基于中位数绝对偏差的异常训练样本探测方法

龚循强[1,2*]，张方泽[1,2]，鲁铁定[1,2]，陈志高[2]

[1] 东华理工大学放射性地质与勘探技术国防重点学科实验室，江西 南昌 330013；

[2] 东华理工大学测绘工程学院，江西 南昌 330013

摘要 遥感图像的监督分类技术在信息提取和变化检测领域中具有广泛的应用，其中训练样本的选择至关重要，训练样本的好坏直接决定分类精度的高低。然而，受到条件的限制和人为的错误均可能导致一些不纯或错选的异常训练样本被选取，从而造成分类精度的降低。为了解决这个问题，采用中位数绝对偏差法，根据图像的光谱信息探测和剔除遥感图像监督分类任务中不纯和错选的训练样本。选取由 Landsat-8 获取南昌市部分地区的光学遥感图像数据，使用支持向量机对含有异常训练样本和剔除异常训练样本的两种情况进行监督分类，并对分类结果进行比较。实验结果表明，剔除异常训练样本的分类精度明显优于含异常训练样本。

关键词 测量；遥感图像；光谱信息；监督分类；中位数绝对偏差；异常训练样本探测

中图分类号 O433.4 **文献标志码** A **doi**：10.3788/LOP57.231202

Abnormal Training Samples Detection Method Based on Median Absolute Deviation

Gong Xunqiang[1,2*], Zhang Fangze[1,2], Lu Tieding[1,2], Chen Zhigao[2]

[1] *Fundamental Science on Radioactive Geology and Exploration Technology Laboratory, East China University of Technology, Nanchang, Jiangxi* 330013, *China*;

[2] *Faculty of Geomatics, East China University of Technology, Nanchang, Jiangxi* 330013, *China*

Abstract The supervised classification technology of remote sensing images is widely used in the field of information extraction and change detection, in which the selection of training samples is very important, and the quality of training samples directly determines the accuracy of classification. However, due to the limitation of conditions and human error, some impure or wrong training samples may be selected, resulting in a decrease in classification accuracy. In order to solve this problem, the median absolute deviation method is used to detect and eliminate impure and wrong training samples in the supervised classification of remote sensing images based on the spectral information of the image. The optical remote sensing image data obtained from Landsat-8 in some areas of Nanchang city is selected, the support vector machine is used to supervise and classify the two situations that contain abnormal training samples and eliminate abnormal training samples, and compare the classification results. Experimental results show that the classification accuracy of removing abnormal training samples is significantly better than that of abnormal training samples.

Key words measurement; remote sensing image; spectral information; supervised classification; median absolute deviation; abnormal training samples detection

OCIS codes 120.0280; 090.6186; 110.4234; 100.4145

收稿日期：2020-03-17；**修回日期**：2020-04-08；**录用日期**：2020-04-15

基金项目：国家自然科学基金(41806114)、江西省自然科学基金(20181BAB216031)、抚州市社科规划(19sk08)、东华理工大学放射性地质与勘探技术国防重点学科实验室开放基金(RGET1905)

* **E-mail**：xqgong1988@163.com

1 引 言

遥感图像分类是遥感数字图像处理的一个重要内容，其广泛应用于土地利用或覆盖、树种识别、植被区分和变化检测等领域[1-2]。对于大多数遥感图像分类任务来说，在监督分类过程中施加额外的先验信息，所以其分类精度通常优于非监督分类。典型的监督学习框架可以提供一定数量的训练样本用于训练分类器，然后使用训练得到的分类器将目标图像分为不同的类别。因此，监督分类精度的高低在很大程度上取决于训练样本质量的好坏[3-5]。

然而，由于受到条件限制以及人为错误，用于训练分类器的训练样本经常会被污染，被污染的异常训练样本通常分为训练样本不纯和训练样本错选两种情况。当训练样本不纯时，其光谱值的标准差不同于同一地物类别的其他训练样本。当选取某种地物类别的训练样本时，若错误地将其他地物类别的训练样本归为该地物类别，其光谱值的均值往往不同于该地物类别的其他训练样本。针对训练样本中含有异常值的问题，相关学者已经提出了许多方法。一种常见的处理策略是设计不受异常训练样本影响的复杂模型，比如集成了几种分类器优点的集成学习方法，该方法对异常训练样本具有鲁棒性[3,6]。尽管采用集成学习方法能够获得较好的结果，但是大多数现有的集成学习方法仅在训练样本中含有少量的异常训练样本才有效果。另一种处理策略是先识别和剔除异常训练样本，然后使用提纯的训练样本来训练分类器，进而得到更精确的分类结果[7-9]，但是当探测和剔除异常训练样本时，均需要大量训练样本的支撑。

