一种基于预分类的高效最近邻分类器算法*)

王卫东[1,2] 郑宇杰[2] 杨静宇[2] 杨 健[2]

(江苏科技大学电子信息学院 镇江 212003)[1] (南京理工大学计算机系 南京 210094)[2]

摘 要 本文的最近邻分类器算法是采用多分类器组合的方式对测试样本进行预分类,并根据预分类结果重新生成新的训练和测试样本集。对新的测试样本采用最近邻分类器进行分类识别,并将识别结果与预分类结果结合在一起进行正确率测试。在ORL人脸库上的实验结果说明,该算法对小样本数据的识别具有明显优势。

关键词 最近邻分类器,预类别,多分类器组合,小样本问题,人脸识别

An Efficient Nearest Neighbor Classifier Algorithm Based on Pre-classify

WANG Wei-Dong[1,2] ZHENG Yu-Jie[2] YANG Jing-Yu[2] YANG Jian[2]

(School of Electronic and Information Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003)[1]

(Department of Computer Science, Nanjing University of Science & Technology, Nanjing 210094)[2]

Abstract A novel nearest neighbor classifier algorithm is proposed in this paper. The algorithm adopts classifiers combination to pre-classify test samples, and then reproduces new training and test sample sets. The new test samples are classified by nearest neighbor classifier. The results combined with pre-classifying ones are tested for recognition rate. In ORL face database, the experimental results prove that the algorithm outperforms traditional methods in small sample size problem.

Keywords Nearest neighbor classifier, Pre-classify, Classifiers combination, Small sample size problem, Face recognition

1 引言

人脸识别属于高维小样本问题。通常将人脸图像向量变换到低维的特征子空间中,再进行特征提取及分类识别。对于高维小样本数据的特征提取,目前,已提出了许多经典算法,主要有主成分分析[1]、fisherface[2]方法、独立成分分析[3]等。但是对于适合小样本数据的分类器方法,所做的研究相对较少。通常采用传统的分类器方法进行小样本数据的识别,如采用贝叶斯分类器[4]、最小距离分类器、最近邻(nearest neighbor,NN)分类器[5]等。

在传统的分类器方法中,最近邻分类器是一种被广泛使用的方法。它是将模式判别为与离它最近的样本同类,其性能与各类模式在子空间中的分布特征有关。因此,若能增强各类模式在子空间中的分布特征,将会提高最近邻分类器的性能。基于上述思想,Li[6,7]等提出了最近特征线(nearest feature line,NFL)方法,而Chien[8]等则进一步提出了最近特征面(nearest feature plane,NFP)方法。NFL是采用每个类内的任意两个样本组成特征线,然后将模式判别为与离它最近的特征线同类。NFP则是采用三个样本组成特征面,将模式判别为与离它最近的特征面同类。这两种方法都提高了最近邻分类器的分类能力,但是,显然它们大大增加了最近邻分类器的时间复杂度。为了克服NFL及NFP的时间复杂度问题,Wenming Zheng[9]等提出了最近邻线(nearest neighbor line,NNL)和最近邻面(nearest neighbor plane,NNP)方法。NNL或NNP取各类与未知样本最近的两个或三个样本组成特征线或特征面。该方法仍属于NFL及NFP方法,但其算法的效率得到了提高。

本文采用多分类器组合的方法对未知样本进行预分类,将各基分类器预分类结果相同的样本加入到训练样本集中,再采用最近邻分类器对各基分类器分类结果不同的样本进行分类识别。由于将大量的未知样本加入到训练样本集中,使得新训练样本集中的样本数量明显增加,而未知样本的数量明显减少。这样不但增强了各类别的模式分布特征,提高了最近邻分类器的性能。同时,由于未知样本数量的减少大大降低了最近邻分类器的时间复杂度。因此,本文算法不但克服了小样本数据对各类模式分布特征的影响,而且算法的效率很高。

2 基于预分类的高效最近邻分类器

2.1 最近邻分类器

设模式类别有C个$\omega_1,\omega_2,\cdots,\omega_c$,每类有标明类别的训练样本$N_i$个,$i=1,2,\cdots,C$。最近邻分类器是采用各类中全部样本都作为代表点,将未知样本X判别为与离它最近的样本同类。因此,最近邻分类器可在一定程度上克服各类样本均值向量的偏差所造成的影响。设ω_i类的判别函数为:

$$g_i(X)=\min_k \| X-X_i^k \| ,k=1,2,\cdots,N_i$$

其中X_i^k的i表示ω_i类,k表示ω_i类中的第k个样本。则

如果$g_j(X)=\min_i g_i(X),i=1,2,\cdots,C$

则判$X\in\omega_j$

2.2 本文算法的基本思想

本文采用多分类器组合的方式,对未知样本进行预分类。若各基分类器对某一未知样本的预分类结果相同,则将该未知样本加入到训练样本集中。反之,将预分类结果不同的未知样本加入到新的测试样本集。最后,利用新的训练和测试

