Towards a Theory of Optimal Similarity Measures

A D Griffiths

Department of Computer Science, University of York, York YO1 5DD, UK
Email: tony@minster.york.ac.uk

D G Bridge

Department of Computation, University College, Cork, Republic of Ireland
Email: d.bridge@cs.ucc.ie

Abstract. The effectiveness of a case-based reasoning system is known to depend critically on its similarity measure. However, it is not clear whether there are elusive and esoteric similarity measures which might improve the performance of a case-based reasoner if substituted for the more commonly used measures. This paper therefore deals with the problem of choosing the best similarity measure, in the limited context of instance-based learning of classifications of a discrete example space. We consider both 'fixed' similarity measures and 'learnt' ones. In the former case, we give a definition of a similarity measure which we believe to be 'optimal' w.r.t. the current prior distribution of target concepts and prove its optimality within a restricted class of similarity measures. We then show how this 'optimal' similarity measure is instantiated by some specific prior distributions, and conclude that a very simple similarity measure is as good as any other in these cases. In a further section, we then show how our definition leads naturally to a conjecture about the 'best' way of learning a similarity measure from the training sample.

1 Introduction

Experiments by Globig and Wess show empirically that the effectiveness of an instance-based learning (IBL) algorithm depends on the similarity measure used [7]. Their experiments show that a similarity measure may maximise the accuracy of a learner in one experiment but be surpassed by a different similarity measure in another experiment. Similarly, in [6], Globig and Lange show that the choice of similarity measure critically affects the number of examples needed to represent a particular function, but that there is no 'universal' similarity measure which will efficiently represent *any* target function. It is well known, therefore, that a similarity measure must be biased in anticipation of some target functions which are expected to be more probable than others.

This paper considers the problem of choosing the 'best possible' similarity measure for an instance-based learner, given some expectation about those target functions which are likely to be encountered by the learner, represented by

a *prior distribution* over the space of possible target functions. §2 gives the definitions that are necessary for the presentation. In §3 we then define what we believe to be the optimal similarity measure w.r.t. some prior distribution of target functions, in the case where the instance-based learner uses only a single, fixed similarity measure. We prove its optimality within a restricted set of similarity measures; its optimality in the general case remains unproven, but stands as a conjecture. We then consider the actual functions that are specified if the definition is instantiated with specific distributions. §4 then shows how this definition can be transferred to the case of an instance-based learner which changes its similarity measure as part of the process of learning. This paper describes work in progress and so §5 concludes with several comments about how the work can be progressed further. This work is an overview of results presented in [8]; full details of all the claims and notation may be found there.

2 Definitions

The paper is concerned with the problem of classifying instances drawn from a finite, discrete example space. For ease of presentation, we assume (w.l.o.g. in this case) that examples are represented by a number of binary (0 or 1) valued features. The *example space*, or space of possible problem descriptions, is therefore the space of N-bit binary vectors, referred to as D_N and defined $D_N \hat{=} \{0,1\}^N$. The i-th bit of a description $d \epsilon D_N$ is written $(d)_i$. In addition, we consider only $\{0,1\}$-valued target functions representing *binary* classifications of the example space and use our usual notation for these target 'concepts'. The space of all concepts defined on D_N is called B_N; $B_N \hat{=} D_N \to \{0,1\}$. In particular, this paper refers to the behaviour of instance-based learning algorithms on *monomial* or conjunctive target concepts. A monomial concept can be represented in the propositional calculus by a simple conjunction of literals; e.g. $u_1\overline{u}_2\overline{u}_3$ may represent the concept on D_5 whose positive instances are 10000, 10001, 10010 and 10011. The set of monomial concepts defined on D_N is referred to as M_N. Further, $M_{N,k}$ is defined as the set of monomials with exactly k literals; u_1 represents a concept in $M_{N,1}$, while $u_1\overline{u}_2u_3\overline{u}_4$ represents a concept in $M_{N,4}$. The decision function for the concept h_U represented by monomial expression U (treated as a set of literals) is defined formally as follows:

$$h_U(d) = \begin{cases} 1 \textit{ if } \forall i \cdot (u_i \epsilon U \to ((d)_i = 1)) \wedge (\overline{u}_i \epsilon U \to ((d)_i = 0)) \\ 0 \textit{ otherwise} \end{cases} \quad (1)$$

Simple IBL algorithms learn by adding cases to a case-base CB and by adjusting a similarity measure σ. A case-base CB is a set of *exemplars*, each of which is a pair $(d,n) \epsilon (D_N \times \{0,1\})$ consisting of a problem description $d \epsilon D_N$ along with a single bit, 0 or 1. Normally, a case-base is *compatible* with some target concept $t \epsilon B_N$ s.t. for each exemplar $(d,n) \epsilon CB$, $t(d) = n$. The similarity measure σ is a total function in $D_N \times D_N \to [0,1]$ which returns a real value indicating the degree of similarity between its two arguments. The pair $\langle CB, \sigma \rangle$

is interpreted as the *representation* of a concept on D_N as follows:

$$h_{\langle CB,\sigma\rangle}(d) = \begin{cases} 1\ if\quad \exists(d_{pos},1)\,\epsilon\, CB \cdot \forall(d_{neg},0)\,\epsilon\, CB \cdot \sigma(d,d_{pos}) > \sigma(d,d_{neg}) \\ 0\ otherwise \end{cases} \tag{2}$$

In other words, a point $d \epsilon D_N$ is positively classified by $h_{\langle CB,\sigma\rangle}$ if and only if there is a stored positive exemplar d_{pos} which is strictly more similar to d according to the chosen similarity measure σ than any of the stored negative exemplars d_{neg}. Many IBL algorithms use a *weighted* similarity measure which counts the 'overlap' between two descriptions but assigns weights to each of the variables of the representation to indicate their importance or 'relevance'. The simplest form of these measures may be defined by a 'weight vector' $\overline{w}\,\epsilon\,[0,1]^N$:

$$\sigma_{\overline{w}}(d_1,d_2) = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \times (1 - |(d_1)_i - (d_2)_i|) \tag{3}$$

If the weight vector $\overline{w}$ has weight 1 in all elements then $\sigma_{\overline{w}}$ treats all dimensions of the representation equally and is analogous to the Hamming distance between the two descriptions. This special case will be written σ_H. $\sigma_{\overline{w}}$, or indeed any other σ, is used by equation (2) only to rank the exemplars of the case-base to identify the 'near neighbours' which classify a test instance. This same information is expressed abstractly (non-numerically) by the *retrieval relations* defined by the measure.[1] The retrieval orderings defined on D_N by a similarity measure σ are the set of orderings $\{\succeq_d^{\sigma}\}_{d\,\epsilon\, D_N}$ s.t. for each problem description $d \epsilon D_N$ the relation $d' \succeq_d^{\sigma} d''$ denotes that d is as similar to a case with description d' as to a case with description d'':

$$\forall d,d',d'' \,\epsilon\, D_N \cdot d' \succeq_d^{\sigma} d'' \leftrightarrow \sigma(d,d') \geq \sigma(d,d'') \tag{4}$$

In general, a description $d \epsilon D_N$ will have a number of descriptions which are equally similar to it. For example, $\sigma_H(1111,1110) = \sigma_H(1111,0111) = \frac{3}{4}$ and therefore both $0111 \succeq_d^{\sigma_H} 1110$ and $1110 \succeq_d^{\sigma_H} 0111$ for the description $d = 1111$. We can therefore say that the descriptions 1110 and 0111 are members of an *equivalence class* of the preference relation defined by d and σ_H.

