That Elusive Diversity in Classifier Ensembles

Ludmila I. Kuncheva

School of Informatics, University of Wales, Bangor
Bangor, Gwynedd, LL57 1UT, United Kingdom
l.i.kuncheva@bangor.ac.uk

Abstract. Is "useful diversity" a myth? Many experiments and the little available theory on diversity in classifier ensembles are either inconclusive, too heavily assumption-bound or openly non-supportive of the intuition that diverse classifiers fare better than non-divers ones. Although a rough general tendency was confirmed in our previous studies, no prominent link appeared between diversity of the ensemble and its accuracy. Diversity alone is a poor predictor of the ensemble accuracy. But there is no agreed definition of diversity to start with! Can we borrow a concept of diversity from biology? How can diversity, as far as we can define and measure it, be used to improve the ensemble? Here we argue that even without a clear-cut definition and theory behind it, studying diversity may prompt viable heuristic solutions. We look into some ways in which diversity can be used in analyzing, selecting or training the ensemble.

1 Introduction

Classifier outputs are combined in an attempt to reach a more accurate decision than that of a carefully designed individual classifier. It is curious that the experts in the field hold diametrically opposite views about our current level of understanding of combining classifiers. In his invited lecture at the 3rd International Workshop on Multiple Classifier Systems, 2002, Ghosh proposes that [5]

> "... our current understanding of ensemble-type multiclassifier systems is now quite mature..."

In an invited book chapter, the same year, Ho states that [8]

> "Many of the above questions are there because we do not yet have a scientific understanding of the classifier combination mechanisms."

Ho proceeds to nominate the stochastic discrimination theory by Kleinberg [9] as the only consistent and theoretically sound explanation of the success of classifier ensembles, criticizing other theories as being incomplete and assumption-dependent. However, as the usual practice invariably shows, ingenious heuristic developments are the heart, the soul and the engine in many branches of science and research.

This study advocates one such idea: that of measuring diversity and incorporating it into the process of building of the ensemble. We draw upon the somewhat futile efforts hitherto to define, measure and use diversity (our own research in this number!). We are cautious to note that no strong claims are made based on the small experimentation study reported here. The message of this paper is that there is still much room for heuristic in classifier combination, and diversity might be one of the lines for further exploration.

The paper is organized as follows. Section 2 explains diversity and its reincarnations. Section 3 looks into some ways in which diversity has been used in classifier ensembles. In Section 4, an ensemble building version of AdaBoost is proposed, which involves diversity and Section 5 concludes the paper.

2 Diversity

Classifiers in an ensemble should be different from each other, otherwise there is no gain in combining them. Quantifying this difference, named also diversity, orthogonality, complementarity, has been identified as an important research direction by many authors [2, 11, 14, 15, 20]. Measures of the connection between two classifier outputs can be derived from the statistical literature (e.g., [23]). There is less clarity on the subject when three or more classifiers are concerned. There is no strict definition of what is intuitively perceived as diversity. At least not in the vocabulary of machine learning, pattern recognition and computer science in general. Biologists and ecologists have axiomatized their idea of diversity several decades ago. For example, suppose that we are interested in the height of adult gorillas in a certain region of Africa. Consider a population π with a probability measure P associated with it. The measure P defines the distribution of heights for the population. A comprehensive study on diversity in life sciences by Rao [18] gives the following axiomatic definition of a diversity measure.

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space, and let $\mathcal{P}$ be a convex set of probability measures defined on it.[1] A function $H(.)$ mapping $\mathcal{P}$ onto the real line is said to be a **measure of diversity** if it satisfies the following conditions

C1: $H(P) \geq 0$, for any $P \in \mathcal{P}$ and $H(P) = 0$ iff P is degenerate.
C2: H is a concave function of P. [2]

The concavity condition ensures that any mixture of two populations has a higher diversity than the average of the two individual diversities. $H(P_i)$ is the diversity within a population π_i characterized by the probability measure P_i. Rao defines $H(P_i)$ to be the averaged **difference** ($\zeta(X_2, X_2)$) between two randomly picked individuals in the population π_i according to the probability measure P_i

[1] Convexity means that for any $P_1, P_2 \in \mathcal{P}$, and for any $t \in [0,1]$, $tP_1 + (1-t)P_2 \in \mathcal{P}$.

