Title: An Effective Strategy to Reduce the Labeling Cost in the Definition of Training Sets by Active Learning

This paper appears in: IEEE Geoscience and Remote Sensing Letters

Date of Publication: 2014

Author(s): Begüm Demir, Luca Minello, Lorenzo Bruzzone,

Volume:11, Issue: 1

Page(s): 79-83

DOI: 10.1109/LGRS.2013.2246539

An Effective Strategy to Reduce the Labeling Cost in the Definition of Training Sets by Active Learning

Begüm Demir, *Member, IEEE,* Luca Minello, Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—This letter proposes a novel strategy for reducing the cost of *in situ* sample labeling for the definition of training sets by active learning (AL) in the framework of supervised classification of remote sensing images. AL methods define a training set according to an iterative procedure that at each iteration requires the labeling of a set of new samples selected by the classifier. The proposed strategy can be embedded in any AL method in order to identify the most informative area on the ground where focusing each AL iteration to reduce the overall cost (in terms of time) of labeling. To this end at each iteration the most uncertain unlabeled samples are initially identified. Then the area on the ground (having a size predefined by the user) that has the highest spatial density of informative (i.e. uncertain and diverse) unlabeled samples is selected by the proposed strategy, and the AL technique is applied only to the samples of that area. This results in a decrease of the overall labeling cost with respect to that required by the use of a given technique in a standard way. Experimental results obtained by embedding the presented strategy in different literature active learning methods confirm its effectiveness.**

***Index Terms*— Active learning, automatic classification, *in situ* data collection, training set, clustering, remote sensing.**

I. Introduction

Active learning (AL) has received an increasing interest in the remote sensing community in the context of supervised classification techniques [1]-[7]. AL aims to build up non-redundant and effective training sets according to an iterative process that requires an interaction between a human expert and the automatic classification system. At each iteration, the most informative unlabeled samples for the considered classifier are selected; then the classification system interacts with a human expert to obtain the true class labels for the selected samples. These samples are added to the training set after manually labeling, and the supervised algorithm is retrained with the enriched training set [1]. The iterative process converges to an effective training set that includes a minimum number of very informative labeled samples. This is a very important result as collecting samples' labels is highly costly in terms of human time and effort, and thus also in terms of money.

To select the most informative batch of samples to be labeled, most of the AL papers presented in the remote sensing literature exploit two criteria [1]- [4]: the uncertainty and the diversity criteria. The former aims at selecting the unlabeled samples that have the lowest confidence to be correctly classified by the considered classifier. The latter identifies uncertain samples that are as distant as possible to each other in the feature space (to decrease the redundancy among the selected samples) [1]. The joint use of these two criteria results in the collection of the informative set of samples. A literature survey of AL methods presented in the remote sensing literature can be found in [3]. These methods assume that reduction in the number of samples being labeled guarantees the reduction in the total cost in terms of human time and effort. However, applicability of this assumption directly depends on the approach considered in the label collection process.

In remote sensing, labeling of samples can be achieved by the use of different approaches: 1) *in situ* surveys, 2) image photointerpretation, or 3) hybrid strategies (both photointerpretation and *in situ* surveys) [1]. *In situ* surveys are highly expensive, yet strictly required if detailed land-cover classes should be recognized. The cost in this case is directly related to the traveling time of the human expert to the ground locations of the samples to be labeled. Thus, when very large images are considered and samples geographically distant to each other should be labeled, the cost can become very high. In these cases we can state that using the AL strategy for reducing the number of labeled samples cannot assure to achieve the minimum cost in terms of human time and effort. According to our knowledge, only few methods exist in the remote sensing literature that considers the label acquisition costs within the AL process [5]-[7]. In these works, the labeling cost is measured with respect to the time required to travel during the labeling process. The method presented in [5] initially selects the most uncertain samples, and then defines the shortest path to travel among these samples according to the traveling salesman problem. In [6], the uncertain samples that are closest to each other are selected by solving a traveling salesman problem with profits. In [7], the selection of cost-efficient and informative samples in terms of uncertainty and diversity is achieved by using a simple sequential forward selection strategy. Possible requirements on the different transportation modes (i.e., foot or car) with respect to the distance between samples are also considered. However, these methods are not efficient when very large images are considered due to the fact that they are based on a random initialization of the position on the ground from which

starting the labeling process. This can strongly affect the cost required to converge to an effective training set. Moreover, informative samples in the considered area can be distant to each other, thus increasing the total distance to travel for the labeling process at convergence.

