

手写金融汉字识别中的可信度估计

徐蔚然, 郭 军

(北京邮电大学信息工程学院, 北京 100876)

摘 要: 由于实际票据中的手写金融汉字书写并不规范, 而且还会受到票据的背景底纹和印章等因素干扰, 所以正确识别率很低. 为了达到要求的识别精度, 必须拒识大量样本, 这样就限制了票据识别系统的自动处理率. 本文提出一种基于支持向量机的可信度估计方法, 其特点是针对每个文字类别, 专门构造用于易混淆字判断与可信度估计的支持向量机. 该方法充分利用了支持向量机在少量训练样本情况下对两类判断问题的学习能力, 可以准确地估计可信度, 从而在拒识较少样本的情况下达到要求的识别精度.

关键词: 汉字识别; 支持向量机; 易混淆字; 可信度; 银行票据 OCR

中图分类号: TP391 文献标识码: A 文章编号: 0372-2112(2005)10-1879-04

Confidence Evaluation for Handwritten Chinese Financial Character Recognition

XU Weiran, GUO Jun

(Beijing University of Posts & Telecom, Beijing 100876, China)

Abstract: Because handwritten characters in practical bank check are always interfered by background image and seals, the accuracy rate of OCR is very low. To obtain required precision, most samples have to be rejected, by which the auto processing rate is confined. In this paper, a new confidence evaluation method based on support vector machines is proposed. In this method, special SVMs are constructed for every character class to evaluate confidence and recognize similar character. This method makes full use of SVM's studying ability in the condition of few training samples, and can accurately estimate the confidence. By this method we can reject fewer samples to obtain required precision.

Key words: Chinese character recognition; support vector machine; similar character; confidence; bank check OCR

1 引言

实际银行票据中的手写金融汉字书写并不工整规范, 同时还会受到票据的背景底纹和印章等因素干扰(见图 1 和图 2), 所以 OCR 正确识别率很低. 因此根据识别结果的可信度设置拒识阈值是达到要求识别精度的基本方法.

识别结果可信度的研究一直受到人们的关注, 并提出很多方法. 采用最小距离分类器时, 依照距离的大小排序, 识别器给出前 k 个候选字 $(c_1 c_2 \dots c_k)$ 以及相应的 k 个距离值 $(d_1 d_2 \dots d_k)$. Xu L^[3] 和 Lee Y S^[2] 分别采用式

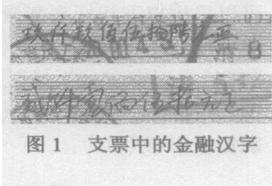
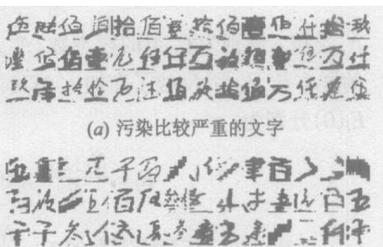


图 1 支票中的金融汉字



(a) 污染比较严重的文字

(b) 污染非常严重或切分错误的文字

图 2 二值化与切分之后的金融汉字

(1)和式(2)中的经验公式估计可信度:

$$CF_{C(k)}(x) = \frac{1/d_k}{\sum_{i=1}^M 1/d_i} \quad (1)$$

$$CF_{C(k)}(x) = \frac{\text{score}_k}{\sum_{i=1}^M \text{score}_M}, \text{score}_k = \frac{1}{d_k - d_1 + 1} \quad (2)$$

[9]则利用首选距离 d_1 及其与二选距离 d_2 的差建立识别系统的误识模型. [8]把这种比较候选字距离测度的方法称为相对尺度; 并把相对尺度和绝对尺度结合起来, 得到综合尺度可信度. 另外, Lin X^[7]通过可信度自适应变换(ACT, Adaptive Confidence Transom)把式(3)的广义可信度转换为可信度. [10]则是通过逻辑回归模型 LRM 把距离测度转换为候选字可信度.

$$e(C(k)|x) = 1 - \frac{d_k}{\min_{k \neq i} d_i} \quad (3)$$

以上方法都是根据文字识别分类器提供的信息计算可信度. 本文提出一种基于支持向量机的可信度估计方法. 该方法在文字识别分类器之后再构造一组支持向量机; 每一个支持

向量机负责识别一种文字类别的错误样本和正确样本;支持向量机输出的测度信息就是广义可信度.该方法充分利用了SVM对小训练样本集的学习能力;同时避免了用SVM处理非两类识别任务.

