基于正交投影的分类器算法

王卫东[1] 苗 帅[1] 杨静宇[2]
(江苏科技大学计算机科学与工程学院 镇江 212003)[1] (南京理工大学计算机系 南京 210094)[2]

摘 要 提出了一种新颖的基于正交投影的分类器算法。该算法将测试样本正交投影到由各类训练样本生成的子空间中,并计算测试样本到各子空间的距离,以此作为分类的依据。该算法不需要计算样本协方差矩阵的逆阵,因此特别适合于小样本问题。在ORL人脸库上的实验结果表明,该算法的模式识别率高于传统分类器方法。

关键词 正交投影,分类器,小样本问题,人脸识别
中图法分类号 TP391.4 **文献标识码** A

Classifier Algorithm Based on Orthogonal Projection

WANG Wei-dong[1] MIAO Shuai[1] YANG Jing-yu[2]
(School of Computer Science and Technology, Jiangsu University of Science and Technology, Zhenjiang 212003, China)[1]
(Department of Computer Science, Nanjing University of Science & Technology, Nanjing 210094, China)[2]

Abstract In this paper, a novel classifier algorithm based on orthogonal projection was presented. This classifier projected testing samples to orthogonal subspaces that were spanned by training samples of per class, and calculated the distance of the testing samples and subspaces, then classified the testing samples according to the distance. The character of this classifier is that it need not calculate the covariance matrices' inverses, so it is very suit for small sample size problem. The experimental results in ORL database prove this classifier algorithm outperforms the traditional classifier algorithm in recognition rate.

Keywords Orthogonal projection, Classifier, Small sample size problem, Face recognition

1 引言

对于高维小样本数据的识别问题,在特征提取方面所做的研究很多。而对适用于小样本数据的分类器算法所做的研究相对较少,通常采用传统分类器算法,如最小距离、最近邻分类器等,这些分类器在解决小样本问题时都有其局限性。

最小距离分类器是取各类均值向量作为该类的代表,将未知样本划分到离它最近的代表点所属的类别。该算法的实质是假设类条件概率密度函数为正态分布,且各类具有相同的协方差结构。小样本数据可能会造成各类均值向量偏差或不稳定,即当训练样本发生变化时,可能会引起类均值向量的急剧变化。因此,类均值向量并不一定能很好地代表各个类,其结果是分类的正确率较低。最近邻分类器是将未知样本判别为与它最近的训练样本同类。该算法是采用各类中的全部样本作为代表点。因此,最近邻分类器可在一定程度上克服各类均值向量的偏差所造成的影响。最近邻分类器的不足是要计算未知样本与所有训练样本的距离并找到最小值,因此,算法的效率较低。另外,将未知样本判别为与离它最近的样本同类,使得最近邻分类器未考虑其他训练样本对模式分布特征的影响,这也影响了该分类器的模式识别率。

为了克服最近邻分类器,仅考虑一个训练样本的局限性,Li[1,2]等提出了最近特征线(NFL, Nearest Feature Line)算法,而文献[3,4]进一步提出了最近特征面(NFP, Nearest Feature Plane)算法。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3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。这两种算法的分类能力都高于传统分类器算法。但是,显然它们大大增加了分类器的时间复杂度。假设某类的训练样本数是 n_i,在NFL算法中,每类的训练样本可以生成的特征线是组合数 $C_{n_i}^2$,NFP算法每类的训练样本可以生成的特征面是组合数 $C_{n_i}^3$,当 n_i 较大时,两种算法的计算代价很大。

对于小样本问题,各类协方差矩阵通常是奇异的,因此不能直接使用贝叶斯分类器。J. H. Friedman提出了正则化判别分析[5](RDA, Regularised Discriminant Analysis),RDA通过对各类样本协方差矩阵与总体协方差矩阵之间的插值运算,解决了各类协方差矩阵的奇异性问题。该算法包含两个参数:复杂参数 λ、收缩参数 γ。由于参数 λ,γ 的取值在[0,1]闭区间内,因此,在运用RDA算法时,对参数 λ 和 γ 的优化选取是比较困难的。文献[6]在模式特征子空间中选取一组标准正交向量来生成虚拟训练样本,从而实现了对总体协方差矩阵的优化,优化后协方差矩阵的特征值均显著变大,从而提高了协方差矩阵的逆阵稳定性。但该算法仍然属于RDA算法,因此,也需要解决RDA算法的参数优化问题。文献[7]

