Pesquisa Operacional (2011) 31(1): 95-111
© 2011 Brazilian Operations Research Society
Printed version ISSN 0101-7438 / Online version ISSN 1678-5142
www.scielo.br/pope

CHOICES AND PITFALLS CONCERNING MIXTURE-OF-EXPERTS MODELING

Denise B. Ferrari and Armando Z. Milioni*

Received May 2009 / Accepted August 2010

ABSTRACT. This paper deals with some important decisions that ought to be made when building a mixture-of-experts model (MEM). Such decisions are related to aspects such as the clustering method and the gating functions used in the model. Depending on how these decisions are made, different mixtures might be formed, yielding different results. In the present study, we investigate the way such decisions affect the performance of MEM's, when using statistical models to regression problem. The famous Boston housing data is used as illustration for the technique.

Keywords: mixture-of-experts model, MEM, clustering, gating function.

1 INTRODUCTION

Researchers from different fields often need to learn and represent phenomena through relationships between variables and use them to predict the phenomena behavior under certain conditions. However, choosing the "best" model in a modeling exercise is always an arduous task, yet because of the various uncertainties associated with the modeling process.

A way of obtaining a better prediction than a single model would provide it may be by combining a number of different model structures (Armstrong, 1986). Each model is adopted at a given observation with a probability that depends on the values of the explanatory variables for that observation. This is the logic under the mixture-of-experts model (MEM).

The MEM technique was introduced by Jacobs *et al.* (1991). In that paper, each expert was a single layer neural network trained using a squared error criterion to perform classification. This method is motivated by a concept widespread in the Computer Science field (Korf, 1987), that is, if a problem may be separated into smaller subproblems, it might be easier to solve the subproblems. Moreover, the prediction accuracy is supposed to be improved through the combination of multiple individual estimates (Waterhouse, 1997). Thus, the general MEM framework specifies that a prediction is made up of a series of predictions from separate models, or experts, each of them weighted by a quantity determined by a so called gating function.

*Corresponding author
Instituto Tecnológico de Aeronáutica. E-mails: denise@ita.br / milioni@ita.br

However, when building a MEM, many important decisions ought to be made. These include determining the number of clusters in to which the global data is to be partitioned and the clustering method to be adopted; the nature of the gating function applied and the criteria to select the experts, for example. Because of so many "degrees- of-freedom", an important thing to be mentioned is that, although the MEM is a very sophisticated technique, it leads not necessarily to better results. Depending on how these decisions are made, different mixtures might be formed, yielding diverse results.

Thus, this work aims at investigating the way the decisions related to the mixture-of-experts modeling process affect the performance of the final model.

In Section 2, the most important issues related to the theoretical aspects of the MEM are introduced. Section 3 focuses on a practical illustration of the technique. Section 4 contains the final comments and concluding remarks.

2 THEORETICAL OVERVIEW

2.1 Mixture-of-experts model

The mixture-of-experts model, introduced by Jacobs *et al.* (1991) provides important paradigms for learning from data for researchers in many different areas, like machine learning and statistics. The mixture-of-exerts model is comprised of a set of models, which perform the role of experts, and a set of mixing weights determined by the gating function.

The objective of this technique is to explain the behavior of some phenomena, under the assumption that there are separate processes involved in the generation of the data under analysis. For this reason, the structure of the mapping varies for different regions of the input space. The use of MEM allows combining many simple models to generate a more powerful one.

The experts are responsible for modeling the generation of outputs, given a certain set of conditions or inputs, and are combined by a set of local mixing weights determined by the gating function, which depends on the input or predictor variables. The gating function, therefore, is responsible for assigning or combining the models to be used in each region, depending on the degree of overlapping on the experts (Alpaydin and Jordan, 1996), *i.e.*, whether the scheme of the mixture under consideration is competitive or cooperative, respectively. In the competitive scheme we have only one expert responsible for each region. Adopting the cooperative scheme, the output of the mixture model corresponds to a linear combination of the different experts' estimates, weighted by the values determined by the gating function. These schemes will be described more carefully ahead.

The general architecture of the mixture-of-experts model for a single output shall be written as the following expression:

$$\hat{y} = \sum_{p=1}^{k} g_p \hat{y}_p \tag{1}$$

where:

$\hat{y}$ = the mixture-of-experts model estimates vector;
K = the number of experts considered in the mixture;
g_p = the gating function vector corresponding to the p-th expert, for $p = 1, 2, \ldots, K$;
$\hat{y}_p$ = the estimates vector for the p-th expert.

As it will be shown ahead, the gating function measures the contribution of each expert to the total estimate of the mixture model and is dependent of the input variables in terms of some similarity measure, usually the Euclidian distance.

The methodology adopted in this work is adapted from that described in Milidiu *et al.* (1999) and used by Melo *et al.* (2004) for performing time series prediction. Its implementation can be summarized into four steps:

i) input space partitioning into clusters;

ii) training expert models for each cluster;

iii) testing and benchmarking in order to assign to each cluster and expert model; and

iv) composing the mixture-of-experts model using a gating function; this will decide how to weight the local expert output for a given input point.

