Minimising Added Classification Error using Walsh Coefficients

Terry Windeatt, Cemre Zor

Abstract. Two-class supervised learning in the context of a classifier ensemble may be formulated as learning an incompletely specified Boolean function, and the associated Walsh coefficients can be estimated without knowledge of the unspecified patterns. Using an extended version of the Tumer-Ghosh model, the relationship between Added Classification Error and second order Walsh coefficients is established. In this paper, the ensemble is composed of Multi-layer Perceptron (MLP) base classifiers, with the number of hidden nodes and epochs systematically varied. Experiments demonstrate that the mean second order coefficients peak at the same number of training epochs as ensemble test error reaches a minimum.

Keywords: Classification Algorithm, Multilayer Perceptrons, Pattern Analysis, Pattern Recognition

1 INTRODUCTION

Walsh coefficients, particularly the Rademacher-Walsh ordering, have previously been used for logic design [1]. In this paper, the second order Walsh coefficients are used for pattern classification, where the goal is to minimise ensemble test error. The motivation will be explained in terms of the meaning of the spectral coefficients, and since the meaning is not dependent on the ordering, we will refer only to the Walsh coefficients. To understand the significance of the coefficients, the Tumer-Ghosh model [2] for ensemble classifiers will be described. This model defines Added Classification Error as the difference between classifier error and Bayes error. The model provides a framework for understanding relationship between classifier correlation and reduction in error due to combining.

An important design issue for Multiple Classifier Systems (MCS) is choice of individual (base) classifier complexity, which is usually set with the help of a validation set or cross-validation techniques [3] [4]. The maximum number of patterns should be reserved for training, which implies that base classifier parameters should ideally be determined from the training set. However, there has been no convincing theory or experimental study to suggest that any measure, computed on the training set, can reliably facilitate optimal ensemble design [5]. It is possible to bootstrap training patterns and use the Ensemble Out-of-Bootstrap error estimate [6], in place of validation, but since each bootstrap replicate uses approximately two-thirds of the patterns, lack of training data can cause degradation of performance. In this paper, the proposed measure based on Walsh coefficients is computed on the training set.

The main contribution is to demonstrate the relationship between second order Walsh coefficients of a Boolean function and Added Classification Error of an ensemble, an issue that has not been addressed in any previous conference or journal publication. First order Walsh coefficients were shown to provide a measure of class separability for selecting optimal base classifiers in [7], in which it is also shown that this does not imply optimality of the ensemble. In contrast, in this paper it is shown that second order Walsh coefficients can be used to determine base classifier complexity for optimal ensemble performance. The motivation for using Walsh coefficients in ensemble design is fully explored in [5] and [7]. The interested reader is further referred to [1] and [8] for an understanding of the meaning and applications of Walsh coefficients.

Section **2** explains the computation of the second order coefficients, and Section **3** discusses their relationship with the model of Added Classification Error. In Section **4**, mean second order Walsh coefficients are computed as the number of nodes and training epochs of MLP base classifiers are systematically varied.

Manuscript received Sept. 1, 2010. This work was supported in part by UK Government, EPSRC under grant number E061664/1

T. Windeatt, C. Zor are with the Centre for Vision Speech and Signal Processing, Faculty of Electronics and Physical Sciences, University of Surrey, Guildford Surrey, GU2 7XH, UK (phone: +44-1483-9286; e-mail: t.windeatt@surrey.ac.uk)

2 WALSH COEFFICIENTS

Consider a two-class supervised learning problem of μ training patterns, with the label given to each pattern X_m denoted by $\Omega_m = \Phi(X_m)$ where $m = 1 \ldots \mu$.. It is assumed that there are N parallel base classifiers and that X_m is an N-dimension vector formed from the decisions of the N classifiers, applied to the original patterns which in general are real-valued and of arbitrary dimension. Therefore, we may represent the *mth* pattern by $(X_{mi}, i = 1 \ldots N)$ a vertex in the N-dimensional binary hypercube. Both pattern features and target label are binary, so that $X_{mi}, \Omega_m \in \{0,1\}$ or $\{1,-1\}$ and Φ is the unknown Boolean function that maps X_m to the target label Ω_m. The Walsh transform is derived from the mapping T_n, that requires $\{1,-1\}$ coding and is defined recursively as follows

