

基于结点敏感度的单隐含层前馈神经网络结构选择

翟俊海^{1,2} 哈明光¹ 邵庆言¹ 王熙熙^{1,2}

(河北大学数学与计算机学院 保定 071002)¹ (河北省机器学习与计算智能重点实验室 保定 071002)²

摘要 提出了一种基于结点敏感度的单隐含层前馈神经网络结构选择方法。该方法从一个隐含层结点数较多的网络开始,首先利用结点敏感度度量隐含层结点的重要性,然后按重要性对隐含层结点由大到小排序,最后逐个剪去不重要的隐含层结点,直到满足预定义的停止条件。该算法的特点是不需要重复训练神经网络,得到的网络结构紧凑,具有较高的泛化能力。在实际数据集和 UCI 数据集上的实验结果显示,提出的算法是行之有效的。

关键词 前馈神经网络,结构选择,敏感度,交叉熵

中图法分类号 TP181 文献标识码 A

Architecture Selection for Single-hidden Layer Feed-forward Neural Networks Based on Sensitivity of Node

ZHAI Jun-hai^{1,2} HA Ming-guang¹ SHAO Qing-yan¹ WANG Xi-xi^{1,2}

(College of Mathematics and Computer Science, Hebei University, Baoding 071002, China)¹

(Hebei Province Key Lab. of Machine Learning and Computational Intelligence, Baoding 071002, China)²

Abstract Based on sensitivity of node, an architecture selection for Single-hidden Layer Feed-forward Neural Networks (SLFNNs) was proposed. Beginning from an initial large number of hidden nodes, the proposed algorithm firstly employs the sensitivity to measure the significance of the hidden nodes, and then the hidden nodes are sorted in descending order by their significance, finally all unimportant nodes are pruned. The algorithm will terminate when a predefined stop condition is held. The main features of the proposed algorithm include the unnecessary of retraining the SLFNN, the compact architecture and the high generalization capacity. We experimented the proposed approaches on real world datasets and UCI datasets, and the experimental results show that the proposed method is effective and efficient.

Keywords Feed-forward neural network, Architecture selection, Sensitivity, Cross-entropy

1 引言

神经网络已广泛应用于函数逼近^[1]、模式识别^[2]、时间序列预测^[3]等多种领域,神经网络的性能与网络结构、损失函数及连接权密切相关^[4]。针对具体应用,如何选择合适的网络结构是机器学习领域的研究热点,神经网络的结构选择就是通过选择隐藏层的层数和隐藏层的结点数来确定网络结构。如果网络结构过于复杂,那么网络容易训练,也能较好地拟合训练数据,但是由于过拟合使得神经网络的泛化能力较低,训练神经网络的时间过长;相反,如果网络结构过于简单,虽然可以缩短训练神经网络的时间,但是网络可能没有一致逼近能力^[5]。

针对神经网络结构选择问题,很多学者开展了深入的研究,提出了一系列的网络结构选择算法,这些算法大致可分为两类:增量方法^[6]和剪枝方法^[7]。增量方法也称为构造方法,这种方法从一个规模较小的网络开始,逐渐增加网络隐含层和隐含层结点,直到满足预定义的停止条件。代表性的工作包括:Redding 等人于 1993 年提出的基于隐含层单元扇入思

想的高阶网络构造方法^[8];Tsoi 等人于 1997 年提出的基于递归技术的神经网络构造方法^[9];Liu 等人于 2002 年提出的基于增量训练的前馈神经网络构造方法^[10];Subirats 等人于 2012 年提出的基于神经元之间竞争思想的神经网络构造方法^[11];Zhang 等人提出的基于 ELM (Extreme Learning Machine) 的自适应隐含层结点增长的 SLFNNs 构造方法^[12]。

增量方法的缺点是大多都需要凭经验预先确定前馈神经网络的最大隐含层数及隐含层结点数。而这两个数字的精确确定往往比较困难,若设置过大,则导致学习时间过长;若设置过小,有可能得不到最优的网络结构。

