Pattern Recognition 51 (2016) 43–58

Contents lists available at ScienceDirect

Pattern Recognition

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

Improved hyperspectral image classification by active learning using pre-designed mixed pixels

Alim Samat [a], Jun Li [b], Sicong Liu [c], Peijun Du [d,*], Zelang Miao [e], Jieqiong Luo [d]

[a] State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi, China
[b] Guangdong Key Laboratory for Urbanization and Geo-Simulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou, China
[c] Department of Information Engineering and Computer Science, University of Trento, Italy
[d] Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Nanjing University, Nanjing, China
[e] Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 10 January 2015
Received in revised form
11 August 2015
Accepted 22 August 2015
Available online 25 September 2015

Keywords:
Sample design
Low-cost
Active learning
Pixel purity index
Support vector machine
Hyperspectral image
Classification

ABSTRACT

Due to the limitation of labeled training samples, computational complexity, and other difficulties, active learning (AL) algorithms aiming at finding the most informative training samples have been an active topic of research in remote sensing image classification in the last few years. Usually, AL follows an iterative scheme, and the search of new samples relies on the whole image, resulting in that an approach may turn out to be prohibitive when the data sets are huge, e.g., hyperspectral data. Large amounts of unlabeled samples are easy to collect indeed, with respect to the cost of labeled sample collection. However, algorithm complexity, data storage capacity and processing times are also limited. Therefore, a sample set smaller in size, and consisting of the most valuable information, is preferable. In this work, we propose a design protocol to generate a more significant candidate sample set for active learning, aiming at reducing the unlabeled sample search complexity, and eventually improving the classification performance. The basic idea is providing the initial labeled and unlabeled samples that are composed of mixed or pure samples for AL heuristics, to find out which one is better for AL from the low-cost sample design point of view. For comparison and validation purposes, six state-of-the-art AL methods (including breaking ties, margin sampling, margin sampling by closest support vectors, normalized entropy query-by-committee, multi-class level uncertainty and multi view adaptive maximum disagreement based active learning) were tested on real hyperspectral images with different resolution both with and without the proposed sample design protocol. Experimental results confirmed the advantages of the proposed technique.

1. Introduction

To fully exploit the huge amount of spaceborne Earth Observation (EO) data, remote sensing science and application communities have been increasingly developing reliable, consistent, and robust approaches to capture spectral and spatial features and meet a range of information needs. Supervised image classification methods that require labeled reference data to deduce a hypothesis (or model) and generate a thematic map (e.g., land use/land cover, land change maps) are among the most commonly undertaken approaches. Nowadays, methods like kernel classifiers [1], ensemble learning (EL) [2], extreme learning machine (ELM) [3] and multinomial logistic regression (MLR) [4] are considered state-of-the-art algorithms for deriving land use/cover maps from satellite imagery. However, like for any other supervised algorithms, their accuracy varies as a function of the training set properties [5,6,48]. Moreover, a training set that could be used for one classifier may lead to undesired lower accuracies with another classifier [7]. Although EL has the advantage of generalizing capability of classifier by fusing the outputs of multiple trained classifiers, sufficient training samples are still needed to reach the diversity condition, which is the cornerstone for the construction of an effective EL system [2,8].

Therefore, one of the major concerns in supervised classification is the training samples should be large enough to provide a representative and unbiased description of the class properties (both statistical and geometrical ones). To this end, a variety of issues have been addressed in many studies, including the size, location and the composition of training samples, with a great deal of attention focused on the issue of the sample size [9]. In some earlier works, under the assumption that the training sample size follows a normal distribution, the required sample size can be

* Corresponding author.
E-mail address: dupjrs@126.com (P. Du).

http://dx.doi.org/10.1016/j.patcog.2015.08.019
0031-3203/

computed by a proportion between a specified half-width of the confidence interval, the estimated population standard deviation and a specified level of confidence [10,11]. Alternatively, if "the larger the better" principle is considered, the rule of thumb is that the training sample number for each class should be at least 10–30 times the number of bands [12].

In summary, even the most popular sampling protocols (e.g., probability sampling, clustering sampling, random sampling and systematic sampling) are not always able to provide the users with an easy way to implement and low cost solution [6], with a limited number of training samples. However, such a solution is very attractive to the remote sensing and pattern recognition communities, considering such factors as positive effects on image interpretation and labeling efficiency, data and field-survey accessibility, data usability, spatial coverage, data process and storage needs, as well as the complexity of the adopted algorithms or models [13–16].

To add a new dimension to this requirement, active learning (AL) has been introduced in recent years, with the aim to exploiting the information available from unlabeled data [17–20]. The protocol to label the originally unlabeled data in AL is usually handled by a user according to an informative (i.e. uncertainty and diversity criterion) measure or/and representative measure [18,20,21]. On the one hand, the search of new informative samples generally relies on the whole available unlabeled samples, i.e., the whole image is considered as the candidate set for AL. Unfortunately, this choice may become prohibitive when the data set is huge. On the other hand, another important issue in AL is how to define the initial training set, as the generalization of the initial training set is essential and critical. This is because that the class uncertainty increases as the number of labeled training samples decreases. Therefore, the initial labeled training set is important for the following AL procedure, which is guided by the initial samples.

