Semi-supervised learning through hierarchical clustering for interactive aerospace image analysis

Sergey Rylov[1,*]

[1]Institute of Computational Technologies of Siberian Branch of the Russian Academy of Sciences, 630090, Academician M.A. Lavrentiev avenue, 6, Novosibirsk, Russia

Abstract. A new semi-supervised classification algorithm based on the non-parametric clustering algorithm HCA is proposed. The algorithm obtains hierarchical segmentation result where additional classes that are not represented in the training samples can be found. High performance of the algorithm allows using it in interactive mode. Experimental studies confirm that the proposed algorithm provides aerospace image classification in conditions of limited number of training samples.

1 Introduction

In many practical problems of data classification, such as satellite images, the process of collecting qualitative training samples (TS) is usually connected with considerable material and time expenditure [1-4]. Therefore, in practice training samples are often limited and unpresentable. Besides, TS can be missing for some classes causing an adverse effect on the classification quality. The use of insufficient training samples does not provide satisfactory quality of segmentation, especially in the analysis of large and complex scenes.

In conditions of insufficient training samples, it is expedient to use classification methods with semi-supervised learning. In the course of their work, not only the information contained in the labelled training samples is used but also unlabelled (unclassified) data is utilized to construct the decision rule. The problem of semi-supervised classification can be considered in two formulations: as a classification that additionally uses information from unclassified data; and as clustering that uses additional conditions from training samples [1, 11]. The use of semi-supervised learning algorithms can reduce manual labour and improve the quality of classification results [3-4].

Over the past few years, works devoted to application of semi-supervised classification methods to satellite images have appeared [5-10], but their practical use is rather difficult due to the large number of configurable parameters and high computational costs [9]. Thus, it is urgent to develop new computationally efficient classification algorithms with semi-supervised learning that will allow processing aerospace images in conditions of small volume of training samples.

* Corresponding author: RylovS@mail.ru

This paper presents a new computationally efficient semi-supervised classification algorithm for aerospace image segmentation, based on the hierarchical grid-based clustering algorithm (HCA) [12], where training samples are taken into consideration at the hierarchy construction stage. The hierarchy itself is based on the set of small groups of data, obtained by nonparametric density estimation in the space of spectral features. The result is represented as a hierarchical segmentation, where, in addition to the required classes, there can be also found other classes not represented in the training samples. Algorithm's high performance provides the expert with the opportunity to use it in an interactive mode, setting extra training samples to get more accurate results if necessary. Experimental results confirm that the proposed algorithm allows classifying images in conditions of small size training samples.

2 Grid-based hierarchical clustering algorithm (HCA)

This section provides a brief description of the clustering algorithm HCA [12], which can be divided into four main stages.

At the first stage, a grid structure is formed in the feature space — it is divided into disjoint cells, the size of which depends on the parameter m. The density of a cell is determined through the number of data elements trapped in this cell.

At the second stage, each non-empty cell is connected to the corresponding neighbouring cell with the highest density. Thus, the set of non-empty cells is divided into connected components $\{G_1,\dots,G_S\}$, which correspond to single-mode clusters. In the HCA algorithm, we take the connected components as the basic elements of the hierarchy. The number of obtained connected components is small relative to the number of all data elements, and therefore the hierarchy construction does not require high computational costs.

At the third stage, the distances between all pairs of adjacent components are found. The distance is determined by estimating the density drop between the components. As a result, the corresponding distance matrix $\{h_{ij}\}$ is formed.

The distances between arbitrary connected components $\{\hat{h}_{ij}\}$ are determined through the distances between adjacent connected components $\{h_{ij}\}$ as follows. Let $\Theta_{ij} = \{Q_{ij}\}$ be the set of all chains of the connected components $Q_{ij} = \langle\, G_i = G_{k_1}, \dots, G_{k_t}, G_{k_{t+1}}, \dots, G_{k_l} = G_j \,\rangle$ such that for all $t = 1,\dots,l-1$, the components $G_{k_t}, G_{k_{t+1}}$ are adjacent. Then the distance between arbitrary connected components G_i and G_j is defined by the formula

$$\hat{h}_{ij} = \min_{Q_{ij} \in \Theta_{ij}} \left[\max_t h_{k_t, k_{t+1}} \right] \tag{1}$$

If the set Θ_{ij} is empty, we assume that $\hat{h}_{ij} = 1$.

