第 12 期
2002 年 12 月

电 子 学 报
ACTA ELECTRONICA SINICA

Vol.30 No.12
Dec. 2002

贝叶斯分类器在手写汉字识别中的应用

蔺志青,郭 军
(北京邮电大学 95 信箱,北京 100876)

摘 要: 由于缺少以较少的存储空间描述汉字特征概率密度函数方法,在汉字识别系统中贝叶斯分类器很少得到应用.本文提出一种只需 6 个数据就能描述一个特征的概率密度函数的方法,构造了一个基于贝叶斯分类器的手写汉字识别系统.实验结果表明,该分类器具有优良的性能.

关键词: 贝叶斯分类器; 汉字识别; 概率密度函数
中图分类号: TP391.1 **文献标识码:** A **文章编号:** 0372-2112 (2002) 12-1804-04

An Application of Bayesian Classifier in the Recognition of Handwritten Chinese Character

LIN Zhi-qing, GUO Jun
(*P. O. Box* 95, *Beijing University of Posts & Telecommunications*, *Beijing* 100876, *China*)

Abstract: Since lack of effective methods to describe the probability density functions of the features of Chinese characters, Bayesian classifiers are rarely used in the Chinese character recognition systems. In this paper, a method, which can describe the probability density function of a feature with only 6 digits, is proposed. Based on this method, a handwritten Chinese character recognition system using a Bayesian classifier is constructed. In our experiments, the classifier showed good performances.

Key words: Bayesian classifier; Chinese character recognition; probability density function

1 引言

现有的汉字识别系统很多都采用模板匹配的方法.在这种系统中,参照特征一般用训练样本特征的平均值来描述,分类器根据输入样本特征与各个文字的参照特征的距离(或相关度)进行识别.显然,只用特征平均值来描述特征是不够的,由于汉字存在各种各样字体,手写汉字中还存在各种各样的变形,因此文字的任何特征都存在一个分布空间.只有把这些分布考虑进去,才能更精确地进行分类识别.为此,在一些分类器中,将特征分布参数(如标准差)引入距离函数[1,2].这种改良明显地提高了识别精度,因而得到了广泛的应用[3~7].显然,仅引入分散性参数还是粗略的,更精确的方法应是给出各个文字特征的概率密度函数,采用贝叶斯分类器进行识别.

但是,识别汉字通常需要较高维数的特征,并且汉字的字符集又很大,从而使得汉字特征的概率密度函数的描述成了问题.由于特征值的分布通常不是某种简单的统计分布,如果没有简单的方法描述这些概率密度函数,则表示所有汉字的各维特征的概率密度函数(参照特征)需要的存储空间将是实用系统所不能承受的.这就是贝叶斯分类器难于应用在汉字识别系统中的主要原因.

本文以手写汉字为对象,分析了其特征分散性的原因及特性,说明了一般情况下不宜采用简单的分布类型,而可以用分段函数进行描述.在此基础上,提出了手写汉字特征分布特性的 4 点假设,及线性分段描述方法.这种描述方法只需 6 个数据就能描述一个特征的概率密度函数,为基于贝叶斯分类器的汉字识别系统提供了一个简单的实现方案.

在采用 HCL2000 手写汉字数据库[8]样本所进行的实验中,分别对 Manhattan 距离分类器、引入标准差的 Manhattan 距离分类器、以及本文提出的贝叶斯分类器的性能进行了测试.结果表明,引入标准差的 Manhattan 距离分类器平均识别率为 87.59%,比一般 Manhattan 距离分类器提高 6.05 个百分点,而贝叶斯分类器的平均识别率达到 89.21%,又进一步提高了 1.62 百分点.并且,贝叶斯分类器具有易于设定拒识条件的优点,实验中,获得了拒识率 17.8%,识别率 97.0%的性能.

