基于次特征值误差补偿和非对称分布的马氏距离改进算法

李国宏, 施鹏飞
(上海交通大学图像处理与模式识别研究所, 上海 200030)

摘 要: 本文提出了一种有限样本集上基于次特征值误差补偿和优势主向量上非对称分布的马氏距离改进算法. 通过改进的马氏距离, 有限样本导致的次特征值误差得到补偿, 样本特征矢量在变换空间的各优势主向量上的投影分布得到更精确的刻画, 因此可以有效地计算最近邻参考矢量. 在 UCI 手写体数字字符数据库上的识别实验结果表明, 该改进算法对于提高识别性能是有效的.

关键词: 特征值; 非对称分布; 马氏距离; 误差
中图分类号: TP391.4 文献标识码: A 文章编号: 0372-2112 (2007) 04-0747-04

Modified Mahalanobis Distance by Compensation for Errors of Non-dominant Eigenvalues and Asymmetrical Distribution

LI Guo-hong, SHI Peng-fei
(Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai 200030, China)

Abstract: A modification on Mahalanobis distance on samples of limited size by compensation for errors of non-dominant eigenvalues and asymmetrical distribution on dominant principle components is proposed. By the introduction of modified Mahalanobis distance, one can compute efficiently the nearest neighbor in transformed space, with compensation for errors of non-dominant eigenvalues and more accurate characterization of the projection distribution of feature vector on every dominant principal component. Experimental results on UCI dataset for handwritten digit recognition indicate that modified algorithm is effective to improve the recognition performance.

Key words: eigenvalue; asymmetrical distribution; Mahalanobis distance; error

1 引言

马氏距离是模式识别中一个有效的相似性测度, 可以采用从主向量分析(PCA) 中得到的特征值-特征向量表达. PCA 对描述数据的原坐标系进行正交变换[1], 变换得到的新坐标值称为主向量, 通常采用主向量描述数据的结构比原坐标系要有效得多.

当可以利用的样本数量有限时, PCA 计算的特征值通常包含误差. 因此, 有必要采用改进的马氏距离来计算未知模式的特征矢量与某类的均值矢量之间的距离, 文献[2] 采用了对特征值加上偏移量的方法对所有的特征值进行误差补偿. 然而, 马氏距离在次主向量(对应于较小的特征值) 上的偏差大于在优势主向量上的偏差, 对主特征值进行误差补偿可能增加类间的相似度, 因此通过以较大的值替换次特征值, 减小次主向量空间的距离偏差, 可以减小总体距离偏差[3]. 传统的马氏距离是在多变量正态分布概率密度函数的假设下推导出来的, 因此, 如果样本的分布服从多变量正态分布, 马氏距离被认为是一个合适的测度指标. 然而, 研究发现样本的分布与正态分布有较大的差异, 主要表现在类内样本分布的非对称特性方面. 本文从白化变换的角度[4] 分析改进的马氏距离, 对次特征值进行误差补偿, 并且以非对称模型描述特征矢量在优势主向量上的分布. 本文给出了一个监督学习领域的实例, 以支持本文马氏距离改进算法的有效性.

2 特征矢量空间的白化变换

给定某类的 M 个训练样本 $\boldsymbol{x}_k$, $k=1, \cdots, M$, 其中 $\boldsymbol{x}_k \in R^N$, $\bar{\mu}=\frac{1}{M}\sum_{k=1}^{M}\boldsymbol{x}_k$, PCA 对协方差矩阵

$$\Sigma_X=\frac{1}{M}\sum_{j=1}^{M}(x_j-\bar{\mu})(x_j-\bar{\mu})^{\mathrm{T}} \tag{1}$$

进行对角化. 为此, 该类的协方差矩阵 Σ_X 被分解为特征值矩阵与特征向量矩阵的乘积:

收稿日期: 2005-01-17; 修回日期: 2007-01-03
基金项目: 国家自然科学基金(No. 60075007); 国家 863 高技术研究发展计划(No. 2006AA01Z119)

$$\Sigma_X = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}} \tag{2}$$

其中, $\boldsymbol{U}$ 是特征向量矩阵, 且 $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}$, $\boldsymbol{D}$ 是由特征值构成的对角矩阵. 令 λ_j 和 ϕ_j 分别是协方差矩阵的第 j 个特征值和特征向量.

