

Classifiers from Dr. Vanathi Gopalakrishnan

December 8, 2003

What are the common methods for estimating accuracy of classifiers induced by supervised learning algorithms? Based on evidence from literature and your hands-on experience (if any), comment on the strengths and weaknesses of each approach and its scope of application.

Answer

Estimating accuracy of learned classifiers is an important problem in machine learning for at least three reasons. First, how accurate a classifier is to a large extent determines how useful it is. An estimate of the accuracy gives an indication of whether to use the classifier for a particular problem and how much confidence to have in its classifications. For example, an accurate classifier can aid doctors in medical diagnosis, while an inaccurate one can be dangerous. Second, accuracy estimates of a learned classifier is part of the learning process of most learning algorithms. Third, a comparing the accuracy of classifiers produced by learning algorithms is important to researchers for understanding and developing learning algorithms.

There are several common methods for estimating accuracy. All such methods are necessarily heuristic, in the sense that they require assumptions about the classifier or the classification problem, and can fail if the assumptions are wrong. [Duda et al 2001, p482]

1 Accuracy

Given a classifier, we want to know how correctly it classifies instances of some population, in order to decide whether to use it for a particular problem. More generally, given a classification learner, we want to know how accurate are the classifiers it learns on a data sample, in classifying instances of the population.

Let X be a population with probability distribution p . Let D be a set of data sampled from X according to p . Let L be a classification learning algorithm whose instance space includes X . Let $L(D)$ be the classifier learned by L on D . If x is in X , let $L(D, x)$ be the class $L(D)$ assigns to x . If x is an instance drawn from S , let $c(x)$ be the actual class of x , and $f(x)$ be the class a classifier f assigns to x . Thus, we are interested in the true accuracy of f , also called the generalization accuracy of f .

Definition 1: The true accuracy a of a classifier f , is the probability that f correctly classifies an instance of X randomly drawn from according to X 's probability distribution. The true error e of f is $e(f, X) = 1 - a$. Usually we cannot measure a directly because X is too large. Instead, we must estimate the accuracy based on a sample S from X , usually a random sample from X .

Definition 2: The *sample error* of f on S is the number of errors of f makes on S , $e(f, S) = \frac{1}{n} \sum_{x \in S} I(f(x), c(x))$, where $I(f(x), c(x)) = 0$ if $f(x) = c(x)$, and 1 otherwise. The *sample accuracy* of f on S , $a(f, S) = 1 - e(f, S)$. Let \hat{a} be our estimate. Thus, we are interested to know two things:

1. The best estimate of the true accuracy p , given a sample error of S , and
2. The probable error in this estimate.

There are two possible sources of error in estimate: *bias* and *variance*. Bias is a systematic under- or over-estimation, while variance is the variability of the estimate about its mean.

2 Methods for Estimating Accuracy

2.1 Cross-Validation Error

Cross-validation (CV) is a very widely used method for estimating the accuracy of a learner L , i.e. the generalization accuracy of the hypotheses it learns. (It cannot be used for a single classifier.) Learning and testing are performed multiple times over different training and test sets.

Given a data set D , it is partitioned into k disjoint sets, D_1, \dots, D_k , of approximately equal size, each D_i randomly drawn from D . The learner is trained once on each D_i , and each time the accuracy of the learned classifier $L(D_i)$ is calculated on the remaining instances, $D \setminus D_i$. The sets $D \setminus D_i$ are called validation sets and calculating their error is called validation. The estimate is the mean of the k test accuracies:

$$\hat{a} = \frac{1}{k} \sum_{i=1}^k a(L(D_i), D \setminus D_i)$$

The major benefit of CV is that it gives accurate estimates without requiring a test data set disjoint from the training set. CV can be applied to any classification learner.

CV is often used to improve the generalization accuracy of the learned classifiers, by adjusting the learning parameters until the learned classifier achieves a high cross-validation accuracy. This is often found to improve generalization accuracy. However, CV is heuristic (like all other estimation methods). For some problems the classifiers with the maximum validation accuracy have the best generalization accuracy [Duda et al 2001, p484]. A heuristic for choosing the proportion of data $|D_i|$ used for training is that it should be greater than

$|D \setminus D_i|$ because during learning it is used for adjusting many parameters, while $|D \setminus D_i|$ is used for adjusting only one (for example, when to stop learning) and therefore this process has fewer degrees of freedom. After the parameters have been adjusted using CV, the validation accuracy is still a good estimate if the true accuracy.

2.2 Disjoint Test Set Error

The most commonly used method for estimating the accuracy of a classifier is the error on a data sample S independently drawn from X , and disjoint from D :

$$\hat{a} = a(f, S) \text{ for } S \cap D = \{\}.$$

Assuming S is drawn from X , this is an unbiased estimator of a . This is a more practical method than (2.1).

The drawback of this estimation method is that it requires a separate test data set. This is a great constraint when data is scarce and the hypothesis space is large, as is very often the case. It requires us to decrease the accuracy we are trying to measure on order to measure it!

Naturally, it is possible to estimate f 's accuracy using the error on a test set that overlaps partially with the training set. The bias would be greater than for the disjoint test set estimator, and smaller than training error estimator. I do not know any formal results for this method.

2.3 Corrected Training Error

If D is the data used to learn f . If D is independently drawn from X , the simplest estimate of true accuracy a is the sample accuracy on D :

$$\hat{a} = a(f, D).$$

For most learning algorithms this is an overestimate of a . In particular, the Hoeffding bounds [Mitchell, p210]

$$P(e(f, X) > e(f, D) + z) \leq \exp(-2|D|z^2)$$

bounds the probability that the sample error on the training set is an underestimate [Mitchell 211]. D is assumed to be independently drawn from X , and f is an arbitrary hypothesis.

Many learners (so-called *agnostic learners*) do not assume the target concept c is in their hypothesis space H , but choose a hypothesis in H with the minimum training error. If f is minimum-error hypothesis, then the bound becomes

$$P((\exists f \in H) (e(f, X) > e(f, D) + z) \leq |H|\exp(-2|D|z^2))$$

where H is the hypothesis space of the learner that learned f . The size of $|D|$ required to guarantee a particular accuracy of the estimate grows linearly in $\frac{1}{z^2}$ and $|H|$ [Mitchell, p210]. Thus, this estimator is useful only when data is

plentiful and the learner considers only a small number of hypotheses. As a result, this is not a commonly used estimator

Some algorithms apply a correction to the training accuracy and use the corrected accuracy as an estimate of a . The correction is supposed to make the estimate more “pessimistic” [Mitchell, p71], making up for the fact that the training accuracy is a biased estimate. For example, C4.5, calculates the estimated standard deviation $\hat{\sigma}(a(f, D))$ of the training accuracy, and uses \hat{a} equal to the lower bound of some confidence interval for $a(f, D)$ [Mitchell, p72]. For example, for a 95% confidence interval,

$$\hat{a} = a(f, D) - 1.96 \hat{\sigma}(a(f, D)).$$

An advantage of this approach is that it saves the effort of computing the sample accuracy on another data sample, since for many algorithms the training accuracy is computed in constructing the hypothesis.

In effect, for large datasets the standard deviation is very small and the pessimistic estimate is close to the training accuracy. As $|D|$ increases, the estimate deviates further from the training accuracy.

Although this estimation method is not statistically valid, it has been found useful in practice. However, much less used than the accuracy of a disjoint test set, or cross-validation.

References

- Mitchell (1996). *Machine learning*.
- Duda, Hart, Stork. *Pattern Classification* (2nd edition, 2001).