In order to deal with the above-mentioned problems, in this letter we present a novel strategy that can be embedded in any AL method. The main goal of the proposed strategy is to reduce the cost of *in situ* sample labeling for the definition of an optimized training set (i.e., a training set that includes samples that are both highly informative and cheap to label) when very large areas are considered. This is achieved by identifying, at the beginning of each AL iteration, the most informative area on the ground that has the highest spatial density of informative (i.e. uncertain and diverse) unlabeled samples. After selecting the small study area by the proposed strategy, the considered AL technique is applied only to the samples of that area.

The paper is organized into five sections. Section II introduces the proposed strategy to AL. Section III describes the considered data sets and illustrates the design of experiments. Section IV shows the experimental results. Finally, Section V draws the conclusion of this work.

II. Proposed Strategy for Active Learning

In this section, we introduce the proposed novel strategy aimed to optimize the cost in terms of human time and effort in the definition of a training set by *in situ* ground survey with AL. Given a generic AL method, the proposed strategy aims at identifying at each AL iteration the area on the ground having the highest density of informative unlabeled samples. This is done to select a small yet informative portion of the image (i.e., the study area) on which running the AL method. The objective of selecting a small portion of the image is to restrict the area to be analyzed on the ground for labeling the samples during the AL process, and thus to reduce the traveling time, while considering very informative samples. This is done by guaranteeing that the selected area contains highly informative samples for the considered classifier. In this way, particularly for very large images, we expect that the total human effort and time for labeling the samples can be significantly reduced. It is worth noting that our aim is not to propose a new AL method, but to present a novel strategy to properly reduce the cost of sample labeling within any AL method.

The proposed strategy assumes that the most informative portion of the image for AL is the one that contains the highest density of both uncertain and diverse samples. To select this portion we define two steps: i) assessment of the spatial density and diversity of the most uncertain samples in different small portions of the image, and ii) selection of the portion in which the density of both diverse and uncertain samples is high. To examine the entire image the steps of the proposed strategy are conducted on a moving window approach. Each step is explained in detail in the following:

Step 1-Assessment of the spatial density and diversity of the most uncertain samples in the entire image: Initially the *m*-most uncertain unlabeled samples are selected by the considered AL method. Then, in order to assess the spatial density of these uncertain samples, thus to localize candidate small areas for AL, we apply to the entire image a square shaped spatial moving window (having a size *w* predefined by the user). Then, for each portion (i.e., window) P_j, $j=1,2,...,n$, where P_j is *j*-th portion of the entire image, the spatial density of uncertain samples D_j^{unc}, $j=1,2,...,n$, is estimated as

$$D_j^{unc} = \frac{|X_j|}{|P_j|} \qquad j=1,2,...,n \tag{1}$$

where X_j is the set of uncertain samples located in the portion P_j, $|\,|$ is the cardinality function and *n* is the total number of portions being analyzed. The portions that have low density of uncertain samples are rejected, whereas the other ones are remained. To this end the density D_j^{unc} of each portion is compared with a threshold *T*. If $D_j^{unc} > T$, P_j will be remained due to its high density on uncertain samples. On the contrary, if $D_j^{unc} < T$, P_j will be rejected. The size *w* of the moving window should be defined by the human expert according to the considered application as well as the geographical characteristics of the study area.

Then, the spatial density D_j^{div}, $j=1,2,...l$ of diverse uncertain samples located in the remained spatial portions P_j, $j=1,2,...,l$ is analyzed ($l \le n$ is the total number of remained portions). This is achieved by using a clustering method due to the fact that samples assigned to different clusters can be considered as diverse samples and vice versa. Accordingly, a clustering method is applied to the most uncertain samples located in the remaining portions of the image. After applying the clustering, the set C_j, $j=1,2,...l$ of cluster labels, where C_j is the set of cluster labels for P_j, are obtained. Then, each portion is analyzed in order to assess the diversity of uncertain samples included in it. This is achieved according to the number $\overline{C_j}$ of different cluster labels in the set C_j. The total number $\overline{C_j}$ of different cluster labels is simply estimated as the sum of the diverse cluster labels in C_j, and thus $D_j^{div} = \overline{C_j}$. If $\overline{C_j}$ is high, the portion P_j is assumed as highly informative and vice versa. Here, we use the kernel *k*-means technique [8] for clustering due to its already proven effectiveness also in AL problems [1], [3]. Nonetheless, any clustering technique can be exploited.

