SAND2009-6940C

Relationships Between Accuracy and Diversity in Heterogeneous Ensemble Classifiers*

Sean A. Gilpin† Daniel M. Dunlavy‡

Abstract
The relationship between ensemble classifier performance and the diversity of the predictions made by ensemble base classifiers is explored in the context of heterogeneous ensemble classifiers. Specifically, numerical studies indicate that heterogeneous ensembles can be generated from base classifiers of homogeneous ensemble classifiers that are both significantly more accurate and diverse than the base classifiers. Results for experiments using several standard diversity measures on a variety of binary and multiclass classification problems are presented to illustrate the improved performance.

Keywords- classification, heterogeneous ensembles, diversity

1 Introduction

The problem of data classification, or data labeling, arises in a wide variety of applications. Examples include detecting spam e-mail messages based on the content of the messages (document classification), labeling cells and tumors as malignant or benign based on the context of MRI scan data (image classification), and identification of individuals based on fingerprints, facial features, and iris patterns (biometric identification). In all of these examples, the goal is to predict a discrete label (e.g., "spam" versus "not spam") for a particular data instance (e.g., a particular e-mail message) based on the attributes of that instance.

More formally, classification is the task of learning a function, f, that maps a set of data instance attributes, $\mathbf{x} = \langle a_1(\mathbf{x}), \ldots, a_m(\mathbf{x}) \rangle$, to one of several predefined class labels from $\mathcal{Y} = \{c_1, \ldots, c_k\}$. The function f is often called a classifier or classifier model, but in this paper we will use the term classifier to designate classification functions and the term classifier model will only be used to describe the structure of such functions. The set of data instances used to learn, or train, a classifier is called the training set; i.e., $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, where n is the number of instances, $\mathbf{x}_i \in \mathbb{R}^m$ is a vector of attributes, or features, for data instance i, and $y_i \in \mathcal{Y}$ is the label for data instance i. In order to validate the models learned, it is common practice to select some of the training data to be used in testing the resulting classifier models. This testing, or validation data, is not used in training the classifier model.

Recent results in solving classification problems indicate that the use of ensembles, or sets, of classifier models, often leads to improved performance over using single classifier models [1, 2, 3, 18]. Much of the previous work on ensemble classifier models (see e.g., [5]) has focused on *homogeneous ensemble classifiers*—i.e., collections of base classifiers of a single model type. In this work, we focus on *heterogeneous ensemble classifiers*, where the classifiers in the collection are not of the same type. Note that such classifier models are also referred to as hybrid ensemble classifiers.

Our work with heterogeneous ensemble classifiers focuses on using diversity measures as a tool to demonstrate how heterogeneous ensemble classifiers can outperform homogeneous ensemble classifiers. Specifically we explore the relationships between accuracy and diversity of the predictions across the base classifiers in each ensemble classifier. Our main goal in this work is to determine if heterogeneous ensemble classifier performance is correlated to diversity of its base classifiers.

A general discussion of the relationship between accuracy and different diversity measures can be found in [15] and references within. Some results on synthetic data are presented in that book, but there is no discussion of the issues associated with heterogeneous ensemble classifiers. In terms of diversity in heterogeneous ensembles, Bian and Wang [3] discussed theoretical connections between different diversity measures and performed several experiments relating accuracy to those measures. However, that work lacked a detailed study on the relationships between accuracy and diversity as a function of the how the heterogeneous ensemble classifier was constructed. In this work, we expand upon

*This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

†Computer Science, University of California at Davis, Email: sagilpin@ucdavis.edu.

‡Sandia National Laboratories, Albuquerque, NM 87123-1318. Email: dmdunla@sandia.gov.

this prior work, investigating more deeply the relationships between accuracy and diversity in heterogeneous ensemble classifiers and provide insight into how such classifiers can be used for improved classification performance.

Our contributions in this paper are summarized as follows:

- We show that a simple approach to creating heterogeneous ensemble classifiers using homogeneous ensemble classifiers leads to improved accuracy for all data sets we experimented with.
- Using empirical results on a variety of standard data sets, we demonstrate that heterogeneous ensemble classifiers are sensitive not only to the types of base classifiers used but the specific composition of base classifier models as well.
- We illustrate how larger ranges of diversity measures are created by varying the percent composition of base classifiers in heterogeneous ensemble classifiers as compared to varying directly the parameters used to generate homogeneous ensemble classifiers.

2 Diversity

The success of ensemble classification models over non-ensemble models is partially dependent on the diversity of the predictions made by its base classifiers [7, 11]. To see this, consider the case when all of the base classifiers make the same predictions. In that case the ensemble classifier would perform no better than any of the base classifiers taken individually and there would be no benefit to using an ensemble classifier. In order to objectively study the relationship between diversity and ensemble classifier performance, we first need to define objective measures of the diversity of predictions made by a set of classifiers.

There are two types of measures that have been used to study the diversity of ensembles: pairwise and non-pairwise. Pairwise measures are designed to compare the differences in predictions of two classifiers. Their interpretation in that setting is clear, but once averaged over all possible pairs in a base classifier set, the interpretation may become less clear. In this paper we only explore the mean of pairwise diversity measures but to get a better understanding of what these measures are trying to tell us it may be useful to study other statistics such as the standard deviation as well. Non-pairwise diversity measures are designed to measure differences in predictions of sets of more than two classifiers. Although their definitions are typically more complex than the pairwise diversity measures, their interpretations are not muddled by the details of working with ensembles of size greater than two.

