Journal of Machine Learning Research 4 (2003) 39-66 Submitted 3/02; Revised 10/02; Published 4/03

Designing Committees of Models through Deliberate Weighting of Data Points

Stefan W. Christensen SWC@SOTON.AC.UK
Department of Chemistry
University of Southampton
Southampton, SO17 1BJ, UK

Ian Sinclair IS1@SOTON.AC.UK
Philippa A. S. Reed PASR1@SOTON.AC.UK
Materials Research Group
School of Engineering Sciences
University of Southampton
Southampton, SO17 1BJ, UK

Editor: Thomas G. Dietterich

Abstract

In the adaptive derivation of mathematical models from data, each data point should contribute with a weight reflecting the amount of confidence one has in it. When no additional information for data confidence is available, all the data points should be considered equal, and are also generally given the same weight. In the formation of committees of models, however, this is often not the case and the data points may exercise unequal, even random, influence over the committee formation.

In this paper, a principled approach to committee design is presented. The construction of a committee design matrix is detailed through which each data point will contribute to the committee formation with a fixed weight, while contributing with different individual weights to the derivation of the different constituent models, thus encouraging model diversity whilst not biasing the committee inadvertently towards any particular data points. Not distinctly an algorithm, it is instead a framework within which several different committee approaches may be realised.

Whereas the focus in the paper lies entirely on regression, the principles discussed extend readily to classification.
Keywords: Neural Networks, Ensembles, Committees, Bagging

1. Introduction

Identifying the true structure of a system via modelling is, in general, not an easy task. Any model will be limited in complexity and though ideally this will match that of the system exactly, in reality it rarely will. Obtaining a model amounts to choosing between two evils: reduce the complexity too much, and the model fails to duplicate the finer aspects of the system; allow too much flexibility, and the model introduces spurious aspects, overfitting the data.

Typically, this is addressed via regularisation, which allows the complexity to be adjusted. Unfortunately however, the architecture of the model may not be commensurate with the system, in which case no amount of regularisation will yield the perfect model. Moreover, even if the architecture is correct, the modelling may fail to provide the globally optimal model, instead settling for a locally optimal one.

An alternative approach is to establish a committee of models, with two main approaches normally being distinguished: ensemble- and modular techniques. In the case of ensembles (Hansen and Salamon, 1990), the predictions of a number of individual models are averaged. The expectation being that where some models will overshoot the target, others will predict too low a value, and so the ensemble prediction will be more accurate than the predictions of the individual models are on average. This has been established by Krogh and Vedelsby (1995). In the case of modular approaches, e.g. mixtures of experts (Jacobs et al., 1991), a number of individual models, each expected to be particularly accurate in certain respects but less so in others, complement one another so that their combination, via a gating network, may attain a high degree of accuracy everywhere.

The typical view of committees is that their prediction error can be decomposed into bias and variance; a more practical decomposition is that into average error of the constituent models, and their ambiguity (Krogh and Vedelsby, 1995). According to the latter, a clear prerequisite for the formation of a successful committee is that the individual models differ significantly from one another. In the case of mixtures of experts this is inevitable, by virtue of their definition. In the case of ensembles, it is not. In fact, it may be quite a formidable task to ensure that the individual models are not mimicking closely each others behaviour. In basic ensembles, where a straight average is taken over a number of generated models, many of these may be very similar. This problem may be reduced by weighting the models differently, e.g. by optimising the weights to obtain the lowest possible error (Perrone and Cooper, 1993), by using singular value decomposition to obtain the unique aspects of the models (Mertz, 1998), or by discarding individual models after an investigation of model collinearity (Hashem, 1997). Alternatively, models can be generated specifically to be different, *either* via encouraging diversity: e.g. bagging (Breiman, 1996), boosting (Shapire, 1990), through input feature grouping (Liao and Moody, 2000), or introduction of noise into the outputs (Breiman, 1998, Raviv and Intrator, 1996) (see e.g. Dietterich, 2000, for an overview of approaches), *or* via strictly enforcing it (Rosen, 1996, Opitz and Shavlik, 1996, Liu and Yao, 1998, 1999). The idea introduced in the present paper is closely related to bagging, and it is positioned firmly within the "encouraging diversity" category.

Following a brief explanatory note on the use of the word "weight", Section 2 makes the case for the designed committee and presents the theory. Section 3 concerns a very real problem that the approach may entail, and a way to circumvent it. Section 4 shows some practical test results, comparing the designed committee approach and bagging. In Section 5, the method and results are discussed, and conclusions, finally, are in Section 6.

1.1 A Note on Weights

In this paper, frequent use will be made of the word "weight". Three separate kinds of weights are to be distinguished; the first of these, the well established "weights" of a neural network, whose final determination concludes model optimisation, will not concern us here. The second kind is the *model* weights that determine the contribution of the individual models to an ensemble's prediction. These will be referred to as w_m, a vector of as many elements as there are models in the ensemble.

The third (and in this paper central) kind are the *data point* weights that determine the influence exercised by the individual data points on the model formation (to be referred to via D, the design matrix), and on the committee formation (referred to as w_p, a vector of as many elements as there are training data).

2. Committee Design[1]

Whilst the weighting of the individual models has been a central issue in committee based modelling so far, the possibility of weighting the influence of individual data points in a rigorous and deliberate manner seems to have been somewhat neglected, if not completely untried. One approach, which depends intrinsically on such weighting is boosting, via which emphasis is put on certain data points that are observed to be particularly difficult to predict accurately. This makes boosting an adaptive technique, i.e. the weighting is a result of the modelling. In the present paper a new, systematic approach to the formation of committees is proposed, whereby these are formally *designed*, by way of *premeditated* weighting of points.

2.1 Non-egalitarian Weighting Schemes

In bagging, or bootstrap aggregating, one of the most popular committee techniques, models are obtained through applying the modelling algorithm several times on data sets, each generated via random sampling, with replacement, of N data points from the original training set, also of N data points. The models are then combined into a committee, which benefits from the diversity introduced by the difference in training data sets. Breiman (1999) argued that "Bagging samples each instance with equal probability - it is perfectly democratic." While it is certainly true that each data point is chosen for the training set of a model with equal probability (provided the selection algorithm used is truly random) and that from this perspective bagging may be said to be entirely democratic, the final committee is unlikely to put equal emphasis on each data point, as all data points are unlikely to be chosen for training equally often - unless an infinite (or at least very large) number of models is included in the committee. Or, in other words, the data distribution (over input space), seen by the committee, may differ from the empirical distribution, given by the original data set, if the number of models included is finite. To illustrate the degree to which actual committees may emphasise certain data points more than others, a series of bootstrapping experiments has been carried out here.

What was sought was an understanding of the extent to which the individual data points are given equal, or at least similar, weights in the committee, as a function of the number of models included in it, and as a function of the number of training data. The bootstrap experiment was as follows:

- *For a given number of models per committee, m, and a given number of data points, n, set up the resampled training sets.*
- *Count the number of instances of every original data point within these and, in turn, over the whole committee.*
- *Calculate, for each data point, the ratio of number of instances to the total number of data points in the committee (i.e. m x n). These ratios are the weights of the individual data points; their sum is unity.*
- *Calculate the relative standard deviation from the mean of these weights (std. dev. divided by the average weight size), A, and the ratio of the largest weight to the smallest, B.*

1. Liao and Moody (2000) suggested a way to generate "designed committees" which, aside from sharing the objective of increasing model diversity, is unrelated to the approach presented here.

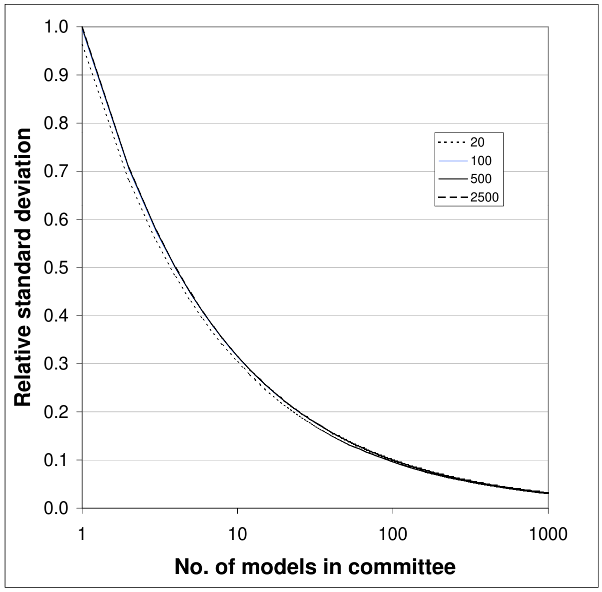

Figure 1: $\langle A \rangle$ vs. number of models in the committee, for committees based on data sets of 20, 100, 500 and 2500 data points, respectively. The curves are almost entirely overlapping; only the curve for the 20 data point case stands slightly out

- *Repeat this procedure 1000 times, and let $\langle A \rangle$ and $\langle B \rangle$ be the averages of the 1000 A and B values respectively.*

The number of models per committee was varied between 1 and 1000, whilst the numbers of training examples in the training sets were 20, 100, 500 and 2500 respectively.

The results appear in Figs. 1 and 2. As can be seen, there is a great deal of difference in the weights, particularly when the number of models per committee is low. For fewer than 10 models per committee, $\langle A \rangle$ is larger than 30% of the average weight size, irrespective of the number of data points available. For 100 models/committee, it is still around 10%. Clearly, therefore, there is a great deal of scatter in the weights.

