A SEMI-SUPERVISED APPROACH TOWARDS LAND COVER MAPPING WITH SENTINEL-2 DESNSE TIME-SERIES IMAGERY

Ting Hu[1], Xin Huang[1], Jiayi Li[1], Jón Atli Benediktsson[2], Jiansi Yang[3], Jianya Gong[1]*

1 School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, P. R. China
2 Department of Electrical and Computer Engineering, University of Iceland, 107 Reykjavik, Iceland
3 School of Urban Design, Wuhan University, Wuhan 430079, P. R. China
Email: huang_whu@163.com

ABSTRACT

This paper presents a new semi-supervised method for land cover classification using Sentinel-2 time-series images, which can deal with the problem of unclear observations. First, the MCCR method, which is constituted by the matrix completion (MC) of unclear observations and feature-adaptive collaborative representation (CR) based classifier, is adopted to handle the data quality problem. Second, by fusing RF, AdaBoost, and MCCR, a tri-training process is proposed to iteratively select the semi-labeled samples, considering the difference of classification certainty in different classifiers and classes. Experiments on two sets of Sentinel-2 images are conducted to validate the effectiveness of the proposed semi-supervised method.

Index Terms—Sentinel-2, tri-training, time series images, land cover classification

1. INTRODUCTION

Accurate land cover mapping is of great importance to our understanding of coupled human-environment systems. Remote sensing data provide a valuable source for land cover mapping. Compared with single-date images, time series imagery is more advantageous in discriminating land cover types, especially for the classes that exhibit different characteristics during different time periods. However, it is difficult to obtain clear, i.e., complete and uncontaminated time series images from optical remote sensing platforms, due to the existence of clouds/cloud shadows and snow/ice cover.

To cope with the unclear observations, some studies generated best-available-pixel (BAP) composites from time series imagery. However, the criterion for the construction of BAP needs to be carefully designed in advance [1]. Zhu et al. used all the clear time-series observations to build a model for each pixel and then took the model coefficients and root mean square error as features to identify land cover types. However, it is not always guaranteed that there are sufficient clear observations for model initialization [2]. In contrast to the methods that totally ignore the unclear observations, approaches that adopt dense time series data investigate discriminative information from the unclear observations. Furthermore, it has been demonstrated that the random forest (RF) classifier outperforms other state-of-the-art classifiers, e.g., support vector machine (SVM), when dealing with observations with contamination [3]. The RF classifier has the ability to tolerate noisy observations, ascribed to the random selection of samples and features. The AdaBoost classifier can also handle the data quality problem by reweighting each training sample in the iterative learning process. However, these methods do not fully consider the different usability of clear and contaminated observations.

In this paper, a new semi-supervised land cover mapping approach is proposed. First, the benefit of recovering unclear observations in the classification is investigated based on the combination of matrix completion (MC) and collaborative representation (CR) classifier. The similarity among the training samples from each class enables the recovery of unclear elements through MC technology. Furthermore, by taking the different mechanisms of the three base classifiers into considerations, a tri-training paradigm is proposed to improve the classification accuracy. According to the different predictions of the three classifiers on unlabeled samples, informative and reliable samples are iteratively selected as newly added semi-labeled samples to retrain the classifiers, and then improve the final classification performance. The Sentinel-2 imagery are widely used as time-series data source due to the high revisit time (5 days at the equator) and multi-spectral bands (13 bands that cover the visible, near infrared, and shortwave infrared domains). In addition, research on the classification of the dense time-series Sentinel-2 images with unclear observations is rare. In this study, we adopt the Sentinel-2 images as unclear time series data to investigate the performance of the proposed semi-supervised approach.

2. PROPOSED METHOD

The proposed approach (see Fig. 1) is composed of two main parts, 1) the construction of an MCCR classifier and 2) the selection of semi-labeled samples by a tri-training learning process.

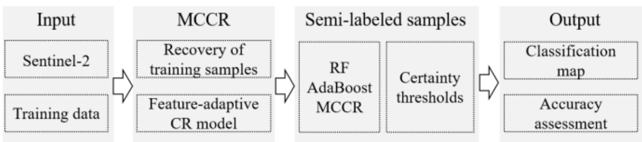


Fig. 1. Methodology flowchart.

