

基于粗糙集和模糊聚类的超谱波段约简¹

石 红 沈 毅 刘志言

(哈尔滨工业大学控制工程系 哈尔滨 150001)

摘 要: 由于超光谱图像数据量大、维数高给分类识别处理带来不便, 该文提出一种可行有效的波段约简方法。通过 FCM 聚类将原始波段划分为若干等价波段组, 然后根据最大隶属度原则只保留每组中具有代表性的波段, 达到维数减小的目的。其中, 模糊聚类中相似度的定义是基于超谱相邻波段间的相关性, 利用粗糙集理论中的处理属性依赖性的方法合理表达出来。实验表明, 这一方法既有效地缩减了高维数据, 又尽可能少地损失有用信息, 保持了原始波段的分类能力。

关键词: 超谱遥感, 粗糙集, 模糊聚类, 波段约简

中图分类号: TP7 **文献标识码:** A **文章编号:** 1009-5896(2004)04-0619-06

Hyperspectral Band Reduction Based on Rough Sets and Fuzzy C-Means Clustering

Shi Hong Shen Yi Liu Zhi-yan

(Dept of Control Eng., Harbin Institute of Technology, Harbin 150001, China)

Abstract A method of hyperspectral band reduction based on Rough Sets (RS) and Fuzzy C-Means (FCM) clustering is proposed, which consists of the following two steps. First, Fuzzy C-Means clustering algorithm is used to classify the original bands into equivalent band groups, which employs the concept of attribute dependency defined in RS to define the distance between a group and the cluster center, viz. the correlatives of adjacent bands. Then the data is reduced by selecting the only one from each group with maximum grade of fuzzy membership. With this approach, great dimension of band is decreased while preserving much wanted information. Simulation results prove the effectiveness of this approach.

Key words Hyperspectral remote sensing, Rough Sets(RS), Fuzzy C-Means(FCM) clustering, Band reduction

1 引言

超光谱分辨率遥感 (Hyperspectral remote sensing) 是指利用很多很窄的电磁波波段从感兴趣的物体获取有关数据。它的基础是测谱学 (Spectroscopy); 20 世纪 80 年代开始建立成像光谱学 (Imaging spectroscopy), 它是在电磁波谱的紫外、可见光、近红外和中红外区域, 获取许多非常窄且光谱连续的图像数据的技术。成像光谱仪为每个像元提供数十至数百个窄波段光谱信息, 能产生一条完整而连续的光谱曲线。与较早的遥感多光谱图像 (4-7 个离散谱带) 技术相比, 它具有两个显著特点: 一是谱分辨率更高 (可达 10nm), 这使得超谱能够解决许多多光谱不能解决的问题; 只是利用高维的遥感数据分类需要大量的样本数据和很长的处理时间; 二是超谱图像相邻谱带间存在较强的相关性, 而多光谱图像的空间相关性要强于谱带间的相关性。

超光谱遥感技术多用于地质调查, 为了从超光谱遥感数据中有效地提取各种地质矿物成分信息, 发展了许多技术方法, 如: 光谱微分技术、光谱匹配技术、混合光谱分解技术、光谱分类技术等。这些分析技术有各自的特点及适用场合。

