文章编号:1001 – 2486(2004)04 – 0112 – 03

基于 DCPM-FGA 的隐马氏模型的参数训练*

韩广军,罗　强,李　兵
(国防科技大学理学院,湖南 长沙　410073)

摘　要:建立了一种将直接比较比例方法与模糊遗传算法结合在一起用于隐马氏模型参数训练的方法。针对不同的码本大小进行了仿真试验。结果表明,在码本较大时该方法效果要好于 B-W 算法和 Viterbi 算法,码本较小时两种方法性能相近。

关键词:隐马氏模型;直接比较比例方法 – 模糊遗传算法;码本
中图分类号:O211.62　　文献标识码:B

Parameters Training of HMM Based on DCPM-FGA

HAN Guang-jun, LUO Qiang, LI Bing
(College of Science, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Using DCPM-FGA Direct comparison-proportional method and Fuzzy genetic algorithm to train the parameters of HMM is a very good idea. Different simulations are made for different codebook sizes. The results of simulations show that when the codebook size is bigger, the DCPM-FGA is better than B-W algorithm, otherwise, it has almost the same performance with the above algorithms.

Key words: HMM; DCPM-FGA; codebook

HMM 是一种重要的信号与系统分析工具。在实际的应用中它扮演着越来越重要的角色。在运用 HMM 技术解决问题的过程中,HMM 的参数训练又是最为关键的一步。由[1]知,传统的训练方法容易陷入局部最优解,所以考虑具有全局搜索功能的遗传算法。而大量的试验证明,DCPM-FGA 与以往那些技术相比无论是最好解的质量还是解的平均质量都显著地超过了那些技术,且 DCPM-FGA 处理约束条件并搜索到高性能解的能力也是令人满意的。所以采用 DCPM-FGA 进行 HMM 的参数训练。

1　方法介绍与 HMM 参数训练

1.1　DCPM-FGA 介绍

将 DCPM(direct comparison-proportional method:直接比较比例方法)与 FGA(fuzzy genetic algorithm:模糊遗传算法)结合即为 DCPM-FGA 方法。它不必显示地利用目标函数的某些性质,可以采用非处处可微的罚函数形式。同时对于 GA 所实行的选择方法来说,适应值提供的惟一功能就是对给定个体进行比较和排序。当采用竞争选择法时,可以完全不必显示地生成新的函数,而只需通过建立基于适应值函数和罚函数的直接比较机制,就可避免选择罚因子,为处理约束问题提供了一种全新的思路。DCPM 方法提供了种群中个体之间的比较准则,并给出在种群中保持一定比例的不可行解的适应性策略。

1.2　运用 DCPM-FGA 进行参数训练的算法流程

(1) 确定实数形式编码;

(2) 确定要使用的选择方法、遗传算子及各参数值,包括 ε, p 和 K 的值。

(3) 设置代数 $t = 0$;

(4) 随机生成初始化种群 $X(0)$;

(5) 计算每个个体的适应值和罚函数值;

* 收稿日期:2004 – 02 – 16
基金项目:国家自然科学基金资助项目(60375023);国防科技大学基础研究资助项目
作者简介:韩广军(1979—),男,硕士生。

(6) 判断有无可行解,有则计算不可行解的比例;

(7) 当终止条件不满足时,令 $t = t + 1$;

(8) 根据适应值、罚值和 ε 的值,按照比较准则用竞争选择法从 $X(t-1)$ 中选择出 $X'(t)$;

(9) 根据交叉概率 p_c 用交叉算子对 $X'(t)$ 中的个体进行交叉操作,得种群 $X''(t)$;

(10) 根据变异概率 p_m 用变异算子对 $X''(t)$ 中的个体进行操作,得到第 t 代种群 $X(t)$;

(11) 计算 $X(t)$ 中个体的适应值和罚值,判断有无可行解,若有则计算不可行解的比例;

(12) 在群体中出现第一个可行解后,每隔 k 代,按照上面所述调整 ε 的值。

1.3 遗传策略

在运用 GA 处理优化问题时,较好的遗传策略往往导致能够以较大的概率较快地搜寻到全局最优解。遗传策略包括很多的内容,简单一点,它主要涉及七大要素 :编码设计、初始种群设定、种群规模设定、适应值函数设计、遗传算子设计、控制参数设计、终止条件设定。

2 仿真试验及结果分析

针对一个简单的 HMM 模型在以下遗传策略下运用以上方法用 Matlab 数学软件进行编程并执行了数次。其中状态集设为 $S = \{1,2,3\}$,码本大小分别为 128 和 4,观察量分别为 130 和 5,最大代数 $T = 2000$。

(1) 适应函数值。在大多数情况下,适应值函数取为目标函数。以基于前向概率的 $P(O|\lambda)$ 为适应值函数。

(2) 编码选择。一般来说,应根据实际问题选择编码方式。GA 对编码要求并不苛刻。大量试验证明,对同一优化问题采用二进制编码和实数编码并不存在显著的差异[3]。这里,选用实数编码。

(3) 种群规模设定。它主要取决于个体中所含变量的个数[4],种群规模分别为 400 和 40。

(4) 初始种群生成。若一开始就让解中每个变量在 [0,1] 中随机产生,则搜索空间太大,花费时间很长,甚至是不可能做到的。所以,就让初始种群中大部分是可行解。

(5) 遗传算子设计。主要采用:

a. 算术交叉,b. 单点随机变异,c. 竞争选择法。

(6) 参数设定。$\varepsilon = 0.0001$, $K = 5$, $p = 0.1$, $p_c = 0.25$, $p_m = 0.012$,竞争规模为 2,精度为 0.0001。

