Korean Journal of Remote Sensing, Vol.22, No.3, 2006, pp.211~219

Application of Random Forests to Assessment of Importance of Variables in Multi-sensor Data Fusion for Land-cover Classification

No-Wook Park† and Kwang-Hoon Chi

Geoscience Information Center, Korea Institute of Geoscience and Mineral Resources

Abstract : A random forests classifier is applied to multi-sensor data fusion for supervised land-cover classification in order to account for the importance of variable. The random forests approach is a non-parametric ensemble classifier based on CART-like trees. The distinguished feature is that the importance of variable can be estimated by randomly permuting the variable of interest in all the out-of-bag samples for each classifier. Two different multi-sensor data sets for supervised classification were used to illustrate the applicability of random forests: one with optical and polarimetric SAR data and the other with multi-temporal Radarsat-1 and ENVISAT ASAR data sets. From the experimental results, the random forests approach could extract important variables or bands for land-cover discrimination and showed reasonably good performance in terms of classification accuracy.

Key Words : Random Forests, Data Fusion, Classification.

1. Introduction

Since the 1990s, the use of multi-sensor remote sensing data has been gaining increased interests in remote sensing communities. The forthcoming ranges of hyperspectral data, recently available high resolution data (e.g. pixel resolution of 1m or less) and polarimetric SAR data from several space-borne platforms, such as KOMPSAT-2, ALOS, Radarsat-2, will provide us with unprecedented opportunity for Earth observation tasks.

To make optimized decisions, better use must be made of all available information acquired from different sensors or sources. Remote sensing data acquired over the same site by different sensors are, in general, partially redundant or complementary, since they have different characteristics and physical interaction. Multi-sensor data fusion may help in the extraction of more information with higher accuracy and less uncertainty. In case of land-cover classification, all land-cover classes in each image may not be distinguishable. If complementary information provided by different sensors can be combined in a data fusion framework, separation between various land-cover classes can be achieved more effectively (Park, 2004).

Received 13 April 2006; Accepted 21 June 2006.

† Corresponding Author: N. – W. Park (nwpark@kigam.re.kr)

Several methodologies for multi-sensor data fusion have been tested and refined with rigorous theoretical backgrounds. The main approaches to data fusion in the remote sensing literature are statistical methods (Solberg *et al.*, 1996), Dempster-Shafer theory of evidence (Le Hégarat-Mascle *et al.*, 1997), fuzzy set theory (Solaiman *et al.*, 1999) and neural network (Serpico *et al.*, 1996).

Unlike the situation where single sensor data are only dealt with, however, one of the most serious problems faced in multi-sensor data fusion is the information content and relative reliability of each sensor (Park, 2004). Since data come from various sensors, the data inevitably have varying degrees of reliabilities for targets. Hence the relative reliabilities and uncertainties of the sensors should be properly accounted for during data fusion processes.

To account for relative reliability or importance, this paper applies random forests to multi-sensor data fusion for land-cover classification. The random forests approach is one of ensemble methods and a non-parametric one and thus it can be effectively applied to multi-sensor data fusion. By randomly deleting the information contained in a certain variable in the out-of-bag samples for each classifier, especially, the importance of variable can be computed. The potentiality of the methods was evaluated from two experiments for land-cover classification with multi-sensor data.

2. Random Forests

Random forests are the general term of an ensemble method for classification and regression and are a combination of tree-structured classifiers such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). Random forests can be categorized into two types: one is a classification and regression tree (CART)-like trees type (Breiman *et al.*, 1984) and the other is a binary hierarchy classifier (Ham *et al.*, 2005). The main difference of those two lies in the splitting manner on each node. The splitting manner of the CART-like trees approach is based on variables or features. Meanwhile, in the binary hierarchy classifier, a split on each node is based on classes or labels. In this paper, the random forests approach based on CART-like trees will be only dealt with. Unless stated otherwise, the description and explanation of random forests used in this paper largely follows Breiman (2001). Detailed explanation of CART-like trees type will not be given in this paper and only brief description and main advantage for data fusion will be discussed.

