Instance-Based Learning of Credible Label Sets

Eyke Hüllermeier

Informatics Institute, Marburg University, Germany
eyke@mathematik.uni-marburg.de

Abstract. Even though instance-based learning performs well in practice, it might be criticized for its neglect of uncertainty: An estimation is usually given in the form of a predicted label, but without characterizing the confidence of this prediction. In this paper, we propose an instance-based learning method that allows for deriving "credible" estimations, namely set-valued predictions that cover the true label of a query object with high probability. Our method is built upon a formal model of the heuristic inference principle underlying instance-based learning.

1 Introduction

The name instance-based learning (IBL) stands for a family of machine learning algorithms, including well-known variants such as memory-based learning, exemplar-based learning and case-based reasoning [14, 10, 8]. As the term suggests, in instance-based algorithms special importance is attached to the concept of an *instance* [2]. An instance, also called a case, an observation or an example, can be thought of as a single experience, such as a pattern (along with its classification) in pattern recognition or a problem (along with a solution) in case-based reasoning.

As opposed to inductive, model-based machine learning methods, IBL provides a simple means for realizing *transductive* inference [15], that is inference "from specific to specific": Rather than inducing a general model (theory) from the data and using this model for further reasoning, the data itself is simply stored. The processing of the data is deferred until a prediction (or some other type of query) is actually requested, a property which qualifies IBL as a *lazy* learning method [1]. Predictions are then derived by combining in one way or the other the information provided by the stored examples, especially by those objects which are *similar* to the new query.

In fact, the concept of similarity plays a central role in IBL whose underlying inference principle corresponds to the well-known *nearest-neighbor* rule, suggesting that "similar objects have similar labels". This assertion, which we shall occasionally call the "IBL assumption", is apparently of *heuristic* nature: It is a rule of thumb that works in most situations but is not guaranteed to do so in every case. This clearly reveals the necessity of taking the aspect of *uncertainty* in IBL into account [4]. Especially, this is true for sensitive applications such as medical diagnosis or legal reasoning and all the more if decisions (classifications) must be made on the basis of sparse experience.

In this paper, we shall propose an instance-based learning method that allows for deriving "credible" predictions. The way in which this approach takes the aspect of uncertainty into account takes its inspiration from statistical methods: The basic idea is to derive a kind of *credible set*[1] for the value (label) to be estimated, that is a subset of values which is likely to contain the true one.

The remaining part of the paper is organized as follows: After some preliminaries, we introduce the concepts of a *similarity profile* and a *similarity hypothesis* (Sections 3–4). These concepts will allow us to propose a formal model of the IBL assumption as well as an instance-based inference scheme that derives predictions in the form of a set of potential labels. Then, a method for learning similarity hypotheses from a memory of cases will be presented, along with theoretical and empirical results on the validity of predictions derived from such hypotheses (Section 5). Finally, in Section 6, we consider the problem of adapting the involved similarity measures so as to optimize our algorithm's performance.

2 Preliminaries

Throughout the paper we proceed from the following setting: $\mathcal{X}$ denotes the instance space, where an instance corresponds to the description x of an object (usually in attribute–value form). $\mathcal{X}$ is endowed with a reflexive and symmetric similarity measure, $\sigma_{\mathcal{X}}$. $\mathcal{L}$ is a set of labels, also endowed with a reflexive and symmetric similarity measure, $\sigma_{\mathcal{L}}$. We assume that $\sigma_{\mathcal{X}}$ and $\sigma_{\mathcal{L}}$ are normalized such that both measures return similarity degrees between 0 and 1, where 1 stands for complete similarity. $\mathcal{D}$ denotes a sample (memory, case base) that consists of n labeled instances (cases) $\langle x_\imath, \lambda_{x_\imath} \rangle \in \mathcal{X} \times \mathcal{L}$, $1 \leq \imath \leq n$. Finally, a novel instance $x_0 \in \mathcal{X}$ (a query) is given, whose label λ_{x_0} is to be estimated.

We do not make any assumptions on the cardinality of the label set $\mathcal{L}$. In fact, we do not even distinguish between the performance tasks of classification (estimating one among a finite set of class labels) and regression (estimating a real-valued output), which means that $\mathcal{L}$ might even be infinite. As concerns classification, however, it deserves mentioning that our method is more suitable for problems involving many labels. This does hardly diminish its practical relevance, since there are enough problems of this type. For example, consider case-based problem solving where an instance corresponds to a problem description, e.g. a set of requirements a technical system has to meet, and the label corresponds to a solution of that problem, e.g. the assemblage of primitive components into a complete technical system [5]. Having to build a new system, one will usually try to exploit experience that has been gained from building systems for similar requirements, relying on the assumption that "similar problems have similar solutions". As another example, consider a problem somewhat more difficult than classification: Rather than predicting one among n labels, we seek a full *ranking* of these labels, that is a complete order relation [3]. Since n "ba-

[1] This term is also common in Bayesian statistics. A related concept in classical statistics is that of a confidence region.

sic" labels can be arranged to $n!$ different rankings, the actual set of potential predictions can be huge.

