Creating and Exploiting Coverage and Diversity

Carla E. Brodley and Terran Lane
School of Electrical
and Computer Engineering
Purdue University
West Lafayette, IN 47906
{brodley,terran}@ecn.purdue.edu

Abstract

In this paper, we illustrate that increasing coverage through diversity is not enough to ensure increased prediction accuracy—if the integration method does not utilize the coverage, then no benefit arises from integrating multiple models. We compare four criteria for selecting base-level classifiers and demonstrate that informed selection can lead to more accurate meta-level classifiers than random selection. In addition, we illustrate empirically that straightforward integration methods fail to utilize the diversity of the base-level classifiers.

Introduction

A fundamental step in forming a classifier that integrates multiple models is selecting the base-level classifiers. Ideally, the classifiers should be *diverse*; each should work well on different parts of the given dataset as no benefit arises from combining the predictions of a set of classifiers that all classify the same portion of the data correctly. A related objective is to maximize the *coverage* of the data, which is the percentage of the data that at least one base-level classifier can classify correctly. Achieving coverage greater than the accuracy of the best base-level classifier requires diversity among the base-level classifiers.

Diversity has been sought either by randomization of learning parameters (Hansen & Salamon, 1990; Heath, Kasif & Salzberg, 1995; Opitz & Shavlik, to appear), by combining multiple models produced by a single learning algorithm (Buntine, 1992; Oliver & Hand, 1995), by forming each base-level classifier using a different representation and search bias (Zhang, Mesirov & Waltz, 1992; Schaffer, 1994), or by training each classifier on different parts of the data (Brodley, 1995b; Chan & Stolfo, 1995). For the last approach, two further distinctions can be made. The first is whether the different parts of the data are chosen randomly or whether they are chosen to be regions that can be described by the attributes. The second distinction is whether the data partition is absolute (i.e., each base-level classifier is trained on its own mutually exclusive part of the data) or whether they are overlapping (such as sampling randomly with replacement would produce).

In this paper, we illustrate that generating base-level classifiers in order to maximize coverage of the data is not enough to ensure increased prediction accuracy—if the integration method does not utilize the diversity of the base-level classifiers, then no benefit arises from integrating multiple models. We illustrate empirically that straightforward integration methods fail to utilize the coverage of the base-level classifiers.

Measuring Coverage and Diversity

One metric for for determining the similarity of the classification decisions of a set of learning algorithms is *classification overlap* (Brodley, 1993). To compute the overlap among a set of classifiers requires counting the number of instances that were classified the same way by each of the classifiers, including instances that were classified incorrectly by all algorithms. To measure the diversity of a set of classifiers one can look at their classification overlap: a set of classifiers S_1 is more diverse than another set S_2 if: overlap(S_1) < overlap(S_2).

Krogh and Vedelsby (1995) measure the ambiguity (diversity) of a classifier as the sum of the squared differences between the output of the ensemble and the output of the classifier. In the case of a discrete classifier, this reduces to the number of instances on which the classifier differs from the ensemble, which is equivalent to the overlap between the base level classifier and the ensemble.

In seeking a diverse set of classifiers one does not want to sacrifice accuracy; ideally, the set of baseline classifiers should be selected such that both diversity and accuracy are maximized. One approach is to select the set of base-level classifiers such that coverage is maximized. However, maximizing coverage may con-

Table 1: Maximizing Coverage

	≥ 2	C_1 only	C_2 only	C_3 only
Scenario 1	80	5	2	3
Scenario 2	60	13	12	10

flict with maximizing each individual classifier's predictive accuracy. In these cases, sacrificing a small amount of accuracy in each base-level classifier may result in increased coverage.

For example, consider the two hypothetical scenarios shown in Table 1. For each scenario the table reports the percentage of the test data that two or more of the classifiers classify correctly, and the additional percentage that each individual classifier classifies correctly. The three classifiers in Scenario 1 have a higher individual classification accuracies (85, 82, and 83 versus 73, 72, and 70), but together they *cover* only 90% of the data, whereas the three classifiers in Scenario 2 cover 95% of the data. Given a combination scheme that can learn the *regions of expertise* of each base-level classifier, it may be desirable to trade individual accuracy for coverage.

