以地物识别和分类为目标的高光谱数据挖掘

王晋年 张 兵 刘建贵 童庆禧 郑兰芬
(中国科学院遥感应用研究所, 北京 100101)

摘 要 高光谱信息挖掘技术是高光谱数据应用延拓与深入的重要环节,其核心在于光谱信息的挖掘。基于高光谱遥感信息的特点,探讨分析以地物识别与分类为目标的高光谱数据挖掘技术,包括基于模式识别的高光谱信息挖掘技术,基于光谱波形特征的挖掘技术,以及亚象元光谱信息挖掘。进一步发展有待于利用挖掘与知识发现的概念与模型,基于地物光谱特征及光谱数据库、知识库,在高光谱超维特征空间充分挖掘地物的光谱信息,以达到地物识别目的,并针对不同的应用目标与领域发展相应的光谱信息挖掘模型与技术。

关键词 高光谱遥感 数据挖掘 超维特征空间 地物识别

0 引 言

高光谱分辨率遥感技术的发展是本世纪末的最后两个 10 年中人类在对地观测方面所取得的重大技术突破之一,是当前乃至下一世纪初的遥感前沿技术。通过高光谱成像所获取的地球表面的图象包含了丰富的空间、辐射和光谱三重信息。进入 90 年代后期,伴随着高光谱遥感应用的一系列基本问题,如高光谱成像信息的定标和定量化、成像光谱图象信息可视化及多维表达、图象-光谱变换、大数据量信息处理等的解决,高光谱遥感已由实验研究阶段逐步转向实际应用阶段。而作为高光谱遥感应用这一热点中的重点就是高光谱数据信息挖掘技术的提高和与之紧密相连的应用领域的扩展。

20 世纪末,面对数字地球和数据爆炸的挑战,数据挖掘成为人们从海量数据库中提取模式和知识的重要手段。本文拟在高光谱图象数据处理和信息提取领域引入数据挖掘与知识发现的概念、模式和方法,探讨针对高光谱图象数据的信息挖掘模型和方法。

1 概 念

1.1 高光谱数据的特点

高光谱遥感数据最主要的特点是:将传统的图象维与光谱维信息融合为一体,在获取地表空间图象的同时,得到每个地物的连续光谱信息,从而实现依据地物光谱特征的地物成份信息反演与地物识别。

高光谱数据是一个光谱图象的立方体(见图 1),它由以下 3 部分组成:

(1) 空间图象维:在空间图象维,高光谱数据与一般的图象相似。一般的遥感图象模式识别算法是适用的信息挖掘技术。

(2) 光谱维:从高光谱图象的每一个像元可以获得一个"连续"的光谱曲线,基于光谱数据库的"光谱匹配"技术可以实现识别地物的目的。同时大多数地物具有典型的光谱波形特征,尤其是光谱吸收特征与地物化学成分密切相关,对光谱吸收特征参数(吸收波长位置、吸收深度、吸收宽度)的提取将成为高光谱信息挖掘的主要方面。

(3) 特征空间维:高光谱图象提供一个超维特征空间,对高光谱信息挖掘需要深切了解地物在高光谱数据形成的 n 维特征空间中分布的特点与行为,研究发现:高光谱的高维空间是相当空的,数据分布不均匀,且趋向于集中在超维立方体空间的角端,典型数据的差异性,可以映射到一系列低维的子空间,因此迫切需要发展有效的特征提取算法去发现保持重要差异性的低维子空间,从而有效地实现信息挖掘。

就数据统计特征而言,在高光谱数据分析中,二次变量(协方差)扮演极为重要的角色,因此在分类与信息挖掘中需同时兼顾一次变量(例如均值)和二次变量,统计距离测量将主要采用 Bhattacharyya

收稿日期:1999-11-03

距离算法。由于高光谱特征空间的不均匀性散布,在分类与信息挖掘中要提高分类识别精度,需要很多的训练样本及更丰富的光谱数据库、知识库,以便准确地确定地物在高光谱超维空间的密度分布函数。

1.2 高光谱数据与信息挖掘

在高光谱遥感技术发展的初期,所有涉及高光谱遥感应用的人无不为这种超多波段和海量的数据所困惑。近年来,随着计算机技术的突飞猛进和诸如 ENVI 以及 TNTmips Hyperspectral Analysis 等专业化高光谱图象处理系统的问世,都对高光谱遥感应用产生了很大的促进作用。尤其是大量的专用算法和模型被开发了出来,它们都与数据挖掘理论和技术紧密相关。数据挖掘中的分类、回归、聚类、概括、依赖模式、变化和偏离检测等概念被有意识或无意识地大量采用。其中比较著名的方法有:基于决策树的分类,人工神经网络分析、贝叶斯概率网络学习、解决不精确和不确定知识问题的粗糙(模糊?)集方法等。可以看出,数据挖掘并不是某一种具体的全新方法,数据挖掘的许多方法在高光谱遥感分析中早已广泛应用。数据挖掘与一般的数据分析概念相比,它更强调基于大量事例的统计分析、挖掘方法的效率、挖掘工具与数据库的集成、挖掘过程的模式和结构等。因此结合高光谱数据的特点和相关概念,如高光谱图象立方体、光谱数据库、光谱波型匹配、光谱特征吸收、图象光谱波形分析等,都非常适合数据挖掘和信息发现技术的应用,在这方面有很多成功的事例。

