

TEN MEASURES OF DIVERSITY IN CLASSIFIER ENSEMBLES: LIMITS FOR TWO CLASSIFIERS

L.I. Kuncheva and C.J. Whitaker¹

Abstract

Independence and dependence of classifier outputs have been debated in the recent literature giving rise to notions such as diversity, complementarity, orthogonality, etc. There seems to be no consensus on the meaning of these notions beyond the intuitive perception. Here we summarize 10 measures of classifier diversity: 4 pairwise and 6 non-pairwise measures. We derive the limits of the measures for 2 classifiers of equal accuracy.

1 Introduction

Classifier combination is an advanced pattern recognition technology gaining increasing attention in the recent literature. Classifier outputs are combined in an attempt to reach a more accurate decision than the best individual classifier in the pool. It is commonly agreed that the classifiers should be *different* from each other, otherwise the overall decision will not be better than the individual decisions. This difference (called also *diversity*) may lead to a better or a worse overall decision, so there is “good” diversity as well as “bad” diversity [6]. It has been recognized that quantifying and studying the dependencies is an important issue in combining classifiers [12].

Let $\Omega = \{\omega_1, \dots, \omega_c\}$ be a set of class labels, and $\mathcal{D} = \{D_1, \dots, D_L\}$ be a set (called also pool, team, ensemble, mixture, etc.) of classifiers. Three common types of classifier outputs are:

1. A class label: $D_i : \mathfrak{R}^n \rightarrow \Omega$, where $\mathbf{x} \in \mathfrak{R}^n$ is an unlabeled vector submitted as the classifier input;
2. Estimates of the c posterior probabilities $P(\omega_i|\mathbf{x})$, $i = 1, \dots, c$;
3. An “oracle” label: The output $D_i(\mathbf{x})$ is 1 if \mathbf{x} is recognized correctly by D_i , and 0, otherwise.

Here we will use **oracle** classifier outputs, i.e., $D_i(\mathbf{x}) \in \{0, 1\}$.

Numerous measures of diversity/dependence have been proposed in the literature. We can summarize the current results as follows:

(1) When classifiers output estimates of the posterior probabilities $\hat{P}(\omega_s|\mathbf{x})$, and the outputs for each class are combined by averaging, the classification error rate above the Bayes error (called the added error of the team E_{add}^{team}) depends on the correlation between the estimates (see [22, 23]). With positively correlated classifiers E_{add}^{team} is only slightly lower than the individual added error E_{add} . For uncorrelated classifiers $E_{add}^{team} = E_{add}/L$, and negatively correlated classifiers reduce the error even further.

(2) When classifiers output class labels, the classification error can be decomposed into bias and variance terms [1, 8] or into bias and spread terms [2]. In both cases the second term accounts for the diversity of the ensemble. These results have been used to study the behavior of classifier ensembles in terms of the bias-variance trade-off.

(3) For the case of classifier outputs in the form of correct/incorrect vote, four levels of diversity are detailed in [19]: Level 1, where no more than one classifier is wrong on each data point. Level 2, where for each data point up to $\lfloor L/2 \rfloor$ could be wrong (the majority is always correct). Level 3, where at least one classifier is correct for each data point, and Level 4, where there might be points for which none of the classifiers is correct.

(4) It is recognized that a negative correlation should be pursued when designing classifier ensembles, and many such design methods have been proposed, predominantly altering the available training set to build the individual classifiers [6, 9, 13, 15, 18, 19, 20, 23].

Practically, there is no unique choice of a measure of diversity or dependence. There are pairwise measures which are calculated for each pair of classifiers in \mathcal{D} and then averaged and non-pairwise measures that either use the idea of entropy or correlation of individual outputs with the averaged output of \mathcal{D} or are based on the distribution of “difficulty” of the data points. In this study we present 10 measures of classifier diversity for oracle classifier outputs: 4 pairwise and 6 non-pairwise. We give the limits of the measures for the case of $L = 2$ classifiers of the same individual accuracy p .

