Probabilistic Indexing for Case-based Prediction

Boi Faltings

Artificial Intelligence Laboratory (LIA)
Computer Science Department (DI)
Swiss Federal Institute of Technology (EPFL)
IN-Ecublens, 1015 Lausanne, Switzerland

Abstract. The main assumption underlying case-based reasoning is that a problem with similar features as an earlier one is likely to have the same solution. However, this assumption has never been formally justified, and one can easily find practical situations where it is not true.
We use probablity theory to show that even if this fundamental assumption can be wrong for particular instances, it is guaranteed to be correct *on the average*, and this no matter what the probability distributions involved are. We define the concept of a *match weight* as a well-justified measure of similarity. We show how it is often possible to effectively compute a lower bound on match weight. We report on the performance of such bounds when used as a similarity measure in a simple example.

1 Introduction

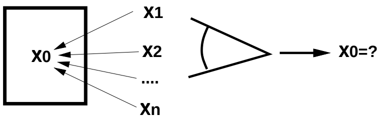

Fig. 1. *Task analyzed in this paper: predict the hidden result X_0.*

In this paper, we address the problem of predicting the value of an unobservable variable X_0 given the values of n observable and related variables $X_1, X_2, ..., X_n$, as illustrated in Figure 1. We call the observable variables *attributes* and the hidden one the *result*. Throughout the paper, we use upper case letters to denote variables, and corresponding lower case for their values; we refer to the observed values by $\hat{x}_1, \hat{x}_2, ..., \hat{x}_n$ and to the true value of the result as $\hat{x}_0$.

The relationship between attributes and result is known only through a set of representative *cases*. Each case is a record $C_i = (x_{i1}, x_{i2}, ..., x_{in}, x_{i0})$ containing the values of all attributes and the result for a particular previous experience.

The goal is to make an optimal prediction of the result given the observed attribute values and the case base. A simple example is prediction of credit risk, where $\{X_i\}$ are attributes describing the applicant and the desired loan and

$X_0 \in \{good, bad\}$. More generally, X_0 and also the attributes could be a vector of more complex values. This task is also assumed in [1, 3], and models most case-based reasoning systems.

We assume that the cases follow the same probability distributions as the scenarios presented for prediction. The optimal prediction is then the *maximum likelihood estimate*, i.e. $\hat{x}_0$ such that

$$pr(X_0 = \hat{x_0} | X_1 = \hat{x_1}, ..., X_n = \hat{x_n})$$

is maximal using the distribution of the precedent cases.

Recognition and Prediction The problem can pose itself in two forms, which we shall call *recognition* and *prediction*. In recognition, the observable attributes are a function of the class and some noise. A typical example of this is recognizing different animals given their features. Many statistical estimation techniques, including k-nearest neighbours, are designed for this model. It also implies that the attribute values are independent of one another given the class, so that the prediction problem can be solved using Bayesian inference ([7, 4]).

In prediction, it is the attributes which determine the result up to some noise. A typical example of this is predicting credit risk: presumably the different caracteristics of the loan will influence the course of events and lead to the credit being good or bad some time in the future. Now, it can no longer be assumed that the attributes are independent given the result: for example, the amount of the requested loan and the income of the applicant might well be statistically independent, but when it is known that the loan has been defaulted on it becomes far more likely that the amount was large with respect to income. A prediction using probabilistic inference would thus require the entire joint probability distribution

$$pr(X_0 | X_1, ..., X_n)$$

In the absence of background information, the number of examples required to estimate this distribution grows exponentially with the number of attributes. We assume here that the given data is insufficient to provide a sufficiently precise estimate of this distribution.

Prediction using base-based reasoning Case-based reasoning avoids the need for explicit probability distributions. Here, we find an earlier case which matches the current observations as closely as possible, and use its value of X_0 as the prediction. Thus, the problem is now no longer to find the maximum likelihood prediction, but to find the precedent which is most likely to have the same result as the observations, i.e. find a case X_i such that

$$pr(x_{i0} = \hat{x}_0 | x_{ij} = \hat{x}_j, x_{ik} = \hat{x}_k, ..., x_{il} = \hat{x}_l)$$

is maximized, where $j, k, ..., l$ are the indices of the matching attributes. We call this probability the *weight* associated with matching this set of attributes. The weight can be used as a similarity measure among cases: the higher the weight of a match, the larger the probability that it makes the correct prediction, and the more similar it is to the current situation. Whether the maximum weight match also gives the maximum likelihood prediction is an open research question.