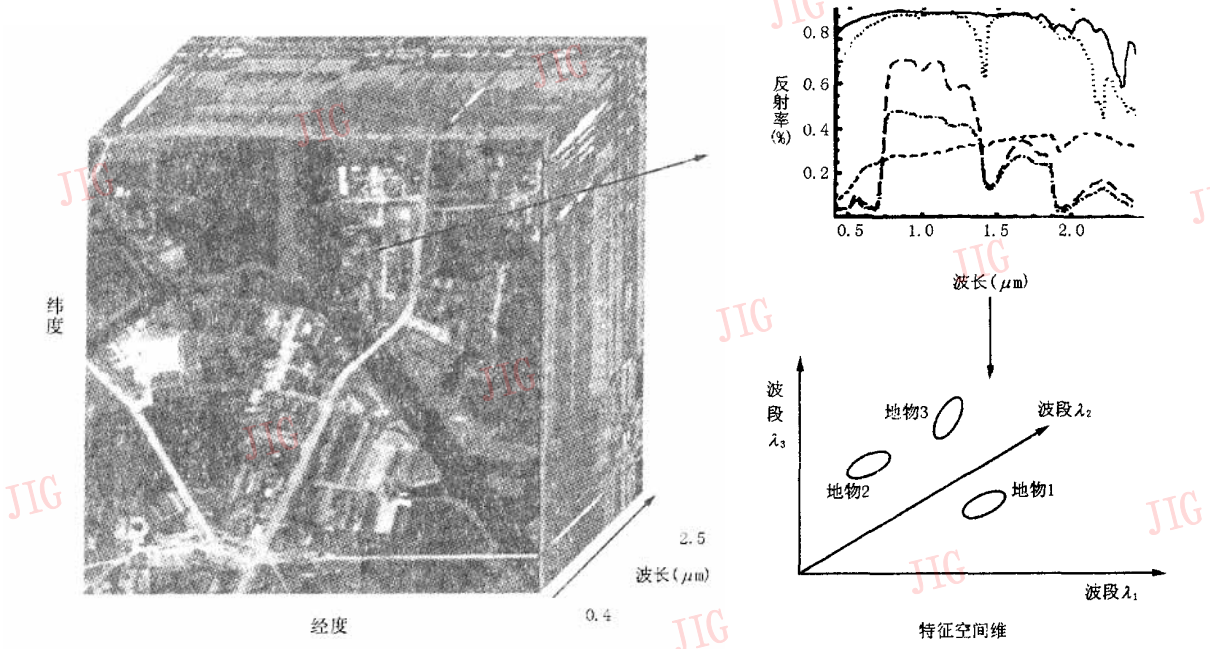

图 1 高光谱数据立方体,光谱信息及 N 维特征空间

2 技 术

从模式识别中发展起来的分类技术在遥感图象中获得了广泛的应用,例如统计模式识别、神经元网络、模糊模式识别等即为几种具有代表性的分类方法。对高光谱图象数据来说,光谱信息挖掘显得非常重要。一方面人们已经获得了大量的实测地面光谱数据,对目标对象的光谱响应积累了越来越多的第一手资料,从中展开信息挖掘,能够帮助人们了解光谱响应特征,获得对分类具有指导性的光谱知识;另一方面,高光谱技术从遥感的角度提供了大尺度获取地面光谱数据的手段,为人们宏观分类识别地物提供了基础。将这两个方面结合起来最大程度地发挥高光谱技术的优势是一个关键性的问题。采用光谱数据库中的标准光谱响应曲线,通过光谱匹配识别对象,牵涉到从图象中选择最终光谱单元、最终光谱单元的识别以及对图象的匹配分类识别等方面。

2.1 基于模式识别的信息挖掘技术

2.1.1 统计模式识别

统计决策分类是模式识别的基本理论之一。按照距离来度量模式的相似性的几何分类法和基于 Bayes 准则的最大似然法是统计模式识别的两种方法。

(1) 几何分类法——最小距离分类

相似性度量的基本假设是：如果两个模式的特征或其简单的组成部分仅有微小的差别，称这两个模式是相似的，微小差别是指距离在一个阈值之下。最简单的方法是以各类训练样本点的集合所构成的区域表示各类决策区，并以点距离作为样本相似度度量的主要依据。这种方法适用于要识别的每一个类都有一个代表向量(均值向量)的情况。先求出未知向量到各代表向量的距离，通过比较将其归为距离最小的一类。一般用广义距离来表述“距离”。广义距离有以下属性：

$D(x,x)=0 \qquad D(x,y)\geq 0$

$D(x,y)=D(y,x) \quad D(x,y)\leq D(x,z)+D(z,y)$

可以根据需要设计出满足上述规则的距离，如明氏距离为：

$$d(\boldsymbol{x},\boldsymbol{y})=\left[\sum_{i=1}^{n}|\boldsymbol{x}_i-\boldsymbol{y}_i|^{\lambda}\right]^{1+\lambda}$$

