ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Regularized Committee of Extreme Learning Machine for Regression Problems

Pablo Escandell-Montero, José M. Martínez-Martínez,
Emilio Soria-Olivas, Josep Guimerá-Tomás, Marcelino Martínez-Sober
and Antonio J. Serrano-López

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

Abstract. Extreme learning machine (ELM) is an efficient learning algorithm for single-hidden layer feedforward networks (SLFN). This paper proposes the combination of ELM networks using a regularized committee. Simulations on many real-world regression data sets have demonstrated that this algorithm generally outperforms the original ELM algorithm.

1 Introduction

Extreme learning machine (ELM) is an efficient learning algorithm for single-hidden layer feedforward networks (SLFN) recently proposed in [1]. It dramatically reduces the learning time by means of randomly selecting weights and biases for hidden nodes instead of adjust them iteratively, the common approach employed by gradient-descent methods. ELM has shown a good generalization performance in several real-world applications. 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 [2]. We propose to use an ensemble of ELM networks whose parameters are initialized independently and combine their predictions to generate a final output.

There are different ways to combine the output of several models [2]. 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. The main difficulty of this approach is the selection of the models that should be part of the committee. This paper aims to investigate the use of regularization methods in order to select automatically the committee members.

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. Section 4 introduces the experiments and the used data sets. Results and discussion are presented in Section 5. Finally, Section 6 summarizes the conclusions of the present study.

2 Extreme Learning Machine

ELM was proposed by Huang et al. [1]. This algorithm makes use of the SLFN architecture. In [1], it is shown that 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 $\mathbf{y_j}$:

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

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 Multilayer Perceptron (MLP), f is an activation function applied to the scalar product of the input vector and the hidden weights. Eq.(1) 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:

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

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 according to the expression $\mathbf{h} = \mathbf{G}^{\dagger} \cdot \mathbf{o}$, where $\mathbf{G}^{\dagger}$ is the pseudo-inverse matrix.

3 Regularized ELM Committee

A committee, also known as ensemble, is a method that consists of taking a combination of several models to form a single new model. 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 $\{\beta_1, \dots, \beta_P\}$ to combine them as $y(x) = \sum_{i=1}^{P} \beta_i s_i(x)$. The output of the committee on instance $\mathbf{x}_i$ is computed as

$$y(\mathbf{x}_i) = \sum_{r=1}^{P} \beta_r s_r(\mathbf{x}_i) = \mathbf{s}_i^T \boldsymbol{\beta}, \tag{3}$$

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

The main idea of the proposed method lies in computing the coefficients that combine the committee members using a regularized version of least squares regression. The regularized regression is useful in this context due to its tendency to prefer solutions with fewer nonzero parameter values, effectively reducing the number of committee members.

3.1 Lasso, Ridge Regression and the Elastic Net

Multiple linear regression is often used to estimate a model for predicting future responses, or to investigate the relationship between the response variable and the predictor variables. For the first goal, the prediction accuracy of the model is important, while for the second goal the size of the model is of more interest. Ordinary Least Squares (OLS) regression is known for often not performing well with respect to both prediction accuracy and model size [6]. Several regularized regression methods were developed the last few decades to overcome these flaws of OLS regression, starting with Ridge regression, followed by Lasso method, and more recently the Elastic net [3, 4].

Ridge regression and the Lasso are regularized versions of least squares regression using penalties on the coefficient vector. Recently, [5] proposed the Elastic net to reach a compromise between the Lasso and Rigde regression. The Elastic net also combines shrinkage and variable selection, and in addition encourages grouping of variables [6].

We consider the usual setup for linear regression with only one output, for the sake of simplicity. We have a response variable $o \in \mathbb{R}$ and a predictor vector $\mathbf{s} \in \mathbb{R}^Q$, and we approximate the regression function by a linear model $E(o|\mathbf{s}) = \beta_0 + \mathbf{s}^T\boldsymbol{\beta}$, where the input to the model, $\mathbf{s}$, is formed by the outputs of the committee members. We consider N observation pairs $(\mathbf{x}_i, \mathbf{o}_i)$.

The three regularization methods for linear models can be described in a generalized way as

$$\min_{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}^{Q+1}} \left[\frac{1}{2N}\sum_{i=1}^{N}(o_i - \beta_0 - \mathbf{s}_i^T\boldsymbol{\beta})^2 + \lambda P_\alpha(\boldsymbol{\beta})\right] \tag{4}$$

where

$$P_\alpha(\boldsymbol{\beta}) = \sum_{j=1}^{Q}\left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right] \tag{5}$$

In the Elastic net penalty P_α is a trade-off between the ridge Regression penalty ($\alpha = 0$) and the Lasso penalty ($\alpha = 1$) [6]. Thus, Ridge Regression introduce a penalty $\|\boldsymbol{\beta}\|^2$, Lasso replaces this penalty by a penalty $\|\boldsymbol{\beta}\|$, and Elastic net introduce a penalty of the form $\lambda_1\|\boldsymbol{\beta}\| + \lambda_2\|\boldsymbol{\beta}\|^2$.

The Elastic net with $\alpha = 1 - \gamma$ for some small $\gamma > 0$ performs much like the Lasso, but removes any degeneracies and wild behavior caused by extreme correlations. More generally, the entire family P_α creates a useful trade-off between Ridge and Lasso.

3.2 Selection of the regularization parameter λ

As mentioned in Section 3.1, the regularized regression methods depend on a parameter λ. This parameter gives information about the relevance of the regularization. Many methods can be applied to solve Eq. (4), here, the Coordinate Descent method proposed in [6] was used due to its high speed compared to other classical approaches. A freeware MATLAB$^{\text{TM}}$ toolbox that implements these algorithms was used[1]. Following the procedure pointed out in [6], a sequence of K values of λ decreasing from λ_{max} to λ_{min} on the log scale is constructed. The values of λ_{max} and λ_{min} are derived from data, and a typical value of K is 100.

