{M} \cup \{\langle x_0, \lambda_{x_0} \rangle\}$
    update empirical profile $h_{\mathcal{M}}(\cdot)$ according to (10)
until no more queries exist
```

Regarding the complexity of CCBI, the estimation step (6) requires computing the similarity between the query x_0 and the cases stored in the memory. Likewise, updating the empirical profile $h_{\mathcal{M}}(\cdot)$ according to (10) again requires scanning the case base. Deriving the prediction itself comes down to intersecting the neighborhood of cases and, hence, strongly depends on the application (structure of $\mathcal{L}$, representation of neighborhoods). Thus, excluding the prediction itself, the overall complexity of a CCBI step is linear in the size of $\mathcal{M}$. As will be seen in the remainder of the paper, several possibilities exist to further increase the efficiency of the basic CCBI scheme.

Needless to say, the strategy of simply adding all observations to the current case base $\mathcal{M}$ will usually not be efficient. In fact, much more sophisticated strategies for maintaining a case base are often used in practice [33], including the possibility of removing or replacing stored cases [31, 27]. Still, the strategy above is sufficient for our purpose here. Besides, it simplifies a theoretical analysis of the prediction performance, as will be seen below.

For obvious reasons, we call the step function $h^*(\cdot)$ defined by the coefficients

$$\beta_k^* \overset{\mathrm{df}}{=} \inf \{ \zeta(\alpha) \,|\, \alpha \in A_k \}, \tag{11}$$

$1 \leq k \leq m$, the *optimal admissible* hypothesis. Since admissibility implies consistency, we have $h^*(\cdot) \leq h_{\mathcal{M}}(\cdot)$. This inequality suggests that the empirical similarity profile $h_{\mathcal{M}}(\cdot)$ will usually overestimate the true profile $\zeta(\cdot)$ and, hence, that $h_{\mathcal{M}}(\cdot)$ might not be admissible (cf. Fig. 2). Of course, the fact that admissibility of $h_{\mathcal{M}}(\cdot)$ is not guaranteed seems to conflict with the objective of providing correct predictions and, hence, gives rise to questions concerning the actual quality of the empirical profile as well as the quality of predictions derived from that hypothesis.

We make the simplifying assumption that the problem space $\mathcal{X}$ is countable. Further, we make the standard assumption that the query problems x_0 (resp. the new cases $\langle x_0, \lambda_{x_0} \rangle$) are chosen at random according to a fixed (not necessarily known) probability distribution $\mu(\cdot)$. Statistically speaking, the observed cases are independent and identically distributed (*iid*) random variables. Note that we can assume $\mu(\{x\}) > 0$ for all $x \in \mathcal{X}$ without loss of generality.

Now, denote by $\mathcal{M}_n$ the case base in the n-th step of the above CBR process, that is, the case base $\mathcal{M}$ such that $|\mathcal{M}| = n$, and by $h_n(\cdot) = h_{\mathcal{M}_n}(\cdot)$ the empirical similarity profile derived from that case base. Since, according to our assumption, the observed cases are random variables, the induced hypotheses $h_n(\cdot)$ are random variables (random functions) as well. As a first important property of the above learning process we can prove that the sequence of hypotheses $h_1, h_2, \ldots$ converges stochastically toward the optimal admissible hypothesis $h^*(\cdot)$.

Theorem 2 *For the sequence $(h_n)_{n \geq 1}$ of empirical similarity profiles it holds true that $h_n \searrow h^*$ stochastically as $n \to \infty$. That is, $h_n \geq h^*$ for all $n \in \mathbb{N}$ and $\mathsf{Pr}(|h_n - h^*|_\infty \geq \varepsilon) \to 0$ as $n \to \infty$ for all $\varepsilon > 0$, where $|h_n - h^*|_\infty \overset{\text{df}}{=} \sup_{0 \leq x \leq 1} |h_n(x) - h^*(x)|$.* □

Proof: Given the definition of $h^*(\cdot)$ and the updating procedure for empirical profiles, it is obvious that $h^*(\cdot) \leq h_n(\cdot)$ for all $n \geq 1$ and, moreover, that the sequence of functions $(h_n)_{n \geq 1}$ is decreasing. Let $\epsilon > 0$ and consider some $1 \leq k \leq m$. According to (11), there is some $\alpha \in A_k$ such that $|\zeta(\alpha) - \beta_k^*| < \epsilon/2$. Then, due to the definition of $\zeta(\alpha)$, there are also $x_{k_1}, x_{k_2} \in \mathcal{X}$ such that $\text{sim}_{\mathcal{X}}(x_{k_1}, x_{k_2}) = \alpha$ and $|\,\text{sim}_{\mathcal{L}}(\lambda_{x_{k_1}}, \lambda_{x_{k_2}}) - \zeta(\alpha)| < \epsilon/2$. Therefore, $|\,\text{sim}_{\mathcal{L}}(\lambda_{x_{k_1}}, \lambda_{x_{k_2}}) - \beta_k^*| < \epsilon$. This implies $|h_n(\alpha) - \beta_k^*| < \epsilon$ as soon as the case base $\mathcal{M}_n$ contains the problems x_{k_1} and x_{k_2}. Since this line of reasoning applies to all $1 \leq k \leq m$, we obtain

$$\|h_n - h^*\|_\infty = \max_{0 \leq \alpha \leq 1} |h_n(\alpha) - h^*(\alpha)| < \epsilon$$

if the case base $\mathcal{M}_n$ contains the (at most $2\,m$) problems x_{k_1}, x_{k_2} ($1 \leq k \leq m$). Since $\mu_{\mathcal{X}}(\{x_{k_1}\}) > 0$ and $\mu_{\mathcal{X}}(\{x_{k_2}\}) > 0$ for all $1 \leq k \leq m$, the probability for this tends toward 1 as $n \rightarrow \infty$. □

Regarding the quality of estimations, we are first of all interested in the probability of incorrect predictions. In this connection, it should be noted that a prediction $C^{est}(x_0)$ might well be correct even if the involved empirical profile $h_{\mathcal{M}}(\cdot)$ is not admissible: Recall that the estimation (6) is derived from a *limited* number of similarity constraints, namely the $\beta_\imath$-neighborhoods associated with known solutions $\lambda_{x_\imath}$. As we cannot exclude that $\beta_\imath = h_{\mathcal{M}}(\mathrm{sim}_{\mathcal{X}}(x_\imath, x_0)) > \zeta(\mathrm{sim}_{\mathcal{X}}(x_\imath, x_0))$, it is true that each of these neighborhoods might be "too small" and, hence, might remove some solutions from the credible solution set $C(x_0)$. Still, this unjustified removal does not necessarily concern the correct label λ_{x_0}. An indeed, we can show the following result:

Theorem 3 *Suppose that observed problems are independent and identically distributed (iid) random variables, generated according to a fixed (not necessarily known) probability distribution $\mu(\cdot)$ over $\mathcal{X}$. Let $C^{est}(x_0)$ be the prediction of the label λ_{x_0} derived from the empirical similarity profile $h_{\mathcal{M}}(\cdot)$. The following estimation holds true:*