(7) 终止条件设定。用最大代数终止法。

(8) 罚函数。采用绝对值罚函数形式,即

$$Viol(X) = \sum_{j=1}^{3} f_j(X)$$

$$f_1(X) = \left| \sum_{i=1}^{N} \pi_i - 1 \right|$$

$$f_2(X) = \sum_{i=1}^{N} \left| \sum_{j=1}^{N} a_{ij} - 1 \right|$$

$$f_3(X) = \sum_{j=1}^{N} \left| \sum_{k=1}^{M} b_j(K) - 1 \right|$$

(9) 加入控制早熟的因子

$$c_k = \min\{1, (F(X'(k))/F(X(k)))^k\}, F(X(k)) = \max\{f(X_i(k)), 1 \leqslant i \leqslant N\};$$

在选择第 $k+1$ 代时,先产生 [0,1] 中的一个随机数 ρ,然后有

① 若 $\rho \leqslant c_k$,则令 $X(k+1) = X'(k)$;

② 若 $\rho > c_k$,则令 $X(k+1) = X(k)$。

由于采用最大代数终止,所以每次得到不同的结果是合理的。做了很多次试验,当码本大小为 128 时,所得最好结果为 $\log P(O|\lambda) = -45.3477$,与 Baum-Welch 公式得到的 $\log P(O|\lambda) = -48.22$ 及 Viterbi

算法的 $\log P(O|\lambda) = -51.55$ 相比，明显提高了很多。当码本大小为 4 时，最好结果为 $\log P(O|\lambda) = -4.9477$，而用 B-W 方法得到的最好的结果为 -4.7531。值得注意的是选择概率，交叉概率以及变异概率的选择也是较困难的，在此仅仅根据经验对它们进行了选择。

-4.9477 对应的 λ 为：

$$\pi:(0.2703 \quad 0.5699 \quad 0.1598)$$

$$A:\begin{bmatrix} 0.2424 & 0.0866 & 0.6710 \\ 0.9928 & 0.0029 & 0.0043 \\ 0.2668 & 0.3614 & 0.3718 \end{bmatrix}$$

$$B:\begin{bmatrix} 0.0226 & 0.7916 & 0.0586 & 0.1271 \\ 0.3779 & 0.0826 & 0.2674 & 0.2721 \\ 0.1157 & 0.1478 & 0.5038 & 0.2327 \end{bmatrix}$$

下面是搜寻到解 λ 的途径的图形：

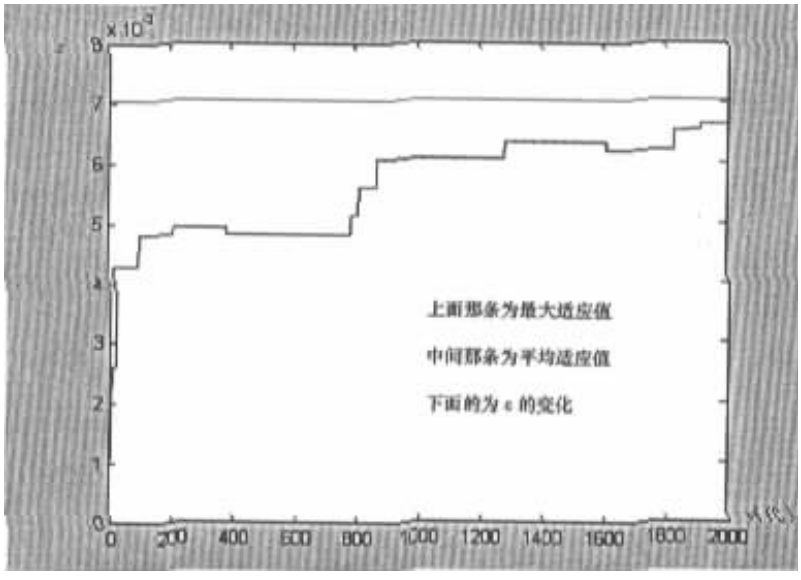

图 1 适应值函数（y 为代数，z 为适应值）

Fig.1 Fitness（y is generation number, z is fitness）

由于采取的是最大代数终止，所以没有能够看到它的收敛。但由于采取了竞争保留策略，故理论上它是一定收敛的。数次试验的图形都是收敛的，这里不再给出。显然，其平均适应值是在上升的，这充分说明了解的进化性。

3 结 论

试验证明 DCPM-FGA 对参数 ε 和 k 的不同选择，计算结果都差异不大。不足之处就在于参数的设定需要特别的谨慎，否则就达不到预期的效果。再者，初始化对结果也有影响。但是，由试验可知它却具有比较的优势。尽管如此，我们的结果也还是不理想，很多地方有待改进。

参 考 文 献：

[1] Janez Kaiser University of Maribor, Faculty of Electrical Engineering and Computer Science, Smetanora 17 Maribor Slovenia, Trainin of HMM with GA[EB]. www.dsplab.uni-mb.si/Clanki/janez/ast98.pdf.

[2] 张文修，梁怡. 遗传算法的数学基础[M]. 西安：西安交通大学出版社，2001.

[3] 云庆夏. 进化算法[M]. 北京：冶金工业出版社，2000.

[4] 李敏强，等. 遗传算法的基本理论与应用[M]. 北京：科学出版社，2002.

[5] Djuric P M, Chun Joon-Hwa. An MCMC Sampling to Estimation of HMMs[J]. IEEE Trans. Signal Processing, 2002, 50(5): 1113 – 1123.