An Enhanced Hierarchical Extreme Learning Machine With Random Sparse Matrix Based Autoencoder

Tianlei Wang[1], Xiaoping Lai[1], Jiuwen Cao[1], Chi-Man Vong[2] and Badong Chen[3]

[1]Institute of Information and Control, Hangzhou Dianzi University, China [2]Faculty of Science and Technology, University of Macau, Macau, China [3]School of Electronic and Information Engineering, Xi'an Jiaotong University, China

Introduction

Recently, by employing the stacked extreme learning machine (ELM) based autoencoders (ELM-AE) and sparse AEs (SAE), multilayer ELM (ML-ELM) and hierarchical ELM (H-ELM) has been developed. Compared to the conventional stacked AEs, the ML-ELM and H-ELM usually achieve better generalization performance with a significantly reduced training time.

However, ML-ELM and H-ELM suffer the following deficiencies:

- The extracted features in ML-ELM tend to be dense and may lead to indistinctive representation.
- The simply stacked AEs in ML-ELM may not well exploit the advantage of ELM.
- The SAE fails to provide analytical solution leading to long training time for big data.
- The ℓ_1-norm based SAE may suffer the overfitting problem.

To address these deficiencies, we propose an enhanced H-ELM (EH-ELM) with a novel random sparse matrix based AE (SMA) in this paper. The contributions are summarized as follows:

- Utilizing the random sparse matrix, the sparse features can be obtained.
- Benefiting from using random sparse matrix, the ℓ_2-norm regularized optimization is formulated in the SMA. The resultant solution can be analytically calculated.
- By virtue of the SMA, the proposed EH-ELM learns faster than ML-ELM and H-ELM.

Proposed SMA

A random matrix projection has been developed based on the Johnson-Lindenstrauss (JL) lemma, which states that after projection, the distance of any pair of two vectors can be preserved within an arbitrarily small tolerance. Based on that, we propose two new random sparse matrices for generation of the hidden-layer parameters in ELM as follows:

$$\text{Scheme 1 } w_{ij} = \begin{cases} 0 & \textit{with probability } \frac{2}{3}, \\ \mathrm{U}(-1,1) & \textit{with probability } \frac{1}{3}, \end{cases} \tag{1}$$

$$\text{Scheme 2 } w_{ij} = \begin{cases} 0 & \textit{with probability } \frac{2}{3}, \\ \mathrm{N}(-1,1) & \textit{with probability } \frac{1}{3}, \end{cases} \tag{2}$$

where $\mathrm{U}(\cdot)$and $\mathrm{N}(\cdot)$are the Uniform and Gaussian distributions, respectively. By virtue of the above described sparse random weight matrix, we proposed a random sparse matrix based AE (SMA). The SMA generates the hidden-layer parameters $\mathbf{W}_M$ and $\mathbf{b}_M$ according to (1) and (2) and solves the output-layer weight $\boldsymbol{\beta}_M$ by the following ℓ_2-regularized nonlinear ELM-AE

$$\min_{\boldsymbol{\beta}_M} \frac{1}{2} C \|\mathbf{H}_M \boldsymbol{\beta}_M - \mathbf{X}^T\|_2^2 + \frac{1}{2}\|\boldsymbol{\beta}_M\|_2^2, \tag{3}$$

$$\mathbf{H}_M = \mathrm{g}(\mathbf{W}_M \mathbf{X} + \mathbf{b}_M \mathbf{1}^T)^T, \tag{4}$$

where $\mathbf{1}$ is an all-one vector of dimension N and $\mathrm{g}(\cdot)$ is the activation function. The solution to problem (3) can be obtained as

$$\begin{aligned} \boldsymbol{\beta}_M &= \mathbf{H}_M^T(\tfrac{\mathbf{I}}{C} + \mathbf{H}_M\mathbf{H}_M^T)^{-1}\mathbf{X}^T \quad \textit{if} \quad N < L, \\ \boldsymbol{\beta}_M &= (\tfrac{\mathbf{I}}{C} + \mathbf{H}_M^T\mathbf{H}_M)^{-1}\mathbf{H}_M^T\mathbf{X}^T \quad \textit{if} \quad N \geq L. \end{aligned} \tag{5}$$

Here, N is the number of samples and L is the number of hidden nodes. Then, the encoded result can be derived as

$$\mathbf{Y} = \mathrm{g}(\boldsymbol{\beta}_M \mathbf{X}). \tag{6}$$

Fig. 1 shows the architecture of SMA.