Step 2-Selection of the most informative portion of the image: The second step of the proposed strategy is devoted to select the final portion of the image in which we have the highest spatial density of the diverse and uncertain samples. Accordingly, the *v*-th portion that has the highest diversity (i.e., maximum number of different cluster labels) is selected as the best portion of the image as

$$v = \arg\max_{j=1,2,...,l} \left\{ D_j^{div} \right\} \tag{2}$$

If there are *t*<*l* portions having the same highest spatial density of the diverse uncertain samples, the *v*-th portion having the

highest density of uncertain samples is selected as the final best portion of the image, i.e.,

$$v = \arg\max_{j=1,2,...t} \left\{ D_j^{unc} \right\} \quad (3)$$

After selecting the best portion of the image to be investigated, the considered AL method is applied for collecting a batch X of h samples from this portion. If the considered AL method only includes an uncertainty criterion, the most uncertain samples are directly selected from this portion according to their already estimated uncertainty values at the first step of the proposed strategy. In other words, it is not required to re-estimate these uncertainty values. However, if the considered AL method includes also the diversity criterion, it will be applied to the most uncertain samples located in the selected portion of the image for the final selection of the diverse and uncertain samples. Then, these samples are labeled by the human expert on the ground and added to the current training set. Algorithm 1 summarizes the single AL iteration considering our proposed strategy:

Algorithm 1: AL with the Proposed Strategy

Input:

m (number of samples selected on the basis of their uncertainty)
h (batch size)
T (threshold)
k (total cluster number)
w (moving window size)

Output:

X (set of unlabeled samples to be included in the training set)

1. Select the m-most uncertain samples on the basis of considered AL method.
2. Analyze the spatial density D_j^{unc}, $j = 1,2,...,n$ of the m-most uncertain samples within n different portions of the image.
3. Select the $l <= n$ portions that have a density higher than T and reject the others.
4. Apply the kernel k-means to the samples included in the remaining $l <= n$ portions.
5. Estimate the density D_j^{div}, $j = 1,2,...l$ of the diverse samples in each of the $l \le n$ selected sub-images.
6. Select the best portion of the image according to (2) or (3).
7. Select the batch X of h samples from the selected portion of the image on the basis of the considered AL method.

The steps of AL are iterated until the desired number of samples is labeled, i.e., the upper bound of the cost (in terms of time spent on the ground) for labeling the samples is achieved. When the AL process is completed, the image is classified by the considered classifier. It is worth noting that proposed strategy is general and can be embedded in any AL method as well as any classifier.

III. Data Set Description and Design of Experiments

A. Data set description

Experimental analyses are conducted on two multispectral data sets. The first data set is an image acquired by the GeoEye multispectral sensor on the city of Trento (Italy) in September 2011 (see Figure 1). The selected test site is a section of 4000×4700 pixels with a spatial resolution of 2 m. For this data set, the available ground reference samples are representative of the five land cover classes (i.e., water, urban, field, forest, shadow). They were randomly divided to derive a validation set of 231 samples (which is used for model selection), a test set of 8210 samples (which is used for accuracy assessment) and an unlabeled samples set of 7638 samples. The 0.8% of the samples of each class in the unlabeled samples set were randomly selected as initial training samples for a total of 63 samples, and the rest were considered as unlabeled samples. The second data set is a pan-sharpened image acquired by the Quickbird multispectral sensor on the city of Trento in October 2005. The selected test site is a section of 2066×3100 pixels with a spatial resolution of 0.7 m. For this data set, the available ground reference samples are representative of the five land cover classes (i.e., water, road, field, forest, bare soil). They were randomly divided to derive a validation set of 195 samples, a test set of 2902 samples, and an unlabeled samples set of 2688 samples. The 2% of the samples of each class in the unlabeled samples set were randomly selected as initial training samples for a total of 57 samples, and the rest were considered as unlabeled samples.