This intuition is supported by Heath et. al's (1995) proof that decreasing the probability that there are instances that none of the base-level methods classifies correctly increases the probability of correct classification by a majority vote scheme. Opitz and Shavlik (to appear) model this tradeoff explicitly in their method for selecting a diverse, yet accurate, ensemble of neural networks.

Defining Areas of Expertise

Generation of diverse base-level classifiers will not, in and of itself, yield an accurate meta-level classifier. It is also necessary to be able to distinguish the data points for which each classifier is accurate and to be trusted. A meta-level classifier incapable of identifying the regions of expertise of its base-level classifiers may well be incapable of exploiting the full data coverage provided by those classifiers. It is our view that if diversity has been created through randomization, then for some datasets it will be difficult to learn or define areas of expertise.

In contrast to randomized methods for creating diversity, top-down methods explicitly assign a different classifier for each mutually exclusive subset of the data (Tcheng, Lambert, C-Y Lu & Rendell, 1989; Utgoff, 1989; Brodley, 1995b). This approach differs from methods that train each base-level classifier on a different, randomly selected subset of the data in two ways: 1) in top-down methods regions of expertise are defined before learning the base-level classifiers rather than afterward and 2) the subsets of instances are typically defined using the attributes rather than randomly.

Top-down methods require a procedure to determine whether one of the base-level classifiers is ideal for a given space (or subspace) or whether the space should be split into a new set of subspaces. For example, the Model Class Selection System (MCS) (Brodley, 1995b) uses characteristics of the given data set, in the form of feedback from the learning process, to guide a search for a tree-structured hybrid classifier. At each stage it chooses among a decision tree, a linear machine and a k-nearest neighbor classifier using heuristic knowledge about the data characteristics that indicate one bias is better than another.

The Difficulty in Exploiting Diversity

Our empirical results show that majority vote and stacked generalization integration schemes may fail to exploit the diversity of a set of base-level classifiers. Our hypothesis is that if the regions of expertise cannot be expressed easily by the input features then a stacked generalization will not utilize the coverage of the base-level classifiers. Empirical support for this hypothesis was generated using the following datasets:

Heart Disease: The Heart data set consists of 303 patient diagnoses (presence or absence of heart disease) described by thirteen symbolic and numeric attributes (Detrano, Janosi, Steinbrunn, Pfisterer, Schmid, Sandhu, Guppy, Lee & Froelicher, 1989).

Landsat: The task for this domain is to predict the type of ground cover from satellite images. Each of 1000 instances is described by seven features (the channels) and labeled with one of four types of ground cover.

Lymphography: This data set consists of 148 instances, each described by nineteen attributes and labeled as one of four classes.

NDVI90: This is a remote sensing dataset for North America at 1 Km resolution, extracted from the 1990 Conterminous US AVHRR Dataset compiled at Eros Data Center. The data consists of a time series of normalized difference vegetation index (NDVI) values representing the maximum NDVI during each of eight one month periods spanning the growing season in 1990. This index is commonly used to infer the amount of live vegetation present within a pixel at the time of data acquisition. There are twelve class labels. The data were extracted at 10,000 random locations, and exclude water bodies.