到稿日期:2010-06-21 返修日期:2010-10-09 本文受国家自然科学基金(60632050)资助。
王卫东(1968—),男,博士,副教授,主要研究方向为模式识别、图像处理、人脸识别,E-mail:wang_wdong@yahoo.com.cn;苗 帅(1986—),男,硕士生,主要研究方向为模式识别、信息系统;杨静宇(1941—),男,教授,博士生导师,主要研究方向为计算机视觉、信息融合、模式识别。

是通过大量生成虚拟训练样本来克服各类协方差矩阵的奇异性问题，该算法解决了文献[5,6]需要进行参数优化的问题，但由于需要生成的虚拟样本数量很多，增加了算法的时间复杂度。

本文提出的基于正交投影的分类器算法是将未知样本正交投影到由各类训练样本生成的子空间中，并计算未知样本到该子空间的距离，再将未知样本判别为与子空间距离最小的那个类。本文算法的特点是由各类的所有训练样本生成子空间，克服了最小距离和最近邻分类器由单特征点代表各类的局限。同时不论训练样本的数量多少，都可以方便地生成子空间。本文算法是对 NFL 及 NFP 算法的推广，当只有两个训练样本时，本文算法退化为 NFL 算法，而当只有 3 个训练样本时本文算法退化为 NFP 算法。

2 向量的正交投影

2.1 正交集

设 z 是 R^n 中的向量，W 是 R^n 中的子空间，则向量与子空间正交及正交补空间的定义如下。

定义 1 如果向量 z 与 R^n 的子空间 W 中的任意向量都正交，则称 z 正交于 W。

定义 2 与子空间 W 正交的向量 z 的全体组成的集合称为 W 的补空间，记作 $W^{\perp}$。

向量 x 属于 $W^{\perp}$ 的充分必要条件是向量 x 与生成空间 W 的任一向量都正交，而 $W^{\perp}$ 是 R^n 的一个子空间。下面给出正交向量集的定义。

定义 3 如果 R^n 中向量集合 $\{u_1,\cdots,u_p\}$ 的任意两个不同向量都正交，即当 $i\neq j$ 时，$u_i\cdot u_j=0$，则称该向量集合为正交向量集。

定理 1[8] 如果 $S=\{u_1,\cdots,u_p\}$ 是由 R^n 空间中非零向量构成的正交集，那么 S 是线性无关集，因此构成所生成的子空间 S 的一组基。

证明见文献[8]。

定义 4 R^n 中子空间 W 的一个正交基是 W 的一个基，且是正交集。

定理 2[8] 假设 $\{u_1,\cdots,u_p\}$ 是 R^n 中子空间 W 的正交基，对 W 中的每个向量 y，线性组合 $y=c_1u_1+c_2u_2+\cdots+c_pu_p$ 中的权值可由下式计算。

$$c_j=\frac{y\cdot u_j}{u_j\cdot u_j}\quad(j=1,\cdots,p)$$

证明见文献[8]。

从定理 2 可以看出，正交基比其他基优越，线性组合中的权值较易计算。如果基不是正交的，则必须解线性方程组才能得到 y 的权值表示。

2.2 正交投影

对 R^n 中给出的非零向量 u，考虑 R^n 中一个向量 y 分解为两个向量之和的问题，一个向量是向量 u 的数量乘积，另一个向量与 u 垂直，如下式所示。

$$y=\tilde{y}+z \tag{1}$$

式中，$\tilde{y}=\alpha u$，α 是一个数，z 是一个垂直于 u 的向量。记 $z=y-\alpha u$，满足式(1)的要求，那么 $y-\tilde{y}$ 和 u 正交的充分必要条件是

$$0=(y-\alpha u)\cdot u=y\cdot u-(\alpha u)\cdot u=y\cdot u-\alpha(u\cdot u)$$

因此，满足式(1)，且 z 与 u 正交的充分必要条件是

$$\alpha=\frac{y\cdot u}{u\cdot u}\text{ 且 }\tilde{y}=\frac{y\cdot u}{u\cdot u}\cdot u$$