2 可信度与支持向量机

可信度本质上是后验概率.虽然大部分测度级分类器(the measurement level)^[3]的输出并不是后验概率,但同样可以用来衡量识别结果的可信程度,如模板匹配分类器的距离测度、神经网络和支持向量机的输出等.因此可以把这种反映识别结果可信程度的变量称为广义可信度^[7,10].广义可信度同样非常重要.在某些应用下,广义可信度可直接替代可信度,如用拒识率换识率等;而且存在ACT变换^[10]、神经网络变换^[4]等多种算法把广义可信度转化为可信度.

支持向量机(Support Vector Machine 简称SVM)也是测度级分类器,它的输出就是一种广义可信度.一个SVM解决的是两类识别问题,它最大的优点是在小训练样本集时能够控制推广能力.在采用SVM时,越能充分利用它的这些特点,就越能充分发挥它的潜力.

3 基于支持向量机的可信度估计方法

3.1 易混淆字判断与可信度估计

为了更清楚地说明本文的方法,这里采用了易混淆字的概念,它的主要含义如下.考虑 M 个类别的分类问题,文字识别分类器 e 把输入的样本 x 归为 M 类中的一种.把样本空间中不属于 C_j 类($j=1, 2, \dots, M$),但却被 e 错分为 C_j 类的模式,称作对于分类器 e 的 C_j 类的易混淆字.

所有被 e 识别为 C_j 类的样本可以分为两种:一种是正确识别为 C_j 类的样本;另一种是错误识别为 C_j 类的样本,即易混淆字样本.如果能够区分这两种样本,那么就可以大大降低误识率;而估计可信度就等价于估计该样本是第一种样本的概率.当采用SVM区分这两种样本时,SVM的输出就是广义可信度.这样就易混淆字判断和可信度估计融合在一起.

3.2 可信度估计方法的结构

方法的结构框图见图3.首先用分类器 e 把输入样本 x 分类到 M 个类别中的某个类别中去.然后采用测度级分类器 e_j 对被识别到 $C_j(j=1, 2, \dots, M)$ 类的样本进行易混淆字判断和可信度估计,得到广义可信度.最后,把广义可信度转化为可信度.

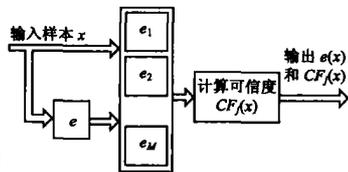


图3 易混淆字判断和可信度估计的结构图

3.3 用于可信度估计的支持向量机

图3中的 e_1, e_2, \dots, e_M 都采用SVM.这里适合采用SVM的原因如下.(1)由于 e 的错误率通常都不会很高,所以即使训练样本数很多,但属于 C_j 类易混淆字的训练样本数量很少,该问题属于小样本集学习问题.SVM是建立在结构风险最小化准则之上的机器学习方法,对小样本情况可以通过控制

学习机器的容量使得支持向量分类器具有较好的推广能力.

(2)SVM是测度级分类器,并且随着SVM测度值从小到大变化,测度值对应的可信度也单调变化,即SVM的输出就是一种广义可信度.(3)由于易混淆字是 e 不能正确分类的样本,所以区分易混淆字通常属于线性不可分问题,而SVM可以实现复杂的非线性分类.(4)SVM专门用于两类问题,对两类分类可以得到最好的效果.

支持向量机的输入:

每个SVM的输入特征包括两部分.一部分特征直接从输入样本中抽取,这里采用 7×7 加 8×8 分块的方向线索特征,共452维^[11].另一部分特征来自分类器 e 的输出信息.由于分类器 e 通常是测度级分类器,这里采用前10个候选字的距离测度作为SVM的特征.

