

利用特征子空间评价与多分类器融合的高光谱图像分类

杨艺, 韩崇昭, 韩德强

(西安交通大学电子与信息工程学院, 710049, 西安)

摘要: 为应对高光谱图像分类中的特征高维度问题,提出一种基于多分类器融合的高光谱图像分类方法. 利用高光谱数据相邻波段的高相关性,通过自适应子空间分解产生多个特征子空间,进而训练生成子分类器;利用 ReliefF-S 算法,对各特征子空间进行评价并生成各子分类器的权重,最终通过加权表决融合实现分类决策. 实验表明,所提方法可有效规避高维特征问题并提升分类性能.

关键词: 高光谱图像;多分类器融合;自适应子空间分解;加权表决

中图分类号: TP391 **文献标志码:** A **文章编号:** 0253-987X(2010)08-0020-05

Hyperspectral Image Classification Based on Feature Subspace Evaluation and Multiple Classifier Fusion

YANG Yi, HAN Chongzhao, HAN Deqiang

(School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: A novel approach for hyperspectral image classification is proposed based on fusion of multiple classifiers to deal with the high dimension in applications of hyperspectral image classification. High correlation of neighboring bands of hyperspectral image data is used to generate feature subsets through adaptive subspace decomposition. A modified ReliefF algorithm (ReliefF-S) is proposed to evaluate feature subsets and to generate their corresponding weight values. Member classifiers are trained based on resulting feature subspaces and their weighted majority voting, and then the fusion of multiple classifiers is accomplished. Experimental results show that the proposed approach reduces the dimension of features effectively, and improves the classification performance.

Keywords: hyperspectral image; multiple classifier fusion; adaptive subspace decomposition; weighted voting.

近年来,高光谱图像分析^[1]在军事和民用等诸多领域都得到了越来越多的重视.高光谱传感器可同时从宽电磁波谱中获取上百个连续的窄频波段信息(即从可见光到红外,0.4 到 2.5 μm).高光谱图像有很高的谱分辨率,频谱波段的增加可有效减少类间重叠,进而提升分类精度^[2-3].然而,高分辨率往往意味着高特征维度.高特征维度是一柄双刃剑,

在带来更丰富信息的同时会增加计算复杂度,当训练样本数不充分时,甚至会引发维数灾难.高维度特征往往含有冗余特征,冗余特征过多有时反而会抑制分类性能.高特征维度问题已成为高光谱图像分析与解译中的一大障碍.

针对高特征维度问题,涌现了许多特征选择方法,包括主成分分析(PCA)、独立分量分析(ICA)、

Relief 方法^[4]、Wrapper 方法及基于智能信息处理的方法等. 基于这些方法, 可获取相对较少的波段(特征维度)用于后续分析与分类. 另一种解决分类中高维特征问题的方法是借助于多分类器融合^[5]将相对复杂的分类问题分解成若干相对简单的分类问题. 高维的高光谱特征可以通过某种方式被分解成多个特征子空间, 分别基于不同子空间设计和训练子分类器, 最终基于某种融合规则实现多分类器融合进而提升分类性能. 其中, 关键因素包括特征子空间生成方法及多分类器融合规则等.

如何选择合适的子空间维度是一个棘手的问题. 随机子空间法(RSM)^[6]是一种典型的特征子空间生成方法, 能有效解决小样本问题. 在原始 RSM 方法中, 子空间随机产生且子空间的维度固定. 对高维特征空间来说, 基于 RSM 获取的子空间数量巨大. 高光谱图像数据的谱相关性强于空间相关性^[7]. 这种强谱相关性是一种可以利用的分解标准, 促使自适应子空间分解(ASD)^[8]等基于谱相关性实现特征空间分解的方法应运而生. 对融合规则来说, 目前应用较为广泛的包括绝对多数表决(majority voting)规则^[9]、Bayes 规则、Dempster 证据组合规则等. 在分类问题中, 多数表决规则简单易行. 如果在表决过程中加入权重信息, 往往可取得更为理想的效果.

