Bayesian Regularized Committee of Extreme Learning Machine

José M. Martínez-Martínez, Pablo Escandell-Montero, Emilio Soria-Olivas, Joan Vila-Francés and Rafael Magdalena-Benedito

IDAL, Intelligent Data Analysis Laboratory, Electronic Engineering Department, University of Valencia, Av de la Universidad, s/n, 46100, Burjassot, Valencia, Spain

Keywords: Extreme Learning Machine, Committee, Bayesian Linear Regression.

Abstract: Extreme Learning Machine (ELM) is an efficient learning algorithm for Single-Hidden Layer Feedforward Networks (SLFNs). Its main advantage is its computational speed due to a random initialization of the parameters of the hidden layer, and the subsequent use of Moore-Penrose's generalized inverse in order to compute the weights of the output layer. The main inconvenient of this technique is that as some parameters are randomly assigned and remain unchanged during the training process, they can be non-optimum and the network performance may be degraded. This paper aims to reduce this problem using ELM committees. The way to combine them is to use a Bayesian linear regression due to its advantages over other approaches. Simulations on different data sets have demonstrated that this algorithm generally outperforms the original ELM algorithm.

1 INTRODUCTION

A simple and efficient learning algorithm for Single-Hidden Layer Feedforward Neural Networks (SLFNs), called Extreme Learning Machine (ELM), has been recently proposed in (Huang et al., 2006). ELM has been successfully applied to a number of real world applications (Sun et al., 2008; Malathi et al., 2010), showing a good generalization performance with an extremely fast learning speed. However, an issue with ELM is that as some parameters are randomly assigned and remain unchanged during the training process, they can be non-optimum and the network performance may be degraded. It has been demonstrated that combining suboptimal models is an effective and simple strategy to improve the performance of each one of the combination members (Seni and Elder, 2010).

There are different ways to combine the output of several models (Seni and Elder, 2010). The simplest way of combining models is to take a linear combination of their outputs. Nonetheless, some researchers have shown that using some instead of all the available models can provide better performance. In (Escandell-Montero et al., 2012), regularization methods are used such as Ridge regression (Hoerl and Kennard, 1970), Lasso (Tibshirani, 1996) and Elastic Net (Zou and Hastie, 2005) in order to select the models, and the proportion of these, that should be part of the committee.

This paper aims to investigate the use of bayesian linear regression regularization in order to build the committee. The use of this kind of regression involves three main advantages (Bishop, 2007):

1. Regularization. This kind of regression involves a regularization term whose associated parameter is calculated automatically.
2. Calculation of the confidence intervals of the output without the need of applying methods that are computationally intensive, e.g. bootstrap.
3. Introduction of knowledge. Bayesian methods allow the introduction of *a priori* knowledge of the problem

The remaining of this paper is organized as follows. Section 2 briefly presents the ELM algorithm. The details of the proposed method are described in Section 3. Results and discussion are presented in Section 4. Finally, Section 5 summarizes the conclusions of the present study.

2 EXTREME LEARNING MACHINE

ELM was proposed by Huang et al. (Huang et al., 2006). This algorithm makes use of the SLFN architecture. In (Huang et al., 2006), it is shown that

M. Martínez-Martínez J., Escandell-Montero P., Soria-Olivas E., Vila-Francés J. and Magdalena-Benedito R. (2013).
Bayesian Regularized Committee of Extreme Learning Machine.
In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods*, pages 109-114
DOI: 10.5220/0004173901090114

the weights of the hidden layer can be initialized randomly, thus being only necessary the optimization of the weights of the output layer. That optimization can be carried out by means of the Moore-Penrose generalized inverse. Therefore, ELM allows reducing the computational time needed for the optimization of the parameters due to fact that is not based on gradient-descent methods or global search methods.

