Diversity in Random Subspacing Ensembles

Alexey Tsymbal[1], Mykola Pechenizkiy[2], Pádraig Cunningham[1]

[1]Department of Computer Science, Trinity College Dublin, Ireland
{Alexey.Tsymbal, Padraig.Cunningham}@cs.tcd.ie
[2]Department of Computer Science and Information Systems,
University of Jyväskylä, Finland
mpechen@it.jyu.fi

Abstract. Ensembles of learnt models constitute one of the main current directions in machine learning and data mining. Ensembles allow us to achieve higher accuracy, which is often not achievable with single models. It was shown experimentally and theoretically that in order for an ensemble to be effective, it should consist of classifiers having diversity in their predictions. One technique, which proved to be effective for constructing an ensemble of diverse classifiers, is the use of different feature subsets as in random subspacing. A number of ways are known to quantify diversity in ensembles, but little research has been done about their appropriateness. In this paper, we compare eight measures of the ensemble diversity with regard to their correlation with the accuracy improvement due to ensembles. We conduct experiments on 21 data sets from the UCI machine learning repository, comparing the correlations for random subspacing ensembles with different ensemble sizes and with six different ensemble integration methods. Our experiments show that the greatest correlation of the accuracy improvement, on average, is with the disagreement, entropy, and ambiguity diversity measures, and the lowest correlation, surprisingly, is with the Q and double fault measures. Normally, the correlation decreases linearly as the ensemble size increases. Much higher correlation values can be seen with the dynamic integration methods, which are shown to better utilize the ensemble diversity than their static analogues.

1 Introduction

A popular method for creating an accurate classifier from a set of training data is to construct several classifiers, and then to combine their predictions. It was shown in many domains that an ensemble is often more accurate than any of the single classifiers in the ensemble. The integration of multiple classifiers, to improve classification results, is currently an active research area in the machine learning and neural networks communities. Dietterich [6] has presented the integration of multiple classifiers as one of the four most important directions in machine learning research. Sharkey [16] gives a good introduction to the area of ensembles and presents a survey of relevant work. While the focus of her paper is on neural networks, most remarks are relevant to any ensemble in general.

Both theoretical and empirical research have demonstrated that an ensemble is good if the base classifiers in it are both accurate and tend to err in different parts of

the instance space (that is have diversity in their predictions). Some studies on boosting [1,5] and random subspacing [9] show that integration of low-accuracy (also called "weak") classifiers can be effective as well. It was shown that the low accuracy of base classifiers in such ensembles is compensated for by the ensemble diversity.

Another important issue in creating an effective ensemble is the choice of the function for combining the predictions of the base classifiers. It was shown that increasing coverage of an ensemble through diversity is not enough to ensure increased prediction accuracy – if the integration method does not properly utilize the ensemble diversity, then no benefit arises from integrating multiple models [3].

One effective approach for generating an ensemble of diverse base classifiers is the use of different feature subsets, or so-called *ensemble feature selection* [13]. By varying the feature subsets used to generate the base classifiers, it is possible to promote diversity and produce base classifiers that tend to err in different sub-areas of the instance space. One efficient way to do ensemble feature selection is the random subspace method or random subspacing [9]. According to this method, the ensemble consists of classifiers constructed in randomly chosen subspaces, that is, classifiers constructed on randomly selected feature subsets.

Measuring diversity is not straightforward – there are a number of ways to measure diversity in ensembles of classifiers, and not much research has been done about the appropriateness and superiority of one measure over another.

In this paper, we consider different measures of the ensemble diversity, which could be used to measure the total ensemble diversity, as a general characteristic of ensemble goodness. The goal of this paper is to compare the considered measures of diversity in the context of random subspacing with different integration methods and with different ensemble sizes. In the existing literature, comparing different measures of the ensemble diversity is normally done by analyzing their correlation with various other ensemble characteristics. Such characteristics are the ensemble accuracy, the difference between the ensemble accuracy and the average base classifier accuracy, and the difference between the ensemble accuracy and the maximal base classifier accuracy [12,17]. In this paper, we compare eight measures of diversity with regard to their correlation with the accuracy improvement due to ensembles.