为了解决少量训练样本中可能存在异常训练样本的问题，本文采用异常值探测能力强和计算效率高的中位数绝对偏差(MAD)法[10-12]探测和剔除异常训练样本，使用常用的支持向量机(SVM)分类器[13]对遥感图像进行分类，通过与异常训练样本分类结果进行比较，验证 MAD 法对提高分类精度具有可行性。

2 实验方法

MAD 法是所提方法的基础，在介绍实验方法前有必要对该方法进行简单介绍。

2.1 MAD 法

给定 n 个观测值$\{x_1,x_2,\cdots,x_n\}$来计算样本中位数，表达式为

$$M=\underset{i=1,\cdots,n}{\text{median}}(x_i)。\tag{1}$$

当 n 是奇数时，则中位数取排序为中间的观测值；当 n 是偶数时，则中位数取排序为 $n/2$ 和 $n/2+1$ 的观测值的平均值。

在中位数的基础上发展而来的 MAD 法是由 Hampel 等[14-15]提出的，表达式为

$$M_{\text{MAD}}=b\times\underset{i=1,\cdots,n}{\text{median}}\left|x_i-\underset{j=1,\cdots,n}{\text{median}}(x_j)\right|,\tag{2}$$

式中：b 表示常数，通常 $b=1.4826$；j 表示第二层循环序号。

为了对观测数据中的异常值进行探测，需要计算每个观测值 x_i 的判定系数 D，表达式为

$$D=\frac{\left|x_i-\underset{j=1,\cdots,n}{\text{median}}(x_j)\right|}{M_{\text{MAD}}}。\tag{3}$$

当 D 值大于给定的阈值时，则认定 x_i 为异常数据。根据大量的科学实验和工程实践结果，选择阈值为 2.5 较为合理[11]。MAD 法能够探测高达 50%的异常数据，所以其具有较强的探测效果。由于 MAD 法具有异常值探测能力强、计算简单和计算效率高等优点，因此实验采用 MAD 法来探测训练样本中可能存在的异常训练样本。

2.2 基于 MAD 的异常训练样本探测方法

在遥感图像的监督分类过程中，如果一个地物类别的训练样本包含属于其他地物类别的像元，那么不纯训练样本的标准差明显不同于该地物类别中的其他训练样本。同理，当选择某一地物类别的训练样本时，若错误地将其他地物类别的训练样本归为该地物类别，其均值则不同于该地物类别中的其他训练样本。根据这一特点，采用 MAD 法来探测地物类别中的不纯(或错选)训练样本。

假设为特定的类别创建 t 个训练样本，并且需要分类的遥感图像中有 d 个波段。采用 MAD 法来探测异常训练样本的过程如下。

1) 对于第 o 个训练样本，分别计算该训练样本中每个波段的标准差 s 或均值 a。

2) S 为 s 所有波段的标准差总和，A 为 a 所有波段的均值总和，即 $S=\sum_{k=1}^{d}s_k$ 或 $A=\sum_{k=1}^{d}a_k$，即构成新的观测值。

3) 对于 t 个训练样本，可以获得 n 个观测值，即观测值为$\{x_1,x_2,\cdots,x_n\}$。

4) 利用(1) ～ (3) 式来探测训练样本中的异常值。

对于其他地物类别，可以重复以上异常值探测步骤。

3 实验设计

为了验证所提方法的有效性，有必要设计合理的实验对所提方法的有效性进行评估。为了模拟异常训练样本，可在一个或多个地物类别中人为地选择一些不纯或错选的训练样本，从而验证所提方法探测异常训练样本的效果。需要指出的是，目前评价训练样本可分离性的指标是计算地物类别之间的 Jeffries-Matusita 距离，其取值范围为[0，2]，并认为该数值大于 1.9 为合格[16]。然而，当数值大于 1.9 时，仍然可能存在不纯或错选的训练样本。为了验证这一问题，实验中所选训练样本的可分离性指标均大于 1.9。下面分别介绍实验数据和分类评价指标。