The following pairwise diversity measures have been used recently to study the diversity of ensembles: disagreement [17], double fault [9, 17], correlation [15], and Yule's Q-statistic [20]. As well, the following non-pairwise diversity measures have been proposed for studying ensembles: entropy [6], Kohavi-Wolpert variance [13], interrator agreement measure [8], measure of difficulty [11], general diversity, and coincident failure diversity[14]. Bian and Wang showed that many of these measures have a level of similarity, and grouped the measures into 3 sets of correlated measures [3]. We chose one diversity measure from each group for use in our experiments: disagreement, double fault, and coincident failure diversity.

To generalize the pairwise diversity measures to an entire ensemble, we took the the average of the measurents over every pair of base classifiers. For a set of base classifiers $\mathcal{B}$ the average pairwise diversity can be calculated using:

$$Average = \frac{2\left(\sum_{i=1}^{|\mathcal{B}-1|}\sum_{j=i+1}^{|\mathcal{B}|} diversity_{i,j}\right)}{(n)(n-1)} \tag{2.1}$$

The following are definitions for the diversity measures that we explored.

2.1 Disagreement Disagreement between a pair of classifiers, f and g, is the proportion of instances for which they predict different class labels. The range of this measure is between 0 (always agree), and 1 (always disagree).

$$D = \frac{1}{n}\sum_{i=1}^{n}\mathcal{I}_{f,g}(\mathbf{x}_i) \tag{2.2}$$

where

$$\mathcal{I}_{f,g}(\mathbf{x}) = \begin{cases} 1 : f(\mathbf{x}) = g(\mathbf{x}) \\ 0 : f(\mathbf{x}) \neq g(\mathbf{x}) \end{cases} .$$

2.2 Double Fault Measure The double fault measure between a pair of classifiers is the proportion of instances for which they both predict the wrong class. The value of this measure is 1 when both of the classifiers are always wrong and 0 when the classifiers are never simultaneously wrong about the same instance. A low double fault measure is desired for an ensemble, because otherwise many of the base classifiers will be making incorrect predictions for the same instances which will increase the chance of that instance being misclassified by the ensemble classifier.

$$DF = \frac{1}{n}\sum_{i=1}^{n}(1 - \mathcal{O}_f(\mathbf{x}_i))(1 - \mathcal{O}_g(\mathbf{x}_i)) \tag{2.3}$$

where

$$\mathcal{O}_f(\mathbf{x}_i) = \begin{cases} 1 : f(\mathbf{x}_i) = y_i \\ 0 : f(\mathbf{x}) \neq y_i \end{cases} .$$

2.3 Coincident Failure Diversity Coincident failure diversity (CDF) is a measure whose value is highest (1), when misclassifications are unique to one base classifier and lowest (0) when all base classifiers always make the same class label predictions.

Let p_i denote the probability that exactly i of the L base classifiers predict the wrong class label for a randomly chosen instance. Then coincident failure diversity is defined as follows:

$$CFD = \begin{cases} 0 & : p_0 = 1 \\ \frac{1}{1-p_0}\sum_{i=1}^{L}\frac{L-i}{L-1}p_i & : p_0 < 1 \end{cases} \tag{2.4}$$

2.4 Examples Listed below is a classification problem where the true label for instance i is y_i. Also listed is an ensemble whose predictions $f*$ along with the base classifier predictions f_j are given as:

i	0	1	2	3	4	5
$f_1(x_i)$	a	a	a	b	b	c
$f_2(x_i)$	c	b	a	c	b	b
$f_3(x_i)$	a	a	a	c	c	c
$f_*(x_i)$	a	a	a	c	b	c
y_i	a	b	a	c	b	c

For diversity measures such as double fault and CFD we are interested only in which predictions are correct and not the actual class labels predicted. The following table, where the value 1 indicates that the prediction was correct and 0 indicates an incorrect prediction, is helpful for calculating those measurements.

i	0	1	2	3	4	5	Accuracy
$O_{f_1}(x_i)$	1	0	1	0	1	1	4/6
$O_{f_2}(x_i)$	0	1	1	1	1	0	4/6
$O_{f_3}(x_i)$	1	0	1	1	0	1	4/6
$O_{f_*}(x_i)$	1	0	1	1	1	1	5/6

To calculate CFD, first calculate values for p_i and then plug them into the equation given in the definition. The value of each p_i will be the proportion of instances that were incorrectly classified by exactly i of the base classifiers: $p_0 = 1/6$, $p_1 = 4/6$, $p_2 = 1/6$, and $p_3 = 0$.

$$CFD = \frac{1}{1-p_0}(\frac{2}{2}p_1 + \frac{1}{2}p_2 + \frac{0}{2}p_3)$$

$$CFD = \frac{6}{5} \times (\frac{4}{6} + \frac{1}{12}) = 9/10$$

The pairwise diversity measurements need to be calculated for each pair of base classifiers, of which there are three in this example. Each row in the following table shows the calculation of the double fault measure, DF, for one pair of base classifiers:

i	0	1	2	3	4	5	DF
f_1 vs f_2	0	0	0	0	0	0	0/6
f_1 vs f_3	0	1	0	0	0	0	1/6
f_2 vs f_3	0	0	0	0	0	0	0/6
f_*	$DF = (0 + 1/6 + 0)/3 = 1/18$						

