一种优化 BP 神经网络训练样本的方法*

刘 彩 红
(重庆师范大学 数学与计算机科学学院, 重庆 400047)

摘 要:BP 神经网络训练样本的选取对网络的泛化能力有较大的影响,特别,怎样从高维大样本数据中选取合适训练样本是一个难点。本文运用因子分析法对大样本数据进行预处理,再利用分析所得的公因子进行聚类分析,这样既可以降低指标的维数,也可以减少样本的数量。实验证明,该方法简化了网络结构、加快了网络的收敛速度,对提高网络的泛化能力有一定的帮助。

关键词:BP 神经网络;因子分析;聚类分析
中图分类号:TP183;O212.1 **文献标识码**:A **文章编号**:1672-6693(2007)03-0051-03

An Approach to Optimizes the Training Samples of BP Neural Network

LIU Cai-hong
(College of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047, China)

Abstract: The selection of BP neural network training samples has a strong impact on the generalization ability of the network, and the selection of training samples from the high-dimensional data is especially difficult. The paper uses the method of factor analysis to pretreat the data from a large samples, and then cluster the result. In this way, both of the dimensions of variables and the number of samples can be reduced. The result of the experiment indicates that the method can improve the convergence speed and simplify the network structure as well as improve the network generalization.

Key words: BP Neural Network; Factor Analysis; Clustering Analysis

神经网络(Neural Network,NN)的泛化能力或推广能力,是指神经网络在训练完成以后输入其训练样本之外的新数据时获得正确输出的能力。所以,泛化能力是神经网络最主要的性能,没有泛化能力的神经网络没有任何使用价值。如何提高 NN 的泛化能力一直是该领域研究者所关注的问题。而神经网络的结构复杂性和样本复杂性是影响神经网络泛化能力的主要因素,所以,人们的研究大都集中于此。例如,正则化[1]、网络集成[2]、输入模糊化[3]、结构优化[4]、PCA 方法对样本预处理[5]等,这些方法不同程度地提高了 NN 的泛化能力。但总体上说,NN 的泛化仍然是一个没有解决或没有完全解决的问题。

在构建网络之前,唯一所知的只有训练样本数据,Partridge[6] 对用于分类的三层 BP 网的研究发现,训练集对泛化能力的影响甚至超过网络结构(隐节点数)对泛化能力的影响。文献[5]利用主成分分析法(Principal Components Analyze,PCA)对训练样本进行预处理,这样既可降噪又可降维,降维对减小网络结构提高网络泛化能力有利。然而,在实际应用中,样本的个数远远大于指标的个数,并且当遇到样本数据很大时,人们往往随机选取其中一部分作为训练样本,这样很可能使训练样本集本身没有包含全部样本的特征,使预测的结果出现较大的误差。文献[7]利用模糊聚类法首先对样本进行分类,然后再从每一类中按一定比例选择训练样本。但是一般聚类算法只擅长处理低维的数据,对高维数据的聚类质量则较差,就会使最终所选取的训练样本有偏差,从而影响到最后预测结果的精度。本文运用因子分析对高维大样本数据先进行预处理,再利用分析所得的公因子进行聚类,这样除了降噪降维外,还可以从大样本集中选出几乎可以包含全

* 收稿日期:2007-01-05 修回日期:2007-04-13
资助项目:重庆市教委项目(No. KJ060818;No. KJ060804)
作者简介:刘彩红(1980-),女,陕西宝鸡人,硕士研究生,研究方向为人工神经网络及其应用。

部样本特性的训练样本,从而可以解决上述问题。

1 因子分析

因子分析(Factor Analysis,FA)就是在处理多指标样本数据时,将具有错综复杂关系的指标(或样品)综合为数量较少的几个因子,以再现原始变量与因子之间的相互关系。是一种应用广泛的多变量统计分析方法,常用来做降维处理。具体讨论参见文献[8],下面仅给出用于BP网络训练的因子分析法的主要步骤。设样本总体为 $X'=(x_{ij})_{n\times p}$(其中 x_{ij} 为要考察的样本 x_i 的第 j 个指标,$i=1,2,\cdots,n$;$j=1,2,\cdots,p$)。

(1)原始数据样本的标准化。实际应用中,数据指标的量纲、数量级往往不同,所以在计算之前先消除这些因素的影响,而需将原始数据标准化。标准化方法可以采用零均值标准差标准化方法,(若进行了这一步,则对以后BP网络训练输出的结果需要进行相反的处理过程,即将输出值还原为原量纲值)。标准化后的样本记为 X_{np}。