剪枝方法首先从一个规模较大的网络开始,然后利用某些度量方法剪去那些对网络性能不利的或者贡献小的结点或连接权,达到压缩网络结构的目的。代表性的工作包括:Karnin 提出的基于全局误差函数敏感性分析的连接权剪枝算法^[13];Hagiwara 提出的基于“Consuming energy”和“Weights power”的剪枝算法,该算法可同时对连接权和隐含层结点进行剪枝^[14];Castellano 等提出的迭代消除隐含层结点并调整连接权的剪枝算法^[15];Zeng 等提出的基于量化敏感性度量

到稿日期:2013-05-20 返修日期:2013-07-22 本文受国家自然科学基金项目(61170040),河北省自然科学基金项目(F2013201110, F2013201220),河北大学自然科学基金项目(2011-228043),河北大学教育教学改革研究项目(JX07-Y-27)资助。

翟俊海(1964—),男,博士,教授,主要研究方向为机器学习, E-mail: mczjh@hbu.cn; 哈明光(1988—),男,硕士生,主要研究方向为机器学习; 邵庆言(1986—),男,硕士生,主要研究方向为机器学习; 王熙熙(1963—),男,教授,博士生导师,主要研究方向为机器学习。

的多层感知器隐含层结点剪枝算法^[16]; Egrioglu 等提出的基于加权信息准则进行剪枝的神经网络模型选择算法^[17]; Xing 等提出的基于互信息的两阶段输入层结点和隐含层结点的剪枝算法^[18]。

近几年有些学者还提出了基于进化计算的剪枝方法,代表性的工作包括: Aladag 提出了基于 tabu 搜索的结构选择算法^[19]; Jeong 等提出了基于广义自适应模型和遗传算法相结合的网络结构剪枝算法^[20]; 郑绪枝等提出了一种快速确定隐含层神经元数的方法^[21]; 姚立忠等根据输入变量和输出变量之间的相关性,提出了一种选择神经网络输入结点的方法^[22]。

剪枝方法的缺点是需要初始化一个规模较大的网络,但相对于增量算法,剪枝算法更容易控制,得到的网络结构也更紧凑。近几年网络结构选择研究主要集中在剪枝算法上。由于理论上证明了单隐含层前馈神经网络具有一致逼近能力^[23],因此本文在单隐含层前馈神经网络的框架下讨论结构选择问题,提出了一种基于结点敏感度的单隐含层前馈神经网络结构选择,该算法的特点是不需要重复训练神经网络,得到的网络结构紧凑,具有较高的泛化能力。在实际数据集和 UCI 数据集上的实验结果显示,本文提出的算法是行之有效的。

2 概率单隐含层前馈神经网络

概率单隐含层前馈神经网络的结构如图 1 所示,它包括输入层、单隐含层和输出层。其中输入层和输出层的激活函数为线性函数,而单隐含层的激活函数为 sigmoid 函数 $\varphi(\cdot) = \tanh(\cdot)$, 软最大化函数为

$$s(o_k) = \frac{\exp(o_k)}{\sum_{i=1}^c \exp(o_i)}$$

对于给定的训练集 $T = \{(x_j, y_j) | j=1, 2, \dots, n\}$, 其中 $x_j \in \mathbb{R}^d$, $y_j \in \{1, 2, \dots, c\}$, 训练集中的样例分为 c 类 $\omega_1, \omega_2, \dots, \omega_c$ 。

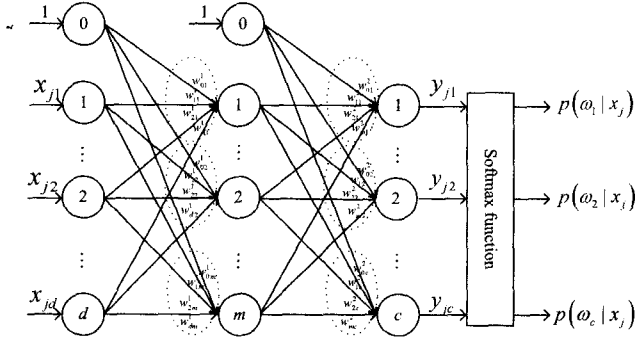


图 1 概率单隐含层前馈神经网络

输入样例 $x_j = (x_{j1}, x_{j2}, \dots, x_{jd})$, $j=1, 2, \dots, n$, 网络隐含层各结点的输出为:

$$v_{jp} = \sum_{s=0}^d x_{js} w_{sp}^1 \quad (1)$$

其中, $p=1, 2, \dots, m$, w_{sp}^1 表示第一层(即输入层到隐含层)的连接权。

网络输出层各结点的输出为:

$$y_{jq} = \sum_{t=0}^m v_{jt} w_{tq}^2 = \sum_{t=0}^m (\sum_{s=0}^d x_{js} w_{st}^1) w_{tq}^2 \quad (2)$$

其中, $q=1, 2, \dots, c$, w_{tq}^2 表示第二层(即隐含层到输出层)的连接权。

经软最大化函数变换后,得到

$$p(\omega_k | x_j) = \frac{\exp(y_{jk})}{\sum_{q=1}^c \exp(y_{jq})} \quad (3)$$

其中, $k=1, 2, \dots, c$ 。

概率单隐含层前馈神经网络的训练采用以交叉熵最小化为学习准则的 BP 算法,交叉熵函数定义如下:

$$E(w) = \sum_{j=1}^n \sum_{k=1}^c [-\delta_k(x_j) \ln p(\omega_k | x_j; w)] \quad (4)$$

其中, w 为连接权参数, δ 为指示函数,定义如下: 当 $y_j = k$ 时, $\delta(x_j) = 1$, 否则 $\delta(x_j) = 0$ 。

当交叉熵函数取得最小值时,联合概率密度 $\prod_{j=1}^n p(x_j, y_j) = \prod_{j=1}^n p(\omega_k | x_j; y_j = k)$ 取得最大值。

假设 w^* 是使 $E(w)$ 取得最小值的 w 值,那么样例 x_j 被分类到 $p(\omega_k | x_j; w^*)$ 取得最大值的那一类里,即:

$$y(x_j) = \arg \max_k p(\omega_k | x_j; w^*) \quad (5)$$

3 基于结点敏感度的单隐含层前馈神经网络结构选择

本节给出基于结点敏感度的单隐含层前馈神经网络结构选择算法,算法的基本思想是将数据集分为训练集 TR 和测试集 TE,在训练集 TR 上用基于交叉熵的 BP 算法训练一个规模较大的单隐含层前馈神经网络,对测试集 TE 中的每一个样例 x ,首先用训练好的概率单隐含层神经网络进行测试,得到该样例隶属于每一类的后验概率 $p(\omega_k | x)$, $k=1, 2, \dots, c$; 然后去掉一个隐含层结点(置相应的连接权为 0),重新测试样例 x ,也得到该样例隶属于每一类的后验概率 $p'(\omega_k | x)$, $k=1, 2, \dots, c$ 。基于这两组后验概率计算每一个隐含层结点的敏感度,将敏感度最小的隐含层结点删除,重复此过程,直到满足预定义的停止条件。隐含层结点 p 的敏感度定义为:

$$s(p) = \sum_{k=1}^c \sum_{x \in Te} |p(\omega_k | x) - p'(\omega_k | x)| \quad (6)$$

其中, $p=1, 2, \dots, m$ 。

下面给出具体的算法步骤。

Step1 将数据集 T 划分为训练集 TR 和测试集 TE;

Step2 在训练集 TR 上用基于交叉熵的 BP 算法训练一个概率单隐含层前馈神经网络,设该网络的隐含层含有 m 个结点;

Step3 重复以下步骤,直到满足预定义的停止条件;

Step4 对于测试集 TE 中的每一个样例 x ,用训练好的神经网络进行测试,得到 $p(\omega_k | x)$, $k=1, 2, \dots, c$;

Step5 逐一去掉隐含层中的每一个结点,并重新测试样例 x ,得到 $p'(\omega_k | x)$, $k=1, 2, \dots, c$;

Step6 对隐含层中的每一个结点 p ,按式(6)计算其敏感度;

Step7 计算 $p^* = \arg \max_p \{s(p)\}$;

