文章编号: 1001-0920(2001)03-296-04

一种印刷体字符识别的新方法: 基于遗传算法的(0, 1, *)-矩阵法

郑朝晖, 裘聿皇, 陈峻峰
(中国科学院 自动化研究所, 北京 100080)

摘 要: 给出一种全新有效的快速算法。该方法通过合理的阈值将模板向量转化为(0, 1, *)-向量, 并充分考虑到代表样本与模板之间相关性的不同因素的不同重要性, 赋以相应的权系数, 并用遗传算法来确定阈值和权系数。印刷体邮政编码的实验结果表明, 该算法在大大缩短识别时间的同时, 识别率可达 98.1%, 而相同实验条件下应用传统模板匹配法时的识别率为 92.1%。

关键词: (0, 1, *)-矩阵; 遗传算法; 印刷体字符识别; 阈值; 权系数; 相关性
中图分类号: TP 18 文献标识码: A

Novel Method for Printed Character Recognition: (0, 1, *)-Matrix Based on GA

ZHENG Zhao-hui, QIU Yu-huang, CHEN Jun-feng
(Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: To reduce the computing complexity of printed character recognition and improve recognition rate as well, a new effective algorithm is proposed. Using two reasonable threshold values, the algorithm transforms real template vectors into (0, 1, *)-ones. Meanwhile, adequately considering the different weightiness of the four different factors denoting the pertinence between template and unknown sample, corresponding weight coefficients are allocated on them. Genetic algorithm is used to decide all these threshold values and weight coefficients. The method produces 98.1% of the recognition rate, which is better than 92.1% of conventional template-matching method under the identical experimental condition.

Key words: (0, 1, *)-matrix; genetic algorithms; printed character recognition; threshold value; weight coefficient; pertinence

1 引 言

印刷体字符识别在邮政系统、银行业务、名片识别和票据识别等方面都有重要应用。目前常用的方法有神经网络法、模板匹配法、子空间法及一些统计方法。这些方法主要是先提取样本的有效特征, 然后依据这些特征进行分类。

本文提出一种新的算法, 与利用神经网络进行特征训练的方法相比, 本文方法不需要提取特征, 也

收稿日期: 2000-01-31; 修回日期: 2000-04-29
基金项目: 国家自然科学基金项目(60075018)
作者简介: 郑朝晖(1976—), 男, 江苏扬州人, 博士生, 从事模式识别、图象处理等研究; 裘聿皇(1942—), 男, 浙江宁波人, 研究员, 博士生导师, 从事系统理论与应用、控制理论等研究。

不需要用大量样本进行反复训练, 而是考虑到印刷体字符相对稳定的特点, 直接基于象素进行整数运算, 从而大大简化了计算复杂性。与模板匹配法相比, 本文方法不涉及求样本间距离的浮点运算, 因而计算简单; 而且更全面地考虑了代表样本间相似性与相异性的因素, 并根据这些因素的不同重要性分别赋以不同的权系数, 因而更加科学、合理。

遗传算法是一种广泛应用的高效的随机化搜索与优化的方法。本文的重要参数设定均采用遗传算法。

2 算　　法

1) 将所有输入的样本图象归一化为 64 × 64, 构成 4 096 × 1 的列向量(实向量)。假设样本类别数为 n, 分别计算训练样本中这 n 类样本向量的象素算术平均值, 得到 n 个实向量 $\rho_1, \rho_2, \cdots, \rho_n$, 其中 $\rho_i (i = 1, 2, \cdots, n)$ 为 4 096 × 1 的列向量。

2) 通过设定两个阈值 $t_1, t_2 (1 > t_1 > 0.5 > t_2 > 0)$, 将实向量 $\rho_1, \rho_2, \cdots, \rho_n$ 转化为(0, 1, *) –向量 $\alpha_1, \alpha_2, \cdots, \alpha_n$。当象素值大于 t_1 时取 "1", 小于 t_2 时取 "0", 否则取 "* "。$\alpha (i = 1, 2, \cdots, n)$ 表示第 i 类样本的模板向量, 为 4 096 × 1 的列向量。

