ENSEMBLES OF HYBRID INTELLIGENT EXPERTS: EXTENDING THE POWER OF OPTIMAL LINEAR COMBINERS

Georgios C. Anagnostopoulos
Dept. of Electrical & Computer Engineering
University of Central Florida
Orlando, FL 32816
Email: gca@ece.engr.ucf.edu

Michael Georgiopoulos
Dept. of Electrical & Computer Engineering
University of Central Florida
Orlando, FL 32816
Email: mng@ece.engr.ucf.edu

David Nickerson
Dept. of Statistics
University of Central Florida
Orlando, FL 32816
Email: nickersn@pegasus.cc.ucf.edu

George Bebis
Dept. of Mathematics & Computer Science
University of Missouri at St. Louis
St. Louis, MI 63121-4499
Email: bebis@mayura.cs.umsl.edu

ABSTRACT

In the present paper we generalize the idea of Optimal Linear Combiners that are used to aggregate information from different sources providing estimates about a specific quantity. Two linear models are introduced, along with their analysis, which combine related components of information when more than one variable is to be predicted. The models' purpose is to produce point estimates of better accuracy in terms of mean squared error. Experimental results dealing with a functional approximation problem demonstrate that the generalized Optimal Linear Combiners suggested yield higher accuracy when compared to other combiners such as the Simple Average, or the conventional Optimal Linear Combiners.

1. INTRODUCTION

Estimation of *variables of interest* (VI) whether involving point estimation or the estimation of an entire distribution, is a fundamental problem with vast numbers of applications such as time series forecasting, pattern recognition and functional approximation, to name a few. Quite often, a *decision maker* (DM) has access to a collection of experts (an *ensemble*) and is faced with the task of obtaining an optimal decision based on the estimates about a particular VI, that these experts supply. Being presented with a plethora of expert opinions, that do not necessarily coincide in all occasions, complicates the process of decision making. Some questions that arise are which experts should be ignored, which should be taken into account, and finally how to end up with a single opinion. For many years the classical approach (the so called *naive approach*) to all these problems was first to choose an estimation performance criterion like the *Mean Square Error*

0-7803-4053-1/97/$10.00 © 1997 IEEE

(MSE), which we will consider henceforth, and then make the data available to the collective of experts. The DM's choice would be the expert featuring the best estimate in terms of smallest MSE.

Over the past 20 years researchers from the field of economic forecasting have addressed the aforementioned issues and demonstrated that aggregating the experts' opinions can lead in many cases to a substantial enhancement in the estimation process (see for example [1], [2]). The idea exploits the fact that experts in the ensemble express independent, or partially independent opinions. This is usually a result of the possibly different data sets processed by them, or due to the different assumptions made by them about the underlying process generating the data relevant to the VI. A particular straightforward approach is to linearly combine the opinions. Models based on this idea are called *MSE optimal linear combiners* (MSE-OLC). An example of such a model is given in Figure 1, where the estimates of 3 experts are linearly combined. These models were first introduced by [3], where it was shown that they outperform the best experts within the ensemble. Although there is much debate on the exact choice of the weights for an OLC model, two versions of this family of combiners are more popular: the *constrained OLC* (MSE-cOLC), where specific linear constraints are imposed on the weights of the linear combination (see [3], [4]), and the *unconstrained OLC* (MSE-uOLC), where there are no constraints on the weights (see [5], [6]). It has been pointed out that even though inter-dependence of the individual members of an ensemble is the cause of multicollinearities among them, it is useful in general to include as many of them as possible in the aggregation procedure by means of an OLC model. In the context of *artificial neural networks* (ANN) the concept of cOLC combiners was introduced by [7] and the uOLC by [8]. In both references it was concluded that ensemble methods like the OLC family minimize up to a certain extent the effect of a potential model overfitting to the training data by performing a smoothing operation in the functional space of estimators.

An extension to the idea of OLC models can be investigated when a DM has to cope with point estimation of a vector rather than a scalar. To derive a vector point estimate employing as many separate OLC models as the number of components in the VI, the DM makes the implicit assumption that the estimation of one component is independent from the estimation of the others. But this assumption does not account for the potential dependence among these components, which is the case in many practical applications. We propose the *generalized* MSE-OLC (MSE-gOLC) models that exploit the aforementioned inter-dependence between vector components in order to enhance individual point estimation.