$$\mathsf{Pr}\left(\lambda_{x_0} \not\in C^{est}(x_0)\right) \;\leq\; \frac{2m}{1+|\mathcal{M}|}, \tag{12}$$

where m is the size of the partition underlying the step function $h_{\mathcal{M}}(\cdot)$. □

Proof: Consider a case base $\mathcal{M}$ with related (empirical) profile $h_{\mathcal{M}}(\cdot)$. We call a problem $x_0 \in \mathcal{X}$ *extremal* (with respect to $\mathcal{M}$) if there is some $1 \leq k \leq m$ and a case $\langle x, \lambda_x \rangle \in \mathcal{M}$ such that $\mathrm{sim}_{\mathcal{X}}(x, x_0) \in A_k$ and $\mathrm{sim}_{\mathcal{L}}(\lambda_x, \lambda_{x_0}) < \mathrm{sim}_{\mathcal{L}}(\lambda_{x_1}, \lambda_{x_2})$ for all $\langle x_1, \lambda_{x_1}\rangle, \langle x_2, \lambda_{x_2}\rangle \in \mathcal{M}$ such that $\mathrm{sim}_{\mathcal{X}}(x_1, x_2) \in A_k$.[4] We first show that, if $\lambda_{x_0} \not\in C^{est}(x_0)$, then x_0 is extremal with respect to the underlying case base $\mathcal{M}$. In fact, if $\lambda_{x_0} \not\in C^{est}(x_0)$, there must be a case $\langle x, \lambda_x\rangle \in \mathcal{M}$ such that $\lambda_{x_0} \not\in \mathcal{N}_{h_{\mathcal{M}}(\mathrm{sim}_{\mathcal{X}}(x,x_0))}(\lambda_x)$. This means that $\mathrm{sim}_{\mathcal{L}}(\lambda_x, \lambda_{x_0}) < h_{\mathcal{M}}(\mathrm{sim}_{\mathcal{X}}(x, x_0))$ and, therefore, $\mathrm{sim}_{\mathcal{L}}(\lambda_x, \lambda_{x_0}) < \mathrm{sim}_{\mathcal{L}}(\lambda_{x_1}, \lambda_{x_2})$ for all $\langle x_1, \lambda_{x_1}\rangle, \langle x_2, \lambda_{x_2}\rangle \in$

[4] This definition of being extremal is to some extent related to the concept of "strangeness" of an observation in the context of so-called confidence machines [18, 28].

$\mathcal{M}$ such that $\mathrm{sim}_{\mathcal{X}}(x_1, x_2) \in A_{\kappa(x,x_0)}$. Thus, x_0 is extremal. This result shows that the probability (12) is upper-bounded by the probability that x_0 is extremal with respect to $\mathcal{M}$. Let $\mathcal{M}_+ = \mathcal{M} \cup \{\langle x_0, \lambda_{x_0}\rangle\}$ and consider $h_{\mathcal{M}_+}(\cdot)$. Obviously, there is a sub-memory $\mathcal{M}_- \subseteq \mathcal{M}_+$ consisting of (at most) $2m$ cases and such that $h_{\mathcal{M}_+}(\cdot) = h_{\mathcal{M}_-}(\cdot)$. Moreover, $\langle x_0, \lambda_{x_0}\rangle \not\in \mathcal{M}_-$ implies that x_0 is not extremal. Therefore, recalling that problems are chosen independently according to $\mu_{\mathcal{X}}(\cdot)$ and noting that $|\mathcal{M}_+| = 1 + \mathcal{M}$, the theorem simply holds due to reasons of symmetry. □

Corollary 4 *The expected proportion of incorrect predictions in connection with the above CBR process converges toward 0.* □

According to the above results, the probability of an incorrect prediction becomes small for large memories, even if the related hypotheses are not admissible. In fact, $\mathsf{Pr}(\lambda_{x_0} \not\in C^{est}(x_0)) \rightarrow 0$ as $|\mathcal{M}| \rightarrow \infty$. In a statistical sense, the prediction $C^{est}(x_0)$ can indeed be seen as *credible solution set*, a justification for using this term not only for $C(x_0)$ but also for $C^{est}(x_0)$. Note that the level of confidence guaranteed by $C^{est}(x_0)$ depends on the number of observed cases and can hence be controlled.

The upper bound established in Theorem 3 might suggest decreasing the probability of an incorrect prediction by reducing the size m of the partition underlying $\mathcal{H}_{step}$. Observe, however, that this will also lead to a less precise approximation of $\zeta(\cdot)$ and, hence, to less precise predictions of solutions. When "merging" two neighbored intervals A_k and A_{k+1}, for instance, the corresponding coefficients β_k and β_{k+1} would be replaced by $\min\{\beta_k, \beta_{k+1}\}$.

It is furthermore interesting to note that the level of confidence does *not* depend on the similarity measures $\mathrm{sim}_{\mathcal{X}}(\cdot)$ and $\mathrm{sim}_{\mathcal{L}}(\cdot)$, i.e., credible predictions can be made for *any* pair of similarity measures. In other words, the suitability of the similarity measures does not influence the confidence of the predicted solution sets $C^{est}(x_0)$. It does influence, however, the *precision* of these predictions: The more suitable the similarity measures are, the "larger" the similarity profile and, hence, the more precise the predictions become. Thus, methods for learning or adapting similarity measures, a topic of general interest in CBR (e.g. [17, 34]), could of course complement CCBI in a reasonable way. Even though this idea will not be developed further in this paper, note that the similarity profile provides an interesting point of departure in this regard: As mentioned above, the precision of predictions is to a large extent dictated by the strength of the similarity profile. Consequently, the latter can be taken as an indicator of the quality of the underlying similarity measures (and, hence, as an optimization criterion).

2.6 Practical Issues

A rather obvious idea in connection with the inference scheme (5) is to take the intersection not over all n cases in the case base $\mathcal{M}$ but only over the $k \ll n$ nearest neighbors of the query x_0. Obviously, this will increase efficiency while preserving the correctness of the prediction. On the other hand, some precision will be lost, but this effect is usually limited due to the fact that less similar problems often hardly contribute to the precision of predictions.

In many applications one is interested in both, a credible solution set and a "point-estimation" of the solution λ_{x_0}, i.e., a distinguished element $\lambda_{x_0}^{est} \in \mathcal{L}$ that can be considered as representative. The latter can be derived from the credible solution set $C^{est}(x_0)$ as a *generalized median*:

$$\lambda_{x_0}^{est} \overset{\mathrm{df}}{=} \arg \max_{\lambda \in C^{est}(x_0)} \sum_{\lambda' \in C^{est}(x_0)} \mathrm{sim}_{\mathcal{L}}(\lambda, \lambda')$$

As can be seen, the generalized median is a kind of center-point, namely the element of the credible solution set which is maximally similar to all other elements.