$\langle B \rangle$ shows a similar pattern, where for fewer than a certain number of models the ratio becomes infinite; i.e. one or more data points are not used at all for training the committee. The number of models required, to ensure all data points are used, depends on the number of data points available; the higher this number, the higher the number of models required.

When 100 models are included, the most frequently used data points are given twice the weight of the least used ones, when the number of data points is 2500; for data sets with only 20 data points, the ratio becomes around 1.5.

The conclusion is clear: unequal weighting is severe when the number of models included in the committee is small, whilst the weights are increasingly harmonised as the number of models included is increased. For the committee sizes most commonly quoted in the literature, typically 5-30 models, the individual data points will typically be receiving very uneven weights, with the complete exclusion of some data points being common amongst the smaller committees.

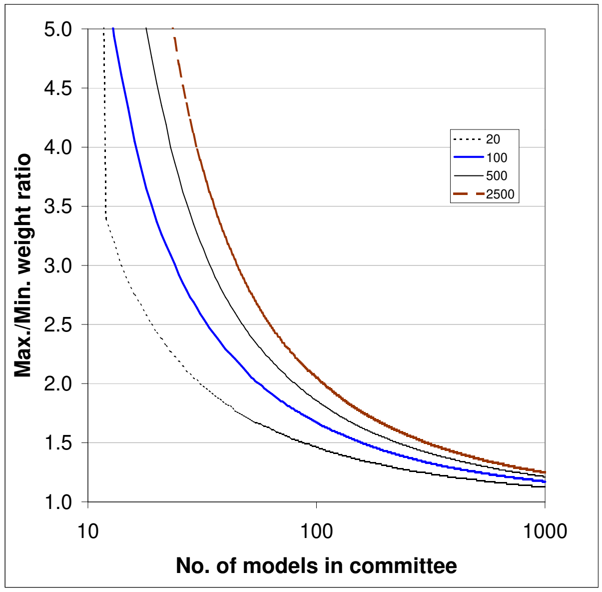

Figure 2: $\langle B \rangle$ vs. number of models in the committee, for committees based on data sets of 20 (lowest curve, dotted, black), 100 (second lowest, thick, blue), 500 (thin, solid black) and 2500 (uppermost, thick, dashed, brown) data points, respectively

Hence, bootstrapping is likely to emphasise certain, randomly chosen, data points.

In contrast, cross-validation used as a sampling principle, by virtue of its use of resampling *without* replacement, includes all data points equally for training. A hybrid version, which also finds widespread use: constructing training sets via random resampling, without replacement, of a pre-defined fraction of points (e.g. 90%) from the original data set, and repeated as many times as one wants models, does, however, also suffer from lack of control over the influence of each data point on the subsequent committee.

In the mixtures of experts approach, all training data points are employed equally; as the individual models are all trained in one go (in effect, as parts of one single, very large network). Thus the weighting problems due to resampling of the original data set do not apply; it is primarily in the ensemble methods, where resampling finds widespread use, that non-egalitarian weighting occurs.

2.2 The Design Matrix

Ideally, each data point should be given a weight in agreement with the prior confidence in it. In the event of all points being accredited with the same reliability, all data points should be weighted equally. In such resampling methods as bagging, equal weighting is only guaranteed in the limit of infinitely many models, though in practise a finite but possibly large number will do. A more efficient way to do this, providing this guarantee for any given number of models, is to use a committee design matrix.

The design matrix proposed here, is a matrix of m columns and n rows, m being the number of models to be created and entered into the committee, whilst n is the number of data points in the training set. The elements in the matrix are the weights for the data points, given the specific model; i.e. all weights are within [0,1] and the sum along each column must be unity. To each model is assigned a specific weight (e.g. a weight of 1/m; it must necessarily also reside within [0,1]); the sum of all model weights must come to 1. Finally, and hereby clearly distinguishing committee design from bagging, the sum, along every row, of elements multiplied by their respective models weight, must equal the overall (i.e. committee) weight for the corresponding data point. In the aforementioned case of all points being treated equally, all such row sums must equal 1/n. In short,

$$D\, w_m \;=\; w_p$$

where D is the design matrix, with w_m and w_p being the model weights, and overall (i.e. committee) data point weights. D is subsequently utilised in the *training* of the individual models via weighting the contributions to the error measure (e.g. the mean squared error, mse) by the corresponding weight from D; i.e. the error measure becomes:

$$L_j \;=\; \sum_{i=1}^{n} d_{i,j}\, L\big(y_i - f_j(x_i)\big) \tag{1}$$

where j is the index of the model, i is the index of the data point, $d_{i,j}$ is the weight from the design matrix, $f_j(x_i)$ is the prediction of model j for data point i, y_i is the true output value for data point i, and L is the loss function, e.g. the 2-norm; the mean squared error thus becomes a *weighted* mean squared error.

Ultimately, when making predictions, the committee is realized through a linear combination of the individual models:

$$y_{pred,\ committee} \;=\; w_m^T\, y_{pred} \;=\; \sum_{i=1}^{m} w_{m_i}\, y_{pred,i} \tag{2}$$

where $y_{pred,committee}$ is the predicted output value, y_{pred} is a vector of the individual model predictions, and w_{m_i} are the model weights.

Bagging committees are, in fact, mathematically extremely similar, and may be realised via this principle through relaxing the constraint that the row sums come to a predetermined amount; i.e. w_p's elements may take on unequal values between 0 and 1 (specifically, they must take on values that are multiples of 1/mn). Boosting committees however, cannot similarly be realised, even though boosting shares the aim of controlling the influence of each data point on each model; it is a truly adaptive technique as opposed to the proposed committee design. Interestingly, mixtures of experts may be nearly realised through careful construction of the design matrix, e.g. ensuring a specific model is particularly accurate in certain areas of input space, by weighting the data points in those areas higher than those elsewhere; the other models may be similarly designed to focus on other areas of input space. The proposed method of designing committees is therefore a very general approach that may encompass both ensemble and modular techniques.

2.3 Setting up a Design Matrix

A pertinent question, arising when designing a committee of m models using n data points, is how to set the weights for each data point, so that the individual models will weight different data points differently, thus encouraging model diversity, while the committee will weight each data point the same, thereby avoiding biasing the committee towards any particular data points. Clearly, this is a case of solving n equations with $m(n+1)$ unknowns which, as $m>1$, has an infinite number of solutions. As it is a linear system, these may be sought via Gauss-Jordan, singular value decomposition, etc. This may prove costly in terms of computer time and memory, however.

2.3.1 CONSTRUCTING QUASI-BAGGING COMMITTEES

A simple, yet functional way of quickly achieving a "randomised" distribution of weights and, as such, being close in spirit to bagging, whilst still adhering to the strictly equal weighting of each data point overall, is to initially give all m models the same weight (i.e. 1/m) while setting up a design matrix of equal weights, i.e. all weights, $d_{i,j} = 1/n$ and then, repeatedly for randomly chosen 2x2 sub matrices from the matrix, make random alterations to 1 of the 4 weights and changing the other 3 so that constant sums along both the 2 rows and the 2 columns are maintained. If repeated sufficiently, the weights seem to be quite "randomised".[2]

Consider this example (Equation 3): there are three data points and 2 models in the committee; i.e. we have a weight matrix with 6 elements, each with the initial weight of 1/3, both models having the same weight of 0.5. If a small change is made to e.g. element (1,1), then a similar change must be made to the element (1,2) but with opposite sign.[3] Also, either element (2,1) or, as chosen here, (3,1) must be changed by the same amount, also with the opposite sign. Finally, the last element, here (3,2), must be changed exactly as was (1,1) in order to maintain the same weight for all three data points within the committee; i.e. leaving w_p undisturbed.

$$D_{original} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} \end{pmatrix} \implies D_{altered} = \begin{pmatrix} \frac{1}{3}+\delta & \frac{1}{3}-\delta \\ \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3}-\delta & \frac{1}{3}+\delta \end{pmatrix} \tag{3}$$

This process is repeated for consecutive 2x2 sub matrices until the weights are adequately "randomised"; within the committee each data point will maintain its initial weight throughout.[4]

2.3.2 CONSTRUCTING CROSS-VALIDATION COMMITTEES

Instead of the randomness inherent to such quasi-Bagging committees, cross-validation may be used as the guiding principle in committee formation; here a binary weight distribution applies to each model - a data point receives a weight (in D) of either 0 or $1/n_j$, where n_j is the number of data points receiving non-zero weights in model j. A data point receiving a non-zero weight in one model must be given zero weight in all the others. As a consequence, this kind of weight distribution is extremely simple (and quick to perform), each element in D being either 0 or $1/n_j$, and each element in w_m given by:

2. No mathematical derivation of the nature of the weight distribution achieved via this approach is known to the authors. At present, we feel this is not of great concern, as the main objective initially is to ensure that each model weights the data points in its own unique way; numerous experiments have shown this to be easily achieved.
3. The constraint that all weights, $d_{i,j}$ must lie within [0,1] must be upheld, making it a constrained problem.
4. This is not very time-consuming; typically the time required for this will be a fraction of the time needed to construct a single model.

$$w_{m_j} = \frac{n_j}{\sum_{k=1}^{m} n_k}$$

The drawback, as is always the case with cross-validation, is the larger number of data points that is required. On the other hand, if many data points are indeed available, then a CV-committee may be advantageous because all the 0's in D are tantamount to calculations that need not be performed - hence model optimisation is quicker.

