

# Active Bias Testing and Adjustment for Concept Learning

Diana F. Gordon (gordon@aic.nrl.navy.mil)

Navy Center for Applied Research in Artificial Intelligence  
Naval Research Laboratory, Code 5514  
Washington, D.C. 20375-5000

## Abstract

Bias is a fundamental aspect of supervised concept learning. Nevertheless, selecting a good bias prior to learning is difficult. In response to this difficulty, systems have recently been developed that dynamically adjust the bias during incremental learning. These systems, however, are limited in their ability to identify erroneous assumptions about the relationship between the bias and the target concept. Without proper diagnosis, it is difficult to identify and then remedy faulty assumptions. We have developed a new approach that, unlike previous approaches, makes these assumptions explicit, actively tests them, and adjusts the bias based on the test results. When bias adjustment is appropriate for learning the target concept, our approach can produce a 10-fold improvement in the rate of convergence to the target concept over a baseline performance.

## Introduction

Concept learning may be viewed as a search through a space of hypotheses to find the *target concept*, i.e. the concept to be learned. If the learning is *supervised*, then training instances that are labeled "positive" or "negative" provide the learner with evidence for learning. The goal of most supervised concept learners is to formulate a hypothesis that is consistent with the instances. Consistency holds if a hypothesis covers all of the positive instances and none of the negative ones. Unfortunately, it is often the case that many hypotheses can be formed that are consistent with the instances. *Bias* is any basis for hypothesis preference other than strict consistency with the instances [Mitchell 1980]. For example, one might prefer simpler hypotheses or hypotheses

expressed in a particular language.

Bias is a fundamental aspect of any concept learning system. Numerous papers have noted this importance (e.g., [Mitchell 1980], [Michalski 1983], [Haussler 1988]). The type of bias that we discuss here is the hypothesis language. This language defines the space of hypotheses. A *strong* bias defines a small hypothesis space, a *weak* bias defines a large hypothesis space, and a *correct* bias defines a space that includes the target concept. Assuming a concept learner effectively selects hypotheses from its space of choices, a strong, correct bias is desirable since it reduces the number of choices and thereby promotes rapid convergence to the target concept.

Traditionally, the bias used by concept learning systems has been designed by the system implementor. There is a recent growing interest, however, in systems that can adjust (shift) the bias by making it stronger or weaker (e.g., [Utgoff 1986], [Rendell 1987]). These systems learn *incrementally*. In other words, not all of the training instances are presented to the system at once. Over the course of receiving instances from which to learn, these systems adjust the bias in response to performance feedback. The bias that they adjust is typically the set of instance features (attributes) in the hypothesis language. The process of dynamic bias adjustment often involves strengthening the bias whenever possible and weakening it to regain correctness.

Although systems exist that dynamically adjust the bias, these systems are limited in their ability to identify erroneous assumptions about the relationship between the bias and the target concept. Proper diagnosis aids in the recovery from faulty assumptions. We have developed a new approach to biasing that addresses the need for proper diagnosis. This