文章编号：1000－5471(2007)04－0090－04

# 基于遗传神经算法优化的汉语分词模型①

陈 琳， 何 嘉

成都信息工程学院 计算机系，成都 610225

**摘要**：针对目前汉语分词系统中 BP 算法收敛速度慢等难题，本文将基于遗传的 BP 神经网络算法用于汉语分词模型，此模型结合了遗传算法和 BP 网络的优点. 实验结果表明：该优化模型完全达到了汉语分词要求，并且在分词速度上也明显优于传统的 BP 神经网络，具有高精确性、收敛速度快等特点.

**关 键 词**：神经网络；汉语分词；遗传算法；BP 算法

**中图分类号**：TP301.6 **文献标识码**：A

汉语自动分词的研究已经有二十多年了，但目前仍然是制约汉语信息处理发展的一个瓶颈. 尹峰等提出了以神经网络理论(BP 模型)为基础的汉语分词模型，为汉语分词研究开辟了新途径. BP 算法是用于多层神经网络训练的著名算法，有理论推导严谨、物理概念清楚、通用性强等优点. 但在实际应用中，BP 算法存在收敛速度慢、易陷入局部最小等缺点[1]，严重妨碍了分词速度. 本文先采用遗传学习算法进行全局寻优，再利用 BP 算法进行精确训练，优化 BP 神经网络权重学习. 通过大量仿真试验，证明该模型明显加快收敛速度.

## 1 基于神经网络汉语分词模型

神经网络模型中的隐含神经元表示一组关联规则，输入的一组编码对应于关联规则的条件，而输出规则对应于关联规则的结果，也就是语句的切分. 具体而言，对刚初始化的神经网络分词模型，可以先输入一定数量的样本进行训练. 输入层每一个神经元均对应固定的字或词，每个样本都有其自身的切分规则. 我们可以把这些规则理解为网络的权重，一旦训练完成，系统就能对这些字词做出正确切分，使神经网络实现自适应和自学习，以获取新的知识. 汉语歧义的规则非常复杂，神经元的学习过程是一个循序渐进的过程，加大样本的训练次数，可以使切分词语的精度得到提高. 神经网络也是一个动态学习的过程，在已得到训练的神经网络中，如果以后输入的语句与原有的切分规则近似，则可以输出与样本近似的切分结果；如果以后输入的语句与原有的规则有较大差别，则神经网络把它看成新的切分规则进行学习[2]. 分词模型流程图见图 1.

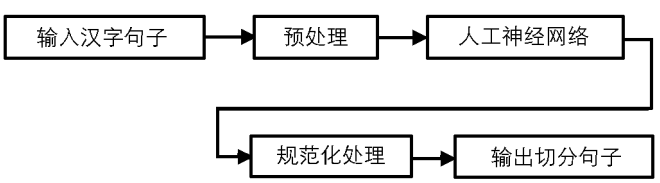


图 1 模型流程图

汉语分词的难点是对歧义字词的切分，神经网络的每一个输出节点代表一种切分方式，可以运用已有的知识切分汉语. 我们从国内公开发表的有关汉语分词的论文所提出的典型歧义语句中挑选出 4 句具有代表性的句子作为试验样本. 例如：样本 1：“物理学起来很难. ”它的切分规则为：如果歧义字段的后键词有趋向动词或助词，则尾字单切，否则该歧义字段成词[3]. 期望输出为 0110100. 其他样本的切分规则见参考文献[3].

① 收稿日期：2006－11－03
基金项目：成都信息工程学院科研基金资助 (CRF200624).
作者简介：陈 琳(1983－)，女，四川成都人，硕士研究生，主要从事人工智能、智能算法的研究.

据此建立相应的神经网络，它具有 13 个输入神经元，10 个输出神经元，隐含层有 60 个神经元. 其训练样本如表 1 所示. 根据大量的实验，我们把网络输出值 $Y_i$($i$=1，…，10)中小于 0.3 大于等于 0 的作为 0 看待，大于等于 0.7 的数值视作 1.

