Computers and Geosciences 112 (2018) 38–46

Contents lists available at ScienceDirect

Computers and Geosciences

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

Case study

An improved optimum-path forest clustering algorithm for remote sensing image segmentation

Siya Chen [a], Tieli Sun [a,b,*], Fengqin Yang [a,**], Hongguang Sun [a,***], Yu Guan [a]

[a] School of Information Science and Technology & School of Geographical Science, Northeast Normal University, Changchun 130024, China
[b] Key Laboratory of Intelligent Information Processing of Jilin Universities, Northeast Normal University, Changchun 130024, China

ARTICLE INFO

Keywords:
Optimum-path forest
Clustering
Remote sensing image segmentation

ABSTRACT

Remote sensing image segmentation is a key technology for processing remote sensing images. The image segmentation results can be used for feature extraction, target identification and object description. Thus, image segmentation directly affects the subsequent processing results. This paper proposes a novel Optimum-Path Forest (OPF) clustering algorithm that can be used for remote sensing segmentation. The method utilizes the principle that the cluster centres are characterized based on their densities and the distances between the centres and samples with higher densities. A new OPF clustering algorithm probability density function is defined based on this principle and applied to remote sensing image segmentation. Experiments are conducted using five remote sensing land cover images. The experimental results illustrate that the proposed method can outperform the original OPF approach.

1. Introduction

Remote sensing images play a significant role in earth science due to their superior ability to express the characteristics of ground objects. These images have been widely applied in various fields, such as environment monitoring (Kang et al., 2015), urban planning (Banerjee et al., 2013) and national defence (Lampropoulos et al., 2008). The number of obtainable remote sensing images has dramatically increased due to the rapid development of remote sensing observation techniques. Large amounts of remote sensing data and growing demands have accelerated the development of remote sensing image processing. An emphasis has been laid on automatically processing and analysing these remote sensing images and extracting useful information from them. Image segmentation is a method used to automatically extract features and distinguish distinct objects in remote sensing images (Zhang et al., 2015).

Numerous remote sensing image segmentation approaches have been utilized. These approaches mainly encompass three strategies: supervised methods, unsupervised methods and semi-supervised methods. The supervised methods require many known pixels with class labels, which are used as a training set to label the unknown pixels. For instance, support vector machine (SVM) (Cortes and Vapnik, 1995; Gomez-chova et al., 2008) is a typical supervised method. Supervised methods are extremely time consuming when applied to hyperspectral or very high resolution images. Furthermore, the class labels are unavailable in many cases.

Most unsupervised methods employ clustering algorithms. Unlike supervised methods, unsupervised methods exploit observation features to segment images and do not require training sets. Unsupervised methods are generally used when the class labels are unknown. Two main clustering approaches are commonly used in the literature: partitioning methods and hierarchical methods. Typical partitioning methods include the minimum spanning forest (Bernard et al., 2012; Tarabalka et al., 2010a), fuzzy c-means (FCM) (Zhong et al., 2014; Alhichri et al., 2014; Li et al., 2013) and associated extension algorithms, k-means (Isa et al., 2009) and associated variant algorithms, iterative self-organizing data analysis (ISODATA) (Ball and Hall, 1965) and spectral clustering techniques (Zhang et al., 2008; Jia et al., 2011). In addition to the basic spectral features of images, hierarchical methods also (Tarabalka et al., 2010b) consider spatial information. Lee (2004) and Lee and Crawford (2004) employed hierarchical clustering and the theory that pixels belonging to the same cluster are spatially contiguous to classify hyperspectral data. Bruzzone and Carlin (2006) and Huo et al. (2015) combined hierarchical segmentation with SVM to classify very high spatial resolution images.

Unsupervised methods include various drawbacks. For instance, the

* Corresponding author. School of Information Science and Technology & School of Geographical Science, Northeast Normal University, Changchun 130024. China.
** Corresponding author.
*** Corresponding author.
E-mail addresses: suntl@nenu.edu.cn (T. Sun), yangfq147@nenu.edu.cn (F. Yang), sunhg889@nenu.edu.cn (H. Sun).

https://doi.org/10.1016/j.cageo.2017.12.003
Received 23 February 2016; Received in revised form 8 December 2017; Accepted 8 December 2017
Available online 9 December 2017

number of resultant clusters remains difficult to determine. Few robust criteria exist to define the appropriate number of clusters. Users generally determine the number of clusters based on their previous knowledge. Furthermore, unsupervised method outputs lack semantic information. Thus, these methods can only group data objects into classes according to their similarity and cannot provide semantic information for these classes or the relationships between classes, often requiring the user to further explain the clustering results.

Semi-supervised methods (Li et al., 2010) combine supervised learning with unsupervised learning. These methods utilize a small number of labelled samples and many unmarked samples for training and classification, providing advantages over both unsupervised and supervised methods. Several strategies have been proposed in the literature (Yang et al., 2013; Tuia and Camps-Valls, 2011.). Some semi-supervised methods use the supervised model to initialize the segmentation algorithms (Tarabalka et al., 2010c). However, these methods require large numbers of labelled pixels and their results rely on the supervised model utilized. Other semi-supervised methods must be accurately tuned and require a large amount of unlabelled data (Munoz-Mari et al., 2012).