本文针对高光谱遥感数据, 利用 ASD 生成特征子空间并生成子分类器. 利用文中提出的 ReliefF-S 方法对特征子空间进行评价, 将所获评价值作为权重, 最终利用加权表决规则实现多分类器融合. 基于高光谱数据集的实验结果表明, 本文所提方法能在有效规避高特征维度问题的同时提升分类性能.

1 多分类器融合

多分类器融合是根据各个分类器的整体性能将每个分类器的输出按某种方式“融合”在一起, 并达到“共识”. 多分类器融合因其在机器学习及模式分类中的出色性能, 受到了广泛的关注. 此外, 在具体的分类问题应用中, 如果特征的维数太大, 仅采用单分类器会比较困难. 此时, 将高维特征向量分解成几个低维向量, 分别作为单分类器的输入再进行融合也是一种很好的选择. 因为多分类器融合对单分类器的性能及运算量要求相对较低, 用分解得到的低维向量作为单分类器的输入, 继而进行多分类器融合, 既可简化单个分类器的构造难度从而降低系统开销, 又可利用不同的特征波段间的互补性提高融

合分类性能. 基于多特征空间的多分类器融合流程如图 1 所示.

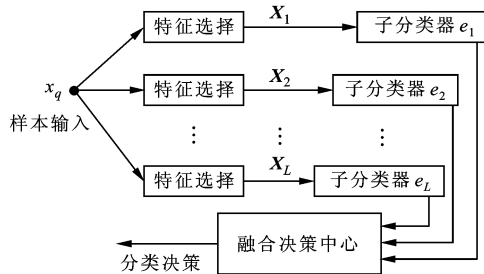


图 1 基于多特征空间的多分类器融合

图 1 中, $X_k (k=1, 2, \dots, L)$ 代表 L 个不同特征空间. 给定样本 x_q , 类别数为 M , 记为 $C_i, \forall i \in \{1, \dots, M\}$. 图 1 中的 $e_j (j=1, 2, \dots, L)$ 代表不同特征空间 X_k 对应的子分类器. 所谓不同特征空间, 可以是高维特征通过某种方式生成的众多特征子空间, 诸如经过基于广义粗集的特征选择^[10]生成的不同特征组合, 经由随机子空间方法(RSM)产生的不同子空间, 也可以是异类异质特征空间. 除采用不同特征空间外, 构造多分类器系统还可以基于不同样本、不同分类器类型等实现^[11].

对分类器融合来说, 子分类器的输出类型是一个关键因素, 直接关系到融合方法的设计和选用. 常见的输出类型包括: 抽象级(Abtract), 输出仅为类别标号; 排序级(Rank), 输出为所有类别标号按可能性大小的排序; 度量级(Measurement), 输出为一系列度量值, 诸如后验概率、隶属度等, 描述样本归属于各类别的程度. 对抽象级和排序级输出, 主要融合方法为表决融合, 且根据算法不同又有多数表决和相对多数表决、加权表决等. 若子分类器输出为度量级, 概率推理、模糊推理以及 D-S 证据理论等都是适合的融合方法^[10-11].

加权表决融合相对于简单表决融合而言, 能进一步区分各个子分类器的差异性, 使各个子分类器在决策的过程中按照其分类能力或重要性发挥不同的作用, 从而提升分类性能, 具体实现如下.

输入样本为 x_q , 子分类器为 $e_j, j=1, \dots, L$. 设 $i, g \in \Delta = \{1, \dots, M\}$ 代表类别标号, W_j 代表不同特征空间(对应不同子分类器)的权重, δ_{ji} 代表子分类器 j 对类别 i 的“硬”支持程度(0 或 1, 即不支持或支持), 其取值与子分类器 j 输出的类别标号相关, 如式(2)所示, 若子分类器 j 输出类别标号为 i , 则 δ_{ji} 取值为 1, 否则取值为 0. 最终融合决策结果类 g 为所有子分类器“硬”支持程度加权和最高的类别.