Let be a set of N patterns, $\mathcal{D} = (\mathbf{x}_i, \mathbf{o}_i); i = 1 \dots N$, where $\{\mathbf{x}_i\} \in \mathbb{R}^{d_1}$ and $\{\mathbf{o}_i\} \in \mathbb{R}^{d_2}$, so that the goal is to find a relationship between $\mathbf{x}_i$ and $\mathbf{o}_i$. If there are M nodes in the hidden layer, the SLFN's output for the j-th pattern is given by $\mathbf{y}_j$:

$$\mathbf{y}_j = \sum_{k=1}^{M} h_k \cdot f(\mathbf{w}_k, \mathbf{x}_j) \tag{1}$$

where $1 \leq j \leq N$, $\mathbf{w}_k$ stands for the parameters of the k-th element of the hidden layer (weights and biases), h_k is the weight that connects the k-th hidden element with the output layer and f is the function that gives the output of the hidden layer; in the case of MLP, f is an activation function applied to the scalar product of the input vector and the hidden weights. The SLFN's output can be expressed in matrix notation as $\mathbf{y} = \mathbf{G} \cdot \mathbf{h}$, where $\mathbf{h}$ is the vector of weights of the output layer, $\mathbf{y}$ is the output vector and $\mathbf{G}$ is given by:

$$\mathbf{G} = \begin{pmatrix} f(\mathbf{w}_1, \mathbf{x}_1) & \dots & f(\mathbf{w}_M, \mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ f(\mathbf{w}_1, \mathbf{x}_N) & \cdots & f(\mathbf{w}_M, \mathbf{x}_N) \end{pmatrix} \tag{2}$$

As mentioned previously, ELM proposes a random initialization of the parameters of the hidden layer, $\mathbf{w}_k$. Afterwards the weights of the output layer are obtained by the Moore-Penrose's generalized inverse (Rao and Mitra., 1972) according to the expression $\mathbf{h} = \mathbf{G}^{\dagger} \cdot \mathbf{o}$, where $\mathbf{G}^{\dagger}$ is the pseudo-inverse matrix.

3 BAYESIAN REGULARIZED ELM COMMITTEE

3.1 Ensemble Methods

A committee, also known as ensemble, is a method that consists in taking a combination of several models to form a single new model (Seni and Elder, 2010). In the case of a linear combination, the committee learning algorithm tries to train a set of models $\{s_1, \dots, s_P\}$ and choose coefficients $\{m_1, \dots, m_P\}$ to combine them as $y(x) = \sum_{i=1}^{P} m_i s_i(x)$. The output of the committee on instance $\mathbf{x}_i$ is computed as

$$y(\mathbf{x}_i) = \sum_{k=1}^{P} m_k s_k(\mathbf{x}_i) = \mathbf{s}_i^T \mathbf{m}, \tag{3}$$

where $\mathbf{s}_i = [s_1(\mathbf{x}_i), \dots, s_P(\mathbf{x}_i)]^T$ are the predictions of each committee member.

The main idea of the proposed method lies in computing the coefficients that combine the committee members using a bayesian linear regression.

3.2 Bayesian Linear Regression

Any Bayesian modeling is carried out in two steps (Congdon, 2006):

- Inference of the posterior distribution of the model parameters. It is proportional to the product of the prior distribution and the likelihood function: $P(\mathbf{w}|D) \propto P(\mathbf{w}) \cdot P(D|\mathbf{w})$ where $\mathbf{w}$ is the set of parameters and D is the data set.
- Calculation of the output distribution of the model, y_{new} (only one output is considered for the sake of simplicity), for a new input $\mathbf{x_{new}}$. It is defined as the integral of the posterior distribution of the parameters $\mathbf{w}$:

$$p(y_{new}|\mathbf{x_{new}}, D) = \int p(y_{new}|\mathbf{x_{new}}, \mathbf{w}) \cdot p(\mathbf{w}|D) \cdot d\mathbf{w} \tag{4}$$

Equation (4) constitutes a natural way of estimating the confidence interval of the model output (Bishop, 2007).