$$T_n = \begin{bmatrix} T_{n-1} & T_{n-1} \\ T_{n-1} & -T_{n-1} \end{bmatrix} \qquad (1)$$

where $T_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ and from (1) second order spectral coefficients are defined [1]

$$s_{ij} = \sum_{m=1}^{\mu} (X_{mi} \oplus X_{mj}) \Omega_m \qquad (2)$$

where $X_{mi}, X_{mj}, \Omega_m \in \{1,-1\}$. The meaning of s_{ij} $i,j = 1 \ldots N$, $i \neq j$ in (2) is that the coefficients represent correlation between $\Phi(X_m)$ and $X_{mi} \oplus X_{mj}$ where $\oplus$ is logic exclusive-OR.

For realistic learning problems, Φ will be incompletely specified and noisy. Relationships for computing spectral

coefficients for incompletely specified Boolean functions, in the context of minimal synthesis of logic circuits, are proved in [9]. Here, we summarise relevant concepts using pattern recognition terminology (patterns for minterms). Although spectral coefficients of any order may be computed using similar formulae, we concentrate on second order.

For binary variables $p, q \in \{0,1\}$ define n_{pq} to be the number of class p patterns (minterms) and define d_q to be the number of unspecified patterns (don't care minterms) that satisfy $X_{mi} \oplus X_{mj} = q$. Note that $X_{mi} \oplus X_{mj} = 1$ implies pair of classifiers *i* and *j* disagree for pattern X_m and $X_{mi} \oplus X_{mj} = 0$ implies classifiers agree. Second order spectral coefficients may then be computed as in [9]

$$s_{ij} = (n_{11} + n_{00}) - (n_{01} + n_{10}) \qquad (3)$$

Since the sum of specified and unspecified patterns of an N-dimensional Boolean function is given by $n_{11} + n_{00} + n_{01} + n_{10} + d_1 + d_0 = 2^N$ substitution into (3) gives various equivalent formulae, for example $s_{ij} = 2(n_{11} + n_{00}) + d_1 + d_0 - 2^N$ The advantage of (3) is that the unspecified patterns (d_1, d_0) do not enter explicitly into the computation.

3 ADDED CLASSIFICATION ERROR MODEL

Figure **1** shows the two class (ω_1, ω_0) model of Added Classification Error (E darkly shaded region) according to [2], which for simplicity is restricted to one dimension (*x*). The optimum (Bayes) boundary in Figure **1** is the loci of all points $\tilde{x} : P(\omega_1 \mid \tilde{x}) = P(\omega_0 \mid \tilde{x})$. The output of the classifier representing class ω_1 is given by

$$\hat{P}(\omega_1 \mid x) = P(\omega_1 \mid x) + \varepsilon_1(x) \qquad (4)$$

where $P, \hat{P}$ are the actual and estimated *a posteriori* probability distributions as shown in Figure **1**, and $\varepsilon_1(x)$ is the difference between them. A similar equation to (4) is obtained for class ω_0 with $P(\omega_0 \mid x), \hat{P}(\omega_0 \mid x)$ and error $\varepsilon_0(x)$. If *b* in Figure **1** is the amount that the *kth* classifier boundary (x_b) differs from the ideal Bayes boundary ($\tilde{x}$), and assuming that b is a Gaussian random variable with mean β and variance σ_b, in [2] it is shown using (4) that Added Classification Error for *kth* classifier is given by

$$E_k = \nabla P(\sigma_b^2 + \beta^2) \qquad (5)$$

where $\nabla P = 0.5(P'(\omega_1 \mid \tilde{x}) - P'(\omega_0 \mid \tilde{x}))p(\tilde{x})$ and P' indicates differentiation.

In this paper, the model is extended to the case of a pair of classifiers *(i,j),* and we assume in the analysis that classifier complexity is varied from under to over-fitting, with respect to optimal. Figure **2** shows decision boundaries of *(i,j)th* classifiers for which it is assumed that the complexity is not sufficient to approximate the Bayes boundary, so that both classifiers under-fit. Note in Figure **2** that estimated probabilities $\hat{P}(\omega_0 \mid x)$ and $\hat{P}(\omega_1 \mid x)$ are omitted for clarity. Mutually exclusive areas under the probability distribution are labelled 1 – 8 in Figure **2**, and denoting the number of patterns in area y by a_y, the contribution from classifiers *i,j* according to area is given in Table 1. For example, a_2, a_5 correspond to areas where classifiers disagree so second subscript is 1.

