Diversity Measures for Multiple Classifier System Analysis and Design

Terry Windeatt
Centre for Vision, Speech and Signal Processing
University of Surrey, Guildford, Surrey, UK, GU2 7XH
Email: t.windeatt@surrey.ac.uk, Phone: 044-01483-689286 Fax: 044-01483-686031

Abstract
In the context of Multiple Classifier Systems, diversity among base classifiers is known to be a necessary condition for improvement in ensemble performance. In this paper the ability of several pair-wise diversity measures to predict generalisation error is compared. A new pair-wise measure, which is computed between pairs of patterns rather than pairs of classifiers, is also proposed for two-class problems. It is shown experimentally that the proposed measure is well correlated with base classifier test error as base classifier complexity is systematically varied. However, correlation with unity-weighted sum and vote is shown to be weaker, demonstrating the difficulty in choosing base classifier complexity for optimal fusion. An alternative strategy based on weighted combination is also investigated and shown to be less sensitive to number of training epochs.

Keywords: decision level fusion, multiple classifiers, ensembles, error-correcting, binary coding

1. Introduction

A method of designing pattern recognition systems, known as the Multiple Classifier System (MCS) or committee/ensemble approach, has emerged over recent years to address the practical problem of designing classification systems with improved accuracy and efficiency. The aim is to design a composite system that outperforms any individual classifier by pooling together the decisions of all classifiers. The rationale is that it may be more difficult to optimise the design of a single complex classifier than to optimise the design of a combination of relatively simple classifiers.

Attempts to understand the effectiveness of the MCS framework have prompted the development of various measures. The Margin (Section 4.1) concept was used originally to help explain Boosting and Support Vector Machines. Bias and Variance (Section 4.2) are concepts from regression theory that have motivated modified definitions for 0/1 loss function for characterising Bagging and other ensemble techniques. Various diversity measures (Section 3) have been studied with the intention of determining whether they correlate with ensemble accuracy. However, the question of whether the information available from any of these measures can be used to assist MCS design is open. Most commonly, MCS parameter values are set with the help of either a validation set or cross-validation techniques [1]. In [2] these measures are described and explained in the context of a vote counting framework. In this paper, in contrast to [2], the proposed measure relaxes the assumption on Hamming Distance (equ. (22)) and is experimentally compared with various pair-wise Diversity Measures.

Although it is known that diversity among base classifiers is a necessary condition for improvement in ensemble performance, there is no general agreement about how to quantify the notion of diversity among a set of classifiers. Diversity measures can be categorised into two types [3], pair-wise and non-pair-wise. In order to apply pair-wise measures to finding overall diversity of a set of classifiers it is necessary to average over the set. Non-pair-wise measures attempt to measure diversity of a set of classifiers directly, based for example on variance, entropy or proportion of classifiers that fail on randomly selected patterns. The main difficulty with diversity measures is the so-called accuracy-diversity dilemma. As explained in [4], as base classifiers approach the highest levels of accuracy, diversity must decrease so that it is expected that there will be a trade-off between diversity and accuracy. There has been no convincing theory or experimental study to suggest that there exists any measure that can reliably predict generalisation error of an ensemble. In [3] the desirability of using negatively correlated base classifiers in an ensemble is recognised, and it is shown experimentally that four pair-wise diversity measures (equations (14) to (17))

are similarly related to majority vote accuracy when classifier dependency is systematically changed. The conclusion in [5] was that the Double Fault measure (equation (17)) showed reasonable correlation with some combination methods. Since there is a lack of a general theory on how diversity impacts ensemble performance, experimental studies provide an important contribution to discovering whether a relationship exists and if so whether it can be quantified and understood.

To be really useful for MCS design, a measure should be capable of extracting relevant information from the training set. Model selection from training data is known to require a built-in assumption, since realistic learning problems are in general ill-posed [6]. The assumption here is that base classifier complexity is varied over a suitable range and that over-fitting of the training set is detected by observing changes in diversity or correlation. It is shown experimentally in Section 6 that, over a range of datasets, some measures are well correlated with base classifier test error when number of training epochs is varied. As with Bias /Variance definitions (Section 4.2) one must assume that the underlying probability distributions are well-behaved, and it is easy to construct examples of probability distributions for which the method fails. The results in Section 6 also demonstrate that correlation with unity-weighting test error is not as strong as with the mean base classifier test error, illustrating the difficulty of choosing base classifier complexity for optimal fusion. An alternative strategy based on weighted combination is also investigated and the sensitivity of combined test error to number of epochs is compared with unity-weighting.

The paper is organised as follows. A measure of correlation, based on a spectral representation of a Boolean function, is defined in Section 2. Conventional pair-wise measures are described in Section 3, which also includes proposal of a new pair-wise measure computed over pairs of patterns rather than pairs of classifiers. Margin and Bias/Variance are discussed in Section 4, and in Section 5 various weighted combination schemes are proposed. Experimental evidence, incorporating Multi-layer Perceptron (MLP) base classifiers in an MCS framework, is presented and evaluated in Section 6.

2. Spectral Measure of Correlation

The architecture envisaged in this paper is a simple MCS framework in which there are B parallel base classifiers whose outputs are combined either by voting or summation. Unity weighting will be designated as MAJ (majority vote) or SUM. The study is restricted to two-class supervised learning problems for which we assume that there are μ patterns with the label given to each pattern X_m denoted by $\omega_m = f(X_m)$ where $m = 1 \ldots \mu$ and f is the unknown function that maps X_m to the target label ω_m. In general the original features associated with each pattern are real-valued but we have no need to refer to them explicitly. Instead, we represent the *mth* pattern by the B-dimensional vector formed from the (real-valued) base classifier outputs ξ given by

$$X_m = (\xi_{m1}, \xi_{m2}, \ldots, \xi_{mB}), \qquad \xi_{mi} \in [0,1], \quad \omega_m \in \{0,1\}, \quad i = 1 \ldots B \tag{1}$$

In this paper we stick to the convention that, where a pair of subscripts refers to pattern and classifier, the first subscript refers to the pattern and the second subscript to the classifier.

If one of two classes is assigned by each of B base classifiers, the *mth* pattern X_m in (1) may be represented as a vertex in the B-dimensional binary hypercube, resulting in a binary-to-binary mapping between classifier outputs and target labels

$$X_m = (x_{m1}, x_{m2}, \ldots, x_{mB}) \qquad x_{mi} \text{ and } \omega_m \in \{0,1\}, \; i = 1 \ldots B \tag{2}$$

In [9], a spectral representation of a Boolean function $f(X)$ is proposed for characterising the mapping in (2) between base classifier outputs and target labels. In contrast to the conventional method of representing a Boolean function as single vertices of the binary hypercube, a spectral representation incorporates global information about the function in each spectral coefficient. A well known property of the transforms that characterise these mappings (e.g. Rademacher-Walsh transform [7]) is that the first order coefficients

represent the correlation between *f(X)* and x_i [8]. First order coefficients provide a unique identifier of *f(X)* if it is linearly separable, and through table-lookup provide weights for a Threshold Logic Unit (TLU) implementation. The meaning of higher order coefficients as measures of correlation is given in [8].

In [2] a method of estimating the first order coefficients is described, which is based on representing *f(X)* using a correlation measure called sensitivity σ that indicates whether a change in binary value x_i gives rise to a change in *f(X).* For a completely specified Boolean function (truth table available), the *mth* pattern component x_{mj} is assigned σ_{mj} (*j=1,2,...B*) as follows

$$\sigma_{mj}^{+} = x_{mj} \oplus x_{nj} = 1, \quad x_{mj} = \omega_m \neq \omega_n, \ \sum_{k=1}^{B} (x_{mk} \oplus x_{nk}) = 1 \tag{3}$$

$$\sigma_{mj}^{-} = x_{mj} \oplus x_{nj} = 1, \quad x_{mj} = \omega_n \neq \omega_m, \ \sum_{k=1}^{B} (x_{mk} \oplus x_{nk}) = 1 \tag{4}$$

where Hamming Distance $D_H(X_m, X_n) = \sum_{j=1}^{B} (x_{mj} \oplus x_{nj})$,

$\oplus$ is logic exclusive-OR

Applying (3) and (4) involves a search in which each pattern *Xm* of one class, is paired with patterns of the other class that are unit Hamming Distance (D_H) apart, and setting $\sigma_{mj}^{+} = 1$ *if* $x_{mj} = \omega_m$ and $\sigma_{mj}^{-} = 1$ otherwise. The search process is identical to the first stage of logic minimisation, a description of which can be found in any standard textbook on combinational logic. In the context here, it may be interpreted as finding a pair of patterns for which all classifiers are identical except one, which is either positively or negatively sensitive to that pattern pair. Each pattern component x_{mj} then has associated σ_{mj}, and these contributions can be added, a technique that is known as spectral summation. [8]. Although spectral summation is not new, the idea of separation into positive and negative contributions is novel and allows Boolean functions to be tested for separability. The technique, along with simple examples, is explained in [2] and [9]. Using spectral summation the difference between excitatory and inhibitory contributions, $\sum_{m=1}^{\mu} \sigma_{mj}^{+}$ and $\sum_{m=1}^{\mu} \sigma_{mj}^{-}$ gives the *jth* first order spectral coefficient. The existence of $\sum_{m=1}^{\mu} \sigma_{mj}^{+} > 0$ and $\sum_{m=1}^{\mu} \sigma_{mj}^{-} > 0$ for given *j* provides evidence that the set of patterns is non-separable in the *j*th component [7].