¹School of Informatics, University of Wales, Bangor, Gwynedd, LL57 1UT, United Kingdom
e-mail: {l.i.kuncheva,c.j.whitaker}@bangor.ac.uk

2 Pairwise diversity measures

Let $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ be a labeled data set, $\mathbf{z}_j \in \mathfrak{R}^n$ coming from the classification problem in question. We can represent the output of a classifier D_i as an N -dimensional binary vector $\mathbf{y}_i = [y_{1,i}, \dots, y_{N,i}]^T$, such that $y_{j,i} = 1$, if D_i recognizes correctly \mathbf{z}_j , and 0, otherwise, $i = 1, \dots, L$.

Consider two classifiers, D_i and D_k , and a 2×2 table that summarizes their outputs as shown below

	D_k correct (1)	D_k wrong (0)
D_i correct (1)	N^{11}	N^{10}
D_i wrong (0)	N^{01}	N^{00}

$N = N^{00} + N^{01} + N^{10} + N^{11}$, where N^{vw} is the number of elements \mathbf{z}_j of \mathbf{Z} for which $y_{j,i} = v$ and $y_{j,k} = w$. By taking $a = N^{11}/N$ to be an estimate of the probability of both classifiers being correct, and proceeding with the other counts N^{vw} , we can rewrite the above table as shown in Table 1.

Table 1: The 2×2 relationship table with probabilities

	D_k correct (1)	D_k wrong (0)
D_i correct (1)	a	b
D_i wrong (0)	c	d

Note that $a + b + c + d = 1$. There are various statistics to assess the similarity of two classifier outputs.

The Q statistic. Yule's Q statistic [24] for two classifiers, e.g., D_i and D_k , is

$$Q_{i,k} = \frac{ad - bc}{ad + bc} \quad (1)$$

For statistically *independent* classifiers, $Q_{i,k} = 0$. 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. For a set of L classifiers, the averaged Q statistics of all pairs is taken. The Q statistics were studied in [11] and related to the limits of P_{maj} .

The correlation coefficient ρ . The correlation between two binary classifier outputs (correct/incorrect), \mathbf{y}_i and \mathbf{y}_k , using probabilities in Table 1 is

$$\rho_{i,k} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \quad (2)$$

For any two classifiers, Q and ρ have the same sign, and it can be proved that $|\rho| \leq |Q|$.

The disagreement measure. This is the ratio between the number of observations on which one classifier is correct and the other is incorrect to the total number of observations, or for $N \rightarrow \infty$, the probability that one of the classifiers is wrong and the other is correct. This measure was used by Skalak [20] to characterize the diversity between a base classifier and a complementary classifier, and also by Ho [7]. In our notation,

$$D_{i,k} = b + c. \quad (3)$$

The double-fault measure. This measure is defined as the proportion of the cases that have been misclassified by both classifiers, or in the limit case, the probability of both classifiers giving a wrong label.

$$DF_{i,k} = d. \quad (4)$$

This measure was used by Giacinto and Roli [4] to form a pairwise diversity matrix for a classifier pool and subsequently to select classifiers that are least related. For all pairwise measures the averaged value over all pairs of classifiers is used. We note that all these pairwise measures have been proposed as measures of (dis)similarity in the numerical taxonomy literature (e.g., [21]).

3 Non-pairwise diversity measures

For the non-pairwise measures we quote the formulae for L classifiers.

The entropy measure E . The highest diversity among classifiers for a particular $\mathbf{z}_j \in \mathbf{Z}$ is manifested by $\lfloor L/2 \rfloor$ of the votes in \mathbf{y}_j with the same value (0 or 1) and the other $L - \lfloor L/2 \rfloor$ with the alternative value. If they all were 0's or all were 1's, there is no disagreement, and the classifiers cannot be deemed diverse. One possible measure of diversity based on this concept is

$$E = \frac{1}{N} \sum_{j=1}^N \frac{1}{(L - \lfloor L/2 \rfloor - 1)} \min \left\{ \sum_{i=1}^L y_{j,i}, L - \sum_{i=1}^L y_{j,i} \right\}. \quad (5)$$

E varies between 0 and 1, where 0 indicates no difference and 1 indicates the highest possible diversity. While value 0 is achievable for any number of classifiers L and any p , value 1 can only be attained for $p \in [\frac{L-1}{2L}, \frac{L+1}{2L}]$.

