'Practical' PAC Analyses of Instance-Based Learning Algorithms [1]

A. D. Griffiths
Department of Computer Science, University of York, York, UK
TONY@MINSTER.YORK.AC.UK

D. G. Bridge
Department of Computation, University College, Cork, Republic of Ireland
D.BRIDGE@CS.UCC.IE

Abstract This paper compares the efficiency of two instance-based learners, one which can change its similarity measure and one which cannot, as they learn monomial target concepts. Using this example, we discuss the methodology of theoretical comparisons of learning algorithms, and describe our approach as a 'practical' PAC analysis. This is contrasted with conventional applications of PAC learning which focus on the properties of the concept space rather than specific learning algorithms. PAC results and informal considerations suggest that a learner with a larger hypothesis space will often be less efficient than a learner with a smaller one, although this is not strictly required by the PAC results. This paper therefore proposes that comparing the effective hypothesis spaces of the different learning algorithms gives useful information about the learning behaviour of those learners. Our results support this; where we have been able to derive results describing the hypothesis space of instance-based learners, a larger hypothesis space is predictive of a less efficient learning algorithm. We conclude that the instance-based learners studied here are examples of learning algorithms where there is a correlation between the upper bounds on the sample complexity of a learner given by PAC results and the (average-case) efficiency, and that the relationship between these two quantities should be investigated further.

1 Introduction

This paper describes the analysis of instance-based learning algorithms (IBL) within the PAC learning framework [13] [2], and discusses the role played in this analysis by knowledge of the hypothesis space of the learner. The PAC framework is concerned with concept learning, and in our work we consider concepts defined on the example space of N-bit binary vectors $D_N \hat{=} \{0,1\}^N$. The set of all concepts that might be defined on D_N is written B_N, $B_N \hat{=} (D_N \rightarrow \{0,1\})$.

The efficiency of a learner is measured in the PAC framework by its sample complexity:

[1] This work was carried out while the first author was studying for a DPhil; the support of an EPSRC research stipend during this time is gratefully acknowledged. In addition, we thank the anonymous reviewer for his/her helpful comments.

Definition 1 Sample Complexity [2]. *The sample complexity $m_L(t, \delta, \epsilon)$ of a learning algorithm L with respect to a* target concept *$t \in B_N$ is the least value of m such that, for any degree of confidence and accuracy $0 < \delta, \epsilon < 1$, the hypothesis inferred by L from a training sample of size m will, with probability $> 1 - \delta$, have an error $< \epsilon$ with respect to the target concept t, using any underlying distribution.*

Additionally, the sample complexity $m_L(C, \delta, \epsilon)$ of a learner L with respect to a *concept space* C ($C \subseteq B_N$) is defined $m_L(C, \delta, \epsilon) = \max_{t \in C} m_L(t, \delta, \epsilon)$. Results in the PAC framework [3] link the sample complexity of a learner to properties of its hypothesis space:

Definition 2 Hypothesis Space. *The hypothesis space of a learner L with respect to a* target concept *t, written H_t^L, is the set of concepts $h \in B_N$ for which there is some sample $\overline{x} \in (D_N)^*$ s.t. the hypothesis output on the sequence $\overline{x}_t$ is precisely h,*

$$H_t^L \hat{=} \{h \in B_N | \exists \overline{x} \in (D_N)^* \cdot L(\overline{x}_t) = h\} \tag{1}$$

where $\overline{x}_t$ is the sample defined $\overline{x}_t = \langle (x_i, t(x_i)) \rangle_{i=1}^m$ and x_i is the i-th element of the sample $\overline{x}$. The hypothesis space H_C^L defined with respect to a concept space *$C \subseteq B_N$ is the set of hypotheses that might be output on training samples for at least one of the target concepts $t \in C$:*

$$H_C^L = \bigcup_{t \in C} H_t^L \tag{2}$$

The result known as the 'Blumer Bound' [3, Thm 2.2] states an upper bound on the sample complexity $m_L(C, \delta, \epsilon)$ of a consistent learning algorithm which is of the order of $(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{\log |H_C^L|}{\epsilon})$, where H_C^L is the hypothesis space of the learning algorithm L w.r.t. the concept space C. The proof of [3, Thm 2.2] can also be modified slightly to give a bound for $m_L(t, \delta, \epsilon)$ of the order of $(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{\log |H_t^L|}{\epsilon})$ c.f. [7, Propn 6.5.11]. A different upper bound on sample complexity [3, Thm 2.1] depends upon the quantity known as the VC dimension[2] of H_C^L.

