On the Representation and Combination of Evidence in Instance-Based Learning

Eyke Hüllermeier[1]

Abstract. In instance-based learning the classification of a novel instance relies upon experience given in the form of similar instances whose labels are already known. Each of these instances can hence be seen as an individual piece of evidence. In this paper, we elaborate on issues concerning the representation and combination of such pieces of evidence. Particularly, we argue that the information provided by similar instances must not be considered as independent. We propose a new inference principle that derives an evidence function specifying the available evidence in favor of each potential label. This principle, which is built upon a probabilistic (random field) model, takes interdependencies between stored instances into account and suggests a generalization of weighted nearest neighbor estimation.

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. 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 [16], that is inference "from specific to specific":[2] 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 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. The major assumption underlying IBL has already been expressed by the philosopher DAVID HUME:[3] "In reality, all arguments from experience are founded on the similarity, which we discover among natural objects, and by which we are induced to expect effects similar to those, which we have found to follow from such objects. ... From causes, which appear *similar*, we expect similar effects. This is the sum of all our experimental conclusions." We shall base our further discussion on the (quite general) *classification* framework where the above assumption translates into the assertion that "similar objects have similar class 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 [7]. 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. Roughly speaking, similar cases should be considered as nothing more than *pieces of evidence*, and the less similar a case, the smaller the associated evidence. Questions concerning the representation and the combination of such pieces of evidence are major topics of this paper.

By way of background, Section 2 gives a concise review of the NEAREST NEIGHBOR principle, which constitutes the core of the family of IBL algorithms. In Sections 3 and 4 we discuss, respectively, alternative approaches to uncertainty (evidence) representation and the problem of interdependence between different pieces of evidence in IBL. In Section 5, an instance-based estimation procedure is introduced, which takes this type of interdependence into account. Rather than simply suggesting a class label for a new instance, this method yields a complete uncertainty measure specifying the available evidence in favor of each potential label.

2 NEAREST NEIGHBOR CLASSIFICATION

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, and $\langle x, \lambda_x \rangle \in \mathcal{X} \times \mathcal{L}$ is called a labeled instance (or a case). In classification, which is the focus of most IBL implementations, $\mathcal{L}$ is a finite (usually small) set comprised of m classes $\{\lambda_1, \ldots, \lambda_m\}$. S denotes a sample that consists of n labeled instances $\langle x_\imath, \lambda_{x_\imath} \rangle$, $1 \leq \imath \leq n$. Finally, a novel instance $x_0 \in \mathcal{X}$ (a query) is given, whose label λ_{x_0} is to be estimated.

The NEAREST NEIGHBOR (NN) principle, which originated in the field of pattern recognition [3], prescribes to estimate the label of the yet unclassified query x_0 by the label of the nearest (most similar) sample instance. The k-NEAREST NEIGHBOR (k-NN) approach is a slight generalization, which takes the $k \geq 1$ nearest neighbors of x_0 into account. That is, an estimation $\lambda_{x_0}^{est}$ of λ_{x_0} is derived from the set $\mathcal{N}_k(x_0)$ of the k nearest neighbors of x_0, usually by means of a

[1] Department of Mathematics and Computer Science, University of Marburg, Germany (eyke@mathematik.uni-marburg.de)

[2] Though IBL yields complete concept descriptions when being applied to all elements of an instance space.

[3] See e.g. [9], page 116.

majority vote:

$$\lambda_{x_0}^{est} = \arg\max_{\lambda \in \mathcal{L}} \mathrm{card}\{x \in \mathcal{N}_k(x_0) \,|\, \lambda_x = \lambda\}. \quad (1)$$

Several conceptual modifications and extensions of the (k)NN principle have been devised, such as distance weighting [6]:

$$\lambda_{x_0}^{est} = \arg\max_{\lambda \in \mathcal{L}} \sum_{x \in \mathcal{N}_k(x_0) : \lambda_x = \lambda} \omega_x, \quad (2)$$

where ω_x is the weight of the instance x. The latter is usually an increasing function of $\sigma_{\mathcal{X}}(x, x_0)$. Note that ω_x can be 0, and that (1) is a special case of (2). When proceeding from (2) – as we shall subsequently do – one can therefore assume $k = n$ without loss of generality.

