引文格式：虞瑶,陈超,李梦梦. 基于验证集信息的高光谱遥感动态集成算法[J]. 地理空间信息, 2024, 22(9): 33-36. doi:10. 3969/j.issn. 1672-4623.2024.09.007

基于验证集信息的高光谱遥感动态集成算法

虞　瑶[1]，陈　超[1]，李梦梦[1]
（1. 江苏省基础地理信息中心，江苏 南京 210013）

摘　要：为了提升动态集成算法的分类精度，提出了一种基于验证集信息的两阶段动态集成算法（TS）。首先通过同时限制分类器精度和相似度自适应调整KNN大小，优先判别光谱性能较优的待分类样本；再以优先判别的待分类样本为新的验证样本补充扩大验证集，从而充分利用验证集信息增加分类的准确性和可靠性。采用两组高光谱遥感影像数据对该算法进行实验，并与3种多分类器动态集成算法进行对比分析。结果表明，与经典动态集成算法相比，TS算法具有更好的分类表现，对于推动高光谱遥感精细分类具有一定的理论和实用价值。

关键词：集成学习；高光谱遥感分类；动态集成；KNN；两阶段
中图分类号：P237　　**文献标志码：**B　　**文章编号：**1672-4623（2024）09-0033-04

Dynamic Integrated Algorithm of Hyperspectral Remote Sensing Images Based on Validation Set Information

YU Yao[1], CHEN Chao[1], LI Mengmeng[1]
(1. Basic Geographic Information Center of Jiangsu Province, Nanjing 210013, China)

Abstract: In order to improve the classification accuracy of dynamic integrated algorithm, we proposed a two-stage dynamic integrated algorithm based on verification set information (TS). Firstly, the classifier accuracy and similarity are limited to change the size of KNN adaptively, and the samples are classified with better spectral performance preferentially. Then, the priority classification sample is used to expand the verification set, and the verification set information is fully utilized to increase the accuracy and reliability of classification. We used two sets of hyperspectral remote sensing image data to verify TS algorithm, and compared with three multi-classifier dynamic integrated algorithms. The results show that compared with the traditional dynamic integrated algorithms, TS algorithm has better performance, which is useful for promoting hyperspectral remote sensing image fine classification.

Key words: integrated learning, hyperspectral remote sensing classification, dynamic integration, KNN, two-stage

高光谱遥感是遥感技术的前沿领域，被广泛应用于地质、环境、生态、海洋、测绘和军事等多个领域[1-3]。虽然高光谱遥感影像在精细分类方面具有明显优势，但光谱数据维度高、波段冗余等问题不仅增加了数据计算成本，而且在影像分类时存在“休斯现象”，导致分类精度受限。通常采取两种方式提升高光谱遥感影像的分类表现：开发新型高光谱遥感影像分类算法和采用集成学习的方式[4]。根据集成过程中是否对待测样本的具体特征进行自适应选取，多分类器集成方法分为静态集成和动态集成。动态集成可根据待分类样本的特征，动态选择部分基分类器参与该样本的标签预测，比静态集成具有更强的针对性和灵活性。1997 年 Woods 等设计了总体局部精度（OLA）和局部类精度（LCA）两种方法计算多分类器的局部精度。2000 年 Kuncheva L I[5]提出了基于聚类与选择的动态集成。2001 年 Giacinto G[6]等提出了基于分类器行为的动态集成算法（MCB），根据 MCB 的信息计算局部类精度可极大减弱能力区域内“假邻居”的影响。2002 年 Smits P C[7]对 Woods 提出的局部精度估计算法进行了改进，根据正确分类的特征空间样本到待测样本的距离计算各分类器的权值，将权值最高的分类器的输出结果作为待测样本的标签值。

动态集成中验证集的选择至关重要，但鲜有相关文献在动态集成过程中利用验证集信息[8-10]。本文综合利用 KNN 和验证集信息，提出了一种基于验证集的两阶段动态集成算法（TS）。首先同时限制分类器精度和相似度自适应调整待分类样本的 KNN 大小，优先判别光谱性能较优的待分类样本；再以优先判别的待

收稿日期：2023-01-03。
项目来源：江苏省自然资源科技资助项目（2022005）。
第一作者简介：虞瑶（1995—），硕士研究生，工程师，主要从事GIS、摄影测量与遥感等方面工作，E-mail：yuyaoyao_yy@163.com。

