Experimental evaluation of nonparametric clustering algorithms for image segmentation

Yuriy Sinyavskiy[*], *Sergey* Rylov , and *Igor* Pestunov

Federal Research Center for Information and Computational Technologies, 630090, Novosibirsk, Russia

Abstract. Experimental evaluation of 12 nonparametric clustering algorithms for image segmentation was made. Algorithms developed in FRC ICT are compared to ones from ENVI, ELKI and Smile software packages. Seven model datasets were generated to estimate clustering accuracy. The computational efficiency was evaluated using digital photographs and fragments of multispectral images obtained from WorldView-2 satellite.

1 Introduction

Image segmentation is required for solution of a number of applied problems. These can be multispectral images obtained from satellites, aircrafts or unmanned aerial vehicles, as well as conventional digital photographs (e.g. medical examinations data). Segmentation has two main goals – splitting the image into parts for further analysis, and grouping pixels into higher-level informative structures [1]. One of the most common approaches to image segmentation is based on the use of data clustering algorithms [2]. Image segmentation is usually performed with neither *a priori* information about the probabilistic characteristics of classes, nor training samples. In these conditions, the most suitable is nonparametric approach to clustering [3]. It allows detecting clusters of complex structure without strict restrictions on probability density function. However, it has not become widespread due to high computational complexity. The use of the grid-based approach makes it possible to achieve high computational efficiency due to processing relatively small number of cells instead of data elements. But the clustering accuracy of detected clusters strongly depends on the grid structure [4]. Efficient density- and grid-based algorithms for multispectral images segmentation have been developed in the FRC ICT. In this paper, an experimental comparison of these algorithms and six most popular clustering algorithms implemented in ENVI [5], ELKI [6] and Smile [7] software packages is performed.

2 Algorithms and datasets

Experimental evaluation was performed using twelve clustering algorithms. Six of ones (HCA_MS [8; 9], HECA_MS [10], ECCA [11], CCAE [11], MeanSC [12], EMeanSC [13])

[*] Corresponding author: yorikmail@gmail.com

has been developed at the FRC ICT. Three efficient nonparametric algorithms (DBSCAN [14], OPTICS [15] and DENCLUE [16]), were taken from ELKI and Smile software packages for data mining. In addition, a parallel implementation of the effective (in terms of clustering quality) iterative density-based MeanShift algorithm [17] was made. The number of iterations in the experiments was limited to ten. Furthermore, software implementations of k-means [18] and ISODATA [18] clustering algorithms from ENVI software package were evaluated (the number of iterations was also limited to ten; cluster centers from the previous iteration were used to initialize the next one, improving the quality of segmentation). These algorithms are included in most popular software packages for satellite image processing, and therefore are often used in practice.

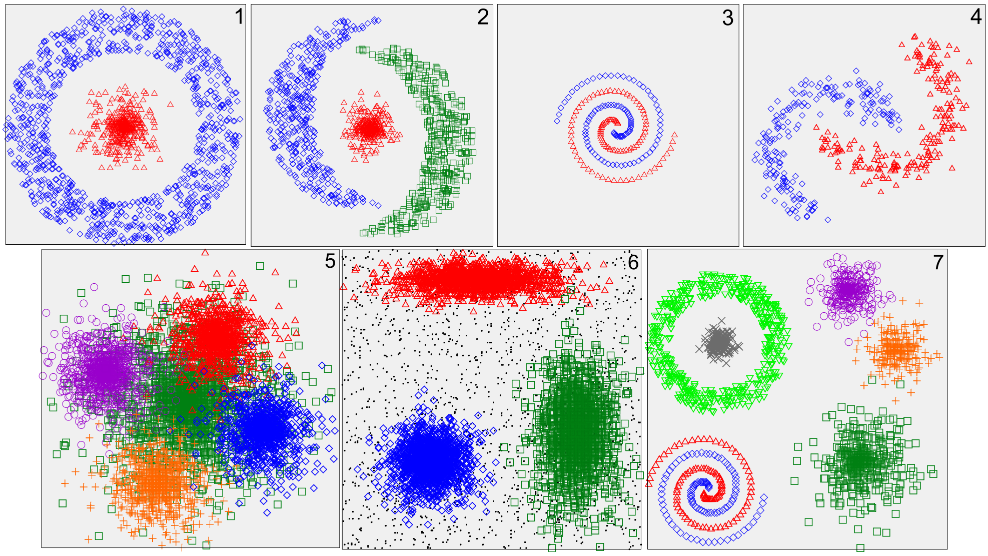

Fig. 1. Reference clustering for model datasets (black dots mark «noise»).