当 $q=1$ 时明氏距离成为曼氏距离，$q=2$ 时，即为欧氏距离。马氏距离考虑了样本的统计特性，形式为：

$$D^2=(\boldsymbol{x}-\boldsymbol{m})^T\sum{}^{-1}(\boldsymbol{x}-\boldsymbol{m})$$

其中 $\boldsymbol{x},\boldsymbol{m}$ 为 n 维特征向量，$\sum^{-1}$ 为协方差矩阵的逆矩阵。马氏距离考虑了各特征参数的相关性，因而比明氏距离更为合理。当各特征间完全不相关，$\sum^{-1}=I$ 时，马氏距离即为欧氏距离。

(2) Bayes 准则——最大似然分类(MLC)

基于 Bayes 准则的判别函数是统计模式识别的参数方法，需要各类的先验概率 $P(\omega_i)$ 和条件概率密度函数 $P(\omega_i|x)$ 已知。$P(\omega_i)$ 通常根据各种先验知识给出或假设它们相等；$P(\omega_i|x)$ 则是首先确定其分布形式，然后利用训练样本估计其参数。一般假设为正态分布，或通过数学方法化为正态分布。其判别函数为：

$$D_i(X)=P(\omega_i)P(\omega_i|x), \qquad i=1,2,\cdots,m$$

若 $D_i(X)>D_j(X)$，$j=1,2,\cdots,m$，$j\neq i$ 则 X 为 ω_i 类。判别函数集有多种导出形式，如最大后验概率准则、最小风险判决准则、最小错误概率准则、最小最大准则、Neyman-Pearson 准则等，是依据不同的规则选择似然比的门限来实现的。这是目前比较成熟的一种分类方法，还在研究中。

(3) Bhattacharyya 距离分类

对于高光谱数据分类而言，除了一次统计变量(例如平均值)外，二次统计变量(协方差等)是分类与地物识别的重要依据。而 Bhattacharyya 距离同时兼顾一次与二次统计变量，因此在测度高光谱超维空间中两类统计距离时，Bhattacharyya 距离是最佳测度。Bhattacharyya 距离可表达为：

$$B=\frac{1}{8}[\mu_1-\mu_2]^T\left[\frac{\sum_1+\sum_2}{2}\right]^{-1}[\mu_1-\mu_2]+\frac{1}{2}Ln\frac{\left|\frac{1}{2}\left[\sum_1+\sum_2\right]\right|}{\sqrt{\left|\sum_1\right|\left|\sum_2\right|}}$$

其中 μ_i 是类别的平均矢量，$\sum_i$ 是类别的协方差矩阵。

2.1.2 神经元网络技术

由于传统统计分类的一些局限性，人们尝试用神经元网络模型来模拟人类对物体的识别机理，于是有关神经网络分类器的研究不断地进行并得到发展。人们发展了各种形式的网络模型和算法，如 Hopfield 网、Hamming 网、CG、单层感知器网、多层感知器网、Kohonen 组织算法等，用得最多的是反向传播算法。

神经元网络包含一个输入层，一个输出层，及一个或多个隐层。输入层结点数与参加分类的特征数相同，输出层结点数与最终类别数相同。而中间隐含层结点数则由实验来确定。以单隐含层的网络为例，其结点数应至少为输入层和输出层结点数中较大者的 2—3 倍。每个结点输入是下层结点输出的加权和：

$$net_j=\sum_i w_{ji}O_i$$

其中 O_i 为下层结点的输出；w_{ji} 是下层结点 i 与相邻上层结点 j 的互联权重，net_j 为该层结点的输入。j 结点通过一个非线性系统函数或驱动函数将 net_j 转化为其输出：

$$O_j=[1+\exp(-net_j+\theta)]^{-1}$$

或

$$O_j=m\tanh(knet_j)$$

其中 θ、m、k 等是由实验得出的常数值。在一次迭代中，求出 O_i 后与期望的输出相比较，根据误差修正权重 w_{ji}，再进入下一次迭代，直到误差达到某个阈值。根据误差按照下式修改权系数：

$$\Delta w_{ji}(n+1)=\eta(\delta_j O_y)+\alpha\Delta w_{ji}(n)$$

其中，$\Delta w_{ji}(n+1)$ 为 $n+1$ 次迭代时连接相邻两层结点 i,j 的加权值的变化；δ_j 为输出结点 j 的误差变化率；η 为训练速度；α 为动量项。要达到一定训练精度，往往需要很多次的迭代，这是非常费时的，然而网络训练一经完成，则可较快地应用于分类识别。神