Figure 1: Coverage – Heart Disease

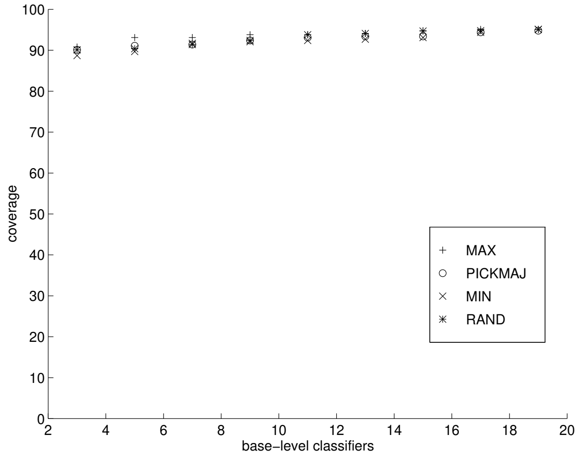


Figure 2: Coverage – Landsat

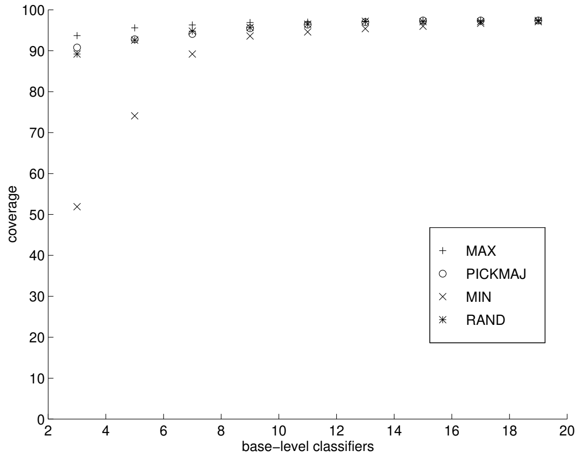


Figure 3: Coverage – Lymphography

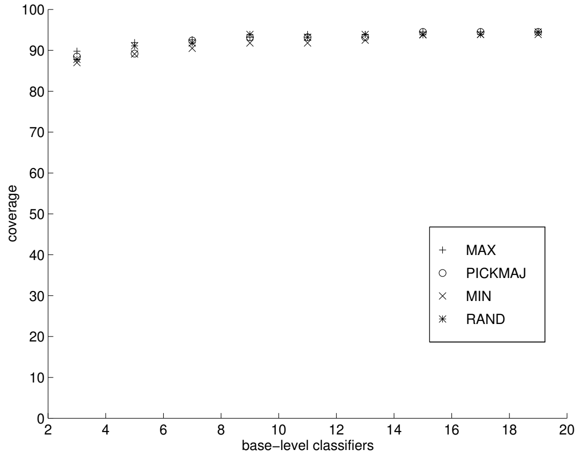
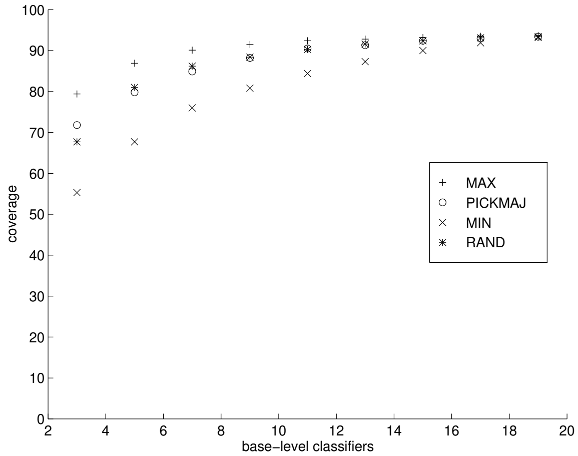


Figure 4: Coverage – NDVI90



In our experiments we used randomization to generate diverse classifiers. We choose linear machines (Nilsson, 1965) as our base-level classifiers and used the thermal training rule (Brodley & Utgoff, 1995a) to update the weights. This rule anneals the amount of correction to the weights, therefore when different random orders are used this is equivalent to random sampling with replacement. For example, for the NDVI dataset, the weights typically converge after approximately 1500 instances have been observed which is about 1/4 of the training data.

For each of ten runs, each dataset was divided into train and test data sets (90% train and 10% test) and the training set was further divided into base-level training and test sets (75% to base-level train and 25% to base-level test). A group of 20 linear machines with randomly selected initial weights was trained on the base-level training set using different random orderings. A majority-vote strategy and the C4.5 decision tree algorithm (Quinlan, 1993) were employed as integration schemes. The majority-vote classifier used the outputs of the linear machines directly, while C4.5 was trained in two ways: 1) using the linear machine decisions plus the original attribute data and 2) using only the linear machine decisions.

