Cho-k-NN: A Method for Combining Interacting Pieces of Evidence in Case-Based Learning

Eyke Hüllermeier
Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik
Universitätsplatz 2, 39106 Magdeburg, Germany
eyke.huellermeier@iti.cs.uni-magdeburg.de

Abstract

The case-based learning paradigm relies upon memorizing cases in the form of successful problem solving experience, such as e.g. a pattern along with its classification in pattern recognition or a problem along with a solution in case-based reasoning. When it comes to solving a new problem, each of these cases serves as an individual piece of evidence that gives an indication of the solution to that problem. In this paper, we elaborate on issues concerning the proper combination (aggregation) of such pieces of evidence. Particularly, we argue that cases retrieved from a case library must not be considered as independent information sources, as implicitly done by most case-based learning methods. Focusing on the problem of prediction as a performance task, we propose a new inference principle that combines potentially interacting pieces of evidence by means of the so-called (discrete) Choquet-integral. Our method, called Cho-k-NN, takes interdependencies between the stored cases into account and can be seen as a generalization of weighted nearest neighbor estimation.

1 Introduction

Case-based or *instance-based learning* algorithms have been applied successfully in fields such as, e.g., machine learning and pattern recognition during the recent years [Aha *et al.*, 1991; Dasarathy, 1991]. The case-based learning (CBL) paradigm is also of central importance in case-based reasoning (CBR), a problem solving methodology which goes beyond the standard prediction problems of classification and regression [Riesbeck and Schank, 1989; Kolodner, 1993].

As the term suggests, in CBL special importance is attached to the concept of a *case*. A case or an instance 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 CBR.

Rather than inducing a global model (theory) from the data and using this model for further reasoning, as inductive, model-based machine learning methods typically do, CBL systems simply store the data itself. The processing of the data is deferred until a prediction (or some other type of query) is actually requested, a property which qualifies CBL as a *lazy* learning method [Aha, 1997]. Predictions are then derived by combining the information provided by the stored cases, primarily by those which are *similar* to the new query.

In fact, the concept of similarity plays a central role in CBL. The major assumption underlying CBL has already been expressed, amongst others, by the philosopher DAVID HUME ([Hume, 1999], page 116): "In reality, all arguments from experience are founded on the similarity ... among natural objects. ... From causes, which appear *similar*, we expect similar effects." This commonsense principle, that we shall occasionally call the "similarity hypothesis" [Rendell, 1986], serves as a basic inference paradigm in various domains of application. For example, in a classification context, it translates into the assertion that "similar objects have similar class labels", and in CBR it suggests that "similar problems have similar solutions".

The similarity hypothesis, which is apparently of a heuristic nature, has been put into practice by means of different inference schemes that combine similarity and frequency information in one way or the other. For example, the well-known nearest neighbor (NN) classifier first selects a neighborhood around the query, consisting of the k most similar cases, and then counts the occurrence of the different class labels. Despite of their simplicity, inference methods of such kind have proved to be quite successful in practical applications.

In this paper, we suggest a further improvement based on the idea of taking interdependencies between neighbored cases into account. In fact, in standard NN methods, these cases are implicitly considered as *independent* information sources. We argue that a corresponding assumption of independence is not always justified and propose a new inference principle that combines *interacting* pieces of evidence in a more thorough way. This principle, which makes use of non-additive measures for modeling interaction between cases and employs the so-called (discrete) Choquet-integral as an aggregation operator, can be seen as a generalization of weighted NN estimation.

By way of background, section 2 gives a concise review of the NN principle, which constitutes the core of the family of CBL algorithms. In section 3, we discuss the problem of interaction between cases in CBL. A new approach to CBL, which takes such interactions into account, is then introduced in section 4 and evaluated empirically in section 5.

2 Nearest Neighbor Estimation

The well-known nearest neighbor (NN) estimation principle is applicable to both classification problems (prediction of class labels) and regression (prediction of numeric values).