¹ 2002-12-13 收到, 2003-04-28 改回

国家自然科学基金 (69904004)、教育部跨世纪优秀人才培养计划和高等学校骨干教师资助计划资助课题

本文提出的波段约简方法,目的是利用光谱分类技术进行地物分类。近年来关于超谱图像融合、分类、识别及相关的研究在国际上不断增多,已经公开发表了一定数量的文献,其中 Purdue 大学的 David Landprebe 等做了大量的研究;但总的来说,这方面的研究仍处于起步阶段,许多技术难点仍待解决^[1,2]。人们已经研究了许多多光谱图像的分类技术,如最常用的最大似然分类法(MLC)。MLC 法可谓是经典的分类方法,它主要根据相似的光谱性质和属于某类的概率最大的假设来指定每个像元的类别。MLC 法最大优点是能快速指定被分类像元到若干类之中的一类中去。但是由于超谱图像数据量大、维数高,使通常的多光谱图像分类方法对于超谱图像的应用有较大的限制。因此不少人考虑通过某种预处理(或者说是特征提取)减小数据维数,使通常的多光谱图像处理技术可以对超光谱数据进行分析 and 处理。有人从理论上论证了减小数据维数,同时尽可能多地保留有用信息的可能性^[3]。一种目前主要应用的减小数据维数的方法是主成份分析法(PCA),它是在统计特征(方差)基础上的多变量正交线性变换,使图像能量集中在较少的系数上,可通过抑制具有较少信息量的系数来实现数据维数的减小;但通常 PCA 仅在最小均方误差意义下才是最优的,有人又提出了改进措施^[4,5]。此外,另一种数据维数约简的途径是基于相邻波段间很强的相关性,从所有谱带中选择起主要作用的子集。不同波段对分类的重要性是不同的,相关性强的波段的作用还可以相互近似替代。因此,从众多的光谱波段中选择一组最适合于特定分类任务的优化光谱波段组合,就可以实现数据维数的缩减,同时最多地保留了有用信息。如 Tu 等提出的基于正则分析(CA)的谱带选择(BS)算法^[6]。近年来,人们又尝试把人工智能、模式识别中的一些理论和方法,如遗传算法、统计分析和小波等^[7-10]应用到超谱特征提取当中,出现了一些性能较好,算法较简单的超谱图像融合方法。

本文提出了一种基于粗糙集(Rough Sets, RS)的优化光谱波段组合的选择方法,然后用 MLC 分类法验证了约简后波段组的分类性能。

粗糙集^[11,12]方法把研究对象视作一种信息系统,是一种处理模糊和不确定性的智能信息处理技术。它的优点是不需要预先知道某些特征或属性的数量描述,直接从给定问题的描述集合出发,通过不可分辨关系和不可分辨类确定给定问题的近似域(上、下逼近),描述、揭示出条件属性对于决策属性的重要性,从而明确有用信息和冗余信息,发现隐含在大量的数据背后的内在规律。

根据相邻的波段影像间存在的较大的相关性这一特点,可以对原始的波段集合中的光谱波段进行模糊等价划分,认为同组光谱波段的分类作用近似相同,然后在每个模糊等价波段组中只选择一个光谱波段。其中,光谱波段间的相似性通过粗糙集理论的方法来表达。由于不同的光谱波段对不同的地物的分类识别效果差异较大,因此可视作条件属性对决策属性的重要程度或二者的依赖程度不同;并且,对相同的地物具有近似相同的分类效果的光谱波段被认为是相似的,据此实现模糊等价划分。

2 基于粗糙集和模糊 C-均值(FCM)的近似波段约简

2.1 影像灰度值离散化

遥感影像的灰度值多为连续区间上的数值,如果直接用粗糙集进行等价划分,会导致等价类过多,并容易受噪声影响。离散化的任务就是把属性连续的取值区间划分为若干个小区间,每个区间对应一个既定的符号或整数。离散化效果取决于划分点设定的合理性及离散区间的个数。主要有两大类方法,即非监督和监督离散化。前者不考虑决策信息,效果稍差。本文采用有代表性的监督离散化方法 χ^2 检验法(ChiMerge)^[13]对遥感影像的灰度值进行离散预处理。

χ^2 检验法考虑了决策类别信息,即训练样本所属的类别信息,可以使得分类的精度更高。本离散化方法给出了一个概念称作“相对的类频率”(Relative Frequency of Class, RFC)。举例说明,如果某个属性值区间包含两类样本,其中 28 个正例,12 个反例,则这个区间用 RFC 可以描述成 70% 正且 30% 负。准确的离散化算法,应该能保持同一区间的 RFC 的一致性和相邻不同区间的 RFC 的差异性。ChiMerge 算法用 χ^2 统计度量来判定相邻的区间的 RFC 是差别很大还是足够相近,如为前者,则说明样本所属类别的分布依赖于样本所在的区间和其相邻

