一种模型选择优化准则及其在高光谱图像非监督分类中的应用

吴 昊,郁文贤,匡纲要,李智勇
(国防科技大学电子科学与工程学院,湖南长沙 410073)

摘 要: 选择合适的类别数是非监督分类中的一个关键问题.针对采用高斯混合建模的高光谱图像非监督分类问题,该文提出了一种基于主成分分析(PCA)的最小描述长度(MDL)型模型选择准则(文中简称为 PMDL)来确定分类类别数,即根据 PCA 变换后保留的各主成分表达的数据方差不同而应具有不同的编码长度这一事实,在计算描述长度时对各维进行加权.分类过程中,论文采用期望最大化(Expectation Maximization)算法在合并的策略下对 PCA 变换后的数据求解混合模型,并应用所提出的准则进行模型选择从而确定待分类的类别数.仿真数据实验证实了新准则的有效性和优良的性能,并采用真实数据对该准则和整个算法进行了验证.

关键词: 非监督分类; 高斯混合模型; 期望最大化算法; 主成分分析; 最小描述长度准则
中图分类号: TN958 **文献标识码**: A **文章编号**: 0372-2112 (2003) 12A-2154-04

An Unsupervised Classification Method Based on a Model Selection Criterion for Hyperspectral Data

WU Hao, YU Wen-xian, KUANG Gang-yao, Li Zhi-yong
(School of Electronic Science and Engineering, National University of Defense Technology, Changsha, Hunan 410073, China)

Abstract: A new principal component analysis (PCA)-based minimum description length (MDL) type criterion (termed PMDL) is proposed in this paper to solve the key problem of class number selection in unsupervised classification. It is based on the fact that data of different dimensions after PCA transform should be encoded with different code length, as they represent different amount of data variance. We perform unsupervised classification to hyperspectral image by Gaussian mixture modeling, estimate the parameters of the mixture model using the expectation maximization (EM) algorithm in merged operations to data after PCA linear projection, and select the number of components according to the proposed criterion. Experiments on a set of synthetic data verify the new criterion. The whole algorithm performs quite effectively and gives proper class number without any prior information applied to real data.

Key words: unsupervised classification; gaussian mixture model; expectation maximization algorithm; principal component analysis; principle of minimum description length

1 引言

高光谱图像非监督分类可以揭示数据的固有结构,为监督分类、检测等处理提供必要的信息.本文采用基于高斯混合建模的分类方法对高光谱图像进行非监督分类,并运用了能有效求解混合模型的 EM 算法.混合模型是一种灵活有力的建模工具,在模式识别中混合模型优于一般的非监督学习(聚类)方法[1].然而如何确定模型的最优混合分量数目,在混合模型的应用中是一个中心问题.一旦确定了混合成分数,即可解决类别数的确定这一非监督分类中的关键问题.

俭约原则(Principle of parsimony)或奥坎氏简化论(Occam's razor)指导了数据分析和统计建模过程,是模型选择的核心.为实现俭约原则,必须将观测数据对一个模型的"俭约性"进行量化,根据该测度从各候选值中搜索出能提供数据良好匹配的简洁模型[2].Rissanen 将该思想提炼成 MDL 准则[3]:"选择给出数据最短描述的模型".其中数据的总描述长度由数据模型的描述长度,以及观测数据用该模型进行编码的描述长度两部分组成.模型的简洁指容易描述,良好的匹配指模型捕捉或描述出了数据中的重要特征.MDL 是一种简单有效的模型选择方法,效果在同类型的准则中最好[9].然而在高光谱数据的应用中,经常采用的变换等预处理对数据的结构会产生影响,统计模型选择准则也应进行相应的调整.本文即是针对高光谱数据分类的应用背景,根据降维后数据的特性提出了一种基于 PCA 的 MDL 型准则,简称为 PMDL.