Consider a setting in which an instance space $\mathcal{X}$ is endowed with a similarity measure $\mathrm{sim} : \mathcal{X} \times \mathcal{X} \to [0, 1]$. An instance corresponds to the description x of an object (usually in attribute–value form). In the standard classification framework, each instance x is assumed to have a (unique) label $y \in \mathcal{L}$. Here, $\mathcal{L}$ is a finite (typically small) set comprised of m class labels $\{\ell_1 \dots \ell_m\}$, and $\langle x, y\rangle \in \mathcal{X} \times \mathcal{L}$ is a labeled instance (case).

The NN principle originated in the field of pattern recognition [Dasarathy, 1991]. Given a sample S consisting of n labeled instances $\langle x_\imath, y_\imath\rangle$, $1 \leq \imath \leq n$, and a novel instance $x_0 \in \mathcal{X}$ (a query), this principle prescribes to estimate the label y_0 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 y_0^{est} of y_0 is derived from the set $\mathcal{N}_k(x_0)$ of the k nearest neighbors of x_0, usually by means of a *majority vote*. Besides, further conceptual extensions of the (k-)NN principle have been devised, such as distance weighting [Dudani, 1976]:

$$y_0^{est} = \arg\max_{\ell \in \mathcal{L}} \sum_{\langle x,y\rangle \in \mathcal{N}_k(x_0)} \omega_x \cdot \mathbb{I}(y = \ell) \qquad (1)$$

where ω_x is the weight of the instance x and $\mathbb{I}(\cdot)$ the standard $\{\text{true}, \text{false}\} \to \{0, 1\}$ mapping. (Throughout the paper, we assume the weights to be given by $\omega_x = \mathrm{sim}(x, x_0)$.)

The NN principle can also be used for regression problems, i.e., for realizing a (locally weighted) approximation of real-valued target functions $x \mapsto y = f(x)$ (in this case, $\mathcal{L} = \mathbb{R}$). To this end, one reasonably computes the (weighted) mean of the k nearest neighbors of a new query point:

$$y_0^{est} = \frac{\sum_{\langle x,y\rangle \in \mathcal{N}_k(x_0)} \omega_x \cdot y}{\sum_{\langle x,y\rangle \in \mathcal{N}_k(x_0)} \omega_x} \qquad (2)$$

3 Interaction Between Cases in CBL

Some case-based approaches completely rely on the (supposedly) most relevant piece of evidence, namely the observation which is most similar to the query. In CBR, for example, it is common practice to retrieve just a single case from the case library, and to adapt the corresponding solution to the problem under consideration. On the one hand, ignoring all but the most similar observation is conceptually simple and computationally efficient. On the other hand, this approach does of course come along with a loss of information, because only a very small part of the past experience is utilized.

If several instead of only a single case are retrieved, as e.g. in k-NN, an important question arises: How should the pieces of evidence coming from the different cases be combined? In k-NN classification, the evidences in favor of a certain class label are simply added up (see (1)). Likewise, in regression the estimation is a simple linear combination of the observed outcomes (see (2)). Thus, the neighbored cases are basically considered as *independent* information sources.

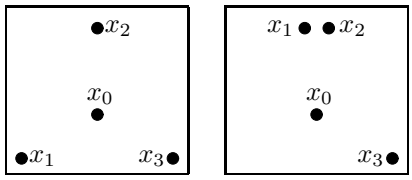

Figure 1: Different configurations of locations in two-dimensional space.

This assumption of independence between case-based evidence can thoroughly be called into question [Hüllermeier, 2002]. Indeed, it is not even in agreement with the similarity hypothesis itself! Namely, if this hypothesis is true, then two neighbored cases that are not only similar to the query case *but also similar among each other* will probably provide similar information regarding the query. In other words, when taking the similarity hypothesis for granted, the information coming from the neighbored cases is at least not independent.

In particular, from a problem solving perspective, one should realize that a set of cases can be *complementary* in the sense that the experiences represented by the individual cases complement or reinforce each other. On the other hand, cases can also be *redundant* in the sense that much of the information is already represented by a smaller subset among them. And indeed, as we said before, the similarity hypothesis suggests that similar cases are likely to be redundant.

