Expected Error Analysis for Model Selection

Tobias Scheffer[1] and Thorsten Joachims[2]

July 13, 1999

[1] Otto von Guericke University, FIN/IWS, Universitätsplatz 2,
39106 Magdeburg, Germany, scheffer@iws.cs.uni-magdeburg.de
Phone no +49391 67 11399
[2] Universitaet Dortmund, LS VIII/Computer Science, 44221 Dortmund,
thorsten@ls8.cs.uni-dortmund.de

Abstract

In order to select a good hypothesis language (or *model*) from a collection of possible models, one has to assess the generalization performance of the hypothesis which is returned by a learner that is bound to use some particular model. This paper deals with a new and very efficient way of assessing this generalization performance. We present a new analysis which characterizes the expected generalization error of the hypothesis with least training error in terms of the distribution of error rates of the hypotheses in the model. This distribution can be estimated very efficiently from the data which immediately leads to an efficient model selection algorithm. The analysis predicts learning curves with a very high precision and thus contributes to a better understanding of why and when over-fitting occurs. We present empirical studies (controlled experiments on Boolean decision trees and a large-scale text categorization problem) which show that the model selection algorithm leads to error rates which are often as low as those obtained by 10-fold cross validation (sometimes even superior). However, the algorithm is much more efficient (because the learner does not have to be invoked at all) and thus solves model selection problems with as many as thousand relevant attributes and 12,000 examples.

A short version of this paper appeared at the International Conference on Machine Learning, 1999.

1 Introduction

In the setting of *classification learning* which we study in this paper, the task of a *learner* is to approximate a joint distribution on *instances* and *class labels* as well as possible. A *hypothesis* is a mapping from instances to class labels; the (generalization, or true) *error rate* of a hypothesis h is the chance of drawing a pair of an instance x and a class label y (when drawing according to the sought target distribution) such that the hypothesis conjectures a class label $h(x)$ which is distinct from the "correct" class label y. This error rate (also referred to as the zero-one loss) is the quantity which we wish to minimize. Unfortunately, however, we cannot determine the error rate because the target distribution is not known to the learner. Instead, the learner is able to perceive a *sample* (*i.e.,* a set of pairs (x_i, y_i) of fixed size) which is drawn according to the target distribution and which allows us to define the *empirical error rate* of a hypothesis which is the frequency of misclassifications with respect to the sample. The learner is provided a sample and is constrained to a *model* – a set of potentially available hypotheses – and can minimize the empirical error rate within that model. One can think of the model as a parametric scheme for a hypothesis while an individual hypothesis is a fully parameterized model. A model might, for instance, consist of all decision trees of depth three, or of all back-propagation networks with a certain fixed architecture. For the latter example, the back-propagation algorithm would be a learner that minimizes the empirical error rate within that model. The choice of the model to which we constrain the learner has a very strong impact on the error rate of the hypothesis that the learner will deliver. For example, if we restrict our learner to decision trees of depth one then, for most learning problems, we can expect even the best hypothesis in that model to incur both a high empirical error rate and a high true error rate. On the other hand, from our understanding of PAC- and VC-style error bounds we know that a low empirical error rate does not imply that the true error rate is also low when the model is very large (or complex, respectively). When we consider very many distinct hypotheses, the chance that at least one of them happens to incur a low empirical error rate (although its true error rate is high) grows rapidly. Therefore, in the worst case, the error rate of even an apparently good hypothesis *might* be large. The problem of selecting a model that will lead to a low error rate of the resulting hypothesis is referred to as *model selection* and it is intimately related to the problem of estimating the error rate of a hypothesis. For our back-propagation example, one possible model selection problem would be to determine the number of hidden units that leads to optimal generalization. For decision trees, possible model selection problems would be to determine the subset of the available attributes, or the depth or structure of a tree that imposes an optimal generalization performance.

Three distinct classes of approaches to the model selection problems can be distinguished: (For a more detailed discussion of these approaches, we refer the reader to Section 6.1.) *Hold-out testing*, or *cross validation* algorithms (*e.g.,* Mosier, 1951; Toussaint, 1974; Kohavi & John, 1997) use independent samples that have not been used for training to compare the apparently best hypotheses of each considered model. Cross validation has proven to be a very general and reliable model selection algorithm but requires repeated invocations of the learner for each model which may, for large-scale applications, require a prohibitively large amount of computation. By contrast, complexity penalization algorithms (*e.g.,* Cun *et al.*, 1989; Mingers, 1989; Vapnik, 1998) try to estimate the true error rate based on only the empirical error rate and some complexity measure of the model. Unfortunately, this information does not suffice to actually determine the generalization error rate and therefore complexity penalization algorithms have to conjecture how the error rate might grow with the model complexity. Inevitably, each conjecture will fail for some model selection problems which leads to a bound on the performance of such algorithms (Kearns *et al.*, 1997).

Bayesian approaches (*e.g.,* Bayes, 1763; Berger, 1985; Rissanen, 1978) exploit additional information on the problem (namely the prior probability of each function being the sought target). When this knowledge is available (which is, unfortunately, a rather strong assumption), then sometimes the *Bayes* hypothesis which is guaranteed to minimize the error rate can be determined. This prior information is also exploited by several other, far less expensive heuristics such as MAP (*maximum a posteriori*) or MDL (*minimum description length;* Rissanen, 1989). Again, we refer the reader to Section 6.1 for a continuation of this discussion.

This paper's scope. In Section 3, we will conduct a new analysis of the generalization error rate of a hypothesis which minimizes the empirical error rate in a given model. The most interesting property of our analysis is that it is grounded on a measurable joint property of the (unknown) learning problem and the given model, namely the distribution of error values of the hypotheses in that model. This distribution "counts" how many hypotheses incur which error rates; we can estimate this distribution from the sample. This immediately leads us to an efficient model selection algorithm which exhibits a number of interesting properties. (a) In order to obtain an estimate of the generalization error of the hypothesis that minimizes the empirical error the learner does not actually have to be invoked (*i.e.,* the empirical error minimizing hypothesis does not have to be found). The model selection algorithm is therefore much more efficient than cross validation while often being at least as accurate. (b) The analysis exploits more information on the learning problem than complexity penalization algorithms and therefore the negative results on complexity penalization by Kearns *et al.* (1997) do not apply. (c) Unlike in Bayesian approaches, no explicit knowledge of the prior probabilities of the hypotheses being the sought target is required. It is not assumed that the hypotheses cover all possible target functions either.

In Section 4, we will evaluate the model selection algorithm empirically. We will refer to Boolean functions and to a large-scale text categorization problem. In Section 5, we will see that our analysis helps us to understand learning curves better. It is tempting to interpret certain PAC results as meaning that increasing the model size imposes a higher error rate because the quality of the error estimates decreases. We will prove that this is not the case. In Section 6, we will discuss our approach as well as other approaches to the model selection problem.

2 Preliminaries

In this paper, we focus on classification learning from labeled examples where the target criterion is the expected zero-one loss.

Instances. We assume that there is a set of *instances* X and a finite set of *class labels* Y. A classification problem is defined by an unknown distribution $D_{XY} = D_{Y|X}D_X$ over labeled instances ($X \times Y$), which we want to approximate as closely as possible. D_X is the distribution of instances X and $D_{Y|X}(y|x)$ is the probability of the class label of an instance x being y. Sometimes, when this is more convenient, we speak of target functions. Note that target distributions can "emulate" target functions ($D_{Y|X}(y|x) = 1$ iff $y = f(x)$, 0 otherwise).

Hypotheses and Error. A *hypothesis* $h : X \rightarrow Y$ is a mapping from instances to class labels. The *true (or generalization) error of a hypothesis*, with respect to the (unknown) distribution D_{XY} is the difference between the predicted value $h(x)$ and all class labels y, weighted with $D_{XY}(x, y)$ – more formally, $E_D(h) = \int_{(x,y)\in X\times Y} \ell(h(x), y) dD_{XY}(x, y)$, where ℓ is the zero-one loss function that returns zero if both its arguments are equal and one otherwise. Let the *sample* S be a sequence of m labeled instances drawn *independently and identically distributed* according to D_{XY}. We then define the *empirical or observed error* as the difference between the predicted value $h(x)$ and the class label observed in the sample, for all examples: $E_S(h) = \frac{1}{m}\sum_{(x,y)\in S} \ell(h(x), y)$.

Hypothesis Language and Model. We assume the existence of a given hypothesis language H which may be infinite and may even have an infinite VC-dimension. A *stratification* of the hypothesis language is a finite collection of models $\langle H_1, \ldots, H_m\rangle$, $H_i \subseteq H$. We do not assume the models to be properly nested, but we require the models H_i to be *finite* subsets of H.

ERM-Learner. A *learner* takes as input a sample S and a model H_i and determines the set $H_i^*(S) = \{h \in H_i : E_S(h) = \min_{h'\in H_i}(E_S(h'))\}$ of hypotheses with least empirical error. There is at least one such hypothesis. If there is more than one, we assume that the learner draws one at random under uniform distribution. We call such a learner an ERM learner (error minimizing learner) and the corresponding hypotheses ERM hypotheses. Note that, when the model selection step is treated explicitly on a separate layer (as is assumed in this paper), it is not necessary for the learner to do anything besides minimizing the empirical error rate. For a discussion of this issue, see Section 6.

Notation. We generally write probability distributions and densities in the form $P_{\{x\}}(f(x) = y)$ where the subscript x indicates that x is a random variable. The distribution of x should become clear in the given context. $P_{\{x\}}(f(x) = y)$ refers to the density of $f(x)$ at y which can be thought of as the chance of drawing an x such that $f(x) = y$. Similarly, we write $\mathbf{E}_{\{x\}}(f(x))$ for the expectation of $f(x)$ over all x (again, the distribution of x becomes clear in the context). We write the binomial distribution as $B[p, n](x)$, denoting the chance of making x mistakes on n trials when the chance of a mistake is p.

3 Expected Error Analysis

We will first sketch the idea of our analysis; in Section 3.1 we will then give our first result which characterizes the generalization error rate of h_L in terms of the distribution of error rates in the given model. Since a straightforward evaluation of Theorem 1 would be too expensive, we discuss how this Theorem can be evaluated efficiently in Section 3.2. In Section 3.3, we argue that the distribution of error rates in the model can be estimated from the sample which leads us to an efficient model selection algorithm in Section 3.4.

Suppose that we have to solve a learning problem D_{XY} for which a collection of possible models $\langle H_1, \ldots, H_k\rangle$ is available. A learning algorithm, when invoked with a sample S and constrained to a model H_i can minimize the empirical error rate $E_S(h)$ and will return a hypothesis h_L with least empirical error rate within H_i. The larger H_i is, the smaller is the empirical error of h_L going to be. But we also know that the difference between true and empirical error rate of h_L is going to increase when $|H_i|$ grows. Why is that? The empirical error rate (given the true error rate) is governed by the binomial distribution. The empirical error rate of each individual hypothesis is an *unbiased* estimate of the corresponding true error rate which means that the expectation of the empirical error rate is just the true error rate. However, with a chance of almost $\frac{1}{2}$, the empirical error of each hypothesis is an *optimistic* estimate of its true error, and with a chance of almost $\frac{1}{2}$ it is a *pessimistic* estimate. An optimistically assessed hypothesis has a greater chance of being selected as h_L than a pessimistically assessed one. In return, the empirical error of h_L is, on average, optimistically biased and the bias grows stronger when we consider more hypotheses. Suppose that we have two models with $|H_1| < |H_2|$. Suppose that the least empirical error rate in H_2 is lower than the least empirical error rate in H_1. Which model should we prefer? The empirical error rate of the best hypothesis from H_2 is lower but is also known to be subject to a stronger optimistic bias than the best hypothesis from H_1. In order to decide for a model H_i, we would like to obtain a reliable estimate of the true error rate $E_D(h_L)$. We will now discuss how such an estimate can be obtained without even invoking the learner.

The target D_{XY} defines an error $E_D(h)$ for each hypothesis $h \in H_i$. These error values define a distribution of error values in H_i, which we write as $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ and which is the "prior" in our analysis. Here, "prior" means prior to observing the sample and minimizing the empirical error. $P_{\{h\}}(E_D(h) = e_D|H_i, D_{XY})$ is the chance of drawing a hypothesis h from H_i (when drawing under uniform distribution) which incurs an error of e_D (since $|H_i|$ is assumed to be finite, we actually have a discrete distribution).