The linear model follows this relationship:

$$y = \mathbf{h}^T \cdot \mathbf{x} + \varepsilon \tag{5}$$

where ε follows a normal distribution with zero mean and variance σ^2, $N(0; \sigma^2)$. Equation (5) leads to the definition of the conditional distribution:

$$p(y|\mathbf{x}, \mathbf{h}, \sigma^2) = \mathcal{N}(\mathbf{h}^T \cdot \mathbf{x}; \sigma^2) \tag{6}$$

In most applications, the parameter distribution is considered to be (Bishop, 2007):

$$p(\mathbf{h}|\alpha) = \mathcal{N}(\mathbf{0}; \alpha^{-1} \cdot \mathbf{I}) \tag{7}$$

where I is the identity matrix and α an hyperparameter. Assuming that the prior distribution and the likelihood function follow Gaussian distributions, the posterior distribution is also Gaussian, with a mean value $\mathbf{m}$ and a variance $\mathbf{S}$ defined as (Bishop, 2007; Chen and Martin, 2009):

$$\mathbf{m} = \sigma^{-2} \cdot \mathbf{S} \cdot \mathbf{X}^T \cdot \mathbf{y} \tag{8}$$

$$\mathbf{S} = \left(\alpha \mathbf{I} + \sigma^{-2} \cdot \mathbf{X}^T \cdot \mathbf{X}\right)^{-1} \tag{9}$$

where $\mathbf{y} = [y_1, y_2, \dots, y_N]$ and $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \dots \mathbf{x_N}]$ are the matrix with the model output vectors and the input vector for those values, respectively.

It is worth noting that the regularization term α in (9) is a natural consequence of the Gaussian approach (Bishop, 2007). This fact differs from other approaches, which requires a term on the cost function being minimized (Deng et al., 2009). Parameters from (9) and (8) are optimized iteratively by means of the ML-II (Berger., 1985) or Evidence Procedure (Barber, 2012). This process applies iteratively expressions (8), (9), (10), (11) and (12); where N is the number of parameters and P is the number of patterns (Bishop, 2007) :

$$\gamma = N - \alpha \cdot trace[S] \tag{10}$$

$$\alpha = \frac{\gamma}{\mathbf{m^T m}} \tag{11}$$

$$\sigma^2 = \frac{\sum_{i=0}^{P} \left(y_i - \mathbf{m^T} \cdot \mathbf{x_i}\right)^2}{P - \gamma} \tag{12}$$

The iterative process is stopped when the difference of the norm of $\mathbf{m}$ between successive iterations falls below a given value. The posterior distribution of the parameters can be applied to (4) in order to obtain the output y_{new} given a new input $\mathbf{x_{new}}$. The output follows a distribution $p(y_{new}|\mathbf{y}, \alpha, \sigma^2) = N\left(\mathbf{h^T} \cdot \mathbf{x_{new}}; \sigma^2\left(\mathbf{x_{new}}\right)\right)$; where the variance is defined as (Bishop, 2007; Chen and Martin, 2009):

$$\sigma^2\left(\mathbf{x_{new}}\right) = \sigma^2 + \mathbf{x_{new}^T} \cdot \mathbf{S} \cdot \mathbf{x_{new}^T} \tag{13}$$

Summing up, using a Gaussian approach in the linear step of the model gives the following advantages (Bishop, 2007; Chen and Martin, 2009; Congdon, 2006):

- *Regularization.* The Bayesian approach involves the use of some parameters (hyperparameters) that allow regularization. This regularization term is obtained from the distribution of the model parameters and helps reducing the overfitting of the model (Bishop, 2007), as we will show in Section 4.
- *Confidence Intervals.* The use of CIs increases the reliability of a model's output. When using neural models, CIs are usually obtained after training the model by means of methods that tend to be computationally costly, e.g. bootstrap (Alpaydin, 2010). The proposed method allows the calculation of CIs and the weight optimization at the same time. This intervals can be calculated easily with the $\mathbf{S}$ matrix (9), the input matrix and with the noise variance that is computed during the iterative calculation of the weights (Barber, 2012).
- *A Priori Knowledge.* A priori knowledge can be introduced in the models by means of error distributions and parameter distributions that must be defined when applying Bayes' theorem. This knowledge can improve the performance of the model.