2 基于模板匹配的汉字识别系统

2.1 参照特征及距离函数

设 $\boldsymbol{v}_i^j=(v_1,v_2,\cdots v_k)^T$ 为从第 $i(1\leqslant i\leqslant c)$个文字的第 $j(1\leqslant j\leqslant t)$个训练样本中抽取的特征向量.这里,$c$ 为系统可识别的文字的个数,t 为计算参照特征时选取的训练样本的个数,k 为特征向量的维数.则第 i 个字的均值型参照特征向量 $\boldsymbol{u}_i$ 用式(1)计算:

收稿日期:2002-02-19;修回日期:2002-05-13
基金项目:国家 863 计划(No.2001AA114080);教育部科学技术研究重点项目(No.02029)

$$\boldsymbol{u}_i = 1/t \sum_j^t \boldsymbol{v}_i^j \quad (1)$$

对应均值型参照特征，分类器采用某种距离函数来计算一个输入样本的特征向量 $\boldsymbol{x}$ 与第 i 个字的参照特征向量 $\boldsymbol{u}_i$ 间的距离 d_i. 式(2)是最简单的距离函数，Manhattan 距离函数.

$$d_i = \sum_{j=1}^k |x_j - u_j^i| \quad (2)$$

式中 x_j 为向量 $\boldsymbol{x}$ 的第 j 个分量，u_j^i 为向量 $\boldsymbol{u}_i$ 的第 j 个分量. 由式(1)可知，通过式(2)得到的 d_i 仅仅是 $\boldsymbol{x}$ 与第 i 个字的特征向量的平均值之间的距离，即第 i 个字的参照特征向量在向量空间中只占据一个点，没有考虑其分散性. 实际上，各个字的特征向量不但具有分散性，而且分散程度也不同. 为了将这个现象反映到距离函数中，人们将特征向量的方差特性引入了距离函数. 例如，一种改良的 Manhattan 距离函数[6]为：

$$d_i = \sum_{j=1}^k \max\{0, |x_j - u_j^i| - a_j^i\} \quad (3)$$

式(3)中，a_j^i 为第 i 个字的参照特征向量第 j 个元素的标准差. 根据式(3)，只要 x_j 与 u_j^i 之差的绝对值不超过 a_j^i，则该维上的距离被完全吸收. 这个距离函数考虑了参照特征的分散性，不再将其作为空间中的一点来计算与输入文字的特征向量之间的距离，而是认为它在向量空间中占据一定的范围.

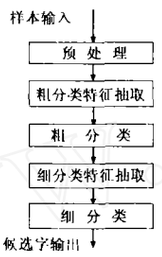

图 1　手写汉字识别系统

2.2　不同分类器的性能

2.2.1　测试系统

图 1 描述了一个有较高性能的手写汉字识别系统. 该系统由预处理、粗分类特征向量抽取、粗分类、细分类特征向量抽取、细分类等几个部分组成.

预处理包括尺寸规整和轮廓线抽取. 如图 2 所示.

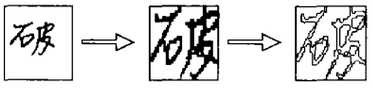

图 2　预处理

设粗分类特征向量 $\boldsymbol{x}^g = (x_1^g, x_2^g, \cdots, x_K^g)^T$，由于粗分类是为了大幅度减少细分类时的参照类，提高识别速度，因此 $\boldsymbol{x}^g$ 的维数 K 不宜过大，一般在 100 至 200 之间.

粗分类时，利用简单距离函数，如式(2)，计算输入文字的特征向量 $\boldsymbol{x}^g$ 同各字的标准特征向量 $\boldsymbol{u}^{gi}$ 之间的距离，选出距离最小的 m 个字，投入细分类.

设细分类特征向量 $\boldsymbol{x}^f = (x_1^f, x_2^f, \cdots, x_L^f)^T$，由于细分类是为了区分相似参照类，因此需要 $\boldsymbol{x}^f$ 能够反映各个相似参照类细节的差别.

细分类时，选择适当的距离函数，计算输入文字的细分类特征向量 $\boldsymbol{x}^f$ 与粗分类后保留下来的 m 个参照类的标准细分类特征向量 $\boldsymbol{u}^{fi}$ 之间的距离，输出具有最小距离的参照类的类名.

特征的选取是影响系统性能的一个重要因素，本文选择了方向线素特征进行分析和测试. 方向线素就是由水平、垂直、+45 度、−45 度某个方向上相邻的两个黑像素所构成[3~6].

