On Concept Space and Hypothesis Space in Case-Based Learning Algorithms

A D Griffiths and D G Bridge

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

Abstract. In order to learn more about the behaviour of case-based reasoners as learning systems, we formalise a simple case-based learner as a PAC learning algorithm. We show that the case-based representation $\langle CB, \sigma \rangle$ is rich enough to express any boolean function. We define a family of simple case-based learning algorithms which use a single, fixed similarity measure and we give necessary and sufficient conditions for the consistency of these learning algorithms in terms of the chosen similarity measure. Finally, we consider the way in which these simple algorithms, when trained on target concepts from a restricted concept space, often output hypotheses which are outside the chosen concept space. A case study investigates this relationship between concept space and hypothesis space and concludes that the case-based algorithm studied is a less than optimal learning algorithm for the chosen, small, concept space.

1 Introduction

The performance of a case-based reasoning system [13] will change over time as new cases are added to the case base by the problem-solving process. A prudent knowledge engineer might wonder whether the performance will necessarily improve, how quickly the performance of the system might change, or how many exemplars would be required to reach some specific level of accuracy in problem solving.

A simple model of a case memory system is presented here as a basis for answering these questions analytically. The model used is a *functional* one in that the knowledge content of the case memory system is modelled as a mapping between input and output domains. The analysis applied to this model is a *probabilistic, worst case* analysis, in that we apply the PAC learning framework [3] [10] to case-based learning.

For the moment, a number of restrictions are made in order to gain leverage on the problems in hand. To focus on the learning behaviour of the systems, the model abstracts away from many aspects of case-based reasoning systems which are of interest in other contexts such as interactive properties, details of the reasoning process at conceptual and implementation levels and knowledge representation issues such as the choice of abstract indices. Additionally, this paper focuses only on case-based classifiers whose task is to decide whether or not the input description is an instance of some concept.

2 Definitions

Our model is of a case-based classifier operating over the space of N-bit binary vectors. Thus the *example space* in the current work will be referred to as $D_N \triangleq \{0,1\}^N$. $(d)_i$ will be used to stand for the i-th bit of a vector $d \epsilon D_N$. The set of $\{0,1\}$-valued total functions defined over this domain will be denoted $B_N \triangleq (D_N \rightarrow \{0,1\})$.

By *hypothesis space* we refer to the set of possible hypotheses that might be output by the case-based learning algorithm over all possible training samples. The term *concept space* on the other hand will be used to refer to some specific subset of B_N from which target concepts for the learning algorithm might be drawn. In particular, section 5 considers the set of *monomial* functions as the concept space for a case-based learning algorithm. A monomial expression U is a combination of no more than N literals chosen without replacement from the set $\{u_1, \ldots, u_N\}$; additionally each chosen literal may be negated before being added to U. The classification function for the expression interprets U as a conjunction of the (possibly negated) literals:

$$h_U^N(d) = \begin{cases} 1 \text{ if } \forall i \cdot (u_i \epsilon U \rightarrow ((d)_i = 1)) \wedge (\overline{u_i} \epsilon U \rightarrow ((d)_i = 0)) \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

The function $h_U^N(d)$ is therefore a $\{0,1\}$-valued function on D_N whose value is decided by a conjunction of the bits of d. The space of such functions will be referred to as M_N. Further, $M_{N,k}$ is defined as the set of monomials with exactly k literals ($\#U = k$).

3 Case-Based Learning Algorithms

Following the work of Jantke [11], a case memory system is modelled as the pair $\langle CB, \sigma \rangle$ where CB is the case-base, or set of stored exemplars, assumed here to be free from observational error, and σ is a similarity measure defined for the space D_N. Using the terminology of Dearden's model [7], the case-base is modelled as a set of pairs of 'descriptions' and 'reports'. As indicated above, a description is an N-bit vector from the space D_N. A report is a single bit denoting the classification of that exemplar, making CB an object of type:

$$CB : \mathcal{P}(D_N \times \{0,1\})$$

The similarity measure σ is a function over pairs of descriptions returning a normalised real value indicating the degree of similarity between the two instances:

$$\sigma : (D_N \times D_N) \rightarrow [0,1]$$

The pair $\langle CB, \sigma \rangle$ is treated as the representation of a function from B_N, according to the following interpretation related to the 'standard semantics' for

a case-based classifier of Jantke and Lange [12]. The function represented by $\langle CB, \sigma \rangle$ is defined as:

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

Informally, a point d from D_N is positively classified by $h^N_{\langle CB,\sigma \rangle}$ 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}. In relation to other semantics discussed by Jantke [11], this interpretation resolves 'ties' between equally similar near neighbours by imposing a preference ordering on the 'report' part of retrieved cases. Negative exemplars are preferred over positive ones in inferring the classification of a new problem instance, i.e. if the set of exemplars which are most similar to d contains both positive and negative exemplars, d will be classified negatively.

Since the interpretation of a case-based representation $\langle CB, \sigma \rangle$ depends on the interaction between the available cases and the similarity measure, a 'case-based' or 'instance-based' learning algorithm may alter its hypothesis by manipulating either of the two components [15, p.79]. The algorithms IB2 [1], VS-CBR [15] and PEBLS [5], for example, each show different ways of adjusting the represented hypothesis via changes to the case-base and/or the similarity measure. In the current paper, we restrict our study to the following family of very simple case-based learning algorithms.

