The subject of this article is land cover classification based on geospatial data. The supervised classification methods are appropriate for most of the thematic tasks of remote sensing because they provide the opportunity to set the characteristics of the initial classes in the form of a training sample set, in contrast to unsupervised methods. There are many approaches to processing such a set; however, their common disadvantage is that they do not consider the factor of training sample separability. This characteristic indicates the extent to which signatures representing different classes do not overlap. A low degree of separability is inherent in high-level training sample mixing. Thus, separability affects classification accuracy. One possible ways to increase separability is training sample clustering. Considering the above, the goal of this study is to develop a training sample clustering technique to improve land cover classification accuracy by increasing the separability of training samples. The tasks of this work are as follows: 1) develop a method for training sample separability assessment; 2) develop a training sample clustering technique based on training sample separability; 3) test the effectiveness of the developed technique by applying it to experimental land cover classification. In the experiments, two land cover classifications were obtained for each of the two selected study areas (i.e., one before and another after training sample clustering. Six land cover classes were defined for each experiment. The training samples were selected for each class. Conclusions. After the application of the developed technique, an increase in the separability of the training samples was evidenced by the developed separability index. In turn, this approach led to an improvement in land cover classification. For the first experiment, this was evidenced by an increase in the overall accuracy and kappa coefficient by 20% (from 63 to 83%) and 21% (from 60% to 81%), respectively. In the second experiment, the increase was 4% (from 77% to 81%) and 5% (from 66% to 71%), respectively.

Keywords: classification; supervised classification; unsupervised classification; clustering; remote sensing; training sample; training sample separability.

1. Introduction

1.1. Motivation

Land cover classification has a broad range of applications in remote sensing [1]. It provides spatially explicit categorized information for environmental monitoring [2, 3], land cover change detection [4], analysis of urban development [5, 6], landmine detection [7], and fossil fuel exploration [8]. Moreover, land cover classification techniques play a crucial role in complex interdisciplinary problems in achieving sustainable development goals [9, 10], primarily combating climate change and its impacts [11], reversing land degradation [12], halting biodiversity loss [13], protecting water-related ecosystems for safe water supply [14], and conducting geoenvironmental hazard assessments [15].

1.2. State-of-the-art

Currently, many classification techniques have been developed, and they are mainly divided into two groups: supervised and unsupervised [16]. Supervised classification and, however, is most suitable for a large number of thematic tasks because it allows setting the characteristics of the original classes, unlike unsupervised classification. Such characteristics are provided by the training sample set.

The training sample set comprises each class sample. In turn, such a sample is presented in the form of corresponding signatures defined in each layer of the input geospatial data. Among the approaches to training sample preprocessing, the following can be distinguished: cluster sampling [17], approaches to reduce the size of the input data [18], noisy image processing [19, 20], and approaches that define mislabeled training data [21]. Along with the above approaches, we highlight the ones that aim to change the data structure, namely, image contour segmentation [22], synthesis of neural network structure [23], ranking and selection of different sampling strategies [24] and iterative clustering for training sample refinement [25].

However, most existing approaches do not consider the factor of training sample separability, which
Intelligent information technologies

67

affects classification accuracy. This characteristic indicates the extent to which signatures representing different classes do not overlap. A low degree of separability is inherent in high-level training sample mixing. In turn, this leads to a significant number of misclassified objects in the classification. Thus, the training sample separability directly affects the classification accuracy [26].

1.3. Objective and Approach

The aim of the present study is to improve land cover classification accuracy. This can be achieved by increasing the separability of training samples. Thus, a method for training sample separability assessment was developed. This method is the basis of the proposed technique, which implies that training sample separability increases via training sample clustering.

Considering the above, the paper structure consists of the following sections.

In the section "Methods", we describe the method of training sample separability assessment. This method is the basis of the proposed training sample clustering technique, which is also presented in this section.

The section "Experiments" implies performing two land cover classifications before and after applying the developed technique to each of the two selected study areas.

Finally, the section "Conclusions" briefly describes the developed methods, obtained experiment results and further research aims.