高光谱数据具有较高的光谱分辨率,产生的波段达数十至数百个,相邻波段间存在着较大的相关性,直接对原始数据建模会使求解混合模型的过程非常复杂,因而降维是必要的.本文首先应用 PCA 线性投影对原始数据进行光谱维降维,然

收稿日期:2003-07-02;修回日期:2003-12-13

后对变换后的数据进行高斯混合建模,对高维空间来说,由于正态假设在其投影子空间比在全维空间中更合理[4],因此 PCA 变换后的数据更符合高斯混合模型假设,可以用 EM 算法来有效求解.同时降维后维数上大幅度地减少使得 EM 算法计算量大大减少;又由于 PCA 变换后数据被去相关,EM 算法求解参数时高斯密度函数被简化,整个算法速度得到了很大提高.因此 PCA 变换不仅使数据更符合高斯混合模型假设,而且简化了模型求解的过程[5].然而变换后各主成分分反映原始数据的能力是不同的[6],在运用 MDL 准则时,由于不同维具有不同的重要性,本文分别用各主成分对原始数据的贡献率对各维的描述长度进行了加权,由此得到 PMDL 准则.在合并的策略下,根据所提出的准则从一系列模型中选择出最佳模型,从而确定了类别数.最后根据得到的模型将每个象点归至各类簇.仿真实验证实,本文提出的 PMDL 准则更适合 PCA 变换后数据的模型选择;将整个算法用于真实高光谱数据,能自动选择分类类别数,并取得了理想的分类结果.

2　理论分析

2.1　高斯混合模型和 EM 算法

对于观测数据集 $X=\{x_1,x_2,\cdots,x_N\}$ 中的单个采样 x_i,其高斯混合模型的概率密度分布函数为:

$$P(x_i|\Theta)=\sum_{k=1}^{K}\alpha_k p_k(x_i|\theta_k) \tag{1}$$

其中,$\Theta=(\alpha_1,\alpha_2,\cdots,\alpha_K,\theta_1,\theta_2,\cdots,\theta_k)$,为各混合成分的参数矢量;$\theta_k=(\mu_k,\sum_k)$,是高斯分布的参数,即均值矢量和协方差矩阵;$\alpha_k$ 是混合系数,表示各成分的先验概率;$p_k(x_i|\theta_k)=\dfrac{\exp[-\frac{1}{2}(x_i-\mu_k)^T\sum_k^{-1}(x_i-\mu_k)]}{(2\pi)^{d/2}|\sum_k|^{1/2}}$ 为各成分的高斯分布密度函数;K 和 d 分别为混合成分数和数据维数.

EM 算法能有效地求解高斯混合模型,它是一种从"不完全"的数据集中求解模型分布参数的极大似然估计的方法[7].这里 EM 算法将样本视为不完全数据,缺失的数据为样本类别标记的集合.

用 EM 算法求解高斯混合模型参数的迭代公式为:

$$\begin{cases}\alpha_k^{t+1}=\dfrac{1}{N}\sum_{i=1}^{N}p(k|x_i,\Theta)\\ \mu_k^{t+1}=\dfrac{\sum_{i=1}^{N}x_i p(k|x_i,\Theta)}{\sum_{i=1}^{N}p(k|x_i,\Theta)}\\ \sigma_k^{t+1}=\dfrac{\sum_{i=1}^{N}p(k|x_i,\Theta)(x_i-\mu_k^{t+1})(x_i-\mu_k^{t+1})^T}{\sum_{i=1}^{N}x_i p(k|x_i,\Theta)}\end{cases} \tag{2}$$

其中 N 为数据采样数目,上标 t 表示迭代次数,$p(k|x_i,\Theta)=\dfrac{\alpha_k p_k(x_i|\theta_k)}{\sum_{j=1}^{K}\alpha_j p_j(x_i|\theta_j)}$ 为单个采样 x_i 当前的迭代阶段属于第 k 类的后验概率.文献[8]证明了 EM 算法的收敛性,但 EM 算法只能保证收敛到局部极值.

该算法每一次迭代的计算复杂度为 $O_1(d^2KN)+O_2(K)$,其中 d 为数据维数,K 为类别数,N 为数据点数,O_1 为算术运算,O_2 为指数运算.可以看到其计算量与数据维数成平方关系,因此降维在这里尤为重要.

