

常见遗传多层前馈神经网络实数编码的不合理性分析和改进*

莫鸿强 罗 飞 侯小梅 毛宗源

(华南理工大学自动控制工程系·广州, 510640)

摘要:对遗传多层前馈神经网络常见的实数编码中不适合于优化权值大小、不利于交叉算子发挥作用等问题,进行了理论分析,提出了“最小模式”等概念,并据此对编码方式做了改进,通过一个实例的仿真结果表明改进后的编码方式具有较好的寻优效果。

关键词:遗传算法;多层前馈神经网络;编码;模式

文献标识码:A

An Analysis and Improvement of the Shortcomings of the Common Real Number Coding Scheme in Genetic Multilayer Feedforward Neural Networks

MO Hongqiang, LUO Fei, HOU Xiaomei and MAO Zongyuan

(Department of Automatic Control Engineering, South China University of Technology, Guangzhou, 510640, P. R. China)

Abstract: This paper theoretically analyses the problems in the common real number coding scheme in genetic multilayer feedforward neural networks, such as being unsuitable for the optimization of the weights, influencing the function of crossing operators, and so on. A concept named as “minimal schema” is proposed, based on which the coding scheme is improved. The simulation results of an example show that the improving coding is effective in optimization.

Key words: genetic algorithms; multilayer feedforward neural networks; coding scheme; schema

1 引言(Introduction)

遗传算法(GA)具有很强的搜索能力,许多学者将它用于神经网络连接权的训练上^[1-7],构成所谓的遗传神经网络^[5]。遗传神经网络权值的编码方式大致可分为两类:一类是二进制数编码方式^[2,6,7];另一类是实数编码方式^[1,3,4]。二进制编码方式的缺点是不直观,且用于较大规模神经网络连接权的编码时,不是因为字符串太长而导致寻优时间过长,就是因为字符串太短,使权值表达精度不够。实数编码方式比较直观,且不会出现精度不足的情况,所以,遗传多层前馈神经网络的连接权多采用实数编码方式。但是,这种以实数作为基因的编码方式其实也不完全合理,下面将对此加以分析。先结合一个实际例子分析该编码方式的不合理的地方,然后从理论上作初步探讨,最后给出编码的改进形式。

2 常见遗传多层前馈神经网络实数编码分析(Analysis of the common real number coding scheme in genetic multilayer feedforward neural networks)

2.1 实例分析(Illustration)

将遗传算法用于一结构已知的多层前馈神经网络的权值优化上。

问题 1 给定一组数据 $\{(x_i, t_i) \mid i = 1, \dots, 16;$
 $x_i \in (0, 2\pi), \text{ 且 } x_{i+1} - x_i = 2\pi/n; t_i = 0.4\sin(x_i) + 0.5\}$ 训练一单输入单输出, BP 网络结构的神经网络。所采用的神经网络隐层数为 3, 隐层神经元数为 6。

用遗传算法解问题 1。采用实数编码方式, 群体规模 100, 基因取值范围 $(-10, 10)$, 变异时, 在该范围内随机产生一实数以代替要变异的基因, 概率 0.004。交叉操作同标准的遗传算法交叉算子, 概率

* 基金项目: 国家自然科学基金(69864001)资助项目。

收稿日期: 1999-10-26; 收修改稿日期: 2000-06-28。

0.4. 训练误差 $E = \frac{1}{2} \sum_{i=1}^{16} (t_i - y_i)^2$, 个体适应度值 e^{-E} . 用赌轮方式选择要复制的个体.

表1 算法改进前适应度分布

Table 1 The fitness values distributions of the unimproved algorithm

代序号	适应度分布	对应训练误差分布
初始群体	最优值 0.4153 ~ 0.8332	0.182 ~ 0.879
	平均值 0.0981 ~ 0.1301	2.039 ~ 2.322
200 代后	最优值 0.5121 ~ 0.8722	0.134 ~ 0.669
	平均值 0.4849 ~ 0.8793	0.134 ~ 0.724
1 000 代后	最优值 0.5126 ~ 0.8793	0.129 ~ 0.668
	平均值 0.5126 ~ 0.8793	0.129 ~ 0.668
5 000 代后	最优值 0.5880 ~ 0.8900	0.117 ~ 0.532
	平均值 0.5880 ~ 0.8900	0.117 ~ 0.532