(2)建立指标的相关系数阵 $R=(r_{ij})_{p\times p}$,并求解 R 的特征根及相应的单位特征向量。

对标准化后的值计算指标之间的相关系数,得到指标的相关系数阵 R,求解出它的 p 个特征根及相应的单位特征向量,分别记为 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p>0$ 和 $\mu_1,\mu_2,\cdots,\mu_p$。根据累计贡献率的要求比如 $\sum_{i=1}^{q}\lambda_i/\sum_{i=1}^{p}\lambda_i\geq 85\%$,取前 q 个特征根及相应的特征向量得到因子载荷阵

$$A=\begin{bmatrix}\mu_{11}\sqrt{\lambda_1} & \mu_{12}\sqrt{\lambda_2} & \cdots & \mu_{1q}\sqrt{\lambda_q}\\ \mu_{21}\sqrt{\lambda_1} & \mu_{22}\sqrt{\lambda_2} & \cdots & \mu_{2q}\sqrt{\lambda_q}\\ \cdots & \cdots & \cdots & \cdots\\ \mu_{p1}\sqrt{\lambda_1} & \mu_{p2}\sqrt{\lambda_2} & \cdots & \mu_{pq}\sqrt{\lambda_q}\end{bmatrix}$$

(3)对因子载荷阵施行因子旋转,并计算因子得分。因子旋转的方法有多种,如正交旋转、斜交旋转等,这里采用方差最大正交旋转。旋转后的因子载荷阵记为 $B_{qp}=A'_{qp}R_{pp}^{-1}$,因子得分的计算公式为 $F_{nq}=X_{np}B'_{qp}$。

2 聚类分析

聚类分析(Clustering Analysis)就是根据研究对象的特征把性质相近的个体归为一类,使得同一类中的个体具有高度的同质性,不同类之间的个体具有高度的异质性的多元分析技术的总称。在应用中先进行因子分析来降低数据的维数,产生新的不相关变量,然后把这些变量作为聚类变量进行聚类分析[8]。

由因子分析所得的因子得分计算各样品间的欧式距离,来进行聚类分析。欧式距离公式为

$$d_{ij}=\left(\sum_{a=1}^{p}(x_{ia}-x_{ja})^2\right)^{1/2}$$

n 个样本通过聚类,被分成 $X_1,X_2,\cdots,X_L$,共 L 个小类,视 n 的大小,以一定比例选取 $x_1,x_2,\cdots,x_m$($m<n$)作为训练样本集。这样对原来的大样本集,通过聚类并选择后,科学地选取适量的训练样本,减少了样本数同时又保证训练样本具有代表性。一般聚类算法只擅长处理低维的数据,而因子分析之后得到的公因子数少于原指标数,对指标进行了降维处理,可使聚类结果更精确可靠。

通过以上分析处理,可使原始大样本数据 X_{np} 优化为训练数据 F_{mq}。

3 对训练样本优化的实现

用本文提出的方法对训练样本进行优化,通过以下步骤实现。

第1步 先对原始数据样本作因子分析,将样本的多个指标综合为数量较少的几个因子,并得到因子得分。

第2步 用因子得分计算各样品间的欧式距离,来进行聚类分析,将样本分成若干小类。

第3步 从每小类中按一定比例科学地选取适量样本。

经过以上分析处理,就得到了优化的训练样本集。

4 应用实例

这里使用的数据集来自于UCI Repository of Machine Learning Databases中的Wine Recognition Database。该数据集由Stefan Aeberhard提供,他使用化学分析的方法确定甜酒之源(the Origin of Wine)。Wine数据集有178个例子,每例包含13个酒的属性,178例共分为3类,第1类有59例,第2类有71例,第3类有48例。实验中将Wine数据集分为两个子集 P 和 Q,P 集的3类例子数分别为40、50和36用作原始训练集,Q 集的3类例子数分别为19、21和12用作仿真测试。实验是对3层BP网络分别用 $P1$ 集、$P2$ 集和 $P3$ 集作训练样本集,Q 集作测试集,设定同一训练函数,同一训练误差,来对比它们达到目标时的训练次数和对 Q 集的错识率。

用本文提出的方法对 P 集进行分析处理后得到训练集 $P1$,简单步骤如下。

(1)因子分析。用SPSS11.5软件对Wine数据集作因子分析,KMO统计值为0.779,大于0.52说明这13个属性是相关的,Wine数据集适合作因子分析。提取特征值大于0.5的因子(注:特征值在某

种程度上可以被看成是表示公因子影响力度大小的指标），得到 7 个公因子，其累计方差贡献率达到了 89.337%。训练网络时，用这 7 个公因子作为网络输入。

（2）聚类分析。用因子分析所得的 7 个公因子作聚类分析，聚类的数目为 5 类，距离测量方法为欧氏距离。

对聚类分析得到的 5 类数据按其前几个公因子值排序，选取每类中公因子值较大的一部分（本例选取 50%）构成最终的训练样本集。对 Wine 数据集的 3 个子集分别作聚类分析，再优选，最后得到 $P1$ 集。