For the purpose of classification of multi-sensor/source data, the random forests can deal with both large data sets and categorical data efficiently. The algorithm is the collection of tree predictors and grows each tree on an independent bootstrap sample from the training data. Suppose there are M variables or sensors for classification. At each node, first, m variables out of all M possible variables are randomly selected. Then the best split on the selected m variables will be found. After that, the tree is grown to maximum depth and a large number of trees vote for the most popular class. Finally, the majority class will be a final output. The way of growing the trees is based on low bias and low correlation in order to improve accuracy. The forests considered consist of using randomly selected inputs or combinations of inputs at each node to grow each tree. By limited use of variables for a split, the computational load can be reduced and as a result large volumes of data sets can be handled (Gislason *et al.*, 2004; Joelsson *et al.*, 2005).

Main advantage of the random forests approach is

that it can provide several analytical results by using out-of-bag samples. As mentioned before, a bootstrap sample for the data is selected for each tree in the forest and then used to grow the tree. The out-of-bag samples mean the remaining samples that are not selected as the bootstrap ones. They can serve as a test set for the tree grown on the bootstrap samples and be used for the estimation of the forest test set error and variable importance. For each tree the out-of-bag samples are put down the corresponding tree and a predicted class that is chosen the most often is obtained. Then, the classification error is computed by comparing the predicted class with the true class. By averaging the classification error over all cases, the overall out-of-bag error or test set error can be obtained (Breiman, 2001).

Especially, variable importance can be computed by using the out-of-bag samples. The computation procedure is as follows: When considering a single tree, first, a tree is used to predict the class of each out-of-bag sample. Then, the values of the variable of interest will be randomly permuted in all the out-of-bag samples and the tree is used to predict the class for these perturbed out-of-bag samples. The variable importance is the increase in the misclassification or error rate between those two steps. It means that by randomly deleting or destroying a certain variable from the whole variables in the out-of-bag samples for each classifier, the increase in the out-of-bag error indicates that the variable deleted is important, since its removal results in the increase of error rate. The importance of a certain variable is averaged over all trees in the forest.

3. Experiments

Two experiments for supervised classification with multi-sensor data sets are carried out to evaluate the applicability of the random forests. The first experiment is done for the fusion of optical and multi-frequency polarimetric SAR data and the second one for the fusion of multi-temporal/polarization SAR data. For the implementation of random forests, a Fortran program that can be freely available from a random forests webpage (http://oz.berkeley.edu/users/breiman/RandomForests/) was used.

1) Fusion of Optical and Polarimetric SAR Data

The multi-sensor data set (grss-dfc-0006) used in the experiment for supervised land-cover classification was provided by the IEEE GRSS Data Fusion Committee (http://www.dfc-grss.org). It includes airborne Thematic Mapper Scanner data with 6 channels and NASA JPL AirSAR data with 9 channels in the C-, L- and P-bands and HH, HV and VV polarizations. Five land-cover classes in an agricultural area are considered: (1) sugar beets, (2) stubble, (3) bare soil, (4) potatoes, (5) carrots. Detailed description of the data sets can be found in Serpico *et al.*(1996) and Park (2004). For the random forests, the number of random splits was set to 4 and the trees in the forest were grown to 50.

The importance of variables is shown in Fig. 1. All importance values were normalized to express the relative ones with respect to the maximum. If the

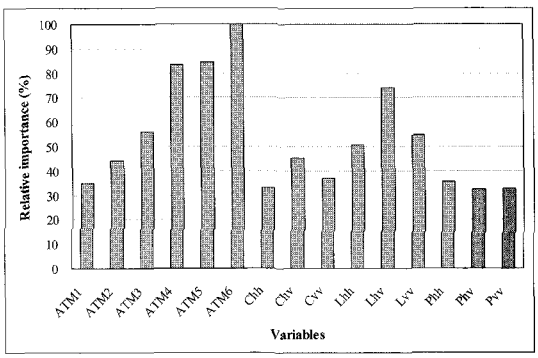


Fig. 1. Results of importance of variable computation in the first experiment.

average increased out-of-bag error value is large, the variable can be considered more important. As expected, infrared bands, ATM4, ATM5 and ATM6, are the most important variables. In addition, L-band data, HV, VV, HH, showed relatively higher values than other C- and P-band SAR data. This outcome resulted in the penetrating depth related to wavelength of the SAR sensor. It means that L-band has the proper amount of penetration power and can reveal better discrimination capability of scattering characteristics between crop classes. Furthermore, the importance of variable for each land-cover class was also computed and is shown in Fig. 2. Like the case of the whole importance values, the contribution of optical data is dominant than those of SAR data. L-band polarimetric data however played a major role in the discrimination of potatoes class. The final classification result is also given in Fig. 3.