**表 1　模型训练样本**

| 样例编号 | 样例内容 | 期望输出 |
|---|---|---|
| A | 物理－学－起来－很难 | 0110100 |
| B | 他－吃－烤白薯 | 11000 |
| C | 他－学会－了－解－方程 | 111110 |
| D | 他－从－车上－下－来 | 110111 |

## 2　基于遗传算法的神经网络训练方法

遗传算法(GA，genetic algorithm)是一类借助于生物界自然选择和自然遗传机制的随机化搜索算法. 它是一种群体性操作，操作对象是群体中的所有个体，通过选择、交叉和变异等操作产生新一代群体. 作为一种求解问题的高效并行的全局搜索方法，其主要特点是群体搜索策略和群体中个体之间的信息交换，能在搜索过程中自动获取和积累有关搜索空间的知识，自适应地控制搜索过程以求得最优解或近似最优解. 它操作的对象为群体中的所有个体，通过对所求问题的解空间进行编码而得到. 主要操作有选择、交叉、变异三种. 鉴于操作个体的适应度，选择操作以一定准则，选取一定数量的个体作父代. 而交叉操作则对父代进行随机配对. 变异操作是以一定概率交换部分遗传信息，其按位进行，改变个体的编码位[4].

将 BP 算法和遗传算法的优点结合就形成一种混合训练算法—GA-BP 算法. GA-BP 算法是在 BP 算法之前，先用 GA 在随机点集中遗传出优化初值，以此作为 BP 算法的初始权值，再由 BP 算法进行训练，这就是 GA-BP 算法的原理.

设有三层 BP 网络，$I_i$ 为输入层中第 $i$ 个结点的输出；$H_i$ 为隐含层中第 $i$ 个结点的输出；$O_i$ 为输出层中第 $i$ 个结点的输出；$\mathrm{WIH}_{ij}$ 为输入层中第 $i$ 个结点与隐含层第 $j$ 个结点的连接权值；$\mathrm{WHO}_{ji}$ 为隐含层中第 $j$ 个结点与输出层第 $i$ 个结点的连接权值[5]. 遗传算法学习 BP 网络的步骤如下：

1) 编码. 采用了压缩二进制编码，汉字二进制码由 16 位减为 13 位.

2) 码串的组成形式. 遗传算法中一个个体可用一个码串表示，将权值向量 $\mathrm{WIH}_{s_1R}$、$\mathrm{WHO}_{1s_1}$ 和阈值向量 $B_{s_11}$、$B_{11}$ 按照顺序编成一个码串，其中 $S_1$ 的值为 1，…，60、$R$ 的值为 1，…，13. 从而产生一个个 $S_k$($k$=1，…，$N$，$N$ 为种群规模).

3) 初始化种群 $P$，包括交叉规模、交叉概率 $P_c$、突变概率 $P_m$ 以及对任一 $\mathrm{WIH}_{ij}$ 和 $\mathrm{WHO}_{ji}$ 初始化. 我们设初始种群是 30，交叉概率 $P_c$ 为 0.95，突变概率 $P_m$ 为 0.09.

4) 适应度函数

计算每一个个体评价函数，并将其排序；可按下式概率值选择网络个体：

$$p_s = f_i / \sum_{i=1}^{N} f_i$$

其中 $f_i$ 为个体 i 的适配值，可用误差平方和 $E(i)$来衡量，即：

$$f_i = 1/E(i) \qquad E(i) = \sum_{P}\sum_{R}(V_K - T_K)^2$$

其中：$i$=1，…，$N$ 为染色体数；$k$=1，…，10 为输出层节点数；$p$=1，…，4 为学习样本数；$V_k$ 为神经元 $k$ 的实际输出，$T_k$ 表示相应的目标输出.

GA-BP 算法的计算过程流程图如图 2 所示.

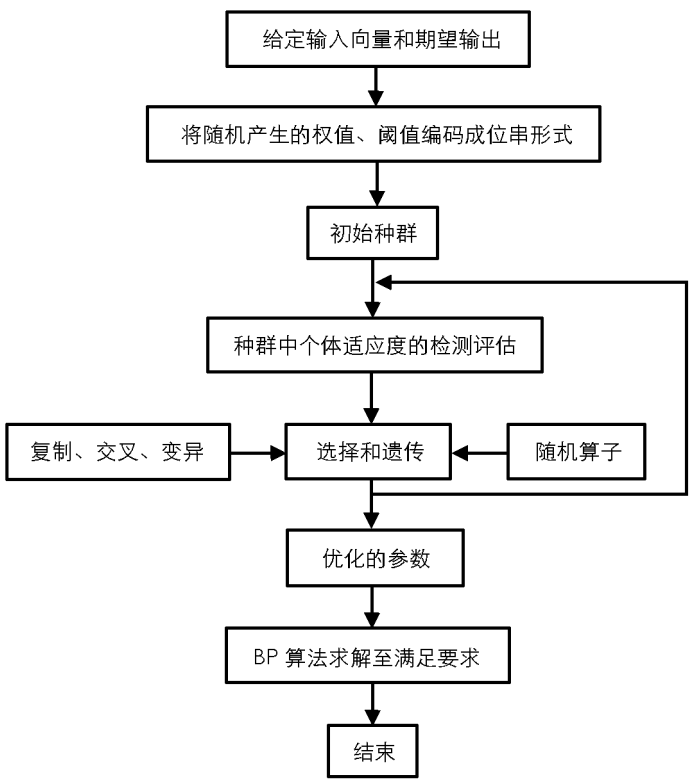


图 2　GA-BP 算法流程图

## 3 试验及结果分析

将“物理学起来很难”. 切分为:“物理”—“学”—“起来”—“很难”. 样本期望值:0110100.