In this work a learning experiment is described by two different probability distributions. A probability distribution on the example space will be referred to as μ, and the probability that a randomly chosen example is a member of some set $X \subseteq D_N$ will be written μX and the *conditional* probability that an example $x \epsilon D_N$ will have some property $p(x)$ given membership of some set $X \subseteq D_N$ will be written $\mu\{x \epsilon X | p(x)\}$. A sequence of examples from $(D_N)^*$ will be written $\overline{x}$, while a training sample $\overline{x}_t \epsilon (D_N \times \{0,1\})^*$ is a sequence of examples from D_N 'labelled' according to the target function t so that for each

[1] These orderings were first noted by Richter, Wess *et al* [16] [14] [19] who refer to these orderings as 'preference relations'. In our work this term is reserved for a different kind of entity [9] [8].

pair of elements $\langle x_i, n_i \rangle$ in $\overline{x}_t$ we have $t(x_i) = n_i$. The probability of an m-sample chosen by independent random sampling from D_N, will be given by the joint probability distribution μ^m defined on the product space $(D_N)^m$, where $\mu^m\{\overline{x}\} = \mu\{x_1\} \times \ldots \times \mu\{x_m\}$. The other distribution is defined on B_N, the space of possible target concepts. It is referred to by π and called the 'prior distribution' of the target concepts.

Systems which learn to classify may be compared in terms of their *accuracy*, meaning simply the probability with which a randomly chosen example will be correctly classified by the chosen classifier. Hence the accuracy of a learning algorithm is the probability $\mu\{x \,\epsilon\, D_N | L(\overline{x}_t)(x) = t(x)\}$ where $L(\overline{x}_t)$ is the hypothesis produced by learning algorithm L given a training sample $\overline{x}_t$ and t is the target concept. Since this will depend on both the training sample and the target concept, expected accuracy is defined with respect to a particular target concept by summing over all possible training samples, weighted by their respective probabilities, and with respect to a particular prior, by calculating the expected accuracy, over all the target concepts, w.r.t that prior distribution:

$$EA^L_{(t,\mu)}(m) = \sum_{\overline{x} \,\epsilon\, (D_N)^m} \mu^m\{\overline{x}\} \cdot \mu\{x \,\epsilon\, D_N | L(\overline{x}_t)(x) = t(x)\} \tag{5}$$

$$EA^L_{(\pi,\mu)}(m) = \sum_{t \,\epsilon\, B_N} \pi\{t\} \cdot EA^L_{(t,\mu)}(m) \tag{6}$$

3 Optimal, 'Fixed' Similarity Measures

$CB1(\sigma)$ is a simple instance-based learning algorithm that learns with a fixed similarity measure and adjusts its hypothesis only by adding cases to the case-base (Figure 1). It is the most straightforward instance-based learning algorithm and is an example of an instance-based learner which learns using a *fixed* similarity measure, since all hypotheses chosen by $CB1(\sigma)$ will use the same similarity measure σ to classify problem instances. Note that Figure 1 actually defines a *family* of learners since the definition is parameterised by the similarity measure σ.

```
set CB = ∅
for i = 1 to m do
     set CB = CB ∪ {(d_i, n_i)}
RETURN CB1(σ)(s̄) = h_⟨CB,σ⟩
```

Fig. 1. $CB1(\sigma)$ **Learning Algorithm [8, Defn 5.3.1].** $\overline{s} = \langle (d_i, n_i) \rangle_{i=1}^m$ is a training sample from $(D_N \times \{0,1\})^m$.

In [8] we took the similarity measure σ_H and measured the accuracy of $CB1(\sigma_H)$ on monomial target concepts. The results of some of these experiments, measuring the accuracy of $CB1(\sigma_H)$ on target concepts in the spaces

$M_{6,1}$, $M_{6,2}$ and $M_{6,3}$, are shown in Figure 2. Methodology for the experiments is described in [8]. The figure shows how concepts from $M_{6,1}$ are learnt more slowly than concepts from $M_{6,2}$, which are learnt more slowly than concepts from $M_{6,3}$.

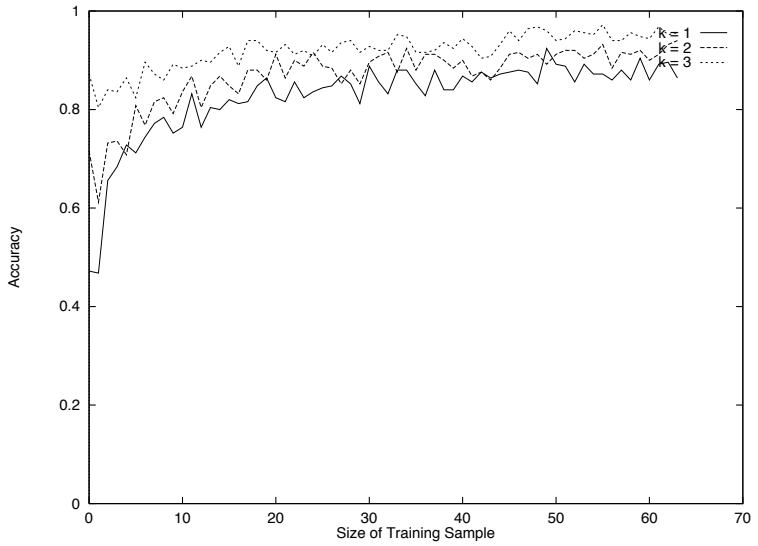

Fig. 2. [8, Fig 5.3] Average Learning Curve for Monomial Concepts $t \in M_{6,k}$ ($CB1(\sigma_H)$)

In addition, we observed that $CB1(\sigma_H)$ was not noticeably more efficient than an extremely simple 'rote-learning' algorithm L_2 [10] [8]. $CB1(\sigma_H)$ is also certainly less efficient for learning monomial target concepts than the 'standard' (non case-based) learning algorithm for monomials [18] [2], and [8] shows that $CB1(\sigma_H)$ is easily out-performed on monomial target concepts by an instance-based learner which changes the weights assigned to a weighted similarity measure (Figure 4, below). The performance of $CB1(\sigma_H)$ is therefore relatively poor.

