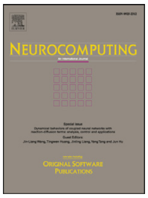
Contents lists available at ScienceDirect

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 xxx

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 feedforward 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/© 2018 Elsevier B.V. All rights reserved.

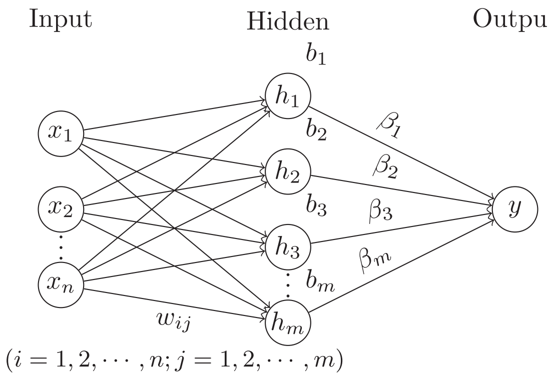


Fig. 1. A simple ELM structure.

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

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

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

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