As it is possible to realize, this is a general framework in which many different decisions are involved, whether concerning the choice of the number of clusters the data set should be partitioned into, or the clustering method to be applied, the models used as the experts, and the gating function to be applied in the mixture. A number of such decisions is explored in the following sections.

2.2 Expert models

The expert models considered in this work were: (i) minimum sum of absolute errors (MSAE) regression; (ii) bayesian (BAYES) regression with priors based on the results of an ordinary least squares regression; and (iii) artificial neural network (ANN), trained using a least square error measure.

The ordinary least squares (OLS) estimation criterion is one of the most often used, for its simplicity and highly desirable statistical properties, under the assumptions of the classical linear regression model (Gujarati, 2000; Pindyck & Rubinfeld, 1998). The least squares estimators enjoy very special properties such as unbiasedness, minimum variance and efficiency under Gaussian conditions. However, the outliers which may arise from heavy-tailed distributions, have a large influence on the resulting estimates, in such a way that they may be far from optimal when errors do not follow the normal distribution, as well as in many other non-Gaussian situations (Narula, 1987).

A possible alternative more resistant to outliers is the MSAE regression, also known as L1-norm, least (or minimum) absolute deviation, among others. The MSAE estimators are not affected by changes in the value of the response variable associated with the non-zero residuals observations, as long as they remain on the same side of the MSAE fitted hyperplane (Narula, 1987).

Moreover, the MSAE regression remains unchanged even if the values of the predictor variables for the observations with non-zero residuals vary, as long as they are maintained within certain intervals, to which the fitted MSAE regression is resistant (Narula & Wellington, 2002). In spite of these desirable properties, a serious restriction to a more expressive application of the MSAE regression has been the poor offering of statistical inference tools, in comparison to the consolidated OLS procedure. However, based on a great number of Monte Carlo studies on small sample properties and asymptotic distribution results, statistical properties and inference procedures have been proposed by Narula (1987).

Another alternative to the traditional OLS regression is the Bayesian Regression. Because of its intensive computational requirements, until recently the lack of fast computational tools prevented a more widespread use of Bayesian treatments (Bishop & Tipping, 2003).

For the traditional regression, the distribution of explanatory variables is assumed to provide no information about the conditional distribution of the response given the explanatory variables. That is, the standard regression approach can be viewed as a Bayesian posterior inference based on a non-informative prior distribution for the parameters of the linear model (Gelman *et al.*, 2004). In this work, the estimates for the Bayesian regression are based on the expected values for the posterior distributions. Therefore, from a Bayesian perspective, the use of standard non-informative priors would lead to Bayesian estimates and standard errors equal to those from the OLS regression approach. Here, the OLS estimates are used as prior information for the Bayesian regression.

Both the MSAE and the Bayesian regressions are in the class of linear models. Statistical linear models have been widely used because of their simple formulation and easy interpretation. There are many real problems, however, in which a linear fit to the data may lead to a model showing poor performance, unable to explain nonlinearities present in the analyzed data. In order to be able to obtain a non-linear fit, we considered an ANN model.

The ANN modeling has recently attracted much attention as a technique for estimating and forecasting. One of its main advantages lies in its freedom from the assumption of linearity, although it does not imply necessarily more satisfactory results, when compared to the traditional techniques (Makridakis *et al.*, 1998).

The theory behind the ANN models is well covered in the literature (see, for example Haykin, 1999). So, in order to set up the ANN configuration for this, we opted for simplicity. The ANN model we adopted is a feedforward network with two hidden neurons and hyperbolic tangent as the activation function. The network output is a linear combination of the hidden layers' outputs and the network is trained using the mean squared error cost function.

2.3 Clustering

The process of clustering consists in grouping individuals in a population in order to learn the structure of the data. The partitions or groups formed can be called "clusters". It is desirable that elements within the cluster show strong similarity to one another, whereas elements in different clusters should be, in some sense, dissimilar. The cluster analysis results in the identification of structures in the data set may be used to generate hypotheses, which ought to be tried out on a separate data set.

There exist many techniques that can be used in cluster analysis. "However, different methods will often yield different groupings since each implicitly imposes a structure on the data" (Webb, 2002).

Clustering processes include: hierarchical methods, quick partitions, mixture models, sum-of-squares methods, among others. Each of them has distinct properties and may be applied to different situations. However, for each clustering procedure, a measure of similarity among patterns and a criterion function to be optimized must be specified (Duda *et al.*, 2001).

In this work, we considered the k-means clustering procedure, a sum-of-squares method and one of the most widely used clustering techniques. It aims at partitioning the data into K mutually exclusive subsets such that minimum variance is achieved within the clusters. This is the reason why it is often called a minimum variance method. The k-means measure of similarity is based on the Euclidian distance and the function to be minimized is the sum-of-squared errors within the clusters. Bishop (1995) presents the following formulation for the k-means clustering algorithm:

Let $D = \{x_1, \ldots, x_n\}$ be a set comprised of n vectors representing samples to be partitioned into exactly K mutually exclusive subsets $D_1, \ldots, D_K$. Let n_i be the number of samples in D_i and let m_i be its sample mean:

$$m_i = \frac{1}{n_i} = \sum_{x \in D_i} x \tag{2}$$

The sum-of-squared error is given by:

$$J_e = \sum_{i=1}^{k} \sum_{x \in D_i} \|x - m_i\|^2 \tag{3}$$

where $x - m_i$ are the error vectors.

