高技术通讯 2012 年 第22 卷 第3 期:249 ~255

doi:10.3772/j.issn.1002-0470.2012.03.004

相似模式类鉴别分析方法①

王言伟② 刘长松 丁晓青

(清华大学电子工程系 清华信息科学与技术国家实验室 智能技术与系统国家重点实验室 北京 100084)

摘 要 针对多类识别时原始特征空间中相近的类经过线性鉴别分析(LDA)降维后,在低维空间中易被混淆,不利于识别的问题,提出了一种通过对相似类对抽取鉴别向量构成特征变换矩阵的相似模式类鉴别分析(SPDA)方法,并将该方法与 LDA 降维相结合,应用于级联改进二次鉴别函数(MQDF)分类器中,实现了对手写汉字识别性能的进一步提高。在脱机手写汉字字符集2000(HCL2000)上的识别率为 98.82%,识别结果高于可查文献中相应的识别结果,这表明该方法是有效的。

关键词 线性鉴别分析,相似模式鉴别,级联分类器,脱机手写汉字识别

0 引 言

基于统计的模式识别,特别是多类识别,为了获得较高的识别性能,提取特征的维数往往较高,这对分类器的训练是一个巨大的挑战。模型参数的数量随着特征维数的增加而增加,而现实中得到的样本却相对较少。当特征维数相对于训练样本的规模仍然较高时,分类器训练时容易导致过学习,参数估计误差大。最终训练出来的分类器泛化性能差,识别性能反而不好。为了解决此问题,研究者将 Fisher 线性鉴别分析(linear discriminant analysis, LDA)[1]扩展为多类的 LDA[2],用于多类分类问题中的特征降维。经过特征降维后,过学习的问题可以得到缓解。

近年来的一些研究结果表明,多类 LDA 降维应用于分类还有一些局限性。多类 LDA 降维关注的是所有类之间的区分性,是一个全局的优化过程。在降维过程中,全局协方差矩阵中较小的特征值对应的投影方向将被丢弃;而原始特征在其它方向上的投影则可能造成低维空间中相近的类分布重叠更加严重[3],最终,就会导致这些类的错误率较高。Luis[4]直接将多类分类问题转化为两两的分类问题,并用异方差方法进行鉴别分析。对具有 C 个类的分类系统而言,则需要解决 $C(C-1)/2$ 个两类的分类问题。可以看到,该方法对类别数量较少的识别系统如数字、英文字符等比较实用,而对于汉字等超多类的分类问题,转化为两类的分类将会产生大量的两类问题。分类器数量的急剧增长和由此带来的匹配难的问题限制了该方法的应用。当分类器性能达到一定程度后,这些少数的相似类之间的误识占了整个分类器错误的大部分比例,因此,降低相似类识别错误率是进一步提升分类器性能的关键。鉴于多类 LDA 降维对相似模式类鉴别的不足,本文提出了一种相似模式类鉴别分析(similar pattern discriminant analysis, SPDA)方法。SPDA 方法通过构建相似类的鉴别投影矩阵,对相似模式类有针对性地提取鉴别特征。本文将传统 LDA 方法与 SPDA 方法相结合,分别应用于级联的改进二次鉴别函数(modified quadratic discriminant function, MQDF)分类器[5,6]中的第一级分类器和第二级分类器,进一步提高了手写汉字识别率。

1 SPDA 方法

SPDA 方法是一种能够对相似类给予有效的鉴别特征分析的方法,相似模式鉴别分析矩阵的 n 个投影向量是由相似度最高的相似类对的鉴别向量组成。相似类对的选取将在第2 节详细介绍。具体方法为:针对每个相似类对进行 Fisher LDA,抽取该相

① 国家自然科学基金(60933010)资助项目。

② 男,1984 年生,博士生;研究方向:模式识别,图像处理,汉字识别;联系人,E-mail: wyw636@gmail.com

(收稿日期:2010-09-15)