2.3.3 CONSTRUCTING MIXTURES OF EXPERTS-LIKE COMMITTEES

It is of course not merely random, and binary, weight-distributions that may be achieved with the committee design principle; distributions may be set up to accommodate diverse interests. This can be accomplished by introducing constraints in the process of assigning weights rather than, as in the case of quasi Bagging, allowing random weight distributions or, as in the case of CV-committees, fixing all weights according to a very simple system. A modular committee can for example be realised through specifying individual regions in the input space (of any dimensionality up to that of the input space) for each model so that data points in the vicinity of the region of a model are assigned a higher weight in that model than those data points which are not in the same vicinity. This can obviously be implemented in many ways. The result is a committee, very similar in spirit to the mixtures of experts committees (Jacobs et al., 1991), bar the fact that the latter are obtained in an adaptive way; i.e. the regions are obtained as a result of the modelling, rather than specified in advance of it.

A possible algorithm for constructing such a committee is the following:

1. *Specify the focus regions (points, curves, etc.); one for each model in the committee.*

2. *Specify the functions, one for each model, that control the weight distribution over the different data points. Typically, these may involve some measure of distance between data point and relevant regions, such that points close to the focus region get higher weights than those more distant. There is, however, no need for this particular choice, indeed, any kind of function varying over the input space may be employed.*

3. *Calculate the initial design matrix, D_{ini}, similar to the conventional design matrix, except each element gives the weight, assigned by the appropriate weight distribution function, for the particular data point and model. Note, that the sums along the m columns need <u>not</u> be unity.*

4. *Used by itself, D_{ini} would probably fail to ensure overall committee weighting in agreement with w_p. The reason for this is that, by mere chance, certain data points may receive little weight in all models, if they happen to lie distant from all focus regions, whereas others that happen to lie close to one, or more, of these regions will receive larger weights in that/those corresponding models. To counterbalance this inequality, each row in D_{ini} must be multiplied by a individual factor so that, ultimately, a "fair" D is obtained:*

$$D = (g^T \, I) \, D_{ini}$$

Here, g is a point-weight-multiplicator vector of n elements, one for each data point, and I is the identity matrix of size n x n. Overall, the linear system to solve becomes:

$$D\, w_m \;=\; ((g^T\, I)\, D_{ini})\, w_m \;=\; w_p$$

where D_{ini} and w_p are known (and w_m can be prespecified as well) and g is the vector to determine.

5. *Again, training is conducted as before (Equation 1), with eventual committee predictions given by Equation 2.*

3. Enlarging an already Existing Committee

An apparent drawback with the designed committee is the apparent necessity of knowing the number of data points, and number of models to be included, in advance of actually performing the modelling. In reality, this is not the case however. Additional data points can be included, provided the committee is enlarged by additional models which include these new data points. Additional models can also be included, preferably not individually however; at least two must be added to the committee at a time, if the individual weighting of data points within each model is to be conserved.

3.1 Adding-in New Models, Based on the Original Data Points

A new super committee can always be formed from two existing committees (or, indeed, from two models, or from one committee and a single model) via linearly combining the outputs of the two committees; i.e.:

$$y_{super} \;=\; \rho\, y_1 \;+\; (1-\rho)\, y_2$$

where ρ is a number between 0 and 1, and y_1 and y_2 are the predictions of the two committees respectively. However, for the new super committee to observe the same overall weighting of the data points, both constituent committees must observe this, if at least one of them does. In other words, if the existing committee has a overall data point weighting vector of w_p, then the second committee, to be added in, must have that same weighting vector also, the w_p of the super committee being:

$$w_{p_{super}} \;=\; \rho\, w_{p_{original}} \;+\; (1-\rho)\, w_{p_{new}} \tag{4}$$

If just a single model is to be added into the committee, then this must observe the same overall data point weighting as the existing committee; however, if this is done repeatedly; i.e. one new model is added in after another, then these new models will all share the same design matrix (which in that case is actually a vector), and the encouragement to develop different models, inherent to the designed committee approach, is lost for these new models. In other words, it is better to add extra models in bulkwise, rather than one at a time.

The choice of ρ will not influence the overall weighting of data points, nor the model diversity, and therefore should reflect other concerns; these are beyond the scope of the present paper.

3.2 Adding-in Extra Data Points

Again, Equation 4 applies, but $w_{p,new}$ and $w_{p,original}$ are not of the same dimensionality. This is handled by recognising that all the extra data points received a weight of zero in the original committee; the new augmenting committee therefore must compensate for this by putting extra emphasis on these so that $w_{p,super}$ is in agreement with prior confidence in the (newly enlarged) data set. For the special case, where all data points are considered equal, symmetry makes finding $w_{p,new}$ particularly simple. Equation 4 takes on the following appearance:

$$w_{p_{super}} = \rho \begin{pmatrix} 1/n_1 \\ \dots \\ 1/n_1 \\ 0 \\ \dots \\ 0 \end{pmatrix} + (1-\rho) \begin{pmatrix} a \\ \dots \\ a \\ b \\ \dots \\ b \end{pmatrix} = \begin{pmatrix} 1/n \\ \dots \\ \dots \\ \dots \\ \dots \\ 1/n \end{pmatrix} \tag{5}$$

where n_1 is the number of data points in the original committee, and n is the corresponding number in the super committee. There are, therefore, only three variables to be determined: ρ, a and b. We get the following two equations from Equation 5.:

$$\rho \frac{1}{n_1} + (1-\rho)a = \frac{1}{n} \tag{6}$$

and

$$(1-\rho)b = \frac{1}{n} \tag{7}$$

Two equations with three unknown variables indicates some freedom of choice, however, there are three criteria that we must impose:

1. $\rho \in [0;1]$, giving the weighting of the original and new committees
2. $a \geq 0$, no data point should have a negative weight in a committee
3. $b > 0$, new data points must receive positive weight.

3.2.1 CRITERION 1

From Equation 7 we have:

$$\rho = 1 - \frac{1}{bn} \tag{8}$$

And, if $b > 0$, then $1/(bn) > 0$ and, in turn, $\rho < 1$. Moreover, we see that to ensure $\rho \geq 0$, we must demand that $1/(bn) \leq 1$, which is true if $b \geq 1/n$.

3.2.2 CRITERION 2

Through combining Equations 6 and 8 we get

$$a = (1 - \frac{n}{n_1})\, b + \frac{1}{n_1}$$

which we after a few trivial manipulations can see is larger than, or equal to, zero, provided $b \leq 1/(n-n_1)$.

3.2.3 CRITERION 3

This is fulfilled by definition.

Overall, we get the following three criteria for ρ, a and b:

$$\rho \in [0; n_1/n]$$
$$a \in [0; 1/n]$$
$$b \in [1/n; 1/(n-n_1)]$$

Any choice of ρ, a and b within these ranges, constrained by Equations 6 and 7, will lead to a correct $w_{p,super}$, and, as before, the specific choice of ρ (and hence a and b) shall not concern us here.

4. Experimental Results

In order to test the designed committee principle, a series of experiments was carried out comparing the proposed approach using the quasi-bagging weight distribution with bagging itself. The performance on five different problems was investigated, two of which are well-established artificial problems with considerable noise, from (Friedman, 1991). Two further problems are real life problems, and the final one is another artificial problem, but noiseless, adapted from Schwefel (1981).

4.1 Friedman problems

1. $y = 10sin(\pi x_1 x_2) \ + \ 20(x_3 - 0.5)^2 \ + \ 10x_4 \ + \ 5x_5 \ + \ n$
2. $y = (x_1^2 + (x_2 x_3 - (\frac{1}{x_2 x_4}))^2)^{0.5} + n$

where, in no. 1, n is N(0,1), whereas in no. 2 n is adjusted to give 3:1 ratio of signal power to noise power. In no. 1, there are five more variables, x_6 to x_{10}; i.e. five redundant inputs. All variables in no. 1. are uniformly distributed in [0,1], in no. 2 they are all uniformly distributed in the following ranges:

$$0 \leq x_1 \leq 100$$
$$20 \leq (x_2/(2\pi)) \leq 280$$
$$0 \leq x_3 \leq 1$$
$$1 \leq x_4 \leq 11$$

For both problems, three different sub problems were studied, differing by the number of data points included for training: 10, 40 and 70. These data set sizes were chosen to help elucidate whether there is a clear influence of the amount of information available about the underlying system from the data on the performance difference (if any exists) between the designed and bagged committees. There was no overlap between the data sets; i.e. the systems were resampled anew for each sub problem.