分类样本为新的验证样本扩大验证集的规模。本文利用HyMap和AVIRIS两个传感器的高光谱遥感影像进行实验，并与经典动态集成算法进行对比分析。结果表明，TS算法比经典动态集成算法的分类表现更优。

1 动态集成算法

1）基于局部精度的动态集成算法。Woods等首次提出了基于局部精度的动态集成算法。首先在验证集中选取 K 个验证样本构成能力区域，再选择能力区域内分类精度最高的基分类器，将该基分类器的输出结果作为待分类样本的标签值。该算法设计了OLA和LCA两种局部精度计算方法。Smits P C[7]提出改进的局部精度算法（MLA），主要考虑到能力区域内每个验证样本的可靠性不同，验证样本与待分类样本的距离越近，相似性越大，可靠性越大，所占的权重越大。

2）MCB算法。Giacinto G[6]等提出了MCB算法。其核心思想是：若多个分类器判定该待分类样本与能力区域内某些验证样本不属于同一类别，则这些验证样本很有可能不属于同一类。为形成有效的能力区域，有必要剔除能力区域内这些“假邻居”。

3）基于Oracle的动态集成方法。Oracle概念认为只要有一个基分类器能全部正确分类测试样本，则该集成系统就能将测试样本全部正确分类[11]。Oracle的性能上限过于理想，是很难达到的。因此，将其改进为若某些基分类器可将 K 个邻域样本全部正确分类，则采用简单投票法的方式决定测试样本的最终类别；若不存在这样的基分类器，则不断减小 K 值，直至出现能将邻域样本全部正确分类的分类器为止。

2 TS算法

2.1 基本原理

通常基于KNN准则的动态集成算法受邻域样本数目 K 值的影响较大[12-13]，其中基于Oracle的动态集成方法构建KNN时仅考虑分类器的分类精度，而MCB算法则通过考虑分类器间的相似度来去除可靠性较差的验证样本，但分类精度受限于预设的相似度阈值。本文提出的TS算法同时考虑了分类器的精度和相似度，自适应地改变测试样本的KNN，使得KNN的设置更加合理有效。验证集质量同样对基于KNN准则的动态集成算法效果具有一定影响，因此本文提出TS算法。

2.2 算法步骤

假设验证集 $V=\{(x_1,l_1),(x_2,l_2),\cdots,(x_{N_v},l_{N_v})\}$，测试样本为 y，分类器集合为 $C=\{c_1,c_2,c_i,\cdots,c_M\}$，测试样本在验证集中的K邻域为 $\theta=\{x_1,x_2,\cdots,x_K\}$，基分类器的能力估计值为 δ_i，$i=1,2,\cdots,M$，基分类器集成 C 对测试样本的分类结果为 $\tilde{y}=\{\tilde{y}_1,\tilde{y}_2,\cdots,\tilde{y}_M\}$。TS算法的具体步骤为：①输入训练集 X，验证集 V，测试样本 y，设置邻域值为 K；②利用Bagging的集成方式，生成 M 个TCRC分类器；③在 V 中找到 y 的 K 邻域 $\theta=\{x_1,x_2,\cdots,x_K\}$，作为能力区域；④对各分类器意见一致的测试样本直接赋予分类器的结果，并在分类结果中进行标记，作为标记点；⑤同时考虑分类器的分类精度和分类器行为，要求局部分类总体精度高于0.9和分类器行为相似度高于0.7；⑥若没有满足条件的 K 值，则去除KNN内最不相似的样本，邻域大小变为 $\theta'=\{x_1,x_2,\cdots,x_{K-1}\}$，$K$ 变为 $K-1$，若存在满足条件的 K 值，则选择KNN中精度最高的分类器，对 y 赋与标签值，至此一部分测试样本被分类为标记点，另一部分测试样本未找到满足条件的分类器，仍未被分类；⑦将标记点与原始验证集合并构成新的验证集 V'；⑧在 V' 中找到 y 的K邻域 $\theta''=\{x_1,x_2,\cdots,x_K\}$，作为能力区域；⑨重复步骤⑤、⑥；⑩若不存在满足条件的KNN，则采用简单投票法的方式确定待分类样本的标签值。

