ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

An affinity matrix approach for structure selection of extreme learning machines

David Pinto, Andre P. Lemos and Antonio P. Braga

Federal University of Minas Gerais - Dept of Electronics,
Belo Horizonte, Minas Gerais - Brazil

Abstract. This paper proposes a novel pruning approach for Extreme Learning Machines. Hidden neurons ranking and selection are performed using *a priori* information expressed by affinity matrices. We show that the similarity between the affinity matrix of the input patterns and the affinity matrix of the hidden layer output patterns can be seen as a measure of the data structural retention through the network. However, from a certain similarity level, adding new hidden nodes will have small or no effect on the amount of information propagated from the input. The proposed approach automatically determines this level and hence the suitable number of hidden nodes. Experiments are performed using classification problems to validate the proposed approach.

1 Introduction

The Extreme Learning Machine (ELM) [1] is a learning algorithm for single layer feedforward network (SLFNs) with a low time complexity able to deal with massive datasets at a fast training convergence and with a good generalization performance. The random definition of the hidden layer parameters and the estimation of the output weights by Ordinary Least Squares (OLS) allow training in a single iteration.

The original ELM formulation does not establish any methodology for SLFN structure definition. The hidden layer dimension is often inferred by trial and error. Several modifications to the ELM algorithm have already been proposed to cope with this issue [2, 3, 4]. However, most of the existing methods demands the use of a validation dataset and/or the computation of the network parameters for each candidate structure.

This paper proposes a novel modification to the ELM learning algorithm, the Affinity Based Pruned ELM (AFP-ELM). Basically, we assume that the SLFN predictive performance is associated not only with the hidden layer linearization capability, but also with the amount of input information it can propagate to the network output. Moreover, we consider that the structural retention of the input patterns in the hidden layer output patterns is associated with the proximity degree between their affinity matrices [5]. This degree depends on the hidden layer dimension. The greater the number of hidden nodes, the higher the structural retention degree. However, from a certain similarity level between those affinity matrices, adding new hidden nodes will have small or no effect on the amount of information propagated from the input. The proposed method automatically determines this level. The symmetry between the affinity matrices of the input data and the hidden layer output data is quantified by the *empirical alignment* [6]. This metric is also used to rank and select the hidden nodes. Since no output information is needed, hidden neurons are selected without any beforehand network parameter tuning and using only training data.

To check whether our *a priori* information based pruning provides good predictive performance, we test the equivalence between the AFP-ELM and the original ELM algorithm with the number of hidden neurons estimated via cross-validation (CV-ELM). Nine classification datasets are considered.

The remainder of this paper is organized as follows. Section 2 presents a brief review of the ELM learning algorithm. Next, section 3 presents the main concepts regarding affinity matrices and empirical alignment, and then the proposed learning algorithm. Section 4 presents the numerical experiments for classification problems and a statistical analysis of the results. Finally, discussion and conclusion are presented in section 5.

2 The Extreme Learning Machine

Given a set of N distinct observations $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{im}]^T \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$ for $i = 1, \ldots, N$, a SLFN can be used to model these observations as follows:

$$\hat{y_i} = \sum_{j=1}^{k} \beta_j g(\mathbf{v}_j \mathbf{x}_i + b_j), \quad i = 1, \ldots, N \tag{1}$$

where k is the number of hidden layer neurons, $g()$ is an activation function, $\mathbf{v}_j = [v_{j1}, v_{j2}, \ldots, v_{jm}]^T$ is the weight vector connecting the inputs to the j^{th} neuron, β_j is the weight connecting the j^{th} neuron to the output and b_j is the j^{th} neuron bias term, for $j = 1, \ldots, k$.

In order for a SLFN composed by k neurons to be able to approximate N observations with a null error there must exist $\mathbf{v}_j$, β_j and b_j, for $j = 1, \cdots, k$ such that:

$$H\beta = Y \tag{2}$$

where $H_{N \times k}$ is the hidden layer output.

ELM [1] is a SLFN learning algorithm where $\mathbf{v}_j$ and b_j for $j = 1, \ldots, k$ are randomly assigned and the output weights are computed using the Moore-Penrose pseudo-inverse:

$$\hat{\beta} = (H^T H)^{-1} H^T Y = H^+ Y \tag{3}$$

3 Node Selection by Alignment Ratio

3.1 Affinity Matrices and Empirical Alignment

Given a m-dimensional dataset $X = \{\mathbf{x}_i\}_{i=1}^N$, where N is the number of patterns, the elements s_{ij} of the Affinity Matrix $S = [s_{ij}]$ contain a measurement or estimation of the affinity of the pair of patterns $(\mathbf{x}_i, \mathbf{x}_j)$, where affinity is defined as a likeness based on relationship or causal connection [5]. Alternatively, a Dissimilarity Matrix $D = [d_{ij}]$ contain a measurement or estimation of the distance of a given pair of patterns.