Definition 1. $CB1(\sigma)$ **Learning Algorithm for Case-Based Classifiers**

set $CB = \emptyset$
for i = 1 to m do
    set $CB = CB \cup \{(d_i, b_i)\}$
set $CB1(\sigma)(\overline{s}) = h_{\langle CB,\sigma \rangle}$

$CB1(\sigma)$ learns by adding each and every member of the training sample $\overline{s}$ (a series of m pre-classified examples (d_i, b_i)) to the case base, and constructs each hypothesis using a single, fixed similarity measure σ. Clearly the usefulness of $CB1(\sigma)$ will depend on the choice of σ; a similarity measure that assigns high similarity to *arbitrary* pairs of descriptions will not be of much use in defining a viable learning algorithm.

The best understood learning algorithms are those which consistent, i.e. those which are able to classify correctly at least the exemplars in their training sample. In the following section we demonstrate precisely which choices of similarity measure allow $CB1(\sigma)$ to behave consistently.

4 Consistency of $CB1(\sigma)$

Theorem 4 below gives necessary and sufficient conditions over σ to make $CB1(\sigma)$ a consistent consistent learning algorithm. Results elsewhere [12, Lemma 3] [14,

Lemma 7] formalise the intuition that a 'reasonable' similarity measure [14], which recognises that an object is more similar to itself than any other object, will be sufficient for consistency. This property is here called 'definiteness' after Day and Faith [6, p.183].

Definition 2. Definiteness of a Similarity Measure. A similarity measure σ is definite iff the comparison of two distinct objects yields a score strictly less than the score given to the comparison of an object to itself.

$$\forall d, d' : \ D_N \cdot d \neq d' \rightarrow \sigma(d, d') < \sigma(d, d) \tag{3}$$

This property ensures a consistent hypothesis since any exemplar in the case base will be judged strictly most similar to itself, and therefore those exemplars at least will be classified correctly by equation (2). Definiteness is not however a necessary condition for consistency. The exemplars in the case base will still be classified correctly as long as the most similar object to a positive exemplar is *any* positive exemplar and the most similar object to a negative exemplar is *any* negative one. In other words, two distinct objects may be assigned maximal similarity only if they are classified the same by all relevant classification functions f. This is recognised informally as a necessary condition by Wess and Globig [15, p.86]. We express it within our framework in our definition of predictivity and prove it a necessary and sufficient condition over σ to make $CB1(\sigma)$ a consistent learning algorithm.

Definition 3. Predictivity of a Similarity Measure with respect to a concept space C. A similarity measure is predictive of a concept space C iff, for any concept $c \in C$:

1. When d is a positive instance of c, the comparison of d and d' yields a score at least as large as the comparison of d to itself only if d' is also a positive instance.

$$\forall c \in C \cdot \forall d, d' \in D_N \cdot \sigma(d, d') \geq \sigma(d, d) \rightarrow c(d) = 1 \rightarrow c(d') = 1 \tag{4}$$

2. When d is a negative instance of c, the comparison of d and d' yields a score strictly greater than the comparison of d to itself only if d' is also a negative instance.

$$\forall c \in C \cdot \forall d, d' \in D_N \cdot \sigma(d, d') > \sigma(d, d) \rightarrow c(d) = 0 \rightarrow c(d') = 0 \tag{5}$$

Note how this relates to equation (2) in that the property of definiteness is relaxed precisely where no misclassification will occur under our chosen classification function (2). The asymmetry in equations (4) and (5) reflects the preference given to negative exemplars in the classification function. Hence we emphasise that choosing a different semantics in (2) would entail a slightly different form of the following theorem.

Theorem 4. Consistency of $CB1(\sigma)$. *For any concept space $C \subseteq B_N$, $CB1(\sigma)$ is a consistent learning algorithm for C if and only if the chosen similarity measure σ is predictive of C.*

Proof. Let $CB1(\sigma)$ infer a hypothesis from some training sample $\overline{s} = \langle (d_i, b_i) \rangle$ for a target concept c. According to the definition of $CB1(\sigma)$, the case-base will contain exactly those labelled examples presented in the training sample; in the absence of observational error we can assume $(d_i, n) \in CB \rightarrow c(d_i) = n$, for $n \in \{0, 1\}$. *a) Sufficiency:* Assume σ is predictive of the concept space C. Taking positive and negative exemplars in the case base separately, consider first d_i such that $b_i = 1$. For any negative exemplar $(d_{neg}, 0) \in CB$, we have $c(d_{neg}) = 0$ and hence by equation (4) $\forall d \in D_N \cdot \forall (d_{neg}, 0) \in CB \cdot \sigma(d, d_{neg}) < \sigma(d, d) \vee c(d) = 0$. Since $c(d_i) = 1$, we conclude $\forall (d_{neg}, 0) \in CB \cdot \sigma(d_i, d_{neg}) < \sigma(d_i, d_i)$, and thus $h_{\langle CB,\sigma \rangle}(d_i) = 1$ by equation (2). By a similar argument, for some d_i such that $b_i = 0$ we derive from equation (5), $\forall (d_{pos}, 1) \in CB \cdot \sigma(d_i, d_{pos}) \leq \sigma(d_i, d_i)$, and hence $h_{\langle CB,\sigma \rangle}(d_i) = 0$ by equation (2). Thus for any example d_i in an arbitrary training sample, $h_{\langle CB,\sigma \rangle}(d_i) = b_i$, making $CB1(\sigma)$ a consistent learning algorithm. *b) Necessity.* It will be shown that for any similarity measure σ' which violates either of equations (4) & (5), there is a target concept c' from the specified concept space for which a training sample can be constructed which will be mis-classified by $CB1(\sigma)$. The consistency of $CB1(\sigma)$ would therefore require a similarity measure satisfying both equations. If equation (4) does not hold, then there must be two descriptors d_1 and d_2 and a target concept c' such that:

$$\sigma'(d_1, d_2) \geq \sigma'(d_1, d_1) \wedge c'(d_1) = 1 \wedge c'(d_2) = 0 \tag{6}$$

Thus $\langle (d_1, 1), (d_2, 0) \rangle$ will be a training sample for c'. Given the case base CB constructed by $CB1(\sigma')$ from this sample, note that $h_{CB,\sigma'}(d_1) = 0$ since equation (6) indicates that the negative exemplar d_2 will be at least as similar to d_1 as d_1 is to itself. Hence $h_{\langle CB,\sigma \rangle}$ disagrees with the training sample. In a similar way, if it assumed that equation (5) is relaxed, then there is a training sample $\langle (d_1, 0), (d_2, 1) \rangle$ resulting in a hypothesis such that $h_{\langle CB,\sigma' \rangle}(d_1) = 1$. Thus $CB1(\sigma)$ will be a consistent learning algorithm for a concept space C if and only if σ is predictive of C. □

The close relationship between definition 2 (definiteness) and definition 3 (predictivity) means that the following additional result can be easily established:

Corollary 5. *$CB1(\sigma)$ is a consistent learning algorithm for the space B_N of all total functions on D_N if and only if σ is a definite similarity measure.*

As a further corollary, we can also state the following.

Corollary 6. *Given a similarity measure σ which is predictive of a concept space C, then for any target concept $c \in C$ there is a case-base CB s.t. $h_{\langle CB,\sigma \rangle} = c$.*

Proof. For some σ and C s.t. σ is predictive of C, take any $c \in C$ and any training sample $\overline{s}$ for c which contains an exemplar for every point in the example space D_N. Since Theorem 4 guarantees that the output of $CB1(\sigma)$ will be consistent with $\overline{s}$, clearly the function $h_{\langle CB,\sigma \rangle}$ output by $CB1(\sigma)$ on $\overline{s}$ will be exactly c. □

Finally we observe that it is a basic result in the PAC framework that a learning algorithm which is consistent with respect to some concept space and which learns using a *finite* hypothesis space is a PAC-learning algorithm for that concept space [3, p.41]. Since the number of distinct boolean functions that can be defined on D_N is 2^{2^N} the hypothesis space of $CB1(\sigma)$ must be finite. Hence, trivially, a similarity measure predictive of any concept space $C \subseteq B_N$ is sufficient to make $CB1(\sigma)$ a PAC learning algorithm for C (c.f. PAC-Learnability results for case-based classifiers for concepts defined on real-valued attributes in [1] [2]). PAC learnability answers one of our original questions (§1): the performance of a consistent case-based reasoning system will eventually improve if enough exemplars are presented. What is more interesting however, is to ask *how many* examples must be processed to guarantee a good hypothesis.

5 Sample Complexity in Case-Based Learning

The sample complexity of a learning algorithm with respect to some concept space is defined within the PAC learning framework as the size of training sample which will ensure, to some level of confidence and accuracy, that the hypothesis chosen by the learning algorithm is a good approximation, for any target concept in the chosen concept space. Theorem 7 gives an upper bound on sample complexity in terms of the VC dimension of the hypothesis space used by an algorithm. The VC dimension of a space of $\{0,1\}$-valued functions is a quantity related to the size of the function space, being defined as the size of the largest possible sample from the example space for which every possible dichotomy into positive and negative examples can be generated by some function in the set ('shattering') [4, p.934] [9, p.189]. Note the relationship of this theorem to the results of the previous section in that it refers specifically to consistent learning algorithms.

Theorem 7. *[4, Thm 2.1(ii)(a)] [9, Theorem 4.4] Suppose that an hypothesis space H has finite VC dimension $d_{VC}(H)$. Then any consistent learning algorithm L which uses hypothesis space H is PAC with sample complexity:*

$$m_L(H,\delta,\epsilon) \leq \left\lceil \frac{k_1.d_{VC}(H)}{\epsilon} \log_2\left(\frac{k_2}{\epsilon}\right) + \frac{k_3}{\epsilon}\log_2\left(\frac{k_4}{\delta}\right) \right\rceil$$

where δ & ϵ are the required levels of confidence and accuracy, and k_i constant.

In giving an upper bound on sample complexity, Theorem 7 shows that the size of training sample that can be processed before a consistent learning algorithm *necessarily* outputs a good hypothesis with high probability will increase with the VC dimension. In what follows, we assume that the converse also holds, and that as the VC dimension increases, the sample complexity of the learning algorithm also must increase. Although strictly this depends on the specific properties of the learning algorithm using the hypothesis space, we hold that, in general, the larger the hypothesis space, the more training examples the

learner must see in order to discriminate between the available hypotheses, and choose a hypothesis that is accurate with high probability [10, p.1103].

