Study of Improved BP Algorithm based on Gradient Descent and Numerical Optimization

Qiuhong Sun[1, 2], Weihong Bi[1] and Xinhang Xu[3]
[1]YanShan University, Qinhuangdao 066004, China
[2]Hebei University of Science and Technology, Shijiazhuang 050000, China
[3]State Grid Hebei Electric Power Research Institute, Shijiazhuang 050000, China
sunqiuhong@hebust.edu.cn

Keywords: Gradient Descent, Numerical Optimization, Improved BP Algorithm.

Abstract: Studied limitations exist in BP model, and discussed the proposed improved algorithm based on BP neural network. Respectively, from the third of the aspects discussed based on improved gradient descent algorithm and improved algorithm based on numerical optimization. Research results showed that the comprehensive method is better than the standard BP algorithm in terms of the number of iterations, the training time and the mean square error and the like, of additional momentum and adaptive outstanding performance parameter method. Researches showed that the Marquardt-Levenberg algorithm neural network convergence fastest training times at least.

1 INTRODUCTION

Along with the application of information management system indifferent fields, the data are continuously stored in the database. Among them, people expect to find out potential knowledge that will help them make decisions. The emergence and development of data mining are just based on this expectation. As a midpoint of different subject studies, data mining involves a great deal of subjects such as database, statistic, machine learning, artificial intelligence, high performance computing, pattern recognition and data visualization etc. Among them, artificial neural network is widely used due to its ability of inherent non-linear processing, adaptive learning and high fault-tolerance.

BP neural network is a feedforward network, the most representative type of network. This class is a multilayer neural network model that map neural network, using a minimum variance of learning, is one of the most widely used neural network model. Multilayer Perceptron network is a kind of three or more sectors of neural networks. A typical multilayer perceptron network is three, feedforward class network, namely: input layer, hidden layer (also called intermediate layer), the output layer. Each neuron between adjacent layers achieves full connection, that is, each neuron and the next layer on layer of each neuron to achieve full connection, and each connection between neurons is not. In practical applications, BP network can be used for classification, regression and time series forecasting and other data mining applications, and pattern recognition problem research, nonlinear mapping problem, such as handwriting recognition, image processing, predictive control, function approximation, data compression and so on.

2 LIMITATIONS OF THE ALGORITHM

In data mining field, although BP network model is currently the most widely used network model, gets good benefits in practical application, and is maturing in theory, but it also has its own limitations and shortcomings, its uncertainty performance in the training process. These limitations are mainly concentrated in three areas: slow convergence and easy to fall into local minima and completely unable to train a network phenomenon.

In recent years, many researchers made many useful improvements on BP network, put forward a number of algorithms to improve the program, such as rapid BP propagation algorithm, the extended Kalman filter algorithm, the second-order

Xu X., Bi W. and Sun Q.
Study of Improved BP Algorithm based on Gradient Descent and Numerical Optimization.
DOI: 10.5220/0006028404520456
In *Proceedings of the Information Science and Management Engineering III (ISME 2015)*, pages 452-456
ISBN: 978-989-758-163-2
Copyright © 2015 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

Study of Improved BP Algorithm based on Gradient Descent and Numerical Optimization

optimization algorithm, the optimal filtering method, and so on. Its main purpose is to avoid falling into local minima and improve the convergence rate. Improve people based on the standard BP algorithm performed faster convergence than the standard gradient method of dozens or even hundreds of times. But they still are not universally applicable, both have their own advantages or defects.

Because a good prospect of artificial neural networks in data mining, and success stories in the practical application, at home and abroad have carried out the neural network data mining for theoretical research. In artificial neural networks, back-propagation algorithm based on BP neural network has occupied a very important position, so the last decade, many researchers have done a thorough study and proposed many improved algorithm and architecture optimization strategy. Which means learning algorithms based on neural network model, to find a neural network connection weights adjustment algorithm, and meet the requirements of the study sample, but also has a faster learning speed. The main goal of improved algorithm is how to reduce the limitations of BP network, speed up training networks in the actual application of data mining process, to avoid falling into local minima, while enhancing the ability of other important network. The structure optimization strategy refers to specific application problems to find an optimal network structure, mainly on how to determine the number of hidden layers and the number of hidden nodes of each hidden layer, and enhance the generalization capability of the network.