Let us finally make a note on the specification of the similarity measures $\mathrm{sim}_{\mathcal{X}}(\cdot)$ and $\mathrm{sim}_{\mathcal{L}}(\cdot)$. Usually, the definition of the latter is uncritical, especially since only the *ordinal* structure of this measure is important: A (strictly) monotone transformation of $\mathrm{sim}_{\mathcal{L}}(\cdot)$ will not change the inference result, i.e., the credible solution set! (As can be shown formally, it only changes the similarity bounds $\beta = h(\alpha)$ but not the neighborhoods $\mathcal{N}_{\beta}(\lambda)$). For example, in the case where solutions are real numbers (as in regression), this means that $\mathrm{sim}_{\mathcal{L}}(\cdot)$ can be defined by any monotone decreasing function of Euclidean distance.

Regarding the definition of $\mathrm{sim}_{\mathcal{X}}(\cdot)$, the cardinal structure of this measure is important in so far as it has an influence on the assignment of similarity pairs (the black points in Fig. 2) to the bins of the (fixed) partition underlying the specification of the similarity profile. Still, our experiments so far have shown that the profile is rather robust toward variations of $\mathrm{sim}_{\mathcal{X}}(\cdot)$. This can be explained by the fact that moving a similarity pair from one bin to another does only have an effect if this pair is a "critical" one that determines the similarity bound in one of the bins.

3 Extensions of CCBI

One disadvantage of credible case-based inference as outlined above concerns one of its key ingredients, the concept of a similarity profile. Due to the fact that a similarity profile provides worst case estimations in the form of lower similarity bounds, it is rather sensitive toward outliers, i.e., similarity pairs

$$(\alpha, \beta) = \big(\mathrm{sim}_{\mathcal{X}}(x_\imath, x_\jmath), \mathrm{sim}_{\mathcal{L}}(\lambda_{x_\imath}, \lambda_{x_\jmath})\big) \tag{13}$$

with comparatively small β. In fact, as $\zeta(\alpha)$ is a lower bound to the similarity of solutions that belong to α-similar problems, even the existence of a single pair of α-similar problems having rather dissimilar solutions entails a small lower bound $\zeta(\alpha)$. Small bounds in turn will obviously have a negative effect on the precision of predictions (5). This problem is illustrated in Fig. 3 for the `auto-mpg` data set, a benchmark from the UCI repository.[5] The picture clearly reveals the aforementioned outlier effect: The similarity profile is "pressed down" by a relatively small number of similarity pairs (13). In this section, some extensions of CCBI will be introduced in order to overcome this drawback.

[5] See section 4 for details concerning the specification of the underlying similarity measures.

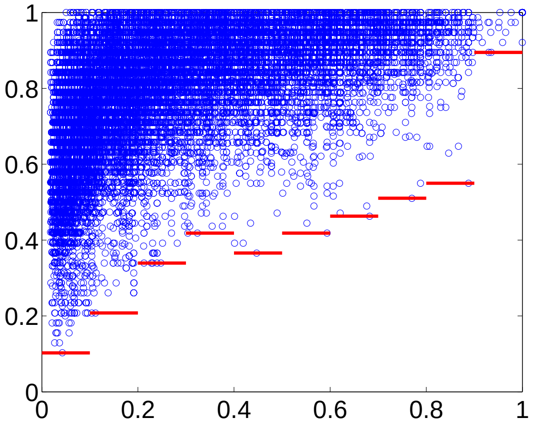

Figure 3: Empirical similarity profile for the `auto-mpg` data (step function). Each point corresponds to a pair (α, β) with $\alpha = \text{sim}_{\mathcal{X}}(x, y)$ (abscissa) and $\beta = \text{sim}_{\mathcal{L}}(\lambda_x, \lambda_y)$ (ordinate).

3.1 $\mathcal{M}$-Similarity Profiles

As already mentioned above, typically not all encountered cases are stored in the case base $\mathcal{M}$. In many applications, the case base will even remain more or less fixed. Under these conditions, not all potential similarity tuples of the form (13) are actually relevant. In fact, in the course of case-based inference, such tuples are only derived for pairs of cases

$$(\langle x_\imath, \lambda_{x_\imath} \rangle, \langle x_\jmath, \lambda_{x_\jmath} \rangle) \in \mathcal{M} \times (\mathcal{X} \times \mathcal{L}),$$

i.e., one of the cases is always an element of the case base. One may benefit from this fact by adjusting the similarity profile (and the corresponding inference scheme) to the case base $\mathcal{M}$. This leads to the concept of an $\mathcal{M}$-similarity profile, which is defined as follows: For all $\alpha \in [0, 1]$,

$$\zeta(\alpha) \overset{\text{df}}{=} \inf_{\langle x, \lambda_x \rangle \in \mathcal{M}, y \in \mathcal{X}, \text{sim}_{\mathcal{X}}(x,y)=\alpha} \text{sim}_{\mathcal{L}}(\lambda_x, \lambda_y) \, . \tag{14}$$

The definition of the empirical similarity profile is changed correspondingly: in the definition (9) of the coefficients β_k, the minimum is taken over those problems where $x_\imath$ is in the case base $\mathcal{M}$, and $x_\jmath$ is any problem that has been encountered so far. In this connection, it can be shown that inequality (12) in Theorem 3 can be generalized as follows:

$$\mathsf{Pr}\left(\, \lambda_{x_0} \notin C^{est}(x_0) \,\right) \;\leq\; \frac{2m}{1+N},$$

where N is the number of problems that have been encountered so far. (The proof is omitted here as it is quite similar to the proof of Theorem 3.)

3.2 Local Similarity Profiles

The idea of adapting a similarity profile to a given case base, as discussed above, can even be carried one step further: a similarity profile can be maintained for each individual case in the case base. We define the profile $\zeta_\imath(\cdot)$ of the $\imath$-th case $\langle x_\imath, \lambda_{x_\imath} \rangle$ as in (4), except that the infimum is taken only over those pairs of cases that involve the $\imath$-th case itself:

$$\zeta_\imath(\alpha) \overset{\mathrm{df}}{=} \inf_{y \in \mathcal{X}, \mathrm{sim}_{\mathcal{X}}(x_\imath, y) = \alpha} \mathrm{sim}_{\mathcal{L}}\left(\lambda_{x_\imath}, \lambda_y\right) \tag{15}$$

Thus, $\zeta_\imath(\cdot)$ allows for statements of the following kind: If a problem is α-similar to $x_\imath$, its solution is at least $\zeta_\imath(\alpha)$-similar to $\lambda_{x_\imath}$. Since $\zeta_\imath(\cdot)$ is adapted to a single case and, hence, to a local subregion of the problem space, we call it a *local similarity profile*. Note that a local profile can be seen as a suitable scaling of the neighborhood of a case and, hence, is in line with the idea of using locally adaptive metrics in nearest neighbor methods [14]. In the inference scheme (5), the neighborhoods $\mathcal{N}_{\zeta(\mathrm{sim}_{\mathcal{X}}(x_\imath, x_0))}(\lambda_{x_\imath})$ are now replaced by the neighborhoods $\mathcal{N}_{\zeta_\imath(\mathrm{sim}_{\mathcal{X}}(x_\imath, x_0))}(\lambda_{x_\imath})$. We refer to this type of local inference as CCBI-L.