3.1 实验数据

实验采集的时间为 2017 年 9 月 14 日，实验数据是由 Landsat-8 获取南昌市部分地区的光学遥感图像。通过图像融合可以获取空间分辨率为 15 m 的多光谱图像，选取的图像尺寸为 1000 pixel×1000 pixel，即 15 km×15 km 作为实验区域，其中包含建筑物、水体、植被和裸地 4 种地物类型。为了合理地对实验结果进行比较，所有验证样本都是固定不变的，验证样本如图 1 所示。设计的训练样本如下。

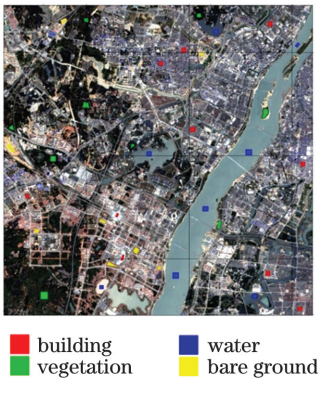

图 1 验证样本

Fig. 1 Verified sample

1）建筑物、水体、植被和裸地所选取的训练样本数量分别为 9、7、6 和 6，其中建筑物中含有三个不纯训练样本，水体中含有一个不纯训练样本，如图 2(a)所示。

2）建筑物、水体、植被和裸地所选取的训练样本数量分别为 8、6、5 和 5，其中建筑物中含有三个错选训练样本，水体中含有一个错选训练样本，如图 2(b)所示。

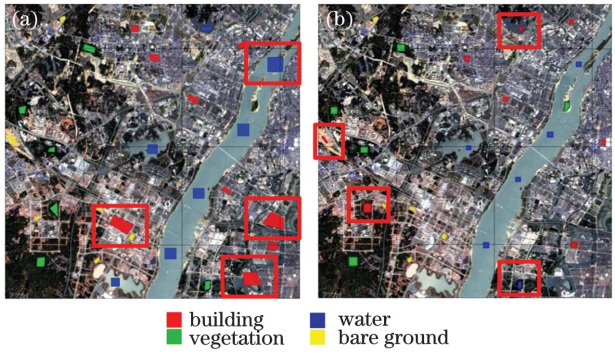

图 2 不同类型的训练样本。(a)不纯样本；(b)错选样本

Fig. 2 Different types of training samples. (a) Impure sample; (b) wrong choice sample

3.2 SVM 法

目前，遥感图像的监督分类方法主要有最大似然法、最小距离分类法、马氏距离分类法和 SVM 法等，其中 SVM 法由于具有较好的分类效果而经常被采用[17]。在机器学习中，SVM 法是一种有监督学习模型的算法，为此可以用于分析分类和回归分析中的数据。给定一组训练样本，每个训练样本被标记为属于两个类别中的一个或另一个，采用 SVM 法创建一个模型，该模型将新样本分配给两个类别中的一个，使其成为非概率二元线性分类器。SVM 模型是将样本表示为空间中的点，以最大间隔来分离各个类别的样本，然后将新样本映射到相同的空间中，并根据其所处间隔的哪一侧来预测类别。

3.3 分类评价指标

遥感图像监督分类后需要对其进行精度评定，目前一般采用混淆矩阵来评价分类精度的好坏[18]。在生成混淆矩阵的基础上，通过生产者精度、用户精度、总体精度和 Kappa 系数等对异常训练样本剔除前后的分类结果进行评价[19-21]。

1）生产者精度指某类别被正确分类的样本数目(对角线值)与该类别真实参考样本总数(混淆矩阵中某类列的总和)的比值。

2）用户精度指某类别被正确分类的样本数目(对角线值)与被分为该类别的样本总数(混淆矩阵中某类行的总和)的比值。

3）总体精度指被正确分类的样本总数与总样本数的比值，被正确分类的样本数沿着混淆矩阵的对角线分布，总样本数等于每个类别真实参考样本总数之和，表达式为

$$P_0=\frac{1}{L}\sum_{l=1}^{L}x_{ll}, \tag{4}$$

式中：x_{ll} 表示第 l 类别被正确分类的样本数目；L 表示总样本数。

4）Kappa 系数是一个用于评价遥感图像分类结果的一致性检验指标[12,22]，表达式为

$$K=\frac{L\sum_{l=1}^{L}x_{ll}-\sum_{l=1}^{L}(x_{l+}\times x_{+l})}{L^2-\sum_{l=1}^{L}(x_{l+}\times x_{+l})}, \tag{5}$$