Finally, the performance of the random forests approach was compared with previous results

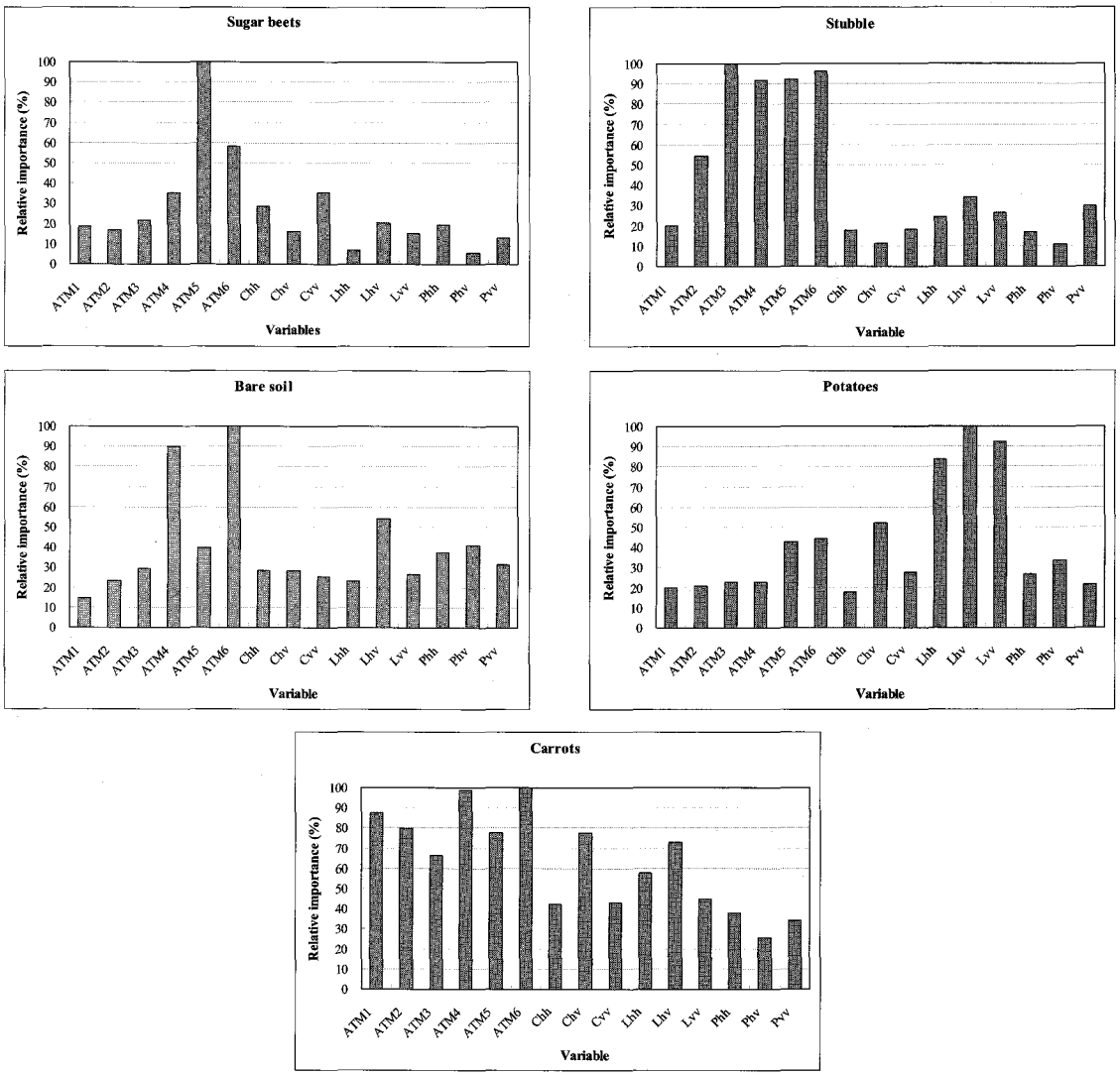


Fig. 2. Results of importance of variable for each class in the first experiment.