*)国家自然科学基金资助(编号:60503026)。王卫东 讲师,博士生;郑宇杰 博士生;杨静宇 教授,博士生导师;杨 健 教授,博士后。

样本集，并采用最近邻分类器进行分类识别。将识别结果与预分类结果融合在一起，进行正确率测试并输出。算法思想如图 1 所示。

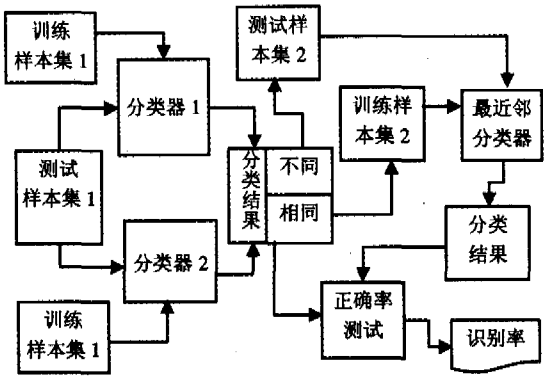

图 1　算法的原理框图

图中训练样本集 2 是由训练样本集 1 的全部样本加上所有预分类结果相同的未知样本组成的。而测试样本集 2 是由所有预分类结果不同的未知样本生成的。在采用最近邻分类器对测试样本集 2 进行分类后，将分类结果与预分类结果一起进行正确率测试。最后以识别率的形式输出。

该算法的显著特点是在动态变化的训练及测试样本集中进行分类识别。采用预分类的方式大大增加了训练样本的数量，使最近邻分类器的模式分类正确率得到了极大的提高。而未知样本数量的减少又提高了算法效率。

2.3　算法步骤

设模式类别有 C 个 $\omega_1,\omega_2,\cdots,\omega_c$，每类有标明类别的训练样本 N_i 个，$i=1,2,\cdots,C$。则高效最近邻分类器的算法步骤如下：

(1)模式特征提取：本文主要采用 PCA 及 Fisherfaces 两种方法进行特征提取。得到特征子空间中的训练样本集 1 及测试样本集 1。

$Y_P^1=\{(y_P)_i^k;i=1,2,\cdots,c,k=1,2,\cdots,N_i\}$

$X_P=\{(x_P)_i;i=1,2,\cdots,S\}$

$Y_F^1=\{(y_F)_i^k;i=1,2,\cdots,c,k=1,2,\cdots,N_i\}$

$X_F=\{(x_F)_i;i=1,2,\cdots S\}$

其中 Y_P^1，X_P 表示采用 PCA 进行特征提取后的训练及测试样本集。而 Y_F^1，X_F 表示采用 Fisherfaces 方法进行特征提取。S 为测试样本数。

(2)预分类：采用多分类器对不同特征子空间的未知样本进行预分类。

$X_P^1=\{(x_P)_i^k;i=1,2,\cdots,S,k=1,2,\cdots,c\}$

$X_F^1=\{(x_F)_i^j;i=1,2,\cdots,S,j=1,2,\cdots,c\}$

其中 $(x_P)_i^k$，$(x_F)_i^j$ 表示将第 i 个未知样本判别为第 k 及第 j 类。

(3)融合分类结果：对各分类器的分类结果进行融合，得到分类结果相同和不同的测试样本集 X_S 及 X_d。对于未知样本 $(x_P)_i^k$，$(x_F)_i^j$

若 $k=j$，将 $(x_P)_i^k$ 加入到 X_S 中

否则，将 $(x_P)_i^k$ 加入到 X_d 中。

则 $X_S=\{(x_P)_i^k;(x_P)_i^k\in X_P^1,(x_F)_i^j\in X_F^1$ 且 $k=j\}$，

$X_d=\{(x_P)_i;(x_P)_i^k\in X_P^1,(x_F)_i^j\in X_F^1$ 且 $k\neq j\}$

(4)生成训练样本集 2 及测试样本集 2：将预分类结果相同的测试样本和原训练样本一起组成新的训练样本集 Y_P^2。再用预分类结果不同的未知样本生成新的测试样本集 X_P^2。