以 $\boldsymbol{U}$ 为变换矩阵, 一个 N 维矢量 $\boldsymbol{X}$ 可以线性变换为另一个 N 维矢量 $\boldsymbol{Y}$, 相应地, $\boldsymbol{Y}$ 可以表示为一个 $\boldsymbol{X}$ 的函数:

$$\boldsymbol{Y} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{X} \tag{3}$$

于是有,

$$\Sigma_Y = \boldsymbol{U}^{\mathrm{T}}\Sigma_X\boldsymbol{U} = \boldsymbol{D} \tag{4}$$

PCA 是一种正交变换, 因此保持欧氏距离特性, 即

$$\| \boldsymbol{Y} \|^2 = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \| \boldsymbol{X} \|^2 \tag{5}$$

正交变换后, 可以再对 $\boldsymbol{Y}$ 进行另一个 $\boldsymbol{D}^{-1/2}$ 变换, 使得协方差矩阵变换为单位矩阵 $\boldsymbol{I}$.

$$\boldsymbol{Y} = \boldsymbol{D}^{-1/2}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X} = (\boldsymbol{U}\boldsymbol{D}^{-1/2})^{\mathrm{T}}\boldsymbol{X} \tag{6}$$

$$\Sigma_Y = \boldsymbol{D}^{-1/2}\boldsymbol{U}^{\mathrm{T}}\Sigma_X\boldsymbol{U}\boldsymbol{D}^{-1/2} = \boldsymbol{D}^{-1/2}\boldsymbol{D}\boldsymbol{D}^{-1/2} = \boldsymbol{I} \tag{7}$$

变换矩阵 $\boldsymbol{U}\boldsymbol{D}^{-1/2}$ 称为白化变换或白化处理. 变换矩阵 $\boldsymbol{D}^{-1/2}$ 的目的是以 $1/\sqrt{\lambda_i}$ 为比例因子改变主向量的尺度, 该变换意味着在主向量上对样本参数进行规范化. 变换后的距离以下式定义:

$$\begin{aligned} d(\boldsymbol{x}) &= \| (\boldsymbol{U}\boldsymbol{D}^{-1/2})^{\mathrm{T}}(\boldsymbol{x} - \bar{\mu}) \| \\ &= ((\boldsymbol{U}\boldsymbol{D}^{-1/2})^{\mathrm{T}}(\boldsymbol{x} - \bar{\mu}))^{\mathrm{T}}((\boldsymbol{U}\boldsymbol{D}^{-1/2})^{\mathrm{T}}(\boldsymbol{x} - \bar{\mu})) \\ &= (\boldsymbol{x} - \bar{\mu})^{\mathrm{T}}\boldsymbol{U}\boldsymbol{D}^{-1/2}\boldsymbol{D}^{-1/2}\boldsymbol{U}^{\mathrm{T}}(\boldsymbol{x} - \bar{\mu}) \\ &= (\boldsymbol{x} - \bar{\mu})^{\mathrm{T}}\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{U}^{\mathrm{T}}(\boldsymbol{x} - \bar{\mu}) \\ &= (\boldsymbol{x} - \bar{\mu})^{\mathrm{T}}\Sigma_X^{-1}(\boldsymbol{x} - \bar{\mu}) \end{aligned} \tag{8}$$

白化变换不是正交变换, 这是因为:

$$(\boldsymbol{U}\boldsymbol{D}^{-1/2})^{\mathrm{T}}\boldsymbol{U}\boldsymbol{D}^{-1/2} = \boldsymbol{D}^{-1/2}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{D}^{-1/2} = \boldsymbol{D}^{-1} \neq \boldsymbol{I} \tag{9}$$

因此, 该变换不保持欧氏距离特性, 即

$$\| \boldsymbol{Y} \|^2 = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{D}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{X}^{\mathrm{T}}\Sigma_X^{-1}\boldsymbol{X} \neq \| \boldsymbol{X} \|^2 \tag{10}$$

由式(8)可以看出, 白化变换空间的距离实际上就是未知模式的特征矢量 $\boldsymbol{x}$ 与某类的均值矢量 $\bar{\mu}$ 之间的马氏距离, 马氏距离是一个用于模式识别的距离测度, 可以采用特征值-特征向量的方式表达如下:

$$d(\boldsymbol{x}) = \sum_{j=1}^{N} \frac{1}{\lambda_j}(\boldsymbol{x} - \bar{\mu}, \phi_j)^2 \tag{11}$$

有时, 某些类的协方差矩阵是奇异矩阵, 因此, 计算马氏距离时, 首先需要计算该协方差矩阵的伪逆矩阵. 本质上, 采用式(11)计算马氏距离将不可避免地遇到 $(\boldsymbol{x} - \bar{\mu}, \phi_j)^2/(\lambda_j) = A/0$ 的情况.