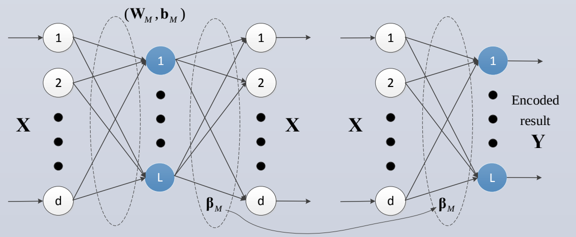

Fig. 1 The architecture of SMA

Proposed EH-ELM

By incorporating the H-ELM learning framework with the SMA described, an EH-ELM is developed. Fig. 2 shows the architecture of EH-ELM which consists of a feature extraction with stacked SMAs and a classification layers with ELM. Assume K SMA layers are used and $\mathbf{Y}^{(k-1)}$is the output of $(k-1)$-th layer with $\mathbf{Y}^{(0)} = \mathbf{X}$, the output $\mathbf{Y}^{(k)}$of the k-th layer is

$$\mathbf{Y}^{(k)} = \mathrm{g}\left(\boldsymbol{\beta}_M^{(k)}\mathbf{Y}^{(k-1)}\right), k = 1, \cdots, K, \tag{7}$$

where $\boldsymbol{\beta}_M^{(k)}$is the output weight of the k-th SMA. The supervised ELM classifier in the last layer is trained as

$$\min_{\boldsymbol{\beta}} \frac{1}{2} C \left\|\mathrm{g}\left(\mathbf{W}\mathbf{Y}^{(K)} + \mathbf{b} \cdot \mathbf{1}^T\right)^T \boldsymbol{\beta} - \mathbf{T}\right\|_2^2 + \frac{1}{2}\|\boldsymbol{\beta}\|_2^2, \tag{8}$$

where $\mathbf{W}$ and $\mathbf{b}$ are the orthogonal random input weights and bias, $\mathbf{T}$ is the desired output matrix of training data. The output weight $\boldsymbol{\beta}$ in the last hidden layer is computed by

$$\begin{aligned} \boldsymbol{\beta} &= \mathbf{H}^T(\tfrac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^T)^{-1}\mathbf{T} \quad \textit{if} \quad N < L, \\ \boldsymbol{\beta} &= (\tfrac{\mathbf{I}}{C} + \mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{T} \, \textit{if} \quad N \geq L, \end{aligned} \tag{9}$$

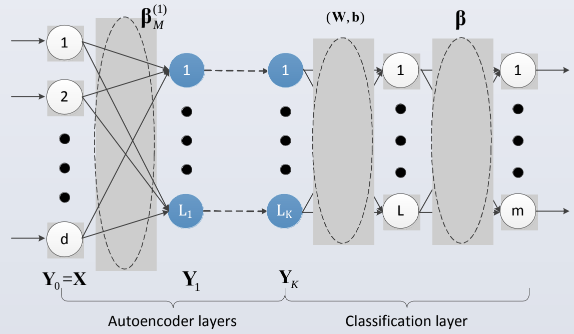

Fig. 2 The architecture of EH-ELM

Experiments

The first experiment is conducted on the real-world NORB dataset to compare the sparsity of the proposed SMA and the existing ℓ_1-norm based SAE. Both the Uniform distribution and the Gaussian distribution are tested to generate the random sparse matrix. The corresponding SMAs are denoted as SMA^U and SMA^G , respectively. The criterion $m_s = (\sqrt{card(\boldsymbol{\beta})} - \|\boldsymbol{\beta}\|_1/\|\boldsymbol{\beta}\|_2)/(\sqrt{card(\boldsymbol{\beta})} - 1)$ ($card(\boldsymbol{\beta})$ is the number of elements in $\boldsymbol{\beta}$) is employed for the sparsity evaluation of the output weight. Different number of hidden nodes of the AE ranging from 100 to 3000 are tested. The curves of sparsity in Fig. 3 show that, the proposed SMA with both random sparse matrix generating methods is effective in sparse encoding.

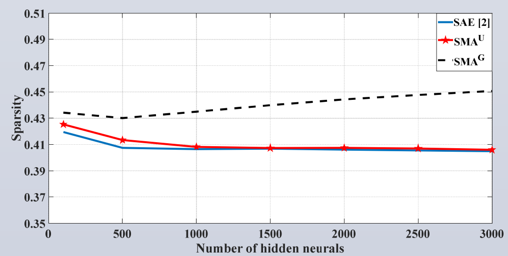

Fig. 3 Sparsity comparison among different sparse AEs.

