Mixture Regression Estimation based on Extreme Learning Machine

Wentao Mao, Yali Wang, Xizheng Cao
School of Computer and Information Engineering, Henan Normal University, China
Email: maowt.mail@gmail.com, yaliwan_g@163.com, caoxizheng@126.com

Yanbin Zheng
School of Computer and Information Engineering, Henan Normal University, China
Engineering Center of Computing Intelligence and Data Mining, Henan, China
Email: zybcgf@163.com

***Abstract*—Recently, Extreme Learning Machine(ELM) has been a promising tool in solving a large range of regression applications. However, to our best knowledge, there are very few researches applying ELM to estimate mixture regression model. To improve the estimation performance, this paper extends the classical ELM to the scenario of mixture regression. First, based on the idea of fuzzy clustering, a set of fuzzy factors are introduced in ELM to measure the degree of membership for a specific class. Furthermore, a new regularization problem is constructed and then the optimal fuzzy factors can be calculated after multiple iterations. Experiments conducted on toy regression data and a structural response prediction data set show the effectiveness of the proposed algorithm compared to the Support Vector Machine-based algorithm in terms of estimation accuracy and computational cost.**

***Index Terms*—extreme learning machine, Mixture regression, regularization, fuzzy clustering**

I. INTRODUCTION

In past decade, extreme learning machine(ELM) [1], has received broad attention in the fields of machine learning and computer vision. Specifically speaking, ELMs extends single-hidden layer feedforward neural network(SLFN) to "generalized" hidden node case. With randomly initialized input weights and hidden layer biases, ELMs seek the minimum of the norm of output weights. Different from the conventional gradient-based learning algorithms, e.g., back-propagation(BP) methods, ELMs analytically calculate the output weights via a simple matrix inversion procedure which results in very high learning speed as well as good generalization performance [2]. Recently, ELMs have shown impressive performance in solving a wide range of real-life problems [3-5].

However, according to our empirical study, ELMs tend to get unsatisfactory results when facing non-linear data sets with large scale. There are two main reasons. One is that a large data set generally contains different models. Another is that measuring error is inevitable in practical experiments. To overcome this shortcoming, a possible solution is using multiple models to describe the data set. This kind of approach is called mixture model estimation. Different methods, e.g., clustering and fuzzy computing, were introduced to solve this problem. As a typical approach, maximum likelihood method [6] generally obtains good results at a great computational cost. Moreover, the applicable probability density function set is very limited. On the contrast, expectation maximization [7] is computationally inexpensive in spite of low convergence speed. Fuzzy clustering is another kind of useful approaches. Xue [8] proposed a multiple-model identification method based on satisfactory fuzzy C clustering. Menard [9] adopted two concepts, ambiguity and distance rejects, to switch various regression models. But this kind of approaches heavily depend on the initial values, and local minimum is unavoidable. Another typical approach is to combine machine learning algorithm with fuzzy estimation. Sun [10] utilized support vector machine(SVM) to conduct fuzzy regression estimation, but the decision model's generalization ability and learning speed are two key parts blocking the further development.

According to the above discussion, the most important factor of mixture model estimation is developing the generalization ability and learning speed of basic learning algorithm. If a more efficient learning machine is introduced, the performance of mixture regression estimation will be improved. Fortunately, ELM can meet these requirements. Inspired by [10], this paper tries to employ ELM as the basic algorithm, and proposes a new algorithm for mixture regression estimation via adding fuzzy clustering. This algorithm can identify various regression models in high speed and precision. To our best knowledge, this research serves as the first attempt to generalize ELM from single model to mixture models.

The paper is organized as follows. In section 2, a brief review to ELM is provided. In section 3, we describe the extension of ELM to the paradigm of mixture models. Section 4 is devoted to computer experiments on two

Corresponding Author: Wentao Mao

doi:10.4304/jcp.8.11.2925-2933

different types of data sets, followed by a conclusion of the paper in the last section.

II. Brief Introduction of ELM

As the theoretical foundations of ELM, [11] studied the learning performance of SLFN on small-size data set, and found that SLFN with at most N hidden neurons can learn N distinct samples with zero error by adopting any bounded nonlinear activation function. Then based on this concept, Huang [2] pointed out that ELM can analytically determine the output weights by a simple matrix inversion procedure as soon as the input weights and hidden layer biases are generated randomly, and then obtain good generalization performance with very high learning speed. Here a brief summary of ELM is provided.

Given a set of i.i.d. training samples $\{(\mathbf{x}_1,\mathbf{t}_1),\cdots,(\mathbf{x}_N,\mathbf{t}_N)\}\subset\mathbb{R}^n\times\mathbb{R}^m$, standard SLFNs with $\tilde{N}$ hidden nodes are mathematically formulated as [2]:

$$\sum_{i=1}^{\tilde{N}}\beta_i g_i(\mathbf{x}_j)=\sum_{i=1}^{\tilde{N}}\beta_i g_i(\mathbf{w}_i\cdot\mathbf{x}_j+b_i)=\mathbf{o}_j,\quad j=1,...,N \tag{1}$$

