Machine Learning, 48, 5–7, 2002

Guest Introduction: Special Issue on New Methods for Model Selection and Model Combination

A classical challenge in fitting models to data is managing the complexity of the models to simultaneously avoid under-fitting and over-fitting the data; fulfilling the goal of producing models that generalize well to unseen data. The papers in this special issue present recent developments in model-complexity control for supervised learning. These thirteen papers represent three areas of significant current research on this subject: *model selection* (explicitly choosing model complexity), *sparse models* (reducing complexity by enforcing sparse representations), and *model combination* (combining multiple models to improve generalization).

Model selection

The first four papers in this special issue discuss recent advances in model selection methods. The basic objective of model selection is to choose a model class that has appropriate complexity for the data, knowing that a class that is too constrained will under-fit the data while a class that is too complex will over-fit. The novelty and effectiveness of the methods presented in this special issue appears to be surprising given the relative maturity of this field. The first two papers, by Chapelle et al. and Surgiyama and Ogawa, present new analyses of generalization error for regression that are based on small sample size bounds rather than asymptotic analyses. These estimates lead to new model selection criteria that extend traditional techniques. Although the two analyses are similar, Sugiyama and Ogawa achieve exact estimates under certain assumptions whereas Chapelle et al. make fewer assumptions but require certain quantities to be empirically approximated. The third paper, by Schuurmans and Southey, presents a geometric view of model selection that yields alternative estimators for expected generalization error, and suggests alternative model selection and regularization strategies. All three of these papers demonstrate positive empirical comparisons to standard model selection criteria such as AIC, BIC and cross validation.

The fourth paper, by Bartlett et al., presents a general analysis of the error of model selection methods that employ data-dependent complexity penalization. They propose new penalization methods that are more data-dependent than structural risk minimization, based on examining training error differences on two half-samples. Bartlett et al. also investigate the relationship between generalization error estimation and model selection quality. Overall, these four papers contribute new ideas to a very old subject.

Sparse models

Motivated by the success of support vector mechines (SVMs), recent work on kernel methods for generalization has focused on finding sparse kernel representations to fit data. The next group of papers investigates learning methods that explicitly seek sparseness in model representations. Sparseness constitutes another form of Occam's razor, where sparse representations in parameter space are considered to be simpler than non-sparse representations. Although these techniques are similar in motivation to model selection, techniques that enforce sparsity are usually not driven by direct attempts to estimate generalization error. Rather, for these methods, generalization analyses are most often conducted post hoc and external to the basic development of the algorithm.

The first paper, by Lin et al., presents a review of the statistical aspects of SVMs and kernel methods for classification, shedding light on why these methods generalize well. The specific techniques they investigate achieve sparseness by using GACV estimates to select hyperparameters in a model-selection-like manner. Gunn and Kandola then investigate SVMs with ANOVA kernels (a polynomial basis) on which it is practical to achieve sparseness by explicitly performing basis selection. They show that sparsity improves the interpretability of the learned model and has the potential to improve generalization. They also present techniques for visualizing the selected basis.

The third paper, by Vincent and Bengio, presents a greedy algorithm that incrementally adds basis functions in a boosting-like manner, but under a squared error loss—showing a connection between boosting and matching pursuit. Sparseness is obtained by early stopping with a validation set. In the case where the basis functions are kernels, their algorithm gives representations in an SVM form with some relaxed constraints. The final paper on sparse models, by Rätsch et al., tackles representations that involve a large or even infinite number of basis functions. They apply important ideas from mathematical programming, such as column generation, to learn sparse representations in these large parameter spaces. As a group, these four papers establish important connections between sparsity, SVMs, boosting and mathematical programming, and propose new learning algorithms based in these ideas.

Model combination

The last group of papers explores issues of model combination. Recently, theoretical and empirical research has demonstrated that generlization can be improved by combining multiple learned models. The first two papers on model combination are related to the previous two papers on sparse models in that they present recent developments on boosting methods. The first paper, by Mannor and Meir, presents new theoretical results on the existence of weak learners, which is a strong precondition for boosting to be applicable. They also prove new generalization error bounds which could be effectively used for early stopping in boosting training. Collins et al. then introduce a new framework that unifies logistic regression and Adaboost by casting both training criteria as optimizations of certain Bregman distances. Their paper proposes a new family of algorithms that interpolate between the behavior of standard logistic regression and boosting training algorithms. General proofs of convergence are provided for the complete family of algorithms they analyze.

The next two papers investigate other types of model combination methods beyond boosting. First, Bylander considers the problem of estimating the generalization error of bagging (a very simple and popular combination method) and introduces a new correction for the bias of out-of-bag estimates. The resulting estimation technique is more computationally efficient that using cross validation for the same purpose. Chipman et al. then present a Bayesian approach to combining tree models. Their work extends previous research by building local linear regression models at the leaves of a decision tree. They propose a particular prior over tree models and prescription for choosing hyperparameters which leads to stable estimates and a robustly applicable regression method.

The final paper in this special issue, by Melnik, analyzes the decision boundary structure of learned classifiers. His analysis is independent of the type of learner and allows one to compare and contrast alternative classifiers and qualitatively understand how they generalize—with the possibility of using this information to select or combine classifiers. Melnik's emphasis on interpretability of the classifier is related to the goals of Gunn and Kandola.

In sum, the papers comprising this special issue offer a glimpse of the state of the art in model-complexity control methods for machine learning. The breadth and recency of the results in each of the areas covered attests to the vitality of research currently taking place in model selection, sparse models and model combination. We hope that this focused collection of papers will prove useful in stimulating future research on these topics which are of central importance to machine learning.

Yoshua Bengio
Dale Schuurmans