的区间, 因此保留这两个区间不动; 如为后者, 说明样本所属类别的分布独立于样本所在的区间和其相邻的区间, 可以将这两个区间合并为一个区间。由此可见, 这种离散化方法尽可能地保持了原有样本数据的类别信息, 这一点对于保留那些具有特殊吸收谱的地物所属类别的信息尤为重要。

2.2 FCM 等价波段划分

对原始光谱波段进行近似等价划分, 与一般的分类任务不同的是划分的类是未知的。因此, 选择聚类的方法, 将上百个光谱波段分组为多个类或簇, 使得同一个簇中的波段之间具有较高的相似性, 而不同的簇中的波段差别较大。

FCM 聚类算法^[14,15]通过对目标函数极小化的迭代优化实现集合划分, 它可以表示出各个元素属于不同类别的程度。目标函数和迭代公式表示为

$$J(U, V) = \sum_{k=1}^n \sum_{i=1}^c u_{ik}^m d_{ik}^2 \quad (1)$$

$$v_i = \sum_k u_{ik}^m x_k / \sum_k u_{ik}^m \quad (2)$$

$$u_{ik} = \begin{cases} 1 / \sum_{j=1}^c (d_{ik}/d_{jk})^{2/(m-1)}, & I_k = \phi \\ 0, & \text{其他} \end{cases} \quad (3)$$

设定迭代停止阈值为 ε_L , l 为迭代次数, 当 $\|U^{(l+1)} - U^{(l)}\| \leq \varepsilon_L$ 则停止迭代, 以上式中 n 为待聚类波段数; c 为类别数; u_{ik} 是第 k 个波段在第 i 个类中的模糊隶属度函数值; v_i 为聚类中心; U 是模糊隶属度集合; V 是聚类中心集合; m 为模糊加权指数; $I_k = \{i | 1 \leq i \leq c; d_{ik} = 0\}$, $\sum_{i \in I_k} u_{ik} = 1$; ϕ 为空集。

特别地, $d_{ik} = \|x_k - v_i\|$, 是波段与聚类中心的距离, 是波段间的相似性的度量, 需要通过光谱波段的属性值 x_k (特征向量) 来计算。

2.3 基于粗糙集的光谱特征向量选取

x_k 和 $d_{ik} = \|x_k - v_i\|$ 的选取, 直接影响 FCM 聚类算法的有效实现, 一种直观的想法是 x_k 取每个光谱波段影像像素的灰度值。假设是 $N \times N$ 个像素的图像, 这样 x_k 的维数过高, 导致算法的处理时间太长。

本文利用粗糙集理论中的基本思想来处理这个问题。

粗糙集理论是一种基于“知识”的处理模糊和不确定信息的新方法。它把研究对象视做一个知识表达系统 $S = (U, R)$, 也称信息系统。这里的“知识”指信息系统中属性集 R 的分类能力。因此, 知识就可以理解为: 等价关系 R 对论域 U 的划分, 记作: $U/R = \{X_1, X_2, \dots, X_n\}$ 。

设 R 是论域 U 上的若干等价关系的交集, R 对 U 的划分得到的各子集 X_i 称为基本概念或基本范畴。

当任一 $X \subseteq U$ 能表达成某些 R 基本范畴的并集时, 称 X 为 R 可定义的或精确的 (Exact sets), 否则 X 为 R 不可定义的, 也称非精确集或粗糙集。

实际上, 基本范畴就是粗糙集理论用来表达一个特定的概念 X 的基本知识模块。精确集是可以用已知知识确定地表达的精确范畴; 粗糙集则只能用知识通过两个精确范畴“上、下近似集” (Upper/lower approximation set) 粗略地定义。

若 $X \subseteq U$ 为粗糙集, 定义两个子集: $R_-(X) = \cup\{X_i \in U/R : X_i \subset X\}$, $R_+(X) = \cup\{X_i \in U/R : X_i \cap X \neq \phi\}$, 分别称为 X 的 R 下近似和 X 的 R 上近似。