表1给出了50组计算的统计结果.由表1知其优化效果并不理想,找到的较优点适应度值低且优化时间长.同时,各次重复计算的结果相互间差别很大,平均适应度的分布范围较宽.从其典型进化曲线可见,绝大部分时间内,群体平均适应度都等于最优适应度.观察相应群体中各个体,绝大多数个体已完全相同.问题本身的复杂性是优化效果差的主要原因,而下面将会看到,编码方式不合理也影响了算法优化性能.典型的进化曲线如图1所示,其中,横坐标为代数,纵坐标为群体适应度平均值和最优值.

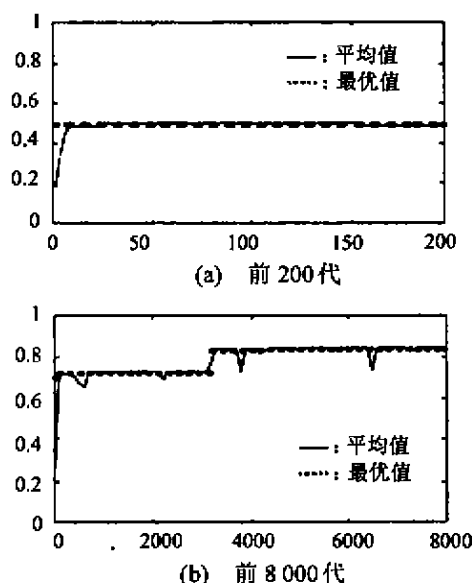


图1 算法改进前典型进化曲线

Fig. 1 Typical evolution curve of the unimproved algorithm

2.2 编码分析(Coding scheme analysis)

本节拟从两方面进行分析,先说明该实数编码方式不适合于优化权值大小,然后说明该编码方式

不利于交叉算子发挥作用.先引入GA空间的定义.所谓GA空间是指由基因型个体所组成的空间.

2.2.1 实数编码方式与权值优化(Real number coding and weight optimization)

为分析编码方式的合理性,先要讨论GA起作用的关键.GA要起作用,首先要对被优化的参数进行编码,然后通过遗传操作对群体中具有一定组合结构形式的个体作结构重组处理,从而不断地搜索出群体中个体间的结构相似性,形成并优化积木块以逼近最优解.GA究其本质而言就是求取最佳基因组合的过程,寻求合理的基因组合方式是其中关键.

遗传多层前馈神经网络权值的优化可以分成“权值之间的合理组合”和“权值自身取值”两部分.两方面需同时编码,同时优化.而前边所采用的实数编码方式,实际上只对权值的组合作了编码,并没有对权值参数进行编码;其优化对象主要还只是权值的组合.权值参数的改变只靠漫无目的随机变异,没有体现出GA“搜索方向性强”的特点来.

多层前馈神经网络(特别是前边所举的BP结构的网络)的结构一般并不复杂.就一 p 输入、 q 输出、 n 隐层神经元,BP结构的三层神经网络而言,网络权值数为 $(p+q+1)n+q$.不考虑变异,规模为100的群体即使不同个体间对应基因座上的基因各不相同,经交叉最多也只能生成 $100^{((p+q+1)n+q)}$ 个个体,对前例遗传神经网络而言是 $100^{19} = 10^{38}$,远小于GA空间所包含的个体数 (5.243×10^{100}) .可以说,问题的主要方面不在于权值之间的排列,而在于权值的取值本身.该编码方式只对权值的组合编码,而不对权值参数本身进行编码,忽略了问题的主要方面,因此说它是不合理的.

2.2.2 实数编码方式与交叉算子(Real number coding and crossover operators)

交叉算子通过改变基因之间的组合起作用,群体必须具有一定程度的多样性,交叉操作才能更好发挥作用.下面对此进行讨论.先介绍几个定义.

设染色体中每一基因均为字符集合 C 中的元素.现令集合 $\bar{C} = C + \{*\}$,其中“*”为通配符,值可取为集合 C 中任一元素.

