Computers & Geosciences 94 (2016) 131–139

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Case study

Remote sensing clustering analysis based on object-based interval modeling

Hui He [a,b], Tianheng Liang [a], Dan Hu [a,*], Xianchuan Yu [a,*]

[a] College of Information Science and Technology, Beijing Normal University, Beijing 100875, China
[b] College of Information and Technology, Beijing Normal University at Zhuhai, Zhuhai 519087, Guangdong Province, China

ARTICLE INFO

Article history:
Received 3 June 2015
Received in revised form
11 April 2016
Accepted 6 June 2016
Available online 22 June 2016

Keywords:
Object-based adaptive clustering
Interval-valued data
High-resolution remote sensing imagery
Pattern recognition

ABSTRACT

In object-based clustering, image data are segmented into objects (groups of pixels) and then clustered based on the objects' features. This method can be used to automatically classify high-resolution, remote sensing images, but requires accurate descriptions of object features. In this paper, we ascertain that interval-valued data model is appropriate for describing clustering prototype features. With this in mind, we developed an object-based interval modeling method for high-resolution, multiband, remote sensing data. We also designed an adaptive interval-valued fuzzy clustering method. We ran experiments utilizing images from the SPOT-5 satellite sensor, for the Pearl River Delta region and Beijing. The results indicate that the proposed algorithm considers both the anisotropy of the remote sensing data and the ambiguity of objects. Additionally, we present a new dissimilarity measure for interval vectors, which better separates the interval vectors generated by features of the segmentation units (objects). This approach effectively limits classification errors caused by spectral mixing between classes. Compared with the object-based unsupervised classification method proposed earlier, the proposed algorithm improves the classification accuracy without increasing computational complexity.

1. Introduction

Clustering analysis is a useful tool in remote sensing applications. However, there exists uncertainty in classifications of remotely sensed imagery. For example, there may be a series of uncertainties in the spectral signatures between classes and spectral variation within classes, because of the inherent uncertainty of remote sensing and the many sources of interference (Cheng et al., 2004). These uncertainties indicate that conventional, crisp clustering algorithms do not achieve a correct classification in most cases. Since the 1980s, fuzzy clustering has been extensively studied and successfully applied to remote sensing classification (Ibrahim et al., 2005; Schowengerdt, 2006; Ghosh et al., 2011). The most commonly utilized fuzzy clustering algorithm is the fuzzy c-means (FCM) algorithm. Many researchers have applied FCM to remotely sensed image analyses (Hasi et al., 2004; Xu et al., 2005; Qin and Xu, 2008; Yu et al., 2008; Joel et al., 2011; Bai and Zhao, 2013), and have achieved more satisfactory results than hard classification methods such as k-means and maximum likelihood classification. Standard FCM is based on image pixels, but high-resolution remote sensing images have smaller targets and more information, which leads to greater uncertainties than lower resolution images from the standpoint of land cover classification. Because more details in the high-resolution (more than 10 m) images often make it more difficult to describe a ground object. Therefore, as a pixel-based method, FCM cannot obtain the desired land cover classification results for high-resolution remote sensing images. Consequently, we expect a clustering algorithm that is more resistant to noise and can take advantage of more detailed information. Object-based classification methods for medium to high-resolution remote sensing images can provide a valid alternative to pixel-based methods (Geneletti and Gorte, 2003; Guo et al., 2007; Yang and Zhou, 2011). Yu et al. (2012) recently proposed an unsupervised classification method that adopts the object-based concept to automatically and effectively classify high-resolution remote sensing images. However, it is difficult to extract effective and stable features from the segmentation units, which directly affects the accuracy and stability of the classification results. For instance, the mean spectral signature is typically used to describe a segmentation unit, but this may not appropriately partition two different objects with the same mean value. Intervals are not only utilized to describe the uncertainties in the observed samples, but also utilized to represent a feature's uncertainty in the segmentation units. Partition clustering is a useful tool for interval-valued data analysis, and has been studied in depth in the literature. Adaptive and non-adaptive

* Corresponding authors.
E-mail address: chuan.yu@ieee.org (X. Yu).

http://dx.doi.org/10.1016/j.cageo.2016.06.006
0098-3004/

FCM clustering algorithms for intervals based on different dissimilarity metrics have been proposed (Gao et al., 1999; de Carvalho, 2007; Xie and Wang, 2012), but are not generally put into use. Moreover, de Carvalho (2007) verified that a clustering algorithm with adaptive distances based on a weighted version of the distances (such as the Hausdorff or city-block distances (de Carvalho, 2006)) outperformed the non-adaptive FCM algorithm. To obtain better results for high-resolution, remotely sensed image clustering analysis, we propose an object-based interval modeling method and an adaptive fuzzy clustering algorithm. We have also improved the distance metric for the intervals using the multiband character of remote sensing data. The structure of this paper is as follows. We introduce the interval-valued data model and dissimilarity metric for comparing the interval-valued data in Section 2, and discuss the proposed method in Section 3. Section 4 demonstrates the improved results by our new algorithms, and Section 5 presents our conclusions.