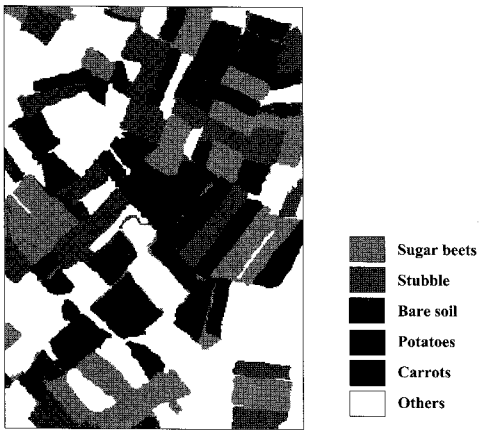


Fig. 3. Classification result of the first experiment.

Table 1. Accuracy statistics of the first experiment. PNNs and *k*-nn results are quoted from Serpico *et al.*(1996).

Class		Random forests	PNN	*k*-nn
User's accuracy	Sugar beets	0.938	0.978	0.974
	Stubble	0.948	0.824	0.884
	Bare soil	0.705	0.796	0.760
	Potatoes	0.869	0.818	0.864
	Carrots	0.971	0.893	0.871
Overall accuracy		0.907	0.886	0.898
Average accuracy		0.886	0.862	0.871
Kappa statistic		0.877	0.850	0.869

obtained by different classification or fusion methods to the same data sets and training/reference samples (Serpico *et al.*, 1996). In Serpico *et al.*(1996), the probabilistic neural network(PNN) and *k*-nn method were applied. For accuracy statistics, overall accuracy, user's accuracy, average accuracy and Kappa statistic were computed by constructing a confusion matrix (Table 1). The improvements for the stubble and carrots classes were significant but the performance for the bare soil class was poorer than PNN and *k*-nn algorithms. Overall, random forests show similar or improved classification performances.

2) Fusion of Multi-temporal Radarsat-1 and ENVISAT ASAR Data

The second experiment was carried out in the Yedang plain, Korea. The data sets considered in the second experiment include multi-temporal C-band Radarsat-1 data (HH polarization) and ENVISAT ASAR data (VV polarization) (Table 2). The Radarsat-1 data acquired from April, 2005 to October, 2005 were used and ENVISAT ASAR data span a whole year from October, 2004 to October, 2005. Dual polarization data (VH and VV polarizations) of ENVISAT ASAR acquired on a single date were especially considered.

Table 2. Multi-temporal SAR data sets used in the second experiment.

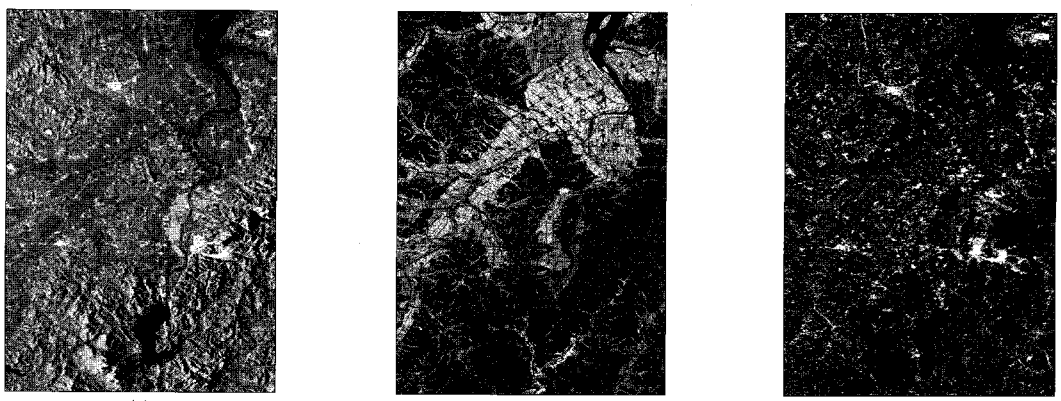
Sensor	Date	Mode	Polarization
Radarsat-1	01.04.2005	Ascending F2	HH
	25.04.2005		
	19.05.2005		
	12.06.2005		
	06.07.2005		
	30.07.2005		
	23.08.2005		
	16.09.2005		
	10.10.2005		
ENVISAT ASAR	31.10.2004	Descending IS2	VV
	09.01.2005		
	13.02.2005		
	20.03.2005		
	24.04.2005		
	29.05.2005		VV & HH
	17.06.2005	Descending IS1	VV & VH
	03.07.2005	Descending IS2	VV
	07.08.2005		
	11.09.2005		
	16.10.2005		