3 次特征值的误差补偿

在许多模式识别问题中, 维数 N 的取值可能非常大, 而只有较少的特征值是主要的, 即

$$\lambda_1 + \cdots + \lambda_N \cong \lambda_1 + \cdots + \lambda_k \quad (k \ll N) \tag{12}$$

研究表明, 从有限样本集上计算的特征值通常包含误差. 因此, 有必要采用改进的马氏距离来计算未知模式的特征矢量 $\boldsymbol{x}$ 与某类的均值矢量 $\bar{\mu}$ 之间的距离, 文献[2]采用了对特征值加上偏移量 b 的方法, 即

$$d(\boldsymbol{x}) = \sum_{j=1}^{n} \frac{1}{\lambda_j + b}(\boldsymbol{x} - \bar{\mu}, \phi_j)^2 \tag{13}$$

然而, 研究发现, 马氏距离在次主向量的偏差大于在优势主向量上的偏差. 考虑到对主特征值进行误差补偿可能同时增加类间的相似度, 因此本文只对相应的次特征值的误差进行补偿. 通过以较大的值替换次特征值, 可以减小次主向量空间的总体距离偏差. 当采用式(14)

$$\begin{aligned} k = \{ n \mid & (\lambda_1 + \lambda_2 + \cdots + \lambda_{n-1})/(\lambda_1 + \lambda_2 + \cdots + \lambda_N) < thr, \\ & (\lambda_1 + \lambda_2 + \cdots + \lambda_n)/(\lambda_1 + \lambda_2 + \cdots + \lambda_N) \geqslant thr \} \end{aligned} \tag{14}$$

选定某类 k 个主特征值后(其中 $0 < thr \leqslant 1.0$ 为主特征值选择阈值), 便可采用 λ_k 替换其余 $N-k$ 个次特征值 $\lambda_{k+1}, \lambda_{k+2}, \cdots, \lambda_N$, 即,

$$\bar{\lambda}_i = \begin{cases} \lambda_i, & \text{if} \quad i \leqslant k \\ \lambda_k, & \text{if} \quad i > k \end{cases} \tag{15}$$

修改后的特征值矩阵采用下式表示:

$$\bar{\boldsymbol{D}} = \mathrm{diag}\{\bar{\lambda}_1, \bar{\lambda}_2, \cdots, \bar{\lambda}_N\} \tag{16}$$

然后, 修改后的协方差矩阵可以按下式计算:

$$\bar{\Sigma}_X = \boldsymbol{U}\bar{\boldsymbol{D}}\boldsymbol{U}^{\mathrm{T}} \tag{17}$$

显然 $\bar{\Sigma}_X$ 必定是非奇异矩阵, 因此改进的距离测度则可以定义如下:

$$\bar{d}(\boldsymbol{x}) = (\boldsymbol{x} - \bar{\mu})^{\mathrm{T}}\bar{\Sigma}_X^{-1}(\boldsymbol{x} - \bar{\mu}) \tag{18}$$

上式也可以采用下列的特征值-特征向量表达方式:

$$\bar{d}(\boldsymbol{x}) = \sum_{j=1}^{N} \frac{1}{\bar{\lambda}_j}(\boldsymbol{x} - \bar{\mu}, \phi_j)^2 \tag{19}$$

式(19)同时从本质上避免了计算时出现 $(\boldsymbol{x} - \bar{\mu}, \phi_j)^2/(\bar{\lambda}_j) = A/0$ 的情况.