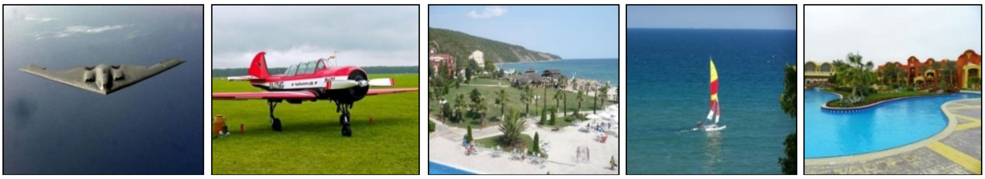

Fig. 2. Test images No. 1-5 (digital photos). Image sizes are 0.3, 1.0, 2.2, 5.0 and 13.8 millions of pixels, respectively.

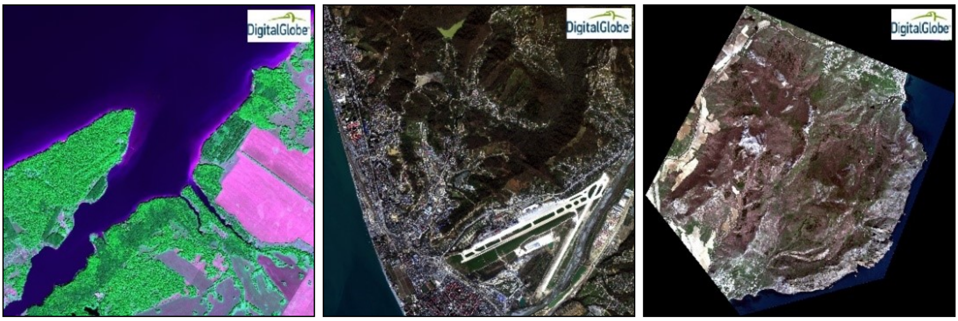

Fig. 3. Test images No. 6-8 (fragments of multispectral images obtained from WorldView-2 satellite). Image sizes are 4.2, 9 and 12 millions of pixels, respectively.

In this work, seven model datasets (Figure 1) and eight test images were used – five digital photos (Figure 2) and three fragments of multispectral images obtained from the WorldView-2 satellite (Figure 3). Three spectral channels (red, green and blue) were used for processing digital photos, and four (red, green, blue and near-infrared) – for satellite images. The model datasets and RGB composites for the test images are available at [19]. The experiments were performed on a personal computer with an Intel Core i7 CPU (4 cores, 2.3 GHz each) and 8 GB RAM.

3 Experimental evaluation

In the first experiment, the clustering accuracy was estimated according to the following definition. Suppose that for a dataset $X = \{x_1, \dots, x_N\}$ of size N, the reference partition g^* into M classes $\{G_0^*, \dots, G_M^*\}$ is known. Then, for an arbitrary partition (clustering) g into K clusters $\{G_0, \dots, G_K\}$, the correspondence function $\gamma(G)\colon \{G_0^*, \dots, G_M^*\} \to \{G_0, \dots, G_K, \emptyset\}$, can be established performing condition $\forall(G \neq \bar{G})\colon \gamma(G) = \gamma(\bar{G}) \Leftrightarrow \gamma(G) = \emptyset$ and the highest value of the expression

$$n_\gamma = \sum_{i=1}^{M} [|G_i^* \cap \gamma(G_i^*)| \cdot I(\gamma(G_i^*) \neq \emptyset)],$$

where $I(\cdot)$ is the characteristic function. Then the clustering accuracy (acc) is determined as:

$$acc = \frac{n_\gamma}{N} * 100\%.$$

The goal of tuning the algorithm parameters was to obtain the maximum clustering accuracy. The MeanShift, k-means and DENCLUE algorithms do not allow detecting “noise”, therefore, the acc values for model dataset No. 6 were calculated without taking the “noise” class into account. The segmentation accuracy values obtained on the model datasets are presented in Table 1, and the processing time is shown in Table 2.