加权表决融合规则为

$$\sum_{j=1}^L W_j \delta_{jg} = \max_{i \in \Lambda} \sum_{j=1}^L W_j \delta_{ji} \geq 0.5 \tag{1}$$

$$\delta_{jg} = \begin{cases} 1, & e_j \text{ 判决为类别 } g \\ 0, & \text{其他类别} \end{cases} \tag{2}$$

本文基于加权表决融合实现高光谱遥感地物分类. 各子分类器基于不同特征子空间进行训练, 权重基于对各个特征子空间的评价生成, 以下将分别介绍自适应特征子空间分解及特征空间评价方法.

2 自适应子空间分解

在高光谱图像中, 谱相关性强于空间相关性, 连续谱段相关性很强. 一些学者研究并提出了基于波段相关性分析的自适应子空间分解(ASD)方法. 不同波段相关程度不同, 因此, ASD 获取的各特征子空间大小未必一致, 其实现依赖于不同波段间的相关矩阵 $\mathbf{R}$ 的计算. $\mathbf{R}$ 的元素由式(3)求解

$$r_{ij} = \frac{\sum_p (x_{pi} - \bar{x}_{pi})(x_{pj} - \bar{x}_{pj})}{\sum_p (x_{pi} - \bar{x}_{pi})^2 \sum_p (x_{pj} - \bar{x}_{pj})^2} \tag{3}$$

式中: p 为像素点的标号; x_{pi} 为像素点 p 的第 i 维取值.

相关矩阵 $\mathbf{R}$ 中的元素 $r_{ij} \in [0, 1]$ 越靠近 1, 波段 i 和 j 间的相关性就越大. 在 ASD 中, 当两个相邻波段的相关性较高时, 它们自然地合并成一组, 直到出现一个波段与起始波段的相关系数小于设定阈值为止, 如此继续, 则特征空间 S 被划分成若干个相邻的波段组. 通过这种方式, 同一特征子空间内所有波段间的相关程度接近. 如图 2 所示, 维度为 N 的特征空间 S 被分解成 L 个大小不同的特征子空间.

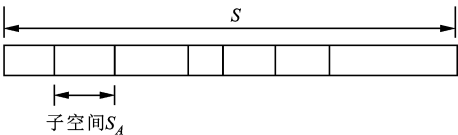


图 2 自适应子空间分解

显然, 有如下关系成立

$$S = \cup S_A, \quad A = 1, \dots, L, \quad L \leq N \tag{4}$$

3 基于 ReliefF-S 的子空间评价

ASD 方法所获取的特征子空间的分类能力各不相同, 在多分类器融合中, 如果能依据各个特征子空间的分类能力或贡献分别赋予不同的权重, 将会进一步提升融合分类性能. 权重的构造可借助于特征评价法来实现. 在众多特征评价方法中, Relief 简

单易行, 其核心思想是根据特征对近距离样本的区分能力来评估特征, 好的特征应该使同类的样本接近, 而使不同类的样本彼此远离. Kononenko 在 Relief 的基础上进一步研究提出 ReliefF 算法, 可解决多类问题及回归问题, 并补充了对数据缺失情况的处理办法. ReliefF 处理多类问题时, 从每个不同类别的样本集合中选择 k 个最近邻样本. ReliefF 平均了 k 个近邻的作用, 这样可以在一定程度上抑制冗余点与噪声点的干扰. ReliefF 是对全特征空间中的所有单特征逐一进行评价的, 本文提出一种改进的 ReliefF-S 算法(S 的含义为 Subset, 即特征子集), 直接对特征子空间进行评价. ReliefF-S 算法的伪代码见图 3.