Clearly, for a realistic learning problem the unknown binary-to-binary function *f* will not be completely specified. Therefore two changes to (3) and (4) are required for a problem in which patterns may be noisy, incompletely specified and perhaps contradictory. First it is assumed that all pattern pairs drawn from different classes contribute, rather than just the nearest neighbours. Second, with no evidence to the contrary, the contribution from a pattern pair is assumed to be inversely proportional to D_H and shared equally between all pattern components that differ. Then the *mth* pattern component x_{mj} is assigned σ_{mj} (*j=1,2,...B*) as shown in the following two equations, which are modifications to equ.(3) and (4)

$$\sigma_{mj}^{+} = \sum_{n=1}^{\mu} \frac{x_{mj} \oplus x_{nj}}{D_H(X_m, X_n)}, \qquad x_{mj} = \omega_m \neq \omega_n \tag{5}$$

$$\sigma_{mj}^{-} = \sum_{n=1}^{\mu} \frac{x_{mj} \oplus x_{nj}}{D_H(X_m, X_n)}, \qquad x_{mj} = \omega_n \neq \omega_m, \tag{6}$$

After applying (5) and (6) the *jth* component x_{mj} of a pattern pair has associated σ_{mj}^{-} only if the *jth* base classifier mis-classifies both patterns. Therefore we expect that a pattern with relatively large $\sum_{j=1}^{B} \sigma_{mj}^{-}$ is likely to come from regions where the two classes overlap. Now we would like to define a measure for each pattern that is based on a summation of contributions. For any pattern, say the *nth* pattern consider a measure that uses (5) and (6) σ_n that looks at the difference between relative difference between excitatory and inhibitory contributions $\sum_{j=1}^{B} \sigma_{nj}^{+}$ and $\sum_{j=1}^{B} \sigma_{nj}^{-}$ normalised so that $-1 \leq \sigma_n \leq 1$

$$\sigma_n = \frac{1}{K} \text{ x } \sum_{j=1}^{B} \left(\frac{\sigma_{nj}^{+}}{\sum_{m=1}^{\mu} \sigma_{mj}^{+}} - \frac{\sigma_{nj}^{-}}{\sum_{m=1}^{\mu} \sigma_{mj}^{-}} \right) \tag{7}$$

$$\text{where} \quad K = \sum_{j=1}^{B} \left(\frac{\sigma_{nj}^{+}}{\sum_{m=1}^{\mu} \sigma_{mj}^{+}} + \frac{\sigma_{nj}^{-}}{\sum_{m=1}^{\mu} \sigma_{mj}^{-}} \right)$$

In [9] σ_n is interpreted as indicating how well the *nth* pattern is separated from patterns of the other class by the set of B classifiers. It may be compared with the Margin for the *nth* pattern which represents the confidence of classification. It is possible to define Cumulative Distribution graphs for σ_n [2] similar to Cumulative Distribution graphs for Margin (Section 4.1), that is $g(\sigma_n)$ versus σ_n where $g(\sigma_n)$ is the fraction of patterns with value at least σ_n. Areas under the distribution are plotted in [2]. In this paper, as base classifier complexity is varied, we plot number of positively correlated patterns (n^+) as well as the mean over positively correlated patterns (σ) given by

$$n^{+} = \frac{1}{\mu} \sum_{n=1}^{\mu} I(\sigma_n), \tag{8}$$

where $I(z)=1$ *if* $z > 0$ and 0 otherwise

$$\sigma = \frac{1}{\mu} \sum_{n=1}^{\mu} \sigma_n \qquad \sigma_n > 0 \tag{9}$$

3. Diversity Measures

Various approaches to defining diversity, and to determining the relationship between diversity and accuracy, have been proposed. For our study we consider pair-wise diversity measures, and follow the notation used in [3], in which the output of a classifier is defined to be 1 if a pattern is correctly classified and 0 otherwise. Let the jth classifier output under this labelling scheme be a μ-dimensional binary vector given by y_{mj} *where* $m = 1,\ldots \mu$.

3.1 Pair-wise diversity over classifiers

The following counts are defined for *ith* and *jth* classifiers

$$N^{11}{}_{ij} = \sum_{m=1}^{\mu} y_{mi} \wedge y_{mj} \tag{10}$$

$$N^{00}{}_{ij} = \sum_{m=1}^{\mu} \bar{y}_{mi} \wedge \bar{y}_{mj} \tag{11}$$

$$N^{10}{}_{ij} = \sum_{m=1}^{\mu} y_{mi} \wedge \bar{y}_{mj} \tag{12}$$

$$N^{01}{}_{ij} = \sum_{m=1}^{\mu} \bar{y}_{mi} \wedge y_{mj} \tag{13}$$

where $\wedge$ is logical AND and $\bar{y}$ is the logical complement of y

The Q statistic, Correlation coefficient (ρ), and Double Fault (F) measures defined in [3], all increase with decreasing diversity. Here an Agreement (A) measure is defined as (1 – Disagreement) to make it also increase with decreasing diversity so that

$$Q_{ij} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \tag{14}$$

$$\rho_{ij} = \frac{N^{11}N^{00} - N^{01}N^{10}}{\sqrt{(N^{11} + N^{10})(N^{01} + N^{00})(N^{11} + N^{01})(N^{10} + N^{00})}} \tag{15}$$

$$A_{ij} = 1 - \frac{N^{01} + N^{10}}{N^{11} + N^{00} + N^{01} + N^{10}} \tag{16}$$

$$F_{ij} = \frac{N^{00}}{N^{11} + N^{00} + N^{01} + N^{10}} \tag{17}$$

where the ij subscripts for N in (14) to (17) have been omitted for convenience

The summation of diversity measure $\Delta \in \{Q, \rho, A, F\}$ over B classifiers is given by

$$\Delta_{sum} = \sum_{i=1}^{B-1} \sum_{j=i+1}^{B} \Delta_{ij} \tag{18}$$

so that multiplying by $\frac{2}{B(B-1)}$ gives mean diversity.

3.2 Pair-wise diversity over patterns

Although diversity measures are conventionally calculated over base classifiers as in equations (14) to (17)., it is also possible to compute them over patterns [10].By analogy with the spectral measure of correlation, (7) we propose to calculate diversity measures over patterns between the two classes as follows

Let $\tilde{N}^{11}{}_{mn} = \sum_{j=1}^{B} y_{mj} \wedge y_{nj} = \sum_{j=1}^{B} x_{mj} \wedge \bar{x}_{nj}$,

$\tilde{N}^{00}{}_{mn} = \sum_{j=1}^{B} \bar{y}_{mj} \wedge \bar{y}_{nj} = \sum_{j=1}^{B} \bar{x}_{mj} \wedge x_{nj}$,

$\tilde{N}^{10}{}_{mn} = \sum_{j=1}^{B} y_{mj} \wedge \bar{y}_{nj} = \sum_{j=1}^{B} x_{mj} \wedge x_{nj}$,

$\tilde{N}^{00}{}_{mn} = \sum_{j=1}^{B} \bar{y}_{mj} \wedge \bar{y}_{nj} = \sum_{j=1}^{B} \bar{x}_{mj} \wedge \bar{x}_{nj}$,

where x_{mj} is defined in (2) , *mth* pattern has $\omega=1$and *nth* pattern has $\omega=0$

Then

$$Q'_{mn} = \frac{\tilde{N}^{11}\tilde{N}^{00} - \tilde{N}^{01}\tilde{N}^{10}}{\tilde{N}^{11}\tilde{N}^{00} + \tilde{N}^{01}\tilde{N}^{10}} \tag{19}$$

Now define the summation for the *nth* pattern over patterns

$$Q'_n = \sum_{m=1}^{\mu} Q'_{mn} \ , \ \omega_n \neq \omega_m \tag{20}$$

Using (20) each pattern may be considered to have an associated diversity leading to a distribution similar to the margin and σ distributions. We compute the mean over those patterns with positively valued coefficient as in (9)

$$Q' = \sum_{n=1}^{\mu} Q_n{}', Q_n > 0 \tag{21}$$

ρ', A', F' are similarly defined as in (15) to (17)

Also we define a new measure for each pattern similar to (7), normalised so that $-1 \le \sigma'_n \le 1$

$$\sigma_n' = \frac{1}{\tilde{K}} \left(\frac{\tilde{N}_n^{11}}{\sum_{m=1}^{\mu} \tilde{N}_m^{11}} - \frac{\tilde{N}_n^{00}}{\sum_{m=1}^{\mu} \tilde{N}_m^{00}} \right) \tag{22}$$

where $\tilde{K} = \left(\frac{\tilde{N}_n^{11}}{\sum_{m=1}^{\mu} \tilde{N}_m^{11}} + \frac{\tilde{N}_n^{00}}{\sum_{m=1}^{\mu} \tilde{N}_m^{00}} \right)$ and $\tilde{N}_n^{11} = \sum_{m=1}^{\mu} \tilde{N}_{mn}^{11}$, $\omega_n \neq \omega_m$

The difference between (22) and (7) is that each individual element x_{mj} is not assigned a value depending on Hamming Distance. In Section 6 results suggest that σ_n and σ_n' are both well correlated with base classifier test error.

4. Other Vote Counting measures

4.1 Margin

The Margin (M_n) of the *nth* training pattern is defined as the difference between the weight given to the correct class and the maximum weight given to any of the other classes [11], as follows

$$M_n(X_n, f(X_n)) = \frac{f(X_n)\sum_{j=1}^{B} \alpha_j x_{nj}}{\sum_{j=1}^{B} |\alpha_j|}$$

M_n is a number between –1 and +1, positive for a correct classification, and its absolute value representing confidence of classification. A useful visualisation is a plot of Margins as cumulative distribution graphs, that is *g(z)* versus z where z is the Margin and *g(z)* is the fraction of patterns with Margin at least z. In [2], area under cumulative distribution is plotted as a way of quantifying confidence of classification for a set of patterns. In this paper, we simply plot the mean over positive margins similar to (9).

$$M = \frac{1}{\mu} \sum_{n=1}^{\mu} M_n \qquad M_n > 0 \tag{23}$$

In [11] it is proved that larger Margins are associated with superior upper bounds on the generalisation error, but also it is recognised that the bounds are not necessarily tight and therefore of limited practical usefulness (tighter bounds are claimed in [12]). It appears that for some problems outliers and misclassifications can distort the Margin distribution, causing over-fitting even though Margins are increased.

