

A Definition of Learner Uncertainty

Chris Thornton

Cognitive and Computing Sciences
University of Sussex
Brighton
BN1 9QH
UK

Email: Christopher.Thornton@firenet.uk.com

Tel: (44)1273 678856

May 21, 2003

Abstract

The paper considers the quantification of inductive bias in concept learning. It argues that some of the best known measures of bias (eg. Vapnik-Chervonenkis dimension) are *context-free* in the sense that they measure the effectiveness of bias in general, rather than with respect to a particular problem. This approach has many advantages but in cases where we want to evaluate the performance of a particular algorithm on a particular problem we need a context-sensitive measure, ie. a way of measuring bias that takes into account the characteristics of the relevant problem. The paper describes a generalization of the ID3 ‘information-needed’ concept that provides such a measure. It also shows how the measure can be used to predict learning performance in simple concept learning experiments.

1 Introduction

In general, learning mechanisms require some background knowledge in order to be able to operate successfully [1]. However, when we are trying to evaluate the performance of a given learner on a given task, we need to know how much of the generated solution is due to the algorithm and how much to the background knowledge. In particular, we want to be able to decide whether the solution provided by the learner is implicitly contained in the background knowledge since in this case no real learning would have taken place. [2]

In the familiar case of concept learning it is conventional to think of the learner as searching through a space of hypotheses for one that satisfies the

constraints¹ imposed by the training set [3]. In this model, the only way background knowledge can affect the learning is by causing the learner to prefer certain hypotheses over others, ie. by causing it to search the space of hypotheses in a directed way. If the background knowledge is effective, the learner will be guided speedily towards a satisfactory hypothesis. A learner that is affected by background knowledge in this way is said to have an *inductive bias*. A strong bias is one that causes the learner to focus on a relatively small number of hypotheses. A correct bias is one which causes the learner to focus on (or move towards) satisfactory hypotheses.

Measuring the effectiveness of background knowledge in concept learning (or any learning that can be viewed as hypothesis space search) is thus largely the same thing as measuring the inductive bias of the relevant learner. A general framework that enables us to quantify inductive bias has been provided by Haussler [4,5,6]. This framework makes use of the PAC learning model of Valiant [7,8] and the concept of the VC (Vapnik-Chervonenkis) dimension [9]. A remarkable attribute of Haussler's framework is that it is essentially 'context-free'. It allows us to quantify the inductive bias of an algorithm in general, rather than with respect to some specific problem. Thus it allows us to measure the general effectiveness of background knowledge rather than its effectiveness with respect to a particular learning problem.

Clearly, context-free measures of inductive bias are exactly what we want where we are interested in the general properties of learning algorithms, or where the aim is to derive generic learnability results. The power of the context-free approach in achieving this end has been amply demonstrated [6,10]. But, of course, context-free measures cannot help us in the case where we need a *context-sensitive* estimate of bias, ie. a measure that tells us to what extent the background knowledge yields a solution to a *particular* problem. Haussler has suggested that obtaining a context-sensitive measure might involve further extensions of the PAC learning model. However, the present paper explores the possibility of deriving a measure from information theory, which is similar to the method for computing the 'information deficit' of decision-tree nodes (as used in, e.g., [11,12]). It shows how we can modify the method to produce a measure of 'hypothesis uncertainty' and then derive a 'learner uncertainty' measure as the mean of the uncertainties for the relevant hypothesis set. This uncertainty measure forms an inverse measure of the effectiveness of the learner's inductive bias on the problem in question. The paper also presents an example in which the learner uncertainty measure is used to predict the positions of learning curves.

2 A review of Haussler's framework

Haussler's framework is based on the familiar concept learning scenario and in particular on Valiant's PAC learning model. In this section we present a

¹We assume noise-free training data.

thumbnail sketch of this framework and show briefly how it is used to provide context-free measures of bias.

Let X be an instance space based on a fixed set of attributes and let I be a finite set of instances in X . Let H be a hypothesis space defined on X . Each hypothesis h covers a certain set of instances and excludes all others. A *dichotomy* of I is a way of partitioning up all the instances in I into positives and negatives. The *dichotomy of I induced by h* is the partitioning that we get if we label all the instances that are covered by h as positive and all the rest as negative. The *dichotomies of I in H* are simply those dichotomies that are generated by hypotheses in H .

