第 33 卷 第 9 期 计 算 机 工 程 2007 年 5 月
Vol.33 No.9 Computer Engineering May 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)09—0204—03 文献标识码：A 中图分类号：TP391.41

基于混合概率 PCA 模型高光谱图像本征维数确定

普 鑫
(中国科学院研究生院，北京 100072)

摘 要：如何有效实现降维是现代成像光谱仪辨识地物类别的一个难点所在。该文在已知高光谱图像地物类别数的情况下，提出了一种采用混合最小描述长度(MMDL)模型选择准则确定高光谱图像本征维数的方法。该方法在期望最大化算法框架下同时实现混合 PPCA 降维和聚类，并根据 MMDL 准则确定数据降维维数，可以得到数据在概率意义下的精确的降维表征。仿真数据和真实数据进行的比较实验表明，该方法能精确地选择数据的本征维数。

关键词：降维；本征维数；混合概率主成分分析；混合最小描述长度准则；期望最大化算法

Intrinsic Dimensionality Determination for Hyperspectral Image Based on Mixture of Probabilistic PCA Model

PU Xin
(Graduate School, Chinese Academy of Sciences, Beijing 100072)

【Abstract】 An intrinsic dimensionality determination method for hyperspectral image with known class number is proposed, which is based on mixture model of probabilistic PCA. Different from common methods that determine the number of dimensionality reduction by setting the number or by eigenvalue thresholding, the algorithm simultaneously conducts dimensionality reduction and clustering under the frame of EM algorithm; and retrieves the intrinsic dimensionality according to the MMDL principle with probabilistically accurate reduced representation of the data. The method can achieve precise results applied to simulated data and real data.

【Key words】 Dimensionality reduction; Intrinsic dimensionality; Mixture of probabilistic principal component analysis (PPCA); Principle of mixture minimum description length (MMDL); Expectation maximization(EM) algorithm

现代成像光谱仪较高的光谱分辨率为辨识地物类别提供了足够的信息，然而其数据的高维特性使得降维成为高光谱图像解译的一个重要部分[1]，因此有效降维成为其中的关键。通常用的方法是PCA变换，通过设定信息量门限来确定降至几维[2]，然而对于不同的数据，门限如何设定是一个难点。本文从PCA方法入手，提出一种基于混合概率PCA模型的高光谱图像本征维数确定方法。不同于传统PCA的方法，基于混合概率PCA模型具有概率模型[3,4]的诸多优点，而且能进一步扩展至混合模式，这种方法扩展了PCA的应用范围，而且可以采用Bayes模型选择方法[5]来更精确地确定数据的本征维数。

1 理论分析

1.1 概率 PCA 模型

对于 d 维观测数据集 $\{t_n, n=1,2,...,N\}$ 中的单个样本矢量 t，传统 PCA 通过与一个变换矩阵 $W: x=W^T(t-\bar{t})$ 相乘来得到数据的降维表示，其中 $\bar{t}$ 为样本均值，S 为样本协方差矩阵，W 由 S 的本征向量组成：$W=(w_1,w_2,...,w_q)$，其中 $Sw_j=\lambda_j w_j$，λ 为 S 的本征值，q 为降维后的维数，即主子空间维数。

隐变量模型与PCA之间有紧密的联系[3]。隐变量模型给出了观测数据t和隐变量x之间的关系，其中最常用的是因子分析，它描述的是线性关系：

$$t=Wx+\mu+\varepsilon \tag{1}$$

式中，W 为因子载荷，μ 为模型均值，ε 为误差项。通常定义 x 和 ε 服从高斯分布，即 $x\sim N(0,I)$ 以及 $\varepsilon\sim N(0,\Psi)$，且 Ψ 为对角阵，则 $t\sim N(\mu,C)$，模型协方差 $C=WW^T+\Psi$。