In the following experiments, a comparison among ML-ELM, H-ELM and the proposed EH-ELM is made. Both the random sparse matrices generated by the Uniform and Gaussian distributions (denoted as EH-ELM^U and EH-ELM^G) are tested. Experiments are conducted on 12 high-dimensional and 11 low-dimensional benchmark classification datasets, as well as the MNIST and NORB datasets.

Tables 2 and 3 show the recognition rates and training time of ML-ELM, H-ELM and the proposed EH-ELM. As highlighted in boldface, EH-ELM obtains higher recognition rate with lower training time than ML-ELM and H-ELM.

Experiments on the datasets, MNIST and NORB, are also carried out to verify the superiority of EH-ELM. Table 4 shows the experimental results. It is obvious that EH-ELM wins the best recognition rate, and also learns faster than ML-ELM and H-ELM.

Table 1. Recognition rates (%) and training time (s) comparisons on high-dimensional datasets

Dataset	ML-ELM [1]		H-ELM [2]		EH-ELM^U		EH-ELM^G
	Rate	Train time	Rate	Train time	Rate	Train time	Rate
ALLAML	89.66	21.33	83.45	16.45	**96.55**	**10.87**	96.55
arcene	82.5	21.46	80.5	18.80	**85.5**	**11.32**	85.5
Carcinom	91.3	21.40	94.78	18.18	**95.36**	**11.09**	95.07
COIL20	99.59	22.83	97.38	12.33	**99.66**	**11.29**	98.97
gisette	96.27	32.93	95.72	21.11	96.66	**19.49**	**96.89**
GLIOMA	62.86	21	62.86	14.27	**71.43**	**10.43**	66.67
HistALL	88.1	30.74	89.45	19.61	93.18	**17.41**	**93.25**
ORL64	89.88	21.10	96.88	14.09	**97.5**	**10.69**	96.75
PCMAC	74.13	23.79	83.55	14.79	**84.94**	**12.59**	84.07
TOX171	74.2	20.84	81.16	15.23	**82.61**	**10.68**	82.61
Yale64	78.33	20.70	87.33	13.98	**88.33**	**10.49**	88.33
YaleB32	93.66	24.03	84.11	12.99	98.04	**12.26**	**98.06**

Table 2. Recognition rates (%) and training time (s) comparisons on low-dimensional datasets

Dataset	ML-ELM [1]		H-ELM [2]		EH-ELM^U		EH-ELM^G
	Rate	Train time	Rate	Train time	Rate	Train time	Rate
BreastTissue	61.29	20.22	51.61	10.59	**80.65**	**9.96**	74.19
bupa	66.90	0.005	65.38	0.003	71.59	**0.003**	**72.28**
mfeatall	98.65	23.62	98.07	12.47	**99.15**	**11.72**	99.1
Cardiotocography	84.33	24.75	88.85	12.60	**91.34**	**12.26**	90.31
diabetes2	64.06	0.05	69.48	0.02	**70.31**	**0.02**	67.4
randomfaces4ar	94.88	25.81	94.8	13.58	**98.92**	**13.04**	98.8
Diabetic	68.75	0.12	66.72	**0.06**	72.41	0.07	**73.16**
randomAR	78.35	25.33	84.45	13.31	89.5	**12.72**	**89.93**
magic	85.85	0.33	84.28	**0.19**	85.98	0.20	**86.17**
pcaAR	86.29	21.85	91.43	11.42	**92.49**	**10.92**	92.23
wine	97.69	20.31	97.44	10.53	**100**	**10.04**	98.72

Table 3. Comparisons on MNIST and NORB datasets

Dataset	ML-ELM [1]		H-ELM [2]		EH-ELM^U		EH-ELM^G
	Rate	Train time	Rate	Train time	Rate	Train time	Rate
MNIST	99.00	281.71	98.99	101.39	**99.01**	78.23	99.00
NORB	88.27	251.59	90.65	165.72	**91.80**	150.34	91.77

Conclusions

- Instead of using the ℓ_1-norm optimization based sparse AE, a novel random sparse matrix based AE (SMA) has been proposed in this paper.
- The proposed SMA is able to provide analytical solutions for the sparse feature encoding.
- An enhanced hierarchical extreme learning machine algorithm (EH-ELM) has been developed by stacking the SMAs.
- Experimental results have been presented to verify the superiorities of the proposed EH-ELM.