In concept learning, we search H for a hypothesis that agrees with some target concept c on all instances (where c is a partitioning of the instances into positives and negatives). In general, H is large. But we assume that the learner has some background knowledge that enables it to move in a more directed way towards a satisfactory hypothesis. In fact, for most purposes it is more convenient to adopt a slightly different perspective. We view learning in the presence of background knowledge as a *blind search* of a hypothesis space that contains only those hypothesis that are compatible with the background knowledge. The great advantage of this construal is that it presents a picture of learning in which there is a direct connection between the effectiveness of the background knowledge (ie. the inductive bias) and the *size* of the hypothesis space.

If the inductive bias is relatively effective, we will expect H to be smaller and to contain hypotheses that are relatively likely to be satisfactory. If the bias is relatively ineffective then we will expect the opposite. Thus we see that the effectiveness of the bias is reflected in the size and composition of the hypothesis space. And, indeed, Haussler has shown that in some situations we can get an indication of the strength of the bias simply by measuring the absolute size of the hypothesis space.

A more sophisticated approach involves measuring the way in which the number of dichotomies of I that can be induced by members of H increases as we increase the size of I . This relationship is called the *growth function* of H in I . If we have a very weak bias, then, as we increase the size of I , the number of dichotomies we can induce using members of H will stay close to the theoretical maximum (the total number of ways of splitting I into positives and negatives). If, on the other hand, we have a very strong bias (ie. a restricted hypothesis space) then the number of dichotomies we can induce will quickly drop below the theoretical maximum. Thus the shape of the growth function gives us another indication of the strength of the bias.

Ideally, we want to have the bias quantified as a single, numeric value. And for this purpose we can use the Vapnik-Chervonenkis dimension or VC dimension. This is simply the highest point on the growth-function curve at which the number of dichotomies of I in H is equal to the total number of possible dichotomies of I ; ie. it is the point at which the bias actually begins to make itself felt. Haussler has proposed various other ways of quantifying the bias but the VC dimension measure has significant attractions. Not only does it provide

us with a single, numeric value but it also enables performance and learnability bounds to be derived for a number of learning algorithms [6].

3 Context-free versus context-sensitive bias measures

As we noted, in Haussler's analysis the measures of bias are context free. They take into account the properties of the instance space and the hypothesis space, but they are not founded on any particular learning problem (ie. any particular way of labelling instances as positive and negative). This means that, in general, we cannot use these measures in order to show why a particular learner solves problem P_1 more easily than problem P_2.² In some circumstances this can be a limitation since situations often arise in practice where it is necessary to critically evaluate the performance of a particular algorithm on a particular problem. We often want to know whether the performance involves learning, or is, in fact, a trivial consequence of the available background knowledge. The very real nature of this problem is demonstrated by the controversies that have raged over the learning feats of such as the NETtalk system [13] and the AM and EURISKO systems [14,15].

There are a number of ways in which we might try to derive a context-sensitive measure of bias. The one described here is based on information theory and has been derived as a generalization of the standard method for computing information deficits. The mathematical basis of the measure is Shannon's information formula. In its most basic form, this states that the amount of information contained in any message (or in general, any event) is inversely and logarithmically related to the a priori probability of that message [16]. The basic equation has the form

$$I = -\log P_i$$

Here, I is the information content and P_i is the a priori probability of the i th message.³ In the case where there is no fixed a priori probability, it is appropriate to use the probability as it appears to the receiver [17].

The meaning of the equation is roughly summarised by saying that the amount of information is equal to the receiver's uncertainty that the message in question will be received. The general equation in information theory allows us to measure the receiver's uncertainty with respect to a whole set of possible messages, provided that the perceived probabilities of the messages sum to 1. The uncertainty is defined by the *entropy* equation:⁴

$$-\sum_{i=1}^n P_i \log P_i$$

²Of course if the two problems are drawn from different classes then Haussler's approach can be used.

³If we want to measure information in bits we should take logs to base 2.

⁴We have omitted the constant factor here.

given that

$$\sum_{i=1}^n P_i = 1$$

This measures the overall uncertainty (or expected surprise) with respect to a forthcoming message [18].

The information measures have been used for various purposes in machine learning. But here we focus on the way in which Quinlan used information theory in formulating the splitting heuristic of the ID3 learning algorithm [12]. In ID3 decision trees are constructed by repeatedly splitting the available instances on the features of selected attributes. The information theoretic heuristic is applied at the point where an attribute must be selected on which to split the instances at a given leaf node. The essential idea is that an ideal split should create a set of child nodes whose average information deficit⁵ is minimized [11].

