第32卷 第23期 计 算 机 工 程 2006年12月
Vol.32 No.23 Computer Engineering December 2006
·人工智能及识别技术· 文章编号：1000—3428(2006)23—0199—03 文献标识码：A 中图分类号：TP391.43

ICA和改进的SVM在有限集字符识别中的应用

鹿晓亮，陈继荣，黄戈祥
(中国科学技术大学电子工程与信息科学系，合肥 230027)

摘 要：介绍了独立分量分析(ICA)基本原理和算法，提出了一种基于独立分量分析和支持向量机的有限集字符识别新方法。对传统向量机解决多分类问题的“一对一”模式进行了改进，将传统向量机的“一对一”模式存在的不可分区域减小到可以忽略的程度，克服了不可分区域的影响。该算法可应用于车牌字符、手写体英文字母、手写体数字、印刷体字母、印刷体数字等有限集字符的识别。在大量的车牌汉字和手写体英文字母自动识别实验中，取得了高于95%的识别结果，证明该算法在有限集字符识别应用中的优越性。
关键词：独立分量分析；支持向量机；特征提取；字符识别

Application of Independent Component Analysis and Improved Support Vector Machines on Finite Set Character Recognition

LU Xiaoliang, CHEN Jirong, HUANG Gexiang
(Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027)

【Abstract】This paper introduces the principle and algorithm of independent component analysis (ICA) and then puts forward a new method to recognize finite set characters utilizing ICA and support vector machines(SVM). It develops the “one-versus-one” model of multi-classification of traditional SVM and at the same time reduces the influence of undivided area. The presented algorithm can be applied to the recognition of license plate characters, handwritten English letters, handwritten numbers, printed letters, printed numbers and other finite set characters. The experiments to recognize Chinese characters on license plates and handwritten English characters can achieve a recognition rate of more than 95%. The results show that this algorithm holds advantage in the finite set characters recognition.
【Key words】Independent component analysis; Support vector machines; Feature extraction; Character recognition

1 概述

有限集字符识别主要有牌照的自动识别、银行票据的自动识别、手写体英文字母(或数字)的识别等应用领域，这些应用可以节省大量的人力物力，有着广泛的应用前景。跟其他多分类问题相比较，有限集字符识别的明显特征是类别数目比较小，仅有常用的几个到几十个汉字、字母、数字或其他字符。由于其应用受到环境的制约，比如牌照的识别，车牌字符可能存在一定的变形和缺损；而银行票据及手写体字母(或数字)的识别，由于每个人的书写风格不同，随机性大，断笔、连笔、字符形变因人而异。这些不确定性因素给自动识别带来了困难，使有限集字符识别成为模式识别中的难题。

独立分量分析 (Independent Component Analysis, ICA)是近年来由盲信源分解技术发展起来的信号处理技术。其基本思想就是将多道观察信号根据统计独立的原则通过优化算法分解成相互独立的成分，从而实现信号的增强和分解。其在图像识别方面应用的基本过程是对样本图像进行ICA分解，分解成相互独立的基图像元，然后对于任何一个待分类图像，用分解出来的基图像元线性组合得到，各个基图像元的组合系数(即待分类图像在基图像元空间的投影系数)就构成了待分类图像的特征向量，从而完成了待分类图像的特征提取。本算法中采用独立分量分析的方法进行字符特征提取。

支持向量机 (Support Vector Machines, SVM)也是近年来由统计学理论发展起来的一种新的机器学习方法，它采用了统计学习理论中结构风险最小化的处理原则，在处理高维空间分类问题中显示出很强的优越性。它是为解决小样本问题学习和分类提出的，有很强的非线性分类能力，并能克服神经网络所固有的过学习和欠学习的缺点。因而本算法中采用支持向量机方法对字符特征进行分类。

在大量的车牌汉字和手写体英文字母识别实验中，取得了高于95%的识别结果，证明该算法在有限集字符识别应用中的优越性。

2 基于独立分量分析ICA的字符特征提取

独立分量分析 ICA 的模型定义在文献[1,3]中有详细介绍，本文不再赘述。ICA 分析的目的在于根据观察信号 x，估算未知的混合矩阵 A 和源信号 s。也等效于寻找分离矩阵 W，满足

$$y=Wx \tag{1}$$

使 y 各分量相互独立，或者说 y 逼近 s。

本文采用基于最大熵理论的负熵估计，得到如下的目标函数：

$$J_G(w)=[E\{G(w^Tx)\}-E\{G(v)\}]^2 \tag{2}$$

式中 x 为观察信号，w 为权值向量，v 为零均值、单位方差的高斯变量，G 为非线性函数，可采用如下形式：

$$G(u)=\log\cosh(u) \tag{3}$$

作者简介：鹿晓亮(1979－)，男，硕士生，主研方向：模式识别，图像处理；陈继荣，副教授；黄戈祥，硕士生
收稿日期：2005-12-22 **E-mail：**grantlu@ustc.edu

