Neurocomputing 313 (2018) 386–391

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Sensitivity of data matrix rank in non-iterative training

Zhiqi Huang[a,b], Xizhao Wang[a,*]

[a] Computer Science and Software Engineering, Shenzhen University, China
[b] Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, China

ARTICLE INFO

Article history:
Received 5 April 2018
Revised 8 June 2018
Accepted 24 June 2018
Available online 30 June 2018

Communicated by Dr. Nianyin Zeng

Keywords:
Neural network
Extreme learning machines
Rank of matrix
Generalizedinverse

ABSTRACT

This paper focuses on the parameter pattern during the initialization of Extreme Learning Machines (ELMs). According to the algorithm, model performance is highly dependent on the matrix rank of its hidden layer. Previous research has already proved that the sigmoid activation function can transform input data to a full rank hidden matrix with probability 1, which secures the stability of ELM solution. In recent study, we notice that, under full-rank condition, the hidden matrix possibly has very small eigenvalue, which seriously affects the model generalization ability. Our study indicates such a negative impact is caused by the discontinuity of generalized inverse at the boundary of full and waning rank. Experiments show that each phase of ELM modeling possibly leads to this rank deficient phenomenon, which harms the test accuracy.

1. Introduction

Introduced by Huang et al. [1,2], the Extreme learning machines (ELMs) as a type of single hidden layer feed-forward neural network (SLFNs) with non-iterative algorithm, the training process contains two parts: first, the weights and bias between input and hidden layers are randomly assigned; second, the weights between hidden and output layers are obtained by solving a system of linear equations using generalized inverse.

In the recent decade, ELM has been studied by many researches: deep learning techniques have been used to improve the ELM performance [3]. Incorporating with other algorithms, hybrid ELMs were proposed by Wang et al. [4,5]. And ELM has been used to solve different problems in multiple areas [6], such as imbalance problem [7], image processing[8] and time series forecasting [9,10]. Also, [11] demonstrated its big data performance. Comparing with the typical back-propagation (BP) algorithm for training feed-forward neural networks, the ELM's non-iterative training mechanism gives it speed and efficiency in most of the cases [12]. Different from BP algorithm where the hidden layer keep tuning in iteration, the hidden matrix of ELM is decided once by the weights between input and hidden layers. And the tuning phase of ELM is to solve a system of linear equations, so the structure and values of hidden matrix play a critical role in model performance. For example, [13] already proved that the sigmoid transformation lead to a full-rank hidden matrix with probability 1. And the stability of solution depends on whether the hidden matrix has full column rank. By looking deep into this full rank transformation, We find that with wide initial range, increasing number of hidden node, particular pre-training method or special pattern of training data, the hidden layer matrix could be weakly linear correlated. That means, the matrix is still full-rank but can be viewed as a perturbation from rank deficient matrix. And due to the discontinuity of generalized inverse, the coefficients between hidden and output layers will have large absolute value and variance which leads to robustness problem of ELM solutions [14].

In this paper, we first point out that the training of ELM is sensitive to the rank of hidden layer matrix, and give a detailed proof on discontinuity of generalized inverse under waning rank matrix. Then based on theoretical analysis, we are going to investigate the following questions: how and why initial range, number of hidden nodes, outliers in training data and unsupervised pre-training affect the model performance respectively.

The rest of this paper is organized as follows. Section 2 gives a brief review on the related works. Section 3 investigates the relationship between rank of matrix and its generalized inverse. Based on the theoretical result, some examples and experiments on different initial methods and network structures are shown in Section 4. And in Section 5, we conclude this paper.

2. Extreme learning machine

ELM means a three layer feed-forward networks with single hidden layer in which the weights and bias between input layer

* Corresponding author.
E-mail addresses: huangzhiqi@szu.edu.cn (Z. Huang), xizhaowang@ieee.org (X. Wang).

https://doi.org/10.1016/j.neucom.2018.06.055
0925-2312/

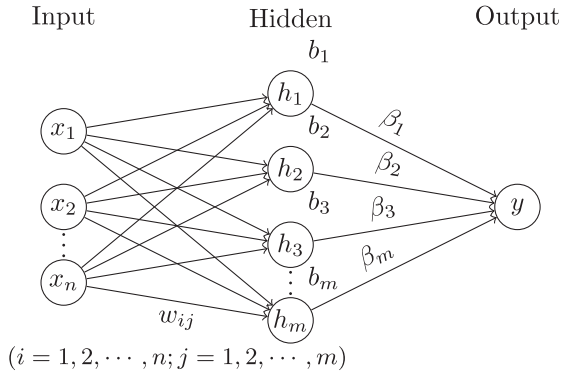


Fig. 1. A simple ELM structure.

and hidden layer are randomly assigned and the weights between hidden layer and output layer are solved by a system of linear equations. A simple structure of ELM for regression problem is shown in Fig. 1 with n nodes in input layer, m nodes in hidden layer and only one node in output layer, while the classification problem, number of output node equals to the number of categories.