[2] The concavity here means that for any $P_1, P_2 \in \mathcal{P}$, and for any $t \in [0,1]$, $H(tP_1 + (1-t)P_2) \geq tH(P_1) + (1-t)H(P_2)$

$$H(P_i) = \int \zeta(X_1, X_2) P_i(\partial X_1) P_i(\partial X_2). \tag{1}$$

If the two individuals are drawn from two different populations π_i and π_j, then the total diversity will be

$$H(P_i, P_j) = \int \zeta(X_1, X_2) P_i(\partial X_1) P_j(\partial X_2). \tag{2}$$

The **dissimilarity** between the two populations π_i and π_j is then

$$D_{ij} = H(P_i, P_j) - \frac{1}{2}(H(P_i) + H(P_j)). \tag{3}$$

The concavity of H guarantees that D_{ij} will be positive for any two populations and their probability measures. This dissimilarity is based on taking out the diversity coming from each population and leaving only the "pure" diversity due to mixing the two populations.

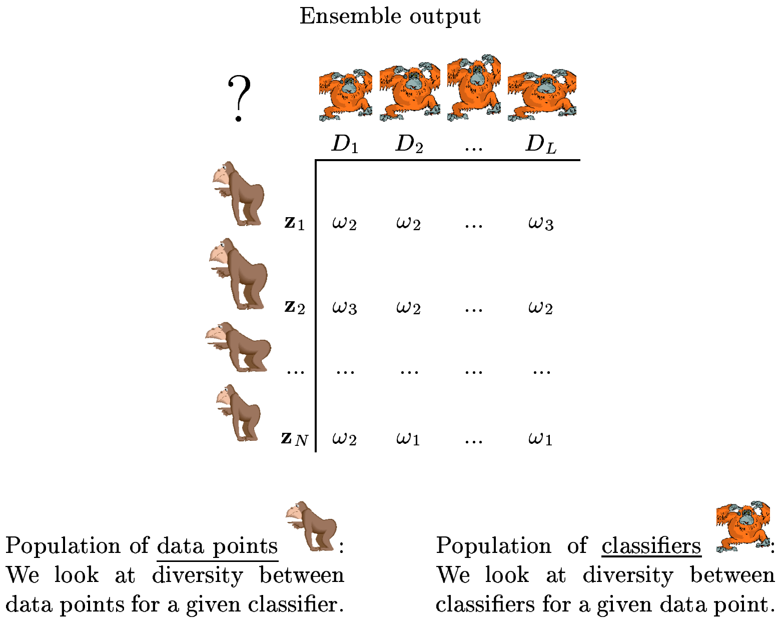
The distance ζ could be any function that satisfies the axioms for distance (nonnegativity, symmetry and a version of the triangle inequality). We can use the Euclidean distance for quantitative variables and a "matching" type of function for qualitative variables, i.e., $\zeta(X_1, X_2) = 1$ if the two variables have different values and 0, otherwise.

The most useful ideas often drift across sciences and branches thereof. How otherwise would neural networks and evolutionary computation become the powerful algorithmic tools they currently are? Many more algorithms have come about as a mathematical allegory for the underlying biological or physical processes. The question is how can we translate the notion of diversity used successfully in biology, ecology, economics, etc., into the mathematical concept needed in our classifier combining niche?

Our problem can be approached from two different angles as shown in Figure 1, depending on what we decide to be our "gorillas". The variable of interest here is the class label taking values in the set $\Omega = \{\omega_1, \ldots, \omega_c\}$. We suppose we have a data set $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ on which the L classifiers in the ensemble, $\mathcal{D} = \{D_1, \ldots, D_L\}$, are tested. Each classifier suggests a class label for every data point $\mathbf{z}_j$. Thus a *population* will be a collection of objects (classifiers or data points?) with the respective values of the class label.

