模式识别的改进算法研究

沈学利，张家明
（辽宁工程技术大学 电子与信息工程学院，辽宁 葫芦岛 125105）

摘 要：针对模式识别问题，设计了一种基于字符串编码的匹配算法，实现了对样本的二次筛选和两种不同的模式匹配算法的交叉匹配，不仅可以有效地减少“洞”的存在，而且可以生成高质量的检测样本，从而增强识别效率、降低时间消耗。对数字识别的实验结果表明，该匹配算法具有很高的识别率，能有效降低错误率，得到稳定的满意解，具有良好的适应性、学习性、开放性和鲁棒性。

关键词：二次筛选；交叉匹配；识别效率；学习性
中图分类号：TP391.4 **文献标识码**：A

Study of improvement algorithm for pattern recognition

SHEN Xue Li, ZHANG Jia Ming
(School of Electronic & Information Engineering, Liaoning Technical University, Huludao 125105, China)

Abstract: In this paper,we designed a kind of matching algorithm based on string coding,to realize the sample quadratic sieve and the cross-matching of two different matching algorithms.This designed algorithm could avoid the “hole”effectively,generate detection samples of high quality,strength the identification efficiency and reduce the time wasted.Experiments result shows that this algorithm has good adqptability,openess and robustness.

Key words: guadratic sieve; alternate to mate; distinguish efficiency; study

人工智能的迅速发展引起了众多学科和不同专业背景学者们的日益关注，成为一门广泛交叉的前沿科学。模式识别是人工智能研究的重要领域之一，模式识别是指计算机代替人类或帮助人类感知模式，是对人类感知外界功能的模拟。它研究的是计算机模式识别系统，也即使一个计算机系统具有模拟人类通过感官接受外界信息、识别和理解周围环境的感知能力。模式识别是一个不断发展的新学科，它的理论基础和研究范围也在不断发展。模式识别的方法很多，例如决策理论方法、句法方法，大多通过样本的相似程度进行识别。本文设计的样本是用固定长度的二进制字符串来模拟，样本识别通过海明距离和R连续位匹配实现。

1 模式识别算法分析

在模式识别中，匹配规则是一个关键点。匹配规则分为完全匹配和部分匹配。如果2个等长字符串的每个对应位上的符号都相同，那么这样的匹配称为完全匹配。然而实际的模式识别中，完全匹配只是其中的一个特例，大多数情况下人们通过不完全匹配进行模式识别，这样可以在尽可能短的时间内识别尽可能多的字符串。最常用的模式匹配规则有海明距离和R连续位匹配。

海明距离有多种实现方式，例如相对海明距离、加权海明距离、相对加权海明距离，这些都是在基本海明距离的基础上改进而来的，没有实质的变化[1]。海明距离的公式如下：

$$D=\sum_{i}^{l}\delta_i \qquad \delta_i=\begin{cases}1, b_i=c_i\\ 0, \text{otherwise}\end{cases}$$

2个样本分别用$S=b_1b_2b_3\cdots b_L$和$T=c_1c_2c_3\cdots c_L$表示，即海明距离的大小与样本的匹配程度成正比。

R连续位匹配规则是指，对于任意2个长度为L的字符串a和b，如果在相对应的位置上至少有R个连续的位相同，那么a和b相匹配。

例如，长为8的位串a=10010110，b=11010100，若定义R=4，则串a与串b匹配，因为位串的第3至第6位对应相等。显然，当R<4时，位串a、b仍然匹配。