Any such discussion, however, requires us to characterise the hypothesis space of our case-based learners. The hypothesis space of $CB1(\sigma)$ with respect to some concept space C will be referred to as $H_C^{CB1(\sigma)}$. The simplicity of $CB1(\sigma)$ means that, for a given target concept t, all possible case-bases $CB \subseteq t$ are reachable by the learning algorithm. This allows the following to be stated about the hypothesis space of $CB1(\sigma)$:

Proposition 8. *A function f is a member of the hypothesis space of $CB1(\sigma)$ with respect to the concept space $C \subseteq B_N$ if and only if there is some target concept $c \in C$ for which there is a case base $CB \subseteq c$ s.t. $h_{\langle CB,\sigma\rangle} = f$.*

$$\forall C \subseteq B_N \cdot \forall f \in B_N \cdot f \in H_C^{CB1(\sigma)} \leftrightarrow \exists c \in C \cdot \exists CB \subseteq c \cdot h_{\langle CB,\sigma\rangle} = f$$

As a corollary of proposition 8:

$$H_C^{CB1(\sigma)} = \bigcup_{t \in C} hyp_{t,\sigma} \tag{7}$$

where $hyp_{t,\sigma} = \{h_{\langle CB,\sigma\rangle} | CB \subseteq t\}$

These statements show how the hypothesis space of $CB1(\sigma)$ depends on the choice of both the similarity measure and the concept space. For smaller concept spaces, since we restrict the possible target concepts and hence the allowable training samples, only a restricted number of the possible functions in B_N may be output as hypotheses. On the other hand, it will not be uncommon that a case base CB which is extensible to some target concept $c \in C$ will be interpreted by equation (2) as a function from outside of the concept space C. In general then, the hypothesis space $H_C^{CB1(\sigma)}$ does not necessarily contain all functions $f \in B_N$, but may well contain functions from outside the chosen concept space.

Our contribution in the remainder of the paper is to establish a lower bound on the VC dimension of the hypothesis space of $CB1(\sigma)$ for particular instances of C and σ. Specifically, we will consider the (highly restricted) set of functions $M_{N,k}$ as concept space and the 'unweighted feature count' σ_F defined in equation (8) as similarity measure.

$$\sigma_F(d_1, d_2) = \frac{1}{N}\#\{i | 1 \leq i \leq N \wedge (d_1)_i = (d_2)_i\} \tag{8}$$

Proposition 15 below reports a surprising result which partly characterises the space of functions $H_{M_{N,k}}^{CB1(\sigma_F)}$. Corollary 16 re-expresses this result in terms of the VC dimension to give the promised lower bound. The following definitions and results, whose proofs are omitted for brevity, are given as necessary preliminaries to Proposition 15 and its corollary.

Definition 9. Extrapolations of monomial function. The extrapolations of a monomial function $h_U^N \in M_N$ are the functions $h_{U'}^{N+1} \in M_{N+1}$ such that $U' \in \{U, U \cup \{u_{N+1}\}, U \cup \{\overline{u_{N+1}}\}\}$.

$$h_{U'}^{N+1} \in \text{extr}_N(h_U^N) \leftrightarrow (U' = U \vee U' = U \cup \{u_{N+1}\} \vee U' = U \cup \{\overline{u_{N+1}}\})$$

Proposition 10. *The union of the extrapolations of the functions $f \in M_N$ is equal to the class of functions M_{N+1}.*

$$\forall N \geq 0 \cdot \bigcup_{f \in M_N} extr_N(f) = M_{N+1}$$

Definition 11. Projections of a description. The projections of a description are constructed by extending the description by a single new bit.

$$\forall d \in D_N, d' \in D_{N+1} \cdot d' \in \text{proj}_N(d) \leftrightarrow \forall 1 \leq i \leq N \cdot (d)_i = (d')_i$$

Definition 12. Projections of a case-base. The projections of a case-base are constructed by adding a new bit, set to one specified value, to the description of each exemplar in the case-base.

$$P_i^N(CB) = \{(d', n) | (d, n) \in CB \wedge d' \in \text{proj}_N(d) \wedge (d')_{N+1} = i\}$$

Proposition 13. *Given a function $f^{N+1} \in B_{N+1}$ defined on D_{N+1} and a second function $h_U^N \in M_N$ defined on D_N, it is concluded that $f^{N+1} = h_U^{N+1}$, i.e. f^{N+1} is the function on D_{N+1} represented by the same monomial expression U, if it can be shown that for any description $d \in D_N$, $h_U^N(d)$ will return the same value as $f^{N+1}(d')$, where d' is either of the projections of d in D_{N+1}.*

$$\forall N \geq 1 \cdot \forall f^{N+1} \in B_{N+1}, h_U^N \in M_N \cdot \forall d \in D_N, d' \in D_{N+1} \cdot$$
$$d' \in proj_N(d) \rightarrow [(h_U^N(d) = 1 \leftrightarrow f^{N+1}(d') = 1)) \rightarrow f^{N+1} = h_U^{N+1}]$$