似类对的有效鉴别投影向量;由对 n 个相似类对抽取的有效鉴别向量构成一个相似模式鉴别分析投影矩阵。

1.1 单个相似类对鉴别投影向量

单个相似类对鉴别投影向量的抽取方法采用 Fisher LDA 的方法

$$J(\boldsymbol{W}^*) = \arg\max_{\boldsymbol{W}} \left| \frac{\boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W}}{\boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W}} \right| \tag{1}$$

其中:

$$\boldsymbol{S}_b = \sum_{i=1}^{2} p(w_i)(\boldsymbol{m}_i - \boldsymbol{m}_0)(\boldsymbol{m}_i - \boldsymbol{m}_0)^{\mathrm{T}}$$

$$\boldsymbol{S}_w = p(w_1)\boldsymbol{S}_{w_1} + p(w_2)\boldsymbol{S}_{w2}$$

$\boldsymbol{S}_b$ 和 $\boldsymbol{S}_w$ 分别为相似类对的类间散度矩阵和类内散度矩阵。$\boldsymbol{m}_1$,$\boldsymbol{m}_2$ 分别为两个类的均值,$\boldsymbol{m}_0$ 为两个相似类的总体均值。$p(w_i)$ 为类先验概率,文中取0.5,$\boldsymbol{S}_{w_1}$,$\boldsymbol{S}_{w2}$ 分别为两类的协方差矩阵。上述的优化问题通常被转化为特征值的求解问题,即

$$\boldsymbol{S}_b \boldsymbol{W} = \lambda \boldsymbol{S}_w \boldsymbol{W} \tag{2}$$

两类问题中 $\boldsymbol{S}_b$ 的秩为 1,因此,只能得到一个有意义的投影方向,即 $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ 最大的特征值对应的特征向量。该方向是在 Bayesian 准则下最具有鉴别性的投影方向。

1.2 相似模式类鉴别投影矩阵

n 个相似类对对应 n 个两类 LDA 变换,能够得到 n 个投影方向,即 $\boldsymbol{W}_i, i = 1,2,\cdots,n$。将这些特征向量进行组合得到一个综合的投影矩阵

$$\boldsymbol{W}_s^* = [\boldsymbol{W}_1, \boldsymbol{W}_2, \cdots, \boldsymbol{W}_i, \cdots, \boldsymbol{W}_n]$$

n 表示选取 n 个相似类对抽取投影向量,n 的大小对特征的鉴别性能有一定的影响。n 过小时,投影后得到的特征维数低,只对少数相似类产生鉴别作用;$n = 1$ 时,退化为 Fisher LDA。n 过大时,特征维数过高,训练分类器的样本数相对不足,容易造成过学习。因此,n 的大小一般通过实验方法选定。

对于参与鉴别分析向量抽取的相似类对,组合后的投影矩阵中至少包含其一个最具有鉴别力的投影方向。该鉴别分析方法是专门针对相似类进行的,在本文中称为相似模式类鉴别分析(SPDA)方法。

2 相似类对的选取

在众多的模式识别问题中都存在着相似类的鉴别问题,本文以手写汉字识别为例说明相似类对的选取并验证相似类鉴别分析方法。在汉字识别问题中,相似类对即相似字符对。相似字符对可以通过手工的方法选取,但工作量很大而且选取的相似字符是代表人对字符识别的能力而非分类器的识别能力。不同分类器对相似字符的鉴别性能是不同的,本文利用与最终分类器相同的改进的二次鉴别函数(MQDF)分类器自动选取相似字符对。如此选出的相似字符对是在该分类器下,分类不稳定和易混淆的字符对。利用分类器对训练集进行识别,对识别结果进行统计并建立混淆度矩阵,从混淆度矩阵中选取相似字符对。在统计意义上,混淆度矩阵能够描述分类器的分类性能及在该分类器上容易发生误识的相似字符的对应关系。一个 $C \times C$ 的混淆度矩阵如式

$$\boldsymbol{F} = \begin{bmatrix} N_{11} & N_{12} & \cdots & N_{1C-1} & N_{1C} \\ N_{21} & N_{22} & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ N_{C-11} & \cdots & \cdots & N_{C-1C-1} & N_{C-1C} \\ N_{C1} & \cdots & \cdots & N_{CC-1} & N_{CC} \end{bmatrix} \tag{3}$$