定义1 由集合 $\bar{C}$ 中元素为基因构成,串长等于染色体长度的字符串称为模式.

定义2 任一次交叉操作都无法破坏和生成的模式称为最小模式.

以串长为2的标准遗传算法的编码方式为例, $** , *0 , *1 , 0* , 1* , 00 , 01 , 10 , 11$ 都是最小模式.所有最小模式构成的集合 S 称为完全最小模式

集. 设 GA 进化过程中第 t 代群体为集合 G_t (以其中个体为元素). 则第 t 代最小模式集为 $S_t = S \cap G_t$.

交叉操作引起 GA 空间中个体的迁移, 是否具备全局搜索性是指这种个体的迁移是否能遍历 GA 空间中的任意点. 因此, 有下述定理.

定理 1 只有交叉算子起作用时, 交叉操作具有全局搜索性的必要条件是群体 G_t 的最小模式集 $S_t = S$.

证 由定义知, $S_t \subset S$ 必成立.

现假设存在最小模式 $s = * \cdots * ch_i * \cdots *$ (i 表示第 i 个基因座, $ch_i \in \bar{C}$), $s \in S$, 且 $s \notin S_t$. 交叉操作能生成 GA 空间中任一点. 则任意与该最小模式匹配的点均不属于 G_t . 因作用在群体 G_t 上的交叉操作无法生成该最小模式, 因此, $s \notin S_k, \forall k \geq t$. 任意与该最小模式匹配的点同样均不属于 $G_k, \forall k \geq t$, 即仅靠交叉无法搜索到这些点, 与假设相矛盾. 故假设不成立, 即 $\forall s \in S$, 成立 $s \in S_t$. 因此, $S_t = S$ 成立. 命题得证.

对前述遗传多层前馈神经网络, 设 O 为其满足最优性能指标的点集 (以下简称最优点集), S_G 为其完全最小模式集, 则最优点集 O 对应的最优最小模式集为 $S_O = S_G \cap O$. GA 空间中的每一点 x , 相应地也有点 x 的最小模式集:

$$S_x = \{ \underbrace{* \cdots * ch_{i+1} * \cdots *}_{i+1} \mid ch_{i+1} \in \{x \text{ 上第 } i+1 \text{ 个基因座上值, } *\}; i, j = 0, \cdots, N-1; i+j = N-1 \}.$$

定理 2 只有交叉算子起作用时, 交叉操作能搜索到最优点的必要条件是: 对群体 G_t 的最小模式集 S_t , 存在至少一点 $x \in O$, 使 $S_x \subset S_t \cap S_O$ 成立.

证 假设对群体 G_t 的最小模式集 S_t , O 中不存在使 $S_x \subset S_t \cap S_O$ 成立的点 x 时, 交叉操作能搜索到最优点. 则对任一点 $x \in O$, 都存在一最小模式 $s = * \cdots * ch_i * \cdots *$ (i 表示第 i 个基因座, $ch_i \in \mathbb{R}$), 满足 $s \in S_x$, 且 $s \notin S_t$. 又仅交叉操作不能生成该最小模式 s , 所以必有 $\forall k \geq t, s \notin S_k$, 因此与该模式 s 匹配的点 $x \notin G_k, \forall k \geq t$, 即仅靠交叉操作无法搜索到任一最优点. 与假设相矛盾. 故假设不成立. 命题得证.

对任意确定的问题, 最优点集的基数是确定的, 它不随编码方式的改变而变化. 但是, 编码方式无论对最优最小模式集还是对第 t 代最小模式集的基数都有很大影响, 这种影响主要通过改变字符集 C 实现. 例如对区间 $[0, 1]$ 上的点编码, 要求精度达到 ± 0.001 . 十位二进制数编码时, 其完全最小模式集