Finally, note that no kind of transitivity is assumed for the similarity measures, which means that the structure of $\mathcal{X}$ and $\mathcal{L}$ is weaker than that of a metric space. This excludes the application of several standard methods from statistics.

3 Similarity Profiles

A basic idea of our approach is to proceed from a formal model of the heuristic IBL assumption, that is a formalization of this otherwise vague principle. As will be seen later, this formalization provides the basis of a sound inference procedure and will allow us to make assertions about the confidence of predictions.

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

$$\forall\, x_1, x_2 \in \mathcal{X} \,:\, \sigma_{\mathcal{X}}(x_1, x_2) \;\leq \sigma_{\mathcal{L}}\left(\lambda_{x_1}, \lambda_{x_2}\right). \tag{1}$$

In words: The similarity between two labels is always lower-bounded by the similarity between the corresponding instances, or, roughly speaking, the more similar two instances are, the more similar are the corresponding labels.

If the similarity constraint (1) does indeed hold true for the application at hand, then one can reason as follows: Given a query x_0 and an observed case $\langle x_1, \lambda_{x_1}\rangle$ such that x_1 is α_1-similar to x_0, the unknown label λ_{x_0} must be an element of the α_1-neighborhood of the label λ_{x_1}, i.e., of the set of labels λ such that $\sigma_{\mathcal{L}}(\lambda, \lambda_{x_1}) \geq \alpha_1$. Moreover, given a second case $\langle x_1, \lambda_{x_1}\rangle$, the same kind of reasoning applies, and we can conclude that λ_{x_0} must be an element of a certain α_2-neighborhood of λ_{x_2}. And we can even come up with a more precise prediction by combining the two constraints: λ_{x_0} must belong to the intersection of the two neighborhoods (see Fig. 1).

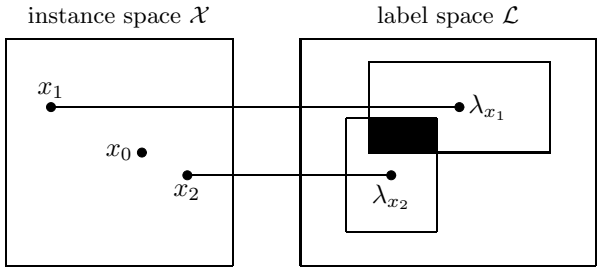

Fig. 1. Each case puts a constraint on the label λ_{x_0} by virtue of property (1).

Needless to say, the similarity constraint (1) will usually not be satisfied for a practical application (and given similarity measures $\sigma_{\mathcal{X}}$, $\sigma_{\mathcal{L}}$). Therefore, let us

consider a relaxation of this constraint:

$$\forall\, x_1, x_2 \in \mathcal{X} \,:\, \zeta\left(\sigma_{\mathcal{X}}(x_1, x_2)\right) \,\leq \sigma_{\mathcal{L}}\left(\lambda_{x_1}, \lambda_{x_2}\right), \tag{2}$$

where ζ is a function $\mathcal{A} \to [0,1]$ with $\mathcal{A} =_{\text{def}} \{\sigma_{\mathcal{X}}(x, x') \,|\, x, x' \in \mathcal{X}\}$. This function assigns to each similarity degree between two instances, α, the largest similarity degree $\beta = \zeta(\alpha)$ such that the following property holds:

$$\forall\, x_1, x_2 \in \mathcal{X} \,:\, \sigma_{\mathcal{X}}(x_1, x_2) = \alpha \;\Rightarrow\; \sigma_{\mathcal{L}}(\lambda_{x_1}, \lambda_{x_2}) \,\geq\, \zeta(\alpha).$$

We call ζ the *similarity profile* of the application at hand. More formally, ζ is defined as follows: For all $\alpha \in \mathcal{A}$,

$$\zeta(\alpha) \,=_{\text{def}} \inf_{x,x' \in \mathcal{X}, \sigma_{\mathcal{X}}(x,x') = \alpha} \sigma_{\mathcal{L}}\left(\lambda_x, \lambda_{x'}\right).$$

Note that the similarity profile conveys a precise idea of the extent to which the application at hand actually meets the IBL assumption. Roughly speaking, the larger ζ is, the better this assumption is satisfied.

Using the relaxed constraint (2), we can perform the same kind of reasoning as before. We only have to replace the $\alpha_\imath$-neighborhoods of the known labels $\lambda_{x_\imath}$ by the corresponding $\beta_\imath$-neighborhoods, where $\beta_\imath = \zeta(\alpha_\imath)$. Thus, the following inference scheme is obtained: $\lambda_{x_0} \in C(x_0)$ with $C(x_0) =_{\text{def}} \mathcal{L}$ if $\mathcal{D} = \emptyset$ and

$$C(x_0) =_{\text{def}} \bigcap_{\imath=1}^{n} \mathcal{N}_{\zeta(\sigma_{\mathcal{X}}(x_\imath, x_0))}(\lambda_{x_\imath}) \tag{3}$$

otherwise, where the β-neighborhood of a label λ is given by

$$\mathcal{N}_\beta(\lambda) =_{\text{def}} \{\lambda' \in \mathcal{L} \,|\, \sigma_{\mathcal{L}}(\lambda, \lambda') \geq \beta\}. \tag{4}$$

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 ζ. We call $C(x_0)$ a *credible label set*, or simply a credible set.