Suppose that H_i contains two hypotheses (as an easy example). The prior $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ specifies which error values occur in H_i. There are either two values with a chance of $\frac{1}{2}$ or one value with a chance of 1 (if both hypotheses have equal errors). Let us invent names h_1 and h_2 for the hypotheses and let $E_D(h_1)$ and $E_D(h_2)$ be the two occurring true error rates. But note that assigning hypothesis names to the known error values is only a notational "trick" that will make life easier for us during the main proof; we do not actually know (and our derivation does not exploit) *which* individual hypothesis incurs a particular error value. Now the following happens. On a sample S of size m, each hypothesis incurs an empirical error rate (in the case of two hypotheses $E_S(h_1)$ and $E_S(h_2)$, respectively), governed by the binomial distribution $B[E_D(h_1), m]$ and $B[E_D(h_2), m]$, respectively. (Each example is classified correctly or wrongly, the chance of a wrong answer being $E_D(h_1)$ and $E_D(h_2)$, respectively. This results in a binomial distribution.) Let us now select the hypothesis with the smaller empirical error, call it h_L. In the general case, there might be a set $H_i^*(S)$ of ERM hypotheses and the learner is then assumed to draw a hypothesis h_L at random under uniform distribution from this set. The chance that h_L has a particular error value e_D is now no longer $P_{\{h\}}(E_D(h) = e_D|H_i, D_{XY})$, because h_L is not a randomly drawn hypothesis. It is, instead, the hypothesis which minimizes the empirical error. Assume that the sample size m is fixed and given *a priori*. By contrast, the sample S itself is a random variable, governed by the distribution $(D_{XY})^m$. This implies that $H_i^*(S)$ is a random variable (as it depends on S) and so is h_L; h_L is drawn randomly from $H_i^*(S)$. This leads us to the posterior distribution $P_{\{S,h_L\}}(E_D(h_L) = e_D|H_i, D_{XY}, m, h_L \in H_i^*(S))$ which is the chance of drawing a sample S (of fixed size m) and, consequently, a hypothesis h_L from $H_i^*(S)$, such that the true error of h_L is e_D. The principle difference between the prior and posterior distribution is that the prior gives the distribution of error values of hypotheses which are drawn uniformly from H_i, whereas the posterior gives the distribution of error values for hypotheses which have been generated by an error minimization process. We will see in Section 3.3 that the prior distribution of error rates can be estimated from the sample. This, together with Theorem 2, will immediately lead to an efficient model selection algorithm in Section 3.4.

3.1 Generalization Error Rate of h_L

We quantify the expected true error of h_L, $\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, D_{XY}, m, h_L \in H_i^*(S))$ in Theorem 1. The crucial part of the proof is how to determine the least empirical error e_S and the number of hypotheses $|H_i^*(S)|$ which achieve this error. The idea is that we can determine the chance that $H_i^*(S)$ is a particular subset H^* by factorizing the least error e_S and calculating the chances that each hypothesis in H^* has an empirical error of e_S and each hypothesis outside incurs a strictly greater error. This, however, imposes another difficulty. We know that the empirical error of a hypothesis is distributed binomially, given the true error. But here we have to determine the chance of *several hypotheses* incurring a certain empirical error e_S. The empirical error rates of several hypotheses are not identically distributed because each hypothesis has its individual true error rate. Furthermore, the empirical error rates of two arbitrary hypotheses are not independent because they are measured on the same sample. Unfortunately, in order to determine the probability of no hypothesis incurring an empirical error rate of less than some e_S, we have to assume that the empirical errors of distinct hypotheses to be independent *up to the true error rates.* This independence assumption is often made implicitly; for instance, the calculation of p-values which is required to compare n-fold cross validation results (the p-value gives the chance that one learner does better than another learner for some problem, given the cross validation results) is based on the assumptions that the holdout errors are independent estimates. The empirical error rates do depend on the corresponding true error rates and are not identically distributed. We do not make any assumptions on the true error rates. Dependencies between true error rates may impose dependencies between empirical error rates of distinct hypotheses; such dependencies do *not* violate our assumption.

Assumption 1 (Independence Assumption) *The empirical errors of hypotheses $h \in H_i$ are independent, given the true error rates.* $P(E_S(h_1), \ldots, E_S(h_{|H_i|})|H_i, D_{XY}, m, E_D(h_1), \ldots, E_D(h_{|H_i|})) = \prod_{j=1}^{|H_i|} P(E_S(h_j)|H_i, D_{XY}, m, E_D(h_j))$.

Theorem 1 (Expected Error of an Error Minimizing Hypothesis) *For a distribution D_{XY} of labeled instances and a finite model H_i, let m be the fixed sample size. Let h_L be a hypothesis which is drawn uniformly from the set $H_i^*(S)$ of ERM hypotheses in H_i with respect to a sample S, drawn according to $(D_{XY})^m$. Then, under the independence assumption 1, the expected error of h_L, $\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, D_{XY}, m, h_L \in H_i^*(S))$, is uniquely determined by (and can be computed from) (a) $|H_i|$, (b) m, and (c) the distribution of error rates in H_i, $P_{\{h\}}(E_D(h)|H_i, D_{XY})$.*

The equation which actually determines the error of h_L is given together with the proof in Appendix A. Equation 4 of Appendix A can, in principle, be evaluated given the distribution of true error values $P_{\{h\}}(E_D(h)|H_i, D_{XY})$, $|H_i|$, and the sample size m.

Equation 4 of Theorem 1 (see Appendix A) can be tweaked to quantify $P_{\{S,h_L\}}(E_D(h_L)|H_i, D_{XY}, m, h_L \in H_i^*(S))$, the distribution of error values of the ERM hypothesis, rather than just the expectation of this distribution. This is useful when the target criterion is, for instance, the error ϵ which is not exceeded by h_L with confidence $1-\delta$, rather than the expected error. This error can easily be determined since $P_{\{S,h_L\}}(E_D(h_L) \geq \epsilon| H_i, D_{XY}, m, h_L \in H_i^*(S)) = \int_{e_D=\epsilon}^{1} dP_{\{S,h_L\}}(E_D(h_L) = e_D|H_i, D_{XY}, m, h_L \in H_i^*(S))$.

3.2 Determining the Generalization Error Efficiently

A straightforward implementation of Equation 4 of Theorem 1 would run in time exponential in $|H_i|$ which is clearly prohibitive. Therefore, we make an additional technical assumption.

Assumption 2 *We assume that, for all h_i, h_j,*

$$P\left(|H^*| = n \middle| h_i \in H^*, H_i, m, D_{XY}\right) = P\left(|H^*| = n \middle| h_j \in H^*, H_i, m, D_{XY}\right).$$

Assumption 2 means that the chance of the set of hypotheses with least empirical error being of size m when it is known that a hypothesis h_i belongs to this set is not dependent on *which* hypothesis is known to be in this set. This is always true when H_i is "large". This assumption is reasonable in all practical cases as $|H_i|$ grows doubly exponentially for Boolean functions and at least singly exponentially for languages such as monomials.

Theorem 2 *For a distribution D_{XY} of labeled instances and a finite model H_i, let m be a fixed sample size. Let h_L be drawn under uniform distribution from the ERM hypotheses $H_i^*(S)$ where S is governed by $(D_{XY})^m$. Under assumptions 1 and 2, the expected error of h_L is*

$$\begin{aligned} &\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, m, D_{XY}, h_L \in H_i^*(S)) \\ &= \frac{\int_{e_D} e_D P_{\{h\}}(E_D(h) = e_D|H_i, D_{XY}) dP_{\{S\}}(h_{e_D} \in H_i^*(S)|H_i, m, E_D(h_{e_D}))}{\int_{e_D} P_{\{h\}}(E_D(h) = e_D|H_i, D_{XY}) dP_{\{S\}}(h_{e_D} \in H^*(S)|H_i, m, E_D(h_{e_D}))} \end{aligned} \tag{1}$$

where

$$\begin{aligned} P_{\{S\}}(h_{e_D} \in H_i^*(S)|H_i, m, E_D(h_{e_D})) &= \sum_{e_S} B[e_D, m](e_S) \prod_{e'_D} \left(\sum_{e \geq e_S} B[e'_D, m](e) \right)^{f(e_D, e'_D)} \\ f(e_D, e'_D) &= \begin{cases} |H_i| P_{\{h\}}(E_D(h) = e'_D|H_i, D_{XY}) & \text{iff } e_D \neq e'_D \\ |H_i| P_{\{h\}}(E_D(h) = e'_D|H_i, D_{XY}) - 1 & \text{iff } e_D = e'_D \end{cases} \end{aligned} \tag{2}$$

and h_{e_D} is an arbitrary hypothesis with true error $E_D(h_{e_D}) = e_D$. Furthermore, the expected error rate can be computed in $O(m^2)$.

The proof is given in Appendix B. Note that the only input to Theorem 2 is $|H_i|$, m, and the error prior $P_{\{h\}}(E_D(h)|H_i, D_{XY})$. Theorem 2 solves the primary complexity problem by removing the product over all subsets of H_i from Equation 4. A careful implementation of the formula now runs in $O(m^2)$ (see Appendix B or contact the authors for copies of the implementation).

In order to implement the expected error analysis into a model selection algorithm, we need to find an effective way to estimate $P_{\{h\}}(E_D(h)|H_i, D_{XY})$. Once we have solved this last problem, we can conduct model selection very effectively by enumerating all models, estimating the error prior for each model and, consequently, determine the expected error of the ERM hypothesis of that model.

3.3 Estimating the distribution of error rates

As the distribution of error rates in the model, $P_{\{h\}}(E_D(h)|H_i, D_{XY})$, depends on D_{XY}, it cannot be determined exactly. All information on D_{XY} which we can access is contained in

S. We have to find an efficient way of obtaining an estimate of the distribution of error rates based on the sample.

One possible way of estimating this distribution is to measure its empirical counterpart, the distribution of empirical error rates of hypotheses in the model, $P_{\{h\}}(E_S(h)|H_i, S)$, and use it as an estimate of $P_{\{h\}}(E_D(h)|H_i, D_{XY})$. As the sample size grows, $P_{\{h\}}(E_S(h)|H_i, S)$ converges towards $P_{\{h\}}(E_D(h)|H_i, D_{XY})$; so, for a reasonably large S a good estimate of $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ can be obtained. We will see in the experimental section that small samples impose an optimistic bias that vanishes as the sample size grows. In fact, the experiments reported in the following sections show that even samples of size 50 or more allow for reasonably accurate estimates. Note that this is a one-dimensional distribution only; the dimensionality does not increase when H_i grows.

How many hypotheses do we have to draw under uniform distribution from H_i in order to obtain a good estimate of this distribution? $P_{\{h\}}(E_S(h)|H_i, S)$ is a discrete distribution with m individual probabilities. Therefore, if we draw $\frac{1}{2\varepsilon^2} \log \frac{m}{\delta}$ hypotheses, the chance of mis-estimating $P_{\{h\}}(E_S(h) = e_S|H_i, S)$ for at least one e_S by more than ε is at most δ. This statement, however, is not strong enough because it does not say anything about how strongly a mis-estimation of the error distribution will influence the quality of the estimate of h_L's generalization error. Let us consider the worst possible case that can occur. Suppose that a hypothesis space contains one single hypothesis with error rate zero, and an exponentially fast growing number of hypotheses with an error rate of one. If we fail to "hit" the extremely good hypothesis, then our estimate of $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ will be a single point with mass one. Although this estimate is not far from the true density (which converges towards a single point exponentially fast), it will impose a strong inaccuracy on the estimated error rate of h_L: while the true error rate of h_L is zero with certainty (provided that the learner is exhaustive and finds the isolated good hypothesis in an exponentially fast growing set of bad hypotheses), Theorem 2 will estimate this error as one. Hence, in order to avoid such failures with high confidence, we need to draw an exponentially fast growing number of hypotheses to estimate the distribution of error rates.

Many hypothesis languages, however, have a certain property of symmetry which we can exploit to achieve this in linear time. Suppose that we have a decision tree with n leaf nodes and assume that these leaves are unlabeled yet. By assigning combinations of the class labels zero and one to the leaves we can generate 2^n distinct trees which have equal "stems" and differ in the labelings of their leafs. We can exploit this property of symmetry and construct an algorithm that prints the distribution of the corresponding 2^n empirical error rates in only $O(n)$. For details on the algorithm, see "Algorithm Estimate-$P_{\{h\}}(E_S(h)|H_i, D_{XY})$" in Appendix E.

Unfortunately, the estimator for $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ discussed above is not unbiased. If, for some model, $P_{\{h\}}(E_D(h) = 0|H_i, D_{XY})$ is zero, the chance of some hypothesis incurring an empirical error of zero is still greater than zero which imposes a bias. Fortunately, though, this bias vanishes when the sample size grows. Two unbiased estimators exist, but both have considerable disadvantages (one estimator has a prohibitive variance and the second relies on a distributional assumption which may easily fail). See (Scheffer & Joachims, 1998a) for a detailed discussion.