2.1 用定点算法(FastICA)求解分离矩阵 W

定点算法(FastICA)也叫快速 ICA，其基本原理是通过随机梯度法调节矩阵 W 来优化目标函数。式(2)中使 $J_G(w)$ 最大的最优解满足下述方程：

$$E\{xg(w^T x)\} - \beta w = 0$$
$$E\{(w^T x)^2\} = \| w \|^2 = 1 \quad (4)$$

式中 g 为 G 函数的导数，$\beta = E\{w_0 x g(w_0{}^T x)\}$，$w_0$ 是 w 在最优化时的取值。用牛顿迭代法求解这个方程，用 w 的瞬时值代替 w_0，则变成

$$w^+ = E\{xg(w^T x)\} - E\{g^{'}(w^T x)\}w$$
$$w^+ = w^+ / \| w^+ \| \quad (5)$$

定点算法的优点是收敛速度快，计算量小，不必选择收敛步长。

2.2 字符的特征提取

特征提取是模式识别的最关键环节之一，本算法先对字符样本进行 ICA 分离，求解分离矩阵 W，得到字符的基图像元。对于待分类字符，由基图像元通过线性组合而成，将各基图像元的权值作为待分类字符的特征向量。

2.2.1 预处理

对字符样本进行 ICA 分离前，必须做一些必要的预处理。首先将字符归一化成 32*32 大小，再将各字符图像按行叠加成一个列向量 f_i(i=1…m)，构成输入信号 f，即

$$f = [f_1, f_2, ... f_m]^T \quad (6)$$

然后对输入信号 f 作零均值处理：

$$f_i = f_i - E(f)$$
$$E(f) = \frac{1}{m}\sum_{i=1}^{m} f_i \quad (7)$$

最后对输入信号 f 作白化处理：

$$x = Vf, \quad E(xx^T) = I \quad (8)$$

白化处理的目的就是使白化信号 x 的各分量相互正交，简化问题的处理。白化矩阵 V 的确定方法如下：

$$\mathrm{cov}(f, f) = E(ff^T) - [E(f)]^2 = E(ff^T) = EDE^T$$
$$V = D^{-\frac{1}{2}}E^T \quad (9)$$

式中 D 矩阵为 f 的协方差矩阵特征值构成的对角阵，E 矩阵为协方差矩阵特征值对应的特征向量构成的矩阵。

经零均值、白化预处理后，输入信号变成具有零均值、单位方差的信号，满足了 ICA 模型定义对输入信号的要求。

2.2.2 FastICA 求解分离矩阵 W

我们采用定点算法对分离矩阵 W 进行估算，估算迭代进行，一次仅估算一个独立分量 w_i。

(1)随机初始化 w_i 并对 w_i 进行标准化：

$$w_i = \frac{w_i}{\|w_i\|} \quad (10)$$

(2)对 w_i 正交化去相关并标准化：

$$w^* = w_i - \sum_{j=1}^{i-1} w_i^T w_j w_j \quad (11)$$

对 w^* 标准化：

$$w^* = \frac{w^*}{\|w^*\|} \quad (12)$$

(3)若 $w^* \approx w_i$，回到步骤(2)计算 w_{i+1}；当所有的 w_i 计算结束时，跳到步骤(5)。

(4)若 $w^* \neq w_i$，则对 w_i 进行如下的调整：

$$w_i = E\{xg(w_i^T x)\} - E\{g'(w_i^T x)\}w_i \quad (13)$$

式中 g 函数为非线性函数 G 的导数，g' 函数为 g 函数的导数，回到步骤(2)。

(5)所在的权值向量 w_i 调整完毕，计算分离矩阵 $\boldsymbol{W}$。

$$\boldsymbol{W} = \begin{bmatrix} w_1^T V \\ w_2^T V \\ \cdot \\ \cdot \\ \cdot \\ w_m^T V \end{bmatrix} \quad (14)$$

