ENSEMBLE MARGIN BASED SEMI-SUPERVISED RANDOM FOREST FOR THE CLASSIFICATION OF HYPERSPECTRAL IMAGE WITH LIMITED TRAINING DATA

Wei Feng[1], *Wenjiang Huang*[1,2], *Gabriel Dauphin*[3], *Junshi Xia*[4], *Yinghui Quan*[5], *Huichun Ye*[1,2], *Yingying Dong*[1]

[1]Key laboratory of Digital Earth Science, Aerospace Information Research Institute,
Chinese Academy of Sciences, Beijing 100094, China
[2]Key Laboratory for Earth Observation of Hainan Province,
Sanya Institute of Remote Sensing, Sanya 572029, China
[3]Laboratory of Information Processing and Transmission,
L2TI, Institut Galilée, University Paris XIII, France
[4]Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan
[5]Key Laboratory for Radar Signal Processing, Xidian University, Shaanxi 710071, China
E-mail: huangwj@radi.ac.cn, fengwei@radi.ac.cn

ABSTRACT

In this paper, we propose a novel ensemble margin based semi-supervised random forest (EMRF) algorithm for the classification of the hyperspectral image with limited training data. The proposed method tries to improve the effectiveness of the ensemble model via adaptively labeling the unlabeled instances with high classification probability then adding them into the training set. The classification probability of a training instance is reflected by the unsupervised margin value of this instance. The higher ensemble margin of an instance, the higher probability the instance being classified correctly and added into to the training set in the next iteration.

Index Terms— Random forest, semi-supervised learning, ensemble margin, classification, hyperspectral image.

1. INTRODUCTION

Due to the great potential to provide distinct abilities to object recognition and detection, the classification of hyperspectral remote sensing imagery becomes one of the most attractive research fields. Supervised learning is an important method in the remote sensing community. This method strongly relies on the quantity and quality of the reference samples. However, the ground campaign of labeling samples is expensive for most real remote sensing image [1]. When the size of the training set is very limited in a comparison of hundreds or thousands of dimensions in hyperspectral data, it frequently results in lower performances for most supervised classifiers.

Ensemble learning is an effective method to develop accurate classification systems [2, 3]. Random Forest (RF) [4], a powerful ensemble learning method, has shown to be particularly competitive with state-of-the-art learning methods, such as boosting [5]. Important advantages such as running efficiently on large databases, handling thousands of input variables without variable deletion and low time cost make RF widely attract the interest of researchers in remote sensing fields [6, 7, 8, 2]. However, RF is one of the supervised methods, i.e. its performance is also affected by the size of the training set. To increase the accuracy of RF, the information of the unlabeled hyperspectral instances have to be explored.

Ensemble margin plays a crucial role in modern machine learning research, and has been demonstrated effective in data selection, noise removal and imbalance learning [9, 6]. In this paper, we propose an ensemble margin based semi-supervised random forest (EMRF) algorithm for the classification of the hyperspectral image with limited training data. This method improves the effectiveness of the RF via adaptively labeling

the unlabeled instances with high classification probability then adding them into the training set.

2. METHOD

2.1. Margin theory and classification probability

There are two kinds of ensemble margins: supervised and unsupervised margins [9]. The first one, introduced by Shapire et al., is the difference between the number of votes for the true class and the maximal number of votes for any other class [9, 3]. The second one, which is the unsupervised version of Schapire's margin, uses the votes number of the most class to instead the votes number of the true class. When a set of M unlabeled samples is denoted as $U = \{(x'_i)\}_{i=1}^{M}$, the classification probability $\rho(x'_i)$ of an instance x'_i could be computed based on the unsupervised margin definition (equation 1).

$$\begin{aligned}\rho(x'_i) &= margin(x'_i) \\ &= \frac{v_{c_1}(x'_i) - v_{c_2}(x'_i)}{\sum_{c=1}^{L}(v_c(x'_i))} \quad \forall(x'_i) \in U \end{aligned} \tag{1}$$

where v_{c_1} is the votes number of the most voted class c_1 for sample x', v_{c_2} is the votes number of the second most popular class c_2, and L represents the number of classes. The higher $\rho(x'_i)$ of an instance x'_i means that the data has a high probability being classified correctly and added into to the training set in the next iteration.

2.2. Ensemble Margin based semi-supervised Random Forest

The process of the proposed ensemble margin based semi-supervised random forest (EMRF) method is detailed in Algorithm 1. Let us denote a training dataset as $S = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is a vector with feature values, y_i is the class label of the vector and N is the number of the instances. The ensemble size is denoted as T. The proposed method first constructs a robust ensemble classifier *random forest* E with the whole training set S. Then the ensemble model E is applied on the unlabeled dataset U to calculate the classification probability of each training instance of U. The unlabeled instances are ordered, in descending order, according to their ρ values. The first 1% instances are labeled with the predicted class of the ensemble model E to form a data set U'. The instances of U' are removed from the unlabeled data set U and added into S to form a new training set S', then $S = S'$. We construct a random forest E' with all the training data in S, and update E with E'. The aforementioned steps are repeated until the maximum iteration number I is reached.