所示,其中,N_{ij} 为第 i 类识别为第 j 类的样本数,N_i 为第 i 类的总样本数。字符 i 类和 j 类之间的相似度 SD_{ij} 定义为

$$SD_{ij} = \frac{N_{ij} + N_{ji}}{N_i + N_j} \tag{4}$$

SD_{ij} 越大表明两个类的相似程度越大。通过对所有字符对的相似度按照降序排列得到相似字符对列表,记为 SL。后续的实验中将选取列表中相似度相对较大的一部分字符对进行相似模式类鉴别分析,这些选取的相似字符对对应的列表记为 TSL。部分 TSL 示例如图 1 所示。

曰日 0.1040	鸣鸟 0.0429	孪孪 0.0250	票栗 0.0221
晴睛 0.0800	已己 0.0343	谓渭 0.0250	汁计 0.0221
千干 0.0671	拨拔 0.0314	请清 0.0243	未末 0.0214
竞竟 0.0607	戍戌 0.0300	淮淮 0.0243	挟抉 0.0214
已巳 0.0550	谭潭 0.0293	束柬 0.0236	气乞 0.0214
孟孟 0.0550	漫漫 0.0293	论沦 0.0236	刁刀 0.0214
抉扶 0.0479	乌鸟 0.0293	洋详 0.0229	设没 0.0214
菜菜 0.0443	涌诵 0.0271	打订 0.0221	谁准 0.0207

图 1 HCL2000 训练集相似字列表 *TSL* 及其相似度

3 级联分类器

通过多类 LDA,降维后的特征对大多数字符类具有较强的鉴别性,但同时降维过程也会加剧相似

字符类之间的混淆。SPDA 是以最大化区分相似字为目标,能够有效地弥补多类 LDA 降维的不足。本文通过 MQDF 级联分类器将这两种鉴别分析方法有效地结合起来。级联分类器训练流程图如图 2 所示。

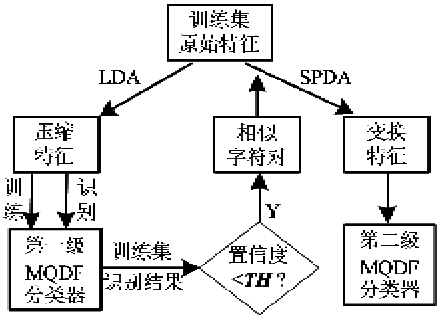

图 2　级联分类器训练流程图

分类器训练时参数估计的方法均采用最大似然估计。第一级 MQDF 分类器是在多类 LDA 降维后的特征上训练的,主要针对大部分字符类的识别;训练第二级分类器时,首先用第一级分类器对训练集进行测试,根据识别结果统计生成混淆矩阵 $\boldsymbol{F}$。然后根据生成的相似字符对列表利用 SPDA 抽取特征。最后,在抽取的特征上训练 MQDF 分类器。该方法的优点在于第一级分类器分类效果不好的字符类在第二级分类器进行重新学习、分类。这种重新学习的方法能够加强分类器对易识别错误字符类的关注,因此,第二级分类器更加具有针对性。

级联分类器在对未知字符图像进行识别时,首先通过第一级分类器进行识别。然后通过广义识别置信度(RC)[7]将识别不稳定的字符提取出来,送入第二级分类器进行识别,最后将识别结果融合起来。级联分类器识别流程图如下图 3 所示。