Note that taking the intersection over $k < n$ of the cases in $\mathcal{D}$ comes along with a loss of precision but preserves correctness of the prediction (3). Since less similar instances will often hardly contribute to the precision of predictions, it might indeed be reasonable to derive $C(x_0)$ from the $k \ll n$ instances maximally similar to x_0, all the more if computing the intersection of neighborhoods (4) is computationally complex.

An apparent disadvantage of a similarity profile concerns its sensitivity toward outliers or, say, "exceptional" cases. In fact, recall that $\zeta(\alpha)$ is a *lower bound* to the similarity of labels that belong to α-similar instances. Thus, the existence of only one pair of α-similar instances having rather dissimilar labels entails a small lower bound $\zeta(\alpha)$. Small bounds in turn will obviously have a negative effect on the precision of (3).

One way to avoid this problem is to maintain an individual similarity profile for each case in the memory $\mathcal{D}$. This approach is somehow comparable to the use

of *local metrics* in kNN algorithms and IBL, e.g., metrics which allow feature weights to vary as a function of the instance [11]. The *local similarity profile* of the $\imath$th case $\langle x_\imath, \lambda_{x_\imath}\rangle$ is defined as follows:

$$\zeta_\imath(\alpha) =_{\text{def}} \inf_{x\in\mathcal{X},\sigma_\mathcal{X}(x,x_\imath)=\alpha} \sigma_\mathcal{L}\left(\lambda_x, \lambda_{x_\imath}\right),$$

where $\inf\emptyset = 1$ by definition. Thus, $\zeta_\imath(\alpha)$ is a lower bound on the similarity between $\lambda_{x_\imath}$ and the label λ_{x_0} of an instance x_0 which is α-similar to $x_\imath$. A local profile indicates the validity of the IBL assumption for *individual* cases. The inference scheme (3) now becomes

$$C(x_0) =_{\text{def}} \bigcap_{\imath=1}^{n} \mathcal{N}_{\zeta_\imath(\sigma_\mathcal{X}(x_\imath,x_0))}(\lambda_{x_\imath}). \tag{5}$$

As can be seen, a case with a poorly developed profile hardly contributes to precise predictions. The local similarity profile might hence serve as a (perhaps complementary) criterion for selecting "competent" cases to be stored in the memory $\mathcal{D}$ [13].

4 Similarity Hypotheses

The application of the inference scheme (3) requires the similarity profile ζ to be known, a requirement that will usually not be fulfilled. This motivates the related concept of a *similarity hypothesis*, which is thought of as an approximation of a similarity profile. A similarity hypothesis can thus be seen as a formal model of the IBL assumption, adapted to the application under consideration.

Formally, a similarity hypothesis is identified with a function $h : \mathcal{A} \to [0, 1]$. The intended meaning of the hypothesis h is that

$$\forall\, x_1, x_2 \in \mathcal{X} \,:\, \sigma_\mathcal{X}(x_1, x_2) = \alpha \;\Rightarrow\; \sigma_\mathcal{L}(\lambda_{x_1}, \lambda_{x_2}) \geq h(\alpha). \tag{6}$$

A hypothesis h is called *stronger* than a hypothesis h' if $h' \leq h$ and $h \not\leq h'$. We say that h is *admissible* if $h(\alpha) \leq \zeta(\alpha)$ for all $\alpha \in \mathcal{A}$.

It is obvious that using an admissible hypothesis h in place of the true similarity profile ζ within the inference scheme (3) leads to correct predictions. That is, the estimation

$$C^{est}(x_0) =_{\text{def}} \bigcap_{\imath=1}^{n} \mathcal{N}_{h(\sigma_\mathcal{X}(x_\imath,x_0))}(\lambda_{x_\imath}) \tag{7}$$

is guaranteed to cover the unknown label λ_{x_0}. Indeed, $h \leq \zeta$ implies

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

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

Yet, assuming the profile ζ to be unknown, one cannot guarantee the admissibility of a hypothesis h and, hence, the correctness of (7). 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 is indeed a good approximation of ζ, it seems reasonable to derive $C^{est}(x_0)$ according to (7) as an approximation of $C(x_0)$, that is, to realize instance-based learning as a kind of *approximate* reasoning. Our results in the next section, showing how to derive a suitable hypothesis from the data given and how to estimate the probability that predictions obtained from such hypotheses are correct, will provide a formal justification for this approach.

Before proceeding, let us note that an approximate version of the local inference scheme (5) can of course be realized as well. In this case, an individual hypothesis $h_\imath$ has to be specified (or induced from data) for each case $\langle x_\imath, \lambda_{x_\imath} \rangle$.

5 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 method thus obtained can be seen as a combination of instance- 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 IBL assumption, whereas storing new cases in the memory corresponds to instance-based learning.

