Enhancement of Land Cover Classification by Training Samples Clustering

Artem Andreiev
Dept. of Geoinformation Technologies in Remote Sensing of the Earth Scientific Centre for Aerospace Research of the Earth
Institute of Geological Sciences
National Academy of Sciences of Ukraine
Kyiv, Ukraine
a.a.andreev@casre.kiev.ua
ORCID 0000-0002-6485-449X

Anna Kozlova
Dept. of Geoinformation Technologies in Remote Sensing of the Earth Scientific Centre for Aerospace Research of the Earth
Institute of Geological Sciences
National Academy of Sciences of Ukraine
Kyiv, Ukraine
ak@casre.kiev.ua
ORCID 0000-0001-5336-237X

Abstract. **In this study, a hybrid approach is proposed to enhance land cover classification accuracy by clustering training samples into homogenous subclasses. The proposed approach implies the integration of both supervised and unsupervised classification methods into a holistic framework. A criterion of training sample separability is developed as separability index of training samples. The approach was applied to enhance the land cover classification of the highly heterogeneous natural landscapes by the case of the Shatsky National Natural Park.**

Keywords: **land cover classification, clustering, hybrid approach, training samples separability**

I. INTRODUCTION

Land cover classification is a key research field in remote sensing, which is still challenging in heterogeneous landscapes [1]. The problem mainly arises from the mixing of land cover classes. Nowadays, the solution of the problem is seen in the application of hybrid classification models based on combining both supervised and unsupervised learning [2].

As known, the classification process implies that expert selects the training samples of each land cover class. Hence, it is necessary to obtain a description of each class. However, due to the human factor, selected training samples tend to be inaccurate as well as presented classes are subjective, which, in turn, decreases classification accuracy.

The study aims to enhance land cover classification accuracy by clustering training samples into homogenous subclasses. For this, an approach to land cover classification is developed as a hybrid of supervised and unsupervised methods. The potential of the proposed approach is explored by mapping the highly heterogeneous natural landscapes of the Shatsky National Natural Park.

II. METHODOLOGY

A. Hybrid approach to classification

The hybrid approach to classification is developed to reduce the impact of problems caused by the high heterogeneity of land cover classes [3]. The core of that approach is the integration of both supervised and unsupervised classification methods into a holistic framework. This conception aims to lessen the subjectiveness of expert-selected classes and the mixing of training samples. This point is reached by clustering of classes training samples with the unsupervised methods. Another point is to provide a reasonable interpretation of classes, which is inherent in supervised methods.

Input data of the proposed approach to classification consists of an image and training samples of each class. The training samples should satisfy such requirements as completeness, sufficiency, and purity [4]. The algorithm of the proposed approach is described in Fig. 1.

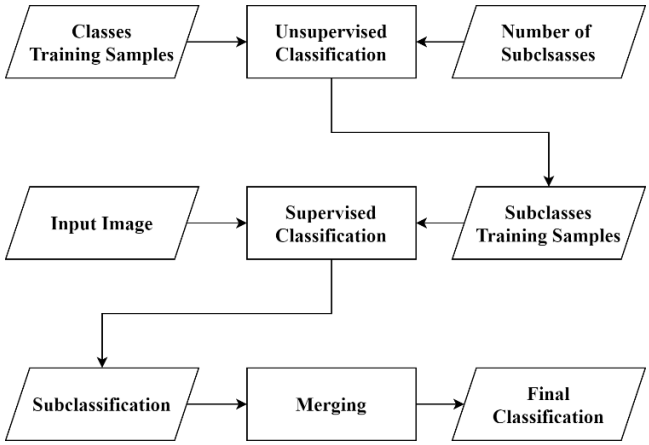

Fig. 1. The scheme of hybrid approach to classification

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

The first step of this algorithm implies training samples clustering via unsupervised classification. Firstly, initial classes are subdivided into subclasses. Secondly, classes training samples are clustered into subclasses training samples in the form of subclusters. The method of unsupervised classification could be specified individually for each class since the clustering is performed for each class separately.

The second step engages the supervised classification. The subclusters obtained in the previous step are used as subclasses training samples. Whereas input image is divided into subclasses, but not to classes, the result of this procedure is named subclassification instead of classification. The method of supervised classification should be assigned taking into account the features of an input image and training samples.

The final step is to merge the subclasses of subclassification into initial classes. Since this procedure is pixelwise, each pixel of subclassification requires identifying the initial class of its subclass. This step is necessary to transform subclassification to classification.