$Y_P^2=\{(y_P)_i^k;y_p\in Y_P^1$ 或 $y_p\in X_S\}$

$X_P^2=\{(x_P)_i;(x_P)_i\in X_d,i=1,2,\cdots,t\}$

其中 t 为预分类结果不同的未知样本数。

(5)采用最近邻分类器进行分类：利用训练样本集 Y_P^2，采用最近邻分类器对测试样本集 X_P^2 进行分类识别。

(6)正确率测试：将测试样本集 X_P^2 的分类结果 X_S 结合预分类结果一起进行正确率测试，并将测试结果作为最终结果输出。

2.4　多分类器组合

由于某个分类器的错分样本集与另一个的错分样本集可能不同，即对某个样本，A 分类器将其错分为 i 类，而 B 分类器可能将其错分为就 j 类，因此，不同的分类器可能给出互补的信息。这样将这些分类器组合起来就是有益的。当选择了适当的多个分类器后，就可从测试样本中分离出可能分类正确的测试样本及可能被错分的测试样本。

对多分类器的选取原则是要保证各基分类器的准确性及多样性。准确性是要求各个分类器的单独使用时识别率较高，而多样性是当多个分类器在预测样本 X 的类别时产生不同的错误。因此，本文在模式的特征提取及分类器的选择两个方面进行设计，来满足准确性及多样性的要求，即在不同的特征子空间中分别采用不同的分类器进行模式分类。

特征提取分别采用主成分分析及 Fisherface 两种方法。主成分分析是在在最小协方差意义对原模式特征的最佳逼近，即在特征子空间中对原模式的最优重建。而 Fisher 线性鉴别是将模式向最具有鉴别信息的方向上进行投影。显然，在进行线性鉴别分析时，Fisher 方向包含了更多的鉴别信息。由于最小距离分类器是一种分段线性分类器，当采用 Fisherface 方法进行特征提取时，使用该分类器将得到较好的分类结果。最近邻分类器是将未知样本判别为与它最近的样本同类，因此，其判别结果与样本在子空间中的分布结构密切相关。故采用主成分分析进行特征提取时，最近邻分类器将会取得较好的分类效果。

2.5　算法的错误率分析

本文提出的高效最近邻分类器的算法实质是通过预类别，将一部分测试样本加入到训练样本集中，即增加了训练样本集中的样本数量。下面就两类问题分析训练样本数量的变化对算法错误率的影响。

对于一般的最近邻分类器，设未知样本 x 的最近邻为 x'，则 x 被错分的概率是：

$$\begin{aligned}&P_N(e|x)\\&=P(\omega_1|x)P(\omega_2|x')+P(\omega_2|x)P(\omega_1|x')\\&=P(\omega_1|x)[1-P(\omega_1|x')]+P(\omega_1|x')P(\omega_2|x)\\&=2P(\omega_1|x)P(\omega_2|x)+[P(\omega_1|x)-P(\omega_2|x)]\\&[P(\omega_1|x)-P(\omega_1|x')]\end{aligned}\tag{1}$$

因 x' 是 x 的最近邻，则概率密度 $P(x'|x)$ 在 x 附近应是尖峰突起，而在其它地方则较小，即在已知 x 的条件下，x 的最近邻 x' 在 x 附近的概率密度最大。假定对于给定的 x，$P(x)$ 是连续且非零的，则任何样本落在以 x 为中心的一个超球 S 里的概率为一正数，记为 P_s，且

$P_s=\int_{x'\in s}p(x')\mathrm{d}x'$

因此，一个样本落在 S 外的概率为 $(1-P_S)$，N 个独立样本落在 S 外的概率为

$$P(x_1,x_2,\cdots,x_N)=(1-P_S)^N\tag{2}$$

当 N 变大时，若 P_S 不变，则(2)式的概率将变小。即当 P_S 不变时，N 变大将导致超球 S 变小。而超球 S 变小，将使 x' 更加趋近于 x。因此，当 $x\in\omega_i$ 时，概率 $P(\omega_i|x')$ 将变大。由式(1)知，随着样本数 N 变大，概率 $P(\omega_i|x')$ 的变大，会使错误率 $P_N(e|x)$ 减小。也就是说，增加训练样本数将减小最近邻分类器的模式分类错误率。

3 实验及分析

本文采用 ORL 人脸库进行对比实验，该库由 40 人，每人 10 幅 92×112 图像组成，其中有些图像是拍摄于不同时期的；人的脸部表情和脸部细节有着不同程度的变化。先对样本进行预处理，预处理采用两次小波变换。在去掉高频成份后，将图像变换为 23×28 像素。则样本特征向量降为 644 维。特征提取方法采用基于类间离散矩阵 S_b 的主成分分析及 Fisherfaces 方法。两种方法均将模式投影到 39 维的特征子空间中。

3.1 各基分类器的实验及分析

实验以每人的前 5 幅图像作为训练样本，后 5 幅作为测试样本。则训练和测试样本均为 200 幅。对于距离的度量采用欧氏距离。表 1 给出了四种单分类器系统的模式识别率。