This section will informally argue that the experimental results for $CB1(\sigma_H)$ on training samples for monomial target concepts are unlikely to be improved by *any* choice of similarity measure for $CB1(\sigma)$, and therefore that instance-based learners such as $CB1(\sigma)$ which use only a single, fixed definition of similarity are limited in the efficiency they can achieve. The section will define what is meant by an 'optimal' similarity measure. We will then define what we believe to be the optimal similarity measure as a function of the prior distribution π and prove its optimality within a restricted set of similarity measures. (We believe this to be indicative of its general optimality.) Finally, we instantiate the measure for a number of specific distributions, including those defining the monomial learning experiment.

Optimality is defined here in terms of the average case accuracy of the instance-based learner:

Definition 1. Optimal Similarity Measure. An optimal similarity measure with respect to a prior distribution π of target concepts in B_N and some family Σ of similarity measures, is a function $\sigma \epsilon \Sigma$ s.t. for any other similarity measure $\sigma' \epsilon \Sigma$, the accuracy of the instance-based learner $CB1(\sigma)$ will be no less than the accuracy of $CB1(\sigma')$.

$$\exists m_0 \epsilon \mathbb{N} \cdot \forall m \geq m_0 \cdot EA^{CB1(\sigma)}_{(\pi,\mu)}(m) \geq EA^{CB1(\sigma')}_{(\pi,\mu)}(m)$$

Richter [15] suggests that the similarity between two objects should be defined as the probability that the two objects share the same classification. He argues that this definition allows a meaningful semantics for 'similarity', and it is intuitively appealing that a case-based classifier should choose as near neighbours those exemplars which are most likely to share the same classification as the target object[2]. We apply Richter's observations by assuming the existence of a prior distribution π defined on the space B_N of possible target concepts which quantifies the *a priori* likelihood of the possible target concepts. The following similarity measure can then be defined in terms of this distribution π:

Definition 2. 'Canonical' Similarity Measure σ_π. The canonical similarity measure σ_π, defined with respect to a prior distribution π over the space B_N of possible target concepts, is the function over pairs of descriptions $d_1, d_2 \epsilon D_N$ whose value is precisely the probability according to π of choosing a target concept which assigns the same classification to d_1 and d_2.

$$\sigma_\pi(d_1, d_2) = \pi\{t \epsilon B_N | t(d_1) = t(d_2)\} \tag{7}$$

We conjecture that the canonical similarity measure σ_π will be the best choice of similarity measure for $CB1(\sigma)$ in the case where the prior distribution π holds; σ_π will be an optimal similarity measure w.r.t. the distribution π in the sense of Definition 1. Since it is not the numeric scores defined by σ_π but rather the retrieval ordering defined by the similarity measure that is important for classification, the definition of σ_π can be generalised to give the following condition which is conjectured to be sufficient for optimality:

Definition 3. Well-ordered Similarity Measure. A similarity measure $\sigma \epsilon (D_N \times D_N) \rightarrow [0,1]$ is well-ordered w.r.t. the prior distribution π iff, given three descriptions $d, d', d'' \epsilon D_N$, the similarity of d to d' is as great as the similarity of d to d'' only if the prior probability of a target concept assigning the same classification to d and d' is as great as the probability of equal classification of d and d'':

$$\forall d, d', d'' \epsilon D_N \cdot$$
$$\sigma(d, d') \geq \sigma(d, d'') \rightarrow \pi\{t \epsilon B_N | t(d) = t(d')\} \geq \pi\{t \epsilon B_N | t(d) = t(d'')\}$$

[2] Very similar comments precede Faltings' recent formalisation [4].

Using the 'retrieval ordering' notation $\succeq_d^\sigma$ and recalling the definition of σ_π, this condition can equivalently be written as:

$$\forall d, d', d'' \epsilon D_N \cdot d' \succeq_d^\sigma d'' \rightarrow d' \succeq_d^{\sigma_\pi} d''$$

The canonical similarity σ_π will of course be well-ordered w.r.t. π, and might be seen as the prototype of the set of similarity measures which are well-ordered with respect to that distribution. There will be many other functions which are also well-ordered for that distribution. Some of these, although computing different numerical scores, will define exactly the same retrieval ordering as σ_π; $\succeq_d^\sigma \equiv \succeq_d^{\sigma_\pi}$ or $d' \succeq_d^\sigma d'' \leftrightarrow d' \succeq_d^{\sigma_\pi} d''$. Others will, however define a different retrieval ordering, but only in the sense that they will define an ordering that is more strict than $\succeq_d^{\sigma_\pi}$; $d' \succeq_d^\sigma d'' \rightarrow d' \succeq_d^{\sigma_\pi} d''$, but not $d' \succeq_d^{\sigma_\pi} d'' \rightarrow d' \succeq_d^\sigma d''$ for some descriptions $d, d' \epsilon D_N$.

Unfortunately, it has not been possible to prove in the general case that 'well orderedness' is either necessary or sufficient for optimality. It has been possible, however, to define a special case where well-orderedness is a necessary and sufficient condition. The proof is straightforward if we consider only a *special subset of the space of all possible similarity measures.* This space is defined as follows:

Definition 4. Separated Simialrity Measure. For any problem description $d \epsilon D_N$, a separated similarity measure σ assigns a different similarity score to the similarity of d to each description $d' \epsilon D_N$:

$$\forall d, d', d'' \epsilon D_N \cdot \sigma(d, d') = \sigma(d, d'') \rightarrow d' = d''$$

No retrieval ordering defined by a separated similarity measure will have an equivalence class of size greater than one; the retrieval ordering will therefore choose an unique maximum element from any set of descriptions. This restriction may seem harsh, but note that any similarity measure σ can be 'separated' simply by adding small increments to the similarity scores in order to break up any equivalence classes of size greater than one appearing in a retrieval ordering $\succeq_d^\sigma$. If σ is a well-ordered similarity measure, then any equivalence classes in a retrieval ordering $\succeq_d^\sigma$ will contain descriptions which are equally likely to have the same classification as d. In this case, the single nearest neighbour chosen by σ', a 'separated' version of σ, will therefore be just as likely to have the same classification as the test instance as any of the descriptions in the nearest neighbour set defined by the original well-ordered measure σ. $CB1(\sigma)$ and $CB1(\sigma')$ will *not* necessarily make the same classifications however. $CB1(\sigma')$ will classify a test instance positively iff the nearest neighbour d^{NN} according to the separated measure σ' is a positive instance of the target concept. $CB1(\sigma)$ on the other hand classifies the same instance positively only if *all* near neighbours calculated by σ are positive instances. This set will include d^{NN}, but may also include a number of other descriptions. The overall relationship between the usefulness in classification of a similarity measure and its separated version is therefore not clear.

However, precisely because a separated similarity measure will always define a singleton nearest neighbour set, it is possible to show that a similarity measure which is both separated and well-ordered will give $CB1(\sigma)$ a better accuracy than any other separated similarity measure.