2.2　模型选择

模型选择方法从算法角度可以分为两类[9]:基于 EM 算法的技术以及随机技术.随机技术包括 Markov 链 Monte Carlo (MCMC) 采样,基于重采样的方法,以及交叉确认等方法,这类技术的计算量比 EM 大得多,不适于模式识别应用,因此主要考虑第一种技术.

基于 EM 算法的技术的基本思路是,对每个固定的混合成分数 k 应用 EM 算法得到一系列的参数估计,最小化某代价函数的 k 即为估计值.目前已经提出了多种准则来确定该代价函数.一类准则是使用 Bayes 因子(正确 Bayes 模型选择准则)的近似形式,例如基于证据的 Bayes 准则(EBB),近似的证据权值(AWE),以及 Schwarz 的 Bayes 推论准则(BIC).另一类是基于描述复杂性的方法,例如最小消息长度(MML)准则,Akaike 的信息准则(AIC),以及 Bezdek 的划分参数(PC)等方法.虽然是从不同的框架中推导出来的,BIC 与 MDL 在求解混合模型类别数的情况下是一致的.实验表明,EBB,MDL/BIC,以及 MML 效果相当;而 MDL/BIC 比 AWE 对真实的 Bayes 因子近似更好,并且 AIC 和 PC 准则的效果差于 MML 和 MDL/BIC.因此可以说,EBB,MDL/BIC 或 MML 效果是最好的,其中 MDL/BIC 的计算最为简单.本文 PMDL 准则即是在 MDL 的基础上针对我们的应用背景提出的.

MDL 准则可以由最小消息长度(MML)准则推导出来[10],MML 准则为:

$$\hat{\theta}=\arg\min_{\theta}\{-\log p(\theta)-\log p(y|\theta)+\frac{1}{2}\log|I(\theta)|+\frac{N(k)}{2}(1+\log K_{N(k)})\} \tag{3}$$

其中,$I(\theta)\equiv -E[\dfrac{\partial^2}{\partial\theta^2}\log p(y|\theta)]$ 为 Fisher 信息阵,$|I(\theta)|$ 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).约去 $-\log p(\theta)$、$\log|I^{(1)}(\theta)|$ 和 $\frac{N(k)}{2}(1+\log K_{N(k)})$ 几项,并用 $\log p(y|\hat{\theta})$ 近似 $\log p(y|\theta)$,得到 MDL 准则下的代价函数为:

$$L(\hat{\theta},y)=-\log p(y|\hat{\theta})+\frac{N(k)}{2}\log n \tag{4}$$

在求解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)[9],因此估计混合模型中第 m 个模型中参数的 Fisher 信息应该为

$$I(\theta_m)=n\alpha_m I^{(1)}(\theta_m) \tag{5}$$

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.则用于混合模型的 MDL 准则(MMDL)的代价函数为:

$$L(\hat{\theta},y)=-\log p(y|\hat{\theta})+\frac{N(k)}{2}\log n+\frac{N(1)}{2}\sum_{m=1}^{k}\log\alpha_m \tag{6}$$

我们提出了针对 PCA 降维后数据模型选择的 PMDL 准则.原始数据经过 PCA 变换后,数据的不同维反映原始数据的能力是不同的,因此在求解最小描述长度时,需要考虑到不

同维的差别. 由此在计算 Fisher 信息量时用各主成分的贡献率进行了加权. 另外, PCA 变换后数据各维不相关, 协方差阵为对角阵, 此时独立变量的个数变为: $N(k) = (k-1) + 2kd$

PCA 变换矩阵是由原始数据协方差矩阵的本征向量构成的, 各本征值的大小对应了变换后各主成分的方差大小, 反映了原始数据的散布情况. 注意到数据各维不相关, Fisher 信息阵为分块对角阵, 考察第 m 个模型的 Fisher 信息阵, 其中的每个分块为各维单一高斯分布对应的 Fisher 信息阵. 我们用各主成分的贡献率 λ_i'[6]对其所对应的 Fisher 信息阵进行加权. 则对应第 m 个模型单个观测的 Fisher 信息阵为:

$$I^{(1)}(\theta_m) = block\text{-}diag\{\lambda_1' I_1^{(1)}(\theta_m) \cdot \lambda_2' I_2^{(1)}(\theta_m) \cdots \lambda_d' I_d^{(1)}(\theta_m)\} \quad (7)$$

其中贡献率 $\lambda_i' = \lambda_i / \sum_{r=1}^{d}\lambda_r$, λ_i 为本征值, d 为数据维数, $I_i^{(1)}(\theta)$, $i = 1, \cdots, d$ 为对应第 m 个模型单个观测各维参数的 Fisher 信息阵.