The model assumptions are the same as used in (4) and (5), namely that the *a posteriori* probability distributions are approximated by base classifier outputs and are locally monotonic around the Bayes boundary. While a Gaussian Distribution satisfies these properties, it is not necessary to assume overlapping Gaussians in the Tumer-Ghosh model [2]. A further assumption in this paper is that the area under the tails of the distribution, represented by (a_4, a_5, a_6, a_7) contain equal number of ω_1 and ω_0 patterns. By substituting the areas from Table 1 representing $n_{11}, n_{00}, n_{01}, n_{10}$ into (3) (e.g. number of ω_1 patterns in $a_2 + a_5$ is n_{11})

$$s_{ij} = a_2 - (a_1 + a_3 - a_8) \qquad (6)$$

since patterns in a_4, a_5, a_6, a_7 cancel. From Figure **2** $a_1 + a_2 + a_3$ and a_8 are fixed, and represent the patterns above the tails of the distributions. From (6)

$$s_{ij} = 2a_2 - \gamma \qquad (7)$$

where additive constant is given by

$$\gamma = a_1 + a_2 + a_3 - a_8 \qquad (8)$$

If we assume that the Bayes rate applies equally to the two classes, that is according to the prior probabilities, the constant in (8) can be easily estimated, by separately summing the number of ω_0 and ω_1 patterns. If p_0 is prior probability class ω_0 and B is estimated Bayes error, a_8 is the total number of ω_0 patterns minus number in $(a_4 + a_5 + a_6)$ minus number in a_7

$$a_8 = p_0\mu - Bp_0\mu - (1 - p_0)B\mu = (p_0 - B)\mu \qquad (9)$$

Similarly, summing the ω_1 patterns in $a_1, \ldots a_7$

$$a_1 + a_2 + a_3 = (1 - p_0 - B)\mu \qquad (10)$$

From (8) (9) and (10), and after normalisation with respect to total number of patterns μ

$$\gamma = 1 - 2p_0 \qquad (11)$$

The difference in Added Classification Error of *ith* and *jth* classifiers is given by $E_{ij} = E_i - E_j$ defined in (5), and shown in Figure **2** as a_2. Therefore from (7)

$$E_{ij} = E_i - E_j = 0.5(s_{ij} + \gamma) \qquad (12)$$

From (11) and (12) it may be stated that, if $p_0 = 0.5$ then $\gamma = 0$ and the difference in Added Error of an arbitrary pair of classifiers is half the second order Walsh coefficient. Note that (6) – (12) rely on perfect model assumptions, otherwise we could use approximations ($\approx$), rather than equality.

Averaging over all pairs of classifiers in (12) the mean difference in added error is given by

$$\Delta\overline{E} = \sum_{i,j,i\neq j} E_{ij} \qquad (13)$$

As complexity of classifiers is increased, the boundaries of classifiers *i,j* in Figure **2** are expected to move closer to the Bayes boundary. When classifiers are on opposite sides of the Bayes boundary, a similar analysis of areas under distribution reveals that E_i cancels E_j. (In Table 1, a_3 is modified to n_{01}, a_6 is modified to n_{11}, n_{01} and $a_2 = E_i$, $a_3 = E_j$). In Section **4**, this will be used to explain why $\Delta\overline{E}$ reaches a peak and reduces when classifiers straddle the Bayes boundary.

Consider now the effect of classifier correlation on the reduction in Added Classification Error of the ensemble. We know from [2] that when classifiers are i.i.d and $\beta = 0$, average added error

$$\overline{\mathrm{E}} = {}^{1}\!/_{N}\, E \qquad (14)$$

In (14) the ensemble added error $\overline{\mathrm{E}}$ has decreased the average individual added error E by the factor ${}^{1}\!/_{N}$. However, when the i.i.d. assumption is relaxed, there is a well-known trade-off between accuracy and diversity [5]. When classifier errors are correlated the error depends on the linear correlation δ averaged over all classifier pairs [2]

$$\frac{\overline{E}}{E} = \left(\frac{1+\delta(N-1)}{N}\right) \qquad (15)$$

with $\delta = 0$ in (15) corresponding to (14).