We can regard the classifier outputs for a given data point $\mathbf{z}_j$ as a population. The diversity within the population will be the diversity of the ensemble with respect to the particular point in the feature space. The within-population diversity $H(P)$ can be measured by the entropy of the distribution of class labels among the classifiers or by the Gini index. Let P_k be the probability that a randomly chosen member of the population outputs label ω_k ($\sum_{k=1}^{c} P_k = 1$). Then the Gini diversity within a population of L classifiers is

$$H(P) = G = 1 - \sum_{k=1}^{c} P_k^2. \tag{4}$$



	D_1	D_2	...	D_L
$\mathbf{z}_1$	ω_2	ω_2	...	ω_3
$\mathbf{z}_2$	ω_3	ω_2	...	ω_2
...	...	...	...	...
$\mathbf{z}_N$	ω_2	ω_1	...	ω_1

Population of data points : We look at diversity between data points for a given classifier.

Population of classifiers : We look at diversity between classifiers for a given data point.

Fig. 1. Two perspectives on diversity in classifier ensembles

This concept underlies the variance component of the error suggested by Kohavi and Wolpert [10] and is also suggested as a measure of diversity within population by Rao [18]. The data set consists of N such populations, one for each data point. Therefore, the average diversity across the whole feature space is calculated as the average G over the data set $\mathbf{Z}$.

The alternative view is to consider the data points as the elements of the population and the classifier as the environment responsible for the distribution of the class labels. In this case, the within-population diversity is not of much use to us; we are interested in the diversity between populations, i.e., between classifiers. Most often we calculate some pairwise measure of diversity and average it across all pairs to get a value for the whole ensemble.

An immediate equivalent of the total diversity $H(P_i, P_j)$, assuming that π_i and π_j are two populations produced by classifiers D_i and D_j is the *measure of disagreement* Dis [7, 13, 22]. We consider the oracle type of outputs from classifiers D_i and D_j, i.e., for every object in the data set, the classifier is either correct (output 1) or wrong (output 0). Then the populations of interest consist of 0's and 1's. We do not assume that the new distribution is simply a mixture of the two distributions. Instead we consider a new space with 4 elements: 00, 01, 10, and 11. Denote the probabilities for these joint outputs of D_1 (first bit) and D_2 (second bit) as follows $Pr(11) = a; Pr(10) = b; Pr(01) = c$ and $Pr(00) = d$. The typical choice for the distance as mentioned before is $\zeta(m, n) = 1$, iff $m \neq n$, and 0, otherwise. Then

$$H(P_i, P_j) = \zeta(1,1) \times a + \zeta(1,0) \times b + \zeta(0,1) \times c + \zeta(0,0) \times d = b + c = Dis. \quad (5)$$

This is the expectation of the disagreement between classifiers D_i and D_j in the space of their joint oracle outputs.

However, the disagreement measure does not take out the individual diversities of π_i and π_j as does D_{ij} in (3). An analogue (in spirit) of the dissimilarity measure would be the **kappa** statistic, κ. It measures the agreement between two categorical variables while correcting for chance [4]. For c class labels, κ is defined on the $c \times c$ coincidence matrix M of the two classifiers. The entry $m_{k,s}$ of M is the proportion of the data set (used currently for testing of both D_i and D_j) which D_i labels as ω_k and D_j labels as ω_s. The agreement between D_i and D_j is given by

$$\kappa = \frac{\sum_k m_{kk} - \text{ ABC}}{1 - \text{ ABC}}, \quad (6)$$

where $\sum_k m_{kk}$ is the observed agreement between the classifiers and 'ABC' is "agreement-by-chance"

$$\text{ABC } = \sum_k \left(\sum_s m_{k,s}\right) \left(\sum_s m_{s,k}\right). \quad (7)$$

Low values of κ signify higher disagreement and hence higher diversity. If calculated on the 2×2 joined oracle output space,

$$\kappa = \frac{2(ac - bd)}{(a+b)(c+d) + (a+c)(b+d)}, \quad (8)$$

The bad news is that despite the large number of proposed measures and formalizations, there is no consensus on what diversity of a classifier ensemble is, which approach should be used to measure it (gorillas = classifiers or gorillas = data points) and what is a good measure of diversity. We will leave this question unanswered here, just acknowledging the "diversity of diversity", and will abstain from strongly advocating one measure or definition over another. In our previous studies we sightly favored the Q statistic (for oracle outputs) [13] because of its: (a) potential sensitivity to small disagreements; (b) value 0 indicating statistical independence; and (c) the relatively small effect of the individual accuracies on the possible range of values of Q. In the rest of this study we draw upon the existing literature and in particular kappa-error plots proposed by Margineantu and Dietterich [16], hence out choice of κ.