To calculate disagreement, the actual predicted labels must be consulted (when the number of classes is greater than two). When two base classifiers are wrong about the same instance, they may still disagree with each other. For example, if the true class is a and the base classifiers predict b and c, they are both wrong, but they also disagree. So it is necessary to look at the predicted class labels to determine the proportion of instances they disagree about. Below is a table containing those proportions for each pair of base classifiers. A value of 1 indicates disagreement between the pair of base classifiers and a value of 0 indicates agreement.

i	0	1	2	3	4	5	D
f_1 vs f_2	1	1	0	1	0	1	4/6
f_1 vs f_3	0	0	0	1	1	0	2/6
f_2 vs f_3	1	1	0	0	1	1	4/6
f_*	$D = (4/6 + 2/6 + 4/6)/3 = 5/9$						

3 Numerical Results

In this section, we present numerical results illustrating the relationship between diversity and accuracy in ensemble classifiers as well as comparisons of the performance of heterogeneous versus homogeneous ensemble classifiers. All experiments were performed using Hemlock (Heterogeneous Ensemble Machine Learning Open Classification Kit) [10], a Java software library designed for investigating ensemble learning models. The base classifiers used in the experiments here were created using the Hemlock interface to the WEKA machine learning library [19].

Each of the ensemble classifiers created consisted of 100 base classifiers which were trained using bagging [4]. The use of bagging ensured that each base classifier was trained with a subsample of the training

set, and is a common method for increasing the diversity of ensemble classifiers [4]. Homogeneous ensemble classifiers consisted of base classifiers of support vector machines (SVMs) [16], random trees (RT) [5], and naive Bayes (NB) [12] classifiers as base classifier models that are implemented in WEKA. Default WEKA model parameters were used in each case. The heterogeneous ensemble classifiers consisted of combinations of the three classifiers types in pairs and in varying proportions. Each ensemble model was used in conjunction with 5-fold stratified cross validation to create a total of 5 ensemble classifiers. Using 5-fold cross validation allowed us to evaluate the diversity and accuracy measures using a test set not used in the training of the classifiers. We used 5-fold cross validation rather than 10-fold cross validation, because it lead to larger testing sets which we believe may be important when measuring diversity measures. In order to make up for the smaller number of training instances used in 5-fold cross validation, we repeated every experiment 10 times using different random seeds for either the fold creation or bagging or both (see the specific experiments later in this section for more details). Each heterogeneous ensemble classifier was created using a percentage of base classifiers from each of two homogeneous ensemble classifiers. The proportion of base classifiers (in 1% increments) led to different heterogeneous ensemble classifiers. This lead to a total of 297 different ensemble models for each experiment, 3 of which were homogeneous and 294 of which were heterogeneous.

Eight data sets representing a wide range of classification problems were used in the experiments. Table 1 presents the characteristics of these data sets; note the varying numbers of instances, classes, and features across the set. These data sets are a subset of the data used in previous work on analyzing performance of ensemble classifier models [1]. The performance measure computed for all experiments was *accuracy* (i.e., proportion of instances with correctly predicted labels) as it easily generalizes to multi-class classification problems. Majority voting [15] was used as the fusion method for combining the base classifiers to create the ensemble classifiers. Using voting as the fusion function seemed appropriate, because it uses the same type of output from the base classifiers (class labels) as the diversity measurements.

Table 2 shows the accuracy of ensemble models averaged over a total of 50 ensemble classifiers (corresponding to the 10 independent runs of 5-fold cross validation). Accuracy measures for the homogeneous ensemble classifiers consisting of SVM, NB, and RT base classifiers and the best heterogeneous ensemble classifier are presented. Note that across all data sets, at least one of the heterogeneous ensemble classifiers outperforms all homogeneous ensemble classifiers. Table 3 presents results of the average CFD measures corresponding to the accuracy results in Table 2. In 6 out of the 8 data sets, a heterogeneous model generated on average more diversity (as measured with CFD) than any of the homogeneous models. Table 4 shows the average disagreement measures for the different ensemble models. The RT ensemble classifier models dominate the other homogeneous models but the heterogeneous models still do better for 4 of the 8 data sets. Table 5 differs from the previous tables, because the double fault measure corresponds to more diverse ensembles when the value is low. So for the heterogeneous models we found the model that had the minimum average double fault measure.

Plots of accuracy and diversity measures for the *bupa* and *yeast* data set can be found in Figure 6 in Appendix A. The plots illustrate several easily identifiable trends between different measures across the data sets, but the combination of noise and nonlinear

Table 1: Characteristics of the experimental data sets.

Name	*Instances*	*Classes*	*Cont. Attr.*	*Nom. Attr.*
abalone	4177	29	7	1
bupa	345	2	6	0
dna	3186	3	0	180
glass	214	6	9	0
ion	351	2	34	0
promoters	106	2	0	57
sonar	208	2	60	0
yeast	1484	10	8	0

Table 2: Average accuracy over ten experiments repeated with different random seeds using 5-fold cross validation and bagging. At least one heterogeneous ensemble classifier outperformed all of the homogeneous ensemble classifiers across the data sets.

Average Accuracy				
Data set	*NB*	*SVM*	*RT*	*Het.*
abalone	0.2371	0.2532	0.2419	0.2555
bupa	0.5594	0.5809	0.7272	0.7333
dna	0.9396	0.9452	0.9501	0.9582
glass	0.5039	0.5820	0.7741	0.7791
ion	0.8253	0.8824	0.9367	0.9398
promoters	0.8874	0.9223	0.8994	0.9324
sonar	0.6899	0.8289	0.82886	0.82889
yeast	0.5792	0.5703	0.6080	0.6193

Table 3: Average coincident failure over ten experiments repeated with different random seeds using five fold cross validation and bagging. A heterogeneous ensemble classifier on average outperformed all of the homogeneous ensemble classifiers.