为了便于说明,我们先把本文提出的基于遗传神经网络的模型称为模型 1(Model1),把经典 BP 神经网络模型称为模型 2(Model2). 用 Matlab 编程,GA 进行了约 70 代的遗传操作达到了目标值 $\varepsilon_{GA}$;BP 算法进行了 45 步收敛到指定精度 $\varepsilon_{BP}$ ($\varepsilon_{BP}=0.001$),总运行时间 elapsed_time= 26.2340s. 模型 1 训练过程误差曲线分别如图 3 所示. 误差平方和曲线和适应度曲线见图 4. 回想结果 TT=[0.003 0 0.982 1 1.004 3 0.009 4 0.991 2 0.008 3 0.017 6 0.000 7 0.000 7 0.000 7].

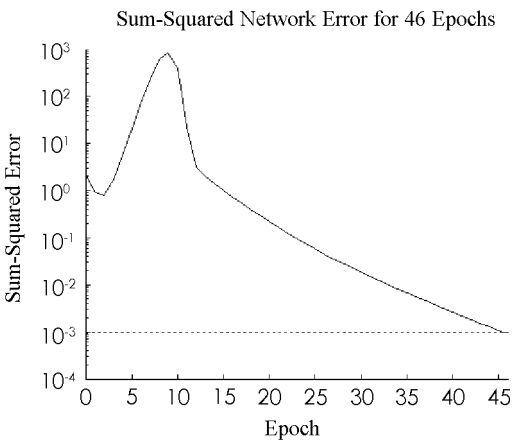


图 3 模型 1 训练过程误差曲线

为便于比较,图 5 给出了纯 BP 算法的训练目标曲线,这里 err_goal=0.001, lr=0.01. 其回想结果 TT=[0.078 6 0.818 5 0.903 1 0.123 9 0.883 1 0.100 2 0.134 2 0.045 1 0.045 1 0.045 1]. 运行时间 elapsed_time = 80.453 0 s. 由图 3 所示,经过加速的模型 1 在第 45 步就已经收敛,网络的误差就达到了 0.001 的要求.

而由图 5 所示,模型 2 经过 3000 次的训练以后,曲线只收敛于 0.01,远远不能满足之前误差要求(0.001). 虽然,网络初始权值对网络的收敛速度有一定影响,但是经过近 50 次的试验,模型 1 训练步数基本维持在 50 次以下,而模型 2 的训练步数在设置的训练步数(3000 次)中,均没有达到收敛. 由此,可见模型 1 无论是在收敛速度还是在运行时间上,都较为完美地解决了模型 2 收敛速度慢的难题,收敛速度提升了 600 多倍.

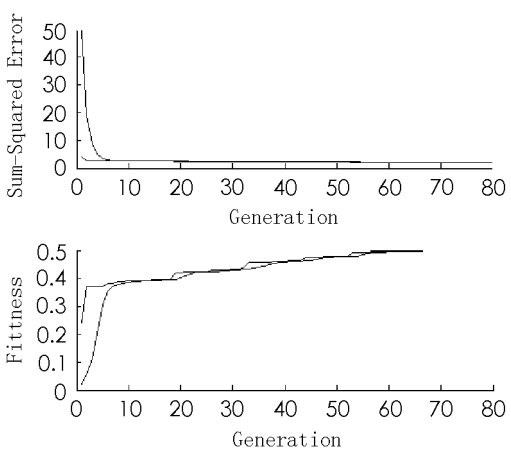


图 4 遗传算法误差平方和曲线、适应度曲线

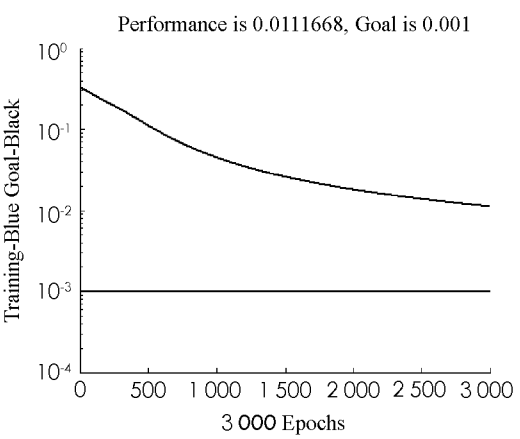


图 5 模型 2 训练过程误差曲线

此外,还作了各种测试(比如增加减少隐含节点数、增加初始种群数等),来检验经过优化的模型 1 的网络适应性. 对比模型 1 和模型 2 的测试结果,表明模型 1 对各种网络的误差平方函数同样具有快速收敛性. 各样本网络输出精度比见表 2(限于篇幅,表 2 中网络输出值 $Y_i$ 中 $i=1,\cdots,8$).