4 EXPERIMENTAL EVIDENCE

Natural two-class benchmark problems selected from [10] and [11] are shown in Table 2. Datasets *dermo2, ecoli2, iris2, vehicle2* are multiclass but the class with most patterns is re-labelled ω_1 and remaining patterns ω_0. *Twonorm* is well-known artificial data using overlapping Gaussian from [12], and 3000 patterns are randomly generated each repetition.

The original features are normalised to mean 0 std 1, and for datasets with missing values the scheme suggested in [10] is used. Random perturbation of the MLP base classifiers is caused by different starting weights on each run. The number of hidden nodes and training epochs of homogenous (same number of nodes and epochs) MLP base classifiers are systematically varied. The experiments are performed with one hundred single hidden-layer MLP base classifiers, using the Levenberg-Marquardt training algorithm with default parameters. Combining uses majority vote (error rates were compared with Sum rule using soft outputs, with no significant difference). The random train/test split is 50/50 (except *twonorm*) and experiments are repeated twenty times and averaged, with tests for significance based on McNemar [13]).

We need to estimate the Bayes classifier for the significance test, and to compute δ using (15). The Bayes estimation is performed for 90/10 split using original features, and a Support Vector Classifier (SVC) with polynomial kernel run 100 times. The polynomial degree is varied as well as the regularisation constant, and lowest test error found is given in Table 2.

Figure **3** gives mean results over the first ten datasets, which clearly indicates the overall trend. Figure **3** (a) (b) shows base and ensemble test error rates with Bayes error subtracted. Figure **3** (c) shows mean linear correlation coefficient between pairs of classifiers computed on training set. Figure **3** (d) gives McNemar coefficient for Ensemble classifier compared with Bayes prediction, where the solid horizontal line is the threshold (3.84), indicating difference at ninety-five percent confidence. Figure **3** (e) is the difference between Figure **3** (c) and the value of δ computed using (15). Figure **3** (f) shows $\Delta\overline{E}$ from (13) computed on the training set, with additive constant γ from (11) removed. In (12), s_{ij} is computed using (3) and normalised by the total number of patterns ($n_{11} + n_{00} + n_{10} + n_{01}$). Note that, for each dataset the class with most patterns is assigned ω_0 to give the same sign to γ in (11).

Figure **3** is intended to show that the relationship between second order Walsh coefficients and Added Error, given in (12), enables the base classifier complexity to be selected for optimal ensemble performance. The additive constant γ defined in (11), has the effect of shifting the curves in Figure **3** (f), but does not change the shape of the curve. Figure **3** (b) (f) demonstrate that mean pair-wise difference of Added Classification Error $\Delta\overline{E}$ reaches a maximum at three to four epochs, the same number for which ensemble test error reaches minimum. For one and two epochs, the classifier boundaries are more constrained, and $\Delta\overline{E}$ is reduced. As number of epochs is increased beyond four, where the ensemble over-fits, Figure **3** (f) indicates a decrease in $\Delta\overline{E}$, since classifiers are either side of Bayes boundary, and E_i and E_j cancel as explained in Section **3**. Note from Figure **3** (a) that the mean base classifier test error is higher than ensemble error, and reaches minimum at seven epochs, indicating that classifiers are sub-optimal with non-zero β defined in (5).

From Figure **3** (c), it is evident that mean linear correlation coefficient is lower at three compared to seven epochs and reflects an increase in diversity. Correlation is maximum (diversity is minimum) at seven epochs, when base classifier test error in figure **3** (a) is minimum. The ensemble error is lower at three epochs due to the effect of (15), showing relationship between δ and reduction in error, which is an example of the accuracy/diversity trade-off [5]. Figure **3** (e) shows that the error in correlation estimate is less than 0.1 for two to seven epochs, and therefore (15) represents a reasonable model of error reduction. For two to seven epochs, the ensemble error is closest to Bayes error, as shown by the McNemar coefficient in Figure **3** (d). From Figure **3**, it may be concluded that second order coefficients reach a peak when the accuracy/diversity trade-off is optimal.