4.2 Bias/Variance

Bias and Variance are concepts from regression theory that, for 0/1 loss functions, are intended to quantify the difference between classifier and Bayes decision boundary. However there are some difficulties with the various Bias/Variance definitions as reported in [13]. First, no single definition can satisfy both zero Bias/Variance for Bayes classifier and additive decomposition of error (as in regression theory). Secondly, it is easy to think of pathological probability distributions for which the bias and variance effects are non-intuitive, for example bias or variance decreasing while error rate is increasing [13] [15]. Thirdly, there is the practical difficulty that the Bayes classification needs to be known or estimated, although some definitions [14] do not account for the Bayes error. In our experiments, we use Breiman's definition [15] which is based on defining Variance as the component of classification error that is eliminated by aggregation. Patterns are divided into two sets, the Bias set containing patterns for which the Bayes classification disagrees with the aggregate classifier and the Unbias set containing the remainder. Bias is computed using patterns from the Bias set and Variance is computed from the Unbias set, but both Bias and Variance are defined as the difference between the probabilities that the Bayes and base classifier predict the correct class label. This definition has non-zero variance with zero bias for the Bayes classifier, but does satisfy the error decomposition property, so that base classifier error can be decomposed into additive components of Bayes error, Bias and Variance.

5. Weighted Combination

The principle behind weighted combination is to reward classifiers that perform well, and there are a variety of methods for determining the size of the weights. In this paper we experimentally investigate the use of the correlation and diversity measures introduced in Sections 2 and 3 for setting the weights. All of the weights in this study are fixed in the sense that none change as a function of the particular pattern being classified. In [16] this is categorized as implicit versus explicit data-dependence and in [17] as constant versus non-constant weighting, the latter implying a partitioning of the input space. Input space partitioning is also implied by classifier selection, and a useful discussion of selection versus fusion appears in [18]. As an example, the logarithmic weights used in Adaboost, although fixed/constant, are a logarithmic function of the errors on the training set. An analysis of weighting is undertaken in [19], assuming unbiased and uncorrelated estimation errors, and the conclusion is that it is difficult to outperform unity weighting. It is generally recognized that a weighed combination may in principle be superior, but it is not easy to estimate the weights. A method of estimating weights for a linear combination of neural networks, based on minimizing classification error, is presented in [20].

We propose fixed weighting coefficients *(α_j j=1…B)*, all estimates made from training data, as follows

$$\alpha_j^{\sigma'} = \frac{1}{\tilde{K}_{\alpha^\sigma}} \sum_{m=1}^{\mu} \sum_{n=1}^{\mu} \{(y_{mj} \wedge y_{nj}) - (\bar{y}_{mj} \wedge \bar{y}_{nj})\} \tag{24}$$

$$\alpha_j^{N^{11'}} = \frac{1}{\tilde{K}_{\alpha^Q}} \sum_{m=1}^{\mu} \sum_{n=1}^{\mu} (y_{mj} \wedge y_{nj}) \tag{25}$$

$$\alpha_j^{A'} = 1 - \frac{1}{\tilde{K}_{\alpha^A}} \sum_{m=1}^{\mu} \sum_{n=1}^{\mu} \{(y_{mj} \wedge \bar{y}_{nj}) + (\bar{y}_{mj} \wedge y_{nj})\} \tag{26}$$

$$\alpha_j^{F'} = \frac{1}{\tilde{K}_{\alpha^{F'}}} \sum_{m=1}^{\mu} \sum_{n=1}^{\mu} (\bar{y}_{mj} \wedge \bar{y}_{nj}) \tag{27}$$

where *mth* and *nth* patterns in (24) - (27) are selected from different classes, $\omega_n \neq \omega_m$ and normalization constants $\tilde{K}$ are chosen so that weights sum to 1. The idea is to give more weight to classifiers that separate the two classes well. These weighting functions are compared in Section 6 with two schemes that use single-layer perceptron (slp) training to combine classifier outputs. Both schemes utilise

the full training set, the first being unconstrained slp (α_{slp})and the second having orientation weights fixed (α^{σ}) so that only slp bias is trained (α_{slp}^{σ}). It is to be expected that unconstrained slp will lead to over-fitting and this is confirmed in Section 6.

6. Experimental Evidence

The purpose of these experiments is to determine performance as the number of hidden nodes and number of training epochs of multi-layer perceptron (MLP) base classifiers are systematically varied. It is not the objective here to determine if combining classifiers with different complexity improves performance, and each node-epoch combination is repeated ten times with the same number of nodes and epochs. All other parameters of the base classifier MLPs are fixed at the same values over all runs. The number of hidden nodes is varied over (2-16) and number of training epochs over (1-32) and random perturbation of the MLP base classifiers is caused by different starting weights on each run.

Natural two-class benchmark problems have been selected from [21] and [22], and the experiments use random 50/50 or 20/80 training/testing splits. The natural datasets, including number of patterns, are Diabetes (768), Cancer (699), Ion (351), Heart (920), Vote (435), Credita (690), Card (690). For datasets with missing values the scheme suggested in [21] is used. The artificial data is Ringnorm, Twonorm, Threenorm from [15], and uses 300 training patterns and 3000 test patterns. All experiments are performed with one hundred single hidden-layer multi-layer perceptron (MLP) base classifiers, using the Levenberg-Marquardt training algorithm with default parameters (μ_{init}=0.001, μ_{dec}=0.1, μ_{inc}=10).

Figure 1 to Figure 5 shows Diabetes 50/50, a dataset that is known to over-fit with Boosting and other methods. Figure 1 shows MAJ, SUM and base classifier error rates; comparison of test and train error rates demonstrates that over-training of the majority vote classifier begins at fewer number of epochs compared with the base classifier. SUM and MAJ test error appear similar and the mean difference between SUM and MAJ over all datasets will be shown in Figure 11. Also in Figure 1 is shown the number (percentage out of ten runs) of significant differences of majority vote (McNemar 95%) with respect to best majority vote test error rate (2 nodes at 8 epochs). Note that all error rates are mean over ten runs and base classifier is also averaged over one hundred base classifiers.

Figure 2 shows mean values of Diversity measures Q, ρ, A, F defined in (14) to (18), along with N^{11} and $N^{11}*N^{00}*4$ (multiplied by 4 since max $N^{11}*N^{00}$ is 0.25) defined in (10) and (11). The decrease in N^{00} (F) over 8 to 32 epochs compared with the increase in N^{11} explains the decrease in $N^{11}*N^{00}$ and hence the peaking of Q at 8 epochs. Figure 3 shows Q', ρ', A', F' defined in (20), (21), and Figure 4 shows n^{+}, σ, σ', M, defined in (8), (9), (22), (23). Comparison of Figure 2 and Figure 4 with Figure 1 indicates that Q, σ, σ' may be correlated with over-fitting. On the other hand margin M, as expected, appears not to detect over-fitting. In order to quantify the ability of the measures to predict generalisation, correlation coefficients with respect to test errors are tabulated below in Table 1 to Table 8.

Figure 5 (a-b) shows Bias and Variance (Breiman definition Section 4.2) calculated on the test set. Since we need to know the Bayes classification to compute Bias and Variance, we make the optimistic assumption that the lowest majority vote test error rate (for this problem 2 nodes at 8 epochs) corresponds to Bayes classification. Bias is high for 1 to 2 epochs and low for 4-32 epochs. At 8 epochs variance is not affected much by number of nodes but increases 16-32 epochs, particularly for 16 nodes where over-fitting is most pronounced. Bias and Variance for the other datasets are given in [2], including artificial data which uses the true Bayes classification. Also shown in Figure 5 (c) is the standard deviation of MAJ test error rate and in (d) the standard deviation of σ'.

Figure 6 and Figure 7 show test errors, σ, σ' M, Q for Diabetes 20/80 and Cancer 50/50. Cancer 50/50 is shown not to over-fit even at 16 nodes, but Q still peaks indicating that it is not well correlated with test error for this problem. Most of the two class problems investigated in this paper, with the exception of Diabetes, appeared resistant to over-fitting for the number of nodes and epochs over which the experiments were run. This was even true for 20/80 training/testing split. To encourage over-fitting the experiments were repeated for 20% classification noise, in which 20% of patterns from each class are chosen at random and the target labels toggled. Figure 8 shows test errors, σ, σ' M, Q for Diabetes 20/80 with 20% classification noise, indicating over-fitting at 4, 8, 16 nodes. Similar curves were also produced for Ion, Heart, Vote, Credita, Card, ion for 50/50 and 20/80 splits both with and without 20% classification noise. Figure 9 demonstrates over-fitting of the base classifier for all the 20/80 datasets with 20% classification noise. In all natural and artificial datasets, when over-fitting occurred, MAJ over-fitted at a lower number of epochs compared with the base classifier. Measures σ, σ', which are best correlated with test error (see below Table 8), indicate that MAJ is optimised when there are more negatively correlated (equivalently fewer positively correlated) patterns compared with optimal base classifier. Curves for Twonorm, shown in Figure 10, demonstrate very well the difficulty of choosing base classifier complexity for optimal fusion. In particular, from 2 to 4 epochs the base classifier is close to optimal while Majority Vote error is increasing.

The correlation coefficients for base classifier test error are given in Table 1 to Table 6 for individual datasets. Each coefficient represents correlation with respect to number of epochs and is averaged over the four node settings. The mean correlation with respect to base classifier, MAJ and SUM over all datasets is given in Table 7 and Table 8. As can be seen from these tables, MAJ and SUM test errors appear to be less well correlated compared with the base classifier.

For Diabetes 50/50 (Table 1) the highest correlation with respect to base classifier is for σ, σ' Q. For this dataset, only these three measures showed up as significantly correlated (95%) when compared with random chance. However the mean for 50/50 datasets (Table 7) shows that Q overall is poorly correlated. In fact the mean over 50/50 datasets indicates that all except Q are well correlated. It is believed that this is due to the datasets being resistant to over-fitting. For example M for Diabetes 50/50 is relatively low compared with mean value of M over all 50/50 datasets (Table 7). For the situation where over-fitting is most prevalent, that is 20/80 datasets with 20% classification noise (Table 8), σ, σ' are the most highly correlated.

