The technology of the informative features searching method applying for the feature space dimension reduction in the problem of classifying areas of natural hyperspectral images

M.I. Khotilin

khotilin.mi@ssau.ru

Introduction

Hyperspectral images are three-dimensional data arrays that include spatial information about an object, supplemented with spectral information for each spatial coordinate. Currently, processing and analysis of hyperspectral images are popular research topics in the field of image processing and computer vision. Within the framework of this article, the technology of applying the method of searching for informative features of a hyperspectral image for the clustering problem, on the example of a separate area, is considered.

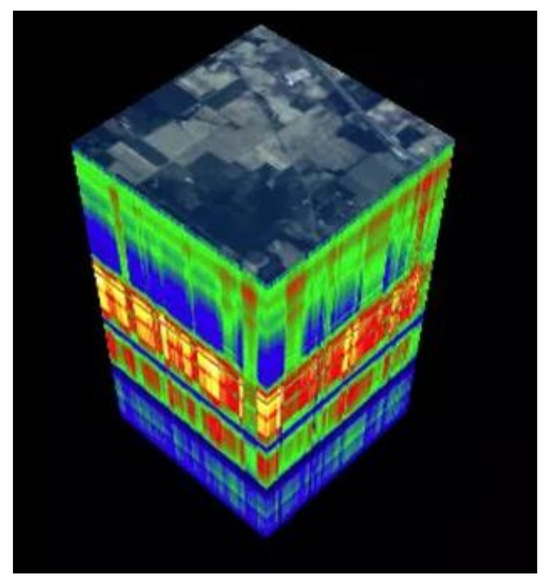

Fig 1. Hyperspectral image sample

Formulation of the problem

The entire course of work can be conditionally divided into sequentially performed stages. At the first stage, using the preprocessing module, a small area, which is used for research and processing, is allocated. Further, a set of all two-dimensional sections by planes is selected from the hypercube of the researched area. This stage is necessary because in advance it is not possible to say which layers are significant, and also because the use of existing methods and means of features calculating is difficult for the considered hyperspectral images.

To study and process the sections obtained above, it was decided to use the well-established MaZda software, which allows calculating various groups of features, as well as the high-level Python programming language. As a result of the work of this software product, we obtain a set of texture and brightness features, which will be used in the future.

The studied images may contain a significant amount of noise components, and it is necessary to perform processing to smooth them out. Further, by explicitly setting the number of clusters, it is possible to separate the data under study, thus obtaining data sets (features) for training, grouped according to certain criteria.

The resulting feature sets are large in size and contain data that may not carry meaningful information that is important in classification. Due to this, it is necessary to reduce the dimension and search for signs that are informative.

Various algorithms can be used to search for informative features. In the framework of this work, the method of sequential addition of features was used. Further, using various classification algorithms (LDA, SVM), it is possible to classify the received data.

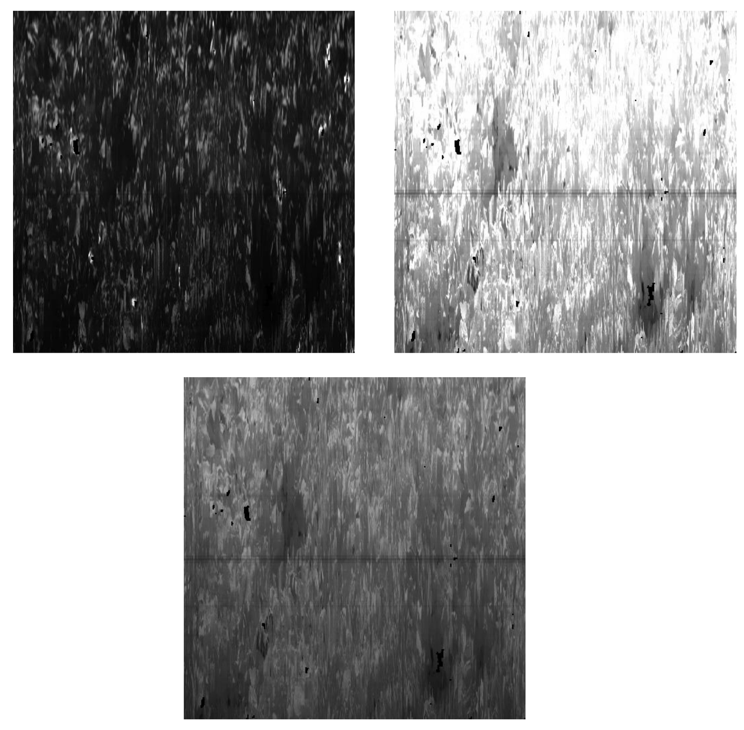

Fig 2. Sample layers of research areas

Practical research and software implementation

As initial data, a set of hyperspectral images of plant leaves was considered, with the number of spectral channels 242 and wavelengths from 436 nm to 965 nm. This set contained images of various classes of crops, such as: tomato, pepper, cabbage, carrot and others.

To research the algorithm described above, small areas of 10 × 10 pixels in size were cut out from the original image using a preprocessing module implemented in Python. The resulting new set of images is divided into training and test sets used in further research.