PCA可视为因子分析的特殊情形，当误差为各向同性，即 $\varepsilon\sim N(0,\sigma^2 I)$ 时，式(1)可以建立一个由隐变量空间至观测数据主子空间的映射，将因子分析与PCA联系起来[3]：

$$p(t\mid x)=(2\pi\sigma^2)^{-d/2}\exp\left\{-\frac{1}{2\sigma^2}\|t-Wx-\mu\|^2\right\} \tag{2}$$

可推导出隐变量 x 关于观测变量 t 的后验概率密度分布：

$$p(x\mid t)=(2\pi)^{-q/2}\left|\sigma^{-2}M\right|^{1/2}\exp\left\{-\frac{1}{2}\left[x-M^{-1}W^T(t-\mu)\right]^T\left(\sigma^{-2}M\right)\left[x-M^{-1}W^T(t-\mu)\right]\right\} \tag{3}$$

其中 $M=W^TW+\sigma^2 I$，维数为 $q\times q$，而 $C=WW^T+\sigma^2 I$，为 $d\times d$。

由此得到单一的概率 PCA(PPCA)模型，在该模型下观测数据的对数似然函数为

$$L(t)=\sum_{n=1}^{N}\ln\{p(t_n)\}=-\frac{N}{2}\left\{d\ln(2\pi)+\ln|C|+tr(C^{-1}S)\right\} \tag{4}$$

各参数的最大似然估计为

$$\mu_{ML}=\frac{1}{N}\sum_{n=1}^{N}t_n,\quad S_{ML}=\frac{1}{N}\sum_{n=1}^{N}(t_n-\mu)(t_n-\mu)^T$$

将式(4)最大化可得

$$W_{ML}=U_q(\Lambda_q-\sigma^2 I)^{1/2}R$$

作者简介：普 鑫(1974—)，男，工程师、硕士生，主研方向：计算机应用

收稿日期：2006-10-12 **E-mail：**xin_001@sina.com

其中 $d\times q$ 矩阵 U_q 的列矢量为 S 的本征向量，对应 $q\times q$ 对角矩阵 Λ_q 中的特征值，R 为任意 $q\times q$ 正交旋转矩阵，实际应用中可以简化地取 $R=I$ ；$\sigma_{ML}^2=\frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j$ 。由此用最大似然参数估计代替了通常的样本协方差矩阵特征值分解。另外 W 和 σ^2 除了用以上公式显示计算，还可以用 EM 迭代算法高效地求解。

1.2 混合 PPCA 模型

将以上的 PPCA 扩展到混合形式，则观测数据集的混合模型对数似然函数为

$$L=\sum_{n=1}^{N}\ln\{p(t_n)\}=\sum_{n=1}^{N}\ln\left\{\sum_{i=1}^{M}\pi_i p(t_n\mid i)\right\} \tag{5}$$

其中 $p(t\mid i)$ 是前述的单一PPCA模型，M为混合成分数，π_i 为混合比例，$\pi_i\ge 0$ 且 $\sum\pi_i=1$，每个混合成分的参数为 μ_i、W_i 和 $\sigma_i{}^2$ 。则可由E-step和M-step推导出用EM算法求解混合PPCA的迭代公式[4]：

$$\begin{cases}\tilde{\pi}_i=\frac{1}{N}\sum_n R_{ni}\\ \tilde{\mu}_i=\frac{\sum_n R_{ni}(t_{ni}-\tilde{W}_i\langle x_{ni}\rangle)}{\sum_n R_{ni}}\\ \tilde{W}_i=\left[\sum_n R_{ni}(t_{ni}-\tilde{\mu}_i)\langle x_{ni}\rangle^T\right]\left[\sum_n R_{ni}\left\langle x_{ni}x_{ni}{}^T\right\rangle\right]^{-1}\\ \tilde{\sigma}_i^2=\frac{1}{d\sum_n R_{ni}}\left\{\sum_n R_{ni}\|t_{ni}-\tilde{\mu}_i\|^2-2\sum_n R_{ni}\langle x_{ni}\rangle^T\tilde{W}_i^T(t_n-\tilde{\mu}_i)+\sum_n R_{ni}tr\left(\left\langle x_{ni}x_{ni}{}^T\right\rangle\tilde{W}_i^T\tilde{W}_i\right)\right\}\end{cases} \tag{6}$$