3 EVIDENCE REPRESENTATION IN IBL

The NN principle merely provides a point-estimation or, say, a decision rule. In order to represent the uncertainty related to a decision, one possibility is to derive a *probability distribution* over $\mathcal{L}$. In fact, this is a quite obvious idea since NN techniques have originally been employed in the context of non-parametric density estimation [11]. Using this type of estimation, which assumes a related statistical setting of the classification problem (with continuous instance space), the following probability distribution can be deduced:

$$p_{x_0} : \lambda \mapsto k^{-1} \cdot \mathrm{card}\left\{x \in \mathcal{N}_k(x_0) \,|\, \lambda_x = \lambda\right\}. \quad (3)$$

As can be seen, the label estimated by the (majority vote) k-NN rule is just the one of maximal (posterior) probability. Still, one should be cautious with the distribution (3). Particularly, it is not clear how reliable the estimated probabilities $p_{x_0}(\lambda)$ actually are. It is possible to construct corresponding confidence intervals, but these are only asymptotically valid. In fact, k is generally small and, hence, (3) not very confident.[4] Apart from that, the underlying NN (density) estimation techniques suffer from some further difficulties.[5]

3.1 Representation of Ignorance

At least two different types of uncertainty can occur in IBL, namely *ambiguity* and *ignorance*. To illustrate, Fig. 1 shows two classification problems (similarity is inversely related to Euclidean distance). The novel instance x_0 is represented by a cross, and black and light circles correspond to instances of two different classes, respectively. In both cases, the k-NN rule (1) with $k = 5$ suggests `black` as a label for x_0. As can be seen, however, this classification is everything but reliable: In the above setting, the proportion of black and light examples is almost balanced (apart from that, the closest points are light). This is a situation of *ambiguity*. The setting below illustrates a problem of *ignorance*: It is true that all neighbors are black, but even the closest among them are actually quite distant.

The probabilistic approach (3) can handle the problem of *ambiguity*. However, a probability measure is not really able to represent *ignorance*, due to the fact that probability degrees always add up to 1. In the second situation in Fig. 1, for instance, (3) yields $p_{x_0}(\texttt{black}) = 1$ and $p_{x_0}(\texttt{light}) = 0$. In other words, the derived probability distribution suggests that the unknown label is `black` with certainty! More generally, this prediction merely reveals that all objects observed so far do have black labels. Still, it provides neither an indication of the actual number of observed objects, nor of the resemblance of these objects to x_0.

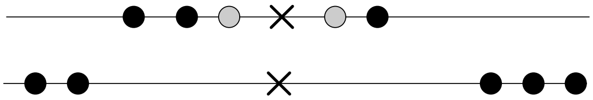

Figure 1. Two situations of uncertainty in connection with the basic k-NN rule, caused by the existence of more than one frequent class label among the nearest neighbors (above) and the absence of any near neighbor (below).

Needless to say, the ability to represent ignorance is quite important from a knowledge representational point of view. In probability theory, the uniform measure is often advocated as a model of complete ignorance, justified by a so-called *principle of insufficient reason*. However, since that measure does only display the *relative* but not the *absolute* amount of available information, this solution appears rather questionable. For example, telling a patient that your experience does not allow any statement concerning his prospect of survival is very different from telling him that his chance is $1/2$.

3.2 Evidence Theory

The above problems motivate the use of alternative, more expressive uncertainty frameworks, such as models based on *belief functions* [13].[6] Belief functions or, equivalently, the dual concept of *plausibility functions*, can be introduced in a simple and intelligible way through the concept of a *basic belief assignment*, also called a mass function.

