DOI:10.13203/j.whugis20160531　　文章编号:1671-8860(2019)03-0405-08

利用随机森林回归进行极化 SAR 土壤水分反演

李平湘[1]　刘致曲[1]　杨　杰[1]　孙维东[1]　黎旻懿[2]　任烨仙[1]

1　武汉大学测绘遥感信息工程国家重点实验室,湖北 武汉,430079

2　德清数联空间信息技术有限公司,浙江 德清,313200

摘　要:全极化合成孔径雷达影像能够提供地物丰富的极化信息,挖掘这些信息在地表参数反演中的作用是目前相关领域的研究趋势之一。针对冬小麦区域的不同植被覆盖情况,利用随机森林回归对常用极化特征在土壤水分反演中的重要性进行评估,并在此基础上进行特征选择,挑选优化的极化特征组合,构建了高精度的土壤水分反演模型。实验结果显示,由重要性评分较高的极化特征所组成的反演模型能得到均方根误差(root mean square error, RMSE)小于 6% 的反演精度,比只输入传统线极化后向散射系数的模型在不同时相、不同数据集的精度都有所提高。与支持向量回归和人工神经网络模型进行比较,利用随机森林回归进行重要性评分与土壤水分反演的效果更好。

关键词:极化 SAR;土壤水分;随机森林回归;支持向量回归;人工神经网络

中图分类号:P237　　**文献标志码**:A

土壤水分是地球生态系统中一个十分重要的组成部分,它是许多水文模型、气候模型、生态模型等的重要输入参数[1],也是研究植物水分胁迫,进行旱情监测、农作物估产等的一个重要指标[2]。随着卫星遥感技术的发展,利用高时空分辨率的多源遥感技术进行大范围土壤水分信息的获取成为了可能。其中,微波由于较强的穿透性以及与表层土壤物理性质的强相关性,被大量运用于土壤表层水分反演中。但被动微波遥感由于空间分辨率较低,在小尺度的土壤水分监测上受到了一定限制;而合成孔径雷达(synthetic aperture radar, SAR)作为主动微波遥感手段不仅具有较高的空间分辨率,也能够提供有利于土壤水分反演的极化信息。因此,利用极化 SAR 信息的土壤水分反演是目前该领域的研究趋势之一。

传统的微波土壤水分反演方法大多在后向散射强度特征与土壤介电、几何特性之间关系的基础上展开。鉴于全极化 SAR 影像能够提供更丰富的极化信息,极化特征在地表参数反演中的相关应用问题也已被不少研究者们进行了不同程度的讨论。文献[3]指出少数极化指标在一定程度上能够提高深层土壤水分反演的精度;文献[4]分析了几种主要的极化特征对春小麦、大豆、玉米和油菜生物量的敏感性,肯定了极化 SAR 反演农作物生物量的潜力;文献[5]提出部分极化特征对地表粗糙度与残茬覆盖具有较高敏感性,这些特征对土壤水分反演精度的提高有潜在研究价值;文献[6]提取多波段 SAR 影像的 Cloude-Pottier 极化分解[7]特征对裸土区域土壤水分与地表粗糙度进行了统计分析,认为这些特征对地表参数敏感性不高。

由此可见,在不同地物、不同观测条件下,极化特征对地表参数的表现存在一定差异,因此研究极化特征在地表参数反演中的作用仍具有实际意义。此外,上述研究大多采用简单的线性回归分析,无法整体而系统地衡量大量极化特征在反演中的作用;而采用机器学习的方法,可以不受制于输入参数的类型与个数。目前在地表参数反演的研究中运用较多的机器学习算法有人工神经网络(artificial neural networks, ANN)[8]、支持向量回归机(support vector regression, SVR)[9]与随机森林回归(random forest regression, RFR)[10]。其中 ANN 与 SVR 在土壤水分反演中的应用包括前向理论模型的反演[11]、经验数据的分析学习[12]以及不同分辨率影像信息的降尺度研究[13],展现了这两种机器学习方法学习速度

收稿日期:2018-03-06

项目资助:国家自然科学基金(41771377, 41601355, 91438203, 41501382);国家国防科技工业局高分项目(03-Y20A10-9001-15/16)。

第一作者:李平湘,教授,主要从事极化 SAR 解译的理论与方法研究。pxli@whu.edu.cn

通讯作者:刘致曲,博士生。meloqu@qq.com

快、反演精度高的优势；RFR则多用于森林生物量及其他植被参数的反演[14]中，对于土壤水分反演的适用性仍然缺乏验证。

与ANN、SVR相比，RFR具有训练参数较少、计算开销小、能生成特征的重要性度量等优点，因此被选为本文利用极化特征进行土壤水分反演研究的建模方法。本文以河北省保定市定兴县的冬小麦农田区域为实验对象，利用C波段Radarsat-2影像提取极化特征研究基于RFR的土壤水分反演问题，通过对极化特征的重要性评估，选择合适的极化特征组合对冬小麦区域进行了土壤水分反演与精度评价，同时与SVR和ANN进行比较，验证RFR对土壤水分反演的适用性以及重要性度量对特征选择的有效性。