Given a set of samples $\mathbf{S} = \{(\mathbf{x}_i, \mathbf{t}_i) | \mathbf{x}_i \in \mathbf{R}^d, \mathbf{t}_i \in \mathbf{R}^t\}_{i=1}^n$, training process of ELM is to determine model parameters $\{w_{ij}, b_j, \beta_j\}$. Since the weights w_{ij} and bias b_j are randomly selected, the training process is only about determining the connections β_j between hidden layer and output layer. Let

$$\mathbf{G}_{n\times m} = \begin{bmatrix} \mathbf{w_1 x_1} + b_1 & \cdots & \mathbf{w_m x_1} + b_m \\ \mathbf{w_1 x_2} + b_1 & \cdots & \mathbf{w_m x_2} + b_m \\ \vdots & \ddots & \vdots \\ \mathbf{w_1 x_n} + b_1 & \cdots & \mathbf{w_m x_n} + b_m) \end{bmatrix} \tag{1}$$

be the middle matrix, where $\mathbf{w_j}$ is the *jth* column of the weight matrix $\mathbf{W}$ between input layer and output layer. Let $g(\cdot)$ be the sigmoid function and $\mathbf{H}$ be hidden layer matrix, then

$$\mathbf{H}_{n\times m} = (g(\mathbf{G}))_{n\times m} = (h_{ij})_{n\times m} \tag{2}$$

Suppose the target matrix is $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_n]^T$, then the training of ELM is transferred to solve the system of linear equations $\mathbf{H}\boldsymbol{\beta} = \mathbf{T}$. In general, the solution $\mathbf{H}^-$ is not unique. [2,12] suggested to use the minimum-norm least square solution. Instead of solving the system of linear equations, the optimization problem change to:

$$\min_{||\beta||} (\min_{\beta \in \mathbf{R}^m} ||\mathbf{T} - \mathbf{H}\boldsymbol{\beta}||^2) \tag{3}$$

the solution of (3) is the Moore–Penrose pseudo-inverse of matrix $\mathbf{H}$, represented as $\mathbf{H}^\dagger$.

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{T} \rightarrow \hat{\boldsymbol{\beta}} = \mathbf{H}^\dagger \mathbf{T} \tag{4}$$

The Moore–Penrose pseudo-inverse and solution has the following properties:

1. $m = n$, $\mathbf{H}^\dagger = \mathbf{H}^-$ if A is full rank. But most of cases in ELM, the number of hidden node is smaller than the number of observations.
2. $m > n$ (kinematically insufficient manipulator), This is the case there are more constraining equations than there are free variables. Hence, it is not generally possible find a solution to these equations. The pseudo-inverse gives solution such that $\mathbf{H}^\dagger\mathbf{T}$ is closest (in a least-squared sense) to the desired solution vector $\mathbf{T}$.
3. $m < n$ (kinematically redundant manipulator), then the Moore–Penrosesolution minimizes the norm of $\boldsymbol{\beta}$. In this case, there are generally an infinite number of solutions, and the Moore–Penrose solution is the particular solution whose 2-norm is minimal.

Now the training process of an ELM can be divided into three steps:

1. Dimension increases from input $\mathbf{S}$ to middle matrix $\mathbf{G}$. Generally, the number of hidden nodes m is greater than number of input attributes d;
2. The sigmoid function transfers middle matrix $\mathbf{G}$ to hidden layer matrix $\mathbf{H}$ with rank increased;
3. Solving a system of linear equations with full rank of coefficient matrix.

Furthermore, the activation function in step 2 not only increases the rank of middle matrix to hidden layer matrix, but also guarantee full column rank of hidden layer matrix with the following proposition.

Proposition 1. *Assume that* $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, $v_i = \{v_{i1}, v_{i2}, \ldots, v_{in}\}$, $i = 1, 2, \ldots, N$ *denotes a set of n-dimensional vectors, such that* $1 \le rank(\mathbf{V}) \le n$. *Then with probability 1, the sigmoid transformation will transfer V in to a set of vectors of full rank.*

$$rank(\mathbf{H}) = n \quad w.p.1 \tag{5}$$

where $\mathbf{H} = \{h_1, h_2, \ldots, h_N\}$, $h_i = \{h_{i1}, h_{i2}, \ldots, h_{in}\}$, $h_{ij} = \text{sigmoid}(v_{ij}) = 1/(1 + e^{v_{ij}})$, $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, n$.

Remark 1. The proof of Proposition 1 can be found in [13]. In step 2, the middle matrix $\mathbf{G}$ is coming from input data $\mathbf{S}$ via a linear transformation and is generally waning rank. Proposition 1 guarantees the sigmoid transformation will transfer a waning rank matrix $\mathbf{G}$ to a full rank matrix $\mathbf{H}$. In the next section, we investigate the relationship between full rank and generalized inverse.