When $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ is assumed to be a normal distribution, the parameters μ and σ can be determined very easily from an observed $P_{\{h\}}(E_S(h)|H_i, S)$. This assumption holds when the target is a Boolean function over 5 or more attributes and the instances are

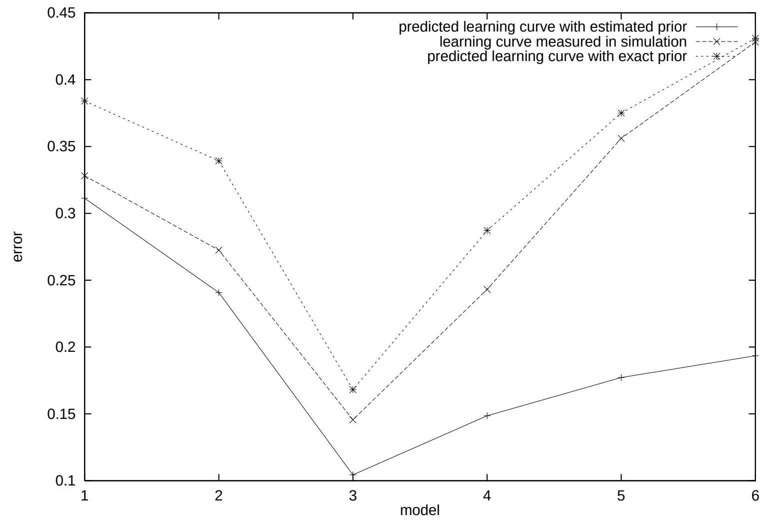

Figure 1: Expected learning curves for Boolean functions, 3 attributes are relevant, sample size 10

governed by the uniform distribution. However, experiments on the artificial and the text categorization problem have shown that this assumption fails frequently.

3.4 Efficient model selection algorithm

Now we can implement Theorem 2 into the following algorithm: (a) For all models H_i, record $P_{\{h\}}(E_S(h)|H_i, S)$ by drawing a small number of hypotheses at random under uniform distribution from H_i and measuring the empirical error. (b) Use $P_{\{h\}}(E_S(h)|H_i, S)$ as an estimate of $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ and use Theorem 2 (as implemented in Appendix B) to determining the expected error of the ERM hypothesis of H_i. (c) After the expected error has been estimated for all models, select the model with the lowest estimated error and (d) invoke the learner with the selected model and the sample S.

4 Experiments

In this section, we evaluate our analysis empirically. Using Boolean functions as targets, we study how the predicted learning curves differ from curves measured in simulations and how expected error analysis based model selection compares to 10-fold cross validation. We also discuss results on a large-scale text categorization problem.

Learning Boolean decision trees. In this set of experiments, we used many randomly drawn Boolean functions as targets. Here, H_i consists of all Boolean functions over i attributes. Unlike some UCI repository repository data sets, randomly drawn Boolean functions have the nice property of not having any special properties that could make an arbitrary heuristic learning technique perform well just by chance. The learner minimizes the empirical error rate within that model. Note that the first No-Free-Lunch Theorem (Wolpert, 1995)

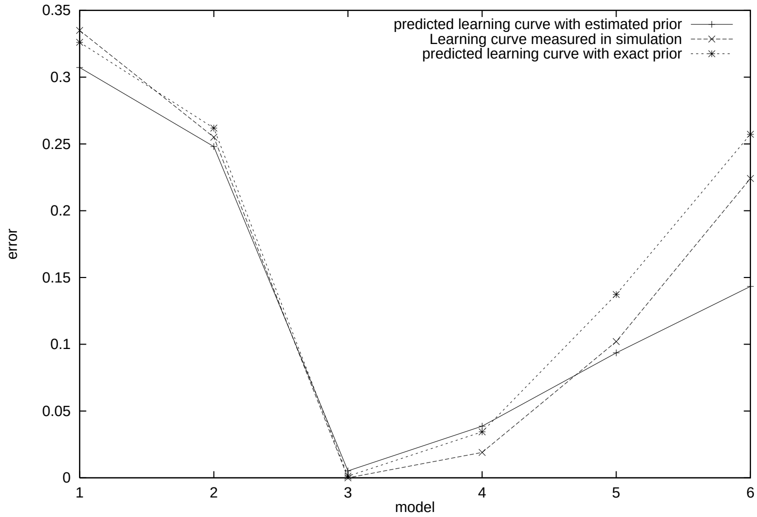

Figure 2: Expected learning curves for Boolean functions, 3 attributes are relevant, sample size 50.

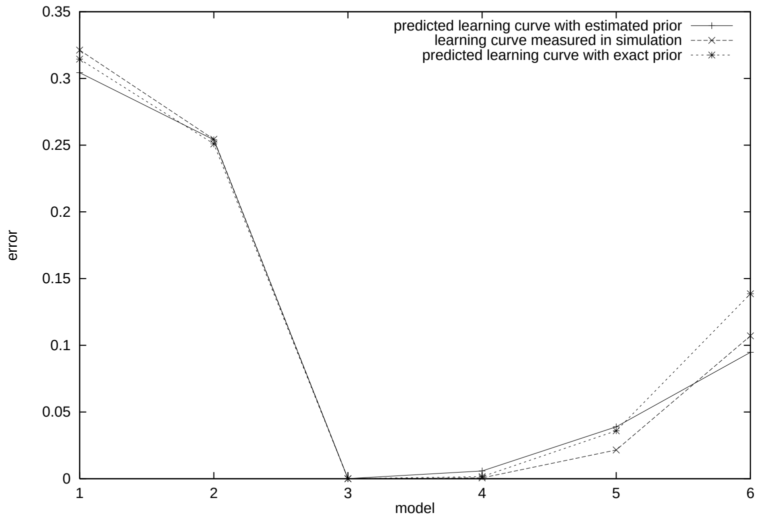

Figure 3: Expected learning curves for Boolean functions, 3 attributes are relevant, sample size 100.

m	10	20	30	40	50	100
EEA	$.29 \pm .001$	$.15 \pm .001$	$.07 \pm .001$	$.05 \pm .001$	$.04 \pm .000$	$.006 \pm .000$
EEA-1000	$.28 \pm .011$	$.16 \pm .001$	$.10 \pm .001$	$.09 \pm .001$	$.06 \pm .001$	$.004 \pm .001$
10-CV	$.27 \pm .001$	$.17 \pm .011$	$.10 \pm .001$	$.06 \pm .001$	$.04 \pm .000$	$.004 \pm .000$

Table 1: True error of the returned hypotheses; target models drawn uniformly

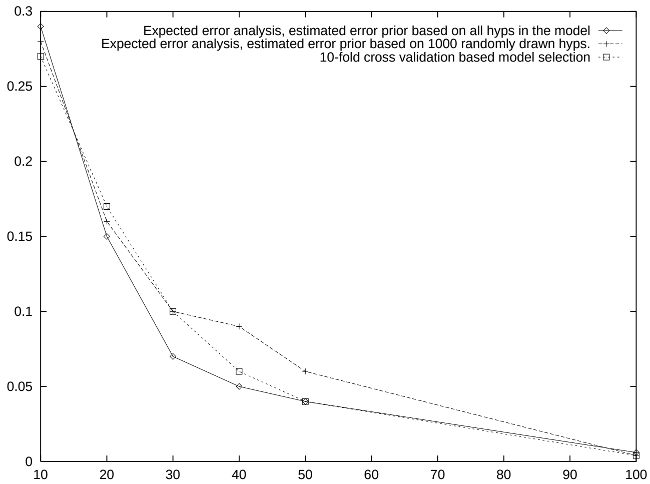

Figure 4: True error of hypothesis returned by model selection based learning. Average over 200 randomly drawn target functions with i attributes (i distributed uniformly). (horizontal axis: sample size, vertical axis: error.)

applies here and claims that all possible learners that minimize the empirical error lead to identical expected learning curves. Figures 3, 4, and 5 compare some predicted learning curves (where the prior has been estimated from the sample) to the average of 200 learning curves measured empirically. As we expected, the quality of the prediction depends crucially on how well the prior distribution of error rates in the model has been estimated. Sample sizes of 10 and 20 lead to unsatisfactory results; when the sample size is 50, the estimated curve approaches the measured curve and for $m = 100$ the predicted curve is a fair estimate of the measured curve. We now want to compare this means of error estimation to 10-fold cross validation. We use the following setting: First, the depth of the target tree is drawn uniformly between 2 and 6 attributes, the target tree is then determined at random with the chosen number of attributes. Next the sample is drawn. We then compare a 10-fold cross validation based model selection algorithm to expected error analysis based model selection. In both cases, models H_1 through H_6 are available where H_i contains all decision trees over i attributes. The cross validation based algorithm uses the averaged hold-out errors to select a model while the expected error analysis based algorithm makes a decision on grounds of the predicted generalization error for each model. After the model is selected, a learner is invoked which uses the whole sample and the true error of the resulting hypothesis is determined. The true error can be determined because the target function is known (but unknown to the learner). Figure 6 shows the results; each point is averaged over 200 distinct target functions. Table 1 presents the same results numerically. The resulting error decreases, of course, with growing sample size. For a sample size of 30, expected error analysis does significantly better than 10-fold cross validation (p-value is .002); for $m = 40$ cross validation does better than expected error analysis when the prior is only estimated by drawing 1000 hypotheses. The error rate achieved by expected error analysis is at least as accurate as by cross validation based model selection. The principle advantage of expected error analysis is that no learning has to be done, the estimate is obtained in an extremely efficient manner.

Scaling up: text categorization. Now we want to demonstrate that the expected error analysis easily scales up to large learning problems with as many as thousand relevant attributes. Text categorization is the problem of mapping texts to semantic categories. Interesting applications of this are classification of newspaper articles and classification of web pages. The documents are represented in terms of a vector of the words occurring in them (*i.e.,* the word orderings are ignored). We used the Reuters corpus of newspaper articles which contains roughly 12,000 documents with 10,000 features and learned the ten most frequent categories. Experiments by Joachims (Joachims, 1998) have shown that decision trees learned on this corpus are highly imbalanced, so we decided to use 2-DL (decision lists with two literals per conjunction) (Rivest, 1987) as hypothesis language. We used a very efficient greedy learner (based on the coverage approach) and used an inverse indexing technique to be able to determine empirical errors quickly. The decision list learner of Rivest (1987) turned out to be too slow for such a large sample, so we modified it slightly. The learner enumerates all monomials with 2 literals; this is done in $O(n^2)$ (where n is as large as 1000). All monomials which cover at least a certain number of instances (we refer to this number as the pruning threshold) and do not exceed a certain empirical error (the error threshold) are appended to the list and the covered instances are removed. If some examples remained after all monomials have been enumerated, the error threshold is incremented and the algorithm recurs. This greedy learner cannot be guaranteed to find the decision list which really minimizes the empirical error – only a locally optimal hypothesis is found.

Assessing a model by means of expected error analysis requires approximately 2 minutes on a PC while running even hold-out testing for one fixed model and one category still takes about 2-3 hours. As we wanted to compare the predicted error rates to cross validation error rates, we had to settle for a small number of models. More details on the experimental setting as well as the tabulated results can be found in Appendix F.

Figure 5 shows the predicted error rates and the error rates estimated by hold-out testing, averaged over the ten most frequent categories. The predicted values are subject to a pessimistic bias of .03 (which we already observed in the previous experiments), but the shapes of the learning curves are fairly similar. Note that a constant bias is undesirable when one wants to know just how accurate a particular hypothesis is; but it is harmless when one wants to know which of several hypotheses is the best one. At first blush, model 20 appears to incur a lower averaged hold-out error while model 30 is, on average, predicted to incur a lower generalization error by expected error analysis. At a closer look, it turns out that, for 8 of 10 categories, expected error analysis and hold-out testing agree on which of the two models incurs a smaller error. Furthermore, a paired t-test shows that, averaged over all categories, model 20 is unlikely to be superior to model 30 (p=.176) although the differences between the hold-out error rates of these models are significant for the individual categories. These results are very promising; expected error analysis provides a good estimate of the generalization error. The estimate is obtained extremely efficiently. Assessing a large number of models for this problem would not be feasible by means of 10-fold (or even 1-fold) cross validation.

5 What is Over-Fitting?

PAC theory is sometimes mis-interpreted as implying that increasing the size or complexity of the model causes the learning curve (*i.e.,* the error rate of the resulting hypothesis) to grow. In this Section, we show that this is not the case.