Weighted combination test errors, defined in Section 5, are shown in Figure 11 and Figure 12. Each plot shows the mean difference over all datasets between weighted and MAJ test error rate, so that negative values indicate that the respective weighted combination is superior. The mean is taken over all nodes and all datasets for the specified number of epochs. For 50/50 datasets all weighted combination schemes except α_{slp} and α^F give lower test error than MAJ, with α_{slp} appearing to over-fit as number of epochs is increased. The improvement over MAJ and SUM is quite dramatic even at 1 epoch. For example, the weighted combination for Diabetes 50/50 at 1 epoch is within one percent of the best MAJ test error (23% at 2 nodes 8 epochs in Figure 1). Generally as number of epochs is increased the improvement of weighted combination over MAJ disappears. A comparison with Adaboost weighting scheme is given in [23].

7. Conclusion

The experiments reported in this paper demonstrate how various measures and test error vary with complexity of MLP base classifier. In order to quantify this relationship, correlation coefficients with respect to test error for varying number of training epochs was calculated for each dataset. The mean correlation coefficients for artificial and real data show that the proposed pair-wise measure, calculated over patterns rather than classifiers, is well correlated with base classifier test error and warrants further investigation. The results suggest that it may be possible to select base classifier complexity to minimise

mean base classifier test error based on information extracted from the training set. MAJ or SUM is generally optimised at fewer number of epochs compared with base classifier, but design of optimal fusion based on the proposed measure may be feasible if a relationship between base classifier and fused test error can be established. An alternative strategy is to use a weighted combination, which is shown in this paper to be much less sensitive to the number of training epochs. Further work is aimed at applying these measures to multi-class problems, by incorporating (Error Correcting) Output Coding [24], which decomposes multi-class into a set of complementary two-class problems.

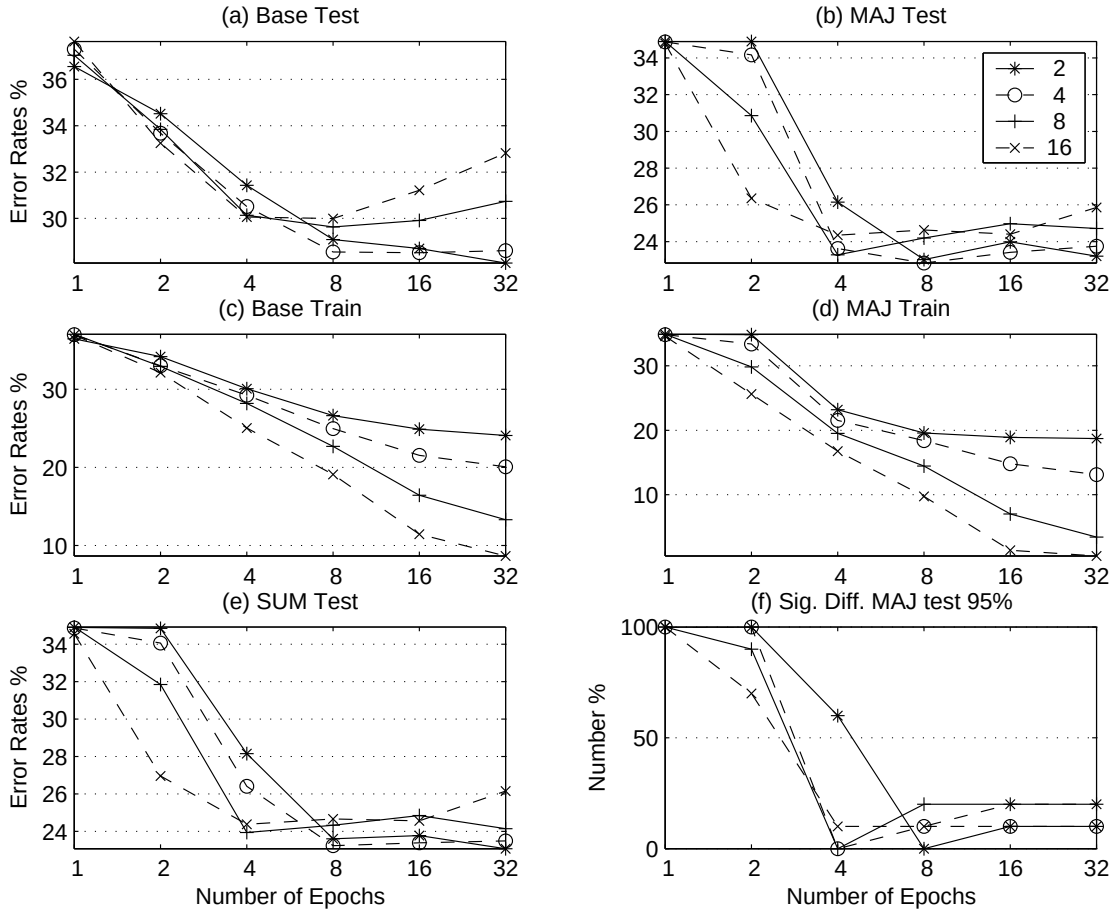

Figure 1: Diabetes 50/50 (a-d) Mean Error rates Test and Train for MAJ and Base Classifier (e) Mean SUM and (f) Sig. Differences for MAJ test error with respect to MAJ error at 2 nodes and 8 epochs (all graphs show 2-16 nodes)

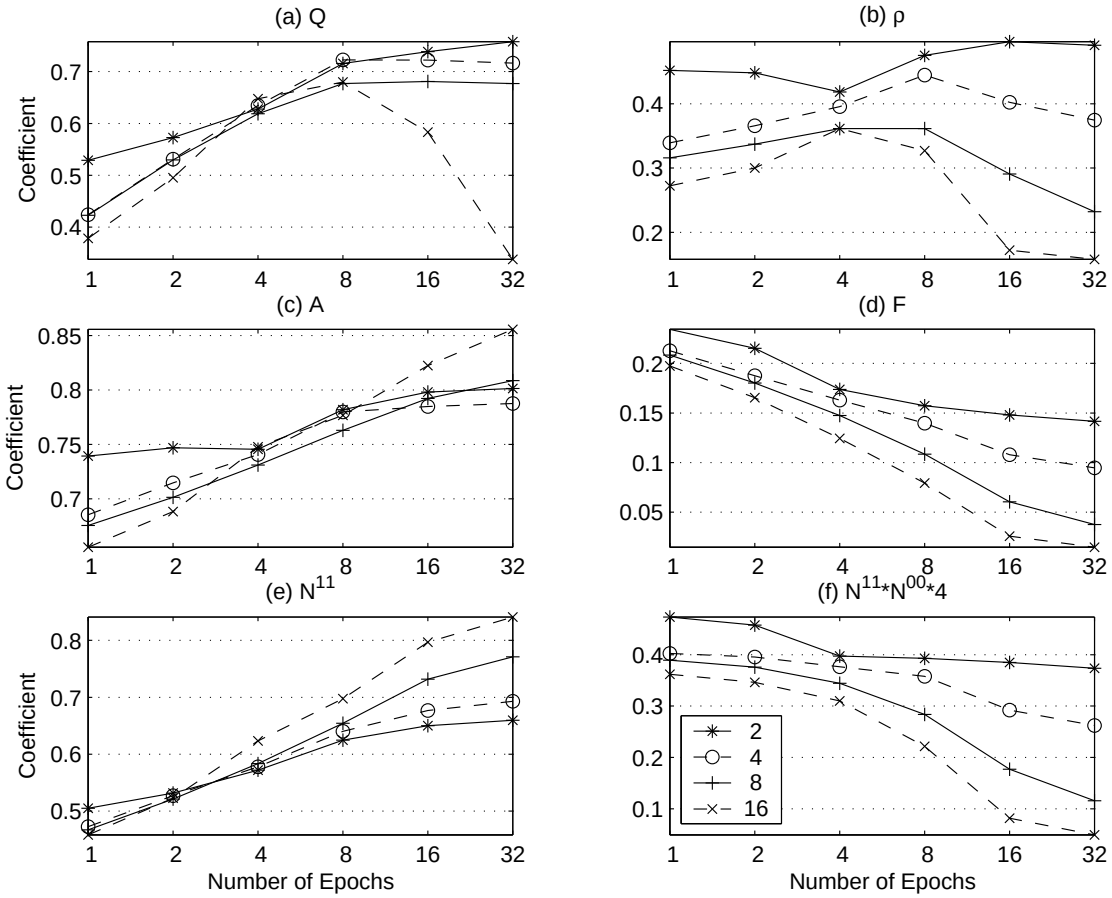

Figure 2: Diabetes 50/50 (a-d) conventional Diversity measures Q, ρ, A, F (e) N^{11} and (f) $N^{11}*N^{00}*4$ (all graphs show 2-16 nodes)