输入: D 为训练样本集, L 为特征子空间数,
 m 为迭代次数, k 为近邻样本数.
输出: $\mathbf{W}=[W_1, W_2, \dots, W_L]$ 为特征评价(或权重)矢量

1) 将所有权重初值设为 0,
 $W_A=0, A=1, \dots, L;$

2) for $i=1$ to m do begin

3) 从样本集 D 中挑选样本 $P_s;$

4) 找出 P_s 的 $k_{\text{nearest Hits}} H_j;$

5) 对于每个类 $C \neq \text{class of } P_s,$
找出 $k_{\text{nearest Misses}} M_j(C);$

6) for $A=1$ to L do

$$W_A = W_A - \sum_{j=1}^k \text{diff}(A, P_s, H_j)/(mk) +$$
$$\frac{\sum_{C \neq \text{class}(P_s)} \frac{p(C)}{1-p(\text{class}(P_s))} \sum_{j=1}^k \text{diff}(A, P_s, M_j(C))}{mk}$$

7) end

8) end

(*)

图 3 ReliefF-S 流程

如图 3, ReliefF-S 随机选取样本 P_s , 找出其同类的 k 个近邻样本(记为 nearestHits, NH). 同时, 在每个不同类的样本集中也各找出 k 个近邻样本(记为 nearestMisses, NM). 按图 3 中式(*)计算不同特征子空间上的权值 $W_A, A=1, \dots, L$. 上述过程迭代 m 次. 式(*)中, $\text{diff}(A, P_s, P')$ 表示样本 P_s 与 P' 在第 A 个特征子空间上的距离(P' 表示 NH 或 NM). 最终可求得各个特征子空间的权重矢量 $\mathbf{W}$. 算法中, k, m 的值由人为设定.

ReliefF-S 的思想是某特征空间上同类样本越接近(即 $\sum_{j=1, \dots, k} \text{diff}(A, P_s, H_j)/(mk)$ 取值越小), 同时不同类样本之间越彼此远离(即 $\sum_{j=1, \dots, k} \text{diff}(A, P_s, M_j)/(mk)$ 取值越大), 该特征子空间权值也就越大, 分类能力越强.

生成各个特征子空间及其各自权重之后,就可以基于加权表决融合来实现分类。

4 基于特征子空间分解与加权表决融合的高光谱图像分类

本文采用自适应子空间分解方法、ReliefF-S 特征子空间评价方法及多分类器加权表决融合技术来完成高光谱图像地物分类。

具体流程如图 4 所示,首先通过对全特征空间使用 ASD 生成 L 个特征子空间 $S_1, \dots, S_L$,并训练相应的 L 个子分类器 $e_1, \dots, e_L$,同时利用 ReliefF-S 获取 L 个特征子空间相应的权值 $W_1, \dots, W_L$,最终按照式(1)、(2)中的加权表决融合规则实现多分类器融合。

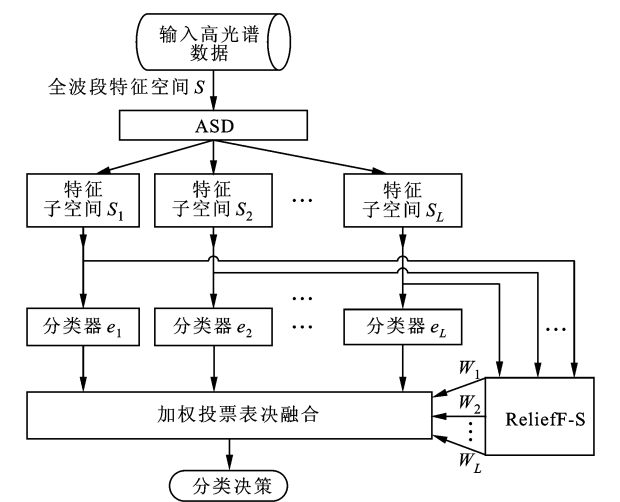


图 4 融合分类流程

5 实 验

为了验证所提方法的有效性和合理性,本文基于两组高光谱数据集进行实验。

5.1 基于 Jasper 数据集的实验

在 Jasper 数据集中选取如下类别的样本:类 1 为常青林;类 2 为孤树;类 3 为灌木丛;类 4 为非螺旋纹草地。样本特征为 60 维。Jasper 高光谱数据集的地面资料见文献[12]。实验步骤如下。