A closer look at similarity The assumption underlying case-based reasoning is that the more similar a case is to the current problem, to more likely it is to provide the correct solution. Most measures of similarity have an ad-hoc character without formal justification, although analyses of certain measures have been proposed ([5, 9]). However, it is not always the case that higher similarity translates into higher prediction accuracy.

As an example, consider predicting the sucess of a company, characterized by two attributes:

- $X_1 = 1/0$ indicates that the company has/does not have debt
- $X_2 = 1/0$ indicates that it is a high-tech company or not

We would like to predict the value of attribute X_0 which is equal to 0 if the company is a failure, and 1 if it will be successful. Let the probability that $X_0 = 1$ given a certain combination of attributes values be given by the following table:

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	0.9	0.6
$X_1 = 1$	0.4	0.5

and note that $pr(X_0 = 0) = 1 - pr(X_0 = 1)$. In a real application, this table would of course be unknown. Assume furthermore that all combinations of attribute values are equally likely, and that we have two sets of cases:

- set C_1: all cases where $X_1 = 0, X_2 = 1$
- set C_2: all cases where $X_1 = 0$

so that $C_2 \supseteq C_1$. We would like to predict the success of a company which has no debt and is a high-tech company, characterized by the vector $(\hat{x}_1 = 0, \hat{x}_2 = 1)$. A case in set C_1 provide a perfect match. Let's see if it also has the highest probability of a correct prediction. Using the table, we see that for our company, the real probability of success is 0.6, the probability of failure is 0.4. A case in C_1 predicts success and failure with the same probabilities, and thus the probability of its x_{10} making the correct prediction is:

$$pr(\hat{x}_0 = x_{10} = 0) + pr(\hat{x}_0 = x_{10} = 1) = 0.6^2 + 0.4^2 = 0.36 + 0.16 = 0.52$$

Now consider a arbitrary case in C_2. It predicts success with probability $0.5 \cdot (0.9+0.6) = 0.75$, failure with probability $1-0.75 = 0.25$. Hence, the probability of making the correct prediction when using an arbitrary case in C_2 is:

$$pr(\hat{x}_0 = x_{20} = 0) + pr(\hat{x}_0 = x_{20} = 1) = 0.6 \cdot 0.75 + 0.4 \cdot 0.25 = 0.45 + 0.1 = 0.55$$

Thus, using a case in C_2, obviously a worse match than a case in C_1, will on average result in a better prediction! Note that this characteristic is incompatible with many attempts at defining well-justified similarity metrics which are sensitive to particular values.

Notation Since the scenarios presented for prediction follow the same distribution as the cases themselves, they can be considered as additional members of the case base for estimating the required probabilities. For the remainder of this paper, we use underlines as a shorthand denoting matching arguments and overlines as a shorthand for non-matching arguments of probablity distributions:

- $X_j = x_{ij}$ is written as $\underline{x_{ij}}$.
- $X_j = x_{ij}$ averaged over all possible values of x_{ij} is written as $\underline{X_j}$

The average is always taken over all value combinations of capitalized variables in the formula. Thus, for example, we write:

$$pr(\underline{x_{50}}|\underline{x_{51}}, ..., \underline{x_{5n}}) \text{ for } pr(\hat{x}_0 = x_{50}|\hat{x}_1 = x_{51}, ..., \hat{x}_n = x_{5n})$$

$$pr(\underline{X_0}|\underline{X_1}, \underline{X_2}) \text{ for}$$

$$\sum_{x_{i0}\in X_0} \sum_{x_{i1}\in X_1} \sum_{x_{i2}\in X_2} pr(x_{i0}, x_{i1}, x_{i2}) \cdot pr(\hat{x}_0 = x_{i0}|\hat{x}_1 = x_{i1}, \hat{x}_2 = x_{i2})$$

2 Statistically independent attributes

Since it is unlikely that we will find a precedent which matches the observations exactly, we will need to determine which partial match is most likely to give us the best estimate of the classification. We define:

Definition 1. The *weight* of an attribute value $X_j = x_{ij}$ or a combination of attribute values is the increase in the probability of correct prediction when the attributes match over its a-priori value:

$$w(\underline{x_{ij}}) = pr(\underline{X_0}|\underline{x_{ij}}) - pr(\underline{X_0})$$

$$w(\underline{x_{ij}}, ..., \underline{x_{il}}) = pr(\underline{X_0}|\underline{x_{ij}}, ..., \underline{x_{il}}) - pr(\underline{X_0})$$