The measure of difficulty θ . The idea for this measure came from a study by Hansen and Salomon [5]. We define a discrete random variable X taking values in $\{\frac{0}{L}, \frac{1}{L}, \dots, 1\}$ and denoting the proportion of classifiers in \mathcal{D} that correctly classify an input \mathbf{x} drawn randomly from the distribution of the problem. To estimate the probability mass function of X , the L classifiers in \mathcal{D} are run on the data set Z .

Figure 1 shows three possible probability mass functions of X for $L = 2$ classifiers.

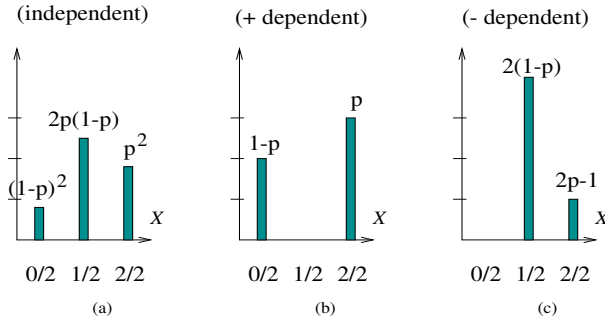


Figure 1: Three patterns of “difficulty” for $L = 2$ classifiers.

Plot (a) shows independent classifiers. If the same points are *difficult* for all classifiers, and the other points are *easy* for all classifiers, we obtain a plot similar to plot (b) (identical classifiers, no diversity in the team). If the points that are difficult for some classifiers are easy for other classifiers, the distribution of X is as (c). Diverse teams of classifiers \mathcal{D} will have smaller variance of X (plot(c)). Teams of similar classifiers will have high variance (plot b) and the variance for independent classifier will be in-between. Therefore we define the measure of *difficulty* θ to be

$$\theta = Var(X). \quad (6)$$

The higher the value of θ , the worse the classifier team. Ideally, $\theta = 0$, but this is an unrealistic scenario. More often, real classifiers are positively dependent and will exhibit patterns similar to plot (b) in Figure 1.

Kohavi-Wolpert variance. Denote by $l(\mathbf{z}_j)$ the number of classifiers from \mathcal{D} that correctly recognize \mathbf{z}_j , i.e., $l(\mathbf{z}_j) = \sum_{i=1}^L y_{j,i}$. Taking the formula for the variance from [8] and applying simple manipulations, the diversity measure becomes

$$KW = \frac{1}{NL^2} \sum_{j=1}^N l(\mathbf{z}_j)(L - l(\mathbf{z}_j)) \quad (7)$$

This expression is the averaged variance of the Bernoulli variable \mathbf{y} with values 0 for incorrect and 1 for correct classification of each object in Z . Interestingly, KW differs from the averaged disagreement measure D_{av} by a coefficient, i.e.,

$$KW = \frac{L-1}{2L} D_{av}. \quad (8)$$

(The proof of the equivalence is given in [10].)

Measurement of interrater agreement κ . A statistic developed as a measure of interrater reliability, called κ , is given in [3]. It is used when different rater (here classifiers) assess subjects (here \mathbf{z}_j) to measure the level of agreement while correcting for chance (see [3] for details). It has links for the intraclass correlation coefficient ([3]) and the significance test of Looney [14].