粗糙集方法能够揭示出属性间的依赖性关系。设两个属性集 (通常是条件属性和决策属性) $P \subseteq R, Q \subseteq R$ 对 U 的划分分别是: $U/P = \{X_1, X_2, \dots, X_l\}, U/Q = \{Y_1, Y_2, \dots, Y_m\}$, Q 与 P 之间的依赖关系 $\lambda_P(Q)$ 定义如下:

$$\lambda_P(Q) = \sum_i^m \text{card}(P_-(Y_i)) / \text{card}(U)$$

表示了应用知识 P 能确切地划入分类 U/Q 的对象的百分比, 同时也反映了不同的属性对论域对象的分类能力; $\text{card}(\cdot)$ 表示集合的基数。

本文中, 把不同的光谱波段视作条件属性集 $P = \{c_1, c_2, \dots, c_n\}$, 地物的类别作为决策属性 $Q = d$, 则 FCM 算法中光谱波段的特征向量一般地可以定义为各谱段的分类能力, 即

$$x_k = \lambda_{c_k}(d), \quad k = 1, 2, \dots, n$$

显然, $\lambda_{c_k}(d)$ 反映的是一个波段总的分类能力; 而不同光谱波段对不同的地物的分类识别效果是有差异的, 用 $\lambda_{c_k}(d)$ 来度量波段的相似度不够准确。因此, 改进以上定义为

$$x_k = \{x_{k1}, x_{k2}, \dots, x_{km}\}$$

其中 $x_{kj} = \text{card}(c_{k-}(Y_j)) / \text{card}(Y_j)$, $j = 1, 2, \dots, m$, 反映了每一类地物能被第 k 个波段确定地识别出来的比例, 亦即反映了不同类别地物对某一波段的光谱特性。其中包括某些地物与众不同的特性, 因为如果某类地物对某一波段呈与众不同的吸收特性, 则被该波段正确识别的比例就大, 这个聚类向量在对应该类地物的分量上与其他向量的就会有很大不同。

对相同的地物具有近似相同分类效果的光谱波段被认为是相似的, 通过对 $x_k (k = 1, 2, \dots, n)$ 聚类划分至同一组。然后根据最大隶属度原则同时考虑相邻波段相关性更强, 从各组中只选其一, 得到原始高维光谱波段的一个近似等价约简。

2.4 计算复杂性分析

粗糙集方法部分需要求 m 个决策类分别由 n 个条件属性所决定的下近似。首先, 求属性对论域的划分的的最坏的复杂度为 $O(|A||U|^2)$ 。其中, A 为属性集合, U 为对象集合。如果按给定属性集先对对象排序, 它的复杂度就降低为 $O(|A||U|\lg|U|)$ 。本文中每次 A 只有一个属性, 因此复杂度为 $O(|U|\lg|U|)$ 。接下来, 在已给定划分的情况下求下近似的复杂度为 $O(|U|)$, 这样求一个决策类下近似的整个的复杂度就是 $O(|U|\lg|U|)$ 。所以, 这部分的复杂度就是 $O(nm|U|\lg|U|)$ 。

属性离散化部分的复杂度为 $O(|U|^2)$, 因为对象排了序, 复杂度降低为 $O(|U|\lg|U|)$ 。

模糊聚类部分, 目标函数在有限步内收敛, 其复杂度为 $O(c|U|)$, c 为聚类数。

3 实验及结论

实验数据来自 AVIRIS 超光谱影像 (1992, 美国印第安那州实验区)。样本数据 2528, 包括 16 类地物, 总共 220 个光谱波段, 去除其中易于被水吸收和噪声污染的波段 103-111 及 148-166 波段, 剩余 192 个波段, 设聚类数为 9。通过聚类, 然后根据最大隶属度原则挑选得到的约简为

$$V_1 = \{9, 15, 31, 50, 56, 65, 96, 136, 186\};$$

实验还得到另外几组约简的波段:

$$V_2 = \{5, 7, 32, 38, 59, 63, 97, 116, 195\}; V_3 = \{13, 17, 18, 27, 28, 56, 65, 68, 120\};$$

V_2 : 模糊聚类后, 由每组选择的 $\delta_{c_k}(d)$ 最大的波段组成; V_3 : 聚类的特征向量为改进前的一维量 $\delta_{c_k}(d)$, 由每组隶属度最大的波段组成。