在利用R连续位匹配的时候会出现“洞”，既能与待识别对象匹配，也能和非识别对象匹配，无法生成一个

有效的检测器。随着 R 的减少，“洞”的数量会增加，R 取值 8 时，每一个串只能匹配自身，不存在“洞”，但是增加了时间复杂度降低了检测效率。传统的避免“洞”的方法是采用多种表示法。它通过一个随机产生的掩码来过滤引入的字符串。例如，给定 2 个字符串 S_1=01101011，S_2=00010011，以及一个掩码 $\wedge$，通过随机产生，如 $\wedge$ =1–6–2–5–8–3–7–4(置换顺序，新串相应在原串中的位置)，那么 $\wedge(S_1)$=00111110，$\wedge(S_2)$=00001011。使用连续位规则 R=3，那么 S_1 匹配 S_2，而 $\wedge(S_1)$不匹配 $\wedge(S_2)$。这个通过转变表示方式的方法虽能避免漏洞，但是并没有在这个转变过程中实现样本的优化，而且增加了时间的消耗[2]。

同样，使用海明距离匹配规则也存在漏洞。事实上，目前使用的所有带有一定匹配率的匹配规则都避免不了漏洞。鉴于此，本文提出一种改进的算法。在改进算法中为了提高识别效率、降低时间消耗，使用海明距离和 R 连续位匹配 2 种不同的匹配方式交叉进行模式识别，这样在不同规则下产生的不同漏洞可以互相弥补，不仅可以有效减少“洞”的存在，而且还可以避免使用其他表示方式消耗时间。所以，交叉使用 2 种不同的匹配规则不仅可以减少“洞”的数量，而且可以生成高质量的样本。

另外，结合生物学上基因突变的思想，大量复制和变异与识别对象匹配度高的样本，而对于匹配度低的样本重复这个变异和选择的过程，保证样本和识别对象的匹配度逐步增大[3]。

2 改进算法的实现

该改进算法的实现步骤如下：

(1)随机产生 N 个样本集合；

(2)对这个样本集合与识别对象进行海明距离计算；

(3)选择海明距离最大的 n_1 个样本，组成样本集合 M_1；

(4)转到步骤(1)，生成样本集合 $M_2\cdots M_n$，直到生成 N 个样本；

(5)把海明距离相同的样本与识别对象进行 R 连续位匹配。R 连续位匹配，即当 2 个字符串至少存在连续 R 位相同时才发生匹配；

(6)根据海明距离和 R 连续位匹配对样本进行降序排列；

(7)对 N 个样本根据排列顺序进行复制，复制的数量与海明距离和 R 连续位匹配成正比。

样本复制的总数量公式：

$$N_d=\sum_{i=1}^{N}(\frac{N}{\lambda}-i)$$

N_d 是复制的总数量，λ 是给定的参数。

(8)对复制产生新样本的非 R 连续位匹配的部分变异，变异的位数与海明距离成反比，形成新的集合 G；

样本变异的总数量公式：

$$C=\sum_{i=1}^{N}(\frac{N}{\eta}+i)$$

C 是变异的总数量，η 是给定的参数。

(9)计算集合 G 中的样本和识别对象的海明距离；

(10)把集合 G 中海明距离大于记忆集合 U 的样本，增加到记忆集合中；

(11)如果没有达到识别要求则转到步骤(1)。

3 实例验证

3.1 问题描述

利用上述算法识别，用位图表示“1”和“7”这 2 个数字，如图 1 所示。所有的数字都用 12 bit 的二进制串表示，每个像素用 1 个二进制数表示，其中 1 表示对应的像素为黑色，0 表示对应的像素为白色。1 和 7 构成的模式集合可用矩阵 A 表示：

$$A=\begin{bmatrix}a_1\\a_2\end{bmatrix}=\begin{bmatrix}010010010010\\111001001001\end{bmatrix}$$