3 Using diversity

The general anticipation is that diversity measures will be helpful in designing the individual classifiers, the ensemble, and the combination method. For this to be possible, there should be a relationship between diversity and the ensemble performance. However, the results from our experiments so far have been disheartening, to say the least [13, 21]. We did not find the desired strong and consistent relationship to guide us into building better ensembles. Although the suspected relationship appears on a large scale, i.e., when diversity spans (uniformly) the whole range of possible values, in practice we are faced with a different picture. Usually the candidates for the ensemble are not very different from one another. This leads to small variations of diversity and also small variations of the accuracy of the ensemble about the individual accuracies. Unfortunately, none of the various diversity measures that we investigated previously (10 measures: 4 pairwise and 6 non-pairwise [13]) appeared to be sensitive enough to detect the changes in the accuracy. This phenomenon is illustrated in Figure 2 showing a typical graph of "accuracy" versus "diversity". A scatterplot of this type was obtained when we simulated classifier outputs with preassigned accuracy (approximately equal for all ensemble members) and preassigned diversity (approximately equal diversity for all pairs). The relationship can easily be spotted on the plot. However, when diversity only varies in a small range, this relationship is blurred (the gray dot and the cloud of classifiers in it).

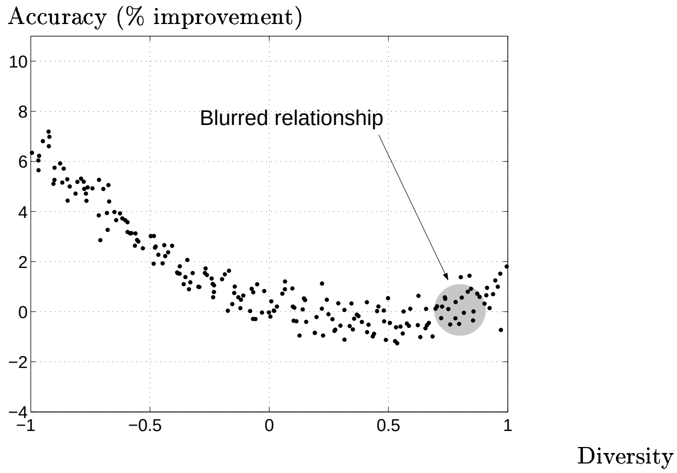


Fig. 2. A typical accuracy-diversity scatterplot. Each point corresponds to an ensemble. The gray dot shows a hypothetical area where ensembles appear most often in real problems.

If we do not enforce diversity, the ensemble is most likely to appear as a dot towards the right side of the graph. For these ensembles, the improvement on the individually best accuracy is usually negligible.

Note that the neat relationship in Figure 2 was obtained under quite artificial circumstances. When the members of the ensemble have different accuracies and different pairwise diversities, such a relationship has not been found. Then is measuring and studying diversity a wasted journey? Several studies which explicitly use diversity to help analyze or build the ensemble offer answers to this skeptical and provocative question.

3.1 Diversity for finding bounds and theoretical relationships

Assume that classifier outputs are estimates of the *posterior probabilities*, $\hat{P}_i(\omega_s|\mathbf{x})$, $s = 1, \ldots, c$, so that the estimate $\hat{P}_i(\omega_s|\mathbf{x})$ satisfies

$$\hat{P}_i(\omega_s|\mathbf{x}) = P(\omega_s|\mathbf{x}) + \eta_s^i(\mathbf{x}), \tag{9}$$

where $\eta_s^i(\mathbf{x})$ is the error for class ω_s made by classifier D_i. The outputs for each class are combined by averaging, or by an order statistic such as minimum, maximum or median. Tumer and Ghosh [24] derive an expression about the added classification error (i.e., the error above the Bayes error) of the team under a set of assumptions

$$E_{add}^{ave} = E_{add}\left(\frac{1+\delta(L-1)}{L}\right), \tag{10}$$

where E_{add} is the added error of the individual classifiers (all have the same error), and δ is a correlation coefficient (the measure of diversity of the ensemble)[3].