3 实验结果与分析

3.1 实验设置

本文基于Purdue Campus、Salinas两幅高光谱遥感影像数据，将TS算法与OLA、MLA、MCB等动态集成算法进行对比分析。基分类器选择TCRC[14]，利用测试样本的切平面估计区域流形，在测试样本的切空间中使用协同表示算法。采用Bagging的集成方式生成分类器集合，集成方式选择静态集成中的多数投票法，记为MV方法[15]。本文采用的分类性能评价指标包括总体精度（OA）、平均精度（AA）和Kappa系数。所有实验重复运行10次，并将结果取平均作为实验结果。

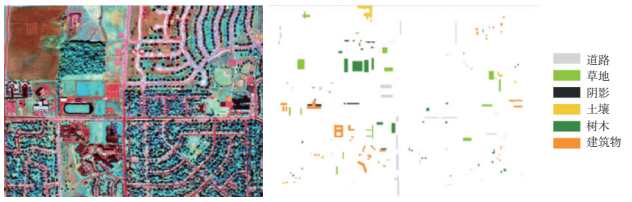
3.2 Purdue Campus高光谱影像实验

Purdue Campus西拉斐特分校校区高光谱遥感数据的光谱区间为450~2 480 nm，空间分辨率为3.6 m；共提供126条光谱波段，空间尺寸为377像素×512像素（图1）。根据解译图，实验选取90个像元作为训练样本，平均每类选取15个，每类选取15个像元作为验证样本，其余像元用于精度评定。6种方法的分类结果见表1，其中OLA、MLA、MCB和TS四种动态集成方法的 K 值设置为5，可以看出，TS算法的

OA、*AA* 和 Kappa 系数均优于另外3种动态集成算法；从各地类的分类准确度来看，TS算法有4种地物类别的分类准确度均高于其他算法，其中第1、2、3类地物的分类准确度提升明显；OLA、MCB 和 MLA 算法的 *OA* 低于 MV 算法，原因可能是训练集规模不足，影响了动态集成的分类效果。6种算法的分类结果见图2，可以看出，TS算法的分类效果图更清晰、更平滑、细节更丰富。

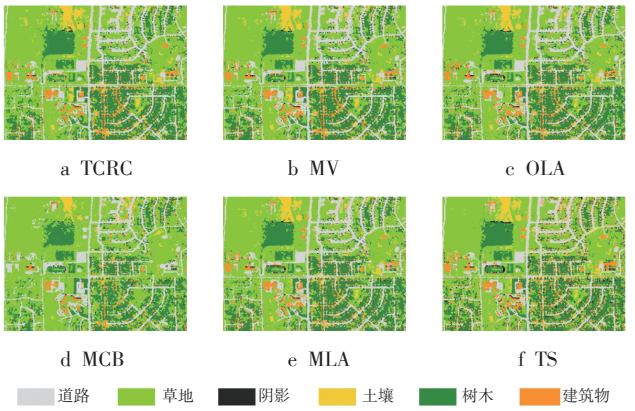
表1　Purdue Campus 分类精度统计表

类别	TCRC	MV	OLA	MCB	MLA	TS
1	89.66	90.04	96.56	96.26	96.56	97.31
2	84.62	86.04	96.57	95.77	96.49	97.60
3	91.10	87.93	63.70	59.36	64.76	84.34
4	86.20	87.18	95.42	94.50	95.42	98.40
5	98.24	98.63	98.99	98.30	99.00	99.68
6	94.14	96.83	77.07	79.38	79.14	88.54
OA	91.04	92.11	91.45	91.38	91.96	95.55
AA	90.65	91.11	88.05	87.26	88.56	94.31
Kappa 系数	0.886 6	0.900 1	0.892 1	0.891 0	0.898 4	0.943 7