2. NOTATION

Before we start presenting the analytical details of the linear combiner models, let us first introduce some useful notation. Let $\mathbb{R}$ be the set of real numbers. Then $\mathbb{R}^M$ denotes the set of all column vectors with M real components and $\mathbb{R}^{M\times N}$ is the set of all real matrices with M rows and N columns. Lower case letters stand for positive integers, unless otherwise specified. A symbol that we will use frequently, when it comes to defining various quantities, is $\triangleq$, which stands for "is defined as".

Lower case underscored letters represent column vectors and upper case letters are used for matrices, unless, again, specified otherwise. The transpose of a vector $\underline{x}$ is denoted by $\underline{x}^T$, and the same notation is used for transposes of matrices as well. Some special vectors are: $\underline{e}_m$ is the m^{th} member of the standard orthonormal basis vector for $\mathbb{R}^M$, so it's components are all zero except the one in the m^{th} row which is equal to one; $\underline{1}_N$ is a column vector with N rows filled with ones; $\underline{0}_N$ is the "zero" column vector with N zeros as components. Also, some special matrices that will be used are the following: I_L is the L by L identity matrix and $O_{M\times L}$ is the "zero" matrix with M rows and L columns containing zeros. Continuing with the vector-matrix notation, we denote the inverse for a square non-singular matrix C by C^{-1}. If $S \in \mathbb{R}^{L\times L}$ is a symmetric positive definite matrix, we will denote this as $S > 0$. Assuming that S is non-negative definite we will define the vector norm $\|\underline{x}\|_S \triangleq \sqrt{\underline{x}^T S \underline{x}}$, which coincides with the standard Euclidian vector norm if S is equal to the proper identity matrix.

To conclude our introduction of notation we will define some of the operators related to random variables (RV). If the vectors $\underline{x}$ and $\underline{y}$ are RVs, by $E\{\underline{x}\}$, $Cov(\underline{x},\underline{y})$ and $Var(\underline{x}) \triangleq Cov(\underline{x},\underline{x})$ we mean the expectation, the covariance and the variance, respectively, of the quantities involved. Finally, $E\{\underline{y}\,|\,\underline{x}\}$ denotes the conditional expected value of $\underline{y}$ conditioned on the vector $\underline{x}$.

3. ANALYSIS

We assume that the DM consults a *module ensemble* (ME), which is defined as being a collection of $N \geq 2$ processing units called *experts* or *modules*. It has to be stressed here that the nature of the experts is of no importance; a mathematical model, an expert system or even a human could serve as a module. The n^{th} member of the ME is of the form illustrated in Fig. 2. It receives as input a vector RV $\underline{x} \in \mathbb{R}^P$, for some integer P, which is common for all modules participating in the

ME. It delivers as output another vector RV $\hat{\underline{y}}_n(\underline{x}) \in \mathbb{R}^M$ with M components, viz.

$$\hat{\underline{y}}_n(\underline{x}) \triangleq \begin{bmatrix} \hat{y}_{n1}(\underline{x}) \\ \vdots \\ \hat{y}_{nM}(\underline{x}) \end{bmatrix} \quad (1)$$

It is assumed that this vector is a point estimator of the vector RV $\underline{y}(\underline{x}) \in \mathbb{R}^M$ which represents the DM's VI. The quantity $\hat{y}_{nm}(\underline{x})$ stands for the point estimator for the m^{th} component of $\underline{y}$ supplied by the n^{th} ME member. Without sacrificing generality we hypothesize that every module belonging to the ME responds with a vector of common dimensionallity M. We underline the fact that the exact nature of the mechanisms that lie behind the point estimation of each module will not be of our concern.