向量 $\tilde{y}$ 称为 y 在 u 上的正交投影，向量 z 称为 y 垂直 u 的分量。从上述定义中可以看出，正交投影可由 u 所生成的子空间 L 所确定，用 $proj_L y$ 来表示 $\tilde{y}$，并称之为 y 在 u 上的正交投影，即

$$\tilde{y}=proj_L y=\frac{y\cdot u}{u\cdot u}\cdot u \tag{2}$$

式(2)中的正交投影与定理 2 中每一项形式一致，这样定理 2 将向量 y 分解为子空间上的正交投影之和。

当 y 表示成 R^n 空间中基 $u_1,\cdots,u_n$ 的线性组合时，y 中和的各项可以分为两部分，使得 y 写成 $y=z_1+z_2$。此处，z_1 是其中一些 u_i 的线性组合，z_2 是其余 u_i 的线性组合，当 $u_1,\cdots,u_n$ 是正交基时，若由 u_i 组成子空间 W，$z_1\in W$，则 $z_2\in W^{\perp}$。这也就是下面的正交分解定理。

定理 3[8] 若 W 是 R^n 的一个子空间，那么 R^n 中每一个向量 y 可以唯一表示为

$$y=\tilde{y}+z \tag{3}$$

此处 $\tilde{y}$ 属于 W 且 z 属于 $W^{\perp}$，若 $\{u_1,\cdots,u_p\}$ 是 W 的任意正交基，那么

$$\tilde{y}=\frac{y\cdot u_1}{u_1\cdot u_1}\cdot u_1+\cdots+\frac{y\cdot u_p}{u_p\cdot u_p}\cdot u_p \tag{4}$$

且

$$z=y-\tilde{y} \tag{5}$$

证明见文献[8]。

下面的定理证明了，向量 $\tilde{y}$ 是子空间 W 中元素对 y 的最佳逼近。

定理 4[8] 假设 W 是 R^n 空间中的一个子空间，y 是 R^n 中的任意向量，$\tilde{y}$ 是 y 在 W 上的正交投影，那么 $\tilde{y}$ 是 W 中最接近 y 的点，也就是

$$\|y-\tilde{y}\|\leqslant\|y-v\| \tag{6}$$

对所有属于 W 又异于 $\tilde{y}$ 的 v 都成立。

证明见文献[8]。

不等式(6)同时给出了一个结论，$\tilde{y}$ 的计算不依赖于特定正交基。如果 W 的一个不同的正交基被用于构造 y 的正交投影，那么这个投影就是向量 y 在 W 中的最近点。

采用式(4)计算正交投影时比较烦琐，当 W 的基是单位正交集时，正交投影的计算可以得到简化。

定理 5[8] 如果 $\{u_1,\cdots,u_p\}$ 是 R^n 中子空间 W 的单位正交基，那么

$$proj_w y=(y\cdot u_1)u_1+\cdots+(y\cdot u_p)u_p$$

如果 $U=[u_1,\cdots,u_p]$，则

对所有 $y\in R^n$，$proj_w y=UU^T y$。

证明见文献[8]。

当得到 W 的一组基后，可以对这组基进行 Gram-Schmidt 标准正交化，从而得到一组单位正交基。利用定理 5 就可以得到向量 y 在子空间 W 中的正交投影。利用式(5)，可以将任意向量 y 分解为在子空间 W 上的正交投影 $\tilde{y}$ 及垂直于子空间 W 的向量 z。将向量 z 的某个范数的大小 $\|z\|$ 作为向量 y 到子空间 W 的距离测度。