Average Coincident Failure Diversity				
Data set	*NB*	*SVM*	*RT*	*Het.*
abalone	0.1745	0.2238	0.1971	0.2475
bupa	0.5375	0.4897	0.6256	0.6253
dna	0.4748	0.8225	0.7337	0.9125
glass	0.4367	0.4849	0.6328	0.6324
ion	0.3880	0.6193	0.8546	0.8583
promoters	0.7709	0.8004	0.6433	0.8665
sonar	0.5125	0.6588	0.6921	0.7417
yeast	0.4015	0.3081	0.4917	0.5425

Table 4: Average disagreement diversity over ten experiments repeated with different random seeds using five fold cross validation and bagging.

Average Disagreement Diversity				
Data set	*NB*	*SVM*	*RT*	*Het.*
abalone	0.2346	0.2677	0.7589	0.7595
bupa	0.2647	0.1328	0.3873	0.4056
dna	0.0213	0.0828	0.3797	0.3777
glass	0.3134	0.3239	0.3930	0.4463
ion	0.0558	0.0661	0.1558	0.1656
promoters	0.1376	0.1230	0.4437	0.4423
sonar	0.1268	0.1818	0.3735	0.3724
yeast	0.1668	0.1229	0.5027	0.5019

Table 5: Average double fault diversity over ten experiments repeated with different random seeds using five fold cross validation and bagging. Lower measures correspond to more diverse models.

Average Double Fault Diversity				
Data set	*NB*	*SVM*	*RT*	*Het.*
abalone	0.5787	0.5475	0.1809	0.1799
bupa	0.2438	0.3469	0.18702	0.18697
dna	0.0509	0.0396	0.0612	0.0359
glass	0.3240	0.2754	0.1228	0.1229
ion	0.1454	0.0947	0.0505	0.0506
promoters	0.0725	0.0543	0.1413	0.0544
sonar	0.2555	0.1528	0.1280	0.1248
yeast	0.3310	0.3626	0.1844	0.1843

interactions prevents deeper understanding of how to make use of such information when creating ensemble classifiers.

The plots in Figures 1–3 help further illustrate several trends identified in the results. These plots are taken from three different data sets and each of them is a plot of a different homogeneous ensemble pairing. Although, it is not entirely clear what the relationships between accuracy and diversity are in these plots, they do suggest that the most successful heterogeneous ensemble classifiers for a particular data set generally involve a large proportion of RT base classifiers. To explain the success of the RT classifiers, it is helpful to note that while the disagreement measures are usually high (adding to the diversity) the double fault measure is usually relatively low. So the diversity that is created by RT classifiers, is created in a way that doesn't reduce total ensemble accuracy.

Both of the heterogeneous ensemble classifiers that involve RT base classifiers show that the best performance is usually achieved through including a larger number of RT base classifiers than SVM or NB base classifiers. The plots also show that there tends to be a wide range in the diversity values and the plots of the diversity measures are usually arching towards more diversity. The plot in Figure 2 is representative of heterogeneous ensemble classifiers that include NB and SVM base classifiers, in that they often show large improvement, in terms of accuracy, over what SVM or NB homogeneous ensemble classifiers demonstrate. However the pointed shape of the accuracy curve is unusual and demonstrates that the performance of the heterogeneous ensemble classifier can be very sensitive to the exact composition of its base classifiers.

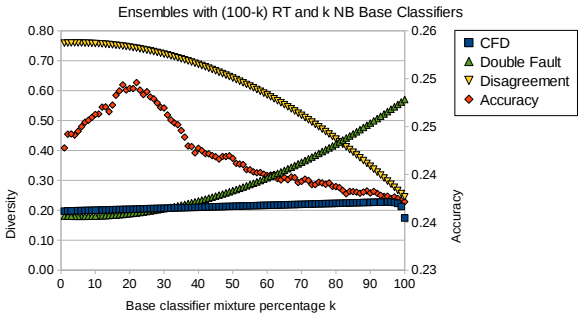

Figure 1: Comparison of homogeneous and heterogeneous ensemble classifiers using the *abalone* data set.

Each of the heterogeneous ensemble classifier experiments showed how two homogeneous ensemble classifiers of different types would perform when mixed to-

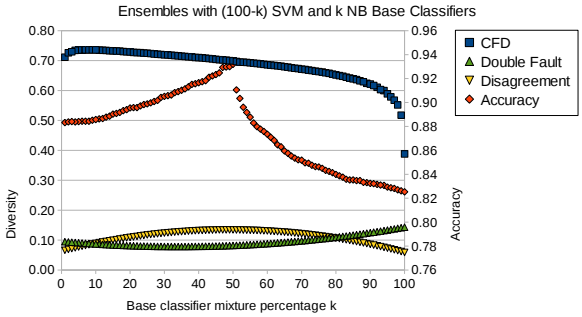

Figure 2: Comparison of homogeneous and heterogeneous ensemble classifiers using the *ion* data set.