The concept of a local similarity profile allows one to distinguish between typical and untypical cases. In fact, due to the enforced global validity of a standard similarity profile, untypical cases and outliers are treated in the same way as typical cases that are representative of their neighborhood. Consequently, the derivation of tight bounds from typical cases is prevented: in the inference scheme (5), the neighborhoods of such cases are as large as those of untypical cases. This effect is obviously avoided by using local similarity profiles. In this connection, we note that a local similarity profile, as a

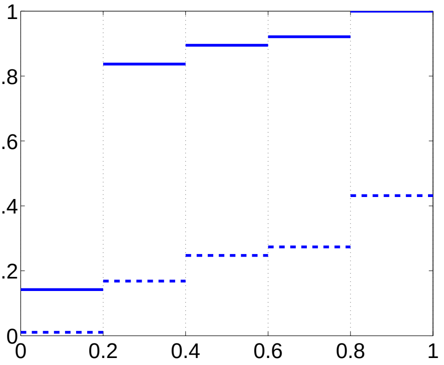

Figure 4: Local empirical similarity profiles of the 9-th (solid line) and 323-rd (dashed line) car in the `auto-mpg` data, using an equi-width partition of size 5.

quantification of the typicality of a case, might serve as a good criterion for selecting "competent" cases to be stored in the case base $\mathcal{M}$ [32]. In the `auto-mpg` data set, for example, the 9-th car (a pontiac catalina) has a much stronger developed (local) empirical profile than the 323-rd car (a mazda glc), as shown in Fig. 4. Thus, the former seems to be much more typical than the latter.

Mathematically speaking, a similarity profile $\zeta(\cdot)$ is the lower envelope of all local profiles $\zeta_\imath(\cdot)$. Consequently, CCBI-L will usually yield predictions that are more precise than those of CCBI. The price to pay is a higher computational complexity, since a profile must be maintained for every case in the case base $\mathcal{M}$. Moreover, it is to be expected that a prediction in the form of a credible solution set becomes less confident or, stated differently, that more cases are needed in order to achieve the same level of confidence. Indeed, it can be proved that inequality (12) in Theorem 3 can now be generalized as follows:

$$\mathsf{Pr}\left(\lambda_{x_0} \not\in C^{est}(x_0)\right) \;\leq\; \frac{|\mathcal{M}|m}{1+N},$$

where N is again the number of problems that have been encountered so far. As can be seen, the factor 2 in the upper bound is now replaced by $|\mathcal{M}|$ and, hence, increases linearly with the size of the case base. Anyway, the probability of an incorrect prediction still becomes arbitrarily small if the number of observed cases (N) is large enough in comparison with the number of cases in the case base.

3.3 Probabilistic Similarity Profiles

Another idea to increase the robustness of similarity profiles is to replace the deterministic similarity bounds $\zeta(\alpha)$ by *probabilistic* bounds. Such probabilistic bounds can be expressed in terms of (cumulative) probability distribution functions $F_\alpha(\cdot)$, with $F_\alpha(\beta)$ being the probability that $\mathrm{sim}_{\mathcal{L}}(\lambda_x, \lambda_y) \leq \beta$ for α-similar problems $x, y \in \mathcal{X}$ [20]:

$$F_\alpha(\beta) \overset{\mathrm{df}}{=} \mathsf{Pr}\left(\mathrm{sim}_{\mathcal{L}}(\lambda_x, \lambda_y) \leq \beta \,|\, \mathrm{sim}_{\mathcal{X}}(x, y) = \alpha\right).$$

In practice it will usually be sufficient to approximate a distribution function by a finite set of quantiles.

The representation of hypotheses in the form of step functions can easily be extended to the probabilistic setting. Let A_k be an interval in the representation (8) of hypotheses. Moreover, let S_k be the set of similarity degrees $\mathrm{sim}_{\mathcal{L}}(\lambda_{x_\imath}, \lambda_{x_\jmath})$ such that $\mathrm{sim}_{\mathcal{X}}(x_\imath, x_\jmath) \in A_k$. Rather than assigning to β_k the minimum of S_k, as in (9), we now define this bound by the $(1-p)$-quantile of S_k, where p is a usually small value such as 0.05. As an empirical quantile, β_k is hence an estimation of the corresponding true quantile of $F_\alpha(\cdot)$. We call the step function $h^p(\cdot)$ given by $h^p(\alpha) = \beta_k$ for $\alpha \in A_k$, with β_k as defined above, the *empirical p-profile.*

Now, suppose that we employ $h^p(\cdot)$ in order to derive a prediction

$$C^{est}(x_0) = \bigcap_{\imath=1}^{k} \mathcal{N}_{h^p(\mathrm{sim}_{\mathcal{X}}(x_\imath, x_0))}(\lambda_{x_\imath}),$$

where $x_1 \ldots x_k$ are the k nearest neighbors of the query problem x_0. What is the level of confidence of this prediction? Unfortunately, we do not have enough information to compute the probability of an incorrect prediction exactly. Still, by making a simplifying independence assumption à la naïve Bayes, the confidence level $(1-p)^k$ can be justified. Our practical experience has shown that this level still underestimates the true confidence level in almost any application (cf. section 4).

Of course, probabilistic estimations of the above type can be derived for different values

$p_1 < p_2 < \ldots < p_\ell$. Thus, by using this probabilistic variant of CCBI, that we call CCBI-P, one obtains a nested sequence

$$C^{p_\ell}(x_0) \subseteq C^{p_{\ell-1}}(x_0) \subseteq \ldots \subseteq C^{p_1}(x_0)$$

of credible solution sets with associated confidence levels. As an advantage of this kind of "stratified" prediction note that it differentiates between predicted solutions better than a single credible solution set does: The solutions in $C^{p_\ell}(x_0)$ are the *most likely* ones, those in $C^{p_{\ell-1}}(x_0) \setminus C^{p_\ell}(x_0)$ are somewhat less likely, and so on.

4 CCBI for Prediction Problems

As mentioned above, an especially simple yet relevant problem class for CBR is given by prediction problems, including classification and regression as special cases. In this context, CBR is typically referred to as *instance-based learning* (IBL) [4, 2]. From a machine learning point of view, IBL is an interesting alternative to inductive, model-based methods. Rather than inducing a general model (theory) from the data, IBL algorithms simply store the data itself [35]. The processing of the data is deferred until a prediction is actually requested, a property that qualifies IBL as a *lazy* learning method [3]. Predictions are then derived by combining the information provided by the stored examples in one way or other.

Typically, IBL is applied to classification problems, where predictions are derived from the query's k nearest neighbors through majority voting. Still, by combining the neighbors' predictions using a weighted sum rather than majority voting, IBL can also be employed for the estimation of numeric values [5]. In [22], the predictive performance of (numeric) IBL was found to be quite able to compete against linear regression (LR) as a representative of classical statistical approaches. More importantly, the authors correctly emphasized a key advantage of IBL, namely the fact that it does not assume strong (structural) properties of the data-generating process, such as linearity in LR.