This is a combinatorial optimization problem (Webb, 2002), since we search for a non-trivial partition of n elements into K groups that minimizes the sum-of-squared errors function. Although the optimal partition may be obtained by examining every possible partition, this would be a very arduous task, since the number of such partitions is approximately $Kn/K!$ and increases exponentially with n (Duda *et al.*, 2001).

The optimal partitions are frequently obtained through iterative optimization. The basic idea behind it is to guess an initial partition and then sequentially move elements from one cluster to another, until there are no more changes in the value of the criterion function. Duda *et al.* (2001) summarized this procedure in some steps:

```
Step 1: Initialize the parameters $n, K, m_1, m_2, \ldots, m_K$;
Step 2: Randomly select a sample $x$;
Step 3: Classify $x$, by computing $i$ such that $i = \arg\min i' \|m'_i - x\|$;
Step 4: Until no change in $J_e$, repeat:
    If $n_i \neq 1$, compute $\rho_k = \frac{n_j}{n_j + 1}\|x - m_j\|,\ j \neq 1$ or
                                  $= \frac{n_j}{n_j - 1}\|x - m_j\|,\ j = i$
    If $\rho_k \leq \rho_j$ for all $j$, then transfer $x$ to $D_k$;
    Recompute $J_e, m_i, m_k$;
Step 5: Return $m_1, m_2, \ldots, m_K$.
```

One drawback of this algorithm is that, like other hill-climbing procedures, it guarantees local but not global optimality. Different starting points may yield to diverse solutions and it is impossible to determine whether the best solution was achieved (Duda *et al.*, 2001).

With almost all clustering procedures, it is presumed that the number of clusters on the data set is known. However, it may not be a reasonable assumption when the properties of the data set under analysis are unknown. For this reason, a difficult question that may arise is how to correctly determine the number of clusters in the analyzed space.

When optimizing a criterion function, as in the case of k-means, does clustering a common approach is to solve the problem repeatedly for different numbers of clusters and then compare the results of each clustering by means of some criterion.

2.4 Gating functions

As mentioned previously, the gating function is responsible for assigning or combining the predictions of each expert in order to generate the overall prediction of the mixture model (Duda *et al.*, 2001). In other words, the gating function represents the relative weights of each expert for the model prediction at a given observation.

Thus, it is possible to interpret the gating functions as probability classifiers (Waterhouse, 1997), where a certain input is assigned to one or more experts, depending on whether the learning scheme adopted is competitive or cooperative, respectively. In this sense, the values g_p in equation (1) can be thought as the posterior probability that a certain input is taken care of by expert p, given the cluster to which it is allocated to.

In the cooperative scheme, each expert estimates a corresponding output. Then, a weighted sum of the experts' outputs is computed as in the equation above, where the weights are turned into probabilities through the use of transformations, which standardize the weights' values.

All these values are standardized in order to assure that the gating values of all experts sum up to one. Consider, for example, a certain non-overlapping partition $D_1, \ldots, D_k$ of the points of a data set $D = \{x_1, \ldots, x_n\}$. Let $\Phi(G_p)$ be the basis function that defines a certain transformation, for the p-th subset. Dividing this value by the sum of the basis functions applied over all the subsets in the partition, *i.e.*, $\sum_{p=1}^{k} \Phi(G_p)$ leads to the standardized gating function value g_p. Straightforward conclusions are that $0 \leq g_{ip} \leq 1$ and $\sum_{p=1}^{k} g_{ip} = 1$.

Some commonly used such transformations are

- Inverse Distance

$$g_{ip} = \frac{1/G_{ip}}{\sum_{p=1}^{k} 1/G_{ip}}\,, \quad \text{for } i = 1, 2, \ldots, n \tag{4}$$

- Softmax

$$g_{ip} \frac{\exp[-G_{ip}]}{\sum_{p=1}^{k} \exp[-G_{ip}]}\,, \quad \text{for } i = 1, 2, \ldots, n \tag{5}$$

- Gaussian

$$g_{ip} \frac{\exp\left\{-\frac{1}{2} G_p^T S_p^{-1} G_p\right\}}{\sum_{p=1}^{K} \exp\left\{-\frac{1}{2} G_p^T S_p^{-1} G_p\right\}} \tag{6}$$

  where

$$G_{ip} = \|x_i - \mu_p\|\,, \quad \text{for } i = 1, \ldots, n$$

  is the Euclidean distance between x_i and μ_p; and S_p is the covariance matrix computed on the cluster's p data.

All of these transformations are radial basis functions (Bishop, 1995; Webb, 2002), since they are functions of a distance metric and are radially symmetric with respect to the clusters' centroids. Some options for the distance metric were adopted: the Euclidean distance and its negative and the Mahalanobis distance, which includes a spread parameter. They are found in equations 4, 5 and 6, respectively.