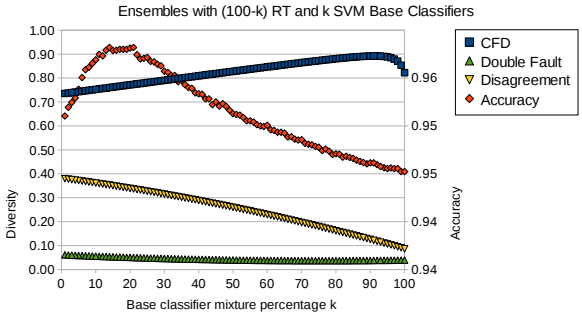

Figure 3: Comparison of homogeneous and heterogeneous ensemble classifiers using the *dna* data set.

gether with different levels of composition. The question we had when we saw the arcs in the heterogeneous ensemble classifier plots was: do the arcs represent extra performance or extra diversity that is caused by combining different types of base classifiers into one ensemble classifier, or are the arcs just a byproduct of combining base classifiers from ensemble classifiers with large differences in performance and diversity? One might assume that if two homogeneous ensemble classifiers consisting of base classifiers of the same type with different levels of accuracy were were blended together, that the plots of those measures would result in a straight line between the end points (or more of a straight line than exhibited by the plots of the heterogeneous ensemble mixtures). To test that, we used five fold cross validation ten times to create ensemble classifiers of bagged random trees. We used the same random seed for the cross validation each time, but used ten different seeds for the bagging procedure. For each of the five folds we selected the ensemble classifiers that had the best and worse performance based on one of the measures: accuracy, disagreement, double fault, CFD. We then gradually blended the base classifiers from the two ensembles together, and plotted how the performance changed as the composition of base classifiers changed. The plots showed that for most of the measures, the composition of two homogeneous ensembles of the same type lead to values closer the line than the values measured for the heterogeneous ensemble classifiers.

For each of the data sets, we calculated the average distance above the line, $a_{model,measure}$, and the maximum distance above the line, $m_{model,measure}$, for the composition of two RT ensemble classifiers, and for the different heterogeneous ensemble classifiers. We then calculated $a_{heter,measure} - a_{homo,measure}$ for each of the heterogeneous models and each of the measures. The average difference is important because it may not be practical to search for the best composition in which case an arbitrary heterogeneous composition will be chosen. In that case the differences in the average distances will represent the expected advantage of using a heterogeneous ensemble classifier. We used the Wilcoxon signed rank test to calculate p-values for the hypothesis:

$$H_a : Median(a_{het.,measure} - a_{hom.,measure}) > 0$$
$$H_0 : Median(a_{het.,measure} - a_{hom.,measure}) \leq 0$$

Where the median refers to the median value of the difference over all data sets. For this test to be relevant we must assume that the eight data sets used in this study are representative of the greater population of data sets. If the tests ends up lending support to the alternative hypothesis, then more times than not, a data set will be expected to have a larger measure value above the line for the heterogeneous model when the composition is randomly chosen, as compared to a homogeneous ensemble classifier that is a random mixture of the best and worst ensemble classifiers.

We also ran the same tests using only the compositions that had the maximum values for the measures, rather than the average over all compositions. This was done to test the potential for heterogeneous ensemble classifiers to create an advantage over homogeneous ensemble classifiers. The following were the hypothesis used in our Wilcoxon signed rank tests:

$$H_a : Median(m_{het.,measure} - m_{hom.,measure}) > 0$$
$$H_0 : Median(m_{het.,measure} - m_{hom.,measure}) \leq 0$$

Test that lend support to the alternative hypothesis indicate that when a random data set is chosen, more often than not, the best composition for the heterogeneous will be further from the line than the best composition for the homogeneous mixture for a data set

These results lend support to the theory that the extra performance and diversity are caused by the

Table 6: P-values for significance tests using Wilcoxon signed rank tests. Tests with low p-values imply that the interactions of base classifiers of different models, found in heterogeneous ensemble classifiers, lead to higher values of the corresponding measure than would be expected with homogeneous ensemble classifiers.

		p-value	
Measure	*Model*	*Average*	*Max*
Accuracy	NB_RT	1.000	0.926
	NB_SVM	0.027	0.875
	SVM_RT	0.629	0.727
Disagreement	NB_RT	0.004	0.004
	NB_SVM	0.004	0.004
	SVM_RT	0.004	0.004
Double Fault	NB_RT	0.012	1.000
	NB_SVM	0.004	1.000
	SVM_RT	0.055	1.000
CFD	NB_RT	0.004	0.004
	NB_SVM	0.004	0.004
	SVM_RT	0.004	0.004

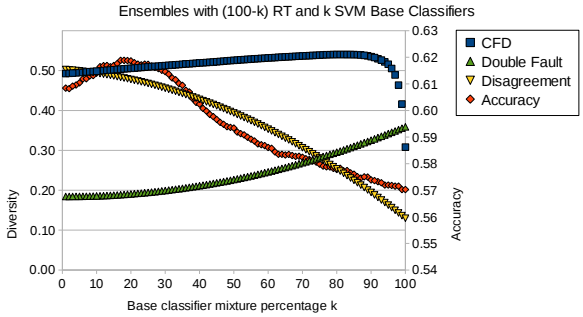

Figure 4: Heterogeneous ensemble classifiers using the *yeast* data set. Ensemble classifiers are composed of SVM and RT base classifiers. The left most value on the x-axis represents homogeneous RT ensemble classifiers, the right most value represents homogeneous SVM ensemble classifiers, and everything in between is representative of compositions of both SVM and RT base classifiers, and are therefore heterogeneous ensemble classifiers.