4 EXPERIMENTAL RESULTS

Several benchmark problems were chosen for the experiments. The data sets were collected from the University of California at Irvine (UCI) Machine Learning Repository[1] and they were chosen due to the overall heterogeneity in terms of number of samples and number of variables. The different attributes for the data sets are summarized in Table 1.

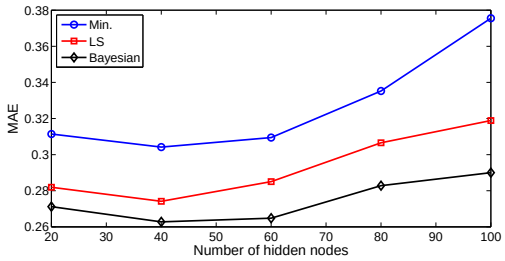
Table 1: Data sets used for the experiments.

Data sets	Samples	Attributes
Housing	506	13
Delta elvevators	9517	6
Abalone	4177	8
Auto	392	7
Autoprice	159	15
Parkinson	5875	21
Add10	9792	10

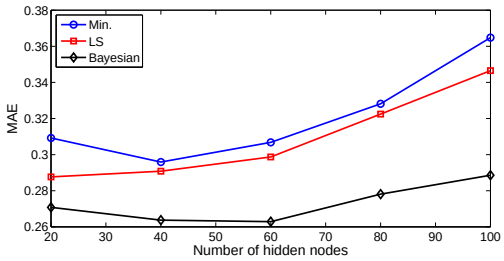
The performance of the proposed approach was evaluated for the previous data sets. We used the following methodology in order to achieve a relative comparison among the several methods:

1. The parameters of the hidden layer of the SLFN were obtained randomly in 50 experiments. The inputs and outputs of the model were standardized (zero mean and unity variance).
2. The number of hidden neurons of each ELM was varied from 20 to 100 neurons with an increment of 20 neurons.
3. For the number of committee members, the same strategy was carried out; committee members from 20 to 100 ELMs were considered with increments of 20 for every different ELM architecture.
4. Two kinds of committees have been tested in this work. On one hand, a linear combination has been proposed, where the coefficients are calculated by least squares. On the other hand, the bayesian linear combination previously stated.
5. In the seven tackled problems, the training data set was formed by 50% of the patterns, and the remaining 50% were used for validation purposes.

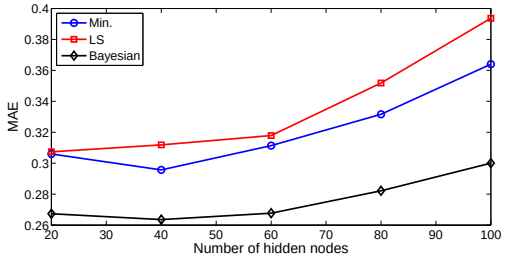
[1] http://archive.ics.uci.edu/ml/



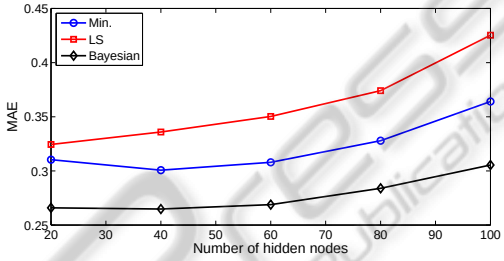
(a) MAE for a committee composed of 20 members.



(b) MAE for a committee composed of 40 members.

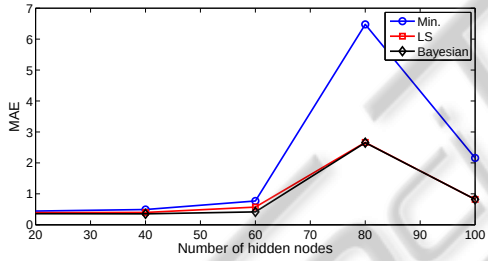


(c) MAE for a committee composed of 60 members.