1 利用随机森林回归的土壤水分反演

1.1 随机森林回归

随机森林方法在以决策树为基学习器构建Bagging集成的基础上，进一步在决策树的训练中引入随机属性选择，使得最终集成的泛化性能可通过个体学习器之间的差异度的增加而进一步提升。简单地说，随机森林是以决策树为基本分类或预测器的一个集成学习模型，每一个决策树是由分类回归树(classification and regression tree，CART)算法构建的未剪枝的决策树。而RFR的基本思想是基于统计学理论，利用Bootstrap抽样方法从原始样本中有放回地抽取多个样本，对每个Bootstrap样本集构建决策树，将所有决策树预测平均值作为最终预测结果。算法具体流程参见文献[10]。

由于Bagging方法每次从原样本集中随机抽取Bootstrap训练样本时，每棵树中约有37%的样本没有被选中，这一部分未被选中的袋外数据(out of bag，OOB)可用于估计随机森林的预测效果。文献[10]指出OOB估计是无偏估计，与用同训练集一样大小的测试集进行估计的精度是一样的。RFR中变量的重要性评分就是一种基于OOB误差的衡量方法，也是本文研究不同极化特征在土壤水分反演中的作用的基础，也称为平均下降精度(mean decrease accuracy，MDA)。其基本思路是在利用OOB测试模型中的每棵树得到OOB误差后，随机打乱OOB中某一变量的值并重新测试每棵树的OOB误差，两次OOB误差差值的平均值即为该变量的重要性评分值。

1.2 主要极化特征

综合已有的极化特征与土壤水分、地表粗糙度和植被参数等相关关系的研究[4,15-16]，本文选取其中应用较多的27个极化特征作为实验的输入参数，包括线极化后向散射系数(linear backscatter coefficients，LBC)(σ^0_{hh}、σ^0_{vv}、σ^0_{hv})、圆极化后向散射系数(circular backscatter coefficients，CBC)(σ^0_{LL}、σ^0_{RR}、σ^0_{LR})、后向散射总功率(σ^0_{span})、线极化相关系数幅度(correlation coefficients，COR)($|\rho_{hhvv}|$、$|\rho_{hhhv}|$、$|\rho_{hvvv}|$)、圆极化相关系数(circular correlation coefficients，CCC)$|\rho_{RRLL}|$、线极化相位差(phase difference，PHA)(φ_{hhvv}、φ_{hhhv}、φ_{hvvv})、雷达植被指数(radar vegetaion index，RVI)、线极化强度比(linear polarimetric ratio，LPR)(r_{hhvv}、r_{hvhh}、r_{hvvv})、圆极化强度比(circular polarimetric ratio，CPR)(r_{LLRR}、r_{LRLL}、r_{LRRR})、Cloude-Pottier分解(Cloude-Pottier decomposition，CPD)特征(H、A、α)以及Freeman-Durden分解[17](Freeman-Durden decomposition，FDD)特征(P_{surf}、P_{dbl}、P_{vol})。

1.3 利用随机森林回归的特征选择与土壤水分反演

RFR算法能够通过已有的训练数据归纳规则，得到输入与输出数据之间的对应关系。本文中随机森林回归的输入为极化特征，输出为土壤体积含水量，由此即可通过模型训练建立SAR影像极化特征与土壤水分参数之间的联系，最终进行土壤水分的反演。考虑到极化特征的量纲存在差异且绝大部分不服从正态分布，因此本文采用线性最小最大值方法对算法中输入的极化特征进行归一化处理。算法中抽取变量个数m_{try}、叶节点最小尺寸nodesize与回归树数量n_{tree}需要在实验前进行设置。