This advantage, however, does not come for free. For methods that dispose of an underlying (statistical) model it is usually much simpler to quantify the *credibility* of a prediction. In LR, for example, an estimated model can be used for deriving a *confidence interval* covering a predicted output with a certain probability. Roughly speaking, this becomes possible by transferring the credibility of the model itself, estimated on the basis of the data in conjunction with the model assumptions, to predictions thereof. Interestingly enough, by deriving predictions in the form of credible sets, CCBI combines advantages from both instance-based and model-based learning: As an instance-based approach it requires fewer structural assumptions than (parametric) statistical methods, and yet it allows for specifying the uncertainty related to predictions.

This section is meant to convey a first idea of how CCBI can be applied to prediction problems and how it performs in practice. To this end, we present some experiments, in which we compared our approach to standard IBL (nearest neighbor estimation [12]). It should be noted in advance, however, that a fair comparison is difficult, especially since the methods provide predictions of different kind. For example, the main purpose of CCBI is to derive estimations in the form of *credible sets*, whereas IBL aims at producing good *point estimations* in the first place. As a consequence, standard IBL and CCBI are not directly comparable. And indeed, the main purpose of our studies is not to show that one approach is better than the other one, but instead that CCBI can reasonably complement standard IBL. Besides, the experiments are intended to support the theoretical results of the previous sections and to underpin our claim that CCBI combines advantages from both instance-based and model-based learning.

We performed experiments for regression problems, that is, prediction of numerical outputs. In this case, a training example is a tuple $\langle x, \lambda_x \rangle$, where $x = (x_1 \ldots x_m)$ is a vector of values for the input attributes, numerical or nominal, and λ_x is a value for the (numerical) output attribute. As a similarity measure, we used

$$\mathrm{sim}_{\mathcal{X}}(x,y) \overset{\mathrm{df}}{=} \exp\left(-\gamma \frac{1}{m} \sum_{\imath} d(x_\imath, y_\imath) \right), \tag{16}$$

where the distance $d(\cdot)$ is defined as $|x_\imath - y_\imath|$ for numerical attributes and assumes values

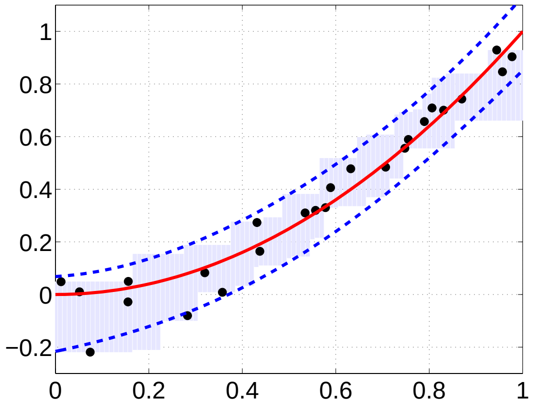

Figure 5: Approximation of $x \mapsto x^2$ (solid line) in the form of a confidence band, using CCBI (shaded region) and linear regression (region between dashed lines). The examples are indicated by black points.

0 and 1 for ordinal features (i.e., $d(x_\imath, y_\imath) = 0$ if $x_\imath = y_\imath$ and $= 1$ otherwise). To guarantee that all attributes do approximately have the same influence – a point of critical importance in IBL [22] – each input attribute is first re-scaled linearly to the unit interval. To facilitate the interpretation of quality measures, we re-scaled the output attribute in the same way.

Since our main objective is to compare IBL and CCBI under equal conditions, we refrained from "tuning" both methods. Particularly, we neither included feature selection nor feature weighting.[6] Besides, we did not put much effort in optimizing the constant γ in (16); $\gamma = 5$ seemed to produce reasonable results, and we used this value throughout our experiments. The partition of the unit interval underlying the similarity hypothesis in CCBI was always defined as a simple equi-width partition of size 10 for the global version and (since there are less training examples in the local approach) of size 5 for CCBI-L.

4.1 An Illustration using Artificial Data

The first example is a simple regression problem and mainly serves as an illustration. The function to be learned is given by the polynomial $x \mapsto x^2$. Moreover, n training examples $\langle x_\imath, \lambda_{x_\imath} \rangle$ are given, where the $x_\imath$ are uniformly distributed in $\mathcal{X} = [0, 1]$ and the

[6] It is well-known that irrelevant features can badly deteriorate instance-based learning methods and, on the other hand, that feature weighting can greatly improve performance [36].

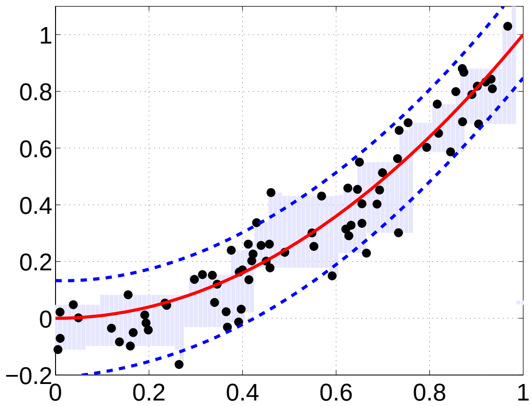

Figure 6: Approximation of $x \mapsto x^2$ (solid line) in the form of a confidence band, using CCBI-L and linear regression (region between dashed lines).

$\lambda_{x_\imath}$ are normally distributed with mean $(x_\imath)^2$ and standard deviation $1/10$. As mentioned above, we employed (16) with $\gamma = 5$ as a similarity measure for instances. Given a random sample (case base) $\mathcal{M}$, we first induce a similarity hypothesis for an underlying equi-width partition of size $m = 5$. Using this hypothesis and the sample $\mathcal{M}$, we derive a prediction λ_x for all instances $x \in [0, 1]$ (resp. for the discretization $\{0, 0.01, 0.02 \ldots 1\}$). Note that such a prediction is simply an interval. The union of these intervals yields a *confidence band* for the true mapping $x \mapsto x^2$. Fig. 5 shows a typical inference result for $n = 25$. Moreover, Fig. 6 shows a result for $n = 75$, using local similarity profiles (CCBI-L).

According to our estimation (12), the degree of confidence for $n = 25$ is $16/26$. This, however, is only a lower bound, and empirically (namely by averaging over 1,000 experiments) we found that the level of confidence is almost 0.9. To draw a comparison with standard statistical techniques, the figures also show the 0.9-confidence band obtained for the regression estimation (and the same samples). As can be seen, CCBI yields predictions of roughly the same precision, CCBI-L is even slightly more precise. This finding was also supported for estimation problems with other functions and input spaces of higher dimension.

In this connection, it should again be mentioned that linear resp. polynomial regression makes many more assumptions than CCBI. Especially, the type of function to be estimated must be specified in advance: Knowing that this function is a polynomial of degree 2 in our example, we took the model $x \mapsto \beta_0 + \beta_1 x + \beta_2 x^2$ as a point of departure and estimated the

	name	size	# var.