Theorem 5. *Let π be any prior distribution on B_N. A separated similarity measure σ will be optimal with respect to π and the space of separated similarity measures if and only if σ is well-ordered with respect to π.*

Proof. The proof is given in [8] (Theorem C.0.1). It outline, the proof considers only $E(L, \overline{x}, x)$, defined $E(L, \overline{x}, x) = \pi\{t \epsilon B_N | L(\overline{x}_t)(x) \neq t(x)\}$. That is, the proof is achieved by fixing the training sample $\overline{x}$ and the test instance x, and summing the expected error as the target concept is varied. Sufficiency is almost directly guaranteed by Definition 3. Necessity is proven by showing that any separated similarity measure σ which is not well-ordered will have accuracy strictly less than that of a 'separated' version σ' of the canonical similarity measure σ_π, as follows. From the definition of σ_π, then $E(CB1(\sigma), \overline{x}, x) \geq E(CB1(\sigma'), \overline{x}, x)$ for any $\overline{x}$ and x. In addition, since σ is not well-ordered, there must be some description x where $E(CB1(\sigma), \overline{x}, x) > E(CB1(\sigma'), \overline{x}, x)$ and hence, overall, the accuracy of $CB1(\sigma)$ must be strictly less than that of $CB1(\sigma')$. □

Assuming for the moment that σ_π and all other well-ordered similarity measures are optimal or close to optimal for instance-based learning, the remainder of this initial section will explore the different forms that σ_π takes as π is instantiated as different prior distributions. Specifically, we will consider the prior distribution which is uniform on B_N, and those which are uniform on M_N and $M_{N,k}$ respectively, but zero elsewhere[3]. This will provide a tentative set of answers to the question, 'What is the best similarity measure for learning B_N, M_N and $M_{N,k}$?' and, in particular, suggest that in fact σ_H is as good as any other similarity measure that might be chosen for those concept spaces.

Proposition 6. *Given the 'maximum entropy' prior distribution U which is uniform over all possible target concepts $t \epsilon B_N$, the canonical similarity measure w.r.t. U is defined by the following equation:*

$$\sigma_U(d_1, d_2) = \begin{cases} 1 \text{ if } d_1 = d_2 \\ \frac{1}{2} \text{ if } d_1 \neq d_2 \end{cases}$$

Proof. Consider separately the two cases where $d_1 = d_2$ and $d_1 \neq d_2$. When $d_1 = d_2$, then $\pi\{t \epsilon B_N | t(d_1) = t(d_2)\} = 1$ trivially, since $t(d_1) = t(d_2)$ for all $t \epsilon B_N$. Where d_1 and d_2 are distinct, however, consider that there are four distinct classifications for the pair of variables $\langle d_1, d_2 \rangle$, one where both are 0,

[3] We consider monomial target concepts here out of continuity from our previous work [10] [11]. We do not claim that learning monomial concepts is a serious or 'realistic' problem for case-based learners, but rather that monomials provide a sufficiently constrained domain of study where methods of analysis can be developed before applying them to more developed models of case-based learning.

one where d_1 is 0 and d_2 is 1, and so on. For each classification there are $2^{|D_N|-2}$ concepts in B_N consistent with that classification, since there are $(|D_N|-2)$ remaining descriptions, each of which may be assigned either 0 or 1 by a target concept. Thus there are 2^{2^N-1} out of 2^{2^N} concepts in B_N which assign the same classification to d_1 and d_2 $(t(d_1) = t(d_2))$, giving a value of $\frac{1}{2}$ for σ_π on distinct descriptions d_1 and d_2 in the case where all concepts in B_N are equally likely. □

Proposition 7. *Given the prior distribution π_M which is uniform on monomial target concepts $t \epsilon M_N$ and zero elsewhere, the canonical similarity measure w.r.t. π_M is defined by the following equation:*

$$\sigma_{\pi_M}(d_1, d_2) = 1 - \frac{(2^{N+1} - 2^{N-\delta+1})}{3^N}$$

where δ is the number of bits of the representation on which d_1 and d_2 differ; $\delta = (N - N.\sigma_H(d_1, d_2))$.

Proof. Let δ be the number of bits on which d_1 and d_2 differ. Consider that out of 3^N monomial concepts, there will be 2^N concepts which have value 1 on the description d_1. Of these, any function defined by a monomial description which only contains literals corresponding to the $N - \delta$ bits on which d_1 and d_2 agree will also have value 1 on d_2. Hence the number of monomial concepts which have value 1 on d_1 and value 0 on d_2 will be $(2^N - 2^{N-\delta})$. Since the set of concepts which have value 1 on d_1 and value 0 on d_2 is disjoint from the set of concepts which have value 0 on d_1 and value 1 on d_2, this gives a total of $(3^N - 2(2^N - 2^{N-\delta}))$ monomial concepts on N variables which assign the same classification to d_1 and d_2. Hence result. □

Proposition 8. *Given the prior distribution π_{Mk} which is uniform on k-literal monomial target concepts $t \epsilon M_{N,k}$ and zero elsewhere, the canonical similarity measure w.r.t. π_{Mk} is defined by the following equation:*

$$\sigma_{\pi_{Mk}}(d_1, d_2) = 1 - \frac{1}{2^{k-1}} \cdot \left(1 - \frac{\binom{N-\delta}{k}}{\binom{N}{k}} \right)$$

where δ is the number of bits of the representation on which d_1 and d_2 differ.

Proof. The derivation is very similar to that of Proposition 7. In total there are $2^k \binom{N}{k}$ k-literal monomial expressions e.g. [8, Propn 5.1.2], while $\binom{N}{k}$ of these will be positive on a specific description $d_1 \epsilon D_N$. Of these, exactly $\binom{N-\delta}{k}$ concepts will also classify d_2 positively. Hence result. □

Propositions 6, 7 and 8 carry a number of implications. Proposition 6 suggests that in the maximum entropy case, then the 'nearest neighbour' assumption of

case-based reasoning is invalid, and that apart from an exemplar whose description exactly matches a problem instance, *all* exemplars are equally (un)informative about the correct classification of the problem instance, whether or not their descriptions are syntactically similar to the problem instance. This illustrates the fundamental idea from learning theory that all generalisation strategies are equivalent when all possible classifications of the example space are equally likely c.f. [21] [17].

Proposition 7 on the other hand describes a similarity measure which correlates negatively with δ, the number of values on which two descriptions disagree. The measure σ_{π_M} has value 1 when $\delta = 0$, has value $1 - 2.\frac{(2^N-1)}{3^N}$ when $\delta = N$ and strictly decreases for values of δ between these two limits. Since the 'feature count' similarity measure σ_H also has this property of decreasing strictly in δ (and, like σ_{π_M}, the value of σ_H is *determined* only by N and δ), then $d' \succeq_d^{\sigma_H} d'' \leftrightarrow d' \succeq_d^{\sigma_{\pi_M}} d''$. The orderings defined on the example space by the two measures are identical, and $h_{\langle CB,\sigma_H\rangle} = h_{\langle CB,\sigma_{\pi_M}\rangle}$ on all arguments and for all case-bases CB. *A fortiori,* σ_H is a well-ordered (and hence, according to our conjecture, optimal) similarity measure for π_M.