The modelling technique employed for both approaches, and all sub problems was, in summary:

Architecture: *MLP with 2 hidden layer neurons (a single hidden layer), and* tanh *activation function.*

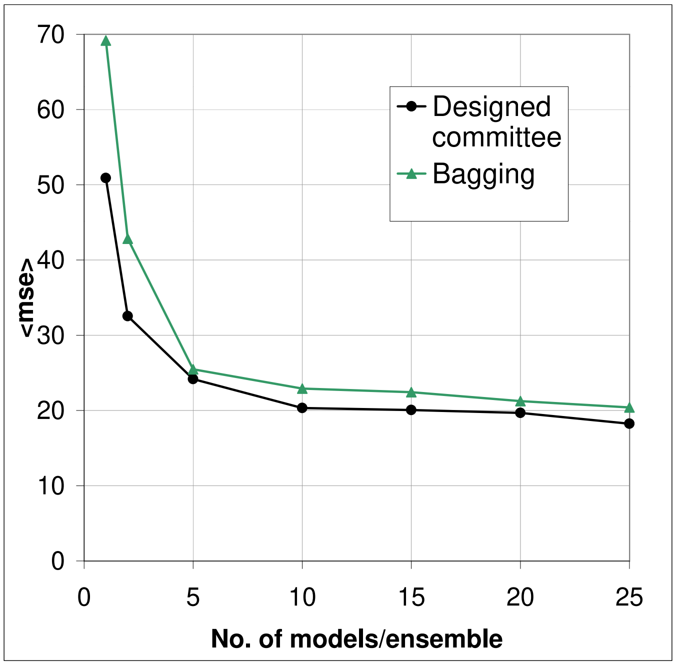

Figure 3: $\langle mse \rangle$ for "Friedman no. 1 (10 pts)"

Optimisation: *Stochastic hill-climber (to avoid landing in poor local optima).*

Training-error: *Mean-squared-error (weighted, in the case of the designed committee). No regularisation was applied.*

Test-error: *Mean-squared-error on a separate set of test data (5000 data points, uniformly sampled, noiseless).*

The entire series of experiments was, in summary:

Problems: *Friedman nos. 1 and 2*

Sub problems: *10, 40, and 70 data points respectively*

Committee-sizes: *2, 5, 10, 15, 20, and 25 models/committee*

Number of runs: *100 (for the 2 models/committee case), 40 (for the 5 models/committee case), 20 (for all the other committee sizes). These were chosen to obtain stable average test errors.*

Results reported: *Average test error, $\langle mse \rangle$ (over the specific number of runs performed). Standard dev. of test errors from $\langle mse \rangle$ (over same number of runs).*

The results are given in Figures 3-8 (Friedman no. 1) and 9-14 (Friedman no. 2).

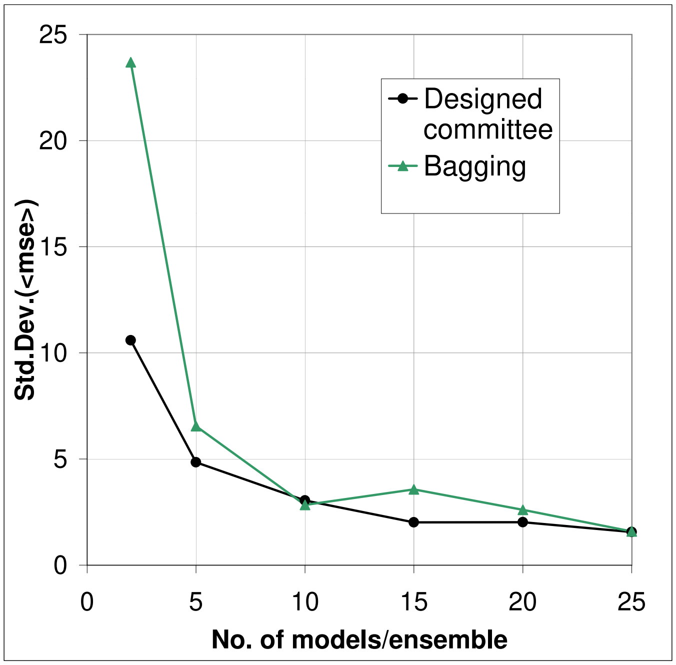

Figure 4: Std. Dev.($\langle mse\rangle$) for "Friedman no. 1 (10 pts)"

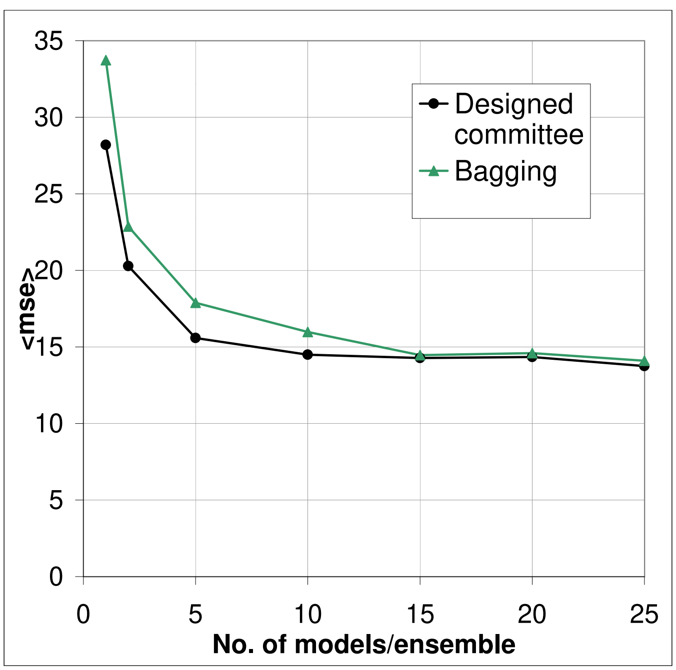

Figure 5: $\langle mse\rangle$ for "Friedman no. 1 (40 pts)"

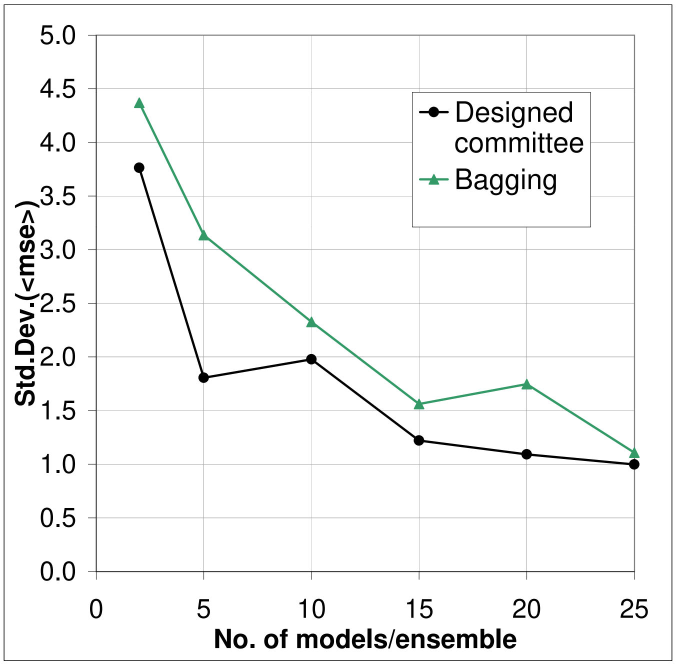

Figure 6: Std. Dev.($\langle mse \rangle$) for "Friedman no. 1 (40 pts)"

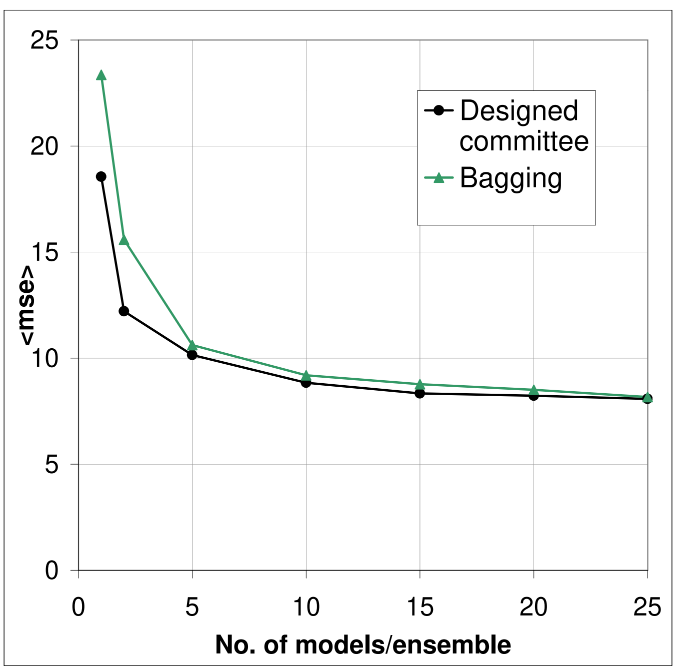

Figure 7: $\langle mse \rangle$ for "Friedman no. 1 (70 pts)"

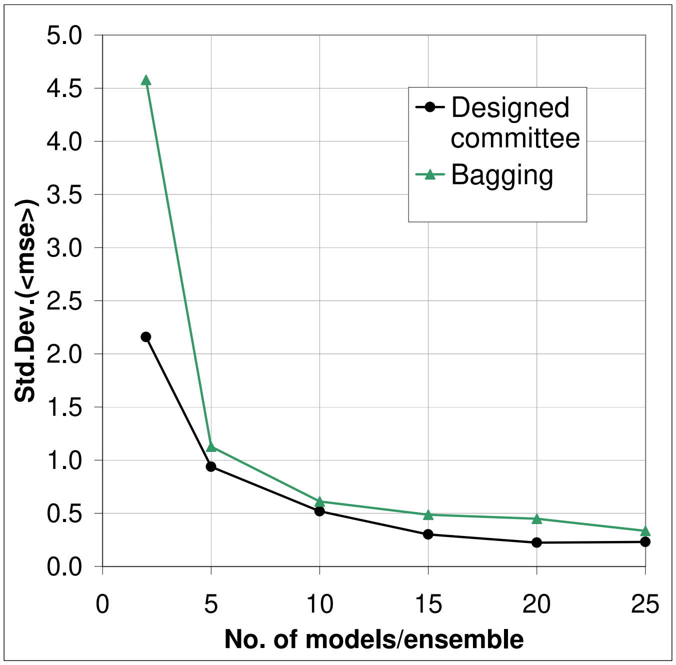

Figure 8: Std. Dev.($\langle mse \rangle$) for “Friedman no. 1 (70 pts)”