It is easy to argue informally that the larger the hypothesis space, the more training examples the learning algorithm must see in order to discriminate between the available hypotheses [9]. The PAC results described above formalise this link between the efficiency and the hypothesis space of a learner. The link is not conclusive, however, since the formal results we have stated are only *upper* bounds. The aim of this paper is therefore to explore, using examples drawn from our studies of instance-based learning algorithms, the relationship between hypothesis space and sample complexity, and in particular to explore just how much information about the learning curve of an algorithm is given by knowledge of its hypothesis space. §2 below discusses in more detail work which is related to our own, while §3 introduces the instance-based

[2] A set $X' \subseteq X$ is *shattered* by the set $F \subseteq (X \to \{0,1\})$ iff there is a concept $c \in F$ s.t. $X'' = \{x \in X' | c(x) = 1\}$ for *all* subsets $X'' \subseteq X'$. The *VC Dimension* of the set $F \subseteq (X \to \{0,1\})$, written $d_{VC}(F)$, is the size of the largest subset of X which is shattered by F. In addition to defining an upper bound on sample complexity, the VC dimension also gives a lower bound in some circumstances and is said to characterise learnability in that a concept space will have a polynomial VC dimension if and only if there is a learning algorithm for that space with polynomial sample complexity [3] [4].

learners which provide our examples. §4 and §5 then present our results concerning these algorithms which form the main data for our presentation and §6 summarises our observations.

2 Related Work

The work described in this paper continues previous studies of IBL within the PAC Learning Framework. In the best known study in this field, Albert and Aha [1] analyse a simple instance-based learner which can learn any concept defined on a space of N real-valued attributes. Since no learning algorithm can learn arbitrary classifications of this representation, the authors have to express the sample complexity as a function of the length of the boundary of the target concept and in this way are able to specify finite sample sizes sufficient for PAC learning broad classes of target concept.

The study presented in [1] has since been extended in a number of ways. For example Albert and Aha consider only a fixed similarity measure, while IBL algorithms which are able to adjust their similarity measure have since been studied, for example in [11] [8]. In addition, Albert and Aha consider problem instances represented by vectors of *real* values and their solution, a sample complexity which is a function of the length of the concept boundary, does not transfer from this setting to one where problem instances are represented by discrete (non-scalar) values. Instance-based learning over problem representations based on binary valued 'boolean' variables is studied in [6] [7].

In this paper we use as a motivating example the *comparison* of an instance-based learner with a fixed similarity measure, with an instance-based learner which adjusts its similarity. In addition, this work differs from both [1] and [6] since we address specifically the question of how to analyse a learning algorithm to discover its learning properties on a *particular* target concept, rather than looking for general 'learnability' properties which hold over a broad class of target concepts.

Our focus distinguishes this work not only from other papers specific to IBL [1] [6], but also from the 'traditional' view taken in the study of PAC learning. We believe that our use of PAC learning results is a more 'practical' or concrete application of the theory than is usual in the field. In its original conception, e.g. [3], 'traditional' PAC learning typically asks questions such as whether an efficient learning algorithm exists for a particular concept space. As a result, it is usually assumed that the concept space is precisely the set of concepts representable by the learner, presumably because the learner has been carefully chosen to match the intended concept space ($H_C^L = C$). In contrast, we are interested in a more 'practical' application of the results, starting off with a *particular* learning algorithm and attempting to determine its sample complexity w.r.t. a particular target concept or concept space. In the work we describe below, the typical question is rather whether one learning algorithm is more efficient than another in the context of a particular learning problem. We would like to establish, for example, whether an instance-based learner which changes its similarity measure requires fewer examples to reach a certain level of accuracy than one which does not. Since the algorithm is no longer assumed to 'fit' the concept space, but rather the concept space is simply a parameter of the analysis, it will now generally

be the case that the hypothesis space contains concepts which lie outside the original concept space ($C \subset H_C^L$). The algorithms such as the instance-based learners which we study in our 'practical' application of PAC learning are not, generally, optimally biased to the concept spaces we choose as examples.

This creates some difficulties for a rigourous application of PAC results. The known lower bounds on sample complexity [4] are given in terms of C rather than H_C^L and are intended to indicate the intrinsic difficulty of learning the concept space C rather than the sample complexity of any single learning algorithm. The Blumer Bound [3, Thm 2.2] meets these lower bounds only when $\log |H_C^L| = O(d_{VC}(C))$ [4] (where $d_{VC}(C)$ is the VC dimension [3] of the concept space C). $\log |H_C^L| = O(d_{VC}(C))$ will sometimes hold when $H_C^L = C$, since $d_{VC}(C) \leq \log_2 |C|$, but it is unlikely to hold if H_C^L is significantly larger than C. Therefore, while $|H_C^L|$ provides an upper bound, there are no known general results which determine the actual sample complexity of the algorithm. Similarly, $|H_t^L|$ provides an upper bound for $m_L(t, \delta, \epsilon)$ but there is no lower bound suitable for comparison. One of the issues in our work has therefore been to observe whether the upper bounds given by $|H_C^L|$ and $|H_t^L|$ provide any useful information about the actual learning behaviour of L.