where $g(x)$ is activation function, $\mathbf{w}_i=[w_{i1},w_{i2},...,w_{in}]^T$ is input weight vector connecting input nodes and the ith hidden node, $\boldsymbol{\beta}_i=[\beta_{i1},\beta_{i2},...,\beta_{im}]^T$ is the output weight vector connecting output nodes and the ith hidden node, b_i is bias of the ith hidden node. Huang [3] has rigorously proved that then for N arbitrary distinct samples and any $(\mathbf{w}_i,b_i)$ randomly chosen from $\mathbb{R}^n\times\mathbb{R}$ according to any continuous probability distribution, the hidden layer output matrix $\mathbf{H}$ of a standard SLFN with N hidden nodes and is invertible and $\|\mathbf{H}\boldsymbol{\beta}-\mathbf{T}\|=0$ with probability one if the activation function $g:\mathbb{R}\mapsto\mathbb{R}$ is infinitely differentiable in any interval. Then given $(\mathbf{w}_i,b_i)$, training a SLFN equals finding a least-squares solution of the following equation [2]:

$$\mathbf{H}\boldsymbol{\beta}=\mathbf{T} \tag{2}$$

where:

$$\mathbf{H}(\mathbf{w}_1,...,\mathbf{w}_{\tilde{N}},b_1,...,b_{\tilde{N}},\mathbf{x}_1,...,\mathbf{x}_{\tilde{N}})=\begin{bmatrix} g(\mathbf{w}_1\cdot\mathbf{x}_1+b_1) & \cdots & g(\mathbf{w}_{\tilde{N}}\cdot\mathbf{x}_1+b_{\tilde{N}}) \\ \vdots & \cdots & \vdots \\ g(\mathbf{w}_1\cdot\mathbf{x}_N+b_1) & \cdots & g(\mathbf{w}_{\tilde{N}}\cdot\mathbf{x}_N+b_{\tilde{N}}) \end{bmatrix}_{N\times\tilde{N}}$$

$$\boldsymbol{\beta}=[\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_{\tilde{N}}]^T$$

$$\mathbf{T}=[\mathbf{t}_1,...,\mathbf{t}_N]^T$$

Considering most cases that $\tilde{N}\ll N$, $\boldsymbol{\beta}$ cannot be computed through the direct matrix inversion. Therefore, Huang [2] calculated the *smallest norm* least-squares solution of (2):

$$\hat{\boldsymbol{\beta}}=\mathbf{H}^{\dagger}\mathbf{T} \tag{3}$$

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose generalized inverse of matrix $\mathbf{H}$ [2]. According to Bartlett's theory [12] that the generalization performance of SLFN will be improved by minimizing training errors as well as the norm of output weights, the solution $\hat{\boldsymbol{\beta}}$ can pledge the generalization ability of SLFN in the theory.

Based the above analysis, Huang [3] proposed ELM whose framework can be stated as follows [2]:

Step 1. Randomly generate input weight and bias $(\mathbf{w}_i,b_i)$, $i=1,\cdots,\tilde{N}$.

Step 2. Compute the hidden layer output matrix $\mathbf{H}$.

Step 3. Compute the output weight $\hat{\boldsymbol{\beta}}=\mathbf{H}^{\dagger}\mathbf{T}$.

Therefore, the output of SLFN can be calculated by $(\mathbf{w}_i,b_i)$ and $\hat{\boldsymbol{\beta}}$:

$$f(\mathbf{x}_j)=\sum_{i=1}^{\tilde{N}}\hat{\beta}_i g_i(\mathbf{w}_i\cdot\mathbf{x}_j+b_i)=\hat{\boldsymbol{\beta}}\bullet h(\mathbf{x}_j)$$

The rigorously theoretical analysis can be found in [2].

III. Mixture Regression Estimation based on ELM

In this section we will first build up the regularization formulation for ELM, and then estimate the the mixture ELM models. As a result, a new approach of mixture model estimation based on ELM is derived. Although the work is somewhat similar to [10], this approach is still specific for ELM.

Huang [2] showed that ELM can be extended to SVM in linear case with less optimization constraints and simpler random kernel. As stated in Section Introduction, the classical ELM seeks the solution with zero training error, e.g., $\|\boldsymbol{\beta}\cdot h(\mathbf{x}_i)-t_i\|=0$. However, considering the acceptable minimal training error, e.g., $\boldsymbol{\beta}\cdot h(\mathbf{x}_i)=t_i-\varepsilon_i$, which exists in many applications, the objective of ELM can be rewritten as minimizing the training error as well as the norm of output weights [2]:

$$\min\ \sum_{i=1}^{N}\|\boldsymbol{\beta}\cdot h(\mathbf{x}_i)-t_i\| \tag{4}$$

and

$$\min\ \|\boldsymbol{\beta}\| \tag{5}$$

From the regularization method point of view, (4) and (5) can be combined into the following formulation:

$$\begin{aligned}&\min\ \frac{1}{2}\|\boldsymbol{\beta}\|^2+C\sum_{i=1}^{N}\varepsilon_i^2\\ &s.t.\quad \varepsilon_i=t_i-\boldsymbol{\beta}\cdot h(\mathbf{x}_i)\end{aligned} \tag{6}$$

where C is regularization parameter which controls the tradeoff between the training error and generalization ability. Note that (6) is similar to LS-SVM except $h(\mathbf{x}_i)$

is generated randomly and the bias b in LS-SVM is not required here.