Under the assumption that the distribution of the precedents is the same as the distribution of the observations, the weight can be computed from these as follows:

$$\begin{aligned} w(\underline{x_{ij}}, ..., \underline{x_{il}}) &= pr(\underline{X_0}|\underline{x_{ij}}, ..., \underline{x_{il}}) - pr(\underline{X_0}) \\ &= \sum_{x_{i0}\in X_0} pr(\underline{x_0}|\underline{x_{ij}}, ..., \underline{x_{il}})^2 - pr(\underline{x_0})^2 \end{aligned}$$

Applying these definitions to the example in the introduction, we find:

$$pr(\underline{X_0}) = (0.6^2 + 0.4^2) = 0.52$$

$$w(x_1 = 0) = 0.55 - 0.52 = 0.03$$

$$w(x_1 = 0, x_2 = 1) = 0.52 - 0.52 = 0$$

so that for these particular values the weight of only matching X_1 is indeed higher than that of matching both X_1 and X_2!

Average weights A more intuitive result can be obtained if instead of considering the weights of matches or mismatches with *particular values*, we only consider the *average* weights of matches or mismatches in a certain attribute. We thus define:

Definition 2. The *average weight* $W(\underline{X})$ of matching an attribute X is the average of $w(X = x_i)$ over all possible values x_i of the attribute:

$$W(\underline{X}) = \sum_{x_i \in X} pr(X = x_i) w(X = x_i)$$

Applying these definitions to the example in the introduction, we now have:

$$\begin{aligned} W(\underline{X_1}, \underline{X_2}) = & \sum_{x_1, x_2 \in \{0,1\}} pr(X_1 = x_1, X_2 = x_2) w(X_1 = x_1, X_2 = x_2) \\ & = 0.25 \cdot (0.9^2 + 0.1^2) + 0.25 \cdot (0.6^2 + 0.4^2) + 0.25 \cdot (0.4^2 + 0.6^2) + \\ & \quad 0.25 \cdot (0.5^2 + 0.5^2) - (0.6^2 + 0.4^2) \\ & = 0.205 + 0.13 + 0.13 + 0.125 - 0.52 = 0.59 - 0.52 = 0.07 \end{aligned}$$

as the average weight for matching both attributes, and

$$\begin{aligned} W(\underline{X_1}) = & \sum_{x_1 \in \{0,1\}} pr(X_1 = x_1) w(\underline{x_1}) \\ & = 0.5 \cdot (0.75^2 + 0.25^2 - 0.52) + 0.5 \cdot (0.45^2 + 0.55^2 - 0.52) \\ & = 0.045 \end{aligned}$$

as the average weight for matching X_1 only, so that *on average* also matching X_2 does produce better results!

Case-based reasoning works on the average It turns out that in fact, the intuition underlying case-based reasoning is in fact always correct as long as only *average* weights are considered:

Theorem 3. *Independently of the probability distributions, the average weight* $W(\underline{X})$ *of any attribute or combination of attributes* X *is always non-negative.*

Proof:
Observe that by Jensen's inequality and the convexity of squaring, for any x_0:

$$\begin{aligned} \sum_{x_i \in X} pr(X = x_i) pr(X_0 = x_0 | X = x_i)^2 \geq & \left[\sum_{x_i \in X} pr(X = x_i) pr(X_0 = x_0 | X = x_i) \right]^2 \\ & = pr(X_0 = x_0)^2 \end{aligned}$$

so that:

$$\begin{aligned} W(\underline{X_i}) = & \sum_{x_0 \in X_0} \left[\sum_{x_i \in X} pr(X = x_i) pr(X_0 = x_0 | X = x_i)^2 \right] - pr(X_0 = x_0)^2 \\ \geq & \sum_{x_0 \in X_0} pr(X_0 = x_0)^2 - pr(X_0 = x_0)^2 = 0 \end{aligned}$$

and the theorem is proven.
QED

For independent attributes, we can show an even stronger theorem:

Theorem 4. *Given a sets of attributes A and an attribute B which is* statistically independent *of all attributes in A:*

$$W(\underline{A}) + W(\underline{B}) \leq W(\underline{A}, \underline{B})$$

Proof: Without loss of generality, we assume A to be single attribute; if A is a set of attributes, it can be regarded as a single vector-valued attribute. We prove the theorem by induction over the sets of possible values that the attributes A and B can take. For the base of the induction, assume that A can take only a single value, whereas B can take n values. Then, attribute A is always matched, thus $W(\underline{A}) = 0$ and $W(\underline{A}, \underline{B}) = W(\underline{B})$, and the theorem is true.

