

文章编号: 1001-0920(2004)08-0877-04

基于进化规划的神经网络群的设计方法

刘芳¹, 李人厚¹, 梅时春^{1,2}

(1 西安交通大学 系统工程研究所, 陕西 西安 710049; 2 第二炮兵工程学院 三系, 陕西 西安 710025)

摘要: 提出一种基于进化规划的神经网络群的自动设计方法。该方法不仅使得神经网络群中的个体网络倾向于完成不同的子任务, 同时各神经网络个体在进化过程中不断寻找最好的协作关系, 而且神经网络群的规模和结构不需预先设定。仿真试验表明, 该算法是有效的。

关键词: 神经网络群; 进化规划; 负相关学习

中图分类号: TP18

文献标识码: A

Designing neural networks ensembles based on evolutionary programming

LIU Fang¹, LI Ren-hou¹, MEI Shi-chun^{1,2}

(1 Systems Engineering Institute, Xi'an Jiaotong University, Xi'an 710049, China; 2 Third Department, Second Artillery Engineering College, Xi'an 710025, China Correspondent: LIU Fang, E-mail: liufang108@sohu.com)

Abstract: Based on the evolutionary programming, a method to automatically design neural networks (NNs) ensembles is proposed. Different individual NNs in the ensemble can learn to subdivide the task and thereby solve it more efficiently and elegantly. The architecture of each NN in the ensemble and the size of the ensemble need not to be predefined. Simulation results show that the proposed method is valid.

Key words: neural networks ensembles; evolutionary programming; negative correlation learning

1 引言

近年来, 人们对神经网络群的研究越来越重视。神经网络群由一组神经网络个体组成, 其中每个子网络分别承担复杂任务的一个子任务, 因而神经网络群比单神经网络具有更高的效率和更好的算法性能。相对于单神经网络, 神经网络群具有如下主要优点: 1) 可以完成更加复杂的任务; 2) 使得整个系统更易于理解和修正; 3) 具有更强的鲁棒性和泛化性能。显然, 神经网络群方法将成为重要且具普遍意义的解决问题的方法。然而, 神经网络群很大程度上依赖于专家的经验以及问题的先验知识, 所以设计神经网络群是一个非常困难的任务。

设计一个由多个子系统组成的学习系统最早可追

溯到1958年。自1990年以来, 人们基于这种思想提出了许多互相类似的算法, 如神经网络群 (NN ensembles)、专家混合方法 (mixtures of experts) 和偏置推动方法 (various boosting and bagging methods)^[1,2]。然而, 所有这些方法都依赖于人工设计。神经网络群中的成员网络个数常常预先设定并固定不变, 这样分配给各成员网络的子任务往往是不清楚的。文献[1]中提出一种基于负相关学习的进化神经网络方法, 该方法一方面利用负相关学习保持种群中个体网络的多样性, 另一方面利用进化算法来优化各子网络的权值, 并对学习好的各子网络进行组合, 从而实现神经网络群的自动设计。然而, 该算法中神经网络群的规模和结构需预先设定。

收稿日期: 2003-08-04; 修回日期: 2003-10-23

作者简介: 刘芳 (1971—), 女, 陕西长安人, 博士生, 从事进化计算与神经网络的研究; 李人厚 (1935—), 男, 浙江宁波人, 教授, 博士生导师, 从事智能控制、计算机协同工作等研究。

本文在文献[1]的基础上,提出一种各子网络结构及其线性组合的系数同时进化学的神经网络群设计方法,同时在进化过程中设计个体的适应度函数不仅是负相关学习目标项,而且包括网络结构复杂度。另外,本文中各子网络通过线性组合方式构成神经网络群,每个子网络对应的线性组合系数也参与进化和优化。因此,本文算法不但使得神经网络群中的个体倾向于完成不同的子任务,同时在进化过程中不断寻找最好的协作关系。另外,本文算法中神经网络群的规模和结构不需预先设定。

本文利用进化规划进行神经网络群的自动设计。进化规划是进化算法的一种,它与遗传算法的不同主要在于:遗传算法关注的是个体行为的遗传,通过个体之间的信息交换来产生下一代个体;而进化规划则侧重于群体的行为变化,只采用变异一个操作算子。本文选择进化规划来设计神经网络,主要基于以下两点考虑:1)由于进化规划采用实数编码方案,省略了二进制数编码方案复杂的网络表示和解码计算,更适合表示和操作神经网络群;2)进化规划只需变异算子,避免了遗传算法中交叉算子在操作时易破坏已进化出的网络结构问题。