The *Bayesian Information Criterion* (BIC) [7] was used in order to select the best model. This criterion introduces a penalty term for the number of parameters in the model, the BIC criterion is defined as

$$BIC = -2 \cdot ln(L) + M \cdot ln(N) \tag{6}$$

where:

N = the number of observations, or the sample size.

M = the number of parameters to be estimated (the number of committee members).

L = is the value of the likelihood function for the estimated model.

4 Experiments

A total of 13 benchmark regression problems were chosen to study the performance of the proposed approach. They were chosen due to the overall heterogeneity in terms of number of samples and number of variables. All data sets were collected from LIACC[2] repository, except *Concrete compressive* data set, that can be found in UCI[3] repository. Data sets were standardized to zero mean and unit variance. One-third of each data set was selected randomly for validating, and the remaining for training. Some statistics of the data sets are shown in Table 1.

The proposed approach was compared with standard ELM network and with a committee without regularization. In the ELM network, the number of hidden nodes was varied in order to select the optimal architecture. For each architecture, it was carried out a total of 100 different initializations of network parameters (randomly generated within the range [-1 1]). The sigmoidal additive activation function was used.

The number of committee members was fixed to 20 for all data sets. Therefore, both committees are initially formed by 20 ELM networks. However, in the regularized committee, some of the members will be discarded during the learning process. In each committee, the member architecture is the same, but their weights and biases are different (randomly generated). On the other hand, for

[1] www-stat.stanford.edu/∼tibs/glmnet-matlab

[2] http://www.liaad.up.pt/∼ltorgo/Regression/DataSets.html

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

Data set	# Attri.	Samples		Data set	# Attri.	Samples	
		Train	Val.			Train	Val.
Abalone	8	2784	1393	Machine CPU	6	139	70
Ailerons	40	9166	4584	Delta ailerons	5	4752	2377
Auto price	15	106	53	Delta elev.	6	6344	3173
Bank	8	5461	2731	House Census	8	15189	7595
Boston h.	13	337	169	Kinematics	8	5461	2731
California h.	8	13760	6880	Triazines	60	124	62
Concrete c.	8	686	344				

Table 1: Information about the selected data sets. Number of attributes and number of samples for both training (two-thirds of the training data) and validation (remainder third of the data) sets.

each data set the member architecture varies and coincides with the architecture employed by the standard ELM. Several values of the parameter α corresponding with the three regularization methods were tested. Finally, α was set to 0.2 because this value provided the best overall performance.

The performance was measured in terms of the RMSE in the validation set and the experiments were repeated 50 times. The averaged predictive error is shown in Table 2.

5 Results

All results reported are for the validation set. For each data set, the minimum RMSE is highlighted in bold face. As it can be observed in Table 2, the proposed approach obtains the best general performance. In 12 of the 13 data sets, the regularized committee provides better results than the other methods. As expected, both committee methods improve the error obtained with the standard ELM. Moreover, in the majority of cases, the fact of employing some instead of all the committee members provide better performance.

The improvement achieved by the regularized committee varies with the data set. For example, compared with ELM, the RMSE improvement of the datasets *Boston housing*, *Concrete compressive*, *Machine CPU* and *Kinematics* is more than 15%. Regarding the regularization process, the average (over all data sets) number of committee members discarded during the learning process is 6.83, which corresponds with a 34.13% of the members.

6 Conclusions

In this paper, we have proposed a regularized committee formed by ELM networks. Based on the results, we can conclude that the performance of the standard ELM network can be outperformed using a regularized committee. Furthermore, given a set of networks, it is better to build a committee that contains

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Data set	# Hidden nodes	Method		
		ELM	Linear committee	Regularized committee
Abalone	40	0.6557	0.6532	**0.6468***
Ailerons	600	0.4524	0.4183	**0.4179***
Auto price	20	0.5567	0.5699	**0.4947***
Bank	400	0.2047	0.1954	**0.1948***
Boston housing	80	0.4834	0.4230	**0.4088***
California housing	400	0.5072	0.4863	**0.4855***
Concrete compressive	140	0.4506	0.3723	**0.3654***
Machine CPU	25	0.3927	0.3155	**0.2666***
Delta ailerons	80	0.5329	0.5269	**0.5247***
Delta elevators	100	0.6033	0.5998	**0.5990***
House census	400	0.6159	0.5930	**0.5920***
Kinematics	400	0.4608	**0.3882**	0.3906
Triazines	10	1.0428	1.0460	**1.0395**

Table 2: Average validation RMSE (50 experiments) obtained with standard ELM, linear committee and regularized committee for each data set. Also it is shown the number of hidden nodes employed for each data set.

some instead of all networks. The difficulty here is to decide what networks should be part of the committee. It has been presented a method based on elastic net regularization in order to select the committee members.

The proposed method has been compared with standard ELM network and a committee in 13 benchmark regression problems, and the results indicate that the algorithm produces better results.

References

[1] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3):489–501, December 2006.

[2] Giovanni Seni and John Elder. *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. Morgan and Claypool Publishers, February 2010.

[3] Anita J. Van der Kooij. *Prediction accuracy and stability of regression with optimal scaling transformations*. PhD thesis, Leiden University, 2007.

[4] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, corrected edition, July 2003.

[5] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67:301—320, 2005.

[6] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1), January 2010.

[7] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

*Regularized committee RMSE is significantly lower ($p < 0.001$, one-tailed paired t-test) than linear committee RMSE.