The paper is organized as follows. In Section 2 we consider the general task of constructing an effective ensemble, and review ensemble feature selection and random subspacing. In Section 3 we consider the question of integration of an ensemble of classifiers and review different integration methods. In Section 4 we present eight different measures for diversity in classification ensembles. In Section 5 we present our experiments with these measures and conclude in the next section with a summary and assessment of further research topics.

2 Ensemble Feature Selection and Random Subspacing

The task of using an ensemble of models can be broken down into two basic questions: (1) what set of learned models should be generated?; and (2) how should the predictions of the learned models be integrated? [6,16].

One way for building models with homogeneous representations, which proved to be effective, is the use of different subsets of features for each model. Finding a set of feature subsets for constructing an ensemble of diverse base models is also known as *ensemble feature selection* [13]. While traditional feature selection algorithms have the goal of finding the best feature subset that is germane to both the learning task and the selected learning algorithm, the task of ensemble feature selection has the additional goal of finding a set of feature subsets that will promote diversity among the base classifiers [13].

Ho [9] has shown that simple random selection of feature subsets may be an effective technique for ensemble feature selection because the lack of accuracy in the ensemble members is compensated for by their diversity. This technique is called the random subspace method or simply Random Subspacing (RS).

Instead of selecting a fixed number of features as in [9] (she used approximately half of the features for each base classifier) we use probabilistic feature selection in our implementation of RS. We consider all the features as having equal probability of being selected to the feature subset. This probability is selected randomly from the interval (0,1) before defining each feature subset. Thus, the initial feature subsets include different numbers of features. It was shown in experiments in [20] that this implementation of RS provides ensembles with greater diversity, and consequently, greater accuracy.

By constructing classifiers in random subspaces one may solve the small sample size problem, because the training sample size relatively increases. Thus, Ho [9] shows that while most other classification methods suffer from the curse of dimensionality, this method can be a good solution to solve this problem.

RS has much in common with bagging [19], but instead of sampling instances, one samples features. Like bagging, RS is a parallel learning algorithm, that is, the generation of each base classifier is independent. This makes it suitable for parallel implementation for fast learning that is desirable in some practical applications. It was shown that, like in bagging, the ensemble accuracy could be only increased with the addition of new members, even when the ensemble complexity grew [9].

3 Integration of an Ensemble of Models

Brodley and Lane [3] have shown that simply increasing coverage of an ensemble through diversity is not enough to insure increased prediction accuracy (coverage is defined there as the percentage of instances on which at least one base classifier is correct). If an integration method does not utilize coverage, then no benefit arises from integrating multiple classifiers. Thus, the ensemble diversity and coverage are not in themselves sufficient conditions for the ensemble accuracy. It is also important for ensemble accuracy to have a good integration method that will utilize the diversity of the base models.

The challenging problem of integration is to decide which one(s) of the classifiers to select or how to combine the results produced by the base classifiers. A number of *selection* and *combination* approaches have been proposed in the literature.

One of the most popular and simplest techniques used to combine the results of the base classifiers, is simple voting (also called majority voting and select all majority (SAM)) [1]. In the voting technique, the classification of each base classifier is considered as an equally weighted vote for that particular class value. The class value that receives the biggest number of votes is selected as the final classification (ties are solved arbitrarily). Weighted Voting (WV), where each vote receives a weight proportional to the estimated generalization performance of the corresponding classifier, works usually better than simple majority voting [1].

A number of selection techniques have also been proposed to solve the integration problem. One of the most popular and simplest selection techniques is Cross-Validation Majority (CVM, also called Single Best, we call it simply Static Selection, SS, in our experiments) [15]. In CVM, the cross-validation accuracy for each base classifier is estimated using the training set, and then the classifier with the highest accuracy is selected (ties are solved using voting).

The described above approaches are *static*. The select one "best" model for the whole data space or combine the models uniformly. In *dynamic* integration each new instance to be classified is taken into account. Usually, better results can be achieved if integration is dynamic.

We consider in our experiments three dynamic integration techniques based on the same local accuracy estimates: Dynamic Selection (DS) [14], Dynamic Voting (DV) [14], and Dynamic Voting with Selection (DVS) [21]. The three dynamic integration techniques contain two main phases [14,21]. First, at the learning phase, the local classification errors of each base classifier for each instance of the training set are estimated according to the 1/0 loss function using cross validation. The learning phase finishes with training the base classifiers on the whole training set. The application phase begins with determining *k*-nearest neighbourhood for a new instance using a distance metric. Then, weighted nearest neighbour regression is used to predict the local classification errors of each base classifier for the new instance.