To illustrate this point by a simple example, consider the problem of predicting student Peter's grade in computer science, knowing that he has an A in French. The latter information (i.e. the case $\langle\text{French}, A\rangle$) clearly suggests that Peter is an excellent student and, hence, supports predicting an A or maybe a B. Yet, one cannot be sure that this prediction is correct. In this situation, the additional information that Peter has a B in mathematics is probably more valuable than the information that he has an A in Spanish as well. In fact, the cases $\langle\text{French}, A\rangle$ and $\langle\text{Spanish}, A\rangle$ are partly redundant, since the two subjects are quite similar by themselves. As opposed to this, the case $\langle\text{mathematics}, B\rangle$ is complementary in a sense, as it suggests that Peter is not only good in languages.

Now, suppose that you know all three grades (mathematics, Spanish, French). Is A or B the more likely grade? Of course, mathematics is more similar to computer science than Spanish and French. However, depending on the concrete specification of similarity degrees between the various subjects, it is quite possible that the weighted k-NN rule favors grade A since the two moderately similar A's compensate for the more similar B. Of course, this result might be judged critically, and one might wonder whether the grade A should really count twice. In fact, one reasonable alternative is to consider the two A's as only a single piece of evidence, telling something about Peter's achievements in languages, instead of two pieces of distinct information. In any case, the close connection between the two A's should not be ignored when combining the three observations.

As a second example, consider the problem of predicting the yearly rainfall at a certain location (city). For instance,

given the rainfall $y_\imath$ at location $x_\imath$ ($\imath = 1, 2, 3$), what about the rainfall at location x_0 in the two scenarios shown in Fig. 1? The important point to notice is that even though the individual distances between x_0 and the $x_\imath$ are the same in both scenarios, the $y_\imath$ should not be combined in the same way. In this example, this is due to the different *arrangements* of the neighbors [Zhang *et al.*, 1997]: Simply predicting the arithmetic mean $(y_1 + y_2 + y_3)/3$ seems to be reasonable in the left scenario, while the same prediction appears questionable in the scenario shown in the right picture. In fact, since x_1 and x_2 are closely neighbored in the latter case, information about the rainfall at these locations will be partly redundant. Consequently, the weight of the (joint) evidence that comes from the observations $\langle x_1, y_1\rangle$ and $\langle x_2, y_2\rangle$ should not be twice as high as the weight of the evidence that comes from $\langle x_3, y_3\rangle$.

The above examples show the need for taking interdependencies between observed cases into account and, hence, provide a motivation for the method that will be proposed in the following section. Before proceeding, let us make two further remarks: First, the above type of interaction between cases seems to be less important if the sample size is large and even becomes negligible in asymptotic analyses of NN principles. In fact, strong results on the performance of NN estimation can be derived [Dasarathy, 1991], but these are valid only under idealized statistical assumptions and arbitrarily large sample sizes. Roughly speaking, if the sample size n tends to infinity, the distance between the query and its nearest neighbors becomes arbitrarily small (with high probability). This holds true even if the size k of the neighborhood is increased too, as a function $k(n)$ of n, provided that $k(n)/n \to 0$ for $n \to \infty$. Moreover, if the individual observations are independent and identically distributed in a statistical sense, the neighborhood becomes "well distributed". Under these assumptions, it is intuitively clear that interdependencies between observations will hardly play any role. On the other hand, it is also clear that statistical assumptions of such kind will almost never be satisfied in practice.

The second remark concerns related work. In fact, there are a few methods that fit into the CBL framework and that allow for taking certain types of interaction between observations into account. Particularly, these are methods that make assumptions on the statistical correlation between observations, depending on their distance [Lindenbaum *et al.*, 1999]. For example, in our rainfall example one could employ a method called *kriging*, which is well-known in geostatistics [Oliver and Webster, 1990]. Usually, however, such methods are specialized on a particular type of application and, moreover, make rather restrictive assumptions on the mathematical (metric) structure of the instance space. Our approach, to be detailed in the next section, is much more general in the sense that it only requires a similarity measure $\mathrm{sim}(\cdot)$ to be given. We do not make any particular assumptions on this measure (such as symmetry or any kind of transitivity), apart from the fact that it should be normalized to the range $[0, 1]$. From an application point of view, this seems to be an important point. In CBR, for example, cases are typically complex objects that cannot easily be embedded into a metric space.