Next, by stacking N module responses into a single column vector we can define a new vector $\hat{\underline{Y}}(\underline{x}) \in \mathbb{R}^L$ as

$$\hat{\underline{Y}}(\underline{x}) \triangleq \begin{bmatrix} \hat{\underline{y}}_1(\underline{x}) \\ \vdots \\ \hat{\underline{y}}_N(\underline{x}) \end{bmatrix} \quad (2)$$

with $L = NM$. Finally, the model produces a point estimator $\tilde{\underline{y}}$ by multiplying $\hat{\underline{Y}}$ with a *weight matrix* W and adding a *bias correction term* $\underline{b}$ will be called a *generalized Linear Combiner (gLC)* with parameters $W \in \mathbb{R}^{LxM}$ and $\underline{b} \in \mathbb{R}^M$. Expressing it as an equation we have

$$\tilde{\underline{y}}(\underline{x}; W, \underline{b}) = W^T \hat{\underline{Y}}(\underline{x}) + \underline{b} \quad (3)$$

Fig. 2 illustrates a block diagram of this combiner. If we define the *gLC estimation error vector* $\tilde{\underline{\varepsilon}} \in \mathbb{R}^M$ as

$$\tilde{\underline{\varepsilon}}(\underline{x}) \triangleq \underline{y}(\underline{x}) - \tilde{\underline{y}}(\underline{x}; W, \underline{b}) \quad (4)$$

then the MSE expression for the m^{th} gLC output is given by

$$MSE_m(W, \underline{b}) = \underline{e}_m^T E\{\tilde{\underline{\varepsilon}}(\underline{x}) \tilde{\underline{\varepsilon}}^T(\underline{x})\} \underline{e}_m \quad (5)$$

A gLC model that minimizes the MSE above will be called *generalized MSE Optimal Linear Combiner (MSE-gOLC)*. Generalized OLC models can be divided into two major categories as is demonstrated below.

1) *Unconstrained MSE-gOLC models (MSE-ugOLC).* For this type of gOLC combiner the DM imposes no restrictions on the values the model parameters may acquire. The optimization procedure for the m^{th} component is stated below

$$\{\underline{w}_m^o, b_m^o\} \triangleq \underset{\underline{w}_m, b_m}{arg\,min}\, MSE_m(\underline{w}_m, b_m) \quad (6)$$

where $\underline{w}_m$ and b_m are the m^{th} column of matrix W and the m^{th} component of the bias correction term, respectively. We define now the matrices $C \in \mathbb{R}^{LxL}$ and $D \in \mathbb{R}^{LxM}$ as

$$C \triangleq Cov(\hat{\underline{Y}}) = E\{\hat{\underline{Y}} \hat{\underline{Y}}^T\} - E\{\hat{\underline{Y}}\} E\{\hat{\underline{Y}}^T\} \quad (7)$$

$$D \triangleq Cov(\hat{\underline{Y}}, \underline{y}) = E\{\hat{\underline{Y}}\, \underline{y}^T\} - E\{\hat{\underline{Y}}\} E\{\underline{y}^T\} \quad (8)$$

By taking appropriate gradients and equating them to zero we can demonstrate that the optimal parameters are given by

$$W^o \triangleq C^{-1} D \quad (9)$$

$$\underline{b}^o \triangleq E\{\underline{y}\} - D^T C^{-1} E\{\hat{\underline{Y}}\} \quad (10)$$

Here we made the reasonable assumption that $C>0$, so that the inverse exists. The minimum MSE value attained for each component of the ugOLC estimator is

$$MSE_m(\underline{w}_m^o, b_m^o) = Var(y_m) - \|\underline{d}_m\|^2_{C^{-1}} \quad (11)$$

where $\underline{d}_m$ and y_m is the m^{th} column of matrix D and the m^{th} component of the VI, respectively. The resulting ugOLC estimator $\tilde{\underline{y}}$ turns out to be unbiased, meaning that

$$E\{\tilde{\underline{\varepsilon}}\} = 0 \quad (12)$$

This has been achieved through the utilization of the bias correction term in Eq. (10).