支持向量机的训练方法:

在本文中,第 j 个支持向量机 e_j 的正训练样本来自 $S_{j\text{-correct}}$ (被 e 正确识别为 C_j 的训练样本);负训练样本来自 $S_{j\text{-error}}$ (被 e 错误识别为 C_j 的训练样本).

4 实验结果

4.1 测试样本库和评价指标

测试样本库中的样本来自实际支票,由18个金融汉字的样本和噪声样本构成.这18个金融汉字分别为:“零”,“壹”,“贰”,“叁”,“肆”,“伍”,“陆”,“柒”,“捌”,“玖”,“拾”,“佰”,“仟”,“万”,“元”,“角”,“整”和“正”.实际支票识别系统的待识别样本中存在不属于这18个金融汉字的样本.这些样本有可能是其他汉字、切分错样本或污染导致的人工无法辨识样本等,它们都属于噪声样本(见图2b).

识别系统的性能可用拒识率和误识率之间的关系表示.由于本文采用了多个SVM进行易混淆字判断和可信度估计,因此需要通过测试每个SVM的指标来得到系统的整体性能.可以采用以下指标来衡量支持向量机 $e_j(j=1, 2, \dots, M)$ 的性能:分类器 e 的错误率 $E_e(j)$;阈值为0时的拒识率 R_j ;阈值为0时的错误率 $E_j(0)$;错误率为1%时的拒识率 $R_j(0.01)$;错误率为0.1%时的拒识率 $R_j(0.001)$.分类器 e 对第 j 个类别的识别错误率 $E_e(j)$ 为:

$$E_e(j) = \frac{N_{j\text{-error}}}{N_j} \times 100\% \quad (4)$$

N_j 是被 e 识别为 C_j 类的样本数; $N_{j\text{-error}}$ 表示被 e 错识成 C_j 类的样本数. $E_e(j)$ 反映了在加入SVM之前分类器的性能. R_j 和 $E_j(0)$ 分别为:

$$R_j = \frac{N_{j\text{-reject}}}{N_j} \times 100\% \quad (5)$$

$$E_j(0) = \frac{N_{j\text{-accept-error}}}{N_j - N_{j\text{-reject}}} \times 100\% = \frac{N_{j\text{-accept-error}}}{N_{j\text{-accept}}} \times 100\% \quad (6)$$

其中 $N_{j\text{-reject}}$ 表示当阈值为0时,被 e_j 判断为不属于 C_j 类,因而被拒绝的样本个数; $N_{j\text{-accept-error}}$ 表示被 e 和 e_j 同时错判为 C_j 类的样本个数.SVM的默认阈值就是0,因此 R_j 和 $E_j(0)$ 就反映了对精度没有特殊要求时的系统性能.对于精度要求严格的情况,如识别银行支票,就需要适当调整每个SVM的阈值,从

而得到 $R_j(0.01)$ 和 $R_j(0.001)$ 等指标.

4.2 测试结果分析

利用测试样本得到表 1 的测试结果. 由于包含噪声样本, 所以理想情况是 $R_j = E_e(j)$, $E_j(0) = 0$, 即 SVM 仅把所有 e 识别错误的样本正确拒绝掉. 在表 1 中可以看到, 当阈值为 0 时, 对于任何一个类别, R_j 和 $E_e(j)$ 都比较接近, 而同时 $E_j(0)$ 也非常低. 也就是说, SVM 几乎只拒绝掉 e 的错识样本, 而且拒绝掉了大部分的错识样本, 从而大大降低了文字识别系统的误识率. 从总体来说, 整个系统在拒绝掉 18.60% 样本后, 把误识率从 18.53% 降低到 4.44%.