The Optimum-Path Forest (OPF) (Rocha et al., 2009) pattern recognition algorithm has recently attracted extensive attention of researchers and has been widely applied to image segmentation. The OPF classifier includes supervised (Papa et al., 2009) and unsupervised versions (Papa and Falcao, 2008). Pisani et al. (2014) introduced the OPF algorithm for land cover classification. Filho et al. (2013) applied OPF operators to segment sandstone thin section images. Cappabianco et al. (2012) handled MR-image brain tissue segmentation via OPF clustering. Nakamura et al. (2014) combined OPF with evolutionary algorithms to extract spectral features and improved the speed and accuracy of segmentation. Iwashita et al. (2014) adopted a path- and label-cost propagation approach to accelerate the OPF classifier training. Costa et al. (2015) introduced a nature-inspired approach to estimate the probability density function (PDF) and increase the speed of the clustering algorithm based on OPF.

This paper proposes a new OPF clustering algorithm and applies it to land cover classification. The new, improved algorithm utilizes more detailed attributes of cluster centres than does the original OPF clustering algorithm. A new probability density function is defined in our proposed algorithm. We consider that the characteristics of cluster centres lie not only in the density but also in the distance between cluster centres and in higher density samples. The experiments show that the proposed algorithm performs better than the original OPF clustering. The remainder of this paper is organized as follows. Section 2 reviews the OPF theory. Section 3 presents the new algorithm. Section 4 discusses the experimental results. Finally, the conclusions are stated in Section 5.

2. The optimum-path forest clustering algorithm

OPF is a clustering algorithm based on a graph structure that represents the feature space. It divides the graph into several optimum-path trees, which each denote a cluster. The advantage of the OPF clustering algorithm is that it does not require assumptions regarding the shape/separability of the feature space. The nodes represent the samples in the dataset and the arcs connecting the nodes denote specific adjacency relations between the samples. Both the nodes and the arcs in the graph can be weighted. The arc weights are computed according to a distance function and the node weights are defined as the probability density values obtained from the arc weights. We defined a path as a series of adjacent nodes and nodes with maximum weights deemed key nodes. A connectivity function is used to evaluate the connectedness between terminal nodes within a path. An optimum-path tree is a tree whose root node is a key node, while other nodes are connected to the root node by paths based on the maximum connectivity function. Each optimum-path tree represents a cluster. The entire optimum-path forest is the output of the clustering algorithm. Our goal is to compute the optimum-path forest in the graph.

2.1. Weighted graphs and PDF estimation

Let N denote the dataset and each sample $s \in N$ define a vector $\vec{v}(s)$. The function $d(s,t)$ denotes the distance between samples s and t. In many cases, we exploit the Euclidean distance function to compute the distance $d(s,t) = \|\vec{v}(t) - \vec{v}(s)\|$.

If samples t and s satisfy a certain adjacency relation in the graph structure $G = (N, A)$, then t is adjacent to s and defined as $t \in A(s)$ or $(s,t) \in A$. Two methods are generally used to define the adjacency relation between samples. The first method states that $t \in A(s)$ if $d(s,t) \le d_f$, while the other states that $t \in A(s)$ if t is one of k-nearest neighbours of s, where $d_f > 0$ and $k > 0$ are real and integer parameters, respectively.

The node weights are defined as the probability density function (PDF) values $\rho(s)$.

$$\rho(s) = \frac{1}{\sqrt{2\pi\sigma^2}|A(s)|} \sum_{\forall t \in A(s)} \exp\left(\frac{-d^2(s,t)}{2\sigma^2}\right) \tag{1}$$

$$\sigma = \max_{\forall (s,t) \in A} \left\{\frac{d(s,t)}{3}\right\} \tag{2}$$

where $|A(s)|$ is the number of nodes adjacent to node s. This choice of parameterσguarantees that all nodes are considered in the density computation, because a Gaussian function encompasses the majority of samples within $d(s,t) \in [0, 3\sigma]$. However, the definition of σ is not fixed and may differ when using other adjacency relations.

The value of k must be determined when the k-nearest neighbour method is used to construct the graph G. One method for determining k is to define a variable $1 \le k_{\max} \le |N|$ and scale $k \in [1, k_{\max}]$. The optimal value of k is then computed in this interval according to the method suggested by Shi and Malik (2000). This method utilizes the minimum value of $C(k)$ to determine the optimal clustering results for $k \in [1, k_{\max}]$ according to the graph-cut measure for multiple clusters. The value of k corresponding to the best clustering result is based on the best value in this range. $C(k)$ is defined as follows.

$$C(k) = \sum_{i=1}^{c} \frac{W_i^{'}}{W_i + W_i^{'}} \tag{3}$$

$$W_i = \sum_{\forall (s,t) \in A | \lambda(s) = \lambda(t) = i} \frac{1}{d(s,t)} \tag{4}$$

$$W_i^{'} = \sum_{\forall (s,t) \in A | \lambda(s) = i, \lambda(t) \neq i} \frac{1}{d(s,t)} \tag{5}$$

where $\lambda(s)$ is the label of samples, $W_i^{'}$ computes the distances between samples in cluster i and samples in other clusters, and W_i computes the distances between samples in the same cluster.

2.2. Data clustering using OPF

A path π_t in the graph is defined as a series of adjacent nodes. The path begins with root $R(t)$ and ends with node t. $R(t)$ is a key node based on the maximum probability density value. A path $\pi_t = \langle t \rangle$ is a trivial path and $\pi_t = \pi_s \cdot \langle s,t \rangle$ is the concatenation of a path π_t and an arc (s,t). A cost function $f(\pi_t)$ is assigned to each path π_t and represents the connection strength between the node $t$and the root node $R(t)$. A path π_t is deemed the optimum path when $f(\pi_t) \ge f(\tau_t)$, where τ_t represents any other path containing node$t$.

The main objective of the clustering is to determine the path with the maximum path-cost value among all possible paths π_t connecting node t and roots based on the maximum of the PDF. Each root node and the samples that are more strongly connected to the current root node than to any other root node are defined as a cluster. The path-cost value of a path