抽取粗分类特征向量时，对文字区域进行 5×5 的分割，在得到的 25 个小区域内，累计 4 种方向线素的数量，从而得一个 100 维的粗分类特征向量 $\boldsymbol{x}^g = (x_1^g, x_2^g, \cdots, x_{100}^g)^T$.

抽取细分类特征向量时，利用 8×8 + 7×7 的二重分割[5,6]，将文字区域分割为 113 个小区域，在这些小区域中统计 4 种方向线素的数量，得到含有 452 个元素的细分类特征向量 $\boldsymbol{x}^f = (x_1^f, x_2^f, \cdots, x_{452}^f)^T$.

2.2.2　性能测试

实验数据来自 HCL2000 手写汉字样本数据库. HCL2000 是在国家 863 计划支持下建立的一个脱机手写汉字数据库系统，含有 GB2312－80 中所规定的一级汉字 3,755 个，每个汉字有 1,600 个样本. 在实验中，每字随机选出 400 个样本，其中 300 个作为训练样本，另 100 个作为测试样本. 各个字的参照特征向量 $\boldsymbol{u}^{gi}$ 和 $\boldsymbol{u}^{fi}$ 是利用这些训练样本根据式(1)计算的.

为了考察将特征的方差引入距离函数的效果，分别测试了用式(2)作为细分类距离函数的分类器 1 和用式(3)作为细分类距离函数的分类器 2 的识别率. 表 1 给出了主要结果.

从表 1 看出，系统 2 的结果明显优于系统 1. 识别率提高了 6.05 个百分点. 识别率等于 100% 的字增加了 9 个，识别率在 90－99 区间的字增加了 1144 个. 而识别率小于 90% 的各个区间的文字数都不同程度地减少、并且识别率越低的区间减少的倍数越大.

表 1　不同距离函数时的识别率

距离函数	平均识别率	识别率分布（文字数）					
		100	90－99	80－89	70－79	60－69	<60
分类器 1	81.54	1	606	1799	1051	245	53
分类器 2	87.59	10	1750	1522	383	75	15

3　贝叶斯分类器及其参照特征

为了进一步提高识别精度，可以考虑采用贝叶斯分类器. 采用贝叶斯分类器，首先要确定其参照特征向量. 设系统选取每个文字的 m 个特征进行识别，各输入样本的特征向量可表示为 $\boldsymbol{x} = (x_1, x_2, \cdots x_m)^T$，文字类 w_i 的特征向量可表示为：

$$\boldsymbol{r}^i = [p_1(x_1|w_i), \cdots p_k(x_k|w_i), \cdots, p_m(x_m|w_i)] \quad (4)$$

其中 $p_k(x_k|w_i)$ 为 w_i 的第 k 个特征的先验概率密度函数.

根据贝叶斯法则，对一个特征向量为 $\boldsymbol{x}$ 的输入样本 Z，有如下决策规则：

assign Z to w_j if

$p(w_j)\ p(\boldsymbol{x}|w_j) = \max_i p(w_i)\ p(\boldsymbol{x}|w_i)$.

式中，$p(w_i)$ 为 w_i 的出现概率，$p(\boldsymbol{x}|w_i)$ 为 w_i 的样本的特征向量为 x 的概率. 在各个特征相互独立的情况下，$p(\boldsymbol{x}|w_i)$ 可以用 $\boldsymbol{r}^i$ 各分量的乘积来表示，即：

$$p(\boldsymbol{x}|w_i) = \Pi_k p_k(x_k|w_i).$$

因此，用 $\boldsymbol{r}^i$ 描述参照特征，在各维特征值相互独立情况下，分类器规则为：

assign Z to w_j if

$p(w_j)\ p(\boldsymbol{x}\mid w_j)=\max_i p(w_i)\ \Pi_k p_k(x_k\mid w_i)$.

或者

assign Z to w_j if

$\ln[p(w_j)\ p(\boldsymbol{x}\mid w_j)]=\max_i\{\ln[p(w_j)]+\sum_k\ln[p_k(x_k\mid w_i)]\}$.

