A Conservation Law for Generalization Performance

Cullen Schaffer
Department of Computer Science
CUNY/Hunter College
695 Park Avenue, New York, NY · 10021
schaffer@roz.hunter.cuny.edu

Abstract

Many aspects of concept learning research can be understood more clearly in light of a basic mathematical result stating, essentially, that positive performance in some learning situations must be offset by an equal degree of negative performance in others. We present a proof of this result and comment on some of its theoretical and practical ramifications.

1 Introduction

In the machine learning subfield of classification or concept learning, bias has long been understood to play a central role. Here we present a mathematical result that makes precise some of our intuitions about bias and discuss the implications for various aspects of classification research. Roughly speaking the result indicates that generalization is a zero-sum enterprise—for every performance gain in some subclass of learning situations there is an equal and opposite effect in others.

2 Conservation of Generalization Performance

Cases in a classification problem are described by attribute vectors. For simplicity, we assume that each component of such a vector may take on only a finite number of values, continuous attributes being discretized to any arbitrary degree of accuracy. This implies that there is a finite set of m possible attribute vectors $\{A_1, \ldots, A_m\}$.[1] For simplicity we also assume exactly two classes. The relationship between attribute vectors and classes may then be specified by an m-component class probability vector C, where C_i is the probability that a case with attribute vector A_i is of class 1. Deterministic relationships and relationships involving attribute noise, class noise or a combination of the two may all be represented in this manner.

Data is presumed to be generated in the same way for both training and testing. Attribute vectors are sampled with replacement according to an arbitrary distribution D and a class is assigned to each case using C—class 1 with probability C_i and class 0 with probability $1 - C_i$ for a case with attribute vector A_i.

We define a *learning situation* S as a triple (D, C, n) where D and C specify how data will be generated and n is the size of the training set. Given a learner, we may speak of its *accuracy* in a learning situation—expected prediction performance of models produced from training sets of size n when data is generated as specified. Following Wolpert (1992b), however, we concentrate instead on *generalization accuracy*—expected prediction performance on cases with attribute vectors not represented in the training set. In measuring both accuracy and generalization accuracy, test cases are generated using D and C as just described. The only difference is that, for generalization accuracy, we ignore cases with attribute vectors seen in training.

For a two-class problem, the accuracy or generalization accuracy of a random guesser is .5, for every D and C. Using this as a baseline for null performance, we define *generalization performance* as .5 less than generalization accuracy. Generalization performance greater than zero indicates accuracy better than chance on cases with unseen attribute vectors; generalization performance less than zero indicates generalization accuracy worse than chance.

We use the notation $\mathrm{GP}_L(S)$ to denote the generalization performance of a learner L in learning situation S, omitting the subscript when L is implied by context. Applying this notation, we may concisely state the following:

[1] It is worth stressing that each A_i is an attribute *vector*, not a component of such a vector.

Law of Conservation of Generalization Performance: For any learner L,

$$\sum_S \mathrm{GP}(S) = 0$$

for every D and n.

That is, total generalization performance over all learning situations is null; positive performance in some situations must be exactly balanced by negative performance in others.

Two points are worth noting before we present a proof. First, the given equation holds for any arbitrary choice of D and n, but these are fixed when the sum is computed. For fixed D and n, summing over S is equivalent to summing over C. Second, if we exclude the possibility of noise, components of C are drawn from {0,1}, there are 2^m possible class probability vectors and the law involves a sum, as written, of the 2^m corresponding terms. If we allow noise, however, components of C are drawn from the real interval [0,1] and there are an infinite number of possible class probability vectors. In this case, the law is properly written

$$\int_S \mathrm{GP}(S)dS = 0$$

where the integral is over the space $[0,1]^m$ of class probability vectors. We prefer to state the result as a sum to emphasize the analogy with physical laws of conservation, but we will prove the more general version.

For clarity in the proof, we omit references to D, L and n; training sets are understood to be of size n and probabilities and expectations to be calculated with respect to D. We use T to denote a training set. When T is fixed, we use $\overline{C}$ to denote the "complement" of C outside T. That is, $\overline{C}_i = C_i$ for $A_i \in T$ and $1 - C_i$ otherwise. When the learner's predictions are determined by context, we extend the notation GP and use $\mathrm{GP}(A_i, C)$ to denote the expected generalization performance—accuracy minus .5—of the prediction for A_i when the class probability vector is C.

