第44卷 第2期　　复 旦 学 报 (自然科学版)　　Vol.44 No.2
2005年4月　　Journal of Fudan University (Natural Science)　　Apr. 2005

Article ID: 0427-7104(2005)02-0318-05

PCA Approach to BP Learning

RUAN Qing, WANG Yi-qiang

(Institute of Mathematics, Research Center and Laboratory of Mathematics for Nonlinear Science, Fudan University, Shanghai 200433*, China)*

Abstract: The idea of the PCA to solve two major problems in BP algorithm is adopted. One is to decide the number of the neurons in the hidden layer ,and the other is to choose the initial values of parameters. 58 subjects offered by Wuhan Tongji Hospital were chosen as training smaples. Through some experiments of comparison ,it is concluded that the improved algorithm is effective and efficient.

Keywords: neural network; BP algorithm; PCA

CLC number: O 29　　**Document code:** A

Nowadays there are many kinds of neural networks which are applied in different fields to solve different problems effectively. Among them ,a backprogpagation (BP) network is the most commonly used. A BP network is a multi-layer feedforward network that adopts the BP algorithm in its learning[1] . After this algorithm was put forward ,more and more researchers were attracted by it. For example ,in 1997 ,it was used to classify the totally unconstrained handwriting numerals[2] . The result showed that the BP network produced 97. 35 % of the recognition rates ,which were better than those of several previous methods reported in the literature on the same database.

Principal component analysis (PCA) is one of the most important methods in multivariate analysis. We adopt a concept of information to describe and to simplify a data matrix ,which is represented by some m-dimensional vectors ,by some optimization technique according to an algebra or a geometry criterion. On knowing the eigenvectors ,one can transform the vectors into the eigenbasis ,i. e. the components of the new vectors are the projections of the old ones onto the eigenvectors. If one decides that the p most relevant eigenvectors are enough to describe the signal ,one just truncates the vectors after the p-th component. Thus one has an p-dimensional embedding of the data[3] .

Here we apply the idea of PCA to BP algorithm to improve the training procedure of a neural network. We take data of 58 HRV samples given by Wuhan Tongji Hospital as an example. In a former research[4] ,it was concluded that it was enough to use four or five neurons in the hidden layer to train the neural network , and now we bring forward the verification of this conclusion. Furthermore ,the problem of choosing initial values in BP algorithm is settled by computation of eigenvectors in PCA ,which effectively shortens the training time of neural network and helps it converge to some better extrema.

1 Backpropagation algorithm

A feedforward neural network consists of an input layer ,an output layer and several hidden layers. Each layer consists of multiple neurons that are connected to all the neurons in adjacent layers. In our research ,the number of the hidden layer is set as one.

Received date: 2003-10-22

Biography: RUAN Qing(1980 —) ,male ,graduate student.

This paper was recommended by Prof. Ruan Jiong and Prof. Gu Fanji.

Before a feedforward network is applied ,it should be trained with a set of subjects. The aim of training is to study the regularity of the variability of a given data set through adjusting the parameters of the neural network (connection weights and node biases) . The parameters will be adjusted iteratively by a process of minimizing the forecast errors. For each training iteration ,all the input vectors included in the training set are submitted to the imput layer in order. The output of each neuron is propagated forward through each layer of the network ,using the following equations

$$\begin{cases} y_l = f\left(\sum_{k=0}^{n_2-1} w'_{kl} x'_k - \theta'_l \right) & l = 0,1,\cdots,m-1, \\ x'_k = f\left(\sum_{i=0}^{n_1-1} w_{ik} x_i - \theta_k \right) & k = 0,1,\cdots,n_2-1, \end{cases} \tag{1}$$

where x_i is an input data from the ith neuron in input layer to neurons in hidden layer ; x'_i is the output of the ith neuron in the hidden layer ; w'_{kl}, w_{ik} are the connection weights ; θ'_l, θ_k are the thresholds of neurons in the hidden layer and the output layer ,respectively. The transformation f is expressed as follows ,

$$f(x) = \frac{1 - \mathrm{e}^{-x}}{1 + \mathrm{e}^{-x}}, \tag{2}$$

which satisfies $f'(x) = \frac{1 - f^2(x)}{2}$. Denoted by θ'_l, θ_k the weights $w'_{n_2 l}$ and $w_{n_1 k}$ respectively , t_l the desired activity of the output unit. Thus ,the total error function can be expressed in terms of w and w',

$$E_{\text{total}} = \frac{1}{2} \sum_{P_1=1}^{P} \sum_{l=0}^{m-1} (t_l^{P_1} - y_l^{P_1})^2, \tag{3}$$

where P is the number of subjects in training set.