估算得到分离矩阵 $\boldsymbol{W}$ 后，可以由式(1)得到基图像元 Y，计算过程如下：

$$Y = W(f + E(f)) \quad (15)$$

式中，$Y = [y_1, y_2, ... y_m]^T$。由 ICA 模型定义可知，任何一个字符样本图像 f_i 都可以由基图像元 Y 线性组合而成：

$$f_i = \sum_{j=1}^{m} a_{ij} y_j = \mathrm{a}_i^{\mathrm{T}} Y \quad (16)$$

式中 a_{ij} 为字符图像 f_i 在基图像空间的投影系数，$\mathrm{a}_i^{\mathrm{T}} = (a_{i1}, a_{i2}, ..., a_{im})$，$\mathrm{a}_i^T$ 就是字符图像 f_i 的特征向量。原样本图像 f 投影在基图像元 Y 空间中，得如下投影方程：

$$f = AY \quad (17)$$

式中，$A = (\mathrm{a}_1, \mathrm{a}_2, ..., \mathrm{a}_m)^{\mathrm{T}}$，f 和 Y 已知，可以由最小二乘法求得混合矩阵 A。

同样地，对于任一待分类字符图像，都可以用同样的方法投影到基图像元 Y 空间，求得各投影系数，得到特征向量。

3 支持向量机

支持向量机理论基本思想就是寻找最优分类超平面，将数据分成两类，其分类平面既要保证准确分类，又要最大化分类平面两侧空白区域。其基本原理请参考文献[2]，这里不做详细介绍。

3.1 传统支持向量机的多分类算法及其存在的问题[4]

传统的支持向量机在处理多类别问题时，通常采用“一对一”或“一对多”模式来处理。所谓“一对一模式”，就是从 n 类问题中任意抽取两类模式进行学习，形成 $n(n-1)/2$ 个判决函数；而“一对多”模式将第 i 类问题与剩余的问题形成两类进行学习，总共得到 n 个判决函数。在非线性分类性能上，“一对一”模式要优于“一对多”模式。因为“一对一”模式任意两类之间都存在一个分类超平面，而“一对多”模式仅用一个分类超平面将某类与其他各类分离。本算法中考虑到类别数目不大，因此我们采用“一对一”模式。传统“一对一”模式描述如下：

第 i 类与第 j 类的判断函数为

$$D_{ij}(x) = w_{ij}x + b_{ij} \quad (18)$$

式中 w_{ij} 为 n 维向量(即第 i 类与第 j 类分类超平面的法向量)的一个分量，b_{ij} 为标量。对于一个待分类的向量 x，进行如下计算：

$$D_i(x) = \sum_{j=1, j\neq i}^{n} \mathrm{sgn}(D_{ij}(x)) \quad (19)$$

将向量 x 划分为第 k 类别，$D_k(x) = \max(D_i(x))$。当符合这一条件的 k 不止一个时，将不可分类。如图 1 所示。

图 1 所示，有 3 个类别，在阴影区域内，有 $D_i(x) = 1$，(i=1,2,3)。此时，不知道将向量 x 该划分为哪一类。

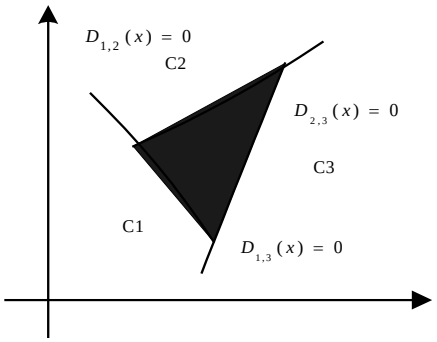

图 1“ 一对一 ”模式中不可分区域示意图

3.2 改进的“ 一对一 ”模式

从“ 一对一 ”模式的式(19)可以发现，传统“ 一对一 ”模式划分规则只是将向量x归类入两两分类中被分类到的次数最大的那一类，而无视向量x距离各分类面的距离远近程度。本文对传统的“ 一对一 ”模式稍加改进，目的在于减小不可分区域到可以忽略的程度，使不可分区域的影响最小。改进的“ 一对一 ”模式仍沿用式(19)的分类函数，当 $D_k(x)=\max(D_i(x))$ 的k值不止 1 个时，我们对符合 $D_k(x)=\max(D_i(x))$ 的$D_k(x)$按下式重新计算:

$$D_i(x)=\sum_{j=1,j\neq i}^{n}\{\mathrm{sgn}(D_{ij}(x))*g_{ij}(|D_{ij}(x)|) \tag{20}$$

式中，函数 $g_{ij}(u)$ 定义如下：

$$g_{ij}(u)=\begin{cases}1-\exp[-\frac{(u-t_{ij})^2}{\sigma_{ij}^2}] & u\geq 0\\ 0 & u<0\end{cases} \tag{21}$$

