A New Error Bound for the Classifier Chosen by Early Stopping

Eric Bax*, Zehra Cataltepe, and Joe Sill
California Institute of Technology

Key words machine learning, learning theory, validation, early stopping, Vapnik-Chervonenkis.

Introduction

Training with early stopping is the following process. Partition the in-sample data into training and validation sets. Begin with a random classifier g_1. Use an iterative method to decrease the error rate on the training data. Record the classifier at each iteration, producing a series of snapshots $g_1, \ldots, g_M$. Evaluate the error rate of each snapshot over the validation data. Deliver a minimum validation error classifier, g^*, as the result of training.

The purpose of this paper is to develop a good probabilistic upper bound on the error rate of g^* over out-of-sample (test) data. First, we use a validation-oriented version of VC analysis [6, 7] to develop a bound. Because of the nature of VC analysis, this initial bound is based on worst-case assumptions about the rates of agreement among snapshots. In practice, though, successive snapshots are similar classifiers. We exploit this feature to develop a new bound. Then we test the bound on credit card data.

VC-Style Bound

Framework

Our machine learning framework has the following structure. There is an unknown boolean-valued target function and an unknown distribution over its input space. For example, the input distribution could be typical data about credit card applicants, and the target function could be 1 if the applicant defaults within 5 years of being issued a credit card and 0 otherwise.

We have a sequence of snapshot classifiers $g_1, \ldots, g_M$. We have d validation examples which were not used to train the classifiers. We also have d' test inputs (but not the corresponding outputs). The validation and test inputs were drawn independently at random according to the underlying input distribution. The validation outputs were determined by the target function. We desire a bound on the error rate over the test inputs of a classifier $g^* \in \{g_1, \ldots, g_M\}$ that has minimum error rate over the validation data. (The error rate of a classifier over a data set is the rate of disagreement over the inputs between the classifier and the target function.)

Single-Classifier Bound

The first step to develop a VC-style bound for the test error of g^* is to develop a bound for an arbitrary snapshot g_m chosen without reference to validation error. Let ν_m be the validation error of g_m, and let ν'_m be the test error. Let $n = d + d'$, the number of inputs in the validation and test data combined. The probabilities in our error bounds are over partitions of the n inputs into d validation examples and d' test examples. Since the inputs are drawn i.i.d., each partition is equally likely.

Let w be the number of the n inputs for which classifier g_m produces the incorrect output. The probability that the validation error is $\frac{k}{d}$ is

$$\binom{n}{d}^{-1} \binom{w}{k} \binom{n-w}{d-k} \tag{1}$$

If the validation error is $\frac{k}{d}$, then the test error is $\frac{w-k}{d'}$. So

$$\Pr\{\nu'_m \geq \nu_m + \epsilon | w\} = \sum_{\{k | \frac{w-k}{d'} \geq \frac{k}{d} + \epsilon\}} \binom{n}{d}^{-1} \binom{w}{k} \binom{n-w}{d-k} \tag{2}$$

Bound by maximizing over w.

$$\Pr\{\nu'_m \geq \nu_m + \epsilon\} \leq \max_{w \in \{0,\ldots,n\}} \Pr\{\nu'_m \geq \nu_m + \epsilon | w\} \tag{3}$$

We refer to the bound as $B(\epsilon)$.

Initial Test Error Bound for g^*

The single-classifier bound

$$\Pr\{\nu'_m \geq \nu_m + \epsilon\} \leq B(\epsilon) \tag{4}$$

is based on probabilities over random partitions of the n inputs into validation and test sets. Classifier g^* is chosen according to validation error. To compute validation error, we implicitly use information about which inputs are in the validation set. So g^* is chosen by reference to the partition at hand, and hence the single-classifier bound is not valid for g^*.

However, the snapshot sequence $g_1, \ldots, g_M$ is chosen without reference to the partition since training references neither validation nor test data. We develop a uniform bound over the $g_1, \ldots, g_M$. The uniform bound includes a bound on g^* since $g^* \in \{g_1, \ldots, g_M\}$.

To obtain a uniform bound, consider the probability of failure for at least one single-classifier bound.

$$\Pr\{\nu'_1 \geq \nu_1 + \epsilon \text{ or } \ldots \text{ or } \nu'_M \geq \nu_M + \epsilon\} \tag{5}$$

*CalTech 256-80, Pasadena, CA 91125 (eric@cs.caltech.edu).