Then, DS simply selects a classifier with the least predicted local classification error. In DV, each base classifier receives a weight that is proportional to the estimated local accuracy of the base classifier, and the final classification is produced by combining the votes of each classifier with their weights. In DVS, the base classifiers with highest local classification errors are discarded (the classifiers with errors that fall into the upper half of the error interval of the base classifiers) and locally weighted voting (DV) is applied to the remaining base classifiers.

4 Measures of the Ensemble Diversity

In this section we consider eight different measures of the ensemble diversity, six of which are pairwise as they are able to measure diversity in predictions of a pair of classifiers (plain disagreement, fail/non-fail disagreement, the double fault measure, the *Q* statistic, the correlation coefficient, and the kappa statistic). The total ensemble diversity is the average of the diversities of all the pairs of classifiers in the ensemble.

The two non-pairwise measures evaluate diversity in predictions of the whole ensemble (entropy and ambiguity).

The *plain disagreement* measure is probably the most commonly used measure for diversity in the ensembles of classifiers with crisp predictions. For example, in [9] it was used for measuring the diversity of decision forests, and its correlation with the forests' accuracy. In [20] it was used as a component of the fitness function guiding the process of ensemble construction. For two classifiers i and j, the plain disagreement is equal to the proportion of the instances on which the classifiers make different predictions:

$$div\_plain_{i,j} = \frac{1}{N}\sum_{k=1}^{N} \mathrm{Diff}(\mathrm{C}_i(\mathbf{x}_k), \mathrm{C}_j(\mathbf{x}_k)), \qquad (1)$$

where N is the number of instances in the data set, $\mathrm{C}_i(\mathbf{x}_k)$ is the class assigned by classifier i to instance k, and Diff(a,b)=0, if a=b, otherwise Diff(a,b)=1. The plain disagreement varies from 0 to 1. This measure is equal to 0, when the classifiers return the same classes for each instance, and it is equal to 1 when the predictions are always different.

The *fail/non-fail disagreement* was defined in [18] as the percentage of test instances for which the classifiers make different predictions but for which one of them is correct:

$$div\_dis_{i,j} = \frac{N^{01} + N^{10}}{N^{11} + N^{10} + N^{01} + N^{00}}, \qquad (2)$$

where N^{ab} is the number of instances in the data set, classified correctly (a=1) or incorrectly (a=0) by the classifier i, and correctly (b=1) or incorrectly (b=0) by the classifier j. The denominator in (2) is equal to the total number of instances N. (2) is equal to (1) for binary classification problems, where the number of classes is 2. It can be also shown that $div\_dis_{i,j} \le div\_plain_{i,j}$, as the instances contributing to this disagreement measure form a subset of instances contributing to the plain disagreement. The fail/non-fail disagreement varies from 0 to 1. This measure is equal to 0, when the classifiers return the same classes for each instance, or different but incorrect classes, and it is equal to 1 when the predictions are always different and one of them is correct.

The *Double Fault* measure (*DF*) [8] is the percentage of test instances for which both classifiers make wrong predictions:

$$div\_DF_{i,j} = \frac{N^{00}}{N^{11} + N^{10} + N^{01} + N^{00}}, \qquad (3)$$

where N^{ab} has the same meaning as in (2). In [12,17] *DF* was shown to have reasonable correlation with the Majority Voting and Naïve Bayes integration methods.

The following measure is based on Yule's *Q statistic* used to assess the similarity of two classifiers' outputs [12]:

$$div\_Q_{i,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}, \quad (4)$$

where N^{ab} has the same meaning as in (2) and (3). For statistically independent classifiers, the expected value of Q is 0. Q varies between –1 and 1. Classifiers that tend to recognize the same objects correctly will have positive values of Q, and those which commit errors on different objects will render Q negative [12]. In the case of undefined value with division by zero, we assume the diversity is minimal, equal to 1.

In [12], after comparative experiments on the UCI Breast cancer Wisconsin data set, the Phoneme recognition and the Cone-torus data sets, and two experiments with emulated ensembles (artificially generating possible cases of the base classifiers' outputs), Q was recommended as the best measure for the purposes of developing committees that minimize error, taking into account the experimental results, and especially its simplicity and comprehensibility (or the ease of interpretation).

