Exploiting Multiple Existing Models and Learning Algorithms

Julio Ortega
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120
julio@almaden.ibm.com

Abstract

This paper presents MMM and MMC, two methods for combining knowledge from a variety of prediction models. Some of these models may have been created by hand while others may be the result of empirical learning over an available set of data. The approach consists of learning a set of "Referees", one for each prediction model, that characterize the situations in which each of the models is able to make correct predictions. In future instances, these referees are first consulted to select the most appropriate prediction model, and the prediction of the selected model is then returned. Experimental results on the audiology domain show that using referees can help obtain higher accuracies than those obtained by any of the individual prediction models.

In this paper, two currently active research programs in machine learning, theory revision (Flann & Dietterich 1989) (Mooney 1993) (Ourston 1991) (Baffes & Mooney 1993) (Richards & Mooney 1995) (Towell, Shavlik, & Noordewier 1990) (R. S. Michalski 1993) (Cohen 1992) (Bergadano & Giordana 1988) and bias selection (Merz 1995) (Ho, Hull, & Srihari 1994) (Brodley 1993) (Schaffer 1993), are viewed from a single perspective. Theory revision systems make use of two sources of knowledge: an existing imperfect model of a domain and a set of available data. Bias selection systems, on the other hand, make use of available data for a domain and several empirical learning algorithms. We propose two methods (MMM and MMC) that take a collection of existing models and learning algorithms, together with a set of available data, and create a combined model that takes advantage of all of these sources of knowledge.

MMC/MMM

The MAI (Model Applicability Induction) approach (Ortega 1994; 1996) addresses the theory revision problem by evaluating the effectiveness of the existing model (or domain theory) as a predictor using the available data. The data set available for training is divided into two categories: data on which the existing model is correct, and data on which the existing model is incorrect. This divided data set is used to a build a "Referee" predictor (using our default inductive method, C4.5 (Quinlan 1993)) that provides a mechanism for deciding in which situations the existing model should be chosen for the prediction of future instances. Another predictor, the "Data" predictor is built using induction on the available data. This predictor is used on future instances where the "Referee" predictor indicates that the existing model is incorrect.

This paper reports on MMM/MMC (Multiple MAI Majority/Multiple MAI Confidence), two extensions of the MAI approach to an arbitrary number of models, of which some may be pre-existing models with no ability to learn; some may be constructed by empirical induction from the available data; and still others may even be theory revision methods that make use of both prior knowledge and empirical learning. From now on, we will refer to the pre-existing models, learning algorithms, or theory revision systems used in the MMM and MMC approaches as *component* models. The MAI approach is extended to multiple models by building a "Referee" predictor for each of the component models. Each "Referee" predictor will tell us whether its corresponding component model can be trusted for particular unseen instances.

Constructing referees is a two step process: building a referee training data set and learning a referee based on this set. For components models with no ability to learn, referee training data sets are built by reclassifying the original training data set according to whether these models make correct predictions on each instance of the training data. Referee training data sets for components models with learning capabilities are similarly built using cross-validation: each training instance is evaluated with a model learned over a partition of the training data that did not contain that particular instance. Referees are then learned using a standard inductive method (C4.5 in our case).

We discuss two variations of this approach (MMM and MMC) that differ only in the manner in which the referees are used during classification. In MMM the final classification is decided by majority voting between the component models predicted to be correct by their corresponding referees. In MMC the final classification is that returned by the component model whose correctness can be trusted the most according a confidence level provided by each of the referees. The confidence estimate used in the experiments that follow is the ratio of instances of the majority class at each leave of the decision trees constructed by C4.5. More details about these algorithms can be found in (Ortega 1996).

The idea of separating the training data into subsets where models either succeed or fail to make correct predictions was used in Schapire's boosting algorithm (Schapire 1990). MMM and MMC can also be viewed as a form of probabilistic evidence combination, an idea that has been explored both in the decision tree literature ((Buntime 1991; Kwok & Carter 1990)) and in the neural networks literature (i.e., the "Mixture of Experts" approach of Jordan et al. (Jordan *et al.* 1991; Jordan & Jacobs 1994)). The idea of using auxiliary learners for combining multiple models was also exploited in Wolpert's stacked generalization (Wolpert 1992) and in the meta-learning methods of Chan and Stolfo (Chan & Stolfo 1995). Finally, Merz's DS approach (Merz 1995) can be viewed as a nearest-neighbor equivalent to the symbolic learning approach of MMM and MMC.