其中 $p_k(x_k\mid w_i)$为 $\boldsymbol{r}^i$ 的第 k 个分量.

采用上述贝叶斯分类器,需要各个文字类的各个特征的先验概率密度函数 PDF,这可以用大量训练样本的统计分布来近似.有了各个特征的 PDF 后,还需要有效的描述 PDF 的方法,使其不需要过多的存储空间.

4 特征的概率分布类型及其描述

导致特征分散的原因是笔画的位置、长度、宽度和方向具有分散性,此外,还有噪声的影响.研究结果表明,笔画的位置、长度、宽度和方向的统计分布特性有各自的特点,在它们的共同作用下,手写文字特征的分散特性难以用某种简单的分布(如高斯、瑞利)来描述.因此,为了描述特征的分布,需要采用区间函数逼近的方法.

对特征的分布进行区间函数逼近,需要对特征分布函数 $p(x)$的性质进行一些假设.不失一般性,设特征值 x 对应特定区域的某种笔画信息量,x 为大于等于 0 的整数,我们可以对 $p(x)$提出如下 4 点假设:

(1)$p(x)$在 $x=0$ 点具有特殊的区间性质;

(2)存在一个上界 Γ,当 $x>\Gamma$ 时 $p(x)=0$;

(3)在区间$[1,\Gamma]$之间存在一点 Δ,使得 $p(\Delta)=\max\{p(x)\}$;

(4)$p(x)$在区间$[1,\Delta]$单调递增,在区间$[\Delta,\Gamma]$单调递减.

以上假设是在对大量样本的不同特征进行统计分析的基础上提出的.假设 1 的根据是,$p(0)$表示指定区域无笔画信息的概率,而 $x\neq 0$ 时的 $p(x)$表示指定区域有笔画信息量且该信息量等于 x 的概率,因此 $p(x)$在 $x=0$ 点具有特殊性;假设 2 的根据是,笔画信息量的特征值都是有界的;假设 3 实际上是假设 2 的推论;假设 4 假定特征的概率分布在$[1,\Gamma]$区间是单峰函数,这一点也是符合一般情况的.

图 3 给出了 2.2.2 节所描述的训练样本的细分类特征的 3 种典型统计分布.由图可见,这 3 种典型统计分布都符合上述 4 点假设,且特征在 0 点,$[1,\Delta]$,$[\Delta,\Gamma]$3 个区间的分布特性明显不同.因此,特征的概率分布应采用分段函数来描述.

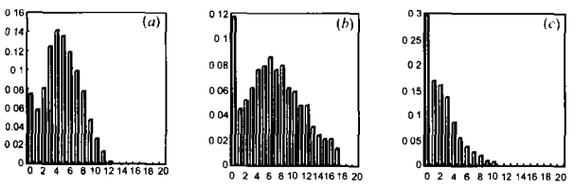

图 3 方向线素特征值的 3 种典型统计分布

利用样本特征的统计分布,可以对概率密度函数在各个区间进行曲线拟合,从而获得各个区间的函数表达式.这是获取分段概率密度函数的一般方法.在许多场合下,线性近似是一种简单有效的方法,它不但可以减少分类器的运算量,还可以减少特征向量的存储空间.本文中,我们定义分段线性函数

$$p'(x)=\begin{cases}p(0), & x=0\\ p(1)+[p(\Delta)-p(1)](x-1)/(\Delta-1), & 1\leqslant x\leqslant\Delta\\ p(\Delta)+[p(\Gamma)-p(\Delta)](x-\Delta)/(\Gamma-\Delta), & \Delta<x\leqslant\Gamma\\ 0, & x>\Gamma\end{cases}\tag{5}$$

来近似描述特征的密度函数.利用函数 $p'(x)$,只用 0,1,峰值点 Δ,上界 Γ 这 4 个点的统计分布值 $p(0)$,$p(1)$,$p(\Delta)$,$p(\Gamma)$,便可通过简单的线性计算获得特征概率密度函数在各点的近似值.

应当注意到,对应 Δ 和 Γ 的不同取值,式(5)会给出多种概率分布类型,图 4 给出了这些不同的类型.