One problem, which we have noticed with this measure in our pilot studies, was its insensitivity on small data sets. For a small number of instances N^{00} is often equal to 0. Q in this case is equal to –1 (maximal diversity) no matter how big the values of N^{01} and N^{10} are, which is not a good reflection of the true differences in classifiers' outputs.

The *correlation coefficient* between the outputs of two classifiers i and j can be measured as [12]:

$$div\_corr_{i,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{\sqrt{(N^{11} + N^{10})(N^{01} + N^{00})(N^{11} + N^{01})(N^{10} + N^{00})}}, \quad (5)$$

where N^{ab} have the same meaning as in (2), (3) and (4). The numerator in (5) is the same as in (4), and for any two classifiers i and j, $div\_corr_{i,j}$ and $div\_Q_{i,j}$ have the same sign, and it can be proven that $\left|div\_corr_{i,j}\right| \le \left|div\_Q_{i,j}\right|$ [12]. This measure, as well as the fail/non-fail disagreement, the DF measure, and the Q statistic were considered among the group of 10 measures in the comparative experiments in [12].

Let N_{ij} be the number of instances in the data set, recognized as class i by the first classifier and as class j by the second one, N_{i*} is the number of instances recognized as i by the first classifier, and N_{*i} is the number of instances recognized as i by the second classifier. Define then Θ_1 and Θ_2 as

$$\Theta_1 = \frac{\sum_{i=1}^{l} N_{ii}}{N}, \text{ and } \Theta_2 = \sum_{i=1}^{l} \left(\frac{N_{i*}}{N} \cdot \frac{N_{*i}}{N}\right), \quad (6)$$

where l is the number of classes and N is the total number of instances. Θ_1 estimates the probability that the two classifiers agree, and Θ_2 is a correction term for Θ_1, which estimates the probability that the two classifiers agree simply by chance (in the case where each classifier chooses to assign a class label randomly). The pairwise diversity $div\_kappa_{i,j}$ is then defined as follows [5]:

$$div\_kappa_{i,j} = \frac{\Theta_1 - \Theta_2}{1 - \Theta_2}. \quad (7)$$

Kappa is equal to 0 when the agreement of the two classifiers equals to that expected by chance, and *kappa* is equal to 1 when the two classifiers agree on every example. Negative values occur when agreement is less than expected by chance – that is, there is systematic disagreement between the classifiers [5]. *Kappa* is able to track negative correlations in a similar manner to *Q* and correlation.

Dietterich [5] used this measure in scatter plots called "κ-error diagrams", where kappa was plotted against mean accuracy of the classifier pair. κ-error diagrams are a useful tool for visualising ensembles.

A non-pairwise measure of diversity, associated with a conditional-entropy error measure, is based on the concept of *entropy* [4]:

$$div\_ent = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{l} -\frac{N_k^i}{S}\cdot \log\left(\frac{N_k^i}{S}\right), \quad (8)$$

where N is the number of instances in the data set, S is the number of base classifiers, l is the number of classes, and N_k^i is the number of base classifiers that assign instance i to class k. To keep this measure of diversity within the range [0,1] the logarithm should be taken to the base l.

This measure was evaluated on a medical prediction problem and was shown to predict the accuracy of the ensemble well [4]. It was also shown that the entropy measure of diversity has the added advantage that it models the change in diversity with the size of the ensemble.

The next non-pairwise measure of diversity is associated with the variance-based measure of diversity proposed for regression problems in [11], called *ambiguity*. This diversity has been proven to have a direct relation with the ensemble error and this motivated us to use an associated diversity measure for classification also. The classification task can be decomposed into l regression tasks, where l is the number of classes. The output in the regression tasks will be the class membership of the instance (binary output 0/1 in the case of crisp classification considered in this paper). The diversity of the classification ensemble can then be calculated as the average ambiguity over these pseudo-regression tasks for each of the instances:

$$div\_amb = \frac{1}{lN}\sum_{i=1}^{l}\sum_{j=1}^{N} Ambiguity_{i,j} = \frac{1}{lNS}\sum_{i=1}^{l}\sum_{j=1}^{N}\sum_{k=1}^{S}\left(\mathrm{Is}(\mathrm{C}_k(\mathbf{x}_j) = i) - \frac{N_i^j}{S}\right)^2, \quad (9)$$

where l is the number of classes, N is the number of instances, S is the number of base classifiers, N_i^j is the number of base classifiers that assign instance j to class i, $\mathrm{C}_k(\mathbf{x}_j)$ is the class assigned by classifier k to instance j, and Is() is a truth predicate.