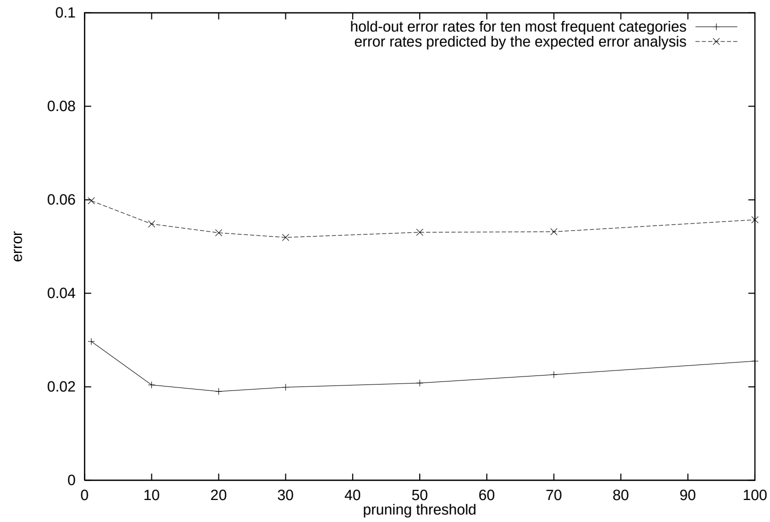

Figure 5: Learning curves (hold-out error rates and expected error analysis results for the text categorization problem)

A learning curve is a function which maps a model index i to the error of the hypothesis which is generated by a given learner on model H_i for a given learning problem. Often, the learning curve grows for large models which is generally referred to as over-fitting. This is frequently considered to be due to the high hypothesis complexity in models with high indexes. However, over-fitting does not necessarily occur. Boosting algorithms have often exhibited a complementary behavior (Schapire *et al.*, 1997), unpruned decision trees have been observed to outperform pruned decision trees (Fisher & Schlimmer, 1988) and Schaffer (1993a) has presented experiments which support his claim that the generalization ability of a learner is a property of the learning problem rather than a property which is intrinsic to the learner. Wolpert (1993) has confirmed this result mathematically.

Using Chernoff bounds, one can guarantee that (with high probability) the difference between true and empirical error of *no* hypothesis in some model H_i exceeds a certain threshold – which immediately leads to worst-case error bounds. This is the way that (agnostic) PAC (Valiant, 1984; Kearns *et al.*, 1992) and VC theory (Vapnik & Chervonenkis, 1971) argue. The empirical error is binomially distributed, so even a poor hypothesis has a small chance (depending on the sample size) of exhibiting a low empirical error. When H_i grows, the chance of *some* hypothesis in H_i possessing a large difference between true and empirical error grows steeply. Therefore, given two hypotheses with equal empirical error which come from distinct models, PAC theory gives better guarantees for the one which comes from the smaller model (*e.g.,* Blumer *et al.*, 1987). But just because *there is* a hypothesis with a large difference between true and empirical error does not mean that the expected error of the returned hypothesis is high. In fact, from the expected error analysis we can derive that when the prior distribution of error values in the model remains constant, the expected error of the returned hypothesis *converges from above* as H_i grows.

To see this, let us look at the expected error of the returned hypothesis h_L when the model

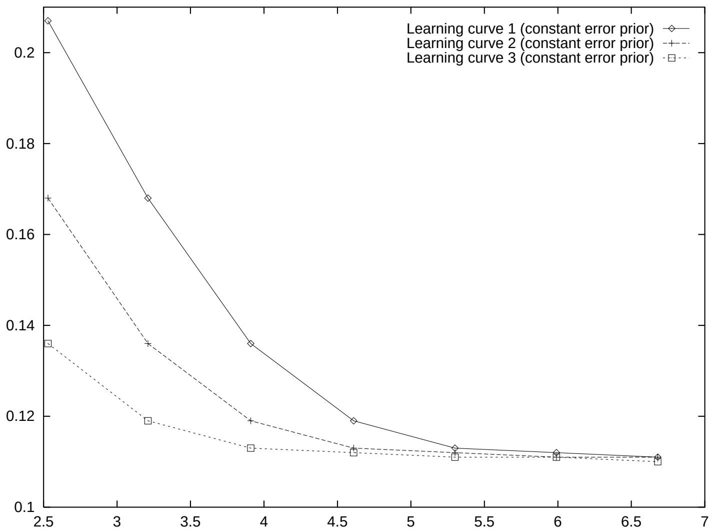

Figure 6: Learning curves; the error prior $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ is constant for all models. The horizontal axis is in units of $\log|H_i|$, the vertical axis shows the error.

size, $|H_i|$, approaches infinity and $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ stays constant.

Theorem 3 *When $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ is constant with $P_{\{h\}}(E_D(h) = 1|H_i, D_{XY}) < 1$, the expected error of an ERM hypothesis h_L converges as $|H_i|$ grows.*

$$\begin{aligned}&\lim_{|H_i|\to\infty} \mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, m, D_{XY}, h_L \in H_i^*(S)) \\ =\ & \frac{\int_{e_D} e_D \times (1-e_D)^m dP_{\{h\}}(E_D(h) = e_D|H_i, D_{XY})}{\int_{e_D} (1-e_D)^m dP_{\{h\}}(E_D(h) = e_D|H_i, D_{XY})} \qquad (3)\end{aligned}$$

The proof is given in Appendix C. Note that Theorem 3 is subject to the assumption that $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ remains fixed while H_i grows. Figure 6 shows three learning curves for distinct learning problems which share the property that the error remains constant while the model size grows (we constructed the models H_i by drawing subsets of the potential hypothesis language H at random).

Let us now study how $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ actually behaves when we want to learn Boolean functions under uniform distribution of the Boolean instances. In this case, the prior distribution of error rates in the model can be determined analytically. When the target function uses attributes x_1 through x_n and the model H_i consists of hypotheses over attributes x_1 through x_i, then the prior is a certain binomial distribution depending on i and n (function and hypothesis must agree on all possible $2^{\max\{i,n\}}$ instances which can be distinguished by target function or hypothesis). By plugging the exact prior into Theorem 1 we can determine the learning curve analytically. Appendix D gives the derivation of the expected learning curve (expected for the uniform distribution over Boolean functions with n attributes).

Figure 7 shows some prior distributions of error values $P_{\{h,D_{XY}\}}(E_D(h)|H_i)$ (D_{XY} is now a random variable as we draw Boolean functions at random) for various models when

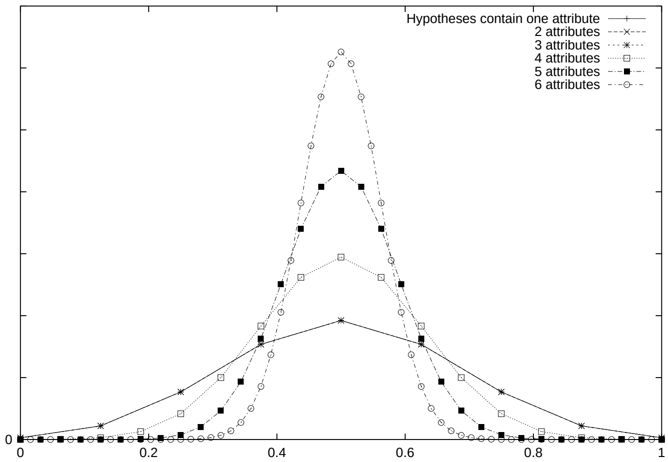

Figure 7: Various shapes of $P_{\{h,D_{XY}\}}(E_D(h)|H_i)$ curves for models which consist of i Boolean attributes when the target function is a randomly drawn function over attributes x_1, x_2, x_3.

three attributes are relevant for the target function. Note that models H_1 through H_3 have identical error priors; the increase in the model size leads to a decrease in error. When irrelevant attributes are added (models H_4 through H_6) the tails of the distribution become skinnier – intuitively, the concentration of really good (and really bad) hypotheses decreases and more hypotheses incur an error of close to $\frac{1}{2}$. This causes the learning curve to rise; Figures 3, 4 and 5 show some predicted learning curves (for various sample sizes) and compare them to learning curves measured in a simulation (the predicted curves are the ones labeled "predicted learning curve, exact prior"). The "simulation" curve shows the error measured in an experiment, averaged over 200 randomly drawn target functions. The deviation originates from three sources: The simulation curve is only measured in an experiment and hence subject to some inaccuracy, the independence assumption 1 on the empirical error causes a modest bias, and the simplification 2 on which the implementation is based incurs a very small error. However, the learning curve is still predicted very accurately. This prediction indicates that the expected error analysis provides a good understanding of the nature of over-fitting.

6 Discussion and Related Work

In Section 6.1, we discuss known approaches to the model selection problem. In Section 6.2, we discuss our results on the expected error analysis and, in Section 6.3, we state some desiderata for forthcoming results. We end this paper with some concluding remarks in Section 6.4.

6.1 Known Approaches to the Model Selection Problem

Virtually any "practical" learning algorithm entails some mechanism for model selection, such as decision tree pruning, neural weight decay, etc. Solutions to this problem can be categorized into *hold-out testing*, *complexity penalization*, and *Bayesian* approaches.

Hold-out testing algorithms split the potential hypothesis language H into subsets ("models") $H_1, H_2, \ldots \subset H$. Starting with the smallest model, the learning algorithm returns one hypothesis from each model. The hold-out set is then used to obtain an estimate of the expected generalization error; the model with the lowest estimate is selected and the learner is invoked for this model with the whole sample. In order to minimize the variance of the estimate (and a small pessimistic bias which is imposed by not using the whole available sample for training) the idea of n-fold cross validation, (*e.g.,* Mosier, 1951; Toussaint, 1974), and bootstrapping (*e.g.,* Efron, 1979) is to average many error measurements which are generated on re-sampled data sets (the instances in the re-sampled data set are drawn without replacement in case of cross-validation and with replacement in the case of bootstrapping). The error which is imposed by a learner that conducts cross-validation based model selection can be bounded by the empirical error of the returned hypothesis, plus a bound on the difference between true and empirical error (which can be chosen according to Wapnik & Tscherwonenkis, 1979; Vapnik, 1982), plus an additional penalty term that accounts for the possibility of cross validation choosing a sub-optimal model (this penalty term depends on the VC dimension of the largest model, and the sample size; Kearns, 1996). When the sample size increases, both the bound on the difference between true and empirical error and the penalty term for choosing a wrong model vanish, so cross validation "always works". For many applications, n-fold cross validation works quite well in practice, but it does not scale well for very large-scale problems as the learner has to be invoked *at least once per model* (*e.g.,* Kohavi & John, 1997).

complexity penalization based methods, by contrast, minimize a demerit criterion which consists of the empirical error and a complexity based penalty term. Empirical error and complexity based penalty have to be weighted according to some factor which has to be adjusted empirically or according to some heuristic. Regularization (Moody, 1992), decision tree pruning (Mingers, 1989), neural weight decay methods (Cun *et al.*, 1989), and Schuurmans' metric based heuristic (Schuurmans, 1997) belong to this class. The Support Vector Machine (Cortes & Vapnik, 1995) which is based on the Structural Risk Minimization framework (Vapnik, 1998), is in this class, too. Complexity penalization based model selection algorithms try to reconstruct the learning curve from only the empirical error rate and some complexity measure of the model. Unfortunately, this is not always possible because the (variance term of the) learning curve can grow in many distinct manners, depending on the learning problem (Schuurmans *et al.*, 1997). The learning curve is *bounded* by PAC- or VC-style results but may, for any given model selection problem, lie somewhere between zero and the worst-case bound. All complexity penalization methods have to conjecture how steeply the learning curve grows with the model complexity and no single conjecture is accurate for all model selection problems. Therefore, given an arbitrary complexity penalization based model selection algorithm, one can always construct a pair of model selection problems, such that the algorithm performs well for one of these problems and fails for the other one – "fails" means that the algorithm incurs an additional error $\lambda > 0$ which does not vanish when the sample size grows (Kearns *et al.*, 1997). Most "practical" learning algorithms incorporate some form of complexity penalization (like tree pruning, weight decay) and account for this problem by providing a parameter that can be adapted for each learning problem – *e.g.,* by cross validation (Ng, 1997; Scheffer & Herbrich, 1997).