3 基于正交投影的分类器算法

本文采用训练样本生成各类样本子空间 W_i，然后将未知样本 x 分解到 W_i 上，计算其正交投影 $proj_{w_i}x$ 及垂直于子空

间 W_i 的向量 dx_i。再计算 dx_i 的范数大小 $\| dx_i \|$ 并将其作为未知样本 x 到子空间 W_i 的距离。设模式类别有 C 个，第 i 类的训练样本数为 n_i，则按下列公式将未知样本 x 判别为第 j 类。

$$g_j(x)=\min_{j=1,2,\cdots,c}\| dx_j \| \tag{7}$$

基于正交投影的分类器算法的步骤如下：

(1) 特征提取：将所有样本投影到特征子空间中。

(2) 生成各类样本子空间 W_i：利用各类的训练样本生成子空间 W_i，首先生成矩阵 $U_i=[x_1 \quad x_2 \quad \cdots \quad x_{n_i}]$，求出该矩阵的列空间 $ColU_i=[u_1 \quad u_2 \quad \cdots \quad u_p]$，则第 i 类训练样本生成的子空间 W_i 就是列空间 $ColU_i$。

(3)Gram-Schmidt 标准正交化：对矩阵 $ColU_i=[u_1 \quad u_2 \quad \cdots \quad u_p]$ 中的列向量进行 Gram-Schmidt 标准正交化，得到子空间 W_i 的一组单位标准正交基，采用矩阵 $U_i'=[u_1' \quad u_2' \quad \cdots \quad u_p']$ 表示。

(4)计算正交投影：利用下列公式计算未知样本 x 到各类子空间的正交投影。

$$proj_w x=U_i'(U_i')^T x$$

(5)计算未知样本 x 到各类子空间 W_i 的距离：采用下式计算未知样本 x 与各类子空间 W_i 的垂直分量 dx_i。

$$dx_i=x-proj_{w_i}x$$

(6)分类识别：计算 dx_i 的范数 $\| dx_i \|$，并将其作为未知样本 x 到子空间 W_i 的距离，利用式(7)进行分类识别。

该分类器方法考虑了全部训练样本对未知样本的影响，克服了最小距离和最近邻分类器由单特征点代表各类的局限性。

4 实验及分析

本文在 ORL 人脸库上进行对比实验，ORL 人脸库包含 40 个人不同表情、不同视角下的 400 幅图像，每人 10 幅。先进行两次小波变换，将图像变换为 23×28 像素。

实验 1　采用 PCA 进行特征提取

随机选择每人的 5 幅图像作为训练样本，其余 5 幅作为测试样本，则训练及测试样本集均有 200 个样本。将上述过程重复 10 次，共得到 10 组训练及测试样本。采用训练样本得到总体协方差矩阵 s_t，并计算主成分分析(PCA)的投影矩阵。将本文算法与传统的分类器算法进行比较，其分类结果如表 1 所列。

表 1　采用 PCA 进行特征提取的对比实验

算法名称	识别率
最小距离	87.05±2.58
最近邻	93.7±1.86
RDA(λ=γ=0.7)	91.8±1.82
RDA(λ=γ=0.3)	95.10±2.11
本文算法	95.10±1.82

从表 1 可以看出，本文算法的模式识别率明显优于传统的最小距离及最近邻分类器。而 RDA 算法的模式识别率随参数 λ，γ 取不同的值波动很大。当参数 λ，γ 取 0.3 时识别率较高，而取 0.7 时识别率较低。本文算法的模式识别率与 RDA 在参数最优时的识别率相同，但本文算法没有需要进行优化选取的参数，且其识别率方差最小，说明本文算法比其它算法鲁棒。

实验 2　采用 Fisherface 进行特征提取

实验 2 的实验方法与实验 1 的基本相同，只是采用 Fisherface 进行特征提取，实验结果如表 2 所列。

表 2　采用 Fisherface 进行特征提取的对比实验

算法名称	识别率
最小距离	94.25±2.40
最近邻	93.15±1.78
RDA(λ=γ=0.7)	94.35±2.30
RDA(λ=γ=0.3)	95.10±1.68
本文算法	96.20±1.36

在采用 Fisherface 进行特征提取时，实验 2 与实验 1 的结果类似。本文算法的模式识别率同样优于传统的最小距离及最近邻分类器，同时也优于 RDA 算法在参数 λ，γ 最优时的模式识别率。同样，本文算法的识别率方差也是最小的。

通过实验 1、2，可以看出无论是采用 PCA 还是 Fisherface，本文算法的模式识别率均优于传统分类器算法。同时，与传统分类器相比，本文算法的鲁棒性更好。因此，本文算法更适于小样本问题。