In our experiments we normalize all the measures to vary from 0 to 1, where 1 corresponds to the maximum of diversity for the sake of simplicity and to avoid the unnecessary complication in understanding the results of correlations with different signs.

5 Experimental Investigations

The experiments are conducted on 21 data sets taken from the UCI machine learning repository [2]. These data sets include real-world and synthetic problems, vary in characteristics, and were previously investigated by other researchers.

The main characteristics of the 21 data sets are presented in Table 1. The table includes the name of a data set, the number of instances included in the data set, the number of different classes of instances in the data set, and the numbers of different kinds of features included in the instances of the data set.

Table 1. Data sets and their characteristics

Data set	Instances	Classes	Features	
			Categorical	Numerical
Balance	625	3	0	4
Breast Cancer Ljubljana	286	2	9	0
Car	1728	4	6	0
Pima Indians Diabetes	768	2	0	8
Glass Recognition	214	6	0	9
Heart Disease	270	2	0	13
Ionosphere	351	2	0	34
Iris Plants	150	3	0	4
LED	300	10	7	0
LED17	300	10	24	0
Liver Disorders	345	2	0	6
Lymphography	148	4	15	3
MONK-1	432	2	6	0
MONK-2	432	2	6	0
MONK-3	432	2	6	0
Soybean	47	4	0	35
Thyroid	215	3	0	5
Tic-Tac-Toe Endgame	958	2	9	0
Vehicle	846	4	0	18
Voting	435	2	16	0
Zoo	101	7	16	0

For our experiments, we used an updated version of the experimental setting presented in [20] to test the EFS_SBC algorithm (Ensemble Feature Selection with the Simple Bayesian Classification). We extended it with an implementation of seven new measures of diversity besides the existing plain disagreement.

We used Simple Bayes (SB) as the base classifier in the ensembles. It has been recently shown experimentally and theoretically that SB can be optimal even when the "naïve" feature-independence assumption is violated by a wide margin [7]. Second, when SB is applied to the sub-problems of lower dimensionalities as in random subspacing, the error bias of the Bayesian probability estimates caused by the feature-independence assumption becomes smaller. It also can easily handle missing feature values of a learning instance allowing the other feature values still to contribute. Be-

sides, it has advantages in terms of simplicity, learning speed, classification speed, and storage space. It was shown [20] that only one "global" table of Bayesian probabilities is needed for the whole ensemble when SB is employed in ensemble feature selection (for each feature of the base classifiers the corresponding probabilities from this table are simply taken). We believe that dependencies and conclusions presented in this paper do not depend on the learning algorithm used and would be similar for most known learning algorithms.

To estimate the ensemble performance with random subspacing, we have used 70 test runs of stratified random-sampling cross validation with 70 percent of instances in the training sets. We experimented with five different ensemble sizes: 5, 10, 25, 50, and 100. At each run of the algorithm, we collect accuracies for the six types of ensemble integration: Static Selection (SS), Majority Voting (V), Weighted Voting (WV), Dynamic Selection (DS), Dynamic Voting (DV), and Dynamic Voting with Selection (DVS). In the dynamic integration strategies DS, DV, and DVS, the number of nearest neighbors (k) for the local accuracy estimates was pre-selected from the set of seven values: 1, 3, 7, 15, 31, 63, 127 ($2^n - 1$, $n = 1,...,7$), for each data set separately, if the number of instances in the training set permitted. Heterogeneous Euclidean-Overlap Metric (HEOM) [14] was used for calculation of the distances (for numeric features, the distance is calculated using the Euclidean metric, and for categorical features the simple 0/1 overlap metric is used).