式中，t_{ij} 表示第i类所有样本x_d的距离$|D_{ij}(x_d)|$的平均值，σ_{ij} 为第i类所有样本x_d的距离$|D_{ij}(x_d)|$的方差。该式中描述的分类函数利用了向量x到各类面的距离信息，将落在不可分区域中的向量x进行再分类，使真正不可分的区域减小到可忽略的程度。用该式计算出各$D_k(x)$后，再将待分类向量x归入$\max(D_k(x))$那一类。按该式计算的$D_k(x)$亦可能会出现不可分情况，但由于考虑了向量x到各分类面的距离信息，出现不可分的可能性(即不可分区域)将缩小将近 0。

4 实验结果及分析

我们分别以车牌汉字字符和手写体英文字母作为试测数据对该算法进行测试，实验中首先将字符均归一化成 32*32 大小。在车牌汉字的识别实验中，取 28 个省(由于实验中其余几个省的车牌无法获得)的车牌汉字共280个进行ICA分解(每个汉字取 10 幅字符图像，灰度为 256 级，如图 2 所示)，另外取 3 000 多个车牌汉字作为测试；在手写体英文字母的识别实验中，我们取 10 个人写的 400 个手写英文字母(如图 3 所示)进行 ICA 分解，再分别取这 10 个人写的 5 000 多个英文字母进行测试；另外，对这些实验数据用传统距离分类器进行了分类，实验结果如表 1、表 2 所示。

表 1 车牌汉字的识别结果(单位：%)

分类结果 \ 分类器	距离分类器	传统的“ 一对一 ”模式的 SVM	改进的“ 一对一 ”模式的 SVM
正确率	67.43	93.13	95.28
错误率	32.57	4.69	4.72
拒识率	0	2.18	0.00

表 2 手写体英文字母的识别结果(单位:%)

分类结果 \ 分类器	距离分类器	传统的“ 一对一 ”模式的 SVM	改进的“ 一对一 ”模式的 SVM
正确率	74.02	94.06	96.20
错误率	25.98	3.79	3.80
拒识率	0	2.15	0.00

图 2 部分车牌汉字图像

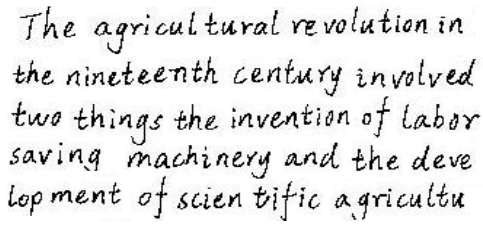

图 3 部分手写体英文字母

在对车牌汉字的识别测试中，传统的“ 一对一 ”模式有 67 个汉字被拒识，而在改进的“ 一对一 ”模式中仅有 3 个汉字被拒识；在手写体英文字母识别实验中，改进的“ 一对一 ”模式将拒识的字符数降低到 7 个。由表 1、表 2 数据对比可知，支持向量机 SVM“ 一对一 ”模式分类器的识别率远高于距离分类器，而改进的“ 一对一 ”模式 SVM 分类器跟传统的“ 一对一 ”模式 SVM 分类器相比，在拒识率上降低到 0.001%级别。该算法在以上两种应用中均获得了高于 95%的识别率，而在手写体英文字母识别中，获得了 96.20%的识别率，与文献[5]效果相当，但在抗噪声能力方面优于文献[1,5]所提出的算法。

5 结论

本文提出的算法利用了独立分量分析的噪声消除优点以及支持向量机良好的非线性分类能力，并对传统“ 一对一 ”模式 SVM 进行了改进，将传统“ 一对一 ”模式的不可分类区域减小到可以忽略的程度。实验结果表明，该算法非常适用于有限集字符的识别。本文提出的算法也适用于其他小类别分类的场合。

参考文献

1 Ozawa S, Tsujimoto T. Application of Independent Component Analysis to Handwritten Japanese Character Recognition[C]. Proc. of International Joint Conference on Neural Networks, 1999: 2867-2871.

2 孙　亮. 支持向量机及其在目标识别中的应用[J]. 信息与控制, 2003, 32(1).

3 范　羚, 吴小培, 龙　飞. 基于独立分量分析的图像特征提取及去噪[J]. 计算机工程与应用, 2003, 39(9).

4 Wu Jing, Zhou Jianguo. Incremental Proximal Support Vector Classifier for Multi-class Classification[EB/OL]. 2004:3201-3206. http://ieeexplore.ieee,org/iel5/9459/30022/01378587.pdf?tp=& arnumber =1378587&isnumber=30.

5 沈淑娟, 姜建国. 手写体字符识别的多特征多分类器设计[J]. 计算机工程与应用, 2004, 40(16).