We use gradient descent algorithm gradient method to adjust the parameters and let the value of E_{total} decrease. First ,select the initial values of w and w' randomly and input the P subjects into the networks. Secondaly ,calculate the values of $x'^{P_1}_k$, $y^{P_1}_l$ and the errors of every unit $\delta^{P_1}_{kl}$, $\delta^{P_1}_{ik}$ with

$$\begin{cases} \delta_l^{P_1} = (t_l^{P_1} - y_l^{P_1})(1 - (y_l^{P_1})^2)/2, \\ \delta'^{P_1}_k = \sum_{l=0}^{m-1} \delta^{P_1}_{kl} w'_{kl} (1 - (x'^{P_1}_k)^2)/2. \end{cases} \tag{4}$$

Thirdly ,we modify the weights and the threshold values with

$$\begin{cases} w'_{kl}(n+1) = w'_{kl}(n) - \eta \frac{\partial E_{\text{total}}}{\partial w'_{kl}}, \\ w_{ik}(n+1) = w_{ik}(n) - \eta \frac{\partial E_{\text{total}}}{\partial w_{ik}}, \end{cases} \tag{5}$$

where

$$\begin{cases} \frac{\partial E_{\text{total}}}{\partial w'_{kl}} = \sum_{P_1=1}^{P} \delta_l^{P_1} x'^{P_1}_k, \\ \frac{\partial E_{\text{total}}}{\partial w_{ik}} = \sum_{P_1=1}^{P} \delta'^{P_1}_k x'^{P_1}_i, \end{cases} \tag{6}$$

and η is the step length ,which will be changed during the training. We should go on repeating the last two processes above until E_{total} is smaller than the value we set initially.

Because the gradient descent algorithm is adopted in the BP algorithm ,it cannot be avoided that the result of the training will converge to a local extremum. And also ,the training is probably inefficient because of the ill parameters. Thus ,the selection of the parameters is a very important step in the training ,for instance ,the number of neurons in the hidden layer. If there are too many neurons in the hidden layer of a network ,it will take too much time to train the network. But if there are too few neurons, E_{total} will not converge to a gratifying extremum. Moreover, how to choose the initial values of w' and w is also worthy of investigating. Some suitable initial value will save a lot of time in training and help the network converge to a better extremum. In order to avoid some of these disadvantages, we design an optimized algorithm to amend the traditional algorithm.

2 Principal component analysis

The major task of PCA is to summarize the information reflected by the original data matrix we get, which is always large, through a certain optimum method, and thus to reveal its major structure by simplifying the data matrix and decreasing its dimensions. Meanwhile, we put forward some reasonable explanations about the information provided by the data matrix in order to solve the problems we are facing.

An idea of fitting data points is introduced in 5 as follows. Su ppose X is a data matrix measured in n samples on p indexes, i.e. X is an n by p matrix. Let $x_1, \cdots, x_n$ stand for n subjects; $x_{(1)}, \cdots, x_{(p)}$ stand for p indexes, then

$$X = \begin{bmatrix} x_1' \\ \cdots \\ x_n' \end{bmatrix} = (x_{(1)}, \cdots, x_{(p)}). \tag{7}$$

If there are some obvious structure relations in these n points and they are basically in a k-dimension space, our task is to find the k-dimension space and n points in it, which are closest to the original points. These n k-dimensional points can reflect n original p-dimensional points, while they lose as little information as possible and show the structure of the original data matrix. Geometrically, the problem is to search in R^n for k orthogonal vectors. These vectors can produce a k-dimension space on which the projecgtions of n p-dimensional points are closest to the original points. Thus the aim of simplifying the date matrix is reached.

We begin from searching for a 1-dimension space of this kind, that means we search for a straight line through the origin on which the projections of the n points in R^n is closest to the original points. The unit vector u_1 stands for the direction of the line l, then the projections of $x_1, \cdots, x_n$ on l is $x_1' u_1, \cdots, x_n' u_1$, the n components of matrix Xu_1. We adopt the least square method as fitting standard, that is to search for u_1 which makes the error square sum

$$\| x_1 - (x_1' u_1) u_1 \|^2 + \cdots + \| x_n - (x_n' u_1) u_1 \|^2 \tag{8}$$

reach its minimum value. The problem equals to searching for $u_1 (u_1' u_1 = 1)$ which makes the quadratic form $u_1' X' X u_1$ reach its maximum. It is easy to prove that u_1 is the standard eigenvector corresponding to λ_1, the largest eigenvalue of $X'X$, and $u_1' X' X u_1 = \lambda_1$.

Similarly, suppose that eigenvalues of $X'X$ are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and the corresponding standard orthogonal vectors are $u_1, u_2, \cdots, u_p$, then the best k-dimension space to fit these data points is the space spanned by $u_1, u_2, \cdots, u_k$. $Xu_j = (x_1' u_j, \cdots, x_n' u_j)'$ is named the j-th principal component of p indexes $x_{(1)}, \cdots, x_{(p)}$, $j = 1, 2, \cdots, p$.