The test environment was implemented within the MLC++ framework (the machine learning library in C++) [10]. A multiplicative factor of 1 was used for the Laplace correction in SB as in [7]. Numeric features were discretized into ten equal-length intervals (or one per observed value, whichever was less), as it was done in [7]. Although this approach was found to be slightly less accurate than more sophisticated ones, it has the advantage of simplicity, and is sufficient for comparing different ensembles of SB classifiers with each other, and with the "global" SB classifier. The use of more sophisticated discretization approaches could lead to better classification accuracies for the base classifiers and ensembles, but should not influence the main findings and conclusions presented in the paper.

In Figure 1 the correlations between diversity and improvement in the classification accuracy due to ensembles for the eight diversities and five ensemble sizes averaged over the data sets and integration methods are shown (Pearson's correlation coefficient r is used). It can be seen from the picture that the highest correlation is with *div_dis*. *Div_plain*, *div_ent*, and *div_kappa* are very close to the best *div_dis* (the difference is at most 0.03 for each ensemble size). The lowest correlation values are with *div_corr*, *div_Q*, and *div_DF*. Surprisingly, the worst correlations are with the *div_Q* and especially *div_DF* measures. As could be seen from the results, *Div_Q* and *div_corr* behave in a similar way, which reflects the similarity in their formulae. Another interesting finding is that the correlations decrease approximately linearly with the increase in the ensemble size. The best correlations are shown for 5 base classifiers. This is not true for the *div_DF* measure, where there is no clear pattern in the change (probably because the correlation is not significant for this measure).

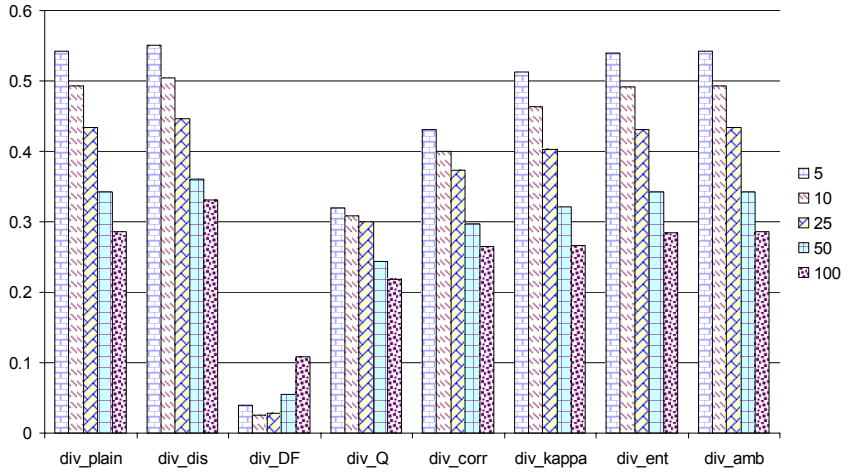

Fig. 1. The correlations for the eight diversities and five ensemble sizes averaged over the data sets and integration methods

In Figure 2 the correlations between diversity and improvement in the classification accuracy due to ensembles for the eight diversities and six integration methods averaged over the data sets and ensemble sizes are shown using Pearson's *r* correlation coefficient. The ranking of the diversities is the same as previously reported (in order of goodness): *div_dis*, *div_plain*, *div_amb*, *div_ent*, *div_kappa*, *div_corr*, *div_Q*, and *div_DF*. We also noticed that the correlation values are almost the same for *div_plain* and *div_amb*. The difference was at most 0.001, probably due to rounding in the computations. Supposedly, this similarity can be explained theoretically. The correlations differ significantly with the six integration methods. Always the dynamic methods (DS, DV, and DVS) have better correlations than the static ones (SS, V, and WV). Normally WV has better correlations than the other two static methods (SS and V). We believe that these differences can be explained by the fact that the dynamic methods make better use of diversity than the static methods, and WV makes better use of diversity than SS and V. These dependencies do not hold true for *div_DF* again, because of the same reason of low correlations.

To check the presented dependencies we recalculated the correlations using Spearman's rank correlation coefficient (*RCC*) as suggested in [12]. All the trends remained the same, and the difference in the averaged values was at most 0.01 and in the particular correlation values - at most 0.05.