Now assume that the theorem holds for any A with up to k attribute values, and B takes n values. Let A be an attribute with $k+1$ values, labelled v_1 through v_{k+1}. Define a new attribute A' with k values v'_1 through v'_k such that:

$$\begin{aligned} v'_i &= v_i, i = 1..k-1 \\ v'_k &= v_k \vee v_{k+1} \end{aligned}$$

so that A' takes the same values as A except that it takes values v'_k whenever A takes values v_k or v_{k+1}. We define the following shorthand notation:

$$\begin{aligned} p &= pr(A = v_k) \\ q &= pr(A = v_{k+1}) \\ r &= pr(\underline{X_0} | A = v_k, B = x_j) \\ s &= pr(\underline{X_0} | A = v_{k+1}, B = x_j) \\ t &= \sum_{x_j \in B} pr(B = x_j) pr(\underline{X_0} | A = v_k, B = x_j) \\ u &= \sum_{x_j \in B} pr(B = x_j) pr(\underline{X_0} | A = v_{k+1}, B = x_j) \end{aligned}$$

Since most terms in the weight calculations involving A and A' are identical, the following differences only involve the terms referring to v_k and v_{k+1}; we have:

$$\begin{aligned} & W(\underline{A}, \underline{B}) - W(\underline{A'}, \underline{B}) \\ &= \sum_{x_0 \in X_0} \sum_{x_j \in B} pr(B = x_j) pr^2 + qs^2 - \frac{1}{p+q}[(pr)^2 + (qs)^2] \\ &= \sum_{x_0 \in X_0} \sum_{x_j \in B} pr(B = x_j) \frac{1}{p+q} \left\{ p^2 r^2 + pqr^2 + pqs^2 + q^2 s^2 - p^2 r^2 - q^2 s^2 \right\} \\ &= \sum_{x_0 \in X_0} \sum_{x_j \in B} pr(B = x_j) \frac{pq}{p+q} (r^2 + s^2) \end{aligned}$$

and:

$$\begin{aligned} W(\underline{A}) - W(\underline{A'}) &= \sum_{x_0 \in X_0} pt^2 + qu^2 - \frac{1}{p+q}[(pt)^2 + (qu)^2] \\ &= \sum_{x_0 \in X_0} \frac{pq}{p+q}(t^2 + u^2) \end{aligned}$$

so that:

$$\begin{aligned} &[W(\underline{A}, \underline{B}) - W(\underline{A}) - W(\underline{B})] - [W(\underline{A'}, \underline{B}) - W(\underline{A'}) - W(\underline{B})] = \\ &\sum_{x_0 \in X_0} \frac{pq}{p+q} \left\{ [\sum_{x_j \in B} pr(B = x_j)(r^2 + s^2)] - (t^2 + u^2) \right\} = \\ &\sum_{x_0 \in X_0} \frac{pq}{p+q} \left\{ [\sum_{x_j \in B} pr(B = x_j) r^2] - t^2 + [\sum_{x_j \in B} pr(B = x_j) s^2] - u^2 \right\} \\ &\geq 0 \end{aligned}$$

since by Jensen's inequality:

$$\begin{aligned} &[\sum_{x_j \in B} pr(B = x_j) r^2] - t^2 \\ &= [\sum_{x_j \in B} pr(B = x_j) r^2] - \left\{ \sum_{x_j \in B} pr(B = x_j) pr(\underline{X_0} | A = v_k, B = x_j) \right\}^2 \\ &\geq [\sum_{x_j \in B} pr(B = x_j) r^2] - [\sum_{x_j \in B} pr(B = x_j) pr(\underline{X_0} | A = v_k, B = x_j)^2] \\ &= \sum_{x_j \in B} pr(B = x_j)(r^2 - r^2) = 0 \end{aligned}$$

and similarly

$$[\sum_{x_j \in B} pr(B = x_j) s^2] - u^2 \geq 0$$

Since the theorem is satisfied for A', which has only k values, it is also satisfied for A. This completes the induction, and the theorem is proven.
QED

3 Dealing with dependent attributes

With the exception of Theorem 3, all results so far have assumed that attributes are statistically independent of one another. In reality, such independence will occur only very rarely. This can have dramatic effects. For example, let X and Y be two attributes such that always $X = Y$. Thus, whenever X matches, Y will also match, and the weight $W(\underline{X}, \underline{Y}) = W(\underline{X}) = W(\underline{Y})$!