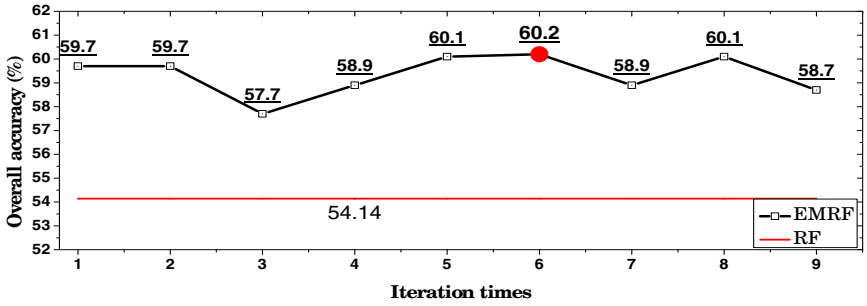
Algorithm 1 Ensemble Margin based semi-supervised Random Forest (EMRF)

1: **Input** $S = \{(x_i, y_i)\}_{i=1}^{N}$: training set; $U = \{(x'_i)\}_{i=1}^{M}$: unlabeled data set; I: iteration times; $E = \varnothing$: an ensemble.
2: **Initialization**
3: Construct a random forest E with all the training data $(x_i, y_i) \in S$
4: **for** i=1:I **do**
5: Use the ensemble E to predict the unlabeled data set U then compute the margin value for each training instance x' ($x' \subseteq U$).
6: Order the instances of U according to their classification probability P (equation 1), in descending order
7: Label the first $\theta\%$ instances of U to form a data set S' ($S' \subseteq U$)
8: Remove the instances of S' from the unlabeled data set U, and update the unlabeled data set $U = U - S'$
9: Add the instances of S' into S to form a new training set S_{new}, then $S = S_{new}$
10: Build an ensemble model E on the dataset S
11: **end for**
12: **Output** The ensemble E

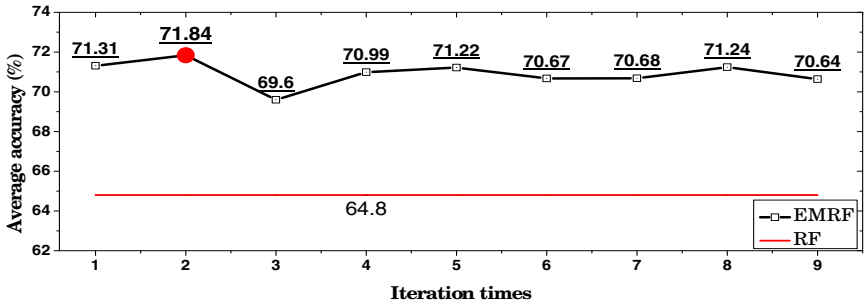
3. EXPERIMENTAL RESULTS

3.1. Datasets

The EMRF is evaluated on two standard hyperspectral images *AVIRIS Indian Pine* and *University of Pavia ROSIS*. The first image is composed of 145*145 pixels, with a spatial resolution of 20 m/pixel, 16 classes, and 200 spectral bands. The second one is an urban area consisting of 610*340 pixels with a spatial resolution of 1.3 m/pixel, nine different classes, and 103 spectral bands. For each hyperspectral image, the reference dataset is divided into three independent parts: the training set, unlabeled data set, and test set. Initially, we randomly selected 20 labeled samples per class as the training set, half of the other as the unlabeled dataset and the rest as the test set.



(a) Overall accuracy



(b) Average accuracy

Fig. 1. The accuracy results of the data *AVIRIS Indian Pine* achieved by the proposed EMRF using multiple iterations.