Now consider that the whole data set $\{(\mathbf{x}_1,\mathbf{t}_1),\cdots,(\mathbf{x}_N,\mathbf{t}_N)\}\subset\mathbb{R}^n\times\mathbb{R}^m$ is composed of multiple models. Without loss of generality, we focus on regression model where the output dimension is one here. The scenario of multiple-dimensional output can be naturally extended. Denote by M the number of models. To estimate various models, we construct the following optimization target based on (6):

$$\begin{aligned}&\min\ L=\frac{1}{2}\sum_{i=1}^{M}\|\boldsymbol{\beta}_i\|^2+C\sum_{i=1}^{M}\sum_{j=1}^{N}\mu_{i,j}^m\varepsilon_{i,j}^2\\&s.t.\quad \varepsilon_{i,j}=t_{i,j}-\boldsymbol{\beta}_i\cdot h_i(\mathbf{x}_{i,j})\\&\qquad\sum_{i=1}^{M}\mu_{i,j}=1,\quad i=1,2,\cdots,M,\quad j=1,2,\cdots,N\end{aligned}\tag{7}$$

In (7), $\mu\in\mathbb{R}^{M\times N}$, and $\mu_{i,j}$ means the membership value of the jth sample belonging to the ith model. $m>1$ controls the ambiguity of clustering results. Generally speaking, $m=2$ is applicable in most applications.

Applying the method of Lagrange multipliers to this problem, we arrive at the following dual optimization problem:

$$\begin{aligned}\max\min\ L=&\frac{1}{2}\sum_{i=1}^{M}\|\boldsymbol{\beta}_i\|^2+C\sum_{i=1}^{M}\sum_{j=1}^{N}\mu_{i,j}^m\varepsilon_{i,j}^2\\&-\sum_{i=1}^{M}\sum_{j=1}^{N}\alpha_{i,j}\left(\boldsymbol{\beta}_i\cdot h_i(\mathbf{x}_{i,j})+\varepsilon_{i,j}-t_{i,j}\right)\\&-\sum_{j=1}^{N}\eta_j\left(\sum_{i=1}^{M}\mu_{i,j}-1\right)\end{aligned}\tag{8}$$

Getting the partial derivative of the parameters respectively, we can obtain:

$$\frac{\partial L}{\partial\boldsymbol{\beta}_i}=0\Rightarrow\boldsymbol{\beta}_i=\sum_{j=1}^{N}\alpha_{i,j}h_i(\mathbf{x}_{i,j})\tag{9}$$

$$\frac{\partial L}{\partial\varepsilon_{i,j}}=0\Rightarrow 2C\mu_{i,j}^m\varepsilon_{i,j}-\alpha_{i,j}=0\tag{10}$$

$$\frac{\partial L}{\partial\mu_{i,j}}=0\Rightarrow Cm\mu_{i,j}^{m-1}\varepsilon_{i,j}^2-\eta_j=0\tag{11}$$

According to (11), we have:

$$\mu_{i,j}=\left(\frac{\eta_j}{Cm\varepsilon_{i,j}^2}\right)^{1/(m-1)}\tag{12}$$

Substituting (12) into the constraint of (7), we have:

$$\eta_j=\left(\frac{1}{\sum_{i=1}^{M}\left(\frac{1}{Cm\varepsilon_{i,j}^2}\right)^{\frac{1}{m-1}}}\right)^{(m-1)}\tag{13}$$

Then the following formula can be obtained:

$$\mu_{i,j}=\frac{1}{(\varepsilon_{i,j})^{2/(m-1)}\sum_{i=1}^{M}(\varepsilon_{i,j})^{2/(1-m)}}\tag{14}$$

On the other side, we have the following formula according to (10):

$$\varepsilon_{i,j}=\frac{\alpha_{i,j}}{2C\mu_{i,j}^m}\tag{15}$$

Therefore, according to (9), (15) and the constraint of (7), we have:

$$\sum_{k=1}^{N}\alpha_{i,k}h_i(\mathbf{x}_{i,k})\cdot h_i(\mathbf{x}_{i,j})+\varepsilon_{i,j}=t_{i,j},\ i=1,2,\cdots,M,\ j=1,2,\cdots,N$$

After introducing ELM kernel [2]: $K_{ELM}(\mathbf{x}_{i,k},\mathbf{x}_{i,j})=h_i(\mathbf{x}_{i,k})\cdot h_i(\mathbf{x}_{i,j})$, the above equation can be rewritten as the following linear case:

$$\left[\mathbf{K}_i+\frac{1}{2C\boldsymbol{\mu}_i^m}I\right]\boldsymbol{\alpha}_i=\mathbf{T}_i\tag{16}$$

where $\mathbf{K}_i$ are the ELM kernel matrix, $\boldsymbol{\mu}_i$, $\boldsymbol{\alpha}_i$, $\mathbf{T}_i$ are the vectors of $\mu_{i,j}$, $\alpha_{i,j}$ and $t_{i,j}$ for the ith class, respectively, and I is the $N\times N$ identity matrix.

Based on the discussions above, we try to estimate different regression models via alternating optimization. This approach includes two steps. First is calculating the Lagrange multipliers $\boldsymbol{\alpha}_i$ for each class using the fixed membership $\boldsymbol{\mu}_i$. Second is obtaining the value of $\boldsymbol{\mu}_i$ by means of the errors between real and predictive outputs. These two steps continue in turn until convergence occurs.

The algorithm of mixture regression estimation for ELM is listed as follows:

Algorithm 1 Mixture regression estimation based on ELM

Input: Class number M, m, regularization parameter C, number of hidden neurons $\tilde{N}$

Output: $\boldsymbol{\mu}_i$, $\boldsymbol{\alpha}_i$, $i=1,2,\cdots,M$

Step 1. Initialize $\boldsymbol{\mu}_i$ and the input weights $\mathbf{w}_i$. Calculate the ELM kernel matrix $\mathbf{K}_i$, $i=1,2,\cdots,M$

Step 2. For $i=1$ to M

Fix $\boldsymbol{\mu}_i$. Calculate the Lagrange multipliers $\boldsymbol{\alpha}_i$ by solving (16)

End for

Step 3. Identify the membership $\boldsymbol{\mu}_i$ using (14) while fixing $\boldsymbol{\alpha}_i$ as:

$$\mu_{i,j} \leftarrow \frac{1}{\left(\varepsilon_{i,j}\right)^{2/(m-1)} \sum_{i=1}^{M} \left(\varepsilon_{i,j}\right)^{2/(1-m)}}$$

Step 4. Go to Step 2 or stop if the change of either $\boldsymbol{\mu}_i$ or $\boldsymbol{\alpha}_i$ between two consecutive iterations is less than a threshold.