3. Continuity of generalized inverse

In this section, we will first proof the generalized inverse is continuous if $\mathbf{H}$ is a full-rank matrix. Along with Proposition 1, these two properties guarantee the stability of ELM solution. Thus, the full-rank matrix $\mathbf{H}$ is insensitive to the perturbation and can get the more stable solution for $\mathbf{H}\boldsymbol{\beta} = \mathbf{T}$. Then, we discuss a special case which the perturbation increases the rank of matrix and discontinuity of generalized inverse under this circumstances. We use the notation $\delta\mathbf{A}$ to represent a perturbation of matrix $\mathbf{A}$.

Proposition 2. *The generalized inverse* $\mathbf{A}^\dagger$ *is continuous if* $\mathbf{A}$ *is a full-rank matrix.*

Proof. Assume $rank(\mathbf{A}) = n$, then $\mathbf{A}^T\mathbf{A}$ is a $n \times n$ non-singular matrix. In fact, it is a symmetric and positive matrix and $\mathbf{A}^\dagger = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$, then we have

$$(\mathbf{A} + \delta\mathbf{A})^T(\mathbf{A} + \delta\mathbf{A}) = \mathbf{A}^T\mathbf{A} + (\mathbf{A} + \delta\mathbf{A})^T\delta\mathbf{A} + (\delta\mathbf{A})^T\mathbf{A}$$

According to Banach theorem, we know that $(\mathbf{A} + \delta\mathbf{A})^T(\mathbf{A} + \delta\mathbf{A})$ is a non-singular matrix if $||(\mathbf{A}^T\mathbf{A})^{-1}[(\mathbf{A} + \delta\mathbf{A})^T\delta\mathbf{A} + (\delta\mathbf{A})^T\mathbf{A}]|| < 1$. This inequality will holds if we take the $||\delta\mathbf{A}||$ small enough. So there exists a small positive η such that the inequality holds if $||\delta\mathbf{A}|| \le \eta$. Now, the generalized inverse matrix is

$$(\mathbf{A} + \delta\mathbf{A})^\dagger = [(\mathbf{A} + \delta\mathbf{A})^T(\mathbf{A} + \delta\mathbf{A})]^{-1}(\mathbf{A} + \delta\mathbf{A})^T$$

Let $||\delta\mathbf{A}|| \to 0$, we have

$$\lim_{||\delta\mathbf{A}||\to 0} [(\mathbf{A} + \delta\mathbf{A})^T(\mathbf{A} + \delta\mathbf{A})]^{-1} = (\mathbf{A}^T\mathbf{A})^{-1}$$

$$\text{and} \quad \lim_{||\delta\mathbf{A}||\to 0} (\mathbf{A} + \delta\mathbf{A})^T = \mathbf{A}^T$$

which implies $\lim_{||\delta \mathbf{A}||\to 0}(\mathbf{A}+\delta\mathbf{A})^{\dagger} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = A^{\dagger}$, the proposition is proved. □

Example 1. Let $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$, then $rank(\mathbf{A}) = 1$, which is not full-rank. It is easy to calculate that $\mathbf{A}^{\dagger} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$. Suppose that $\delta\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & \epsilon \\ 0 & 0 \end{bmatrix}$, $\epsilon \neq 0$. then $\mathbf{A}+\delta\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & \epsilon \\ 0 & 0 \end{bmatrix}$. Noting that the rank is increase from 1 to 2 and $\mathbf{A}+\delta\mathbf{A}$ is full-rank. we get $(\mathbf{A}+\delta\mathbf{A})^{\dagger} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \epsilon^{-1} & 0 \end{bmatrix}$. It is easy to see that limit of $(\mathbf{A}+\delta\mathbf{A})^{\dagger}$ does not exists when $\epsilon \to 0$. So the generalized inverse $\mathbf{A}^{\dagger}$ is discontinuous if $\mathbf{A}$ is waning rank. Next, we will give a theoretical proof about this property.

Proposition 3. *Suppose the singular values of* $\mathbf{A}^{m\times n}$ *are* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0$, *then*

$$||\mathbf{A}|| = \lambda_1 \quad \text{and} \quad ||\mathbf{A}^{\dagger}|| = \lambda_k^{-1} \tag{6}$$

Proof. The definition of norm

$$||\mathbf{A}|| = \max_{||\mathbf{x}||=1} ||\mathbf{A}\mathbf{x}||, \quad \mathbf{x} = \mathbf{R}^n$$

According to the definition of Euclidean norm

$$||\mathbf{A}\mathbf{x}||^2 = \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}$$

The eigenvalue of $\mathbf{A}^T\mathbf{A}$ are $\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_k^2$ and eigenvector $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$, so

$$\begin{aligned} \max_{||\mathbf{x}||=1} ||A\mathbf{x}||^2 &= \max_{||\mathbf{x}||=1} (\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}) \\ &= \max_{||\mathbf{x}||=1} (\mathbf{x}^T \sum_{1}^{k} \lambda_i \mathbf{v}_i\mathbf{v}_i^T)\mathbf{x} \\ &= \max_{||\mathbf{x}||=1} \sum_{1}^{k} \lambda_i^2 (\mathbf{x}^T\mathbf{v}_i^T)^2 \end{aligned}$$

with $\sum_1^k (\mathbf{x}^T\mathbf{v}_i)^2 \leq 1$, then $\max_{||\mathbf{x}||=1} ||\mathbf{A}\mathbf{x}||^2 \leq \lambda_1^2$. If let $\mathbf{x} = \mathbf{v}_1$, then

$$\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \lambda_1^2 \leftrightarrow \max_{||\mathbf{x}||=1} ||\mathbf{A}\mathbf{x}||^2 = \lambda_1^2 \leftrightarrow ||\mathbf{A}|| = \lambda_1$$