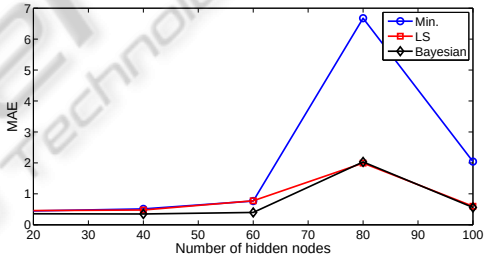


(d) MAE for a committee composed of 80 members.

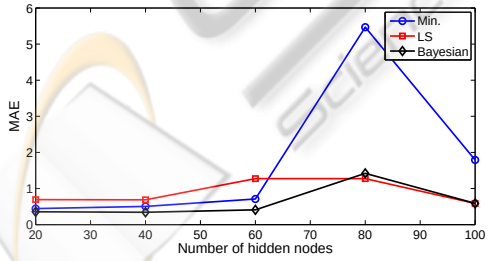
Figure 1: Performance in terms of MAE in the validation set of the proposed algorithm compared with the linear comittee of ELMs and with the member of the committee which presented the minimum error for *Abalone* data set.



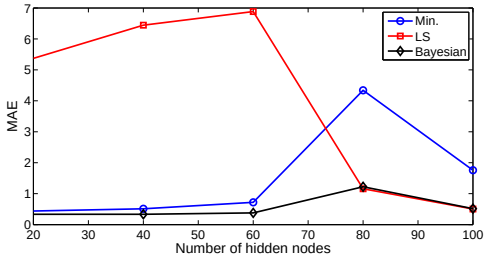
(a) MAE for a committee composed of 20 members.



(b) MAE for a committee composed of 40 members.



(c) MAE for a committee composed of 60 members.



(d) MAE for a committee composed of 80 members.

Figure 2: Performance in terms of MAE in the validation set of the proposed algorithm compared with the linear comittee of ELMs and with the member of the committee which presented the minimum error for *Autoprice* data set.

Each pattern was assigned randomly to one of the two sets (either training or validation) for each experiment.

Table 2 shows the performance in terms of MAE (Mean Absolute Error) in the validation set of the proposed method, the bayesian regularized ELM committee (Bayes.), in comparison with the results of the linear committee (LS), where the coefficients are calculated by least squares as previously mentioned. The MAE was computed according with (14):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (14)$$

where y_i is the observed output and $\hat{y}_i$ is the output predicted by the model.

Another model is included in Table 2 (Min.), which refers to the minimum MAE of the ELM network inside the whole committee of ELMs, that is, the member of the committee which presented the minimum MAE. In this way, we obtain a reference value that indicates if the committee provides best results as a whole than the best ELM network that takes part of the committee.

Table 2 shows the median value of the 50 experiments that were considered for each configuration of ELM-committee. Moreover, only three of the five values of the different number of neurons in the hidden layer tested are presented for the sake of simplicity. This values are 20, 60, 100 initial number of neurons, which corresponds with first, second and third rows of each method respectively (Table 2). The columns corresponds to the several number of committee members tested, which were varied from 20 to 100 with increments of 20, as mentioned previously.

To summarize the information contained in Table 2, a comparison of three methods has been carried out. To do such comparison, the value of the MAE in each method, for each number of committee tested (columns) and each number of hidden neurons tested, is compared. The result of this comparison shows that, in general terms, the proposed method outperform the other two methods. Specifically, the 71.43% of the times the proposed method won, the 9.52% of the times the linear committee won and the 19.05% of the times they tied. The third method, the member of the committee which presented the minimum MAE, never was the best method.