Breiman [1] derives an upper bound on the generalization error of random forests (ensembles of decision trees built according to a simple randomization technology, one possible variant of which is bootstrap sampling) using the averaged pairwise correlation between the ensemble members. The classifiers produce class labels and majority vote is assumed as the combination method. The bound is given by

$$Pr(\text{ generalization error of the ensemble }) \quad \leq \quad \bar{\rho}(1-s^2)s^2, \tag{11}$$

where $\bar{\rho}$ is the averaged pairwise correlation (our diversity measure)[4], and s is the "strength" of the ensemble. The strength is a measure of accuracy based on the concept of margin. Admittedly the bound is not very tight as it is based on the Chebyshev's inequality but nonetheless it shows the tendency: the higher the diversity (small $\bar{r}ho$), the lower the error.

Both results can be viewed as pieces of that yet missing more general theory of diversity.

[3] Averaged pairwise correlations between $P_i(\omega_s|\mathbf{x})$ and $P_j(\omega_s|\mathbf{x})$, $i, j = 1, \ldots, L$ are calculated for every s, then weighted by the prior probabilities $\hat{P}(\omega_s)$ and summed.

[4] Since the classifier outputs are labels, therefore categorical, the correlation is calculated between the two oracle outputs. For every data point $\mathbf{z}_j \in \mathbf{Z}$, the output of D_i is taken to be 1 if the suggested label for $\mathbf{z}_j$ matches the true one, and -1, otherwise.

3.2 Diversity for visualization

Diversity measures have been used to find out what is happening in the ensemble. Pękalska and coauthors [17] look at a two-dimensional plot derived from the matrix of pairwise diversity. Each classifier is plotted as a dot in the 2-d space found by Sammon mapping which preserves the "distances" (diversities in our case). The ensemble is a classifier itself and can also be plotted. Any method of combination of the individual outputs can also be mapped. Even more, the oracle classifier (all objects correctly recognized) can be plotted as a point to complete the picture.

Margineantu and Dietterich suggest the kappa-error plots as shown in Figure 3 [16]. Every pair of classifiers is plotted as a dot in a two-dimensional space. The pairwise measure kappa (6) is used as the x-coordinate of the point and the average of the individual training errors of the two classifiers is used as the y-coordinate. Thus there are $L(L-1)/2$ points in the scatterplot. The best pairs are situated in the left bottom part of the plot: they have low error and low kappa (low agreement = high diversity).

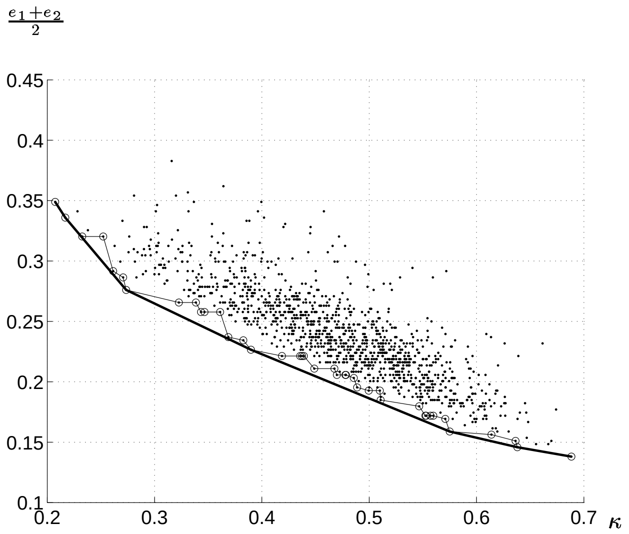


Fig. 3. Kappa-error plot, the convex hull, and the Pareto optimal set of pairs of classifiers

The cloud of points shows the pairwise diversity in one ensemble. Margineantu and Dietterich use it to verify that AdaBoost generates more diverse classifiers than Bagging. The example in the figure corresponds to an ensemble of 50 classifiers for the *glass* data set from UCI Machine Repository Database [5]. The shape

[5] http://www.ics.uci.edu/~mlearn/MLRepository.html

of the cloud indicates that there is a certain trade-off between the accuracy of the pair and its κ-diversity.