Since the actual sample complexity of learners such as IBL algorithms is not accessible without more powerful methods of analysis, we can only validate the upper bounds provided by $|H_C^L|$ and $|H_t^L|$ indirectly. We do this by comparing the upper bound with the *average-case* accuracy of the learner, measured by empirical experiment. The number of examples needed to reach a certain level of accuracy in the average case will be quite different from the numeric value of the sample complexity; instead we are interested in whether the *qualitative* relationships between the average-case accuracies of learners in different experiments will be predicted by differences in the quantities $|H_C^L|$ and $|H_t^L|$. This paper explores the hypothesis that, where there are learning algorithms L and L' and target concepts t and t' s.t. $|H_t^L| < |H_{t'}^{L'}|$ or $d_{VC}(H_t^L) < d_{VC}(H_{t'}^{L'})$, then the average-case accuracy of L on the target concept t is (in some sense) likely to be greater than that of L' on t' (c.f. similar statements in H_C^L). This will be valid only if two assumptions hold. Firstly, it is assumed that the accuracy of a learning algorithm increases as the the sample complexity decreases, i.e. that average case accuracy correlates (negatively) with worst-case sample complexity. Secondly, it is assumed that the sample complexity of the learner increases with the upper bounds. This paper provides some examples where these conditions evidently do hold and which demonstrate correlation between increased hypothesis space and decreased accuracy.

3 Case Studies: $CB1$ and VS-CBR

Our questions concerning the 'practical' use of the PAC learning framework will be illustrated by results concerning the hypothesis spaces of simple instance-based learners when learning *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.

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

Fig. 1: $CB1$ **Learning Algorithm [7, Defn 5.3.1].** $\overline{s} = \langle (d_i, n_i) \rangle_{i=1}^{m}$ **is a training sample from** $(D_N \times \{0,1\})^m$**.**

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 i-th bit of the representation is said to be *relevant* to a monomial concept $t \epsilon M_N$ if the literal u_i or $\overline{u}_i$ appears in the expression representing t, and *irrelevant* if not.

Simple instance-based learning algorithms learn by adding cases to a case-base CB and by choosing a similarity measure σ. A case-base CB is a set of *exemplars*, each of which is a pair $(d, n) \epsilon (D_N \times \{0,1\})$. Normally, a case-base is *compatible* with some target concept $t \epsilon B_N$ s.t. for each exemplar $(d, n) \epsilon CB$, $t(d) = n$. This is written $CB \subseteq t$; $CB \subseteq t \hat{=} (\forall (d, n) \epsilon CB \cdot t(d) = n)$. The similarity measure σ is a total function in $D_N \times D_N \rightarrow [0,1]$ 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 $\{0,1\}$-valued function defined on D_N as follows (c.f. the 'standard semantics' of [5]):

$$h_{\langle CB,\sigma \rangle}(d) = \begin{cases} 1 & \textit{if } \exists (d_{pos}, 1) \epsilon CB \cdot \forall (d_{neg}, 0) \epsilon CB \cdot \sigma(d, d_{pos}) > \sigma(d, d_{neg}) \\ 0 & \textit{otherwise} \end{cases} \tag{3}$$

The learners studied here use a *weighted* similarity measure; in this paper, this measure is simply a sum of the bits of the representation on which two descriptions agree, weighted according to a 'weight vector' $\overline{w} \epsilon [0,1]^N$:

$$\sigma_{\overline{w}}(d_1, d_2) = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \times (1 - |(d_1)_i - (d_2)_i|) \tag{4}$$

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. In this case, the function will be written σ_H, and is the similarity measure used by the most straightforward instance-based learner, $CB1$ (Figure 1). $CB1$ learns by accumulating all the available cases into the case-base CB and choosing the hypothesis represented by $\langle CB, \sigma_H \rangle$.

$CB1$ is limited by its use of just a single measure of similarity; this is known to reduce the effectiveness of an instance-based learner [5], and particularly to reduce the small sample accuracy where some bits of the representations are irrelevant to the target concept [15]. The instance-based learner $VS\text{-}CBR$ [14, Fig 4] overcomes this limitation by learning whether or not each bit of the representation is relevant to a monomial target concept and adjusts the weights of the similarity measure $\sigma_{\overline{w}}$ accordingly. The operation of $VS\text{-}CBR$ is described in detail in [14]. In outline, $VS\text{-}CBR$ differs from $CB1$ by *selectively* adding exemplars to the case-base (all positive