B. Number of subclasses

According to the algorithm of the hybrid approach to classification, the parameter "Number of subclasses" is not defined a priori. This parameter sets the number of subclasses for each initial class. The most appropriate value of this parameter is the one that maximizes the separability of training samples, thereby minimizing their mixing.

Fig. 2 describes the algorithm of selection of the most appropriate number of subclasses. Presented algorithm iterates over combination sets, which contains a number of subclasses for each class. Such a combination set is illustrated in Fig. 2 as "Number of subclasses".

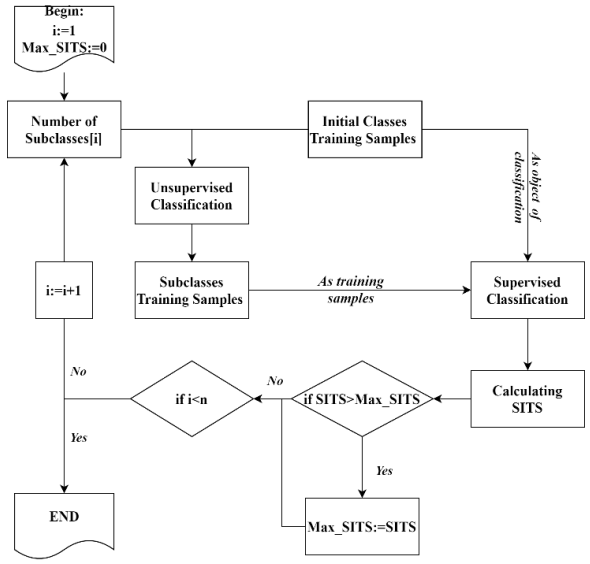

Fig. 2. The algorithm of selection of the most appropriate number of subclasses

In order to limit the iterative process, the maximum number of subclasses should be assigned for each class. An expert makes this decision taking into account the available computational power or some features of training samples, namely their size, density, etc. Thereafter, the number of iterations is calculated by following combinatorial formula:

$$n = \prod_{i=1}^{K} sub_i \quad , \tag{1}$$

where K is the number of initial classes, sub_i is the maximum assigned number of subclasses for i class.

The first step of this algorithm also involves the training samples clustering, as the algorithm of the hybrid approach to classification. However, it is worth noting that the number of subclasses varies with each iteration.

The second step is to perform supervised classification for initial classes training samples and their subclasses training samples as an object of classification and training samples, respectively. The method of supervised classification cannot differ from the one chosen for the hybrid approach to classification.

Estimation of training sample separability is engaged at the third step. In order to carry out this task, the separability index of training samples (SITS) is proposed. SITS quantifies the separability of training samples by measuring the ratio of the number of correctly classified training samples to the total number of training samples. Calculation of SITS is similar to the standard calculation of classification overall accuracy [5]:

$$SITS = \frac{\sum_{i=1}^{K} TS_{\mathrm{corr}\,i}}{\sum_{j=1}^{K} TS_{\mathrm{total}\,j}} \quad , \tag{2}$$

where K is the number of initial classes, TS_{corri} is the number of correctly classified training samples of i class, TS_{totali} is the total number of training samples of i class.

In order to calculate SITS, the number of correctly classified training samples is provided by classification obtained in the second step. Since training samples and their total number are specified a priori as input data, the calculation of SITS could be performed automatically, unlike the calculation of classification overall accuracy.

After all n iterations, the highest value of SITS will be determined. This value refers to the most appropriate combination set among others. Thus, returning to the hybrid approach to classification, this set assigns the number of subclasses for each class.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

III. EXAMPLE

A. Study Area

The proposed hybrid approach was tested at the area of the Shatsk National Natural Park. It is situated in northwest Ukraine, within Volyn' oblast, between 51º 28'25"N and 23º 49'29"E (Fig. 3). Lying in the vast wetland region of West Polissia, the Park encompasses diverse forests, peat bogs, transitional mires, meadows, and lakes. As a study area, the Park was chosen due to the high heterogeneity of its natural landscapes. During land cover classifications, it often results in the subjectiveness of expert-selected classes and the mixing of training samples.

Since 2007, the Park belongs to the Ukrainian network of the test sites for satellite-based products validation [6]. Above 100 georeferenced sample plots were set here to provide comprehensive ground truth information about the representative landscapes of the West Polissia region.