The disagreement measure mentioned before, $Dis = b+c$, was used by Skalak [22] to characterize the diversity between a base classifier and a complementary classifier, and then by Ho [7] for measuring diversity in decision forests.

3.3 Overproduce and select

Several studies try the method of producing a pool of classifiers, usually by bagging (taking bootstrap samples from the data sets and building a classifier on each sample) or boosting (modifying the training set for every new member of the ensemble by putting more "emphasis" on the hard objects). Then a selection procedure is suggested to pick the members of the team which are most diverse or most diverse and accurate.

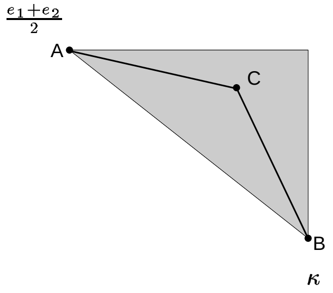
Giacinto and Roli [6] use the *double fault* measure (probability of both classifiers being incorrect, $DF = d$) and also the Q statistics [19], to form a pairwise diversity matrix for a classifier pool and subsequently to select classifiers that are least related. The selection is carried out using a search method through the set of all pairs of classifiers until the desired number of ensemble members is reached.

Margineantu and Dietterich [3, 16] use kappa to select the ensemble out of the set of classifiers produced by AdaBoost. They call this "ensemble pruning". One proposed technique matches the work by Giacinto and Roli. The pairwise κ's are calculated for the whole ensemble. The pruned ensemble is created by progressively selecting pairs with lowest kappas (highest diversity) until the desired number of classifiers is reached. Since both studies apply greedy algorithms, optimality of the selected ensemble is not guaranteed.

Another interesting strategy of selection is to use the kappa-error plots. As the most desirable pairs of classifiers are situated toward the lower left corner of the plot, Margineantu and Dietterich use the convex hull [16], called kappa-error convex hull pruning. The convex hull of points is depicted in Figure 3 with a thick line.

It might happen that the convex hull contains only a few classifiers on the frontier. Small variations of the estimates of κ and $\frac{e_1+e_2}{2}$ might change the whole frontier, making convex-hull pruning overly sensitive to noise. The number of classifiers in the pruned ensemble cannot be specified in advance. This lack of control on the ensemble size is seen as a defect of the method [16].

Therefore we may look at **Pareto optimality** as an alternative to the convex hull approach. Let $A = \{a_1, \dots, a_m\}$ be a set of alternatives (pairs in our case) characterized by a set of criteria $C = \{C_1, \dots, C_M\}$ (low kappa and low error in our case), The Pareto-optimal set $S^* \subseteq S$ contains all non-dominated alternatives. An alternative a_i is non-dominated iff there is no other alternative $a_j \in S$, $j \neq i$, so that a_j is better than a_i on *all* criteria. For the two criteria in our example, the Pareto optimal set will be a superset of the convex hull. The concept is illustrated in Figure 4.



Suppose that points A and B are in the convex hull. Point C is not in the convex hull because it is "behind" the segment AB. However, C is better than A on the error criterion and better than B on the kappa criterion. Therefore C is nondominated, so it belongs in the Pareto optimal set.

Fig. 4. Illustration of Pareto optimality

The Pareto-optimal set for the glass data example is depicted in Figure 3 by a thin line joining the circled points in the set.

4 Diversity for building the ensemble

Here we only sketch a possible use of diversity during *the process of building of the ensemble.* The motivation is that diversity should step out of the passive role of being only a tool for monitoring and should help actively at the design stage. The overproduce-and-select approach discussed earlier is a step in this direction. However, we need to overproduce first. An alternative approach would be to stop the growing of the ensemble when diversity and accuracy satisfy a certain condition.

We take as the starting point the kappa-error diagram and run AdaBoost[6]. The first and the second classifier (D_1 and D_2) will define one single point on the diagram. This point will be the convex hull and the Pareto optimal set of itself. The third classifier, D_3,will place two more points: one for (D_1, D_3) and another for (D_2, D_3). At this step we recalculate the Pareto optimal set. If the points by new classifier have not changed the previous Pareto optimal set, then this classifier is not accepted. Another training set is generated with the same distribution and a new classifier is attempted on it. We run the acceptance check again, and proceed in this manner. A pre-specified parameter T defines the limit number of attempts from the same distribution. When T attempts have been made and a classifier has not been accepted, the procedure stops and the classifier pairs in the last Pareto optimal set are declared to be the ensemble.