Let Ω be a finite set and let $\omega_0 \in \Omega$ be unknown. Partial knowledge about ω_0 can then be modeled by means of a basic belief assignment $m : 2^{\Omega} \rightarrow [0, 1]$, where $m(\emptyset) = 0$ and $\sum_{A \subseteq \Omega} m(A) = 1$. The subsets $A \subseteq \Omega$ with $m(A) > 0$ are called *focal sets*. The value $m(A)$ corresponds to the specific support that can be assigned to A (i.e. to the event $\omega_0 \in A$) on the basis of the given evidence, but not to any proper subset $B \subsetneq A$.

A situation of complete ignorance is adequately captured by the mass function m with $m(\Omega) = 1$ and $m(A) = 0$ for all $A \subsetneq \Omega$. Perfect knowledge corresponds to the case where $m(\{\omega\}) = 1$ for some $\omega \in \Omega$ and $m(A) = 0$ for all $A \neq \{\omega\}$. If all focal sets are singletons, then m is actually equivalent to a probability distribution.

A belief function $\mathsf{Bel} : 2^{\Omega} \rightarrow [0, 1]$ is a normalized, non-additive uncertainty measure. An underlying mass function m induces the belief function $A \mapsto \sum_{B \subseteq A} m(B)$. The related plausibility function $\mathsf{Pl} : 2^{\Omega} \rightarrow [0, 1]$ is defined as $A \mapsto \sum_{B \cap A \neq \emptyset} m(B)$. $\mathsf{Bel}(A)$ is the mass *necessarily* covered by A, i.e. the *guaranteed* support of A. $\mathsf{Pl}(A)$ is the *potential* support, namely the mass that would be covered by A if the masses in all sets B with $B \cap A \neq \emptyset$ were shifted to $B \cap A$ (on the basis of more specific information). The belief and plausibility function are related through the following equation: $\mathsf{Pl}(A) = 1 - \mathsf{Bel}(\Omega \setminus A)$. That is, a set A is plausible in so far as its complement is not guaranteed.

3.3 Evidence Theory in IBL

The application of evidence theory in instance-based learning has been advocated in [4] and [8]. In both approaches, the basic idea is

[4] An estimated probability is always a multiplicity of $1/k$. Particularly, $p_{x_0}(\lambda) \in \{0, 1\}$ in the special case $k = 1$, i.e. for the 1-NN rule.

[5] For example, the estimation of a density function is generally not normalized.

[6] Indeed, several frameworks based on belief functions do coexist.

to model "instance-based" evidence by means of a plausibility (belief) function defined over the set $\mathcal{L}$ of labels: Let x_0 be a query and let $\langle x_\imath, \lambda_{x_\imath} \rangle$ be an observed case. In [4], the evidence that comes from this case (in favor of label $\lambda_{x_\imath}$) is modeled by the basic belief assignment $m_{x_0} : 2^{\mathcal{L}} \rightarrow [0,1]$ with

$$m_{x_0}(\{\lambda_{x_\imath}\}) = 1 - \delta, \quad m_{x_0}(\Omega) = \delta, \tag{4}$$

where $\delta \in [0,1]$ is inversely related to the similarity between x_0 and $x_\imath$: The larger $\sigma_{\mathcal{X}}(x_0, x_\imath)$, the stronger the label $\lambda_{x_\imath}$ is supported. For the induced plausibility function one has

$$\mathsf{Pl}_{x_0}(\{\lambda\}) = \left\{ \begin{array}{lll} 1 & \text{if} & \lambda = \lambda_{x_\imath} \\ \delta & \text{if} & \lambda \neq \lambda_{x_\imath} \end{array} \right. . \tag{5}$$

That is, $\lambda_{x_\imath}$ is fully plausible as a label for x_0. Still, all other labels remain plausible to a certain extent as well, depending on the similarity between $x_\imath$ and x_0. In fact, δ specifies the degree of ignorance expressed by (5). Complete ignorance corresponds to $\delta = 1$ ($x_\imath$ is not at all similar to x_0). Note that the plausibility function (5) reflects *absolute* evidence, which depends on the absolute similarity between x_0 and its neighbors, whereas a probability distribution models *relative* evidence. For example, a probability distribution remains unchanged when doubling the similarities between x_0 and all of its neighbors.