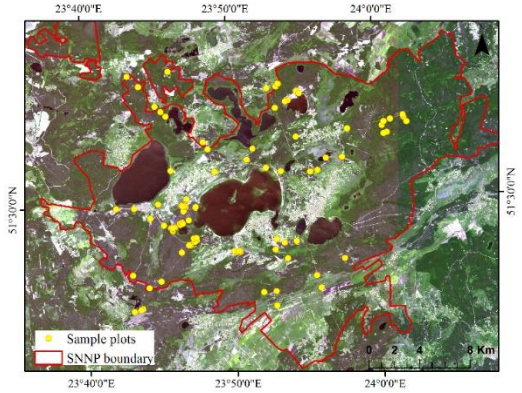

Fig. 3. Location of the study area and sample plots within the Shatsk National Natural Park. They are shown on the fragment of the true-colored composite of the Sentinel-2 Multispectral Instrument (MSI) image acquired on 1 June 2018

B. Input Image

A cloud-free Sentinel-2A multispectral instrument (MSI) image acquired on 01 June 2018 was downloaded from the U.S. Geological Survey (USGS) archive (https://earthexplorer.usgs.gov). The image was obtained at the top of the atmosphere reflectance (TOA, Level 1C) and then atmospherically corrected to the bottom of the atmosphere reflectance (BOA, Level 2A) using the Sen2Cor tool (https://step.esa.int/main/snap-supported-plugins/).

During the processing, Sen2Cor discarded the three bands (B1, B9, and B10) that consider the effects of aerosols and water vapour on reflectance. Then, the Sentinel-2 bands acquired at 20 m data were resampled using the nearest neighbour method to obtain a layer stack of 10 spectral bands at 10 m. Finally, the obtained image was resized to the extent of the study area and account for 2284x1554 pixels.

C. Training samples

Six broad land cover classes were the focus, as follows: artificial surfaces, forest, natural grassland, agricultural areas, water bodies, and inland wetlands.

As it was mentioned, completeness, sufficiency, and purity are the key requirements for training samples. An extensive analysis of representative features of each class all over the study area provided the satisfaction of the requirements. Updated information from the georeferenced sample plots has also contributed to initial training sample completeness and purity.

The given classes varied considerably both in spatial extent and heterogeneity. The relatively small class included diverse features (e.g. agricultural areas) while the bigger one could be quite homogenous (e.g. natural grassland). Therefore, the number of training pixels of each class also varied disproportionally. The overall number of all training pixels accounted for 2684. Table I shows labels, description, and training pixel amount for the land cover classes assigned for the experiment.

TABLE I. THE CLASSIFICATION SCHEME USED IN THE EXPERIMENT

Land Cover Class	Description	Training pixels
Artificial surfaces	Urban public and industrial built-up areas, transport units, and construction sites	370
Forest	Broadleaved, coniferous, and mixed forests, roadside tree lines, areas with tree cover more than 30%	611
Natural grassland	Natural herbaceous vegetation, permanent grasslands of natural origin, pastures	544
Agricultural areas	Arable land, permanent crops, fallow lands, heterogeneous agricultural areas, open soils	313
Water bodies	Lakes, rivers and streams of natural origin, including man-made reservoirs and canals.	403
Inland wetlands	Non-forested areas of peat bogs, transitional mires, eutrophic marshes, and reed beds	438

D. Classifications

The initial set of training samples were used to obtain a land cover map (Fig. 4a) applying Mahalanobis distance as a method of supervised classification [7]. To estimate the separability of this set, the value of its SITS was calculated by the formula (2):

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

$$\text{SITS}_{\text{initial}} = \frac{TS_{\text{corr}_1} + TS_{\text{corr}_2} + TS_{\text{corr}_3} + TS_{\text{corr}_4} + TS_{\text{corr}_5} + TS_{\text{corr}_6}}{TS_{\text{total}_1} + TS_{\text{total}_2} + TS_{\text{total}_3} + TS_{\text{total}_4} + TS_{\text{total}_5} + TS_{\text{total}_6}} = \frac{339 + 525 + 538 + 307 + 403 + 356}{370 + 611 + 544 + 313 + 403 + 438} = 0.92$$

Another land cover map was obtained using the hybrid approach, which provides clustering of the initial training samples. The application of this approach involves a definition of subclasses number for each initial class of the training samples. According to the proposed algorithm, an expert should set the maximum number of subclasses for each initial class. Taking into account the size of the initial set of training samples, the maximum number of subclasses for each class was set to 10. Under the formula (1), there are 10^6 different combinations of training samples, derived by their clustering.