2) *Constrained MSE-gOLC models (MSE-cgOLC).* Let us assume that the experts participating in a ME provide the DM with unbiased estimates of the VI having the form

$$\hat{\underline{y}}_n = E\{\underline{y} \mid \underline{x}\} + \hat{\underline{\varepsilon}}_n \quad (13)$$

$$E\{\hat{\underline{\varepsilon}}_n\} = \underline{0}_M \quad (14)$$

where $\hat{\underline{\varepsilon}}_n$ is the estimation error of the n^{th} module output. Also, an unbiased gOLC estimate (see Eq. (15) and (16)) would be another desirable feature for the DM, i.e.,

$$\tilde{\underline{y}}(\underline{x}) = E\{\underline{y} \mid \underline{x}\} + \tilde{\underline{\varepsilon}} \quad (15)$$

$$E\{\tilde{\underline{\varepsilon}}\} = \underline{0}_M \quad (16)$$

This *apriori* knowledge/hypothesis can be incorporated into a gOLC model in the form of constraints imposed on the parameters. It can be shown after manipulating Eq. (3) and Eq. (13)-(16) that the following constraint has to be met by the weight matrix

$$AW = I_M \tag{17}$$

For the above equation we define the *constraint matrix* $A \in \mathbb{R}^{MxL}$ as the following Kronecker tensor product

$$A \triangleq \underline{1}_N^T \otimes I_M \tag{18}$$

Eq. (17) implies that the constraint matrix should be a right inverse of the weight matrix. Taking into account these constraints the DM is faced with the following M decoupled constrained minimization problems

$$\{\underline{w}_m^{co}, b_m^{co}\} \triangleq \underset{A\underline{w}_m=\underline{e}_m, b_m}{arg\,min} MSE_m(\underline{w}_m, b_m) \tag{19}$$

If we first define the symmetric matrix $\Gamma \in \mathbb{R}^{MxM}$ as

$$\Gamma \triangleq AC^{-1}A^T \tag{20}$$

then by using Lagrangian multipliers we obtain that in this case, the optimal parameters are

$$W^{co} \triangleq C^{-1}D + C^{-1}A^T\Gamma^{-1}(I_M - AC^{-1}D) \tag{21}$$

$$\underline{b}^{co} \triangleq \underline{0}_M \tag{22}$$

Because of the special structure of the constraint matrix given by Eq. (18), it can be proven that matrix Γ will be invertible if $C>0$. The fact that all the bias correction terms are identically zero (Eq. 22) should not be a surprise, since we hypothesized the unbiasedness of all the involved estimators. The resulting minimum MSE's are displayed below

$$MSE_m(\underline{w}_m^{co}) = MSE_m(\underline{w}_m^{o}, b_m^{o}) + \left\|C^{-1}A^T\Gamma^{-1}(A\underline{w}_m^{o} - \underline{e}_m)\right\|_C^2 \tag{23}$$

We observe that for the constrained combiner the lowest MSE value is greater than the one attained when employing an unconstrained model instead. They differ by a factor, whose magnitude depends on how close the weight matrix W^o complies to the condition described by Eq. (17).

When attempting a comparison of OLC and gOLC models, it is quite straightforward to demonstrate that the former family of models is just a special case of the latter ones. This can be derived with the help of Fig. 3, where a DM consults with 2 experts receiving common input to obtain a point estimator of a two dimensional VI. If the DM chooses to assume that these two components are statistically independent, he/she would apply an OLC model (either unconstrained or constrained depending if there are any further assumptions about unbiasedness of the available estimators). For this particular case the weights used by the OLC model correspond to the solid lines that connect the module outputs with the outputs of the combiner. However, if there is indeed inter-dependence among the two VI components and the DM applies a gOLC model then additional weights would be available to further minimize the MSE. So, OLC models would be a suboptimal special case of gOLC combiners. It also becomes obvious that an OLC model (either constrained or unconstrained) could be derived from its gOLC counterpart by setting the weights that correspond to the dashed lines equal to zero.