The advantage of the introduced distance $\{\hat{h}_{ij}\}$ is that it has ultrametric property [13], i.e. it is a metric that satisfies the strong triangle inequality: $\hat{h}_{ij} \le max\,(\hat{h}_{ik}, \hat{h}_{kj}),\ \forall\ i, j, k$. It is known that there exists a unique correspondence between the distance matrices with ultrametric property and dendrograms [14]. Therefore, such matrix describes some hierarchical partition.

At the final stage, the single linkage algorithm (SLINK) for dendrogram construction is applied to the distance matrix $\{h_{ij}\}$. As a result, we obtain a dendrogram corresponding to the ultrametric $\{\hat{h}_{ij}\}$.

The clustering algorithm HCA allows obtaining hierarchical clustering structure of the data, while it is able to separate clusters intersecting in the feature space. The use of non-parametric density estimation allows distinguishing clusters of complex shape. The implemented algorithm is capable of processing data with up to 8 dimensions and up to

several hundred millions of elements. At the same time, its computational efficiency allows clustering multispectral four-band images up to 100 million pixels in size within one second on a regular PC.

3 Semi-supervised hierarchy construction

In the proposed approach, the training samples are taken into account at the stage of hierarchy construction. When clustering with HCA algorithm, the hierarchy is formed with SLINK dendrogram construction method, which is applied to the distance matrix between adjacent connected components. The SLINK method for an $n \times n$ size matrix consists of (n-1) iterations. At each iteration, two nearest elements are united, while the distance to the new combined element is defined as the minimum of the distances to the elements to be united. There exists an algorithm [15] based on the use of an array of partial minima (containing indices and values of minimal elements in each row of the distance matrix), which allows finding the minimum matrix element at each iteration with $O(n)$ operations. Thus, the overall achieved computational complexity of the SLINK algorithm is $O(n^2)$.

Within semi-supervised approach, before the dendrogram construction, we establish classes for the connected components due to the available training samples. The component is assigned to the class of the training samples, which fall into this component. If the component contains training samples from different classes, then this component is marked as a `conflicting` one, and its class is established by voting of the TS elements. Components that do not contain a single TS element remain unlabelled. At this stage, the distances between the `non-conflicting` components belonging to the same class are set to zero. Thus, at the stage of hierarchy construction, such components are immediately united into one class.

Component's class labels are considered at the stage of dendrogram construction. At each iteration of the SLINK algorithm, the class labels of the elements to be united are checked. If they belong to different classes, then the union does not take place. Instead, the distance between them is set to one, and the minima in the corresponding rows of the matrix are updated. In other cases, the union takes place as usual. At the same time when a labelled element is united with an unlabelled one, the combined element inherits the class of the labelled element. Thus, the unlabelled components are eventually united with the closest labelled components in the metric (1). As a result, the top of the hierarchy consists of the combined components belonging to the required classes, as well as the groups of unlabelled components that turned out to be at distance 1 from all the marked ones (meaning they represent their own classes).

It is possible to retrieve data partitioning with various detail degree from the given hierarchical structure by varying the dendrogram cut value. Moreover, the classes known from training samples will stay separated at all levels. After obtaining the result, the user can interactively add new elements to the training samples, and the algorithm will only require performing anew the dendrogram construction step to update the result, which does not take significant time.

4 Experimental studies

In this section, the results of the proposed semi-supervised classification algorithm on model data and images are presented. It is shown that with the use of small size training samples (TS), the proposed approach allows distinguishing similar classes successfully, while avoiding unnecessary fragmentation.

Fig. 1,*a* represents the two-dimensional model dataset consisting of five classes with normal distribution, simulating vegetation classes (birch forests, coniferous forests, meadow vegetation, wetlands, agricultural land) on Landsat satellite image in red and green spectral features. Training samples (8 points) are marked on the image with black dots. With the use of these limited TS, classification algorithms make significant mistakes on the red class. For example, overall accuracy of Minimum Distance classification is 93,32 % (Fig. 1,*c*). At the same time, clustering algorithms have a problem to separate strongly intersecting purple and blue classes. Expectation-maximization clustering overall accuracy reaches only 89,28 % (Fig. 1,*d*). Nevertheless, the proposed semi-supervised classification algorithm managed to successfully extract all classes with 96,24 % overall accuracy (Fig. 1,*b*).