Experimental Design

MMM and MMC differ only in the methods used for combining the predictions of referees. To assess whether the accuracy improvements obtained with MMM and MMC are due to either the use of referees or the model combination methods it is useful to compare these approaches with two earlier ones, SAM (Select All Majority) and CVM (Cross-Validation Majority), discussed by Merz (Merz 1995). In the SAM approach, the prediction of each component model is an equally weighted vote for that particular prediction. The prediction with most votes is selected. CVM is a small variation of an algorithm suggested by Schaffer (Schaffer 1993). In CVM, the cross-validation accuracy for each model is estimated with the training data, and the model of highest accuracy is selected for use with all of the test data.

The comparison between the performance of the SAM, CVM, MMM, and MMC is interesting because these four systems can be characterized as varying across two dimensions:

		Model Applicability Induction	
		No	Yes
Model Combination	Voting	SAM	MMM
	Selection	CVM	MMC

Figure 1: SAM, CVM, MMM, and MMC approaches characterized by method combination and use of referees constructed by model applicability induction.

1. The first dimension considers the model combination method: voting or selection. Given a set of models, SAM and MMM choose the classification receiving the most votes, while CVM and MMC choose the prediction returned by the single model selected as the most "trustworthy".
2. The second dimension considers whether or not referees are learned (using model applicability induction) and later used to filter out component models that should not be trusted for specific test examples. Referees are used with our approaches, MMM and MMC, but not with SAM and CVM.

The characteristics of the SAM, CVM, MMM, and MMC approaches along these dimensions are summarized in Figure 1. In the experiments that follow, some differences in performance between these four approaches may vary across one dimension but not the other, permitting the assignment of credit or blame for such differences.

The experiments discussed in this paper were conducted with the audiology dataset from the UCI (University of California at Irvine) Machine Learning repository (Murphy & Aha 1992). For the approaches that use cross-validation (CVM, MMM, and MMC) the number of folds was set to 4. As component models in SAM, CVM, and MAI we use three non-learning models named *am10*, *am25*, and *am50*. These models were created following Mooney's approach (Mooney 1993) for generating models of varying degrees of imperfection. A perfect model, i.e., one that correctly classifies 100% of all audiology examples, was first constructed by running C4.5 on the complete data set of audiology examples with all pruning disabled. The result is a model that contains 86 rules with an average of 7.79 conditions per rule. The *am10*, *am25*, and *am50* models then were created by randomly adding and then randomly deleting a percent of all conditions from the rules of the perfect model (a corresponding *10%, 25%, and 50%*). This creates contaminated models with errors both of omission and commission. Adding con-

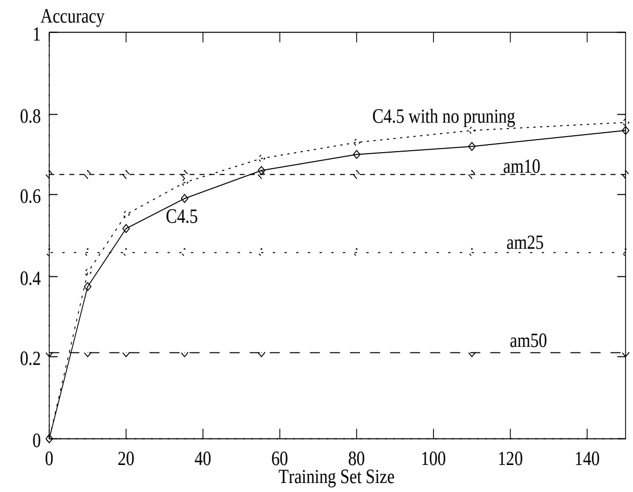

Figure 2: Baseline accuracies of the *am10*, *am25*, and *am50* non-learning models, and the *standard C4.5* and *C4.5 with no pruning* learning models.