4. EXPERIMENTS

In order to demonstrate the superiority of gOLC models over other existing combiners a series of experiments was performed on a function approximation task. More specifically, the comparison was made between the "Naive" model (the output of the combiner coincides with the one belonging to the expert with the lowest MSE), the "Simple Average" model (all the outputs estimating the same component of the VI are simply averaged), the OLC and the gOLC family models. The problem consisted of approximating the noisy contour of a circle with radius R residing on a plane. Three feedforward ANN's dedicated to the estimation of the X and Y coordinates of the circle were separately trained on a common training set of patterns; then, their responses were fused using the aforementioned combiners. After obtaining the optimal weights for these combiners, their responses were compared to a testing set that consisted of patterns not used during the training procedure of the individual ANN's. This was done in order to draw conclusions about the generalization properties of each class of linear combiners.

The experiments were divided in 4 major sets reflecting 4 different noise levels by varying the standard deviation σ of the random radial component. For each set the procedure of training the three ANN's, combining them and then evaluating the resulting MSE for the various combiners was repeated for 1000 times. During the derivation of the combiners' optimal parameters (training phase of the combiners) gOLC models were exhibiting the lowest MSE, which was something to be expected, since they are optimal among all other possible linear combiners. The results illustrated in Fig. 4 through Fig. 7 were generated for $R = 10$ and for σ taking the values 0.0, 0.5, 1.0, 2.0 . The figures show the percentage out of the 1000 cases related to each set (noise level) for which each model or family of models yielded the minimum MSE score on the testing set. Note that in the these figures "avg" stands for the Simple Average models. For this specific approximation task it is evident that gOLC combiners maintain their superiority even when they are presented with data not used in the derivation of their parameters,

that is, they exhibit good generalization. Another observation is that their performance degrades gradually with increasing noise levels.

5. CONCLUSIONS

Although slightly more complex than conventional OLC models, gOLC combiners are able to relent more accurate point estimates when components of a vector VI share information about each other. Assuming that the training sets used for training gOLC models are representative enough of the entire pattern domain (including the patterns in all possible testing sets), their superiority becomes more dominant when there is little noise or uncertainty involved in the data.

6. REFERENCES

[1] C.W.J. Granger, "Combining Forecasts - Twenty Years Later", Journal of Forecasting, Vol. 8, 1989, pp. 167-173

[2] R.A. Jacobs, "Methods for Combining Experts' Probability Assessments", Neural Computation, Vol. 7, 1995, pp. 867-888

[3] J.M. Bates and C.W.J. Granger, "The Combination of Forecasts", Operational Research Quarterly, Vol. 20, 1969, pp. 319-25

[4] J.P.Dickinson, "The combination of short term forecasts", Proceedings Univ. of Lancaster Forecasting Conference, 1972

[5] C.W.J. Granger and R. Ramanathan, "Improved methods of combining forecasts", Journal of Forecasting, Vol. 3, 1984, pp. 197-204

[6] D. Bunn, "Forecasting with more than one model", Journal of Forecasting, Vol. 8, 1989, pp. 161-166

[7] M.P. Perrone and L.N. Cooper, "When Networks Disagree: Ensemble Methods for Hybrid Neural Networks". In "Artificial Neural Networks: Forecasting Time Series", V. Vemuri editor, IEEE Computer Society Press, Los Alamitos, CA, 1994, pp. 126-142

[8] S. Hashem and B. Schmeiser, "Improving Model Accuracy Using Optimal Linear Combinations of Trained Neural Networks", IEEE Transactions on Neural Networks, Vol. 6, 1995, pp. 792-794