In Figure **3**, all ten datasets have lowest ensemble test error at three to four epochs, so taking an average is meaningful. In contrast Figure **4** shows ensemble error and $\Delta\overline{E}$ for one dataset *vehicle2*, which has minimum ensemble error and maximum $\Delta\overline{E}$ at seven epochs. The

corresponding artificial *twonorm* graphs are shown in Figure **5**, as number of training patterns is varied [10 20 30] percent. This is equivalent to 300, 600, 900 training patterns and shows similar trend to real datasets.

5 DISCUSSION & CONCLUSION

The definition of bias and variance (β, σ_b) in (5) is useful for understanding Added Classification Error and relationship to combining [14] [15], but the subtleties of applying bias and variance to 0/1 loss function have prompted many different definitions [16], none universally accepted [17] [18]. Finding the relationship between these definitions is the subject of future research.

It has been shown that second order spectral coefficients of incompletely specified Boolean functions have an important role to play in designing ensemble classifiers. The results indicate that the unknown *a posteriori* distributions of the benchmark datasets, do satisfy the assumptions outlined in Section 3. The estimation of the coefficients in this paper is based on the principle of minimal logic synthesis, and future direction may consider how alternate methods of estimation lead to different generalizations. Finally, it is worth noting that although MLP base classifiers are considered here, the techniques are applicable to any base classifier that outputs a binary decision.

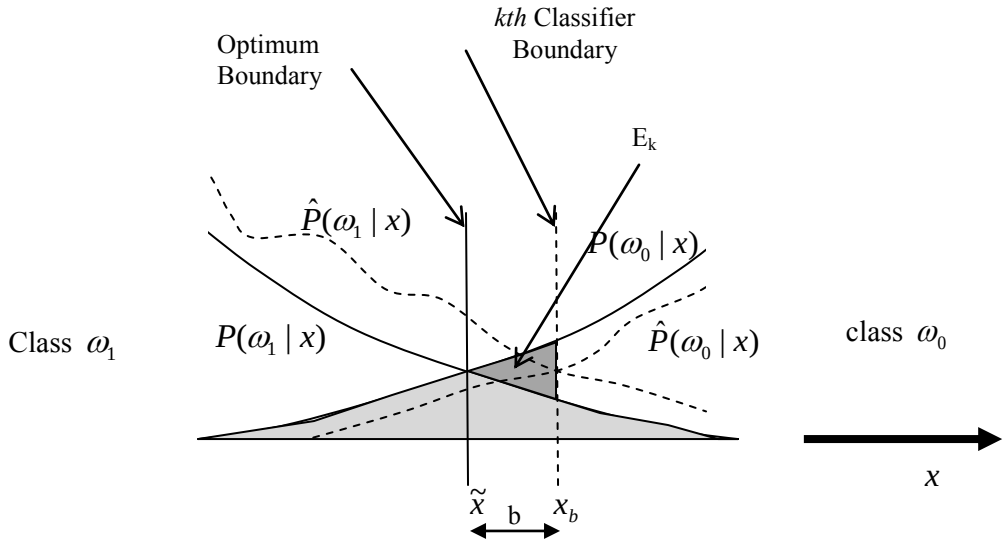

Figure 1: Model of error region associated with *a posteriori* probabilities showing optimum (Bayes) boundary, kth classifier boundary with Added Classification Error (E_k)

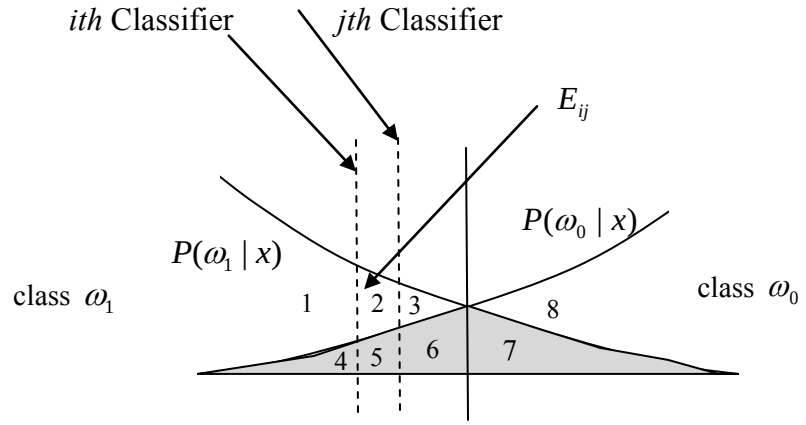

Figure 2: Model showing pair of classifier boundaries and the difference in Added Classification Error between *ith* and *jth* classifiers E_{ij} (area 2)