The proof runs as follows. We want to evaluate

$$\begin{aligned}
&\int_S \mathrm{GP}(S)dS \\
&= \int_C \mathrm{E}_T(\mathrm{E}_A(\mathrm{GP}(A, C)))dC \\
&= \int_C \sum_j (\mathrm{P}(T_j \,|\, C) \sum_i \mathrm{P}(A_i)\mathrm{GP}(A_i, C))dC \\
&= \sum_j I_j
\end{aligned}$$

where

$$I_j = \int_C \mathrm{P}(T_j \,|\, C) \sum_i \mathrm{P}(A_i)\mathrm{GP}(A_i, C)dC$$

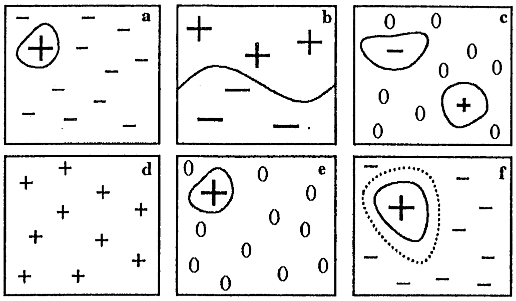

Figure 1: Possible and Impossible Biases.

By a change of variables, we have, also

$$I_j = \int_C \mathrm{P}(T_j \,|\, \overline{C}) \sum_i \mathrm{P}(A_i)\mathrm{GP}(A_i, \overline{C})dC$$

Adding the two formulas for I_j, and since $\mathrm{P}(T_j \,|\, C) = \mathrm{P}(T_j \,|\, \overline{C})$, we have

$$I_j = \frac{1}{2}(\int_C \mathrm{P}(T_j \,|\, C) \sum_i \mathrm{P}(A_i)(\mathrm{GP}(A_i, C) + \mathrm{GP}(A_i, \overline{C}))dC$$

But one of the GPs is $C_i - \frac{1}{2}$ and the other is $(1 - C_i) - \frac{1}{2}$, so each element of the sum is 0. Hence $I_j = 0$ for all j. □

3 Possible and Impossible Learners

The conservation law imposes limitations on the performance of any classification learner. Figure 1 illustrates schematically some examples of what we may and may not expect from such a learner in light of the law. Each box in the figure represents the *regularity space*—$[0,1]^m$ or $\{0,1\}^m$—of possible class probability vectors C and indicates where a learner achieves positive, negative or null generalization performance.

The top row illustrates the performance of three plausible learners. The first, in Figure 1a, achieves strong positive performance in a narrow class of problems at the expense of mildly worse-than-chance performance for all other problems. The second, in Figure 1b, achieves strong performance over a broad class of problems, balanced by performance much worse than chance over the equally broad complement. The last, in Figure 1c, balances moderate positive performance in a narrow class of problems against moderate negative performance in another narrow class, achieving null performance everywhere else.

The diagrams encapsulate the *biases* of the three learners—the compromises they make in trading off positive generalization performance in some learning situations for negative performance in others. Whether a particular bias is appropriate for an intended application will depend on the likely distribution of problems over regularity space. The top row in Figure 1 simply illustrates the wide range of biases consistent with conservation of generalization performance.

The bottom row, by contrast, shows some biases ruled out by the conservation law. Consider, for example, a learner like the one pictured in Figure 1d that achieves at least mildly better-than-chance performance, on average, regardless of the underlying regularity. Unfortunately, a weak, general learner of this kind is like a perpetual motion machine—conservation of generalization performance precludes it. Another attractive goal is the targeted learner of Figure 1e that achieves positive performance for a narrow class of problems—k-DNF, say—and null performance outside this class. Again, the implied imbalance of generalization performance makes this learner impossible. Finally, given a learner that performs well for a narrow class like k-DNF and not too much worse than chance elsewhere, it is natural to consider extending the target class. But this cannot be accomplished, as in Figure 1f, by improving performance in some regions of regularity space without degrading performance in others. The conservation law rules out strict improvements of any kind for all learners, so far as generalization is concerned. This observation is formalized below in Section 5.2.

Since machine learning researchers *have* often set sights on general-purpose learners, targeted learners and improved versions of established learning algorithms, these points may seem unintuitive. Analysis of a simple example helps show how the conservation law applies to a real learner.