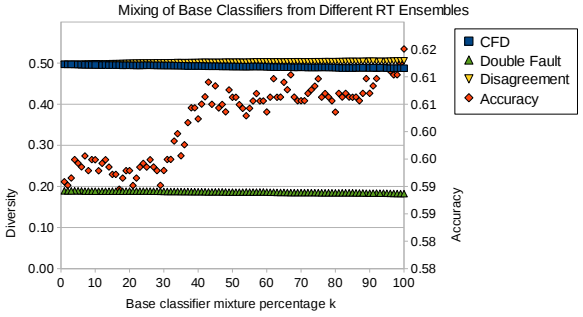

Figure 5: Homogeneous ensemble classifiers using the *yeast* data set. The ensemble classifiers are created using a mixture of of the best and worst performing ensembles for the given measure. The left most value on the x-axis represents the worst ensemble classifier for a measure, the right most value represents the best ensemble classifier, and everything in between is some combination of the two.

interactions in the heterogeneous ensemble classifier, for the tests where the p-values are small. These results only hold when the number of base classifiers for the ensemble classifiers is arbitrarily fixed (in this case 100).

One thing that was very clear from the plots from the above experiments was that heterogeneous ensemble models can have a much larger range of diversity than homogeneous ensemble models. The plots in Figures 4 and 5 can be compared to show the differences between heterogeneous mixtures and homogeneous mixtures, respectively, of base classifier sets. These plots are specifically taken from the experiments with the *yeast* data set but the trends found in these two graphs are representative of what is found in other data sets. What we see is that the range of accuracy in the heterogeneous and homogeneous mixtures are about the same. However the range in diversity measures for the heterogeneous mixtures are much larger than in the homogeneous mixtures. So even though the homogeneous ensemble classifiers show a wide range of accuracy measures, they show relatively little difference in the amount diversity they display. This may explain why the heterogeneous models are often able have better accuracy than homogeneous models. During training, homogeneous models search through a smaller space in terms of diversity.

4 Conclusions and Future Work

Our experiments showed that heterogeneous ensemble classifiers are capable of being more accurate and more diverse than homogeneous ensemble classifiers. The method we used performed a search over all possible heterogeneous model compositions involving only two base classifier models. In practice this may not be possible, but also may not be necessary using other fusion functions that re-weight the predictions generated by base classifiers. So in the future we intend to perform follow-up experiments using different fusion functions.

Another question that arose during our experimen-

tation was whether or not the diversity measures we used were adequate for for multiclass diversity problems, and whether or not they were appropriate for use with fusion functions that use prediction outputs other than just target class labels. If the fusion function is re-weighting the the outputs of the base classifier predictions, it seems logical that the diversity measures should re-weight the amount that each base classifier affects the diversity measure of the ensemble classifier. The tables in Appendix A show the amount of improvement over the value at the line for each of the ensemble types.

We would also like to run these test on more data sets to further support our conjecture that heterogeneous ensemble classifiers result in increased performance and diversity. We also need to extend these results to the case where the number of base classifiers is not fixed. Ideally, out of bag (OOB) error [4] could be used as a stopping criteria for the set of base classifiers for each ensemble classifier. We would then still be interested if performance and diversity could increase in heterogeneous ensemble classifiers, but also if heterogeneous ensemble classifiers of significantly smaller sizes could achieve the same performance and diversity as the best homogeneous ensembles classifier.

5 Acknowledgments

We would like to thank Philip Kegelmeyer for providing the data sets for our testing and for helpful suggestions throughout the project. We also thank the developers of the WEKA and JAMA libraries used in HEMLOCK.

References

[1] R. Banfield, L. Hall, K. Bowyer, and W. P. Kegelmeyer. A comparison of decision tree ensemble creation techniques. *IEEE Trans. Pat. Recog. Mach. Int.*, 29(1):173–180, 2007.

[2] J. Basilico, D. Dunlavy, S. Verzi, T. Bauer, and W. Shaneyfelt. Yucca mountain LSN archive assistant. Technical Report SAND2008-1622, Sandia National Laboratories, 2008.

[3] S. Bian and W. Wang. On diversity and accuracy of homogeneous and heterogeneous ensembles. *Intl. J. Hybrid Intel. Sys.*, 4:103–128, 2007.

[4] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[5] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[6] P. Cunningham and J. Carney. Diversity versus quality in classification ensembles based on feature selection. In *ECML '00: Proceedings of the 11th European Conference on Machine Learning*, pages 109–116, London, UK, 2000. Springer-Verlag.

[7] T. G. Dietterich. Ensemble methods in machine learning. In *Proc. International Workshop on Multiple Classifier Systems*, pages 1–15, 2000.

[8] J. Fleiss. *Statistical Methods for Rates and Proportions*. John Wiley and Sons, New York, 1981.

[9] G. Giancinto, F. Roli, and P. F. Roli. Design of effective neural network ensembles for image classification purposes. *Image Vision and Computing Journal*, 19:699–707, 2001.

[10] S. Gilpin and D. Dunlavy. HEMLOCK v1.0 user guide: A software package for creating and evaluating heterogeneous ensemble classification models. Technical report, Sandia National Laboratories, 2009. In preparation.

[11] L. Hansen and P. Salamon. Neural network ensembles. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(10):993–1001, Oct 1990.

[12] G. H. John and P. Langley. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, 1995.

[13] R. Kohavi and D. Wolpert. Bias plus variance for zero-one loss functions. In *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 275–283, 1996.

[14] W. Krzanowski and D. Partridge. Software diversity: Practical statistics for its measurement and exploitation. *Information & Software Technology*, 39:39–707, 1996.