式中：x_{l+} 和 x_{+l} 分别表示第 l 类别所在列和行的样本数之和。

4 实验结果与分析

4.1 不纯训练样本的结果与分析

在建筑物类别中 9 个训练样本的观测值分别为 8439.798、10317.971、8779.719、16017.037、8629.521、8832.289、21726.055、13354.052 和 7827.259，可以得到相应的判定系数为 0.263、0.997、0.035、4.822、0.136、0、8.654、3.035 和 0.675，大于 2.5 的训练样本为不纯训练样本。同样可以得到水体类别中 7 个训练样本的判定系数分别为 1.319、0.674、1.285、0、0.121、0.174 和 59.003，表明水体类别中存在一个不纯训练样本。所探测的 4 个不纯训练样本与实验设计部分的不纯训练样本保持一致，具体位置如图 2(a)所示。

不纯训练样本剔除前后的分类结果，如图 3(a)所示。从图 3(a)可以看到，由于建筑物样本中存在大量的不纯训练样本，从而造成其他地物类别的大量像素被错分成建筑物。同时，在部分建筑物周围的很多植被均被错分成水体，这是由于在水体样本中存在一个含有植被像素的不纯训练样本，剔除不纯训练样本后的分类结果则较为准确，如图 3(b)所示。相应的分类精度如表 1 所示。从表 1 可以看到，当训练样本中存在不纯训练样本时，植被、水体和裸地的生产者精度分别为 68.037%、71.030% 和 81.883%，明显低于剔除不纯训练样本后的生产者精度 95.658%、99.807% 和 98.573%；建筑物、水体和裸地的用户精度分别从 64.125%、84.295% 和 86.316% 提高到 94.571%、97.097% 和 99.711%；剔除所有不纯训练样本的总体精度和 Kappa 系数分别为 97.639% 和 0.968，明显高于不纯训练样本的总体精度 79.562% 和 Kappa 系数 0.726，分别提高 18.077 个百分点和 0.242。

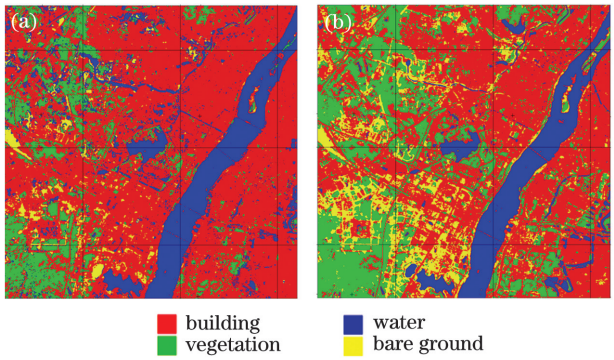

图 3 不纯训练样本剔除前后的分类结果。(a)剔除前；(b)剔除后

Fig. 3 Classification results before and after removal impure training samples. (a) Before removal; (b) after removal

表 1 不纯训练样本剔除前后的分类精度

Table 1 Classification accuracy before and after removal impure training samples

Class	Before removal		After removal	
	Producer/%	User/%	Producer/%	User/%
Building	99.494	64.125	96.866	94.571
Vegetation	68.037	100.000	95.658	99.807
Water	71.030	84.295	99.807	97.097
Bare ground	81.883	86.316	98.573	99.711

4.2 错选训练样本的结果与分析

建筑物类别中 8 个训练样本的观测值为 112990.212、88510.148、88608.985、90845.565、70938.949、81853.891、65344.236 和 87350.853，可以得到相应的判定系数分别为 3.760、0.087、0.102、0.437、2.549、0.912、3.389 和 0.087，判定系数大于 2.5 的样本是建筑物类别中的异常训练样本。同样可以求得 6 个水体训练样本的判定系数为 0.262、0.843、0.262、2.342、10.499 和 0.506，说明第 5 个训练样本为错选训练样本。