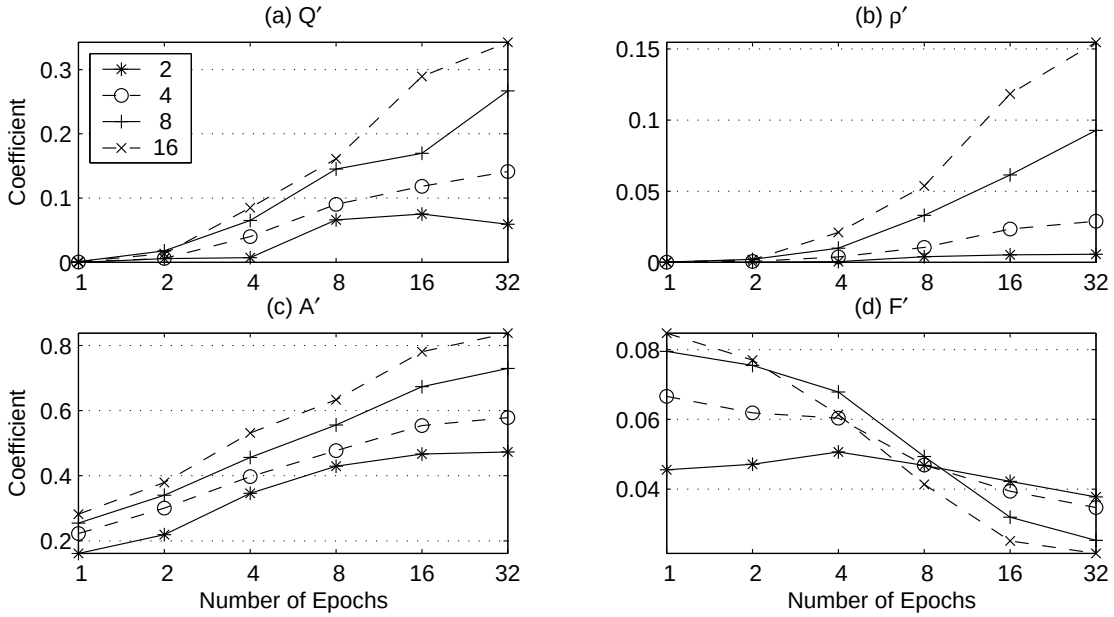

Figure 3: Diabetes 50/50 Diversity measures computed over patterns Q′ ρ′, A′, F′ (all graphs show 2-16 nodes)

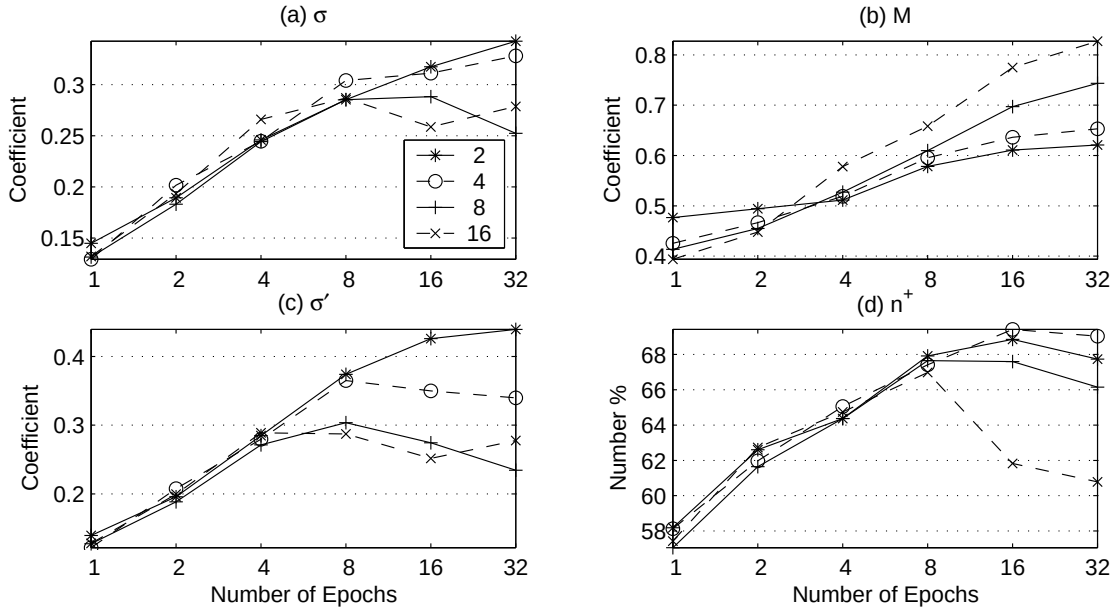

Figure 4: Diabetes 50/50 (a) spectral measure σ and (c) σ′ without Hamming Distance assumption, b) Margin M and (d) number of positively correlated patterns n^+ (all graphs show 2-16 nodes)

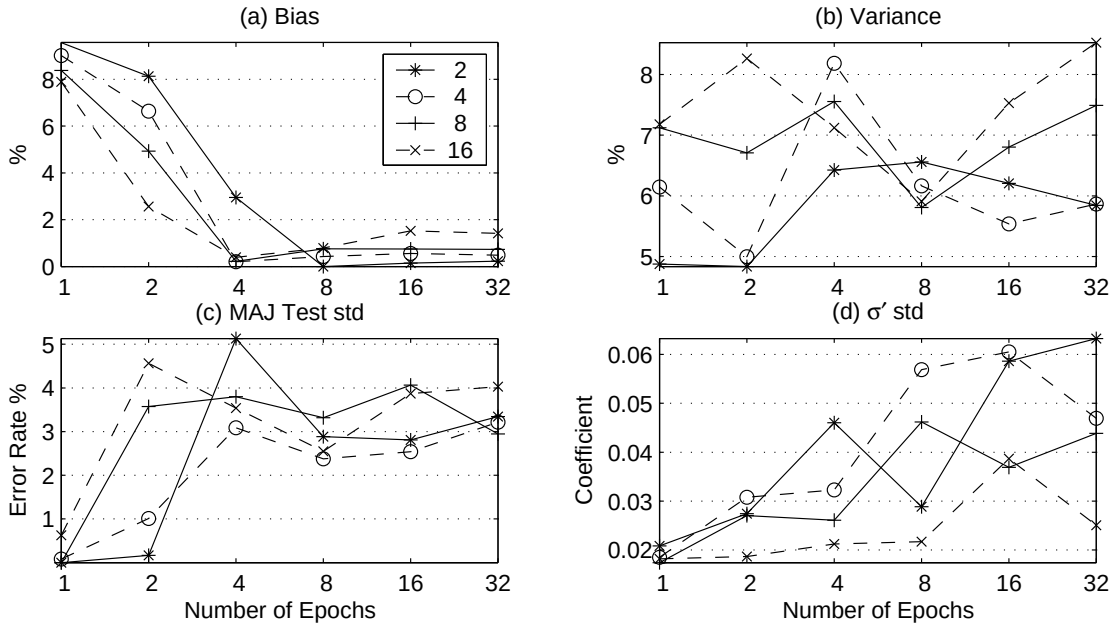

Figure 5: Diabetes 50/50 (a-b) Bias, Variance (c-d) standard deviation for MAJ test error and σ′ (all graphs show 2-16 nodes)

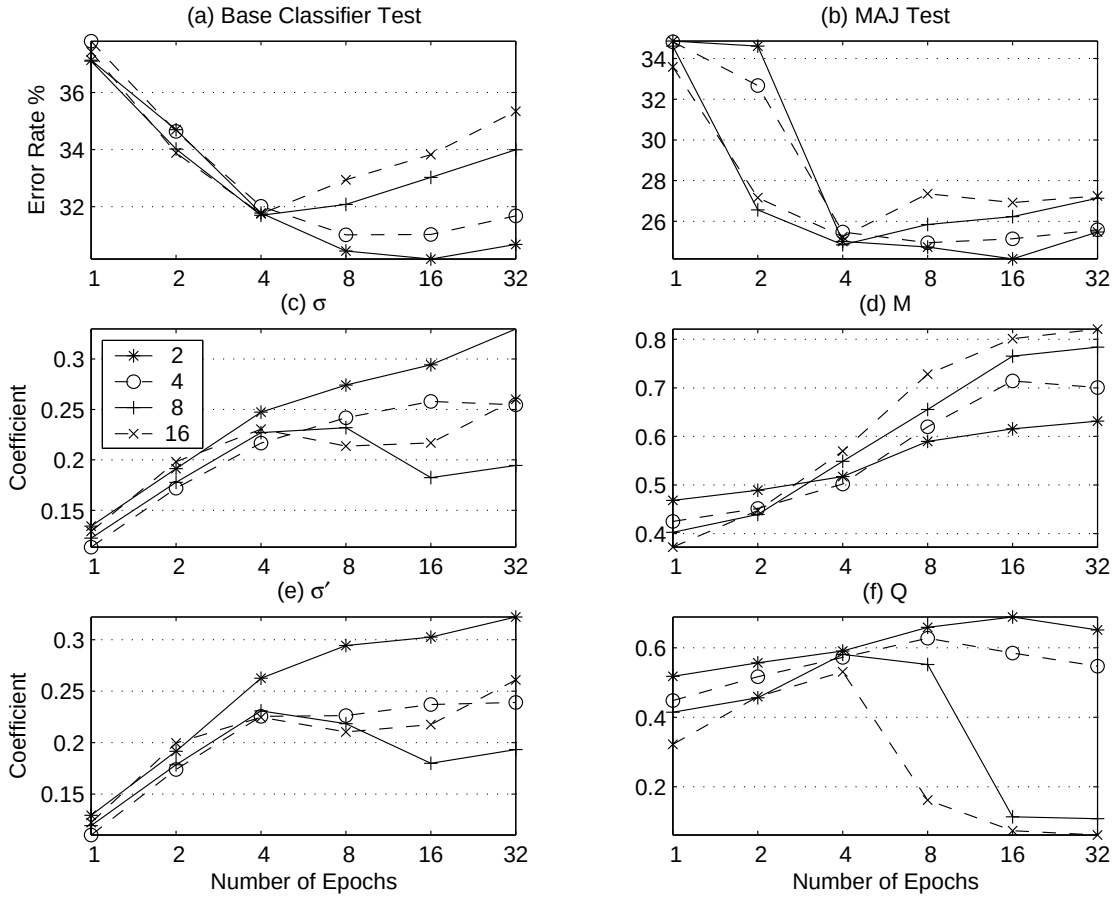

Figure 6: Diabetes 20/80 Summary graphs (a-b) Test error for MAJ and base classifier (c) spectral measure σ (d) Margin M (e) spectral measure σ′ and (f) Diversiy Measure Q (all graphs show 2-16 nodes)