In order to illustrate the performance of the proposed algorithm, graphically, compared with the linear comittee of ELMs and with the member of the committee which presented the minimum error, we present Figures 1 and 2. These figures show the results for *Abalone* (Figure 1) and *Autoprice* (Figure 2) data sets. Each figure presents four cases that correspond to several committees. The first case corresponds to a committee composed of 20 members, the second case to one of 40 members, the third case to one of 60 members and, finally, the latter case corresponds to a committee composed of 80 members.

Notice that for *Abalone* data set, Figures 1a and 1b the bayesian regularized ELM committee presents always the minimum MAE; the second best method is the linear committee of ELM. However, Figures 1c and 1d show that the proposed method presents always the minimum MAE again, but in these cases (60 and 80 committee members) the second best method is Min. (the member of the committee which presented the minimum MAE).

For *Autoprice* data set, Figures 2a and 2b show

Table 2: Performance in terms of MAE in the validation set of the proposed method (Bayes.), in comparison with the results of the linear committee (LS) and the member of the committee which presented the minimum MAE.

Data set	Method	C1	C2	C3	C4	C5
Abalone	Min.	0,311	0,309	0,306	0,310	0,309
		0,309	0,307	0,311	0,308	0,312
		0,376	0,365	0,364	0,364	0,364
	LS	0,282	0,288	0,307	0,324	0,371
		0,285	0,299	0,318	0,350	0,388
		0,319	0,347	0,394	0,425	0,486
	Bayes.	0,271	0,271	0,267	0,266	0,262
		0,265	0,263	0,268	0,269	0,273
		0,290	0,289	0,300	0,305	0,309
Add10	Min.	0,497	0,491	0,494	0,493	0,495
		0,425	0,425	0,427	0,427	0,425
		0,407	0,407	0,407	0,407	0,407
	LS	0,408	0,398	0,394	0,390	0,384
		0,375	0,361	0,348	0,339	0,331
		0,353	0,336	0,323	0,313	0,301
	Bayes.	0,408	0,398	0,394	0,389	0,383
		0,375	0,361	0,348	0,338	0,331
		0,353	0,336	0,323	0,313	0,301
Auto	Min.	0,311	0,309	0,306	0,310	0,309
		0,309	0,307	0,311	0,308	0,312
		0,376	0,365	0,364	0,364	0,364
	LS	0,282	0,288	0,307	0,324	0,371
		0,285	0,299	0,318	0,350	0,388
		0,319	0,347	0,394	0,425	0,486
	Bayes.	0,271	0,271	0,267	0,266	0,262
		0,265	0,263	0,268	0,269	0,273
		0,290	0,289	0,300	0,305	0,309
Autoprice	Min.	0,444	0,448	0,444	0,438	0,424
		0,769	0,768	0,710	0,720	0,732
		2,159	2,043	1,791	1,758	1,773
	LS	0,391	0,459	0,692	5,373	1,850
		0,571	0,776	1,273	6,883	1,791
		0,822	0,599	0,587	0,504	0,505
	Bayes.	0,359	0,356	0,356	0,335	0,341
		0,416	0,399	0,410	0,379	0,397
		0,822	0,560	0,587	0,515	0,505
Delta elev.	Min.	0,480	0,477	0,479	0,477	0,477
		0,463	0,462	0,463	0,462	0,461
		0,460	0,459	0,460	0,459	0,459
	LS	0,459	0,457	0,457	0,456	0,456
		0,455	0,454	0,455	0,455	0,455
		0,453	0,454	0,455	0,455	0,456
	Bayes.	0,459	0,456	0,456	0,454	0,453
		0,454	0,453	0,453	0,452	0,451
		0,453	0,453	0,453	0,452	0,452
Housing	Min.	0,439	0,440	0,440	0,441	0,438
		0,415	0,408	0,410	0,403	0,406
		0,422	0,412	0,420	0,419	0,416
	LS	0,374	0,376	0,383	0,390	0,388
		0,330	0,331	0,348	0,346	0,361
		0,309	0,317	0,327	0,341	0,359
	Bayes.	0,364	0,351	0,348	0,339	0,327
		0,319	0,308	0,310	0,302	0,300
		0,301	0,286	0,289	0,285	0,287
Parkinson	Min.	0,753	0,755	0,754	0,752	0,750
		0,702	0,704	0,705	0,702	0,700
		0,669	0,669	0,671	0,669	0,669
	LS	0,704	0,678	0,658	0,647	0,631
		0,633	0,616	0,605	0,592	0,589
		0,594	0,581	0,574	0,565	0,560
	Bayes.	0,705	0,679	0,659	0,648	0,633
		0,633	0,617	0,606	0,592	0,589
		0,593	0,582	0,575	0,565	0,558