The behaviour in δ of the similarity measure described by Proposition 8 is a little more subtle. As long as $\delta \leq N - k$ then the value of the similarity measure $\sigma_{\pi_{Mk}}$ decreases strictly in δ. If instead $\delta > N - k$ then no function defined by a k-literal monomial expression can classify both d_1 and d_2 positively. In this case $\binom{N-\delta}{k}$ has value 0 and the value of $\sigma_{\pi_{Mk}}$ is given by $\frac{2^{k-1}-1}{2^{k-1}}$ which is independent of δ. This suggests that all exemplars whose descriptions are sufficiently dissimilar to d_1 will be equivalent for the purposes of classifying d_1. Provided d_1 and d_2 agree on most features however, the usefulness of an exemplar with description d_2 for classifying d_1 will *increase* as δ decreases. Although $\sigma_{\pi_{Mk}}$ decreases only non-strictly in δ, this is still sufficient to ensure $d' \succeq_d^{\sigma_H} d'' \rightarrow d' \succeq_d^{\sigma_{\pi_{Mk}}} d''$, and σ_H is also a well-ordered similarity measure for the distribution π_{Mk}.

This discussion therefore supports (but does not strictly prove) the assertion that *no* similarity measure can do better than σ_H in the experiments described in [8] and that the learning curves shown in Figure 2 represent the limit of efficiency that can be achieved by $CB1(\sigma)$, using *any* possible similarity measure, when learning monomial target concepts.

4 Optimal 'Learnt' Similarity Measures

§3 showed there are experiments, such as learning monomial target concepts, where an instance-based learner with a fixed measure of similarity has apparently poor performance which cannot be overcome by any choice of similarity measure. On the other hand, several theoretical studies [12] [5] show that the power of instance-based learning is extended if the learner is allowed to change its definition of similarity. This section therefore considers the 'best' representation for similarity in this class of instance-based learners which are able to learn

a definition of similarity from the training sample rather than using a single, fixed definition of similarity. The section proceeds as follows:

1. Firstly, we illustrate the gains in efficiency achieved by allowing the similarity measure to vary by describing a 'similarity learning' IBL algorithm called $VS\text{-}CBR$, which learns monomial target concepts more efficiently than $CB1(\sigma)$, even using an 'optimal measure' for $CB1(\sigma)$ such as σ_H.
2. Secondly, we consider ways in which $VS\text{-}CBR$ itself might be inefficient, and in response define $VS\text{-}CBR3$.
3. Finally, we apply the approach of §3, which says that 'two descriptions are similar if they are likely to have the same classification' and from this derive an apparently new IBL algorithm called $VS\text{-}CBR2$.

This note describes work in progress and we have not yet been able to carry out any evaluation of $VS\text{-}CBR2$ and $VS\text{-}CBR3$. Instead, we conjecture a number of hypotheses which we expect to be confirmed in our forthcoming evaluation of these algorithms.

Figure 3 shows the IBL algorithm $VS\text{-}CBR$. This algorithm is *specifically* tailored to monomial target concepts; it will not correctly identify non-monomial targets. Its operation is straightforward [19] [11]:

- Only the first positive example in the training sample is added to the case-base. All other positive examples are discarded.
- All negative examples in the training sample are added to the case-base.
- Only binary weights are assigned to $\sigma_{\overline{w}}$; the algorithm searches the space of 2^N similarity measures corresponding to the weight vectors $\overline{w} \epsilon \{0,1\}^N$.
- All weights are 1 initially. A weight changes to zero iff two *positive* examples are observed which disagree on that bit of the representation.

Figure 4 shows the learning curve of $VS\text{-}CBR$ for target concepts in $M_{6,1}$, $M_{6,2}$ and $M_{6,3}$. Experimental conditions can again be found in [8]. The figure shows the learning curves starting from the same initial accuracies as observed in Figure 2. The learning curves however achieve a close to perfect classification accuracy much more quickly than the learning curves for $CB1(\sigma_H)$ shown in that figure. This emphasises the benefit of changing the similarity measure during learning. Figure 4 also shows that, in contrast to Figure 2, the learning curve for target concepts in $M_{6,1}$ approaches perfect accuracy more rapidly than the curve for target concepts in $M_{6,2}$ which in turn converges more rapidly than the curve for target concepts in $M_{6,3}$.

$VS\text{-}CBR$ is an example of the large family of IBL algorithms which change their similarity measure but only in a limited way, namely through choosing different weights for a similarity measure similar to equation (3) [20]. However, we may question whether a weighted similarity measure is actually the best or 'optimal' way to express knowledge about similarity inferred from the training sample, just as in §3 we questioned whether σ_H was the best similarity measure for $CB1(\sigma)$ in experiments such as those shown in Figure 2. This premise certainly seems questionable in the case of $VS\text{-}CBR$ since this algorithm clearly

```
forall 1 ≤ i ≤ N, n ϵ {0,1} set f[i,n] = 1
set CB = ∅
for i = 1 to m do
    if n_i = 1 then
        if ¬∃d ϵ D_N · (d,1) ϵ CB then set CB = CB ∪ {(d_i,1)}
        for j = 1 to N do
            set f[j, 1 − (d_i)_j] = 0
    else
        set CB = CB ∪ {(d_i,0)}
forall 1 ≤ i ≤ N
    if f[i,0] = 1 ∨ f[i,1] = 1 then
        set w_i = 1
    else
        set w_i = 0
RETURN VS-CBR(s̄) = h_⟨CB,σ_w̄⟩
```

Fig. 3. VS-CBR **Learning Algorithm for Concepts in** M_N **[19, Fig 4].** $\overline{s} = \langle (d_i, n_i) \rangle_{i=1}^{m}$ is a training sample from $(D_N \times \{0,1\})^m$.