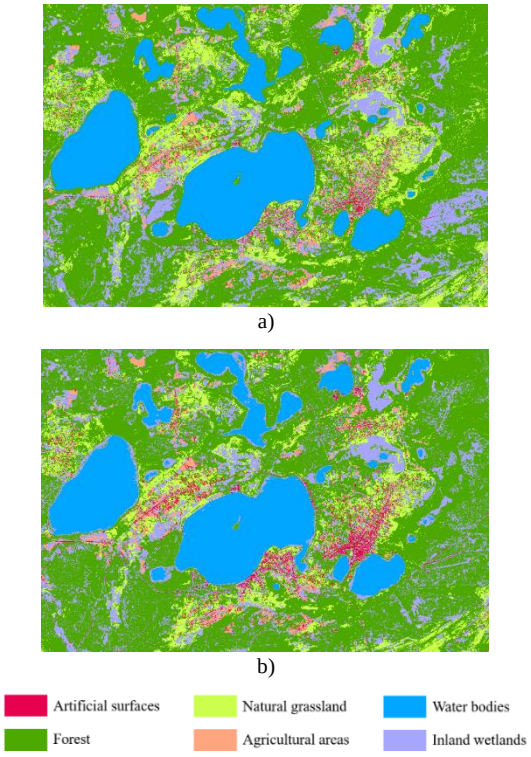

Fig. 4. Land cover maps of the study area were obtained using (a) initial training samples and (b) training samples after clustering

Among all iterations of the K-Means clustering [8], the maximum value of SITS is achieved by subdividing initial classes into the combination of subclasses, which is shown in Table II.

TABLE II. THE DETERMINED NUMBER OF SUBCLASSES FOR EACH INITIAL CLASS

Land Cover Class	1	2	3	4	5	6
Number of subclasses	10	3	1	4	4	6

The SITS value of obtained set of training samples is calculated herein (2):

$$\text{SITS}_{\text{final}} = \frac{TS_{\text{corr}_1} + TS_{\text{corr}_2} + TS_{\text{corr}_3} + TS_{\text{corr}_4} + TS_{\text{corr}_5} + TS_{\text{corr}_6}}{TS_{\text{total}_1} + TS_{\text{total}_2} + TS_{\text{total}_3} + TS_{\text{total}_4} + TS_{\text{total}_5} + TS_{\text{total}_6}} = \frac{363 + 609 + 542 + 308 + 403 + 428}{370 + 611 + 544 + 313 + 403 + 438} = 0.99$$

After that, the final classification (Fig. 4b) is carried out using the determined set of training samples.

E. Accuracy Assessment

The initial and final land cover maps were verified independently from each other using proportionate stratified random samplings. Such sampling technique produces sample set sizes that are directly related to the size of the classes. It is widely used in assessing the classification accuracy of heterogeneous landscapes. To determine the required total sample size, the minimum sample size was set to 0.01% of the total number of the input image pixels. Therefore, validation samples were equal to 355 pixels for each land cover map.

As a primary source of reference data, high spatial resolution satellite images (QuickBird) available in Google Earth TM for 2018 were used for verification.

Confusion matrices were constructed to assess overall accuracy (OA), producer's accuracy (PA), and user's accuracy (UA) of the land cover maps.

Table III shows the confusion matrix of the initial land cover map. Its overall accuracy was 77%. Both producer's and user's accuracy were very low for the classes #1 artificial surfaces (PA – 22%, UA – 50%) and #6 'inland wetlands' (PA – 52%, UA – 29%). User's accuracy was also low for class #3 'natural grassland' (63%), while producer's accuracy was very low for class # 4 'agricultural areas' (40%). Only classes #2 'forest' and # 5 'water bodies' had high both producer's and user's accuracy.