0-7803-3905-3/97/$10.00©1997 IEEE

Bound the probability of the union event by the sum of event probabilities.

$$\leq \Pr\{\nu'_1 \geq \nu_1 + \epsilon\} + \ldots + \Pr\{\nu'_M \geq \nu_M + \epsilon\} \tag{6}$$

Use the single-classifier bound for each probability.

$$\leq MB(\epsilon) \tag{7}$$

Subtract $MB(\epsilon)$ from one to bound the probability of the complement of (5).

$$\Pr\{\nu'_1 < \nu_1 + \epsilon \text{ and } \ldots \text{ and } \nu'_M < \nu_M + \epsilon\} \geq 1 - MB(\epsilon) \tag{8}$$

This uniform bound applies to g^* since it is a snapshot.

$$\Pr\{\nu'_* < \nu_* + \epsilon\} \geq 1 - MB(\epsilon) \tag{9}$$

where ν'_* and ν_* are the test and validation error rates of g^*.

Central Classifier Bound

Choose a set of "central" classifiers $c_1, \ldots, c_S$ without reference to the partition of inputs into validation and test sets. For example, select central classifiers by sampling the snapshots at intervals of 100: $c_1 = g_{100}, \ldots, c_{10} = g_{1000}$.

Let c^* be a central classifier which may be chosen with reference to the partition. Let ν'_+ and ν_+ be the test and validation error rates of c^*. Since the central classifiers are chosen without reference to the partition, we can use a uniform bound over them as a bound for c^* in the same manner as we used a uniform bound over the snapshots as a bound for g^* in (9).

$$\Pr\{\nu'_+ < \nu_+ + \epsilon\} \geq 1 - SB(\epsilon) \tag{10}$$

As before, let ν'_* and ν_* be the test and validation error rates of g^*. Add $\nu'_* - \nu'_+$ to both sides of the inequality in the event.

$$\Pr\{\nu'_+ + (\nu'_* - \nu'_+) < \nu_+ + (\nu'_* - \nu'_+) + \epsilon\} \geq 1 - SB(\epsilon) \tag{11}$$

This implies

$$\Pr\{\nu'_* < \nu_+ + (\nu'_* - \nu'_+) + \epsilon\} \geq 1 - SB(\epsilon) \tag{12}$$

Note that the difference in error rates between any two classifiers can be no greater than the rate of disagreement. Let δ be the rate of disagreement between g^* and c^* over the test inputs. Since $\delta \geq \nu'_* - \nu'_+$,

$$\Pr\{\nu'_* < \nu_+ + \delta + \epsilon\} \geq 1 - SB(\epsilon) \tag{13}$$

Let $\beta = \nu_* - \nu_+$. Rewrite ν_+ as $\nu_* - \beta$.

$$\Pr\{\nu'_* < \nu_* + \beta + \delta + \epsilon\} \geq 1 - SB(\epsilon) \tag{14}$$

This is the central classifier bound, in which the test error of g^* is bounded by reference to a central classifier c^*. Note that the bound is valid for c^* chosen according to the partition. So it is valid to use the central classifier that minimizes $\beta + \delta$ as c^* in the bound (14). However, the set of central classifiers $c_1, \ldots, c_S$ must be chosen without reference to the partition. Hence, the set cannot be chosen to minimize $\beta + \delta$ directly.

Selecting Central Classifiers

We may use the validation and test inputs to select the set of central classifiers as long as we do not differentiate between validation and test inputs. In this way, we choose the same set of central classifiers regardless of the partition. Since the probabilities of bound (14) are over partitions, the bound is valid.

Let r_{ms} be the number of validation and test inputs for which g_m and c_s disagree. Note that the difference in validation error rates β is no greater than the rate of disagreement over validation inputs. So $\beta + \delta$ is no greater than the sum over validation and test examples of disagreement rates between g^* and c^*. The sum of rates is maximized when the disagreements are concentrated in the smaller data set. Note that g^* could be any g_m, and we choose c^* to minimize $\beta + \delta$.

$$\beta + \delta \leq \max_m \min_s \frac{r_{ms}}{\min(d, d')} \tag{15}$$

Refer to the bound as γ.

We can choose bounding methods and select central classifiers using any approximation of $\beta + \delta$ that neither references validation and test outputs nor differentiates between validation and test inputs. We can approximate $\beta + \delta$ by altering the bound (15). The average rate of disagreement in each data set is $\frac{r_{ms}}{n}$, so substitute $\frac{r_{ms}}{n}$ for $\frac{r_{ms}}{min(d,d')}$. We still have the rate of disagreement over validation inputs bounding the difference in validation errors β. Scale the disagreement to reflect any *a priori* beliefs about the relationship between disagreements and error rate differences. For example, to express a belief that, on average, the validation error difference is half the rate of disagreement, replace $\frac{r_{ms}}{n}$ by $\frac{r_{ms}}{n}(\frac{1}{2}\frac{d}{n} + \frac{d'}{n})$. Finally, instead of maximizing over classifiers g_m, take an average, weighted according to any *a priori* beliefs about which classifier is g^*. For example, if the initial classifiers have high training error, then give them less weight.