4 The Cho-k-NN Method

4.1 Non-Additive Measures

Let X be a finite set and $\nu(\cdot)$ a measure $2^X \to [0, 1]$. For any $A \subseteq X$, we interpret $\nu(A)$ as the *weight* or, say, the *degree of relevance* of the set of elements A.

A standard assumption on the measure $\nu(\cdot)$, which is made e.g. in probability theory, is additivity: $\nu(A \cup B) = \nu(A) + \nu(B)$ for all $A, B \subseteq X$ s.t. $A \cap B = \emptyset$. Unfortunately, additive measures cannot model any kind of interaction between elements: Extending a set of elements A by a set of elements B always increases the weight $\nu(A)$ by the weight $\nu(B)$, regardless of A and B.

Suppose, for example, that the elements of two sets A and B are *complementary* in a certain sense. To illustrate, one might think of X as a set of workers in a factory, and of a weight $\nu(A)$ as the productivity of a team of workers A. In that case, complementarity means that the output produced by two teams A and B is higher if they cooperate. Formally, this can be expressed as a positive interaction: $\nu(A \cup B) > \nu(A) + \nu(B)$. Likewise, elements can interact in a negative way. For example, if two sets A and B are partly *redundant* or *competitive*, then $\nu(A \cup B) < \nu(A) + \nu(B)$.

The above considerations motivate the use of non-additive measures, also called fuzzy measures, which are simply normalized and monotone [Sugeno, 1974]:

- $\nu(\emptyset) = 0, \nu(X) = 1$
- $\nu(A) \leq \nu(B)$ for $A \subseteq B$

In order to quantify the interaction between subsets of a set X, as induced by a fuzzy measure $\nu(\cdot)$, several indices have been proposed in the literature. For two individual elements $x_\imath, x_\jmath \in X$, the *interaction index* is defined by

$$I_\nu(x_\imath, x_\jmath) \stackrel{\mathrm{df}}{=} \sum_{A \subseteq X \setminus \{x_\imath, x_\jmath\}} \frac{(|X| - |A| - 1)!\,|A|!}{(|X| - 1)!} (\Delta_{\imath\jmath}\nu)(A),$$

where $|A|$ denotes the cardinality of A and

$$(\Delta_{\imath\jmath}\nu)(A) \stackrel{\mathrm{df}}{=} \nu(A \cup \{x_\imath, x_\jmath\}) - \nu(A \cup \{x_\imath\}) - \\ - \nu(A \cup \{x_\jmath\}) + \nu(A).$$

The latter can be seen as the *marginal interaction* between $x_\imath$ and $x_\jmath$ in the context A [Murofushi and Soneda, 1993]. The above index can be extended from pairs $x_\imath, x_\jmath$ to any set $U \subseteq X$ of elements [Grabisch, 1997]:

$$I_\nu(U) \stackrel{\mathrm{df}}{=} \sum_{A \subseteq X \setminus U} \frac{(|X| - |A| - |U|)!\,|A|!}{(|X||U| + 1)!} (\Delta_U \nu)(A),$$

where

$$(\Delta_U \nu)(A) \stackrel{\mathrm{df}}{=} \sum_{V \subseteq U} (-1)^{|U| - |V|} \nu(V \cup A).$$

4.2 Modeling Interaction in Case-Based Learning

Now, recall our actual problem of combining evidence in case-based learning. In this context, the set X of elements corresponds to the neighbors of the query case x_0:

$$X = \mathcal{N}_k(x_0) = \{x_1, x_2 \ldots x_k\} \tag{3}$$

Our comments so far have shown that fuzzy measures can principally be used for modeling interaction between cases. The basic question that we have to address in this connection is the following: What is the *evidence weight* or simply the *weight*, $\nu(A)$, of a subset A of the neighborhood (3)?