All of the algorithms, except for DENCLUE, k-means, and MeanShift, allow obtaining a reference partition for model datasets No. 1-3. Processing of model dataset No. 4 by the MeanSC and EMeanSC algorithms leads to one misclassified data element. The DBSCAN and OPTICS algorithms make it possible to obtain a slightly less accurate result. All algorithms, except for DBSCAN and OPTICS, allowed to obtain a clustering accuracy of about 85% for model dataset No. 5. Errors are caused by significant overlap of model classes. Model dataset No. 6 contains «noise» that could not be detected by the MeanShift, k-means and DENCLUE algorithms. The algorithms developed at FRC ICT made it possible to obtain a clustering accuracy higher than 84% for model dataset No. 6. The rest of the algorithms achieved less than 80% accuracy. When processing model dataset No. 7, only the HCA_MS, HECA_MS, ECCA, CCAE, MeanSC and EMeanSC algorithms allow to detect all classes. Applying DBSCAN, OPTICS, DENCLUE, and k-means algorithms leads to correct separation only for the normally distributed classes. The results of the experiment demonstrate that the algorithms developed at the FRC ICT are superior to the known nonparametric algorithms in terms of clustering quality and/or processing time.

In the second experiment, the considered algorithms were applied to test images and the processing time was compared. The results are presented in Table 3. Dashes in the table correspond to unacceptably high processing times (more than 18 hours). Analysis of the results shows that the DBSCAN, OPTICS and DENCLUE algorithms do not allow efficiently handle large images. In addition, their processing time increases significantly with increasing number of channels. Algorithms from the ENVI package are better adapted to image analysis, but the processing time for images larger than 9 million pixels exceeds 5 minutes. On the other hand, algorithms HCA_MS, HECA_MS, ECCA, CCAE, MeanSC and EMeanSC allow interactive segmentation of large multispectral images.

Table 1. Clustering accuracy (acc) for model datasets (percentage).

Algorithm	Model dataset						
	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7
HCA_MS	100	100	100	100	83.13	84.43	93.26
ECCA	100	100	100	100	84.83	85.65	98
CCAE	100	100	100	100	84.83	85.65	98
HECA_MS	100	100	100	100	85	86.57	93.26
MeanSC	100	100	100	99.75	86.7	89.18	98.6
EMeanSC	100	100	100	99.75	86.7	89.16	98.7
MeanShift	51.79	90.36	50	53.75	86	79.53	79.1
k-means	47.79	64.73	53	47.75	84.67	79.92	78.99
DBSCAN	100	100	100	99	65.2	84.56	90.63
OPTICS	100	100	100	99	65.2	84.56	90.63
DENCLUE	65.14	99.36	50	50.5	85.9	79.92	83.8

Table 2. Processing time for model datasets (in seconds).

Algorithm	Model dataset						
	No. 1	№ 2	No. 1	№ 4	No. 1	№ 6	No. 1
HCA_MS	0.008	0.001	0.002	0.001	0.003	0.004	0.001
ECCA	0.003	0.005	0.003	0.003	0.001	0.028	0.013
CCAE	0.006	0.011	0.003	0.008	0.002	0.15	0.019
HECA_MS	0.006	0.008	0.004	0.003	0.006	0.03	0.021
MeanSC	0.002	0.005	0.004	0.002	0.01	0.014	0.017
EMeanSC	0.018	0.017	0.007	0.007	0.13	0.19	0.17
MeanShift	0.034	0.019	0.003	0.005	0.321	0.259	0.017
k-means	0.016	0.015	0.015	0.016	0.016	0.085	0.017
DBSCAN	0.031	0.016	0.016	0.016	0.374	1.17	0.031
OPTICS	0.09	0.074	0.048	0.065	1.131	2.461	0.14
DENCLUE	0.178	0.09	0.06	0.14	0.184	3.26	0.721

Table 3. Processing time for test images (in seconds).