Coverage

To examine whether performance could be improved by careful selection of the base-level classifiers we compared the following four criteria: 1) random selection, 2) maximization of coverage, 3) minimization of coverage and 4) maximization of the accuracy when the LMs are used in a majority vote integration scheme (i.e. select the subset of LMs that yields the highest accuracy majority-vote classifier for the base-level test set). The four criteria are referred to as RAND, MAX, MIN, and PICKMAJ, respectively. Each selection cri-

Figure 5: Majority Vote – Heart Disease

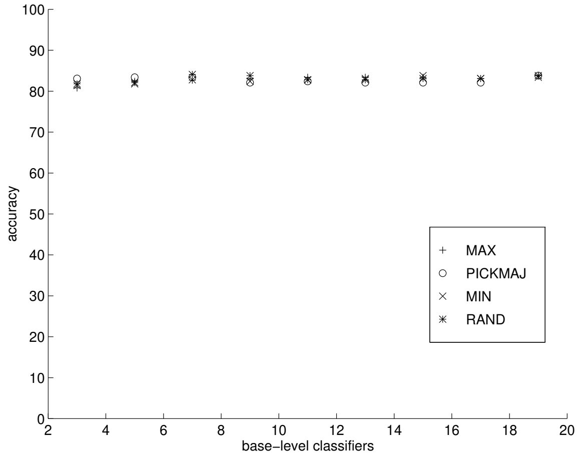
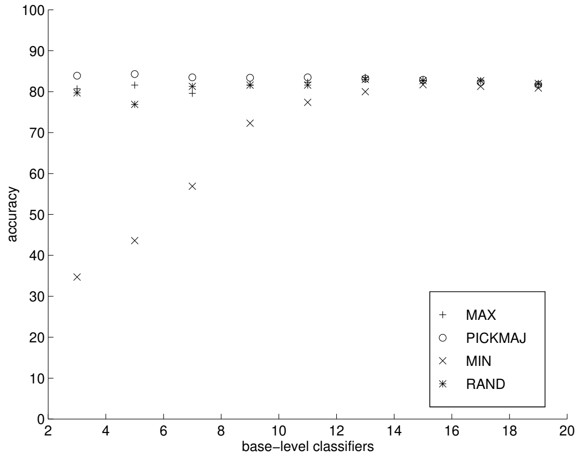


Figure 6: Majority Vote – Landsat



terion (other than RAND) chose base-level classifiers using a greedy stepwise forward selection procedure. The particular criterion was computed from the instances in the base-level test set.

Figures 1 to 4 show the coverage of the base-level classifiers selected by each criterion for subsets ranging from 3 to 19 LMs. The coverage was computed using the base-level test set. The figures illustrate that when only a few base-level classifiers are selected, the choice of criterion can lead to significantly different levels of coverage for the Landsat and the NDVI90 datasets.

Majority Vote

To evaluate the impact that increasing coverage has on the predictive accuracy of the meta-level classifiers we selected subsets, ranging from 3 to 19, of the 20 linear machines and applied a majority vote integration scheme. Ties were broken randomly. The results for

Figure 7: Majority Vote Ensemble –Lymphography

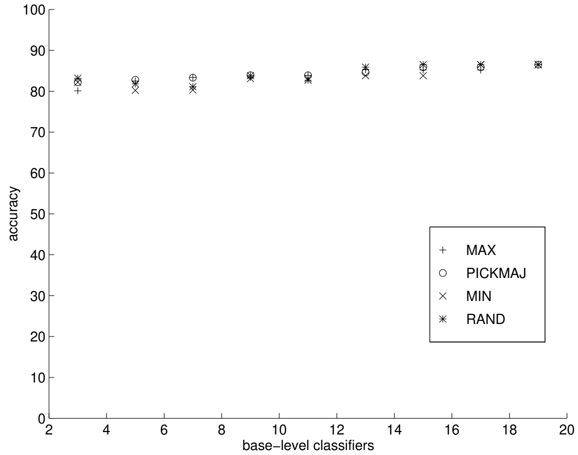
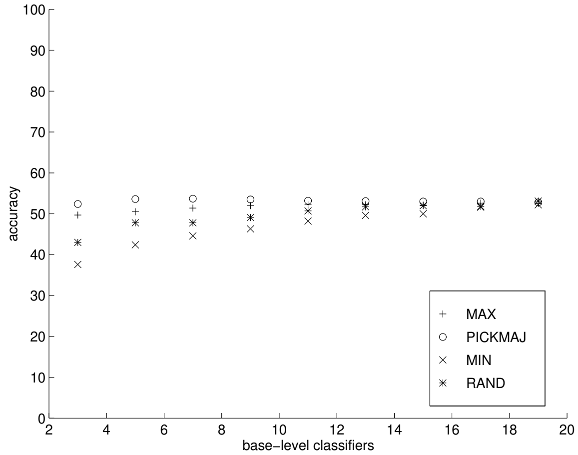