表 1 不同的特征提取方法与分类器的组合方式

特征提取	PCA		Fisherfaces	
分类器	最小距离	最近邻	最小距离	最近邻
识别正确率(%)	89.5	94.5	93	93.5

正如本文前面的分析，当特征提取方法采用主成分分析，而分类器采用最近邻分类器时，模式识别率较高。而用 Fisherfaces 方法进行特征提取时，最小距离分类器的识别率得到了很大的提高。

3.2 对比实验及分析

实验 1：以每人的前 5 幅图像作为训练样本，后 5 幅作为测试样本。本文算法采用 PCA+最近邻分类器与 Fisherfaces+最小距离分类器组合进行预分类。各基分类器预分类结果相同的样本集为 X_S，不同的样本集为 X_d。表 2 为本文算法的识别结果。

表 2 本文算法的识别结果

X_s 样本数	X_s 错分数	X_d 样本数	X_d 错分数	算法的错分数
184	2	16	4	6

从表 3 中可以看出，在 200 个测试样本中，预分类结果相同的样本有 184 个，基分类器将其中 2 个样本错分为同一类。预分类结果不同的样本有 16 个，再次采用最近邻分类器对这些样本进行分类识别，错分样本数为 4 个，结合 X_S 错分样本数，则算法最后的错分数为 6 个。由于经过预分类后，只需再对 16 个样本进行分类，因此，极大地提高了最近邻分类器的效率。

实验 2：分别取每人的前 2、3、4、5 幅图像作为训练样本，对应的取每人后 8、7、6、5 幅作为测试样本。则测试样本集分别有 320、280、240、200 个测试样本。对本文算法同两个单分类器系统进行对比实验。

表 3 取不同训练和测试样本集的识别率(%)

训练样本数	PCA+最近邻	Fisherfaces+最小距离	本文算法
2	84	81.88	89
3	87.5	87.86	91.4
4	90.42	91.25	95
5	94.5	93	97

从表 3 中可以看出，当训练样本的数量减少，而测试样本的数量增加时，两种单分类器系统的识别率急剧下降。但本文算法识别率的波动却较小，即使每人只取前 2 幅图像作为训练样本，本文算法的识别率也可达到 89%。这是因为通过预分类，本文算法将大量的测试样本加入到训练样本集中。这样就增加了训练样本的数量。因此，其模式识别率的变化较小。而当取每人的前 5 幅图像作为训练样本时，识别率可高达 97%。通过对不同的训练及测试样本集的对比实验，说明本文算法非常适用于解决小样本问题。

实验 3：以每人的前 5 幅图像作为训练样本，后 5 幅作为测试样本。将本文算法与 NN、NFL、NFP、NNL 及 NNP 进行对比实验。其中 NN、NFL、NFP、NNL 及 NNP 的实验结果取自文[9]。

表 4 本文算法与其它近邻算法的识别率(%)

NN	NFL	NFP	NNL	NNP	本文
94.65	95.63	95.80	95.18	95.75	97

与其它近邻算法的对比实验可以看出，本文提出的高效最近邻分类器算法的识别能力明显高于其它算法。说明增加训练样本的数量可以增强模式在子空间中的分布特征，从而提高最近邻分类器的性能。因此，本文算法非常适合于小样本数据的识别问题。

结束语 本文采用多分类器组合的方式，对测试样本进行预分类，并利用预分类的结果重新分配训练和测试样本，由此增加了训练样本的数量并减少了测试样本的数量，克服了小样本数据的局限性。本算法的特点是在动态变化的训练及测试样本集进行分类识别。这就打破了传统的在静态的训练和测试样本集上进行特征提取和分类识别的方法。

参 考 文 献

1 Kirby M, Sirovich L. Application of the KL procedure for the characterization of human faces. IEEE Trans. Pattern Anal. Machine Intelligence, 1990,12(1):103～108

2 Belhumeur P N, et al. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Anal. Machine Intell., 1997,19(7):711～720

3 Comon P. Independent component analysis: A new concept. Signal Processing, 1994,36 (3): 287～314

4 Anderson T W. An Introduction to Multivariate Statistical Analysis,2nd ed. New York: Wiley,1984

5 Cover T W, Hart P E. Nearest Neighbor Pattern Classification. IEEE Trans, On Information Theory, 1967,13(6):21～27

6 Li S Z. Face Recognition Based on Nearest Linear Combinations. In:Proceedings of CVPR, 1998. 839～844

7 Li S Z, Lu J W. Face Recognition Using the Nearest Feature Line Method. IEEE Trans. on Neural. Networks, 1999, 10(2):439～443

8 Chien J T, Wu C C. Discriminant Waveletfaces and Nearest Feature Classifiers for Face Recognition. IEEE Trans. on PAMI, 2002,24(12):1644～1649

9 Zheng Wenming, Zou Cairong, Zhao Li. Face Recognition Using Two Novel Nearest Neighbor Classifiers. In: Proceedings of ICASSP, 2004. 725～728