As for the SLC format data preprocessing, coregistration, multi-looking, speckle filtering and geocoding were carried out. Since the Radarsat-1 and ENVISAT ASAR data were acquired from different orbits, they showed the different imaging geometry. To reduce the effect of topography on the backscattering coefficient, geocoding with DEM extracted from a 1:25,000 scale digital topographic map of the study area was done. The lay over and

shadow zones extracted during geocoding with DEM were masked out throughout the data processing for classification. Finally, the study area consists of 660 by 1300 pixels with a pixel size of 25m by 25m. Five land-cover classes such as paddy fields, dry fields, forest, water and urban were considered. For supervised classification, training and reference sets were collected during field survey from July, 2005 to October, 2005. High-resolution optical data acquired in April and June, 2006 were also used for construction of the training and reference sets.

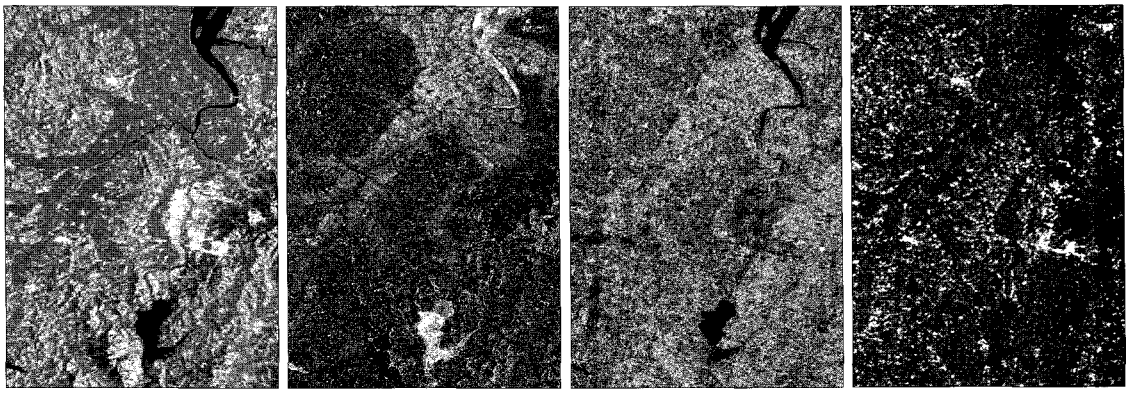
Before classification, first, a feature extraction step was applied by considering the scattering properties of multi-temporal SAR data. This study considered three features including average backscattering coefficient, temporal variability and long-term coherence like previous researches (Strozzi *et al.*, 2000; Park *et al.*, 2005). From the average backscattering coefficient, water and urban areas can be discriminated due to their very low and high backscattering coefficient, respectively. Temporal variability can be used to discriminate cultivated and water areas from forest and urban areas in which the temporal condition is relatively stable and thus the temporal variability is low. It should be noted that the relative intensity of temporal variability depends on the frequency, polarization incidence angles of the SAR sensor considered. In previous researches (Strozzi *et al.*, 2000; Bruzzone *et al.*, 2004), the cultivated areas included paddy and dry fields. Though the paddy and dry fields generally show higher temporal variability than those of other classes, intensity or property of temporal variability between those two cultivated areas may be quite different due to the different cultivation mechanism during the plant growth cycle. In this experiment, the paddy and dry fields were considered as the separated two land-cover classes. The standard deviation values were used as temporal variability factors and the paddy and dry fields were considered as independent two land-cover classes. As a final feature, long-term coherence can be used for the discrimination of urban areas where there are many permanent scatterers from other classes.

The whole 9 Radarsat-1 data were used for the extraction of the average backscattering coefficient and temporal variability (Fig. 4. (a) and (b)). The coherence maps of Radarsat-1 data were also extracted from two interferometric pairs with 24 days intervals and three pairs with 48 days intervals. Finally, those coherence maps were averaged (Fig. 4. (c)). Since ENVISAT ASAR data sets include



(a) (b) (c)

Fig. 4. Features extracted from multi-temporal Radarsat-1 data sets: (a) average backscattering coefficient, (b) temporal variability, (c) coherence.