In the case of competitive scheme, only the weights corresponding to one certain model equals to one, while the others are zero. This can be obtained through the "winner-take-all" (WTA) transformation:

$$\begin{aligned} g_p \;=\; & 1\,, \text{ if } G_p = \min_p\{G_p\}\,, \quad \text{for } p = 1, \ldots, K \\ & 0\,, \text{ otherwise} \end{aligned}$$

With the WTA transformation, only one expert (the one which achieves the best fit over the data points of the subset under consideration) is completely responsible for the overall estimate of a certain observation lying within that subset. The other experts do not have any contribution to the final estimate for the given observation. The assignment of observations to the subsets is determined by the least Euclidean distance relative to the clusters centroids.

3 CASE STUDY

According to what was mentioned before, the mixture-of-experts approach may constitute a good alternative for modeling data originated from a diversity of processes. However, many decisions ought to be made when using this technique and, it is important to mention that, different choices may lead to different estimates and performances.

The dataset used for the analysis is the famous Boston Housing data, by Harrison & Rubinfeld (1978), obtained from the UCI Machine Learning Repository (Asuncion & Newman, 2007). We chose to work with a well known dataset precisely because the purpose of this article is not to obtain new results regarding this specific problem – something that has already been exhaustively done (Asuncion & Newman, 2007) – but, again, to illustrate the many decisions when dealing with MEM and their consequences.

The regression problem is to predict the median house price, given a number of demographic variables for a set of houses in suburbs of Boston. There are 506 observations and 8 continuous explanatory variables are used. The response variable is the median house price in the range 0 to 50. The validation data set is comprised of 100 observations randomly selected from the original data set. The remaining 406 observations are used as a calibration data set.

3.1 Methodology

In a general way, the common procedure below was adopted for all cases.

First, the global models were obtained by applying the MSAE, BAYES and ANN models to a calibration (or global) data set. For the BAYES model, the prior information come from the estimates of an OLS regression. The best fitted model was, then, chosen according to some performance indicators.

Then, the input space of the global data was separated into a certain number of clusters, employing the k-means clustering algorithm. Having the observations assigned to the proper clusters, the MSAE, BAYES and ANN models were applied locally to each cluster or local data set. A criterion was again used to establish the best fitted local models (the local experts). The mixture-of-exerts models were obtained by combining the local experts' estimates using different gating functions.

After that, each observation within the validation data set was allocated to a convenient cluster, determined by the least distance of such observation to the clusters centroids. The global and local calibrated models were applied to the validation data set, as well as the resulting mixture-of-experts models, and tests were performed.

The analysis of the different aspects was carried out sequentially, in such a way that the results achieved in a certain step were adopted in the subsequent ones.

The OLS regressions were estimated using standard functions available in the R Statistical Software (R Development Core Team, 2008). The BAYES models were implemented in WinBugs (Lunn *et al.*, 2000), a general-purpose Bayesian computing environment, and the ANN models were built with the aid of the R package AMORE (Limas *et al.*, 2007).

3.2 Results

i. Number of clusters:

Here, we considered the problem of partitioning the calibration data set into different numbers of clusters. In order to analyze the influence of the number of subsets and, consequently, the number of experts that are to form the mixture, we performed the mixture-of-expert procedure using 2, 3, 4 and 5 clusters. Table 1 shows the distribution of data points, according to the clusters they belong to, for all clustering cases.

Table 1 – Number of observations assigned to each cluster for the calibration data set.

Data set	Number of clusters			
	2	**3**	**4**	**5**
Cluster 1	33	213	29	103
Cluster 2	373	33	158	29
Cluster 3	–	160	196	149
Cluster 4	–	–	23	102

The MSAE, BAYES and ANN models were applied to the global and local calibration data sets, in order to compute the global and local estimates. For the local estimates, the different numbers of partitions were adopted. The performance of global and local models was evaluated by means of the least mean absolute deviation (MAD). The results are summarized in Table 2.

From Table 2 it is possible to observe that according to the criterion adopted, the global expert is the ANN model, which achieved the least MAD value (equal to 2.78), among the models applied to the global data set, *i.e.* when all observations are gathered into a single cluster.

Concerning the local estimates, we have distinct results for experiments using different numbers of clusters. It happens because the clusters formed do not necessarily contain the same observations; they do not even have to contain elements in common.

Thus, according to the stated criterion, for the experiment with 2 clusters, for instance, the local experts for clusters 1 and 2 are, respectively, a MSAE regression (MAD equal to 1.57) and an ANN (2.78). For the experiment with 3 clusters we obtained ANN experts for clusters 1 and 3, and a MSAE regression as the expert for cluster 2. It is important to notice that, although the

Table 2 – MAD values for global and local estimates.