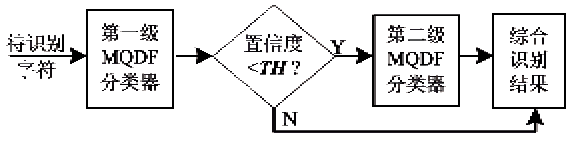

图 3　级联分类器识别流程图

广义识别置信度是识别结果的一种有效的度量,仅与最小的两个识别距离 d_1 和 d_2 有关,如式

$$RC = 1 - d_1/d_2 \tag{5}$$

所示。

在某种程度上广义识别置信度越高,识别结果的准确度越高。当识别结果的前两选为相似字符对时,有 $d_1 \approx d_2$,因此,$RC \approx 0$,即相似字符的广义识别置信度一般较低。文中将小于一定常数阈值 TH 的字符确定为相似字符。识别广义置信度高于 TH 的字符不需要进行再识别。这样不仅能够保证识别的准确度而且能够大大提高级联系统的整体运行速度。

鉴于上述的分析,分类器识别结果的融合规则如下:

$$R = \begin{cases} R_{11}, & RC_1 > TH \\ R_{21}, & C_1 \in TSL \ \text{或}\ C_2 \in TSL, \\ & RC_2 > RC_1 \\ R_{11}, & \text{其他} \end{cases} \tag{6}$$

其中,R_{ij} 为第 i 个分类器的第 j 个识别候选。第一级分类器的识别结果可以分为两个部分,一部分是广义识别置信度 RC_1 较高的识别结果,这部分结果直接作为最终的识别结果;另一部分为广义识别置信度较低的识别结果,如果其前两选识别结果在 TSL 中并满足 $RC_1 \leqslant TH$,则利用第二级分类器进行再识别。识别后若识别结果的广义置信度满足 $RC_2 > RC_1$,最终的识别结果取第二级分类器的识别结果,否则保持第一级分类器的识别结果。

虽然 SPDA 是针对相似类进行的,但在其投影向量的数量达到一定程度后,即提取适当特征维数的特征,可能对其他的字符也具有鉴别作用,因此,引入式

$$R = \begin{cases} R_{11}, & RC_1 > TH \\ R_{21}, & RC_2 > RC_1 \\ R_{11}, & \text{其他} \end{cases} \tag{7}$$

的融合规则与前述的规则进行对比。式(6),(7)中的融合规则分别记为 $R1$ 和 $R2$。

4　实验结果

为验证算法的有效性,提出的方法在已公开的 HCL2000(Handwritten Character Library 2000)测试集[8]上进行测试。常用训练集 700 套(xx001-xx700)和测试集 300 套(hh001-hh300),每套样本中包含 GB-2312 中 3755 个汉字字符图像。其部分字符样本如图 4 所示。

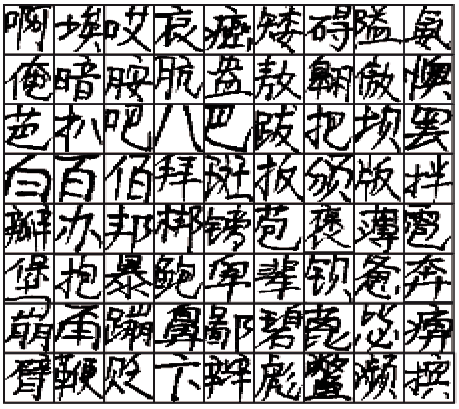

图 4 HCL2000 字符集示例

4.1 *TH* 的选择

TH 的大小决定了第一级分类器的识别结果是否需要进行重新识别。本文中 *TH* 的大小参考文献[6]确定。*TH* 的选择需要尽量使得大于该阈值的样本的识别率足够高,而小于该阈值的识别结果中识别正确的样本尽可能地少。这样能够有效地降低将原来识别正确的样本被重新识别错误的风险。训练集上,将所有广义识别置信度大于 *TH* 的测试样本挑选出来,设其总数为 N_l,其中识别正确的字符数为 N_{lr},在挑选出来的样本中识别准确率为 $(N_{lr}/N_l)\times 100\%$,其与广义识别广义置信度的关系如图 5 所示。在 $RC = 0.23$ 时,识别率达到了99.99%,表明如果一个字符的广义识别置信度大于0.23,则其识别结果非常可靠。因此,在级联分类器中选取 $TH = 0.23$。