It is worth noting that, although this algorithm has similar derivation process with [10], there still exist two apparent differences. One is that the Lagrange multipliers $\boldsymbol{\alpha}_i$ are directly calculated via the linear system (16) rather than the quadratic problem in [10]. Another is that the proposed approach has more simple form which doesn't have the bias parameter b and the kernel matrix $\mathbf{K}$ can be computed directly from inner product between any two nonlinear piecewise continuous functions rather than a fixed kernel function. Therefore, the proposed approach applies especially to ELM.

IV. Experiments

In this section, we run experiments to test the proposed method. For better comparison, we examine two types of regression data sets whose samples come from different models in simulation case. One is toy data and another is an engineering data set, named as structural response prediction data. Our goal is to demonstrate the effectiveness of the proposed algorithm. For comparison, we choose two baselines. First is the traditional fuzzy clustering, called as FC, proposed in [13]. The second is the fuzzy clustering SVM proposed in [10]. For short, it is called FC-SVM in this paper. Correspondingly, the proposed algorithm is named as fuzzy clustering ELM, FC-ELM for short. In two baselines, the Gaussian RBF kernel is used and defined as $K(\mathrm{x}_i, \mathrm{x}_j) = \exp\left(\left\|\mathrm{x}_i - \mathrm{x}_j\right\|^2 / \sigma^2\right)$ [14], where σ is kernel parameter and set 0.2. In FC-SVM and FC-ELM, m is set 2, the regularization parameter C is set 100, and the number of classes is set 2. The number of hidden neurons is set 10. Each of input and output variables are rescaled linearly to the range $[-1,+1]$. All programs are carried out in MATLAB2010a environment running in a Core 2, 2.66GHz CPU and 3.37GB RAM.

A. Toy Data

The first data set is constructed using the following mixture function:

$$f_1(x) = \begin{cases} \operatorname{sinc}(x) + e, & x \in [-2,2] \\ 0.1\sin(\pi x) + 0.1x + 0.4 + e, & x \in [-2,2] \end{cases}, \quad e \sim N(0, 0.02)$$

And the second data set comes from the following mixture function:

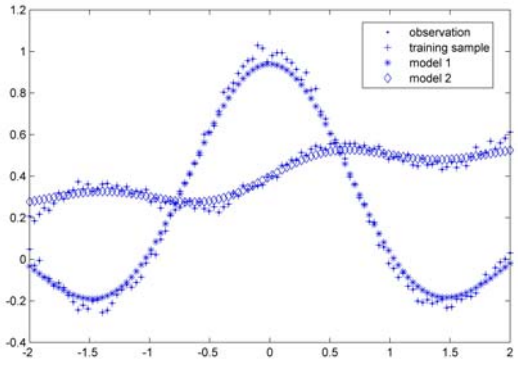
$$f_2(x) = \begin{cases} 10\sin x + e, & x \in [0,10] \\ 0.1\sin\left(\dfrac{\pi x}{2}\right) + 2 + e, & x \in [0,10] \end{cases}, \quad e \sim N(0, 0.5)$$

For better illustration, we choose the single-dimensional input. In each data set, the training set of 140 data points $\{(x_i, y_i)\}$ is generated with x_i drawn uniformly from input space. e_i is a Gaussian noise term. The other 60 data points are randomly selected as test set. According to our empirical study and the results in [10], mixture regression estimation is a NP-hard problem. So, we run experiments 30 times. Here two error evaluations are introduced. One is RMSE, and another is correct classification ratio that means the percentage of getting right classification results. All the experimental results are the mean values of 30 trails.

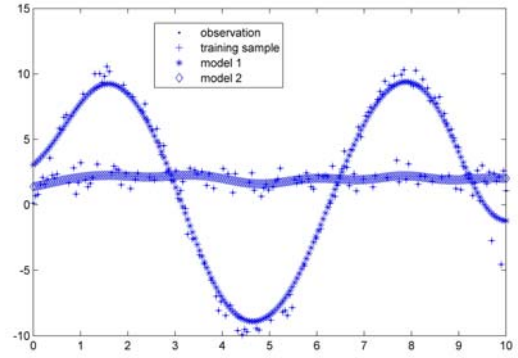
The modeling performance of the proposed algorithm is given in Fig.1.

Obviously, the proposed algorithm can estimate accurately two regression models in each data set. In Fig.1, we use two symbols '*' and 'diamond' to describe the different models. According to the illustration, two regression models are nearly identified, which demonstrates that the proposed algorithm can effectively solve the mixture regression estimation using ELM and fuzzy clustering. Note that in our experiments, we observed in some trials the experiments fail which had also been reported in [10]. For example, Fig.2 provides a incorrect modeling results.