经元网络具有以下优点:不需对原始类别做概率分布假设,不存在求解概率分布参数的问题,是一种无参分类器;输入与输出结点之间通过隐含层,结点之间通过权重来连接,因而这种方法可以将多种数据,如纹理信息、地形信息等,方便而有效地融合到分类中来,加强分类能力;输出结果的驱动函数是非线性的,因此系统也是非线性系统,这样可以在特征空间构造出分类界面比较复杂的子空间,这对非线性可分的特征子空间尤为有效。然而,训练参数如初始权重、收敛速度、对输入数据的预处理等,对分类都有重要的影响,表现在:网络结构的隐层数越多,结点数越多,即网络结构越复杂,越可以精确地对训练数据进行分类,但也使系统失去较好的概括性,对以后的分类精度有不利影响;为达到一定的分类精度,每个类别至少有 10—30 倍于波段数的训练样本点,而训练样本的选择有一定的困难;为使结点输入 net_j 的变化与输出 O_j 的变化有相近的百分比而不至于使驱动函数饱和、使训练在错误的水平上滞留,应对输入特征进行规范化预处理,使驱动函数保持在不饱和的状态;由于驱动函数的高度非线性,容易使网络陷入小输入变化引起大输出变化的不稳定状态。

2.1.3 模糊模式识别

遥感图象象元所描述的对象由于各种原因往往也具有模糊的特性。例如,混合象元如果从精确的角度出发不应当被划归为某一个类别。因此在遥感界也有大量的研究人员进行模糊分类的研究。

给定论域 U 上的一个模糊集合 F 是指:对于任意的 $x\in U$,确定了一个数 $\mu_F(x)$,$\mu_F(x)\in[0,1]$,其中 $\mu_F(x)$ 为 x 对 F 的隶属度函数。当 $\mu_F(x)\in\{0,1\}$ 时 F 退化成普通集合。隶属度函数 $\mu_F(x)>0$,则 x 就是模糊集合的一个元素,尽管由于隶属度不同它们对外界的作用不同。而 x 则既可归属于 F,也可归属于 F 的补集。

模糊集合中也定义了类似于普通集合的各种运算如相等、包含、并、交、余、差集等以及各种运算的性质,可以用来操作和使用模糊集合。模糊集合的核心是隶属度函数的确定,隶属度函数对模糊集合的应用效果有很大的影响。确定隶属度函数的过程与实际的应用背景有很强的关联性,没有通用的方法。几种常用的方法是:

(1) 模糊统计法

$$\mu_F(u_0)=\lim_{n\to\infty}(u_0\in F*\text{ 的次数})/n$$

其中 $F*$ 是与模糊集合相联系的普通集合。

(2) 二元对比排序

对论域 U 中的元素 x_i 按照某些特性在两两对比中建立比较值,然后在相对比较取值的基础上通过某些计算方法确定总体隶属度。

(3) 推理法

某些场合可以利用相应的数理知识计算出隶属度函数,然后在实践中检验与调整。

3 高光谱数据挖掘

地物覆盖由于化学成份差异形成可诊断的典型光谱吸收特征,这成为地物光谱识别的理论基础。

3.1 光谱吸收指数(SAI)

任一光谱吸收特征可由光谱吸收谷点 M 与光谱吸收两个肩部 S_1 和 S_2 组成。根据遥感图象光谱分辨率和中心波长位置,S_1,S_2,M 可以分别位于一个波段,也可以是几个波段的线性组合。吸收谷点 M 与两个肩端组成的“非吸收基线”的距离可以表征为光谱吸收深度(H),令:ρ_{s1}、λ_{s1}为吸收左肩端 S_1 的反射率和波长位置;ρ_M、λ_M 为吸收点的反射率和波长位置;ρ_{s2}、λ_{s2}为吸收右肩端 S_2 的反射率和波长位置。这样,吸收肩部的波长差即为吸收波段宽度:$W=\lambda_{s1}-\lambda_{s2}$,吸收的对称性参数 d 可表达为 $d=(\lambda_{s1}-\lambda_{s2})/w$,而吸收肩端反射率差为 $\Delta\rho_s=\rho_{s2}-\rho_{s1}$,则“非吸收基线”方程为:

$$W\cdot\rho-\Delta\rho_s\cdot\lambda=W\rho_{s1}-\Delta\rho_s\cdot\lambda_{s1}$$

它表达了无光谱吸收特征地物的光谱贡献与光谱行为。实际上“非吸收基线”为一曲面方程,将问题简化为直线,吸收位置的光谱值与相应基线值的比值的倒数可定义为光谱吸收指数(SAI),表达为:

$$SAI\ \frac{d\rho_{s1}+(1+d)\rho_{s2}}{\rho_m}$$

将 SAI 转化为单次散射反照率的函数,对于光谱识别、光谱混合分析具有重要意义。研究表明,光谱反射率 $\rho(\lambda)$不能直接线性混合,难于进行混合光谱分解与成份丰度反演,而平均单次散射反照率 $\bar{\omega}$ 则主要依赖于成份含量,而且可以线性混合。光谱吸收指数可以表达为:

$$SAI=(d\bar{\omega}_{s1}+(1-d)\bar{\omega}_{s2})/\bar{\omega}_m$$

$\bar{\omega}$ 的线性混合反演为:

$$\bar{\omega}=\sum_i\frac{m_i\bar{\omega}_i}{\delta_iD_i}\Big/\sum_i\frac{m_i}{\delta_iD_i}$$