Practical experience has shown that in many practical problems, attribute dependence can be modelled using probabilistic networks ([7, 2]). A probabilistic network is a directed acyclic graph whose nodes are attributes and whose arcs indicate statistical dependencies between nodes. Nodes are statistically dependent only on their direct parents, i.e. if node x has parents $y_0, .., y_k$, then for any other set of nodes Z:

$$pr(x|y_0, ..., y_k, Z) = pr(x|y_0, ..., y_k)$$

More important than the links which are present in the network are those which are absent; these indicate independence relations. More precisely, any pair of nodes x_1 and x_2 with parents Z_1 and Z_2 and not having a path between them is statistically independent given $Z_1 \cup Z_2$, i.e.:

$$pr(x_1|x_2, Z_1 \cup Z_2, Y) = pr(x_1|Z_1 \cup Z_2)$$

This result was proven by Olmsted ([6]); it also figures as Corrolary 4 on page 120 of [7]. Thus, two variables X and Y which have no direct link between them are *conditionally independent* given the values of their common ancestors Z:

$$pr(X = x, Y = y|Z) = pr(X = x|Z) \cdot pr(Y = y|Z)$$

Conditional weights We define the *conditional weight* of a set of matching or mismatching attributes A given a set of matching or mismatching attributes B:

$$W(\underline{A}|\underline{B}) = W(\underline{A}, \underline{B}) - W(\underline{B})$$

We can now prove:

Theorem 5. *Assume that X and Y are conditionally independent given an attribute or set of attributes Z. Then:*

$$W(\underline{X}|\underline{Z}) + W(\underline{Y}|\underline{Z}) \leq W(\underline{XY}|\underline{Z})$$

Proof: We apply the following transformations:

$$\begin{aligned} W(\underline{X}|\underline{Z}) &= W(\underline{X}, \underline{Z}) - W(\underline{Z}) \\ &= \sum_{z \in Z} pr(z) \cdot [pr(\underline{X_0}|\underline{X}, \underline{z}) - pr(\underline{X_0}|\underline{z})] \\ W(\underline{Y}|\underline{Z}) &= \sum_{z \in Z} pr(z) \cdot [pr(\underline{X_0}|\underline{Y}, \underline{z}) - pr(\underline{X_0}|\underline{z})] \\ W(\underline{X}, \underline{Y}|\underline{Z}) &= \sum_{z \in Z} pr(z) \cdot [pr(\underline{X_0}|\underline{X}, \underline{Y}, \underline{z}) - pr(\underline{X_0}|\underline{z})] \end{aligned}$$

We now prove the theorem by showing that the inequality holds for every z, i.e. we show that:

$$pr(\underline{X_0}|\underline{X}, \underline{z}) - pr(\underline{X_0}|\underline{z}) + pr(\underline{X_0}|\underline{Y}, \underline{z}) - pr(\underline{X_0}|\underline{z}) \leq pr(\underline{X_0}|\underline{X}, \underline{Y}, \underline{z}) - pr(\underline{X_0}|\underline{z})$$

Let $p_z(\cdot)$ be the probability distribution $pr(\cdot)$ conditioned on $\underline{z}$. By the assumption of conditional independence, $p_z(\underline{X_0}, \underline{X})$ and $p_z(\underline{X_0}, \underline{Y})$ are independent probability distributions. Also define $W_z(\cdot) = p_z(\underline{X_0}|\cdot) - p_z(\underline{X_0})$, so that we have:

$$W_z(\underline{X}) + W_z(\underline{Y}) \leq W_z(\underline{X}, \underline{Y})$$

Since X and Y are independent with respect to this weight calculation, this inequality is true by Theorem 4, and so the theorem is proven. QED.