We denote p principal components of $x_{(1)}, \cdots, x_{(p)}$ by $w_{(j)} = Xu_j$, $j = 1, 2, \cdots, p$. In matrix, we have

$$W = (w_{(1)}, \cdots, w_{(p)}) = XU, \tag{9}$$

where $U = (u_1, \cdots, u_p)$.

Since $W'W = U'X'XU = U'(U\Lambda U')U = \Lambda$, p principal components are orthogonal to each other, and the norm of the i-th principal components is $\sqrt{\lambda_i}$.

Also, because of the orthogonality of the matrix $U = (u_{ij})$, we have

$$X = XUU' = WU' \tag{10}$$

that is
$$\mathrm{x}_{(i)} = \sum_{j=1}^{p} u_{ij}\mathrm{w}_{(j)} \qquad i = 1,2,\cdots,p.$$

This shows that we can restore the original data by some linear combinations of principal components vectors. If some part of the linear combinations can approximately denote the original data, for some definite $k(1 \le k \le p)$, we have
$$\mathrm{x}_{(i)} \approx \sum_{j=1}^{k} u_{ij}\mathrm{w}_{(j)} \qquad i = 1,2,\cdots,p.$$

Let $\alpha_k := \sum_{i=1}^{k}\lambda_i$ and $\tau_k := \sum_{i=1}^{k}\lambda_i \Big/ \sum_{i=1}^{p}\lambda_i$ be total variance contribution to X and variance contribution ratio, respectively, of $w_{(1)}, \cdots, w_{(k)}$. Since tr $X'X = \sum_{i=1}^{p}\lambda_i$, the total variance of X is fixed, the former one reflects the absolute contribution of how well the subspace can fit the original data points, while the latter one reflects the relevant contribution ratio, i. e. the importance of the principal components. The larger the ratio is, the better the fitting effect is. In practice, we always hope k is as small as possible while τ_k is as large as possible.

In order to simplify the computation, a usual way is as follows. When the first k principal components occupy more than 85 % of the total information, i. e. k is chosen based on the variance contribution ratio $\tau_k >$ 85 %, and then we just analyze these k principal components.

3 The improvement of BP algorithm

Now we apply the idea of PCA to the BP network training to improve the algorithm. The improved algorithm settles two problems in training a BP network: to choose the number of the neurons in the hidden layer and to choose the initial values of the parameters (connection weights and node biases). In the following experiment, we take a series of HRV data given by Wuhan Tongji Hospital as the training sample. The data contains 58 HRV subjects, each of which contains 44 indexes (valued between - 1 to 1 after standardized), and a clinical diagnose (valued 1 or - 1 : 1 shows the subject has the cardiovascular disease while - 1 the opposite). We train a three-layer feedforward network to study these subjects. In the network there are 44 neurons in the input layer, corresponding to the 44 indexes, and there is one neuron in the output layer. By training the network, we hope to get a result close to the clinical judgement when 44 indexes of any subject are input.

We consider the information transfer from the input layer to the hidden layer with $x'_k = f\left(\sum_{i=0}^{n_1-1} w_{ik}x_i - \theta_k\right)$. To each neuron in the hidden layer, the input value is a linear combination of the n_1 indexes of the subject. In view of the whole network, the procedure from the hidden layer to the output layer is actually to identify and classify the subjects by these "linear combinations". Thus, a "reasonable" neural network must reserve the useful information and remove the redundant information reflected by the original indexes in the hidden layer or in these "linear combinations". On the other hand, the "linear combination" just coincides with the PCA. Therefore, a natural idea is to choose the parameters according to the PCA. In Section 2 we have introduced a method by which the data were transformed from the original high dimensional index space to the low dimensional principal component space. We use the method to choose the least k which satisfies that the error contribution ratio τ_k is more than 95 %. In our opinion, the first k principal components are enough to contain the information reflected by the original n_1 indexes. In other words, it is enough to reach a satisfactory training result when there are k neurons in the hidden layer.

A former research[4] shows, with the same data used, that 3 to 4 neurons in the hidden layer are enough to get a satisfactory study result. This result is shown in Tab. 1. Here by our calculation the result is $k = 4$, which supports the result obtained in [4]. Furthermore, we found that the lar gest eigenvalue, 439. 616 7, occupies

81.7 % of the total variance. This can just explain the reason why we can also get a good study effect when there 's only one neuron in the hidden layer.

After we decide the construction of the network ,another problem is to choose the initial values of the parameters (connection weights and node biases). Since the first k principal components bear most information reflected by the original data ,we choose the eigenvectors corresponding to the first k eigenvalues of the relevant matrix $R = X'X$ as the initial values of the weights from the input layer to the hidden layer.