结合式(5), 按照推导标准 MDL 准则的方式, 同样约去 $-\log p(\theta)$、$\log|I^{(1)}(\theta)|$ 和 $\frac{N(k)}{2}(1+\log K_{N(k)})$ 几项, 由此得到 PMDL 准则下的代价函数:

$$L(\theta, y) = -\log p(y|\theta) + \frac{k-1}{2}\log n + \sum_{m=1}^{k}\sum_{i=1}^{d}\log(n\alpha_m\lambda_i') = -\log p(y|\theta) + \frac{N(k)}{2}\log n + d\sum_{m=1}^{k}\log\alpha_m + k\sum_{i=1}^{d}\log\lambda_i' \quad (8)$$

3 算法实现

算法首先应用 PCA 线性投影对数据进行光谱维降维, 变换后前几个主成分形成的特征数据包含了大部分的谱带间差别信息, 因此大大减少了数据量. 实验发现高光谱数据波段间的相关性较强, 选定待保留主成分的累积贡献率[6]门限为 98 %, 即可满足保留信息分量、消除噪声分量的要求.

然后对变换后的数据进行高斯混合建模, 在合并的策略下估计出多个模型: 指定一个最大簇数 K_{max}, 用 EM 算法估计出各个簇的分布(均值、方差), 合并当前聚类模型中最近的两个聚类中心, 得到的新的中心作为下一阶段的初始估计, 如此循环直到类别数减少到 2. 由于 EM 算法只能收敛到局部极值, 这里采用随机选取初始聚类中心的 K-Mean 算法估计一个粗略聚类, 作为 EM 算法的初始估计, 虽然这并不能保证 EM 收敛到全局最优, 但已大大加快了 EM 的收敛速度. 然后进行下一步的模型选择.

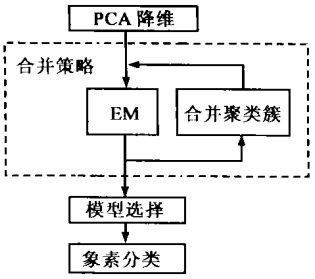


图 1 算法流程

根据本文提出的 PMDL 准则, 计算出每个 EM 收敛后的描述长度值, 将该值视为聚类簇数 k 的函数, 作出函数曲线. 在应用模型选择准则时, 由于 EM 仅保证收敛到局部极值, 即计算出的每个描述长度只能代表这个聚类簇数对应的这类模型的一个局部最优收敛结果的好坏, 不能代表这类模型的整体最优收敛结果的好坏, 因此分类过程中没有简单地取其最小值对应的模型, 而是选择了描述长度的所有局部极小值中的最小值所对应的模型. 大量实验显示, 准则 MDL、MMDL, 及本文提出的准则, 由于受局部收敛性的影响, 得到的准则曲线往往是类别数越大准则值越小, 选取曲线局部极小值中的最小值能合理地解决该问题.

最后根据求解出的模型用最小欧氏距离法将每个象点归至各类别. 整个算法的流程如图 1 所示.

4 实验结果

我们先用合成数据来验证所提出的 PMDL 准则的有效性, 然后运用整个算法对 OMIS 真实数据进行分类.