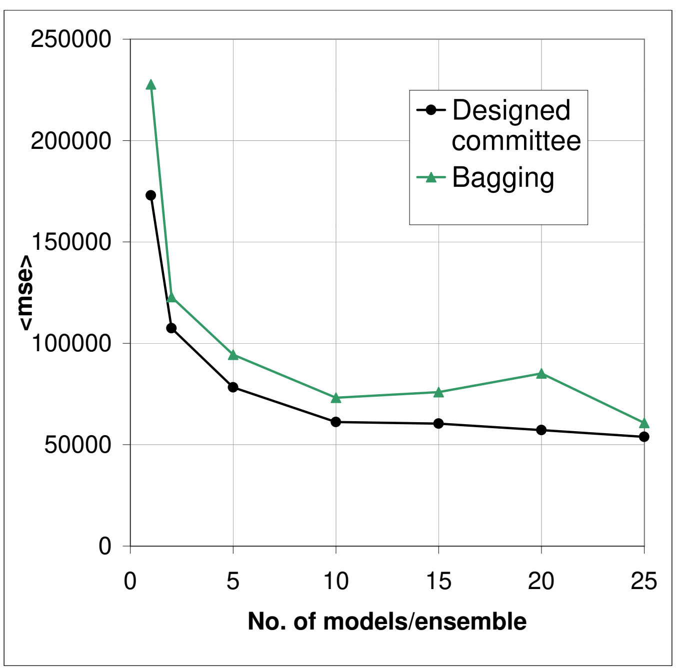

Figure 9: $\langle mse \rangle$ for “Friedman no. 2 (10 pts)”

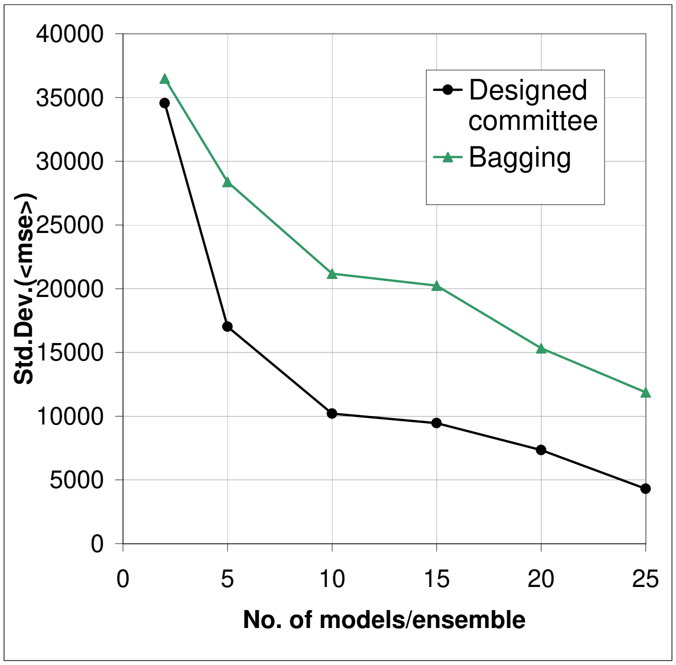

Figure 10: Std. Dev.($\langle mse \rangle$) for “Friedman no. 2 (10 pts)”

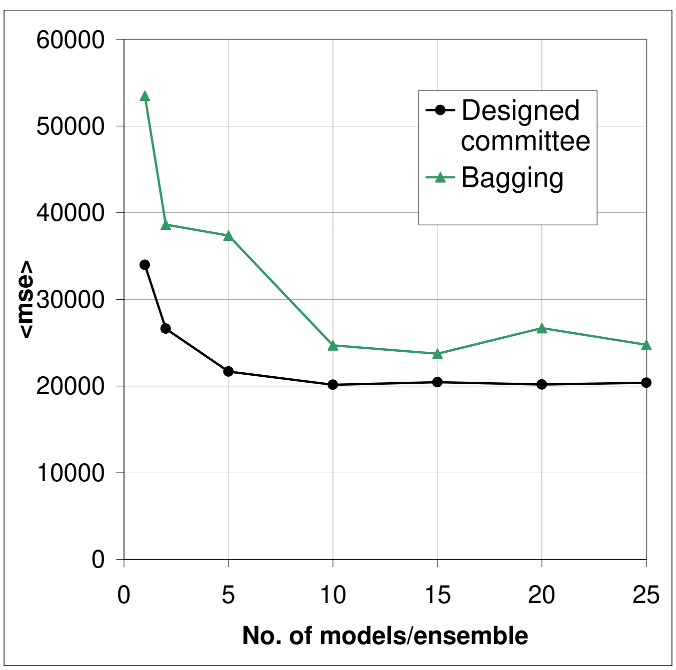

Figure 11: $\langle mse \rangle$ for “Friedman no. 2 (40 pts)”

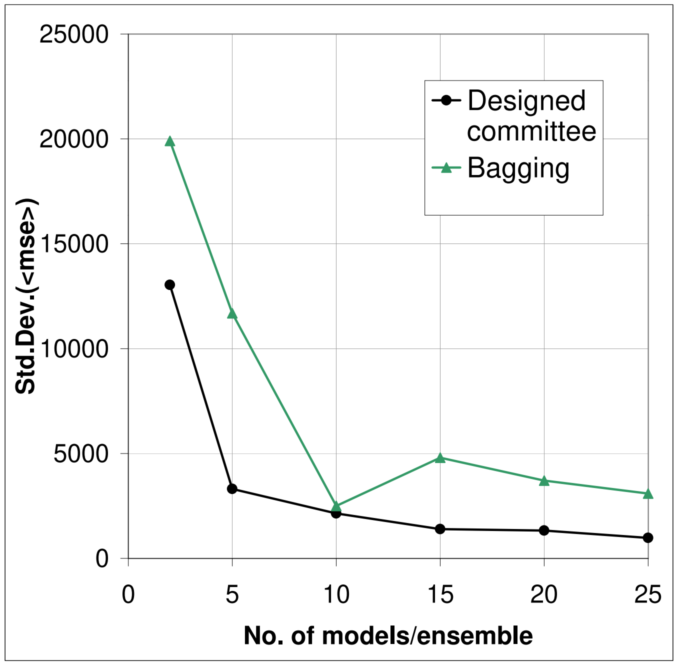

Figure 12: Std. Dev.($\langle mse \rangle$) for "Friedman no. 2 (40 pts)"

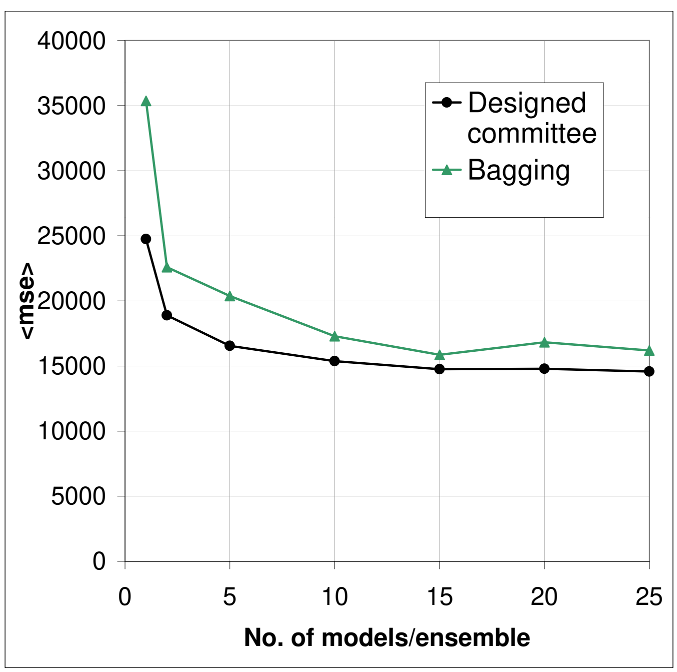

Figure 13: $\langle mse \rangle$ for "Friedman no. 2 (70 pts)"

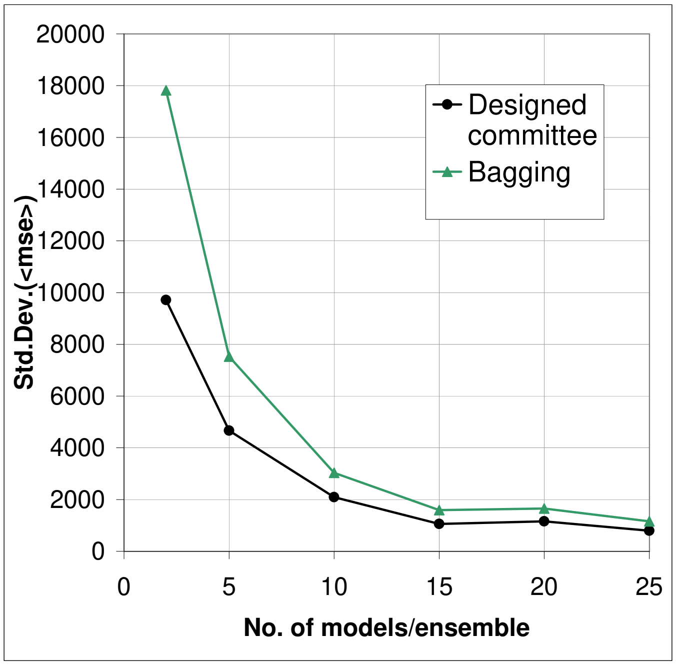

Figure 14: Std. Dev.($\langle mse \rangle$) for “Friedman no. 2 (70 pts)”

All curves, for both $\langle mse \rangle$ and the standard deviation, show the same qualitative behaviour; very nearly monotonicly dropping with increasing committee size. For the $\langle mse \rangle$, this is in agreement with expectation, theory, and with results from the literature. The drop in scatter is a less well established result.