01	breast-tumor	277	1/8
02	cholesterol	297	6/7
03	cleveland	297	6/7
04	cpu	209	6/1
05	housing	506	12/1
07	pharynx	193	1/10
08	sensory	576	0/11
09	strike	625	5/1
10	bodyfat	252	14/0
11	pollution	60	15/0
12	pw-linear	200	10/0
13	auto-price	159	15/0
15	bolts	40	7/0
16	cloud	108	4/2
18	fruitfly	125	2/2
19	lowbwt	189	7/2
20	fishcatch	71	5/2
21	echo-months	61	6/3
22	quake	2178	3/0
23	auto-mpg	392	4/0

Table 1: Data sets used in the experiments: name, number of examples, number of predictor variables (numerical/nominal).

coefficients $\beta_\imath$. Usually, however, such knowledge will not be available. For instance, the performance of LR becomes much worse due to typical overfitting effects when adapting a polynomial of degree $k > 3$ to the data. Moreover, the confidence band for LR is only valid if the error terms follow a normal distribution (as they do in our case but not in general).

4.2 Real-World Regression Problems

We also applied CCBI to several real-world data sets from the UCI repository and the Statlib archive.[7] The data is summarized in table 1.

In order to test the effectiveness of the probabilistic strategy for CCBI, we have applied this approach to the data sets with different values for the parameter p (namely $p = 0, 0.02, 0.04$). The following performance measures were derived by means of a leave-one-

[7] `http://www.ics.uci.edu/~mlearn`, `http://lib.stat.cmu.edu/`

out cross-validation:

(1) The *correctness* or empirical confidence (CONF) measured in terms of the relative frequency of correct predictions (predicted interval covers true value).

(2) The *precision* of predictions (PREC) measured in terms of the average length of a predicted interval.

(3) The *mean absolute error* (MAE) measured in terms of the average distance between the true value and the point estimation (center of the interval).

As a neighborhood size for CCBI-P we used $k = 20$. Again, note that this parameter is less important in CCBI than in k-NN estimation. As mentioned previously, dissimilar neighbors will often hardly influence the prediction in terms of a credible set. And indeed, we observed that even though varying this parameter has an effect for small k, increasing k beyond ≈ 15 hardly changed the results.

The results for this series of experiments are summarized in table 2. As can be seen, the use of probabilistic bounds yields an extreme gain of precision at the cost of a mostly slight deterioration of the confidence. This finding, which basically holds true for all data sets, clearly provides strong evidence for the effectiveness of the probabilistic extension of CCBI: By varying the parameter p, a smooth tradeoff between confidence and precision can be achieved. Regarding the quality of the CCBI point estimation, the influence of p is less strong, though in general, more precise estimations come along with a slightly more accurate point estimation.

Admittedly, there are some data sets for which CCBI performs poorly, either in terms of confidence or in terms of precision or both. Looking at the characteristics of these data sets, there are two plausible explanations. Firstly, confidence and precision is weak if the size of the data set is too small. Of course, this is natural, since statistically confident and precise predictions cannot be made on the basis of sparse data. Secondly, CCBI seems to have problems with data sets in which nominal attributes prevail. As a plausible explanation, note that in this case there exist only a small number of different similarity degrees $\mathrm{sim}_{\mathcal{X}}(x, y)$. If these degrees are not well distributed over the unit interval, an

	CONF	PREC	MAE	CONF	PREC	MAE	CONF	PREC	MAE
01	0.9856	.8817	.1680	.8159	.5879	.1696	.7329	.4041	.1714
02	0.9663	.5918	.1271	.8013	.3310	.0909	.6599	.2398	.0909
03	1.0000	.8695	.3136	.9024	.5800	.2555	.7576	.3981	.2344
04	0.9809	.0665	.0187	.8278	.0404	.0177	.7608	.0274	.0187
05	1.0000	.6134	.0904	.8538	.3185	.0689	.7787	.2146	.0643
07	0.9896	.7895	.2069	.8083	.5596	.1807	.7202	.4670	.1791
08	1.0000	.9184	.1194	.8333	.4118	.1250	.7326	.2826	.1222
09	0.9888	.7758	.3506	.8368	.1551	.0727	.7296	.1104	.0608
10	0.9802	.3946	.0607	.8333	.2095	.0620	.6984	.1881	.0663
11	0.9500	.4974	.1173	.7500	.3267	.1232	.6333	.2682	.1140
12	0.9800	.5526	.0955	.8200	.3267	.0908	.7300	.2727	.0921
13	0.9623	.2484	.0583	.7547	.1404	.0509	.6792	.1223	.0544
15	0.9250	.5802	.2021	.6250	.3903	.1801	.4750	.1965	.1483
16	0.9537	.2956	.0714	.7963	.2157	.0739	.7315	.1848	.0779
18	0.9520	.8692	.2411	.7760	.5743	.2004	.6240	.4549	.1947
19	0.9577	.5292	.0934	.7937	.3281	.0950	.6508	.2584	.0983
20	0.9437	.2014	.0506	.8732	.1876	.0544	.7183	.1330	.0487
21	0.9344	.7245	.2520	.8033	.6601	.2542	.7049	.5443	.2322
23	0.9923	.6389	.1180	.8316	.3775	.0956	.7015	.2855	.0905

Table 2: Results for CCBI-P: Confidence, precision, and mean absolute error of predictions for $p = 0$ (left), $p = 0.02$ (middle), and $p = 0.04$ (right).

equi-with partition is likely to produce a poor and unbalanced similarity profile. In this case, the use of an *adaptive* partition (in line with equi-frequency histograms) seems to be advised, an option that we did not exploit so far but that should definitely be given a try.

We also found that the CCBI point estimations are on average slightly inferior to the point estimations produced by standard k-NN estimation (see also table 3 below), even though there are some exceptions where the former are even better than the latter. Nevertheless, the investigation of the statistical (Pearson) correlation between the precision (PREC) of CCBI estimations and the mean absolute error of the standard k-NN estimations (results omitted due to space restrictions) showed a significant positive correlation between these two quantities. This finding suggests that the width of the CCBI confidence interval is a good indicator of the accuracy of a k-NN prediction. Consequently, it might be an interesting idea to complement the latter by the former, i.e., to take the k-NN prediction as a point estimation and the CCBI prediction as a confidence interval.