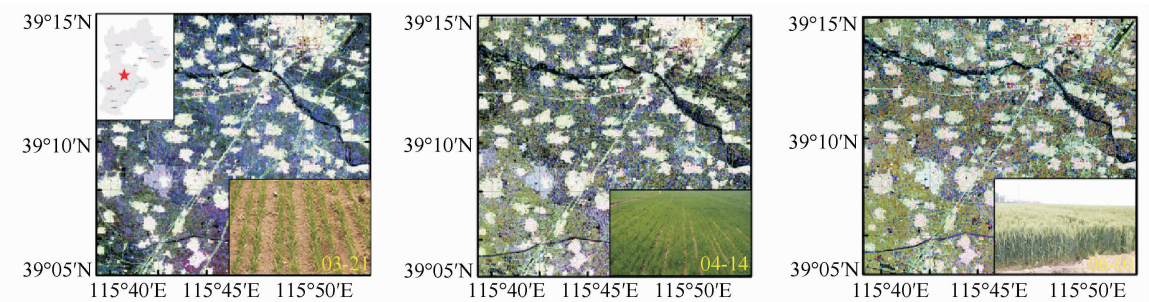
本文实验共包含27个极化特征，通过特征选择能够去掉其中冗余或不相关的特征，从而在提高计算效率的同时保证模型反演精度。一种直接的特征选择方法是根据模型训练时得到的特征重要性评分进行选择，本文以此为标准对极化特征进行组合，利用RFR构建不同的土壤水分反演模型，对比各模型的精度以验证极化特征对于土壤水分反演的贡献。进一步为验证RFR方法的可靠性，本文利用SVR与ANN方法对其反演精度进行了比较。其中，SVR算法采用的核函数为径向基函数(radial basis function，RBF)，通过重复网格参数寻优实验设置损失系数为0.000 01，惩罚系数为10，宽度系数为0.003 9。同样，通过重复实验后，选定最优的ANN网络参数设置为隐

含层两层,每层各 5 个神经元,传递函数均为 tansigmod 型函数,训练目标为 0.01,最小梯度为 0.01,最大迭代次数为 100 次。SVR 与 ANN 输入的训练样本、极化特征及其归一化方式与 RFR 保持一致且在重复实验时保持不变,每种算法均重复 10 次计算最终的均方根误差(root mean square error, RMSE)与决定系数 R^2。

2　实验区与数据集

2.1　实验区

实验区位于河北省保定市定兴县,地处北纬 39°05′39″～39°20′00″,东经 115°30′37″～115°58′06″。该区域属东部暖温带半干旱季风性气候地区,地势平坦开阔,主要农作物为小麦和玉米。河北区域是我国典型的农业家庭承包责任制管理模式,相关农业活动常常在一个较短的时间内完成,所以同一时期地块间往往具有相似的物候阶段。本文获取数据时间为 2013 年 3 月到 6 月冬小麦经历返青、拔节、孕穗、乳熟至成熟几个阶段,其中 3 月 21 日为返青期,4 月 14 日为拔节期,6 月 1 日为乳熟期,植被覆盖度与植株密度有较大差异,实地照片如图 1 所示。



(a)2013-03-21　　(b)2013-04-14　　(c)2013-06-01

图 1　Radarsat-2 全极化 PauliRGB 影像与对应地面实景

Fig. 1　PauliRGB Images of Quad-Polarimetric Radarsat-2 and Photos of Scenes

2.2　Radarsat-2 数据集

本文实验采用在 2013 年获取的 3 景 C 波段 Radarsat-2 精细四极化模式 SLC 影像(如图 1 所示),标称分辨率为 8 m,轨道均为升轨且视向为右视,中心入射角为 45.08°。三景影像的观测模式及中心入射角相同,因此本文实验中雷达入射角对土壤水分反演的影响可以忽略。Radarsat-2 影像在 NEST 软件中进行几何校正后,通过 PolSARpro v4.2 软件做 5×5 像素窗口的 Refined Lee 滤波处理,最终对各采样点取 3×3 像素窗口平均得到后向散射系数以及极化特征。

2.3　土壤水分数据集

土壤水分数据采集在卫星过境时同步进行,采样期间研究区无降雨,各时期采样时间与采样点分布如表 1 所示。为减少采样过程带来的数据不确定性,去掉存在灌溉情况的采样点,由此 3 次实验共得采样点 122 个。实验采用土盒法烘干称重土壤样本获得各验证点的土壤重量含水量,经过土壤容重(假定壤土为 1.4 g/cm^2)转换为对应的体积含水量。转换后的采样点土壤体积含水量取值范围为 9.6%～44.9%。

表 1　土壤水分同步采样信息

Tab. 1　Information of Soil Moisture Samples

采样信息	2013-03-20 至 2012-03-21	2013-04-14 至 2013-04-15	2013-05-31 至 2013-06-01
物候阶段	返青	拔节	乳熟
采样个数	47	32	43
土壤体积含水量/%	15.8～44.9	9.6～41.1	16.7～33.0
植株高度/cm	4.8～10.6	17.2～26.5	62.7～85.3