Data set	Model	Global set	Number of clusters			
			2	3	4	5
Cluster 1	MSA	3.48	**1.57**	4.01	**1.42**	2.09
	BAY	3.71	1.89	4.35	1.59	2.22
	ANN	**2.78**	1.80	**3.17**	1.47	**1.84**
Cluster 2	MSA	–	3.36	**1.57**	2.10	**1.42**
	BAY	–	3.63	1.89	2.15	1.59
	ANN	–	**2.78**	1.80	**1.75**	1.47
Cluster 3	MSA	–	–	2.13	4.08	4.41
	BAY	–	–	2.17	4.49	4.86
	ANN	–	–	**1.77**	**3.39**	**3.77**
Cluster 4	MSA	–	–	–	1.30	1.88
	BAY	–	–	–	1.42	1.92
	ANN	–	–	–	**0.95**	**1.53**
Cluster 5	MSA	–	–	–	–	1.30
	BAY	–	–	–	–	1.42
	ANN	–	–	–	–	**0.95**

best fitted models for clusters 1 and 3 were of the same structure, *i.e.* both ANN's, they differ from each other by their distinct set of coefficients (or weights).

Similarly, for the experiment with 4 clusters, the expert for cluster 1 is a MSAE regression, while the experts for clusters 2, 3 and 4 are ANN's. For the experiment with 5 clusters, the ANN model wins at clusters 1, 3, 4 and 5, while for cluster 2 the expert is a MSAE regression.

The mixture-of-experts models were built by combining the local experts obtained for each of the partition cases, using the WTA transformation as the gating function. Later, the global expert and the MEM's previously calibrated were applied to the validation data set. Table 3 below shows MAD values achieved by the global expert and the mixture-of-expert models, for both calibration and validation data sets, for all clustering cases.

Table 3 – MAD for final model estimates over calibration and validation data sets.

Data set	Number of clusters				
	Global set	2	3	4	5
Calibration	2.78	2.68	2.49	2.47	2.39
Validation	3.59	3.01	2.84	2.91	2.76
Relative difference	29%	12%	14%	18%	16%

Comparing the values presented in the Global Set column of Table 3 to the subsequent ones, it is possible to observe that the MEM for every cluster configuration performed better than the global expert, for the calibration data set. Also, the results indicate that a better fitted MEM is obtained as the number of clusters increases, with respect to the calibration data set; however, the same does not occur to the validation data set. See also Figure 1.

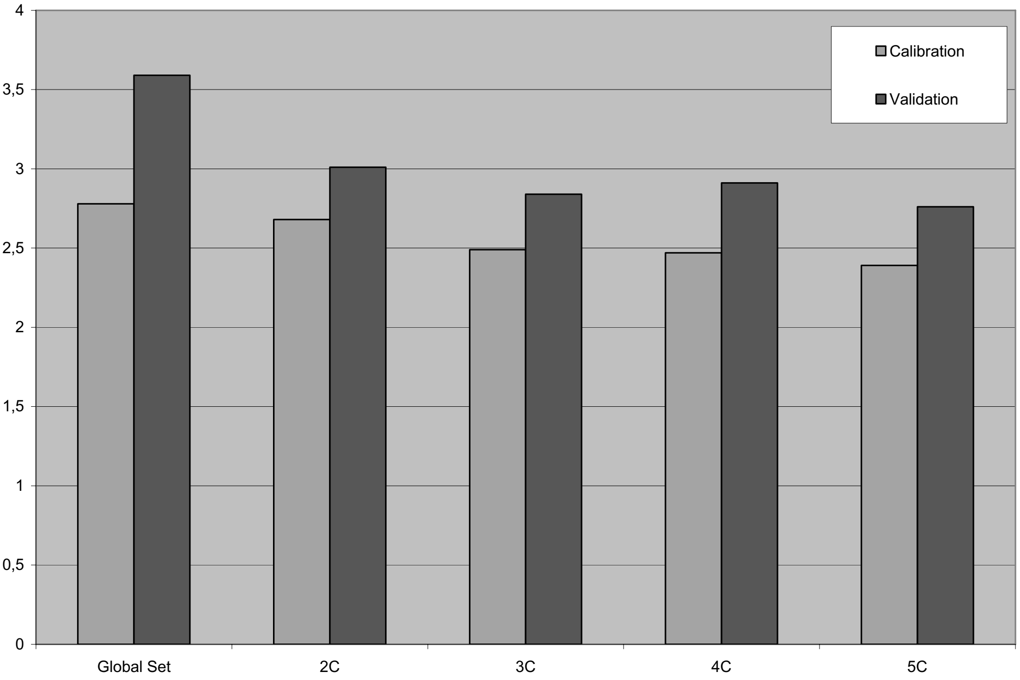

Figure 1 – MAD values for final model estimates.

The last row in Table 3 shows the percent differences in MAD values between calibration and validation data sets, according to the number of clusters adopted. We see that the most consistent configuration, that is, the one which led to the least divergence between MAD values for the calibration and validation estimates, is achieved when the global data set is partitioned into 2 clusters. However, we have better fit in both calibration and validation for any other configuration.