Now consider the $||\mathbf{A}^{\dagger}||$. Assume $\mathbf{A}$ has singular value decomposition (SVD) $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ then $\mathbf{A}^{\dagger} = \mathbf{V}\mathbf{\Sigma}^{\dagger}\mathbf{U}^T$, where

$$\mathbf{\Sigma} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_k \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma}^{-1} = \begin{bmatrix} \lambda_1^{-1} & & \\ & \ddots & \\ & & \lambda_k^{-1} \end{bmatrix}$$

$$\begin{aligned} ||\mathbf{A}^{\dagger}||^2 &= \max_{||\mathbf{x}||=1} ||\mathbf{A}^{\dagger}\mathbf{x}||^2 \\ &= \max_{||\mathbf{x}||=1} \{(\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{x})^T(\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{x})\} \\ &= \max_{||\mathbf{y}||=1} \mathbf{y}^T\mathbf{\Sigma}^{-2}\mathbf{y} \end{aligned}$$

Same as the norm of $\mathbf{A}$, the norm of $\mathbf{A}^{\dagger}$ is the square root of the largest eigenvalue of $\mathbf{\Sigma}^{-2}$ which is λ_k^{-1}. Now, suppose a small perturbation $\delta\mathbf{A}$ and $\mathbf{B} = \mathbf{A}+\delta\mathbf{A}$. Regarding to the singular values of $\mathbf{A}$ and $\mathbf{B}$, we have the following □

Proposition 4. *Suppose* $rank(\mathbf{A}) = rank(\mathbf{B}) = k$ *and the singular values of* $\mathbf{A}$ *are* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$, *because* $\mathbf{B}$ *has the same rank with* $\mathbf{A}$, $\mathbf{B}$ *has singular values* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$. *Then*

$$\sigma_i \leq \lambda_i + ||\delta\mathbf{A}|| \tag{7}$$

Proof. According to the singular value decomposition (SVD), $\mathbf{A}^T\mathbf{A}$ has the eigenvalue $\lambda_1^2, \lambda_2^2, \ldots, \lambda_k^2$ and eigenvector $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$, then apply the Courant-Fischer minimax theory [15,16], we have

$$\begin{aligned} \sigma_{r+1}^2 &\leq \max_{\substack{||\mathbf{x}||=1 \\ \mathbf{x}^T\mathbf{p_i}=0}} \mathbf{x}^T\mathbf{B}^T\mathbf{B}\mathbf{x} \\ &= \max_{\substack{||\mathbf{x}||=1 \\ \mathbf{x}^T\mathbf{p_i}=0}} \mathbf{x}^T(\mathbf{A}+\delta\mathbf{A})^T(\mathbf{A}+\delta\mathbf{A})\mathbf{x} \\ &\leq \max_{\substack{||\mathbf{x}||=1 \\ \mathbf{x}^T\mathbf{p_i}=0}} \{(\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x})^{\frac{1}{2}} + (\mathbf{x}^T(\delta\mathbf{A})^T(\delta\mathbf{A})\mathbf{x})^{\frac{1}{2}}\}^2 \\ &\leq \{\max_{\substack{||\mathbf{x}||=1 \\ \mathbf{x}^T\mathbf{p_i}=0}} (\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x})^{\frac{1}{2}} + \max_{\substack{||\mathbf{x}||=1 \\ \mathbf{x}^T\mathbf{p_i}=0}} (\mathbf{x}^T(\delta\mathbf{A})^T(\delta\mathbf{A})\mathbf{x})^{\frac{1}{2}}\}^2 \\ &\leq (\lambda_{r+1} + ||\delta\mathbf{A}||)^2, \quad r = 1, 2 \ldots, k-1. \end{aligned}$$

Thus

$$\sigma_{r+1} \leq \lambda_{r+1} + ||\delta\mathbf{A}|| \leftrightarrow \sigma_r \leq \lambda_r + ||\delta\mathbf{A}||$$

Also called the singular perturbation theory, Proposition 4 establishes a relationship between original matrix and its perturbation. And gives a perturbation bounds to singular values. According to Propositions 3 and4, we can conclude the discontinuity of generalized inverse in waning rank matrix. □

Proposition 5. *If the* $m\times n$ $(m<n)$ *matrix* $\mathbf{A}$ *is waning rank,* $rank(\mathbf{A}) = k < n$, *the small perturbation* $\delta\mathbf{A}$ *increases the rank of* $\mathbf{B} = \mathbf{A}+\delta\mathbf{A}$.