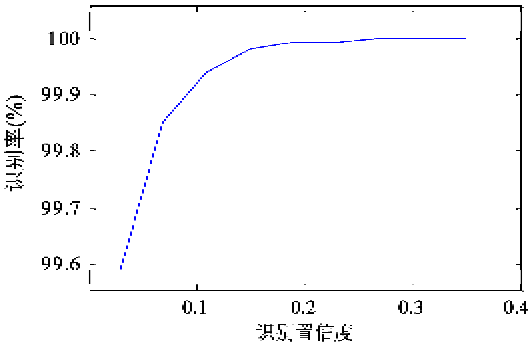

图 5 识别率与广义置信度的关系

4.2 单级分类器

本文采用 8 方向 392 维梯度特征[9,10]。第一级 MQDF 分类器通过传统的训练方法获得,原始 392 维梯度特征经过多类 LDA 降维到 128 维,测试集上的识别率为98.53%。第一级分类器训练过程中 3755 个字符类均参与了 LDA。第二级分类器采用了不同的特征维数。第二级分类器在测试集上的识别率如表 1 所示。

表 1 第二级分类器在 HCL2000 测试集上的识别性能

n	128	200	320	400	600	800
识别率(%)	97.36	98.07	98.38	98.44	98.49	98.09

第二级分类器特征维数为 n,参加相似类鉴别分析的字符类的数量最多为 $2n$ 类。以 $n = 600$ 为例,参加 LDA 变换的字符类最多为 1200 个类,接近于多类 LDA 中 3755 个类的 1/3,而其识别率仅降低 0.04%。这说明在鉴别分析中,一部分字符类起到绝大部分的作用。其原因主要在于对参加 LDA 变换的相似字符,其特征经过投影后得到的 n 维特征中,必然有一个最具鉴别性的特征;而对于没有参加 LDA 变换的字符而言,如果一个字符类本身容易区别于其它类,那么,适当维数的特征对其鉴别同样有效。另外,从上表中也能够发现,特征维数 n 对分类器本身的鉴别性能是有影响的。随着 n 的增加,识别率会出现一个峰值,在 $n = 600$ 达到最高。

4.3 级联分类器

在级联分类器中,将表 1 中多个分类器分别与第一级分类器进行级联。第一级分类器均为 128 维特征,32 维截断维数的 MQDF 分类器。

表 2 级联分类器识别结果

n	128	200	320	400	600	800
*R*1 识别率(%)	98.58	98.63	98.68	98.69	98.69	98.67
*R*2 识别率(%)	98.23	98.60	98.79	98.82	98.80	98.74

表 2 中,级联的识别结果基本都高于第一级分类器的识别率 98.53%。*R*2 的最高识别率为 98.82%,在可查的文献[6,10-13]中是最高的,这说明相似类鉴别分析方法对相似字符的鉴别是有效的。基于 *R*1 的融合方法中,第二级分类器对将要识别的字符类都在训练时进行了重新学习,因此,级联后的结果相对第一级分类器的识别结果都有一定的提高。但在基于 *R*1 的规则下,进入第二级分类器的字符的数量是受 *TSL* 中相似字符数量的限制,因此,级联分类器的性能提升是有上限的。基于 *R*2 的方法舍弃了进入第二级分类器的字符必须在 *TSL* 中的原则,即在第一级分类器中识别广义置信度低于一定阈值的字符,都需要经过第二级分类器再识别。级联结果中,第二级分类器在 128 维时,低于

$R1$ 的识别率,甚至低于第一级分类器的识别率;而在较高的特征维数上其识别率超过了基于 $R1$ 的方法和第一级分类器的识别结果。这是因为特征维数较低时,对于不在 TSL 中的字符,其投影后的特征鉴别性不足造成的误识字符数远大于第二级分类器对 TSL 中字符校正的字符数,因此,其总的识别率低于第一级分类器的识别率。随着特征维数的增加,特征的鉴别性能得到提高,其对总体字符类的鉴别性能也得到提升,因此,级联后的性能得到很大的提升。在特征维数足够高时 $n = 400$,级联识别率最高,相对错误率下降了 19.73%。