另外, 从表 1 中可发现分类器 e 对不同类别文字的错误率相差很大, 例如, “万”“角”“正”的错误率在 35% 以上; 而“贰”的错误率仅为 3%; 大部分文字种类的错误率分布在 5~20% 之间. 这是由于噪声样本与“万”“角”“正”等汉字更相似

所致(见图 2). 通过 SVM 的处理后, 不同类别文字的错误率变得相差不大, 都分布在 2~9% 之间.

A2iA 公司^[12]认为银行中人工处理的误识率为 1% (整个文字串的处理错误率). 在表 1 中, 把单个字的错误率设置为 1% 时, 系统整体的拒识率为 35.12%; 把单个字的错误率设置为 0.1% 时, 系统整体的拒识率为 59.37%. 由于文字串中的文字会同时受到污染或出现切分错误, 从而导致同时被识别错, 所以字串的识别正确率远远大于单个文字识别正确率的乘积. 如果保守地假设单字错误率低于 0.1% 时, 字串错误率低于 1%, 则系统的自动处理率为 40.63%. 在国内的标准支票中, “金额”会分别以大写、小写和打印磁码等形式书写 3 次, 有效利用这一信息后, 实际系统的自动处理率会更高. 例如, 采用本文方法后北京邮电大学模式识别实验室构造的“银行票据 OCR 系统”的自动处理率可达到 70% 以上.

表 1 实验结果

($E_e(j)$): 分类器 e 的错误率; R_j : 阈值为 0 时的拒识率; $E_j(0)$: 阈值为 0 时的错识率;

$R_j(0.01)$: 错误率为 1% 时的拒识率; $R_j(0.001)$: 错误率为 0.1% 时的拒识率)

	零	壹	贰	叁	肆	伍	陆	柒	捌	玖
$E_e(j)$	20.64%	7.00%	3.00%	14.46%	9.10%	10.99%	14.16%	10.49%	29.38%	11.64%
R_j	26.04%	2.78%	1.34%	14.97%	7.37%	8.30%	12.38%	9.63%	33.44%	12.47%
$E_j(0)$	3.20%	5.02%	2.08%	4.35%	3.74%	5.85%	5.36%	4.08%	6.05%	3.44%
$R_j(0.01)$	35.24%	17.01%	8.22%	24.36%	16.96%	39.27%	25.65%	22.80%	53.95%	27.45%
$R_j(0.001)$	56.41%	51.03%	31.07%	47.41%	27.60%	41.75%	39.86%	47.86%	78.37%	61.09%
	拾	佰	仟	万	元	角	整	正	系统整体性能	
$E_e(j)$	8.63%	19.66%	5.40%	39.06%	26.21%	46.75%	20.67%	37.07%	18.53%	
R_j	5.54%	21.80%	3.35%	38.48%	28.45%	51.86%	20.83%	39.47%	18.60%	
$E_j(0)$	4.67%	5.36%	3.27%	5.16%	2.03%	8.77%	5.80%	5.17%	4.44%	
$R_j(0.01)$	22.58%	30.49%	14.85%	61.09%	48.49%	68.50%	47.20%	51.28%	35.12%	
$R_j(0.001)$	46.91%	71.53%	37.53%	69.93%	92.73%	71.54%	70.61%	73.81%	59.37%	

53% 降低到 4.44%; 同时给出准确的可信度.

4.3 同其他方法比较

采用相同的测试样本库, 对[3]、[2]和[7]介绍的 3 种文字识别中常用可信度估计方法进行测试(它们的计算公式分别为公式 1, 2 和 3, 在表 2 中用 M1, M2 和 M3 标识), 得到表 2. 采用 SVM 方法得到的拒识率比原有方法中拒识率最低的方法分别降低了 11.58 和 5.15 个百分点.

表 2 不同方法比较

	M1	M2	M3	SVM
错误率 1% 拒识率	46.70%	48.25%	48.01%	35.12%
错误率 0.1% 拒识率	64.52%	66.73%	65.86%	59.37%

5 结论

本文提出一种基于支持向量机的易混淆字判断和可信度估计方法. 该方法的特点是: 把易混淆字判断和可信度估计融合在一起, 并针对每个文字类别, 专门构造用于易混淆字判断与可信度估计的支持向量机. 该方法充分利用了支持向量机在少量训练样本情况下对两类别判断问题的学习能力, 因而可以准确估计可信度. 对于支票中的金融汉字识别问题, 该方法可以在拒识率仅为 18.60% 的情况下, 把识别错误率由 18.