3) 建立模板矩阵 A (A 为 $n \times 4\,096$ 的矩阵), 其每一行依次为 $\alpha_1^{\mathrm{T}}, \alpha_2^{\mathrm{T}}, \cdots, \alpha_n^{\mathrm{T}}$。记待识别的印刷体字符所构成的 4 096 × 1 的列向量为 β。

4) 定义矩阵与向量的运算 $\otimes$。令矩阵 $D_{p\times q}$ 的元素为 $d_{ij} (i = 1, 2, \cdots, p, j = 1, 2, \cdots, q)$, 向量 $\epsilon_{q\times 1}$ 的元素为 $b_k (k = 1, 2, \cdots, q)$, 则定义

$$D \otimes \epsilon = \left[\sum_{j=1}^{q} d_{1j} \otimes b_j, \sum_{j=1}^{q} d_{2j} \otimes b_j, \cdots, \sum_{j=1}^{q} d_{nj} \otimes b_j \right]^{\mathrm{T}} \quad (1)$$

其中 $\otimes$ 的运算规则为

$$0 \otimes 1 = 1 \otimes 0 = -1 \quad (2)$$

$$a \otimes * = * \otimes a = 0, \quad a = 0, 1 \text{ 或 } * \quad (3)$$

$$b \otimes b = 1, \quad b = 1 \text{ 或 } 0 \quad (4)$$

取反规则为

$$\bar{1} = 0, \quad \bar{0} = 1, \quad \bar{*} = * \quad (5)$$

矩阵 A 取反 $\bar{A}$, 即将矩阵的每一个元素按规则(5) 取反。同样, 向量 β 取反 $\bar{\beta}$ 亦如此计算。

定义 $\gamma^{(0,0)}, \gamma^{(0,1)}, \gamma^{(1,0)}, \gamma^{(1,1)}$ 如下

$$\begin{cases} \gamma^{(0,0)} = \bar{A} \otimes \beta, & \gamma^{(1,1)} = A \otimes \beta \\ \gamma^{(1,0)} = A \otimes \bar{\beta}, & \gamma^{(0,1)} = \bar{A} \otimes \bar{\beta} \end{cases} \quad (6)$$

其中, A 是模板矩阵, β 是待识别样本向量, $\gamma^{(\cdots)} (i, j = 0, 1)$ 为 $n \times 1$ 列向量。

$\gamma^{(0,0)}$ 是模板与待测样本间相似性的度量。其元素 $c_i^{(0,0)} (i = 1, 2, \cdots, n)$ 为非负整数, 表明模板向量 $\alpha (i = 1, 2, \cdots, n)$ 与待测样本向量 β 对应位都含有 "0" 的个数[1, 2]。

$\gamma^{(1,1)}$ 是模板与待测样本间相似性的度量。其元素 $c_i^{(1,1)} (i = 1, 2, \cdots, n)$ 为非负整数, 表明模板向量 $\alpha (i = 1, 2, \cdots, n)$ 与待测样本向量 β 对应位都含有 "1" 的个数。

$\gamma^{(1,0)}$ 是模板与待测样本间相异性的度量。元素 $c_i^{(1,0)} (i = 1, 2, \cdots, n)$ 为非正整数, 其元素绝对值表明模板向量 $\alpha (i = 1, 2, \cdots, n)$ 中元素为 "1", 而 β 中对应元素为 "0" 的元素个数。

$\gamma^{(0,1)}$ 是模板与待测样本间相异性的度量。元素 $c_i^{(0,1)} (i = 1, 2, \cdots, n)$ 为非正整数, 其元素绝对值表明模板向量 $\alpha (i = 1, 2, \cdots, n)$ 中元素为 "0", 而 β 中对应元素为 "1" 的元素个数。

