

基于 SVM 的脱机手写汉字机器学习识别方法研究

王建平, 陈 军, 徐晓冰, 王熹徽

(合肥工业大学 电气与自动化工程学院, 安徽 合肥 230009)

摘 要:提出了一种模糊统计方法的脱机手写体汉字特征提取方法, 结合小波网格方法和汉字笔画密度特征方法对汉字进行特征提取, 并运用支持向量机方法, 通过机器学习对脱机手写汉字识别。仿真实验表明, 支持向量机方法在脱机手写汉字识别中有良好的识别性能及模糊统计方法是有效的。

关键词:支持向量机; 脱机手写汉字; 模糊统计特征; 汉字识别

中图分类号:TP182

文献标识码:A

文章编号:1673-629X(2006)10-0104-04

Research on Method of Off-Line Handwritten Chinese Characters Recognition Based on SVM

WANG Jian-ping, CHEN Jun, XU Xiao-bing, WANG Xi-hui

(School of Electric Engineering and Automation, Hefei University of Technology, Hefei 230009, China)

Abstract: In this paper, a new feature extraction method based on fuzzy statistic feature was proposed. SVM, The theory of small-sample statistical learning proposed by Vapnik, was used for off-line handwritten Chinese characters recognition. The feature date was extracted by three methods they are the density of Chinese characters stokes, wavelet transform and elastic meshing, and fuzzy statistic feature. The result of recognition shows that the SVM method can be used practically in off-line handwritten Chinese characters recognition and the new feature extraction method is effective and scientific.

Key words: SVM; off-line handwritten Chinese characters; fuzzy statistic feature; Chinese characters recognition

0 引言

20 世纪 60 年代以来, 学者们对汉字识别做了许多研究工作, 提出了很多方法。如: 模板匹配法、基于 K-L 数字变换的匹配方法、小波分析法等等。这些方法虽然大大地推动了汉字识别的发展。但是由于脱机手写汉字受到每个个体写字的差异性的影响, 所以脱机手写汉字的识别仍然是文字识别中的一个难题^[1]。

基于统计学的模式识别方法解决汉字识别问题有个基本前提, 只有在学习样本趋于无穷的时候, 识别性能才能在有理论上达到识别要求。这样就给实际的识别问题提出了很大的困难, 于是人们转向小样本机器学习的汉字识别。

Vapnik 等人在有限样本的机器学习问题的研究上取得了很大的进展, 并建立了一整套完整的理论体系——支持向量机 (Support Vector Machines, SVM)。这一新的理论方法在解决模式识别中小样本、非线性、及高维识别问题中表现出了独特的优势和应用前景^[2,3]。

文中应用提出的模糊统计的特征提取方法, 结合小波

网格的特征提取方法和笔画密度的特征提取方法对脱机手写汉字进行特征提取, 在小样本学习的情况下, 采用支持向量机方法进行脱机手写体汉字识别取得了良好的识别效果。

1 脱机手写汉字识别的 SVM 算法

1.1 问题概述

支持向量机算法是在结构风险最小化基础上, 对两种不同类别的样本数据找到一个最优分类面的方法。

对于支持向量机的两类模式识别问题, 考虑这样的样本: 它包含 n 个指标 ($x \in R^n$) 和 l 个样本点的集合, 记这 l 个样本点的集合为:

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\} \in (X \times Y)^l \quad (1)$$

其中: $x_i \in X = R^n$ 是输入指标向量, $y_i \in Y = \{1, -1\}$ 是输出指标, $i = 1, 2, \dots, l$ 。这 l 个样本组成的集合称为训练集。问题是对任意给定的一个新的模式 x , 根据训练集, 推断它所对应的输出 y 是 1 还是 -1。

1.2 脱机手写汉字识别的 SVM 分类定义

对于脱机手写汉字的两类模式识别问题, 考虑这样的手写汉字样本: 它包含 n 个指标 ($x \in R^n$) 和 l 个手写汉字样本点的集合, 汉字识别的分类问题: 记这 l 个手写汉字

样本点的集合为:

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\} \in (X \times Y)^l$$

其中: $x_i \in X = R^n, y_i \in Y = \{1, -1\}, i = 1, 2, \dots, l$ 。

寻找 $X = R^n$ 上的一个实值函数 $g(x)$, 以便决策函数

$$f(x) = \text{sgn}(g(x)) \quad (2)$$

推断任一模式(手写汉字) x 相对应的 y 值。也就是说手写汉字分类问题实质上就是寻找一个能将 R^n 上的点分成两部分的规则。

1.3 脱机手写汉字识别的 SVM 分类方法

脱机手写汉字识别的 SVM 的线性最优分类基本思想可用图 1 的两维情况说明。

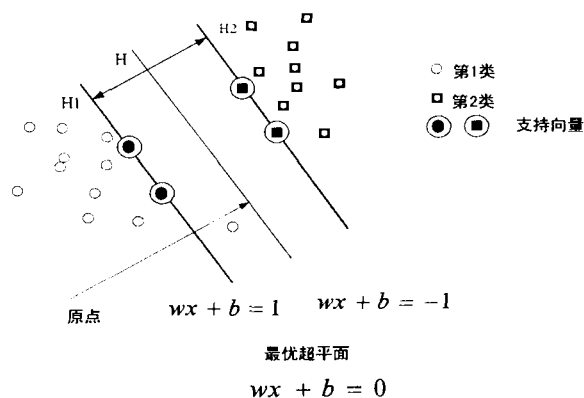


图 1 SVM 线性最优分类示意图

其中:圆点和方点代表两类手写汉字样本, H 为分类线, H_1, H_2 分别为过各类中离分类线最近的样本且平行于分类线的直线,它们之间的距离叫做分类间隔(margin)。所谓最优分类线就是要求分类线不但能将两类手写汉字正确分开(训练错误率为 0),而且使手写汉字分类间隔最大。手写汉字分类线方程为:

$$w \cdot x + b = 0 \quad (3)$$

基于最优分类线,可以找到两条平行于最优分类线的两条极端分类线使得两类手写汉字样本分别位于最优分类面的两侧。即可以得到 H_1, H_2 的方程为:

$$\begin{cases} w \cdot x + b = 1 \\ w \cdot x + b = -1 \end{cases} \quad (4)$$

此时,两条极端分类线之间的距离为 $\frac{2}{\|w\|}$,使分类间隔最大实际上就是解最优化问题:

$$\min_{w, b} \frac{1}{2} \|w\|^2 \quad (5)$$

(约束条件) s.t. $y_i((w \cdot x_i) + b) \geq 1, i = 1, \dots, l$ (6)

求得最优解 w^*, b^* ;

当手写汉字训练样本线性不可分时,即某些样本不满足式(6)时可以在条件中增加一个松弛项 $\xi_i \geq 0$,约束条件变为:

$$y_i((w \cdot x_i) + b) + \xi_i \geq 1, i = 1, \dots, l \quad (7)$$

此时最优分类面由下式确定:

$$\min_{w, b} \frac{1}{2} \|w\|^2 + C \left(\sum_{i=1}^l \xi_i \right) \quad (8)$$

其中: C 为某个指定的常数,通常称为惩罚因子,它实际上是对错分手写汉字样本惩罚作用,它的大小反应了惩罚的程度。

对于式(7), (8) 确定的具有线性约束的二次规划问题,采用拉格朗日乘子法求解其对偶问题:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) - \sum_{j=1}^l \alpha_j \quad (9)$$

$$\text{s.t.} \sum_{i=1}^l y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C \quad (10)$$

其中: C 为某个指定的常数, $i = 1, \dots, l; j = 1, \dots, l$ 。

考虑手写汉字识别是一类非线性分类问题,可使用一个非线性函数 ϕ 把样本数据映射到一个高维特征空间,再在该空间建立最优超平面,此时式(9)可变换为:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j \alpha_i \alpha_j K(x_i \cdot x_j) - \sum_{j=1}^l \alpha_j \quad (11)$$

其中 $K(x_i \cdot x_j) = \phi(x_i) \cdot \phi(x_j)$ 称为核函数,作为核函数它必须满足 Mercer 条件^[4](如:线性核函数、多项式核函数、高斯径向基核、多层感知器核等)。

对于手写汉字识别问题,笔者通过较多的实验仿真计算比较,采用高斯径向基核或多项式核函数的效果明显。

2 脱机手写汉字识别的 VC 维与期望风险

VC 维是统计学习理论里面的一个重要的概念^[5]。其定义为:假设集 F 是一个由 x 上取值为 1 或者 -1 的函数值组成的集合。定义 F 的 VC 维为:

$$\text{VC dim}(F) = \max \{m : N(F, m)\} = 2^m \quad (12)$$

即 F 的 VC 维就是它能打散的 x 中的点的最大个数,若任意数目的手写汉字样本都有函数能将它打散,那么函数的 VC 维就是 ∞ 。

对有限个数的手写汉字样本,仅仅用经验风险来近似期望风险是行不通的,统计理论有如下基于 VC 维的期望风险估计:

记 h 为 F 的 VC 维,实际风险 $R(f)$ 和经验风险 $R_{\text{emp}}(f)$ 之间至少以 $1 - \eta$ 的概率满足。

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{8}{l} (h (\ln \frac{2l}{h} + 1) + \ln \frac{4}{\eta})} \quad \eta \in (0, 1] \quad (13)$$

由上式可以看出机器学习不仅要使经验风险最小,而且也要使 VC 维尽可能的小。然而对一个实际手写汉字识别问题,这两者有相互矛盾的倾向。

文中在设计手写汉字识别的分类器时采用了基于结构风险最小化原则。先选择适当的分类器模型,使 VC 维较小,然后对模型进行参数估计,使经验风险最小。实验仿真结果表明,此方法设计的分类器可以很好地满足实际风险较小。

3 SVM 的手写汉字识别特征样本提取

手写汉字样本集的选取合理性关系到 SVM 分类效果。为了使 SVM 对手写汉字具有更好的识别效果,文中

采用 3 种方法对手写汉字特征提取进行样本数据融合。这 3 种方法融合的样本数据对手写汉字整体与笔画和统计与结构特征具有正交互补性表征,可以更好地用于小样本集的 SVM 识别^[6-8]。

3.1 二维小波分析的弹性网格特征样本提取

对汉字图像进行二维小波变换可以较好地得到汉字的结构特征。变换后得到的低频系数图像保持了图像轮廓的主要信息,高频图像反应了原图在不同方向的细节信息。其中水平高/低频图像 HL 反映了图像在水平方向的信息;垂直低/高频图像 LH 反映了图像在垂直方向上的信息;斜向高/高频图像 HH 反映了图像在斜向 45°/135° 方向的信息。利用二维小波分解提取较工整手写体汉字的横、竖和撇捺方向的结构特征。

考虑到手写汉字的不规则性,尽可能地利用图像的结构特征,采用的基于图像质心为中心的每副变换子图弹性 4 网格划分法来统计每个网格的灰度平均值作为手写汉字的特征样本之一。其计算方法如下:

设一幅汉字图像的像素点灰度值为 $c(i, j)$, $i = 1, 2, \dots, m$; $j = 1, 2, \dots, n$ 。那么有:

$$\left. \begin{aligned} G_i &= \sum_{i=1}^m \sum_{j=1}^n i \cdot c(i, j) / \sum_{i=1}^m \sum_{j=1}^n c(i, j) \\ G_j &= \sum_{i=1}^m \sum_{j=1}^n j \cdot c(i, j) / \sum_{i=1}^m \sum_{j=1}^n c(i, j) \end{aligned} \right\} \quad (14)$$

对 G_i, G_j 四舍五入取整,可得图像的质点坐标 (p, k) , $1 < p < m, 1 < k < n$ 。

每幅变换子图的 4 个网格的灰度平均值为:

$$\left. \begin{aligned} H_{11} &= \sum_{i=1}^p \sum_{j=1}^k c(i, j) / p \cdot k & t = 1, 2, 3, 4 \\ H_{12} &= \sum_{i=1}^p \sum_{j=k}^n c(i, j) / p \cdot (n - k) & t = 1, 2, 3, 4 \\ H_{13} &= \sum_{i=p}^m \sum_{j=1}^k c(i, j) / (m - p) \cdot k & t = 1, 2, 3, 4 \\ H_{14} &= \sum_{i=p}^m \sum_{j=k}^n c(i, j) / (m - p) \cdot (n - k) & t = 1, 2, 3, 4 \end{aligned} \right\} \quad (15)$$

其中:低频、水平高/低频、垂直低/高频、斜向高/高频系数图像分别以 $t = 1, 2, 3, 4$ 表示。

可得脱机手写汉字的小波结构特征样本集为:

$$H = [H_{11} \ H_{12} \ H_{13} \ H_{14} \ H_{21} \ H_{22} \ H_{23} \ H_{24} \ H_{31} \ H_{32} \ H_{33} \ H_{34} \ H_{41} \ H_{42} \ H_{43} \ H_{44}] \quad (16)$$

3.2 脱机手写汉字的笔画密度特征提取

笔画密度特征能较好地反映手写汉字的整体和笔画分布特征。在一幅 $m \times n$ 的汉字点阵中,分别进行 0°, 90°, 45° 和 135° 方向线扫描投影,并对扫描线上的笔画灰度值累加统计,构成汉字在 4 个方向上的笔画密度直方图。文中对 4 个方向分别选取 24 根扫描线进行扫描,得到汉字的 96 维笔画特征(如“辶”字手写体的垂直直方图如图 2 所示)。