3　土壤水分反演与结果分析

3.1　模型参数设置

本文实验将包含 122 个采样点的样本集随机分为训练数据集(80 个,约占 65%)与验证数据集(42 个,约占 35%)两个部分,分别用作随机森林回归的模型训练与精度验证。随机森林中 nodesize 取值选取算法默认值 5;m_{try} 取值选取此前相关研究的推荐设置,即输入变量个数的 1/3;随着取值的增大,n_{tree} 为 2 000 时 OOB 误差的减少趋于稳定,为兼顾训练精度与时间,本文中 n_{tree} 取值均为 2 000。

3.2　极化特征重要性与特征选择结果

对不同时相的训练样本,由 OOB 误差计算得到的极化特征重要性评分如图 2 所示。在随机森林的训练过程中计算 OOB 误差时,若打乱某

一特征的取值使得 OOB 误差显著增加，则认为该特征对反演正确的贡献较大，特征重要性即由平均精度下降进行度量。观察特征重要性结果可以看出，后向散射功率信息对各时期的冬小麦土壤水分反演精度贡献都很突出。对于3月返青期的数据，LBC 与 CBC 的重要性明显高于其他特征：LBC 是目前大多数反演模型的输入参数，其重要性已被此前的研究广泛认可；而 CBC 是所有3个线极化及其相位的组合，因此 LBC 与 CBC 之间具有高度的相关性，也显示出了较高的重要性。该时期小麦植被较为稀疏，包含土壤与植被散射信息的极化分解特征并不占绝对优势，但 FDD 的3个分量以及 CPD 中的 α 参数依然显示了不可忽视的贡献：FDD 的3个分量分别代表了表面、二次及体散射机制的强度，对于在反演中区别出植被的贡献具有一定帮助，且该3分量在训练样本中与线极化具有较高相关性，同样能够引起 FDD 3分量的高重要性评分结果；CPD 中的 α 表征了地表由表面散射到二面角散射的变化过程，较低的 α 将显示更多土壤贡献的成分，因此也会对返青期表面散射占优的土壤水分反演起到一定作用。相较而言，极化通道间的 COR、PHA 以及 LPR 并没有显示出较高的重要性，其中 $|\rho_{hhvv}|$ 与区域的匀质性有关，若匀质程度没有较大差异，该特征也无法提供有益信息；φ_{hhvv} 与表面的粗糙程度相关，而实验区的耕作模式基本一致，粗糙度特征的区分对土壤水分信息的提取帮助不大。拔节期的结果与返青期略有不同，差异主要在于 σ^0_{hv} 与 P_{vol} 的重要性有了显著提高，原因在于拔节期小麦植被层的影响更强，σ^0_{hv} 与 P_{vol} 有助于去除植被层的影响。乳熟期的结果则显示各种极化特征对于土壤水分反演精度的提升并不如前两个时期明显，可能的原因在于该时期样本的土壤水分分布范围相对较窄，且植被层的一致性较高（见表1中土壤水分与株高信息），使得该时期只需要少量极化特征就能对土壤水分进行较好的拟合（图3的乳熟期验证集反演精度在所有组合中均为最高）。总体样本的结果与各时期基本相同，总功率、LBC 与 CBC 依然占据着重要的位置，它们之间的相关性本身也较高；φ_{hhvv} 虽然在单一时期内对土壤水分反演的贡献并不突出，但总体样本非均匀程度的差异使其表现出了较显著的重要性，描述散射随机程度的 H 以及散射机制变化的 α 重要性的提升也有相似的原因；P_{surf} 能够将土壤的表面散射信息从植被中分离出来，因此也具有较高的重要性，而描述植被散射贡献的 P_{vol} 与 RVI 能够帮助在总体样本中量化不同时期植被影响的差异，同样显示了高重要性。