4.1 准则比较

参照文献[9]我们随机生成了包含 8 个成分的高斯混合数据, 数据维数为 3. 每个成分的均值为:

$$\mu_1 = \delta[0\ 0\ 0]^T, \mu_2 = \delta[1\ 0\ 0]^T, \mu_2 = \delta[0\ 1\ 0]^T, \cdots$$
$$\mu_7 = \delta[0\ 1\ 1]^T, \mu_8 = \delta[1\ 1\ 1]^T$$

各维的标准差为 1, δ 确定了类间距, 三次实验分别取 δ 为 4, 3.5, 3. 直观地讲, 数据是以三维空间中边长为 δ 的立方体的 8 个顶点为均值生成的高斯混合. 每次实验我们都生成了 50 个数据集, 每个数据集含 1200 个样本.

对以上数据运用 PCA 变换, 然后对变换数据进行高斯混合建模. 为了在相同的条件下比较各准则, 而排除分类算法局部收敛等其它因素的影响, 将最大类别数 K_{max} 取为 8, 初始聚类中心就取为各类簇的均值. 将我们提出的 PMDL 准则与标准 MDL, 以及 MMDL[9]进行比较, 图 2 显示了根据各准则选择的成分数的累积结果. 可以看到本文提出的模型选择准则的准确率最高.

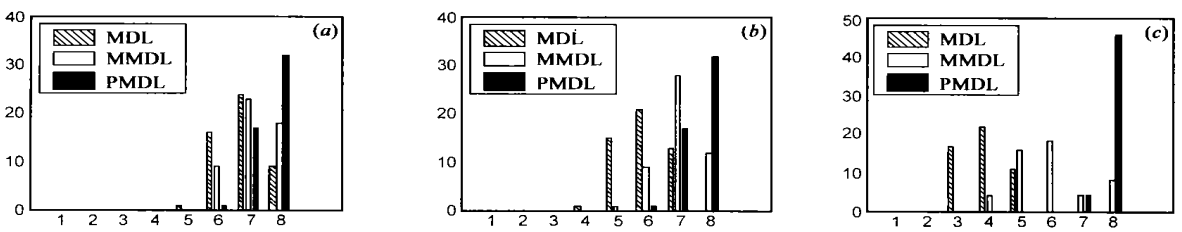


图 2 δ 分别取 4(a), 3.5(b), 3(c) 时, 依照各准则进行模型选择的统计结果(k 的真实值为 8). 三种准则依次为 MDL 准则, MMDL 准则以及本文所提出的 PMDL 准则

4.2 PHI 真实数据分类

我们采用的是 PHI 数据,PHI(Pushbroom Hyperspectral Imager)是中国科学院上海技物所研制生产的推帚式 CCD 面阵成像光谱仪,光谱范围为可见光-近红外波段(0.40～0.95μm),可选波段 244.数据含 80 个波段,图像尺寸为 150 × 160.为加快运算速度,我们对建模数据进行了抽样,抽样前对数据进行高斯滤波,以保证抽样数据仍能很好地表征数据的主要分布,我们将图像 x 和 y 方向的采样间隔都取为 4,相应地,高斯滤波窗口尺寸取为 4 ×4,一系列实验显示该尺度下高斯滤波后的采样并不影响分类精度.

将 PCA 累积贡献率门限选为 98%,由于场景中的地物种类不多,我们将合并算法的最大类别数 K_{max} 设为 12.图 3 显示了 PCA 变换得到的前 3 个主成分的结果,经过 PCA 变换之后,数据从原始的 80 个波段降为 3 维,大大减少了分类算法的计算量.

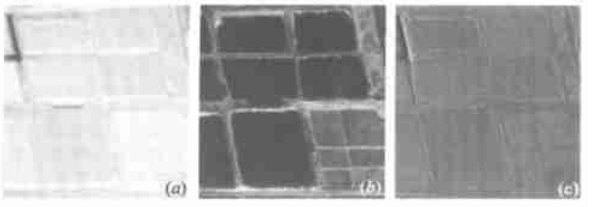
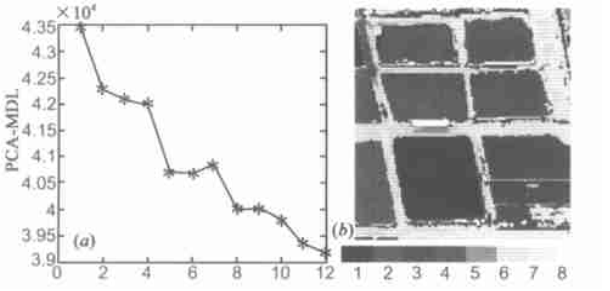