Given a hypothesis space $\mathcal{H}$, i.e. a class of functions $h : \mathcal{A} \to [0, 1]$, learning amounts to choosing one among these hypotheses on the basis of the given data. But which of the hypotheses are interesting candidates? Of course, first of all a hypothesis h should be consistent with the data given, that is, (6) should be satisfied for all cases in $\mathcal{D}$:

$$\forall\, x, x' \in \mathcal{D} \,:\, \sigma_{\mathcal{X}}(x, x') = \alpha \;\Rightarrow\; \sigma_{\mathcal{L}}(\lambda_x, \lambda_{x'}) \geq h(\alpha). \tag{8}$$

Denote by $\mathcal{H}_C \subseteq \mathcal{H}$ the set of hypothesis that are consistent in this sense. Among two consistent hypothesis h and h', where h is stronger than h', we should prefer the former since it leads to more precise predictions.[2] Thus, we call a hypothesis h_* optimal if $h_* \in \mathcal{H}_C$ and if there is no hypothesis $h \in \mathcal{H}_C$ such that h is stronger than h_*. 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 $\mathcal{H}_C$ is given by the set $\{h \in \mathcal{H} \,|\, h \leq h_*\}$.* □

Given the assumptions of this observation, IBL can be realized as a *candidate-elimination* algorithm [9], where h_* 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.

[2] Note that the extreme hypothesis $h \equiv 0$ is always consistent but leads to the trivial prediction $C^{est}(x_0) = \mathcal{L}$.

Note that (8) guarantees consistency in the "empirical" sense that $\lambda_{x_\imath} \in C^{est}(x_\imath)$ for all $\langle x_\imath, \lambda_{x_\imath}\rangle \in \mathcal{D}$. One might think of further demanding a kind of "logical" consistency, namely $C^{est}(x) \neq \emptyset$ for all $x \in \mathcal{X}$. Of course, this additional requirement makes the testing of consistency more difficult and would greatly increase the complexity of learning.

5.1 Hypotheses as step functions

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

$$h : x \mapsto \sum_{k=1}^{m} \beta_k \cdot \mathbb{I}_{A_k}(x), \tag{9}$$

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 $[0, 1]$. The class $\mathcal{H}_{step}$ of functions (9), defined for a fixed partition, does obviously satisfy the assumptions of Observation 1. The optimal hypothesis h_* is defined by the values

$$\beta_k \;=_{\text{def}}\; \min\{\, \sigma_{\mathcal{L}}(\lambda_x, \lambda_{x'}) \mid \langle x, \lambda_x\rangle,\, \langle x', \lambda_{x'}\rangle \in \mathcal{D},\, \sigma_{\mathcal{X}}(x, x') \in A_k \,\} \tag{10}$$

for $1 \leq k \leq m$, where $\min \emptyset = 1$ by definition. We call h_* the *empirical similarity profile*.

Now, suppose that the case base is to be extended, i.e. that a newly observed case $\langle x_{n+1}, \lambda_{x_{n+1}}\rangle$ is to be added to the current sample $\mathcal{D}$. Updating the empirical similarity profile h_* can then be accomplished by passing the iteration

$$\beta_{\kappa(x_{n+1}, x_\jmath)} \;\leftarrow\; \min\{\, \beta_{\kappa(x_{n+1}, x_\jmath)}, \sigma_{\mathcal{L}}(\lambda_{x_{n+1}}, \lambda_{x_\jmath}) \,\} \tag{11}$$

for $1 \leq \jmath \leq n = |\mathcal{D}|$. The index $1 \leq \kappa(x, x') \leq m$ is defined for instances $x, x' \in \mathcal{X}$ by $\kappa(x, x') = k \Leftrightarrow \sigma_{\mathcal{X}}(x, x') \in A_k$. As can be seen, the time complexity of updating the empirical profile is linear in the size of the memory.

5.2 The learning process

The updating scheme (11) suggests a process in which prediction and learning are repeated alternately in the style of incremental supervised learning:

- At each point of time, we dispose of a sample $\mathcal{D}$ with an associated empirical similarity profile h_*.
- Having to predict the label of a new instance x_0, an estimation $C^{est}(x_0)$ is derived from $\mathcal{D}$ and h_* according to (7).
- The system learns the correct label λ_{x_0} from the teacher.
- $\langle x_0, \lambda_{x_0}\rangle$ is added as the $(n+1)$th case $\langle x_{n+1}, \lambda_{x_{n+1}}\rangle$ to the memory and the empirical profile h_* is updated.

Needless to say, the strategy of simply adding all observations to the current case base $\mathcal{D}$ will usually not be efficient. In fact, much more sophisticated strategies for maintaining a case base are often used in practice, including the possibility of removing or replacing stored cases [12]. 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 $h^* \in \mathcal{H}_{step}$ defined by the values

$$\beta_k^* =_{\text{def}} \inf \{ \zeta(x) \mid x \in \mathcal{A} \cap A_k \}, \tag{12}$$

$1 \leq k \leq m$, the *optimal admissible* hypothesis. Since admissibility implies consistency, we have $h^* \leq h_*$. This inequality suggests that the empirical similarity profile h_* will usually overestimate the true profile ζ and, hence, that h_* might not be admissible. And indeed, the constraints imposed by the observed cases will usually not "press" the step function h_* below the profile ζ (see Fig. 2 for an illustration). Of course, the fact that admissibility of h_* 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. In the sequel, we shall present first answers to these questions.