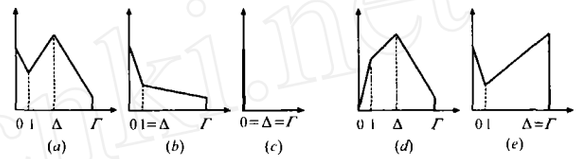

图 4 式(5)所对应的不同的概率分布类型

图 4(a)对应笔画信息易变的区域,是最常见的分布类型.图 4(b)对应笔画不是完全不可能出现的区域.图 4(c)对应笔画完全不可能出现的区域.图 4(d)对应一定会出现笔画的区域.图 4(e)对应的情况在实际中是罕见的.

5 贝叶斯分类器的性能测试

式(5)为我们提供了一个描述手写汉字特征概率密度函数的方法,有了它,第 3 节所提出的贝叶斯分类器便有了一个完整的实现方案.为了检验这个分类器的性能,依旧采用第 2.2 节所描述的系统及方向线素特征进行测试,并将贝叶斯分类器用于细分类.具体地,贝叶斯分类器规则采用对数形式,即对于一个输入文字 Z 的特征向量$\boldsymbol{x}=(x_1,x_2,\cdots x_{452})^T$

assign Z to w_j if

$\ln[p(w_j)\ p(\boldsymbol{x}\mid w_j)]=\max_i\{\ln[p(w_i)]+\Sigma_k\ln[pk(x_k\mid w_i)]\}$.

在样本测试中,$p(w_i)=p(w_j)(i,j=1,2,\cdots 3755)$,因此,

$\operatorname{arcmax}_i\{\ln[p(w_i)]+\Sigma_k\ln[p_k(x_k\mid w_i)]\}$

$=\operatorname{arcmax}_i\{\Sigma_k\ln[p_k(x_k\mid w_i)]\}$

为了防止出现计算 0 的对数的错误,在贝叶斯分类器中,假设每个特征值的出现概率 $p_k(x_k)$的最小值 $\mathrm{Min}Prob>0$,当 $p_k(x_k)$的计算值 $<\mathrm{Min}Prob$ 时,令其等于 $\mathrm{Min}Prob$.$\mathrm{Min}Prob$ 的取值可参考训练时 $p(x)$最小的非零值.

用式(5)来近似描述特征的概率密度函数,每个特征需要存储 Δ、Γ、$p(0)$、$p(1)$、$p(\Delta)$、$p(\Gamma)$共 6 个数值.每个文字的特征向量含有 452 个特征,因此它的存储需要 452 * 6 个数据.选用两位小数来表示 $p(x)$,然后将其扩大 100 倍化成 100 以内的整数,则每个数据只占 1 个字节.当文字数为 3755(国标一级汉字)时,特征向量所需的存储空间不到 10M.这是很容易满足的要求.

采用与第3.3节相同的训练和测试数据进行实验，贝叶斯分类器的识别率及其分布如表2所示.

表2　不同距离函数时的识别率

	平均识别率	识别率分布(文字数)					
		100	90-99	80-89	70-79	60-69	<60
PDFFC分类器	89.21%	11	2050	1452	215	22	5
与式(3)相比	+1.62%	+1	+300	-70	-168	-53	-10

通过表2可知,贝叶斯分类器的平均识别率达到89.21%,与改良的Manhattan距离分类器相比提高了1.62个百分点.从识别率的分布性对比结果看,贝叶斯分类器大幅度地减少了识别率小于80%的文字的数目.由原来的473个,占总数的12.6%,降至242个,占总数的6.4%.

贝叶斯分类器的另一个优点是容易得到有效的拒识条件:用 F 表示第1候选文字类,S 表示第2候选文字类,由于输入样本属于各个文字类的可能性 P_i($P_i = \Sigma_k \ln[p_k(x_k \mid w_i)]$)的值域是同一个区间(在本实验中该区间为[452Min*Prob*,452 ln100]),因此可以根据 $\Delta P = P_F - P_S$ 设定分类器的拒识条件.原因在于,ΔP 越小,判断其属于第1文字类的风险越大.而且 ΔP 是大还是小,可以与文字类无关地进行判断.表3给出了在设定拒识条件情况下,测试贝叶斯分类器性能的一个结果.