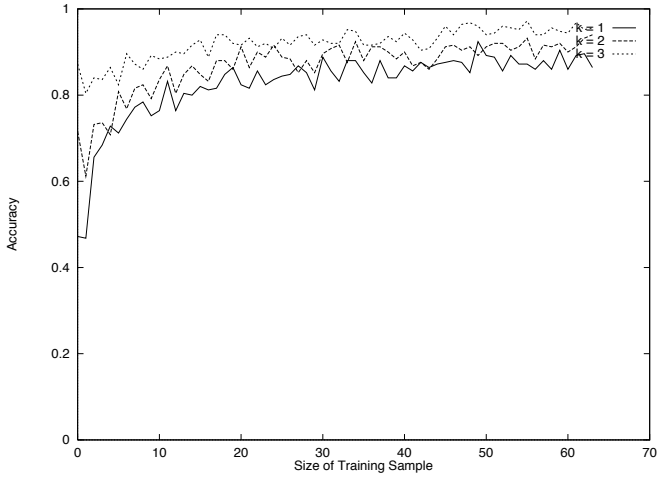

Fig. 2: **[7, Fig 5.3] Average Learning Curve for Target Concepts** $t \epsilon M_{6,k}$ **(**$CB1$**)**

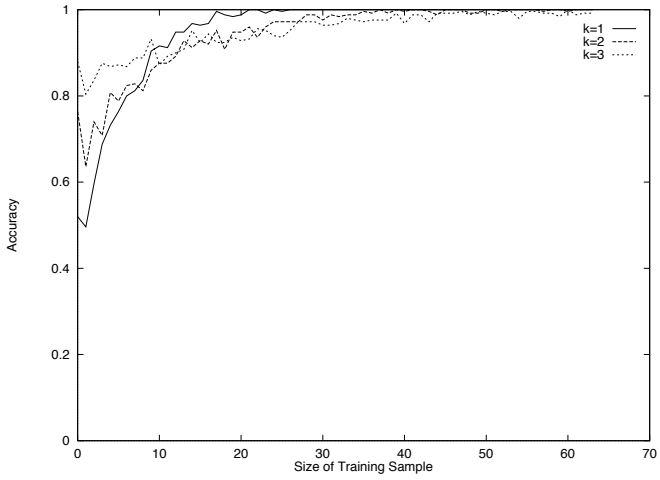

Fig. 3: **[7, Fig 6.1] Average Learning Curve for Target Concepts** $t \epsilon M_{6,k}$ **(**VS-CBR**)**

exemplars other than the first are discarded) and by adjusting the weights of the similarity measure $\sigma_{\overline{w}}$. All weights of $\overline{w}$ are 1 initially, but a weight changes to zero iff two *positive* examples are observed which disagree on that bit of the representation.

Figures 2 and 3 show the results of simple experiments measuring the accuracy of $CB1$ and VS-CBR w.r.t. target concepts $t \epsilon M_{6,k}$ for values for k of 1, 2 & 3. The figures show the average accuracy of the algorithms for a target concept from each of the concept spaces plotted against the size of the training sample.

Figure 2 shows the results of this experiment for $CB1$. The graph shows how the accuracy of the learner starts at approximately $(1 - 2^{-k})$ for a target concept $t \epsilon M_{N,k}$ (corresponding to the probability of randomly sampling a negative instance of the target concept), shows a slight initial decrease and then increases steadily as additional exemplars are added to the case-base. The figure shows that $CB1$ learns concepts from $M_{6,1}$ more slowly than concepts from $M_{6,2}$, which are learnt more slowly

N	$\lvert H^{CB1}_{M_{N,k}} \rvert$			$\lvert B_N \rvert = 2^{2^N}$
	$k=1$	$k=2$	$k=3$	
1	4	—	—	4
2	10	10	—	16
3	96	60	60	256

Table 1: [7, Table 5.3] Values of $\lvert H^{CB1}_{M_{N,k}} \rvert$ for $1 \leq k \leq N \leq 3$

than concepts in $M_{6,3}$.

Figure 3 shows the average learning curve for $VS\text{-}CBR$ on target concepts on the same concept spaces $M_{6,k}$. The figure shows the learning curves start from the same initial accuracies as observed in Figure 2, but achieve a close to perfect classification accuracy much more quickly than in Figure 2. Additionally, in Figure 3, 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 $M_{6,3}$: the opposite to the pattern shown in Figure 2.