These results show that it is hard to tell which number of clusters would lead to the best fit. In this example, using partitions up to 5 clusters, we observe that the increase in fit is greater if we compare the configuration with 2 and 3 clusters than that between configurations with 3 and 4 or 4 and 5 clusters. That is, the decrease in MAD values is more significant if we move from 2 to 3 clusters (MAD varies 0.19 and 0.17, respectively, for the calibration and the validation data sets) than if we move from 3 to 4 clusters (respectively, 0.02 and -0.07, *i.e.*, in the validation data set the result with 4 clusters is worse than with 3), or from 4 to 5 clusters (0.08 and 0.15). Thus, the maximum increase in the final model performance is achieved for the 3 clusters configuration for this example and therefore this will be our choice in this case study. Nevertheless, there is no guarantee that this is the best choice. With a different criteria (say, for instance, minimum MAD on the validation data set), other choices could be made (5 clusters, in the example), not to mention the fact that, using the same data set, we could have tested for configurations with 6, 7 or more clusters as well.

ii. Criterion for choosing the experts:

Having determined an adequate number of partitions, another factor that was investigated was the criterion adopted to select the expert models, *i.e.*, the models that will compose the final mixture-of-experts model.

We considered the calibration results already obtained for global and local models for the partition into 3 clusters. Here, the models' performances were evaluated in terms of various indicators: the least MAD (mean absolute deviation – the only one we had used up to here), WAPE (weighted absolute percent error), MAPE (mean absolute percent error) and RMSE (root mean squared error).These measures were selected because they weight different characteristics when choosing the best fitted model. MAD, for example, may favor MSAE regression, for it minimizes the sum of absolute deviations. On the other hand, RMSE may favor the ANN model because of the minimization of the sum of square errors. The choice of the gating function remains unchanged.

From Table 4 we observe the performance of the three models applied to the global data set, evaluated in terms of the various criteria. According to all indicators, the global expert is the ANN, for it achieved the least values with respect to all measures.

Table 4 – Evaluation of global estimates.

Indicator	Model		
	MSAE	BAYES	ANN
MAD	3.48	3.71	**2.78**
WAPE	0.15	0.16	**0.12**
MAPE	0.18	0.19	**0.13**
RMSE	5.51	5.26	**4.15**

Table 5 presents the results for the local models. For clusters 1 and 3 the local models were ANN's; they achieved the best performance with respect to all indicators. For cluster 2, however, the criterion for assessing the models' performance has influence on the choice of the local expert. According to MAD, WAPE and MAPE, for instance, the best fitted model is the MSAE regression. On the other hand, the RMSE criterion indicates a different expert for cluster 2; in this case, the ANN.

Because of such divergences in results, we propose an alternative indicator. Thus, (i) the results for all models are evaluated with respect of each indicator and to the best fitted model according to that indicator is assigned a "victory"; (ii) the number of victories of each model is computed; and, finally, (iii) the winner expert is the model with the greatest number of victories.

Table 5 – Evaluation of local estimates.

	Indicator	MAD	WAPE	MAPE	RMSE	Victories
Clust 1	MSAE	4.01	0.19	0.20	6.26	**0**
	BAYES	4.35	0.21	0.23	5.92	**0**
	ANN	**3.17**	**0.15**	**0.16**	**4.22**	**4**
Clust 2	MSAE	**1.57**	**0.12**	**0.13**	2.65	**3**
	BAYES	1.89	0.15	0.15	2.46	**0**
	ANN	1.80	0.14	0.15	**2.45**	**1**
Clust 3	MSAE	2.13	0.08	0.08	2.85	**0**
	BAYES	2.17	0.08	0.08	2.77	**0**
	ANN	**1.77**	**0.07**	**0.07**	**2.38**	**4**

According to this criterion, the resulting local experts are, then, ANN's for clusters 1 and 3 and a MSAE regression for cluster 2. Just for coincidence, this result is identical to that obtained by using only the MAD indicator, as done previously. If another criterion were used in the previous analysis, we could be facing now a change in decision.

Again, the local experts were employed for building the mixture-of-experts model, using the WTA transformation as the gating function. Then, the global expert and the MEM were applied to the validation data set, just like in the former case.

Table 6 summarizes the results achieved by these models, when applied to both the calibration and validation data sets. Also, Table 6 shows columns with values corresponding to the relative differences between the results for these data sets. It corresponds to the consistency of the models, that is, the ability of the models to perform uniformly on both calibration and validation. The best fitted model was, then, chosen according to the number of victories relative to its consistency; thus, the winning model was the mixture-of-experts.

Table 6 – Evaluation of global model and WTA-MEM.

Indicator	Calibration		Validation		Differences	
	Global	MEM	Global	MEM	Global	MEM
MAD	2.78	**2.49**	3.59	**2.84**	29%	**14%**
WAPE	0.12	**0.11**	0.16	**0.13**	28%	**13%**
MAPE	0.13	**0.12**	0.19	**0.14**	41%	**16%**
RMSE	4.15	**3.49**	4.73	**4.12**	**14%**	18%
Victories	**0**	**4**	**0**	**4**	**1**	**3**

iii. Gating function:

The next aspect to be investigated was the influence of different gating functions to the final mixture-of-experts model. The gating functions evaluated were WTA, Inverse Distance, Softmax and Gaussian Transformations.