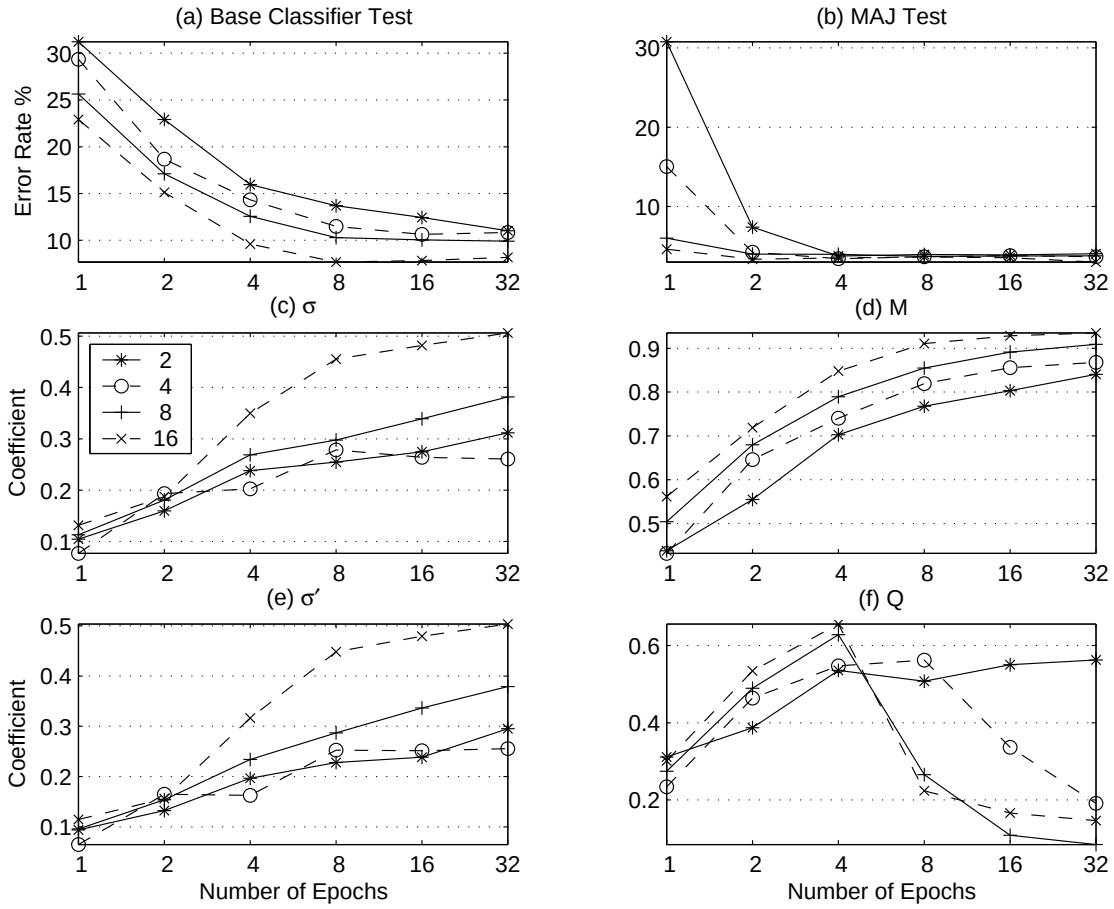

Figure 7: Cancer 50/50 Summary graphs (a-b) Test error for MAJ and base classifier (c) spectral measure σ (d) Margin M (e) spectral measure σ′ and (f) Diversiy Measure Q (all graphs show 2-16 nodes)

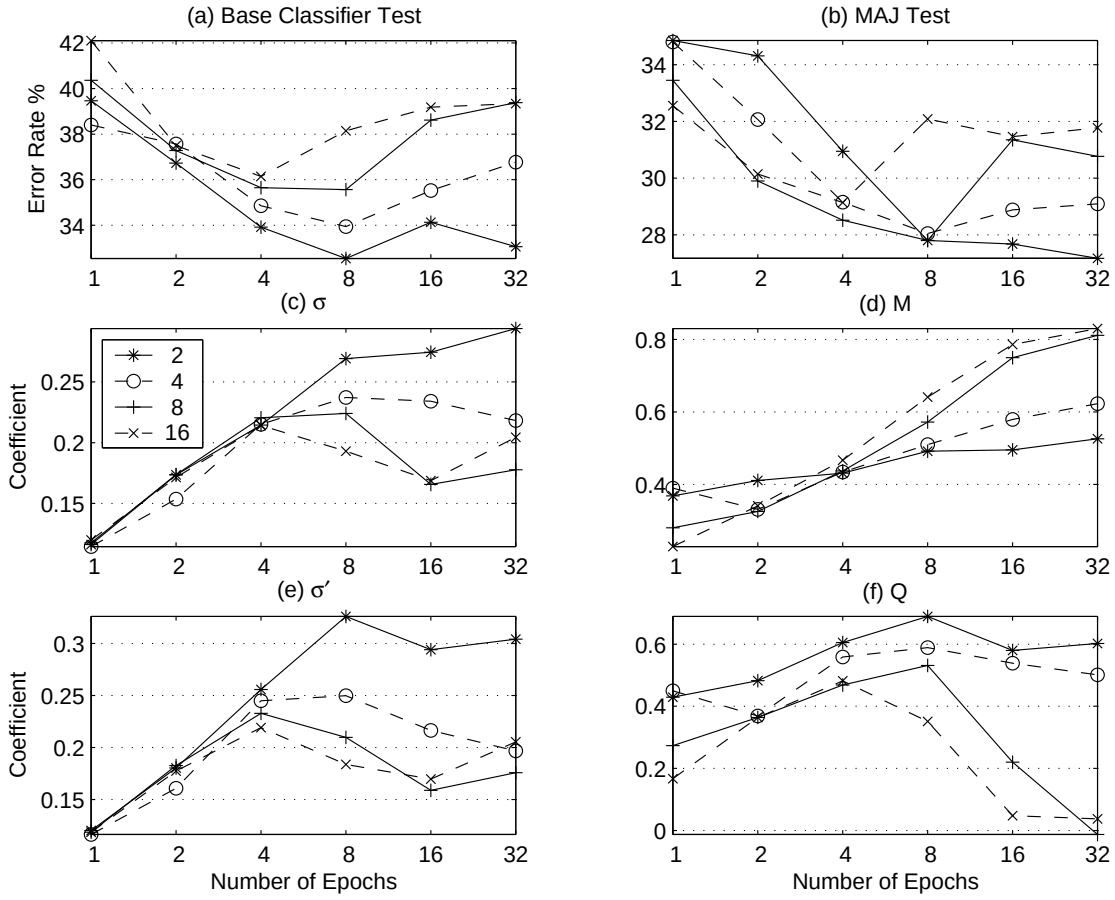

Figure 8: Diabetes 20/80 + 20% classification noise Summary graphs (a-b) Test error for MAJ and base classifier (c) spectral measure σ (d) Margin M (e) spectral measure σ′ and (f) Diversiy Measure Q (all graphs show 2-16 nodes)

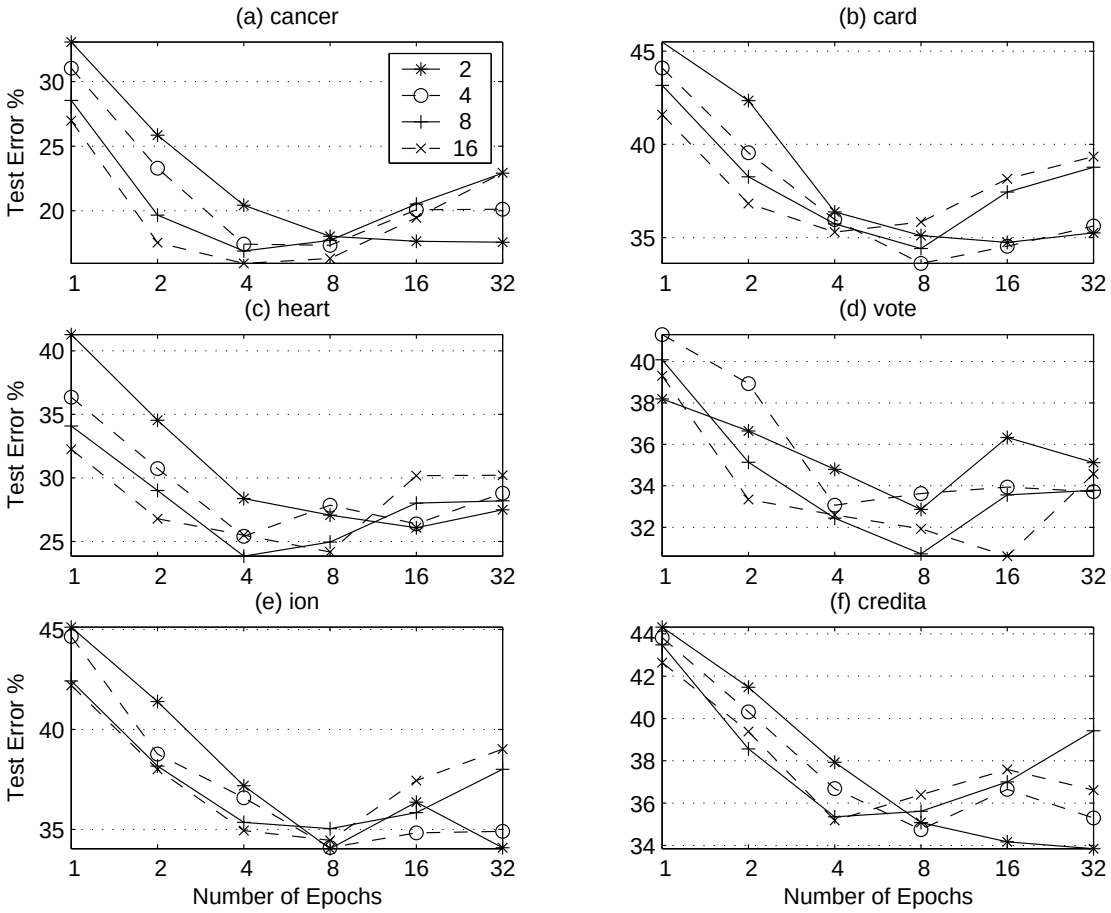

Figure 9: Mean Base Classifier test error for 20/80 datasets with 20% classification noise (all graphs show 2-16 nodes)

