Predicting the Relation between Model Class, Domain, and Error Rate

Tobias Scheffer
University of Magdeburg
FIN/IWS, Universitätsplatz 2, 39106 Magdeburg, Germany
scheffer@iws.cs.uni-magdeburg.de

Extended Abstract

One of the questions central to the field of Meta-Learning is how domain properties, the model class that a learner uses, and the resulting error rate relate. The histogram of error rates in a given model class is a joint property of model class and domain. For a certain class of learners, this property (together with the sample size) can be shown to approximately determine the resulting generalization error. In many cases, this result can be exploited to determine a model class which is close to optimal for a given domain. In my talk, I will summarize the analysis that quantifies the generalization error in terms of this histogram, sketch applications of the analysis to model selection problems, and discuss a list of open questions.

Sketch of the analysis. The analysis of the resulting generalization error is based on the assumption that the model class contains only finitely many hypotheses. The learning algorithm minimizes the empirical error rate within this model class, although it is not necessary that the learner finds the hypothesis with globally smallest error rate. When there are several hypotheses with the same small empirical error e, then it is assumed that the learner will find them with equal prior probability. Given the resulting empirical error, what can we say about the resulting generalization error? In a crude simplification of the analysis, consider the equation $P(\text{generalization error}|\text{empirical error}) = c \times P(\text{empirical error}|\text{generalization error})P(\text{empirical error})$ which says that we can quantify the generalization error in terms of the likelihood $P(\text{empirical error}|\text{generalization error})$ (which is simply the binomial distribution) and the prior $P(\text{empirical error})$ which is the histogram of error rates in the model class (with respect to the given domain). See [7] for the derivation.

When the learner is exhaustive and can hence be guaranteed to minimize the empirical error in the model class, then we do not even need to know the resulting empirical error. When we know the number of hypotheses in the model class and make the simplifying approximation that the empirical error rates of distinct hypotheses are independent (given the generalization error rates), then we can determine the distribution that governs the resulting empirical error. The resulting empirical error is e if none of the hypotheses incur a lower empirical error. Since the histogram tells us how many hypotheses have which generalization error we can calculate $P(\text{least empirical error is } e) = \prod_h P(\text{empirical error} \geq e|\text{generalization error of } h)$. The analysis is described in [10, 6, 7].

Applying the analysis to the model selection problem. Given the histogram of error rates in the model class, the sample size (and, for non-exhaustive learners, the empirical error), the analysis yields the expected generalization error of a learner that uses the given model class for the given domain. In a typical model selection scenario, we have a list of potentially available model classes (perhaps decision trees with i leaf nodes) and would like to select the model class (the number of leaves i) that minimizes the resulting generalization error. A possible heuristic for the solution of the model selection problem would be to estimate the error histogram for each model class, and to use the model class with least predicted generalization error for learning. In principle, the error histogram can be estimated by recording the empirical error rate of some hypotheses in the model class. Some results on the complexity of estimating the error histogram can be found in [6].

The resulting model selection algorithm has been applied to the problem of selecting the optimal number of leaves of decision trees [7] and to a large text categorization problem [10, 6].

Quantifying the error rate of cross validation. Similarly to the generalization error of the hypothesis which is returned by a learner, we can quantify the generalization error of the hypothesis which is returned by a cross validation wrapper. The wrapper invokes the learner once for each considered model class and uses the n-fold cross validation error to select the model class which is finally used for learning. When the error histograms of all model classes are known, we can determine the resulting error rate [8]. We can use this result as a heuristic to decide how the sample should be split into training and holdout set (for holdout testing), or to decide whether, in this domain, holdout testing or n-fold cross validation should be considered *at all*, or whether simple empirical error minimization is likely to lead to a better result.

Open problems. Many questions in this framework are left to be answered. One of the most fundamental question is whether there are useful confidence bounds for the predictions. Ideally, the predicted error should come with a confidence interval which is based on the sample size and number of hypotheses which were used to estimate the error histogram. One result in this direction was obtained in [3]. A second fundamental question involves handling infinite hypothesis languages and learners that do not break ties at random. Finally, the analysis does not predict the effect of *local* changes of hypotheses (*e.g.,* tree pruning) well.

Related Work. Average-case analyses have been presented for decision stump learners [2], k-nearest neighbor [5], naive Bayesian classifiers [4], and some other learners. Analyses in the same spirit but with stronger simplifications have been presented in [9, 1].

References

[1] P. Domingos. Process-oriented estimation of generalization error. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligenct*, 1999.

[2] W. Iba and P. Langley. Induction of one-level decision trees. In *Proceedings of the Ninth International Conference on Machine Learning*, pages 233–240, 1992.

[3] J. Langford and D. McAllester. Computable shell decomposition bounds. In *Proceedings of the International Conference on Computational Learning Theory*, 2000.

[4] P. Langley and S. Sage. Tractable average case analysis of naive bayes classifiers. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 220–228, 1999.

[5] S. Okamoto and Y. Nobuhiro. An average-case analysis of the k-nearest neighbor classifier for noisy domains. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 238–243, 1997.

[6] T. Scheffer. *Error Estimation and Model Selection.* Infix Publisher, Sankt Augustin, 1999.

[7] T. Scheffer. Nonparametric regularization of decision trees. In *Proceedings of the European Conference on Machine Learning*, 2000.

[8] T. Scheffer. Predicting the generalization performance of cross validatory model selection criteria. In *Proceedings of the International Conference on Machine Learning*, 2000.

[9] T. Scheffer and T. Joachims. Estimating the expected error of empirical minimizers for model selection (abstract). In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 1998.

[10] T. Scheffer and T. Joachims. Expected error analysis for model selection. In *Proceedings of the Sixteenth International Conference on Machine Learning*, 1999.