$P2$ 集的构造：对 P 集作因子分析，然后从其 3 个子集中随机选取 50% 的样本构成训练集 $P2$。

$P3$ 集的构造：分别从 P 集的 3 个子集中随机选取 50% 的样本，就构成了训练集 $P3$。

$P1$ 集、$P2$ 集、$P3$ 集都是从 P 集的 3 个子集中选取 50% 的样本，所以，它们的 3 类例子数分别都为 20、25、18。

用 $P1$、$P2$ 做训练集时输入层为 7 个节点（因子分析得到的 7 个公因子），用 $P3$ 做训练集时输入层为 13 个节点（酒的 13 个属性），输出层都为 3 个节点。训练采用 LM 算法，训练误差取 0.001，隐含层节点分别取 6、8、10。对训练好的网络，用 Q 集作仿真测试。由于初始化方法的随机性，使 BP 网络的工作结果也是相应变化的，下列表中的数据是 20 次运行结果的平均值。

表 1　隐节点取 6 时的实验结果

训练集	训练次数	错识率
$P1$	9.7	1.8/52
$P2$	10.2	3.35/52
$P3$	27.3	4.8/52

表 2　隐节点取 8 时的实验结果

训练集	训练次数	错识率
$P1$	8.4	2.1/52
$P2$	9.25	3.75/52
$P3$	21.9	5.6/52

表 3　隐节点取 10 时的实验结果

训练集	训练次数	错识率
$P1$	9	1.7/52
$P2$	8.35	4/52
$P3$	12.8	5.75/52

从训练次数项可以看到，训练样本集为 $P3$ 时的训练次数远多于训练样本集为 $P1$ 和 $P2$ 时。而且在实验过程中，以 $P3$ 做训练样本集进行训练时，超出最大训练次数而没有收敛的次数也较多，由此可以说明，对训练样本集运用因子分析，可以加快网络的收敛速度。从错识率项可以看到，以 $P1$ 做训练样本集时的错识率最小，也就是泛化能力最好。$P1$ 集和 $P2$ 集的不同在于，对 $P1$ 集作了聚类分析，这样就保证了在对样本数量降维的同时，还可以使选出的样本几乎包含全部样本的特性，这些有助于网络泛化能力的提高。在用 $P1$、$P2$ 做训练集时输入层为 7 个节点，用 $P3$ 做训练集时输入层为 13 个节点，所以说，该方法简化了网络结构。由此可知，本文提出的方法可以简化网络结构，加快网络的收敛速度，提高网络的泛化能力。

本实验采用的 Wine 数据集有 178 个例子，每个例子有 13 个属性，数据量不算多。实际应用中的数据量远远多于此，若是对高维大样本数据用本文的方法优化训练样本，效果会更明显。

5 结论

本文运用因子分析对高维大样本数据进行预处理，再利用其结果进行聚类分析，这样即可以降低指标的维数，也可以减少样本的数量，并且最终的训练样本还可包含几乎全部样本的信息，从而达到对样本和指标的降维。实验证明，该方法简化了网络结构，加快了网络的收敛速度，对提高网络的泛化能力有一定的帮助。

参考文献：

[1] HINTON G E. Connectionist Learning Procedures [J]. Artificial Intelligence, 1989, 40: 185-234.

[2] HANSEN L K, SALAMON P. Neural Network Ensembles [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(10): 993-1001.

[3] ISHIBUCHI H, NII M. Fuzzification of Input Vector for Improving the Generalization Ability of Neural Networks [C]. Alaska: The Int '1 Joint Conf on Neural Networks, Anchorage, 1998.

[4] 唐万梅. BP 神经网络网络结构优化问题的研究[J]. 系统工程理论与实践, 2005, 10(10): 95-100.

[5] 陈小前，罗世彬，王振国，等. BP 神经网络应用中的前后处理过程研究[J]. 系统工程理论与实践, 2002, 22(1): 65-70.

[6] PARTRIDGE D. Network Generalization Differences Quantified. Neural Networks, 1996, 9(2): 263-271.

[7] 何勇，项利国. 基于模糊聚类的 BP 神经网络模型研究及应用[J]. 系统工程理论与实践, 2004, 24(2): 79-82.

[8] 袁志发，周静芋. 多元统计分析[M]. 北京：科学出版社, 2002.

[9] 吕佳. 可能性 C-Means 聚类算法的仿真实验[J]. 重庆师范大学学报（自然科学版）, 2005, 22(3): 129-132.

（责任编辑　游中胜）