2 算法

进化算法是一种基于群体的学习算法,具有许多优良特性(不需要导数信息、适用面广、简单易行、具有较强的鲁棒性等),因而得到了广泛的应用。近年来,将进化算法用于神经网络设计愈来愈受到人们的重视,并取得了很多研究成果。然而,具有最小的训练误差的神经网络不一定具有最好的泛化能力^[3],而种群中的其他个体也可能包含泛化能力的有用信息。因此,由一组神经网络组成的神经网络群比单神经网络具有更好的泛化能力^[1]。但神经网络群设计的关键是解决任务的分配问题,即应使各子网络分别完成不同的任务,才能使其各司其职,最终以最小的网络规模和神经网络群完成给定的任务。

2.1 适应度函数设计

2.1.1 利用负相关学习为各子网络分配学习任务

对随机向量 $(F_i(n), F_j(n))$,如果 $D(F_i(n)) \neq 0, D(F_j(n)) \neq 0$,则称

$$\rho_{F_i(n)F_j(n)} = \frac{\text{cov}(F_i(n), F_j(n))}{\sqrt{D(F_i(n))} \sqrt{D(F_j(n))}} = \frac{E[F_i(n) - E(F_i(n))][F_j(n) - E(F_j(n))]}{\sqrt{D(F_i(n))} \sqrt{D(F_j(n))}} \quad (1)$$

为 $F_i(n)$ 与 $F_j(n)$ 的相关系数,其中 $D(\bullet)$ 表示方差。如果 $\rho_{F_i(n)F_j(n)} < 0$,则称 $F_i(n)$ 与 $F_j(n)$ 负相关。

负相关学习方法提供了一种为神经网络群中的不同网络个体分配不同子任务的方法(文献[4,5]从理论和试验上都证明了该方法的有效性)。本文基于文献[1]的方法构造神经网络的目标函数如下:

设训练样本集为 $\{(x(1), d(1)), \dots, (x(N), d(N))\}$,其中: $x \in X^p, d$ 表示标准输出, N 表示样本集的大小。则第 i 个神经网络学习的目标函数为

$$E_i = \frac{1}{N} \sum_{n=1}^N E_i(n) = \frac{1}{N} \sum_{n=1}^N \frac{1}{2} (\alpha F_i(n) - d(n))^2 + \frac{1}{N} \sum_{n=1}^N \lambda p_i(n), \quad (2)$$

$$p_i(n) = (\alpha F_i(n) - F(n)) \sum_j (\alpha F_j(n) - F(n)). \quad (3)$$

其中: $F(n) = \frac{1}{M} \sum_{i=1}^M \alpha_i(n) * F_i(n)$ 表示神经网络群对第 n 个样本的输出; M 表示神经网络群的规模; N 表示样本集的大小; α_i 表示神经网络群中第 i 个子神经网络对应的线性组合系数,该系数表示第 i 个子神经网络对整个神经网络群输出的贡献,也可理解为采用一种“类似投票”的方式决定各个子神经网络对整个神经网络群最终输出的影响; $F_i(n)$ 表示第 i 个子神经网络对第 n 个样本的输出; p_i 表示相关性惩罚函数; λ 表示惩罚系数, $0 \leq \lambda \leq 1$ 。

式(2)对应的目标函数实质上表示一个多目标优化问题,其中第1项表示神经网络群的输出误差。从 $p_i(n)$ 的计算公式(3)并结合式(1),可以看出 $p_i(n)$ 可用来度量第 i 个神经网络与神经网络群中其他神经网络的相关程度的累计和。当 $p_i(n)$ 取正值时,式(2)的第2项作为“惩罚项”;当 $p_i(n)$ 取负值时,式(2)的第2项作为“奖励项”。因而,上述神经网络学习的目标函数鼓励第 i 个神经网络的输出与神经网络群中的其他神经网络的输出尽量负相关。

2.1.2 适应度函数

考虑如下适应度函数:

$$f(i, t) = \frac{1}{E_i + aH(i, t)} \quad (4)$$

其中: $f(i, t)$ 表示第 t 代的第 i 个网络个体的适应度; $H(i, t)$ 表示第 t 代的第 i 个网络个体的结构复杂度,它通过网络的节点个数与连接个数之和进行评价; a 为惩罚系数。

2.2 利用小生境机制保持种群的多样性

在进化算法中, 保持种群的多样性具有重要的意义。本文利用小生境(生物学中小生境指特定环境下的一种生存环境) 机制来进一步保持种群的多样性。在进化规划中, 通过反映个体之间相似程度的共享函数来调整群体中各个个体的适应度, 从而在此后的群体进化过程中, 算法能够依据调整后的新适应度进行选择运算, 以维护种群的多样性, 创造出小生境的进化环境。

通过反映个体之间相似程度的共享函数^[6] 调整群体中各个个体的适应度的方法如下:

设第 l 个训练样本能被神经网络群中的 q 个网络个体正确学习, 这 q 个网络个体标记为 $q_1, q_2, \dots, q_q$, 则该 q 个网络个体对样本 l 的共享度 S_{lj} ($j = q_1, q_2, \dots, q_q$) 均为 $1/q$, 其他网络个体对样本 l 的共享度为 0。考虑所有样本的情况, 则网络个体 j 对所有样本的共享度为 $S(j) = \sum_{l=1}^N S_{lj}$, 其中 N 表示样本集的大小。在计算出群体中各个个体的共享度后, 依据

$$f(i, t) = f(i, t) * S(j) \quad (5)$$

调整各个个体的适应度。

2.3 选择机制

本算法中的选择机制为确定性选择方法与精英选择策略相结合^[7], 其具体选择过程如下: 随机选择子神经网络进行结构变异; 然后用权值进化算法对其进行训练, 并用确定性选择策略选取适应值高的个体作为新的网络群体, 反复进行权值训练, 直到满足一定的算法终止条件; 最后将种群中的各子网络进行组合, 并计算神经网络群的输出误差。如果此时获得的神经网络群优于结构变异前的神经网络群, 则用当前神经网络群替代原神经网络群, 否则保留原神经网络群。

2.4 网络连接权进化

进化规划的实施过程类似于遗传算法, 包括产生初始群体、计算适应度、执行突变、选择、反复迭代, 直到获得满意结果。进化规划只有变异算子, 绝大多数进化规划都采用高斯变异算子。近年来, 有人提出用柯西(Cauchy) 变异算子代替高斯变异算子^[8]。柯西分布和高斯分布很相似, 但它具有较高的两翼概率特性。柯西分布容易产生一个远离原点的随机数, 如果用柯西变异替换原进化规划中的高斯变异来产生后代, 则这种变异方式跳出局部极小的能力较强。但柯西分布较小的中央部分是它的弱点。本文在自适应进化规划方法的基础上, 将柯西变异算子与高斯变异算子相结合, 同时引入搜索指导方

向, 提出一种新的进化规划变异规则:

如果 $f(i, t) > f(i, t-1)$, 则

$$\text{dir}(W_j(i, t)) = \text{sgn}(W_j(i, t) - W_j(i, t-1)),$$

$$W_j(i, t) = W_j(i, t-1) + \text{dir}(W_j(i, t)) \times$$

$$\sigma_j(i, t-1)C_j(0, 1)k_1;$$

否则

$$W_j(i, t) = W_j(i, t-1) + \sigma_j(i, t-$$

$$-1)C_j(0, 1)k_2,$$

$$\forall i \in \{1, 2, \dots, N_p\}, \forall j \in \{1, 2, \dots, n\};$$

$$\sigma_j(i, t) = \sigma_j(i, t-1)\exp(\tau N(0, 1) +$$

$$\tau N_j(0, 1)).$$

其中: $\tau' = \left| \sqrt{2\sqrt{n}} \right|^{-1}$, $\tau = \left| \sqrt{2n} \right|^{-1}$; n 是 $W(i, t)$ 的维数; $N(0, 1)$ 表示高斯随机变量; $C(0, 1)$ 表示柯西随机变量; $W_j(i, t)$ 表示第 t 代种群中的第 i 个个体的第 j 个参数; N_p 表示种群的大小; $\text{dir}(W_j(i, t))$ 表示 $W_j(i, t)$ 的进化方向; $\text{sgn}(\bullet)$ 表示符号函数; $f(\bullet)$ 表示适应度函数; $\sigma_j(i, t)$ 表示 $W_j(i, t)$ 的进化步长; k_1 和 k_2 为可调参数。

2.5 网络结构进化和种群规模进化

网络结构进化包括 3 个基本算子: 隐单元增添变异, 隐单元删减变异, 连接权修剪变异^[9]。经过多次网络权值进化和网络结构进化后, 若神经网络群误差仍未达到误差上限, 则随机增添神经网络个体数目, 扩大种群规模; 反之, 若神经网络群误差已达到误差上限, 则删减种群中对神经网络群输出影响不大的个体, 以精简神经网络群中的个体数目。

3 仿真试验

选用著名的 Iris(蝴蝶花) 数据测试本文算法的有效性, Iris 数据经常被用来检验聚类(无监督)或分类(有监督)模式识别方法的性能。Iris 共有 3 个可分的种类, 分别称为 Virginica(第 1 类)、Setosa(第 2 类)和 Versicolor(第 3 类), 具体属于哪一类主要取决于以下 4 个特征: 萼片长度(sepal length), 萼片宽度(sepal width), 花瓣长度(petal length)和花瓣宽度(petal width)。Iris 数据中包含 3 类共 150 个样本, 每类 50 个。对有监督设计标准的错误个数为 0~5, 而无监督设计标准的错误个数在 15 左右。取出 120 个样本(每类 40 个)作为训练样本, 余下 30 个样本作为测试样本。