Example. Consider the *majority learner* that uniformly predicts the class most prevalent in the training set. This is an elementary example of what might be considered a weak, general learning algorithm. Intuitively, when one class is more prevalent than the other, we expect the majority learner to detect the fact and perform better than chance on cases with unseen attribute vectors. When classes are equally likely, the majority learner achieves accuracy of .5—or zero generalization performance—whichever class it predicts. It would seem then that the majority learner yields mild positive performance for a broad class of regularities and null performance elsewhere. But this would imply positive performance overall, a violation of the conservation law. Where is the fallacy?

The problem is in the second prong of the intuitive argument. Even if classes are equally prevalent, sampling variation is likely to yield a majority in the training set. Assuming equal prevalences, however, a majority of one class in the training set implies an expected majority of the other class among cases with unseen attribute vectors. Hence the majority learner will tend to pick the minority class for purposes of generalization, when classes are equally prevalent, yielding negative generalization performance.

Of course, if the training set covers a small fraction of the attribute space, class prevalences among cases with unseen attribute vectors will be very closely balanced and the degree of negative performance may be miniscule. Nevertheless, substantial positive performance in cases with strong majorities is exactly offset by these tiny negatives. Relationships involving equal or nearly equal prevalences so dominate the regularity space that the necessary balance is maintained. □

4 Two Questions about Bias

The conservation law tells us that positive performance in some learning situations must be balanced by negative performance in others. This makes two questions relevant for any concept learner: (1) When does the learner yield generalization performance better than chance? and (2) When does it yield generalization performance worse than chance?

Concentration on the first question within the field of classification research has tended to promote a one-sided, positive perspective on bias. We have asked how bias helps. The conservation law forces us to recognize that bias is always, and equally, both a positive and negative force.

5 Some Implications of the Conservation Law

5.1 Asymptotic Results

It is customary for machine learning researchers to consider how accuracy improves as the training set size, n, increases. Theoreticians prove probabilistic convergence results using PAC (Valiant, 1984) and related frameworks. Empiricists plot learning curves.

As Wolpert has pointed out, however, focusing on generalization rather than accuracy has the effect of "normalizing" differences in training set size. The conservation law tells us that overall generalization performance remains null for every n. As a consequence, performance will increase with increasing n for some regularities only to the extent that it decreases with n for others. If we use generalization accuracy instead of ordinary accuracy in plotting learning curves, some problems may yield familiar trajectories like the one in Figure 2a, but others must then yield the less familiar pattern of asymptotic degradation shown in Figure 2b. Again, an example may aid intuition.

Example. Even if one class is significantly more prevalent than the other, the majority learner will sometimes fail to detect the fact when n is small, since sampling variation may cause the minority class to dominate training data. As n increases, however, the majority learner identifies the true majority class with ever increasing reliability. When the underlying regu-

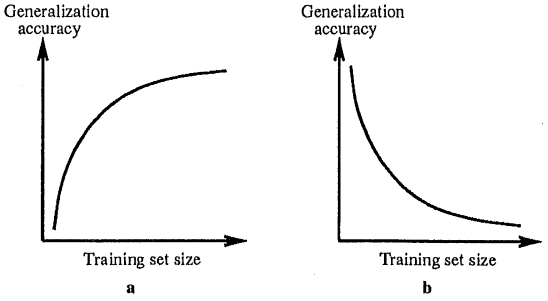

Figure 2: Learning Curves for Generalization Accuracy.

larity entails a strong majority, the generalization accuracy curve will thus follow the pattern of Figure 2a.

When class prevalences are equal, however, generalization accuracy decreases with increasing n, as in Figure 2b. Performance only slightly worse than chance would be expected for small training sets—the majority learner normally chooses the wrong class for generalization, but class prevalences for cases with unseen attribute vectors are nearly equal and the negative effect is small. For larger training sets, the learner continues to choose the wrong class, but the chance of large disparities in class prevalences among cases with unseen attribute vectors increases. If the underlying relationship is deterministic then, in the limit, for training sets that include all but one of the possible attribute vectors, the majority learner is *always* wrong in predictions for cases with the single unseen attribute vector and generalization accuracy reaches its minimum of zero. □

Very Important when Comparing Algorithms –

5.2 Empirical Testing and Comparison

Reports on concept learning algorithms are often supported by empirical results demonstrating either (1) that a single algorithm performs well on a suite of test problems or (2) that some algorithm outperforms competitors on such a suite. The UCI repository (Murphy and Aha, 1992) has become a de facto standard for tests of this kind. In light of the conservation law, two points about empirical tests are worth noting:

1. Every demonstration that the generalization performance of an algorithm is better than chance on a test suite is an implied demonstration that it is worse than chance on an alternative suite.
2. Every demonstration that the generalization performance of one algorithm is better than that of another on a test suite is an implied demonstration that it is worse on an alternative suite.