First, the boundary conditions $\nu(\emptyset) = 0$ and $\nu(X) = 1$ should of course be satisfied, expressing that the full evidence is provided by the complete neighborhood X. Moreover, according to our comments on the similarity hypothesis (section 3), the evidence coming from a set of cases $A \subseteq X$ should be discounted if these cases are similar among themselves. Likewise, the weight of A should be increased if the cases are "diverse" (hence complementary) in a certain sense.[1] To express this idea in a more rigorous way, we define the *diversity* of a set of cases A by the average pairwise dissimilarity:

$$\mathrm{div}(A) \stackrel{\mathrm{df}}{=} \frac{2}{|A|^2 - |A|} \sum_{x_i, x_j \in A} 1 - \mathrm{sim}(x_i, x_j)$$

(By definition, the diversity is 0 for singletons and the empty set.) We furthermore define the *relative diversity* by

$$\mathrm{rdiv}(A) \stackrel{\mathrm{df}}{=} \frac{2\,\mathrm{div}(A)}{1 - \min_{i,j} \mathrm{sim}(x_i, x_j)} - 1 \in [-1, 1]$$

Now, the idea is to modify the basic (additive) measure

$$\mu(A) \stackrel{\mathrm{df}}{=} c^{-1} \sum_{x_i \in A} \mathrm{sim}(x_0, x_i), \tag{4}$$

where $c = \sum_{x_i \in X} \mathrm{sim}(x_0, x_i)$, by taking the diversity of A into account. Of course, this can be done in different ways. Here, we used the following approach:

$$\bar{\nu}(A) \stackrel{\mathrm{df}}{=} \mu(A) \cdot (1 + \alpha\,\mathrm{rdiv}(A)) \tag{5}$$

As can be seen, the original measure $\mu(A)$ of a set of cases A is increased if the diversity of A is relatively high, otherwise it is decreased. The parameter $\alpha \geq 0$ controls the extent to which interactions between cases are taken into consideration: $\mu(A)$ can be modified by at most $(100\,\alpha)\%$. For $\alpha = 0$, interactions are completely ignored and the original measure $\mu(\cdot)$ is recovered.

For the measure $\bar{\nu}(\cdot)$ as defined in (5), the monotonicity condition does not necessarily hold. To remedy this problem, we simply enforce this property by setting

$$\nu(A) \stackrel{\mathrm{df}}{=} \max_{B \subseteq A} \bar{\nu}(B). \tag{6}$$

Finally, the boundary conditions are guaranteed by dividing the measure thus obtained by $\nu(X)$.

To illustrate, consider again the rainfall example in the right picture of Fig. 1 and suppose that $\mathrm{sim}(x_i, x_0) = .5$ ($i = 1, 2, 3$), $\mathrm{sim}(x_1, x_2) = .9$, $\mathrm{sim}(x_1, x_3) = \mathrm{sim}(x_2, x_3) = 0$, With $\alpha = 1/2$ in (5), we obtain the following weights:

A	x_1	x_2	x_3	x_1, x_2	x_1, x_3	x_2, x_3
$\nu(A)$	.27	.27	.27	.33	.83	.83

[1] The idea of "diversity" of a set of cases plays also a role in case-based retrieval [McSherry, 2002]. Here, however, the problem is more to *find* diverse cases, rather than to *combine* them.

As can be seen, the joint weight of $\{x_1, x_2\}$ is relatively low, reflecting the partial redundance of the two cases.

Before going on, let us comment on the derivation of the measure $\nu(\cdot)$ from the similarity function $\mathrm{sim}(\cdot)$. Firstly, even though the measure (6) captures our intuitive idea of decreasing (increasing) the evidence weight of cases that are (dis-)similar by themselves, we admit that it remains ad hoc to some extent, and by no means we exclude the existence of better alternatives. For example, an interesting idea is to derive the measure in an indirect way: First, the interaction indices $I_\nu(\cdot)$ from section 4.1 are defined, again based on the similarity between cases. These indices can be seen as constraints on the measure $\nu(\cdot)$, and the idea is to find a measure that is maximally consistent with these constraints.

