Rolling Force Prediction Algorithm Based on Bayesian Regularization Neural Network

Xiaodan Zhang[1,a], Lu Yao[2,b] and Zhenxiong Zhou[1,c,*]

[1]College of Electrical & Information Engineering, Beihua University ,Jilin, China

[2]Synthetic resin factory of Jilin Petrochemical Co

[a] z1314yo@163.com, [b]1410770136@qq.com, [c]742884852@qq.com

*Corresponding author

Keywords: Hot continuous rolling, Rolling force prediction, Neural network, Bayesian regularization.

Abstract. For obtaining relative accurate rolling-mill model is difficulty by the simple mathematical method, due to the complexity of the actual production scene and the non-linear relationship between variables, this paper firstly proposes an improved Bayesian regularization neural network model according to these measured data of 1580 production line. In this model, the paper constructs the improved Bayesian neural networks by the introduction of bound terms that represents the network complexity in the objective function. At last, the simulation result proves the effectiveness and validity of the model and the prediction accuracy of the model algorithm is superior to the traditional model.

1. Introduction

The characteristics of the rolling process are non-linear, large-delay, strong coupling and parameter variation. Since state parameters of the control system are constant change, the traditional control model cannot well adapted due to its shortcomings. Taking into account of the learning neural network, many studies showed that predictive control effect using traditional BP algorithm or LM algorithm, which use neural networks to create rolling force model and use on-site measured data for training and learning, is remarkable. Since neural network training time is usually too long, and the sample data contains noise, there are problems of training times too much or the network scale too large and other issues, so it tends to make the network to remember unnecessary details when neural network training. If the noise included in the training process of network data are recorded, the new data may result in incorrect output, that is to say the traditional neural network algorithm does not have good generalization function, and there is a few problems such as the difficulty to control complexity degree of the model and the difficulty to overcome over-fitting data and so on. According to the project specific issues, comparing with the BP algorithm, LM algorithm and Bayesian algorithm[1] , this paper proposes an improved Bias method of neural network prediction for the rolling force of the hot rolling mill, so as to obtain the better mathematical model.

2. The principle of Bayesian regularization neural network

The method of Bayesian regularization mainly through modifying the training performance functions of neural network to improve their marketing capabilities. Input variables are two categories in actual system, one can be observed, and the other is also uncontrollable and unobserved. But the two variables have an impact on the output system. Let X be an observable variable, X=[x1, x2…xn]. Then the following relationship between the system output d and the input x:

$$d = f(x) + \varepsilon \tag{1}$$

Where f represents the effect of the unobservable inputs to the outputs in the system; represents the random variable with a distribution.
The training performance function of neural network is generally using the mean square error function. Assuming the error function E0 is:

$$E_0 = \frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{N}(y'_{nk} - \mathrm{y}_{nk})^2 \tag{2}$$

Where N represents the number of samples, K represents the output number of neural network, y'_{nk} represents the expected output, y_{nk} represents the actual output of network.
Although the function that make the above-mentioned objective function reach to minimize has infinite, the neural network has local minimum. Aiming at the above-mentioned problems, it can be solved by regularization theory which adds a constraint term to obtain stable and useful solutions. In general, if F(x) is smooth, it will have the interpolation ability. When the network weight is small, the network output is smoother. So using smoothness constraint as a constraint term, it can effectively reduce the network weight. Then the objective function is:

$$F = \alpha \cdot \frac{1}{2}\sum_{i=1}^{W} w_i^{\,2} + \beta \cdot \frac{1}{2}\sum_{k=1}^{K}\sum_{j=1}^{N}(y'_{nk} - y_{nk})^2 = \alpha J_W(F) + \beta J_D(F) \tag{3}$$

Where JW represents squares of the network weights, JD represents the response the residual squares sum of the response and the target value, α and β represent hyper-parameters which control other parameters (such as weights and thresholds) forms of distribution. If $\alpha << \beta$, the aim of the training algorithm is to make training error of the network as small as possible; if $\alpha >> \beta$, then the purpose of the training algorithm is to minimize valid network parameters ,so that it be able to compensate for the larger networks errors and make the network produce a smoother response.