In order to quantify the degree of input structural retention of the SLFN hidden layer, the *empirical alignment* described in [6] was adopted. Thus the alignment quantity $A(I, P)$ between input and hidden layer projection can be expressed in the form

$$A(I,P) = \frac{\langle I, P \rangle_F}{\sqrt{\langle I, I \rangle_F \langle P, P \rangle_F}} \tag{4}$$

where I and P are affinity matrices of the input and the hidden layer projection, respectively, and $\langle ., . \rangle_F$ is the Frobenius inner product [6].

3.2 Affinity Based Pruned ELM

The node selection approach proposed here creates a ranking of the hidden neurons according to their ability to retain input information. The importance of each neuron in terms of input structural retention is quantified by the so-called *empirical alignment ratio*, a metric proposed in this paper. Considering the i^{th} neuron the *empirical alignment ratio* is defined as follows:

$$r_A^i = \frac{A(I, P_i)}{A(I, P)} \tag{5}$$

where P_i is the Euclidean dissimilarity matrix of the hidden layer projection without the i^{th} neuron. The *empirical alignment ratio* quantifies the negative effect on structural retention when the i^{th} neuron is suppressed. Thus the lower the ratio, the greater is the input information loss brought by leaving out the neuron. The usefulness of a hidden node is therefore inversely proportional to its *empirical alignment ratio.*

Starting from a high dimensional hidden layer, only a few neurons (with the lowest *empirical alignment ratio*) are sufficient to propagate enough input information to estimate the SLFN output. Based on this idea the ranked neurons are inserted one-by-one in the hidden layer in the ascending order of alignment ratio. At each node insertion the *empirical alignment* between SLFN input and hidden layer output is computed. The curve relating the hidden layer dimension with the empirical alignment registered is named here *empirical alignment curve.* For a certain number of nodes this curve saturates, indicating that the remaining neurons have small or no contribution to the structural retention of the input patterns.

Therefore, the compromise point between the SLFN complexity and the hidden layer structural retention degree may be the *empirical alignment curve knee.* As defined in [7], the curve knee is the point of maximal curvature. The method proposed in [8] was adopted to find the knee point of the *empirical alignment curve.* Such method consists in determining the point with the maximum distance d_{max} to the line segment defined by the first and last alignment curve points:

$$d_{max} = max \left[\sqrt{(C - P_x)^2 + (aC + c - P_y)^2} \right] \tag{6}$$

where (P_x, P_y) are the coordinates of any point on the curve, $y = ax + c$ is the equation of the straight line that passes through the first and last curve points, and C is a constant defined as follows:

$$\frac{P_x + aP_y - ac}{a^2 + 1} \tag{7}$$

The proposed method provides a network structure definition based solely on the inputs. No previous parameter adjustment is needed. Furthermore, the method guides to the choice of the most parsimonious model in terms of input data structural retention using only training data.

4 Experimental Results

In this section, the proposed modified ELM learning algorithm is evaluated using supervised classification problems. Nine binary classification datasets collected from the UCI Repository [9] were considered: *Liver Disorders* (LIV), *Breast Cancer Wisconsin* (CAN), *Australian Credit* (AUS), *German Credit* (GER), *Pima Indians Diabetes* (DIA), *Statlog (Heart)* (HRT), *Ionosphere* (ION), *Parkinsons* (PKS) and *Sonar* (SON). All of them have been pre-processed in the same way. Firstly, we took 30 different random permutations without replacement for each dataset. Then, for each permutation, we took 70% of the instances for training and 30% for testing. The variables were normalized to zero-mean and unit standard deviation.

In all experiments we set the initial number of hidden nodes $k \in \{300, 500\}$. The hidden layer parameters were sampled from an uniform distribution $U[-0.1, 0.1]$. Sigmoidal activation functions have been used for all neurons. We compared AFP-ELM with a *Cross-Validation* Based ELM (CV-ELM). In the CV-ELM formulation, the number of hidden nodes is defined using hold-out cross-validation. Additionally, we considered the state-of-the-art machine learning algorithms Vanilla Linear SVM (VAN-SVM), Radial Basis Function SVM (RBF-SVM) and the Real AdaBoost (RE-ADA). For all problems the SVM hyper-parameters were selected using 10-fold cross-validation.

Table 1 shows the average test accuracy rate (standard deviations in brackets) over 30 repetitions.

4.1 Statistical Analysis

Based on the classification results, we are interested in testing whether considering just the input structural retention for adjusting the hidden layer dimension leads to a poor predictive performance. By attesting the statistical equivalence between AFP-ELM and CV-ELM in terms of accuracy, there will be a strong evidence to conclude that the proposed pruning methodology does not lead to