4 Hypothesis Space of $CB1$ and $VS\text{-}CBR$

This section presents our results concerning the hypothesis spaces of $CB1$ and $VS\text{-}CBR$. It is clear from the definition of $CB1$ that exemplars are added indiscriminately to the case-base, and therefore all possible case-bases are reachable by the algorithm. The hypothesis space of $CB1$ w.r.t. a target concept t is simply the set of concepts that are represented by σ_H and a case-base compatible with t (c.f. [7, Propn 5.3.1]):

$$\forall t \,\epsilon\, B_N \cdot H^{CB1}_t = \{h_{\langle CB,\sigma\rangle} | CB \subseteq t\}$$
$$\forall C \subseteq B_N \cdot H^{CB1}_C = \bigcup_{t\,\epsilon\, C} \{h_{\langle CB,\sigma\rangle} | CB \subseteq t\}$$

Table 1 shows the cardinalities of the hypothesis spaces $H^{CB1}_{M_{N,k}}$ for values $1 \leq k \leq N \leq 3$. The table shows that the small values of N and k that we have been able to consider are really too small to demonstrate the structure and relationships of the hypothesis spaces. However, in the case where differences in the hypothesis space start to emerge, i.e. where Table 1 shows there are 96 members of $H^{CB1}_{M_{3,1}}$ but only 60 members of $H^{CB1}_{M_{3,2}}$ and $H^{CB1}_{M_{3,3}}$, then the numbers shown are consistent with the argument that larger hypothesis spaces correlate with slower convergence in learning. That is, the limited picture presented by the table shows that $k \leq k' \rightarrow \lvert H^{CB1}_{M_{N,k}} \rvert \geq \lvert H^{CB1}_{M_{N,k'}} \rvert$, i.e. $\lvert H^{CB1}_{M_{N,k}} \rvert$ decreases (non-strictly) in k, while Figure 2 shows that the accuracy of $CB1$ increases in k. There are 3^{2^N} case-bases for functions defined on N propositional variables and calculating $H^{CB1}_{M_{N,k}}$ requires $6,561$ tests for $N = 3$, but 4.3×10^7 tests for $N = 4$. Extending the table beyond what is shown above would therefore not be practical without much more selective evaluation of the case-bases.

Further statements about H^{CB1} have been elusive, but more success has been forthcoming in the case of $VS\text{-}CBR$[3]. The following result includes at least two main

[3]Details of the analysis summarised here can be found in [7] [8]. Proofs of the new results of this

claims about $H^{VS\text{-}CBR}$. Firstly, it implies that the positive exemplars discarded by $VS\text{-}CBR$ are irrelevant to the hypothesis represented and secondly, it shows that the different weighted similarity measures that might be chosen by $VS\text{-}CBR$ can be ignored in the sense that all hypotheses of $VS\text{-}CBR$ are representable by the uniformly weighted similarity measure σ_H:

Proposition 3 [7, Propn 6.4.2] *The effective hypothesis space of $VS\text{-}CBR$ w.r.t. any target concept $t \epsilon M_N$ is the set of concepts $h_{\langle CB,\sigma_H\rangle}$ where CB is any case-base compatible with t and which in addition has no more than one positive exemplar:*

$$\forall t \epsilon M_N \cdot H_t^{VS\text{-}CBR} = \{h_{\langle CB,\sigma_H\rangle} | CB \subseteq t \wedge \#\{d_{pos} \epsilon D_N | (d_{pos}, 1) \epsilon CB\} \leq 1\}$$

For each monomial concept $t \epsilon M_{N,k}$, there are a number of more specific monomial target concepts which will have all the negative instances of t as negative instances and will still be positive on some single description. The result below therefore follows immediately as a corollary of Proposition 3:

Corollary 4 [7, Cor 6.4.3] *The hypothesis space $H_{M_{N,k}}^{VS\text{-}CBR}$ of $VS\text{-}CBR$ w.r.t. the concept space $M_{N,k}$ is* included by *the hypothesis space $H_{M_{N,k'}}^{VS\text{-}CBR}$, for all $N \geq k' \geq k$.*

$$\forall 1 \leq k \leq k' \leq N \cdot H_{M_{N,k}}^{VS\text{-}CBR} \subseteq H_{M_{N,k'}}^{VS\text{-}CBR}$$

The effective hypothesis space $H_{M_{N,k}}^{VS\text{-}CBR}$ will therefore contain fewest hypotheses for the smallest values of k ($k = 1$) and the most for the largest values of k, as k approaches N. On the other hand, Figure 3 shows that $VS\text{-}CBR$ learns target concepts represented by the smallest monomial expressions more rapidly than those represented by larger monomial expressions. $VS\text{-}CBR$ therefore demonstrates clearly the correlation of hypothesis space and average-case accuracy discussed above. In addition, the contrast between Corollary 4 and Table 1, where $|H_{M_{N,k}}^{CB1}|$ apparently decreases in k, reflects the contrast between Figures 2 and 3, where target concepts in $M_{6,1}$ are apparently easier for $VS\text{-}CBR$ to learn, but harder for $CB1$.