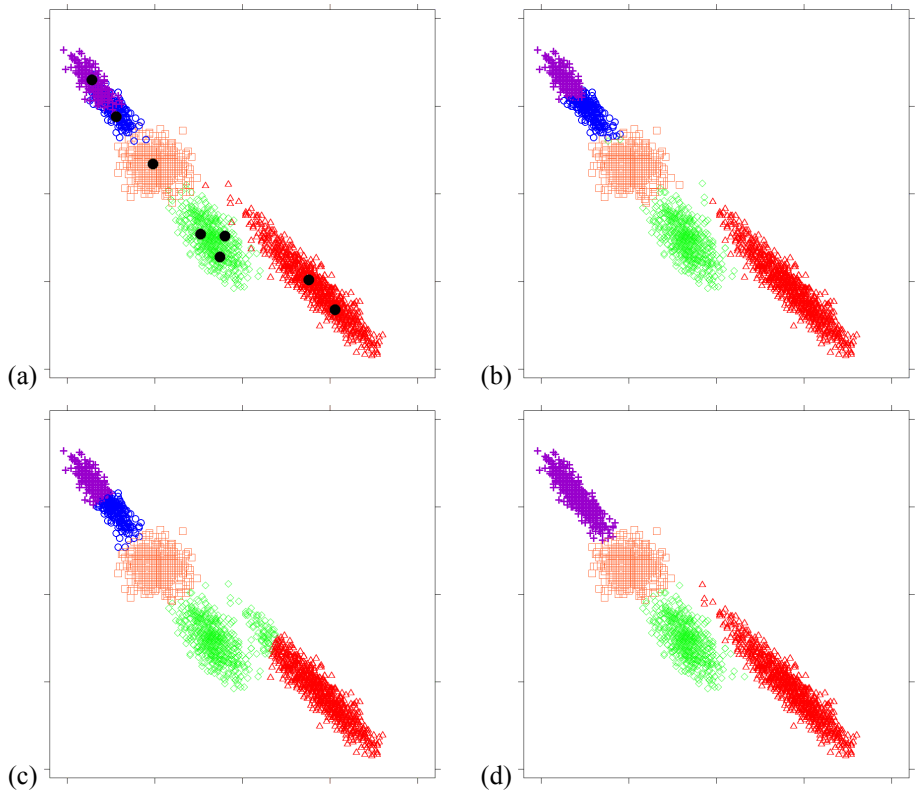

Fig. 1. Model dataset with TS tags (a); semi-supervised classification result (b); Minimum Distance classification result (c); Expectation-maximization clustering result (d)

Fig. 2 presents the model colour image containing the ring and the background, which consist of gradient transitions of various colours. Also, the image contains tags of the training samples (4 points for the ring and 5 points for the background). When clustering this image, a fragmented segmentation result is obtained, where both classes break up into many clusters, otherwise the ring and the background cannot be separated. However, with the given limited training samples, the proposed semi-supervised approach successfully distinguishes required classes (Fig. 2).

Fig. 3 shows the semi-supervised classification result of the high spatial resolution multispectral satellite image with training samples consisting of only 8 points. As a result, all 6 required classes were successfully extracted, including a forest that is extremely heterogeneous in its spectral characteristics. Moreover, the proposed algorithm additionally

marked out a class that is not represented in the training samples, which corresponds to a football field (marked in white colour).

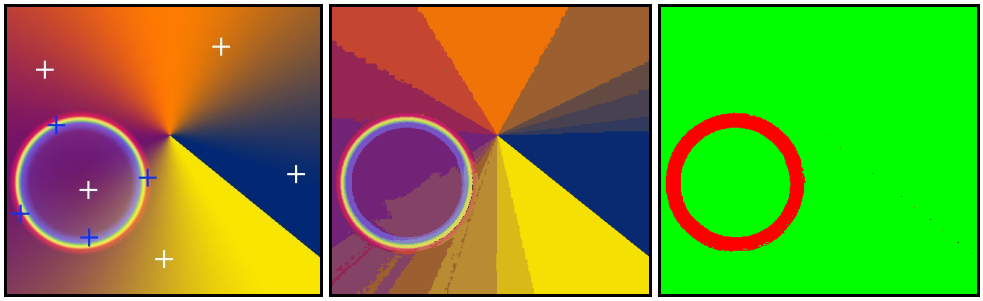

Fig. 2. Model image with TS tags (left); HCA algorithm clustering result (centre); semi-supervised classification result (right)