Concerning the $\langle mse \rangle$, the pattern for all individual problems and sub problems is the same: at the smallest possible committee size, one single model, there is a great deal of difference between the bagging- and designed committee-accuracies, with bagging giving the less accurate results. At the other extreme, for large committees, the two different principles lead to committees of almost identical accuracy. In between, the advantage of the designed committee drops firmly as the committee size increases. For two of the sub problems (Friedman no. 1, 10 and 70 data points, respectively) bagging has all but completely caught up with the designed committee at 5 models per committee, with error curves running very closely, and in parallel, for all larger committees. In one (Friedman no. 1, 40 data points) bagging achieves this at 15 models per committee. For two of the Friedman no. 2 sub problems, bagging never catches entirely up with the designed committee principle, but the difference reaches a very small level at 10 models per committee (40 and 70 data points, respectively) and already at 2 models per committee for the smallest sub problem (10 data points), also for which, at 25 models per committee, the difference between bagging and the designed committee approach does seem to vanish.

There does not seem to be any clear effect of data set size on the difference between the two techniques, in terms of prediction accuracy.

The monotonic character is rigorous for all curves, except bagging in Friedman no. 2 (notably non-monotonic for the 40 data points sub problem).

The monotonicity is broken rather more severely for the scatter curves; for bagging this applies to both the 40 data points sub problems and one of the 10 data point ones (Friedman no. 1). It also applies to the designed committee on Friedman no. 1 (40 data points).

A general pattern, concerning the difference between bagging and the designed committee in terms of the scatter, is more difficult to identify. For both the 70 data point sub problems, the designed committees' have much less scatter at 2 models/committee, but bagging quickly catches up. The same seems to be the case for Friedman no. 1 (10 data points). For Friedman no. 1 (40 data points), however, no trend is obvious; for Friedman no. 2 (40 data points), bagging is only really close to the designed approach at 10 models/committee, with the latter method having much lower scatter for all other committee sizes. Finally, for Friedman no. 2 (10 data points), the two methods seems to have nearly the same scatter at 2 models/committee, with bagging having much higher scatter for all other committee sizes.

Overall it can be said that the designed approach never obtained an $\langle mse \rangle$ larger than the corresponding bagging committee, and only in one case (Friedman no. 1, 10 data points, 10 models/committee) did it have larger scatter (and only marginally so).

4.2 Ozone data set

The Ozone data set has, like the Miles per Gallon set in Section 4.3, been obtained from the UCI Repository Of Machine Learning Databases and Domain Theories (Murphy and Aha, 1994) but it is originally from (Breiman and Friedman, 1985). It contains 13 variables and 366 data points but, owing to missing data in the set, the number of inputs has been reduced to 8 (with one output) and 330 data points. It is a real life data set, concerning climate measurements.

The modelling undertaken here was as follows:

Architecture: *MLP with 3 hidden layer neurons (a single hidden layer), and* tanh *activation function.*

Optimisation: *Downhill simplex with multiple restart (15 runs, best model retained).*

Training-error: *Mean-squared-error (weighted, in the case of the designed committee). No regularisation was applied.*

Test-error: *Cross-validation, with 80% data for training and 20% for testing.*

The entire series of experiments was, in summary:

Committee-sizes: *1, 2, 5, 10, 20, and 40 models/committee*

Number of runs: *200 (for the 1 model/committee case), 100 (for the 2 models/committee case), 40 (5 models/committee), 20 (10 models/committee), 10 (20 models/committee) and 5 (40 models/committee). These were chosen to obtain stable average test errors.*

Results reported: *Average test error, $\langle mse \rangle$ (over the specific number of runs performed). Standard dev. of test errors from $\langle mse \rangle$ (over same number of runs).*

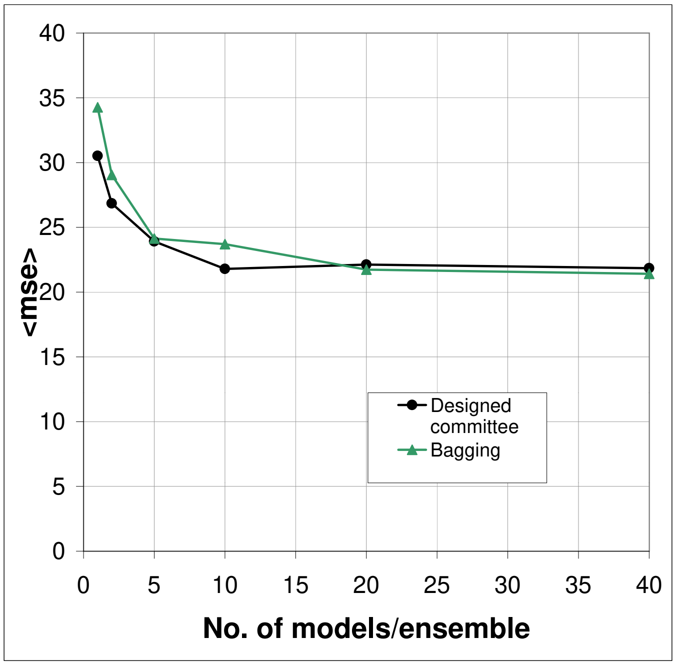

Figure 15: $\langle mse \rangle$ for “Ozone”

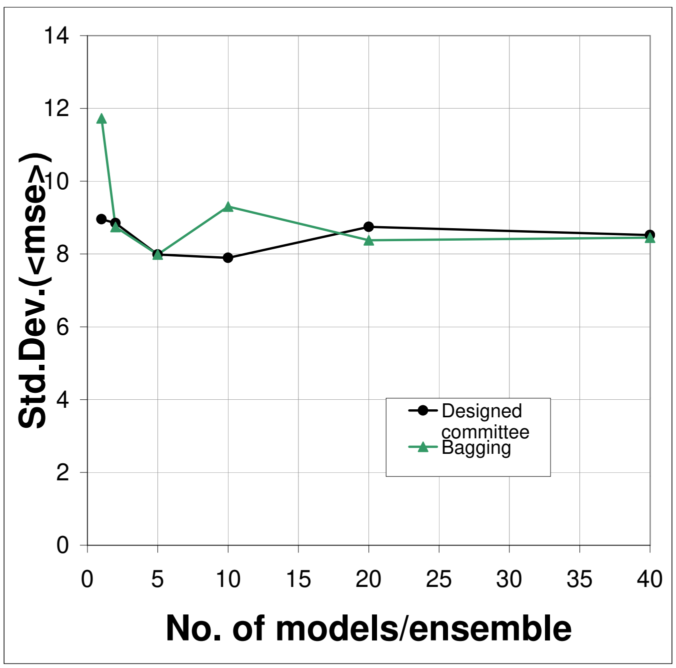

Figure 16: Std. Dev.($\langle mse \rangle$) for “Ozone”

Results appear in Figures 15-16.

The $\langle mse \rangle$ pattern is slightly different from in the Friedman examples. Though the designed committee approach outperforms bagging considerably at committee sizes of 1, 2 and 10 models, bagging actually performs best at 20 and 40 models per committee, albeit only by very little margin. The overall picture is consistent with the Friedman examples: small committees $\Rightarrow$ bagging is inferior; large committees $\Rightarrow$ bagging is at least as good as the designed approach (and here actually better).

For bagging, the scatter in $\langle mse \rangle$ again seems to drop with committee size, but no clear trend appears for the designed committee approach.

4.3 Miles per Gallon data set

The Miles per Gallon data set, in the present version obtained from the UCI Repository, originated in the StatLib library, which is maintained at Carnegie Mellon University. It contains 9 variables and 398 data points but, owing to missing data in the set, the number of inputs has been reduced to 7 (with one output) and 392 data points. It is a real life data set, concerning fuel economy in cars.

The modelling undertaken here was as follows:

Architecture: *MLP with 3 hidden layer neurons (a single hidden layer), and* tanh *activation function.*

Optimisation: *Downhill simplex with multiple restart (15 runs, best model retained).*

Training-error: *Mean-squared-error (weighted, in the case of the designed committee). No regularisation was applied.*

Test-error: *Cross-validation, with 75% data for training and 25% for testing.*

The entire series of experiments was, in summary:

Committee-sizes: *1, 2, 5, 10, 20, and 40 models/committee*

Number of runs: *160 (for the 1 model/committee case), 80 (2 models/committee), 32 (5 models/committee), 16 (10 models/committee), 8 (20 models/committee) and 4 (40 models/committee). These were chosen to obtain stable average test errors.*

Results reported: *Average test error, $\langle mse \rangle$ (over the specific number of runs performed). Standard dev. of test errors from $\langle mse \rangle$ (over same number of runs).*

Results appear in Figures 17-18.

Again at small committees the designed approach fares better in terms of $\langle mse \rangle$, but from 5 models and upwards the two techniques are rather equal though at the very largest committee size (40 models) bagging performs considerably worse than at the 20 model/com-mittee size, as well as worse than the designed approach. The drop in $\langle mse \rangle$ with committee size is not consistent for bagging; it is for the designed committee approach however.