Since the focal sets in (4) are nested, the plausiblity function (5) is actually equivalent to a *possibility distribution* [5]. Again, a possibility distribution is a more flexible concept than a probability distribution, not restricted by a normalization constraint. In fact, the important point in this section is not the specific definition of the measure (5), but rather the insight that specifying experience in the form of a measure of *absolute* evidence appears to be particularly reasonable in IBL, where evidence does not only depend on the *frequency* of observed cases but also on their *closeness* to the query x_0. By an *evidence function* we here simply mean a measure $\eta_{x_0} : \mathcal{L} \rightarrow [0,1]$ such that $\eta_{x_0}(\lambda)$ represents the *absolute* evidence in favor of $\lambda_{x_0} = \lambda$. Particularly, $\eta_{x_0}(\lambda) = 0$ means that no such evidence is available, whereas $\eta_{x_0}(\lambda) = 1$ suggests that enough evidence has been accumulated so as to regard λ as fully possible.

An evidence function can be taken as a point of departure for deciding on the further line of action. For example, on the basis of $\eta_{x_0}(\cdot)$ one might decide whether or not further information should be gathered in order to reduce ambiguity or uncertainty. If not, the maximally supported label might be chosen as an estimation of a class label or, in a problem solving context (where labels correspond to solutions), several well-supported alternatives might be pursued as promising solutions.

4 DEPENDENCE OF EVIDENCE

Some instance-based approaches completely rely on the (supposedly) most relevant piece of evidence, namely the most similar observation. In case-based reasoning, for example, it is common practice to retrieve just the most similar among the stored cases. The representation of evidence is then rather simple and might be realized, for example, by means of a single plausibility function (5).

Ignoring all but the most similar observation is computationally efficient but comes along with a loss of information. If several observations are retrieved, an important question arises: How should the different pieces of evidence be combined? In [4], it is proposed to combine the belief (plausibility) functions induced by different observations by means of DEMPSTER's rule of combination. This yields a new belief function, regarded as a representation of the overall evidence. In [8], this approach has been criticized, since the different pieces of information cannot be assumed to be *distinct* in the sense of [14], as required by DEMPSTER's rule.

Indeed, the independence of instance-based evidence must *not* be taken for granted! On the contrary, the interdependence between pieces of evidence that come from different cases is actually a consequence of the IBL assumption itself: If it is true that similar objects have similar labels, then one should not be surprised to retrieve two objects having a similar label if these objects are similar by themselves.

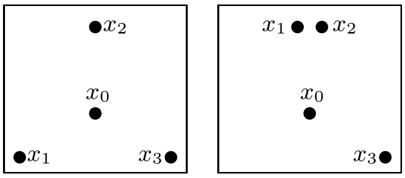

Figure 2. Different configurations of locations in two-dimensional space.

To illustrate this aspect consider the example shown in Fig. 2. Suppose that the instances $x_\imath$ correspond to locations and that $\mathcal{L} = \{\texttt{rainy}, \texttt{sunny}\}$. Moreover, suppose $\lambda_{x_1} = \lambda_{x_2} = \lambda_{x_3} = \texttt{rainy}$. What about the weather in x_0? Clearly, the given information provides evidence for rainy weather in x_0. The important point to notice is that the evidence appears to be larger in the first situation (left), even though the individual similarities $\sigma_{\mathcal{X}}(x_0, x_\imath)$ are the same in both situations. This is due to the different *arrangement* of the neighbors [17]. More generally, it is the similarity among the observed cases themselves which has to be taken into account.[7] For example, consider again Fig. 2 (right) and suppose that we know about the weather in x_1. Since x_2 is very close to x_1, the information $\lambda_{x_2} = \texttt{rainy}$ is then not astonishing. On the contrary, relying upon the IBL assumption, $\lambda_{x_1} = \lambda_{x_2}$ was to be expected! Consequently, the case $\langle x_2, \lambda_{x_2} \rangle$ does hardly provide new evidence. Rather, the first and the second case might even be considered as one *single* piece of information. The situation is completely different in the left picture, where x_1 and x_2 are very dissimilar.