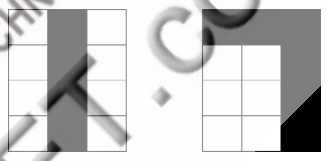

图 1 “1”和“7”的位图

3.2 系统训练

首先对数字“1”的模式作系统训练。随机产生 5 个样本(N=5)，对这些样本与数字 1 的模式进行海明距离计算，选择海明距离最大的 3 个样本(n_1=3)组成样本集合 M_1；然后在重新随机生成的 5 个样本中与数字 1 的模式进行海明距离计算，选择海明距离最大的 2 个样本(n_2=2)组成样本集合 M_2，这样就完成了上面所说的步骤(4)，形成了高质量的匹配度高的样本。把这 5 个高质量的样本与数字 1 的模式串进行 2 个连续位的匹配(R=2)，根据海明距离和 2 个连续位匹配对样本进行降序排列，在海明距离相同的情况下有 2 个连续位匹配的样本放在前面。在利用 R 连续位匹配的时候会出现洞，就是既能与自己匹配，也能和非己匹配无法想成一个有效的检测器。随着 R 的减少，“洞”的数量会增加，R 取值 12 时，每一个串只能匹配自身，不存在“洞”但是增加了时间复杂度降低了检测效率，所以，交叉使用这 2 种不同的匹配规则可以减少“洞”的数量。对这 5 个样本按顺序复制不同的数量，给定 λ 是 0.5，则第一个样本复制 9 个，后面 4 个样本依此类推。对复制后的样本进行变异，变异的位数分别是 0、1、2、3、4，这样保证了变异的方向是趋向于完全匹配识别对象。重新计算变异后的样本和数字 1 的模式串的海明距离。对于完全匹配数字“1”的样本加入到记忆集合中去，保证以后再次识别数字

“1”时不需要学习，直接通过记忆集合的匹配迅速识别出来，而且这种不断更新记忆集合的机制保证了记忆集合的最优，可以有效地降低错误识别的概率。

3.3 数字识别

通过上面的系统训练可以有效地识别数字“1”和“7”，实验结果如图2所示。

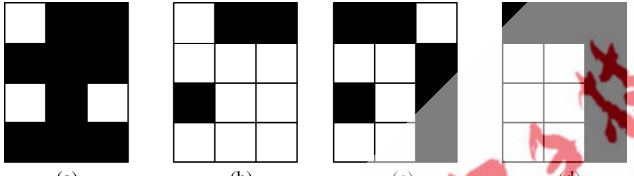

图2 识别数字“7”的过程

图2(a)是随机产生的样本集，图2(b)和图2(c)分别是进化到第15代和第30代的样本集。图2(d)是进化到第40代的样本集。

表1 改进前识别数字“7”时样本海明距离的变化

变异次数	5	15	20	30	35	40	50	60
海明距离	2	4	5	7	7	8	10	12

表2 改进后识别数字“7”时样本海明距离的变化

变异次数	5	10	15	20	25	30	35	40
海明距离	3	3	4	6	7	8	10	12

表1和表2分别为算法改进前后识别数字“7”时样本海明距离的变化。对比实验结果可知，上述的算法可以提高识别效率、减少变异次数，而且成功识别消耗的时间大幅减少。

而且本文在海明距离和R连续位匹配规则的基础上结合生物学基因突变的思想实现对数字的识别。在以下几个方面实现了创新：

(1)在系统训练中2次利用海明距离，有效地淘汰了劣质的随机样本提高了模式识别的效率；

(2)海明距离和R连续位匹配规则的交叉使用，有效地避免了“洞”的出现，提高了模式识别的准确率；

(3)记忆集合的动态更新，保证了模式识别的效率，降低了时间复杂度。

实验结果表明合适的参数使得整个系统在相似性和匹配度之间均衡调节，呈现出更高的检测率。该系统具有良好的适应性、自学习性、开放性、鲁棒性。

参考文献

[1] 宋程，李涛，陈桓，等.基于人工免疫原理的未知病毒检测方法[J].计算机工程与设计，2005，26(3)：125-127.

[2] 李鸿吉.模糊数学基础及实用算法[M].北京：科学出版社，2005.

[3] 王茜，傅思思，葛亮.基于人工免疫的新型检测器生成模型[J].计算机应用，2006，26(11)：2618-2621.

(收稿日期：2009-03-21)