2.1. The construction of MCCR classifier

The locations of unclear observations in Sentinel-2 data are first obtained by the cloud displacement index (CDI) algorithm, which has been integrated into the Fmask 4.0 software [4]. Observations that are recognized as clouds, cloud shadows or snow are all taken as contaminated. The feature vector of each sample is then formed by a time series stacking of multi-spectral bands, and hence the features of training samples with the same label should be in a lower dimensional space. Suppose $\boldsymbol{D}_i$ denotes the dictionary composed of all the training samples from class i, the matrix $\boldsymbol{D}_i$ should be the low-rank if all the matrix elements are reliable. Therefore, the unclear elements in $\boldsymbol{D}_i$ can be recovered by MC technology.

For a completely clear unlabeled sample $\boldsymbol{y}$, it can be linearly represented by the dictionary $\boldsymbol{D} = [\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_p]$, according to the general CR model [5]. However, it is difficult to ensure that all the elements in a feature vector are reliable. Let $\boldsymbol{y}$ is composed of two parts, i.e., clear sub-feature ($\boldsymbol{y}_c$), and unclear sub-feature ($\boldsymbol{y}_{uc}$), the sub-feature $\boldsymbol{y}_c$ can be linearly represented via:

$$\boldsymbol{y}_c = \boldsymbol{D}_c\boldsymbol{\alpha} + \boldsymbol{\varepsilon}_c \tag{1}$$

Through the flexible selection of clear sub-features, the crisp (label) and soft (class posterior probability) outputs of each unlabeled sample are determined based on the representation coefficient and model residual error [6]. In this way, an MCCR classifier has the potential to deal with the unclear features.

2.2 The tri-training selection of semi-labeled samples

Similar to the MCCR, the RF and AdaBoost classifier can also handle contaminated features. However, they have different mechanisms in the classification procedures, resulting in different crisp and soft predictions of each unlabeled sample. In the tri-training algorithm, three base classifiers are initially trained with the pre-collected training samples, and the next step is to select the qualified semi-labeled samples from the unlabeled ones, and the semi-labeled samples are then fed into each classifier to retrain the models. The selection of semi-labeled samples is iteratively implemented until the predefined maximum number is reached. Finally, the prediction of each sample is determined by the probability fusion of each classifier.

In order to improve the discrimination capability of each classifier, it is needed to select samples that are 1) correctly labeled and 2) different from the existing training samples. For each unlabeled sample, the difference between the two highest posterior probabilities is used to measure the classification certainty (*Cer*). Generally, samples with lower classification certainty are more liable likely to be mislabeled, but if the predicted label is correct, these samples can be more informative to the classifier [7]. Therefore, the above two criteria are actually conflicting, and it is necessary to set a tradeoff strategy [8]. The detailed selection procedure is described as follows:

1) Generation of a candidate set ($\mathcal{P}$): An unlabeled sample can be reliably labeled to some extent, when the crisp predictions of the three classifiers are identical. Therefore, the candidate set $\mathcal{P}$ is generated by these samples.

2) Tradeoff selection ($\mathcal{P}_{trim}$): The samples in set $\mathcal{P}$ are not absolutely correct owing to the small size of the training samples and unclear observations. Therefore, further selection from $\mathcal{P}$ is needed. Taking into account the difference in the certainty of different classes and classifiers, the threshold for a further selection is set as:

$$T_{ik} = \beta \times \mathrm{mean}(Cer_k(\mathcal{P}_i)) \tag{2}$$

where $\mathcal{P}_i$ denotes the samples in $\mathcal{P}$ that are labeled as class i, Cer_k means the certainty generated by the classifier k, and β is the tradeoff parameter that balances the reliability and diversity of samples. The unlabeled samples $\boldsymbol{y}$ whose certainties are higher than the corresponding thresholds are first merged class-by-class, and the samples of each classifier are then intersected to form the trimmed semi-labeled sample set:

$$\mathcal{P}_{trim} = \cap_k[\cup_i(Cer(\boldsymbol{y}) > T_{ik})] \tag{3}$$

The confidence that a sample in $\mathcal{P}_{trim}$ is correctly-labeled is higher with a larger β, while the diversity of samples in $\mathcal{P}_{trim}$ is larger with a smaller β. The samples in $\mathcal{P}_{trim}$ are respectively divided into three parts, i.e., $\mathcal{P}_{trim}^k$, where k =

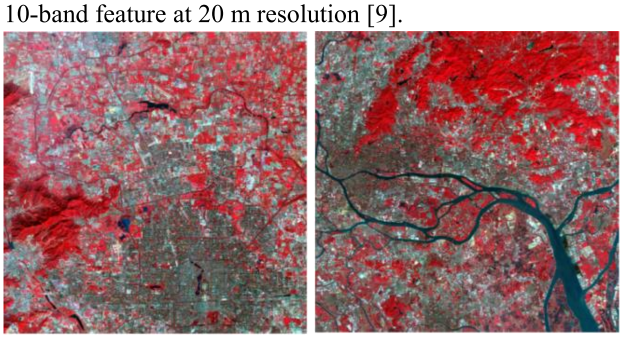
{*RF*, *AdaBoost*, *MCCR*}, and the rule is that, if the certainty of a sample in classifier k is lower than that in the other two classifiers, this sample will be partitioned into $\mathcal{P}_{trim}^{k}$.