Step8 删除 p^* 所对应的隐含层结点。

4 实验结果

为了进一步验证本文算法的有效性,分别在 2 个人工数据集和 5 个 UCI 数据集^[19]上进行了实验,5 个 UCI 数据集分别为 Iris (DS1)、ecoli (DS2)、balance (DS3)、breast (DS4) 和

heart(DS5);两个真实数据集分别是 CT (DS6)和 RenRu (DS7)。CT 图像数据库是将河北大学附属医院的 212 幅脑 CT 图像进行特征提取后得到的,212 幅图像中包含正常 CT 图像 170 幅,病变 CT 图像 42 幅。用于表示图像的特征 35 个,其中 10 个对称特征,9 个纹理特征和包括均值、方差、熵等在内的 16 个统计特征。RenRu 数据库是由河北大学智能图文实验室创建的,由 92 个汉字“人”和 56 个汉字“入”构成,每个汉字用 26 个特征来描述。实验所用的数据集的基本信息如表 1 所列。

表 1 实验所用数据集的基本信息

数据集	样例数	属性数	类别数
DS1	150	5	3
DS2	336	8	8
DS3	625	5	3
DS4	277	10	2
DS5	303	14	2
DS6	212	35	2
DS7	148	26	2

实验采用交叉验证的方法,对于每一个数据集,取 2/3 的样例作为训练集,1/3 的样例作为测试集,实验的停止条件是剪枝后神经网络的测试精度不低于剪枝前神经网络测试精度的 95%。

我们首先对比了网络结构选择前后测试精度的变化情况,实验结果如表 2 所列。

表 2 网络结构选择前后测试集测试精度对比

测试精度	DS1	DS2	DS3	DS4	DS5	DS6	DS7
Ori_acc	0.9000	0.8125	0.8900	0.6774	0.8020	0.9200	0.8400
Best_acc	0.9600	0.8482	0.8995	0.7097	0.9307	0.9333	0.8600

其中,Ori_acc 表示网络结构选择之前的测试精度,而 Best_acc 表示结构选择之后的测试精度。网络结构选择前后测试精度的变化情况如图 2 所示。

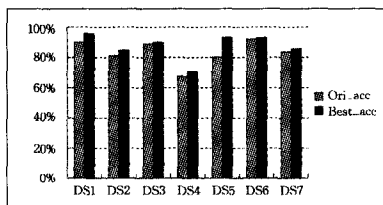


图 2 网络结构选择前后测试精度变化情况

网络结构选择前后隐含层结点个数的对比如表 3 所列。

表 3 网络结构选择前后隐含层结点个数对比

隐含层结点数	DS1	DS2	DS3	DS4	DS5	DS6	DS7
n	30	35	30	40	35	40	30
best_node_num	17	19	18	19	13	2	5

其中,n 表示网络结构选择之前的隐含层结点数,而 best_node_num 表示网络结构选择后的隐含层结点数。网络结构选择前后隐含层结点个数的变化情况如图 3 所示。

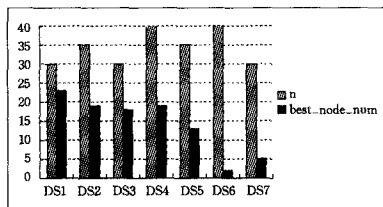


图 3 网络结构选择前后隐含层结点个数变化情况

从实验结果可以看出,应用本文提出的算法对隐含层结点较多的神经网络进行剪枝之后,网络的结构有了比较大的变化,有些网络可以去掉 1/2 甚至多于 1/2 的结点,而精度也得到了保证。

在实验中,我们发现学习速率 η 对隐含层结点个数的选择有很大影响。当 η 选取不同的值时,网络结构选择会有不同的效果。然而在确定了隐含层结点的个数后,选取不同的学习速率 η 对于测试精度却没有特别大的影响。为了选取合适的 η 值,我们分别实验对比了选择不同的 η 值对隐含层结点个数的选择以及选择好之后对网络精度的影响。

为了便于观察,我们仅以前 3 个数据集 (DS1、DS2、DS3) 为例,其他 4 个数据集上的实验结果类似。各个数据集在选择不同的 η 值时,对于选择网络隐含层结点个数的影响如图 4 所示。选择不同的 η 对选择好网络结构的神经网络的测试精度的影响如图 5 所示。