It should be noticed that the above type of interdependence might disappear under certain (additional) statistical assumptions, and is often negligible in asymptotic analyses of NN principles. Still, the example shows that it is of great practical importance for standard IBL applications.

5 A NEW EVIDENCE MODEL

5.1 Random Fields

The mutual dependence between (random) variables in probabilistic models is completely determined by their *joint* probability distribution, but is also (partially) characterized through statistics of that distribution. Important information is provided by the *covariance* between two random variables X and Y:

$$\mathrm{Cov}[X,Y] = \mathrm{E}\left[\,(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\,\right], \tag{6}$$

where $\mathrm{E}[\cdot]$ denotes the expected value operator. If X and Y are independent, then $\mathrm{Cov}[X,Y] = 0$. If $\mathrm{Cov}[X,Y] > 0$, then X and

[7] Reasoning on the basis of the arrangement of neighbors is more involved, especially if $\mathcal{X}$ is a non-metric space.

Y are positively correlated: Roughly speaking, X and Y have a tendency to deviate from their expected values in the same direction. A corresponding (reverse) statement holds for $\mathrm{Cov}[X,Y] < 0$.

Expressing the dependence between variables by means of their covariance is a central idea of so-called *random field* models [15]. A random field is an indexed class $\{X(t)\}_{t \in T}$ of random variables. An index t is also called a *location* and the value $X(t)$ the *state* of the random field at that location. The index set T is generally endowed with a metric. In homogeneous isotropic random fields, the covariance between two random variables $X(t_1)$ and $X(t_2)$ is assumed to be a (decreasing) function of the *distance* between t_1 and t_2:

$$\mathrm{Cov}[X(t_1), X(t_2)] \doteq \gamma\left(\Delta(t_1, t_2)\right). \tag{7}$$

Random fields are apparently very interesting for IBL, especially in connection with the above mentioned problem of evidence combination. In fact, an obvious idea is to interpret t_1 and t_2 in (7) as instances and the related random quantities $X(t_1)$ and $X(t_2)$ as labels. Still, this approach leads to some technical difficulties, especially due to the mathematical structure of a random field. For example, the set $\mathcal{L}$ of labels is generally not a metric space. Consequently, the covariance between two (random) labels is not well-defined. Apart from that, a probabilistic model again suffers from the inability to represent ignorance. In order to avoid these problems, we are now going to employ the random field model in a slightly different way.

5.2 Evidence Estimation

Recall the basic setting introduced at the beginning of Section 2. Particularly, let S be a sample comprised of n labeled instances $\langle x_\imath, \lambda_{x_\imath} \rangle$ and let $x_0 \in \mathcal{X}$ be a query whose label λ_{x_0} is to be estimated. For all instances $x \in \mathcal{X}$, let

$$\theta_x \doteq \begin{cases} 1 & \text{if} \quad \lambda_x = \lambda \\ 0 & \text{if} \quad \lambda_x \neq \lambda \end{cases}, \tag{8}$$

where $\lambda \in \mathcal{L}$ is a fixed label. On the basis of the "auxiliary" labeling (8), the set $\mathcal{X}$ is partitioned into two new classes: Those instances with label λ and those instances with a different label. Note that the set $\mathcal{L}' = \{0,1\}$ of auxiliary labels has a trivial metric structure, determined by the distance function Δ with $\Delta(0,1) = \Delta(1,0) = 1$ and $\Delta(1,1) = \Delta(0,0) = 0$. This is in agreement with the usual classification framework, where $\mathcal{L}$ is a nominal scale.

The IBL assumption is now expressed in terms of the covariance function: For all $x, x' \in \mathcal{X}$, we postulate the equality

$$\mathrm{Cov}[\theta_x, \theta_{x'}] = \gamma\left(\sigma_{\mathcal{X}}(x, x')\right), \tag{9}$$

where γ is a non-decreasing function $[0,1] \rightarrow \mathfrak{R}$. That is, $\{\theta_x\}_{x \in \mathcal{X}}$ is considered as a special type of random field, namely a *binary* random field, also called a *random partition of space*. More specifically, γ should satisfy $\gamma(0) = 0$, thereby expressing that labels of completely dissimilar instances are unrelated.