The main reason is the membership factors are initialized randomly. It will bring about the incorrect

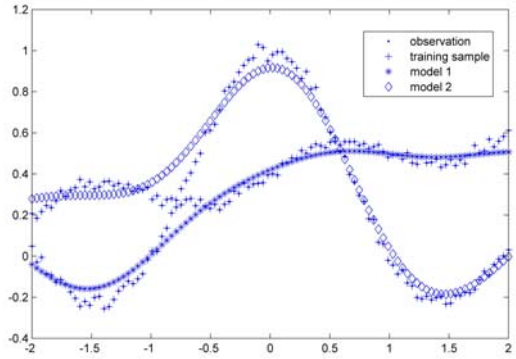


(a) $f_1(x)$

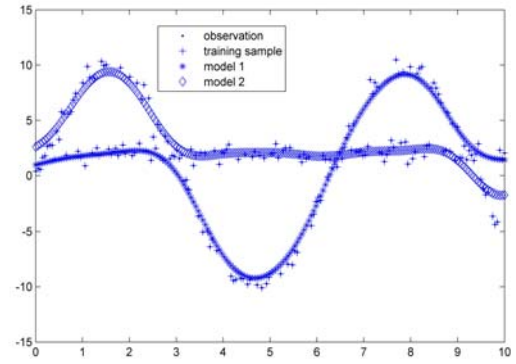


(b) $f_2(x)$

Figure 1. Modeling performance of FC-ELM on two simulation data sets.



(a) $f_1(x)$



(b) $f_2(x)$

Figure 2. Incorrect modeling results of FC-ELM on two simulation data sets

classification results. However, we observed that these wrong results also emerged more frequently in other two baselines than FC-ELM. It can be explained that, in spite of random initial membership factors, the good generalization ability of ELM can adjust the negative affect of randomly initial values. On the contrary, the performance of SVM heavily depends on the choosing of hyper-parameters, i.e., regularization parameter and kernel parameters. It makes FC-SVM cannot improve the results in a flexible manner like ELM. The numerical results of test error also reflect these comparison, as listed in Table I and Table Ⅱ.

The results reported in Table I and Table Ⅱ have show the comparative advantage of the proposed algorithm. At three error indexes, the proposed algorithm all gets better predictive performance than other two approaches. Especially speaking, FC-SVM and FC-ELM both improve the identification performance than the traditional FC approach. The reason is that these two data sets contain the nonlinear regression model which can not be tackled effectively by the traditional FC approach. On the contrast, SVM and ELM are both good at establishing the nonlinear model, which will provide more useful information or domain knowledge for training. This comparison shows the benefits using nonlinear learning algorithm to estimate the mixture regression models. Moreover, FC-ELM gets higher precision and less computational time than FC-SVM, which demonstrates the effectiveness of ELM in mixture regression estimation problem. Especially at computational time, it will give us an engineer choice---an algorithm with high precision and very little computational burden.

TABLE I.
TEST ERRORS ON THE f_1 DATA SET WITH STANDARD DEVIATION IN BRACKET

	FC	FC-SVM	FC-ELM
RMSE	0.6251(0.31)	0.2561(0.23)	**0.0574(0.048)**
Correct ratio(%)	26.67	60.00	**96.67**
Time(s)	2.853	0.538	**0.151**

TABLE Ⅱ.
TEST ERRORS ON THE f_2 DATA SET WITH STANDARD DEVIATION IN BRACKET

	FC	FC-SVM	FC-ELM
RMSE	8.651(4.18)	5.392(2.73)	**3.764(2.14)**
Correct ratio(%)	33.33	73.33	**86.67**
Time(s)	3.145	0.721	**0.244**

B. Structural Resoponse Prediction Data

Our previous work [15] has proved theoretically that the response signals of a structure can be predicted by the response of same structure with different boundary condition. Here this method is called as structural response prediction. It is of important significance in the fields of mechanical manufacturing and aero-craft design, etc. The procedure can be summarized in three steps. First, it needs to select several groups of response signals at different measuring nodes. Second, the training samples should be constructed using the response under one boundary condition as input and another boundary condition as output. Finally, a regression model is established to predict the response by means of the response under other boundary conditions. Note that this method can work in time and frequency domain.

If the amount of response signals is low, the prediction performance cannot be guaranteed due to less domain information. From another perspective, we can gather some neighboring points to establish regression model together. In this scenario, domain knowledge will be increased than single point. According to the dynamic theory, the response at one point will be affected by the neighboring points, which indicates multiple regression models exist in the response at these points. Therefore, the mixture regression method is applicable to solving this problem because it can not only reduce predictive error but also shorten prediction time.

Simplified from many engineering structure, e.g., aircraft and car, cylindrical shell is widely used as a typical structure. Same with our previous work [15], a cylindrical shell vibration system was established to research structural response prediction. The finite element model and corresponding stochastic vibration experimental setup of this cylindrical shell system are shown in Fig.3, which is same with [15].