If we denote \bar{p} to be the average individual classification accuracy, then

$$\kappa = 1 - \frac{\frac{1}{L} \sum_{j=1}^N l(\mathbf{z}_j)(L - l(\mathbf{z}_j))}{N(L - 1)\bar{p}(1 - \bar{p})} \quad (9)$$

and so κ is related to KW and D_{av} as follows

$$\kappa = 1 - \frac{L}{(L - 1)\bar{p}(1 - \bar{p})} KW = 1 - \frac{1}{2\bar{p}(1 - \bar{p})} S_{av}. \quad (10)$$

Fleiss [3] defines the pairwise κ_p as

$$\kappa_p = \frac{2(ad - bc)}{(a + b)(c + d) + (a + c)(b + d)} \quad (11)$$

However, it can be shown that the (non-pairwise) κ (9) is not obtained by averaging κ_p .

Generalized diversity. This measure has been proposed in [17]. Let Y be a random variable expressing the proportion of classifiers (out of L) that **fail** on a randomly drawn object $\mathbf{x} \in \mathfrak{R}^n$. Denote by p_i the probability that $Y = \frac{i}{L}$. (Not that $Y = 1 - X$ introduced for θ). Denote by $p(i)$ the probability that i randomly chosen classifiers will fail on a randomly chosen \mathbf{x} . Then

$$p(1) = \sum_{i=1}^L \frac{i}{L} p_i, \quad \text{and} \quad p(2) = \sum_{i=1}^L \frac{i}{L} \frac{(i-1)}{(L-1)} p_i. \quad (12)$$

The generalization diversity measure GD is

$$GD = 1 - \frac{p(2)}{p(1)}. \quad (13)$$

Coincident failure diversity. This is a modification of GD proposed in [16].

$$CFD = \begin{cases} 0, & p_0 = 1.0; \\ \frac{1}{1-p_0} \sum_{i=1}^L \frac{L-i}{L-1} p_i, & p_0 < 1 \end{cases} \quad (14)$$

4 Results and Conclusions

Table 2 summarizes the 10 measures for the case of $L = 2$ classifiers. The expressions use the probabilities as in Table 1, a : both correct, b : first correct second wrong, c : first wrong second correct, and d : both wrong. The limits were calculated for two classifiers of the same individual accuracy p ranging from 0.5 to 1 and sketched as plots in Table 2.

First, Q is the only measure whose limits and independence value do not depend on p .

Second, only Q , ρ , and the pairwise interrater agreement, κ_p , have a constant value (zero) for statistically independent classifier outputs.

Third, there are groups of measures with similar behavior for the considered special case:

- ρ and the pairwise κ_p (it can be shown that the two are equivalent for classifiers of the same accuracy);
- Disagreement, KW variance, and Entropy;
- GD and CFD

Little is published about using measures of diversity for *designing* classifier ensembles [4, 18] or about the relationship between the classification potential of the team \mathcal{D} and a specific measure of diversity. Theoretical results have been derived for special cases, e.g., for classifiers of the same accuracy and pairwise dependency [22, 11]. Further investigations need to be performed to examine the relationship between diversity/dependency of the classifiers, the differences in the accuracy of the individual classifiers, methods to form the team of classifiers, and the accuracy of the team.

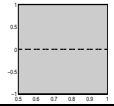
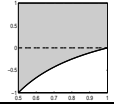
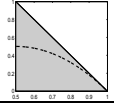
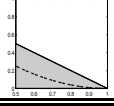
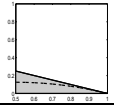
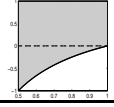
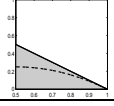
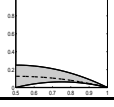
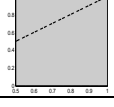
This paper has examined ten measures of diversity/dependency for the two-classifier case. Once there is no agreement on a ‘best’ measure in the literature, how can we choose among such measures? For quantitative classifier outputs, the correlation coefficient is the natural measure of diversity/dependency used by many

authors. Features of the correlation coefficient are a 0 value for independence and a range of ± 1 for the extremes of positive and negative dependency. Accuracy and diversity can be thought of as ‘conceptually orthogonal’, and so, a measure of diversity should not depend on p . The results in Table 2 show that the Q statistic is the only measure which satisfies the above features and so we commend its use.