Table 1: Classification Accuracy

	VAN-SVM	RBF-SVM	RE-ADA	CV-ELM	AFP-ELM
LIV	68.33(3.68)	68.59(3.77)	71.63(3.43)	67.21(4.11)	67.66(4.11)
CAN	97.29(1.04)	97.54(1.44)	95.89(1.24)	95.46(1.93)	95.38(1.48)
AUS	83.79(2.27)	84.15(2.00)	85.52(2.05)	84.15(2.25)	84.78(2.29)
GER	76.93(2.10)	75.94(1.92)	75.81(1.59)	75.33(2.39)	75.83(1.94)
DIA	76.91(2.04)	76.14(2.18)	76.19(2.39)	76.48(2.20)	76.74(2.10)
HRT	82.51(3.40)	83.13(4.28)	81.85(3.76)	82.47(4.52)	82.80(3.58)
ION	87.80(2.36)	94.84(2.07)	93.49(2.48)	84.69(3.01)	85.85(3.30)
PKS	86.21(4.08)	87.64(3.17)	88.33(3.47)	85.98(3.91)	85.80(3.89)
SON	74.89(5.11)	81.34(4.84)	81.08(5.29)	71.83(6.08)	73.98(5.53)

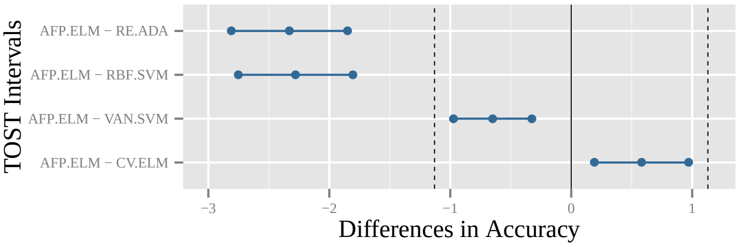

Fig. 1: TOST 90% confidence intervals.

loss of performance. To guide this conclusion we adopted an equivalence test procedure, called *two one-sided test* (TOST) [10].

The design of an equivalence test requires the definition of an acceptance criterion δ. It is the limit outside which the difference in mean values should be considered statistically significant. The TOST consists of rejecting the null hypothesis of dissimilarity, leading to conclude equivalence of performance, if and only if a $(1 - 2\alpha)100\%$ equal-tailed confidence interval of the differences is completely contained in the interval $[-\delta, +\delta]$.

Figure 1 shows the 90% confidence intervals of the differences in performance between AFP-ELM and the other algorithms. The equivalence interval (shown in dotted lines) was defined using a closed-form expression proposed by Limentani et al. [11]. Choosing a significance level $\alpha = 0.05$ and a power $1 - \beta = 0.80$, we found $\delta = 1.13$. Since the 90% confidence intervals for the differences between AFP-ELM and both CV-ELM and VAN-SVM are completely contained within $[-\delta, +\delta]$, there is enough evidence, for a significance level $\alpha = 0.05$, to reject the null hypothesis of dissimilarity between those algorithms. On the other hand, there is enough evidence to conclude that RBF-SVM and RE-ADA perform better than AFP-ELM because the respective TOST confidence intervals besides being strictly negative are completely outside the equivalence margin.

5 Conclusions and Future Work

The statistical equivalence between AFP-ELM and CV-ELM assures the efficacy of the network structure selection approach. The proposed method, unlike those based on *cross-validation*, makes no use of output information to select a network structure which improves the SLFN predictive performance. Nevertheless, as confirmed by the results, by ensuring the propagation of enough input information the AFP-ELM achieves classification performance as high as CV-ELM does.

Therefore, we can conclude that the proposed approach is a promising alternative to perform the ELM pruning. Moreover, no validation set and network parameter fitting are required to specify the hidden layer dimension.

Future work shall address generalization to multi-output problems and comparison with other ELM pruning methodologies.

Acknowledgments.

This work was supported by the Brazilian agencies CNPq and CAPES.

References

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

[2] Hai-Jun Rong, Yew-Soon Ong, Ah-Hwee Tan, and Zexuan Zhu. A fast pruned-extreme learning machine for classification problem. *Neurocomputing*, 72:359–366, 2008.

[3] Yoan Miche, Antti Sorjamaa, Patrick Bas, Olli Simula, Christian Jutten, and Amaury Lendasse. Op-elm: optimally pruned extreme learning machine. *Trans. Neur. Netw.*, 21(1):158–162, jan 2010.

[4] Yoan Miche, Mark van Heeswijk, Patrick Bas, Olli Simula, and Amaury Lendasse. Trop-elm: A double-regularized ELM using LARS and tikhonov regularization. *Neurocomputing*, 74(16):2413–2421, 2011.

[5] Y. Weiss. Segmentation using eigenvectors: a unifying view. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 975–982 vol.2, 1999.

[6] Nello Cristianini, John Shawe-Taylor, and Jaz Kandola. On kernel target alignment. In *Proceedings of the Neural Information Processing Systems, NIPS'01*, pages 367–373. MIT Press, 2002.

[7] S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, pages 576–584, 2004.

[8] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *Proceedings of the 2011 31st International Conference on Distributed Computing Systems Workshops*, ICDCSW '11, pages 166–171, Washington, DC, USA, 2011. IEEE Computer Society.

[9] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[10] D. J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, 15(6):657–680, dec 1987.

[11] Giselle B. Limentani, Moira C. Ringo, Feng Ye, Mandy L. Berquist, and Ellen O. McSorley. Beyond the t-test: statistical equivalence testing. *Analytical Chemistry*, jun 2005.