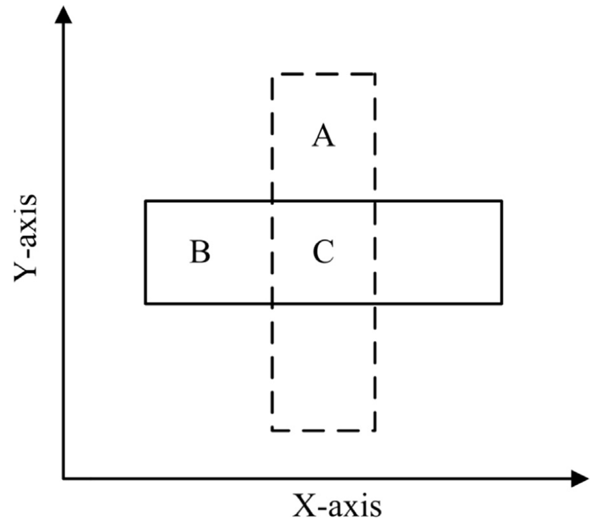

Fig. 1. Schematic of different interval objects based on the same mean.

2. Interval modeling and the dissimilarity metric

2.1. Definition of closed interval-valued data

The term "closed interval-valued data" means that the sample data are within a certain range. A more formal definition is (Moore, 1966):

Definition 1 $$\tilde{a} = [a^-, a^+] = \{x \in \Re : a^- \le x \le a^+\}, \quad (1)$$

Where a^+ *and* a^- are real numbers representing the lower and upper bounds of the interval-valued data denoted by $\tilde{a}$.Thus, $d = a^+ - a^-$ is the width and $med = (a^+ + a^-)/2$ is the mid value of the subinterval, which are both important features of interval-valued data. Many connections between interval algebra and fuzzy theory have been presented in the literature, especially describing the properties of datasets affected by uncertainties expressed as intervals (Irpino and Verde, 2008).

2.2. Definition of distance for interval-valued data

Distance is a dissimilarity and triangle inequality for different datasets and plays an important role in clustering analysis. Clustering is the process of dividing data elements into classes or clusters according to some distance measure, such that elements in the same class are as similar as possible, and elements in different classes are as dissimilar as possible. As in classic multivariate data clustering analysis, dissimilarities and distance measures for interval-valued data are crucial, because their appropriateness determines whether the clustering analysis is meaningful. There are many distance metrics for interval-valued data (Liem and Lucien, 2002; de Souza et al., 2004; de Souza and de Carvalho, 2004; Chavent et al., 2006; de Carvalho and de Souza, 2010; Hani and Chantal, 2011; Irpino et al., 2014). An overview of dissimilarity and distance metrics for interval data (e.g., city-block, Euclidean, Mahalanobis, Hausdorff, Wasserstein) can be found in Peng and Li (2006), Irpino and Verde (2008), and de Carvalho and Camilo (2010). However, the optimal dissimilarity metric depends on the application.

Let $\tilde{a} = [a^-, a^+]$, $\tilde{b} = [b^-, b^+]$ and $\tilde{c} = [c^-, c^+]$ be three intervals. Then, $d(\tilde{a}, \tilde{b})$ is a distance measure if it satisfies:

reflexivity: $d(\tilde{a}, \tilde{a}) = 0$,symmetry: $d(\tilde{a}, \tilde{b}) = d(\tilde{b}, \tilde{a})$, and the triangular inequality: $d(\tilde{a}, \tilde{b}) \le d(\tilde{a}, \tilde{c}) + d(\tilde{c}, \tilde{b})$.

3. Methodology

We present an adaptive fuzzy clustering analysis scheme based on object-based interval modeling, to improve the unsupervised classification accuracy for high-resolution remote sensing images. The process flow is shown in Fig. 2. The method includes the following steps.

(1) Segmentation and object interval modeling. The remotely sensed image is segmented to obtain a series of units with a spatial neighborhood and high homogeneity. The resulting data are utilized for interval-valued modeling.
(2) Fuzzy clustering analysis. The interval model of the segmentation units is clustered with adaptive interval-valued fuzzy c-means clustering (AIV-FCM).
(3) Post-processing. The classification results are post-processed using spot filtering and class combination methods.

3.1. Object-based interval modeling

3.1.1. Image segmentation

Although image segmentation is a key process in object-based classification, it is not the focal point of our discussion in this article. In this study, we applied the watershed-based algorithm for segmentation given its better computational efficiency, which is the patent for invention proposed by our team and experiments show an excellent effect in segmenting of high-resolution remote sensing imagery. The main steps are as follows (Yu and Kang, 2009).

Step 1: Compute the gradient of the multispectral image to obtain the gradient image.
Step 2: Segment the gradient image with the watershed algorithm.
Step 3: Merge regions according to the similarity between an undetermined region and its adjacent regions (see the definition in Eq. (2)). If the largest similarity value is greater than some threshold (e.g., 0.8), the two regions are merged to avoid over-segmentation. This process continues until no undetermined region remains. Note that an undetermined region is defined as a segmented region with an area smaller than the preset threshold (in pixels), which is dependent on the image resolution. In this study, the threshold $t = 200$.

For the similarity measure in Step 3, we first transform the multispectral image from SPOT5 to RGB space using band selection (Eiumnoh et al., 2012) , where band 1, 2 and 3 is related to R, G and B, respectively. Then the color space is transformed from RGB to the CIE (1976) color space, commonly known by its abbreviation