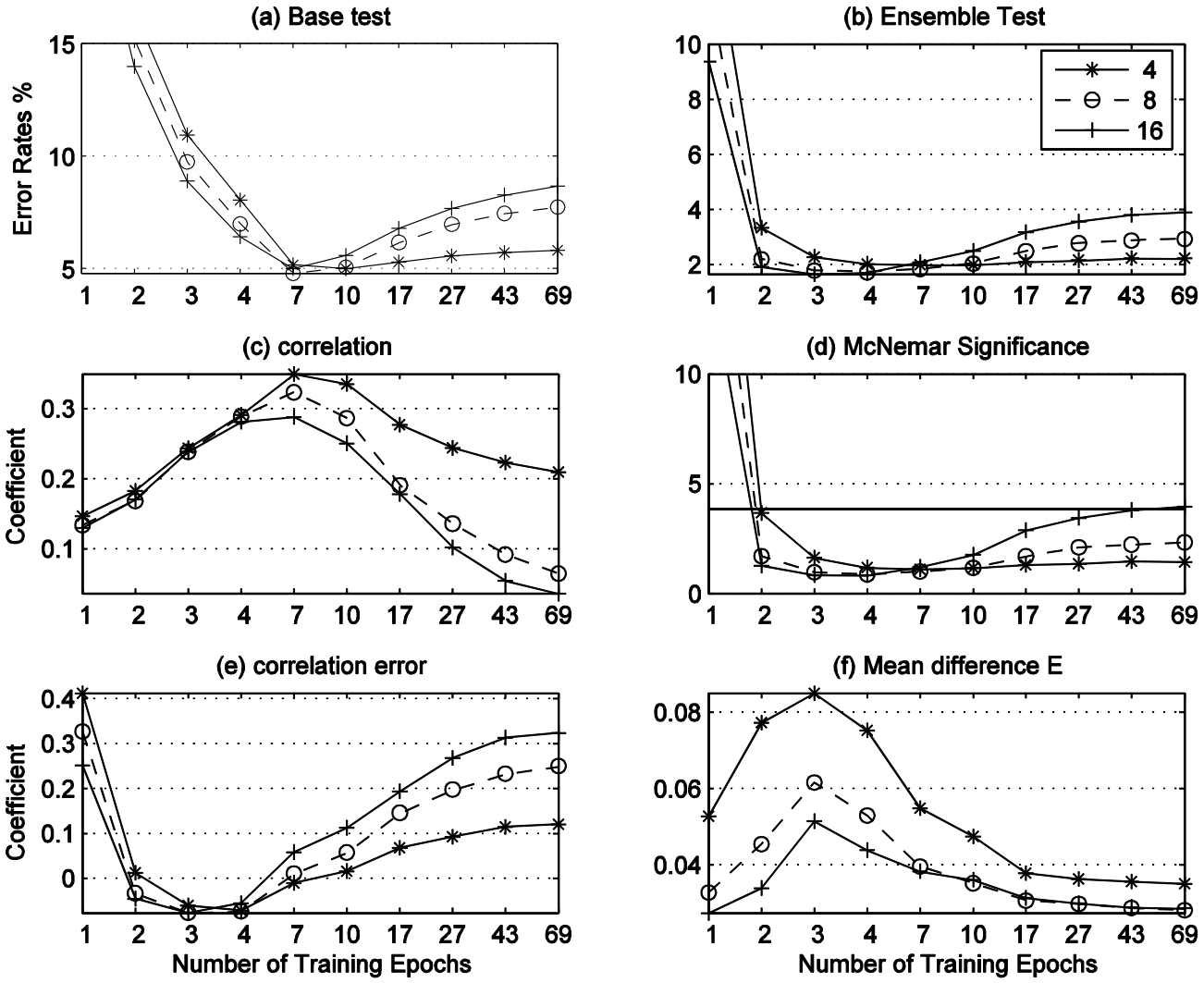

Figure 3: Mean over ten 2-class datasets for [4,8,16] nodes (a) (b) test error rates with Bayes estimate subtracted (c) Linear pair-wise correlation coefficient (d) McNemar Significance versus Bayes (e) Error in correlation estimate and (f) $\Delta\overline{E}$ the mean second order coefficients with additive constant subtracted (12)