2. Methods

Training sample separability assessment. Because this technique is the optimization based on training sample separability, the object function is the developed separability index of the training sample (SITS).

SITS is the result of the training sample separability assessment, and its stepwise algorithm is described in Figure 1.

The first step involves classifier training on the input training sample set. Notably, the supervised classification method must be the same as that selected for further land-cover classification. This is because the separability depends not only on the training sample structure but also on the supervised classification method.

In the second step, each signature in the training sample set is classified by the obtained classifier.

The third step is the formation of a confusion matrix [27] for the classification obtained in the previous step.

The fourth and final step is the SITS calculation. This index can be calculated for the entire training sample and two separate classes.

The SITS of the two classes (SITS pair) is the average arithmetic value of the sensitivity and specificity indicators [28]. Sensitivity was calculated using the following formula:

\[
\text{sensitivity} = \frac{x_{ji}}{x_{ii} + x_{ji}},
\]

where \(x_{ji}\) is the number of class \(j\) signatures classified as class \(i\).

The following formula corresponds to the specificity calculation:

\[
\text{specificity} = \frac{x_{ij}}{x_{jj} + x_{ij}}.
\]

Fig. 1. The algorithm of the training sample separability assessment
Thus, the SITS\textsubscript{pair} calculation formula has the following formula:

\[
\text{SITS}_{\text{pair}} = \frac{\text{sensitivity} + \text{specificity}}{2}.
\]  

(1)

SITS of the entire training sample set (SITS\textsubscript{overall}) was calculated using the following formula:

\[
\text{SITS}_{\text{overall}} = \sum_{i=1}^{K} \frac{\sum_{j=1}^{K} x_{ij}}{\sum_{i=1}^{K} \sum_{j=1}^{K} x_{ij}},
\]  

(2)

where \(K\) is the number of classes.

This index quantifies the separability of the training sample set by measuring the ratio of correctly classified signatures to the total number of signatures. In other words, the SITS\textsubscript{overall} equals the overall accuracy (OA) [27] based on the confusion matrix obtained in the previous step.

The values of the considered indices range from 0 to 1. At the same time, value 0 indicates that the training sample is entirely mixed (minimum separability), and value 1 corresponds to the training sample, which is entirely separable (maximum separability).

\textbf{Training sample clustering technique.} The developed technique assumes that only centroid methods of unsupervised classification (i.e. K-means, K-medians, along with others) [29] are considered. Thus, the goal of the proposed technique can be defined as finding the optimal number of clusters for each class of the training sample. The optimal number of clusters is one that provides the corresponding clustered set of the training sample with the highest value of the SITS\textsubscript{overall} among all other options. This index is calculated using equation (2).

The algorithm of the developed technique is illustrated in a flowchart (Figure 2).

This algorithm is an iterative procedure. In turn, each iteration contains two steps.

The first step of the iteration is to calculate SITS\textsubscript{pair} for each pair of classes of the training sample using the formula (1). Then, the pair with the lowest value of this indicator is selected.

The second step involves finding the optimal number of clusters for pairs of classes selected in the previous step. For this reason, the number of clusters of these two classes increases from 1 to that number, at which the SITS\textsubscript{overall} value increase stops.

The iterations are then repeated without considering the pair selected for each iteration.

Such two-step iterations continue until at least one of the following stopping criteria is met:

1) if the SITS\textsubscript{overall} value equals 1;
2) searching for the optimal number of clusters for all consecutive pairs of classes indicates no increase in the SITS\textsubscript{overall}.

This procedure results in the optimal structure of the training sample with the highest separability among all considered options. The obtained training sample set was used for further classification.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2.png}
\caption{The algorithm of the training the sample clustering technique}
\end{figure}
3. Experiments

Two experiments were conducted to test the effectiveness of the developed technique. Each stage consisted of performing two land cover classifications before and after applying the obtained technique. The first experiment’s study area was the Ivano-Frankivsk region’s test site (Figure 3, b), and the second experiment’s study area was Shatsk National Natural Park (Figure 3, e).