The first point follows directly from the conservation law—positive performance in one learning situation must be balanced by negative performance in another. The second point is an easy consequence:

> **Corollary (No Strict Improvement):** Since $\sum_S \mathrm{GP}_{L_1}(S) = \sum_S \mathrm{GP}_{L_2}(S) = 0$, we have
>
> $$\begin{aligned} &\sum_S [\mathrm{GP}_{L_1}(S) - \mathrm{GP}_{L_2}(S)] \\ &\quad = \sum_S \mathrm{GP}_{L_1}(S) - \sum_S \mathrm{GP}_{L_2}(S) \\ &\quad = 0 \end{aligned}$$
>
> The equation $\sum_S [\mathrm{GP}_{L_1}(S) - \mathrm{GP}_{L_2}(S)] = 0$ tells us that the superiority of any learner L_1 over another L_2 in some learning situations must be balanced exactly by the superiority of L_2 in others. If the difference $\mathrm{GP}_{L_1}(S) - \mathrm{GP}_{L_2}(S)$ is positive in some learning situations, it must be negative in others to yield an overall total of zero. □

To take an extreme example, consider an empirical comparison of two decision tree induction algorithms, one using information gain in attribute selection and another using information loss. Suppose, as we might expect, that the first algorithm proves superior in generalization on a suite of problems chosen from the UCI repository. Then the information-loss-based algorithm must be superior for some other suite. Empirical testing * cannot prove that information gain is inherently a better criterion than information loss, but only that it is better for some problems and worse for others.

In short, empirical success in generalization is always due to problem selection. Although it is tempting to ** interpret such success as evidence in favor of a learner, it rather shows that the learner has been well applied.

5.3 Enhancing Learners

As just noted, the conservation law rules out strict improvements to any learner. Generalization performance is neither created nor destroyed, but only transferred from one learning situation to another. This fact provides a useful perspective on several approaches researchers have considered for enhancing existing learning algorithms.

5.3.1 Overfitting Avoidance

Induction techniques can often be applied to yield a range of models that increase in complexity as they fit training data with ever greater precision. In several subfields of machine learning—including decision tree, rule set and neural network induction—researchers have first proposed a basic technique that fits training data as far as possible and then, later, considered methods intended to limit complexity appropriately.

These overfitting avoidance methods *may* help to reduce uncertainty about induced models, but the conservation law makes it clear that they can produce no domain-independent increase in expected generalization performance. Given a large decision tree that fits training data well and a pruned version that fits somewhat less well, we may often be more sure of how the performance of the smaller tree will carry over to fresh data. Unless we know something a priori about the true regularity, however, there is no reason to believe the smaller tree will outperform the larger one.

5.3.2 Constructive Induction

Another general approach to enhancing existing learners is to provide means for creating new attributes from old. By increasing the number of prediction models a learner can represent easily, methods for constructive induction (Birnbaum and Collins, 1991, Part III) broaden the range of problems on which it will perform well. According to the conservation law, however, benefits of this kind must entail compensating costs, as far as generalization is concerned. As the region of positive performance is extended, either the degree of positive performance within the region must decrease or the degree of negative performance must increase outside it (see Figure 1f).

5.3.3 Multiple Models and Meta-Level Learners

A third approach to building on existing work is to combine learners or the models they produce. Cross-validation may be used to select one of several learners (Schaffer, 1993b) or it may be used to induce an arbitration scheme of greater complexity for pooling predictions (Wolpert, 1992c). Data diagnostics may be used to select a learner (Brodley and Utgoff, 1993). Multiple models produced by a single learner may also be combined (Kwok and Carter, 1990; Gams, 1989; Buntine, 1990).

No matter how elaborate the multiple model or meta-level learning method it employs, however, any algorithm implementing a fixed mapping from training sets to prediction models is subject to the limitations imposed by the conservation law. In particular, though a multiple model or meta-level learner may be superior in some learning situations to the algorithms on which it builds, this superiority must be balanced exactly by inferiority in other learning situations. Like other enhancement schemes, multiple model and meta-level methods simply shift generalization performance between regions of regularity space.