3.2. Results

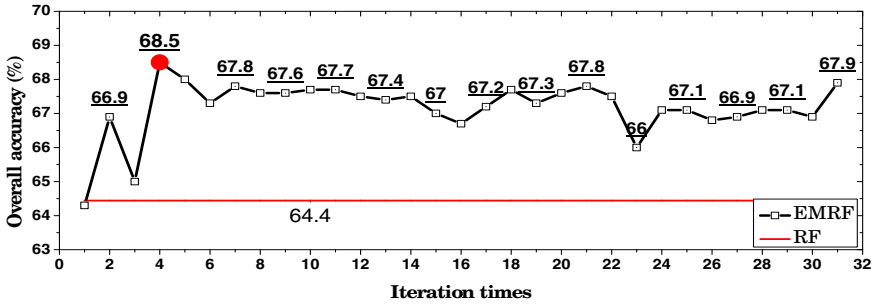
For the performance evaluation of the proposed method, the state-of-the-art learning method RF [4] is utilized in the comparative analysis. All ensembles are implemented with 100 trees. All the reported results are mean values of 30-time calculations. Figures 1 and 2 respectively illustrate the performance comparison of the proposed method and the classic RF by increasing the number of iteration on the data *AVIRIS Indian Pine* and the data *University of Pavia ROSIS*. As expected, the proposed method provides the best results, both in terms of general average accuracy (AA) and overall accuracy (OA). When compared with the supervised RF, the proposed algorithm improves the AA about **6%** (data *AVIRIS Indian Pine*) and about **4%** (data *University of Pavia ROSIS*). The achieved OA improvement of EMRF is about **7%** for dataset *AVIRIS Indian Pine* with respect to RF. In addition, both figures show accuracy curves of the proposed method are with respect to the iteration time I. By increasing I to an optimum value for different data sets, the classification accuracy is also improved.

4. CONCLUSION

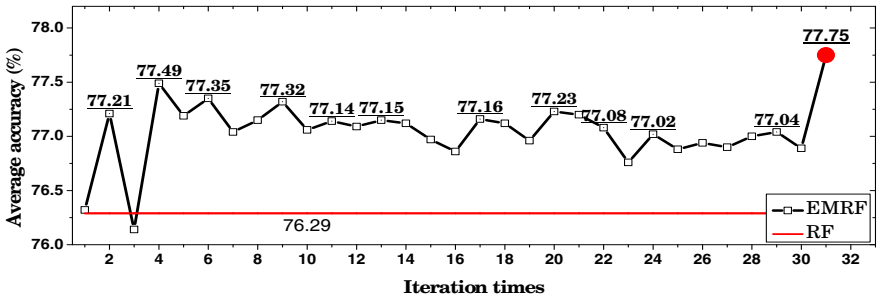
In this paper, we propose an EMRF algorithm which combines RF and margin theory for the training data limitation problem of the hyperspectral image. The proposed method effectively uses unlabeled samples in the learning stage to increase the number of training instances. The margin theory is employed to mining the informative samples from the unlabeled dataset to improve the quality of the training set. Experimental results on the overall accuracy and average accuracy indicate the superiority of EMRF over classic RF. In our future work, the parameter of the iteration time I will be estimated using out-of-bag samples.

5. ACKNOWLEDGMENTS

This work is supported by the National special support program for high-level personnel recruitment (Wenjiang Huang), Ten-thousand Talents Program (Wenjiang Huang), Hainan Provincial Key R&D Program of China(ZDYF2018073), the National Natural Science Foundation of China (41601466, 41871339, 61461003).



(a) Overall accuracy



(b) Average accuracy

Fig. 2. The accuracy results of the data *Pavia University ROSIS* achieved by the proposed EMRF using multiple iterations.

6. REFERENCES

[1] X. Ma, H. Wang, and J. Wang, "Semisupervised classification for hyperspectral image based on multi-decision labeling and deep feature learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 120, pp. 99 – 107, 2016.

[2] J. Xia, N. Falco, J. A. Benediktsson, J. Chanussot, and P. Du, "Class-separation-based rotation forest for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 4, pp. 584–588, 2016.

[3] W. Feng and W. Bao, "Weight-based rotation forest for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 11, pp. 2167–2171, Nov 2017.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[5] S. Bernard, S. Adam, and L. Heutte, "Dynamic random forests," *Pattern Recognition Letters*, vol. 33, no. 12, pp. 1580 – 1586, 2012.

[6] W. Feng, W. Huang, and J. Ren, "Class imbalance ensemble learning based on the margin theory," *Applied Sciences*, vol. 8, no. 815, 2018.

[7] B. Fu, Y. Wang, A. Campbell, Y. Li, B. Zhang, S. Yin, Z. Xing, and X. Jin, "Comparison of object-based and pixel-based random forest algorithm for wetland vegetation mapping using high spatial resolution gf-1 and sar data," *Ecological Indicators*, vol. 73, pp. 105 – 117, 2017.

[8] F. Ctanovas-Garctia, F. Alonso-Sarrtia, F. Gomariz-Castillo, and F. Onate-Valdivieso, "Modification of the random forest algorithm to avoid statistical dependence problems when classifying remote sensing imagery," *Computers & Geosciences*, vol. 103, pp. 1 – 11, 2017.

[9] W. Feng and S. Boukir, "Class noise removal and correction for image classification using ensemble margin," in *2015 IEEE International Conference on Image Processing (ICIP)*, Sept 2015, pp. 4698–4702.