其中 $R_{ni}=\frac{p(t_n\mid i)\pi_i}{p(t_n)}$ 表示第 i 个混合成分生成第 n 个数据点 t_n 的后验概率。

1.3 混合 MDL 模型选择准则

引人概率模型后，即可对混合 PPCA 模型应用 Bayes 方法进行模型选择，方法是对每个降维维数应用 EM 算法得到一系列的参数估计，最小化某代价函数的维数即为本征维数估计值。这里代价函数可以由 MDL 准则确定，对于混合情形，可应用 MMDL 准则。

MDL 准则可以由最小消息长度(MML)准则推导出来，MML 准则为

$$\hat{\theta}=\arg\min_{\theta}\{-\log p(\theta)-\log p(y\mid\theta)+\frac{1}{2}\log\left|I(\theta)\right| +\frac{N(k)}{2}(1+\log K_{N(k)})\} \tag{7}$$

其中，$I(\theta)\equiv -E[\frac{\partial^2}{\partial\theta^2}\log p(y\mid\theta)]$ 为 Fisher 信息阵，$\left|I(\theta)\right|$ 是其行列式；$K_{N(k)}$ 为 $N(k)$ 维空间的最优量化栅格常数；$N(k)=(k-1)+k(d+d(d+1)/2)$ ，即独立参数的个数。而 $I(\theta)=nI^{(1)}(\theta)$ ($I^{(1)}(\theta)$ 是对应单个观测样本的 Fisher 信息阵，n 为样本尺寸)。对其进行简化，用 $\log p(y\mid\hat{\theta})$ 近似 $\log p(y\mid\theta)$ ，得到 MDL 准则下的代价函数为

$$L(\hat{\theta},y)=-\log p(y\mid\hat{\theta})+\frac{N(k)}{2}\log n \tag{8}$$

由于混合模型中所有的数据点并非具有相同的重要性，每个数据点在估计不同的参数时具有不同的权值(混合系数)[5]，因此估计混合模型中第m个模型中参数的Fisher信息应该为

$$I(\theta_m)=n\alpha_m I^{(1)}(\theta_m) \tag{9}$$

其中 $I^{(1)}(\theta_m)$ 为第 m 个模型产生的单个观测所对应的 Fisher 信息。由此混合模型的 MDL 准则(MMDL)的代价函数为

$$L(\hat{\theta},y)=-\log p(y\mid\hat{\theta})+\frac{N(k)}{2}\log n+\frac{N(1)}{2}\sum_{m=1}^{k}\log\alpha_m \tag{10}$$