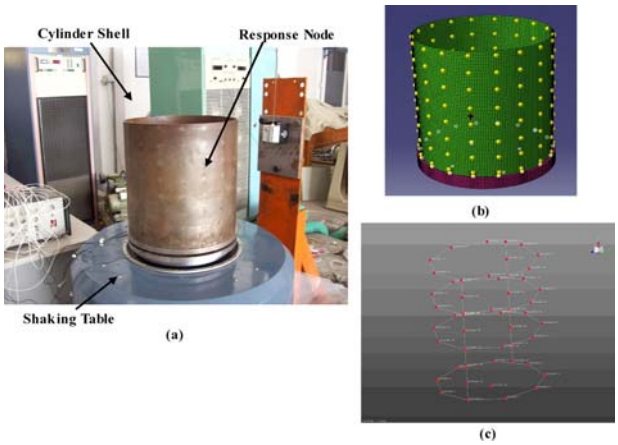


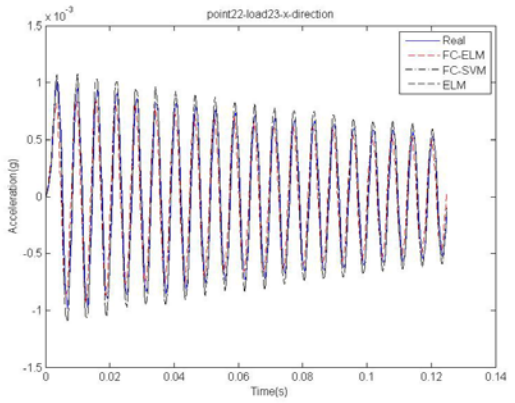
Figure 3. Cylindrical vibration system containing (a)stochastic experimental setup, (b) finite element model and (c)map of measuring points

Our previous work [15] provides the specific parameters of this system. Here a brief introduction is given. The cylindrical shell in Fig.3 is assembled at bottom by a clamp. Specifically, steel and magnesium aluminum (Mg-Al) alloy clamps(shown in Fig.3(a)) are provided to represent two different boundary conditions, respectively. As shown in Fig.3(c), 144 measuring points are placed at the shell's surface in 8 turns. The '+' mark in Fig.3(b) is impulse point, where 30 groups of half a sine pulse loads are acted to produce simulation response in time domain. Note that all simulation response are added by a Gaussian noise term with signal-to-noise ratio 3:1. To imitate the true engineering environment, we also set up an experimental platform (Fig.3(a)). This system is used to generate stochastic vibration data in frequency domain. Correspondingly, 30 groups of simulation impulse in time domain and test acceleration response signals in frequency domains are collected in X, Y, Z directions, respectively. We use LMS Virtual.Lab to pre-process the collected signals.

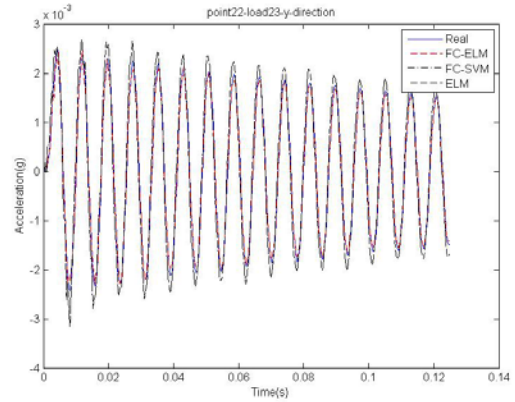
For better illustration, we choose randomly the 23th group of response in total 30 groups for prediction. That means this group is used as test sample while the rest 29 groups of response are used as training samples. We set the response under Mg-AI clamp is prediction target by means of the response under steel clamp. As analyzed above, we choose three neighboring points to construct training samples with the input dimension is 1. Experiments are repeated 30 trials to exclude the stochastic effect.

For better comparison, we randomly choose one measuring point(on circle 4) to draw the prediction curve. We fist check the effectiveness of the proposed approach in time domain. To evaluate the performance, we add the tradition ELM for comparison. The ELM code is at http://www3.ntu.edu.sg/home/egbhuang/elm_codes.html. The comparative results on X, Y, Z directions as shown in Fig. 4.

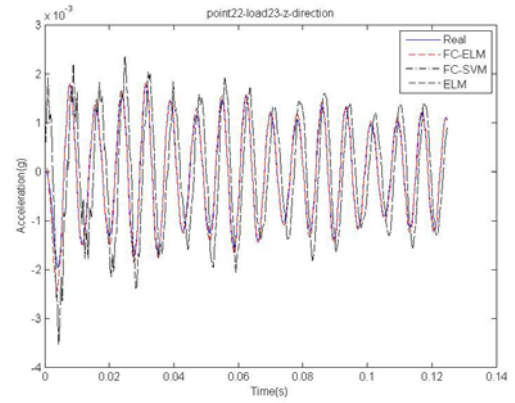
Obviously, FC-ELM gets very similar curve with the true prediction curve, which demonstrates the proposed approach works efficiently. Note that FC-ELM is generalized from classical ELM and FC-SVM is extended from traditional SVM. In Fig.4, FC-ELM works better



(a) X direction



(b) Y direction



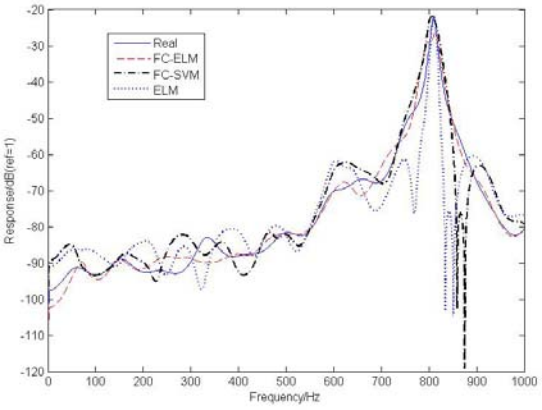
(c) Z direction

Figure 4. Comparative results on shock acceleration response of cylindrical shell in (a) X direction, (b) Y direction and (c) Z direction.