3) Reduction of selected samples: To further select representative samples from a large number of semi-labeled samples, the k-means method is adopted with the cluster number is set as 20% of the current number of training samples in each class, and the resulting center patterns are then taken as semi-labeled samples.

3. EXPERIMENT AND ANALYSIS

3.1 Dataset and experimental setup

In this study, we selected the Sentinel-2 (S2) images located in the two cities of China (Beijing and Guangzhou) as the datasets. All the available Level-1C top-of-atmosphere (TOA) images acquired in 2017 with less than 60% contaminated observations were downloaded from the USGS website. They were corrected to Level-2A bottom-of-atmosphere (BOA) reflectance using Sen2Cor plugin provided by the European Space Agency. In the Beijing (BJ) dataset, 20 Sentinel-2 images were stacked to construct the features, and 11 Sentinel-2 images were stacked for Guangzhou (GZ) dataset. In each city, a subset of size 2000×2000 was clipped as the study area (Fig. 2). For each image. the bands 2-8, 8A, 11 and 12 were used to produce 10-band feature at 20 m resolution [9].



(a) Beijing (b) Guangzhou

Fig. 2. The two Sentinel-2 datasets covering 40 km × 40 km (R: 8a; G: 4; B: 3). (a) Beijing in 20171005 (b) Guangzhou in 20171027.

The reference samples for each dataset are randomly collected in polygon by manually interpreting S2 time series images together with Google Earth imagery of the same period respectively. Half of the reference polygons are selected for training and the others are taken for test. One sample was randomly chosen from each polygon to generate the training and test set, and the experiment for each dataset was carried out five times independently. The numbers of training/test pixels for each dataset are listed in Table I.

The parameters in all the base classifier are optimized by 10-fold cross validation based on the training samples. The maximum iteration number in the proposed tri-training approach is set as 3, and the tradeoff parameter β is 1.0 in the two datasets.

TABLE I THE NUMBERS OF TRAINING&TEST SAMPLES

		Built-up	Cropland	Forest	Water	Soil
BJ	Training	217	135	89	110	50
	Test	217	135	89	110	50
GZ	Training	172	112	150	143	125
	Test	172	112	150	142	125

3.2 Results and analysis

The classification results of the RF, AdaBoost, MCCR, and the tri-training method for the two datasets are presented in Table II, in terms of the average accuracies for the 5 experiments. In addition to the three base classifiers, the probability fusion of each classifier (each pixel is assigned to the label with the maximum sum of the class probability of the three classifiers) is also selected as comparisons (hereafter referred to as 'Fuse').

TABLE II THE CLASSIFICATION RESULS OF THE THREE BASE CLASSIFIERS AND TRI_TRAINING METHOD.

		RF	AdaBoost	MCCR	Fuse	Tri-training
OA (%)	BJ	89.58	89.26	90.07	90.42	92.35
	GZ	93.71	93.44	93.58	94.31	95.28
KC	OA	0.866	0.863	0.869	0.871	0.899
	KC	0.922	0.920	0.921	0.926	0.939

With respect to the overall accuracy (OA), all the three base classifiers achieve relatively good results in the BJ and GZ datasets, indicating their capability in dealing with the observation noise (clouds/cloud shadows, snow/ice, etc.) in S2 multi-temporal images. The three classifiers also perform similar to each other, both in terms of OA and the Kappa coefficient (KC). The 'Fuse' classifier slightly outperforms each individual classifier with an increase of about 0.005 in KC. It is important note that the difference between the 'Fuse' classifier and the tri-training approach only originates from the introduction of the iterative semi-labeled samples based on classification certainties of diverse classifiers. It can be seen that the OA and KC values of the tri-training method in both experiments are higher than the 'Fuse' method. For the BJ dataset, OA values are increased by nearly 2.0% and the KC values improves by 0.028. The improvement in GZ is slightly lower, with about 1.0% and 0.013 in terms of OA and KC, respectively.