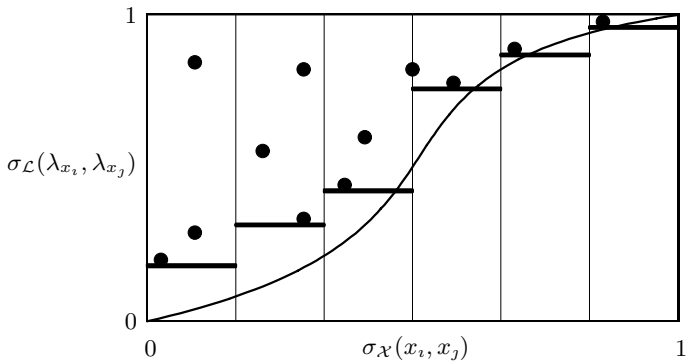

Fig. 2. Similarity profile (solid line) and empirical similarity profile (step function). Each point is induced by a pair of observed cases. By the definition of the similarity profile, all points are located above the graph of that function.

5.3 Properties of the learning process

We make the simplifying assumption that the instance space $\mathcal{X}$ is countable. Further, we make the standard assumption that the query instances 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 μ. In other words, the observed cases are

independent and identically distributed (*i.i.d*) random variables, i.e. $\mathcal{D}$ is an *i.i.d* sample. Note that we can assume $\mu(x) > 0$ for all $x \in \mathcal{X}$ without loss of generality.

Now, denote by $\mathcal{D}_n$ the case base in the nth step of the above learning process, that is the sample $\mathcal{D}$ such that $|\mathcal{D}| = n$, and by h_n the empirical similarity profile derived from that sample. Since, according to our assumption, the observed cases are random variables, the induced hypotheses h_n 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^*.[3]

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 $\mathsf{Pr}(|h_N - h^*|_\infty \geq \varepsilon) \to 0$ for all $\varepsilon > 0$.* □

As concerns the quality of estimations, we are first of all interested in the probability of incorrect predictions. Denote by

$$q_{n+1} \;=_{\text{def}}\; \mathsf{Pr}\left(\, \lambda_{x_0} \notin C^{est}(x_0) \,|\, \mathcal{D}_n, h_n \,\right) \tag{13}$$

the probability that the $(n+1)$th prediction, i.e. the prediction derived from $\mathcal{D}_n$ and h_n, is incorrect. In this connection, it should be noted that a prediction might well be correct even if the involved empirical profile h_* is not admissible: Recall that the estimation (7) is derived from a *limited* number of constraints (4), namely the $\beta_\imath$-neighborhoods associated with known labels $\lambda_{x_\imath}$. As we cannot exclude that $\beta_\imath = h_n(\sigma_\mathcal{X}(x_\imath, x_0)) > \zeta(\sigma_\mathcal{X}(x_\imath, x_0))$, it is true that each of these neighborhoods might be "too small" and, hence, might remove some labels from the credible 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 interesting result:

Theorem 3 *The following estimation holds true for the probability (13):*

$$q_{n+1} \;\leq\; 2m/\,(1+n)\,, \tag{14}$$

where m is the size of the partition underlying $\mathcal{H}_{step}$. □

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

According to the above results, the probability of an incorrect prediction becomes small for large memories, even if the hypotheses h_n are not admissible. In fact, this probability tends toward 0 with a convergence rate of order $O(1/n)$. In a statistical sense, the predictions $C^{est}(x_0)$ can indeed be seen as *credible sets*, 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.

[3] All proofs, omitted here due to reasons of space, can be found in [6].

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 ζ and, hence, to less precise predictions of labels. "Merging" two neighbored intervals A_k and A_{k+1}, for instance, means to define a new hypothesis h with $h|(A_k \cup A_{k+1}) \equiv \min\{\beta_k, \beta_{k+1}\}$.

It is interesting to note that the confidence of a prediction does not depend on the similarity measures $\sigma_{\mathcal{X}}$ and $\sigma_{\mathcal{L}}$. In other words, our method works for *any* pair of such measures. Yet, the similarity measures will strongly influence the *precision* of predictions. Indeed, one cannot expect precise predictions if $\sigma_{\mathcal{X}}$ and $\sigma_{\mathcal{L}}$ are not suitably defined (in which case the IBL hypothesis is hardly satisfied). Therefore, the adaptation of these measures to the application at hand is clearly advised. In this connection, an interesting idea is to take the empirical similarity profile induced by the measures as an indicator of their suitability: Define $\sigma_{\mathcal{X}}$ and $\sigma_{\mathcal{L}}$ such that the induced profile becomes "large" in a certain sense, since large profiles yield precise predictions. This problem will be discussed in more detail in Section 6 below.