Two caveats apply to our claims about Corollary 4. Firstly, it was necessary to take a value of $N = 6$ in order to demonstrate that the inclusion shown in Corollary 4 can be proper in some cases. That is, [7, Example 6.4.10] shows a concept which is not in $H_{M_{6,1}}^{VS\text{-}CBR}$ and is a member of $H_{M_{6,2}}^{VS\text{-}CBR}$, but no such examples exist for smaller values of N; for $N \leq 5$, $H_{M_{N,k}}^{VS\text{-}CBR} \supseteq H_{M_{N,k'}}^{VS\text{-}CBR}$ also, for all values $1 \leq k \leq k' \leq N$. The other question that arises is whether a similar result could also be proven for the effective hypothesis space of $VS\text{-}CBR$ with respect to a *single* target concept, $H_t^{VS\text{-}CBR}$ for some $t \epsilon M_N$ (c.f. Proposition 6, below). We might expect, as with $H_{M_{N,k}}^{VS\text{-}CBR}$, that $H_t^{VS\text{-}CBR}$ w.r.t. $t \epsilon M_{N,k}$ would be smallest for the least values of k and that the size of the hypothesis space would increase with k. This is not directly implied by Proposition 3 however. For small values of k, monomials in $M_{N,k}$ will have *more* positive instances, which would tend to increase the number of different case-bases satisfying Proposition 3. However, fewer negative instances will reduce the number of representable concepts, and in addition there might be considerable

paper are given in [7].

```
set CB = ∅
for i = 1 to m do
    set CB = CB ∪ {(d_i, n_i)}
RETURN CB2(s̄) = h_⟨CB,σ_w̄t⟩
```

Fig. 4: CB2 Learning Algorithm for Concepts in M_N. $\overline{s} = \langle (d_i, n_i) \rangle_{i=1}^{m}$ is a training sample from $(D_N \times \{0,1\})^m$ and weight vector $\overline{w}_t$ is defined in equation (5).

overlap between the concepts representable by case-bases containing different positive exemplars; the balance of these factors is not at all clear.

However, since all concepts in $H_t^{VS\text{-}CBR}$ are represented by case-bases compatible with t (c.f. Proposition 3), the following corollary can be trivially proven and, in contrast to Corollary 4, definitely holds at the level of the hypothesis space w.r.t. a single target concept:

Corollary 5 *For any value of N, and any monomial target concept $t \epsilon M_N$, the effective hypothesis space $H_t^{VS\text{-}CBR}$ of $VS\text{-}CBR$ w.r.t. the target concept t is included by the hypothesis space H_t^{CB1} of $CB1$ w.r.t. the same target concept.*

$$\forall 1 \leq N \cdot \forall t \epsilon M_N \cdot H_t^{VS\text{-}CBR} \subseteq H_t^{CB1}$$

This result would also therefore apply to H_C^L; $H_{M_{N,k}}^{VS\text{-}CBR} \subseteq H_{M_{N,k}}^{CB1}$ for all values of N and k. Corollary 4 therefore gives a description of $H_C^{VS\text{-}CBR}$ which is consistent with Figure 3, while Corollary 5 shows that the relationship between $H_t^{VS\text{-}CBR}$ and H_t^{CB1}, and between $H_{M_{N,k}}^{VS\text{-}CBR}$ and $H_{M_{N,k}}^{CB1}$, is consistent with Figures 2 & 3, which show that $VS\text{-}CBR$ is more efficient for the target concepts tested than $CB1$. Further results showing similar correlations between hypothesis space and efficiency are shown in the next section, which considers two further algorithms, related to $VS\text{-}CBR$, which we have found useful for understanding that algorithm.

5 'Constituents' of $VS\text{-}CBR$

It is suggested in [7] [8] that a similarity learning IBL algorithm which manipulates the weighted similarity measure $\sigma_{\overline{w}}$ might be considered as two constituent processes, one of which manipulates the case-base element of the representation $\langle CB, \sigma \rangle$ and one of which manipulates the similarity measure. We argue that the part of $VS\text{-}CBR$ which constructs the case-base can be studied via the IBL algorithm CB2 (Figure 4). $CB2$ is essentially similar to $CB1$ except, in place of the similarity measure σ_H, it uses the measure $\sigma_{\overline{w}_t}$, which weights the bits of the representation according to their relevance to the target concept t:

$$(w_t)_i = \begin{cases} 1 & \text{if bit } i \text{ is relevant to } t \\ 0 & \text{if not} \end{cases} \tag{5}$$

Since $\sigma_{\overline{w}_t}$ is the similarity measure that will eventually be chosen by $VS\text{-}CBR$ if enough positive examples are available [14] [8], then the sample complexity of $CB2$ gives an indication of the number of examples needed by $VS\text{-}CBR$ to adequately

populate its case-base. Similarly, in [8] we note that $VS\text{-}CBR$ will choose the 'ideal' weight vector $\overline{w}_t$ on precisely those training samples from which the 'standard learning algorithm for monomials' [12] [2], which we call M, can correctly identify a monomial target concept. The sample complexity of M will therefore give an indication of the number of examples needed by $VS\text{-}CBR$ to converge to a 'good' similarity measure.