错选训练样本剔除前后的分类结果，如图 4 所示。从图 4 可以看到，由于在水体样本中存在一个

将植被错选为水体的训练样本，从而造成部分植被被错分成水体，而剔除单个错选训练样本后的分类结果则不存在这一情况。此外，图 4(a)中有大量的植被、水体和裸地被错分成建筑物，这是由于在建筑物样本中存在含有植被、水体和裸地像素的错选训练样本，而图 4(b)的分类结果则较好。相应的分类精度如表 2 所示。从表 2 可以看到，当训练样本中存在错选训练样本时，建筑物、植被、水体和裸地的生产者精度为 94.439%、79.676%、79.788% 和 67.475%，明显低于剔除所有错选训练样本的 97.573%、99.307%、100.000% 和 98.288%；建筑物和水体的用户精度分别从 62.329% 和 85.508% 提高到 98.469% 和 98.674%；含有错选训练样本的总体精度和 Kappa 系数分别为 81.294% 和 0.748，而剔除所有错选训练样本的总体精度和 Kappa 系数分别为 98.859% 和 0.985，剔除所有错选训练样本的总体精度和 Kappa 系数较剔除前分别提高 17.565 个百分点和 0.237。

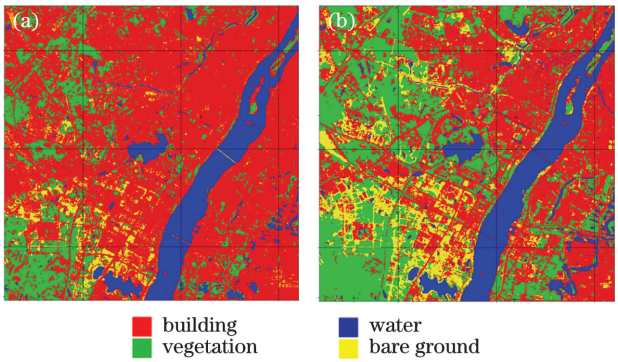

图 4 错选训练样本剔除前后的分类结果。(a)剔除前；(b)剔除后

Fig. 4 Classification results before and after ramoval of wrongly selected training samples. (a) Before removal; (b) after removal

表 2 错选训练样本剔除前后的分类精度

Table 2 Classification accuracy before and after removal of wrongly selected training samples

Class	Before removal		After removal	
	Producer/%	User/%	Producer/%	User/%
Building	94.439	62.329	97.573	98.469
Vegetation	79.676	99.884	99.307	99.583
Water	79.788	85.508	100.000	98.674
Bare ground	67.475	98.542	98.288	98.569

5 结 论

基于像素的遥感图像监督分类任务中，可能会遇到所选取的训练样本中存在部分异常训练样本的情况。为了解决这一问题，提出一种基于 MAD 的异常训练样本探测方法。采用该方法分别对不纯和错选训练样本进行探测和剔除，通过对剔除异常训练样本前后的分类结果进行比较。实验结果表明，所提方法能够准确探测遥感图像监督分类任务中训练样本不纯和错选的情况，从而有效消除异常训练样本对分类结果的影响；剔除异常训练样本后的总体精度和 Kappa 系数明显优于异常训练样本，分类精度提高显著，充分说明所提方法的有效性。

参 考 文 献

[1] Chen X, Ma J W, Dai Q. Remote sensing change detection based on Bayesian networks classifications [J]. Journal of Remote Sensing, 2005, 9(6): 667-672.
陈雪, 马建文, 戴芹. 基于贝叶斯网络分类的遥感影像变化检测[J]. 遥感学报, 2005, 9(6): 667-672.

[2] Sukawattanavijit C, Chen J, Zhang H S. GA-SVM algorithm for improving land-cover classification using SAR and optical remote sensing data[J]. IEEE Geoscience and Remote Sensing Letters, 2017, 14(3): 284-288.

[3] Frenay B, Verleysen M. Classification in the presence of label noise: a survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2014, 25(5): 845-869.

[4] Pelletier C, Valero S, Inglada J, et al. Effect of training class label noise on classification performances for land cover mapping with satellite image time series [J]. Remote Sensing, 2017, 9(2): 173.