Bayesian learners (Bayes, 1763; Berger, 1985) focus on the posterior probability of a function f having generated the sample S: $P(f|S)$. Under certain ideal conditions, one can, under high computational effort, derive the Bayes hypothesis from $P(f|S)$ which is guaranteed to have the least generalization error. The Bayes hypothesis is a weighted majority (where $P(f|S)$ defines the weights) over all hypotheses. Usually, the posterior is used for less expensive heuristics such as MAP (the *Maximum A Posteriori* hypothesis maximizes the chance of having generated the data) or MDL (*Minimum Description Length;* Rissanen, 1985). The more distinct hypotheses a model contains, the greater is the description length required for each hypothesis. On the other hand, a more complex hypothesis is more likely to classify all examples properly while the sample is likely to contain more exceptions to a very simple hypothesis. The MDL heuristic chooses a model such that the returned hypothesis minimizes the sum of description lengths required for the hypothesis and the exceptions to the hypothesis in the data. Often, MDL is considered to be a complexity penalization based model selection technique. This is essentially right, because MDL tries to reconstruct the learning curve from the empirical error rate and a complexity measure of the model. However, since MDL requires the prior $P(f)$ to be known (the prior is required to determine the description length of a hypothesis in an optimal code) but returns a hypothesis which is distinct from the Bayes hypothesis, one might also consider it to be a heuristic within the Bayesian framework. For all Bayes-based model selection algorithms, the prior distribution $P(f)$ is assumed to be known in advance – which is indeed a very strong assumption. The No-Free-Lunch Theorems (Wolpert, 1992) explain that Bayesian learners perform better than randomly guessing only if the actual and the assumed prior are "not completely unaligned". Bayesian learning under uncertain knowledge on the prior $P(f)$ is referred to as *robust Bayesian learning* (for an overview, see Berger, 1993).

6.2 Expected Error Analysis

Related work. Many attempts have been made to find an efficient means of assessing models. The potential benefits of analyses of $P_{\{h_L\}}(E_D(h_L)|H_i, D_{XY}, m, L)$ have been discussed by Wolpert (1995). Recently, Langley and Sage (1999) have presented an analysis of Naive Bayes classifiers which follows the same spirit as the analysis presented here. Langley and Sage find a mathematical model that characterizes the *behavior* of the learning algorithm and is able to predict learning curves with an accuracy that compares to the accuracy achieved by our analysis. Unfortunately, however, the target function has to be known in order to apply their analysis whereas our analysis can be applied after the distribution of error rates has been estimated from the sample.

"Bias-variance decompositions" are analyses of the error of hypothesis h_L (returned by learner L) which split the expected loss into two or three intuitively meaningful terms: the intrinsic target noise ("σ^2") is the expected loss of the Bayes optimal hypothesis; the bias quantifies how well the learner "guesses" the target on average; the variance measures how much the returned hypotheses h_L differ over all possible samples when the learner is invoked repeatedly. Bias-variance decompositions are known for the quadratic loss function (Geman *et al.*, 1992) and the zero-one loss (Kong & Dietterich, 1995; Dietterich & Kong, 1995; Kohavi & Wolpert, 1996). The bias-variance decomposition serves as a tool for the analysis of learning algorithms rather than an efficient method of estimating the error incurred by a learner for a particular problem. Estimating the terms of the decomposition requires repeated runs of the learner and is not easier than n-fold cross validation. By contrast, the analysis presented in this paper leads to a solution which can be evaluated very efficiently.

Domingos (1998) presented a solution which is based on several strong assumptions (*e.g.,* it is assumed that the hypotheses which are contained in one model have equal true errors. Scheffer and Joachims (1998b, 1998a) have found a solution which is based on two independence assumptions, one of which is still rather strong. Recently, Domingos (1999) has proposed a new analysis, which deviates from the first analysis of Scheffer and Joachims (1998b) only in some technical details. In particular, it exploits the same assumption that hypotheses are drawn at random and independently by the learner and thus have independent true error rates. The proximity between the analyses of Scheffer and Joachims (1998a) and Domingos (1999) can best be understood by comparing Equation 10 of Domingos (1999) to Theorem 3 of Scheffer and Joachims (1998a). Note further that Theorem 1 of Scheffer and Joachims (1998a) generalizes Theorem 3 by taking nonzero empirical errors into account. By contrast, in this paper and in its corresponding extended abstract (Scheffer & Joachims, 1999), we presented a very general solution which is only based on the weaker assumption that the empirical errors of distinct hypotheses are independent, given the true error rates.

Learning curves. We presented experiments which showed that, when the prior distribution of error values $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ is known (as is the case when the target is a Boolean function and the instances are uniformly distributed), our analysis predicts and thus explains learning curves with a precision that has not been achieved by PAC theory (*e.g.,* Valiant, 1984), or statistical physics (*e.g.,* Tishby, 1995). One reason for that is that our analysis is not a worst-case analysis; instead, we consider the distribution of error values which is a joint property of the learning problem and the hypothesis language. Our explanation for over-fitting is substantially different from the explanation provided by the PAC framework. The largest difference between true and empirical error of any hypothesis in the hypothesis language grows necessarily when the hypothesis language grows. In order to account for the possibility of worst-case choices of h_L from the empirical error minimizing hypotheses, PAC theory gives weaker guarantees for hypotheses which have been learned from larger models (Blumer *et al.*, 1987). These guarantees hold even if the learner is somehow able to always return the worst possible hypothesis from the set of empirical error minimizing hypotheses. Theorem 3, however, proves that an increase of the number of hypotheses in the model does not cause the expected learning curve to grow when the hypothesis is drawn at random from the set of empirical error minimizing hypotheses. But when the model grows such that the variance of the distribution of error rates in the model decreases (*i.e.,* the tails of the distribution grow skinnier), *this* actually causes the error rate to grow. The variance of the error prior decreases, for instance, when irrelevant attributes are added to the model.

Model selection. Our experiments show that when the prior is estimated from the sample, our error estimate is poor when the sample is too small to estimate the prior distribution well. For sample sizes of 50 or above, expected error analysis based model selection becomes quite accurate. For large sample sizes, the error estimate has turned out to be pessimistically biased (because we assumed that the empirical errors of distinct hypotheses are independent of each other); the pessimistic bias has been almost at a constant level of 0.03. This makes n-fold cross validation appear preferable when the task is to determine the precise accuracy of some hypothesis. For large samples, cross validation is almost unbiased and comes with (approximate) confidence bounds. However, when the task is to determine which of several hypotheses (learned from distinct models) is better, an almost constant bias is not harmful. Therefore, in our experiments on randomly drawn Boolean function, expected error analysis estimation based model selection has turned out to be at least as accurate as 10-fold cross validation based model selection – even for small samples. However, the expected error analysis based algorithm is much more efficient than cross validation because no learning has to be done. For the text categorization problem with 10,000 attributes and 12,000 examples, we obtained quite accurate estimates of the learning curve in an extremely efficient manner while even hold-out testing based assessment of a single model for a single category required hours. Therefore this approach to model selection seems to be particularly interesting for large scale model selection problems (such as text categorization or knowledge discovery in databases) with (tens of) thousands of relevant attributes and many (tens of) thousands of examples.

6.3 Limitations and Future Work

Our analysis is subject to two primary restrictions. First, the model size is required to be finite. Decision trees, decision lists, k-DNF, and similar languages lie within the scope of our approach but, as yet, neural networks lie outside. Second, the loss function is restricted to the expected zero-one loss (or generalization error). We have found a solution for linear cost functions but the corresponding algorithm is too slow for practical purposes. So far, we do not have solutions for further loss functions, such as the quadratic loss. Besides these cases in which the analysis cannot be applied, there are situations in which it should not be applied. Primarily, this is the case when additional background knowledge makes automatic model selection unnecessary. When the prior distribution of target functions or distributions $P(f)$ is known, choosing any other than the Bayes hypothesis would be sub-optimal. Even when the Bayes hypothesis is computationally intractable, heuristics like the MAP or MDL hypothesis exploit this additional knowledge and promise better results, provided that $P(f)$ is actually known rather than being conjectured.

Our analysis is furthermore restricted to "ERM learners" which minimize the empirical error rate and do nothing else. Many practical learners, however, employ regularization mechanisms which might even result in selecting hypotheses with a higher empirical error rate (*e.g.,* decision tree pruning, weight decay, ...). We should note, however, that such regularization mechanisms are means of conducting *model selection.* Since we assume that the model selection step is conducted explicitly on a separate layer, there is no need to incorporate complexity regularization mechanisms in the actual learner. Therefore, we do not see this as a restriction. It would, however, be a most interesting and most ambitious goal to extend this analysis to cover greedy learners. Applying the analysis "as is" to greedy learners (which cannot be guaranteed to minimize the empirical error rate) requires some faith in the learner's ability to find a hypothesis that is reasonably close to having a minimal empirical error rate (note that we did actually use a greedy learner for the text categorization problem). The problem with greedy learners is that the choice of hypotheses which they consider is driven by the data. Hence, the observed empirical error rates are not a reasonable estimate of the distribution of true error rates of the considered hypotheses.

Automatic model selection cannot *per se* be proven to always improve the generalization performance – *i.e.,* there are learning problems for which simply minimizing the empirical error rate rather than constraining the learner to a restricted model is optimal. But several conditions have been identified under which model selection can in fact be shown to be beneficial, see Schaffer (1993a, 1993b), Wolpert (1993), Scheffer (1999).

6.4 Concluding Remarks

We have demonstrated that the expected error rate of a hypothesis that minimizes the empirical error rate within a given model depends on the prior distribution of error rates of hypotheses in that model. This distribution can be estimated from the data and, when it is known, the distribution of error rates (and hence the expected error rate) of an ERM hypothesis can be determined analytically. Our analysis predicts the error rate of a hypothesis that would be the result of a learning algorithm without the learning algorithm actually having to be invoked. We have demonstrated that this analysis leads to accurate error estimates and can therefore be used to construct a very efficient model selection algorithm that can be used to solve large-scale model selection problems. Since the analysis is based on a joint property of the learning problem and the model it allows us to make much stronger claims on the error rate than worst-case bounds. We can derive from the analysis that it is not the increase in the number of hypothesis in the model (or the complexity of the model) that causes the learning curve to grow and impose over-fitting, but rather a shrinking ratio of hypotheses with extremal error rates in the model that can be caused by irrelevant attributes.

A Appendix: Proof of Theorem 1

Theorem 1 (Expected Error of an ERM Hypothesis) *For a distribution D_{XY} of labeled instances and a finite model H_i, let $E_D(h)$ be the true error of each hypothesis $h \in H$. Let m be the fixed sample size. Let h_L be a hypothesis which is drawn uniformly from the set $H_i^*(S)$ of ERM hypotheses in H_i with respect to a sample S, drawn according to $(D_{XY})^m$. Then, under the independence assumption 1, the expected error of h_L is given by*

$$\begin{aligned}
&\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, D_{XY}, m, h_L \in H_i^*(S)) \\
&= \sum_{j=1}^{|H_i|}\Bigg(E_D(h_j)\sum_{e_S}\Bigg(B[E_D(h_j), m](e_S)\sum_{n=1}^{|H_i|}\frac{1}{n}\sum_{\substack{H^*\subseteq H_i\setminus\{h_j\}\\ |H^*|=n-1}} \\
&\qquad \prod_{h^*\in H^*} P_{\{S\}}(E_S(h^*) = e_S|E_D(h^*), m) \prod_{h\in H\setminus\{h_j\}\setminus H^*} P_{\{S\}}(E_S(h) > e_S|E_D(h), m)\Bigg)\Bigg)
\end{aligned} \tag{4}$$

Proof. For the sake of readability, we will not write dependencies on H_i and D_{XY} explicitly. For example, we will write $P_{\{h\}}(E_D(h))$ rather than $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ but keep in mind that virtually all probabilities depend on these fixed quantities.

The expected error $\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, D_{XY}, m, h_L \in H_i^*(S))$ can be expressed as the sum over all errors $E_D(h_j)$ times the chance that h_j is selected by the learner.

$$\begin{aligned}
&\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|m, h_L \in H_i^*(S)) \\
&= \sum_{j=1}^{|H_i|} E_D(h_j) P_{\{S,h_L\}}(h_L = h_j|m, h_L \in H_i^*(S))
\end{aligned} \tag{5}$$

The chance of a hypothesis being selected which is not in $H_i^*(S)$ (the set of ERM hypotheses) is zero (Equation 6). We now factorize the number of ERM hypotheses n (Equation 7).