(a) (b) (c) (d)

Fig. 5. Features extracted from multi-temporal ENVISAT ASAR data sets: (a) average backscattering coefficient, (b) temporal variability, (c) VH/VV ratio, (d) coherence.

different polarization and/or mode data, some data were included or excluded. The average backscattering coefficient and temporal variability were computed from 10 data acquired from descending orbits with IS2 mode (Fig. 5. (a) and (b)). Only VV polarization channel among dual polarization data acquired on May 29, 2005 was used for those features. The ratio of VH and VV channels (VH/VV) acquired on June 17, 2005 was considered as another feature (Fig. 5 (c)). For the ENVISAT ASAR data sets, two interferometric pairs with 35 days intervals were extracted and then averaged (Fig. 5 (d)). All features used in random forests classification are listed in Table 3. The number of random splits was set to 2 and the trees in the forest were grown to 50 like the first experiment.

Fig. 6 shows the relative importance values computed form the out-of-bag sample errors. The final classification result is also shown in Fig. 7. The temporal variability and average backscattering coefficient of Radarsat-1 data sets showed the highest importance values. Those features generally had discrimination capabilities of paddy and water classes from other ones. As shown in Fig. 7, the paddy class occupied the large portion of the study area and thus the contribution of the temporal variability and

Table 3. Features extracted for the second experiment.

Feature	Abbreviation
Average backscattering coefficient of Radarsat-1	R_avg
Temporal variability of Radarsat-1	R_temp
Coherence of Radarsat-1	R_coh
Average backscatteing coefficient of ENVISAT ASAR	E_avg
Temporal variability of ENVISAT ASAR	E_temp
Coherence of ENVISAT ASAR	E_coh
VH/VV ratio of ENVISAT ASAR	E_ratio

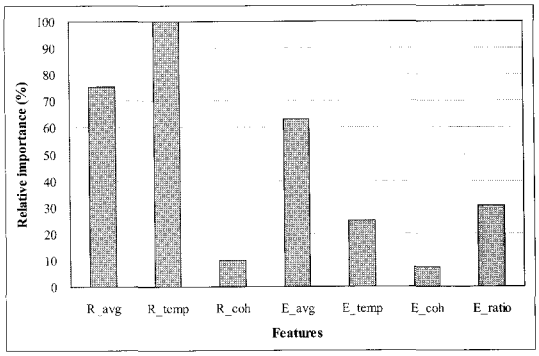


Fig. 6. Relative importance values of all features used in the second experiment.

average backscattering coefficient to discrimination of the paddy class resulted in the highest importance values. The average backscattering coefficient of ENVISAT ASAR data sets is the next. The superiority of the features from Radarsat-1 data to those from ENVISAT ASAR data however does not

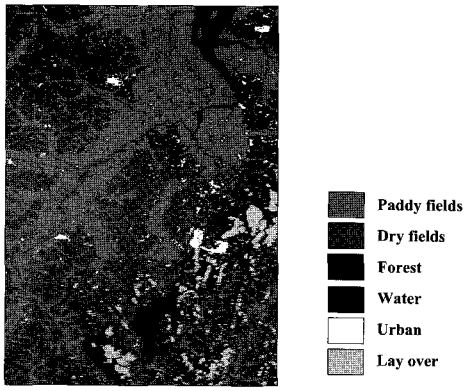


Fig. 7. Classification result of the second experiment.

explain that HH polarization data always give us more information than the VV polarization data in an agricultural area. It should be noted that the results are the combined results of various parameters such as acquisition dates, incidence angle, mode, polarization state, etc. The coherence values of both Radarsat-1 and ENVISAT ASAR data sets have little contributions to discrimination of the land-cover classes in the study area. It is worth noting that the ratio between VH and VV polarization, which is a feature to characterize the level of polarization dependence, showed higher relative importance values than those from the coherence values. In general, the VH/VV ratio can give different information on agricultural fields and forests. That is, Agricultural fields show an intermediate VH/VV ratio and forests the highest VH/VV ratio of the natural targets (Wegmüller *et al.*, 2003). Though it was extracted from one data acquired in July, the VH/VV ratio could affect the discrimination between dry fields and forests that showed mixed characteristics in the average backscattering coefficient. This experimental results confirm the effectiveness of the multi-polarization capability of ENVISAT ASAR data.

The classification accuracy results are given in Table 4. The improvement of classification accuracy

Table 4. Accuracy statistics of the second experiment.