a 假彩色影像　　b 地面样本数据分布

图1　Purdue Campus假彩色影像图与地面样本数据分布图



a TCRC　b MV　c OLA

d MCB　e MLA　f TS

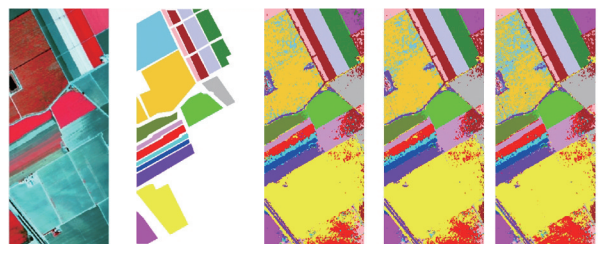
图2　Purdue Campus 数据6种算法分类土地覆盖图

3.3　Salinas 高光谱影像实验

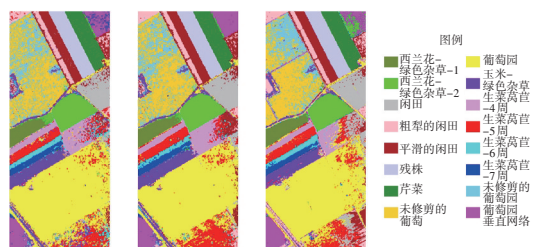
Salina 数据集是由 AVIRIS 传感器于1998年在美国加州地区的一块测试地上采集的，影像大小为512像素×217像素，空间分辨率为3.7 m，波长范围为0.4~2.5 um（图3），实验选取其中的204个波段。6种算法的分类结果见表2，可以看出，OLA、MCB、MLA 和 TS算法的分类准确度明显高于静态集成的 MV 算法，TS算法的 *OA*、*AA* 和 Kappa 系数比 OLA 算法分别提高了 0.11%、4.34% 和 0.06；比 MCB 算法分别提高了 0.61%、4.82% 和 0.68；比 MV 算法分别提高了 1.03%、5.4%和0.0116；但 *OA* 比 MLA 算法略低，*AA* 则提高了约4%。从各地物类别的分类准确度来看，TS算法有4种地物类别的分类准确度均高于其他算法。6种算法的分类结果见图3，可以看出，TS算法的分类效果图噪声更少、边界更清晰。

表2　Salinas 数据各算法的分类精度统计表

类别	TCRC	MV	OLA	MCB	MLA	TS
1	99.91	99.94	99.24	97.65	99.24	99.53
2	99.21	99.24	99.97	99.48	99.97	99.95
3	97.04	97.62	97.55	93.72	97.80	98.05
4	98.19	98.20	93.61	93.77	93.41	94.80
5	94.97	94.47	98.40	97.38	98.37	98.17
6	99.89	99.96	100	100	100	100
7	99.79	99.79	99.98	99.88	99.96	99.88
8	65.68	68.27	68.73	68.51	71.14	69.26
9	94.89	94.83	99.91	99.53	99.91	99.92
10	91.00	91.23	87.57	87.91	88.59	87.85
11	96.03	96.32	98.09	97.42	98.21	98.30
12	95.76	95.36	85.56	91.00	87.98	94.43
13	76.05	77.49	92.19	91.96	91.47	91.92
14	84.10	85.58	93.49	94.11	93.96	93.60
15	59.17	60.06	60.35	57.97	60.29	57.62
16	100	100	97.48	97.39	97.72	97.56
OA	84.51	85.19	86.11	85.61	86.75	86.22
AA	90.73	91.15	92.21	91.73	92.38	96.55
Kappa 系数	0.827 6	0.835 2	0.846 2	0.840 0	0.852 6	0.846 8



a 假彩色影像图　b 地面样本数据分布图　c TCRC　d MV　e OLA



f MCB　g MLA　h TS

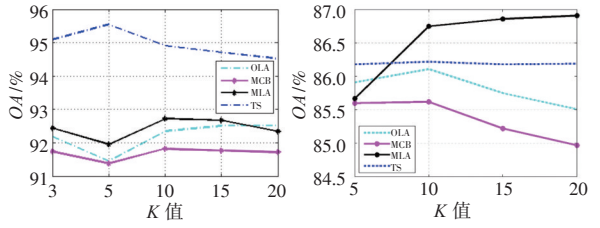
图3　Salinas 数据算法分类土地覆盖图

3.4　参数分析

1）邻域参数。OLA、MCB、MLA 和 TS 四种动态集成算法的 *OA* 值随邻域参数的变化情况见图4，可以

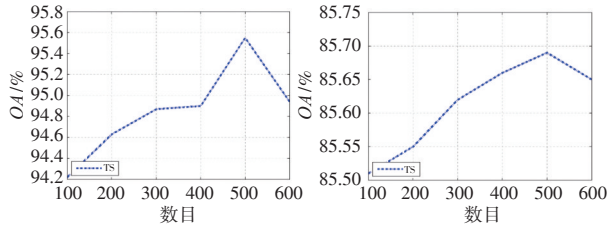
看出，对于Purdue Campus高光谱数据，4种动态集成算法分类精度随 K 值的变化趋势不同，TS算法的 OA 随 K 值的增加而降低，最佳的邻域参数为5；对于Salinas高光谱数据， K 值对TS算法的影响不大，精度基本稳定在86.2%。