其中 m_i 为类别 i 的百分含量;ω_i 为其单次散射反照率,σ_i 为密度,D_i 为粒度。获得一系列典型吸收特征

的 SAI 图象以后，可以用最小二乘法反演各种地物的光谱混合成份的含量。

3.2　导数光谱波形匹配技术与植被分类

传统植被指数可以表达为光谱反射率的一阶或者二阶导数乘以系数 K 的形式，不同研究者提出的植被指数可以认为是反映波形形态变化的反射光谱的 n 阶导数，而这种光谱 n 阶导数实质上表达了植被叶绿素、水、氮等生物化学元素吸收波形的变化，是这些吸收物质的丰度与状态的光谱指标。实际上植被光谱随波长变化是连续的过程，例如可以将归一化植被指数 $NDVI$ 这种离散形式变为连续的形式，在 $\Delta\lambda\to 0$ 时取极限：

$$NDVI = \frac{1}{2R(\lambda)} \cdot \frac{\mathrm{d}R(\lambda)}{\mathrm{d}\lambda}$$

同样，其它形式的植被指数也可变为连续的形式，即光谱导数与一系数的乘积。在鄱阳湖湿地植被遥感实验中，就采用了导数光谱波形匹配模型和光谱夹角填图的方法。

导数光谱波形分析能够部分消除大气效应、植被环境背景影响，能够反映植物的本质特征，被利用来提取植被生物化学信息。

3.3　二值编码匹配

对光谱库的查找和匹配过程必须是有效和快速的，而且，对成象光谱数据大程度的光谱冗余度来说，为实施匹配，全部光谱数据的原始形式可能并不必要。Goetz 提出了对光谱进行二值编码的建议，使得光谱可用简单的 0—1 序列来表述。一旦完成编码，就可利用基于最小汉明距离的算法来进行匹配识别。目前已经提出了一系列的二值编码方法，最简单的方法是：

$$h(n) = \begin{cases} 0, & \text{如果 } x(n) \le T \\ 1, & \text{如果 } x(n) > T \end{cases} \quad n = 1,2,\cdots,N$$

其中 $x(n)$ 是象元第 n 通道的亮度值，$h(n)$ 是其编码，T 是选定的门限值，一般选为光谱的平均亮度，这样每个象元灰度值变为 1 比特，象元光谱变为一个与波段数长度相同的编码序列。然而有时这种编码不能提供合理的光谱可分性，也不能保证测量光谱与数据库里的光谱相匹配，所以需要更复杂的编码方式。

(1) 分段编码：对编码方式的一个简单变形是将光谱通道分成几段进行二值编码，这种方法要求每段的边界在所有象元矢量都相同。为使编码更有效，段的选择可以根据光谱特征进行，例如在找到所有的吸收区域以后，边界可以根据吸收区域来选择。

(2) 多门限编码：采用多个门限进行编码可以加强编码光谱的描述性能。例如采用两个门限 T_a、T_b 可以将灰度划分为 3 个域：

$$h(n) = \begin{cases} 00, & \text{如果 } x(n) \le T_a \\ 01, & \text{如果 } T_a < x(n) \le T_b \\ 11, & \text{如果 } x(n) > T_a \end{cases} \quad n = 1,2\cdots,N$$

这样象元每个通道值编码为 2 位二进制数，象元的编码长度为通道数的两倍。事实上，两位码可以表达 4 个灰度范围，所以采用 3 个门限进行编码更加有效。

(3) 仅在一定波段进行编码：这个方法仅在最能区分不同地物覆盖类型的光谱区编码。如果不同的波段的光谱行为是由不同的物理特征所主宰，我们可以仅选择这些波段进行编码，这样既能达到良好的分类目的又能提高编码和匹配识别效率。

3.4　光谱角度匹配(SAM)

当模式类的分布呈扇状分布时，定义两矢量之间的广义夹角余弦为相似函数，即为用得较为广泛的广义夹角匹配模型。将象元 N 个波段的光谱响应作为 N 维空间中的矢量，则可通过计算它与最终光谱单元之间的广义夹角来表征其匹配程度：夹角越小，说明越相似。两矢量广义夹角余弦为：

$$\cos(a) = \frac{X \cdot Y}{\|X\|Y\|}$$

最终光谱单元可从光谱库中得来，也可直接由图象中通过选择训练区抽取出来。

3.5　光谱相似度测定

光谱匹配需要一个指标来衡量在整个测量的波长范围内光谱的相似程度，可以用相关系数进行测度。相关系数定义为：

$$r_{xy} = \frac{\sigma_{xy}^2}{\sigma_{xx}\sigma_{yy}} = \frac{\sum_i (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_i (x_i - \bar{x}_i)^2}\sqrt{\sum_i (y_i - \bar{y}_i)^2}}$$

其中 σ_{xy}^2 为协方差，σ_{xx}，σ_{yy} 为标准差。

也可以用两光谱曲线的均方差 D 来定义这一个指标：

$$D = \left[\frac{1}{\lambda_b - \lambda_a}\int_{\lambda_a}^{\lambda_b} [S_1(\lambda) - S_2(\lambda)]^2 d_\lambda\right]^{1/2}$$