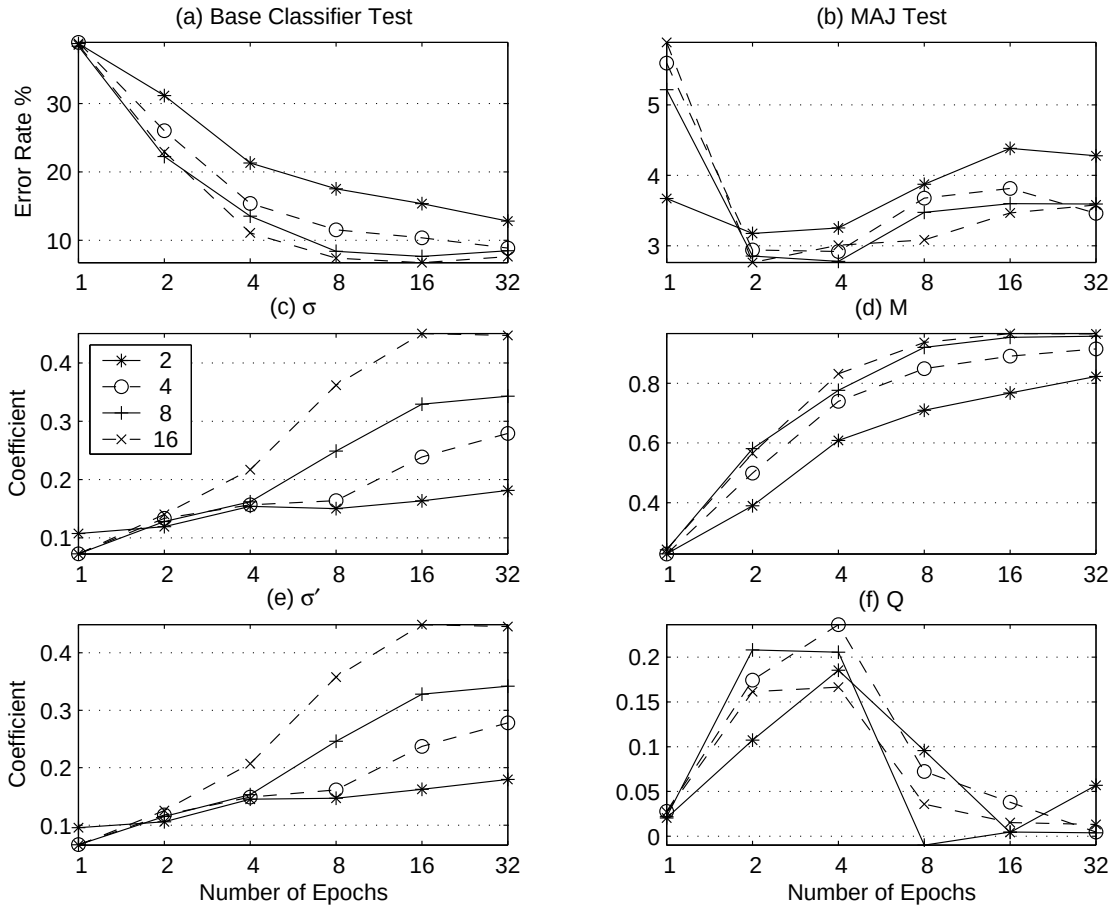

Figure 10: Twonorm Summary graphs (a-b) Test error for MAJ and base classifier (c) spectral measure σ (d) Margin M (e) spectral measure σ′ and (f) Diversiy Measure Q (all graphs show 2-16 nodes)

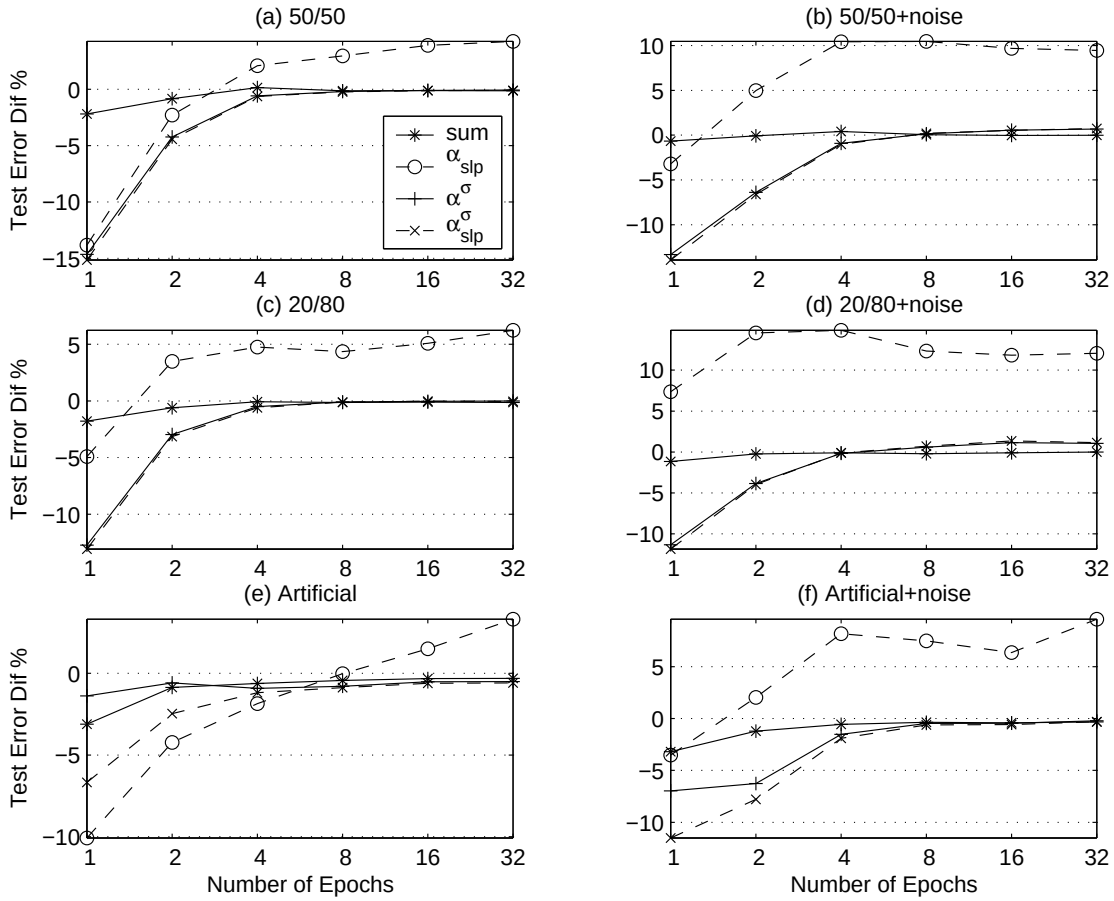

Figure 11: Mean Difference between (SUM, α_{slp}-wtd, α^{σ}-wtd, $\alpha_{slp}{}^{\sigma}$-wtd) and MAJ test error over all 50/50, 20/80 and Artificial datasets with and without 20% classification noise

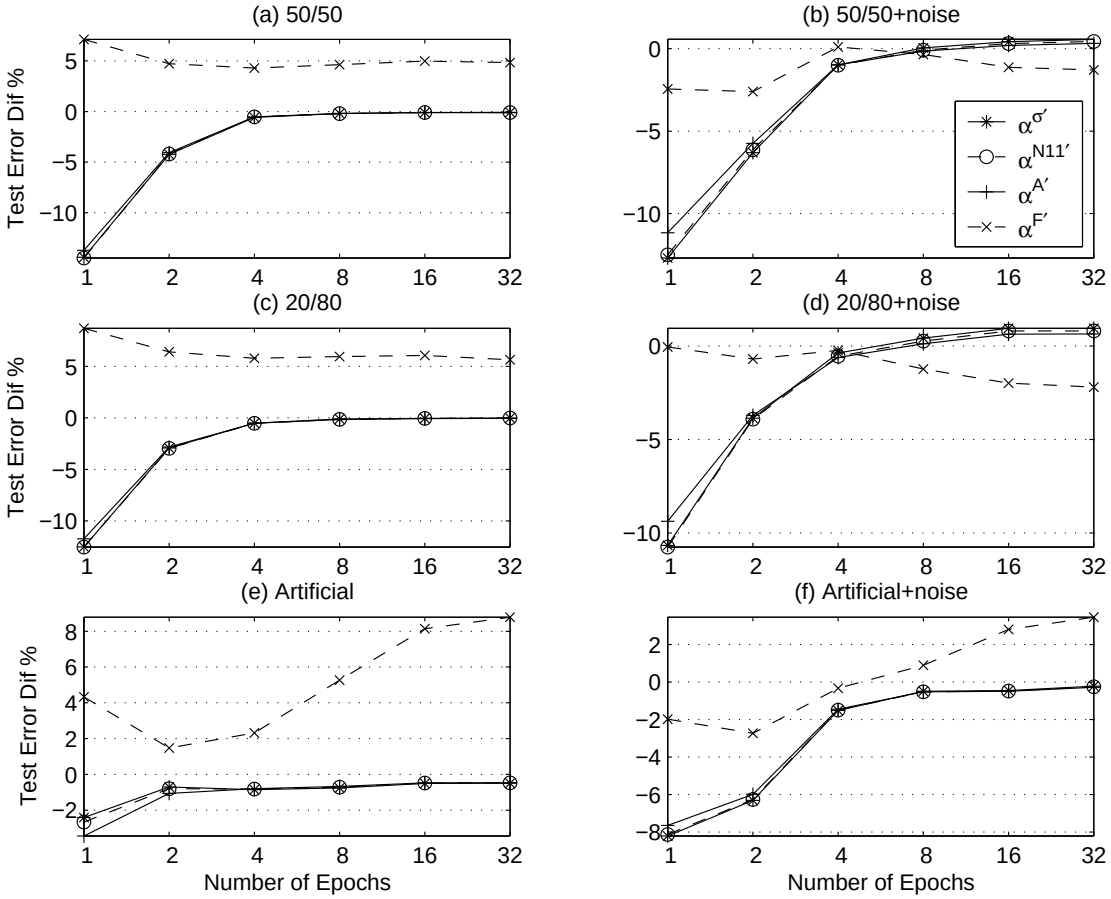

Figure 12: Mean Difference between various weighed vote schemes ($\alpha^{Q'}$-wtd, $\alpha^{n11'}$-wtd, $\alpha^{A'}$-wtd, $\alpha^{F'}$-wtd) and MAJ test error over all 50/50, 20/80 and Artificial datasets with and without 20% classification noise