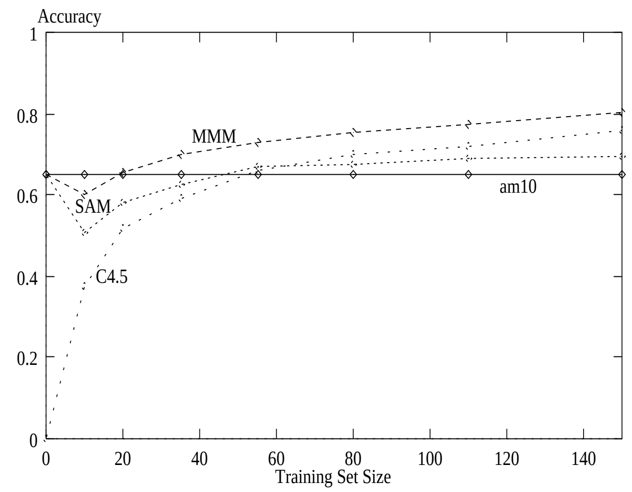

Figure 3: Accuracy of SAM and MMM with the *am10* model and C4.5 as component models.

ditions results in a model which is overly specific: for some examples, no rule of the model will be satisfied. Deleting conditions causes some rules to be overly general, thus producing incorrect predictions for some examples. As learning models we use the standard C4.5 learning algorithm in default mode and C4.5 with all pruning disabled.

Figure 2 shows the average accuracy of the models described above. Obviously, the accuracy of the non-learning models (*am10*, am25, and *am50*) is constant at their initial level for any size of the training set. The accuracies over the complete data set of the imperfect models *am10, am25,* and *am50* are *65%, 46%,* and *22%*, respectively. Accuracies for learning models depend on the size of the training set. With the complete training data set of 150 examples, *C4.5* and *C4.5 with no pruning* achieve accuracies of *76%* and *77%*, respectively. There is only a slight but consistent difference: the accuracy of C4.5 with no pruning stays above that obtained with the default C4.5 settings for any size of the training set.

Experimental Results with One Model and One Learning Algorithm

The first experiment examines the learning behavior of SAM, CVM, MMM, and MMC when the set of component models include just the *am10* model and the C4.5 algorithm with standard settings. For clarity, the results of this experiment are presented in two separate graphs: Figure 3 (SAM and MMM) and Figure 4 (CVM and MMC).

Figure 3 shows that SAM's accuracy appears to be an average of the accuracies of the component models and is very close to:

$$SAM's\text{accuracy} = \frac{\text{am10's accuracy} + \text{C4.5's accuracy}}{2}$$

In contrast, with training sets larger than 40 examples, MMM's accuracy is higher than that of both *am10* alone and C4.5. As noted earlier, SAM and MMM make predictions using a voting strategy. However, voting in MMM is restricted to the models whose referees predict them to be correct for a specific instance. Thus, the higher performance of MMM can be attributed to the referees' ability to appropriately decide whether the *am10* model or the tree build by C4.5 can be trusted for particular test examples.

As with SAM, CVM's accuracy is always a value between the accuracy of the component models, i.e., between the accuracies of *am10* model and C4.5 in Figure 4. CVM's mean accuracy is usually very close to that obtained by the best of the component models. CVM's accuracy is identical or very close to that of the best component model when the difference in accuracy between these models is large, and slightly lower than that of the best component model when these models have close, but not identical, mean accuracy values. As Figure 4 shows, CVM's accuracy is close to that of the *am10* model with 0 or 10 training examples, close to that of C4.5 with 150 training examples, and very similar to both with 55 training examples. Note that 55 is the number of training examples at which the accuracies of the *am10* model alone and that of C4.5 alone are approximately the same. CVM's worst performance occurs for the intermediate values of train-

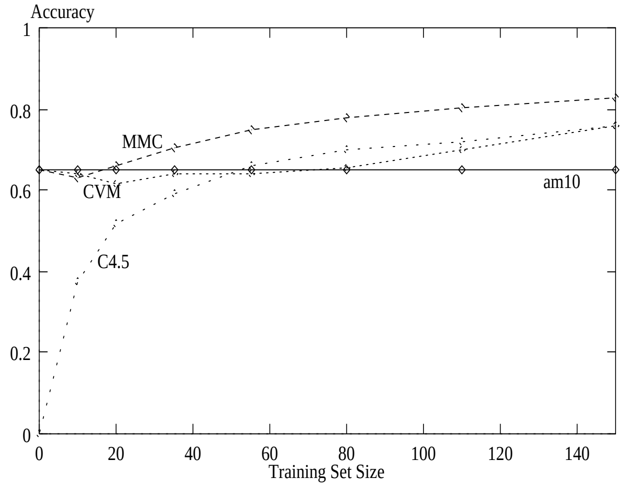

Figure 4: Accuracy of CVM and MMC with the *am10* model and C4.5 as component models.