The estimates for global and local models are the same as those for the 3 clusters configuration. The local experts adopted are those chosen according to the number of victories criterion, presented in Table 5.

The performance of the different mixture models is presented in Table 7. There we observe the indicators of performance for each type of mixture, as well as for the global expert, all applied on both the calibration and validation data sets. Table 7 also shows the relative differences between the calibration and validation values achieved by each model, for all the different performance indicators.

Table 7 shows interesting results. First, if we try to evaluate the different experts based solely in the calibration results, we see that both the WTA and the Softmax are equally good and both achieve the best fit with respect to every performance indicator. The fact that these two models achieve the same performance may seem surprising at a first sight. While the WTA transformation is a hard classifier, so that each locally winning model is completely responsible for the estimation in that subset, the Softmax transformation, on the other hand, leads to a mixture model in which for a given cluster more than one local expert model contribute to the final estimate of an observation in that cluster. However, for this data set, the distances from the observations to centroids of clusters other than that the observation belongs to are so big that the Softmax transformation yields the same weights as those of the WTA transformation.

Table 7 – Evaluation of global estimates and MEM with different gating functions.

Data set	Indicator	Global	EXPERT			
			WTA	Inverse	Softmax	Gaussian
Calibration	MAD	2.78	**2.49**	2.56	**2.49**	3.95
	WAPE	0.12	**0.11**	0.11	**0.11**	0.18
	MAPE	0.13	**0.12**	0.14	**0.12**	0.21
	RMSE	4.15	**3.49**	3.62	**3.49**	5.68
Victories		**0**	**4**	**0**	**4**	**0**
Validation	MAD	3.59	2.84	**2.76**	2.84	4.17
	WAPE	0.16	0.13	**0.12**	0.13	0.18
	MAPE	0.19	0.14	**0.14**	0.14	0.21
	RMSE	4.73	4.12	**4.05**	4.12	5.61
Victories		0	0	4	0	0
Differences	MAD	29%	14%	8%	14%	**6%**
	WAPE	28%	13%	7%	13%	**5%**
	MAPE	41%	16%	3%	16%	**2%**
	RMSE	14%	18%	12%	18%	**1%**
Victories		**0**	**0**	**0**	**0**	**4**

If one decides to evaluate the experts based on the validation results, the winning expert should be the mixture that uses the Inverse transformation gating function. This one achieves the best performance according to all indicators in the validation data set. Moreover, the Inverse Distance mixture is only slightly inferior to the WTA or Softmax experts for the calibration data set.

On the other hand, adopting the consistency criterion, in which one looks for the least divergence between calibration and validation performances, the decision should be favorable to the Gaussian expert. This doesn't seem to be a wise decision, though, since this option leads to a "consistently" bad model. In fact, for both calibration and validation, this model yields the poorest fit according to all indicators.

4 CONCLUDING REMARKS

This work dealt with the investigation of some of the decisions that ought to be made in order to build a mixture-of-experts model. We presented the mixture-of-experts modeling technique applied to solve a regression problem, using different models of different natures as experts.

An illustration of the technique was provided in order to show the consequences of different decisions involved in the process. It employed the Boston housing data and considered the estimation of the median housing price, given a number of demographic variables for a set of houses in suburbs of Boston.

The aspects analyzed in this case study were: (i) the number of clusters into which the data is partitioned; (ii) the criterion used for the selection of experts; and (iii) the gating functions applied for building the mixture-of-experts models. An attempt was also made in order to show how such factors may affect the performance of the final model.

From the results it was possible to notice some remarkable observations:

i) Although the mixture-of-experts model constitutes a sophisticated technique, it does not necessarily lead to more accurate estimates. It would be the case if one decided to use the Gaussian mixture model, for example, which shows poorer performance than even the global expert. In the present case study, the best partition obtained was that of 3 clusters, for it proved to generate the greatest increase in performance, against the modeling effort. This was observed by comparing the results for adjacent clustering configurations.

ii) Depending on the criterion used for evaluating the performance of the candidate models, different experts may be chosen, yielding distinct mixtures. It happens because the each criterion measures distinct fitting characteristics. Thus, this is a difficult decision to make and there is no conclusive choice. For this reason, an alternative indicator was proposed. In order to capture different aspects of models' adherence, the expert was selected as the model to achieve more "victories" with respect to the different performance indicators.

iii) As it was shown in the study case, different gating functions led to different combinations of models and, consequently, to distinct performances of the resulting mixtures. According to the criterion established for measuring the models' performance (*i.e.* the number of victories with respect to consistency), every mixture performed better than the global expert, and the Gaussian mixture had the least discrepancy between calibration and validation for all indicators. However, choosing the Gaussian mixture as the expert proved to be a not very wise decision, since it led to the "consistently" worst model. In this case, taking into account the clusters' variances for weighting the local models did not lead to better results. In fact, a better option would be selecting the Inverse distance mixture, since it had the best performance for validation, while still performing well for calibration.