**表 2 模型 1 网络样本输出精度比**

| 样例编号 | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ | $Y_5$ | $Y_6$ | $Y_7$ | $Y_8$ |
|---|---|---|---|---|---|---|---|---|
| A | 0.003 0 | 0.982 1 | 1.004 3 | 0.009 4 | 0.991 2 | 0.008 3 | 0.017 6 | 0.000 7 |
| B | 0.999 4 | 0.996 9 | 0.001 1 | 0.001 1 | 0.001 1 | 0.001 5 | 0.001 5 | 0.001 5 |
| C | 0.984 9 | 0.957 3 | 1.003 8 | 0.916 1 | 1.013 7 | 0.002 6 | 0.009 3 | 0.003 2 |
| D | 0.984 5 | 0.912 4 | 0.005 4 | 1.004 0 | 0.976 6 | 0.900 7 | 0.002 0 | 0.002 4 |

**表 3　模型 2 网络样本输出精度比**

| 样例编号 | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ | $Y_5$ | $Y_6$ | $Y_7$ | $Y_8$ |
|---|---|---|---|---|---|---|---|---|
| A | 0.078 6 | 0.818 5 | 0.903 1 | 0.123 9 | 0.883 1 | 0.100 2 | 0.134 2 | 0.045 1 |
| B | 0.814 9 | 0.940 8 | 0.047 6 | 0.029 3 | 0.041 5 | 0.029 2 | 0.029 2 | 0.029 2 |
| C | 0.889 5 | 0.922 9 | 0.871 8 | 0.815 3 | 0.996 2 | 0.040 6 | 0.060 5 | 0.060 5 |
| D | 0.950 1 | 0.887 5 | 0.009 2 | 0.871 4 | 0.923 0 | 0.933 1 | 0.045 1 | 0.065 1 |

## 4　结束语

本文讨论了基于遗传算法的 BP 神经网络算法来改进神经网络分词模型的原理和构建，并通过 MATLAB 做了大量的仿真试验. 从仿真试验结果可以看出：改进后的神经网络与传统的神经网络模型相比，改进后的神经网络既能避免 BP 算法陷入局部最小点、收敛速度慢的问题，又能克服 GA 以类似穷举的形式寻找最优解而引起的搜索时间长、速度慢、易陷于局部极小等缺点. 因此，进一步提高该模型在分词领域中的实用性和分词效率，对于中文信息的自动化处理具有非常重要的意义.

**参考文献：**

[1] 阎平凡，张永久. 人工神经网络与模拟进化计算[M]. 北京：清华大学出版社，2000：291－301.
[2] 林亚平，李　彦，童调生等. 汉语自动分词中的神经网络技术研究[J]. 湖南大学学报，1997，24(6)：95－101.
[3] 尹　锋. 基于神经网络的汉语自动分词系统的设计与分析[J]. 情报学报，1998，17(1)：41－49.
[4] 严柏军，宋承天，王克勇，等. BP 神经网络和遗传算法在货车车锁检测中的应用[J]. 计算机辅助设计与图形学学报，2002，20(2)：1179－1183.
[5] 飞思科技产品研发中心. MATLAB6.5 辅助优化计算与设计[M]. 北京：电子工业出版社，2003：155－217.

# Optimization of Chinese Word Segmentation Based on Neural Network and Genetic Algorithm

CHEN　Lin，　HE　Jia

*Dept of Computer，Chengdu University of Information Technology，Chengdu 610225，China*

**Abstract**：For the Chinese word segmentation problem，the BP algorithm has relatively slow convergence speed. This paper introduces genetic algorithm to optimize Chinese word segmentation system based on neural network and the model has the advantages of both neural network and genetic algorithm. Experiment results show that the optimized model perfectly meets the requirement of the Chinese word segmentation system and can do word segmentation faster compared with the conventional BP neural network. The optimized model has the advantages of high accuracy and convergence speed.

**Key words**：neural network；Chinese word segmentation；Genetic Algorithm (GA)；BP genetic algorithm

**责任编辑**　张　枸