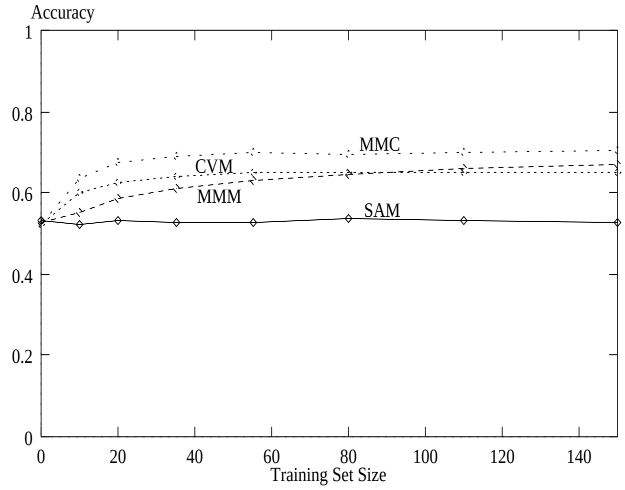

Figure 5: Accuracy of SAM, CVM, MMM, and MMC with the *am10, am25* and *am50* non-learning models.

ing set sizes approximately halfway between 0 and 55 (around 20 examples in Figure 4), and approximately halfway between 55 and 150 examples (i.e. around 80 examples in Figure 4).

When the mean accuracy of the two component models is similar, CVM's accuracy tends to be an average of the component models' accuracies. This occurs because the accuracy estimates employed by CVM may not be precise enough to discriminate between two models with similar accuracy. A large variance in these estimates causes CVM to often miss the correct choice for the best model. If the accuracy of the two component models is identical, nothing is lost. However, if the average of the these models is different enough, then CVM's accuracy tends toward the average of the two and, as happens with SAM, will be lower than that of the best component model.

In contrast to CVM, MMC can obtain accuracies that are higher than those of its component models. As Figure 4 shows, with 40 or more training examples, MMC's mean accuracy is higher than that of CVM, the *am10* model alone, or C4.5 alone. The increase in accuracy can be attributed to the use of referees to indicate which model can be trusted the most in specific instances, as both CVM and MMC use a selection strategy for combining the predictions of the component models.

Finally, it is worth noting from both Figures 3 and 4 that the approaches using a selection strategy (CVM and MMC) outperform their corresponding models that use a voting strategy (SAM and MMM). CVM matches or outperforms SAM for all training set sizes and MMC matches or outperforms MMM for all training set sizes. Thus, one can also conclude that selection appears to be a more appropriate strategy than voting for combining the predictions of component models in this particular learning situation.

Experimental Results with Multiple Models and Learning Algorithms

Figure 5 illustrates the different behavior of SAM, CVM, MMM, and MMC using the *am10*, *am25*, and *am50* as components. Since none of the component models can learn in this experiment, the available training data is never used by SAM. As a result, SAM's accuracy is fixed, regardless of the number of examples available for training. SAM's accuracy (53%) is higher than that of the two worst component models, *am25* (47%) and *am25* (22%), but lower than that of the best component model, *am10* (65%). As before, SAM's accuracy appears to be an average of the accuracy of the non-learning component models. With only a few component models of very different quality, as is the case in this experiment, SAM's majority voting strategy results in values lower than those of the best component model.