$$rank(\mathbf{A}+\delta\mathbf{A}) > rank(\mathbf{A}) > k \tag{8}$$

Then we have the inequation:

$$||(\mathbf{A}+\delta\mathbf{A})^{\dagger}|| \geq \frac{1}{||\delta\mathbf{A}||} \tag{9}$$

Proof. Assume $rank(\mathbf{A}+\delta\mathbf{A}) = r > k$, then the *rth* singular value of matrix $\mathbf{A}$ is $\lambda_r = 0$. According to Proposition 4, the *rth* singular value of $\mathbf{A}+\delta\mathbf{A}$, σ_r has

$$\sigma_r \leq ||\delta\mathbf{A}||$$

Meanwhile, apply Proposition 3, the norm of $(\mathbf{A}+\delta\mathbf{A})^{\dagger}$ has

$$||(\mathbf{A}+\delta\mathbf{A})^{\dagger}|| \leq \frac{1}{\sigma_r}$$

Therefore

$$||(\mathbf{A}+\delta\mathbf{A})^{\dagger}|| \geq \frac{1}{||\delta\mathbf{A}||}$$

□

Remark 4. In fact, this conclusion is related to the continuity of singular value. As we can see, for diagonal matrix $\mathbf{\Sigma}$, the generalized inverse is calculated by taking the reciprocal of each non-zero element on the diagonal, leaving the zeros in place, and then transposing the matrix. The discontinuity is coming from taking the reciprocal of matrix elements.

The continuity of generalized inverse plays an important role for getting a stable solution in ELM. Moreover, from the above propositions, we know that full rank hidden layer matrix cannot secure the model performance because the full rank could be a consequence of matrix perturbation and generalized inverse will not be continuous from waning rank to full rank. In the following section, some numerical experiments were carried out from different perspectives to show this special hidden layer matrix pattern and its final impact on model performance.

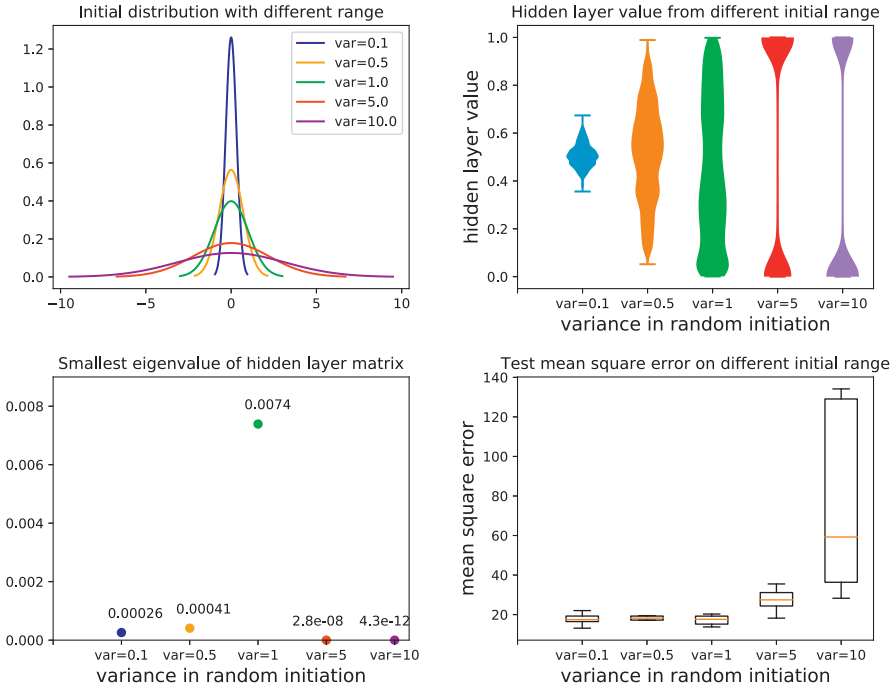


Fig. 2. Model comparison between different initial ranges.

4. Experiments

4.1. Different random initial range

During the training of ELM, weights and bias between input and hidden layers are random selected and a common choice is sampling from standard normal distribution. Because the sigmoid function is bounded between 0 and 1, if we change the random initial range by using different variance in normal distribution, the full rank hidden layer could move close to a waning rank matrix which will eventually harm the model generalization ability. This experiment is based on the House Prices dataset with 50 hidden layer nodes and results are visualized in Fig. 2

From upper two graphs in Fig. 2, we can see the different distributions adopted by random initialization give different range to weights and bias. And wider range of initialization gives more separated value in hidden layer. With initial distribution following *Normal*(0, 5) and *Normal*(0, 10), most of the hidden layer values are either 0 or 1. Such pattern in hidden layer matrix create a high possibility of collinearity among columns. In this circumstance, the full column rank of hidden layer still holds but with tiny eigenvalue (almost zero eigenvalue in Fig. 2 lower left). Therefore, the ELM model runs into a perturbed matrix rank situation. And the discontinuity of generalized inverse lead to unstable model performance (large range of mean square in Fig. 2 lower right).