基数为 21; 而实数编码时, 其完全最小模式集基数达 1001. 设某问题的最优点为 $x = 0.500$, 最优最小模式集对二进制编码为 $S_{Ob} = \{1 * * * * * * * *, * 0 * * * * * * *, \cdots, * * * * * * * * 0\}$, 对实数编码为 $S_{Od} = \{0.500\}$. 因为最优点只有一个, 所以, $S_{xb} = S_{Ob}, S_{xd} = S_{Od}$. 对规模为 100, 随机生成的初始群体 G_{Ob} (或 G_{Od}) (对应群体最小模式集为 S_{Ob} (或 S_{Od})), 若二进制数编码每位为 0 或 1 的概率都 0.5, 则 $S_{Ob} \cap S_{Ob} \supset S_{xb}$ 的概率为 $(1 - (1 - 0.5)^{100})^{10} > 0.99999$, 而实数编码 $S_{Od} \cap S_{Od} \supset S_{xd}$ 的概率 ≤ 0.1 (个体两两相异时取等号). 可以说, 在这个问题上, 二进制数编码要比实数编码更适合于交叉算子发挥作用.

如果能对实数编码方式作适当的调整, 使其既能保持精度的优点, 又能更好地发挥交叉算子的作用, 就有望改善遗传多层前馈神经网络的寻优性能. 由定理 2 可知, 这就是要增加 $S_x \subset S_t \cap S_O$ 成立的概率. 一个可行的调整方向是适当减小字符集合 C 的“规模”, 使得 GA 空间的完全最小模式集 S 的“规模”尽可能减小, 这样与空间中不同点匹配的最小模式重叠的概率就会随之增大, 从而 $S_x \subset S_t \cap S_O$ 的可能性增大.

3 算法改进 (Algorithm improvements)

算法改进主要针对编码方式, 但是, 改进要取得效果, 交叉、变异算子也必须作出相应调整. 根据第四节的两点结论, 并兼顾计算精度, 现将实数编码改为十进制数编码. 将每个权值看作 4 个数字符按一定顺序的排列, 例如, 3.146 看作排列“3”“1”“4”“6”. 交叉位置不再限制在权值之间, 权值内部也可以交叉, 交叉方式同前. 变异操作的对象变为十进制数字符, 数字符的变异为在字符集合 C 中任选一数字符来代替原数字符. 其余同前.

算法用 C 语言编程在 64M 内存的 P II 350 PC 上实现. 改进前后算法对应的可执行文件大小几乎相等, 都约为 71K 字节. 程序运行时, 除这 71K 外, 为满足运算等操作的需要, 还需占用不少内存, 且对内存的这部分需要不随算法改进而增加或减少. 整型、实型变量在内存中分别占 2 和 4 个字节, 每个权值看作 4 个数字符的排列, 且群体规模为 100, 所以, 算法改进后所需内存只比改进前约多 7600 字节, 远少于程序运行时所需内存量, 因而存储量增加不多. 即使对权值数为 200 的较大规模遗传神经网络, 其内存增加量也才 80K, 目前一般的微机都能承受. 改进后, 交叉时要交换的基因座数增多, 待判断是否变

异的基因座数也同时增大,两者的平均数都是改进前的4倍,这将增加算法的运算量.但由于遗传神经网络运算量最大的部分在于求个体适应度值,且算法的改进对其没有影响.与适应度值的计算相比,交叉和变异操作所需时间不多.实际运行结果表明,算法改进后运行时间约为改进前的1.2~1.4倍.PⅡ 350 PC的计算性能较好,存储量和运算量增加不大时实际运行的效果不会受到很大影响,因而算法的改进可行.

改进后,GA空间的完全最小模式集 S 的基数由原来的减少为950.因此, $S_e \subset S_i \cap S_o$ 的概率要比改进前的大很多.表2和图2分别给出了其50组计算的统计结果和典型的进化曲线.

表2 算法改进后适应度分布

Table 2 The fitness value distributions of the improved algorithm

代序号	适应度分布	对应训练误差分布
初始群体	最优值 0.4153 ~ 0.8332	0.183 ~ 0.879
	平均值 0.0981 ~ 0.1301	1.973 ~ 2.322
200代后	最优值 0.6779 ~ 0.9146	0.089 ~ 0.389
	平均值 0.6690 ~ 0.8978	0.108 ~ 0.402
1000代后	最优值 0.7904 ~ 0.9425	0.059 ~ 0.235
	平均值 0.7747 ~ 0.9271	0.075 ~ 0.255
5000代后	最优值 0.9356 ~ 0.9622	0.038 ~ 0.067
	平均值 0.8993 ~ 0.9600	0.041 ~ 0.106

