F-043

Feature selection using $\ell_{2,1-2}$ regularized deep neural network

Zheng Zide*　Washizawa Yoshikazu†

1 Introduction

The back propagation neural network is one of the most popular and widely used models because of the excellent data fitting ability. However it is prone to over-fit. The main reasons for the over-fitting problem are noise and using too much irrelevant features in the training process. As sensors of the internet of things (IoT) play an important role in data collection and the limitation of environmental conditions, noise is inevitable in the process of data collection. We can avoid the over-fitting problem by reducing impact of noise and selecting relevant features.

Feature selection aims to select relevant subset features from the original feature set. It can help speeding up the learning process, reducing the data storage cost, and relaxing noise influence and the over-fitting problem. Sparse feature selection uses a row sparse matrix to select features, and the $\ell_{2,1}$ norm is widely used to obtain the row sparse matrix. In [1], the $\ell_{2,1-2}$ norm was proposed to select features and achieved excellent classification performance. Comparing with the $\ell_{2,1}$ norm, the $\ell_{2,1-2}$ norm is likely to obtain sparser solution.

In this paper, motivated by the advantage of the $\ell_{2,1-2}$ norm, we apply it to the back propagation neural network model to select relevant features and build a robust model. In details, we use the $\ell_{2,1-2}$ norm in the input layer to select features, and use the Frobenius ($\ell_{2,2}$) norm regularization in the rest of the layers to enhance the robustness of the model. As a result, the proposed method is considered to be better classification performance than the Frobenius norm regularization.

2 Neural network

In the forward propagation of the neural network, neurons deliver data from the former layer to the next layer. Let $l = 0, 1, \ldots, L$ be the index of the layer, then we have

$$\begin{aligned} \mathbf{z}^l &= \mathbf{W}^l \mathbf{a}^{l-1} + \mathbf{b}^l \\ \mathbf{a}^l &= \sigma^l(\mathbf{z}^l), \end{aligned} \tag{1}$$

where $\mathbf{W}^l$, $\sigma^l(\cdot)$, $\mathbf{z}^l$, $\mathbf{a}^l$, and $\mathbf{b}^l$ are the weight matrix, the activation function, the linear result, the output, and the bais of the *l-th* layer respectively.

The weight which minimizes the error of the network can be obtained by the back propagation algorithm. With the gradient descent, let $\boldsymbol{\delta}^l = \frac{\partial L(\mathbf{W})}{\partial \mathbf{z}^l}$, the gradient between two layer is

$$\boldsymbol{\delta}^l = (\mathbf{W}^{l+1})^T \boldsymbol{\delta}^{l+1} \odot \sigma'^l(\mathbf{z}^l), \tag{2}$$

and the gradients of W^l_{jk} and the bias b_j are given by

$$\frac{\partial L(\mathbf{W})}{\partial W^l_{jk}} = \delta^l_j a^{l-1}_k, \tag{3}$$

and

$$\frac{\partial L(\mathbf{W})}{\partial b^l_j} = \delta^l_j. \tag{4}$$

2.1 $\ell_{2,1-2}$ norm regularization

Sparse feature selection aims to get a row sparse marix $\mathbf{W}$ to select features. Mathematically, it is described as

$$\min_{\mathbf{W}} L(\mathbf{W}) + \alpha R(\mathbf{W}), \tag{5}$$

where $L(\mathbf{W})$ is the loss function, $R(\mathbf{W})$ is the regularization term, and $\alpha > 0$ is the regularization parameter.

For $\mathbf{W} \in \mathbf{R}^{d \times c}$, the definition of the $\ell_{2,1-2}$ norm is

$$\|\mathbf{W}\|_{2,1-2} = \|\mathbf{W}\|_{2,1} - \|\mathbf{W}\|_F, \tag{6}$$

where the $\ell_{2,1}$ norm and the Frobenius norm are defined as

$$\|\mathbf{W}\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{c} \mathbf{W}^2_{i,j}}, \ \|\mathbf{W}\|_F = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{c} \mathbf{W}^2_{i,j}}. \tag{7}$$

And $\mathbf{W}_{i,j}$ is the element of matrix $\mathbf{W}$ (*i-th* row *j-th* column), the gradient of the $\ell_{2,1}$ norm is

$$\frac{\partial}{\partial \mathbf{W}}(\|\mathbf{W}\|_{2,1}) = [\phi(\mathbf{w}_1)^T, \phi(\mathbf{w}_2)^T, \ldots, \phi(\mathbf{w}_c)^T]^T, \tag{8}$$

where

$$\phi(\mathbf{w}_i) = \begin{cases} \frac{\mathbf{w}_i}{\|\mathbf{w}_i\|_F}, & \text{if } \mathbf{w}_i \neq 0 \\ 0, & \text{if } \mathbf{w}_i = 0, \end{cases} \tag{9}$$

and $\mathbf{w}_c$ is the *c-th* row of matrix $\mathbf{W}$. The gradient of the Frobenius norm is

$$\frac{\partial}{\partial \mathbf{W}}(\|\mathbf{W}\|_F) = \begin{cases} \frac{\mathbf{W}}{\|\mathbf{W}\|_F}, & \text{if } \mathbf{W} \neq 0 \\ 0, & \text{if } \mathbf{W} = 0. \end{cases} \tag{10}$$