In the field of data mining, although BP network model has been currently the most widely used network model, made in the practical application of the good benefits, but in theory are maturing, but it also has its own limitations and shortcomings, its uncertainty performance in the training process. These limitations are mainly concentrated in the following three aspects:

(1) slow convergence

(2) easy to fall into local minimum point

(3) there is a phenomenon that the network is completely unable training

3 IMPROVED BP ALGORITHM BASED ON GRADIENT DESCENT

In recent years, many researchers BP network made many useful improvements, put forward a number of algorithms to improve the program, such as rapid BP propagation algorithm, the extended Kalman filter algorithm, the second-order optimization algorithm, the optimal filtering method, and so on. Its main purpose is to avoid falling into local minima and improve the convergence rate. Improve people based on the standard BP algorithm performed faster convergence than the standard gradient method of dozens or even hundreds of times. But they still are not universally applicable, both have their own advantages or defects.

Improved algorithm based on gradient descent. Gradient descent method is standard on the basis of the objective function by calculating the gradient network weights and closing values corrected, generally used only third-order gradient information on weights and thresholds of the objective function.

Weights standard gradient descent values and thresholds amended iterative process can be expressed as:

$$W^{(k+1)} = W^{(k)} - \alpha \nabla f\left(W^{(k)}\right) \tag{1}$$

Among them, $W^{(k)}$ by the value of the ownership of the network consisting of a vector and threshold, α is the learning rate, $f\left(W^{(k)}\right)$ as the objective function (instead of using the objective function using the error function performance because sometimes contains error than other items), $\nabla f\left(W^{(k)}\right)$ represents the gradient of the objective function. Although the standard BP algorithm to train the network provides a simple and effective method, but because of the training process as a smaller permanent α constant, and thus slow convergence and local minimum problems. To solve these problems, people made a lot of improvements in the study algorithm BP learning algorithm on the basis of its application. More representative are the following:

(1) Additional Momentum Method.

when updating their weight and closing value , additional momentum, not only to consider the role of error in the gradient, but also consider the impact of trends in the error surface. It allows to ignore small changes in properties on the network. In the absence of additional momentum effect, the network may be caught in shallow local minima, with additional momentum effect is likely to slip these minima.

This method is on the basis of the standard BP algorithm, in every weight and the closing value plus a proportional change in weight and the note of the previous amount of change in value, and the reverse propagation method to generate a new right value change. Weights and closing iterative process value

453

correction can be expressed as:

$$\Delta W_{ij}(k+1)=(1-mc)\eta\delta_i p_j+mc\Delta W_{ij}(k) \quad (2)$$

$$\Delta b_i(k+1)=(1-mc)\eta\delta_i+mc\Delta b_i(k) \quad (3)$$

Where k is the number of training, mc is the momentum factor, and generally is about 0.95.

Substantive additional momentum law will affect a recent change in weight, to pass through a momentum factor. When the momentum factor value is zero, change the weight is only produced according to the gradient descent method; when the momentum factor value is 1, the new weight change is the last weight change, and produced according to the gradient method changing portions were ignored. To do this adjustment when the change will increase the weight of the bottom surface toward the mean direction error after momentum. When the network weights into the flat area at the bottom of the error surface, δ_i will become very small, so, $\Delta W_{ij}(k+1)\approx\Delta W_{ij}(k)$, thereby preventing $\Delta W_{ij}(k+1)=0$ situation appears this will help the network to jump out of local minima in the error surface.

According to the design principles of additional momentum, the weight correction cause too much growth results in an error when the new weights was canceled without being adopted, and stopped the momentum effect, so the network does not enter the larger error surface; when the new error rate of change of its old value exceeds the maximum error rate of change of a pre-set time, we have the right to cancel the calculated value change. The maximum error rate of change is generally get the value of 1.04, thus making additional momentum method of training program design, you must add a condition to determine the proper use of weights correction formula. General training program for using the momentum method to determine the conditions as follows:

$$mc=\begin{cases}0, \text{当} SSE(k+1)>SSE(k)\bullet 1.04\\ 0.95, \text{当} SSE(k+1)<SSE(k)\\ mc, \text{其他情况}\end{cases} \quad (4)$$

Wherein, SSE (k) represents the output error of the network timing k.

Additional momentum method can solve some extent local minimum, appropriate to speed up the convergence, but if improper initial iteration direction, the train will be easy access to the platform area, in the calculation process, also because of some wrong adjustments while adding momentum to adjust the rear misleading; the size of the momentum factor which should be selected as the training problem complexity increases and decreases training otherwise prone to oscillation and divergence, for practical applications, we should take a smaller momentum factor.

(2) Adaptive Learning Rate Method.