Secondly, even though the assumption that similar cases provide redundant information is supported by the similarity hypothesis, one might of course argue that the similarity between the predictive parts of two cases, x_i and x_j, is not sufficient to call them redundant. Rather, the associated output values y_i and y_j should be similar as well. Indeed, if y_1 differs drastically from y_2, the first two measurements in our rainfall example might better be considered as non-redundant. (In that case, the two measurements in conjunction suggest that there is something amiss ...) This conception of redundancy can easily be represented by deriving $\nu(\cdot)$ from an extended similarity measure $\mathrm{sim}'(\cdot)$ which is defined over $\mathcal{X} \times \mathcal{L}$.[2]

Anyway, the important point of this section is not so much the specification of a *particular* evidence measure, but rather the insight that non-additive measures can *in principle* be used for modeling the interaction between cases in CBL.

4.3 Aggregation of Interacting Pieces of Evidence

So far, we have a tool for modeling the interaction between different pieces of evidence in case-based learning. The next question that we have to address is how to *combine* these pieces of evidence, i.e., how to aggregate them in agreement with the evidence measure $\nu(\cdot)$.

For the time being we focus on the problem of regression. Recall that in the standard approach to NN estimation, an aggregation of the output values $f(x_i) = y_i$ is realized by means of a simple weighted average:

$$y_0^{est} = \sum_{x_i \in X} \mu(\{x_i\}) \cdot f(x_i), \tag{7}$$

where $\mu(\{x_i\}) = \mathrm{sim}(x_0, x_i) \left(\sum_{x_i \in X} \mathrm{sim}(x_0, x_i)\right)^{-1}$. Interestingly, (7) is nothing else than the standard Lebesgue integral of the function $f : X \to \Re$ with respect to the additive measure (4):

$$y_0^{est} = \int_X f\, d\mu$$

In order to generalize this estimation, an integral with respect to the non-additive measure $\nu(\cdot)$ is needed: the Choquet integral, a concept that originated in capacity theory [Choquet, 1954].

[2] We didn't explore this alternative in detail so far.

Let $\nu(\cdot)$ be a fuzzy measure and $f(\cdot)$ a non-negative function.[3] The Choquet integral of $f(\cdot)$ with respect to $\nu(\cdot)$ is then defined by

$$\int^{ch} f \, d\nu \stackrel{\text{df}}{=} \int_0^\infty \eta([f > t]) \, dt$$

where $[f > t] = \{x \,|\, f(x) > t\}$. The integral on the right-hand side is the standard Lebesgue integral (with respect to the Borel measure on $[0, \infty)$). In our case, where X is a finite set, we can refer to the *discrete* Choquet integral which can be expressed in a rather simple form:

$$y_0^{est} = \sum_{\imath=1}^{k} f(x_{\pi(\imath)}) \cdot \big(\nu(A_\imath) - \nu(A_{\imath-1})\big) \,, \qquad (8)$$

where $\pi(\cdot)$ is a permutation of $\{1 \dots k\}$ such that $0 \leq f(x_{\pi(1)}) \leq \ldots \leq f(x_{\pi(k)})$, and $A_\imath = \{x_{\pi(1)} \dots x_{\pi(\imath)}\}$.

The discrete Choquet integral (8) can be seen as a special type of aggregation operator, namely a generalized arithmetic mean. Indeed, (8) coincides with (7) if $\nu(\cdot)$ is an additive measure (i.e. if $\nu(\cdot) = \mu(\cdot)$). Otherwise, it is a proper generalization of the standard (weighted) NN estimation.

Coming back to our running example, suppose that we have measured the following rainfalls: $y_1 = 100$, $y_2 = 120$, $y_3 = 200$. According to (8), we then obtain the estimation $y_0^{est} \approx 157$. As the joint weight of the two locations with less rainfall, x_1 and x_2, has been decreased, this estimation is higher (closer to y_3) than the standard weighted NN estimation given by $y_0^{est} = 140$.

So far, we have focused on the problem of regression. In the case of classification, the Choquet integral cannot be applied immediately, since an averaging of class labels $y_\imath$ does not make sense. Instead, the Choquet integral can be derived for each of the indicator functions $f_\ell : y \mapsto \mathbb{I}(y = \ell)$, $\ell = \ell_1 \dots \ell_m$. As in (1), the evidence in favor of each class label is thus accumulated separately. Now, however, the interaction between cases is taken into account. As usual, the estimation is then given by the label with the highest degree of accumulated evidence.