表3　有拒识情况下的贝叶斯分类器的性能

拒识条件	拒识率	识别率	识别率的提高
$\Delta P < 5$	4.9%	92.1%	2.9%
$\Delta P < 10$	9.4%	94.2%	5.0%
$\Delta P < 15$	13.6%	95.8%	6.6%
$\Delta P < 20$	17.8%	97.0%	7.8%

从表3可以看出,ΔP 这一拒识条件与拒识率几乎成线性关系,ΔP 每增加5,拒识率上升4.2至4.9.同时,随着拒识率(定义为:正确识别的样本数/(总样本数-被拒识的样本数))的上升,识别率单调增加.当拒识率为17.8%时,识别率已达到97.0%,与不设拒识的贝叶斯分类器相比,提高了7.8个百分点.这些结果说明 ΔP 这一拒识条件是合理和有效的,同时也使我们看到,有拒识的贝叶斯分类器可以在较低的拒识率的条件下达到很高的识别精度.这种有拒识的贝叶斯分类器易于与其他分类器进行组合构成更高性能的识别系统.

6　结论

本文提出一种用分段线性函数描述手写汉字特征的概率密度函数的方法,并在此基础上实现了一个识别手写汉字的贝叶斯分类器.在测试实验中,该分类器表现出了很好的性能,在不设拒识时,对未经训练的取自HCL2000的测试样本的识别精度达到89.2%.在设拒识的条件下,可以用较小的拒识率换取识别率的明显提高,例如,在拒识率为17.8%的情况下,识别率达到了97.0%.这种特点使得它易于同其他分类器组合形成识别精度更高的系统.

本文提出的贝叶斯分类器采用的决策规则假设各个特征之间相互独立,因此,分类器的性能与所采用特征的独立性关系密切.本文所采用的测试系统用的是方向线素特征,这种特征的独立性并不高.因此,可以相信,如果选取更好的特征,该贝叶斯分类器的性能还会进一步提高.这也是我们下一步的研究课题.

参考文献:

[1] S Mori, K Yamamoto, M Yasuda. Research on machine recognition of handprinted characters [J]. IEEE Trans, 1984, PAMI-6(4): 386-405.

[2] T W Hildebrand, W Liu. Optical recognition of handwritten Chinese characters: advances since 1980 [J]. Pattern Recognition, 1993, 26(2): 205-225.

[3] N Kato, M Suzuki, S Omachi, H Aso, Y Nemoto. A handwritten character recognition system using directional element feature and asymmetric mahalanobis distance [J]. IEEE Trans, 1999, PAMI-21(3): 258-262.

[4] J Guo, N Sun, Y Nemoto, M Kimura, H Echigo, R Sato. Recognition of handwritten characters using pattern transformation method with cosine function [J]. IEICE Trans, 1993, J76-D-Ⅱ(4): 835-842.

[5] J Guo, N Sun, Y Nemoto, R Sato. Recognition of handwritten character database ETL9B using pattern transformation method [J]. IEICE Trans, 1993, J76-D-Ⅱ(5): 1015-1022.

[6] N Sun, J Guo, Y Nemoto, R Sato. A new algorithm of handwritten character recognition by estimating the standard deviation of input pattern [J]. IEICE Trans, 1994, J77-D-Ⅱ(1): 79-90.

[7] 郭军,马跃,盛立东,钟义信.发展中的文字识别理论与技术[J].电子学报,1995,23(10):184-187.

[8] 郭军,蔺志青,张洪刚.一个新的脱机手写汉字数据库模型及其应用[J].电子学报,2000,28(5):115-116.

作者简介:

蔺志青　女,1959年4月生于河北省石家庄市,1982年7月毕业于北京邮电学院计算机与通信专业,1992年3月至1993年3月在日本东北学院大学留学,1993年6月至今在北京邮电大学任教,现任该校电信工程学院副教授.研究领域主要包括计算机应用、计算机与通信、中文信息处理等.