Next we give some experimental results with the proposed ensemble construction method. Four data sets were used, three from the UCI and one called "cone-torus"[7]. The characteristics of the data sets are summarized in Table 1.

Table 2 shows the errors (in %) and the number of classifiers in the ensembles for the standard AdaBoost (run up to 50 classifiers), the kappa-error selection

[6] We use AdaBoost in its resampling version: the likelihood of data poins to be selected is modified.

[7] Available at http://www.bangor.ac.uk/~mas00a/Z.txt and Zts.txt, [12]

Table 1. Characteristics of the data sets used

Name	Objects	Classes	Features	Availability
glass	214	6	9	UCI[5]
cone-torus	800	3	2	see footnote 7
liver	345	2	6	UCI[5]
pima	768	2	8	UCI[5]

from the final ensemble of 50 using Pareto optimality, and the proposed diversity-incorporation method ($T = 5$ and $T = 10$). All the results are the testing averages from 20 runs. At each run we split randomly the data set into 90% training and 10% testing. The same 20 splits were used with each method. The "winner" (smallest error) for each data set is marked in boldface.

Table 2. Testing error in % (and ensemble size L), average from 20 splits into 90/10 training/testing

Data set	Classical		Select from 50		Incorporate		Incorporate	
	AdaBoost		(Pareto)		diversity, $T = 5$		diversity, $T = 10$	
glass	23.18	(50)	25.00	(28.55)	**22.95**	(24.85)	24.09	(34.95)
cone-torus	12.38	(50)	12.56	(33.70)	12.81	(37.00)	**12.25**	(41.50)
liver	30.29	(50)	32.14	(16.45)	32.71	(7.8)	**28.43**	(9.10)
pima	**26.43**	(50)	28.05	(16.40)	29.09	(6.30)	28.64	(8.85)

As we all know, miracles rarely happen in pattern recognition. If our results appear dramatically better than everybody else's then better double check the code! The results here show that we can sometimes achieve better performance than standard AdaBoost with smaller number of ensemble members. In any case, this part of the paper was not intended as a consistently examined new ensemble building technology. It is instead an illustration of the potential of diversity as an ensemble building aid.

5 Conclusions

This talk looks into diversity again, asking the same old awkward question: do we need to measure and exploit diversity at all? We try to relate the notion of diversity, which appears well channeled in biological and ecological studies, to diversity in combining classifiers. Unfortunately, a straightforward translation is not apparent at this stage, so some analogues from the field of combining classifiers are presented.

Subsequently, the usage of diversity is summarized into four main bullets: for theoretical relationships and limits, for monitoring, for selection from a given pool of classifiers, and for direct use in building the ensemble. The fourth direction seems to be the least researched. The lack of effort can be explained by the discouraging results trying to link unequivocally diversity with the ensemble error for practical purposes. If there is no proven link, then why bother incorporating diversity into the building of the ensemble? For regression and approximation such a link exists, and there are increasing amount of studies on training the members of the ensemble by enforcing negative correlation between the classifier outputs.

However, as argued at the beginning, heuristics sometime produce a surprising escape from what seems to be a dead end. Even without an agreed upon definition, based upon intuition only, diversity can be put to work. Once successful, there should be explanations and maybe a theory that will tell us when, where and how we can make the best of diversity. Will the quest for diversity, now a marginal streak in the research on classifier combination, resurface one day as a major theory for new ensemble creating methods?

Let us return to the experts' disagreement about where we are. My personal view is that we have accumulated quite a lot of unstructured insight on classifier combination. We have a good critical mass of experimental studies and some patchy but exciting theory on different ensemble building and combination methods. So, yes, we know a lot and, no, we don't have the all-explaining theory. This makes our field of research what it is – challenging and entertaining.

6 Acknowledgements

I wish to thank my colleagues Chris Whitaker, Tim Porter and Catherine Shipp for the inspirational and enjoyable discussions on the subject and in particular on the proposed method.