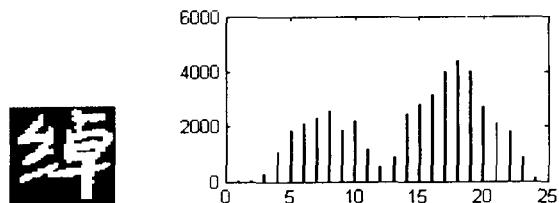


图 2 汉字‘辶’的垂直直方图

可得脱机手写汉字的笔画密度特征样本集为:

$$M = \begin{bmatrix} Mh_1 & Mh_2 & \cdots & Mh_{23} & Mh_{24} \\ Ms_1 & Ms_2 & \cdots & Ms_{23} & Ms_{24} \\ Mp_1 & Mp_2 & \cdots & Mp_{23} & Mp_{24} \\ Mn_1 & Mn_2 & \cdots & Mn_{23} & Mn_{24} \end{bmatrix} \quad (17)$$

其中: $Mh_1 \sim Mh_{24}, Ms_1 \sim Ms_{24}, Mp_1 \sim Mp_{24}, Mn_1 \sim Mn_{24}$ 分别表示 0°, 90°, 45° 和 135° 方向线扫描投影笔画灰度累加值。

3.3 汉字轮廓结构模糊统计特征提取

文中提出了一种基于汉字轮廓结构直方图的模糊统计特征。汉字水平和垂直密度的直方图可以较好地反应汉字的结构及笔画特征,如直方图的波谷点(见图 2),对水平直方图和垂直直方图上汉字的波谷个数、波谷点位置、直方图的均值及直方图的方差进行统计可以得到汉字的结构及笔画的统计特征。采用二维小波变换的低频/低频图像统计其轮廓点数作为手写汉字的轮廓统计特征。最后形成一组汉字直方图 11 维的统计特征,其中波谷点位置用 2 维表示,文中实验用的汉字归一化为 24×24 点阵的灰度图,通过大量的观察和比较,给出如下定义和结论:

定义 1 设直方图的包络曲线满足 $H_i = f(i)$, 其中 $i = 1, 2, \dots, 24$, $H_M = \max\{H_1, H_2, \dots, H_{24}\}$, 其中 $H_i, i = 1, 2, \dots, 24$ 为各灰度的累加值。满足以下条件的点 (j, H_j) ($0 < j \leq 24$) 为谷点:

$$1) \ H_{j-1} - H_j > N, H_{j+1} - H_j > N$$

如果 $H_{j-1} = H_j$, 则比较 H_{j-2}, H_j , 此时需满足: $H_{j-2} - H_j > N, H_{j+1} - H_j > N$, 若还有 $H_{j-2} = H_j$, 比较 H_{j-3}, H_j , 依次类推。同理如果 $H_{j+1} = H_j$, 比较 H_{j+2}, H_j 。其中 N 为给定的阈值。 $0 < N < 255$ 。

$$2) \ H_j < H_{M/3}。$$

结论 1 谷点数可以反映汉字的结构特征,水平直方图中单体字的谷点数为 0,上下结构的汉字的谷点数为 1,上中下结构的汉字的谷点数为 2。同样垂直直方图中的谷点数反映汉字的左右结构特征。

在一幅 24×24 的手写字符图像中对各汉字的谷点处统计其模糊特征,并给出如下定义。

定义 2 对汉字的水平直方图,峰值点横坐标小于 6 定义其为‘上’的隶属度为 1,‘中’和‘下’的隶属度为 0;峰值点横坐标大于 10 小于 14 定义其为‘中’的隶属度为 1,‘上’和‘下’的隶属度为 0;当峰值点横坐标大于 18 定义其为‘下’的隶属度为 1,‘上’和‘中’的隶属度为 0;形

成一个梯度隶属度函数。如图 3 所示。

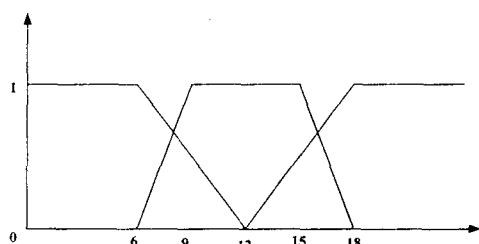


图 3 汉字的隶属度函数

对汉字的垂直直方图,其左右结构也用以上梯度隶属度函数(见图 3)表示。于是可以得到汉字的谷点位置的隶书度的信息,文中用一个 6 维的特征向量表示。

提取汉字轮廓结构的模糊统计特征最后得到一个 17 维向量:

$$F = [F_1, F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9, F_{10}, F_{11}, F_{12}, F_{13}, F_{14}, F_{15}, F_{16}, F_{17}]$$