The information deficit of node X is simply the amount of information required to produce a true classification of an arbitrary instance that currently classifies at X . To compute this, we look at the relative frequencies of positive and negative instances at X and derive probabilities for the unseen instance being positive and negative. By feeding these probability estimates into the general information formula we derive the information deficit or uncertainty value. A high information deficit value means that there is a roughly equal number of positives and negatives at X and that we are therefore relatively uncertain as to whether an unseen instance classifying at X will be positive or negative. If the instances at X are mainly positives (negatives) then the probability that the unseen instance is positive (negative) is high. In this case we have a low information deficit. To a first approximation the splitting heuristic causes selection of an attribute that minimizes the average information deficit (uncertainty) of child nodes.

4 Hypothesis and learner uncertainty

We now turn to the main task of deriving a context-sensitive measure of inductive bias. The essential idea is to use the information-deficit measure in a novel way. Rather than using it to derive the uncertainty levels arising from decision tree leaf nodes, we will use it to derive the uncertainty levels arising from complete hypotheses. The underlying procedure however will remain the same. From the relative frequencies of positives and negatives covered by the hypothesis we will derive probability values. And from these probability values we will derive an uncertainty value.

We think of each hypothesis in H as a distinct way of representing the target concept c . Each hypothesis partitions the set of instances in a different way. So using a particular hypothesis, we must represent the division of instances into positive and negative examples of the concept in terms of the partition defined

⁵Quinlan uses the term *information needed*.

by the hypothesis. Some hypotheses will agree with the target concept on all instances. But most will disagree in some cases. Such a hypothesis will cover (and exclude) a mix of positive and negative examples. So we can compute the level of uncertainty (with respect to the classification of an unseen instance) of any mechanism using h as a way of representing c by deriving the relevant probabilities.

We first compute the probability that an instance covered by h is positive by looking at the relative frequency of positives covered by h . We then compute the probability that an instance covered by h is negative by looking at the relative frequency of negatives covered by h . Feeding these two probability values (that necessarily sum to 1) into the information value gives us an uncertainty value for a classification of an arbitrary instance covered by h . We can then repeat the procedure for instances excluded by h and derive an overall uncertainty value by averaging the two results.

In the context of this model, the learner’s uncertainty can be measured by applying the entropy formula to the relevant probabilities. Formally, we define the hypothesis uncertainty associated with h as

$$U_h = \frac{-h_{inc}^+ \log h_{inc}^+ - h_{inc}^- \log h_{inc}^-}{2} + \frac{-h_{exc}^+ \log h_{exc}^+ - h_{exc}^- \log h_{exc}^-}{2}$$

Here, h_{inc}^+ is the probability that an instance covered by h is a positive (ie. the relative frequency with which instances covered by h are positives) and h_{inc}^- is the probability that an instance covered by h is a negative. h_{exc}^+ and h_{exc}^- are the corresponding values for excluded instances. In measuring hypothesis uncertainty we are effectively computing two ID3 information-deficit values: (1) the information deficit for that part of the instance space that is covered by the hypothesis and (2) the deficit value for that part of the space that is excluded.

Having defined hypothesis uncertainty, it is now straightforward to derive a measure for learner uncertainty; ie. a measure of the learner’s uncertainty given a particular target concept and a space of hypotheses with respect to the classification of an arbitrary instance. We define the *learner uncertainty* U of some learner L to be the average of the hypothesis uncertainties for every hypothesis in L ’s hypothesis space. The disadvantage of this is that the measure is undefined in the case where the hypothesis space is infinite. However, in these cases we might use sampling techniques to derive a best guess.⁶

The fact that learner uncertainty provides an effective, context-sensitive measure of inductive bias can be seen by considering the way in which learner uncertainty varies as we manipulate the hypothesis space. Consider the task of learning concept c . As we have noted, background knowledge in the learner can make itself felt in two ways: as a reduction in the overall size of H or as an increase in the quality of the hypotheses in H (ie. their chances of being

⁶As is often the case with entropy-based measures, the quantitative interpretation of U (ie. the meaning we should assign to particular values of U) is problematic. We do not address this issue here.

satisfactory). High-quality hypotheses will obviously tend to split the instances up in the same way as the target concept. But if they do this, the derived probability values will tend to be extreme and hypothesis uncertainties will tend to be very low. On the other hand, if the quality of hypotheses is low, they will tend to split the instances up in ways that disagree with the target concept. If they do this the derived probability values will tend to be intermediate and the hypothesis uncertainties will be high. Thus the learner uncertainty measure appears to be sensitive to the degree to which background knowledge affects the quality of hypotheses in a particular problem scenario.