下面依据相关度向量

$$\begin{aligned} \sigma = & [\gamma^{(1,1)}, \gamma^{(0,0)}, \gamma^{(0,1)}, \gamma^{(1,0)}] \times \\ & [w^{(1,1)}, w^{(0,0)}, w^{(0,1)}, w^{(1,0)}]^{\mathrm{T}} = \\ & [c_1^{(1,1)} w^{(1,1)} + c_1^{(0,0)} w^{(0,0)} + c_1^{(0,1)} w^{(0,1)} + \\ & c_1^{(1,0)} w^{(1,0)}, \cdots, c_n^{(1,1)} w^{(1,1)} + c_n^{(0,0)} w^{(0,0)} + \\ & c_n^{(0,1)} w^{(0,1)} + c_n^{(1,0)} w^{(1,0)}]^{\mathrm{T}} \end{aligned} \quad (7)$$

对样本进行分类。其中, $w^{(1,1)}, w^{(0,0)}, w^{(0,1)}, w^{(1,0)}$ 分别对应于 $\gamma^{(1,1)}, \gamma^{(0,0)}, \gamma^{(0,1)}, \gamma^{(1,0)}$ 的权系数。

σ 为 $n \times 1$ 列向量, 其元素 $s_i (i = 1, 2, \cdots, n)$ 为待测样本 β 与第 i 类模板向量 α 的相关度。根据本文算法, s_i 越大表示待测样本 β 与第 i 类模板向量 α 越相关, 也就越有可能被归为 i 类。取

$$j = \arg \operatorname{Max} s_i, \quad i = 1, 2, \cdots, n \quad (8)$$

可将待测样本 β 归于 j 类; 若 j 不止一个, 则拒识之。

3 用遗传算法确定权系数与阈值

由上述算法分析可知, 阈值 t_1, t_2 以及权系数 $w^{(1,1)}, w^{(0,0)}, w^{(0,1)}, w^{(1,0)}$ 是本文算法的待求参数, 对最终识别率有至关重要的影响。因此, 采用有效合理的搜索算法来求解这些参数是非常必要的。

遗传算法是一种以自然选择和遗传理论为基础, 将生物进化过程中的适者生存规则与种群内部染色体的随机信息交换机制相结合的全局最优化搜索算法[3]。下面利用遗传算法来求解这 6 个参数。

1) 编码: 采用二值编码。设定权系数[0, 15] 之间的整数, 阈值 t_1, t_2 分别为[0. 6, 0. 95] 和[0. 05,

0.4] 之间的实数。故权系数 $w^{(1,1)}, w^{(0,0)}, w^{(0,1)}, w^{(1,0)}$ 可用 4 位二进制串来表示。其映射精度 Π_w 为

$$\Pi_w = \frac{15-0}{15-0} = 1$$

而阈值可用 3 位二进制串来表示, 其映射精度 Π_1, Π_2 分别为

$$\Pi_1 = \frac{0.95-0.6}{7-0} = 0.05$$

$$\Pi_2 = \frac{0.4-0.05}{7-0} = 0.05$$

2) 适应度的选取与定标: 以识别率作为衡量适应度的唯一指标。在每个个体的适应度正比于该个体对应的阈值和权值分布下测试样本集的识别率。为了加大选择力度, 必须对适应度进行定标。记识别率为 r, 则适应度函数为

$$f(r) = \frac{1}{0.21-(r-0.8)u(r-0.8)} \quad (9)$$

其中 $u(\cdot)$ 为阶跃函数。

3) 遗传算子: 采用常用的选择、交叉和变异算子。选择采用赌轮选择机制, 交叉采用单点交叉, 变异采用小概率常规位突变。具体步骤如下:

① 随机产生一定数目的染色体串长为 22 位的初始群体;

② 判断是否满足终止准则, 若满足则停止;

③ 解码个体, 产生每一个体对应的阈值 t_1, t_2 和权系数 $w^{(1,1)}, w^{(0,0)}, w^{(0,1)}, w^{(1,0)}$; 根据阈值可求得该个体对应的模板矩阵, 结合权系数按上述算法将所有训练样本归类, 计算出该个体对应下的识别率, 进而求出适应度;