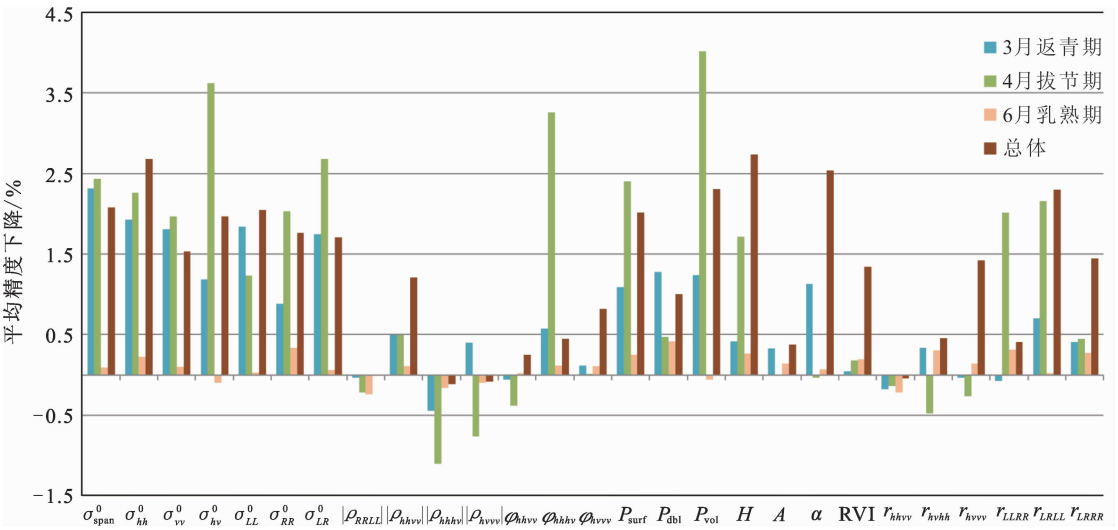


图2 SAR影像主要极化特征的重要性评分

Fig. 2 Importance Scores of Main Polarimetric Features of SAR Images

根据以上结果，本文分别选取各时相及总体重要性前3（most important 3，MI03）、前6（MI06）的特征建立反演模型，同时与传统反演模型输入中的线极化特征（LBC）和全部27个极化特征（MI27）所构建的模型进行反演精度对比，验证依据随机森林重要性评分的特征选择的效果。总体来说，CBC、FDD、CPD 及 CPR 在不同时期的重要性评分都较高，因此本文以传统模型输入特征＋一组重要性较高的极化特征 的形式进行特征组合，构建多种反演模型，包括 LBC-FDD、LBC-CPD、LBC-CPR、L-CBC 共4种特征组合模型来验证极化特征对土壤水分反演的贡献。另一

方面,部分重要性评分较高的特征具备与 LBC 或 CBC 的高相关性,如 FDD 与 CPD 分解特征。因此,为了防止 RFR 由于特征间的高相关性而在评分时出现高估的情况,考虑加入与其他特征相关性较低的特征组合进行对比,包括 LBC-COR、LBC-PHA、LBC-LPR 3 种组合,最终验证各模型的反演精度。

3.3　土壤水分反演结果

根据上述特征组合分别构建基于 RFR、SVR 与 ANN 的土壤水分反演模型,以 RMSE 与 R^2 评估各模型反演精度,结果如图 3 所示。从总体样本的反演结果来看,RFR 对训练数据能够进行较好的拟合,训练数据集的 RMSE 均小于 5%,且 R^2 均高于 0.80;验证数据集的 R^2 普遍较低,说明 RFR 在训练时存在过拟合的问题,但总体样本 RMSE 都在 6%左右,还是能够证明这种算法对于土壤水分反演的有效性。在所有模型中,MI27 的精度在训练与验证数据集中都不是最高的,说明输入特征越多并不意味着更高的反演精度,特征选择有其必要性;MI03 与 LBC 均只包含 3 个