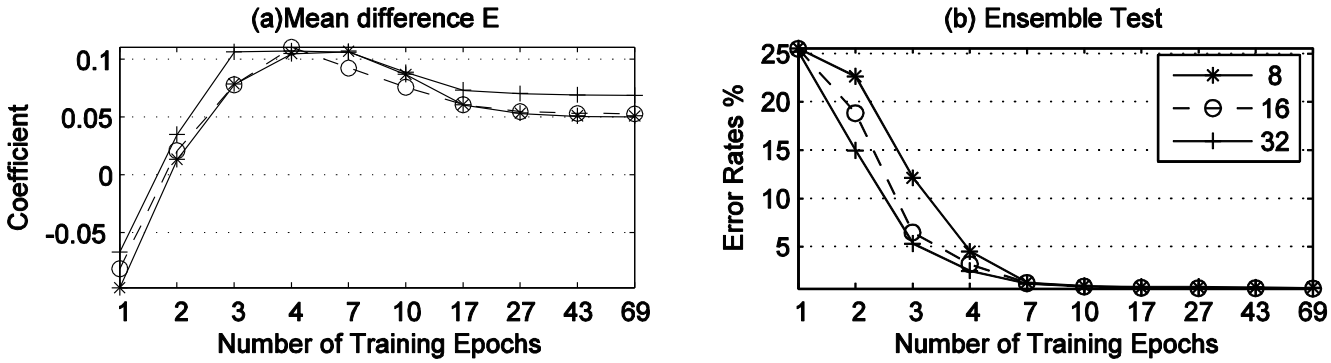

Figure 4: *Vehicle2* dataset for [8,16,32] nodes (a) $\Delta\overline{E}$ the mean second order coefficients with additive constant subtracted (b) ensemble error rate with Bayes estimate subtracted

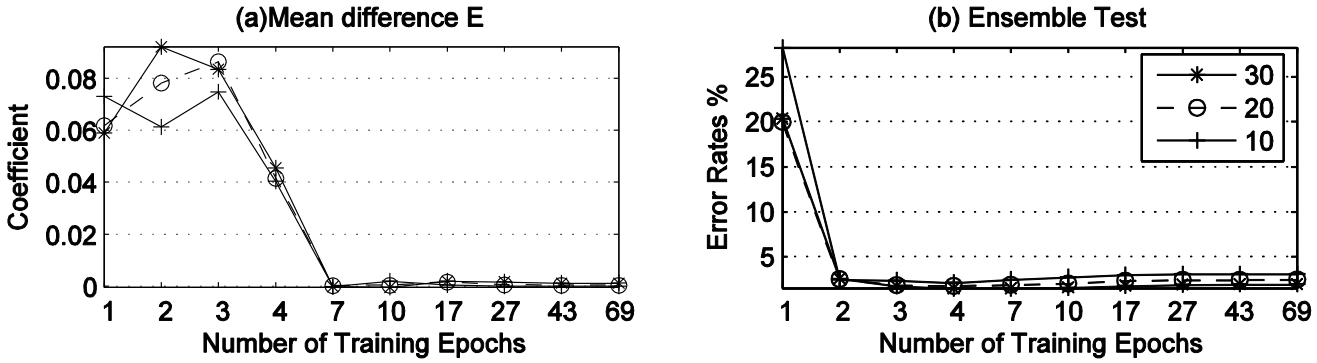

Figure 5: *Twonorm* dataset for 8 nodes [10.20,30]% training (a) $\Delta\overline{E}$ the mean second order coefficients with additive constant subtracted (b) ensemble error rate with Bayes estimate subtracted

Table 1: **Areas under Distribution defined in Fig. 2, showing corresponding number of class ω_1, ω_0 patterns (1st subscript) for which the pair of classifiers agree or disagree (2nd subscript)**