实验 3　与相关分类器算法的对比实验

与 NFL，NFP 及文献[7]的二次判别分析方法进行对比实验，采用 PCA 进行特征提取，实验方法与实验 1 相同，结果如表 3 所列。

表 3　本文算法与相关算法的对比实验

算法名称	识别率	时间复杂度(分)
NFL	95.20±2.0	1.33
NFP	95.20±1.82	1.30
二次判别分析	95.10±1.86	1.21
本文算法	95.10±1.82	1.08

从表 3 可以看出，4 种分类器算法的模式识别率基本相同。但正如本文前面的分析，NFL 算法每个类的训练样本可以生成的特征线是组合数 $C_{n_i}^2$，而 NFP 算法每个类的训练样本可以生成的特征面是组合数 $C_{n_i}^3$。由于每人的训练样本数是 5 个，则 NFL 及 NFP 算法均可生成 10 个特征线及特征面，因此，这两种算法的时间复杂度基本相同。而 NFL 算法采用点到直线的距离公式进行计算，NFP 采用本文的正交投影公式进行计算，实验数据共有 10 组，2000 个测试样本，因此，NFL 比 NFP 用时稍多。这两种分类器大大增加了算法的时间复杂度。

而文献[7]的二次判别分析方法是通过大量生成虚拟训练样本来克服各类协方差矩阵的奇异性问题，因此，其算法的时间复杂度也明显高于本文算法。由于分类器的时间复杂度是分类器的重要性能指标，故可以认为本文算法更适于小样本的识别问题。

实验 4　采用 Fisherface 的相关实验

实验 4 的方法与实验 3 完全相同，只是采用 Fisherface 进行特征提取，结果如表 4 所列。

表 4　采用 Fisherface 的相关实验

算法名称	识别率	时间复杂度(分)
NFL	96.35±1.0	1.33
NFP	96.20±1.36	1.30
二次判别分析	95.10±1.82	1.21
本文算法	96.20±1.36	1.08

实验 4 与实验 3 的结果基本相同，只是二次判别分析方法的模式识别率稍低，这种分类器比较适合于在 PCA 特征子空间中使用。而其余 3 种分类器算法的识别率相差不多，均高于在 PCA 子空间中的模式识别率，这说明在 Fisherface 子

空间中，使用这3种分类器的分类效果较好。而算法的时间复杂度由于与特征提取方法无关，因此，实验4与实验3的结果完全相同。

结束语 本文提出的基于正交投影的分类器算法，其特点是由各类的所有训练样本生成子空间，克服了最小距离和最近邻分类器由单个特征点代表各类的局限。同时不论训练样本的数量多少，该算法都可以方便地生成子空间，避免了贝叶斯分类器中各类协方差矩阵因小样本问题而引起的奇异性问题。本文算法由于计算非常简单且易于理解，因此非常适合于小样本数据的识别问题。

参 考 文 献

[1] Li S Z. Face Recognition Based on Nearest Linear Combinations [C]// Proceedings of CVPR. 1998:839-844

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

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

[4] Zheng Wen-ming, Zou Cai-rong, Zhao Li. Face Recognition Using Two Novel Nearest Neighbor Classifiers[C]// Proceedings of ICASSP. 2004:725-728

[5] Friedman J H. Reguralized discriminant analysis[J]. J. Amer. Statistic. Assoc., 1989, 84(405):165-175

[6] 王卫东，郑宇杰，杨静宇. 采用虚拟训练样本优化正则化判别分析[J]. 计算机辅助设计与图形学学报，2006，18(9)：1327-1331

[7] 王卫东，杨静宇. 采用虚拟训练样本的二次判别分析方法[J]. 自动化学报，2008，34(4)：400-407

[8] Lay D C. Linear Algebra and Its Applications (Third Edition) [M]. Pearson International Edition, 2003

(上接第144页)

3) 相反，如果在n次计算中某一次计算得到的操作符路径与前一次相比发生了改变，则说明数据流的特性在发生变化，应该提高系统的灵敏度，即减小路径计算的时间间隔，因此将τ减小一半，$\tau=\tau/2$。