This experiment can be repeated based on other distributions with different ranges, for example uniform distribution or student's-t distribution. It is worth mentioning that [17,18] also pointed out this phenomenon related to Moore–Penrose pseudo-inverse and Random Vector Functional Link Networks (RVFL). With the proof in Section 3 and visualization in Fig. 2, we can have a more comprehensive understanding of this issue.

4.2. Increasing hidden layer nodes

The choice of network structure in ELM, especially the number of hidden nodes, requires a balance between training accuracy and model efficiency. In this part, we show that because the number of hidden node is exactly the number of columns in hidden layer matrix, the more number of hidden nodes the model has, the closer hidden layer columns to linear correlation. The following experiment will demonstrate this phenomenon.

Training House Price dataset with increasing number of hidden nodes, each time recorded the test mean square error and smallest eigenvalue of hidden layer matrix. From Fig. 3 right, the mean square error first decreases and then increases with number of nodes increasing from 10 to 150. On the left, the smallest eigenvalue decreases to 7.5×10^{-6}. When decreasing, the hidden layer is moving close to the boundary of full and waning rank. With number of hidden nodes greater than 80, the analytical part of ELM already starts to suffer from the matrix rank perturbation effect.

4.3. Training set with outliers

Training set with outliers could effect most of the machine learning algorithms. For ELM, outliers will cause the rank perturbation problem. To verify, we create an artificial dataset with significant outlier for ELM training. Suppose we have a two dimension structural dataset with 500 instances, and the data is following a normal distribution with low variance except one outlier. The construction of this dataset is shown in 10.

$$\begin{cases} x_{1,j} \sim \text{Normal}(1, 0.1) & j = 1, 2, \ldots, 499 \quad \text{and} \quad x_{1,500} = 10 \\ x_{2,j} \sim \text{Normal}(3, 0.1) & j = 1, 2, \ldots, 499 \quad \text{and} \quad x_{2,500} = 30 \end{cases} \tag{10}$$

For simplicity, we first re-scale the input range in [0,1], then apply them to a SLFN with number of hidden nodes $m = 5$, and randomly assign weights and bias between input and hidden layers. The setting of outliers will cause robustness problem. The rank of $\mathbf{H}$ is 5 which means it is a full rank hidden layer matrix. Yet the column-wise variances are all near zero which indicates the columns are actually close to each other and the full-rank is just a perturbation from waning rank. In fact, the smallest eigenvalue of $\mathbf{H}^T\mathbf{H}$ is 6.08×10^{-6}. When computing the generalized inverse, it

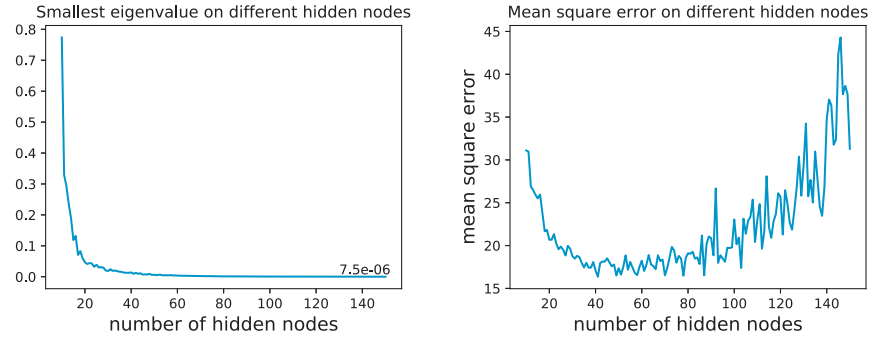
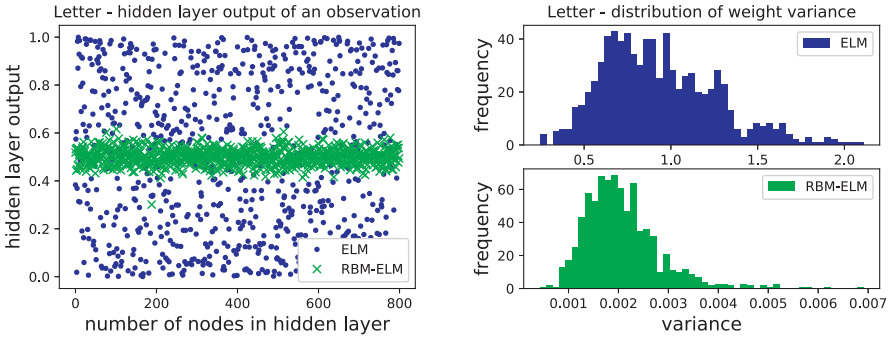


Fig. 3. Model comparison between different number of hidden nodes.