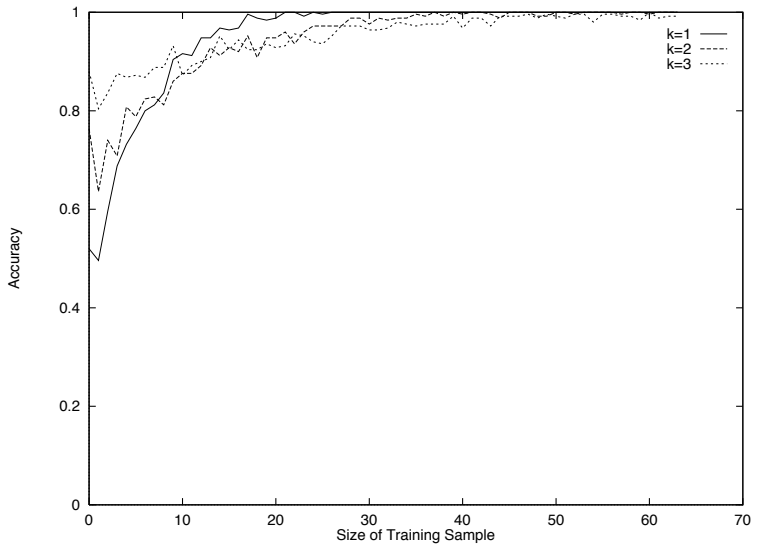

Fig. 4. [8, Fig 6.1] Average Learning Curve for Monomial Concepts $t \epsilon M_{6,k}$ (VS-CBR)

discards information about the target concept when it computes the weight vector $\overline{w}$. After processing any sample $\overline{s}$, $f[i,n] = 1$ only if no positive exemplar d_{pos} has been processed such that $(d_{pos})_i = (1-n)$ and therefore all observed positive examples have value n on bit i of the representation. $\sigma_{\overline{w}}$ has a value 1 in a certain position iff that bit of the representation has the same value in all positive examples in the training sample, while f encodes the information of what that value is. There is therefore a loss of information when f is translated to the similarity measure $\sigma_{\overline{w}}$; this is also shown by the fact that the state of the array f determines a single corresponding weight vector $\overline{w} \epsilon \{0,1\}^N$, while the converse is not true.

A better approach might be if the information in f is used directly. In [11], we noted that the information held in the f array directly corresponds to the monomial hypothesis chosen by the standard algorithm for monomials [18] [2] (which we call 'M'). This information can be employed more effectively by the following similarity measure σ_U^M which is defined for a single monomial function with monomial representation U, and which states that two descriptions are similar iff they are given the same classification by the concept represented by U:

$$\sigma_U^M(d_1, d_2) = \begin{cases} 1 \; if \quad h_U(d_1) = h_U(d_2) \\ 0 \; otherwise \end{cases} \tag{8}$$

Figure 5 shows the algorithm $VS\text{-}CBR3$ which computes this similarity measure in a very similar fashion to $VS\text{-}CBR$. This algorithm will now be very similar to M; in fact the hypothesis will be exactly the same as that of M provided the training sample contains at least one negative and one positive example.

```
forall 1 ≤ i ≤ N, n ϵ {0,1} set f[i,n] = 1
set CB = ∅
set U = ∅
for i = 1 to m do
    if n_i = 1 then
        if ¬∃d ϵ D_N · (d,1) ϵ CB then set CB = CB ∪ {(d_i,1)}
        for j = 1 to N do
            set f[j, 1 − (d_i)_j] = 0
    else
        set CB = CB ∪ {(d_i,0)}
forall 1 ≤ i ≤ N
    if f[i,0] = 1 then set U = U ∪ {ū_i}
    if f[i,1] = 1 then set U = U ∪ {u_i}
RETURN VS-CBR3(s̄) = h_⟨CB,σ_U⟩
```

Fig. 5. $VS\text{-}CBR3$ **Learning Algorithm for Concepts in** M_N**.** $\overline{s} = \langle (d_i, n_i) \rangle_{i=1}^m$ is a training sample from $(D_N \times \{0,1\})^m$.

$VS\text{-}CBR3$ suggests one way of improving on the performance of $VS\text{-}CBR$. However, §3 argued that the 'meaning' of the similarity of two objects should

be the probability that the two objects are assigned the same classification; this can be developed to give yet another view of the problem of learning similarity measures. Definition 2 defined a 'canonical' similarity measure, based upon this intuition, which makes use of *a priori* probabilities of target concepts. An *adaptive* definition for similarity can be based instead on *a posteriori* probabilities of target concepts, given what is actually known about the target concept from the training sample. That is, instead of the similarity measure σ_π defined as the unconditional probability $\pi\{t \epsilon B_N | t(d_1) = t(d_2)\}$, an adaptive similarity measure can be defined by the *conditional* probability of two objects being classified the same *given that the target concept is one which produces the labellings observed in the training sample:*

$$\begin{aligned} \sigma_\pi^{\overline{s}}(d_1, d_2) &= \pi\{t \epsilon B_N \wedge \overline{x}_t = \overline{s} | t(d_1) = t(d_2)\} \\ &= \frac{\pi\{t \epsilon B_N | (t(d_1) = t(d_2)) \wedge (\overline{x}_t = \overline{s})\}}{\pi\{t \epsilon B_N | \overline{x}_t = \overline{s}\}} \end{aligned} \tag{9}$$

where $\overline{x} \epsilon (D_N)^*$ is the (unlabelled) sample of the example space corresponding to $\overline{s}$ (i.e. the sample $\overline{x} \epsilon (D_N)^*$ s.t. $\overline{x}_t = \overline{s}$ for some concept t consistent with the labellings on $\overline{s}$.)

As with §3, the prior distribution π can be instantiated in order to consider what the best 'instance-based' learner for a particular prior might look like. Taking again the special case of monomial target concepts and the distribution π_M which is uniform on the monomials and zero on all other concepts, then the conditional probability is easily reckoned as the proportion of monomials consistent with the training sample which agree on the two descriptions. Hence, if a set VS of concepts is defined as the set of monomials which are consistent with the training sample $\overline{s}$, then $\sigma_{\pi_M}^{\overline{s}}$ is instantiated as the measure σ_{VS}, defined as follows:

$$\sigma_{VS}(d, d') = \frac{|\{t \epsilon VS | t(d) = t(d')\}|}{|VS|}$$

```
set CB = ∅
set VS = M_N
for i = 1 to m do
    if n_i = 1 then
        set VS = VS \ {t ϵ M_N | t(d_i) = 0}
    else
        set VS = VS \ {t ϵ M_N | t(d_i) = 1}
    if ∀(d, b) ϵ CB · σ_VS(d, d_i) < 1 then
        set CB = CB ∪ {(d_i, n_i)}
RETURN VS-CBR2(s̄) = h_⟨CB,σ_VS⟩
```

Fig. 6. VS-$CBR2$ **Learning Algorithm for Concepts in** M_N **[8, Defn 8.3.1].** $\overline{s} = \langle (d_i, n_i) \rangle_{i=1}^m$ is a training sample from $(D_N \times \{0,1\})^m$.

Figure 6 defines the IBL algorithm VS-$CBR2$ which calculates the set VS naïvely. It is presented for illustration; a more efficient way of calculating the set VS must be found, since Figure 6 as it stands defines an algorithm which has exponential execution time in the worst case. (It should be possible to calculate the similarity measure σ_{VS} more efficiently from properties of the training sample or from a more efficient representation such as a version space [13].)