2）数目。两组数据中TS算法的 OA 值随数目的变化趋势见图5，可以看出，对于Purdue Campus高光谱数据，数目为100~500时， OA 值随数目的增加而增加，但数目继续增大， OA 值反而降低，其原因可能是随着数值的增大，验证集里加入了过多的可靠性差的验证样本，最佳的数目为500；对于Salinas高光谱数据，数目设置为500最佳。



a Purdue Campus b Salinas

图4 两组数据算法领域参数–OA变化曲线图



a Purdue Campus b Salinas

图5 两组数据算法数目–OA变化曲线图

4 结 语

本文提出了TS算法，通过优先分类的待分类样本扩大验证集的规模，并利用两套高光谱遥感影像数据进行了实验验证。结果表明，对于不同的实验数据，TS算法的精度提升幅度虽有不同，但在实验中均取得优于经典动态集成算法的结果。该算法还存在一定的不足，如两阶段寻找KNN不可避免地导致计算复杂度升高、计算效率下降，因此如何减小该算法的计算复杂度有待进一步研究。

参考文献

[1] 童庆禧,张兵,郑兰芬. 高光谱遥感[M]. 北京:高等教育出版社,2006

[2] 杜培军,夏俊士,薛朝辉,等. 高光谱遥感影像分类研究进展[J]. 遥感学报,2016,20(2):236-256

[3] Moon H, Ahn H, Kodell R L, et al. Ensemble Methods for Classification of Patients for Personalized Medicine with High-dimensional Data[J]. Artificial Intelligence in Medicine,2007,41(3):197-207

[4] Chi M,Kun Q,Benediktsson J A,et al. Ensemble Classification Algorithm for Hyperspectral Remote Sensing Data[J]. IEEE Geoscience and Remote Sensing Letters, 2009, 6(4): 762-766

[5] Kuncheva L I. Clustering-and-selection Model for Classifier Combination[C]. International Conference on Knowledge Based and Intelligent Information and Engineering Systems,2000:185-188

[6] Giacinto G, Roli F. Dynamic Classifier Selection Based on Multiple Classifier Behavior[J]. Pattern Recognition, 2001, 34(9):1 879-1 881

[7] Smits P C. Multiple Classifier Systems for Supervised Remote Sensing Image Classification Based on Dynamic Classifier Selection[J]. IEEE Transactions on Geoscience and Remote Sensing,2002,40(4):801-813

[8] Dos Santos E M, Sabourin R, Maupin P. A Dynamic Overproduce-and-choose Strategy for the Selection of Classifier Ensembles[J]. Pattern Recognition,2008,41(10):2 993-3 009

[9] Hou C, Xia Y, Xu Z, et al. Learning Classifier Competence Based on Graph for Dynamic Classifier Selection[C]. The 12th IEEE International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, 2016:1 164-1 168

[10] Ko A H R, Sabourin R, Britto A S, et al. From Dynamic Classifier Selection to Dynamic Ensemble Selection[J]. Pattern Recognition,2008,41(5):1 718-1 731

[11] Cruz R M O, Sabourin R, Cavalcanti G D C. Dynamic Classifier Selection: Recent Advances and Perspectives[J]. Information Fusion,2018,41:195-216

[12] Didaci L, Giacinto G. Dynamic Classifier Selection by Adaptive K-nearest-neighbourhood Rule[C]. International Workshop on Multiple Classifier Systems, Springer, Berlin, Heidelberg,2004:174-183

[13] Jr A S B, Sabourin R, Oliveira L E S. Dynamic Selection of Classifiers: A Comprehensive Review[J]. Pattern Recognition,2014,47(11):3 665-3 680

[14] Su H J, Zhao B, Du Q, et al. Tangent Distance-based Collaborative Representation for Hyperspectral Image Classification[J]. IEEE Geoscience and Remote Sensing Letters, 2016,13(9):1 236-1 240

[15] Su H, Yu Y, Du Q, et al. Ensemble Learning for Hyperspectral Image Classification Using Tangent Collaborative Representation[J]. IEEE Transactions on Geoscience and Remote Sensing,2020,58(6):3 778-3 790