Proposition 14. *For a given case base CB containing exactly one positive exemplar, if there is a function $f \in M_{N,k}$ s.t. $CB \subseteq f$, then for any larger k' s.t. $k \leq k' \leq N$, there is some $f' \in M_{N,k'}$ so that also $CB \subseteq f'$.*

$$\forall N \geq 1 \cdot \forall 1 \leq k \leq N \cdot \forall CB \in \mathcal{P}\,(D_N \times \{0,1\}) \cdot$$
$$(\#\{d_{pos} : D_N | (d_{pos}, 1) \in CB\} = 1 \rightarrow$$
$$\forall f \in M_{N,k} \cdot CB \subseteq f \rightarrow \forall k \leq k' \leq N \cdot \exists f' \in M_{N,k'} \cdot CB \subseteq f')$$

The following result can now be established:

Proposition 15. *The effective hypothesis space $H_{M_{N,k}}^{CB1(\sigma_F)}$ of the case-based learning algorithm $CB1(\sigma_F)$, defined with respect to the 'unweighted feature count' similarity measure σ_F and the set of k-literal monomial functions $M_{N,k}$, contains the set of all monomial functions M_N defined on D_N.*

$$\forall N \geq 1 \cdot \forall 1 \leq k \leq N \cdot M_N \subseteq H_{M_{N,k}}^{CB1(\sigma_F)}$$

Proof. By induction on N. Proposition 8 shows that the required result is equivalent to requiring that for each $f \in M_N$, there is a 'target concept' $t \in M_{N,k}$ for any value $1 \leq k \leq N$, and some case base $CB \subseteq t$, such that $h^N_{\langle CB,\sigma_F\rangle} = f$. Therefore, it will be sufficient to show $\forall N \geq 1 \cdot H(N)$, defining H as below. Introducing the extra restriction that case bases contain a single positive exemplar will allow reference to proposition 14 in subsequent argument:

$$H(N) \triangleq \forall f \in M_N \cdot \forall 1 \leq k \leq N \cdot \exists t \in M_{N,k} \cdot \exists CB \subseteq t \cdot p^+(CB) \wedge h^N_{\langle CB,\sigma_F\rangle} = f$$

where $p^+(CB) \triangleq \#\{d_{pos} : D_N|(d_{pos}, 1) \in CB\} = 1$.

Base Case $H(1)$. $M_1 = \{\{\}, \{u_1\}, \{\overline{u_1}\}\}$. $h_{\langle\{(1,1)\},\sigma_F\rangle} = h_{\{\}}$, $h_{\langle\{(1,1),(0,0)\},\sigma_F\rangle} = h_{\{u_1\}}$ and $h_{\langle\{(1,0),(0,1)\},\sigma_F\rangle} = h_{\{\overline{u_1}\}}$. Hence $H(1)$.

Inductive Step $H(p) \rightarrow H(p+1)$. We make the inductive hypothesis $H(p)$:

$$\forall f \in M_p \cdot \forall 1 \leq k \leq p \cdot \exists t \in M_{p,k} \cdot \exists CB \subseteq t \cdot p^+(CB) \wedge h^p_{\langle CB,\sigma_F\rangle} = f \quad (9)$$

Proposition 10 indicates that it will be sufficient to infer from equation (9) that for any monomial function $f \in M_p$ each extrapolation of f is a member of the hypothesis space with respect to $M_{p+1,k}$ for values $1 \leq k \leq p+1$. Proposition 14 in turn shows that it will be sufficient to derive from the inductive hypothesis that for each $f' \in \text{extr}_p(f)$ there is a $t \in M_{p+1,1}$ and a case-base $CB \subseteq t$ containing just one positive exemplar which represents f', which will entail the results for all other values of k.

Hence it will be shown equation (9) entails that for each $h^p_U \in M_p$ there are functions t_1, t_2 and t_3 and case bases CB_1, CB_2 and CB_3 satisfying:

$$\forall h^p_U \in M_p \cdot \exists t_1 \in M_{p+1,1} \cdot \exists CB_1 \subseteq t_1 \cdot p^+(CB_1) \wedge h^{p+1}_{\langle CB_1,\sigma_F\rangle} = h^{p+1}_U \quad (10)$$

$$\forall h^p_U \in M_p \cdot \exists t_2 \in M_{p+1,1} \cdot \exists CB_2 \subseteq t_2 \cdot p^+(CB_2) \wedge h^{p+1}_{\langle CB_2,\sigma_F\rangle} = h^{p+1}_{U\cup\{u_{p+1}\}} \quad (11)$$

$$\forall h^p_U \in M_p \cdot \exists t_3 \in M_{p+1,1} \cdot \exists CB_3 \subseteq t_3 \cdot p^+(CB_3) \wedge h^{p+1}_{\langle CB_3,\sigma_F\rangle} = h^{p+1}_{U\cup\{\overline{u_{p+1}}\}} \quad (12)$$

For any function $h^p_U \in M_p$, equation (9) asserts there must be some case base CB s.t. there is some $h^p_T \in M_{p,1}$ where $CB \subseteq h^p_T$ and $h^p_{\langle CB,\sigma_F\rangle} = h^p_U$. It will be shown that there are case-bases defined in terms of CB and T which will satisfy each of equations (10) to (12):

a) Case-based representation of h^{p+1}_U.