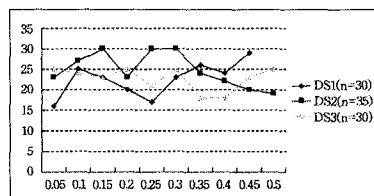


图 4 不同的 η 值对选择隐含层结点的个数的影响

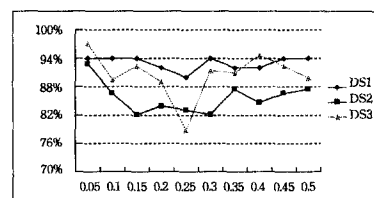


图 5 不同的 η 值对选择好网络结构的神经网络的数据集测试精度影响

从实验结果可以看出,不同的学习速率 η 对选择的隐含层结点的个数有着比较大的影响,但是当选择好网络结构之后,不同的学习速率 η 对新的神经网络的分类精度没有太大的影响,分类精度始终都保持在一个比较稳定的范围内。

我们还对网络结构选择前后神经网络训练以及测试过程所用时间进行了对比,时间对比的结果如图 6 所示。

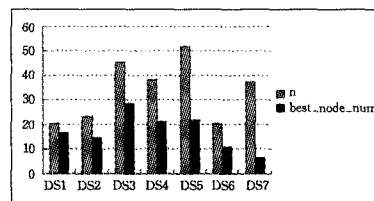


图 6 神经网络结构选择前后对数据分类的时间对比

通过实验对比可以看出,在精度基本不变甚至有些提高的前提下,通过去掉部分隐含层结点可使神经网络的训练和测试时间缩短。

结束语 本文针对单隐含层前馈神经网络,提出了一种基于结点敏感度的剪枝算法,该算法利用交叉熵作为误差函数训练神经网络,用敏感度度量隐含层结点的重要性,将不重要的结点逐个删除,直到满足预定义的停止条件。我们通过实验对比了剪枝前后神经网络隐含层结点的个数、网络的测试精度及所用的 CPU 时间;我们还实验对比了神经网络的学习

习速率 η 对剪枝算法以及测试精度的影响。实验结果显示, 本文提出的算法是行之有效的, 用该算法对网络进行剪枝不需要重复训练神经网络, 得到的网络结构紧凑, 具有较高的泛化能力。