The chance of h_j being chosen as h_L when h_j is a minimum error hypothesis is $\frac{1}{n}$ (Equation 8). Now we factorize the empirical error e_S of h_j.

$$
\begin{aligned}
& P_{\{S,h_L\}}(h_L = h_j | m, h_L \in H_i^*(S)) \\
&= P_{\{S,h_L\}}(h_L = h_j | h_L \in H_i^*(S), h_j \in H_i^*(S), m) \times P_{\{S\}}(h_j \in H_i^*(S)|m) && (6) \\
&= \sum_{n=1}^{|H_i|} \Bigg(P_{\{S,h_L\}} \left(h_L = h_j \middle| |H_i^*(S)| = n, h_j \in H_i^*(S), h_L \in H_i^*(S), m \right) && (7) \\
&\qquad P_{\{S\}} \left(h_j \in H_i^*(S), |H_i^*(S)| = n \middle| m \right) \Bigg) \\
&= \sum_{n=1}^{|H_i|} \frac{1}{n} P_{\{S\}} \left(h_j \in H_i^*(S), |H_i^*(S)| = n \middle| m \right) && (8) \\
&= \sum_{e_S} \Bigg(\left(\sum_{n=1}^{|H_i|} \left(\frac{1}{n} P_{\{S\}} \left(h_j \in H_i^*(S), |H_i^*(S)| = n \middle| E_S(h_j) = e_S, m \right) \right) \right) \\
&\qquad P_{\{S\}}(E_S(h_j) = e_S | E_D(h_j), m) \Bigg) && (9) \\
&= \sum_{e_S} \left(\sum_{n=1}^{|H_i|} \left(\frac{1}{n} P_{\{S\}} \left(h_j \in H_i^*(S), |H_i^*(S)| = n \middle| E_S(h_j) = e_S, m \right) \right) B[E_D(h_j), m](e_S) \right) && (10)
\end{aligned}
$$

The empirical error (given the true error) is binomially distributed, so $P_{\{S\}}(E_S(h_j) = e_S|E_D(h_j), m)$ in Equation 9 equals $B[E_D(h_j), m](e_S)$ in Equation 10. Now we need to determine the unknown term $P_{\{S\}}\left(h_j \in H_i^*(S), |H_i^*(S)| = n \middle| E_S(h_j) = e_S, m\right)$. In Equation 11, we factorize all possible $H_i^*(S)$ of size n. When the hypotheses in $H_i^*(S)$ incur an empirical error of e_S, all other hypotheses have to incur a strictly higher error (according to the definition of $H_i^*(S)$). In Equation 12, we exploit the independence assumption to resolve the quantifiers.

$$
\begin{aligned}
& P \left(h_j \in H_i^*(S), |H_i^*(S)| = n \middle| E_S(h_j) = e_S, m \right) && (11) \\
&= \sum_{\substack{H^* \subseteq H_i \setminus \{h_j\} \\ |H^*| = n-1}} P \left(\forall h^* \in H^* \colon E_S(h^*) = e_S, \forall h \in H_i \setminus H^* \colon E_S(h) > e_S \middle| E_D(h^*), E_D(h), E_S(h_j) = e_S, m \right) \\
&= \sum_{\substack{H^* \subseteq H_i \setminus \{h_j\} \\ |H^*| = n-1}} \Bigg(\prod_{h^* \in H^*} P(E_S(h^*) = e_S | E_S(h_j) = e_S, E_D(h^*), m) \\
&\qquad \prod_{h \in H \setminus \{h_j\} \setminus H^*} P(E_S(h) > e_S | E_S(h_j) = e_S, E_D(h), m) \Bigg) && (12) \\
&= \sum_{\substack{H^* \subseteq H_i \setminus \{h_j\}, \\ |H^*| = n-1}} \left(\prod_{h^* \in H^*} B[E_D(h^*), m](e_S) \prod_{h \in H_i \setminus \{h_j\} \setminus H^*} \sum_{e > e_S} B[E_D(h^*), m](e) \right) && (13)
\end{aligned}
$$

This completes the proof. ∎

B Proof of Theorem 2

Theorem 2 (Efficient Implementation of Theorem 1) *For a distribution D_{XY} of labeled instances and a finite model H_i, let m be a fixed sample size. Let h_L be drawn under uniform distribution from the ERM hypotheses $H_i^*(S)$ where S is governed by $(D_{XY})^m$. Under assumptions 1 and 2, the expected error of h_L is*

$$\begin{aligned}
&\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, m, D_{XY}, h_L \in H_i^*(S)) \qquad (14)\\
&= \frac{\int_{e_D} e_D P_{\{h\}}(E_D(h) = e_D|H_i, D_{XY}) dP_{\{S\}}(h_{e_D} \in H_i^*(S)|H_i, m, E_D(h_{e_D}))}{\int_{e_D} P_{\{h\}}(E_D(h) = e_D|H_i, D_{XY}) dP_{\{S\}}(h_{e_D} \in H^*(S)|H_i, m, E_D(h_{e_D}))}
\end{aligned}$$

where

$$\begin{aligned}
P_{\{S\}}(h_{e_D} \in H_i^*(S)|H_i, m, E_D(h_{e_D})) &= \sum_{e_S} B[e_D, m](e_S) \prod_{e'_D} \left(\sum_{e \geq e_S} B[e'_D, m](e) \right)^{f(e_D, e'_D)} \qquad (15)\\
f(e_D, e'_D) &= \begin{cases} |H_i| P_{\{h\}}(E_D(h) = e'_D|H_i, D_{XY}) & \text{iff } e_D \neq e'_D \\ |H_i| P_{\{h\}}(E_D(h) = e'_D|H_i, D_{XY}) - 1 & \text{iff } e_D = e'_D \end{cases}
\end{aligned}$$

and h_{e_D} is an arbitrary hypothesis with true error $E_D(h_{e_D}) = e_D$. Furthermore, the expected error rate can be computed in $O(m^2)$.

Proof. Remember that the learner draws a hypothesis from $H_i^*(S)$ under uniform distribution and that assigning error values to individual hypothesis names was a notational trick in the first place. This means, if $E_D(h_i) = E_D(h_j)$ then $P_{\{S\}}(h_L = h_i|H_i, m) = P_{\{S\}}(h_L = h_j|H_i, m)$. Let h_{e_D} be an arbitrary hypothesis with $E_D(h_{e_D}) = e_D$. Then, Equation 5 becomes

$$\begin{aligned}
&\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|m, h_L \in H_i^*(S)) \qquad (16)\\
&= \int_{e_D} e_D dP_{\{S,h_L\}}(E_D(h_L) = e_D|m, h_L \in H_i^*(S))\\
&= \int_{e_D} e_D P_{\{S,h_L\}}(h_L = h_{e_D}|m, h_L \in H_i^*(S)) dP_{\{h\}}(E_D(h) = e_D) \qquad (17)
\end{aligned}$$

In order to understand Equation 17 note that we essentially rearranged Equation 5 such that all hypotheses with equal true error rates are tagged together. Now we have to take care of $P_{\{S,h_L\}}(h_L = h_{e_D}|m, h_L \in H_i^*(S))$. We insert Equation 10 into Equation 17. (We also split up the conjuction "$|H_i^*(S)| = n, h_{e_D} \in H_i^*(S)$".)

$$\begin{aligned}
&\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|m, h_L \in H_i^*(S))\\
&= \int_{e_D} e_D \left(\sum_{e_{min}} B[e_D, m](e_S) \sum_{n=1}^{|H_i|} \left(\frac{1}{n} P(h_{e_D} \in H_i^*(S)|E_S(h_{e_D}) = e_S, m) \right.\right. \qquad (18)\\
&\qquad \left.\left. P_{\{S\}} \left(|H_i^*(S)| = n \middle| h_{e_D} \in H_i^*(S), E_S(h) = e_S, m \right) \right) \right) dP_{\{h\}}(E_D(h) = e_D)
\end{aligned}$$

Exploiting assumption 2 we can claim that

$$const = \sum_{n=1}^{|H_i|} \frac{1}{n} P \left(|H_i^*(S)| = n \middle| h \in H_i^*(S), E_S(h) = e_S, m \right) \qquad (19)$$

is constant for all hypotheses h. $P_{\{S,h_L\}}(E_D(h_L)|m, h_L \in H_i^*(S))$ should integrate to 1. This fixes the scaling factor *const* to

$$const = \left(\int_{e_D} \sum_{e_S} B[e_D, m](e_{min}) \right. \tag{20}$$

$$\left. P(h_{e_D} \in H_i^*(S) | E_S(h_{e_D}) = e_S, m) dP_{\{h\}}(E_D(h) = e_D) \right)^{-1}.$$

We write $\sum_{e_{min}} B[e_D, m](e_{min}) P(h_{e_D} \in H_i^*(S) | E_S(h_{e_D}) = e_{min}, m)$ as $P_{\{S\}}(h_{e_D} \in H_i^*(S) | m, E_D(h_{e_D}))$ and thus arrive at Equation 14. Now we will focus on Equation 15. Equation 21 follows from the straightforward observation that h_{e_D} is in $H_i^*(S)$ iff h_{e_D} incurs an empirical error of e_S and all other h_j incur at least an error of e_S. Note that this equation exploits assumption 1. In Equation 22, we group all hypotheses with equal true error e'_D into one factor and take this factor to the number of hypotheses with that error. Note that one hypothesis (h_{e_D}) has already been assigned an empirical error and is thus not included in the product.

$$P_{\{S\}}(h_{e_D} \in H_i^*(S) | m, E_D(h_{e_D}))$$

$$= \sum_{e_S} B[e_D, m](e_S) \prod_{\substack{j=1 \\ h_j \neq h_{e_D}}}^{|H_i|} \left(\sum_{e \geq e_S} B[E_D(h_j), m](e) \right) \tag{21}$$

$$= \sum_{e_S} B[e_D, m](e_S) \prod_{e'_D} \left(\sum_{e \geq e_S} B[e'_D, m](e) \right)^{|\{h : h \in H \setminus \{h_{e_D}\}, E_D(h) = e'_D\}|} \tag{22}$$

Now Equations 18, 20, and 22 can be rewritten as Equation 15. In order to complete the proof, we still have to show that the expected error can be evaluated in $O(m^2)$. For copies of the implementation, please contact the authors.

Implementation of Theorem 2. *Input*: $P_{\{h\}}(E_D(h)|H_i, D_{XY})$, m, $|H_i|$. *Output:* Expected error rate of h_L.

1. Initialize all variables to zero.

2. **For** $\mathtt{e_D} = 0 \ldots m$, **For** $\mathtt{e_S} = m \ldots 0$, **Let** $\mathtt{P\_E_S\_geq\_e_S\_under\_e_D[eS][e_D]} = B[m, \frac{\mathtt{e_D}}{\mathtt{m}}](\mathtt{e_S}) + \mathtt{P\_E_S\_geq\_e_S\_under\_e_D[e_S + 1][e_D]}$. (make a table of $P_{\{S\}}(E_S(h) \geq e_S | E_D(h) = e_D)$.)

3. **For** $\mathtt{emin} = 0 \ldots m$

   (a) **Let** $\mathtt{prod[emin]} = 0$, (the product $\prod(...)$ in Equations 21 and 22)

   (b) **For** $\mathtt{e_D} = 0 \ldots m$, **Let** $\mathtt{prod[e_D]} = \mathtt{prod[e_D]} \times pow(\mathtt{P\_E_S\_geq\_e_S\_under\_e_D[emin][e_D]}, |H_i| \times \mathtt{P\_E_D[e_D]})$.

4. **For** $\mathtt{e_D} = 0 \ldots m$

   (a) **For** $\mathtt{eS} = 0 \ldots m$, **Increment** $\mathtt{P\_h\_in\_Hstar[e_D]}$ by $B[\frac{\mathtt{e_D}}{\mathtt{m}}, \mathtt{m}](\mathtt{eS}) \times \mathtt{prod[eS]}$ ($P_{\{S\}}(h \in H_i^*(S) | H_i, m, E_D(h))$).

5. **For** $\mathtt{e_D} = 0 \ldots m$

(a) **Increment** `numerator` by $\frac{\mathtt{e_D}}{\mathtt{m}}$ × `P_E`$_\mathtt{D}$`[e`$_\mathtt{S}$`]` × `P_h_in_Hstar[e`$_\mathtt{D}$`]`.

(b) **Increment** `denominator` by `P_E`$_\mathtt{D}$`[e`$_\mathtt{S}$`]` × `P_h_in_Hstar[e`$_\mathtt{D}$`]`.

6. **Return** `Exp_E`$_\mathtt{D}$`[i]` $= \frac{\mathtt{numerator}}{\mathtt{denominator}}$ (the estimated expected true error of model i).