Let us finally mention that results similar to the above theorems can also be obtained for the case of *local* similarity profiles [6]. Usually, local profiles yield predictions that are more precise but less confident. This finding can also be grasped intuitively: The level of confidence decreases since one has to learn more similarity profiles from the same amount of data, and the precision increases because local profiles are much more tolerant toward outliers.

5.4 Examples

This section is meant to convey a first idea of the practical performance of our method, without laying claim to providing an exhaustive experimental evaluation. (As an aside, let us note that a comparison with standard IBL, or machine learning methods in general, is difficult anyway. The main reason is that our method provides a different type of prediction, namely credible label sets rather than point-estimations.)

Artificial data. As a first example, let us consider a simple regression problem.[4] More specifically, let the function to be learned be given by the polynomial $x \mapsto x^2$. Moreover, suppose n training examples $\langle x_\imath, \lambda_{x_\imath} \rangle$ to be given, where the $x_\imath$ are uniformly distributed in $[0, 1]$, and the $\lambda_{x_\imath}$ are normally distributed with mean $(x_\imath)^2$ and standard deviation $1/10$. As a similarity measure for both instances (inputs) and labels (outputs) we employ the function $(u, v) \mapsto \exp(-2|u - v|)$. Given a random sample $\mathcal{D}$, we first induce a similarity hypothesis for an underlying equi-width partition of size $m = 5$. Using this hypothesis and the sample $\mathcal{D}$, we derive a prediction λ_x for all instances x (resp. for

[4] Strictly speaking, since our theoretical results above assume a countable instance space, they do not apply to regression proper. They can be generalized to this case, however.

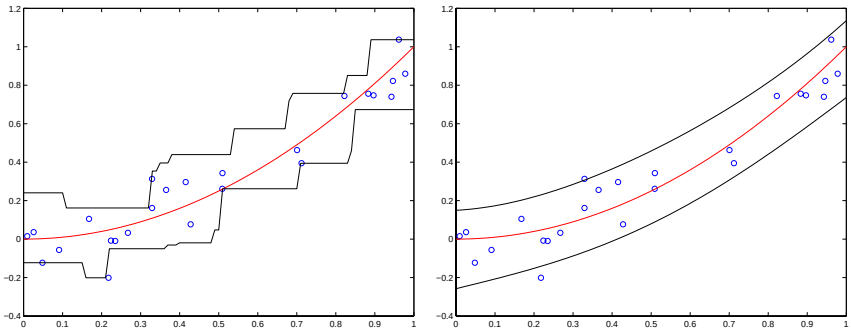

Fig. 3. Approximation of the function $x \mapsto x^2$ in the form of a "confidence band"; left: our instance-based approach, right: linear regression.

the discretization $\{0, 0.01, 0.02, \ldots, 1\}$). Note that such a prediction is simply an interval. Hence, what we obtain is a lower and an upper approximation of the true mapping $x \mapsto x^2$.

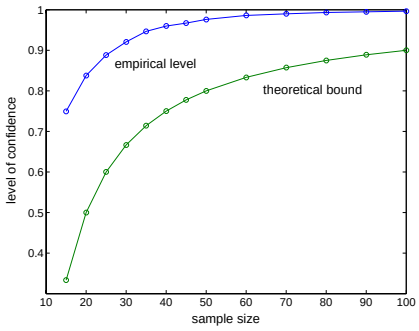

Fig. 4. Confidence levels of predictions: Theoretical bound and empirical level for different sample sizes.

Fig. 3 (left) shows a typical inference result for $n = 25$. According to our estimation (14), 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 (see Fig. 4).

To draw a comparison with standard statistical techniques, Fig. 3 (right) shows the 0.9-confidence band obtained from a linear regression estimation for the same sample. In general, it turned out that linear regression yields slightly more precise predictions. However, in this connection one should realize that this method makes much more assumptions than our instance-based approach. Especially, the type of function to be estimated must be specified in advance: Knowing that this function is a polynomial of degree 2, we estimated the coefficients β_i in the mapping $x \mapsto \beta_0 + \beta_1 x + \beta_2 x^2$ in our example, but usually such

knowledge will not be available (results already become worse when estimating a polynomial of degree $k > 2$). Moreover, the confidence band is valid only if the error terms follow a normal distribution (as they do in our case).

The housing data. We also applied our method to several real-world data sets, not fully discussed here due to reasons of space. For example, in connection with the HOUSING DATABASE,[5] the problem is to predict the price of houses which are characterized by 13 attributes. To apply our method, we simply defined similarity as an affine-linear function of the distance between (real-valued) attribute values (see Section 6 below for the acquisition of such similarity measures). For 30 randomly chosen sample cases we have learned corresponding local similarity hypotheses, using 450 cases as training examples. Using these (local) hypotheses, we derived predictions for the prices of the 56 houses that remain of the complete data. The precision of the predictions was approximately 10,000 dollars with a confidence level of 0.85. Taking the center of an interval as a point-estimation, one thus obtains predictions of the form $x \pm 5,000$ dollars. As can be seen, these estimations are quite reliable but not extremely precise (the average price of a house is approximately 22,500 dollars). In fact, this example clearly points out the practical limits of an inference scheme built upon the IBL assumption: A similarity-based prediction of prices cannot be confident and extremely precise at the same time, simply because the housing data meets the IBL assumption but moderately. Our approach takes these limits into account and makes them explicit.