5 Empirical Validation

In order to validate the extension of NN estimation as proposed in the previous section, we have performed several experimental studies using benchmark data sets from the UCI repository[4] and the StatLib archive.[5]

Experiments were performed in the following way: A data set is randomly split into a training and a test set of the same size. For each example in the test set, a prediction is derived using the training set in combination with weighted k-NN resp. Cho-k-NN. In the case of regression, an estimation y_0^{est} is evaluated by the relative estimation error $|y_0^{est} - y_0| \cdot |y_0|^{-1}$, and the overall performance of a method by the mean of this error over all test examples. In the case of classification, we simply took the misclassification rate as a performance index. Moreover, we derived statistical estimations of the *expected* performance of a method by repeating each experiment 100 times.

data set	k	weighted k-NN	Cho-k-NN
auto-mpg	5	12.21 (0.05)	11.56 (0.05)
	7	12.18 (0.06)	11.53 (0.05)
bolts	5	47.07 (1.19)	38.77 (0.71)
	7	51.36 (1.25)	39.94 (0.81)
housing	5	14.83 (0.08)	14.48 (0.08)
	7	14.99 (0.09)	14.62 (0.08)
detroit	5	16.02 (0.55)	14.90 (0.50)
	7	15.93 (0.55)	14.71 (0.54)
echomonths	5	97.77 (7.17)	72.87 (3.87)
	7	99.03 (8.98)	74.80 (7.55)
pollution	5	4.12 (0.05)	4.05 (0.05)
	7	4.22 (0.05)	4.18 (0.05)

Table 1: Estimation of expected relative estimation error and its standard deviation.

For the purpose of similarity computation, all numeric attributes have first been normalized to the unit interval by linear scaling. The similarity was then defined by $1-$distance for numeric variables and by the standard 0/1-measure in the case of categorical attributes. The overall similarity $\mathrm{sim}(\cdot)$ was finally obtained by the average over all attributes. As the purpose of our study was to *compare* – under equal conditions – weighted k-NN with Cho-k-NN in order to verify whether or not taking interactions into account is useful, we refrained from tuning both methods, e.g. by including feature selection or feature weighting (even though it is well-known that such techniques can greatly improve performance [Wettschereck *et al.*, 1997]). Results have been derived for neighborhood sizes of $k = 5$ and $k = 7$; the parameter α in (5) has always been set to $1/2$.

The application of Cho-k-NN for regression has shown that it consistently improves weighted k-NN, sometimes only slightly but often even considerably. Some results are shown in table 1. In particular, it seems that the smaller the size of the data set, the higher the gain in performance. This finding is intuitively plausible, since for large data sets the neighborhoods of a query tend to be more "balanced"; as already said, the neglect of interaction is likely to be less harmful under such circumstances.

For classification problems, it is also true that Cho-k-NN consistently outperforms weighted k-NN; see table 2. Usually, however, the gain in classification accuracy is only small, in many cases not even statistically significant. Again, this is especially true for large data sets, and all the more if the classification error is already low for standard k-NN. Nevertheless, one should bear in mind that, in the case of classification, the final prediction is largely insensitive toward modifications of the estimated evidences in favor of the potential labels. In fact, in this study we only checked whether the final prediction is correct or not and, hence, used a rather crude quality measure. More subtle improvements of an estimation such as, e.g., the enlargement of an example's margin [Schapire *et*

[3] The Choquet integral can be extended to any real-valued function through decomposition into a positive and negative part.

[4] `http://www.ics.uci.edu/~mlearn`

[5] `http://stat.cmu.edu/`

data set	k	weighted k-NN	Cho-k-NN
glass	5	33.58 (0.34)	33.23 (0.35)
	7	34.54 (0.37)	33.77 (0.39)
wine	5	3.52 (0.20)	3.48 (0.20)
	7	3.27 (0.20)	3.25 (0.20)
zoo	5	10.43 (0.60)	10.08 (0.60)
	7	11.65 (0.51)	11.35 (0.52)
ecoli	5	16.30 (0.25)	16.29 (0.24)
	7	16.17 (0.24)	16.11 (0.24)
balance	5	15.62 (0.14)	15.23 (0.16)
	7	13.28 (0.16)	13.15 (0.16)
derma	5	3.75 (0.13)	3.68 (0.13)
	7	3.57 (0.12)	3.46 (0.11)

Table 2: Estimation of expected classification error and its standard deviation.

al., 1998], are not honored by this measure. For the future, we therefore plan to complement our results by more sophisticated experimental comparisons.