It will be shown that either projection of CB (definition 12), $P^p_0(CB)$ and $P^p_1(CB)$ is a case-based representation of h^{p+1}_U; clearly either projection also has a single positive exemplar. It will first be shown that for any function $h^N_{\langle CB,\sigma_F\rangle}(d)$ defined on D_N the functions represented by the projections of CB will classify the projections of d positively iff $h^N_{\langle CB,\sigma_F\rangle}(d) = 1$:

$$\forall N \geq 1 \cdot \forall i \in \{0,1\} \cdot \forall h^N_{\langle CB,\sigma_F\rangle} \in B_N \cdot$$
$$\forall d \in D_N, d' \in \text{proj}_N(d) \cdot (h^N_{\langle CB,\sigma_F\rangle}(d) = 1 \leftrightarrow h^{N+1}_{\langle P^N_i(CB),\sigma_F\rangle}(d') = 1) \quad (13)$$

Assume there is some $d \in D_N$ such that $h_{\langle CB,\sigma_F\rangle}(d) = 1$, and let d' be a projection of d in D_{N+1}. There must be a positive exemplar in CB satisfying equation (2). For any $d_1, d_2, d_3 \in D_N$ where $\sigma_F(d_1,d_2) > \sigma_F(d_1,d_3)$, consider the projections of d_1 in D_{N+1}, $d_1' \in \text{proj}_N(d_1)$. Consider also projections of d_2 & d_3, $d_2' \in \text{proj}_N(d_2), d_3' \in \text{proj}_N(d_3)$, such that $(d_2')_{N+1} = (d_3')_{N+1}$. Let $\gamma_{i,j}$ stand for the number of bits which d_i and d_j agree on; similarly, let $\gamma_{i',j'}$ stand for the number of bits agreed on by d_i' and d_j'. Since the extending bit $(d_1')_{N+1}$ will either agree or disagree with the bit extending d_2 and d_3, we have $\gamma_{1',2'} - \gamma_{1,2} = \gamma_{1',3'} - \gamma_{1,3} = \delta$, where $\delta \in \{0,1\}$. Therefore we also have $\sigma_F(d_1',d_2') > \sigma_F(d_1',d_3')$, and, letting $d' = d_1'$, any d_{pos} from the projection of the case-base $= d_2'$ and any $d_{neg} = d_3'$:

$$\forall i \in \{0,1\} \cdot \exists (d_{pos},1) \in P_i^N(CB) \cdot \forall (d_{neg},0) \in P_i^N(CB) \cdot \sigma_F(d',d_{pos}) > \sigma_F(d',d_{neg}) \tag{14}$$

and $h^{N+1}_{\langle P_i^N(CB),\sigma_F\rangle}(d') = 1$, $i \in \{0,1\}$. Similarly $h^N_{\langle CB,\sigma_F\rangle}(d) = 0 \rightarrow$ $h^{N+1}_{\langle P_i^N(CB),\sigma_F\rangle}(d') = 0$; hence (13).

We have shown that for any function $h^N_{\langle CB,\sigma_F\rangle}$, either projection of CB (definition 12) will represent a new function defined on D_{N+1} which classifies the projections of d positively iff $f^N(d) = 1$. Therefore by proposition 13:

$$h^{p+1}_{\langle P_0^p(CB),\sigma_F\rangle} = h^{p+1}_{\langle P_1^p(CB),\sigma_F\rangle} = h_U^{p+1} \tag{15}$$

It remains only to show $P_i^p(CB) \subseteq h_T^{p+1}$. For any $(d',n) \in P_i^p(CB)$, there is a unique d such that $d' \in \text{proj}_p(d)$ and $(d,n) \in CB$ (definition 12). Since $CB \subseteq h_T^p$, $(d,n) \in CB \rightarrow h_T^p(d) = n$. Since d & d' agree on their first p bits and also $h_T^p \in M_{p,k}$ so that T refers only to the first p bits of representation, $h_T^p(d) = 1 \leftrightarrow h_T^{p+1}(d') = 1$. Hence also $h_T^{p+1}(d') = n$ and therefore $(d',n) \in P_i^p(CB) \rightarrow h_T^{p+1}(d') = n$. Hence the following result, concluding (10):

$$P_i^p(CB) \subseteq h_T^{p+1} \tag{16}$$

b) Case-based representation of $h^{p+1}_{U\cup\{u_{p+1}\}}$. It will be shown that the case base $P_1^p(CB) \cup \{(d_{new},0)\}$ is a case-based representation of $h^{p+1}_{U\cup\{u_{p+1}\}}$, where d_{new} is defined as follows:

$$\begin{aligned} (d_{new})_x &= |1 - (d_{pos})_x| \\ (d_{new})_i &= (d_{pos})_i \ \textit{where} \ \ 1 \le i \le p \ \wedge \ i \neq x \\ (d_{new})_{p+1} &= 0 \end{aligned}$$

d_{pos} is the description of the unique positive exemplar in $P_1^p(CB)$, inherited from CB, and x is the smallest value s.t. $u_x \in T \vee \overline{u_x} \in T$, T being the representation of the target function $h_T^p \in M_{p,1}$.