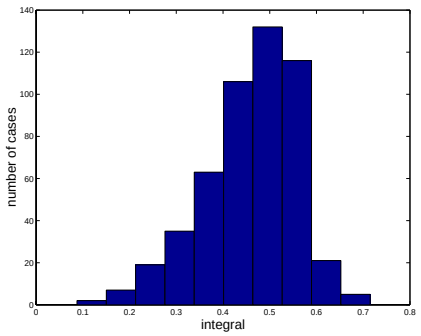

Fig. 5. Distribution of the quality of cases for the housing data, measured in terms of the integral of similarity profiles.

In connection with the housing data, let us recall that a local similarity profile can serve as an indicator of the "quality" of a case. For example, suppose that we measure this quality by means of the integral of the profile (which is easy to compute since the latter is a step-function, see Section 6 below). Fig. 5 shows the distribution of this quality measure for the housing data in the form of a histogram. As can be seen, there are a few cases of rather high

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

quality. The corresponding houses are "typical" in the sense that their prices are representative of the prices for similar houses, and deriving predictions based on these cases will usually be better than gathering a case base at random.

6 Adaptation of Similarity Measures

As already mentioned above, an intuitively reasonable principle for adapting the similarity measures $\sigma_{\mathcal{X}}$ and $\sigma_{\mathcal{L}}$ is to define these functions such that the induced (empirical) similarity profile becomes "large" in some sense. Of course, in order to realize this idea one first has to clarify the meaning of "large".

Recall that empirical profiles are specified as step functions for a given partition of $[0,1]$. This partition is determined through $m+1$ points $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_m = 1$. Since no complete order is defined on the class of step functions $\mathcal{H}_{step}$ in a natural way, such an order has to be imposed somehow. For example, one possibility is to associate with each function h, specified through coefficients $\beta_1, \ldots, \beta_m$, its integral. We thus obtain the optimization criterion

$$I(h) =_{\text{def}} \sum_{k=1}^{m} (\alpha_k - \alpha_{k-1}) \cdot \beta_k \quad \rightarrow \quad \max \tag{15}$$

Instead of the width $\alpha_k - \alpha_{k-1}$ of the interval $A_k = [\alpha_{k-1}, \alpha_k)$ other weights could be used as well. For instance, a reasonable idea is to weigh β_k by the probability that the similarity between two instances lies in the interval A_k. This probability can be estimated from the sample $\mathcal{D}$ by the corresponding relative frequency.

Needless to say, a suitable method for adapting similarity measures can be developed only on the basis of some assumptions concerning the structure of these measures. Here, we proceed from the following assumption which is often satisfied in practice: Instances x are characterized by means of a fixed number of attribute values, and the similarity $\sigma_{\mathcal{X}}$ is a convex combination of individual similarity measures defined for the different attributes. The same assumption is made for the measure $\sigma_{\mathcal{L}}$:

$$\sigma_{\mathcal{X}} = v_1 \sigma_{\mathcal{X}}^1 + v_2 \sigma_{\mathcal{X}}^2 + \ldots + v_p \sigma_{\mathcal{X}}^p,$$
$$\sigma_{\mathcal{L}} = w_1 \sigma_{\mathcal{L}}^1 + w_2 \sigma_{\mathcal{L}}^2 + \ldots + w_q \sigma_{\mathcal{L}}^q.$$

The task of adapting $\sigma_{\mathcal{X}}$ and $\sigma_{\mathcal{L}}$ can then be specified as determining the coefficients $v_\imath$ and $w_\jmath$ in an optimal way.

Now, consider a sample $\mathcal{D}$ consisting of n cases $\langle x_\imath, \lambda_{x_\imath} \rangle$. We denote by $\alpha_{\imath\jmath}^k = \sigma_{\mathcal{X}}^k(x_\imath, x_\jmath)$ the similarity degree between the kth attribute values of $x_\imath$ and $x_\jmath$. Likewise, $\beta_{\imath\jmath}^k = \sigma_{\mathcal{L}}^k(\lambda_{x_\imath}, \lambda_{x_\jmath})$ denotes the similarity degree between the kth attribute values of the labels $\lambda_{x_\imath}$ and $\lambda_{x_\jmath}$.