4 性能测试与评价

本文将MultiFactor算法与FIFO，Greedy算法进行了比较。实验环境为Pentium4 1.7GHz的CPU，内存256M，操作系统是Windows 2000 Professional。与其它的调度算法相比，MultiFactor查询优化策略不仅考虑了对系统内存的需求，而且考虑了系统时间延迟，是对二者进行的一种综合考虑。

图2为各个算法在运用于单流上的查询时系统内存使用量的比较。我们利用系统中元组的数量来代表内存使用量。从该图可以看出，在内存使用量方面，MultiFactor策略要好于其它两种策略。

图3为各算法在运用于单流上的查询时时间延迟的比较。可以看出MultiFactor策略与其它两种策略相比，时间延迟较小，并且随着时间的变化优势更加明显。并且MultiFactor调度策略根据查询的截止时间来得到调度时间片，而查询中的截止时间为600ms。所以从该图中可以看到MultiFactor的时间延迟是小于600ms的。

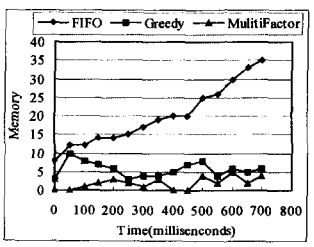

图2 单流上3种算法内存使用量比较

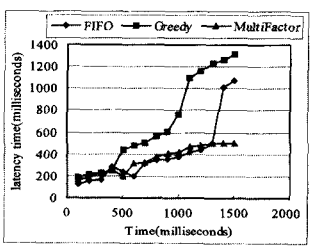

图3 单流上3种算法时间延迟的比较

图4是各个算法在包含连接操作符的查询上进行内存使用量的比较。从图中可以看出，具有连接操作符的查询比不含有连接的查询内存使用量有了明显的增加。MultiFactor策略在加入连接操作符之后，在内存使用量方面仍好于其它两种策略。

图5为各个算法在包含连接操作符的查询上时间延迟的比较。可以看出，具有连接操作符的查询比不含有连接的查询时间延迟上有了明显的增加，MultiFactor策略在加入连接操作符之后，在时间延迟方面仍好于其它两种策略。

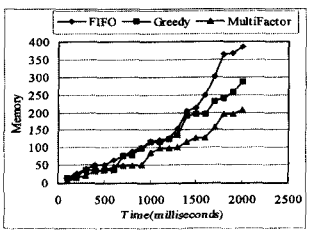

图4 多流上3种算法内存使用量比较

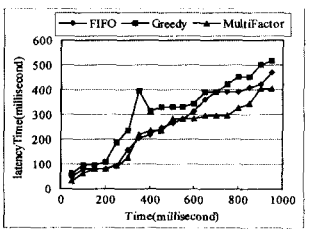

图5 多流上3种算法时间延迟比较

结束语 本文根据数据流的种种特性，考虑了系统的内存需求和查询的响应时间，提出了一种适应性查询优化及调度策略MultiFactor。该策略根据各操作符消耗系统中元组数量的快慢来动态调整操作符的调度次序，按查询的截止时间来确定各操作符的调度时间，并依据数据流各个特性因素的变化来适时调整系统的查询计划，是一种具有适应性的查询优化策略。

参 考 文 献

[1] Abadi J, Cherniack M, Christian C, et al. Aurora: a new model and architecture for data stream management[J]. the VLDB Journal, 2003, 12(2):120-139

[2] Carney D, Çetintemel U, et al. Operator Scheduling in a Data Stream Manager[C]// Proc. of the 29th International Conference on Very Large Data Bases (VLDB'03). Berlin, Germany, September 2003:213-216

[3] Das A, Dally W J. Stream Scheduling: A Framework to Manage Bulk Operations in a Memory Hierarchy[C]// Proc. of the 16th International Conference on Parallel Architecture and Compilation Techniques. 2007, 405:253-262

[4] Munagala K, Srivastava U, Widom J. Optimization of Continuous Queries with Shared Expensive Filters[C]// Proceedings of the twenty-sixeh ACM Sigmod-Sigct-Sigart Symposium on Principles of Database Systems. 2007:215-224

[5] Chandrasekaran S, Cooper O, Deshpande A, et al. TelegraphCQ: Continuous dataflow processing for an uncertain world[C]// Proc. of Conference on Innovative Data Systems Research. Asilomar, CA, 2003:269-280