4.4 算法复杂度

相对于基本的 MQDF 分类器,级联的分类器增加了对相似字的区分。在提升分类性能的同时必然增加了算法复杂度。程序运行环境为 PC Intel Core (TM) 2, 3GHz。基本 MQDF 识别的速度为 212 字/s。级联的分类器识别速度如表 3 所示。

表 3 算法识别速度比较

n	128	200	320	400	600	800
级联分类器 $R1$ (字/s)	92	75	62	54	46	32
级联分类器 $R2$ (字/s)	88	68	60	52	45	36

从表 3 中可以看出,随着 n 的增加识别时间逐渐地增加。这一点是易于理解的,特征维数升高后,计算复杂度必然相应地增加。基于 $R2$ 准则的方法稍微比基于 $R1$ 的准则慢一些。级联后分类器的识别速度降低很多,但在主观上每秒 50 字左右的速度是可以接受的。

4.5 相似字的鉴别

字符 i 类和 j 类相互混淆导致错误的字符数为

$$\#err_{ij} = N_{ij} + N_{ji} \tag{8}$$

相似字符对是一个相对的概念,无法给出一个绝对的定义。本文中将训练集上相互识别错误最多的字符对进行统计。图 6 给出了在第一级分类器下训练集上错误最多的前 1000 对相似字符错误的统计情况。

字符的相似性不会随着字符集的不同而有较大的变化,即训练集和测试集上的相似字符对应该具有一致性。测试集上相似度最大的前 n 对相似字符对的识别率统计如表 4 所示。

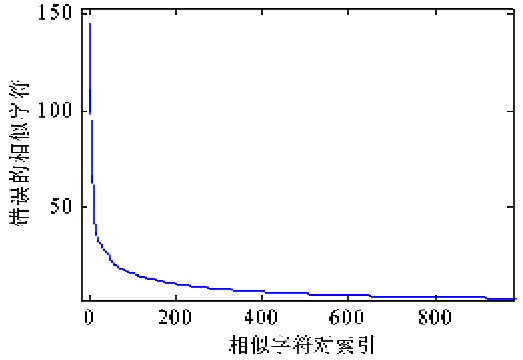

图 6 HCL2000 训练集上相似字符读对的统计结果

表 4 HCL2000 测试集相似字符对的识别结果

n	特征维数	LDA(%)	SPDA(%)	级联(%)
128	128	95.06	96.00	95.66
200	200	95.75	96.28	96.62
320	320	96.39	96.98	97.24
400	400	96.63	97.22	97.43
600	600	97.10	97.55	97.68

从表 4 可以看出,基于 SPDA 的识别率均高于相应的基于 LDA 的方法,最终级联的识别结果大都有了进一步的提高。其中,特征维数在 128 时,级联的性能相对基于 SPDA 的方法有所下降。这主要是由于特征维数较低,特征鉴别能力下降所导致的。

设计第二级分类器的主要目的是为了对第一级分类器识别不稳定和错误的样本进行重新鉴别,以弥补第一级分类器对相似字鉴别的不足。图 7 中给出了 HCL 测试集上部分被第一级分类器误识,经过第二级的分类器被纠正的测试集样本。

图 7 中,$h11$,$h12$ 为第一级分类器的前两选的识别距离,$h21$,$h22$ 为第二级分类器前两选的识别距离。第一级分类器的识别结果显示,其首选与第二选的识别距离非常接近;而在第二级分类器的识别结果中,首选与第二选的识别距离相对于第一级分类器前两选之间的距离拉大了。经过多类 LDA 及降维后,这些字符类在特征空间中被混淆了;而经过相似类鉴别分析后,拉大了相似字符对之间的距离,因此能够对第一级分类器的部分识别结果起到有效的校正作用。