其中 F_1 表示手写体汉字的轮廓特征信息; $F_2 \sim F_{11}$ 为汉字的结构特征信息; $F_{12} \sim F_{17}$ 为谷点的模糊特征信息。

4 识别试验

文中以金连文实验室的手写字体为实验样本,每个汉字有样本 40 个,其中 30 个作为训练样本,10 个该汉字样本加上其他 10 个汉字样本共 20 个样本作为待识别样本,实验选取了 10 个汉字共 400 个样本。特征提取前先对样本预处理,其中归一化采用插值法,把样本归一化为 2424 点阵。去噪采用中值滤波,特征提取结合小波网格、笔画密度特征和直方图的模糊统计特征方法提取汉字的特征,并加入了部分冗余维特征。最后形成了一个 140 维特征。

采用 Matlab 的支持向量机工具箱:OSU-SVM3.00。分别选取高斯径向基核和多项式核作为核函数进行训练识别,其识别结果如表 1 所示。

5 结论

从识别结果可以看出,支持向量机对较规则手写字符有比较好的识别结果,对于小样本识别问题,用支持向量

机训练识别是可行的,文中所采用的笔划密度特征和直方图统计特征及模糊统计特征方法对手写汉字的特征提取及数据融合是科学有效的。另外对于手写字符撇捺方向的变形,可能会导致识别率的下降,如何解决此问题还有待进一步的研究。

表 1 基于支持向量机方法的手写汉字识别结果

字符	识别样本个数	多项式核识别率	多项式核误识率	径向基核识别率	径向基核误识率
大	20	0.90	0.10	1.00	0.00
吹	20	1.00	0.00	1.00	0.00
处	20	0.90	0.10	0.85	0.15
川	20	1.00	0.00	1.00	0.00
计	20	0.90	0.10	0.95	0.05
崔	20	0.95	0.05	0.95	0.05
呆	20	1.00	0.00	1.00	0.00
寸	20	1.00	0.00	0.95	0.00
从	20	0.80	0.20	0.85	0.15
括	20	1.00	0.00	0.90	0.10

参考文献:

- [1] 张中. 汉字识别技术综述[J]. 语言文字应用, 1997(2): 77-86.
- [2] Vapnik V N. The Nature of Statistical of Learning Theory [M]. New York Spring, 1995.
- [3] 柳回春, 马树元. 支持向量机的研究现状[J]. 中国图像图形学报, 2002(6): 619-623.
- [4] 王晓光, 王群. 用于车牌字符识别的 SVM 算法[J]. 现代电子技术, 2004(8): 8-10.
- [5] 张学工. 关于统计学习与支持向量机理论[J]. 自动化学报, 2000(1): 32-42.
- [6] 金连文, 彭秀兰, 尹俊勋. 一种手写体汉字特征提取新方法—小波变换及弹性网络技术的应用[J]. 中国图像图形学报, 1998(7): 499-552.
- [7] 王志红. 小波和神经网络模式识别技术及其在车牌识别中的运用[D]. 合肥: 合肥工业大学, 2003. 30-55.
- [8] 边肇祺, 张学工. 模式识别[M]. 北京: 清华大学出版社, 2002. 284-325.

(上接第 103 页)

变步长法、加动量项法、串连法、利用遗传算法优化 BP 算法、采用模拟退火算法改进 BP 算法。其中, 前三种方法在加快网络收敛速度方面比较显著, 后两种方法主要是避免网络学习陷入局部极小点从而能够收敛到全局最小点。在实际应用中表明, 将以上几种方法结合使用, 网络的学习效果将更理想。

参考文献:

- [1] Weymace N, Martens J P. A fast and robust learning algorithm for feedforward neural networks[J]. Neural Networks, 1991, 4(3): 363-369.

- [2] Chan L W, Fallside F. An adaptive training algorithm for back propagation networks[J]. Computer Speech and Language, 1987, 2: 205-218.
- [3] Fahlman S C, Lebiere C. The Cascade Correlation Learning Architecture[J]. Advance in neural information processing systems, 1990, 2: 524-532.
- [4] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagation errors[J]. Nature, 1986, 323: 533-536.
- [5] 阎平凡, 张长水. 人工神经网络与模拟进化计算[M]. 北京: 清华大学出版社, 2001.
- [6] Kirkpatrick S, Gelatt C D, Vecchi M P. Optimization by simulated annealing[J]. Science, 1983, 220: 671-680.