Many efforts have been devoted to the use of a few predefined labeled and a large amount of new samples [17,18,20]. Although large amount of unlabeled samples are cheap to collect indeed, they are highly dependent on data storage space, RAM, data stream speed and computational capability, which are usually limited. Therefore, performing techniques with high efficiency and a much smaller amount of unlabeled samples should be interesting options. In this regard, some works (e.g., [22,23]) devoted to further reduce the costs beyond the fact that unlabeled data are cheaper to collect. Examples of these recent research works are [24], where a self-training approach has been adopted for uncertainty sampling to reduce misclassification costs in AL, and [23], where the active sample selection problem is addressed in the framework of a Markov decision process to optimize the collection of labeled samples by field surveys. Additionally, according to [25], AL techniques can converge to good classification accuracies starting from any initial training set without any assumption on the distribution of the classes, which means the unlabeled samples are more important. However, cost reduction by designing protocols to wisely select labeled and unlabeled samples has not attracted much attention yet.

We should recall that, ideally, if one is allowed to use the available information without any limitation, the best candidate set is the whole image. However, in reality some of the samples are hard to be labeled. For instance, a popular labeling process is using high- and very high spatial resolution satellite data through open systems like Google Earth™, this approach is often limited due to the lack of available image data or the temporal differences.

Another important issue is the computational complexity. In the AL iterative scheme, it is always time consuming to go through the whole image. Instead, an alternative solution is to design a candidate set only with suitable and smaller size. In general, for classifiers such as maximum likelihood classifier, decision tree (DT) and classier ensemble, the design of the training stage calls for a large number of pure pixels. On the contrary, mixed pixels may be able to provide more information than pure pixels in some scenarios. As a matter of fact, training sets containing mixed pixels were used for accurate remote sensing image classification in [26–28]. However, from the sample design point of view, the initial labeled and unlabeled samples should be composed of only pure pixels, or both pure and mixed pixels, are not clear in AL yet.

In this work a novel sampling protocol is proposed to generate the representative and refined candidate sets for the initial labeled training samples in AL procedure. The proposed protocol aims at improving the generalization capacity of the initial labeled samples, and guiding the AL procedure efficiently. In more detail, the proposed protocol, designed for hyperspectral images, consists of three steps:

(1) Initialization: apply noise elimination approach minimum noise fraction (MNF) to the original image.
(2) Pixel purity index (PPI) computation for each pixel.
(3) Selection of the candidate set: according to the PPI values, the entire image is subdivided into two parts, respectively representing pure pixels, mixed pixels.

To evaluate the performance of the proposed protocol, state-of-the-art AL approaches, including breaking ties (BT) [29], margin sampling (MS) [30], normalized entropy query-by-committee (nEQB) [20], multi-class level uncertainty (MCLU) [31], multi view adaptive maximum disagreement based active learning strategy (MVAMD) [32], and margin sampling by closest support vectors (MScSV) [33] were considered. Test results were obtained on three hyperspectral remote sensing images with different spatial resolutions. Please note that we do not compare here the performances of different AL methods, since it is out of the purpose of this work, also exclusive comparison studies can be found in many other works [20,31–33].

The rest of this paper is organized as follows: Section 2 presents the details of the proposed method. Section 3 presents the considered hyperspectral data sets with experimental setup. Section 4 provides the experimental results. Finally, Section 5 concludes the paper with some discussions.

2. The proposed method

Generally, AL methods attempt to generalize the statistical and/or geometrical properties of the samples $\mathbf{X}=\{x_i\}_{i=1}^{l}, x_i \in R^d$ with label $\boldsymbol{y}=\{y_1, y_2, ..., y_l\}$ by selecting the most informative samples from unlabeled samples$\mathbf{U}=\{x_i\}_{i=l+1}^{l+u}, x_i \in R^d$ according to informative measures, i.e. uncertainty and/or diversity criteria. This procedure can be compactly described by

$$f^* = \underset{f}{\operatorname{argmin}} \frac{1}{2}\{f\}^2 + \sum_{i=l+1}^{l+u} \xi(y_i, f(x_i)) \tag{1}$$

where f^* denotes a classification model trained by labeled samples, f is the decision function, and ξ is the loss function based on the unlabeled samples x_i.

In order to make the classification model as efficient as possible, usually the existing labeled sample count is very limited and AL is focused on the most informative samples that really help to improve the performance of the model [33]. Location, properties and number of the initial labeled and unlabeled samples are crucial to any classification approaches. To stress that, it is helpful to briefly recall the basic relationship between sample size, estimated error rate and the upper bound of the noise rate [34].

Let's consider a sequence of m samples drawn from two sets of labeled and unlabeled samples $\left[\mathbf{X}=\{x_i\}_{i=1}^{l}, \mathbf{U}=\{x_i\}_{i=l+1}^{l+u}\right]$. For a given worst-case classification error rate ε, confidence δ, upper bound on the classification noise rate η (<0.5), and N the number