The theory of random fields offers a large repertoire of statistical prediction methods. Within our context, these methods can be used for estimating a label θ_{x_0}, given the labels $\theta_{x_\imath}$ of the sample instances $x_\imath$ [10]. Here, we restrict ourselves to the simplest approach, namely *linear* estimation theory. A linear estimator of θ_{x_0} is a function which is linear in the observations $\theta_{x_1}, \ldots, \theta_{x_n}$. It can be shown that the following estimator is optimal among all linear estimators in the sense that it minimizes the mean squared error $\mathrm{E}[(\theta^{est}_{x_0} - \theta_{x_0})^2]$:

$$\theta^{est}_{x_0} = \mathrm{E}(\theta_{x_0}) + \sum_{\imath=1}^{n} \alpha_\imath \, \theta_{x_\imath}, \tag{10}$$

where the vector $\alpha = (\alpha_1, \ldots, \alpha_n)^\top$ of coefficients is defined as $\alpha = C^{-1} \cdot c$. Here, C is the $n \times n$ covariance matrix with entries[8]

$$C_{\imath\jmath} = \mathrm{Cov}[\theta_{x_\imath}, \theta_{x_\jmath}] = \gamma\left(\sigma_{\mathcal{X}}(x_\imath, x_\jmath)\right), \tag{11}$$

and $c = (c_1, \ldots, c_n)^\top$ is an $n \times 1$ vector with

$$c_\imath = \mathrm{Cov}[\theta_{x_0}, \theta_{x_\imath}] = \gamma\left(\sigma_{\mathcal{X}}(x_0, x_\imath)\right). \tag{12}$$

The term $\mathrm{E}(\theta_{x_0})$ in (10) corresponds to the prior probability of $\theta_{x_0} = 1$, whereas $\theta^{est}_{x_0}$ is an estimation of the posterior probability [12].

Now, we are not interested in an estimation of the probability of $\theta_{x_0} = 1$, but rather in a quantification of the available *evidence* in favor of the label λ. To this end, we modify the estimation (10) in two ways. Firstly, the term $\mathrm{E}(\theta_{x_0})$ will be used for representing "prior evidence" rather than prior probability. If no prior information is available, this means that $\mathrm{E}(\theta_{x_0}) = 0$ rather than $1/m$. This way, it becomes possible to represent ignorance.

Secondly, we derive an estimation for λ not on the basis of the complete sample S, but rather on the basis of the subset $S_\lambda \doteq \{\langle x, \lambda_x \rangle \in S \,|\, \lambda_x = \lambda, \sigma_{\mathcal{X}}(x, x_0) > 0\}$. This way, the *absolute* rather than the relative evidence is measured, and we arrive at the following evidence function $\eta_{x_0} : \mathcal{L} \rightarrow [0,1]$:

$$\eta_{x_0}(\lambda) = \eta_{x_0}(\lambda \,|\, S) \doteq \sum_{\imath=1}^{|S_\lambda|} \alpha_\imath \tag{13}$$

for each label $\lambda \in \mathcal{L}$, where $\alpha = C_\lambda^{-1} \cdot c_\lambda$. Here, C_λ and c_λ denote, respectively, the matrix (11) and the vector (12) restricted to the observations in S_λ. By definition, $\eta_{x_0}(\lambda) \doteq 0$ if $S_\lambda = \emptyset$.

A comparison with (2) shows that (13) can be considered as a weighted NN estimation. Now, however, the weight $\alpha_\imath$ of an instance $x_\imath$ is not determined by the similarities $\sigma_{\mathcal{X}}(x_0, x_\jmath)$ alone. Rather, this weight also takes the similarities among the individual pieces of evidence into account. A simple distance-weighted scheme is obtained for the special case where these pieces are completely independent (C_λ is the unit matrix). We call the decision function

$$(S, x_0) \mapsto \arg\max_{\lambda \in \mathcal{L}} \eta_{x_0}(\lambda \,|\, S) \tag{14}$$

the DNN (DEPENDENT NEAREST NEIGHBOR) classifier. The k-DNN classifier is defined by replacing the set S on the right-hand side of (14) by the set $\mathcal{N}_k(x_0)$ of x_0's k nearest neighbors.