The basic idea of this approach is the learning rate η error changes should be based on adaptively adjusted to make the connection weights to adjust to changes in the direction of error is reduced. Adaptive learning rate adjustment criteria are: correction value check weights really reduce the error function, if it fell, then the selected learning rate value is small, you can increase it by an amount; if not reducing network error, then produced an overshoot, so accordingly you should reduce the value of learning rate. Specific adaptive learning rate adjustment formula is as follows:

$$\eta(k+1)=\begin{cases}1.05\eta(k), \text{当} SSE(k+1)<SSE(k)\\ 0.7\eta(k), \text{当} SSE(k+1)>SSE(k)\\ \eta(k), \text{其他情况}\end{cases} \quad (5)$$

Wherein, SSE (k) represents the output error of the network timing k.

In the standard BP algorithm, the learning rate α in the training process remains constant. The basic idea of adaptive learning rate law: maintaining the stability of the training, so that each correction weights for iterative steps as large as possible. The process can be expressed as

$$W^{(k+1)}=W^{(k)}-\alpha^{(k)}\nabla f(W^{(k)}) \quad (6)$$

Adaptive learning rate is conducive to shorten the learning time. One of the important reasons for the slow convergence standard BP algorithm is the learning coefficient poor choice. The learning rate is selected too small, too slow convergence; learning rate is chosen too large, it is possible to amend overdone, leading to network flapping diverge. The research shows that, in a certain range to increase the learning coefficient, can greatly speed up the efficiency of learning, get more than the standard BP algorithm converges faster. However, if $\nabla f(W^{(k)})$ is small, there are still a small amount of correction weight problems, resulting in low learning rate.

(3) Additional Momentum and Adaptive Learning Rate Binding Method.

Because of the additional momentum method can alleviate network into local minima situation, and adaptive learning rate method and can speed up the convergence of the network, so some

Study of Improved BP Algorithm based on Gradient Descent and Numerical Optimization

scholars both together to form a new algorithm that additional momentum and adaptive learning rate binding assay. Experimental results show that the comprehensive method than the standard BP algorithm in terms of the number of iterations, the training time and the mean square error and the like, of additional momentum and adaptive outstanding performance parameter method. This paper mainly uses this algorithm can improve the performance of the neural network can learn, and confirm the validity of the algorithm in practical application.

4 IMPROVED ALGORITHM BASED ON NUMERICAL OPTIMIZATION

BP network training is essentially a nonlinear objective function optimization problem. People study of nonlinear optimization problem for several hundred years ,and many traditional numerical optimization method is also faster convergence, and therefore it is natural to think of numerical optimization algorithms based on BP network right on the plant for training. Gradient descent method is different, based on numerical optimization algorithms using not only the first-order derivative information of the objective function, and often use the second derivative information of the objective function. Such algorithms include quasi-Newton method, Levenberg-Marquardt method and the conjugate gradient method, which can be described as unified

$$\begin{cases} f\left(W^{(k+1)}\right) = \min f\left(W^{(k)} + \alpha^{(k)} S\left(W^{(t)}\right)\right) \\ W^{(k+1)} = W^{(k)} + \alpha^{(k)} S\left(W^{(t)}\right) \end{cases} \tag{7}$$

Among them, the ownership and the threshold values of the network composed of vector $W^{(k)}$, the search direction $S\left(W^{(t)}\right)$by the composition of each component W vector space, $\alpha^{(k)}$is in $S\left(W^{(t)}\right)$so that $f\left(W^{(k+1)}\right)$ in the direction to achieve a very small step. Thus, network optimization method weights can be divided into the following two steps: a) First, determine the current iteration of the best search direction; b) finding the optimal iterative step in this direction. Three methods discussed below, the difference is different from the direction in the search for the best choice.

(1) Quasi-Newton Method.
Quasi-Newton method is a common method for fast optimization, convergence is faster than a ladder degrees. Quasi-Newton method in the search direction improvement over gradient method, it is not only the use of the criterion function gradient point in the search, but also takes advantage of its second derivative matrix, resulting in increased computational complexity. There are typical BFGS algorithm and step tangent algorithm (One step secant, OSS).
BFGS algorithms usually require very little number of iterations will be able to converge, but it each iteration of the computation and memory requirements greater than the conjugate gradient method. Wherein the BFGS algorithm in training small network better, step algorithm is tangent compromise conjugate gradient method and quasi-Newton method, each step required memory and computation are less than BFGS algorithm.