④ 进行遗传算法迭代: 选择 -交叉 -变异;

⑤返回步骤 ②。

4 实例分析

现以印刷体数字识别为例, 并与传统模板匹配法进行比较, 用以说明本文算法的有效性。传统模板匹配法通过计算待测样本与模板向量之间的距离进行分类, 其中模板向量采用浮点数表示, 因此主要进行浮点运算。本文算法利用遗传搜索所得的阈值将模板向量(0, 1, *) 化, 然后进行加权, 因而大多是整数运算, 大大减少了计算复杂性, 节省了计算时间。更为重要的是, 本文算法考虑了样本相似性与相异性的不同重要性, 并用遗传算法为不同参数赋以不同的权系数。与传统模板匹配法相比, 本文算法在缩短计算时间的同时, 识别率也得到大大提高。

本实验的样本取自中国科学院自动化研究所文字识别工程中心收集的印刷体邮政编码样本库。训练样本为 4 000 个, 测试样本为 3 000 个。对所有原始样本只规格化, 而未进行任何去噪声与平滑等预处理。遗传算法的最大迭代代数为 30 代。交叉概率为 85%, 变异概率为 5%, 种群大小为 50, 染色体串长为 22 位。

实验结果如下: 用遗传算法所求得的 6 个参数分别为

$$t_1 = 0.85, \quad t_2 = 0.05$$

$$w^{(1,1)} = 3, \quad w^{(0,0)} = 1$$

$$w^{(0,1)} = 15, \quad w^{(1,0)} = 9$$

两种识别方法的比较参见表 1, 识别率与遗传代数的关系如图 1 所示。其中 —●— 为群体最大识别率, —■— 为群体平均识别率。

表 **1** 两种识别方法的比较

识别方法	识别率(%)
传统的模板匹配法	92.1
基于遗传算法的(0, 1, *) -矩阵法	98.1

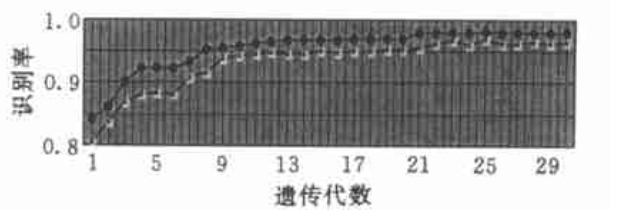

图 **1** 识别率与遗传代数关系

从实验结果看, 所求的权系数分布很不均匀, 说明代表样本间相关性的不同因素在识别中所起的作用各不相同。这也充分说明了加权的必要性。其中, 代表样本间相异性的因素远比代表样本间相似性的因素重要, 因此不应只考虑待测样本与模板之间的相似性因素, 而更应考虑它们之间的相异性因素。传统的方法通常是按距离将样本分类的, 这种方法从根本上忽略了代表样本间相似性和相异性因素的不同重要性, 因而其识别率要低得多。

从图 1 可以看出, 遗传算法经 30 次迭代后, 最终获得了满意的结果。本文算法的思想对手写体字符识别也有重要的参考价值。

(下转第 302 页)