(1)样本选取及分类器训练。从每个类别中随机选取 1 000 个样本作为训练样本,同时随机选取另外 1 000 个样本作为测试样本。子分类器为最小二乘支持向量机(LS-SVM)。

(2)利用 ASD 生成特征子空间。基于 ASD(相关性阈值设为 0.885),原始特征空间被分解为 3 个

特征子空间。子空间 1 的波段为(1, ..., 19),维数为 19;子空间 2 的波段为(20, ..., 48),维数为 29;子空间 3 的波段为(49, ..., 60),维数为 12。

(3)利用 ReliefF-S 进行特征子空间评价。从训练样本集中随机选取 250 个样本来进行基于 ReliefF-S 的特征子空间评价,进而产生各子空间相应的权值:子空间 1 为 3.375 1;子空间 2 为 1.075 5;子空间 3 为 2.248 5。重复进行 30 次实验,每次训练样本、测试样本等均重新随机选取,分类性能取 30 次实验的平均值。为了验证所提方法的有效性,对基于初始全特征空间以及无加权的投票表决等方法进行实验,并对结果进行了比较,如表 1 所示。

表 1 基于 Jasper 数据集的分类性能

方法	分类性能				
	类 1	类 2	类 3	类 4	总体
全空间	78.93	83.47	96.73	99.73	89.72
表决	79.97	79.87	94.10	99.93	88.47
加权表决	83.97	93.10	95.20	99.93	93.05

5.2 基于延安枣园地区数据集的实验

在延安枣园数据集中,选取如下类别的样本:类 1 为玉米地;类 2 为桃园地;类 3 为松树苗地;类 4 为葡萄园地。样本特征为 78 维。子分类器为 LS-SVM,实验步骤如下。

(1)样本选取及分类器训练。从每个类别中随机选取 1 000 个样本作为训练样本,同时随机选取另外 1 000 个样本作为测试样本。

(2)利用 ASD 生成特征子空间。基于 ASD(相关性阈值设为 0.865),原始特征空间被分解为 3 个特征子空间:子空间 1 的波段为(1, ..., 23),维数为 23;子空间 2 的波段为(24, ..., 69),维数为 46;子空间 3 的波段为(70, ..., 78),维数为 9。

(3)利用 ReliefF-S 进行特征子空间评价。从训练样本集中随机选取 250 个样本来进行基于 ReliefF-S 的特征子空间评价,进而产生各子空间相应的权值:子空间 1 为 2.310 9;子空间 2 为 5.577 0;子空间 3 为 1.845 6。重复进行 30 次实验,每次训练样本、测试样本等均重新随机选取,分类性能取 30 次实验的平均值。为了验证所提方法的有效性,对基于初始全特征空间、无加权的投票表决等方法进行了实验,并对结果进行比较,如表 2 所示。

由表 1 和表 2 中的实验结果可知,加权表决融合分类性能最优。相比其他方法,本文提出的基于多

分类器加权表决融合方法可有效提升分类准确率,同时子分类器设计是在相对较低维度的子空间上进行的,例如在 5.1 节的实验中,原始特征空间维度为 60 维,而基于 ASD 分解后得到的特征子空间维数分别为 19、29 和 12;例如在 5.2 节的实验中,原始特征空间维度为 78 维,而基于 ASD 分解后得到的特征子空间维数分别为 23、46 和 9,无需在高维全特征空间上进行分类器设计与训练,基于各特征子空间的各子分类器的设计和训练还可以并行实现,从而降低了问题复杂度。