	-Q	-A	F	-Q′	-A′	F′	-σ	-σ′	-M
Diabetes	94	82	81	72	84	65	96	97	78
Cancer	20	98	99	65	99	99	96	93	99
Heart	74	88	91	76	90	89	94	93	82
Vote	-61	96	99	81	99	99	89	89	98
Ion	-62	97	99	96	99	99	94	94	98
Credita	83	90	96	82	94	92	94	91	89
Card	80	90	96	83	94	92	92	90	89

Table 1: Correlation coefficients (x100) for base classifier test error with respect to epochs and averaged over nodes for 50/50 datasets

	-Q	-A	F	-Q′	-A′	F′	-σ	-σ′	-M
Diabetes	60	60	70	55	68	64	87	88	61
Cancer	-59	99	99	72	99	99	88	87	99
Heart	48	85	93	78	90	90	94	91	83
Vote	-96	98	99	83	99	99	89	89	98
Ion	-81	98	99	96	99	99	88	90	99
Credita	40	88	96	82	93	92	89	86	89
Card	32	92	98	87	96	95	92	89	93

Table 2: Correlation coefficients (x100) for base classifier test error with respect to epochs and averaged over nodes for 20/80 datasets

	-Q	-A	F	-Q′	-A′	F′	-σ	-σ′	-M
Diabetes	96	68	48	39	60	-30	88	97	44
Cancer	93	92	72	64	85	77	81	95	74
Heart	92	70	69	55	73	53	81	85	58
Vote	77	71	74	47	72	72	77	79	58
Ion	83	65	67	51	64	69	89	86	55
Credita	83	71	76	54	77	57	77	75	64
Card	87	74	79	61	79	61	78	77	67

Table 3: Correlation coefficients (x100) for base classifier test error with respect to epochs and averaged over nodes for 50/50 datasets + 20% classification noise

	-Q	-A	F	-Q′	-A′	F′	-σ	-σ′	-M
Diabetes	83	37	31	15	36	9	88	93	19
Cancer	80	76	44	41	59	60	66	81	46
Heart	75	56	70	46	65	61	72	70	53
Vote	75	61	67	41	63	69	75	75	53
Ion	7	74	79	64	75	80	84	80	71
Credita	42	69	82	54	78	73	83	79	68
Card	49	74	86	63	82	75	78	74	73

Table 4: Correlation coefficients (x100) for base classifier test error with respect to epochs and averaged over nodes for 20/80 datasets + 20% classification noise

	-Q	-A	F	-Q′	-A′	F′	-σ	-σ′	-M
Ringnorm	11	96	99	80	99	96	94	92	96
Threenorm	54	90	96	79	94	88	88	86	89
Twonorm	7	98	99	66	99	99	89	89	99

Table 5: Correlation coefficients (x100) for base classifier test error with respect to epochs and averaged over nodes for Artificial datasets

	-Q	-A	F	-Q′	-A′	F′	-σ	-σ′	-M
Ringnorm	48	85	94	68	92	75	91	89	82
Threenorm	63	75	82	57	82	59	84	82	69
Twonorm	78	69	72	48	73	54	74	75	60

Table 6: Correlation coefficients (x100) for base classifier test error with respect to epochs and averaged over nodes for Artificial datasets + 20% classification noise

	BASE CLASSIFIER			MAJ VOTE			SUM		
MEAS.	50/50	20/80	Artificial	50/50	20/80	Artificial	50/50	20/80	Artificial
-Q	33	-8	12	35	2	35	35	3	43
-ρ	21	-21	29	20	-14	45	18	-14	49
-A	92	88	95	76	70	62	74	67	48
F	95	94	98	85	83	76	84	80	62
-Q′	79	79	75	65	61	49	64	58	41
-ρ′	79	79	78	61	58	45	60	55	31
-A′	94	92	97	82	77	69	80	74	55
F′	91	91	95	78	77	65	76	74	52
-σ	94	90	90	81	78	62	78	74	48
-σ′	92	89	89	79	75	60	76	71	45
-M	90	89	94	77	73	64	75	69	49

Table 7: Mean correlation coefficient (x100) over all datasets with respect to base classifier, SUM and MAJ test error

	BASE CLASSIFIER			MAJ VOTE			SUM		
MEAS.	50/50	20/80	Artificial	50/50	20/80	Artificial	50/50	20/80	Artificial
-Q	87	59	63	68	54	59	69	54	55
-ρ	83	53	73	63	42	67	64	42	61
-A	73	64	76	40	13	57	42	8	45
F	69	66	82	37	18	63	38	13	52
-Q′	53	46	58	18	0	38	21	-4	29
-ρ′	44	39	53	12	-5	35	14	-9	26
-A′	73	65	82	41	17	64	43	13	51
F′	51	61	62	17	11	41	19	7	32
-σ	81	78	83	52	34	65	53	30	51
-σ′	85	79	82	57	36	65	57	31	51
-M	60	55	70	28	7	52	29	3	41

Table 8: : Mean correlation coefficient (x100) over all datasets+20% noise with respect to base classifier, SUM and MAJ test error

1 L. K. Hansen, P. Salamon, Neural Network Ensembles, IEEE Transactions on PAMI, 12(10), 1990, 993-1001

2 T Windeatt, Vote Counting Measures for Ensemble Classifiers, Pattern Recognition 36(12), 2003, 993-1001.

3 L. I. Kuncheva, C.J. Whitaker, Measures of diversity in classifier ensembles, Machine Learning 51, 2003, 181-207.

4 L. I. Kuncheva, M. Skurichina, R.P.W. Duin. An experimental study on diversity for bagging and boosting with linear classifiers, Information Fusion, 3 (2), 2002, 245-258.

5 Kuncheva L.I. A theoretical study on six classifier fusion strategies, IEEE Transactions on PAMI, 24(2), 2002, 281-286.

6 A. N. Tikhonov, V. A. Arsenin, Solutions of ill-posed problems, Winston & Sons, Washington, 1977.

7 T. Windeatt, Recursive Partitioning for combining multiple classifiers, Neural Processing Letters 13(3), 2001, 221-236.

8 S. L. Hurst, D. M. Miller, J. Muzio, Spectral Techniques in Digital Logic, Academic Press, 1985.

9 T. Windeatt, R. Tebbs, Spectral technique for hidden layer neural network training, Pattern Recognition Letters 18(8), 1997, 723-731.

10 L. I. Kuncheva, That elusive diversity in classifier ensembles, Proc. Iberian Conf. On Pattern Recognition and Image Analysis, Mallorca, Spain, Lecture Notes in Computer Science, Springer-Verlag, 2003, 1126-1138.

11 R. E. Schapire, Y. Freund, P. Bartlett, Boosting the Margin: a new explanation for the effectiveness of voting methods, The Annals of Statistics 26(5), 1998, 1651-1686.

12 V. Koltchinskii, D. Panchenko, Empirical margin distributions and bounding the generalisation error of combined classifiers, The Annals of Statistics 30(1), 2002, 1-50.

13 G. James, Variance and Bias for General Loss Functions, Machine Learning 51, 2003, 115-135.

14 E. B. Kong, T. G. Dietterich, Error- Correcting Output Coding corrects Bias and Variance, 12th Int. Conf. Machine Learning, San Francisco, 1995, 313-321.

15 L. Breiman, Arcing Classifiers, The Annals of Statistics 26(3), 1998, 801-849.

16 M. S, Kamel, N. M. Wanas, Data Dependence in Combining Classifiers, Proc. of 4th Int. Workshop on Multiple Classifier Systems, Eds: T .Windeatt, F. Roli, Guildford, UK, Lecture Notes in Comp. Science, Springer Verlag, June 2003, 1-14.

17 C. J. Merz, Using correspondence analysis to combine classifiers, Machine Learning, 36(1-2), 1999, 33-38.

18 L. I. Kuncheva. Switching between selection and fusion in combining classifiers: An experiment, IEEE Transactions on SMC, Part B, 32 (2), 2002, 146-156.

19 G. Fumera, F. Roli, Linear Combiners for Classifier Fusion: Some Theoretical and Experimental Results, Proc. of 4th Int. Workshop on Multiple Classifier Systems, Eds: T .Windeatt, F. Roli, Guildford, UK, Lecture Notes in Comp. Science, Springer Verlag, June 2003, 74-83.

20 N. Ueda, Optimal Linear Combination of Neural Networks for Improving Classification Performance, IEEE Trans. PAMI, 22(2), Feb 2000, 207-215.

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

22 C.J. Merz , P. M. Murphy, UCI repository of machine learning databases, 1998 , http://www.ics.uci.edu/~mlearn/MLRepository.html

23 T. Windeatt, R. Ghaderi, G. Ardeshir, Spectral Coefficients and Classifier Correlation, Proc.of 4th Int, Workshop on Multiple Classifier Systems, Eds: T .Windeatt, F. Roli, Guildford, UK, , Lecture Notes in Comp. Science, Springer Verlag, June 2003, 276-285.

24 T. Windeatt, R. Ghaderi, Coding and Decoding Strategies for multiclass learning problems, Information Fusion 4(1), 2003, 11-21.