	CONF	PREC	MAE	1-NN	3-NN	5-NN	7-NN	9-NN
01	0.8159	.3700	.1387	.2349	.1798	.1747	.1736	.1738
02	0.7407	.2159	.0782	.1263	.1020	.0960	.0926	.0899
03	1.0000	.2104	.0934	.1692	.1546	.1549	.1556	.1531
04	0.8038	.0530	.0202	.0165	.0241	.0288	.0306	.0309
05	0.8261	.1416	.0516	.0683	.0573	.0612	.0653	.0680
07	0.7254	.2926	.1135	.1989	.1578	.1429	.1394	.1367
08	0.8281	.1326	.0663	.1431	.1279	.1288	.1236	.1178
09	0.8432	.0913	.0328	.0344	.0279	.0296	.0291	.0293
10	0.8254	.1608	.0580	.0686	.0523	.0533	.0556	.0568
11	0.8333	.2263	.0865	.1234	.1166	.1044	.1098	.1087
12	0.8600	.1899	.0696	.1138	.0879	.0832	.0823	.0851
13	0.7547	.1203	.0424	.0498	.0502	.0539	.0562	.0589
15	0.7000	.1718	.0754	.1482	.1285	.1112	.1311	.1397
16	0.7407	.2432	.0916	.0887	.0805	.0773	.0872	.0963
18	0.7680	.4226	.1458	.2317	.1690	.1641	.1597	.1580
19	0.8201	.2087	.0724	.1074	.0961	.0901	.0889	.0891
20	0.7042	.0928	.0342	.0305	.0396	.0583	.0634	.0639
21	0.7541	.4917	.1660	.1999	.2020	.1965	.1869	.1878
22	0.9752	.4235	.1472	.1710	.1448	.1412	.1402	.1396
23	0.8571	.2757	.0727	.0900	.0765	.0750	.0735	.0742

Table 3: Results for CCBI-L: Confidence, precision, and mean absolute error of predictions; mean absolute error for k-NN point estimations with $k = 1, 3, 5, 7, 9$.

In a second series of experiments, we have employed the local version CCBI-L. The results are summarized in table 3. As it was to be expected from our theoretical analysis, predictions become more precise but less confident in comparison with the global version of CCBI. Apart from that, it is interesting to note that CCBI-L yields extremely good point estimations. In fact, more often than not, these point estimations are better than those of standard k-NN. Recalling that CCBI is actually not intended to produce point estimations, at least not in the first place, this is a surprisingly good an indeed unexpected result.

5 Related Work

As mentioned previously, formal approaches to CBR in general, and the formalization of the CBR hypothesis in particular, have not received very much attention as yet. There are, of course, a few notable exceptions. In [15], for example, it is shown that a special version of the CBR hypothesis is correct *on average*, in the sense that problems with similar features are more likely to have the same solution, given that the similarity measure is appropriately defined. Two important differences to our approach deserve mentioning: Firstly, we assume a similarity measure to be given, that is, our approach does not require the specification of an ideal measure (see also comments below) but remains valid regardless of the similarity measure employed. Secondly, we are not directly concerned with the probability of a correct versus incorrect prediction (which only makes sense if $|\mathcal{L}|$ is small, a requirement rarely fulfilled in CBR), but rather with the derivation of credible sets which are likely to cover the true output (solution).

In [25], the authors consider the problem to quantify the extent to which the CBR hypothesis holds for a particular application at hand. To this end, they propose a measure of the *problem–solution regularity.* In contrast to our concept of a similarity profile, however, this is a one-dimensional measure. Besides, it is not used for the purpose of prediction but rather as a kind of trigger for the maintenance of the CBR system.

Regarding the aspect of uncertainty in CBR, the importance of being able to assign degrees of confidence to predictions has been pointed out by several authors (e.g. [10]). In [11], different confidence measures for case-based (nearest neighbor) predictions are proposed and evaluated, and this work has been continued in [13] in connection with a concrete CBR application (Spam filtering). More generally, the problem to characterize the reliability of an estimation has recently received attention in the machine learning field as well [24, 28]. Again, however, note that assessing the confidence of a single point estimate, as done in the aforementioned papers, is quite different from the goal that we have pursued in the current paper, namely deriving predictions in the form of credible sets.

6 Summary, Conclusions, and Future Work

In this paper, we have proposed a method for supporting the retrieval of candidate solutions in case-based reasoning. Our method, called credible case-based inference (CCBI), exploits problem solving experience in the form of a case base in order to predict a set of promising candidates that might solve a new problem. The corresponding inference scheme is based on a formalization of the heuristic CBR hypothesis suggesting that similar problems have similar solutions. As CCBI has interesting theoretical properties, notably the fact that a prediction covers the true solution with high probability, it contributes to a formal foundation of CBR. Besides, let us highlight the following three points:

(1) As CCBI hardly assumes more than the specification of similarity measures for problems and solutions, it is quite general and widely applicable. In fact, note that no kind of transitivity is assumed for the similarity measures, which means that the structure of the problem space $\mathcal{X}$ and the solution space $\mathcal{L}$ might be weaker than that of a metric space (which is a point of great practical relevance in CBR). Consequently, CCBI predictions can be derived in many situations where standard methods (e.g. from statistics) are not applicable.

In fact, even though in our experiments we employed CCBI only for regression problems, it can be applied to more complex output spaces in exactly the same manner. For example, it could be used to predict *rankings* in so-called label ranking problems [19, 16], or other types of structured output like *trees* or *graphs* that might represent solutions in a CBR context. In these cases, a CCBI prediction would be given by a credible subset of ranking, trees, or graphs, respectively.

(2) CCBI works for any pair of similarity measures, even if these measures are not defined in an optimal way. That is, the predictions remain correct with high probability, even though they might become rather imprecise. This, however, should not be seen as a disadvantage. On the contrary, CCBI does not pretend a precision or credibility of case-based predictions which is actually not justified. Instead, imprecise predictions can be taken as an indication that either CBR is not appropriate for the application, or at least that the similarity measures are not well specified.

(3) As a formalization of the CBR hypothesis, the concept of a similarity profile, which plays a key role in CCBI, is interesting by itself. In particular, a similarity profile clearly shows to what extent the CBR hypothesis actually holds true for the application at hand.

Moreover, the concept of a local similarity profile supports the selection of representative and, hence, useful cases to be stored in the case base.

In section 4, we have applied CCBI to regression problems. In this regard, it is worth mentioning that CCBI in a sense unifies diverse types of prediction problems. Moreover, it combines advantages from both, instance-based and model-based (statistical) learning: As an instance-based approach it requires fewer structural assumptions than (parametric) statistical methods, and yet it allows for specifying the uncertainty related to predictions. Empirical results presented for regression problems suggest that CCBI performs rather well in practice. In particular, even though it is not designed to produce point estimations, it is quite competitive to standard IBL in this regard. Besides, it yields useful predictions in the form of credible sets, which is of course its key advantage.

From a (case-based) problem solving point of view, prediction appears to be the most simple problem class, mainly because there is no need for adaptation. In the experimental part we have focused on this type of problem as it allows for a systematic evaluation of the prediction quality of CCBI, and since corresponding benchmark data is available. Regarding future work, an interesting idea is to go one step further and apply CCBI in the context of search-oriented CBR as briefly touched on in the introduction. We conclude the paper by giving a brief outline of this idea.

According to the view of transformational adaptation taken in [8], case-based problem solving can be cast as a search process. Within the related model, (potential) cases correspond to search states and adaptation operators play the role of search operators. Now, the key idea is to use CCBI in order to complement this model in a reasonable way. In fact, in [8] the authors note that, according to their approach, CBR could principally be realized by enumerating the search space completely. Understandably, they look at this idea with reservation, immediately pointing to the enormous complexity it brings about. Our approach applies exactly to this problem: CCBI supports CBR by predicting a promising ("credible") subset of search states, thereby focusing search to promising cases and providing important information to a search method which is applied for actually finding a solution. From the perspective of CBR, this approach might not merely be seen as an application. In conjunction with the ideas presented in [8], it could contribute in a more general way to a formal framework of CBR in which (transformational) adaptation is realized as a search process and (case-based) experience is used in order to concentrate on promising regions of the related search space.