5 Example

Initial experiments suggest that the learner uncertainty measure provides a fairly good indication of learning performance. To show this, and also to demonstrate how learner uncertainty is worked out in a given case, we present a simple example using the ID3 learning algorithm [11,12] and a very simple set of concept learning problems. All the problems use the same instance space. Instances are 2-dimensional vectors giving a colour attribute and a shape attribute for a blocksworld block; eg. [blue wedge], [red brick] etc. The possible values of the shape attribute are ‘wedge’, ‘brick’ and ‘sphere’; the possible values of the colour attribute are ‘red’, ‘green’ and ‘blue’. Hypotheses are decision trees based on the two attributes.

Since there are only two features we can draw out the attribute space as a 2-dimensional diagram as shown in Figure 1. Each cell in this diagram represents a possible instance so we can represent a particular target concept by labelling all the cells to show whether the corresponding instance is a positive or a negative. In Figure 1 we see three different concepts each represented by a rectangular matrix of cells. The positive cells in a given matrix denote those instances that are positive examples of the corresponding concept, and the negative cells denote the negatives. Under each matrix we see the relevant learner uncertainty value for a learner using decision trees as hypotheses.

To derive the learner uncertainty values we took each hypothesis in turn (ie. each possible decision tree) and computed the hypothesis uncertainty for that tree using the formula shown above. We then averaged over all the hypothesis uncertainties to produce the final learner uncertainty values.

In Figure 2 we see some learning curves that were produced by applying ID3 to the three target concepts. Each curve is labelled with the number of the relevant concept and the relevant uncertainty value. To generate the curves, we took each concept in turn and, for all values of n between 1 and 9 (the total number of instances) generated 50 random training sets of size n . We then took all the training sets of size n , ran ID3 on each one. The average error rate⁷ of the decision tree produced was then derived. This process produced average error rate values for all sample sizes between 1 and 9 and a set of such values

⁷The average error rate in this case is simply the frequency with which the decision tree disagrees with the target concept over the classification of an instance.

	blue	green	red
wedge	+	-	-
brick	+	-	-
sphere	+	-	-

1: $U = 1.39$

	blue	green	red
wedge	-	-	-
brick	+	+	-
sphere	+	-	-

2: $U = 1.43$

	blue	green	red
wedge	+	-	-
brick	-	+	-
sphere	-	-	+

3: $U = 1.49$

Figure 1:

forms one curve in the graph. Note that the curves tend to come together as

.so /pics/uncertainty_curves.picture

Figure 2:

the sample sizes approach the maximum size, but with samples of less than 6 instances, we can see that the learner uncertainty value corresponds fairly well to the north-easterly shift of the curve and therefore to the performance of the algorithm on the relevant problem. It is fairly straightforward to explain these performance variations. Decision trees are essentially disjunctive-normal-form formulae; each term in a conjunction specifies a value for one attribute, eg. *colour = blue*. Geometrically, each individual term generalizes over one row or one column of the matrix. Thus, the more the positive instances tend to confine themselves to some small number of rows and/or columns, the less terms we need in our conjunctions.

In concept 1, we have the ideal situation. The positives are all gathered into a single column. This means that there are a large number of hypotheses in the hypothesis space that yield extremal probabilities and therefore have a low uncertainty value. This implies that the bias is appropriate for this concept and

produces a relatively low learner uncertainty. Concept 3 represents the opposite extreme. Here our three positives have no attribute values in common and therefore never share the same row or column. As a result, many hypotheses will yield intermediate probabilities and therefore have high uncertainty values. The bias is *inappropriate* for this concept, the learner uncertainty is high and the performance is poor. Concept 2 represents the intermediate situation.

6 Discussion

The paper has attempted to relate Haussler’s framework for measuring inductive bias to standard information measures. It has shown that it is possible to derive a context-sensitive measure of inductive bias which is based on the notion of information deficit (ie. information gain). The measure provides a simple metric for quantifying the effectiveness of background knowledge with respect to a particular learning problem. As such it may be useful for the purposes of discriminating between trivial learning (ie. performance that is a straightforward consequence of available background knowledge) and non-trivial learning. Initial experiments using toy learning problems seem to suggest that the measure may be a reasonable indicator of learning performance; but more empirical work is required to decide whether this is the case.