Six broad land cover classes were defined for the first experiment: artificial surfaces, crops, grasslands, tree-covered areas, water bodies, and bare rocks. The following six classes were selected for the second experiment: artificial surfaces, crops, grasslands, tree-covered areas, water bodies, and wetlands. The training samples were selected for each of the aforementioned classes.

Input data. The data set for the first experiment included seven bands of Landsat-OLI8 satellite image (acquired on August 9, 2018) and three spectral indices (namely Normalized Difference Vegetation Index (NDVI), Normalized Difference Build-up Index (NDBI) and Build-Up Index (BUI) [30]). The second experiment data set contained ten spectral bands of the Sentinel-2 satellite image acquired on June 1, 2018.

Technique application. The maximum likelihood and Mahalanobis distance [31] were selected for the first and second experiments, respectively. In addition, K-means [29] was selected as an unsupervised classification method for training sample clustering in both experiments.

First, the initial training sample separability was assessed in each experiment. The SITSoverall values equal 0.91 and 0.92, respectively.

Second, the training sample was clustered. The optimal number of clusters for each input class was determined for the first experiment to be 2, 5, 4, 2, 1, 2; for the second experiment to be 10, 3, 1, 4, 4, and 6. The SITSoverall values of the obtained training samples were 0.95 and 0.99, respectively.

Finally, the corresponding classifications were performed. The initial and final classification maps for the first experiment are shown in Figures 3, a and 3, c, respectively. The classification maps for the second experiment are shown in Figures 3, d and 3, f, respectively.

Accuracy assessment. The assessment of classification accuracy involved independent verification of the initial and final land cover maps. For this purpose, the test sample set for the first experiment contained 60 pixels for each land cover map and 355 pixels for the second experiment. Satellite images (QuickBird) of high spatial resolution were used as reference data.

Metrics such as OA and kappa coefficient were selected for the classification accuracy assessment [25]. The accuracy assessment results and the SITS values are listed in Table 1.

<table>
<thead>
<tr>
<th>#</th>
<th>SITS initial</th>
<th>SITS final</th>
<th>OA initial</th>
<th>OA final</th>
<th>Kappa initial</th>
<th>Kappa final</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.91</td>
<td>0.94</td>
<td>63</td>
<td>83</td>
<td>60</td>
<td>81</td>
</tr>
<tr>
<td>2</td>
<td>0.92</td>
<td>0.99</td>
<td>77</td>
<td>81</td>
<td>66</td>
<td>71</td>
</tr>
</tbody>
</table>

Fig. 3. First experiment: initial classification (a), Landsat-OLI8 image of the study area (b), final classification (c); Second experiment: initial classification (d), Sentinel-2 image of the study area (e) and final classification (f)
4. Discussion

The aim of the developed optimization technique is to improve land cover classification accuracy. This goal is proposed to be achieved by increasing the separability of the training sample. Thus, the experimental result should be considered in terms of classification accuracy and training sample separability.

Two experiments confirmed the improved land cover classification accuracy after applying the proposed technique along with increased training sample separability. The SITS increased by 3% (from 91% to 94%) in the first experiment and by 7% (from 92% to 99%) in the second experiment. The OA and kappa coefficient values. For the first experiment, by 20% (from 63 to 83%) and 21% (from 60% to 81%); for the second experiment, by 4% (from 77% to 81%) and 5% (from 66% to 71%). An increase in the OA and kappa coefficients indicated an improvement in the land cover classification. For the first experiment, by 20% (from 63 to 83%) and 21% (from 60% to 81%); for the second experiment, by 4% (from 77% to 81%) and 5% (from 66% to 71%).

5. Conclusions

In this paper, we have presented a training sample clustering technique. The purpose of this technique was to increase the separability of the training sample set. Because separability directly affects classification accuracy, the technique increases the accuracy along with the separability.

In order to assess separability, an appropriate method for training sample separability assessment was developed and presented.

Further research should aim at applying the proposed technique to other classification methods.