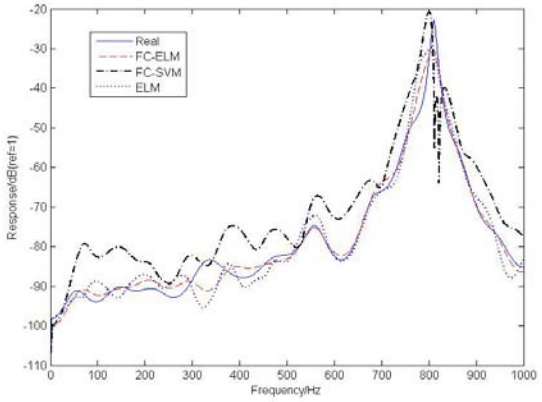
than FC-SVM in all three directions, which means the proposed approach has better generalization ability. This observation also verifies the theoretical analysis in Section Ⅲ. Similar comparative result can be found between FC-ELM and ELM. It is obvious that ELM works unstable due to its randomness. And FC-ELM performs much better than ELM because it can exploit the domain knowledge from the response at other points. Therefore, we can claim the proposed algorithm is a feasible method to solve the problem of structural response prediction. At some peaks, the prediction curve of FC-SVM is even closer to the real curve than ELM. Obviously, FC-ELM and FC-SVM both get significantly

better predictive performance than single learning approach, which also demonstrates the effectiveness of mixture regression approach.. The reason is, as discussed above, when facing few observations, mixture regression approach can exploit much useful information from other models(time points) in order to compensate the negative effect caused by limited number of training samples.

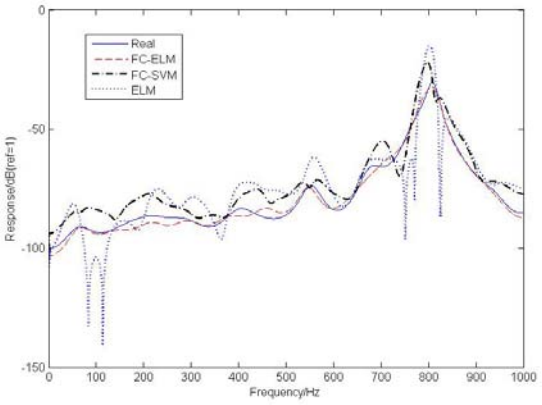
We also evaluate the performance of the proposed approach in frequency domain. In this test, we choose three measuring points at different positions to evaluate, two at circle 3 and one at circle 6. The comparative results are shown in Fig.5. Due to the limitation of test equipment, we can only collect the response in Z directions. Note that in experimental system, additional noise will submerge some response signals at some point near to the shaking table.



(a) First point at circle 3



(b) Second point at circle 3



(c) Third point at circle 6

Figure 5. Comparative results on stochastic vibration response of cylindrical shell at three measuring points.

As shown in Fig.5, at three measuring points the proposed approach all gets lower predictive error than FC-SVM and ELM. At main peak, FC-ELM gets predictive curve more close to the true curve. And ELM generally fluctuates sharply due to its randomness. FC-SVM also gets better performance than ELM, but it still tends to obtain higher errors than FC-ELM. The reason is the model obtained by FC-ELM has lower model freedom and better generalization ability. These comparative results are very similar to the results in Fig.4, which demonstrates the effectiveness of the proposed approach and the significance of mixture regression approach. Note that the comparative results shown in Fig.5 are more distinct than the results in time domain. The reason is in frequency domain the response signals contain more measuring noise while the simulation response signals only contain Gaussian noise. Other measuring points have similar results. Due to the limitation of paper's space, these comparative results are omitted here.

The numerical results also demonstrate the benefit of the proposed approach in terms of mean RMSE and average percentage error(APE) on all 144 measuring points, as listed in Table Ⅲ. It is very obvious that the proposed approach get lowest predictive error in whole test environment. And the mixture regression approach can work better than the single regression approach. The relative improvement reaches 27.89%.

V. CONCLUSIONS

The initial idea of this work is to improve the estimation performance for mixture regression models. In this paper, a new algorithm for mixture regression estimation is proposed. This approach is an extension of classical ELM from single-model to multiple models. The extension includes two steps. The first step is extending classical ELM to the scenario of multiple models in the form of fuzzy clustering. The second step is calculating the membership factors by the error variables which are obtained after estimating the models in each iterative step. These two steps allow us to exploit the inner model's structure hidden in the mixed data through a more effective way. The experimental results on two simulation data sets demonstrate the effectiveness of the

TABLE Ⅲ.
NUMERICAL RESULTS ON ALL MEASURING POINTS

	Simulated shock data			Stochastic vibration data		
	ELM	FC-SVM	FC-ELM	ELM	FC-SVM	FC-ELM
Mean RMSE	2.46e-004	5.82e-005	**3.14e-005**	1.19e-002	7.39e-003	**5.72e-003**
Mean APE	19.46	11.82	**8.65**	27.50	19.72	**14.22**

proposed algorithm.

Totally speaking, this paper has two innovations. In theory, this paper gives an efficient method to solve mixture regression estimation. From engineering perspective, this paper provides a modeling approach for complex regression problem which is generally with small-scale observations. One potential problem is how to overcome the negative effect of randomly initial values of membership factors[16]. This should be achieved by introducing some domain knowledge, and will be studied in our future research.

ACKNOWLEDGMENT

This work was jointly supported by National Natural Science Foundation of China(NO. U1204609) , Foundation and Advanced Technology Research Program of Henan Province, China(NO.122300410111), Key Scientific and Technological Project of Henan Province, China(No. 102102210176, 102102210177) and Young Core Instructor Project from the Higher Education Institutions of Henan Province, China(No.2011GGJS-061).