$\langle Mse \rangle$ scatter drops rapidly for both methods, with bagging having significantly less at 5 models/committee than the designed approach. For smaller committees it is the other way around; for larger ones the two methods attain almost identical levels.

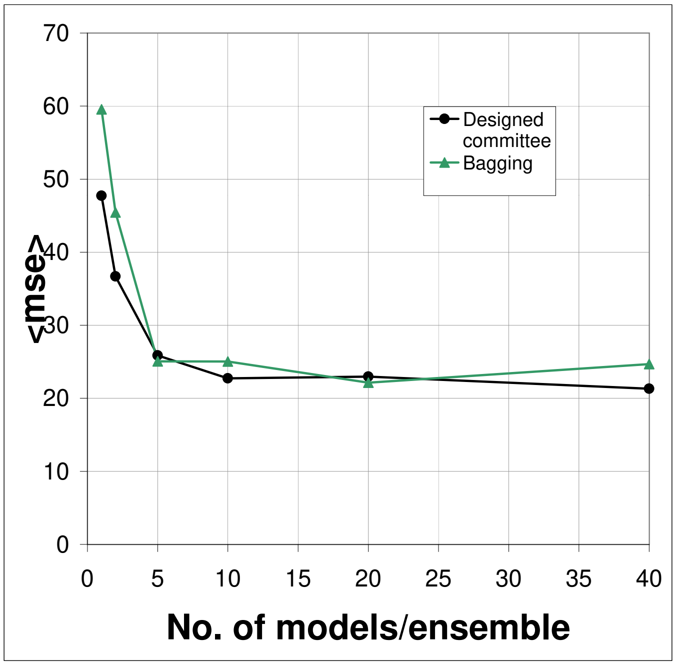

Figure 17: $\langle mse \rangle$ for “Miles per Gallon”

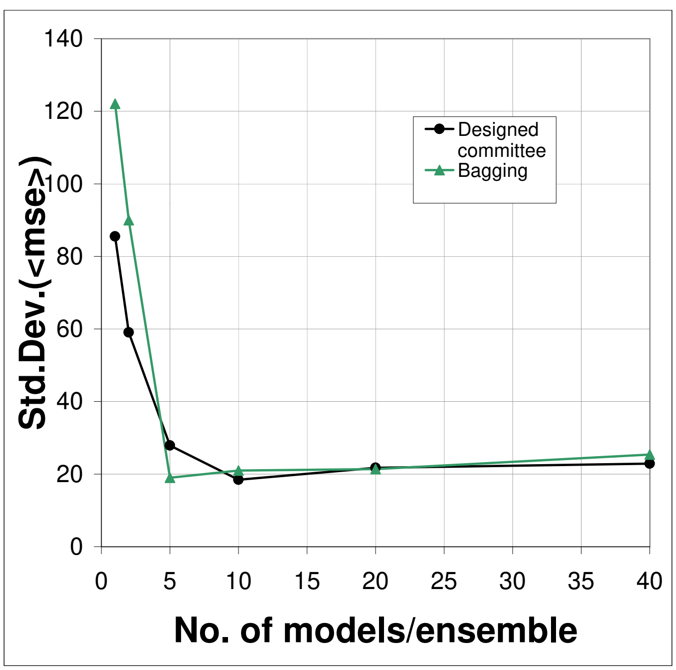

Figure 18: Std. Dev.($\langle mse \rangle$) for “Miles per Gallon”

4.4 Schwefel 3.2 data set

The Schwefel 3.2 data set comes from a function given by Schwefel (1981) and it is less well known in the modelling community. It is an artificial problem, the data arising from sampling of the function:

$$f(x) \;=\; \sum_{i=2}^{3}[(x_1 - x_i^2)^2 \;+\; (1-x_i)^2]$$

For the present use, 100 data points were sampled, input values randomly distributed over the interval [-5,5] for all inputs. To the data sampled was not added any noise.

The modelling undertaken was as follows:

Architecture: *MLP with 7 hidden layer neurons (a single hidden layer), and* tanh *activation function.*

Optimisation: *Downhill simplex with multiple restart (15 runs, best model retained).*

Training-error: *Mean-squared-error (weighted, in the case of the designed committee). No regularisation was applied.*

Test-error: *Cross-validation, with 75% data for training and 25% for testing.*

The entire series of experiments was, in summary:

Committee-sizes: *1, 2, 5, 10, 20, and 40 models/committee*

Number of runs: *160 (for the 1 model/committee case), 80 (2 models/committee), 32 (5 models/committee), 16 (10 models/committee), 8 (20 models/committee) and 4 (40 models/committee). These were chosen to obtain stable average test errors.*

Results reported: *Average test error, $\langle mse \rangle$ (over the specific number of runs performed). Standard dev. of test errors from $\langle mse \rangle$ (over same number of runs).*

Results appear in Figures 19-20.

In this experiment, bagging is outdone rather extensively, in terms of $\langle mse \rangle$, at all committee sizes; even at 40 models/committee has bagging failed to catch up. Also in terms of $\langle mse \rangle$ scatter has the designed committee approach got significantly lower values, except at 10 models/committee, where the two methods attain roughly the same level.

5. Discussion

The usefulness of weighting models individually in committee based modelling, and of taking the confidence in the individual data points into consideration when deriving models, has been understood for decades; however these principles are not usually employed in conjunction. Deliberate weighting of data points is normally not a concern in the construction of committees, and very unequal weights may be given to data points of the same provenance; a fact that will generally meet with little understanding from the data owners, who have gathered the data.

Weighting the contribution of the individual data points, to the error-measure employed in the training phase, so that every data point contributes with a pre-specified weight to the *committees*

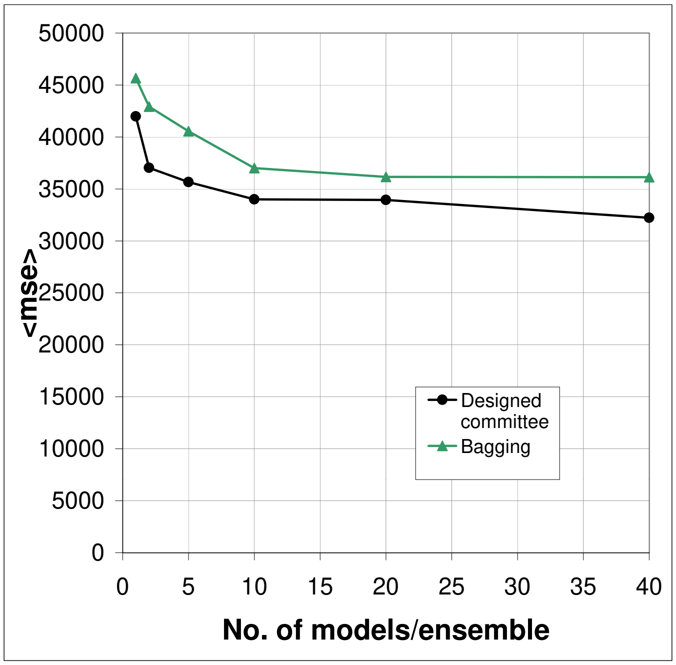

Figure 19: $\langle mse \rangle$ for "Schwefel 3.2"

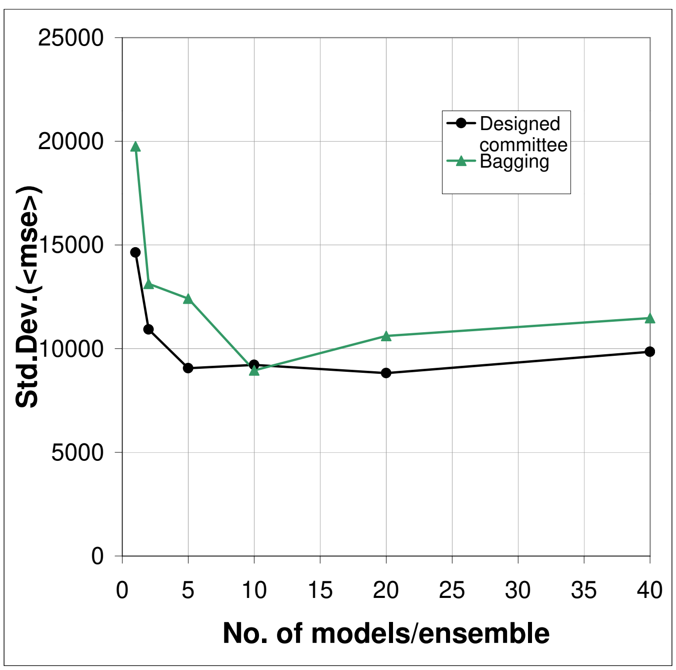

Figure 20: Std. Dev.($\langle mse \rangle$) for "Schwefel 3.2"

measure of error can, however, easily be accomplished, as long as the number of models to be entered into the committee is known in advance of the modelling. The use of a design matrix, as suggested in this paper, illustrates a way to do so.

The extent to which such premeditated weighting is advisable is partly illustrated by the bootstrap experiments conducted here. It will be seen that even when the number of models included in the committee is fairly large, e.g. > 100, there is a great deal of difference between the weights that the individual data points contribute. When using 100 models in the committee, the most influential data point may have a weight more than 50% larger than that of the least influential - even when the number of data points is as small as 20. If the number of models is less than 30, which will often be the case in practical work, this percentage may exceed 400; i.e. certain data points may, in effect, be almost obliterated from the training data set.