表 2 基于延安枣园地区数据集的分类性能

方法	分类性能				
	类 1	类 2	类 3	类 4	总体
全空间	94.10	98.80	93.87	99.53	96.58
表决	92.27	99.20	85.77	99.33	94.14
加权表决	94.10	99.20	96.80	99.53	97.40

需要指出的是,多分类器融合并不能总是提升分类准确率,其中最为关键的因素是各个子分类器间的差异性.本文实验中差异性体现在各子分类器是基于不同的特征子空间生成的,同时利用各子空间的性能评估作为权重参与加权表决融合,进一步增大了差异性,从实验结果来看,加权融合的性能优于无加权的融合,即权重的引入扩大了各子分类器间的差异,从而进一步提升了融合分类效果。

6 结 论

本文提出了一种基于多分类器融合实现高光谱图像分类的方法.利用高光谱数据相邻波段的高相关性,通过特征空间自适应分解算法产生多个特征子空间,并基于此设计训练子分类器.利用针对特征集合的 ReliefF-S 算法,对各个特征子空间进行评价进而生成各子分类器的权重,最终通过加权表决融合实现多分类器决策,提高分类性能.基于高光谱数据的实验表明,本文所提方法合理有效。

多分类器融合的研究是当前的热点问题之一,本文基于多分类器融合技术来应对高光谱图像地物分类中的高特征维度问题.多分类器融合的目的和作用包括降低问题复杂度以及提升分类性能,而真正有效地提高分类准确率还有赖于多分类器间的差异性的增大.差异性的研究是多分类器系统研究的前沿,目前还没有能被广泛接受的理论与方法.定义差异性度量并以此监督多分类器融合系统的构造与

训练,才能从根本上保证和提高多分类器融合的有效性,这也是未来的工作与研究方向。

参考文献:

[1] PLAZA A, BENEDIKTSSON J A, BOARDMAN J W. Recent advances in techniques for hyperspectral image processing [J]. Remote Sensing of Environment 2009, 113: S110-S122.

[2] LANDGREBE D. Hyperspectral image data analysis [J]. IEEE Signal Processing Magazine, 2002, 19(1): 17-28.

[3] LANDGREBE D. On information extraction principles for hyperspectral data: a white paper[R]. West Lafayette, Indianan, USA: Purdue University. School of Electrical and Computer Engineering, 1997.

[4] SUN Yijun. Iterative relief for feature weighting: algorithms, theories, and applications[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(6):1035-1051.

[5] KUNCHEVA L I. Combining pattern classifiers: methods and algorithms[M]. New York, USA: Wiley, 2004.

[6] HO T H. The random subspace method for constructing decision forests [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1998, 20(8): 832-844.

[7] TU T M, CHEN C H, WU J L, et al. A fast two-stage classification method for high-dimensional remote sensing data [J]. IEEE Trans on Geoscience and Remote Sensing, 1998, 36(1): 182-191.

[8] ZHANG Ye, MITA D, ZHANG Junping, et al. Adaptive subspace decomposition for hyperspectral data dimensionality reduction [C]// Proceedings of IEEE International Conference on Image Processing. Piscataway, NJ, USA: IEEE, 1999: 326-329.

[9] BEHROOZ P. Voting algorithms [J]. IEEE Trans on Reliability, 1994, 43(4):617-629.

[10] 孙亮. 高维遥感数据融合与分类的知识发现方法研究[D]. 西安: 西安交通大学电子与信息工程学院, 2006.

[11] 韩德强. 基于证据推理的多源分类融合理论与方法研究[D]. 西安: 西安交通大学电子与信息工程学院, 2008.

[12] ITT. ENVI tutorial: vegetation hyperspectral analysis [EB/OL]. [2009-11-06]. <http://www.ittvis.com/portals/o/tutorials/envi/Vegetation-Analysis.pdf>.

(编辑 武红江)