[5] Yang B, Wang X. Boosting quality of pansharpened images using deep residual denoising network [J]. Laser & Optoelectronics Progress, 2019, 56(16): 161009.
杨斌, 王翔. 基于深度残差去噪网络的遥感融合图像质量提升[J]. 激光与光电子学进展, 2019, 56(16):

161009.

[6] Angelova A, Abu-Mostafam Y, Perona P. Pruning training sets for learning of object categories[C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR' 05), June 20-25, 2005, San Diego, CA, USA. New York: IEEE, 2005: 494-501.

[7] Brodley C E, Friedl M A. Identifying mislabeled training data[J]. Journal of Artificial Intelligence Research, 1999, 11: 131-167.

[8] Büschenfeld T, Ostermann J. Automatic refinement of training data for classification of satellite imagery [J]. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2012, I-7: 117-122.

[9] Chellasamy M, Ferré T P A, Greve M H. An ensemble-based training data refinement for automatic crop discrimination using WorldView-2 imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2015, 8(10): 4882-4894.

[10] Rousseeuw P J, Hubert M. Robust statistics for outlier detection[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011, 1(1): 73-79.

[11] Leys C, Ley C, Klein O, et al. detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median[J]. Journal of Experimental Social Psychology, 2013, 49(4): 764-766.

[12] Gong X Q, Shen L, Lu T D. Refining training samples using median absolute deviation for supervised classification of remote sensing images [J]. Journal of the Indian Society of Remote Sensing, 2019, 47(4): 647-659.

[13] Koda S, Zeggada A, Melgani F, et al. Spatial and structured SVM for multilabel image classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(10): 5948-5960.

[14] Hampel F R. The influence curve and its role in robust estimation [J]. Journal of the American Statistical Association, 1974, 69(346): 383-393.

[15] Huber P J. Robust statistics [M]//Lovric M. International Encyclopedia of Statistical Science. Berlin: Springer, 2011: 1248-1251.

[16] Pei H, Sun T J, Wang X Y. Object-oriented land use/cover classification based on texture features of Landsat 8 OLI image[J]. Transactions of the Chinese Society of Agricultural Engineering, 2018, 34(2): 248-255.
裴欢, 孙天娇, 王晓妍. 基于 Landsat 8 OLI 影像纹理特征的面向对象土地利用/覆盖分类[J]. 农业工程学报, 2018, 34(2): 248-255.

[17] Wang S T, Wu X, Zhu W H, et al. Fluorescence detection of polycyclic aromatic hydrocarbons by parallel factor combined with support vector machine [J]. Acta Optica Sinica, 2019, 39(5): 0530002.
王书涛, 吴兴, 朱文浩, 等. 平行因子结合支持向量机对多环芳烃的荧光检测[J]. 光学学报, 2019, 39(5): 0530002.

[18] Foody G M, Mathur A. The use of small training sets containing mixed pixels for accurate hard image classification: training on mixed spectral responses for classification by a SVM[J]. Remote Sensing of Environment, 2006, 103(2): 179-189.

[19] Wang M, Fan T F, Yun W G, et al. PFWG improved CNN multispectral remote sensing image classification[J]. Laser & Optoelectronics Progress, 2019, 56(3): 031003.
王民, 樊潭飞, 员卫国, 等. PFWG 改进的 CNN 多光谱遥感图像分类[J]. 激光与光电子学进展, 2019, 56(3): 031003.

[20] Liu C R, Frazier P, Kumar L. Comparative assessment of the measures of thematic classification accuracy[J]. Remote Sensing of Environment, 2007, 107(4): 606-616.

[21] Wu B, Lin S S, Zhou G J. Quantitatively evaluating indexes for object-based segmentation of high spatial resolution image [J]. Journal of Geo-Information Science, 2013, 15(4): 567-573.
吴波, 林珊珊, 周桂军. 面向对象的高分辨率遥感影像分割分类评价指标[J]. 地球信息科学学报, 2013, 15(4): 567-573.

[22] Yang Y K, Xiao P F, Feng X Z, et al. Comparison and assessment of large-scale land cover datasets in China and adjacent regions[J]. Journal of Remote Sensing, 2014, 18(2): 453-475.
杨永可, 肖鹏峰, 冯学智, 等. 大尺度土地覆盖数据集在中国及周边区域的精度评价[J]. 遥感学报, 2014, 18(2): 453-475.