[15] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[16] J. C. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods: Support Vector Learning*, pages 185–208. MIT Press, Cambridge, MA, USA, 1999.

[17] D. B. Skalak. The sources of increased accuracy for two proposed boosting algorithms. In *Proc. American Association for Artificial Intelligence (AAAI), Integrating Multiple Learned Models Workshop*, pages 120–125, 1996.

[18] W. Wang, D. Partridge, and J. Etherington. Hybrid ensembles and coincident failure diversity. In *Proc. International Joint Conference on Neural Networks*, 2001.

[19] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, June 2005.

[20] G. U. Yule. On the association of attributes in statistics, with examples from the material of the childhood society. *Proceedings of the Royal Society of London*, 66:22–23, 1899.

A Accuracy and Diversity Measure Results

Table 7: Maximum vertical displacement between disagreement measure between the heterogeneous ensemble classifier and a linear interpolation of the disagreements of two different homogeneous ensemble classifiers.

Accuracy				
Data	*NB_RT*	*NB_SVM*	*SVM_RT*	*RT_RT*
abalone	0.0351	0.0188	0.0408	0.0416
bupa	0.0334	0.0833	0.0460	0.0341
dna	0.0068	0.0110	0.0095	0.0058
glass	0.0695	0.0410	0.0501	0.0359
ion	0.0338	0.0838	0.0163	0.0028
promoters	0.0198	0.0161	0.0282	0.0293
sonar	0.0091	0.0307	0.0007	0.0306
yeast	0.0280	0.0125	0.0304	0.0127
Average	0.0294	0.0371	0.0278	0.0241

Disagreement				
Data	*NB_RT*	*NB_SVM*	*SVM_RT*	*RT_RT*
abalone	0.2344	0.1996	0.1996	0.0004
bupa	0.1692	0.4694	0.2471	0.0041
dna	0.2027	0.2567	0.1115	0.0098
glass	0.2010	0.2257	0.1359	0.0037
ion	0.3174	0.5458	0.1691	0.0116
promoters	0.1343	0.0580	0.1467	0.0050
sonar	0.2124	0.3205	0.0884	0.0013
yeast	0.1713	0.2596	0.2161	0.0035
Average	0.2053	0.2919	0.1643	0.0049

Double Fault				
Data	*NB_RT*	*NB_SVM*	*SVM_RT*	*RT_RT*
abalone	-0.0043	-0.0067	-0.0037	-0.0029
bupa	-0.0012	-0.0029	-0.0017	-0.0085
dna	-0.0006	-0.0003	-0.0004	-0.0095
glass	-0.0019	-0.0021	-0.0012	-0.0147
ion	-0.0009	-0.0015	-0.0004	-0.0411
promoters	-0.0006	-0.0001	-0.0006	-0.0239
sonar	-0.0012	-0.0014	-0.0003	-0.0223
yeast	-0.0015	-0.0013	-0.0018	-0.0011
Average	-0.0015	-0.0020	-0.0013	-0.0155

CFD				
Data	*NB_RT*	*NB_SVM*	*SVM_RT*	*RT_RT*
abalone	0.2298	0.1939	0.1010	0.0024
bupa	0.0275	0.1212	0.1440	0.0025
dna	0.4622	0.3400	0.0885	0.0016
glass	0.0355	0.0287	0.0370	0.0064
ion	0.4061	0.3397	0.1760	0.0074
promoters	0.0981	0.0087	0.0919	0.0038
sonar	0.2032	0.1098	0.1036	0.0056
yeast	0.2381	0.1658	0.3906	0.0004
Average	0.2126	0.1635	0.1416	0.0038

Table 8: Average vertical displacement between disagreement measure between the heterogeneous ensemble classifier and a linear interpolation of the disagreements of two different homogeneous ensemble classifiers.

Accuracy				
Data	*NB_RT*	*NB_SVM*	*SVM_RT*	*RT_RT*
abalone	0.0111	0.0043	0.0218	0.0061
bupa	-0.0047	0.0312	-0.0167	0.0109
dna	0.0016	0.0049	0.0041	0.0012
glass	0.0024	0.0163	0.0289	0.0146
ion	0.0099	0.0259	0.0028	-0.0002
promoters	0.0055	0.0081	0.0140	-0.0156
sonar	-0.0214	0.0061	-0.0071	0.0025
yeast	0.0067	0.0075	0.0099	-0.0016
Average	0.0014	0.0130	0.0072	0.0022

Disagreement				
Data	*NB_RT*	*NB_SVM*	*SVM_RT*	*RT_RT*
abalone	0.1634	0.3544	0.1385	-0.0002
bupa	0.1170	0.3531	0.1762	-0.0021
dna	0.1323	0.1790	0.0754	0.0052
glass	0.1415	0.1591	0.0943	-0.0022
ion	0.2276	0.4224	0.1172	-0.0026
promoters	0.0900	0.0379	0.0987	0.0001
sonar	0.1457	0.2314	0.0587	-0.0040
yeast	0.1179	0.1849	0.1485	0.0013
Average	0.1419	0.2403	0.1134	-0.0006

Double Fault				
Data	*NB_RT*	*NB_SVM*	*SVM_RT*	*RT_RT*
abalone	-0.0043	-0.0067	-0.0037	-0.0029
bupa	-0.0012	-0.0029	-0.0017	-0.0085
dna	-0.0006	-0.0003	-0.0004	-0.0095
glass	-0.0019	-0.0021	-0.0012	-0.0147
ion	-0.0009	-0.0015	-0.0004	-0.0411
promoters	-0.0006	-0.0001	-0.0006	-0.0239
sonar	-0.0012	-0.0014	-0.0003	-0.0223
yeast	-0.0015	-0.0013	-0.0018	-0.0011
Average	-0.0015	-0.0020	-0.0013	-0.0155