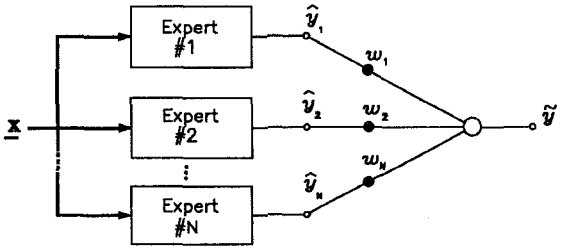

Figure 1

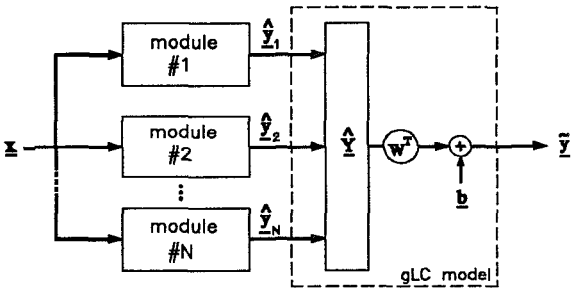

Figure 2

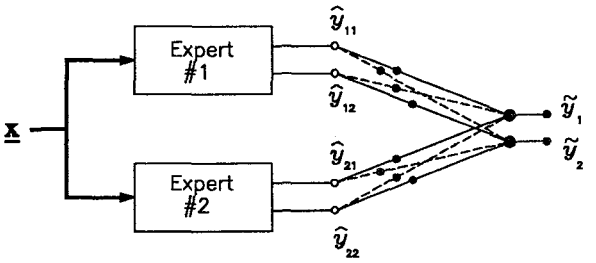

Figure 3

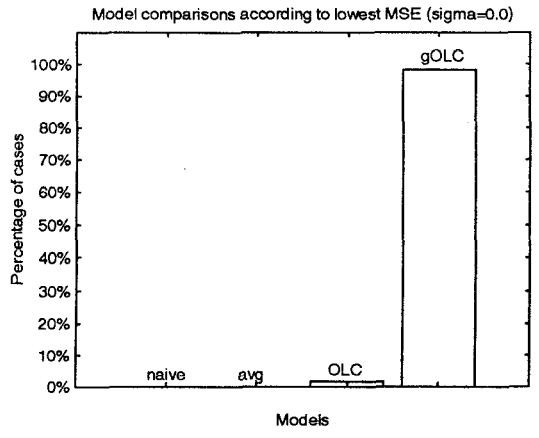

Figure 4

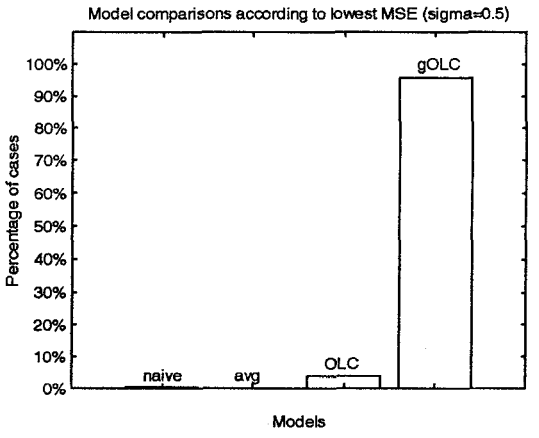

Figure 5

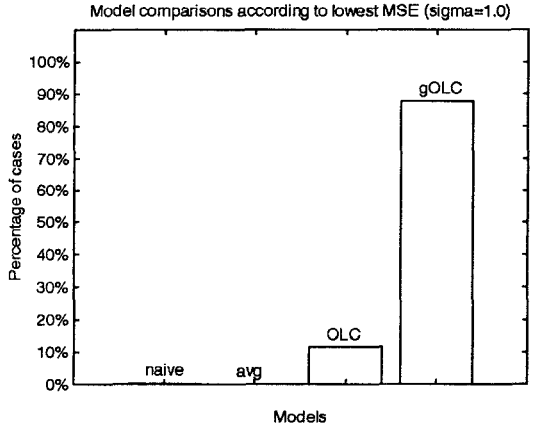

Figure 6

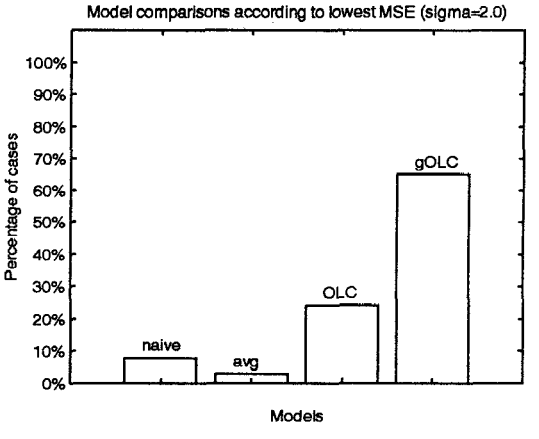

Figure 7