(2) Conjugate Gradient Method.
When using the conjugate gradient vector to determine the conjugate direction, said this algorithm co reels gradient method. 1990 J. Leonard and MA Kramer conjugate gradient method and line search strategies together. In the conjugate gradient method, conjugate direction along the line search, convergence rate than the general gradient descent method is much faster. In a typical training algorithm is the use of learning rate determines the weight and step off the updated values, and in most conjugate gradient algorithm weights each step repeatedly adjust, along with conjugate gradient line search to determine the weights step. Conjugate gradient method may be necessary to calculate or store information on the second derivative has a function of second-order method, compared with the quasi-Newton method, its computational cost is very low, and therefore very useful for large-scale problems.

(3) Levenberg-Marquardt Algorithm.
Marquardt-Levenberg optimization algorithm can be seen as a kind of quasi-Newton approximation method. Conventional BP algorithm correction weight values only when the gradient of the error function of the weights, ie the first derivative information. By the second derivative of the error function were corrected weights, can greatly accelerate the convergence speed of BP network. This is the basic starting point Quasi - Newton Method.

The objective function is assumed to be a network error, so that by adjusting the minimum. By quasi-Newton method, can be expressed as: where is the

455

Hessian matrix for the gradient, if it is assumed as a function of the square and that there are

$$\nabla V(\vec{w}) = J^T(\vec{w})e(\vec{w})$$
$$\nabla^2 V(\vec{w}) = J^T(\vec{w})J(\vec{w}) + S(\vec{w}) \quad (8)$$

On the Gauss-Newton method, $S(\vec{w})$ assuming approximately zero, while Marquardt-Levenberg optimization algorithm further modified Gauss-Newton method, making $\Delta\vec{w} = \left[J^T(\vec{w})J(\vec{w}) + \mu I\right]^{-1} J^T(\vec{w})e(\vec{w})$, where μ is a scalar. It depends on μ, between the gradient descent algorithm ($\mu \to \infty$) and quasi-Newton method ($\mu \to 0$) in two extreme cases change the optimization.

5 CONCLUSION

Research results show that the comprehensive method than the standard BP algorithm in terms of the number of iterations, the training time and the mean square error and the like, of additional momentum and adaptive outstanding performance parameter method.

As seen above, Marquardt-Levenberg algorithm is dynamically adjusting the damping factor, is changed according to the results of the iteration convergence direction, so as to achieve the purpose of decreasing the error. The key step of the algorithm is to calculate the Jacobian matrix. Its advantage is that when there are fewer number of network weights converge very quickly, you can make a shorter study time in practical applications. Research show that the Marquardt-Levenberg algorithm neural network convergence fastest training times at least.

REFERENCES

Li chungui etc. Traffic Flow forcasting Algorithm using Simulated Annealing Genetic BP Network. *Proceedings of 2010 International Conference on Measuring Technology and Mechatronics Automation* 2010.

Shan Li, Haibing Chen, JunXian Yun etc. Optimization for Cyclosporine Blood Concentration Prediction Based on Genetic Algorithm-BP Neutal Network. *Proceedings of Second International Conference on Genetic and Evolutionary Computing.* 2008.

Yu-min Chiang, Huei-min Chiang, Shang-yi Lin. The Application of Ant Colony Optimization for Gene Selection in Microarray-based Cancer Classification. *Proceedings of Seventh International Conference on Machine Learing and Cybernetics.* 2008.

Li Wang, Dong-qing Wang, Ning Ding. Research on BP Neural Network Optimal Method Based on Improved Ant Colony Algorithm. *IEEE Computer Society Proceedings of 2010 Second International Conference on Computer Engineering and Applications.* 2010.

Yan Zhao, Zhongjun Xiao, Jiayu Kang. Optimization Design Based on Improved Ant Colony Algorithm for PID Parameters of BP Neural Network. *Proceedings of 2010 2nd International Asia Conference on Informatics in control, Automation and Robotics.* 2010.

Yuxiang Shao, Qing Chen. Application Ant Colony Neural Network in Lithology Recognition And Prediction:Evidence from China. *Proceedings of 2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Applications.* 2008.

Wu Yuguo, Song Chongzhi. Design Optimization Based on Neural Networks and Ant Colony Optimization. *Proceedings of 2007 Second IEEE Conference on Industrial Electronics and Applications.* 2007.

Yu-min Chiang, Huei-min Chiang, Shang-yi Lin. The Application of Ant Colony Optimization for Gene Selection in Microarray-based Cancer Classification. *Proceedings of Seventh International Conference on Machine Learing and Cybernetics.* 2008.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

456