It is worth stressing that there is no relationship between our U measure and the VC dimension. One is an absolute measure of hypothesis-language power; the other is a relative measure. For a given hypothesis language we have a fixed VC dimension but we have a range of possible U values depending on what we take to be the learning task. In general we would expect to obtain the same U value for a given ‘power/problem ratio’, ie. a given ratio between the power of the hypothesis language and the difficulty of the problem. Thus we would expect get the same U value for an extremely primitive hypothesis language used with a very simply problem as we would from a very rich hypothesis language used with a very hard problem.

In conclusion, we should mention two major difficulties with the current framework. First of all, the U measure as currently defined is very expensive to compute. In practice it may be impossible to compute in non-toy situations. However, it may be that fairly accurate results can be obtained by standard approximation methods. Secondly, from the theoretical perspective, it is not satisfactory that the learner uncertainty measure does not take account of the ‘size’ impact of background knowledge, ie. the way in which it affects the absolute size of the hypothesis space. A possible solution to this problem might be to redefine the learner uncertainty measure in terms of the *union* of the probability sets generated from the various hypotheses. This would involve normalizing probability values to ensure that they sum to 1 but, since, other things being equal, large sets of probability values generate larger entropy values, it would have the advantage of making the impact of hypothesis-space size explicit.

References

- [1] Mitchell, T. (1980). The need for bias in learning generalizations. Technical Report CBM-TR-117, Dept. of Computer Science, Rutgers University.
- [2] Lenat, D. and Brown, J. (1984). Why AM and EURISKO appear to work. *Artificial Intelligence*, 23, No. 3 (pp. 269-294).
- [3] Rendell, L. (1986). A general framework for induction and a study of selective induction. *Machine Learning*, 1, No. 1 (pp. 177-226).
- [4] Haussler, D. (1987). Bias, version spaces and valiant's learning framework. *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 324-336). University of California, Irvine: (June 22-25).
- [5] Haussler, D. (1988). Quantifying inductive bias: AI learning and valiant's learning framework. *Artificial Intelligence*, 36 (pp. 177-221).
- [6] Haussler, D. (1990). Probably approximately correct learning. *Proceedings of the Eighth National Conference on Artificial Intelligence*. Vol. Two, Cambridge, Mass.: MIT Press (July 29 1990-Aug 3 1990).
- [7] Valiant, L. (1984). A theory of the learnable. *Communications of the ACM*, 27 (pp. 1134-42).
- [8] Valiant, L. (1985). Learning disjunctions of conjunctions. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (pp. 560-566). Los Altos: Morgan Kaufmann.
- [9] Vapnik, V. and Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theor. Probab. Appl.*, 16, No. 2 (pp. 264-280).
- [10] Haussler, D., Kearns, M. and Schapire, R. (1992). Bounds on the sample complexity of bayesian learning using information theory and the VC dimension. UCSC-CRL-91-44, University of California at Santa Cruz.
- [11] Quinlan, J. (1983). Learning efficient classification procedures and their application to chess end games. In R. Michalski, J. Carbonell and T. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach*. Palo Alto: Tioga.
- [12] Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, 1 (pp. 81-106).
- [13] Sejnowski, T. and Rosenberg, C. (1987). Parallel networks that learn to pronounce english text. *Complex Systems*, 1 (pp. 145-68).
- [14] Lenat, D. (1982). AM: discovery in mathematics as heuristic search. In R. Davis and D.B. Lenat (Eds.), *Knowledge-Based Systems in Artificial Intelligence* (pp. 1-225). New York: McGraw-Hill.

- [15] Lenat, D. (1983). Theory formation by heuristic search; the nature of heuristics II: background and examples. *Artificial Intelligence*, 21 (pp. 31-59).
- [16] Shannon, C. (1949). A mathematical theory of communication. The Mathematical Theory of Communication, University of Illinois Press.
- [17] Mackay, D. (1969). *Information, Mechanism and Meaning*. London: MIT Press.
- [18] Watanabe, S. (1969). *Knowing and Guessing: A Quantitative Study of Inference and Information*. New York: Wiley.