表1给出了各组约简波段影像最大似然分类的精度, 其中3类地物的样本数太少, 使最大似然分类计算过程矩阵产生奇异, 因此用其余13类进行分类验证。表中 V_4 为文献[16]方法的约简波段组:

$$V_4 = \{25, 56, 86, 117, 145, 186, 207\}.$$

表1 最大似然分类精度

类别名称		RS-FCM(%)			
		V_1	V_2	V_3	V_4
1	Corn-notill	90	92.5	42.5	82.5
2	Corn-min	79.17	91.67	79.17	91.67
3	Corn	100	95.24	95.24	95.24
4	Grass/pasture	100	100	100	97.06
5	Grass/trees	100	100	100	83.87
6	Hay	100	100	100	100
7	Soy-notill	91.53	83.05	94.92	91.52
8	Soy-min	94.52	87.67	75.34	79.45
9	Soy-clean	82.61	82.61	73.91	73.91
10	Wheat	100	95	100	95
11	Woods	100	100	100	100
12	Bl-gr-tr-dr	100	92.31	92.31	69.23
13	Stone	100	77.78	77.78	77.78
	Average	95.22	92.14	87.01	87.48

从表1中数据发现, V_1 对应的波段相似度的定义显然较 V_3 的更合理准确; V_2 选择波段的原则忽视不同波段对地物的不同特性, 因此分类的精度也稍逊于 V_1 ; 而对于13类地物的分类效果 V_1 也好于 V_4 。可见, 本文提出的波段约简方法是有效的。

不足的是, 以上算法, 虽然通过聚类能够分析出哪些波段对识别某一类或几类地物是相似的, 但是因为聚类的准则是向量间的“欧氏距离”, 综合考虑了向量的各个分量, 因而可能会使包含特殊吸收谱信息的向量对于聚类中心的隶属度并非最大。可以通过多次的实验, 找到一种比按照最大隶属度原则来选择波段更好的方法, 从聚类的结果中选择更具代表性的波段。文中之所以给出的第1和第2种约简波段组的比较, 就是基于此目的尝试。只是第2组更注重波段对所有地物的整体分类能力, 忽略了波段对某类地物的特性; 相对第2组, 第1组已经在很大程度上增大了波段对各类地物单独分类能力的重视。当然, 有待于研究的是最好找到一种更合理的方案, 最大程度地减小具有吸收特性的谱段信息的丢失, 进一步提高分类精度。

图1为某一波段的原超谱图像, 图中标出的矩形及多边形为16种地物样本; 图2为利用约简后的波段进行MLC分类后的结果示意图。

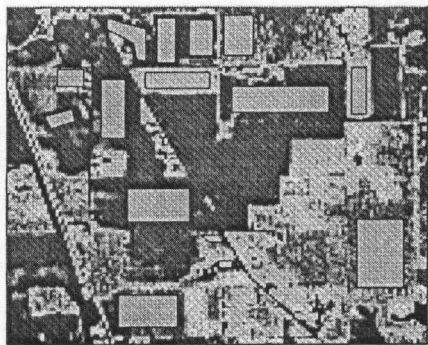


图1 某一波段的超谱图像(黑白)

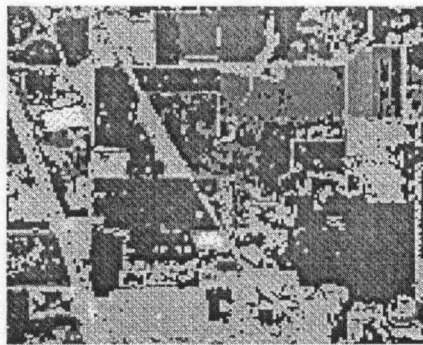


图2 波段约简后 MLC 分类示意图(彩色)