4 Using weights as similarity metrics

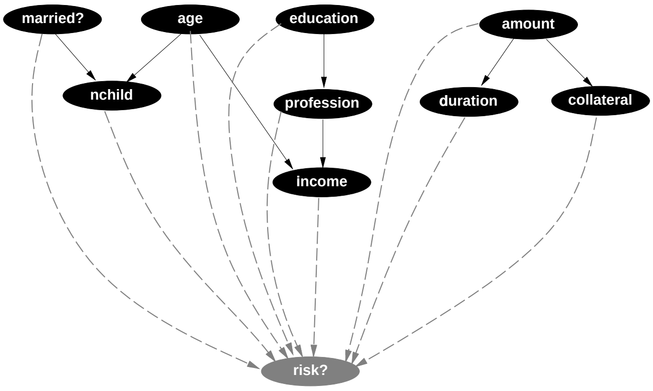

Fig. 2. *Example of assessing credit risk.*

Since match weights give the probability that a case makes the correct prediction, they would make an ideal similarity metric for indexing cases. This requires in particular an efficient way of computing match weights for any combination of features. Note that estimating match weights for large combinations of features directly from the case base will give inaccurate values due to an insufficient number of examples. Conditional independence relations provide a way to obtain a lower bound on the match weight given the match weights for small combinations of features. This lower bound could be used as a similarity metric, and we now give an algorithm for computing it.

A synthetic example of assessing credit risk is used to illustrate. It has 9 attributes whose dependencies are accurately characterized by the graph shown in Figure 2. Credit risk (good/bad) is computed as a deterministic function of all attributes. As an example, assume that we want to know the weight of matching a case in attributes X_3, X_4, X_5, X_6, X_7 and X_8.

Bounding match weights The algorithm for bounding match weights has the following steps:

1. reduce the probabilistic network for all attributes to one which contains only those that participate in the match.
2. perform a topological sort, ordering the nodes in the network so that every node is only connected to predecessors in this ordering.
3. approximate the conditional match weights for all attributes and arcs in the network by counting out in the case base.
4. compute the weight of the combined match by composition.

In general, it will be very rare for *all* attributes in the case description to match the observations. The first step is thus to construct a *reduced* probabilistic network containing only attributes which participate in the match. In this new network, paths through nodes which have been eliminated are replaced by direct links. Such direct links must be created for all paths which are either:

- a sequence of ancestor relations, or
- two sequences of ancestor relations leading to a common ancestor.

but not between nodes which share a common successor (so-called head-to-head nodes in the literature on Bayesian networks).

Combining match weights Theorem 4 now allows us to combine weights estimated from the case base into bounds on weights for simultaneous matches of all features in a network. This computation is very similar to the proapgation of probabilities in probabilistic networks ([7]):

```
1. perform a topological sort on the network, classifying nodes into classes
   $G_0, ..., G_k$ such that all parents of nodes in $G_{k+1}$ are in $G_k, G_{k-1}, ..., G_0$.
2. W $\leftarrow$ 0
3. for i $\leftarrow$ 0 to k do
     – for all X $\in G_i$ with ancestors $Y$: estimate $W(\underline{X}|\underline{Y})$ from the case base
       and set $W \leftarrow W + W(\underline{X}|\underline{Y})$
```

In this example, we have two classes: $G_0 = \{X_3, X_6, X_7\}$ and $G_1 = \{X_4, X_5, X_8\}$. Thus, we require the weights $W(\underline{X_3}), W(\underline{X_6})$ and $W(\underline{X_7})$ as well as the conditional weights $W(\underline{X_4}|\underline{X_3}), W(\underline{X_5}|\underline{X_4}, \underline{X_6})$ and $W(\underline{X_8}|\underline{X_7})$.

Approximating match weights from the case base Average weights can be precomputed by counting out all pairs of matching cases. For computing the weight of the combination $W(\underline{X_k}, \underline{X_l}, ..., \underline{X_m})$, the algorithm would be as follows:

```
1. $W_t \leftarrow 0, W_f \leftarrow 0$
2. for all pairs of cases $C_i, C_j$
   matching in $X_k, X_l, ..., X_m$ do

     if $x_{0i} = x_{0j}$ then $W_t \leftarrow W_t + 1$
     else $W_f \leftarrow W_f + 1$.
```

3. return $\frac{W_t}{W_f+W_t}$

This algorithm can also be used to estimate several match weights in parallel during a single pass over all case combinations. Conditional match weights are best computed using the formula:

$$W(\underline{A}|\underline{B}) = W(\underline{A}, \underline{B}) - W(\underline{B})$$

Dealing with inexact matches Weights are computed only for attributes which match exactly. For attributes with a large number or values, such as numbers, this will rarely be the case. It would be desirable to take into account imprecise matches in such attributes as well.