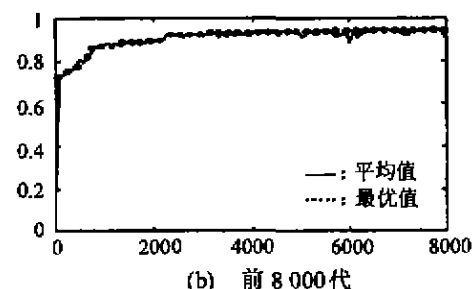
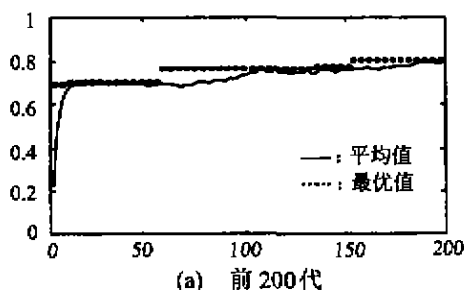


图2 算法改进后典型进化曲线

Fig. 2 Typical evolution curve of the improved algorithm

对照表2和表1,图2和图1,可以明显看出,十进制编码无论是寻优能力还是收敛情况都要比修改前的好,且适应度分布更合理.因此,可以说这种改

进是有效的.

4 结束语(Conclusion)

本文指出了常见遗传多层前馈神经网络实数编码不适合于优化权值大小,且不利于发挥交叉操作的作用;就此提出了最小模式和最小模式集等概念,并根据这些概念修改了原编码方式和算法,提高了算法的优化能力.

在仿真过程中发现,适应度达到0.96以后就基本停滞不前,寻优能力大大减弱.这说明了目前的算法尚缺乏微调能力.该缺点的克服还有赖于结合遗传多层前馈神经网络的特点对遗传算子、适应度函数和各种参数作进一步的调整.

参考文献(References)

- [1] Moutana D and Davis L. Training feedforward neural networks using genetic algorithms [A]. Proceedings of Eleventh International Joint Conference on Artificial Intelligence [C]. San Mateo, CA: Morgan Kaufmann, 1989, 762-767
- [2] Pan Weidong. Design of an artificial neural network aided by genetic techniques [J]. Pattern Recognition and Artificial Intelligence, 1994, 7(1): 72-77 (in Chinese)
- [3] Wang Baozhong, Kang Lishan and He Wei. A multi-layer neural network backpropagation algorithm based on real-coded genetic algorithm [J]. Journal of Wuhan University, 1998, 44(3): 289-291 (in Chinese)
- [4] Li Mingqiang, Xu Boyi and Kou Jisong. On the combination of genetic algorithms and neural networks [J]. System Engineering Theory and Practices, 1999, 19(2): 65-70 (in Chinese)
- [5] Yao X. A review of evolutionary artificial neural networks [J]. Int. J. of Intelligent Systems, 1993, 8(4): 539-567
- [6] Whitley D and Hanson T. Optimizing neural networks using faster, more accurate genetic search [A]. Proceedings of the Third International Conference on Genetic Algorithms and Their Applications [C]. San Mateo, CA: Morgan Kaufmann, 1989, 391-396
- [7] Marshall S J and Harrison R F. Optimization and training of feedforward neural networks by genetic algorithms [A]. The Second IEEE International Conference on Artificial Neural Networks [C]. Stevenage, 1991, 39-43

本文作者简介

莫鸿强 1976年生,华南理工大学自动控制工程系博士生.主要研究方向为智能控制,优化方法等.

罗飞 1957年生,1982年2月和1986年5月在北京钢铁学院获学士和硕士学位,1993年7月在北京科技大学获博士学位.1990年10月至1992年10月作为中日联合培养博士生在日本九州工业大学学习,1996年10月至1997年10月在日本九州工业大学做访问学者.现为华南理工大学自动控制工程系教授.主要研究方向为神经网络,遗传算法,模糊控制,运动控制系统,机器人控制系统.

侯小梅 女,1968年生,华南理工大学博士生.主要研究方向:智能控制,混沌加密等.

毛宗源 见本刊2001年第2期第165页.