CVM, on the other hand, always chooses the best single model, as determined by the accuracy obtained by each component model on the training data. With non-learning models, the optimal choice for CVM is the best available component model, *am10* in this experiment. As a result, the accuracy of the best component model is an upper-bound on the accuracy that CVM can obtain. In the experiment corresponding to Figure 5, the upper-bound on CVM's accuracy is 65%, the accuracy of the *am10* model. However, this opti-

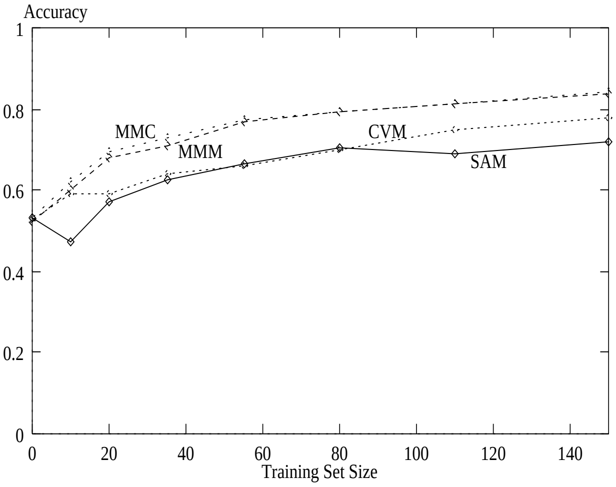

Figure 6: Accuracy of SAM, CVM, MMM, and MMC with several theories and learning algorithms.

mal choice is made consistently by CVM only when the training set is large enough to permit reliable estimates of the accuracy of the component models. In our experiment, this occurs with training sets containing 55 or more examples. With training sets containing fewer examples, CVM's accuracy is lower than that of the best model.

In contrast to CVM and SAM, the accuracy of MMM and MMC is not necessarily bounded by that of the best component model. As shown in Figure 5, MMM's accuracy with 150 training examples is 67%, and MMC's accuracy is 71%, both higher than the 65% accuracy of *am10*, the best of the component models.

Figure 6 shows the average accuracies obtained with SAM, CVM, MMM, and MMC when their component models include three non-learning models (*am10*, *am25*, and *am50*) and two learning models (*C4.5 with default pruning* and *C4.5 with no pruning*). As before, CVM's accuracy tracks closely that of the best component model for each given training set size, and SAM's accuracy appears to be an average of the accuracy of the component models. However, the voting strategy is not as damaging when restricted to the "trustworthy" models: MMM's accuracy is very similar to that of MMC. For training sets larger than 10 examples, both MMM and MMC obtain better accuracies than SAM, CVM, or any of the component models. With 150 examples, both MMM and MMC obtain mean accuracies of 84%, larger than the accuracy of both the best non-learning model, 65% of *am10*, and the best learning model, 77% of *C4.5 with no pruning*. The differences in accuracy are statistically significant at the 99% confidence level.

Other Results

More comprehensive experimental results with audiology models of varying quality and audiology data that includes varying degrees of noise are presented in (Ortega 1996) (WWW address: `http://csftp.vuse.vanderbilt.edu/theses/ortega/`). With sufficiently large degrees of noise in the data and existing models of relatively good quality CVM outperforms MMC. The ability of MMC to learn referees vanishes due to the lack of sufficient information in the noisy data, a factor to which CVM is immune.

Results with domains that use manually constructed models of soybean diseases, illegal chess positions, and DNA promoters are also presented in (Ortega 1996). In the experiments with these domains, the component models provided to SAM, CVM, MMM, and MMC consist of the C4.5 algorithm (using standard settings) and the respective manually constructed model.

In the soybean domain, the results are qualitatively similar to those shown for the audiology domain in Figures 3 and 4.

With the original data representation of the illegal chess position domain none of the SAM, CVM, MMM, and MMC approaches obtain a better accuracy than that obtained by the hand-crafted model alone. This is because C4.5's learning bias, that of constructing the smallest possible tree, is inadequate given the original representation of the data. As with noisy data, the best of these approaches is CVM, since it's accuracy closely matches that of the hand-crafted model. However, if the original data is re-expressed in terms of extended features extracted from the model, very significant improvements in accuracy are then possible, and MMC obtains the highest accuracy.

In the DNA promoters domain, none of the existing approaches obtained better accuracies than C4.5 alone, due to the fact that the predictive ability of the existing hand-crafted model is close to null. It's accuracy is close to that obtained by randomly selecting classifications.