The algorithm initialises VS to the set of all monomial concepts M_N; the initial similarity measure used by VS-$CBR2$ is therefore equivalent to σ_H (w.r.t. the retrieval orderings defined by the two measures; see Proposition 7). On the other hand, if there is only a single monomial concept consistent with the training sample, $VS = \{t\}$ for some $t \epsilon M_N$, then the similarity measure σ_{VS} is reduced to the statement that two descriptions are similar iff they are given the same classification by t. i.e. $\sigma_{VS} = \sigma_t$ where:

$$\sigma_t(d_1, d_2) = \begin{cases} 1 \text{ if } t(d_1) = t(d_2) \\ 0 \text{ otherwise} \end{cases} \tag{10}$$

In addition, it is clear that, if the training sample contains enough examples for the (non case-based) standard monomial learner M to identify the target concept, then VS-$CBR3$ also converges to the same final similarity measure σ_t (c.f. equation (8)). These two extremes have a strong intuitive appeal. However, the trajectory of σ_{VS} between the two is less clear.

We have not yet carried out any evaluation of VS-$CBR2$ and VS-$CBR3$. We would expect the following points to be demonstrated in empirical experiments or average-case analysis comparing $CB1(\sigma_H)$, VS-CBR, VS-$CBR2$ and VS-$CBR3$:

1. Since we argue that the similarity measures used by VS-$CBR2$ and VS-$CBR3$ make better use of the information that may be derived from the training sample than VS-CBR, we would expect both these algorithms to learn with more rapidly increasing accuracy than either VS-CBR or $CB1(\sigma_H)$.
2. Secondly, we would expect that VS-$CBR2$ is at least as accurate as VS-$CBR3$ 'on average'; the similarity measure used by the former is an instantiation of equation (9), which we believe to be the 'best' way of calculating similarity if a suitable prior is known. Comparing VS-$CBR2$ and VS-$CBR3$ therefore gives a very interesting test of the conjectures of optimality put forth in this paper.
3. Finally, we would also expect at least VS-$CBR3$ to follow the trend of Figure 2 rather than Figure 4; we would expect this algorithm to learn monomial target concepts with the smallest monomial expressions more slowly than those represented by a larger monomial expression. This follows from the logic of a 'constituent analysis' [8] [11]. Once the correct similarity measure has been identified by VS-$CBR3$, then the learner needs only one positive exemplar and one negative exemplar in the case-base to correctly identify the target concept. It is therefore easily shown that the number of examples needed in the worst case for VS-$CBR3$ to populate its case-base is constant

i.e. is independent of the choice of monomial target concept. As in VS-CBR however, the number of examples needed by VS-$CBR3$ to infer a good similarity measure increases linearly in the number of irrelevant variables [8] [11]. Thus, overall, we would also expect the sample complexity of VS-$CBR3$ to increase in the number of irrelevant variables, as indicated. The picture for VS-$CBR2$ is less clear, since we have not yet been able to characterise the hypotheses of this learner.

In addition, the comparison of VS-$CBR2$ and VS-$CBR3$ with other 'efficient' algorithms for learning monomial target concepts, such as the 'standard' algorithm [18] and Version Space [13], would provide a very useful comparison of instance-based and non instance-based methods and demonstrate whether inductive learning using the 'case-based representation' $\langle CB, \sigma \rangle$ has any fundamental differences (w.r.t. sample complexity and efficiency) from other (non case-based) learning algorithms.

5 Conclusions

The main results of this paper are presented in §3 and §4. §3 defined the 'canonical' similarity measure σ_π and the property of 'well-orderedness' which we believe to define the 'optimal' similarity measure that can be chosen for the instance-based learner $CB1(\sigma)$ given some prior distribution of target concepts. We then showed, by instantiating σ_π, that the straightforward 'feature-counting' similarity measure σ_H is well-ordered for the distributions U, π_M and π_{M_k} and therefore we would expect that no other choice of similarity measure would improve the performance of $CB1(\sigma)$ on the concept spaces B_N, M_N and $M_{N,k}$ beyond that already measured for $CB1(\sigma_H)$. It has already been shown (e.g. in Figures 2 & 4) that the performance of $CB1(\sigma_H)$ can easily be bettered, at least on monomial target concepts, by VS-CBR. This suggests that the policy of using only a fixed similarity measure, as in $CB1(\sigma)$, places a strong restriction on an instance-based learner and that there is at least one case where an instance-based learner which can change its similarity measure is strictly more accurate in learning than an IBL algorithm with a fixed similarity measure.

Future work might explore whether σ_H is also well-ordered for more general, 'realistic' sets of target concepts. It would be interesting to extend the work presented here and calculate the canonical similarity measure for the prior distributions corresponding to the set of concepts represented, for example, by linear threshold functions, or by k-term DNF, k-DNF, k-term CNF and k-CNF propositional formulae.

§4 shows how the canonical similarity measure of §3 also suggests a 'new' instance-based learning algorithm VS-$CBR2$ for learning monomial concepts which we believe to be more efficient than the algorithm VS-CBR [19] which we have studied previously [8] [11]. We also defined VS-$CBR3$, a further variant of VS-CBR which we also hope to be more efficient than the original algorithm. Unlike §3 where we found evidence suggesting that the 'obvious' similarity measure σ_H is as good as any other for the experiments considered there,

the similarity measures used by VS-$CBR2$ and VS-$CBR3$ are quite different from that used by VS-CBR. Whereas VS-CBR uses a conventional 'weighted' similarity measure which sums similarity along a number of different dimensions, VS-$CBR2$ and VS-$CBR3$ use a 'non-dimensional' or 'context sensitive' similarity measure where the dimensions of the representation do not contribute independently to the calculation of similarity.

Detailed investigations of the learning curves of these new algorithms must await further work. It seems clear however that VS-$CBR2$ and VS-$CBR3$ will both learn more efficiently than VS-CBR. If this is confirmed to be the case, then this raises a number of points about the nature of instance-based learning:

1. The comparison of VS-CBR, VS-$CBR2$ and VS-$CBR3$, where we fully expect VS-$CBR2$ and VS-$CBR3$ to be found to be more efficient than VS-CBR, suggests that in the case of monomial target concepts at least, the weighted similarity measure $\sigma_{\overline{w}}$ is not the most conducive representation for information about similarity between descriptions *in the special case of monomial target concepts*. This raises the question of whether *general purpose* IBL algorithms could also improve their accuracy of learning if they used some other form of similarity measure than the weighted similarity measure $\sigma_{\overline{w}}$. It would be interesting to instantiate equation (9) with a prior distribution corresponding to the bias toward 'naturally occurring classification problems' assumed in general purpose learners such as IB4 [1] or PEBLS [3], and to compare the learner derived that way with existing instance-based learners.
2. However, if weighted, 'dimensional' similarity measures are abandoned in order to achieve better generalisation accuracy and more efficient learning, then VS-$CBR2$ and VS-$CBR3$ demonstrate that some other useful properties of instance-based learning can be lost. IBL, like other forms of case-based reasoning, normally has the advantage that the system can give a straightforward explanation of its inferences; "the solution was chosen because the case which is most similar to the new problem contains that outcome". The similarity measures used by VS-$CBR2$ and VS-$CBR3$ do not however support this kind of explanation since they both have the property, also inherent in equation (9), that all positive exemplars are equally similar to any point in the example space, and all negative exemplars are also equally similar to any point. For example, in the case of VS-$CBR2$, for any case-base CB and the set VS of monomial concepts consistent with CB:

$$\forall d, d_1^+, d_2^+ \epsilon D_N \cdot \\ (d_1^+, 1) \epsilon CB \wedge (d_2^+, 1) \epsilon CB \rightarrow \sigma_{VS}(d, d_1^+) = \sigma_{VS}(d, d_2^+) \qquad (11)$$

$$\forall d, d_1^-, d_2^- \epsilon D_N \cdot \\ (d_1^-, 0) \epsilon CB \wedge (d_2^-, 0) \epsilon CB \rightarrow \sigma_{VS}(d, d_1^-) = \sigma_{VS}(d, d_2^-) \qquad (12)$$

In VS-$CBR2$ and VS-$CBR3$, nearly all the information transferred from the training sample to the hypothesis $h_{\langle CB, \sigma \rangle}$ is expressed through the similarity measure σ rather than the case-base CB. Any of the positive exemplars are

interchangeable with one another, and indeed all but one of the positive exemplars can be deleted without changing the concept represented; the same applies also to the negative exemplars. It is clear that the classifier can no longer be said to reflect a policy of lazy generalisation (since most of the information in the exemplars of CB is redundant and most of the knowledge about the target concept has already been compiled into the similarity measure) and therefore any advantage which follows from storing an extensional representation of the target concept in the case-base may be lost (since the instance-based classifier is no longer able to distinguish between exemplars of the same classification).

The ideas presented in this paper therefore emphasise that there is a tension in instance-based learners between maximising predictive accuracy and exploiting any advantages which follow from lazy generalisation.

References

1. D W Aha. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies*, 36:267–287, 1992.
2. M Anthony and N Biggs. *Computational Learning Theory.* Cambridge University Press, 1992.
3. S Cost and S Salzberg. A weighted nearest neighbour algorithm for learning with symbolic features. *Machine Learning*, 10(1):37–66, March 1993.
4. B Faltings. Probabilistic indexing for case-based prediction. In E Plaza and D Leake, editors, *Case-Based Reasoning Research and Development. Proceedings of the Second International Conference on Case-Based Reasoning, ICCBR-97. Providence, RI, July 1997.*, Lecture Notes in Artificial Intelligence vol. 1266, pages 611–622. Springer Verlag, 1997.
5. C Globig, K P Jantke, S Lange, and Y Sakakibara. On case-based learnability of languages. *New Generation Computing*, 15(1), 1997.
6. C Globig and S Lange. Case-based representability of classes of boolean functions. In *ECAI-96: Proceedings of Twelfth European Conference on Artificial Intelligence*, pages 117–121, 1996.
7. C Globig and S Wess. Symbolic learning and nearest-neighbour classification. In H-H Bock, W Lenski, and M M Richter, editors, *Information Systems and Data Analysis: Prospects, Foundations, Applications. Proceedings of the 17th Annual Conference of the Gesellschaft fur Klassification e.V. University of Kaiserslautern, March 3-5, 1993.* Springer-Verlag, 1994.
8. A D Griffiths. *Inductive Generalisation in Case-Based Reasoning Systems.* PhD thesis, Published as Technical Report YCST-97-02, Department of Computer Science, University of York, York YO1 5DD, UK, 1997.
9. A D Griffiths and D G Bridge. Formalising the knowledge content of case memory systems. In I D Watson, editor, *Progress in Case-Based Reasoning: Proceedings of the First UK Workshop on Case-Based Reasoning, Salford UK, Jan 1995*, Lecture Notes in Artificial Intelligence vol. 1020, pages 32–41. Springer Verlag, 1995.
10. A D Griffiths and D G Bridge. A yardstick for the evaluation of case-based classifiers. In *(Forthcoming) Proceedings of Second UK Workshop on Case-Based Reasoning, Salford UK*, Apr 1996.

11. A D Griffiths and D G Bridge. PAC analyses of a 'similarity learning' IBL algorithm. In E Plaza and D Leake, editors, *Case-Based Reasoning Research and Development. Proceedings of the Second International Conference on Case-Based Reasoning, ICCBR-97. Providence, RI, July 1997.*, Lecture Notes in Artificial Intelligence vol. 1266, pages 445–454. Springer Verlag, 1997.
12. K P Jantke. Case-based learning in inductive inference. In *COLT92: Proceedings of the Fifth ACM Workshop on Computational Learning Theory, July 92, Pittsburgh PA*, pages 218–223. ACM Press, 1992.
13. T M Mitchell. Generalisation as search. *Artificial Intelligence*, 18(2):203–226, 1982.
14. M M Richter. Classification and learning of similarity measures. In *Proceedings of the Sixteenth Annual Conference of the German Society for Classification (Gesellschaft für Klassifikation e.V.)*. Springer Verlag, 1992.
15. M M Richter. On the notion of similarity in case-based reasoning. In R della Garcia, R Kruse, and R Viertl, editors, *Workshop Papers: Mathematical and Statistical Methods in Artificial Intelligence, September 1994, Udina, Italy*, Courses and Lectures: International Centre for Mechanical Sciences, Vol 363, pages 171 – 183. Springer Verlag, 1995.
16. M M Richter and S Wess. Similarity, uncertainty and case-based reasoning in PATDEX. In R S Boyer, editor, *Automated Reasoning – Essays in Honour of Woody Bledsoe*, pages 249–265. Kluwer, 1991.
17. C Schaffer. A conservation law for generalization performance. In *Machine Learning: Proceedings of Eleventh International Conference ML94*, pages 259–265. Morgan Kaufmann, 1994.
18. L G Valiant. Deductive learning. *Philosophical Transactions of the Royal Philosophical Society of London A*, 312:441–446, 1984.
19. S Wess and C Globig. Case-based and symbolic classification - A case study. In S Wess, K-D Althoff, and M M Richter, editors, *Topics in CBR: Selected papers from the First European Workshop on Case-Based Reasoning - EWCBR-93, Kaiserslautern, Germany, November '93*, Lecture Notes in Computer Science vol. 837, pages 77–91. Springer-Verlag, 1994.
20. D Wettschereck, D W Aha, and T Mohri. A review and comparative evaluation of feature weighting methods for lazy learning algorithms. Technical Report AIC-95-012, Navy Center for Applied Research in AI, Naval Research Laboratory, Washington, DC 20375-5337, USA, 1995.
21. D H Wolpert. On the connection between in-sample testing and generalisation error. *Complex Systems*, 6:47–94, 1992.