*電気通信大学修士既卒 Graduates of the University of Electro-Communications
†電気通信大学 The University of Electro-Communications

We apply the $\ell_{2,1-2}$ norm to the input layer for selecting features, and use the Frobenius norm regularization in the other layers to enhance the robustness of the model. The optimization problem is

$$\min_{\mathbf{W}} L(\mathbf{W}) + \alpha\Big\{\|\mathbf{W}^1\|_{2,1-2} + \frac{1}{2}\sum_{l=2}^{L}\|\mathbf{W}^l\|_F^2\Big\}. \quad (11)$$

Let $N(\mathbf{W})$ represents the norm part of (11). The gradient of W_{jk}^l is

$$\frac{\partial L(\mathbf{W})}{\partial W_{jk}^l} + \alpha\frac{\partial N(\mathbf{W})}{\partial W_{jk}^l}, \quad (12)$$

where $\frac{\partial L(\mathbf{W})}{\partial W_{jk}^l}$ is given by Eq. (3) and $\frac{\partial N(\mathbf{W})}{\partial W_{jk}^l}$ can be obtained by using Eqs. (8) and (10). As the gradients are given, we use the gradient descent to update the parameter of the network.

3 Experiment

We used five open datasets (COIL20, Mnist, ORL, USPS, and Yale) to conduct experiments. We set the learning rate as 0.1 and selected the best regularization parameter from the range of {0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05}. We also made minor adjustments to the above parameters. For loss function $L(\mathbf{W})$, we used the cross entropy loss in our experiment.

We conducted the experiment with five fold cross validation, and compared the result with four cases; no regularization, the Frobenius norm, the $\ell_{2,1}$ norm and the $\ell_{2,1-2}$ norm regularization. The results are shown in Table 1. As the $\ell_{2,1}$ norm also obtains a sparse solution, we compared the effect of sparsity between the $\ell_{2,1}$ norm and the $\ell_{2,1-2}$ norm. We verified the effect of sparsity by comparing the sparse rate (ratio of zero row vectors in the sparse matrix) and the sparse classification accuracy (result of using sparse matrix to conduct the experiment). The results are shown in Table 2.

Table 1: Classification accuracy of four methods

Average classification accuracy [%]				
Dataset	no regularization	$\ell_{2,2}$	$\ell_{2,1}$	$\ell_{2,1-2}$
COIL20	90.44	91.18	91.32	**93.65**
Mnist	96.82	97.03	97.60	**97.99**
ORL	93.75	94.25	95.00	**95.25**
USPS	95.92	96.20	96.83	**97.03**
Yale	80.13	81.21	82.40	**85.77**

Table 2: Results of sparsity

Sparse rate and sparse classification accuracy [%]		
Dataset	$\ell_{2,1}$	$\ell_{2,1-2}$
COIL20	12.42/90.76	15.01/92.19
Mnist	39.59/97.21	56.28/97.71
ORL	9.561/90.77	10.29/94.50
USPS	4.981/95.74	5.603/96.38
Yale	7.005/81.95	9.502/83.59

3.1 Discussion

As shown in Table 1, the Frobenius norm regularization showed higher classification accuracy than no regularization. As the $\ell_{2,1}$ norm and the $\ell_{2,1-2}$ norm were used to select features, they both showed higher classification accuracy than the Frobenius norm regularization. The $\ell_{2,1-2}$ norm case showed the best result because of the feature selection effect. To confirm this, we compared the effect of sparsity between the $\ell_{2,1}$ norm and the $\ell_{2,1-2}$ norm. As shown in Table 2, the $\ell_{2,1-2}$ norm showed a higher sparse rate and classification accuracy than the $\ell_{2,1}$ norm. We also did the paired samples student's t-test between the classification accuracy of the Frobenius norm regularization and the $\ell_{2,1-2}$ norm regularization. The results confirmed that the classification accuracy of the proposed method is significantly higher than the Frobenius norm regularization.

4 Conclusion and future work

We applied the $\ell_{2,1-2}$ norm as the regularization term to the back propagation neural network to select features and build a robust model. We proposed it into the input layer, and used the Frobenius norm in the other layers to enhance the robustness of the model. As a result, experiments on five open datasets showed the effectiveness of our method.

For our future work, as the dropout tries to shut-down some neurons randomly in each layer of the network, we try to apply the $\ell_{2,1-2}$ norm to all layers of the network to "shut-down" some neurons discriminantly.

References

[1] Y. Shi, J. Miao, Z. Wang, P. Zhang, and L. Niu. Feature selection with $\ell_{2,1-2}$ regularization. *IEEE Transactions on Neural Networks and Learning Systems*, 29(10):4967–4982, 2018