Figure 8: Majority Vote – NDVI90



the test data are shown in Figures 5 to 8. The accuracies of the meta-level classifiers were not significantly different for the MAX and PICKMAJ criteria. However, when fewer than nine LMs were included, RAND and MIN were in most cases significantly less accurate for the Landsat and NDVI90 datasets. For these datasets, a few of the base-level linear machines obtained much lower accuracies than others, which were selected by the MIN criterion. Because there were relatively few inaccurate classifiers, RAND's performance was closer to MAX. For the other two datasets, all of the base-level linear machines were of roughly equivalent accuracy and diversity, resulting in less disparity among the performance of the four selection criteria. These results illustrate that selecting base-level classifiers such that coverage is maximized can improve performance for some datasets.

Figure 9: Accuracy for Heart Disease

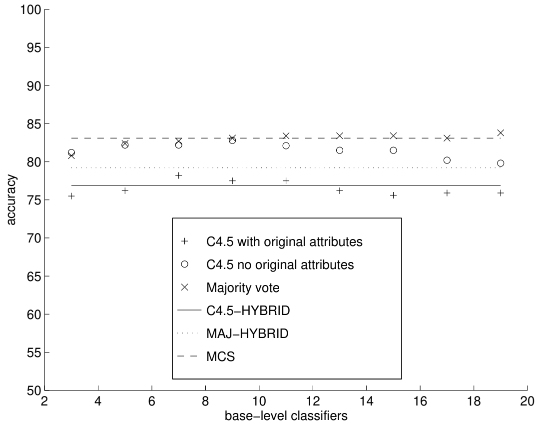


Figure 10: Accuracy for Lymphography

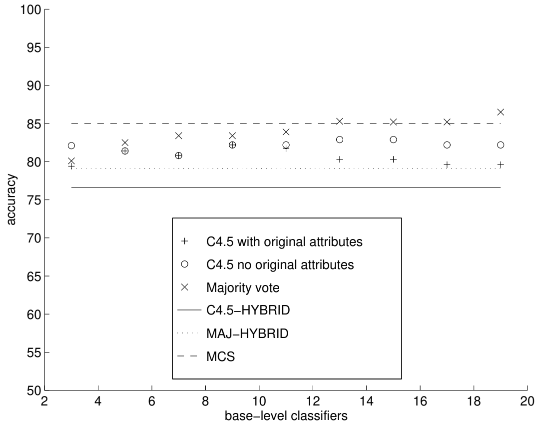


Figure 11: Accuracy for Landsat

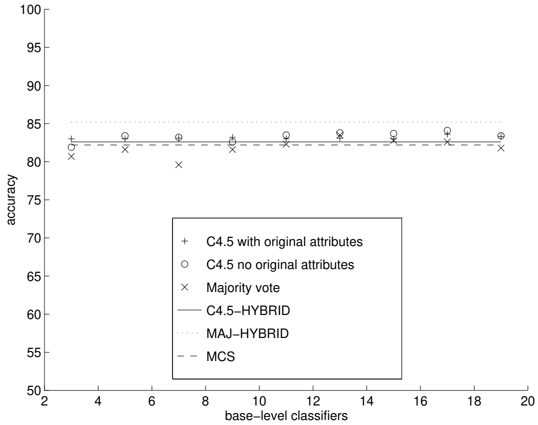
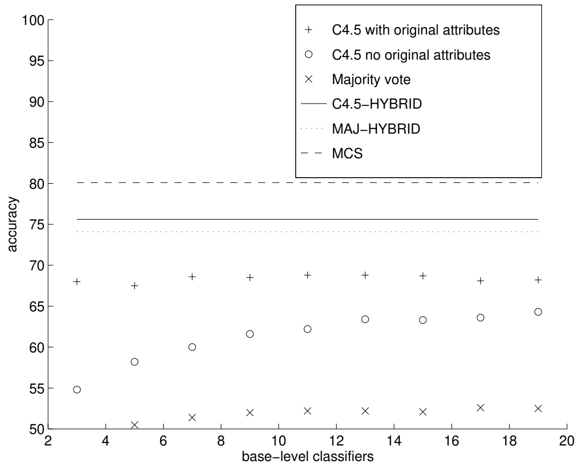