Concluding Remarks

This paper presents two approaches, MMM and MMC, that combine the benefits of an arbitrary number of existing models and learning algorithms. In our experiments, where we use a small number of models/algorithms and good quality data, MMC always obtained better accuracies than MMM, and CVM always obtained better accuracies than SAM. The selection of a single most trustworthy model/algorithm appears to be superior to majority voting in the learning situations we examined. MMC obtained better accuracies than did CVM and any of the models/learning algorithms in experiments where the models/learning algorithms appeared to cover different portions of the description space, given that the representation of the data was adequate for learning referees. CVM obtained better accuracies than MMC in situations where the model and the data were of extremely different quality or where learning was impossible due to poor quality data or an inadequate data representation. In contrast to MMC, CVM never obtained better accuracies than those of the best of the models/learning algorithms.

References

Baffes, P. T., and Mooney, R. J. 1993. Symbolic revision of theories with M-of-N rules. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence.*

Bergadano, F., and Giordana, A. 1988. A knowledge intensive approach to concept induction. In *Proceedings of the Fifth International Machine Learning Conference.*

Brodley, C. E. 1993. Addressing the selective superiority problem: Automatic algorithm/model class selection. In *Proceedings of the Tenth International Conference on Machine Learning*, 17–24.

Buntime, W. 1991. Classifiers: A theoretical and empirical study. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, 638–644.

Chan, P. K., and Stolfo, S. J. 1995. A comparative evaluation of voting and meta-learning on partitioned data. In *Proceedings of the Twelfth International Conference on Machine Learning*, 90–98.

Cohen, W. W. 1992. Abductive explanation-based learning: A solution to the multiple inconsistent explanation problem. *Machine Learning* 8(2):167–219.

Flann, N. S., and Dietterich, T. 1989. A study of explanation-based methods for inductive learning. *Machine Learning* 4(2):187–226.

Ho, T. K.; Hull, J. J.; and Srihari, S. N. 1994. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16(1):66–75.

Jordan, M. I., and Jacobs, R. A. 1994. Hierarchical mixtures of experts and the em algorithm. *Neural Computation* 6:181–214.

Jordan, M. I.; Jacobs, R. A.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural Computation* 3:79–87.

Kwok, S. W., and Carter, C. 1990. Multiple decision trees. In Shachter, R. D.; Levitt, T. S.; Kanal, L. N.; and Lemmer, J. F., eds., *Uncertainty in Artificial Intelligence 4*. North Holland: Elsevier Science Publishers B. V. 327–335.

Merz, C. J. 1995. Dynamic learning bias selection. In *Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics*, 386–393.

Mooney, R. J. 1993. Induction over the unexplained: Using overly-general theories to aid concept learning. *Machine Learning* 10:79–110.

Murphy, P. M., and Aha, D. W. 1992. *UCI Repository of Machine Learning Databases.* Irvine, CA: Department of Information and Computer Science, University of California at Irvine.

Ortega, J. 1994. Making the most of what you've got: using models and data to improve learning rate and prediction accuracy. Technical Report CS-94-01, Computer Science Dept., Vanderbilt University. Abstract appears in *Proceedings of the Twelfth National Conference on Artificial Intelligence*, p. 1483, Seattle, WA.

Ortega, J. 1996. *Making the Most of What You've Got: using Models and Data to Improve Prediction Accuracy.* Ph.D. Dissertation, Vanderbilt University, Nashville, TN.

Ourston, D. 1991. *Using Explanation-Based and Empirical Methods in Theory Revision.* Ph.D. Dissertation, University of Texas, Austin, TX.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann.

R. S. Michalski, e. 1993. Special issue on multistrategy learning. *Machine Learning* 11:109–261.

Richards, B. L., and Mooney, R. J. 1995. Refinement of first-order horn-clause domain theories. *Machine Learning* 19(2):95–131.

Schaffer, C. 1993. Selecting a classification method by cross-validation. In *Preliminary Papers of the Fourth International Workshop on Artificial Intelligence and Statistics*, 15–25.

Schapire, R. 1990. The strength of weak learnability. *Machine Learning* 5(2):197–227.

Towell, G. G.; Shavlik, J. W.; and Noordewier, M. O. 1990. Refinement of approximate domain theories by knowledge-based neural networks. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, 861–866.

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks* 5:241–259.