By equation (2), we have $h^{p+1}_{\langle P_1^p(CB)\cup\{(d_{new},0)\},\sigma_F\rangle} = f^{p+1}$, where:

$$f^{p+1}(d) = \begin{cases} 1 \ \textit{if} \ h^{p+1}_{\langle P_1^p(CB),\sigma_F\rangle}(d) = 1 \wedge \sigma_F(d,d_{pos}) > \sigma_F(d,d_{new}) \\ 0 \ \textit{otherwise} \end{cases} \tag{17}$$

From the definition of d_{new}, we have $\sigma_F(d, d_{pos}) > \sigma_F(d, d_{new})$ iff d agrees with d_{pos} on a strict majority of the bits $\{u_x, u_{p+1}\}$; note $(d_{pos})_{p+1} = 1$ since $(d_{pos}, 1) \epsilon P_1^p(CB)$, while $(d_{new})_{p+1} = 0$ by definition. (All other bits are irrelevant to the comparison since they are common to both d_{pos} and d_{new}). Hence:

$$\sigma_F(d, d_{pos}) > \sigma_F(d, d_{new}) \leftrightarrow ((d)_x = (d_{pos})_x \wedge (d)_{p+1} = 1) \tag{18}$$

Substituting (15) and (18) in (17):

$$h^{p+1}_{(P_1^p(CB) \cup \{(d_{new}, 0)\}, \sigma_F)} = f^{p+1} = h^{p+1}_{U \cup \{u_{p+1}\}} \tag{19}$$

since $h_U^{p+1}(d) = 1$ implies that $(d)_x$ must have the same value as $(d_{pos})_x$.

Clearly, the new case-base still contains a single positive exemplar; to satisfy equation (11), it must only be shown $P_i^p(CB) \cup \{(d_{new}, 0)\} \subseteq h_T^{p+1}$. From (16), we have $P_i^p(CB) \subseteq h_T^{p+1}$. Note also $h_T^{p+1}(d_{new}) = 0$ since by definition, d_{new} will fail to satisfy T. Hence $(d_{new}, 0) \epsilon h_T^{p+1}$ and $P_1^p(CB) \cup \{(d_{new}, 0)\} \subseteq h_T^{p+1}$.

c) Case-based representation of $h^{p+1}_{U \cup \{\overline{u_{p+1}}\}}$. Equally, the case base $P_0^p(CB) \cup \{(d'_{new}, 0)\}$, where $d'_{new} = d_{new}$ as defined as above except $(d'_{new})_{p+1} = 1$, is an equivalent representation to $U \cup \{\overline{u_{p+1}}\}$. Hence (12). □

Corollary 16. Lower bound on VC Dimension of $H^{CB1(\sigma_F)}_{M_{N,k}}$. *The VC Dimension of* $H^{CB1(\sigma_F)}_{M_{N,k}}$, *the effective hypothesis space of* $CB1(\sigma_F)$ *with respect to the concept space* $M_{N,k}$, *is at least* $\mathcal{O}(N)$.

Proof. $H^{CB1(\sigma_F)}_{M_{N,k}}$ contains M_N (proposition 15). Therefore any sample shattered by M_N will be shattered by $H^{CB1(\sigma_F)}_{M_{N,k}}$, and the VC dimension of $H^{CB1(\sigma_F)}_{M_{N,k}}$ will be at least that of M_N, which is $\mathcal{O}(N)$ [3, p.76] [9, p.193]. □

In contrast to corollary 16, note the following result:

Proposition 17. Upper bound on VC Dimension of $M_{N,k}$. *The VC Dimension of* $M_{N,k}$ *is no greater than* $1 + \log_2 \binom{N}{k}$.

Proof. Let $\overline{x}$ be a sample of size v, which orders the set of examples X and is shattered by $M_{N,k}$. Consider that there are 2^{v-1} subsets of X which contain a particular $x_i \epsilon X$, and also that there are exactly $\binom{N}{k}$ functions $f \epsilon M_{N,k}$ that classify x_i positively. Since each subset of X must be labelled by a distinct member of $M_{N,k}$, we have $2^{v-1} \leq \binom{N}{k}$, and hence $v \leq 1 + \log_2 \binom{N}{k}$ □

Hence, while the VC dimension of the hypothesis space of $CB1(\sigma_F)$ with respect to the set of functions $M_{N,k}$ is at least $\mathcal{O}(N)$ (Corollary 16), the VC dimension of $M_{N,k}$ itself is $\mathcal{O}(\log N)$ (Proposition 17). Theorem 7 leads us to

believe that this qualitative difference, reflecting the presence of a number of spurious functions in the hypothesis space of $CB1(\sigma)$ in this instance, indicates that $CB1(\sigma_F)$ is a less than optimal learning algorithm (with respect to sample complexity) for the space $M_{N,k}$. That is, as N increases, we would expect the number of examples $CB1(\sigma_F)$ needs to reach an accurate hypothesis will rapidly outgrow the number needed by a learning algorithm whose hypothesis space represents exactly the functions contained in $M_{N,k}$.