For the time being, suppose the measure $\sigma_{\mathcal{X}}$ to be given. The optimal adaptation of $\sigma_{\mathcal{L}}$ can then be formulated as a linear optimization problem: Choose $\beta_1, \ldots, \beta_m$ and the coefficients $w_1, \ldots, w_q$ so as to maximize (15) subject to the

constraints

$$\beta_{\kappa(x_\imath,x_\jmath)} \leq w_1\beta^1_{\imath\jmath} + w_2\beta^2_{\imath\jmath} + \ldots + w_q\beta^q_{\imath\jmath}, \qquad (1 \leq \imath, \jmath \leq n)$$

$$w_1 + w_2 + \ldots + w_q = 1$$

$$w_1 \geq 0, \ldots, w_q \geq 0$$

Again, the index $\kappa(x_\imath, x_\jmath)$ specifies the interval of the underlying partition that covers the similarity degree between $x_\imath$ and $x_\jmath$: $\sigma_{\mathcal{X}}(x_\imath, x_\jmath) \in A_{\kappa(x_\imath,x_\jmath)}$. This coefficient must be known for writing down the linear inequalities above, which is the main reason why $\sigma_{\mathcal{X}}$ and $\sigma_{\mathcal{L}}$ cannot be optimized simultaneously. As can be seen, however, an optimal $\sigma_{\mathcal{L}}$ can be found in a quite efficient way once $\sigma_{\mathcal{X}}$ is given.[6] This suggests an optimization procedure in which the adaptation of $\sigma_{\mathcal{L}}$ is embedded as a sub-routine. For example, one could apply any local search method that searches the space of similarity measures $\sigma_{\mathcal{X}}$, that is the space of admissible coefficients $v_1, \ldots, v_p$. The quality of a measure $\sigma_{\mathcal{X}}$, e.g. the fitness in genetic algorithms, can then be computed by solving the above linear program, i.e. by deriving the measure $\sigma_{\mathcal{L}}$ that complements $\sigma_{\mathcal{X}}$ in an optimal way.

7 Summary

We have proposed an instance-based learning method that allows for deriving an estimation in the form of a *credible label set* rather than a single label. This set provably covers the true label with high probability. Bearing in mind that the IBL assumption might apply to an application in a limited scope, our inference scheme does not pretend a precision or credibility of instance-based predictions which is actually not justified. At a formal level, uncertainty is expressed by supplementing (set-valued) predictions with a level of confidence.

From a statistical point of view, our method can be seen as a non-parametric approach to estimating confidence regions, which makes it also interesting for statistical inference (cf. the comparison with linear regression in Section 5.4). In [7], an instance-based prediction method has been advocated as an alternative to linear regression techniques. By deriving set-valued instead of point estimations, our approach somehow combines advantages from both methods: Like the instance-based approach it requires less structural assumptions than (parametric) statistical methods. Still, it allows for specifying the uncertainty related to predictions by means of confidence regions.

A main concern in this paper was aimed at the *correctness* of predictions (3). Still, it is also possible to obtain results related to the *precision* of predictions. In [6], for instance, a result similar to the one in [7] has been shown: Provided that the function $x \mapsto \lambda_x$ mapping instances to labels satisfies certain continuity assumptions, it can be approximated to any degree of accuracy. That is, for each $\epsilon > 0$ one can find a finite memory of cases $\mathcal{D}$ such that $\lambda_x \in C^{est}(x)$ for all $x \in \mathcal{X}$ and $\sup_{x\in\mathcal{X}} \mathrm{diam}(C^{est}(x)) < \epsilon$.

[6] Despite its theoretical complexity, linear programming is rather efficient in practice.

Without going into detail, we have proposed the use of *local* similarity profiles in order to overcome the problem that globally admissible hypotheses might be too restrictive for some applications. In this connection, let us also mention a further idea of weakening the concept of globally valid similarity bounds, namely the use of *probabilistic* similarity hypotheses [4].

References

1. D.W. Aha, editor. *Lazy Learning.* Kluwer Academic Publ., 1997.
2. D.W. Aha, D. Kibler, and M.K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.
3. W.W. Cohen, R.E. Schapire, and Y. Singer. Learning to order things. *Journal of Artificial Intelligence Research*, 10, 1999.
4. E. Hüllermeier. Toward a probabilistic formalization of case-based inference. In *Proceedings* IJCAI–99, pages 248–253, Stockholm, Sweden, 1999.
5. E. Hüllermeier. Focusing search by using problem solving experience. In W. Horn, editor, *Proceedings* ECAI–2000*, 14th European Conference on Artificial Intelligence*, pages 55–59, Berlin, Germany, 2000. IOS Press.
6. E. Hüllermeier. Similarity-based inference: Models and applications. Technical Report 00-28 R, IRIT – Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier, October 2000.
7. D. Kibler and D.W. Aha. Instance-based prediction of real-valued attributes. *Computational Intelligence*, 5:51–57, 1989.
8. J.L. Kolodner. *Case-based Reasoning.* Morgan Kaufmann, San Mateo, 1993.
9. T.M. Mitchell. Version spaces: A candidate elimination approach to rule learning. In *Proceedings* IJCAI-77, pages 305–310, 1977.
10. S. Salzberg. A nearest hyperrectangle learning method. *Machine Learning*, 6:251–276, 1991.
11. R. Short and K. Fukunaga. The optimal distance measure for nearest neighbor classification. IEEE *Transactions on Information Theory*, 27:622–627, 1981.
12. 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.
13. B. Smyth and E. Mc Kenna. Building compact competent case-bases. In *Proceedings* ICCBR-99*, 3rd International Conference on Case-Based Reasoning*, pages 329–342, 1999.
14. C. Stanfill and D. Waltz. Toward memory-based reasoning. *Communications of the ACM*, pages 1213–1228, 1986.
15. V.N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, 1998.