Figure 1. True color composite of the Trento GeoEye data set

B. Design of experiments and parameter setting

In the experiments, we used the Support Vector Machine (SVM) classifier with Radial Basis Function (RBF) kernel. We considered an one-against-all architecture of binary SVMs for addressing out multiclass problems [9]. The values for the regularization parameter C and the spread γ of the RBF kernel were chosen performing a grid-search model selection only at the first iteration of the AL process as suggested in [1]. In the experiments, the size w of the moving window was defined by limiting the human expert's traveling area to 1 km^2 for the first data set, whereas that of the second data set was set to 0.1 km^2. As a result, for both data sets the moving window size w was 500×500 pixels. In our experiments, we selected the total cluster number k as the total number of land-cover classes present in the data sets (i.e., k=5 for both data sets). The value of m (number of uncertain samples selected at

the first step of the proposed strategy) was selected equal to 200. In the experiments, at each iteration of the AL process, h=5 samples were added to the current training set, whereas the threshold T is set equal to the batch size h. All experimental results are provided as the average accuracies of ten trials that are obtained according to ten initial randomly selected training sets.

We applied the proposed strategy to two of the most effective state of the art AL methods presented in the remote sensing literature, i.e. the Multiclass-Level Uncertainty with Enhanced Clustering Based Diversity (MCLU-ECBD) technique [1] and the Entropy Query by Bagging (EQB) technique [2]. The MCLU-ECBD technique selects the most informative unlabeled samples by the MCLU strategy, and then assesses the diversity of the most uncertain samples by a kernel-clustering technique. [1]. The EQB technique assesses the uncertainty of samples according to the maximum disagreement between a committee of classifiers. The disagreement among classifiers is evaluated by the entropy of the distribution of the different labels obtained by a committee of classifiers [2]. The results of EQB are obtained fixing the number of EQB predictors to eight and selecting bootstrap samples containing 75% of initial training patterns. These values have been suggested in [2]. It is worth emphasizing that the MCLU-ECBD technique analyzes both the uncertainty and diversity of samples, whereas the EQB technique only evaluates the uncertainty of samples. In our experiments, the results obtained by the ECBD-MCLU and EQB techniques when applied with the proposed strategy are denoted as MCLU-ECBD-S and EQB-S, respectively. In order to show the effectiveness of the MCLU-ECBD-S and EQB-S, we have compared the results with those obtained without using the proposed strategy (i.e., using the standard ECBD-MCLU and EQB). All the results are provided as learning rate curves, which show the average classification accuracy versus the total time spent during the labeling process. The labeling time of all the methods was calculated using the same approach proposed in [6], where the shortest path to travel among the selected samples is defined according to the traveling salesman problem. In our experiments, we set the velocity of traveling by supervisor to 6 km/hours, whereas the time taken by the supervisor to assign a label to each sample was set to 2 minutes.

IV. Experimental Results

A. Results: GeoEye Data Set

The small areas selected by the proposed strategy for this data set at the first five AL iterations are shown in Figure 2. From the figure, one can see that the selected areas contain the highest spatial concentration of samples of the urban, field and forest classes. This is due to the fact that these samples have the lowest confidence on their correct class label (thus are much more informative from the AL point of view) compared to those of water and shadow classes.

Figure 3 shows the behavior of the average (on 10 trials) classification accuracies versus the time obtained by the proposed MCLU-ECBD-S and the standard MCLU-ECBD (see Figure 3.a), as well as the proposed EQB-S and standard EQB (see Figure 3.b). From the figure, one can observe that the MCLU-ECBD-S and EQB-S lead to the highest accuracies for all the iterations and significantly outperforms the MCLU-ECBD and EQB, respectively. As an example, the MCLU-ECBD provides an accuracy of 90.95% when 14 hours are spent for the collection of ground reference data. However, the accuracy obtained by the proposed MCLU-ECBD-S in the same time is 93.73% (see Figure 3.a). Moreover, the proposed MCLU-ECBD-S provides an accuracy of 94.02% spending 23 hours, whereas the standard MCLU-ECBD requires 48 hours to achieve a similar accuracy. The similar performances are also observed when comparing the EQB-S with the standard EQB (see Figure 3.b). Moreover, by analyzing the figure one can observe that the proposed MCLU-ECBD-S and EQB-S reach convergence in a smallest time with respect to MCLU-ECBD and EQB, respectively. These results demonstrates the importance of proposed strategy for optimizing the definition of the training sets with the lowest sample labeling cost in terms of traveling time.