Next, using the same module, the resulting image hypercube - an array of brightnesses - was divided into two-dimensional ones for each pair of coordinates, and the resulting hypercube sections were used to further obtain a set of features, both brightness and texture, extracted using the MaZda software. Aggregating together the obtained features, we receive a set of aggregate features that characterize each image under consideration.

For the studied small areas of the image, the number of total brightness and texture features is 86878 features (24200 brightness features + 62678 texture features) for each image.

After this stage, it is necessary to reduce the dimension of the feature space due to the fact that calculations for all available features can take a lot of time and consume significant computing resources. To reduce the dimension in this work, we used the method which consists of the joint use of linear discriminant and correlation analysis. To search for informative features, the method of sequential addition of features was used.

As a result of using this approach, it was possible to reduce the dimension of the feature space under consideration from 86878 to 39, that is, by more than 2000 times.

Further, various classification algorithms were applied to the obtained data: LDA, SVM Logistic Regression, K-Nearest Neighbors - in order to select the algorithm with the highest classification accuracy. To assess the quality of classification, this paper proposes to use a measure equal to the ratio of the number of incorrectly classified objects to the total number of classified objects.

As a result, we obtain a hyperspectral image processing pattern that has lower resource requirements compared to classical methods. It can also be assumed that the use of this pattern may allow it to be used for various purposes in the field of image analysis, for example, when analyzing on mobile devices or unmanned aerial vehicles or drones.

Mean	Variance	Skewness	Kurtosis	Perc.01%	Perc.10%	Perc.50%	Perc.90%
157.14876	144.06051	0.1639918	0.45355819	131	143	157	173
162.1157	248.63124	-0.66460667	0.28063435	120	142	163	181
153.09917	169.51083	-0.090571516	-0.55467854	126	136	154	169
166.42149	159.45045	-0.0048762354	-0.28754545	137	149	166	183
161.1157	306.18496	-0.39521939	-0.632427	124	135	163	182
126.65289	115.11919	-1.0839733	2.0503131	89	114	129	138
117.47934	172.74544	-0.65975158	0.97799158	85	101	119	133
124.69421	86.212281	-1.8975042	5.231749	92	113	126	133
120.30579	322.22881	-0.94277475	0.33735527	71	93	127	137
103.52066	604.76197	-0.39328448	-0.95836516	52	66	109	133
139.28099	337.42518	0.020848394	-0.96826362	104	118	139	162
149.78512	97.424903	-0.42379834	-0.2208567	133	136	151	162
150.97521	301.85889	-0.33355898	-0.28789795	108	126	152	170
157.41322	299.89536	-1.4137731	1.9655896	103	140	161	175
148.53719	192.36432	-0.38173432	-0.62651238	115	131	152	165
150.71074	155.18079	0.37082298	0.27854873	127	133	151	165
136.14876	57.250598	-1.2862341	3.8945146	113	127	137	144
150.49587	59.687999	0.61356361	0.22847674	137	140	150	160
166.45455	119.40496	0.2603534	0.012614663	143	153	166	180
136.71901	114.16898	-0.47649492	-0.42093704	113	122	138	149
118.13223	296.18086	0.069470425	-0.25491302	81	98	118	141
95.785124	478.31746	-0.36284382	-0.48351579	45	63	100	123
104.7686	452.82248	-0.1997897	0.013403302	55	75	108	128
85.041322	637.60986	-0.011396138	-1.0092859	43	49	89	116
105.94215	402.03798	0.39155934	-0.72735138	74	82	104	135
85.818182	1926.6777	0.25170089	-1.236841	19	29	73	147

Fig 3. Fragment of the array of initial features data

Algorithm name	Accuracy
Linear Discriminant Analysis (LDA)	0,978
Support Vector Machines (SVM)	0,957
Logistic Regression	0,934
K-Nearest Neighbors	0,966

Fig. 4 The classification accuracy of the considered algorithms

Conclusion

Finding features that uniquely determine whether objects (image areas) belong to a certain class is one of the most important tasks of image classification and processing. Existing methods of image processing, determination of their features and classification algorithms work well with relatively small amounts of initial data. The calculations described in this paper were relatively small and were performed on a personal computer. Processing large arrays of source images takes considerable time and computational resources.

Currently, work is underway to study the possibility of constructing an algorithm that can much more efficiently cope with the task of searching for informative features of areas of hyperspectral images and their classification, while maintaining the accuracy of these processes.

The results obtained in this work can be used to develop tools for the intellectual analysis of hyperspectral images of various areas of human life and activity. For example, in agriculture, examining images taken from fields occupied by certain crops, one can find those that are weedy or that are different from those growing originally. Also, when compiling forest maps, it is possible to determine the composition of forests by hyperspectral images of their surface.

САМАРСКИЙ УНИВЕРСИТЕТ
SAMARA UNIVERSITY

SAMARA NATIONAL RESEARCH UNIVERSITY
34, Moskovskoye shosse, Samara, Russia (www.ssau.ru)