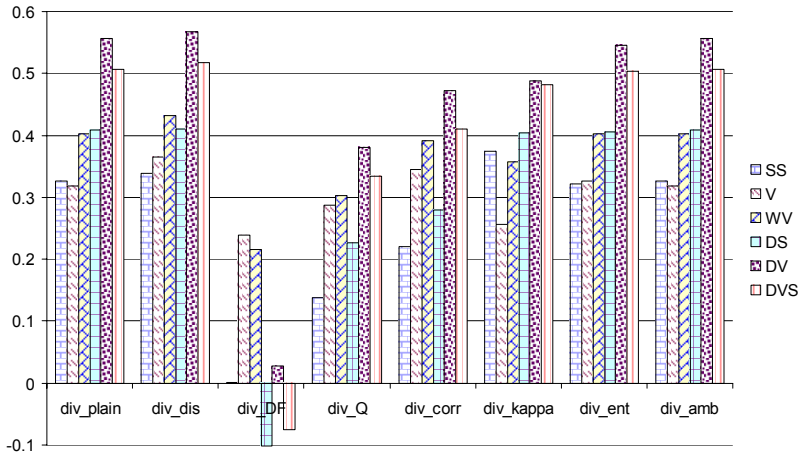

Fig. 2. The correlations for the eight diversities and six integration methods averaged over the data sets and ensemble sizes

To validate the findings and conclusions presented before and to check the dependency of the results on the selection of the data sets, we divided all the data sets into two groups in the following two ways: (1) with greater than the average improvement due to ensembles (10 data sets), and with less than or equal to the average improvement (11 data sets); and (2) with less than 9 features (10 data sets), and with greater or equal to 9 features (11 data sets); and checked all the dependencies for these groups.

The results for these groups supported our previously reported findings in this paper (the ranking of the diversities, the correlation decrease with the ensemble size's increase, and the ranking of the integration methods). Expectedly, the correlations for the group with better improvements were greater than for the other group (by up to 0.15 on average). Unexpectedly, greater correlations (by up to 0.15 on average) were for the group with larger amounts of features than for the group with fewer features. This needs further research.

We noticed also interesting behaviour with the selected k-neighbourhood values for dynamic integration. DS needs higher values of k. This can be explained by the fact that its prediction is based on only one classifier being selected, and thus, it is very unstable. Higher values of k provide more stability to DS. The average selected k is equal to 33 for DS, and it is only 14 for DV. For DVS, as a hybrid strategy, it is in between at 24 (for the ensemble size of 5). The selected values of k do not change significantly with the change of the ensemble size. The only change noticed is that DS with more ensemble members needs even greater k (up to 43 for 100 ensemble members).

6 Conclusions

In our paper, we have considered eight ensemble diversity metrics, six of which are pairwise measures (the plain disagreement, *div_plain*; the fail/non-fail disagreement, *div_dis*; the Double Fault measure, *div_DF*; the *Q* statistic, *div_Q*; the correlation coefficient, *div_corr*; and the *kappa* statistic, *div_kappa*), and two are non-pairwise measures (entropy, *div_ent*; and ambiguity, *div_amb*). To integrate the base classifiers generated with random subspacing, we used six integration methods: Static Selection (SS), Majority Voting (V), Weighted Voting (WV), Dynamic Selection (DS), Dynamic Voting (DV), and Dynamic Voting with Selection (DVS). We considered five ensemble sizes: 5, 10, 25, 50, and 100.

In our experiments, to check the goodness of each measure of diversity, we calculated its correlation with the improvement in the classification accuracy due to ensembles. The best correlations were shown by *div_plain*, *div_dis*, *div_ent*, and *div_amb*. *Div_Q* and *div_corr* behaved in a similar way, supported by the similarity of their formulae. Surprisingly, *div_DF* and *div_Q* had the worst average correlation. The correlation coefficients for *div_amb* were almost the same as for *div_plain*. All the correlations changed with the change of the integration method, showing the different use of diversity by the integration methods. The best correlations were shown with DV. The correlations decreased almost linearly with the increase in the ensemble size. The best correlations are for the ensemble size 5.

It would be interesting to check the presented dependencies and conclusions in other contexts in the future. For example, other ensemble generation strategies and integration methods can be tried.

Acknowledgments: This material is based upon works supported by the Science Foundation Ireland under Grant No. S.F.I.-02IN.1I111. This research is partly supported by the COMAS Graduate School of the University of Jyväskylä, Finland. We would like to thank the UCI machine learning repository of databases, domain theories and data generators for the data sets, and the MLC++ library for the source code used in this study.