6 Concluding Remarks

This paper has motivated the consideration of mutual dependencies between cases that represent past experience in case-based learning. The basic idea is that a combination of two or more cases can provide *complementary* but also *redundant* evidence. In order to model this type of interaction in a formal way, we have proposed the use of non-additive measures. The aggregation of different pieces of evidence can then be accomplished by means of the Choquet integral. The inference scheme thus obtained, referred to as CHO-k-NN, is a direct extension of the standard weighted NN estimation (which is recovered in the case of an additive measure).

Our experimental results have shown that CHO-k-NN consistently outperforms standard (weighted) k-NN on publicly available benchmark data. Sometimes, only marginal improvements can be achieved, particularly in the case of classification and all the more for large, "well-distributed" data sets, but for many problems CHO-k-NN is considerably better than k-NN. All things considered, it can be said that taking interaction between cases in CBL into account is worthwhile by any means (mostly it helps, at worst it remains ineffective).

As already said, the derivation of a non-additive measure $\nu(\cdot)$ from the similarity function $\mathrm{sim}(\cdot)$ as outlined in section 4 is only a first attempt which leaves scope for development. In particular, the degree to which a set of cases is complementary resp. redundant might not only depend on their mutual similarity but also on other aspects.

We have explained how to use the Choquet integral as an aggregation operator in both regression and classification problems. An interesting question concerns the extension of our approach to more general problem solving tasks as they are typically found in case-based reasoning. Even though it is obvious that the approach cannot be transferred immediately, the basic ideas and concepts might still be useful. In any case, the concrete solution will strongly depend on the particular type of application.

References

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

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

[Choquet, 1954] G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295, 1954.

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

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

[Grabisch, 1997] M. Grabisch. k-order additive discrete fuzzy measures and their representation. *Fuzzy Sets and Systems*, 92:167–189, 1997.

[Hüllermeier, 2002] E. Hüllermeier. On the representation and combination of evidence in instance-based learning. In *Proc.* ECAI–2002, pages 360–364, Lyon, France, 2002.

[Hume, 1999] D. Hume. *An Enquiry concerning Human Understanding*. Oxford Univ. Press Inc., New York, 1999.

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

[Lindenbaum *et al.*, 1999] M. Lindenbaum, S. Marcovich, and D. Rusakov. Selective sampling for nearest neighbor classifiers. AAAI-99, pages 366–371, Orlando, 1999.

[McSherry, 2002] D. McSherry. Diversity-concious retrieval. ECCBR–02, pages 219–233. Springer, 2002.

[Murofushi and Soneda, 1993] T. Murofushi and S. Soneda. Techniques for reading fuzzy measures (iii): interaction index. In *9th Fuzzy System Symposium*, pages 693–696. Sapporo, Japan, 1993.

[Oliver and Webster, 1990] M. Oliver, R. Webster. Kriging: a method of interpolation for geographical information system. *Int. J. Geogr. Inform. Syst.*, 4(3):313–332, 1990.

[Rendell, 1986] L. Rendell. A general framework for induction and a study of selective induction. *Machine Learning*, 1:177–226, 1986.

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

[Schapire *et al.*, 1998] RE. Schapire, Y. Freund, P. Bartlett, and WS. Lee. Boosting the margin: a new explanation for the evectiveness of voting methods. *Annals of Statistics*, 26:1651–1686, 1998.

[Sugeno, 1974] M. Sugeno. *Theory of Fuzzy Integrals and its Application*. PhD thesis, Tokyo Inst. of Techn., 1974.

[Wettschereck *et al.*, 1997] 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.

[Zhang *et al.*, 1997] J. Zhang, Y. Yim, and J. Yang. Intelligent selection of instances for prediction in lazy learning algorithms. *Art. Intell. Review*, 11:175–191, 1997.