6 Practical Implications

It is tempting to view the conservation law as theoretically sound, but practically irrelevant. In applications, induction algorithms face only an infinitesimally small subset of all possible regularities. So long as negative generalization performance and the performance degradations that balance improvement are confined to regions of regularity space outside this subset, limitations imposed by the conservation law are of no practical consequence. Since induction algorithms are normally designed with the real world in mind, there is reason to hope that this is the case.

In fact, however, it is not hard to find examples where consequences of the conservation law *are* apparent in practical contexts.

- Accuracy below the default level is extremely common in empirical work. Holte (1993), for example, cites 28 algorithms and variants that perform worse than default on the UCI breast cancer data set. When class prevalences are equal, default accuracy is .5 and performance below the default level constitutes a case of negative generalization performance. When class prevalences are unequal, performance below default indicates that the tested algorithm has improved performance with respect to the majority learner of Section 3 at the expense of degraded performance in other regions of interest.
- Techniques for avoiding overfitting are intended to improve generalization performance, and they do in many cases. These same techniques degrade performance in other cases however, as noted in studies by Fisher (1988) and Schaffer (1993a; 1992a; 1992b). Danyluk and Provost (1993, see Figure 5) report a dramatic example in which accuracy in an important real-world application drops continuously from 85 to below 50 percent as the minimum generality of induced rules is increased—a standard overfitting avoidance measure.
- Aha (1990, p. 69) observed a downward-sloping learning curve (Figure 2b) in applying a nearest neighbor algorithm to heart disease data.
- Several recent research efforts have attempted to extend the range of decision tree induction by allowing linear-discriminant-based tests at decision nodes in place of single-attribute tests (Murthy *et al.*, 1993; Sahami, 1993; Brodley and Utgoff, 1992). Independent tests (Schaffer, 1994b) indicate that the methods are successful, but at the cost of a serious loss of performance in some problems for which ordinary decision tree methods are sufficient.

Cases like these have often been treated as anomalous. The conservation law suggests instead that they ought to be expected, and that, the more we think to look for them, the more often examples will be identified in contexts of practical importance. A rather striking confirmation of this point is the fact that work conducted since the main body of this paper was written uncovered a host of realistic cases in which information loss is substantially superior to information gain as a decision tree splitting criterion (Schaffer, 1994a). As it happens, the "extreme" case of Section 5.2 may arise in quite ordinary circumstances.

7 Memorization and Overall Accuracy

Restricting attention to generalization rather than overall accuracy brings a useful clarity to the issues of bias we have been addressing. Moreover, in many practical applications, attribute vectors from the training set are extremely unlikely to reappear in testing.

Still, it must be recognized that overall accuracy is the usual goal. A few comments about accuracy on cases with attribute vectors represented in the training set are therefore in order. We will refer to this component of overall accuracy as *memorization accuracy*, so that we may speak of concept learning as a combination of memorization and generalization.

Memorization differs sharply from generalization in that some algorithms *are* inherently better for the task than others. When the underlying relationship is deterministic, for example, rote learning is clearly superior to all other approaches. Even when we admit noisy relationships, a categorical statement about the relative value of two algorithms is still sometimes possible. Regardless of the true underlying relationship, a learner that predicts the class most often associated with an attribute vector in the training set when that vector reappears in testing may be expected to perform better than one that predicts the class least often associated with it.

In the non-deterministic case, however, it may be that neither of two learners is strictly superior for memorization. Whether pruned or unpruned decision trees will yield better memorization accuracy, for example, depends on the distribution of problems over regularity space.

In sum, all biases are admissible for generalization—in the sense that none is strictly dominated by another—but only a subset is admissible for memorization. Research characterizing the latter will be an important contribution. But progress along this line has no bearing on the question of generalization. Given a learner L_1 with a bias inadmissible for memorization, we may always patch it by using the model produced by any learner L_2 with an admissible bias for predictions involving cases with familiar attribute vectors. This leaves generalization behavior unaffected and yields an algorithm not dominated by any other in overall accuracy.

8 Related Work

The basic point that methods for induction to unseen cases cannot be justified rigorously dates at least to Hume (1740). See Goodman (1983) for an extended discussion. Mitchell (1980) is normally credited with stressing the consequent necessity of bias in concept learning algorithms. An equivalent point is implicit in the reliance of Bayesian statisticians (Berger, 1985) on information encoded in prior distributions and is treated in detail by Buntine (1990).

The present work draws most heavily on Wolpert (1992b). The conservation law may be viewed as a reinterpretation of Theorem 3 from Wolpert (1992a) or the roughly equivalent statement in Schaffer (1993a, Section 5.1).