4 结束语

本文提出了一种基于粗糙集和 FCM 聚类的遥感超光谱波段约简方法。超谱数据的各个谱带间的相关特性不同,在不同的频率段相关性的差别很大,给依据相邻波段间的相关性进行冗余波段约简的处理带来不便;针对不同的波段对不同的地物的分类效果相差很大这一特点,粗糙集理论能够用属性的分类能力(或属性间的依赖性)定义波段之间的相关性,实现模糊等价聚类,从而进一步实现了一种算法简单,性能良好的高维超光谱波段约简方法;既减小了数据维数又最大地保留了原始光谱的信息。

参 考 文 献

- [1] G. P. Abousleman, *et al.*. Hyperspectral image compression using entropy-constrained predictive trellis coded quantization. *IEEE Trans. on Image Processing*, 1997, IP-6(4): 566-573.
- [2] Ryan M J, Arnold J F. The lossless compression of AVIRIS images by vector quantization. *IEEE Trans. on Geoscience and Remote Sensing*, 1997, GRS-35(3): 546-550.
- [3] Jimenez L O, Landgrebe D. A supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. *IEEE Trans. on System, Man, and Cybernetics-Part C: Applications and Reviews*, 1998, SMC-C-28(1): 39-54.
- [4] Jia Xiuping, Richards J A. Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification. *IEEE Trans. on Geoscience and Remote Sensing*, 1999, GRS-37(1): 538-542.
- [5] Zhang Ye, Desai M D, Zhang Junping, *et al.*. Adaptive subspace decomposition for hyperspectral data dimensionality reduction. International Conference on Image Processing (ICIP99'), Kobe, Japan, 1999: 326-329.
- [6] Tu Te-Ming, Chen Chin-Hsing. A fast two stage classification method for high dimensional remote sensing data. *IEEE Trans. on Geoscience and Remote Sensing*, 1998, GRS-36(1): 182-191.
- [7] Morgan J T, Henneguelle A, Crawford M M, *et al.*. Best bases Bayesian hierarchical classifier for hyperspectral data analysis. Geoscience and Remote Sensing Symposium, IGARSS'02, Toronto, Canada, 2002 IEEE International, 2002, Vol.3: 1434-1437.
- [8] Esposito P G, Bartoloni A. An application of genetic algorithms to the geometric correction of HypSEO hyperspectral data. Geoscience and Remote Sensing Symposium, IGARSS'02, Toronto, Canada, 2002 IEEE International, 2002, Vol.6: 3507-3509.
- [9] Kaewpijit S, Le Moigne J, El-Ghazawi T. A wavelet-based PCA reduction for hyperspectral imagery. Geoscience and Remote Sensing Symposium, IGARSS'02, Toronto, Canada, 2002 IEEE International, 2002, Vol.5: 2581-2583.
- [10] Hsu Pai-Hui, Tseng Yi-Hsing. Feature extraction of hyperspectral data using the best wavelet packet basis. Geoscience and Remote Sensing Symposium, IGARSS'02, Toronto, Canada, 2002 IEEE International, 2002, Vol.3: 1667-1669.
- [11] Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Dordrecht: Kluwer Academic Pub., 1991: Ch.2-5.
- [12] Pawlak Z, *et al.*. Rough sets. *Communications of the ACM*, 1995, 38(11): 89-95.
- [13] Kerber R. ChiMerge: discretization of numeric attributes. in Proc. 10th National Conference on Artificial Intelligence, San Jose, CA, 1992: 123-127.
- [14] Dunn J C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 1973, 3(3): 32-57.
- [15] Bezdek C, Dunn J C. Optimal fuzzy partitions: a heuristic for estimating the parameters in a mixture of normal distributions. *IEEE Trans. on Comput.*, 1986, C-35(8): 935-938.
- [16] 孙立新, 高文. 基于粗糙集的遥感优化分类波段选择. 模式识别与人工智能, 2000, 13(2): 181-186.

石 红: 女, 1975 年生, 博士生, 主要研究方向为人工智能、粗糙集等。

沈 毅: 男, 1965 年生, 教授, 博士生导师, 主要研究方向为智能检测处理与控制。已发表论文 80 余篇。

刘志言: 男, 1936 年生, 教授, 博士生导师, 主要研究领域为工业过程控制。已发表论文 50 余篇。