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
ω_1	n_{10}	n_{11}	n_{10}	n_{10}	n_{11}	n_{10}	n_{10}	
ω_0				n_{00}	n_{01}	n_{00}	n_{00}	n_{00}

Table 2: Datasets showing # patterns, prior probability ω_0, #continuous and discrete features and estimated Bayes error

DATASET	*#pat*	p_0	*#con*	*#dis*	*%Bay*
cancer	699	.655	0	9	3.1
card	690	.555	6	9	12.8
credita	690	.555	3	11	14.1
diabetes	768	.651	8	0	22.0
heart	920	.553	5	30	16.1
ion	351	.641	31	3	6.8
vote	435	.614	0	16	2.8
dermo2	366	.694	1	33	0
ecoli2	336	.574	5	2	2.07
iris2	150	.667	4	0	0
vehicle2	846	.742	18	0	0.238
twonorm	3000	.50	20	0	2.3

References

[1] L. Hurst, D. M. Miller, and J. Muzio, *Spectral Techniques in Digital Logic*, Academic Press, 1985.

[2] K. Tumer and J. Ghosh, "Error correlation and error reduction in ensemble classifiers," *Connection Science*, vol. 8, no. 3, pp. 385-404, 1996.

[3] R. S. Smith and T. Windeatt, "Class-separability Weighting and Bootstrapping in Error Correcting Output Code Ensembles," in *Proceedings of the 9th Workshop on Multiple Classifier Systems*, Cairo, Egypt, Apr. 2010, pp. 185-194.

[4] L. K. Hansen and P. Salamon, "Neural Network Ensembles, " *IEEE Trans. PAMI*, vol. 12, pp. 993-1001, Oct. 1990.

[5] T. Windeatt, "Accuracy/ Diversity and ensemble classifier design," *IEEE Trans Neural Networks*, vol. 17, pp. 1194- 1211, Sept. 2006.

[6] T. Bylander, "Estimating generalisation error two-class datasets using out-of-bag estimate," *Machine Learning* vol. 48, no. 1-3, pp. 287-297, 2002.

[7] T. Windeatt, "Vote Counting Measures for Ensemble Classifiers," *Pattern Recognition, vol.* 36, no. 12, pp. 2743-2756, 2003.

[8] K. G. Beauchamp, *Walsh Functions and their Applications*, Academic Press, 1975.

[9] B. J. Falkowski and M.A.Perkowski, "Effective Computer Methods for the Calculation of Rademacher-Walsh Spectrum for Completely and Incompletely Specified Boolean Functions," *IEEE Trans. on Computer-Aided Design*, vol. 11, pp.1207-1226, Oct. 1992.

[10] L. Prechelt, Proben1: A set of neural network Benchmark Problems and Benchmarking Rules, *Tech Report 21/94*, Univ. Karlsruhe, Germany, Sept., 1994.

[11] C. J. Merz and P. M. Murphy, UCI repository of ML databases, [Online], http://www.ics.uci.edu/~mlearn/MLRepository.html.

[12] L. Breiman, "Arcing classifiers," *The Annals of Statistics*, vol. 26, no. 3, pp. 801-849, 1998.

[13] T. G. Dietterich, "Approx. statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895-1923, 1998.

[14] G. Fumera and F. Roli, "A Theoretical and Experimental Analysis of Linear Combiners for Multiple Classifier Systems," *IEEE Trans. PAMI*, vol. 27, pp. 942-956, June 2005.

[15] L.I. Kuncheva, *Combining Pattern Classifiers*, Wiley, 2004.

[16] G. James, "Variance and Bias for General Loss Functions," *Machine Learning*, vol. 51, no. 2, pp.115-135, 2003.

[17] R. S. Smith and T. Windeatt, "A Bias-Variance Analysis of Bootstrapped Class-Separability Weighting for ECOC Ensembles," in *Proceedings of the 22nd International Conference on Pattern Recognition*, Istanbul, Turkey, Aug. 2010.

[18] C. Zor, T. Windeatt and B. Yanikoglu, "Bias-Variance Analysis of ECOC and Bagging using Neural Nets," in *Proceedings of the Workshop on Supervised and Unsupervised Ensemble Methods and their Applications*, Barcelona, Spain, Sept. 2010, pp. 65-74.