2 具体算法

从上述分析，可以在已知类别数 K 的条件下提出基于混合概率 PCA 模型的本征维数确定方法，算法具体流程如图 1 所示。

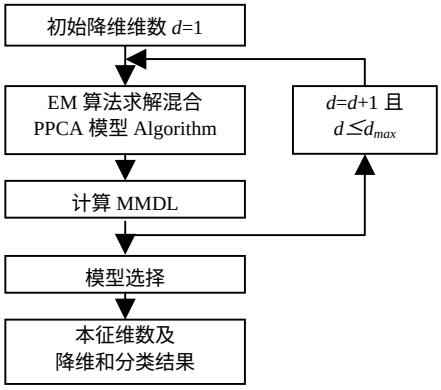

图 1 算法流程

(1)设定最大降维维数d_{max}，初始降维维数d=1；

(2)在当前降维维数下用 K 均值算法对数据进行初始分类，计算混合 PPCA 模型初始参数，并将初始参数代入式(6)的 EM 算法进行迭代直到收敛；

(3)计算当前模型下的 MMDL 准则值；

(4)重复(2)、(3)直到$d=d_{max}$；

(5)根据 MMDL 准则进行模型选择确定合适的降维维数，根据所选择模型的参数按照最大后验概率(MAP)准则对数据进行分类。

3 实验结果

为验证上述方法的可行性，采用仿真数据进行测试，这里参照文献[2]生成仿真数据，数据包含 3 类，每类含 1 000 个样本，共 3 000 个数据点，数据维数为 10，信号维和数据维均为 5，每类数据均服从多维高斯分布，均值分别为 $\mu_1=[0\,0\,0\,0\,0\,0\,0\,0\,0\,0]$，$\mu_2=[\delta\,0\,0\,0\,0\,0\,0\,0\,0\,0]$，$\mu_3=[0\,\delta\,0\,0\,0\,0\,0\,0\,0\,0]$。$\delta=40$ 确定类间距，协方差阵为 $C=diag\{10,8,6,4,2,1,1,1,1,1\}$ ，图 2 显示的是仿真数据前二维的表现，图 3 为算法对仿真数据一次实验所得到的 MMDL 曲线及模型选择结果，可见算法正确选择了数据的本征维数 5。重复进行了 50 次仿真实验，结果均为 5。

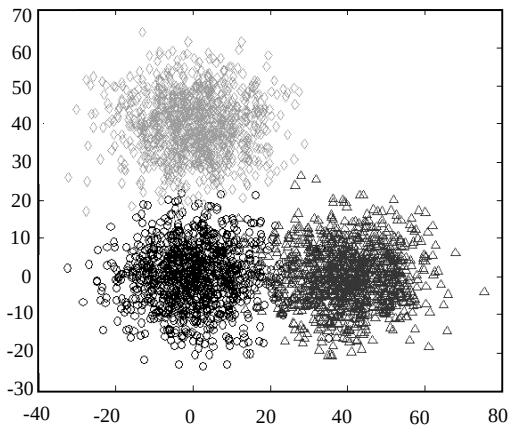

图 2 仿真数据(前二维)

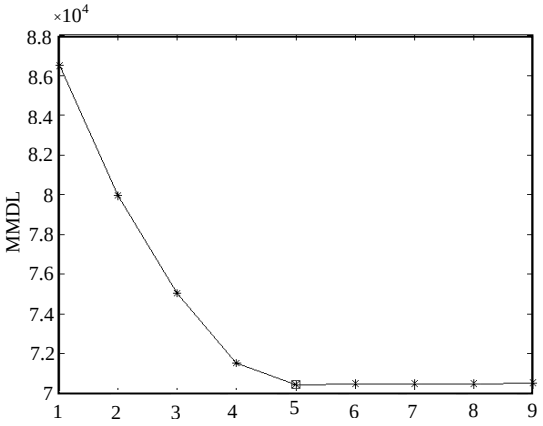

图 3 MMDL 曲线及模型选择结果

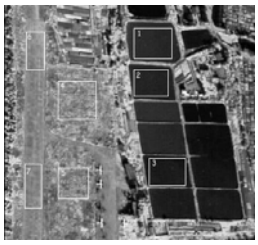

图 4 PHI 真实数据

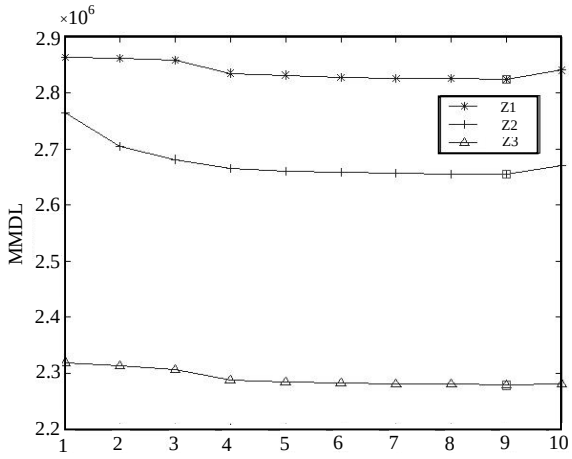

图 5 MMDL 曲线及模型选择结果

另外采用PHI高光谱图像数据进行测试，光谱范围为可见光-近红外波段(0.40μm~0.95μm)，可选波段 244。图像尺寸为 330×311，共 80 个波段，如图 4。图中 1、2、3 为水域，4、5 为草地，6、7 为机场跑道，取图中所示的不同类型的地物分别进行了 3 次实验，分别为区域 1、4、6，共 5 700 点，区域 2、4、7，共 5 351 点，区域 3、5、7，共 4 604 点。由于高光谱图像各波段间的相关性较强，降维的幅度较大，设定最大降维维数d_{max}为 10。图 5 为MMDL曲线及模型选择结果，3 次实验所选择的本征维数均为 9。