Figure 12: Accuracy for NDVI90



A Comparison of Integration Methods

The coverage results shown in Figures 5 to 8 illustrate that while the set of base-level classifiers are covering a large portion of the data space, the meta-level classifiers are unable to take advantage of much of this coverage. To determine whether it was the choice of base-level classifiers or the integration scheme that prohibited taking advantage of the coverage we ran experiments with several other combination schemes. In Figures 9 to 12 we plot the accuracy of each meta-level classifier on the final test set when the base-level classifiers were chosen to maximize coverage (MAX).

We employed C4.5 as an integration scheme with and without the original dataset attributes. We observed that the C4.5 tree built from the LM classification decisions and the original attributes is constructed primarily from the original attributes. Indeed, its accuracy is indistinguishable from the accuracy of C4.5 run on the original data. Excluding the NDVI90 datasets, C4.5 without the original attributes resulted in higher accuracy that C4.5 with the original attributes.

To examine whether it was our base-level classifier construction method that was causing the difficulty, we constructed a majority vote classifier of an LM, a decision tree and a 1-nearest neighbor (1-NN) classifier, shown in the Figures as MAJ-Hybrid. We also ran C4.5 as an integration scheme in conjunction with the LM, 1-NN and decision tree (C4.5-Hybrid).

The average accuracy over the ten runs is shown in Table 2. The accuracies in the table for MAJ-LM and C4.5-LM were the maximum average value observed over all possible numbers of base-level classifiers.[1] In addition to the meta-level classifiers, we included the

[1] The numbers for MAJ-LM were: Heart (19 LMs), Landsat (5 LMs), Lymphography (19 LMs) and NDVI90 (17 LMs). The numbers for C4.5-LM were: Heart (9 LMs), Landsat (13 LMs), Lymphography (13 LMs) and NDVI90 (19 LMs).

Table 2: Test Set Accuracy

	Heart Disease		Landsat		Lymphography		NDVI	
Method	Accuracy	Coverage	Accuracy	Coverage	Accuracy	Coverage	Accuracy	Coverage
MCS	83.1	83.1	82.2	82.2	85.0	85.0	80.1	80.1
1-NN	76.6	76.6	81.7	81.7	80.7	80.7	78.1	78.1
Dtree	77.6	77.6	82.2	82.2	77.1	77.1	71.0	71.0
LM	79.7	79.7	69.5	69.5	73.6	73.6	51.7	51.7
MAJ-LM	83.8	95.1	83.5	97.0	86.5	94.5	52.6	93.5
C4.5-LM	82.8	93.8	83.8	97.0	82.9	93.9	64.3	93.7
MAJ-Hybrid	79.2	90.4	85.1	91.7	79.1	92.8	74.1	89.7
C4.5-Hybrid	76.9	90.4	82.6	91.7	76.6	92.8	75.6	89.7

accuracies of a 1-nearest neighbor, a single linear machine (LM), a decision tree algorithm (Dtree) and of MCS, which integrates base-level classifiers that are each trained on a mutually exclusive part of the data. Recall that a fundamental difference between MCS and the other meta-level classifiers is that MCS explicitly assigns each instance to a particular classifier. The assignment scheme can be described by the attributes and therefore defines a geometric partition of the instance space rather than a random partition. For each method we include the base-level classifiers' coverage of the test data. Note that for MCS, 1-NN, LM, and Dtree coverage is equivalent to accuracy.