Indeed, in [8], the concept of similarity is integrated into problem solving by means of a, say, "ideal" similarity measure. By pointing to optimal initial search states, this measure somehow guarantees the retrieval of cases which can be adapted easily. Needless to say, finding such measures will be difficult in practice, if possible at all. CCBI takes a different (more pragmatic) approach: It takes any similarity measure as a *given* input, even if this measure is not "ideal". It then derives a *set* of *promising* search states rather than *the optimal* initial state, and the precision of this prediction depends on how ideal the similarity measure actually is.

Acknowledgements. The author is grateful to three anonymous referees whose comments helped to improve the paper.

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] D.W. Aha. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies*, 36:267–287, 1992.

[3] D.W. Aha, editor. *Lazy Learning.* Kluwer Academic Publ., 1997.

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

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

[6] E. Balfe and B. Smyth. Case-based collaborative web search. In *ECCBR-2004, 7th European Conference on Case-Based Reasoning*, Madrid, Spain, 2004.

[7] R. Bergmann, K.D. Althoff, S. Breen, M. Göker, M. Manago, R. Traphöner, and S. Wess. *Developing industrial case-based reasoning applications: The INRECA methodology*, volume 1612 of *LNAI*. 2 edition, 2003.

[8] R. Bergmann and W. Wilke. Towards a new formal model of transformational adaptation in case-based reasoning. In H. Prade, editor, ECAI-98*, 13th European Conference on Artificial Intelligence*, pages 53–57, 1998.

[9] S. Brüninghaus and KD. Ashley. Combining case-based and model-based reasoning for predicting the outcome of legal cases. In *Proceedings ICCBR-2003, 5th International Conference on Case-Based Reasoning*, Trondheim, 2003. Springer-Verlag.

[10] W. Cheetham. Case-based reasoning with confidence. In *EWCBR–2000, 5th European Workshop on Case-Based Reasoning*, pages 15–25, Trento, Italy, 2000.

[11] W. Cheetham and J. Price. Measures of solution accuracy in case-based reasoning systems. In *Proc. ECCBR–2004, 7th European Conference on Case-Based Reasoning*, pages 106–118, Madrid, Spain, 2004. Springer-Verlag.

[12] B.V. Dasarathy, editor. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, Los Alamitos, California, 1991.

[13] SJ. Delany, P. Cunningham, D. Doyle, and A. Zamolotskikh. Generating estimates of classification confidence for a case-based spam filter. In *Proc. ICCBR-2005, 6th Int. Conf. on Case-Based Reasoning*, pages 177–190, Chicago, 2005. Springer-Verlag.

[14] C. Domeniconi, J. Peng, and D. Gunopulos. Locally adaptive metric nearest-neighbor classificaton. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:1281–1285, 2002.

[15] B. Faltings. Probabilistic indexing for case-based prediction. In D.B. Leake and E. Plaza, editors, *Case-based Reasoning Research and Developement, Proceedings* ICCBR-97, pages 611–622. Springer-Verlag, 1997.

[16] J. Fürnkranz and E. Hüllermeier. Pairwise preference learning and ranking. In *Proc.* ECML–2003*, 13th European Conference on Machine Learning*, Cavtat-Dubrovnik, Croatia, September 2003.

[17] T. Gabel and A. Stahl. Exploiting background knowledge when learning similarity measures. In *ECCBR-2004, 7th European Conference on Case-Based Reasoning*, Madrid, Spain, 2004.

[18] A. Gammerman and V. Vovk. Prediction algorithms and confidence measures based on algorithmic randomness theory. *Theoretical Computer Science*, 287:209–217, 2002.

[19] S. Har-Peled, D. Roth, and D. Zimak. Constraint classification: a new approach to multiclass classification. In *Proceedings 13th Int. Conf. on Algorithmic Learning Theory*, pages 365–379, Lübeck, Germany, 2002. Springer.

[20] E. Hüllermeier. Toward a probabilistic formalization of case-based inference. In *Proc.* IJCAI–99*, 16th International Joint Conference on Artificial Intelligence*, pages 248–253, Stockholm, Sweden, July/August 1999.

[21] E. Hüllermeier. Instance-based prediction with guaranteed confidence. In *Proc.* ECAI–04*, 16th European Conference on Artificial Intelligence*, pages 97–101, Valencia, Spain, 2004.

[22] D. Kibler, D.W. Aha, and MK. Albert. Instance-based prediction of real-valued attributes. *Computational Intelligence*, 5:51–57, 1989.

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

[24] M. Kukar and I. Kononenko. Reliable classifications with machine learning. In *Proc. European Conference on Machine Learning, ECML*, pages 219–231, 2002.

[25] DB. Leake and DC. Wilson. When experience is wrong: Examining CBR for changing tasks and environments, In *Proc. ICCBR–99*, 218–232, 1999.

[26] T.M. Mitchell. Version spaces: A candidate elimination approach to rule learning. In *Proceedings* IJCAI-77, pages 305–310, 1977.

[27] R. Pan, Q. Yang, JJ. Pan, and L. Li Competence. Driven Case-Base Mining. In *Proc. AAAI–2005*, 228-233. Pittsburgh, USA.

[28] K. Proedrou, I. Nouretdinov, V. Vovk, and A. Gammerman. Transductive confidence machines for pattern recognition. In *Proc. European Conference on Machine Learning, ECML*, pages 381–390, 2002.

[29] C.K. Riesbeck and R.C. Schank. *Inside Case-based Reasoning.* Hillsdale, New York, 1989.

[30] B. Smyth and M.T. Keane. Adaptation-guided retrieval: questioning the similarity assumption in reasoning. *Artificial Intelligence*, 102(2):249–293, 1998.

[31] B. Smyth and T. Keane. Remembering to forget. In C.S. Mellish, editor, *Proceedings International Joint Conference on Artificial Intelligence*, pages 377–382. Morgan Kaufmann, 1995.

[32] B. Smyth and E. Mc Kenna. Building compact competent case-bases. In Proc. 3rd International Conference on Case-Based Reasoning, number 1650 in LNAI, pages 329–342, 1999.

[33] B. Smyth and E. McKenna. Competence models and the maintenance problem. *Computational Intelligence*, 17(2):235–249, 2001.

[34] A. Stahl. Learning of similarity measures: A formal view based on a generalized CBR model. In *Proceedings ICCBR-2005, 6th International Conference on Case-Based Reasoning*, Chicago, Illinois, 2005. Springer-Verlag.

[35] C. Stanfill and D. Waltz. Toward memory-based reasoning. *Communications of the ACM*, pages 1213–1228, 1986.

[36] D. Wettschereck, D.W. Aha, and T. Mohri. A review and empirical comparison of feature weighting methods for a class of lazy learning algorithms. *AI Review*, 11:273–314, 1997.