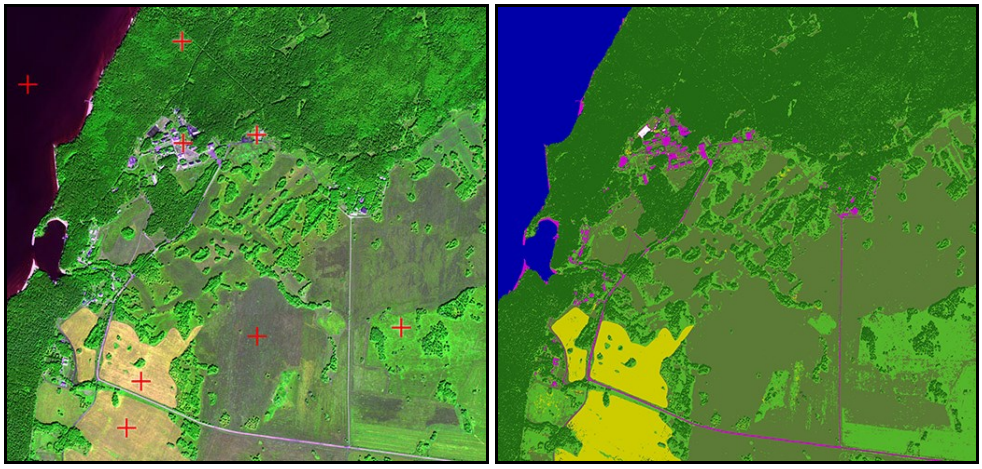

Fig. 3. WorldView-2 satellite image (RGB composite, bands 4, 6, 1) with TS tags and the result of the semi-supervised classification algorithm

5 Conclusion

The proposed semi-supervised classification algorithm can perform aerospace image segmentation under conditions of small and incomplete training samples. This provides the potential to the effective use of data obtained from field studies, which are commonly not used directly in the automated satellite image processing. Hierarchical representation of the segmentation result and high performance of the algorithm significantly facilitate the work of an expert.

For further research, we are planning to implement automatic partition of the `conflicting` components with a view to better fit with the training samples.

This work was supported by the Russian Foundation for Basic Research (No. 18-37-00492 mol_a).

References

1. O.I. Travkin, *Proc. XVIII Intern. Conf.: Analytics and data management in data-intensive areas*, 361-369 (Ershovo, Moscow Region, 2016)

2. Yu.V. Adaskina, A.M. Popov, P.V. Rebrova, *Proc. XVIII Intern. Conf.: IMS*, 15-24 (St. Petersburg, 2015)
3. V.J. Prakash, L.M. Nithya, IJCTT, **8**, 25-29 (2014)
4. M.F.A. Hady, F. Schwenker, *Handbook on Neural Information Processing*, 215-239. (Springer, 2013)
5. B. Banerjee, K.M. Buddhiraju, Journal of the Indian Society of Remote Sensing, **43 (4)**, 719-728 (2015)
6. K. Tan et al., ISPRS Journal of Photogrammetry and Remote Sensing, **97**, 36-45 (2014)
7. R. Luo et al., SPIE Remote Sensing, **10004**, 100040T (2016)
8. L. Yang et al., IEEE Geoscience and Remote Sensing Letters, **11 (3)**, 651-655 (2014)
9. L. Wang et al., ISPRS Journal of Photogrammetry and Remote Sensing, **97**, 123-137 (2014)
10. X. Jing, S.Y. Chen, L.L. Fan, *Proc. Ninth Intern. Conf.: ICDIP 2017*, **10420**, 1042030 (2017)
11. L. Lelis, J. Sander, *Proc. Ninth IEEE Intern. Conf. : ICDM'09*, 842-847 (IEEE, 2009)
12. S.A. Rylov, I.A. Pestunov, Journal of Physics: Conference Series (IOP Publishing, 2018) (to be published)
13. B. Leclerc, Math. Sci. Humaines, **127 (73)**, 5-37 (1981)
14. A. Mirzaei, M. Rahmati, IEEE Tr. Fuzzy Syst, **18 (1)**, 27-39 (2010)
15. C.F. Olson, Parallel computing, **21 (8)**, 1313-1325 (1995)