For the NDVI90 dataset, MCS is the most accurate method. For this dataset, the best individual classifier is a 1-NN. MCS is able to take advantage of situations in which the best approach is a single classifier, whereas a majority vote scheme cannot. For the Heart and Lymphography datasets, the best base-level classifier is a linear machine, which the meta-level classifiers MAJ-LM, C4.5-LM and MCS can all exploit. For the Landsat data, MAJ-Hybrid has the highest accuracy, but the difference between it and the other meta-level classifiers is not statistically significant. Finally, none of the meta-level classifiers can take advantage of the coverage achieved by its base-level classifiers. This analysis does not apply to MCS, for which the accuracy is equivalent to the coverage, but the coverage for MCS is much lower than for the base-level classifiers of the other methods.

Conclusions

In this paper we described a metric for choosing base-level classifiers that maximizes the coverage of a dataset. We illustrated empirically that straightforward integration approaches cannot exploit the diversity of the classifiers. We illustrated that informed selection of base-level classifiers can lead to more accurate meta-level classifiers for some datasets than random selection. Finally, our results indicate that new integration schemes are required to exploit the coverage achieved by the base-level classifiers. One can view this problem in two ways. First, the straightforward integration schemes examined in this article do not handle the type of diversity that cannot be expressed geometrically. An alternative view is that our coverage measure does not generate and measure the types of diversity that our integration methods can utilize.

Acknowledgments

We would like to thank our reviewers for their helpful comments.

References

Brodley, C. E. (1993). Addressing the selective superiority problem: Automatic algorithm/model class selection. *Machine Learning: Proceedings of the Tenth International Conference* (pp. 17-24). Amherst, MA: Morgan Kaufmann.

Brodley, C. E., & Utgoff, P. E. (1995a). Multivariate decision trees. *Machine Learning, 19*, 45-77.

Brodley, C. E. (1995b). Recursive automatic bias selection for classifier construction. *Machine Learning, 20*, 63-94.

Buntine, W. (1992). Learning classification trees. *Statistics and Computing, 2*, 63-73.

Chan, P. K., & Stolfo, S. J. (1995). A comparative evaluation of voting and meta-learning on partitioned data. *Machine Learning: Proceedings of the Twelfth International Conference* (pp. 90-98). Tahoe City, CA: Professional Book Center.

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disese. *American Journal of Cardiology, 64*, 304-310.

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions of Pattern Analysis and Machine Intelligence, 12*, 993-1001.

Heath, D., Kasif, S., & Salzberg, S. (1995). Committees of decision trees. In Gorayska & Mey (Eds.), *Cognitive Technology*. Elsevier.

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross valiation and active learning. *Advances in Neural Information Processing Systems 7*. Cambridge, MA: MIT Press.

Nilsson, N. J. (1965). *Learning machines*. New York: McGraw-Hill.

Oliver, J. J., & Hand, D. J. (1995). On pruning and averaging decision trees. *Machine Learning: Proceedings of the Twelfth International Conference* (pp. 430-437). Tahoe City, CA: Professional Book Center.

Opitz, D. W., & Shavlik, J. W. (to appear). Generating accurate and diverse members of a neural network ensemble. *Advances in Neural Information Processing Systems 8*. Cambridge, MA: MIT Press.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Schaffer, C. (1994). Cross-validation, stacking, and bi-level methods for stacking: Meta-methods for classification learning. In Cheeseman & Oldford (Eds.), *Selecting Models from Data: Artificial Intelligence and Statistics IV*. New York: Springer-Verlag.

Tcheng, D., Lambert, B., C-Y Lu, S., & Rendell, L (1989). Building robust learning systems by computing induction and optimization. *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence* (pp. 806-812). Detroit, Michigan: Morgan Kaufmann.

Utgoff, P. E. (1989). Perceptron trees: A case study in hybrid concept representations. *Connection Science, 1*, 377-391.

Zhang, X., Mesirov, J. P., & Waltz, D. L. (1992). Hybrid system for protein secondary structure prediction. *Journal of Molecular Biology, 225*, 1049-1063.