that the proposed method and the linear committee of ELMs are very similar; although the proposed method presents less MAE in some cases. However, Figures 2c and 2d show that, in general terms, the proposed method presents the minimum MAE again, but in these cases, when the number of hidden neurons is 60, the linear committee of ELMs presents less MAE, although the difference is almost negligible, specially in Figure 2c.

Summarizing, in general terms the proposed method outperforms the linear comittee of ELMs and the member of the committee which presented the minimum MAE. Moreover, it presented more robustness than the mentioned methods. This can be seen in Figure 2d, where the only method that is almost constant is the proposed one.

5 CONCLUSIONS

This paper aims to investigate the use of bayesian linear regression regularization with the intention of building a committee of ELM in order to avoid the problem of local minima found in this emergent neural network. The use of this kind of regression involves three main advantages:

1. Regularization. This kind of regression involves a regularization term whose associated parameter is calculated automatically.
2. Calculation of the confidence intervals of the output without the need of applying methods that are computationally intensive, e.g. bootstrap.
3. Introduction of knowledge. Bayesian methods allow the introduction of *a priori* knowledge of the problem.

A performance comparison of this method with the linear comittee of ELMs and with the member of the committee which presented the minimum error has been carried out on widely used benchmark problems of some real-world regression problems.

Summarizing, the proposed method not only keeps the advantage of extremely fast training speed but also solves the main inconvenient of this technique; the local minima problem. In general terms the proposed method outperforms the linear comittee of ELMs and the member of the committee which presented the minimum MAE. Moreover, it presented more robustness than the mentioned methods. Another advantage is that due to the fact that this method uses a regularization method entails that the generalization ability improves.

REFERENCES

Alpaydin, E. (2010). *Introduction to Machine Learning*. MIT Press, 2nd edition.

Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

Berger., J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer, 1st ed. 2006. corr. 2nd printing edition.

Chen, T. and Martin, E. (2009). Bayesian linear regression and variable selection for spectroscopic calibration. *Anal. Chim. Acta*, 631(1):13–21.

Congdon, P. (2006). *Bayesian Statistical Modelling*. Wiley.

Deng, W., Zeng, Q., and Chen, L. (2009). Proc. ieee symp. comput. intell. data mining.

Escandell-Montero, P., Martínez-Martínez, J. M., Soria-Olivas, E., Guimerá-Tomás, J., Martínez-Sober, M., and Serrano-López, A. J. (2012). Regularized committee of extreme learning machine for regression problems. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2012. ESANN '12*.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Huang, G., Zhu, Q.-Y., and Siew, C. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, 70:489–501.

Malathi, V., Marimuthu, N., and Baskar, S. (2010). Intelligent approaches using support vector machine and extreme learning machine for transmission line protection. *Neurocomputing*, 73(10-12):2160 – 2167.

Rao, C. R. and Mitra., S. K. (1972). *Generalized Inverse of Matrices and Its Applications*. Wiley.

Seni, G. and Elder, J. (2010). *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. Morgan and Claypool Publishers.

Sun, Z.-L., Choi, T.-M., Au, K.-F., and Yu, Y. (2008). Sales forecasting using extreme learning machine with applications in fashion retailing. *Decision Support Systems*, 46(1):411 – 419.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67:301–320.