Tests

This section outlines the results of tests on a set of credit card data. Each example corresponds to a credit card user. There are six inputs that correspond to user traits. The traits are unknown because the data provider has chosen to keep them secret. There is a single output that indicates whether or not the credit card user defaulted. The data were obtained from a machine-learning database site at the University of California at Irvine. The discrete-valued traits were removed, leaving the six continous-valued traits. Of the 690 examples in the original database, 24 examples had at least one trait missing. These examples were removed, leaving 666 examples. The data were cleaned by Joseph Sill. For further information, see [5].

There were 10 tests. In each test, the 666 examples were randomly partitioned into 444 training examples, $d = 111$ validation examples, and $d' = 111$ test examples. In each test, a classifier was trained, producing $M = 1000$ snapshots. The classifiers are artificial neural networks with six input units, six hidden units, and one output unit. The hidden and output units have tanh activation functions. The initial weights were selected independently and uniformly at random from $[-0.1, 0.1]$. The networks were trained by gradient descent on mean squared error over training examples, using sequential mode weight updates with random order of example presentation in each epoch. After each epoch, a snapshot was recorded.

In each test, eight sets of central classifiers were extracted. The first set contains all snapshots. Hence, the error bounds based on the first set of central classifiers are the traditional error bounds. The other sets of central classifiers were drawn from the snapshots at regular intervals of 10, 20, 50, 100, 200, 500, and 1000 classifiers. For example, the set drawn at intervals of 10 contains $S = 100$ central classifiers, snapshots $g_{10}, g_{20}, \ldots, g_{1000}$.

In each test, the validation data was used to determine g^*, the snapshot with minimum validation error, and ν_*, its validation error. For each set of central classifiers, the validation data and the test inputs were used to determine c^*, the best central classifier, ν_+, its validation error, and δ, the rate of disagreement between g^* and c^* over the test inputs. This information was used to derive test error bounds for g^* using formula (14).

Table 1 shows the averages over the 10 tests of the validation error of g^*, the validation error of c^*, the rate of disagreement δ between c^* and g^* over the test inputs, and the difference β between the validation errors of c^* and g^*. In the top line, each snapshot is a central classifier, so c^* is g^*. As the number of central classifiers S decreases, the validation error of the best central classifier increases and its rate of disagreement with the classifier chosen by early stopping also increases.

Table 2 shows the average upper bound on the test error of g^* that is achieved with 90% confidence when a fixed number S of central classifiers are used for all tests.

S	ν_*	ν_+	δ	β
1000	0.198	0.198	0.000	0.000
100	0.198	0.205	0.000	0.006
50	0.198	0.205	0.006	0.006
20	0.198	0.207	0.012	0.009
10	0.198	0.212	0.014	0.014
5	0.198	0.221	0.033	0.023
2	0.198	0.222	0.072	0.023
1	0.198	0.234	0.094	0.036

Table 1: For S central classifiers, validation error ν_* of g^*, validation error ν_+ of c^*, test set disagreement rate δ between c^* and g^*, and validation error difference β between c^* and g^*. (Average over 10 tests.)

S	$\nu_+ + \delta$	$\epsilon_{\min}(S)$	avg. bound
1000	0.198	0.253	0.451
100	0.205	0.208	0.413
50	0.211	0.199	0.410
20	0.219	0.181	0.400
10	0.225	0.163	0.388
5	0.254	0.145	0.399
2	0.294	0.118	0.412
1	0.328	0.091	0.419

Table 2: For S central classifiers, average upper bound on test error of g^* with 90% confidence. (The value $\epsilon_{\min}(S)$ is the minimum ϵ such that $SB(\epsilon) \leq 0.10$.)

To derive the bound, recall formula (14).

$$\Pr\{\nu'_* \geq \nu_+ + \delta + \epsilon\} \leq SB(\epsilon) \qquad (16)$$

Let $\epsilon_{\min}(S)$ be the minimum ϵ such that $SB(\epsilon) \leq 0.10$. The best upper bound with failure probability no more than 10% is $\nu_+ + \delta + \epsilon_{\min}(S)$. At first, the bound improves as the number of central classifiers is decreased. The decrease in $\epsilon_{\min}(S)$ more than offsets the increase in $\nu_+ + \delta$ as fewer central classifiers are used. Eventually, there are too few central classifiers to attain a good match between some central classifier and the classifier chosen by early stopping. After this, the best bound increases as the number of central classifiers is decreased.