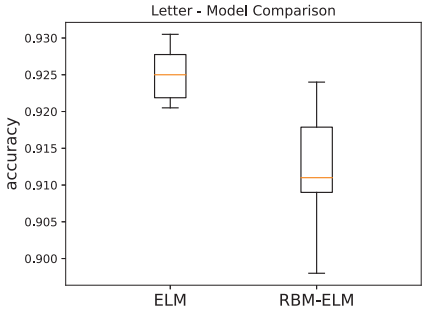


Fig. 4. Model comparison between different initial methods.

will have large norm and variance. In this case, the norm $||\mathbf{H}^{\dagger}|| = 2.41 \times 10^5$ and variance $\mathrm{Var}(\mathbf{H}^{\dagger}) = 2.32 \times 10^7$. With such generalized inverse, the model will fail to learn the real pattern of dataset. In general, the rank of input matrix also plays an important part of the ELM model training [19].

Table 1
Generalized inverse (GI) comparison.

Model	GI-Mean	GI-Variance	GI-Norm
ELM	1.7052×10^{-7}	0.0362	94.62
RBM-ELM	1.5632×10^{-7}	14518.80	2637.16

4.4. Unsupervised pre-training with RBM

Now we consider another ELM approach: instead of random assigning, the Restricted Boltzmann machines (RBMs) [20,21] are used as an unsupervised pre-training phase for weights between input and hidden layers [22]. RBM is a generative stochastic model which can be used to capture the probability distribution over a set of inputs. Recent study and application of RBM can be found in [23]. After RBM pre-training, the network is analytical solved by GI as a supervised fine-tuning phase. Named RBM-ELM, this approach in SLFN is mentioned in [24] and extended to multiple-hidden layer feed-forward neural networks (MLFNs) in [25]. We found for some dataset, the RBM pre-trained hidden matrix could also be a waning rank perturbation. The experiments are based on the Letter Recognition dataset from UCI Machine Learning Repository [26], results are shown in Fig. 4.

First, we train both models with 800 hidden nodes. Taking a random observation, although both hidden layer matrices are full rank, the ELM hidden values are close to a uniform distribution within [0,1], while the RBM hidden values is nearly a perturbation around constant 0.5, see Fig. 4 upper left. Then the values of column-wise variance have different pattern between two hidden matrices, see Fig. 4 upper right. That means, the column vectors of RBM hidden matrix are close to each other. Furthermore, the generalized inverses are compared in Table 1. The large norm and variance are noteworthy. Same as other experiments, this pattern is due to the discontinuity. At last, we compare the test accuracy based on 10-fold cross validation. Fig. 4 lower shows when the full rank matrix is a perturbation from waning rank, the model generalization ability will be reduced.

5. Conclusion

This paper presents a study on sensitivity of hidden layer matrix rank in ELM. We first review the training process of ELM from the matrix transformation standpoint. Then focus on the relationship between rank of matrix and continuity of generalized inverse. The experiments are carried out to visually analyze this issue. The conclusion can be listed as follow:

1. Generalized inverse is continuous with full rank matrix, but discontinuous when waning rank matrix perturbs to full rank or vice versa.
2. Even if the sigmoid function transform input data to a full rank hidden matrix with probability 1, it is possible that the full rank is actually close to a waning rank.
3. Because of the solution of ELM highly depends on the full column rank assumption, the rank degeneration will prevent model from learning the pattern of data.
4. During training of ELM, initial range, initial method, outliers and network structure all could cause the rank perturbation problem.
5. To ensure the generalization ability of ELM, we suggest that special attention should be paid to monitor the data pattern and eigenvalue of hidden matrix.

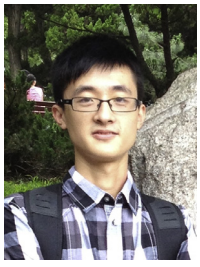
Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant nos. 61772344 and 61732011), in part by the Natural Science Foundation of SZU (Grant nos. 827-000140, 827-000230, and 2017060).