Acknowledgements

Thanks to William Cohen, Tom Ellman, Haym Hirsh, Mike Pazzani, Paul Utgoff and David Wolpert for useful input to the work reported here.

References

Aha, David W. 1990. *A study of instance-based learning algorithms for supervised learning tasks: Mathematical, empirical, and psychological evaluations.* Ph.D. Dissertation, University of California, Irvine. Available as UCI Department of Information and Computer Science Technical Report 90-42.

Berger, J. O. 1985. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York.

Birnbaum, Lawrence A. and Collins, Gregg C., editors 1991. *Machine Learning: Proceedings of the Eighth International Workshop.*

Brodley, Carla E. and Utgoff, Paul E. 1992. Multivariate versus univariate decision trees. Technical Report 92-8, Department of Computer Science, University of Massachusetts.

Brodley, Carla E. and Utgoff, Paul E. 1993. Dynamic recursive model class selection for classifier construction. In *Proceedings of the Fourth International Workshop on Artificial Intelligence and Statistics.*

Buntine, Wray 1990. *A Theory of Learning Classification Rules.* Ph.D. Dissertation, University of Technology, Sydney.

Danyluk, Andrea and Provost, Foster 1993. Small disjuncts in action: Learning to diagnose errors in the local loop of the telephone network. In *Proceedings of the Tenth International Conference on Machine Learning.* Morgan Kaufmann. 81–88.

Fisher, Douglas and Schlimmer, Jeffrey 1988. Concept simplification and prediction accuracy. In *Proceedings of the Fifth International Conference on Machine Learning.* 22–28.

Gams, Matjaz 1989. New measurements highlight the importance of redundant knowledge. In *Proceedings of the Fourth European Working Session on Learning.* Pitman Publishing. 71–80.

Goodman, Nelson 1983. *Fact, Fiction, and Forecast.* Harvard University Press, Cambridge, MA, fourth edition.

Holte, Robert C. 1993. Very simple classification rules perform well on most commonly used datasets. *Machine Learning* 11(1):63–90.

Hume, David 1740. *A Treatise of Human Nature.* Oxford University Press, second edition. Modern edition issued in 1978.

Kwok, S. K. and Carter, C. 1990. Multiple decision trees. In Schacter, R. D.; Levitt, T. D.; Kanal, L. N.; and Lemmer, J. F., editors 1990, *Uncertainty in Artificial Intelligence 4.* North-Holland.

Mitchell, Tom M. 1980. The need for bias in learning generalizations. Technical Report CBM-TR-117, Rutgers University.

Murphy, P. M. and Aha, D. W. 1992. UCI repository of machine learning databases [a machine-readable data repository]. Maintained at the Department of Information and Computer Science, University of California, Irvine, CA. Data sets are available by anonymous ftp at ics.uci.edu in the directory pub/machine-learning-databases.

Murthy, S. K.; Kasif, S.; Salzberg, S.; and Beigel, R. 1993. OC1: Randomized induction of oblique decision trees. In *Proceedings of the Eleventh National Conference on Artificial Intelligence.* 322–327.

Sahami, Mehran 1993. Learning non-linearly separable boolean functions with linear threshold unit trees and madaline-style networks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence.* 335–341.

Schaffer, Cullen 1992a. Deconstructing the digit recognition problem. In *Machine Learning: Proceedings of the Ninth International Conference (ML92).* Morgan Kaufmann.

Schaffer, Cullen 1992b. Sparse data and the effect of overfitting avoidance in decision tree induction. In *Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI-92).*

Schaffer, Cullen 1993a. Overfitting avoidance as bias. *Machine Learning* 10(2):153–178.

Schaffer, Cullen 1993b. Selecting a classification method by cross-validation. *Machine Learning* 13(1).

Schaffer, Cullen 1994a. Conservation of generalization: An extended illustration. In preparation.

Schaffer, Cullen 1994b. Linear discriminant tree induction algorithms compared. In review.

Valiant, L. G. 1984. A theory of the learnable. *Communications of the ACM* 27(11):1134–1142.

Wolpert, David H. 1992a. On overfitting avoidance as bias. Technical Report SFI-TR-92-03-5001, Santa Fe Institute.

Wolpert, David H. 1992b. On the connection between in-sample testing and generalization error. *Complex Systems* 6:47–94.

Wolpert, David H. 1992c. Stacked generalization. *Neural Networks* 5:241–259.