Of course, prediction accuracy being the main interest in most modelling, differences in the data points' weights does not constitute a principal problem *per se* and, as witnessed by boosting, such differences may even be beneficial. But, as opposed to boosting, in which differences in point weights are meticulously obtained, so as to maximise prediction accuracy, in certain other approaches, most famously bagging, weight differences are assigned at mere random, which is much harder to defend.

Certainly, this may be a moot point; clearly the importance of weighting the data points, in agreement with the prior confidence in them, may be negligible if the number of data points, relative to the complexity of the system being modelled, is very large, and the data points are well distributed. In that case, even completely disregarding certain data points may not affect the modelling severely, since there may be other data points nearby (in input space) that will carry much of the information, lost from dropping the others. The information loss is, however, potentially severely damaging when the number of data points is very low and, in effect, there are no other data points in a dropped points vicinity. In that situation, losing the contribution from a single data point may badly affect the committee, in the neighbourhood of that point. Therefore, it is advisable to ensure correct weighting of the data points within the committee, when the amount of data (relative to the complexity of the system being modelled) is low. This will often be found to apply in cases where the cost of gathering data points is very large - typically the case in laboratory experiments. In the converse situation, of plentiful supplies of data, the use of such weighting is presumably of lesser importance.

The results of the problems investigated here, comparing the designed committee principle and bagging, reveal clear advantages of the former over the latter. Generally, bagging prediction accuracy is much lower than that for the designed committee, at the minimum committee size of 1 model/committee. This of course is unsurprising, since the bagging model is obtained from only $\sim 2/3$ of the data points, whilst the designed model has been found using all of them. However, as the committee size increases, the difference in prediction accuracy is generally reduced considerably, although here bagging was rarely found to perform best, and never found to perform significantly better than the designed approach.

The fact that bagging managed to perform better than the designed committee approach at all, even if only rarely and marginally, is probably testimony to the great amount of diversity, amongst *committees*, that bagging gives rise to. This diversity is encouraged through bagged committees all having individual w_p's, whereas designed committees share the same w_p and it is evidenced by baggings generally larger scatter in $\langle mse \rangle$. Presumably, if averages had been produced over larger numbers of committees, and subsequently more stable $\langle mse \rangle$ values had been obtained, a completely consistent pattern would have emerged. This is of course speculation at present, and only larger scale experiments will reveal whether this assumption holds or not.

A very consistent picture emerged of $\langle mse \rangle$ scatter dropping with increasing committee size until levelling off at large committee sizes. What causes this effect, which applies to both methods, is not clear; conceivably it is an artefact reflecting the general modelling schedule used for both techniques. For all experiments in this study, the number of committees over which the averages were obtained dropped with increasing committee size; when large, the committees are likely to be quite similar, leading to similar $\langle mse \rangle$s; when small, the committees are likely to be quite dissimilar, leading to dissimilar $\langle mse \rangle$s. For now, this is our preferred explanation of the effect.

In the Friedman problems, no clear effect could be found, of overall data point density, on the performance difference between the proposed approach and bagging; this runs somewhat counter to expectation but can possibly be explained as a result of the data point density being very low for all the data sets, even the largest. Further experimental work will be required to elucidate this issue properly.

A real concern with the proposed framework, may be that the process outlined is designed; i.e. how many models, and data points, are to be included must be known in advance of the modelling,; this may, or may not, be practical for the modelling task at hand. As shown, this does not prevent the possibility of enlarging an already existing committee; in other words, should a given committee be considered inadequate, it can be augmented by further models at any point, and these may even be obtained on different data from those used originally.

The core concept in the proposed framework is the design matrix; this is bestowed with a great deal of freedom, allowing the information contained in the data to be distributed over the individual models according to a great many principles. "Random" distribution has been shown, as has distribution following a cross-validation scheme. Other principles could be employed as well, whereby the distribution is based on certain properties of the data points. Inherent to these are their locations in input/output space, and from this a distribution not unlike that of mixtures-of-experts can be obtained. Typically, however, there may be more pieces of information stored with the data than those eventually employed in the modelling. Such additional properties may also form the basis of the distribution of information across the committee's models.

An intrinsic obstacle lies herein, in the sense that obtaining an appropriate design matrix may itself require a modelling exercise, though possibly a less complicated one, prior to the actual modelling of the targeted system. The obvious advantage of course, is that tailor-made distributions can indeed be obtained. It must be remembered, however, that the driving force behind the distribution of information across models should at all times be the attempt to maximise the differences between these models, as this is the prime means through which a committee exercises its superiority over a single model, with respect to system approximation.

6. Conclusions

In this paper, a new, systematic approach to the formation of committees, by way of a design matrix, has been introduced. Each data point contributes with a prespecified, overall weight to the training of the committee, whilst contributing with unequal specific weights in the training of the constituent models, thus encouraging model diversity. The overall weight should be assigned based on the prior confidence in the data point; if all are considered equally reliable, each should be given the same overall weight.

Within this framework, different principles may be used to distribute the influence of the individual data points over the committee's models. Thus bagging is a special case, corresponding to a particular type of distribution; the binary distribution inherent to cross-validation may be used, and mixtures-of-experts may be emulated though, owing to the strictly adaptive nature of that approach, it cannot be fully implemented. This same discordance, between adaptive and designed modelling, prevents boosting from being directly implementable under the proposed framework.

In the paper it has been shown how to augment an existing committee by additional models, whether these be built on the original data, or on all/partly new data, such that the resulting committee places the desired emphasis on each data point.

Further, use of the design matrix has been conjectured, on theoretical grounds, to be particularly advisable when the number of data points available, relative to the complexity of the system to be modelled, is low; if the supply of data points is relatively generous, however, it may be a near redundant exercise. Counter to the theoretical argument, the actual experiments carried out here did not show this, possibly because all sub problems studied had very few data points.

Finally, in a direct head-to-head comparison with bagging, the theoretical advantage of the designed committee, due to its "fairer" weighting of the information contained in the data, would seem to be supported by the numerical experiments carried out here. It must be acknowledged that the number and variety of these experiments is insufficient to warrant very strong conclusions, but, on the other hand, the clear pattern that emerged, and the agreement between this and theory, constitutes a strong argument for the case of premeditated weighting of data points during committee training.

Acknowledgments

SWC thanks Dr. Steve Gunn for numerous stimulating discussions in the general area of machine learning.

References

L. Breiman. Bagging predictors. Machine Learning 24 (2):123-140,1996.

L. Breiman. Randomizing outputs to increase prediction accuracy. Technical Report No. 518. Department of Statistics, UC Berkeley, 1998.

L. Breiman. Combining predictors. In Amanda J. C. Sharkey, editor, Combining Artificial Neural Nets. Ensemble and Modular Multi-Net Systems, Perspectives in Neural Computing, 31-50. Springer, 1999.

L. Breiman & J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. Journal of the American Statistical Association, 80, 580-619, 1985.

T. G. Dietterich. Ensemble methods in machine learning. In J. Kittler and F. Roli, editors, Multiple Classier Systems, vol. 1857 of Lecture Notes in Computer Science, pages 1-15, Cagliari, Italy, Springer, 2000.

J. H. Friedman. Multivariate adaptive regression splines. The Annals of Statistics, 19 (1), 1-141, 1991.

L. K. Hansen, & P. Salamon. Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12. No. 10, 993-1000, 1990.

S. Hashem. Optimal linear combinations of neural networks. Neural Networks, vol. 10 (4), 599-614, 1997.

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. Neural Computation 3, 79-87, 1991.

A. Krogh, & J. Vedelsby. Neural network ensembles, cross validation, and active learning. In G. Tesauro, D. S. Touretzky and T. K. Leen, eds. Advances in neural information processing systems 7, 231-238, MIT Press, Cambridge MA, 1995.

Y. Liao, & J. Moody. Constructing Heterogeneous Committees Using Input Feature Grouping: Application to Economic Forecasting. In Advances in Neural Information Processing Systems 12, S. A. Solla, T. K. Leen, K.-R. Mller, eds., MIT Press, 2000.

Y. Liu. & X. Yao. A Cooperative Ensemble Learning System. In Proc. of the 1998 IEEE International Joint Conference on Neural Networks (IJCNN'98), pp.2202-2207, Anchorage, USA, 1998.

Y. Liu, & X. Yao. Ensemble learning via negative correlation. Neural Networks 12, 1399-1404, 1999.

C. J. Mertz. Classification and regression by combining models. Ph.D. thesis, UC Irvine, 1998.

P. M. Murphy & D. W. Aha. UCI Repository of machine learning databases. University of California, Department of Information and Computer Science. Irvine, CA (1994).

D. W. Opitz, & J. W. Shavlik. Actively searching for an effective neural-network ensemble. Connection Sci., 8 (3-4), 1996.

M. P. Perrone, & L. N. Cooper. When networks disagree: Ensemble methods for hybrid neural networks. In R. J. Mammone, editor, Neural Networks for Speech and Image Processing, 1993.

Y. Raviv, & N. Intrator. Bootstrapping with noise: An effective regularization technique. Connection Science, Special issue on combining estimators, 8:356-372, 1996.

B. E. Rosen. Ensemble Learning using Decorrelated Neural Networks. Connection Science, Vol. 8, No. 3-4, pp. 373-384, 1996.

H. Schwefel. Numerical Optimization of Computer Models. Wiley, New York, 1981.

R. Shapire. The strength of weak learnability. Machine Learning, 5 (2), 197-227, 1990.