实验表明与一般的指定降维维数或设定信息量门限的降维方法相比较，本文所提出的基于混合概率 PCA 模型的高光谱图像本征维数确定方法具有如下优点：

(1)具有概率模型，可以采用 Bayes 模型选择方法更精确地确定数据的本征维数，得到数据在概率意义下的精确降维表征；

(2)可以推广至由具有不同维数的模型组合形成的软混合模型，在实际应用中比仅允许选择一个维数的模型更精确。

4 结论

本文提出的基于混合概率 PCA 模型的高光谱图像本征维数确定方法，能在已知高光谱图像地物类别数的情况下有效实现降维，实验表明该方法能准确选择数据的本征维数，得到数据在概率意义下的精确降维表征；同时该方法还可推广至不同维数模型的软混合模型，比仅允许选择一个维数的模型效果更好。

致谢：在此对提供数据的中科院技术物理研究所表示感谢。

参考文献

1 Landgrebe D. Information Extraction Principles and Methods for Multispectral and Hyperspectral Image Data[M]//Chen C H. Information Processing for Remote Sensing. New Jersey: World Scientific Publishing Co., 2000.

2 Minka T P. Automatic Choice of Dimensionality for PCA[C]//Leen T K, Dietterich T G, Trep V. Advances in Neural Information Processing Systems. MIT Press, 2001: 598-604.

3 Tipping M E, Bishop C M. Probabilistic Principal Component Analysis[J]. Journal of the Royal Statistical Society, 1999, 61(3): 611-622.

4 Tipping M E, Bishop C M. Mixtures of Probabilistic Principal Component Analysis[J]. Neural Computation, 1999, 11(2): 443-482.

5 Figueiredo M A T, Leitao J M N, Jain A K. On Fitting Mixture Models[M]//Hancock E, Pellilo M. Energy Minimization Methods in Computer Vision and Pattern Recognition. New York: Springer-Verlag, 1999: 54-69.

6 Figueiredo M A T, Jain A K. Unsupervised Learning of Finite Mixture Models[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2002, 24(3): 381-396.

7 Dempster A P, Laird N M, Rubin D B. Maximum-likelihood from Incomplete Data via the EM Algorithm[J]. J. Royal Stat. Soc. Ser. B, 1977, 39(1): 1-38.

（上接第 203 页）

同类抗原时，可以快速生成与之对应的抗体。算法中的交叉操作增强了抗原识别、记忆功能和调节功能。因此采用本文算法解决 QoS 多播路由问题可达到更高的效率，仿真结果也证实了这一点。

参考文献

1 孙宝林, 李腊元. 一种基于遗传算法的多约束 QoS 多播路由优化算法[J]. 计算机工程与应用, 2003, 39(30): 1-3.

2 许　毅, 李腊元. 基于蚁群算法的 QoS 多播路由优化算法[J]. 计算机应用研究, 2005, 22(1): 183-185.

3 Dorigo M. Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem[J]. IEEE Trans. on Evolutionary Computation, 1997, 1(1): 53-66.

4 Dicaro G, Dorigo M. Ant-net: Distributed Stigmergetic Control for Communications Networks[J]. Journal of Artificial Intelligence Research, 1998, 9(2): 317-365.

5 Jerne N K. Towards a Network Theory of the Immune System[J]. Annual Immunology, 1974: 373-389.

6 莫宏伟. 人工免疫系统原理与应用[M]. 哈尔滨: 哈尔滨工业大学出版社, 2002: 187-236.