The algorithm generates some tables to avoid double computations. It runs in $O(m^2)$. This completes the proof. ∎

C Proof of Theorem 3

In Equation 23, we write the expected error as the integral over all error values; in Equation 24, we factorize the empirical error of h_L. Note that $P(a) = \sum_b P(a|b)P(b)$. In Equation 25, we apply Bayes Theorem: $P(a|b) = P(b|a)\frac{P(a)}{P(b)}$, where $a = \{E_D(h_L) = e_D\}$, $b = \{E_S(h_L) = e, h_L \in H_i^*(S)\}$, and $c = \{E_S(h) = e\}$. What happens in Equation 26 is $\frac{P(a|b)p(c)}{P(a,b)} = \frac{P(a,b),P(c)}{P(b)P(a,b)} = \frac{P(c)}{P(b)}$, where $a = \{E_S(h_L) = e\}$, $b = \{h_L \in H_i^*(S)\}$, and $c = \{E_D(h_L) = e_D\}$. The idea of Equation 27 is that $P(a,b|c) = P(a|c)P(b|a,c)$. Here, $a = \{E_S(h_L) = e\}$, $b = \{h_L \in H_i^*(S)\}$, and $c = \{E_D(h_L) = e_D\}$. In Equation 28, we state that the empirical error is binomially distributed; we can remove the sum since $P_{\{S,h_L\}}(h_L \in H_i^*(S)|E_S(h_L) = e)$ is zero for $e > 0$ and 1 for $e = 0$ when $|H_i| \to \infty$. In Equation 29 we claim that h_L is an ERM hypothesis if and only if it incurs an empirical error rate of zero. Note that every hypothesis with error strictly less than 1 has a nonzero chance of incurring an empirical error rate of 0. As $|H_i|$ approaches infinity, the chance of at least one hypothesis incurring an empirical error of 0 (under assumption 1) approaches 1.

$$
\begin{aligned}
&\lim_{|H_i|\to\infty} \mathbf{E}_{\{S,h_L\}}(E_D(h_L)|m, h_L \in H_i^*(S)) \\
&= \lim_{|H_i|\to\infty} \int_{e_D} e_D dP_{\{S,h_L\}}(E_D(h_L) = e_D|m, h_L \in H_i^*(S)) && (23)\\
&= \lim_{|H_i|\to\infty} \int_{e_D} e_D \Big(\sum_e dP_{\{S,h_L\}}(E_D(h_L) = e_D|m, h_L \in H_i^*(S), E_S(h_L) = e) \\
&\quad P_{\{S,h_L\}}(E_S(h_L) = e|m, h_L \in H_i^*(S))\Big) && (24)\\
&= \lim_{|H_i|\to\infty} \int_{e_D} e_D \sum_e P_{\{S,h_L\}}(E_S(h_L) = e, h_L \in H_i^*(S)|E_D(h_L) = e_D, m) \\
&\quad \frac{P_{\{S,h_L\}}(E_S(h_L) = e|m, h_L \in H_i^*(S))dP_{\{h_L\}}(E_D(h_L) = e_D)}{P_{\{S,h_L\}}(E_S(h_L) = e, h_L \in H_i^*(S)|m)} && (25)\\
&= \lim_{|H_i|\to\infty} \int_{e_D} e_D \Big(\sum_e P_{\{S,h_L\}}(E_S(h_L) = e, h_L \in H_i^*(S)|E_D(h_L) = e_D, m) \\
&\quad \frac{dP_{\{h_L\}}(E_D(h_L) = e_D)}{P_{\{S,h_L\}}(h_L \in H_i^*(S)|m)}\Big) && (26)\\
&= \lim_{|H_i|\to\infty} \int_{e_D} e_D \Big(\sum_e P_{\{S,h_L\}}(E_S(h_L) = e|m, E_D(h_L) = e_D)
\end{aligned}
$$

$$P_{\{S,h_L\}}(h_L \in H_i^*(S)|E_S(h_L)=e, E_D(h_L)=e_D, m)\frac{dP_{\{h_L\}}(E_D(h_L)=e_D)}{P_{\{S,h_L\}}(h_L \in H_i^*(S)|m)}\Bigg) \quad (27)$$

$$= \int_{e_D} e_D B[e_D, m](0)\frac{dP_{\{h_L\}}(E_D(h_L)=e_D)}{P_{\{S,h_L\}}(h_L \in H_i^*(S)|m)} \quad (28)$$

$$= \frac{\int_{e_D} e_D \times (1-e_D)^m dP_{\{h\}}(E_D(h)=e_D)}{P_{\{S,h\}}(E_S(h)=0|m)} \quad (29)$$

$$= \frac{\int_{e_D} e_D \times (1-e_D)^m dP_{\{h\}}(E_D(h)=e_D)}{\int_{e_D}(1-e_D)^m dP_{\{h\}}(E_D(h)=e_D)} \quad (30)$$

Note that $B[e_D, m](0) = (1-e_D)^m$ which completes the proof. ∎

D Expected Learning Curve for Boolean Functions

In this Section we focus on $\mathbf{E}_{\{D_{XY},S,h_L\}}(E_D(h_L)|H_i, m, h_L \in H_i^*(S))$, the expected error of the ERM hypothesis h_L, where D_{XY} is governed by the uniform distribution over Boolean concepts. In this Section, D_{XY} is not fixed any more; we will therefore write down all dependencies on H_i and D_{XY} explicitly again.

Assume that the learning domain is characterized by a prior $P(D_{XY})$ over target distributions. What is now the expected error of h_L? Theorem 1 claims that, when $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ is known, knowledge of the exact D_{XY} is not necessary. Each D_{XY} yields some $P_{\{h\}}(E_D(h)|H_i, D_{XY})$. Let $P(P_{\{h\}}(E_D(h)|H_i, D_{XY})|P(D_{XY}))$ be the probability of a particular error prior $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ under the given distribution of targets $P(D_{XY})$. In order to determine the expected error over all target functions, we integrate over all possible error priors.

$$\mathbf{E}_{\{D_{XY},S,h_L\}}(E_D(h_L)|H_i, m, h_L \in H_i^*(S))$$
$$= \int_p \mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, p, m, h_L \in H_i^*(S))dP(p|P(D_{XY})) \quad (31)$$

$\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, p, m, h_L \in H_i^*(S))$ is the expected error of h_L when the error prior $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ is p and is given in Theorem 1. For Boolean functions, this Equation has an explicit solution.

Let $P(D_{XY}) = P(D_{Y|X}D_X)$ be such that D_X is the uniform distribution over all Boolean instances and $D_{Y|X}$ is governed by the uniform distribution over all Boolean concepts over the Boolean variables x_1 through x_n. Let H_i be the set of all Boolean functions over Boolean variables $x_1, \ldots, x_i$. Let S be a sample of size m and let h_L be drawn uniformly from $H_i^*(S)$, the hypotheses with least empirical error. Two cases have to be distinguished.

(1) $i < n$. D_{XY} splits the Boolean space into 2^n instances whereas the hypotheses h split the space only into 2^i distinguishable subspaces. Hence, 2^{n-i} Boolean instances with potentially distinct class labels fall into each subspace. However, the hypothesis has to assign one class label to each subspace. Since the target function is uniformly distributed, assigning a class label of 0 will mis-classify a number of instances distributed according to $\epsilon = B[\frac{1}{2}, 2^{n-i}]$ while a class label of 1 will incur an error of $2^{(n-i)} - \epsilon$. Let $\epsilon_1, \ldots, \epsilon_{2^i}$ be the number of instances which are mis-classified when the corresponding subspace (1 through 2^i) is assigned a class label of 0 (the ϵ_j lie between 0 and 2^{n-i}). The ϵ_j are the parameters of the distribution $P_{\{h\}}(E_D(h)|H_i, D_{XY})$. Over all target functions, these parameters are distributed according to

$$P_{\{D_{XY}\}}([\epsilon_1, \ldots, \epsilon_{2^i}]) = B\left[\frac{1}{2}, 2^{n-i}\right](\epsilon_1) \times \ldots \times B\left[\frac{1}{2}, 2^{n-i}\right](\epsilon_{2^i}). \tag{32}$$

For a given set of parameters $\epsilon_1, \ldots, \epsilon_{2^i}$, $P_{\{h\}}(E_D(h)|H_i, D_{XY})$, the sum of errors E_j incurred in subspace $j = 1$ through 2^i, is distributed according to

$$P_{\{h\}}(E_D(h) = e_D|H_i, D_{XY}) = P[\epsilon_1, \ldots, \epsilon_{2^i}]\left(\sum_{j=1}^{2^i} E_j = 2^i e_D\right) \tag{33}$$

where

$$\begin{aligned} &P[\epsilon_1, \ldots, \epsilon_k]\left(\sum_{j=1}^{k} E_j = e\right) \qquad (34) \\ &= \frac{1}{2}P[\epsilon_2, \ldots, \epsilon_{2^i}]\left(\sum_{j=2}^{k} E_j = e - \epsilon_1\right) + \frac{1}{2}P[\epsilon_2, \ldots, \epsilon_{2^i}]\left(\sum_{j=2}^{k} E_j = e - 2^{n-i} + \epsilon_1\right) \end{aligned}$$

and

$$P[\epsilon_1](E_1 = e_1) = \begin{cases} 1 & \text{iff } \epsilon_1 = 2^{n-i} - e_1 = e_1 \\ \frac{1}{2} & \text{iff } \epsilon_1 = e_1 \\ \frac{1}{2} & \text{iff } \epsilon_1 = 2^{n-i} - e_1 \\ 0 & \text{otherwise} \end{cases} \tag{35}$$

Equation 34 is recursive; the intuition of this equation is that an error of e is incurred in subspace 1 through k when *either* an error of ϵ_1 (class label 0) is incurred in subspace 1 and an error of $e - \epsilon_1$ is incurred in subspace 2 through k, *or* an error of $2^{n-i} - \epsilon_1$ is incurred in subspace 1 (class label 1) and the remaining error of $e - (2^{n-i} - \epsilon_1)$ is incurred in subspaces 2 through k.

In this case, $\mathbf{E}_{\{D_{XY}, S, h_L\}}(E_D(h_L)|H_i, m, h_L \in H_i^*(S))$ takes the form

$$\begin{aligned} &\mathbf{E}_{\{D_{XY}, S, h_L\}}(E_D(h_L)|H_i, m, h_L \in H_i^*(S)) \\ &= \sum_{(\epsilon_1, \ldots, \epsilon_{2^i})} \mathbf{E}_{\{S, h_L\}}(E_D(h)|H_i, m, P_{\{h\}}(E_D(h)|H_i, D_{XY}), h_L \in H_i^*(S)) \\ &\qquad P_{D_{XY}}([\epsilon_1, \ldots, \epsilon_{2^i}]) \end{aligned} \tag{36}$$

where $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ depends on $[\epsilon_1, \ldots, \epsilon_{2^i}]$ and is given by Equation 33, $dP_{D_{XY}}$ $([\epsilon_1, \ldots, \epsilon_{2^i}])$ is given by Equation 32, and $\mathbf{E}_{\{S, h_L\}}$ $(E_D(h)|H_i, m,$ $P_{\{h\}}(E_D(h)|H_i, D_{XY})$, $h_L \in H_i^*(S))$ is given by Theorem 1.

$i \geq n$. In this case, the target function assigns one class label to 2^{i-n} instances which can be distinguished by the hypothesis. Given a target function D_{XY},

$$P_{\{h\}}(E_D(h) = e_D|H_i, D_{XY}) = B\left[\frac{1}{2}, 2^i\right](2^i e_D). \tag{37}$$

What is the distribution of the distributions $P_{\{h\}}(E_D(h) = e_D|H_i, D_{XY})$ when $P(D_{XY})$ is the uniform distribution over all Boolean functions with n attributes and D_X is the uniform distribution over all Boolean instances? Here, the key observation is that, in this situation, $P_{\{h\}}(E_D(h) = e_D|H_i, D_{XY})$, as given in Equation 37, does not depend on the concrete D_{XY}. Hence, $P_{\{h,D_{XY}\}}(E_D(h)|H_i) = P_{\{h\}}(E_D(h)|H_i, D_{XY})$. $\mathbf{E}_{\{D_{XY},S,h_L\}}(E_D(h_L)|H_i, m, h_L \in H_i^*(S))$ takes the form

$$\begin{aligned}&\mathbf{E}_{\{D_{XY},S,h_L\}}(E_D(h_L)|H_i, m, h_L \in H_i^*(S)) \\ &= \mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, m, P_{\{h\}}(E_D(h)|H_i, D_{XY}), h_L \in H_i^*(S))\end{aligned} \tag{38}$$

where the error prior $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ is given by Equation 37 and $\mathbf{E}_{\{S,h_L\}}(E_D(h_L)|H_i, m, P_{\{h\}}(E_D(h)|H_i, D_{XY}), h_L \in H_i^*(S))$ is given by Theorem 1. Note that the expected error of h_L only depends on n (the number of relevant attributes), i (the number of attributes of model H_i), and the sample size m.