$S_1(\lambda)$、$S_2(\lambda)$ 为两个象元光谱。考虑到每类的光谱响应值的分布存在一个范围，我们可以采用规一化的

方法来给出这一指标，即夹角余弦：

$$\cos(\boldsymbol{S}_1,\boldsymbol{S}_2)=\int S_1(\lambda)S_2(\lambda)d_\lambda\Big/\left[\int S_1(\lambda)^2d_\lambda\right]^{1/2}\left[\int S_2(\lambda)^2d_\lambda\right]^{1/2}$$

3.6 基于光谱特征的分级掩模分类

决策树是多层分类器的一种方法，可以形成比较复杂的决策面，此时一个未知类别的象元可以采用一个或者几个决策函数逐一分级分类成某个特定的类别。决策树分类的模型要求各种类别之间具有内在的等级归属关系，可以按照其特性逐一细化成精细类别，通过加入决策函数一步一步地进行分类。利用传统的单层最大似然法分类器有如下缺点：① 仅可以采用一种可能的特征组合；② 每个被分类的象元都必须与所有类进行比较来产生分类结果；③ 分类所采用的特征子集按照使平均类间分离度最大的标准来选取，所以它不一定对某个特定的类别能达到最优分类。而决策树由于可以在不同的分类等级与节点上采用不同的特征子集，甚至可以采用不同的分类器，所以可以克服这些缺点，并且由于可以选用较少的特征，所以可以避免高维数据带来的系列问题，当然分类性能依赖于合适的决策树设计。而树设计要考虑的主要问题是确定如何在每个节点分离类别，如何控制类重叠，以及树由多少层构成。各种各样不同的决策树可以分成 4 个类别：从上到下，从下到上，混合类型及树增长等方法。

二叉树是一个特殊的决策树类型，它在每一个节点上只在两个类别之间进行选择分类，或者分成左子树，或者右子树。任何树都可以被变换为一个等效的二叉决策树。二叉决策被认为可以提高分类效果。

Jia 提出了一个渐近二叉分类树方法，在每个节点仅考虑一对类别的分离情况。该方法的优点在于：① 在多层决策树的结构设计上，其结果简单明确，完全由图象中的类别数 M 来确定，即总共具有 $M-1$ 层，$M(M-1)/2$ 个节点，在第 k 层具有 k 个节点，这样的结构在软件设计上易于实现。② 在两类分离问题上，由于该方法在每个节点仅考虑两个类别，而两类分离问题是模式识别中相对简单的，可以允许选择适合于这两类分离问题的特征，因此降低了数据维数，这对成象光谱数据是极其有利的。而且更为重要的是可以设计出比较简单的决策函数，例如线性决策面，这样做可以减少计算量。③ 在每个类别对之间的分离问题上，其优点是可以采用独特的分类算法，这一点对分类器的设计尤其重要，因为分类器的设计应当考虑到类别对之间的本质差异，而不同的对之间的差异特征不同，应当采取不同的算法。事实上，如果采用相同的算法、相同的特征到所有的节点上，它和一般的单层分类器并无差别，例如采用相同的波段组合以及最大似然法，两者的结果完全一样。总的来说，只处理两个类别提供了一个最优的分类环境。如果它们在这种情况下不能分离，则在单层分类器的条件下也不能分离。但是，可以看到，这种方法只有到最底层时所有的类别才被分离出来，而中间层并没有分出任何类别。在每个层的一个节点上，摈弃了一个类别最后得到该类的分布。而实际上，有些类别具有明显区别于其它类别的特征(如植被)，有可能经过一层达到分离；还有一些类别的光谱特征非常相似，在分离的某个阶段有可能一步将它们全部分离出去。因此如能考虑光谱特征，不但对于特征选择非常重要，而且能够设计出更为有效的分类器。

测量技术里的一个基本原理是，相对测量比绝对测量具有更高的精度，这使我们可以认为，在进行多光谱或者成象光谱数据分类时，通过相对比较来区分两个类别比直接从众多的类别中间识别出某一个类别要容易一些。这要求我们举出所有可能出现的类别，并采用比绝对分类识别更为有效的分类方法。分析的过程事实上是融合对被分析数据的先验知识。这些知识可能具有主观性，例如分析人员在特定的地学领域作为一个专家所拥有的知识，或者对所研究的特定区域具有详细的地面调查知识，或者对该地区以前所采集的一些定量化的数据。这些信息越多，从数据分析中得出的结论就应当越多。

城市地物及人工目标是熟悉而又复杂的研究对象。我们可以通过实地考察、先验知识以及大量地收集它们的光谱数据，分析得出该地区的地物大致类型以及它们的光谱响应特征，然后利用其光谱特点，通过相对比较的方法达到逐步分类识别的目的。所以我们采用一种分层分类的方法，在每个分类节点上采用相对分类技术，分成两个或者几个类别，然后，针对每个类别，产生一个模板图，在将该类别细分时，在该模板所确定的范围，采用其它特征将该子类进行细分直到所有类别被分离出来为止。这种方法建立在对图象类别与光谱特征的详细了解的基础上。在最后分离出所有类别以后，采用简单的复合技术将各种类别的分布统一到一张图上。