3. Improved Bayesian regularization neural network algorithm

Bayesian neural networks put the probability distribution of weight value (threshold value) in the whole space as the starting point, considers the parameter as a random variable, considers the objective function as the likelihood function of training data, and the right decay term corresponds to the priori probability distribution of the network parameters, and integration the prior probability distribution assumption of the parameters, and the parameters of the posterior distribution can be constantly adjusted after the observing data are given . The prediction results of Bayesian neural network are based on an average of the posterior distribution of the parameters, a single model is mapped to a point in parameter space, and all models are mapped to the entire parameter space, in order to guarantee strong generalization ability of the network in theory [2].
Assuming the network structure H is given (primarily the number of hidden layer neurons) and the network model $di = f(xi, W, H)$ is given. In the absence of the sample data, the prior distribution of the weights (threshold) is $p(w|\alpha, H)$; the posteriori distribution is $p(w|D, \alpha, \beta, H)$ with the sample data set $D = \{x^N, d^N\}$. According to the Bayesian rule [3] is

$$p(w|D, \alpha, \beta, H) = \frac{p(D|w, \beta, H)\, p(w|\alpha, H)}{p(D|\alpha, \beta, H)} \tag{4}$$

Where *p(D|w,β,H)* represents the likelihood function, *p(D|α,β,H)* represents normalization factor, w represents the weight value (threshold) vector. The knowledge on the weights distribution is little when there is no data; therefore the prior distribution is a very wide distribution. It can be converted to a compact posterior distribution when the data are obtained; the weight value only in a very small range will produce consistent with the performance of the network map [4].

3.1 Prior probability

In the absence of the prior knowledge of weights, p(w|α, H) follows the Gauss distribution that the mean is 0 and the variance is 1/α[5]:

$$p(w \mid \alpha, H) = \frac{1}{z_W(\alpha)} \exp(-\alpha \sum_{i=1}^{W} w_i) \tag{5}$$

Thus, the value of the normalization factor ZW(α) is

$$z_W(\alpha) = \int_{-\infty}^{+\infty} \exp(-\alpha \sum_{i=1}^{W} w_i) dw = (\frac{2\pi}{\alpha})^{\frac{W}{2}} \tag{6}$$

3.2 Approximate probability

Assuming the noise smoothing function with a Gauss distribution that the mean is 0 and the variance is 1/ β produces the desired output d, for a given input x, the observed probability of the output d:

$$p(d_n \mid x_n, w, \beta, H) \infty \prod_{k=1}^{K} \exp(\frac{\beta}{2}[d_{nk} - y_{nk}(x_n, w, \beta, H)]^2) \tag{7}$$

If each sample independently selected data,

$$p(D \mid w, \beta, H) = \prod_{n=1}^{N} p(d_n \mid x_n, w, \beta, H) = \frac{1}{z_D(\beta)} \exp(-\beta J_D) \tag{8}$$

The normalization factor $z_D(\beta) = \int_{-\infty}^{+\infty} \exp(-\beta J_D) dD$, $J_D = \frac{1}{2} \sum_{k=1}^{K} \sum_{j=1}^{N} (y_{nk} - c_{nk})^2$,therefore,

$z_D(\beta) = (\frac{2\pi}{\beta})^{\frac{N}{2}}$ .Where N is the input vector dimension.

3.3 Optimized and solved

The Prior Probability function and approximate probability function into equation (8), we can obtain:

$$p(w \mid D, \alpha, \beta, H) = \frac{1}{z_M(\alpha, \beta)} \exp(-\beta J_D - \alpha J_W) = \frac{1}{z_M(\alpha, \beta)} \exp[-M(w)] \tag{9}$$

where $z_M(\alpha, \beta) = \int_{-\infty}^{+\infty} (-\beta J_D - \alpha J_W) dw$.

If the sample data reaches a certain number, the posterior distribution tends to Gaussian distribution. If the posterior distribution curve simultaneously satisfies the sufficiently narrow and the sharply enough peak, you can further simplify the problem, namely using the Taylor expansion obtain $z_M(\alpha, \beta)$. Assume that w_{MP} is the weight value (threshold value) to which B is the minimum value corresponding. The Taylor expansion of $M(w)$ in the vicinity of w_{MP} is

$$M(w) \approx M(w_{MP}) + \frac{1}{2}(w - w_{MP})^T \nabla\nabla M(w_{MP})(w - w_{MP}) \tag{10}$$

Where $\nabla\nabla M(w_{MP}) = \beta \nabla\nabla J_D(w_{MP}) + \alpha \nabla\nabla J_W(w_{MP}) = \beta \nabla\nabla J_D(w_{MP}) + \alpha I$. $\nabla\nabla$ represents the second derivative, therefore

$$z_M(\alpha,\beta)=(2\pi)^{\frac{W}{2}}\{\det[\nabla\nabla M(w_{MP})]\}^{-\frac{1}{2}}\exp[-M(w_{MP})] \tag{12}$$

