ISPRS Journal of Photogrammetry and Remote Sensing 120 (2016) 99–107

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Semisupervised classification for hyperspectral image based on multi-decision labeling and deep feature learning

Xiaorui Ma, Hongyu Wang *, Jie Wang

School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, PR China

ARTICLE INFO

Article history:
Received 18 November 2015
Received in revised form 11 July 2016
Accepted 1 September 2016
Available online 17 September 2016

Keywords:
Hyperspectral image
Semisupervised classification
Deep learning

ABSTRACT

Semisupervised learning is widely used in hyperspectral image classification to deal with the limited training samples, however, some more information of hyperspectral image should be further explored. In this paper, a novel semisupervised classification based on multi-decision labeling and deep feature learning is presented to exploit and utilize as much information as possible to realize the classification task. First, the proposed method takes two decisions to pre-label each unlabeled sample: local decision based on weighted neighborhood information is made by the surrounding samples, and global decision based on deep learning is performed by the most similar training samples. Then, some unlabeled ones with high confidence are selected to extent the training set. Finally, self decision, which depends on the self features exploited by deep learning, is employed on the updated training set to extract spectral-spatial features and produce classification map. Experimental results with real data indicate that it is an effective and promising semisupervised classification method for hyperspectral image.

1. Introduction

Hyperspectral image can provide both spectral and spatial information of the object in the observed scene (Van der Meer et al., 2012; Bioucas-Dias et al., 2013; Willett et al., 2014). With this affluent information, hyperspectral image is widely used for lots of applications, e.g., land-cover investigation, mineral census, and so on. Hyperspectral image classification, which is the basis of variety remote sensing applications, has recently attracted interest and achieved significantly progress (Camps-Valls et al., 2014).

In the last decades, many methods have been developed for hyperspectral image classification. Unsupervised classifier focuses on finding patterns using only unlabeled samples, which is easily to be applied in remote sensing area. Partitioning and hierarchical clustering algorithms, such as fuzzy clustering (Zhong et al., 2014), and fuzzy C-Means (FCM) algorithms (Niazmardi et al., 2013), heavily depend on similarity measurement, which makes these methods sensitivity to noise and spectral variation. Moreover, the mixture resolving clustering algorithms require a description of data using statistical distribution (Zhong et al., 2006). However, samples of the target class are usually too few to estimate the statistical properties of the class, which makes unsupervised classification method hardly achieve acceptable performance.

Supervised classifiers, such as multinomial logistic regression (MLR) (Mahdi et al., 2014), support vector machine (SVM) (Gao et al., 2015), artificial neural network (ANN), and sparse representation based classification (SRC) (Chen et al., 2013; Yuan et al., 2014; Tang et al., 2014), which require labeled samples for training are widely used in hyperspectral image classifications. The quantity and quality of the labeled samples are critical for supervised methods (Persello and Bruzzone, 2014; Naeini et al., 2014; Shao et al., 2014) However, for remote sensing image, the ground campaign of labeling samples is expensive (Dópido et al., 2013) which always needs the assistance of experts and even some devices. These limitations cause instability in model parameter estimation, even incur the occurrence of Hughes phenomenon (Hughes, 1968), thus, foster the development of semisupervised methods which are able to utilize both limited training samples and vast of unlabeled samples to realize classification.

Semisupervised classifiers are able to classify with limited labeled samples together with large amount of unlabeled ones, which could achieve higher accuracy and are of great interest in practice (Khodadadzadeh et al., 2014; Pu et al., 2014; Li et al., 2014). Generally, five models are widely used in semisupervised learning (Zhu, 2005; Persello and Bruzzone, 2014), include: (1) generative models (Jin et al., 2013; Chapel et al., 2014) which

* Corresponding author.
E-mail address: whyu@dlut.edu.cn (H. Wang).

http://dx.doi.org/10.1016/j.isprsjprs.2016.09.001

estimate the conditional density to predict the labels of unlabeled samples. (2) Self-training (Dópido et al., 2013; Wang et al., 2014; Tan et al., 2015a) schemes which use the previous classification map to train the classifier iteratively. (3) Co-training methods (Di and Crawford, 2012; Zhang et al., 2014) that train several classifiers with independent subsets of the training set, then, use the unlabeled samples with high reliability to update the training set and train another classifier. (4) Transductive support vector machines (TSVM) (Maulik and Chakraborty, 2013; Yang et al., 2014) which maximize the margin for both the labeled and unlabeled samples. (5) Graph-based methods (Camps-Valls et al., 2007; Tan et al., 2015b; Im and Taylor, 2015) that spread the label information of each sample to its neighbors until a global stable state is achieved on all samples. Although the aforementioned models can achieve better performance with limited training set, they suffer from some limitations. Specifically, Generative model is based on the assumption that unlabeled data should follow certain distribution; Self-training reinforces poor predictions; Co-training requires that the features can be divided into independent subsets; TSVM is hard to deal with large unlabeled data; and graph-based algorithm is sensitive to graph structure and heavy calculation. This paper presents a self-training related method to suppress the error reinforcement, and a multi-decision labeling method to accurately label test samples in the semisupervised learning process.