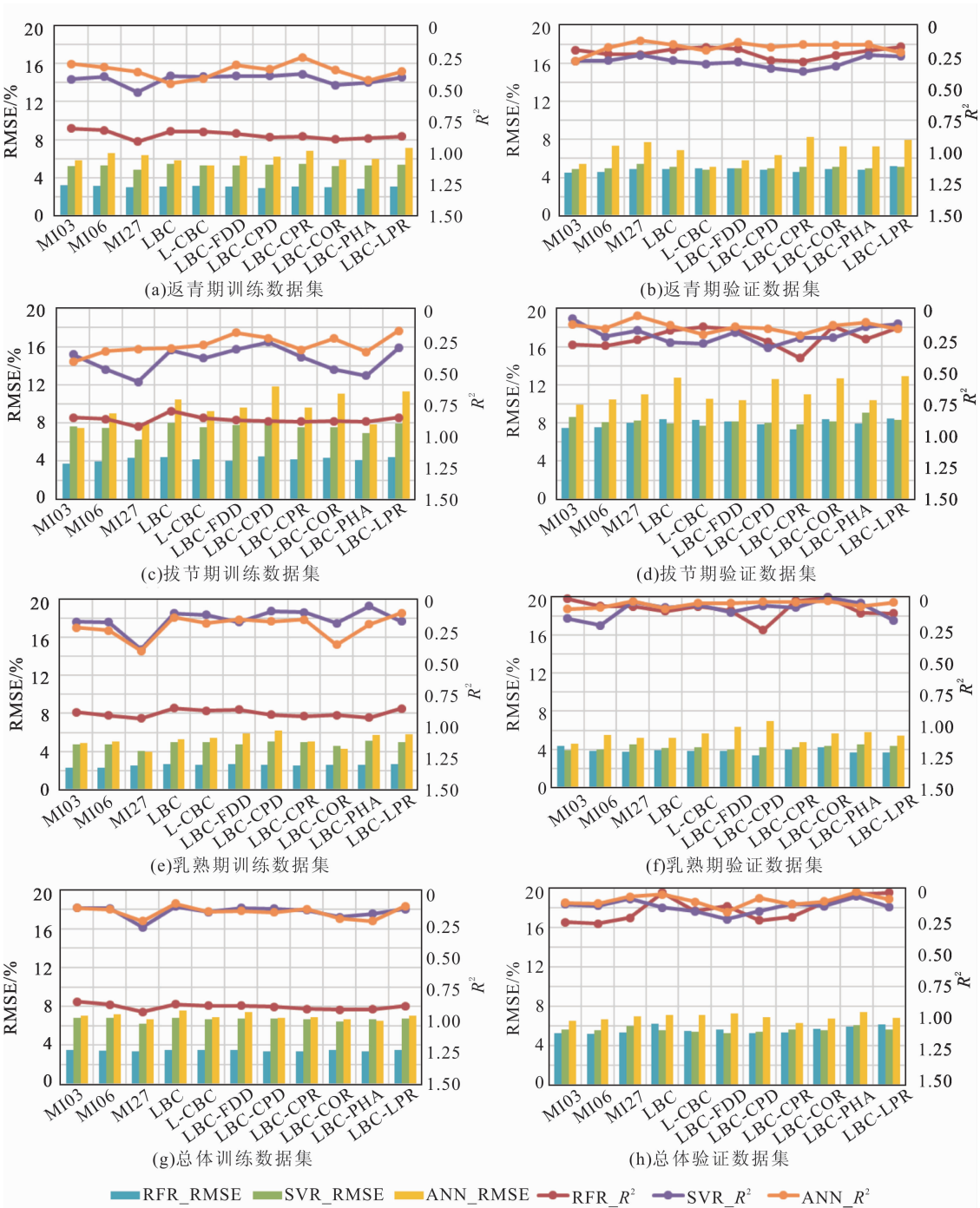


图 3　RFR、SVR 与 ANN 的土壤水分反演精度对比

Fig. 3　Retrieving Accuracy of RFR, SVR and ANN

输入特征，而MI03的验证集RMSE更低且R^2更高，证明了基于重要性评分的特征选择的有效性；MI06相比MI03输入特征增加，RMSE总体得到了一定提高，与6种LBC联合组合模型（输入特征个数均为6个）相比，验证集RMSE也达到了最低的5.23%；6种LBC联合组合模型中，L-CBC、LBC-FDD、LBC-CPD与LBC-CPR的精度提升较大，LBC-PHA与LBC-LPR的提升则并不明显。其中，圆极化散射矩阵中的*LR*对应目标平面分量，*RR*与*LL*则直接代表了Krogager分解[18]中的二面角散射分量。以上分量在冬小麦不同长势条件下的差异是L-CBC模型在总体反演精度上相对LBC有较大提升的可能原因。LBC-FDD则包含了描述表面散射、二次散射与体散射分量的信息，同样能够对不同时期的植被与土壤效应进行一定区分，为总体的反演精度带来提升。LBC-CPD包含了表征散射机制的α参数，同时CPD特征与LBC的线性相关性不高，使得模型中包含的信息冗余度相较其他特征组合更低，因此在各时期的反演结果都对LBC模型有较明显的提升。CPR特征通过比值处理相比CBC降低了与LBC之间的相关性，同样在反演精度上得到了较好的结果。比较不同时期数据的反演结果，乳熟期各模型的精度相对最高，其次为返青期模型。与训练集相比，拔节期的验证集反演精度有较为明显的下降，RFR的过拟合现象相对其他时期更严重，可能的原因是这一时期的土壤水分分布范围相对更宽（见表1），而采样点相对较少，样本数量的不足使得RFR难以找到顾及全局情况的拟合结果，从而导致更严重的过拟合。各时期不同模型反演精度的高低趋势与总体结果近似。值得注意的是，模型的反演精度与所包含的极化特征的重要性评分有一定相关性。例如，拔节期的FDD中体散射分量评分明显高于同时期其他特征，因此LBC-FDD模型的训练集反演精度也相对其他LBC组合模型更高；而其他时期的FDD体散射分量评分则没有明显占优，因此对应训练集反演精度也没有突出表现。这一点一方面证明了RFR重要性评分在参数反演中的参考价值，另一方面也反映了以经验数据为指导的RFR特征选择的局限性。