TABLE III. CONFUSION MATRIX OF THE INITIAL LAND COVER MAP

Class#		**Actual Class**							
		1	***2***	***3***	***4***	***5***	***6***	***Σ***	***UA,%***
Predicted Class	***1***	2	0	0	2	0	0	4	50
	2	3	160	3	6	0	7	179	89
	3	2	8	30	2	0	6	48	63
	4	1	0	0	10	0	0	11	91
	5	0	0	0	0	56	2	58	97
	6	1	24	9	5	0	16	55	29
	Σ	9	192	42	25	56	31	**355**	
	PA, %	22	83	71	40	100	52		***OA, %*** **77**

Table IV shows the confusion matrix of the final land cover map. Its overall accuracy was 81%. Only class #4 'agricultural areas' had low producer's accuracy (31%). However, at the same time its user's

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

accuracy was the highest (100%). For classes #1 ‘artificial surfaces’ and #3 ‘natural grassland’ both producer’s and user’s accuracy were equal (67%) or almost equal (PA – 62%, UA – 66%), but still not high. All other classes had high both producer’s and user’s accuracy.

TABLE IV. CONFUSION MATRIX OF THE FINAL LAND COVER MAP

Class#		Actual Class							
		1	***2***	***3***	***4***	***5***	***6***	***Σ***	***UA,%***
Predicted Class	***1***	6	0	0	3	0	0	9	67
	2	1	168	6	6	0	2	183	92
	3	0	6	23	6	0	0	35	66
	4	0	0	0	11	0	0	11	100
	5	0	0	0	0	55	1	56	98
	6	2	18	8	9	0	24	61	39
	Σ	9	192	37	35	55	27	**355**	
	PA, %	67	88	62	31	100	89		***OA, %*** **81**

IV. DISCUSSION AND CONCLUSION

The experiment has revealed, that the land cover classification of the study area was enhanced by application of developed approach. This is evidenced by a 4% increase in overall accuracy from 77% to 81%.

The most significant enhancement was appeared in two classes, namely #1 ‘artificial surfaces’ and #6 ‘inland wetlands’ (Fig. 5). PA values of both #1 and #6 classes were increased by 3 and 1.7 times, respectively. This indicates that those predicted classes became more referenced to actual ones. Meanwhile, UA values of both #1 and #6 classes were increased by 1.3 each. This points that those predicted classes became less misclassified.

a) b)

Fig. 5. Fragments of initial (a) and final (b) land cover maps illustrating enhancement of artificial surfaces and inland wetlands classification

Described enhancement is reflected by increase of SITS value after training samples clustering. It is significant, that among with increasement of PA and UA values of mentioned classes, number of their correctly classified training samples (TS_{cor1} and TS_{cor6}) increased from 339 and 356 to 363 and 428, respectively.

Thus, the developed approach enhances land cover classification accuracy of heterogenous landscapes by clustering training samples into homogenous subclasses.

Further research should be aimed at approbation of the developed framework with application of other supervised and unsupervised methods. Also, this approach could be extended by additional criteria of training samples separability.

REFERENCES

[1] J. Paneque-Gálvez, J. Mas, G. Moré, J. Cristóbal, M. Orta-Martínez, A. Luz, M. Guèze, M. Macía, V. Reyes-García, “Enhanced land use/cover classification of heterogeneous tropical landscapes using support vector machines and textural homogeneity.” Intern. J. Appl. Earth Obs. Geoinformation, vol.23, pp. 372-383, Aug. 2013.

[2] J. Xiao, Yuhang Tian, X. Ling Xie, Jiang and Jing Huang. “A Hybrid Classification Framework Based on Clustering.” IEEE Transactions on Industrial Informatics, vol. 16, pp. 2177–2188. Jan. 2020.

[3] A.A. Andreiev, “Hybrid approach to classification of remote sensing data.” CERes Journal, vol. 6, issue 2, pp. 32–37. Dec. 2020.

[4] W.G. Cochran, Sampling Techniques. New York: John Wiley & Sons, 1977.

[5] M.O. Popov, “Methodology of accuracy assessment of classification of objects on space images”, J. Autom. Inf. Sci., vol. 39, pp. 1-10. 2007. (In Russian).

[6] V.I. Lyalko, M.A. Popov, S.A. Stankevich, J.I. Zelyk, S.V. Cherny, “Calibration/Validation Test Sites in Ukraine: current state and directions of further research and development.” Ukrainian Metrological Journal, vol. 2, pp. 15-26. 2014. (In Russian)

[7] L. Bruzzone, B.A. Demir, “A review of modern approaches to classification of remote sensing data”, in Land use and land cover mapping in Europe, I. Manakos, M. Braun, Eds. Springer: Dordrecht, Netherlands, pp. 127–143. 2014.

[8] A.K. Jain, R.C. Dubes, “Algorithms for Clustering Data.” Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.