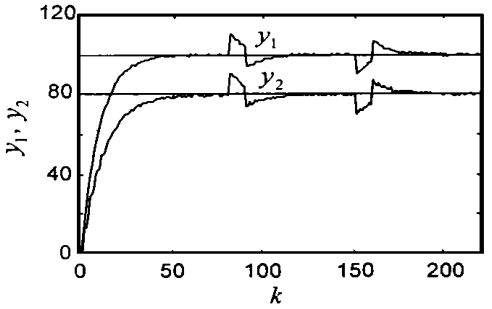

图 1 输出 y_1, y_2

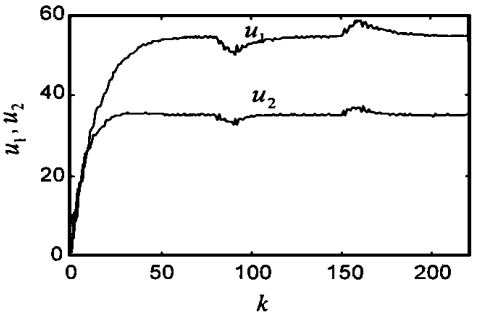

图 2 控制 u_1, u_2

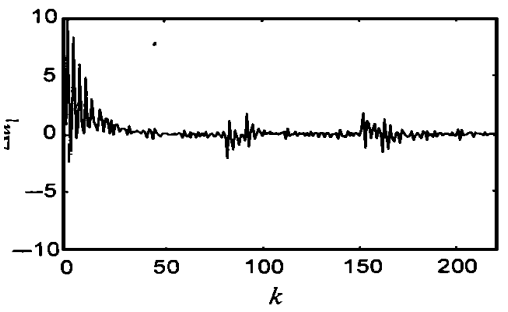

图 3 控制 u_1 变化率

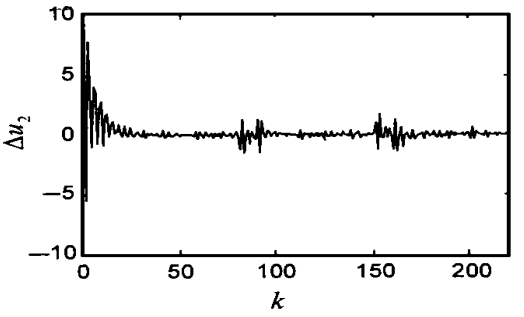

图 4 控制 u_2 变化率

6 结 语

本文通过线性化处理, 把有约束预测控制的滚动优化问题转化为线性规化问题。理论推导及仿真实例均表明, 目的规划是求解多目标、多变量和有约束预测控制滚动优化问题的有效方法。由于采用线性优化策略, 该算法的计算效率高, 数值稳定性好。

参考文献:

[1] 王伟. 广义预测控制理论及其应用[M]. 北京: 科学出版社, 1998. 188-189.

[2] 舒迪前. 预测控制系统及其应用[M]. 北京: 机械工业出版社, 1996. 377-381.

[3] Yang S H, Wang X Z, Mcgreavy C. A multivariable coordinated control system based on predictive control strategy for FCC reactor-regenerator system[J]. Chem Eng Sci, 1996, 51(11): 2977-2982.

[4] Campo P J, Morari M. Robust model predictive control [A]. Proc of the 1987 American Control Conf[C]. Green Valley: American Automatic Control Council, 1987. 1021-1026.

(上接第 298 页)

5 结 语

由于汉字种类繁多, 印刷体汉字识别系统一般采用多级分类器, 在每一级中又采用不同的分类算法。本文侧重于比较两种单一算法的有效性, 而不是对不同算法组合的有效性进行比较。邮政编码是印刷体字符的小子集, 不需要进行多级粗分, 用一种分类算法就能达到较为满意的结果, 因此可更方便有效地进行不同识别算法的优越性比较。

本文提出一种新的模板匹配方法, 通过设定合理的阈值将浮点运算转化为整数运算, 大大减少了识别算法的计算复杂性, 缩短了识别时间。算法中充分考虑了代表样本之间相关性的因素在识别过程中所起的不同作用, 并赋以相应的权系数。利用遗传算法的优良全局搜索能力, 确定上述阈值和权系数, 取得了令人满意的效果。

参考文献:

[1] 裘聿皇. 成组技术与相似性系数[J]. 自动化学报, 1999, 25(2): 275-278.

[2] 裘聿皇. 我国草兔的聚类研究[J]. 兽类学报, 1989, 9(3): 168-172.

[3] 潘正君, 康立山, 陈毓屏. 演化计算[M]. 北京: 清华大学出版社, 1998.