3.4 Approximate calculation of hessian matrix

If you want to optimize the solution, the first is to calculate *Hessian* matrix when $M(w)$ in the minimum point of w_{MP}. The formula (11) shows that the calculation amount of $\nabla\nabla J_D(w_{MP})$ is large. Therefore, *Hessian* matrix can be further simplified to improve computing speed.
Make $\varepsilon_{nk}=(y_{nk}-c_{nk})$, then

$$\frac{\partial J_D}{\partial w}=\{\sum_{k=1}^{K}\sum_{n=1}^{N}\frac{\partial\varepsilon_{nk}}{\partial w_i}\} \tag{13}$$

$$\nabla\nabla(J_D(w_{MP}))_{ij}\approx\sum_{k=1}^{K}\sum_{n=1}^{N}[\frac{\partial\varepsilon_{nk}}{\partial w_i}\cdot\frac{\partial w_{nk}}{\partial w_j}+\varepsilon_{nk}\cdot\frac{\partial^2\varepsilon_{nk}}{\partial w_i\partial w_j}] \tag{14}$$

3.5 Determination of hyper-parameters α and β

Hyper-parameters α and β can be obtained by calculating the posterior distribution:

$$p(\alpha,\beta\mid D,H)=\frac{p(D\mid\alpha,\beta,H)\,p(\alpha,\beta\mid H)}{p(D\mid H)} \tag{15}$$

Assume that the prior distribution $p(\alpha,\beta\mid H)$ meet a very wide distribution function. Because the normalization factor $p(D\mid H)$ has nothing to do with α,β in the above formula, so the problem of obtaining the maximum a posteriori distribution could be transformed into the problem of solving maximum likelihood function. Because the approximate function $p(D\mid\alpha,\beta,H)$ is the normalization factor of the formula (15), then

$$p(D\mid\alpha,\beta,H)=\frac{p(D\mid w,\beta,H)\,p(w\mid\alpha,H)}{p(w\mid D,\alpha,\beta,H)} \tag{16}$$

Uniting the formula (8) and formula (9), we can obtain

$$p(D\mid\alpha,\beta,H)=\frac{z_M(\alpha,\beta)}{z_D(\beta)z_W(\alpha)} \tag{17}$$

For formula (7) taking the logarithm

$$\ln(p(D\mid\alpha,\beta,H)=-\alpha J_W(w_{MP})-\beta J_D(w_{MP})+\frac{N}{2}\ln\beta-\frac{N}{2}\ln(2\pi)-\frac{1}{2}\ln\{\det[\nabla\nabla M(w_{MP})]\}+\frac{W}{2}\ln\alpha \tag{18}$$

If the characteristic value of $\beta\nabla\nabla J_D(w_{MP})$ is $\{\lambda_i\},i=1,2,...W$, we can obtain that the characteristic value of $\nabla\nabla J_D(w_{MP})$ is $\{\lambda_i+\alpha\}$ by the formula (11). Also, because J_D is a normal error term, then

$$\frac{d}{d\alpha}\ln\{\det[\nabla\nabla M(w_{MP})]\}=\frac{d}{d\alpha}\ln[\prod_{i=1}^{W}(\lambda_i+\alpha)]=\sum_{i=1}^{W}\frac{1}{\lambda_i+\alpha}=tr[\nabla\nabla M(w_{MP})]^{-1} \tag{19}$$

Since λ_i and β are proportional, therefore

$$\frac{d\lambda_i}{d\beta}=\frac{\lambda_i}{\beta}\Rightarrow\frac{d}{d\beta}\ln\{\det[\nabla\nabla M(w_{MP})]\}=\frac{d}{d\beta}\ln[\prod_{i=1}^{W}(\lambda_i+\alpha)]=\frac{d}{d\beta}\sum_{i=1}^{W}[\ln(\lambda_i+\alpha)]=\sum_{i=1}^{W}\frac{1}{\lambda_i+\alpha}\cdot\frac{d\lambda_i}{d\beta}=\frac{1}{\beta}\sum_{i=1}^{W}\frac{\lambda_i}{\lambda_i+\alpha} \tag{20}$$

Respectively make the partial derivative of A and B in the formula (20) equal to 0, you can get:

$$\begin{aligned}2\alpha J_W(w_{MP})&=W-\sum_{i=1}^{W}\frac{\alpha}{\lambda_i+\alpha}=\sum_{i=1}^{W}\frac{\lambda_i}{\lambda_i+\alpha}=\gamma\\2\beta J_W(w_{MP})&=N-\sum_{i=1}^{W}\frac{\lambda_i}{\lambda_i+\alpha}=N-\gamma\end{aligned} \tag{21}$$

Therefore, the maximum prominence of α_{MP} and β_{MP} can be obtained:

$$\alpha_{MP}=\frac{\gamma}{2J_W(w_{MP})},\beta_{MP}=\frac{N-\gamma}{2J_D(w_{MP})} \tag{22}$$

Where γ represents the number of parameters works in reducing the network performance index function, $\gamma\in(0,W)$.