Ideally, which class the sample belong to is only depends on its own attributes. However, for hyperspectral image classification, sample attributes are disturbed by several factors. Firstly, the reflectance spectra is determined not only by the object material, but also its surrounding material. Secondly, measurement noise of the remote sensing equipment causes spectral deviation. Thirdly, viewing angle and environmental factors, such as aerosols and moisture, also cause variation of the reflectance spectra. Therefore, which class the sample belong to can not be precisely determined by its spectral features, some other information is necessary. The most common information besides spectral information is spatial information, such as texture, shape, and profile features (Zhang et al., 2012; Zhang et al., 2015; Ghamisi et al., 2015) are widely used in hyperspectral image classification and achieve good performance. In this paper, three decisions, i.e., local decision, global decision, and self decision, based on different information are utilized to decide a sample label. Local decision is based on the surrounding samples, global decision is about the most similar ones, and self decision is depend on its own attributions. First, in order to deal with the limited training set, we prelabel every unlabeled sample with the first two decisions. Then, some unlabeled samples with high confidence are selected to extend the original training set. Finally, a deep network for self decision is trained based on the updated training set to extract features and produce the classification map.

The main contributions of this paper lie in: (1) Combination of three decisions for semisupervised classification, which can exploit, utilize, and optimize the information of hyperspectral image simultaneously, and achieve better performance. (2) Makes local decision by weighted spatial neighborhood information, which can utilize smooth prior in semisupervised classification. (3) Defines an effective similarity measurement between a sample and a class based on deep learning, which is more effective than traditional measurement. (4) Realizes self decision by contextual deep learning, which can better exploit both spatial and spectral features for hyperspectral image.

The rest of this paper is organized as follows. Section 2 gives a brief introduction of deep auto-encoder. In Section 3, the proposed semi-supervised classification is introduced in detail. In Section 4, some experiments are conducted and their results together with relevant discussions are reported. The conclusions are finally summarized in Section 5.

2. Background

Deep network, which is a variation of neural network, has been successfully used in computer vision, even remote sensing area (Zhang et al., 2016), such as hyperspectral or multispectral feature learning and hyperspectral image classification (Zhao and Du, 2016; Tuia et al., 2015; Huang et al., 2014). In traditional neural network, all weights are initialized randomly, and gradient descent is used for tunning all the weights to get better performance, but this works well only if the initialized weights are close to a good solution. Deep network is a neural network with a strategy of weight initialization, which trains each hidden layer by minimizing the discrepancy between original data and its reconstruction. Auto-encoder is an encoding model which the input can be reconstructed from the code with minimum decoding error, and it is a common reconstruction model of deep network. Deep auto-encoder utilizes auto-encoder to initial all hidden layers, every adjacent two layers constitute an auto-encoder, and the output of each hidden layer is the input of the subsequent layer. After initialization, all the parameters are fine-tuned by stochastic gradient descent strategy which can be efficiently implemented using the back propagation algorithm.

More specifically, for a given sample $\boldsymbol{x}$, train the first auto-encoder, which encodes the input sample by $\boldsymbol{r}^{(1)} = \mathrm{s}(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})$, and decodes it by $\tilde{\boldsymbol{x}} = \mathrm{s}(\tilde{\boldsymbol{W}}^{(1)}\boldsymbol{r}^{(1)} + \tilde{\boldsymbol{b}}^{(1)})$. $\boldsymbol{W}^{(1)}, \tilde{\boldsymbol{W}}^{(1)}, \boldsymbol{b}^{(1)}, \tilde{\boldsymbol{b}}^{(1)}$ are initialed randomly, and adjusted by minimizing the reconstruct error. Then, use the code $\boldsymbol{r}^{(1)}$ as the first hidden layer and the input of next auto-encoder, repeat the above procedure for subsequent hidden layers $\boldsymbol{r}^{(h)}$ to get all parameters $(\boldsymbol{W}^{(h)}, \boldsymbol{b}^{(h)})$, where $h = 1, \ldots, H$, and H is the number of hidden layers. This is pre-training, which initializes the weight of each layers in a relative correct way. After pre-training all the hidden layers, an output layer is added on the top to supervise the training. Similar to traditional neural network, a global optimization strategy by minimizing the following energy function is applied to fine-tuning the whole network:

$$J(\boldsymbol{W}, \boldsymbol{b}) = \frac{1}{2L}\sum_{m=1}^{L}\left\|\boldsymbol{y}^{(m)} - \boldsymbol{z}^{(m)}\right\|_2^2 + \frac{\lambda}{2}\sum_{h=1}^{H}\left(\left\|\boldsymbol{W}^{(h)}\right\|_F^2 + \left\|\boldsymbol{b}^{(h)}\right\|_2^2\right), \quad (1)$$

where $\boldsymbol{y}^{(m)}$ is the desired output and $\boldsymbol{z}^{(m)}$ is the real output, the second term is a weight decay term that decreases the magnitude of the weights to prevent over-fitting, and the weight decay parameter λ controls the relative importance of the two terms. The objective of fine-tuning is further minimizing the difference between the desired output $\boldsymbol{y}$ and the real output $\boldsymbol{z}$ by minimizing the energy function $J(\boldsymbol{W}, \boldsymbol{b})$. All the parameters determined by pre-training step are updated by stochastic gradient descent.

3. Methodology

This section presents the proposed semisupervised method for hyperspectral image classification. First of all, we briefly define the notations in this paper. Suppose the set of labeled pairs is denoted as $\{(\boldsymbol{x}^{(m)}, y^{(m)})\}_{m=1}^{L}$, and the set of unlabeled samples is denoted as $\{\boldsymbol{x}^{(m)}\}_{m=L+1}^{L+U}$, where $\boldsymbol{x}^{(m)} \in \mathbb{R}^{N\times 1}, y^{(m)} \in \{1, 2, \ldots, K\}, N$ is the number of attributions or bands, K is the number of classes, L and U are the number of labeled and unlabeled samples respectively, and the total number of samples is $M = L + U$. The goal of classification is learning a hypothesis which assign labels to unlabeled samples.

Theoretically, the sample label only depends on its own attributes. However, for hyperspectral image classification, the spectral information is not enough to decide the label. Firstly, due to the limitation of spatial resolution, the information captured by