采用本文方法可获得由 3 个子网络组成的网络群。子网络结构分别为 4-2-3, 4-4-3, 4-4-3, 非零连接权分别为 6, 13, 11。该神经网络群不仅可对训练样本达到 100% 正确学习, 对测试样本的识别率也达

到了100%。而文献[10]方法的学习结构的网络结构为: 4-8-4-3, 非零连接权个数为128个, 对所有样本的输出误差平方和为3 366 2(最好学习结果)。

由上述试验可看出, 采用本文算法, 每个子神经网络的结构都减小很多, 并提高了整个神经网络群的泛化能力。本文最终实验结果获得的3个子网络, 比文献[10]中的网络数目多([10]只有1个网络)。换言之, 本文将网络的规模大小与问题复杂性的矛盾转化为子网络数目多少与问题复杂性的矛盾。后者比较容易由软件模拟或硬件解决, 而且基于进化规划的进化神经网络群更便于工程上并行实现。

4 结 语

本文提出了一种基于进化规划的神经网络群设计方法。仿真实验表明, 由多个神经网络组成的神经网络群具有更为简洁的子网络结构和更好的算法性能。同时还提出了一种用于网络权值训练的高斯变异与柯西变异相结合的进化规划算法。该算法可以增强进化规划, 具有跳出局部极小的能力, 可改善算法的全局收敛性。但是, 如何进一步发展神经网络群的设计方法及其相关理论尚有待于进一步研究。

参考文献(References):

- [1] Liu Yong, Yao Xin, Tetsuya Higuchi. Evolutionary ensembles with negative correlation learning[J]. *IEEE Trans on Evolutionary Computation*, 2000, 4(4): 308-387.

- [2] Liu Yong, Yao Xin. Simultaneous training of negatively correlated neural networks in an ensemble[J]. *IEEE Trans on Man and Cybernetics — Part B*, 1999, 29(6): 716-725.
- [3] Baldi P. Learning in linear neural networks: A survey[J]. *IEEE Trans on Neural Networks*, 1995, 6(4): 837-858.
- [4] Clemen R T, Winkler R L. Limits for the precision and value of information from dependent sources[J]. *Operartion Reasearch*, 1985, 33: 427-442.
- [5] Perrone M, Cooper L N. When Networks Disagree: Ensemble Methods for Hybrid Neural Networks[M]. London: Chapman & Hall, 1993.
- [6] 周明, 孙树栋. 遗传算法原理及应用[M]. 北京: 国防工业出版社, 1999. 74-77.
- [7] 李人厚. 智能控制理论和方法[M]. 西安: 西安电子科技大学出版社, 1999. 178-179.
- [8] 云庆夏. 进化算法[M]. 北京: 冶金工业出版社, 2000. 157-158.
- [9] 李孝安. 进化神经网络理论及方法研究[D]. 西安: 西北工业大学, 2000. 92-93.
- [10] 王磊, 戚飞虎. 进化计算在神经网络学习中的应用[J]. *计算机工程*, 1999, 25(11): 41-43.
(Wang L., Qi F. H. Applcation of evolutionary computation in neural network [J]. *Computer Engineer*, 1999, 25(11): 41-43.)

(上接第871页)

7 结 语

本文提出的AD-Mine算法, 能够自动判别数据库中的数据特性, 确定挖掘策略, 最大限度地利用计算机的内存资源, 提高数据挖掘的效率, 与现有的算法相比具有更强的适应性, 与Apriori类算法相比具有更高的效率。同时, 本文提出的一种新的数据库划分方法, 是一种高效的方法, 可以显著提高数据挖掘的效率, 特别是对于挖掘较长频繁模式时, 优势更加明显。

参考文献(References):

- [1] Agrawal R, Srikant R. Fast algorithms for mining association rules in large database[A]. *Proc of the 20th VLDB Conf* [C]. Santiago, 1994. 487-499.
- [2] Park J S, Chen M S, Yu P S. An effective hash-based algorithm for mining association rules[A]. *Proc of the*

1995 *ACM SIGMOD* [C]. San Jose, 1995. 175-186.

- [3] Orlando S, Palmerini P, Perego R. Enhancing the apriori algorithm for frequent set counting[A]. *Proc of 3rd Int Conf on DataWAK* 2001 [C]. Munich: Spriger, 2001. 1-17.
- [4] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation[A]. *SIGMOD 00* [C]. Dallas, 2000. 1-12.
- [5] Pei J, Han J, Lu H, et al. H-Mine: Hyper-structure mining of frequent in large database[A]. *Proc 2001 Int Conf on Data Mining (ICDM 01)* [C]. San Jose, 2001. 38-49.
- [6] Savasers A, Omiecinski E, Navathe S. An efficient algorithm for mining association rules in larges dadabase[A]. *In 21st Int Conf on Very Large Databases (VLDB)* [C]. Ztrich, 1995. 432-443.