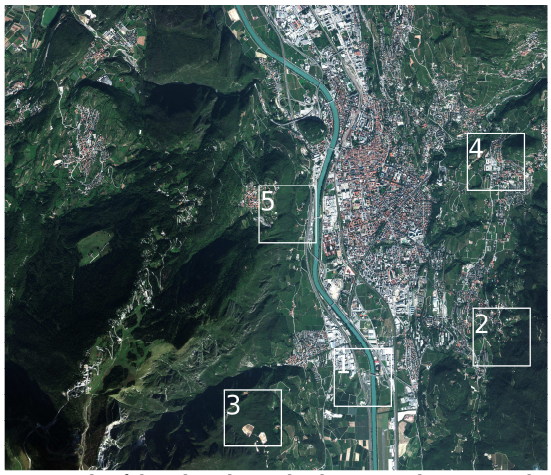

Figure 2. Example of the selected areas by the proposed strategy at the first 5 iterations of AL for one trial. The number of AL iteration is given on the left corner of the selected area.

B. Results: Quickbird Data Set

For this data set, due to the space limitations, we report only the qualitative results. Figure 4 shows the results obtained by the MCLU-ECBD-S and the standard MCLU-ECBD (see Figure 4.a); and the EQB-S and standard EQB (see Figure 4.b), respectively. By analyzing the figure, one can observe that also on this data set the MCLU-ECBD-S and EQB-S result in the highest accuracies for all the iterations. As an example, the MCLU-ECBD provides an accuracy of 90.05% when 4 hours are spent for the collection of ground reference data. However, the accuracy obtained by the proposed MCLU-ECBD-S in the same time is 92.99% (see Figure 4.a). Moreover, the proposed MCLU-ECBD-S provides an accuracy of 94.20% spending 8 hours for the label collection process, whereas the standard MCLU-ECBD requires 19 hours to achieve a similar accuracy. Similar behaviors are also observed when comparing the EQB-S with standard EQB. Moreover, by analyzing the figure one can see that the proposed the MCLU-ECBD-S and EQB-S, again, reach convergence in a smallest time compared to the MCLU-ECBD and EQB, respectively.

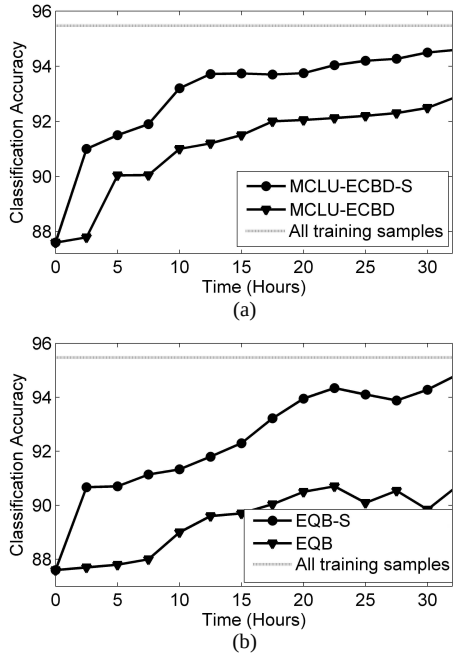

Figure 3. Average overall classification accuracy obtained by (a) the proposed MCLU-ECBD-S and the standard MCLU-ECBD; (b) the proposed EQB-S and the standard EQB for the Trento GeoEye data set. The dashed line "All training samples" shows the accuracy obtained including all unlabeled samples in the training set after manual labeling.