The semi-labeled samples are selected according to the classification certainties of different classifiers and classes. With different values of the tradeoff parameter β, the balance between the confidence and information of selected sample varies, thus leading to different classification results. Taking the GZ dataset as an example, Fig. 3 displays the impacts of different β (ranging from 0.4 to 1.6 with a step of 0.2) on the results in regard to the OA.

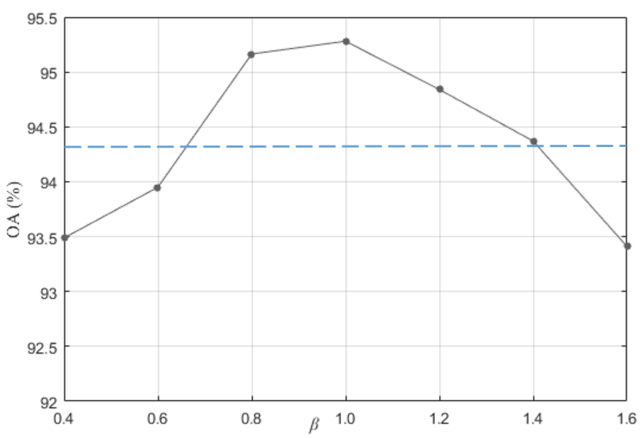


Fig. 3. The effect of the parameter β to overall accuracy of the GZ datasets.

Compared with the benchmark accuracy obtained by probability fusion (the dashed line in Fig. 2), the tri-training module has the ability to further improve the classification performance with a moderate β value (e.g., from 0.8 to 1.2). When the value of β is small, the high confidence of the selected semi-labeled samples cannot be guaranteed, and hence the classification accuracy decreases. For large β values, the similarity between the selected samples and the initial manual training samples are too high for each classifier to improve the classification capability. Therefore, in this study, the adaptive threshold for each class and classifier is suggested as the average certainty of the candidate unlabeled samples in each class and classifier.

4. CONCLUSIONS

In this paper, a new semi-supervised land cover mapping approach with Sentinel-2 time-series imagery is introduced. Considering existing contaminated observations, a method (MCCR), combined with the recovery of unclear observations and feature-adaptive collaborative representation based classifier, is developed to handle the unreliable observations. Moreover, a tri-training paradigm based on the three classifiers (RF, AdaBoost, and MCCR) is proposed to improve the classification accuracy, considering their complement and diversity. The semi-labeled samples are iteratively selected according to the adaptive thresholds for each class and classifier. The effectiveness of our method was validated on two datasets.

5. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grant 41771360, the National Program for Support of Top-notch Young Professionals, and the Hubei Provincial Natural Science Foundation of China under Grant 2017CFA029.

6. REFERENCES

[1] P. Beckschäfer, "Obtaining rubber plantation age information from very dense Landsat TM & ETM + time series data and pixel-based image compositing," *Remote Sens. Environ.*, vol. 196, pp. 89-100, 2017.

[2] Z. Zhu, and C. E. Woodcock, "Continuous change detection and classification of land cover using all available Landsat data," *Remote Sens. Environ.*, vol. 144, pp. 152-171, 2014.

[3] A. Schneider, "Monitoring land cover change in urban and peri-urban areas using dense time stacks of Landsat satellite data and a data mining approach," *Remote Sens. Environ.*, vol. 124, pp. 689-704, 2012.

[4] D. Frantz, E. Haß, A. Uhl, J. Stoffels, and J. Hill, "Improvement of the Fmask algorithm for Sentinel-2 images: Separating clouds from bright surfaces based on parallax effects," *Remote Sens. Environ.*, vol. 215, pp. 471-481, 2018.

[5] J. Li, H. Zhang, L. Zhang, X. Huang, and L. Zhang, "Joint Collaborative Representation With Multitask Learning for Hyperspectral Image Classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5923-5936, 2014.

[6] T. Hu, X. Huang, J. Li, and L. Zhang, "A novel co-training approach for urban land cover mapping with unclear Landsat time series imagery," *Remote Sens. Environ.*, vol. 217, pp. 144-157, 2018.

[7] L. Zhu, P. Xiao, X. Feng, X. Zhang, Y. Huang, C. Li, "A co-training, mutual learning approach towards mapping snow cover from multi-temporal high-spatial resolution satellite imagery," *ISPRS J. Photogram. Remote Sens.* vol. 122, pp. 179–191, 2016.

[8] Z. Zhou, M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Trans. Knowledge Data Engineering,* vol. 17, no. 11, pp. 1529-1541, 2005.

[9] M. Belgiu, C. Ovidiu, "Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis." *Remote Sens. Environ.*, vol. 204, pp. 509-523, 2018.