The hypothesis spaces of $CB2$ and M give further data concerning the link between hypothesis space and average-case efficiency. Average-case experiments for $CB2$ result in graphs similar to Figure 3 [7, Fig 6.2]; target concepts in $M_{6,1}$ are learnt more rapidly than target concepts in $M_{6,2}$, which are learnt more rapidly than target concepts in $M_{6,3}$. However, $CB2$ converges more rapidly than $VS\text{-}CBR$ and requires fewer examples to reach a particular level of accuracy. These two facts, that the accuracy of $CB2$ on training samples for monomial target concepts $t \in M_{N,k}$ decreases with k, and that $CB2$ is apparently more efficient than $VS\text{-}CBR$, are reflected in the relationships of the hypothesis spaces. ($t \sqsupseteq t'$ is read 't is specialised by t'' $(\forall d \in D_N \cdot t'(d) = 1 \rightarrow t(d) = 1)$.):

Proposition 6 [7, Propn 6.5.4] *The effective hypothesis space H_t^{CB2} of $CB2$ with respect to a target concept $t \in M_{N,k}$ is included by the hypothesis space $H_{t'}^{CB2}$ with respect to any target concept $t' \in M_{N,k'}$ where $k' \geq k$ and t' is a concept that specialises t:*

$$\forall 1 \leq k \leq k' \leq N \cdot \forall t \in M_{N,k} \cdot \forall t' \in M_{N,k'} \cdot t \sqsupseteq t' \rightarrow H_t^{CB2} \subseteq H_{t'}^{CB2}$$

Proposition 7 [7, Propn 6.5.6] *The effective hypothesis space H_t^{CB2} of $CB2$ w.r.t. any monomial target concept $t \in M_N$ is included by the hypothesis space $H_t^{VS\text{-}CBR}$ of $VS\text{-}CBR$ w.r.t. the same target concept.*

$$\forall N \cdot t \in M_N \cdot H_t^{CB2} \subseteq H_t^{VS\text{-}CBR}$$

The statements of Propositions 6 and 7 can easily to be extended to refer to entire concept spaces in the style of Corollary 4. The following statements are immediate corollaries of the above results:

$$\forall 1 \leq k \leq k' \leq N \cdot H_{M_{N,k}}^{CB2} \subseteq H_{M_{N,k'}}^{CB2}$$
$$\forall 1 \leq k \leq N \cdot H_{M_{N,k}}^{CB2} \subseteq H_{M_{N,k}}^{VS\text{-}CBR}$$

Similarly, the results we have available for M again show correlation between hypothesis space and average-case accuracy. The algorithm we call M is well known and analysed. Langley and Iba claim that "we have shown analytically that the number of training instances required for [the standard monomial learning algorithm] to achieve a given level of accuracy [in the average case] increases only with the logarithm of the number of irrelevant attributes"[10], that is, depends only on the quantity $N - k$. For a target concept $t \in M_{N,k}$, the sample complexity of M can also be shown to be a function of $(N - k)$:

Proposition 8 [7, Propn 6.5.13] *The hypothesis space of M, the standard learning algorithm for monomial concepts, w.r.t. a k-literal monomial target concept $t \in M_{N,k}$,*

contains all concepts in M_N which specialise the target concept t along with the concept f_0 which has value 0 on all descriptions:

$$H_t^M = \{h \,\epsilon\, M_N | h \sqsubseteq t\} \cup \{f_0\}$$

where $h \sqsubseteq t$ is read 'h specialises t' and f_0 is the concept s.t. $\forall d \,\epsilon\, D_N \cdot f_0(d) = 0$.

Proposition 8 shows that $|H_t^M| = 3^{N-k}+1$ for a target concept $t \,\epsilon\, M_{N,k}$, so that the upper bound on $m_L(t, \delta, \epsilon)$ [7, Propn 6.5.11] increases linearly in $N-k$. Similarly, from equation (2), $H_{M_{N,k}}^M$ is the set of concepts $\{h \,\epsilon\, M_N | \exists t \,\epsilon\, M_{N,k} \cdot h \sqsubseteq t\} \cup \{f_0\}$ containing all the monomial concepts which are at least as specific as the concepts in $M_{N,k}$. The quantity $|H_{M_{N,k}}^M|$ will also increase in $(N-k)$ and in fact $k \leq k' \rightarrow H_{M_{N,k}}^M \supseteq H_{M_{N,k'}}^M$. The hypothesis space of M, reckoned with respect to both a single target concept and also a space of monomial target concepts $M_{N,k}$, therefore also shows a larger hypothesis space which correlates with lower average case accuracy.