图 7 HCL2000 相似字校正的部分结果

5 结 论

经过传统多类线性鉴别分析降维后,原始特征空间中相近的类在低维空间中易被混淆,不利于识别。针对该问题,本文提出了相似模式类鉴别分析的方法,实验结果表明部分字符类在鉴别分析中起主要作用。该方法与传统多类 LDA 特征降维方法相结合,在级联分类器中实现了对手写汉字识别性能的进一步提高。

参考文献

[1] Fisher R A. The statistical utilization of multiple measurements. *Ann Eugenics*, 1938, 8: 376-386

[2] Rao C R. The utilization of multiple measurements in problems of biological classification. *Journal of Royal Statistical Society B*, 1948, 10: 159-203

[3] Tao D C, Li X L, Wu X D, et al. Geometric mean for subspace selection. *IEEE Trans On Pattern Analysis and Machine Intelligence*, 2009, 31(2): 260-274

[4] Rueda L, Oommen B J, Henri'quez C. Multi-class pairwise linear dimensionality reduction using heteroscedastic schemes. *Pattern Recognition*, 2010, 43 (7): 2456 - 2465

[5] Kimura F, Takashina K, Tsuruoka S, et al. Modified quadratic discriminate functions and its application to Chinese character recognition. *IEEE Trans On Pattern Analysis and Machine Intelligence*, 1987, 9(1): 149-153

[6] Fu Q, Ding X Q, Li T Z, et al. An effective and practical classifier fusion strategy for improving handwritten character recognition. In: Proceedings of the International Conference on Document Analysis and Recognition, Curitiba, Brazil, 2007. 1038-1042

[7] Lin X F, Ding X Q, Chen M, et al. Adaptive confidence transform based classifier combination for Chinese character recognition. *Pattern Recognition Letters*, 1998, 19 (10): 975-988

[8] Zhang H G, Guo J, Chen G, et al. HCL2000: a large-scale handwritten Chinese character database for handwritten character recognition. In: Proceedings of the International Conference on Document Analysis and Recognition, Barcelona, Spain, 2009. 286-290

[9] Liu C L, Nakashima K, Sako H, et al. Handwritten digit recognition: investigation of normalization and feature extraction techniques. *Pattern Recognition*, 2004, 37(2): 265-279

[10] Liu H L, Ding X Q. Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes. In: Proceedings of the International Conference on Document Analysis and Recognition, Seoul, South Korea, 2005. 19-23

[11] Liu X B, Jia Y D, Tan M. Geometrical-statistical modeling of character structures for natural stroke extraction and matching. In: Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition, La Baule, France, 2006

[12] Long T, Jin L W. Building compact MQDF classifier for large character set recognition by subspace distribution sharing. *Pattern Recognition*, 2008, 41(9):2916-2925

[13] Zhang H G, Deng W, Guo J, et al. Handwritten Chinese character recognition using local discriminate projection with prior information. In: Proceedings of the International Conference on Pattern Recognition, California, USA, 2008. 1-4

Similar pattern discriminant analysis

Wang Yanwei, Liu Changsong, Ding Xiaoqing

(State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084)

Abstract

In multi-class recognition, neighbor classes in the original feature space are prone to be more confused after feature dimensionality reduction by linear discriminant analysis (LDA). It does not benefit the recognition. To solve this problem, this paper proposes a similar pattern discriminant analysis (SPDA) method, which constructs the feature transformation matrix based on discriminant vectors extracted from similar pattern pairs. The proposed SPDA method was applied together with LDA to the cascade modified quadratic discriminant function (MQDF) classifiers to improve the performance of recognizing handwritten Chinese characters. The results show that the recognition accuracy on handwritten character library 2000 (HCL2000) reaches up to 98.82%, which is higher than the corresponding results found in the literature. The experiment indicates that the proposed method is effective.

Key words: linear discriminant analysis (LDA), similar pattern discriminant, cascade classifier, offline Chinese handwriting recognition