To illustrate, let S contain three instances, x_1, x_2, x_3, with label λ. Let

$$C_\lambda = \begin{pmatrix} 1 & \alpha & 0 \\ \alpha & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad c_\lambda = \begin{pmatrix} \beta \\ \beta \\ \beta \end{pmatrix}$$

with $0 \leq \alpha, \beta < 1$. That is, x_1 and x_2 are similar to some extent, but x_3 resembles neither x_1 nor x_2. Moreover, all three instances resemble the query x_0 to the same degree. According to (13), x_3 supports λ to the degree β, whereas the corresponding supports through x_1 and x_2 are discounted by $1 + \alpha$:

$$\eta_{x_0}(\lambda) = \frac{\beta}{1+\alpha} + \frac{\beta}{1+\alpha} + \beta.$$

Particularly, $\eta_{x_0}(\lambda) = 3\beta$ for $\alpha = 0$ and $\eta_{x_0}(\lambda) \rightarrow 2\beta$ for $\alpha \rightarrow 1$ (compare with the two situations shown in Fig. 2).

[8] The definition of γ must guarantee that the matrix C is positive definite.

5.3 Alternative Estimation Methods

The linear estimator (10) is a simple yet efficient estimation procedure with nice statistical properties. However, as one drawback of (10) let us mention that $\theta^{est}_{x_0}$ can fall outside the scope of the observed values $\theta_{x_1}, \ldots, \theta_{x_n}$. In our case, this means that $\eta_{x_0}(\lambda) \leq 1$ is not guaranteed for the evidence measure (13). A “practical” step is to simply truncate the corresponding value whenever $\eta_{x_0}(\lambda) > 1$. This step can be seen as a “rough” approximation to a saturation effect anyway induced by the probabilistic model: The more evidence has already been accumulated, the smaller the absolute increase in evidence due to the observation of a new case will be. (Besides, it should be noticed that $\eta_{x_0}(\lambda) \leq 1$ is actually not required in (14).)

Without going into detail, let us mention that the above defect can also be overcome by more sophisticated models or estimation procedures. For example, a main reason for the aforementioned problem is that $\{\theta_x\}_{x\in\mathcal{X}}$ is a *binary* random field. For such fields, not all covariance functions are actually feasible. Therefore, one might work with an underlying *real* random field $\{Y_x\}_{x\in\mathcal{X}}$, where the Y_x are, for example, Gaussian random variables. The binary values θ_x can then be derived from these variables by means of a threshold function. This approach allows for a direct calculation of (posterior) probability degrees and, hence, guarantees $\eta_{x_0}(\lambda) \leq 1$.

Apart from the fact that (13) is a measure of *absolute* uncertainty, the separate treatment of labels has further advantages in IBL. For example, consider two similar instances $x_\imath$ and $x_\jmath$ with different labels $\lambda_{x_\imath}$ and $\lambda_{x_\jmath}$. The original probabilistic model would simply “average” between these labels: Roughly speaking, the two cases would wipe out each other, a solution that may lead to undesirable effects and counterintuitive results. For a thorough analysis of formal properties of the uncertainty measure (13), refer to an extended version of this paper.

5.4 Experimental Results

In order to validate our extension of the NN principle, we have performed a large number of experimental studies, using data sets from the UCI repository.It has to be mentioned, however, that these data sets are not optimally suited for our purpose. Firstly, many data sets are large and random sampling leads most probably to more or less “balanced” situations where the neglect of dependence is not as harmful. Secondly, an important aspect of our approach is the faithful representation of the available evidence in favor of the different labels. But this aspect is completely neglected if – as in experimental studies – only the correctness of the final decision (classification accuracy) counts. In fact, the consideration of dependence will often change the *distribution* (13) but not necessarily the *decision* (14).