References

1. L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
2. P. Cunningham and J. Carney. Diversity versus quality in classification ensembles based on feature selection. Technical Report TCD-CS-2000-02, Department of Computer Science, Trinity College Dublin, 2000.
3. T. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization. *Machine Learning*, 40(2):139–157, 2000.
4. J.L. Fleiss. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, 1981.
5. J. Ghosh. Multiclassifier systems: Back to the future. In F. Roli and J. Kittler, editors, *Proc. 3d International Workshop on Multiple Classifier Systems, MCS'02*, volume 2364 of *Lecture Notes in Computer Science*, pages 1–15, Cagliari, Italy, 2002. Springer-Verlag.
6. G. Giacinto and F. Roli. Design of effective neural network ensembles for image classification processes. *Image Vision and Computing Journal*, 19(9-10):699–707, 2001.
7. T.K. Ho. The random space method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

8. T.K. Ho. Multiple classifier combination: Lessons and the next steps. In A Kandel and H. Bunke, editors, *Hybrid Methods in Pattern Recognition*, pages 171–198. World Scientific Publishing, 2002.
9. E.M. Kleinberg. Stochastic discrimination. *Annals of Mathematics and Artificial Intelligence*, 1:207–239, 1990.
10. R. Kohavi and D.H. Wolpert. Bias plus variance decomposition for zero-one loss functions. In L. Saitta, editor, *Machine Learning: Proc. 13th International Conference*, pages 275–283. Morgan Kaufmann, 1996.
11. A. Krogh and J. Vedelsby. Neural network ensembles, cross validation and active learning. In G. Tesauro, D.S. Touretzky, and T.K. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 231–238. MIT Press, Cambridge, MA, 1995.
12. L.I. Kuncheva. *Fuzzy Classifier Design.* Studies in Fuzziness and Soft Computing. Springer Verlag, Heidelberg, 2000.
13. L.I. Kuncheva and C.J. Whitaker. Measures of diversity in classifier ensembles. *Machine Learning*, 51:181–207, 2003.
14. L. Lam. Classifier combinations: implementations and theoretical issues. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 78–86, Cagliari, Italy, 2000. Springer.
15. B. Littlewood and D.R. Miller. Conceptual modeling of coincident failures in multiversion software. *IEEE Transactions on Software Engineering*, 15(12):1596–1614, 1989.
16. D.D. Margineantu and T.G. Dietterich. Pruning adaptive boosting. In *Proc. 14th International Conference on Machine Learning*, pages 378–387, San Francisco, 1997. Morgan Kaufmann.
17. E. Pękalska, R.P.W. Duin, and M. Skurichina. A discussion on the classifier projection space for classifier combining. In F. Roli and J. Kittler, editors, *Proc. 3d International Workshop on Multiple Classifier Systems, MCS'02*, volume 2364 of *Lecture Notes in Computer Science*, pages 137–148, Cagliari, Italy, 2002. Springer-Verlag.
18. C.R. Rao. Diversity: Its measurement, decomposition, apportionment and analysis. *Sankya: The Indian Journal of Statistics, Series A*, 44(1):1–22, 1982.
19. F. Roli, G. Giacinto, and G. Vernazza. Methods for designing multiple classifier systems. In J. Kittler and F. Roli, editors, *Proc. Second International Workshop on Multiple Classifier Systems*, volume 2096 of *Lecture Notes in Computer Science*, pages 78–87, Cambridge, UK, 2001. Springer-Verlag.
20. B.E. Rosen. Ensemble learning using decorrelated neural networks. *Connection Science*, 8(3/4):373–383, 1996.
21. C.A. Shipp and L.I. Kuncheva. Relationships between combination methods and measures of diversity in combining classifiers. *Information Fusion*, 3(2):135–148, 2002.
22. D.B. Skalak. The sources of increased accuracy for two proposed boosting algorithms. In *Proc. American Association for Artificial Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop*, 1996.
23. P.H.A. Sneath and R.R. Sokal. *Numerical Taxonomy.* W.H. Freeman & Co, 1973.
24. K. Tumer and J. Ghosh. Linear and order statistics combiners for pattern classification. In A.J.C. Sharkey, editor, *Combining Artificial Neural Nets*, pages 127–161. Springer-Verlag, London, 1999.