图 3　PCA 运算结果,算法根据信息量门限保留了 3 个主成分

不同于仿真实验,EM 算法对真实数据更容易产生收敛到局部极值的情况,这里选择描述长度的所有局部极小值中的最小值所对应的模型.实验选择的混合模型数为 8,它对应了曲线局部极小值中的最小值,即数据被分为 8 类,如图 4,视觉判断与用 ENVI 软件进行光谱分析的结果表明,分类结果准确并且足够揭示场景中的基本地物类别.



(*a*)针对 PCA 的 MDL 准则曲线(*b*)最终分类结果和类别颜色标记

图 4　准则曲线和分类结果

5　结论

本文研究了高光谱图像的非监督分类问题,为很好地揭示数据的结构,选择合理的类别数,我们根据 PCA 降维后数据的性质提出了一种对不同维进行加权编码的 PMDL 准则,在合并策略下自适应地选择出了混合模型的最优分量数.实验表明,与标准 MDL 以及另一种成功的混合模型选择准则相比,该准则对 PCA 降维数据的模型选择效果更好;整个分类算法能自动确定最优类别数,得到数据最紧凑的表示,并取得了理想的分类结果.

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] A K Jain, R Dubes. Algorithms for Clustering Data [M]. New Jersey: Prentice Hall, 1988.

[2] Mark H Hansen, Bin Yu. Model selection and the principle of minimum description length [J]. Journal of the American Statistical Association, 2001, 96: 746 - 774.

[3] J Rissanen. Modeling by shortest data description [J]. Automatica, 1978, 14: 165 - 471.

[4] Landgrebe D. Information extraction principles and methods for multispectral and hyperspectral image data[A]. Chen C H, Information Processing for Remote Sensing[M]. New Jersey: the World Scientific Publishing Co, 2000.

[5] Wu Hao, et al. An unsupervised classification method for hyperspectral image combining PCA and gaussian mixture model [A]. Proc. of the 3rd Inter. Symposium on MIPPR, SPIE Vol. 5286 [C]. Beijing, MIPPR, 2003. 729 - 734.

[6] 吴翊,李永乐,胡庆军. 应用数理统计 [M]. 长沙:国防科技大学出版社,1995. 285 - 303.

[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 - 38.

[8] Redner R A, Walker H F. Mixture density, maximum likelihood and the EM algorithm [J]. SIAM Review, 1984, 26(2): 195 - 239.

[9] Mario A T Figueiredo, Jose M N Leitao, Anil K Jain. On fitting mixture models [J]. Hancock E, Pellilo M (Edt.), Energy Minimization Methods in Computer Vision and Pattern Recognition, Springer-Verlag, 1999: 54 - 69.

[10] Mario A T Figueiredo, Anil K Jain. Unsupervised learning of finite mixture models [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2002, PAMI-24: 381 - 396.

作者简介:

吴　昊　女,1976 年 5 月出生于湖北武汉,2000 年毕业于国防科技大学电子科学与工程学院,获硕士学位,现为国防科技大学电子科学与工程学院博士研究生,主要从事遥感信息处理,高光谱图像地物分类等方面的研究工作.

郁文贤　男,1964 年 10 月出生于上海,博士,教授,博士生导师,国家 863 专家,主要从事雷达目标识别、神经元网络、信息融合以及遥感应用等,在国内外学术刊物上发表近百篇学术论文.

匡纲要　男,1966 年 7 月出生于湖南衡阳,博士,副教授,硕士生导师,主要从事雷达信号处理、SAR 图像解译、多光谱图像处理等方面的研究工作,在国内外学术刊物上发表 50 余篇学术论文.

李智勇　男,1975 年 3 月出生于辽宁阜新,2000 年毕业于国防科技大学电子科学与工程学院,获硕士学位,现为国防科技大学电子科学与工程学院博士研究生,主要从事遥感信息处理,高光谱图像目标检测等方面的研究工作.