本文实验也将SVR与ANN的反演精度与RFR进行了对比。不难发现，即使通过重复实验设置了SVR与ANN的最优参数，RFR仍然在不同时期、不同模型的训练中取得了最好的训练效果，不仅达到了最低的RMSE，所有训练集的拟合优度R^2都在接近0.90的水平，而SVR与ANN则全部低于0.50。虽然相比其他两种算法的过拟合程度较高，但RFR在验证集中普遍取得了最低的RMSE，也证明了这种算法在本实验区土壤水分反演中的适用性。然而并不能以此证明RFR在机制上对其他算法有优势。此前，这3种算法在植被参数及土壤水分的相关研究中均有一定涉及，大多以两两对比的形式存在：文献[19]在土壤水分产品降尺度研究中应用到了ANN与SVR算法，分别取得了1.1%与1.3%的RMSE；文献[20]则对比了SVR与RFR在森林属性分配中的表现，结果显示两种算法对不同属性的提取精度互有高低。因此，RFR的优势还需要在不同的研究区进行进一步验证。3种算法不同模型的反演精度趋势大致相同，但在拔节与乳熟期出现了部分不一致的情况，表明基于经验数据的RFR重要性评分有一定局限性，所选择的特征在其他算法中的表现不一定最优。综合来看，RFR不管在训练还是验证集中均保持了较高的反演精度，且算法的参数相对较少、设置简单，在土壤水分反演的研究中具备适用性。

4 结 语

本文利用RFR算法能够评估变量重要性程度的特点，针对冬小麦区域土壤水分反演问题分析了常用SAR极化特征的重要性，结果表明除了传统的LBC特征外，CBC、FDD、CPD和CPR对土壤水分反演的重要性也较高，是值得后续继续研究的极化特征。本文在重要性评分的基础上进行特征选择，并利用RFR算法建立不同的反演模型，分析讨论了不同特征组合的土壤水分反演结果：基于重要性评分的特征选择的确能够减少特征个数并保持精度；联合CBC、FDD、CPD和CPR的反演模型精度相较传统LBC模型均有提高，证明了这些重要性较高特征在土壤水分反演中的贡献，展现了利用极化特征进行土壤水分反演的潜力。最终通过与SVR、ANN方法的对比，进一步证明了RFR在本文实验区土壤水分反演研究中的适用性。但由于本文实验是在经验数据的基础上展开的，极化特征在不同时期的表现有一定差异，模型的反演精度与极化特征的重要性评分有相关性，因此反演模型与相关结论只针对本实验区的情况，其可扩展性还需顾及更多情况。

参 考 文 献

[1] Ren Xin. A Surface Moisture Inversion Teclmique Using Multi-Polarization and Multi-Angle Radar Images[D]. Beijing: Institute of Remote Sensing Application, Chinese Academy of Sciences, 2003 (任鑫. 多极化多角度 SAR 土壤水分反演算法研究[D]. 北京:中国科学院遥感应用研究所, 2003)

[2] Wei Xiaolan, Li Zhen, Chen Quan. The Simulation Analysis and Validation of Soil Moisture Retrieval Using S-band Radar[J]. *Geo-information Science*, 2008, 10(1):97-101 (魏小兰, 李震, 陈权. S 波段雷达数据反演土壤水分的模拟分析和验证[J]. 地球信息科学学报, 2008, 10(1):97-101)

[3] Bourgeau-Chavez L L, Leblon B, Charbonneau F, et al. Evaluation of Polarimetric Radarsat-2 SAR Data for Development of Soil Moisture Retrieval Algorithms over a Chronosequence of Black Spruce Boreal Forests[J]. *Remote Sensing of Environment*, 2013, 132(1): 71-85

[4] Wiseman G, McNairn H, Homayouni S, et al. Radarsat-2 Polarimetric SAR Response to Crop Biomass for Agricultural Production Monitoring[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2014, 7(11): 4 461-4 471

[5] Adams J R, Berg A A, McNairn H, et al. Sensitivity of C-band SAR Polarimetric Variables to Unvegetated Agricultural Fields[J]. *Canadian Journal of Remote Sensing*, 2013, 39(1): 1-16

[6] Baghdadi N, Dubois-Fernandez P, Dupuis X, et al. Sensitivity of Main Polarimetric Parameters of Multifrequency Polarimetric SAR Data to Soil Moisture and Surface Roughness over Bare Agricultural Soils [J]. *IEEE Geoscience and Remote Sensing Letters*, 2013, 10(4): 731-735

[7] Cloude S R, Pottier E. An Entropy Classification Scheme for Land Applications of Polarimetric SAR Data[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 1997, 35: 68-78

[8] Notarnicola C, Angiulli M, Posa F. Soil Moisture Retrieval from Remotely Sensed Data: Neural Network Approach Versus Bayesian Method[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2008, 46(2): 547-557

[9] Ahmad S, Kalra A, Stephen H. Estimating Soil Moisture Using Remote Sensing Data: A Machine Learning Approach[J]. *Advances in Water Resources*, 2010, 33(1): 69-80