4 非对称分布

传统的马氏距离是在多变量正态分布概率密度函数的假设下推导出来的, 因此, 如果样本的分布服从多变量正态分布, 马氏距离被认为是一个合适的测度指标. 然而, 在本文的研究过程中, 发现样本的分布与正态分布有较大的差异, 主要表现在类内样本分布的非对称特性方面, 分布的形状由于一些样本变形较大而被不可思议地改变, 非对称分布可以采用图 1 所示的概率密度函数描述.

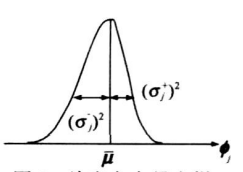

图 1 给定主向量上样本的非对称分布

为了描述非对称分布, 需要估计准方差(见图 1), 为此首先定义下列矢量集合 $\boldsymbol{S}_j^+$ 和 $\boldsymbol{S}_j^-$:

$$S_j^+ = \{\bar{u}_j^i \mid \bar{u}_j^i \geqslant 0\} \quad (20)$$

$$S_j^- = \{\bar{u}_j^i \mid \bar{u}_j^i < 0\} \quad (21)$$

其中 $\bar{u}_j^i = (x_i - \bar{\mu}, \phi_j)$ 表示 $x_i - \bar{\mu}$ $(i = 1, 2, \cdots, M)$ 在特征向量 $\phi_j (j = 1, 2, \cdots, k)$ 上的分量投影, 准方差 $(\sigma_j^+)^2$ 和 $(\sigma_j^-)^2$ 定义如下:

$$(\sigma_j^+)^2 = \frac{1}{|S_j^+|} \sum_{u \in S_j^+} u^2 \quad (22)$$

$$(\sigma_j^-)^2 = \frac{1}{|S_j^-|} \sum_{u \in S_j^-} u^2 \quad (23)$$

采用非对称模型刻画式(14)选定的优势主向量上的特征矢量分布, 对式(19)改进后的马氏距离采用下列的特征值-特征向量表达方式:

$$\bar{d}(x) = \sum_{j=1}^{N} \frac{1}{\sigma_j} (x - \bar{\mu}, \phi_j)^2 \quad (24)$$

其中,

$$\sigma_j = \begin{cases} \sigma_j^+, \text{ if } (x - \bar{\mu}, \phi_j) \geqslant 0 \ \&\& \ j \leqslant k, \\ \sigma_j^-, \text{ if } (x - \bar{\mu}, \phi_j) < 0 \ \&\& \ j \leqslant k, \\ \lambda_j, j > k \end{cases} \quad (25)$$

5　实验结果及分析

本文的算法在 UCI 数据集[5]的手写体数字字符数据库上进行测试. 该数据库共包含 5620 个字符样本, 本文将其中训练集中的 3823 个样本用于训练, 测试集中其余的 1797 个样本用于测试. 图 2(a)和(b)分别给出了用于训练和识别的部分样本图像, 表 1 给出了每类用于训练和识别的样本数量. 为减小图像的尺寸, 共计算 8×8=64 个像素, 其中每个像素值等于原图像中 4×4 块中各像素值之和. 基于训练集样本计算 10 类(0~9)字符的次特征值补偿值及优势主向量的非对称分布参数(准方差). 根据式(24)计算每一个测试样本到所有类别均值矢量的距离, 采用最小距离分类器进行分类决策. 本文的方法在不同的 thr 值(式(14))下进行识别率测试, 结果参见图 3. 表 2 列出了不同 thr 取值条件下各类的 k 值(优势主向量的数量). 本文的方法与现有的一些其它方法进行了识别性能比较, 包括马氏距离, 采用欧氏距离为测度指标的 K-NN 算法. 比较实验的结果列于表 3 中.

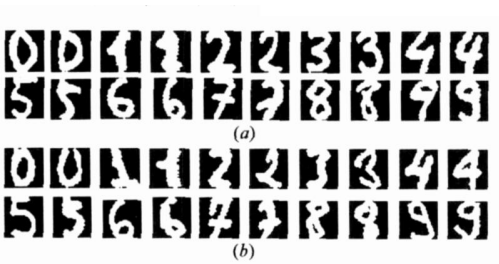

图 2　UCI 数据集中的样本 (a) 训练样本; (b) 测试样本

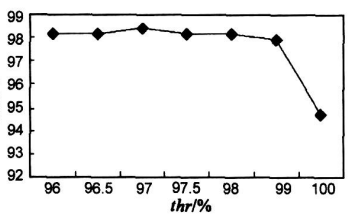

图 3　不同 thr 条件下的识别率结果

表 1　每类用于训练和测试的样本数量

类别	0	1	2	3	4	5	6	7	8	9
训练	376	389	380	389	387	376	377	387	380	380
测试	178	182	177	183	181	182	181	179	174	180

