论文阅读笔记 2018/01/01

论文题目：

基于BP神经网络的文本验证码破解

作者：高原，中图分类号：TP309，A，1007-7820（2012）07-0377-06

论文拟解决主要问题：

以某游戏网站的验证码为例，介绍了基于文本图像验证码的识别和破解过程。

论文主要研究内容：

破解过程分为字符提取、字符修正和字符识别。在字符提取阶段需要对图片进行预处理降低提取难度，提取主要采用了近似颜色统计法；在字符修正阶段对比了传统的旋转算法和改进旋转算法，归一化字符；在字符识别阶段采用了BP神经网络方法，对验证码的识别正确率达70%，每个验证码的平均破解时间为1.625s。

论文使用的所有理论方法：

字符提取

字符修正

字符识别

破解过程：

（1）在字符提取阶段对图片进行预处理降低提取难度，采用颜色统计法提取字符。

（2）在字符修正阶段采用改进旋转算法修正字符，然后归一化。

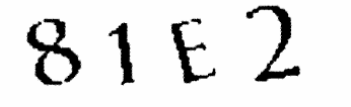
（3）在字符识别阶段，采用BP神经网络算法对单个字符进行识别。

（1）字符提取

通过预处理除去各种干扰点、干扰线后，按照近似颜色统计法和上下扫描法提取每个字符，提取目的为将字符一个个找出，将背景色设置为白色，将前景色设置为黑色。



图1 原图

图2 提取的字符

（2）字符修正

通过对字符的研究发现，当旋转角度不同时，包围字符的矩形框面积也不同。当字符转正时，包围它的矩形面积最小。

根据字符特征分析已经说明了字符的旋转角度在±20°之间，每2°旋转一次，求得包围字符矩形的面积，再比较所有矩形面积找到面积最小时的角度，这就是所要旋转的角度。

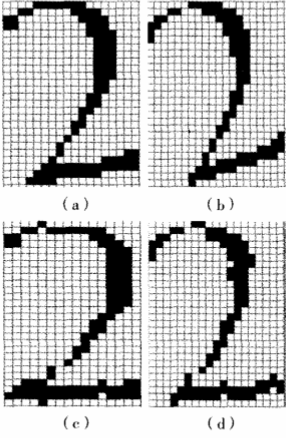


图3 旋转前的字符和旋转后的字符

（3）字符识别

BP神经网络由信息的正向传播和误差的反向传播两个过程组成，具体表现为：输入层接收来自外界的的输入信息，并传递给隐含层；隐含层负责信息变换，然后将处理的信息传递到传输层；进一步处理后，由输出层向外界输出处理结果，完成一次正向传播学习过程。当目标输出与实际输出相差过大时，进入误差的反向传播阶段。

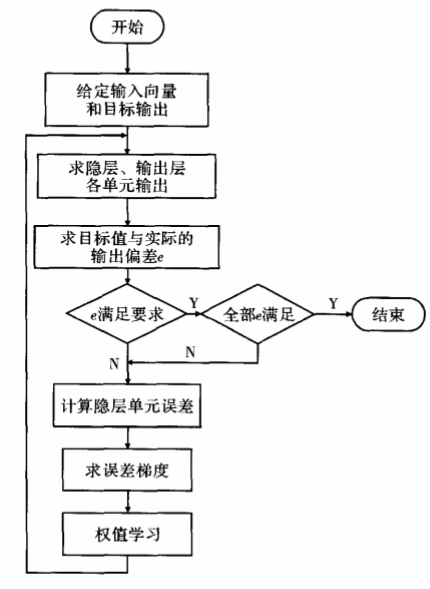


图4 BP神经网络算法框图

将BP神经网络的思想应用于文本验证码的识别上，就是一个寻找待识别的字符属于哪一类别的分类过程。论文采用常见的3层设计，可以有效的提高网络学习速度，并且保持较好的识别速度和精确度。

论文对训练样本的选择原则为：每个类别均要选取样本，个数视情况而定；样本个数应适量，约为网络权总数的5~10倍；样本要求交叉输入，防止网络产生震荡；同类别样本应保证多样性。

在输入层的设计中，通过对样本字符的观察，发现不同字符在上下左右区域以及中心区域的字符点数存在较大区别，所以选取了11个特征输入，在高10和20，宽6和12处将图片分割为9个部分，分别计算每部分的像素个数作为9个特征输入，统计图片中心作为第10个特征输入，再把前9个统计数相加作为第11个特征输入。所以输入层一共有11个神经元（如图5）。在隐含层的设计中根据改进公式并结合实验测试效果，最终选定隐含层的神经元数为7。在输出层的设计中，因为预计输出类别数目为16，所以输出层为 个。

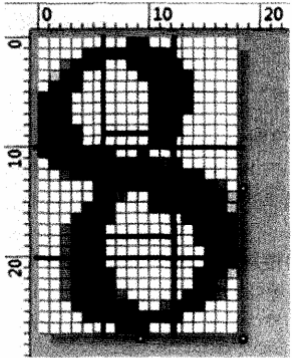


图5 特征选取示例图

论文的创新点：

虽然该论文的破解工作只是针对某个游戏网站，但是对于采用类似验证码机制的网站也适用。本文中涉及的方法和思路也可以为其他类型的验证码破解提供新思路。该论文根据对某游戏网站的100个文本验证码样本进行分析，并总结该验证码的特点，根据分析总结得出的特点，归纳的使用了BP神经网络对验证码进行分析处理，并根据在实际应用中发现的BP神经网络的使用缺点，在实际破解中加入了模拟退火算法，变学习速率算法得到BP神经网络改进学习，提高了识别正确率。最终对该游戏网站的验证码识别率高达70%。

论文的总结：

当前对文本验证码的破解的研究已经相对成熟，研究验证码的破解有助于更好的发现设计漏洞，维护网络安全。