Tables 3 and 4 show the results of tests to select the number of central classifiers using estimates of $\beta + \delta$, as discussed in the previous section. The bound γ, as defined in inequality (15), was computed for each test. This bound proved too loose to be useful because the central classifiers have high rates of disagreement with the initial snapshots in the training sequences. These rates determine the bound since it maximizes over snapshots. However, the initial snapshots are almost never chosen by early stopping.

An alternative estimator, γ_s, was computed by ignor-

S	$\beta+\delta$	γ_s	γ_a
1000	0.000	0.000	0.000
100	0.006	0.042	0.006
50	0.012	0.052	0.009
20	0.021	0.084	0.015
10	0.028	0.095	0.020
5	0.056	0.143	0.037
2	0.095	0.228	0.078
1	0.130	0.273	0.124

Table 3: For S central classifiers, the actual value of $\beta+\delta$ and the estimates γ_s and γ_a. (Average over 10 tests.)

ing the first 10 snapshots. Hence,

$$\gamma_s = \max_{m>10} \min_{s} \frac{r_{ms}}{\min(d, d')} \tag{17}$$

where r_{ms} is the number of validation and test inputs for which g_m and c_s disagree. Another estimator, γ_a, was computed by averaging disagreement rates over snapshots (instead of maximizing).

$$\gamma_a = E_m \min_{s} \frac{r_{ms}}{\min(d, d')} \tag{18}$$

Table 3 compares the average of $\beta + \delta$ to the average of γ_s and γ_a. On average, γ_a is more accurate than γ_s, γ_a underestimates $\beta + \delta$, and γ_s overestimates $\beta + \delta$.

Table 4 compares the bounds derived by choosing the number of central classifiers in four different ways. (The choice is over $S \in \{1, 2, 5, 10, 20, 50, 100, 1000\}$.)

1. Set $S = 1000$. This gives the bound without central classifiers: $\nu_* + \epsilon_{\min}(1000)$.

2. Choose S to minimize $\nu_* + \epsilon_{\min}(S) + \gamma_s$, i.e. use γ_s to estimate $\beta + \delta$.

3. Choose S to minimize $\nu_* + \epsilon_{\min}(S) + \gamma_a$, i.e. use γ_a to estimate $\beta + \delta$.

4. Choose S to minimize $\nu_* + \epsilon_{\min}(S) + \beta + \delta$. In practice, it is not valid to choose S this way, since computing $\beta + \delta$ requires knowledge of the partition of inputs into validation and test sets. (See the previous section.) This is the "ideal" bound that would be achieved by a perfect estimator of $\beta + \delta$.

Table 4 displays the average bound for each method and the standard deviation of the average bound as an estimate of the mean bound over all partitions of the data set into training, validation, and test sets, i.e. over all possible tests. This statistic shows that the average bounds obtained through selecting central classifiers with our estimates are statistically significantly less than the bounds obtained without central classifiers.

method	avg. bound	std. dev. of avg.
traditional	0.451	0.007
estimator γ_s	0.414	0.015
estimator γ_a	0.386	0.016
ideal	0.365	0.016

Table 4: Performance of four bounding methods. Statistics are over 10 tests.

Discussion

We have developed and experimented with a new test error bound for the classifier chosen by early stopping. Further results are available in a technical report [1]. The report introduces alternative types of central classifiers, analyzes the central classifier bound mathematically, and outlines bound procedures for the case in which test inputs are unknown.

Another report [3] develops similar results in the full VC framework. For more advanced applications of bounding by inference, see [2]. Finally, for improved uniform bounds over the central classifiers, see [4].

Acknowledgements

We thank Dr. Yaser Abu-Mostafa and Dr. Joel Franklin for their teaching and advice.

References

[1] E. Bax, Z. Cataltepe, and J. Sill, Alternative error bounds for the classifier chosen by early stopping, CalTech-CS-TR-97-08.

[2] E. Bax, Validation of voting committees, CalTech-CS-TR-97-13.

[3] E. Bax, Similar classifiers and VC error bounds, CalTech-CS-TR-14.

[4] E. Bax, Improved uniform test error bounds, CalTech-CS-TR-15.

[5] J. Sill and Y. Abu-Mostafa, Monotonicity hints, to appear in *Advances in Neural Information Processing Systems, 9.* 1997.

[6] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data* p.31, Springer-Verlag New York, Inc. 1982.

[7] V. N. Vapnik and A. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, *Theory Prob. Appl.*, 16(1971):264-280.