表 2　不同 thr 取值条件下各类的 k 值

thr(%) \ Class	0	1	2	3	4	5	6	7	8	9
96	10	9	9	12	11	12	8	11	14	13
96.5	11	10	10	13	12	13	9	12	15	14
97	13	11	11	15	14	15	10	13	16	16
97.5	14	12	13	16	15	16	11	15	18	17
98	16	14	15	18	17	18	13	17	20	19
99	22	20	21	24	24	24	19	23	26	25
100	64	64	64	47	64	64	45	64	47	64

表 3　不同算法的识别率比较(%)

马氏距离	改进的马氏距离		K-NN(采用欧氏距离测度)		
	$thr = 0.97$	$thr = 0.98$	$K = 1$	$K = 2$	$K = 3$
94.71	98.39	98.16	98.00	97.38	97.83

图 3 中的识别率数据表明, 本文算法的识别性能随补偿阈值 thr 的变化而变化, 当 $thr = 0.97$ 时, 识别率达到最大值, 98.39%. 其中, 对应于 $thr = 1.00$(即未进行次特征值误差补偿)的识别率为未采用非对称分布补偿的识别率, 显然等于采用马氏距离的识别率. 从表 2 容易看出, 各类的优势主向量的个数随 thr 变化, 类别 3, 6, 8 的原始协方差矩阵是奇异矩阵, 因为它们非 0 特征值的个数小于特征矢量的维数, 因此, 在计算输入特征矢量到上述三类的马氏距离时, 需要采用它们各自协方差矩阵的伪逆矩阵. 从表 3 可以看出, 本文改进的马氏距离的分类性能明显优于马氏距离, 而且超过采用欧氏距离为测度指标的 K-NN 算法. 最后应该指出, 在识别阶段本文的算法与其它方法相比, 不需要额外的计算开销.

6　结论

本文提出了一种有限样本集上基于次特征值误差补偿和优势主向量上非对称分布的马氏距离改进算法. 通过改进的马氏距离, 有限样本导致的次特征值误差得到补偿, 样本特征矢量在变换空间的各优势主向量上的投影分布得到更精确的刻画, 因此可以有效地计算最近邻参考矢量. 本文的算法在 UCI 手写体数字字符数据库上进行识别实验, 在单特征矢量单分类器的条件下, 识别率达到 98.39%, 且识别时不需要额外

的计算开销, 表明该改进算法对于提高识别性能是有效的.

参考文献:

[1] Scholkopf B, Smola A, Muller K R. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural Comput, 1998, 10(5): 1299- 1319.

[2] Kato N, Suzuki M, Omachi S, et al. A handwritten character recognition system using directional element feature and asymmetric mahalanobis distance[J]. IEEE Trans PAMI, 1999, 21(3): 258- 262.

[3] Takeshita T, Nozawa S, Kimura F. On the bias of Mahalanobis Distance due to limited sample size effect[A]. Document Analysis and Recognition, Proceedings of Second International Conference[C]. IEEE, 1993. 171- 174.

[4] Fukunaga K. Introduction to Statistical Pattern Recognition (2nd Edition)[M]. New York: Academic Press, 1990.

[5] Blake C, Keogh E, Merz C J. UCI Repository of machine learning database[OL]. http://www.ics.uci.edu/~ mlearn/MLRepository.html, 1998.

作者简介:

李国宏 男, 1969 年生于河南沁阳. 2005 年在上海交通大学图像处理与模式识别研究所获得博士学位, 现为华为技术有限公司工程师, 主要从事无线接入网方面的研究工作.

E-mail: ligh0929@sohu.com

施鹏飞 男, 1940 年生于上海. 1966 年上海交通大学电机系研究生毕业, 现为上海交通大学教授、博士生导师, 主要从事生物特征识别、智能交通系统、医学图像处理等方面的研究工作.

E-mail: pfshi@sjtu.edu.cn