References

- [1] E. Bauer and R. Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, 36:105–142, 1999.
- [2] L. Breiman. Combining predictors. In A.J.C. Sharkey, editor, *Combining Artificial Neural Nets*, pages 31–50. Springer-Verlag, London, 1999.
- [3] J.L. Fleiss. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, 1981.
- [4] G. Giacinto and F. Roli. Design of effective neural network ensembles for image classification processes. *Image Vision and Computing Journal*, 2000. (to appear).
- [5] L.K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.
- [6] S. Hashem. Treating harmful collinearity in neural network ensembles. In A.J.C. Sharkey, editor, *Combining Artificial Neural Nets*, pages 101–125. Springer-Verlag, London, 1999.
- [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] 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.
- [9] 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.
- [10] L.I. Kuncheva and C.J. Whitaker. Measures of diversity in classifier ensembles. (submitted).
- [11] L.I. Kuncheva, C.J. Whitaker, C.A. Shipp, and R.P.W. Duin. Limits on the majority vote accuracy in classifier fusion. (submitted).
- [12] 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.
- [13] Y. Liu and X. Yao. Ensemble learning via negative correlation. *Neural Networks*, 12:1399–1404, 1999.
- [14] S.W. Looney. A statistical technique for comparing the accuracies of several classifiers. *Pattern Recognition Letters*, 8:5–9, 1988.
- [15] D. Opitz and J. Shavlik. A genetic algorithm approach for creating neural network ensembles. In A.J.C. Sharkey, editor, *Combining Artificial Neural Nets*, pages 79–99. Springer-Verlag, London, 1999.
- [16] D. Partridge and W. Krzanowski. Distinct failure diversity in multiversion software. (personal communication).
- [17] D. Partridge and W. J. Krzanowski. Software diversity: practical statistics for its measurement and exploitation. *Information & Software Technology*, 39:707–717, 1997.
- [18] B.E. Rosen. Ensemble learning using decorrelated neural networks. *Connection Science*, 8(3/4):373–383, 1996.
- [19] A.J.C. Sharkey and N.E. Sharkey. Combining diverse neural nets. *The Knowledge Engineering Review*, 12(3):231–247, 1997.
- [20] 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.
- [21] P.H.A. Sneath and R.R. Sokal. *Numerical Taxonomy*. W.H. Freeman & Co, 1973.
- [22] K. Tumer and J. Ghosh. Error correlation and error reduction in ensemble classifiers. *Connection Science*, 8(3/4):385–404, 1996.
- [23] 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.
- [24] G.U. Yule. On the association of attributes in statistics. *Phil. Trans., A*, 194:257–319, 1900.

Table 2: Expressions of the 10 measures of diversity for $L = 2$ classifiers. The limits and the values for independent classifiers when both classifiers have individual accuracy p , are also shown. The last column plots in grey the regions of possible values of the measures versus individual accuracy $p \in (0.5, 1)$. The dashed line corresponds to independent classifiers. $a, b, c,$ and d are the probabilities from Table 1

Measure	General expression	Limits		Independence value	Plot
		Minimum	Maximum		
Q	$\frac{ad-bc}{ad+bc}$	-1	1	0	
ρ	$\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$	$\frac{(p-1)}{p}$	1	0	
Disagreement	$b + c$	0	$2(1 - p)$	$2p(1 - p)$	
Double fault	d	0	$1 - p$	$(1 - p)^2$	
KW	$\frac{b+c}{4}$	0	$\frac{(1-p)}{2}$	$\frac{p(1-p)}{2}$	
κ_p	$\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$	$\frac{(p-1)}{p}$	1	0	
Entropy	$\frac{b+c}{2}$	0	$(1 - p)$	$p(1 - p)$	
Difficulty (θ)	$ad + \frac{(a+d)(b+c)}{4}$	$\frac{(1-p)(2p-1)}{2}$	$p(1 - p)$	$\frac{p(1-p)}{2}$	
GD	$\frac{b+c}{1-a+d}$	0	1	p	
CFD	$\frac{b+c}{1-a}$	0	1	$\frac{2p}{1+p}$	