In summary, the Bayesian neural network is an iterative process, each iteration involves three inferences: the first layer inference is to maximize $P(w|\alpha,\beta,H)$ under the conditions of hyper-parameters; the second layer inference is to optimize hyper-parameters α,β, and to infer the most possible hyper parameters; the third layer inference is the significant degree of calculation model, and select the best model. Fig.1 shows Bayesian neural network training flow chart.

4. Simulation

In the theoretical analysis, the paper made a series of simulation experiments based on the measured data of a 1580 hot strip mill production line which consists of slab yard, furnace, roughing mill group, finishing mill, coiling machines and other equipment. The strip design thickness of 1580 hot-rolling mill is 1.2mm ~ 12.7mm and width is 700mm ~ 1750mm. The main varieties include low-carbon steel, silicon steel, carbon-structural steel, micro-alloy steel, low-alloy steel. The steel strength class are σb≤65kg/mm2, σs≤50kg/mm2.

The seven rack four-roll mill of 1580 finishing mill was arranged in tandem, the seven mills referred to as F1 ~ F7, the distance between each rack is 5800mm. four roller pairs (PC) mill, roll crossing with unilateral transmission form. Four roller pair-cross (PC) mill is used by F2 ~ F7, the form of unilateral cross-use drive is adopted by roll. F1 has the negative roll bending, F2 ~ F7 has the positive bending [6].

To calculate the rolling speed and rolling time frame: According to the rule of volume flow rate, taking into account the distance of the roughing mill exports to the finishing mill entrance is 18m, and the distance between each rack is 5.8m, The two frame transmission time of the rolled piece in each can be calculated, and the total rolling time also can be obtained. The results are shown in Table1. The pre-set roll gap values of each frame are shown in Table 2.

We can know that there are many factors affecting the rolling pressure changes by the mechanism analysis of the rolling process and prior experience, such as the entrance thickness of rolled plate, exit thickness, reduction rate, rolling temperature, rolling speed, roll diameter, chemical composition content and so on. According to the scene measured data, the determine parameters of each layer is as follows:

The input layer: the C content , the Si content, the Mn content, the Cu content, the entry thickness (H), outlet thickness (h1), rolling width (B), the rolling temperature (T), the rolling time (t1), reduction ratio (e1).
The output layer: the rolling force (P).
The number of neurons in the input layer is 10, the number of neurons in the output layer is 1.

The comparing the results show that L-M algorithm need to calculate the Jacobin matrix and the Hessian matrix, and larger storage space. When the number of parameters is very large, L-M algorithm may not be practical. In addition, the approximation precision of neural network trained by train function for learning samples is very high, it is easy to realize "over-match" for the sample data points. But for non-learning samples (such as validation learning-effect sample), the approximation error will appear a singular phenomenon that decreased and then risen along with the increases in the number of neural network training, which cannot guarantee the generalization ability of the network. However, trainer adds the weights of the network to the performance function, select the optimum weights and thresholds groups so can reduce the weight range in order to make the network output smoother, ease the lack of generalization capability, and ensure that the Bayesian training network is stability and robustness.

5. Conclusions

Based on the background of a 1580mm hot strip mill production line, This paper describes the structure and function of modern hot rolling control system, put forward the multi-level control strategy and research on the finishing mill model based on LEVEL 2. Considering the network stability using the BP algorithm is poor and the generalization ability is low, Bayesian neural networks is introduced into the constraint in the traditional neural function according to the complexity of the actual production site. Combined with the non-linear characteristics of the variables, higher precision neural network prediction model has been obtained based on the measured data. At last, Study and the experimental result found that the prediction accuracy of the optimized model has been significantly improved, and the stability of its network, the training speed and the generalization capabilities are superior to the traditional network neural network.

6. Acknowledgement

This research was financially supported by the Scientific and Technological Planning Project of Jilin Province, China (Grant No.20150519023JH), and the Scientific and Technological Research Project of Department of Education of Jilin Province, China (Grant No.2014167).

References

[1] Liyan, Dong, “Research of Application Foundation on Bayesian Networks [D],” *Changchun: Jilin University*, 2007.

[2] Wei Dong, Zhang Minglian, Jiang Zhijian, Sun Ming, “Neural Network Non-linear Modelling Based on Bayesian Methods,” *Computer Engineering and Applications*, 2005, 11(1): 5.

[3] M.N.Haimeer. “Bayesian-neural network approach for probabilistic modeling of bacterial growth/no-growth interface” *International Journal of Food Microbiology*, 2003, (82): 233-243.

[4] Hu Xianlei, Wang Jun, Wang Zhaodong, “Process Model Setup System of 3340mm Plate of Shougang Group Co,” *Steel Rolling*, 2003, 20(1): 42-44.

[5] Man Leung Wong, Shing Yan Lee, Kwong Sak Leung, “Data Mining of Bayesian Network Using Cooperative Coevolution,” *Decision Support Systems*, 2004, (8): 451-472.