Contributions of authors: conceptualization, methodology – Artem Andreiev, Leonid Artiushyn; formulation of tasks, analysis – Leonid Artiushyn; geospatial data processing – Artem Andreiev; writing – original draft preparation, writing – review and editing – Artem Andreiev, Leonid Artiushyn.

Conflict of Interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Financing

This study was conducted without financial support.

Data Availability

Data will be made available upon reasonable request.

Use of Artificial Intelligence

The authors confirm that they did not use artificial intelligence technologies in their work.

All authors have read and agreed to the publication of the finale version of this manuscript.

References


17. Sedgwick, P. Cluster sampling. BMJ, 2014, article no.1215. DOI: 10.1136/bmj.g1215.


27. Popov, M. O. Metodologiya otsenki tochnosti klassifikatsii ob’yektov na kosmicheskih izobrazheniyakh [Methodology of accuracy assessment of classification of objects on space images]. Journal of Automation
ПІДВИЩЕННЯ ТОЧНОСТІ КЛАСИФІКАЦІЇ ЗЕМНИХ ПОКРИВІВ
ШЛЯХОМ КЛАСТЕРИЗАЦІЇ НАВЧАЛЬНОЇ ВИБІРКИ

А. А. Андреєв, Л. М. Артюшин

Предметом вивчення в статті є класифікація земних покривів. Як відомо, саме керовані методи класифікації є актуальними для більшості тематичних задач дистанційного зондування Землі, оскільки вони дають можливість задати характеристики вихідних класів у вигляді навчальної вибірки на відміну від некерованих методів. Існує значна кількість підходів до обробки набору навчальної вибірки, але їхнім спільним недоліком є те, що вони не враховують фактор розділимості навчальної вибірки. Дана характеристика показує, наскільки кількість навчальної вибірки впливає на точність класифікації. Одним з варіантів підвищення точності класифікації є кластеризація навчальної вибірки. Отже, метою даного дослідження була розробка методики кластеризації навчальної вибірки, яка дозволяє підвищити точність класифікації земного покриву безпосередньо за рахунок підвищення розділимості навчальної вибірки. Таким чином завдання цього дослідження наступні: 1) розробити метод оцінювання розділимості навчальної вибірки; 2) розробити методику кластеризації навчальної вибірки; 3) перевірити ефективність розробленої методики, виконавши експериментальну класифікацію земних покривів із застосуванням розробленої методики. В якості експерименту було отримано дві класифікації для кожного з двох обранých територій дослідження: одна класифікація до застосування навчальної вибірки, а друга – після. Було залучено шість класів в кожному з експериментів. Навчальну вибірку було відібрано для кожного з класів. Висновки. Після застосування розробленої методики було зафіксовано підвищення точності класифікації, яке зумовлено зменшенням розділимості. В свою чергу, це призвело до підвищення точності класифікації. Для другого експерименту це зазначене підвищення загальної точності класифікації та кла- керованої на 3% до 8% та 2% (з 60% до 81%) відповідно. А для другого експерименту підвищення загальної точності класифікації та керованої становило 4% (з 77 до 81%) та 5% (з 66 до 71%) відповідно.

Ключові слова: класифікація; керовані класифікації; некеровані класифікації; кластеризація; ДЗЗ; навчальна вибірка; розділимість навчальної вибірки.

Андреєв Артем Андрійович – доктор філософії, молодший науковий співробітник, ДУ "Науковий центр аерокосмічних досліджень Землі ІГН НАН України", Київ, Україна.

Артюшин Леонід Михайлович – д-р техн. наук, проф., головний науковий співробітник, Державний науково-дослідний інститут авіації, Київ, Україна.

Artem Andreiev – PhD, Junior Researcher, Scientific Centre for Aerospace Research of the Earth, Kyiv, Ukraine, e-mail: artem.a.andreev@gmail.com, ORCID: 0000-0002-6485-449X, Scopus Author ID: 57203280237.

Leonid Artiushyn – Doctor of Technical Sciences, Professor, State Research Institute of Aviation, Kyiv, Ukraine, e-mail: artleonid2017@gmail.com, ORCID: 0000-0002-7488-7244, Scopus Author ID: 6602317317.