References

1. Bauer E., R. Kohavi, An empirical comparison of voting classification algorithms: bagging, boosting, and variants, Machine Learning, 36 (1,2) (1999) 105-139.
2. Blake C.L., E. Keogh, C.J. Merz, UCI repository of machine learning databases [http:// www.ics.uci.edu/ ~mlearn/ MLRepository.html], Dept. of Information and Computer Science, University of California, Irvine, CA, 1999.
3. Brodley C., T. Lane, Creating and exploiting coverage and diversity, in: Proc. AAAI-96 Workshop on Integrating Multiple Learned Models, Portland, OR, 1996, pp. 8-14.
4. Cunningham P., J. Carney, Diversity versus quality in classification ensembles based on feature selection, in: R.L. deMántaras, E. Plaza (eds.), Proc. ECML 2000 11th European Conf. On Machine Learning, Barcelona, Spain, LNCS 1810, Springer, 2000, pp. 109-116.

5. Dietterich T.G., An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, Machine Learning 40 (2) (2000) 139-157.
6. Dietterich T.G., Machine learning research: four current directions, AI Magazine 18(4) (1997) 97-136.
7. Domingos P., M. Pazzani, On the optimality of the simple Bayesian classifier under zero-one loss, Machine Learning, 29 (2,3) (1997) 103-130.
8. Giacinto G., F. Roli. Design of effective neural network ensembles for image classification processes. *Image Vision and Computing Journal*, 19(9-10):699–707, 2001.
9. Ho T.K., The random subspace method for constructing decision forests, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20 (8) (1998) 832-844.
10. Kohavi R., D. Sommerfield, J. Dougherty, Data mining using MLC++: a machine learning library in C++, Tools with Artificial Intelligence, IEEE CS Press (1996) 234-245.
11. Krogh A., J. Vedelsby, Neural network ensembles, cross validation, and active learning, In: D. Touretzky, T. Leen (Eds.), Advances in Neural Information Processing Systems, Vol. 7, Cambridge, MA, MIT Press, 1995, pp. 231-238.
12. Kuncheva L.I., C.J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Machine Learning 51 (2) (2003) 181-207.
13. Opitz D., Feature selection for ensembles, in: Proc. 16th National Conf. on Artificial Intelligence, AAAI Press, 1999, pp. 379-384.
14. Puuronen S., V. Terziyan, A. Tsymbal, A dynamic integration algorithm for an ensemble of classifiers, in: Z.W. Ras, A. Skowron (eds.), Foundations of Intelligent Systems: 11th Int. Symp. ISMIS'99, Warsaw, Poland, LNAI 1609, Springer, 1999, pp. 592-600.
15. Schaffer C., Selecting a classification method by cross-validation, Machine Learning 13 (1993) 135-143.
16. Sharkey A.J.C., On combining artificial neural nets, Connection Science, Special Issue on Combining Artificial Neural Networks: Ensemble Approaches 8 (3,4) (1996) 299-314.
17. Shipp C.A., L.I. Kuncheva, Relationship between combination methods and measures of diversity in combining classifiers, Information Fusion 3 (2002) 135-148.
18. Skalak D.B., The sources of increased accuracy for two proposed boosting algorithms, in: AAAI-96 Workshop on Integrating Multiple Models for Improving and Scaling Machine Learning Algorithms (in conjunction with AAAI-96), Portland, Oregon, USA, 1996, pp. 120-125.
19. Skurichina M., R.P.W. Duin, Bagging and the random subspace method for redundant feature spaces, in: J. Kittler, F. Roli (Eds.), Proc. 2nd Int. Workshop on Multiple Classifier Systems MCS 2001, Cambridge, UK, 2001, pp. 1-10.
20. Tsymbal A., S. Puuronen, D. Patterson, Ensemble feature selection with the simple Bayesian classification, *Information Fusion*, Elsevier Science 4 (2) (2003) 87-100.
21. Tsymbal A., S. Puuronen, I. Skrypnyk, Ensemble feature selection with dynamic integration of classifiers, in: Int. ICSC Congress on Computational Intelligence Methods and Applications CIMA'2001, Bangor, Wales, U.K, 2001, pp. 558-564.