Test image	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8
Image size (millions of pixels)	0.3	1	2.2	5	13.8	4.2	9	12
Number of channels	3	3	3	3	3	4	4	4
HCA_MS	0.13	0.33	1.04	0.85	4.75	26.3	6.1	666
ECCA	0.17	0.48	0.66	3.1	5.5	1.3	3.54	3
CCAE	0.05	0.09	0.15	0.32	0.63	7.8	6.5	1
HECA_MS	0.05	0.23	0.74	0.54	2.74	47	9.49	183.3
MeanSC	0.09	0.51	0.86	1.44	8.99	1.44	8.16	4.2
EMeanSC	0.39	2.25	3.16	5.21	31.31	4.97	28.74	10.47
MeanShift	2.91	52	102	67	388	4138	217	62388
k-means	0.5	36	5	17	1196	75	302	588
ISODATA	1	15	5	9	1178	68	332	337
DBSCAN	194	2731	13098	–	–	39965	–	–
OPTICS	638	5244	40013	–	–	–	–	–
DENCLUE	6934	39849	–	–	–	–	–	–

4 Conclusion

An experimental comparison of the HCA_MS, HECA_MS, ECCA, CCAE, MeanSC, and EMeanSC algorithms, developed at the FRC ICT, with the nonparametric algorithms DBSCAN, OPTICS and DENCLUE, as well as the *k*-means and ISODATA algorithms from the ENVI software package, has been performed. It was shown that the developed algorithms are superior to the most popular clustering algorithms in terms of clustering quality and/or processing time. In addition, the *k*-means and ISODATA algorithms which are included in common software packages and, as a consequence, are often used in practice, do not allow interactive processing of large multispectral images. On the other hand, the algorithms HCA_MS, HECA_MS, ECCA, CCAE, MeanSC and EMeanSC, thanks to the use of modern approaches to clustering (density-, grid- and ensemble-based), allow dialog-mode image segmentation.

This research was supported by RFBR (grants No. 18-07-00600, 19-29-01175).

References

1. L. Shapiro, G. Stockman. *Computer vision.* (New Jersey: Prentice Hall, 2001)
2. G. Menardi, Stat. Anal. Data Min.: The ASA Data Sci. J. **13 (1)**. 83-97 (2020)
3. I.A. Pestunov, Yu.N. Sinyavskiy. Bulletin of KemSU. **4/2 (52)**. 110-125 (2012) (In Russ.)
4. D. Krstinic, A.K. Skelin, I. Slapnicar. IET Image Process. **5 (1)**. 63-72 (2011)
5. http://www.harrisgeospatial.com/SoftwareTechnology/ENVI.aspx
6. https://elki-project.github.io
7. https://haifengl.github.com/smile
8. I.A. Pestunov, S.A. Rylov, V.B. Berikov. Optoelectronics, Instrumentation and Data Processing. **51(4).** 329-338 (2015)
9. S.A. Rylov. CEUR Workshop Proceedings. **2033.** 150-155 (2017)
10. S.A. Rylov, I.A. Pestunov. Modern Problems of Earth Remote Sensing from Space. 52 (2017) (In Russ.)
11. I.A. Pestunov, S.A. Rylov, Yu.N. Sinyavskiy, V.B. Berikov. CEUR Workshop Proceedings. **1901.** 194-200 (2017)
12. I.A. Pestunov, Yu.N. Sinyavskiy. Avtometriya. Second IASTED Intern. Multi-Conf. on Automation, Control, and Information Technology – Signal and Image Processing. 5-9 (2005)
13. I.A. Pestunov, V.B. Berikov, Yu.N. Sinyavskiy. Vestnik SibSAU. **5 (31)**. 56-64 (2010) (In Russ.)
14. M. Ester, H.-P. Kriegel, J. Sander, X. Xu. KDD'96. 226-231 (1996)
15. M. Ankerst, M.M. Breunig, H.-P. Kriegel, J. Sander. ACM SIGMOD. 49-60 (1999)
16. A. Hinneburg, D. Keim. Knowl. Inf. Syst. **5 (4)**. 387-415 (2003)
17. Y. Cheng. IEEE Trans. Pattern Anal. **17 (8)**. 790-799 (1995)
18. R.C. Gonzalez, R.E. Woods. *Digital image processing. 2nd ed.* (Addison-Wesley Pub (Sd), 2002)
19. https://drive.google.com/drive/folders/11r-w_RBC38KEYqFJueQ7tlWh6VndDPUn