Anyway, all in all the experiments show that – on average – k-DNN even leads to slightly improved classification performance. The following table shows the (average) percentage of correct classifications for some well-known data sets when using, respectively, one half of the cases as training and one half as test cases. Here, we compared k-DNN (with $\gamma(x) \equiv x$) to a simple weighted k-NN scheme and the original (majority vote) k-NN classifier (with $k = 5$):

data set	k-DNN	weighted k-NN	k-NN
GLASS	64,67	64,01	62,80
WINE	72,58	71,91	72,35
ABALONE	85,27	84,66	84,05
IRIS	96,53	96,53	96,40
PIMA DIABETES	72,10	71,95	71,86
BALANCE SCALE	85,90	85,32	85,10

Many other data sets yield qualitatively similar results. Due to lack of space, however, we refrain from a detailed exposition.

6 Concluding Remarks

We have emphasized two points as important aspects of instance-based learning: Firstly, an estimation should reflect the available evidence supporting the different labels, rather than simply return the final decision in the form of the apparently most probable label. Particularly, we have argued in favor of a measure of *absolute* evidence that is able to represent (partial) ignorance. Secondly, observed cases in IBL must *not* be considered as independent pieces of evidence. In fact, the necessity of taking the mutual dependencies between observed cases into account is a consequence of the underlying IBL principle itself.

These considerations have given birth to a new inference scheme which is built upon a probabilistic random field model. Our estimation procedure can be seen as an extension of weighted NN estimation. Roughly speaking, it shows how to discount the information provided by neighbored and, hence, non-independent cases so as to modulate their influence on the new estimation in a proper way.

The covariance model (9) captures the essence of the basic IBL assumption: The more similar two instances, the more likely they have similar labels. The function γ “parameterizes” this principle. Our experiments have shown that estimations are quite robust toward variations of this function. Still, the problem of how to specify (or learn) γ so as to achieve optimal performance for the application at hand leads to an interesting technical question. Apart from that, a main challange for future work is to integrate our inference scheme into real-world IBL and CBR systems.

REFERENCES

[1] *Lazy Learning*, ed., D.W. Aha, 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] *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, ed., B.V. Dasarathy, IEEE Computer Society Press, 1991.

[4] T. Denoeux, ‘A k-nearest neighbor classification rule based on Dempster-Shafer Theory’, IEEE *Transactions on Systems, Man, and Cybernetics*, **25**(5), 804–813, (1995).

[5] D. Dubois and H. Prade, *Possibility Theory*, Plenum Press, 1988.

[6] S.A. Dudani, ‘The distance-weighted k-nearest-neighbor rule’, IEEE *Trans. Systems, Man, and Cybernetics*, **SMC-6**(4), 325–327, (1976).

[7] E. Hüllermeier, ‘Toward a probabilistic formalization of case-based inference’, IJCAI–99, pp. 248–253, (1999).

[8] E. Hüllermeier, ‘Similarity-based inference as evidential reasoning’, ECAI–2000, pp. 50–54, (2000).

[9] D. Hume, *An Enquiry concerning Human Understanding*, Oxford University Press Inc., New York, (1999).

[10] M. Lindenbaum, S. Marcovich, and D. Rusakov, ‘Selective sampling for nearest neighbor classifiers’, AAAI-99, pp. 366–371, (1999).

[11] D.O. Loftsgaarden and C.P. Quesenberry, ‘A nonparametric estimate of a multivariate density function’, *Annals of Mathematical Statistics*, **36**, 1049–1051, (1965).

[12] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, Inc., 1991.

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

[14] P. Smets, ‘The concept of distinct evidence’, in *Proceedings* IPMU-92, pp. 789–794, Palma de Mallorca, Spain, (1992).

[15] E. Vanmarcke, *Random Fields: Analysis and Synthesis*, MIT Press, Cambridge, Massachusetts, 1983.

[16] V.N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.

[17] J. Zhang, Y. Yim, and J. Yang, ‘Intelligent selection of instances for prediction in lazy learning algorithms’, *Artificial Intelligence Review*, **11**, 175–191, (1997).