4　混合光谱信息与亚象元信息挖掘

由于空间采样间隔往往大于地面目标，因此象元的光谱测定往往是数种地物类别的混合效果，必须采用混合象元分解技术来估计每个象元中各种成分的比例。混合象元解混的主要问题是最终光谱单元的选择以及混合光谱分解模型的建立。

4.1　最终光谱单元的选择

最终光谱单元选择的具体过程如下：

(1) 对图象进行 MNF(Minimum Noise Fraction)变换——一种特殊的主成分变换。这种方法通过两个步骤的主成分变换来实现：第一步变换基于对噪声协方差矩阵的估计，对数据中的噪声去相关和归一化，产生一个图象序列，使得其中的噪声“白化”，即其中的噪声方差为 1、在序列之间是互不相关的；第二步是对第一步产生的图象序列实施标准的主成分变换。变换后各成分按照信噪比而不是方差从大到小的顺序来排列，这对寻找纯象元是重要的。该方法被用来确定数据的内在维数、从图象中分离噪声以及减少后续处理的计算量。因为 MNF 变换的特征是按照信噪比从大到小的顺序排列的，而噪声对纯象元的选择极为不利，所以在 MNF 变换的前几个特征上进行处理来获得纯象元。

(2) 产生象元纯度指数图象。象元纯度指数 PPI(Pixel Purity Index)是在多光谱、高光谱数据中寻找“光谱最纯”象元的方法，而这些象元的光谱一般是混合模型的最终单元。传感器测量的光谱辐射值是非负的，由这些离散的辐射光谱构成的向量位于多维空间中，在该空间中形成一个凸的多维区域。凸多维区域的边界对应于混合模型的最终光谱单元，其它位于内部的向量可以认为是由这些边界单元线性混合而成。因此，确定了边界象元以后，就找到了混合象元解混和图象分类中需要的最终单元。具体做法是，将 N 维散点图重复投影到单位随机向量上去，找出每次投影的极值象元，并对这些极值象元累计计数，产生一个象元纯度指数图，图中象元的值代表该象元在投影过程中为极值的次数。纯象元就是象元纯度指数图中数值较大的那些象元。

(3) 通过散点图确定最终光谱单元。从数据中抽出上一步找出的纯象元，作出它们在 N 维光谱空间中的散点图。相同地物的点在散点图中总是集中分布，由此可以选定训练样本。

4.2　混合光谱分解模型

混合光谱模型用来进行混合象元的分解，采用线性混合光谱模型以及一些地物的已知反射率值和图象 DN 值，也可用来进行反射率反演。有两种混合情形：宏观混合与微观混合。光子只与一种表面物质发生反射或吸收作用的情形属于宏观混合，这时总的反射辐射强度是各种物质反射辐射强度以各自出露表面积百分数为权重的加权和，这是光谱混合的线性模型。当光子与多于一个的表面物质发生作用时，称为微观混合，这时的混合模型是非线性的。表面反射率的混合光谱反演方法假定混合模型是线性的，而且多光谱或者高光谱数据中的所有光谱变化都可以用为数不多几个最终光谱单元来混合。最终光谱单元是一条光谱，图象的最终光谱单元是一个或者多个覆盖一定区域的、能够代表图象中一个独立物质类别的象元点的辐射光谱的编码，处于线性混合模型混合线的端点或者混合空间的角点。

5　讨　论

随着一系列航天高光谱器即将面世，例如美国 EOS 中的 MODIS，所谓新千年计划中的 EO－1，美国海军 HRST 中的 NEMO，Orbimage 的 Orbview－4 以及澳大利亚 ARIES。高光谱遥感将进入航天时代，高光谱遥感数据将成为遥感应用的主要信息源之一，而对高光谱数据的挖掘亦将成为高光谱数据应用的关键环节。本文仅对高光谱信息挖掘进行了初步的探讨，以起到抛砖引玉之效。进一步的发展有待于利用挖掘与知识发现的概念与模型，基于地物光谱特征及光谱数据、知识库，在高光谱超维特征空间充分挖掘地物的光谱信息，以达到地物识别目的，同时需对不同的应用目标与领域发展相应的光谱信息挖掘模型与技术，例如，矿物光谱信息挖掘模型、植被参量模型、水域浮游植物、光谱信息挖掘模型、典型目标的光谱信息挖掘模型等等。

参 考 文 献

1　Wang Xiangjun, Wu Changshan, Liu Jiangui *et al*. Data fusion and spectral analysis based on the hyperspectral image processing and analysis system—— HIPAS. In: Proceedings of the Fourth International Airborne Remote Sensing Conferencc/21st Canadian Symposium on Remote sensing, Vol Ⅰ-543, 1999.

2　张　兵，郑兰芬，童庆禧. 成象光谱技术应用植被精细光谱分析. 遥感信息科学开放研究实验室年报，1997，323～327.