Plaza ([8]) has studied similarity measures where attribute values are grouped into a hierarchy such that all values in a group share some similarities. For example, an attribute taking as values real numbers between 0 and 10 could be represented by a hierarchy of intervals:

- level 0: [0..10]
- level 1: [0..4], [5..10]
- level 2: [0..2],[3..4],[5..7],[8..10]
- level 3: 0,1,2,3,4,5,6,7,8,9,10

Now, a match can occur at different levels. For example, values 2 and 5 would match at level 0, whereas 5 and 7 would match at level 2. Depending on the level of the match, its contribution to the total similarity can be smaller or greater.

Such hierarchies can be applied to weight computations as well and significantly improve accuracy. Now, we consider every level in the hierarchy a separate attribute, and compute weights for matches at all levels in the hierarchy. This results in much higher weights and thus greater confidence in the prediction.

5 Discussion

Most existing theoretical analyses of statistical prediction have considered the *classification* problem, where attributes can be assumed to provide independent evidence to the classification; examples of this are work in Bayesian inference ([7, 2]) as well as k-nearest neighbour classifiers.

In this paper, we have instead considered the *prediction* problem, where the contributions of attributes are not independent. The case-based reasoning approach is promising for this problem class because it does not require any assumptions about the nature of the relationship between attributes and the result. owever, assumptions about the attribute-class relationship are often introduced in the similarity measure used for case indexing. The analysis in this paper does not require any such assumption.

The main novel results of this paper are that on average, increased similarity does indeed lead to improved prediction *independently of how attributes and classes are distributed*, and that furthermore it is possible to compute lower bounds on the true probability of correct prediction.

In practice, these lower bounds seem to provide very powerful similarity metrics, although our experiments are still too rudimentary to give definite conclusions. The sum of the weight and the a-priori probability of correct prediction is equal to the probability that the case makes the correct prediction. If it is close to 1, which is often the case in our synthetic example, this provides a good confidence measure for using the particular case. In applications where bounds are always much smaller than 1, this is an indication that either the case base is really much too sparse, or that the attributes used are not the right ones for classification. In this case, the bounds may also help in guiding the search for attributes which would allow more accurate classification.

Acknowledgements

Part of this work was carried out while the author was on leave at NEC Research Institute, Princeton, NJ; the generous support is gratefully acknowledged. I would also like to thank Marek Blaszczyk for the implementation example.

References

1. **R.H. Creecy, B.M. Masand, S.J. Smith, D.L.Waltz**: "Trading MIPS and Memory for Knowledge Engineering," *Communications of the ACM* **35**(8), August 1992
2. **D. Heckerman**: "Probabilistic Similarity Networks," *MIT Press*, 1990
3. **S. Kasif, S. Salzberg, D. Waltz, J. Rachlin, D. Aha**: "Towards a Framework for Memory-Based Reasoning," NECI Technical Report 95-132, December 1995
4. **P. Myllymäki, H. Tirri**: "Massively Parralel Case-Based Reasoning with Probabilistic Similarity Metrics," *Proceedings of the 1st European Workshop on Case-based Reasoning*, Lecture Notes in Artificial Intelligence 837, pp. 145-154, Springer Verlag, 1993
5. **S. Okamoto, K. Satoh**: "An Average-Case Analysis of k-Nearest Neighbor Classifier," *Proceedings of the 1st International Confernce on Case-based Reasoning*, Lecture Notes in Artificial Intelligence 1010, pp. 253-264, Springer Verlag, 1995
6. **S. Olmsted**: "On representing and solving decision problems," Ph.D. Thesis, Department of Engineering - Economic Systems, Stanford University, 1983
7. **J. Pearl**: "Probabilistic Reasoning in Intelligent Systems," Morgan-Kaufmann, 1988
8. **E. Plaza, R. López de Mántaras, E. Armengol**: "On the Importance of Similitude: An Entropy-Based Assessment," *Proceedings of the 3rd European Workshop on Case-based Reasoning*, Lecture Notes in Computer Science 1168, pp. 324-338, Springer Verlag, 1996
9. **M.M. Richter**: "On the Notion of Similarity in Case-Based Reasoning," in G. della Riccia et al (eds.): *Mathematical and Statistical Methods in Artificial Intelligence*, Springer Verlag 1995, pp. 171-184

This article was processed using the LaTeX macro package with LLNCS style