Class		Random forests
User's accuracy	Paddy fields	0.959
	Dry fields	0.804
	Forest	0.869
	Water	0.999
	Urban	0.957
Overall accuracy		0.943
Average accuracy		0.918
Kappa statistic		0.909

in the dry fields is sustainable, compared with Park *et al.*(2005). The possible explanation is that the use of multi-polarization data could give more information on the discrimination of dry fields from other classes.

4. Conclusions

The random forests classifier for both the classification of multi-sensor data and accounting for the importance of variable has been applied in this paper. The distinguished feature of the random forests approach is its ability to estimate or compute the importance of variable by using out-of-bag samples. In two experiments, the random forests approach could estimate which variable or feature played a major role in discriminating the land-cover classes considered. Also, it indicated a good classification accuracy comparable to other non-parametric data fusion algorithms. It is expected that the feature selection based on the information on the importance of variable in the random forests approach would be effectively incorporated into hyperspectral data classification.

Acknowledgment

The provision of the multi-sensor data set (grss-

dfc-0006) by IEEE GRSS Data Fusion Technical Committee is gratefully acknowledged. Three data among the ENVISAT ASAR data sets used in this study were kindly provided by the Korea Aerospace Research Institute (KARI). We also thank Professor Hoonyol Lee of Kangwon National University for his insightful comments on coherence imagery generation. This work was supported by the public application research of satellite data.

References

Breiman, L., 2001. Random forests, *Machine Learning*, 45(1): 5-32.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone, 1984. *Classification and regression trees*, Wadsworth Inc.

Bruzzone, L., M. Marconcini, U. Wegmuller, and A. Wiesmann, 2004. An advanced system for the automatic classification of multitemporal SAR images, *IEEE Transactions on Geoscience and Remote Sensing*, 42(6): 1321-1334.

Gislason, P. O., J. A. Benediktsson, and J. R. Sveinsson, 2004. Random forest classification of multisource remote sensing and geographic data, *Proceedings of IGARSS 2004*.

Ham, J., Y. Chen, M. M. Crawford, and J. Ghosh, 2005. Investigation of the random forest framework for classification of hyperspectral data, *IEEE Transactions on Geoscience and Remote Sensing*, 43(3): 492-501.

Joelsson, S. R., J. A. Benediktsson, and J. R. Sveinsson, 2005. Random forest classifiers for hyperspectral data, *Proceedings of IGARSS 2005*.

Le Hégarat-Mascle, S., I. Bloch, and D. Vidal-Madjar, 1997. Application of Dempster-Shafer evidence theory to unsupervised classification in multisource remote sensing, *IEEE Transactions on Geoscience and Remote Sensing*, 35(4): 1018-1031.

Park, N.-W., 2004. *Multi-source spatial data fusion with geostatistical uncertainty assessment: applications to landslide susceptibility analysis and land-cover classification*, Ph. D. Thesis, Seoul National University.

Park, N.-W., H. Lee, and K.-H. Chi, 2005. Feature extraction and fusion for land-cover discrimination with multi-temporal SAR data, *Korean Journal of Remote Sensing*, 21(2): 145-162.

Serpico, S. B., L. Bruzzone, and F. Roli, 1996. An experimental comparison of neural and statistical non-parametric algorithms for supervised classification of remote-sensing images, *Pattern Recognition Letters*, 17(13): 1331-1341.

Solberg, A. H. S., T. Taxt, and A. K. Jain, 1996. A Markov random field model for classification of multisource satellite imagery, *IEEE Transactions on Geoscience and Remote Sensing*, 34(1): 100-113.

Solaiman, B., L. E. Pierce, and F. T. Ulaby, 1999. Multisensor data fusion using fuzzy concepts: application to land-cover classification using ERS-1/JERS-1 SAR composites, *IEEE Transactions on Geoscience and Remote Sensing*, 37(3): 1316-1326.

Strozzi, T., P. B. G. Dammert, U. Wegmüller, J. M. Martinez, A. Beaudoin, J. Askne, and M. Hallikainen, 2000. Landuse mapping with ERS SAR interferometry, *IEEE Transactions on Geoscience and Remote Sensing*, 38(2): 766-775.

Wegmüller, U., T. Strozzi, A. Wiesmann, and C. Werner, 2003. ENVISAT ASAR for land cover information, *Proceedings of IGARSS 2003*.