参考文献:

[1] Cheng Lin Liu, Masaki Nakagawa. Precise candidate deletion for large character set recognition by confidence evaluation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(6): 636-642.

[2] Lee Y S, Chen H H. Analysis of error count distributions for improving the postprocessing performance of OCCR[J]. Communications of Chinese Oriental Information Processing Society, 1996, 6(2): 81-86.

[3] Lei Xu, Adam Kzyzak, Ching Y S. Methods of combining multiple classifiers and their applications to handwriting recognition[J]. IEEE Transactions on Systems, Man and Cybernetics, 1992, 22(3): 418-435.

[4] Eki Ishidera, Atsushi Sato. A candidate reduction method for handwritten kanji character recognition[A]. ICDAR' 2001[C]. Seattle, USA: ICDAR, 2001. 8-13.

[5] Luiz S Oliveira, Robert Sabourin, Ching Y Suen. Automatic recognition of handwritten numerical strings: a recognition and verification strategy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(11): 1438-1454.

[6] Gethin Williams, Steve Renals. Confidence measures from local posteriors

- or probability estimates[J]. Computer Speech and Language, 1999, 13(4): 395- 411.
- [7] Xiaofan Lin, Xiaoqing Ding, Ming Chen, Rui Zhang, Youshou Wu. Adaptive confidence transform based classifier combination for Chinese character recognition[J]. Pattern Recognition Letters, 1998, 19(10): 975- 988.
- [8] 张洪刚, 刘刚, 郭军. 一种书写汉字识别结果可信度的测定方法[J]. 计算机学报, 2003, 26(7): 819- 824.
Zhang Hong Gang, Liu Gang, Guo Jun. an algorithm of handwritten numerals segmentation based on multimould and its application [J]. 2003, 26(7): 819- 824. (in Chinese)
- [9] 马少平, 夏莹, 朱小燕, 姜哲. 汉字识别系统的误识模型[J]. 清华大学学报, 1998, 38(S1): 26- 28.
Ma Shao Ping, et al. Mis recognition model of Chinese character recognition system[J]. Journal of Tsinghua University(Sci&Tech), 1998, 38(S1): 26- 28. (in Chinese)
- [10] 李元祥, 丁晓青. 一种利用逻辑回归模型的候选字可信度估计方法[J]. 模式识别与人工智能, 2002, 15(2): 143- 148.
Li Yuanxiang, Ding Xiaoqing. Measuring character candidate confidence using logistic regression model and its performance evaluation[J]. PR & AI, 2002, 15(2): 143- 148. (in Chinese)
- [11] J Guo, N Sun etc. Algorithm for recognition of handwritten characters using pattern transformation with cosine function [J]. IEICE Trans, 1993, J76-D-II(4): 835- 842.
- [12] N Gorski, V Anisimov, E Augustin, O Bardt, S Maximov. Industrial bank check processing: the A2iA CheckReaderTM [J]. International Journal on Document Analysis and Recognition, 2001, 3: 196- 206.
- [13] 徐蔚然. 基于统计分类器的银行票据自动处理系统若干关键技术研究[D]. 北京: 北京邮电大学, 2003, 7.
Xu Weiran. A Research on Key Techniques in Bank Cheque OCR System Based on Statistical Classifier[D]. Beijing: Beijing University of Posts and Telecommunications, 2003. 7. (in Chinese)

作者简介:



徐蔚然 男, 1975 年 4 月生于辽宁省抚顺市, 讲师, 研究方向为模式识别与人工智能, 图像处理. E-mail: xuweiran@263.net.

郭军 男, 1959 年 10 月生于吉林省舒兰县, 北京邮电大学教授, 博士生导师, 研究方向为模式识别, 智能信息处理等.