CFD				
Data	*NB_RT*	*NB_SVM*	*SVM_RT*	*RT_RT*
abalone	0.1258	0.1370	0.0567	0.0000
bupa	0.0164	0.0797	0.0799	0.0000
dna	0.2666	0.2004	0.0544	-0.0008
glass	0.0234	0.0196	0.0236	0.0014
ion	0.2335	0.2692	0.1074	0.0019
promoters	0.0595	0.0054	0.0558	-0.0036
sonar	0.1170	0.0853	0.0619	0.0022
yeast	0.1410	0.1228	0.2266	-0.0010
Average	0.1229	0.1149	0.0833	0.0000

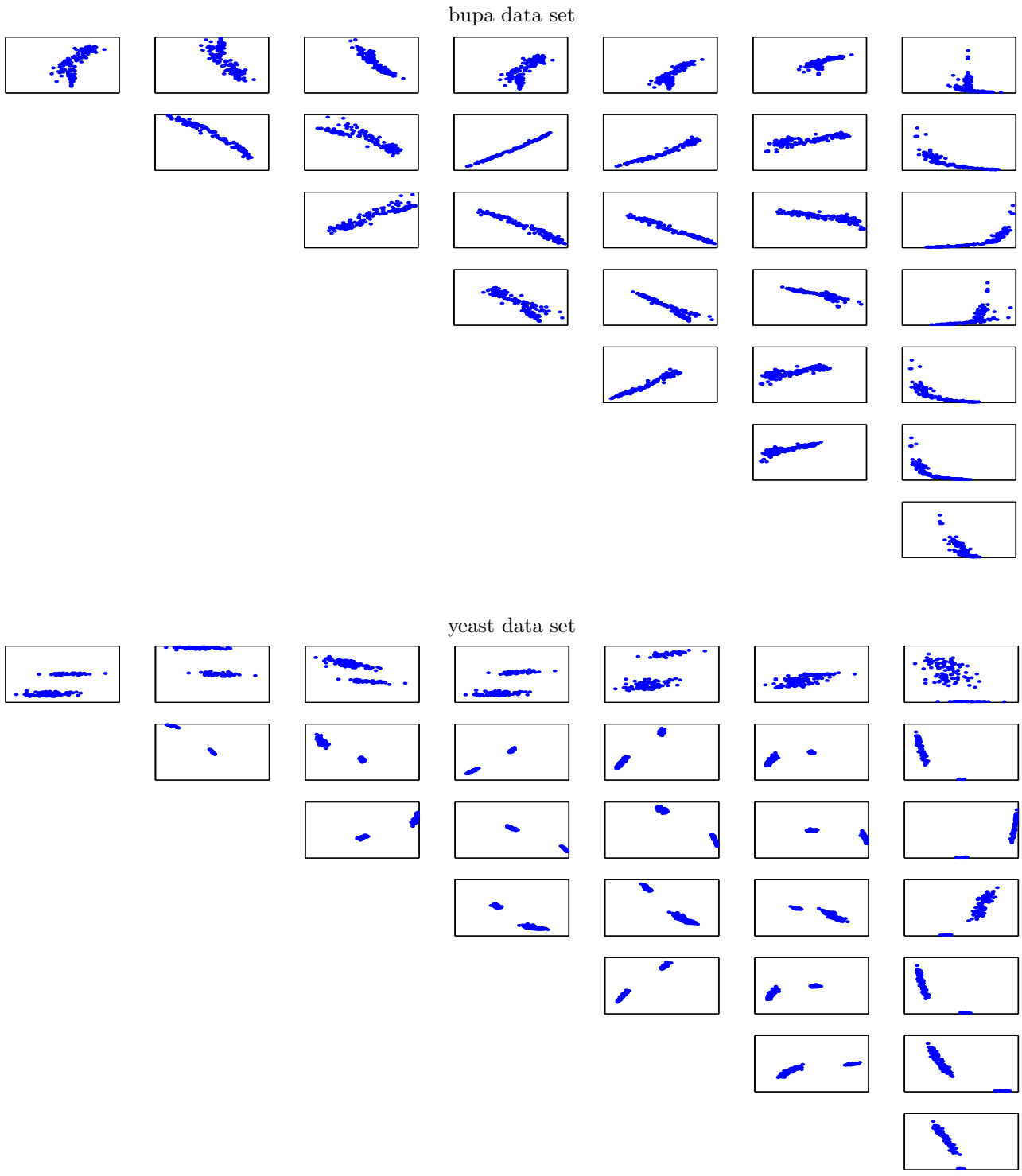

Figure 6: Matrix plots of pairwise comparisons of accuracy and diversity measures using SVM, RT, and NB classifiers for the *bupa* (top) and *yeast* (bottom) data sets. The rows of each plot correspond to 1) accuracy, 2) disagreement, 3) Yule's Q-statistic, 4) double fault, 5) entropy, 6) general diversity, 7) coincident failure, and 8) measure of difficulty as described in Section 2. The columns are ordered the same, but starting with disagreement. Subplot (i, j) in each matrix plot is a plot of the measure in row i (Y-axis) by column j (X-axis).