Finally, we note that Proposition 15 is very much a partial characterisation of the hypothesis space in this instance. In addition to that formal result, direct enumeration establishes the presence of functions such as $u_1 + u_2.u_3$ and $u_1.u_2 + u_1.u_3 + u_2.u_3$ in $H_{M_{3,1}}^{CB1(\sigma_F)}$, and in addition shows that only a fraction of B_N is output as hypotheses on training samples for functions in $M_{N,k}$ and that $|H_{M_{N,k}}^{CB1(\sigma_F)}|$ varies for different values of k.

6 Conclusions

$CB1(\sigma_F)$ is a general purpose learning algorithm with a rich hypothesis language. Specifically, for any fixed *predictive* similarity measure (Definition 2) such as σ_F, corollaries 6 and 5 indicate that there is a case-based representation $\langle CB, \sigma_F \rangle$ for any $\{0,1\}$-valued total function on D_N. In addition we have explored the nature of the hypothesis space of $CB1(\sigma_F)$. Considering the possible hypotheses that might be output on training samples for functions in specific concept spaces $M_{N,k}$, it has been shown here (Proposition 15) that the hypothesis space of $CB1(\sigma_F)$ with respect to the concept space $M_{N,k}$ includes not only $M_{N,k}$ but also *all* monomial functions M_N. Arguments related to Theorem 7 lead us to believe in addition that the presence of these spurious hypotheses will make $CB1(\sigma_F)$ a relatively inefficient learning algorithm for $M_{N,k}$ (with respect to sample complexity) compared to a consistent learning algorithm which can represent *only* the functions $M_{N,k}$. We suggest that this is a natural corollary of the generality of $CB1(\sigma_F)$.

In contrast, Wess and Globig have already pointed out and ably demonstrated that "the [similarity] measure (respectively the way to modify the measure) is the bias of case-based reasoning" [15, p.90]. That is, with some prior knowledge of the concept space to be learnt, the similarity measure can be manipulated so that the hypotheses output by the case-based learner are more likely to be close to the possible target concepts. Such strategies demonstrably improve efficiency with respect to sample size [8] [15], although performance will obviously be degraded outside the chosen concept space.

Where more sophisticated case-based learning algorithms outperform a simple but universal algorithm such as $CB1(\sigma_F)$, this must be seen as the result of some bias in the learning algorithm to the target concepts that the algorithms are being tested on. We believe that the formalisation presented here and its attention to the hypothesis space of the case-based learner provide a tool for the rigorous comparison of the many possible case-based learning algorithms and the different forms of bias they embody. Much work remains in carrying out these

comparisons and in extending the model, for example to allow for the possibility of observational error in the cases of the case base.

Acknowledgements The first author is funded by an EPSRC grant and receives additional support under the CASE award scheme from Logica Cambridge Ltd. The authors would like to thank Robert Dormer for leading the way into Computational Learning Theory.

References

1. D W Aha, D Kibler, and M K Albert. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.
2. M K Albert and D W Aha. Analyses of instance-based learning algorithms. In *AAAI-91: Proceedings of the Ninth National Conference on Artificial Intelligence*, pages 553–558, 1991.
3. M Anthony and N Biggs. *Computational Learning Theory*. Cambridge University Press, 1992.
4. A Blumer, A Ehrenfeucht, D Haussler, and M K Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, Oct 1989.
5. S Cost and S Salzberg. A weighted nearest neighbour algorithm for learning with symbolic features. *Machine Learning*, 10(1):37–66, Mar 1993.
6. W H E Day and D P Faith. A model in partial orders for comparing objects by dualistic measures. *Mathemetical Biosciences*, 78(2):179–192, 1986.
7. A M Dearden and M D Harrison. The engineering of case memory systems. *submitted to the Journal of Intelligent Information Systems.*
8. C Globig and S Wess. Symbolic learning and nearest-neighbour classification. In *Proceedings of the 17th Annual Conference of the Gesellschaft fur Klassification e.V. University of Kaiserslautern, March 3-5, 1993*. Springer-Verlag, 1994.
9. D Haussler. Quantifying inductive bias: AI learning algorithms and Valiant's learning framework. *Artificial Intelligence*, 36:177–221, 1988.
10. D Haussler. Probably approximately correct learning. In *AAAI-90 Proceedings of the Eight National Conference on Artificial Intelligence, Boston, MA*, pages 1101–1108. American Association for Artificial Intelligence, 1990.
11. K P Jantke. Case-based learning and inductive inference. GOSLER report 08/92, FB Mathematik & Informatik, TH Leipzig, 1992.
12. K P Jantke and S Lange. Case-based representation and learning of pattern languages. In *EWCBR-93 Working Notes of the first European Workshop on Case-Based Reasoning*, volume 1, pages 139–144. University of Kaiserslautern, 1993.
13. E L Rissland, J Kolodner, and D Waltz. Case-based reasoning. In *Proceedings of DARPA Case-Based Reasoning Workshop May 1989*, pages 1–13. Morgan Kaufmann, 1989.
14. P Turney. Theoretical analyses of cross-validation error and voting in instance-based learning. Technical Report NRC-35073, Knowledge Systems Laboratory, Institute for Information Technology, National Research Council (Canada), 1993.
15. S Wess and C Globig. Case-based and symbolic classification algorithms - A case study using version space. In *Topics in CBR: Selected papers from EWCBR-93*, LNCS vol. 837, pages 77–91. Springer-Verlag, 1994.