References

[1] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Extreme learning machine: a new learning scheme of feedforward neural networks, in: Proceedings of the IEEE International Joint Conference on Neural Networks, 2, IEEE, 2004, pp. 985–990.
[2] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Extreme learning machine: theory and applications, Neurocomputing 70 (1) (2006) 489–501.
[3] S. Ding, N. Zhang, J. Zhang, X. Xu, Z. Shi, Unsupervised extreme learning machine with representational features, Int. J. Mach. Learn. Cybern. 8 (2) (2017) 587–595.
[4] R. Wang, Y.-L. He, C.-Y. Chow, F.-F. Ou, J. Zhang, Learning elm-tree from big data based on uncertainty reduction, Fuzzy Sets Syst. 258 (2015) 79–100.
[5] J. Zhai, S. Zhang, C. Wang, The classification of imbalanced large data sets based on mapreduce and ensemble of elm classifiers, Int. J. Mach. Learn. Cybern. 8 (3) (2017) 1009–1017.
[6] N.L. Azad, A. Mozaffari, A. Fathi, An optimal learning-based controller derived from hamiltonian function combined with a cellular searching strategy for automotive coldstart emissions, Int. J. Mach. Learn. Cybern. 8 (3) (2017) 955–979.
[7] W. Mao, J. Wang, Z. Xue, An elm-based model with sparse-weighting strategy for sequential data imbalance problem, Int. J. Mach. Learn. Cybern. 8 (4) (2017) 1333–1345.
[8] Y. Luo, B. Yang, L. Xu, L. Hao, J. Liu, Y. Yao, F. van de Vosse, Segmentation of the left ventricle in cardiac MRI using a hierarchical extreme learning machine model, Int. J. Mach. Learn. Cybern. pp (2017) 1–11.
[9] N. Zeng, H. Zhang, W. Liu, J. Liang, F.E. Alsaadi, A switching delayed PSO optimized extreme learning machine for short-term load forecasting, Neurocomputing 240 (2017) 175–182.
[10] X. Luo, X. Yang, C. Jiang, X. Ban, Timeliness online regularized extreme learning machine, Int. J. Mach. Learn. Cybern. 9 (3) (2018) 465–476.
[11] H. Zhao, X. Guo, M. Wang, T. Li, C. Pang, D. Georgakopoulos, Analyze eeg signals with extreme learning machine based on PMIS feature selection, Int. J. Mach. Learn. Cybern. 9 (2) (2018) 243–249.
[12] G.-B. Huang, H. Zhou, X. Ding, R. Zhang, Extreme learning machine for regression and multiclass classification, IEEE Trans. Syst. Man. Cybern., Part B (Cybern.) 42 (2) (2012) 513–529.
[13] A.-M. Fu, X.-Z. Wang, Y.-L. He, L.-S. Wang, A study on residence error of training an extreme learning machine and its application to evolutionary algorithms, Neurocomputing 146 (2014) 75–82.
[14] W. Deng, Q. Zheng, L. Chen, Regularized extreme learning machine, in: Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, CIDM, IEEE, 2009, pp. 389–395.
[15] P. Lancaster, M. Tismenetsky, The Theory of Matrices: With Applications, Elsevier, 1985.
[16] H. Avron, E. Ng, S. Toledo, A Generalized Courant-Fischer Minimax Theorem, Lawrence Berkeley National Lab, 2008 Technical report LBNL-6393E.
[17] M. Li, D. Wang, Insights into randomized algorithms for neural networks: Practical issues and common pitfalls, Inf. Sci. 382 (2017) 170–178.
[18] L. Zhang, P.N. Suganthan, A comprehensive evaluation of random vector functional link networks, Inf. Sci. 367 (2016) 1094–1105.
[19] X. Zhao, W. Cao, H. Zhu, Z. Ming, R.A.R. Ashfaq, An initial study on the rank of input matrix for extreme learning machine, Int. J. Mach. Learn. Cybern. 9 (5) (2018) 867–879.
[20] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[21] G.E. Hinton, A practical guide to training restricted Boltzmann machines, in: Neural Networks: Tricks of the trade, Springer, 2012, pp. 599–619.
[22] M. Liu, B. Liu, C. Zhang, W. Wang, W. Sun, Semi-supervised low rank kernel learning algorithm via extreme learning machine, Int. J. Mach. Learn. Cybern. 8 (3) (2017) 1039–1052.
[23] N. Zeng, Z. Wang, H. Zhang, W. Liu, F.E. Alsaadi, Deep belief networks for quantitative analysis of a gold immunochromatographic strip, Cognit. Comput. 8 (4) (2016) 684–692.
[24] A.G. Pacheco, R.A. Krohling, C.A. da Silva, Restricted boltzmann machine to determine the input weights for extreme learning machines, Expert Syst. Appl. 96 (2018) 77–85.
[25] X.-Z. Wang, T. Zhang, R. Wang, Noniterative deep learning: Incorporating restricted Boltzmann machine into multilayer random weight neural networks, IEEE Trans. Syst. Man. Cybern. Syst. PP (99) (2017) 1–10.
[26] D. Dheeru, E. Karra Taniskidou, UCI machine learning repository (2017). URL: http://archive.ics.uci.edu/ml.

Professor Xizhao Wang received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1998. He is currently a Professor with the Big Data Institute, Shenzhen University, Shenzhen, China. His current research interests include uncertainty modeling and machine learning for big data. He has edited more than ten special issues and published three monographs, two textbooks, and more than 200 peer-reviewed research papers. By the Google scholar, the total number of citations is over 5000. He is on the list of Elsevier 2015/2016 most cited Chinese authors. He is the Chair of the IEEE SMC Technical Committee on Computational Intelligence, the Editor-in-Chief of Machine Learning and Cybernetics Journal, and Associate Editor for a couple of journals in the related areas. He was a recipient of the IEEE SMCS Outstanding Contribution Award in 2004 and a recipient of the IEEE SMCS Best Associate Editor Award in 2006.



Zhiqi Huang received his B.Sc. degree in applied mathematics from the Sun Yat-sen University, Guangzhou, China and M.Sc. degree in statistics from University of Maryland, United States. He is working as research assistant at Big Data Institute, Shenzhen University, Shenzhen, China. His current research interests focus on machine learning algorithms and neural network structure.