REFERENCES

[1] G.-B. Huang, Q.-Y. Zhu and C.-K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," in *Proceedings of International Joint Conference on Neural Networks (IJCNN2004)*, Budapest, Hungary, pp. 985-990, 25-29 July, 2004. DOI: 10.1109/IJCNN.2004.1380068

[2] G.-B. Huang, Q.-Y. Zhu and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, pp. 489-501, 2006. DOI: http://dx.doi.org/10.1016/j.neucom.2005.12.126

[3] F. L. Chen and T. Y. Ou, "Sales forecasting system based on gray extreme learning machine with taguchi method in retail industry," Expert Systems with Applications, vol. 38, no. 3, pp. 1336-1345, 2011. DOI: http://dx.doi.org/10.1016/j.eswa.2010.07.014

[4] R. Minhas, A. Baradarani, S. Seifzadeh and Q. M. J. Wu, "Human action recognition using extreme learning machine based on visual vocabularies," Neurocomputing, vol. 73, pp. 1906-1917, 2010. DOI: http://dx.doi.org/10.1016/j.neucom.2010.01.020

[5] A. A. Mohammed, Q. M. J. Wu and M. A. Sid-Ahmed, "Application of wave atoms decomposition and extreme learning machine for fingerprint classification," Lecture Notes in Computer Science, vol. 6112, pp. 246-256, 2010. DOI: 10.1007/978-3-642-13775-4_25

[6] S. B. Caudill and R. N. Acharya, "Maximum-likelihood estimation of a mixture of normal regressions: starting values and singularities," Comments Statistics-Simulation, vol. 27, no. 3, pp. 667-674, 1998. DOI: http://dx.doi.org/10.1080/03610919808813502

[7] L.Yee, J. H. Ma and W. Zhang, "A new method for mining regression classes in large data sets," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 1, pp. 5-21, 2001. DOI: 10.1109/34.899943

[8] Z. Xue and S. Li, "A multi-model modeling approach to MIMO nonlinear systems," Acta Electronica Sinica, vol. 33, no. 1, pp. 52-56, 2005.

[9] M. Menard, P. Dardignac and Y. Courboyla, "Switching regression models using ambiguity and distance rejects: application to ionogram analysis", in Proceedings of Pattern Recognition IEEE Computer Society Press, Barcelona, pp. 688-691, 2000. DOI: http://dx.doi.org/10.1109/ICPR.2000.906168

[10] Z. Sun, X. Yang and Y. Sun, "Fuzzy regression estimation based on support vector machine," Journal of Zhejiang University (Engineering Science), vol. 39, no. 6, pp. 810-813, 2005.

[11] G.–B. Huang, H.A. Babri, "Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions," IEEE Transactions on Neural Networks, vol. 9, no. 1, pp. 224-229, 1998. DOI: 10.1109/72.655045

[12] P.L. Bartlett, "The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network," IEEE Transactions on Information Theory, vol. 44, no. 2, pp. 525–536, 1998. DOI: http://dx.doi.org/10.1109/18.661502

[13] R.J. Hathaway, J.C. Bezdek, "Switching regression models and fuzzy clustering," IEEE Transactions on Fuzzy Systems, vol. 1, no. 3, pp. 195-204, 1993. DOI: http://dx.doi.org/10.1109/91.236552

[14] J. Bai, L. Yang and X. Zhang, "Parameter Optimization and Application of Support Vector Machine Based on Parallel Artificial Fish Swarm Algorithm," Journal of Software, vol. 8, no. 3, pp. 673-679, 2013. DOI:10.4304/jsw.8.3.673-679

[15] W. Mao, M. Tian and G. Yan, "Research of load identification based on multiple-input multiple-output SVM model selection," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 226, pp. 1395-1409, 2012. DOI: http://dx.doi.org/10.1177/0954406211423454

[16] B. Qiu, J. Liu and W. Ma, "Novel Methods for Intuitionistic Fuzzy Multiple Attribute Decision Making," Journal of Software, vol. 7, no. 11, pp. 2553-2559, 2013. DOI: 10.4304/jsw.7.11.2553-2559

Wentao Mao was born in Xin'xiang, Henan Province in 1980. He received his M.S. degree in Computer Science from Chongqing University of Posts and Telecommunications in 2006, and the Ph.D degree in engineering mechanics from Xi'an Jiaotong University, China, in 2011. He is currently working as a lecturer at Henan Normal University, China. Up to now, he has published 20 papers. He is heading over one China National Nature Science Foundation and one Foundation and Advanced Technology Research Program of Henan Province, China. His current research interests include support vector machine, kernel methods and multiple task learning.

Yali Wang was born in Xin'xiang, Henan Province in 1979. She received her Bachelor degree in Computer Science and Technology from Henan Normal University, China in 2001 and M.S. degree in Computer Science from Suzhou University in 2007. Now she is working as a lecturer at Henan Normal University, China. Her current research interests include pattern recognition and kernel methods.

Xizheng Cao was born in Shandong Province in 1976. He received his M.S. degree from Northeast Normal University in 2004, and the Ph.D degree from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, in 2007. He is currently working as an associate professor at

Henan Normal University, China. His main research interest focuses on digital music and pattern recognition.

Yanbin Zheng, born in 1964, received his Ph.D. degree from Beihang University in 2007. Now, he is a full professor and Master supervisor in School of Computer and Information Engineering, Henan Normal University, China. His research interests include the virtual reality, multi-agent network and pattern recognition. Up to now, he has published more 30 papers. As the main researcher, he has directed and taken part in Key Scientific and Technological Project of Henan Province, China, five times.