3　王向军，张　兵，童庆禧，郑兰芬. 一种基于广义夹角的高光谱

图象分类方法. 遥感信息科学开放研究实验室年报,1997, 328～330.
4 Tong Qingxi, Zheng Lanfen, Wang Jinnian *et al*. Study on the Wetland Environment by hyperspectral remote sensing. In:Proceedings of Third International Airborne Remote Sensing Conference, Vol. I 67-74 Copenhagen Denmark, 7～10 July, 1997.
5 Zhang Bing, Liu Jiangui, Wang Xiangjun *et al*. Urban environmental study with hyperspectal remote sensing incorporating high spatial resolution data. In:Proceedings of the Fourth International Airborne Remote Sensing Conference/21st canadian Symposium on Remote Sensing Vol Ⅱ-657,1999.
6 Wang Jinnian, Zheng Lanfen, Tong Qingxi. Derivative spectra matching for wetland vegetation identification by hyperspectral imagery. SPIE Vol. 3502 0277-786×198. Hyperspectral Remote Sensing and Application, 1998.
7 童庆禧, 郑兰芬, 王晋年等. 湿地植被成像光谱遥感研究. 遥感学报, 1997,1(1).
8 Wang Jinnian, Zheng Lanfen, Tong Qingxi. Spectral absorption identification model and mapping mineral mapping by airborne high spectral rexolution remote sensing data. In:Proceedings of Eleventh Thematic Conferece and Workshop on Applied Geologic Remote Sensing Las Vegas, Nevada, February 1996, 27～29.
9 王晋年, 郑兰芬, 童庆禧. 成像光谱图象光谱吸收, 鉴别模型与矿物填图研究. 环境遥感, 1996,11(1).
10 Adams J B, Smith M O, Gillespie A R. Imaging spectroscopy: Interpretation based on spectral mixture analysis. In:Pieters, C M, Englert P A J (eds), Remote Geochemical Analysis: Elementatl and Mineralogic Composition, UK:Cambridge University Press, 1993, 145～166.
11 Clark R N, Gallagher A J, Swayze G A. Material absorption band depth mapping of imaging spectrometer data using a complete band shape least-squares fit with library reference spectra. In:Proceedings of the Second Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) Workshop, JPL Publication 90-54, 1990,176～186.
12 Cloutis E A. Hyperspectral geological remote sensing:Evaluation of analytical techniques. International Journal of Remote Sensing, 1996,17:2215～2242.
13 Farrand W H, Harsanyi J C. Mapping distributed geologiical and botanical targets through constrained energy minimization. In: Proceedings of the Tenth Thematic Conference on Geological Remote Sensing, San Antonio, Texas, 9～12 May,1994, I-419～I-429.
14 Green A A, Berman M, Switzer P *et al*. A transformation for ordering multispectral data in terms of image quality with implications for noise removal. IEEE Transactions on Geoscience and Remote Sensing,1998,26:65～74.
15 Mustard J F, Sunshine J M. Spectral analysis for earth science: Investigations using remote sensing data. Iin Renz Andrew N (ed), Remote Sensing for the Earth Sciences: Manual of Remote Sensing, Vol 3, New York:John Wiley & Sons, 1999, 251～306.
16 Landgrebe D. On Information Exfroution Principles for Hypers Pectral Data a White Paper,1998.
17 Jimenez L, Landgrebe D A. Supervised classification in high dimensional space:Gep,etrocal, statistical, and asymptotical properties of multivariate data. IEEE Transactions on System, Man, and Cybernetics, January 1998.
18 Goetz A F. Hyperspectral imaging:Advances in a spectrum of applications. In:Proc the 5th Australian Remote Sensing Conference, Perth,8～12 October, 1990.
19 Jia Xiuping, Richards J A. Progresive two-class decision classifier for optimization of class discriminations. Remote Sensing of Environment, 1998,63:289～297.
20 Green A A, Berman M, Switzer P, Craig M D. A transformation for ordering multispectral data in terms of image quality with implications for noise removal. IEEE Transactions on Geoscience and Remote Sensing, 1998,26(1):65～74.
21 Zheng Lanfen, Xiang Yueqin, Liu Weidong. Spectral modeling and crops study for precision farming with high spectral resolution remote sensing. Proc the 4th International Airborne Remote Sensing Conference and Exibition/21st Canadian Symposium on Remote Sensing, 1999, Ⅱ:296～303.

Hyperspectral Data Mining——Toward Target Recognition and Classification

Wang Jinnian, Zhang Bing, Liu Jiangui, Tong Qingxi and Zheng Lanfen
(*Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing* 100101)

Abstract Hyperspectral Data Mining is a key technology to discover spectral information for application of hyperspectral remote sensing data. In this paper, focusing on target recognition and classification, we introduce some hyperspectral data mining algorithm, spectral feature-based data mining algorithm, and sub-pixel spectral information mining algorithm. In advance new algorithms should be developed to explore target spectral information in *N*-dimensional feature space based on spectral knowledge and application modelling.

Keywords Hyperspectral, Data mining, *N*-dimensioal feature, Target recognition