Tab. 1 The relation between the error and the number of the neurons in hidden layer

Neuron	Error
1	4.611 1
2	0.112 8
3	0.111 7
4	0.111 7

In the traditional BP algorithm ,the initial values of the parameters are chosen randomly. We design an experiment to indicate the advantage of the improved algorithm. We take the data given by Wuhan Tongji Hospital as the training sample. According to the former analysis ,we choose the number of the neurons in the hidden layer as 4 ,and other parameters are chosen by two methods below : (1) all weighs and threshold values are chosen randomly ; (2) we take the first k eigenvectors of R as the initial values of the weights from the input layer to the hidden layer ,and all the rest parameters are retained from (1) .

We take 10 groups of the neural network to compare the training iteration times and the error of the convergence after training the network by BP algorithm. The result is shown in Tab. 2.

It clearly shows that the approved BP algorithm is more efficient and more accurate than the traditional BP algorithm. The comparison of the standard deviation also indicates that the improved algorithm is more stable.

We apply the idea of PCA to the traditional BP algorithm. Through the experiments ,we discover that the algorithm is approved not only in the study efficiency but also in the study effect. More important , some opinions are brought forward to decide the construction of the network , i. e. the number of the neurons in the hidden layer. Compared with the result of the method in 4 , this method proves to be more applicable.

Tab. 2 The comparison of the two methods

Group	Method(1)		Method(2)	
	Steps	Error	Steps	Error
1	18 847	0.032	15 693	0.076
2	11 397	0.023	14 868	0.033
3	57 759	0.036	14 600	0.033
4	39 612	0.027	15 267	0.031
5	7 724	0.038	14 100	0.096
6	24 729	0.030	14 051	0.071
7	9 505	0.576	14 536	0.100
8	14 174	0.026	14 600	0.081
9	19 249	0.032	14 771	0.087
10	17 766	0.038	15 827	0.052
Average	22 076.2	0.086	14 831.3	0.066
SD *	14 715	0.163	571.544 6	0.025

* *SD* is the standard deviation of each column.

References :

1 Werbos P J. The roots of backpropagation : From ordered derivatives to neural networks and political forecasting M . New York :John Wile y & Sons ,1994.

2 Cho S B. Neural-network classifiers for recognizing totally unconstrained hand written numerals J . *IEEE Trans on Neural Networks* ,1997 ,**8** :43-53.

3 Kantz H ,Schreiber T. Nonlinear time series analysis M . Cambrid ge :Cambridge University Press ,1997.

4 Ruan J ,Ruan Q ,Wu C. Neural network in signal analysis of heart rate variability A . In : Wan g L ,eds. Proceedings of the 9th International Conference on Neural Information Processing (ICONIP '02) Vol. 4 C . Sin gapore : IEEE Press ,2002. 2037-2040.

5 Joliffe I T. Principal component analysis M . New York : S pringer Verlag ,1986.

(下转第327页)

2 Friedrichs T. Dirac operator in Riemannian geometry M . New York:A M S ,2000.

3 Baum H. An upper bound for the first eigenvalue of Dirac operator on compact spin manifold J . *Math Zeitschrift* ,1991 ,**206** :409-422.

4 Zhang X. Lower bounds for eigenvalues of hypersurface Dirac operators J . *Math Res Lett* ,1998 ,**5** :199-201.

5 Chavel I. Eigenvalues in Riemannian geometry M . New York:Acadamic Press ,1984. 15 - 25.

Estimate of Dirac-Laplace Operator 's Eigenvalue Gap

YAN Yin-hui

(*Institute of Mathematics , Fudan University , Shanghai* 200433 , *China*)

Abstract : With the help of the general Bochner formula and Rayleigh theorem ,the estimate of Dirac-Laplace operator 's eigenvalue gap is obtained for the case that M is an n-dimensional compact Spin-submanifold of the unit sphere of dimension $n + p$. And for the special case that M is a compact minimal Spin-submanifold ,the estimate of eigenvalue gap is given.

Keywords : Dirac-Laplace operator ; eigenvalue ; Spin-submanifold

(上接第 322 页)

主成分分析方法在 BP 学习中的应用

阮 庆 ,王逸蔷

(复旦大学 数学研究所 非线性数学模型与方法开放实验室研究中心 ,上海 200433)

摘 要: 用主成分分析的思想解决 BP 算法中的两个问题. 一是隐层中神经元的个数 ,另一个是训练的初始参数. 为了便于比较 ,采用来自武汉同济医院的 58 个样本作为学习对象. 通过实验比较得知 ,改进后的算法不仅节省了训练时间 ,而且能够得到更好的学习效果.

关键词: 神经网络; BP 算法; 主成分分析

中图分类号: O 29 **文献标识码**: A