6 Conclusions

§2 explained how we have used the PAC learning framework in a slightly unusual way for the analysis of instance-based learning algorithms. In our approach, we attempt to answer 'practical' questions such as whether one learning algorithm is more efficient than another w.r.t. a particular learning problem. §2 also noted that the well-known results of the PAC framework only directly provide *upper* bounds on the sample complexity of learning algorithms in the situations we consider, and that these bounds are dependent on the hypothesis space H_C^L or H_t^L of the learning algorithm. In the absence of formal results characterising when these upper bounds meet the actual sample complexity, we have used our results concerning the hypothesis spaces of instance-based learners to explore the extent to which differences in these hypothesis spaces are predictive of differences in the relative efficiencies of those algorithms.

The results we have presented, for the four algorithms $CB1$, $VS\text{-}CBR$, $CB2$ and M, are consistent with the assertion that a larger hypothesis space implies a less efficient learning algorithm. The only omissions in our presentation are where we have not been able to resolve the difficult problems involved in reasoning about and counting instances of the case-based representation. In the case of $H_{M_{N,k}}^{CB1}$, we have found that a straightforward 'inclusion' relationship such as Corollary 4 or Proposition 6 does not hold [7, Example 5.4.3] [7, Example 5.4.4], but have not been able to demonstrate any other kind of relationship or derive a quantitative result directly. In the case of $VS\text{-}CBR$, Corollary 4 shows that $k \leq k' \rightarrow H_{M_{N,k}}^{VS\text{-}CBR} \subseteq H_{M_{N,k'}}^{VS\text{-}CBR}$ but, again, the difficulties of the case-based representation have prevented us confirming or refuting the equivalent result in terms of H_t^L ($t \sqsupseteq t' \rightarrow H_t^{VS\text{-}CBR} \subseteq H_{t'}^{VS\text{-}CBR}$). Otherwise, the results we have presented provide a handful of examples where knowledge of the hypothesis spaces of a set of learners with respect to different concept spaces provides at least qualitative information about the learning curves of those algorithms. It seems that the instance-based learners we have studied therefore represent the kind of algorithm that can be usefully studied by analysis of the hypothesis space; further work must consider the limits of this approach and attempt to characterise the learning problems for which the approach is applicable.

References

[1] M K Albert and D W Aha. Analyses of instance-based learning algorithms. In *Proc. AAAI-91,* pp 553–558, 1991.

[2] M Anthony & N Biggs. *Computational Learning Theory.* CUP, 1992.

[3] A Blumer, A Ehrenfeucht, D Haussler, & M K Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. of the ACM*, 36(4):929–965, 1989.

[4] A Ehrenfeucht, D Haussler, M Kearns, & L Valiant. A general lower bound on the number of examples needed for learning. *Info. & Computation*, 82:247–261, 1989.

[5] C Globig, K P Jantke, S Lange, & 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 *Proc. ECAI-96,* pp 117–121, 1996.

[7] A D Griffiths. *Inductive Generalisation in Case-Based Reasoning Systems.* PhD thesis, Dept. of Computer Science, University of York, York, UK, 1997.

[8] A D Griffiths & D G Bridge. PAC analyses of a 'Similarity Learning' IBL algorithm. In E Plaza and D Leake, eds, *Case-Based Reasoning Research and Development: Proc. ICCBR-97,* LNAI 1266, pp 445–454. Springer, 1997.

[9] D Haussler. Probably approximately correct learning. In *Proc. AAAI-90,* pp 1101–1108. 1990.

[10] P Langley & W Iba. Average-case analysis of a nearest neighbour algorithm. In R Bajcsy, ed., *Proc. IJCAI-93*, pp 889–894. 1993.

[11] K Satoh and S Okamoto. Towards PAC-learning of weights from qualitative distance information. In D W Aha, ed, *Case-based reasoning: Papers from the 1994 AAAI Workshop*, Technical Report WS-94-01. AAAI Press, 1994.

[12] L G Valiant. Deductive learning. *Philosophical Transactions of the Royal Philosophical Society of London A*, 312:441–446, 1984.

[13] L G Valiant. A theory of the learnable. *Comms. of the ACM*, 27:1134–1142, 1984.

[14] S Wess & C Globig. Case-based and symbolic classification - A case study. In S Wess, K-D Althoff, & M M Richter, eds, *Topics in CBR: Selected papers from EWCBR-93*, LNAI 837, pp 77–91. Springer, 1994.

[15] D Wettschereck, D W Aha, & T Mohri. A review and comparative evaluation of feature weighting methods for lazy learning algorithms. Report AIC-95-012, Navy Center for Applied Research in AI, NRL, Washington DC, 1995.