Other aspect that may highly influence the resulting composition is the clustering method used for partitioning the data set. Soft of fuzzy clustering and Kohonen Self-Organizing Maps (K-SOM) are examples of clustering methods that were not considered in this work and may be possible topics to be investigated in the future. For an application using the K-SOM clustering method, refer to Pinto & Milioni (2004).

As a final remark, it is important to reiterate that the modeling process is always an arduous task and the eager hunt for the "best" model sometimes may not be well rewarded. The mixture-of-experts model's proposal serves as an alternative method for solving complex problems, by allowing a combination of many different simple models to build a more flexible and powerful one. This flexibility makes good promises of improvement in the accuracy of a model, compared to a unique simple model. However, it requires a great number of hypotheses and choices to be made.

Thus, when dealing with the possibility of using a complex modeling technique, it is very important to keep the Occam's razor rule in mind: "Upon those who advocate more complex procedures rest the burden of proof to show that such procedures are worth the added expense" (Armstrong, 1986).

ACKNOWLEDGEMENTS

The authors would like to thank Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for the financial support (grant number 2004/08713-7).

REFERENCES

[1] Alpaydin E & Jordan MI. 1996. Local linear perceptrons for classification. *IEEE Transactions on Neural Networks*, **7**(3): 788–792.

[2] Armstrong JS. 1986. Research on forecasting: a quarter-century review, 1960-1984. *Interfaces*, **16**(1): 89–103.

[3] Asuncion A and Newman DJ. 2007. UCI machine learning repository. http://www.ics.uci.edu/~mlearn/MLRepository.html.

[4] Bishop CM. 1995. *Neural networks for pattern recognition.* Oxford University Press.

[5] Bishop CM and Tipping ME. 2003. Bayesian regression and classification. In: Suykens JA, Horvath G, Basu H, Micchelli C and Vandewalle J (eds.) *Advances in learning theory: methods, models and applications, volume 190 of NATO Science Series III: Computer and Systems Sciences*, chapter 13, pages 267–288. IOS Press.

[6] DUDA RO, HART PE AND STORK DG. 2001. *Pattern classification*. John Wiley and Sons, Inc., USA, 2nd edition.

[7] GELMAN A, CARLIN JB, STERN HS & RUBIN DB. 2004. *Bayesian data analysis*. Chapman and Hall/CRC, USA, 2nd edition.

[8] GUJARATI D. 2000. *Econometria basica*. Makron Books, São Paulo, Brazil, 3rd edition.

[9] HARRISON D & RUBINFELD DL. 1978. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, **5**: 81–102.

[10] HAYKIN S. 1999. *Neural networks: a comprehensive foundation.* Prentice Hall, USA, 2nd edition.

[11] JACOBS RA, JORDAN MI, NOWLAN SJ & HINTON GE. 1991. Adaptive mixture of local experts. *Neural Computation*, **3**(1): 79–87.

[12] KORF RE. 1987. Planning as search: a quantitative approach. *Artificial Intelligence*, **33**: 65–88.

[13] LIMAS MC, MERÉ JBO, GONZÁLEZ EPV, ASCACIBAR FJM DE P, ESPINOZA AVP & ELIAS FA. 2007. *AMORE: A MORE flexible neural network package*. http://wiki.r-project.org/rwiki/doku.php?id=packages:cran:amore. R package version 0.2-11.

[14] LUNN DJ, THOMAS A, BEST N & SPIEGELHALTER D. 2000. WINBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**: 325–337.

[15] MAKRIDAKIS S, WHEELWRIGHT W & HYNDMAN RJ. 1998. *Forecasting Methods and Applications*. John Wiley and Sons, Inc., USA, 3rd edition.

[16] MELO B, NASCIMENTO JR CL & MILIONI AZ. 2004. *Daily Sugar Price Forecasting Using Mixture of Local Expert Models*. Data Mining V: Data Mining, Text Mining and their Business Applications. WIT Press, London, UK, 1st edition.

[17] MILIDIU RL, MACHADO RJ & RENTERIA RP. 1999. Time-series forecasting through wavelets transformation and mixture of expert models. *Neurocomputing*, **28**: 145–156.

[18] NARULA SC. 1987. The minimum sum of absolute errors regression. *Journal of Quality Technology*, **19**(1): 37–45.

[19] NARULA SC & WELLINGTON JF. 2002. Sensitivity analysis for predictor variables in the msae regression. *Computational Statistics and Data Analysis*, **40**: 355–373.

[20] PINDYCK RS & RUBINFELD DL. 1998. *Econometric models and economic forecasts*. Irwin McGraw-Hill, Singapore. 4th edition.

[21] PINTO DB & MILIONI AZ. 2004. Mlem on cross section data. In *Annals of XII Congreso Latino Iberoamericano de Investigacion de Operaciones*, Habana, Cuba.

[22] R DEVELOPMENT CORE TEAM. 2008. R: *A language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. http://www.R-project.org. ISBN 3-900051-07-0.

[23] WATERHOUSE SR. 1997. *Classification and regression using mixtures of experts*. PhD thesis, Department of Engineering, Cambridge University.

[24] WEBB A. 2002. *Statistical pattern recognition*. John Wiley and Sons, Ltd., England, 2nd edition.