E Efficient Estimation of the Error Distribution

For some hypothesis languages, 2^n symmetric variants of a hypothesis of description length n exist that can be constructed by assigning different combinations of class labels to the n "terminal nodes" of the hypothesis. This is, for instance, the case for decision trees and decision lists. In this Section, we will show how to exploit this property to write down the empirical error rates of $c \times 2^n$ hypotheses in $O(c \times n)$. This trick enables us to estimate $P_{\{h\}}(E_D(h)|H_i, D_{XY})$ accurately (based on exponentially many hypotheses) and efficiently (in linear time).

Algorithm Estimate-$P_{\{h\}}(E_S(h)|H_i, S)$.

1. **For** $\mathtt{e} = 0 \ldots m$, **Initialize** `P_Es[e]` to 0. (`P_Es[e]` will contain $P_{\{h\}}(E_S(h)|H_i, S)$.)
2. **Repeat** c **times:**
   (a) Draw a "hypothesis stem" (*e.g.,* a decision tree with missing class labels) at random.
   (b) **For all** `e`, **Initialize** `newcnt[e]` and `oldcnt[e]` to 0.
   (c) Let `pos(i)` and `neg(i)` refer to functions that return number of positive and negative examples covered by the ith terminal node of the hypothesis stem.
   (d) **Increment** `oldcnt[neg(1)]` and `oldcnt[pos(1)]` by 1.
   (At this point, `oldcnt[e]` gives the number of hypotheses which consist of only the first terminal node and incur an error of `e`.)
   (e) **For** `i = 2...len(stem)`
      i. **For** $\mathtt{e} = 1 \ldots m$
         A. **Increment** `newcnt[e + pos(i)]` by `oldcnt[e]`.
         (An error rate of `e + pos(i)` is incurred by the first i nodes when the ith node carries class label 0 and $i - 1$ nodes incur an error rate of `e`.)
         B. **Increment** `newcnt[e + neg(i)]` by `oldcnt[e]`.

(Now, `newcnt[e]` gives the number of hypotheses consisting of the first i terminal nodes with all possible assignments of class labels to these nodes which incur an empirical error of `e`.)

ii. **For** all e, **set** `oldcount[e]` to `newcount[e]`.

(f) **For** $\texttt{e} = 0 \ldots m$, **Increment** $\texttt{P\_E}_\texttt{S}[\texttt{e}]$ by $\texttt{newcount[e]}/(2^{\texttt{len(stem)}} \times c)$.

3. **Return** $\texttt{P\_E}_\texttt{S}$.

F Empirical Data

This section provides further details on the experimental settings an on the obtained results for the text categorization problem.

In a first set of experiments, we wanted to determine how many of the attributes are relevant. We sorted the attributes according to their information gain (in accordance with the results of Yang & Petersen, 1997). We then compared 4 models, H_1 through H_4, containing 250, 500, 750, and 1000 attributes, respectively. We measured the hold-out error of the hypotheses returned by our learner on 3400 examples and determined the error rates predicted by Theorem 1. Surprisingly, both learning curves (the curves of the hold-out errors and the predicted errors averaged over the ten most frequent categories) are almost flat except for some noise. The hold-out error is about 0.02 while the predicted error is about 0.05 (the pessimistic bias of about 0.03 has already been observed earlier). When 500 or more attributes are used, the generalization error is significantly lower than when only 250 attributes are used ($p < 0.05$). The predicted error of the model with 500 attributes is lower than the predicted error for 250 models, too. However, there are no significant differences between the error rates for 500, 750, and 1000 attributes ($p > 0.3$). Expected error analysis prefers 1000 attributes but this is neither significantly better nor worse than 500, or 750 attributes. Unfortunately, we could not increase the number of attributes beyond 1000 because in this case obtaining the hold-out error estimates becomes too expensive. Even though we used a very efficient greedy learner, the complexity grows still quadratically in the number of attributes.

In the next set of experiments, we used a different collection of models. We used 7 models where H_i contains decision lists with monomials that cover at least a certain number of examples. The parameter which is adapted here influences the model by only allowing monomials which are supported by a certain sample size. One could think of this number as a pruning threshold which is adapted. Table 2 shows the hold-out error incurred for the ten most frequent concepts, and Table 3 shows the error rates predicted by expected error analysis. Since the hold-out set is quite large (3400 examples), most differences are significant.

ACKNOWLEDGMENT

Thanks to Claus Weihs and Stefan Wrobel for carefully proof-reading the manuscript. We would like to thank Uschi Sondhauss and Fritz Wysotzki for discussions and helpful comments. This work was partially supported by grant WY 20/1-2 of the German Research Council (DFG), the DFG collaborative research center SFB 475, and an Ernst von Siemens fellowship held by Tobias Scheffer.

category	1	10	20	30	50	70	100
acq	.0907258	.0601959	.055011	.0596198	.0561636	.0648041	.0829493
corn	.0034562	.0014400	.0011520	.0011520	.0011520	.0011520	.0011520
earn	.0578917	.0417627	.0345622	.0406106	.0455069	.0486751	.0538594
crude	.0290899	.0129608	.015265	.014400	.0144009	.0172811	.0221774
grain	.0086405	.0089285	.0095046	.0086405	.0080645	.0103687	.0106567
interest	.0241935	.0253456	.0187212	.0207373	.0213134	.0207373	.0207373
money-fx	.0288018	.0218894	.0207373	.0236175	.0250576	.0256336	.0250576
ship	.015841	.0097926	.0086405	.0080645	.0106567	.0092165	.0112327
trade	.0328341	.0213134	.0244816	.0207373	.0239055	.0256336	.0250576
wheat	.0057603	.0011520	.0023041	.0014400	.0023041	.0025921	.0025921
average	.0297	.0204	.01904	.0199	.0208	.0226	.0255

Table 2: Hold-out error rates depending on the pruning threshold for 500 attributes

category	1	10	20	30	50	70	100
acq	.169624	.164426	.136109	.14765	.146808	.157075	.15685
corn	.0254748	.0199941	.0199937	.0198764	.0194109	.0293709	.0285292
earn	.0800091	.0701208	.080318	.0694885	.0787617	.0781107	.0768116
crude	.0500038	.0354245	.0453292	.0426815	.0410724	.0405766	.0400437
grain	.050358	.0486773	.0462126	.0431106	.0408503	.0405403	.050588
interest	.0504822	.0494526	.0436008	.0447569	.0558017	.0403235	.0524012
money-fx	.0678693	.0599967	.0598764	.0694562	.0682563	.0670629	.0659219
ship	.0299725	.0303424	.030023	.0199194	.0195121	.0194726	.0285667
trade	.0541985	.0499002	.0482594	.0428654	.0411051	.0403724	.0401568
wheat	.0200082	.0199976	.0198373	.0199312	.0188736	.0187006	.017471
average	.059797	.054837	.05295	.05196	.053044	.05316	.05573

Table 3: Predicted error rates depending on the pruning threshold for 500 attributes

References

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. of the Royal Soc. of London*, *53*, 370–418.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag.

Berger, J. (1993). An overview of robust Bayesian analysis. Tech. rep. 93-53C, Department of Statistics, Purdue University.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. (1987). Occam's razor. *Information processing Letters*, *24*, 377–380.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

Cun, Y. L., Denker, J., & Solla, S. (1989). Optimal brain damage. In *NIPS-89*, pp. 598–605.

Dietterich, T., & Kong, E. (1995). Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Tech. rep., Department of Computer Science, Oregon State University.

Domingos, P. (1998). A process-oriented heuristic for model selection. In *ICML-98*, pp. 127–135.

Domingos, P. (1999). Process-oriented estimation of generalization error. In *IJCAI-99*.

Efron, B. (1979). Bootstrap methods: another look at the jacknife. *Annals of Statistics*, *7*(1), 1–26.

Fisher, D., & Schlimmer, J. (1988). Concept simplification and prediction accuracy. In *ICML-88*, pp. 22–28.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1–48.

Joachims, T. (1998). Text categorization with support vector machines. In *Proceedings of the European Conference on Machine Learning*.

Kearns, M., Mansour, Y., Ng, A., & Ron, D. (1997). An experimental and theoretical comparison of model selection methods. *Machine Learning Journal*, *27*, 7–50.

Kearns, M., Schapire, R., & Sellie, L. (1992). Towards efficient agnostic learning. In *Int. Conference on Computational Learning Theory*, pp. 341–352.

Kearns, M. (1996). A bound on the error of cross validation using the approximation and estimation rates, with consequences for the training-test split. In *Advances in Neural Information Processing Systems*, Vol. 8, pp. 183–189.

Kohavi, R., & John, G. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 245–271.

Kohavi, R., & Wolpert, D. (1996). Bias plus variance decomposition for zero-one loss functions. In *ICML-96*.

Kong, E., & Dietterich, T. (1995). Error-correcting output coding corrects bias and variance. In *ICML-95*, pp. 313–321.

Langley, P., & Sage, S. (1999). Tractable average case analysis of naive bayes classifiers. In *ICML-99*, pp. 220–228.

Mingers, J. (1989). An empirical comparison of selection measures for decision-tree induction. *Machine Learning*, *3*, 319–342.

Moody, J. (1992). The effective number of parameters: An analysis of generalization and regularization in non-linear learning systems. In *NIPS-4*.

Mosier, C. (1951). Problems and designs of cross validation. *Educational and Psychological Measurement*, *11*, 5–11.

Ng, A. (1997). Preventing overfitting of cross validation data. In *Proc. Int. Conference on Machine Learning*, pp. 245–253. Morgan Kaufmann.

Rissanen, J. (1978). Modelling by shortest data descriptions. *Automatica*, *14*, 465–471.

Rissanen, J. (1985). Minimum-description-length principle. *Ann. Statist.*, *6*, 461–464.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquery.* World Scientific.

Rivest, R. L. (1987). Learning decision lists. *Machine Learning*, *2*(2), 229–246.

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

Schaffer, C. (1993b). Selecting a classification method by cross-validation. *Machine Learning*, *13*, 135–143.

Schapire, R., Freund, Y., Bartlett, P., & Lee, W. (1997). Boosting the margin: A new explanation for the effectiveness of voting methods. In *Machine Learning: Proceedings of the Fourteenth International Conference*, pp. 322–330.

Scheffer, T. (1999). *Error Estimation and Model Selection.* Ph.D. thesis, Technische Universitaet Berlin, School of Computer Science.

Scheffer, T., & Herbrich, R. (1997). Unbiased assessment of learning algorithms. In *IJCAI-97*, pp. 798–803.

Scheffer, T., & Joachims, T. (1998a). Estimating the expected error of empirical minimizers for model selection. Tech. rep. TR 98-9, Technische Universitaet Berlin.

Scheffer, T., & Joachims, T. (1998b). Estimating the expected error of empirical minimizers for model selection (abstract). AAAI-98.

Scheffer, T., & Joachims, T. (1999). Expected error analysis for model selection. In *Proceedings of the International Conference on Machine Learning (ICML-99).*

Schuurmans, D. (1997). A new metric-based approach to model selection. In *AAAI-97.*

Schuurmans, D., Ungar, L., & Foster, D. (1997). Characterizing the generalization performance of model selection strategies. In *ICML-97*, pp. 340–348.

Tishby, N. (1995). Statistical physics models of supervised learning. In Wolpert, D. (Ed.), *The Mathematics of Generalization.* Addison-Wesley.

Toussaint, G. (1974). Bibliography on estimation of misclassification. *IEEE Transactions on Information Theory IT20*, *4*, 472–479.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, *27.*

Vapnik, V. (1982). *Estimation of Dependencies Based on Empirical Data.* Springer.

Vapnik, V. (1998). *Statistical Learning Theory.* Wiley.

Vapnik, V. N., & Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, *16*(2), 264–280.

Wapnik, W., & Tscherwonenkis, A. (1979). *Theorie der Zeichenerkennung.* Akademie Verlag, Berlin.

Wolpert, D. (1993). On overfitting avoidance as bias. Working paper 93-03-016, The Santa Fe Institute.

Wolpert, D. H. (1995). The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework. In Wolpert, D. H. (Ed.), *The Mathematics of Generalization*, The SFI Studies in the Sciences of Complexity, pp. 117–214. Addison-Wesley.

Wolpert, D. (1992). On the connection between in-sample testing and generalization error. *Complex Systems*, *6*(1), 47.

Yang, Y., & Petersen, J. (1997). A comparative study on feature selection in text categorization. In *ICML-97.*