[10] Breiman L. Random Forest[J]. *Machine Learning*, 2001, 45(1): 5-32

[11] Baghdadi N, Cresson R, El-Hajj M, et al. Estimation of Soil Parameters over Bare Agriculture Areas from C-band Polarimetric SAR Data Using Neural Networks[J]. *Hydrology and Earth System Sciences*, 2012, 16: 1 607-1 621

[12] Pasolli L, Notarnicola C, Bruzzone L, et al. Polarimetric Radarsat-2 Imagery for Soil Moisture Rtrieval in Alpine Areas[J]. *Canadian Journal of Remote Sensing*, 2011, 37: 535-547

[13] Srivastava P K, Han D, Ramirez M R, et al. Machine Learning Techniques for Downscaling SMOS Satellite Soil Moisture Using MODIS Land Surface Temperature for Hydrological Application[J]. *Water Resources Management*, 2013, 27: 3 127-3 144

[14] Karjalainen M, Kankare V, Vastaranta M, et al. Prediction of Plot-Level Forest Variables Using TerraSAR-X Stereo SAR Data[J]. *Remote Sensing of Environment*, 2012, 117: 338-347

[15] Baghdadi N, Cerden O, Zribi M, et al. Operational Performance of Current Synthetic Aperture Radar Sensors in Mapping Soil Surface Characteristics in Agricultural Environments: Application to Hydrological and Erosion Modelling[J]. *Hydrological Processes*, 2008, 22: 9-20

[16] Yang Guijun, Shi Yuechan, Zhao Chunjiang, et al. Estimation of Soil Moisture from Multi-Polarized SAR Data over Wheat Coverage Areas[C]. The First International Conference on Agro-Geoinformatics, Shanghai, China, 2012

[17] Freeman A, Durden S L. A Three-Component Scattering Model for Polarimetric SAR Data[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 1998, 36(3): 963-973

[18] Krogager E. A New Decomposition of the Radar Target Scattering Matrix[J]. *Electronics Letter*, 1990, 26(18): 1 525-1 526

[19] Srivastava P, Han D, Ramirez M R, et al. Machine Learning Techniques for Downscaling SMOS Satellite Soil Moisture Using MODIS Land Surface Temperature for Hydrological Application[J]. *Water Resource Management*, 2013, 27: 3 127-3 144

[20] Shataee S, Kalbi S, Fallah A, et al. Forest Attribute Imputation Using Machine-Learning Methods and ASTER Data: Comparison of K-NN, SVR and Random Forest Regression Algorithms[J]. *International Journal of Remote Sensing*, 2012, 33: 6 254-6 280

Soil Moisture Retrieval of Winter Wheat Fields Based on Random Forest Regression Using Quad-Polarimetric SAR Images

LI Pingxiang[1] *LIU Zhiqu*[1] *YANG Jie*[1] *SUN Weidong*[1] *LI Minyi*[2] *REN Yexian*[1]

1 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

2 Deqing iSpatial Co. Ltd, Deqing 313200, China

Abstract: Soil moisture has great significance in the researches of hydrology, meteorology and agriculture yield estimation. The quad-polarimetric SAR images can provide a lot of polarimetric features, the significance of the features in surface parameter retrieval have attracted attentions in previous researches with no final conclusions because of the complexity of terrain scattering. In this paper, random forest regression (RFR) is used for both soil moisture retrieval and the importance evaluation of polarimetric features of Radarsat-2 images in winter wheat fields. According to the score of importance, feature selection and combination are done for modelling. We evaluate the retrieval accuracy of models with different feature combinations. The results show that models of important features selected by RFR have RMSE(root mean square error) less than 6% which are better results compared to traditional models; when compared with support vector regression and artifical neural networks, the RFR also shows best retrieval accuracies, which proves that RFR is suitable for soil moisture retrieval and feature selection. The high retrieval accuracies of LBC-CPD(linear backscatter coefficients-Cloude-Pottier decomposition) and LBC-CPR(linear backscatter coefficients-circular polarimetric ratio) indicates these features can improve the retrieval accuracy of soil moisture.

Key words: polarimetric SAR; soil moisture; random forest regression; support vector regression; artificial neural networks

First author: LI Pingxiang, professor, specializes in the theories and methods of polarimetric SAR. E-mail: pxli@whu.edu.cn

Corresponding author: LIU Zhiqu, PhD candidate. E-mail: meloqu@qq.com

Foundation support: The National Natural Science Foundation of China, Nos. 41771377, 41601355, 91438203, 41501382; the GF Satellite Program from State Administration of Science, Technology and Industry for National Defense of China, No. 03-Y20A10-9001-15/16.