参 考 文 献

- [1] Kumar S. Neural networkw [M]. Beijing: Tsinghua University Press, 2006
- [2] Bishop C M. Neural networks for pattern recognition [M]. Oxford: Clarendon Press, 1996
- [3] Zhang G P. An investigation of neural networks for linear time-series forecasting [J]. Computers & Operations Research, 2001, 28(12): 1183-1202
- [4] Zanchettin C, Ludermir T B, Almeida L M. Hybrid training method for MLP; optimization of architecture and training [J]. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 2011, 41(4): 1097-1109
- [5] Yang S H, Chen Y P. An evolutionary constructive and pruning algorithm for artificial neural networks and its prediction applications [J]. Neurocomputing, 2012, 86: 140-149
- [6] Kwok T Y, Yeung D Y. Constructive algorithms for structure learning in feedforward neural networks for regression problems [J]. IEEE Transactions on Neural Networks, 1997, 8(3): 630-645
- [7] Reed R. Pruning algorithms-a survey [J]. IEEE Transactions on Neural Networks, 1993, 4(5): 740-747
- [8] Redding N J, Kowalczyk A, Downs T. Constructive higher-order network that is polynomial time [J]. Neural Networks, 1993, 6(7): 997-1010
- [9] Tsoi A C, Tan S. Recurrent neural networks; A constructive algorithm, and its properties [J]. Neurocomputing, 1997, 15(3/4): 309-326
- [10] Liu D R, Chang T S, Zhang Y G. A constructive algorithm for feedforward neural networks with incremental training [J]. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, 2002, 49(12): 1876-1879
- [11] Subirats J L, Franco L, Jerez J M. C-Mantec: A novel constructive neural network algorithm incorporating competition between neurons [J]. Neural Networks, 2012, 26: 130-140
- [12] Zhang R, Lan Y, Huang G B, et al. Universal approximation of extreme learning machine with adaptive growth of hidden nodes [J]. IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(2): 365-371
- [13] Karnin E D. A simple procedure for pruning back-propagation trained neural networks [J]. IEEE Transactions on Neural Networks, 1990, 1(2): 239-242
- [14] Hagiwara M. A simple and effective method for removal of hidden units and weights [J]. Neurocomputing, 1994, 6: 207-218
- [15] Castellano G, Fanelli A M, Pelillo M. An Iterative Pruning Algorithm for Feedforward Neural Networks [J]. IEEE Transactions on Neural Networks, 1997, 8(3): 519-531
- [16] Zeng X Q, Yeung D S. Hidden neuron pruning of multilayer perceptrons using a quantified sensitivity measure [J]. Neurocomputing, 2006, 69(7-9): 825-837
- [17] Egrioglu E, Aladag C H, Gunay S. A new model selection strategy in artificial neural networks [J]. Applied Mathematics and Computation, 2008, 195(2): 591-597
- [18] Xing H J, Hu B G. Two-Phase Construction of Multilayer Perceptrons Using Information Theory [J]. IEEE Transactions on Neural Networks, 2009, 20(4): 715-721
- [19] Aladag C H. A new architecture selection method based on tabu search for artificial neural networks [J]. Expert Systems with Applications, 2011, 38(4): 3287-3293
- [20] Jeong C, Min J H, Kim M S. A tuning method for the architecture of neural network models incorporating GAM and GA as applied to bankruptcy prediction [J]. Expert Systems with Applications, 2012, 39(3): 3650-3658
- [21] 郑绪枝, 雷靖, 夏薇. 基于快速确定隐层神经元数的 BP 神经网络算法[J]. 计算机科学, 2012, 39(26): 432-436
- [22] 姚立忠, 李太福, 易军, 等. 神经网络模型的透明化及输入变量约简[J]. 计算机科学, 2012, 39(9): 247-251
- [23] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators [J]. Neural Networks, 1989, 2: 359-366
- [17] Wright J, Yang A, Ganesh A, et al. Robust face recognition via sparse representation [J]. IEEE Trans. on PAMI, 2008, 31(2): 210-227
- [18] Zheng X, Cai D, He X F, et al. Locality preserving clustering for image database[C]//ACM Int. Conf. Multimedia, 2004: 885-891
- [19] Lin Z, Liu R, Su Z. Linearized alternating direction method with adaptive penalty for low rank representation [C]//NIPS. 2011: 612-620
- [20] 陈丽敏, 杨静, 张健沛. 一种基于加速迭代的大数据集谱聚类方法 [J]. 计算机科学, 2012, 39(5): 172-175
- [21] Huang Y, Liu Q S, Zhang S, et al. Image retrieval via probabilistic hypergraph ranking [C]// Int. Conf. Comput. Vis. Pattern Recog. San Francisco, CA, 2010: 3376-3383
- [22] Zheng M, Bu J, Chen C, et al. Graph Regularized Sparse Coding for Image Representation [J]. IEEE Trans. Image Process, 2011, 20(5): 1057-1149
- [23] Jolliffe I. Principal Component Analysis [M]. New York: Springer-Verlag, 1986: 1580-1584
- [24] Chen S, Donoho D, Saunders D. Atomic decomposition by basis pursuit [J]. Soc. Ind. Appl. Math. Rev., 2011, 43(1): 129-159
- [25] Alpert C J, Kahng A B. Recent directions in netlist partitioning: A survey [J]. Integration: The VLSI Journal, 1995, 19(1/2): 1-81
- [26] Bolla M. Spectra euclidean representations and clustering of hypergraphs [J]. Discrete Mathematics, 1993, 117(1-3): 19-39
- [27] Zhuang L, Gao H, Lin Z, et al. Non-Negative Low Rank and Sparse Graph for Semi-Supervised Learning[C]//Proceedings of IEEE Conference on CVPR. Providence, RI, 2012: 2328-2335

(上接第 148 页)