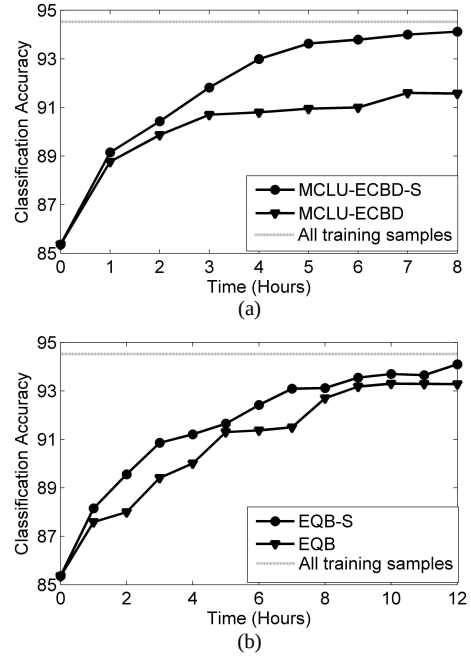

Figure 4. Average overall classification accuracy obtained by (a) the proposed MCLU-ECBD-S and the standard MCLU-ECBD method; (b) the proposed EQB-S and the standard EQB for the Trento Quickbird data set. The dashed line "All training samples" shows the accuracy obtained including all unlabeled samples in the training set after manual labeling.

It is worth noting that on this data set the differences on the accuracies obtained with and without exploiting the proposed strategy are not as high as those obtained on the previous data set. This is mainly due to the fact the proposed strategy is more effective for large images due to larger number of small areas available for AL.

V. Conclusions

In this paper, a novel strategy has been presented for reducing the cost of the definition of training sets by *in situ* survey with AL methods. The proposed strategy identifies at each AL iteration a small portion of the analyzed image (i.e., of the area on the ground) that contains the highest spatial concentration of informative (i.e., uncertain and diverse) unlabeled samples. Then the considered AL method is applied to the samples of that area. In this way, focusing on small areas having the highest density of informative unlabeled samples, it is possible to limit the traveling time required for the human expert for defining the final training set. It is worth emphasizing that the proposed strategy is general and can be used with any AL method as well as with any classifier.

The proposed strategy has been applied to two different data sets simulating the *in situ* label collection process. It has been implemented by using two effective AL methods taken from the literature, i.e. the MCLU-ECBD and the EQB. In all experiments the results show that the proposed strategy allows one to significantly reduce the cost of the collection of reference samples to reach the target classification accuracy.

As a final remark, we point out that the proposed strategy is very promising for possible operational applications due to crucial ability to significantly reduce the sample labeling cost in terms of human time and effort. As a future development of this work, we plan to develop a tool with an interface based on the proposed strategy, to implement it with different AL techniques and then to test it for real label collection on the ground.

References

[1] B. Demir, C. Persello, and L. Bruzzone, "Batch mode active learning methods for the interactive classification of remote sensing images", *IEEE Trans. on Geoscience and Remote Sensing*, vol. 49, no.3, pp. 1014-1031, March 2011.

[2] D. Tuia, F. Ratle, F. Pacifici, M. Kanevski, and W. J. Emery, "Active Learning methods for remote sensing image classification", *IEEE Trans. on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218 -2232, Jul. 2009.

[3] L. Bruzzone, C. Persello, B. Demir, Active Learning Methods in Classification of Remote Sensing Images, in Signal and Image Processing for Remote Sensing, 2nd Edition, Ed: Prof. C.H. Chen, CRC Press - Taylor & Francis, Chapter 15, 2012, pp. 303-323.

[4] S. Patra, L. Bruzzone, "A batch-mode active learning technique based on multiple uncertainty for SVM classifier, " *IEEE Geoscience and Remote Sensing Letters*, vol.9, no.3, pp.497-501, May 2012.

[5] A. Liu, G. Jun and J. Ghosh, "Spatially cost-sensitive active learning", *In SIAM International Conference on Data Mining*, Sparks, Nevada, USA, pp.814-825, 2009.

[6] A. Liu, G. Jun, and J. Ghosh, "Active learning of hyperspectral data with spatially dependent label acquisition costs", *IEEE International Geoscience and Remote Sensing Symposium*, Cape Town, South Africa, pp. V-256 - V-259, 2009.

[7] B. Demir, L. Minello, L. Bruzzone, "A cost-sensitive active learning technique for the definition of effective training sets for supervised classifiers", *International Conference on Geoscience and Remote Sensing Symposium*, Munich, Germany, 2012.

[8] R. Zhang and A. I. Rudnicky, "A Large scale clustering scheme for kernel k-means," *IEEE International Conference on Pattern Recognition*, 2002, Quebec, Canada, pp. 289-292.

[9] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification", *IEEE Trans. on Geoscience and Remote Sensing*, vol. 43, no. 6, pp. 1351-1362, Jun. 2005.