# Thematic processing of the Earth remote sensing multispectral data

V.V. Asmus, A.A. Buchnev, V.P. Pyatkin

## Abstract

The issues connected with the problem of choosing adequate algorithms of the Earth remote sensing multispectral data recognition are considered. The system of supervised classification based on Bayesian strategy of maximum likelihood for normally distributed vectors of features is represented. The system of cluster analysis including the algorithm of $K$ - means and clusterization based on the analysis of multidimensional histogram modes is described.

***Keywords:*** the Earth remote sensing, objects recognition, vector of features, supervised classification, cluster analysis, the algorithm of $K$ -means, the analysis of multidimensional histogram.

The Earth remote sensing (ERS) state-of-the-art is characterized by moving of basic researches from separate scientific-methodology issues to the development of full technologies of thematic problem solution. There is large experience of digital multispectral data ERS processing. However, the accuracy and the cost of the processing is still unsatisfactory from the user point of view.

The main problem of data ERS interpretation - the problem of increasing the quality of decoding - is connected with the problem of choosing of adequate algorithm of recognition. The difficulties arising are due to the following reasons [1]:

1. The structure of real data doesn't correspond to the data model used in the algorithm. For example, the failure of the proposal about normal data vectors distribution or the failure of the condition that the field of the measurements is random. As we can see from the experience, such situations take place when we use data in the format JPEG, as well as when the radiation from the scanned object is outside the limits of dynamic range of survey equipment. In these cases we should refuse from the methods requiring covariance matrices transformations or use the approaches increasing data dispersion (for example, in the classification as follows it is possible to add Gausian noise with zero mean and unity dispersion to spectral canals with zero dispersion).

2. The training sequences are non presentative: deficient number of data to restore the parameters of decision rule; incompatibility of training data and data, presented to the recognition ("pollution" of samples by mix vectors of measurements, i.e. the vectors, which are formed when some nature objects fall into resolution element of survey system, no exact correspondence of the training data got with the use of clusterisation to true thematic classes, noise of the equipment, atmospheric influence and etc.)

So, we can say that modern experience of automated data recognition ERS shows: practically it's impossible to determine beforehand which algorithm would be better from the point of view of correspondance between accuracy of classification and its cost. Therefore it's advantageous to put some algorithms into recognizing system and to make the choosing of optimal algorithm empirically on the step of training on the results of classification of test data. Then the chosen algorithm is used for recognition of all sets of vectors of measurements.

The represented system of *supervised classification (the classification with training)* consists of seven classifiers (one classifier is per-element and other six – object classifiers), based on the use of Bayesian strategy of maximum likelihood for the normal distributed vectors data, and two object classifiers based on the minimum distance [2]. For the element we mean $N$ -dimensional vector of the features $x = (x_1, \ldots, x_N)^T$, where $N$ – the number of spectral bands, and for the object we mean the block of adjacent vectors of square or crosswise form. We supposed that vectors $x$ have a normal distribution $N(m_i, B_i)$ in class $\omega_i$ with mean value $m_i$ and covariance matrix $B_i$ . The decision about the belonging of the object central element to one or another class is made on the base of the classification of the whole object. Classifiers OMARK1, OIND1, OMEANI1, OMEAN1 operate with the block in the form of the cross and the classifiers OMARK3, OIND3, OMEANI3, OMEAN3 operate with square block. Classifiers OMARK1 and OMARK3 incorporate image spatial characteristics on the base of causal Marcov random field model of the first and third order correspondingly. OIND1 and OIND3 are based on the assumption that the vectors inside the block are independent. OMEANI1 and OMEANI3 classify the mean of the block under the assumption that the vectors inside the block are independent. OMEAN1 and OMEAN3 classify the mean of the block under the assumption that the vectors inside the block are independent and with covariance matrix equal identity matrix. When the size of the object is equal to 1, all object classifiers become per-element classifiers based on the strategy of maximum likelihood, except classifiers OMEAN1 and OMEAN3, which become per-element classifiers based on the minimum of distance.

As an example let's take the algorithm of the operation of classifiers OMEANI1 and OMEANI3. Let $\Omega = (\omega_1, \ldots, \omega_m)$ – finite set of classes, $X = (x^1, \ldots, x^k)$ – the object consisting from $k$ $N$ -dimensional vectors $x^l = (x_1^l, \ldots x_N^l)$ . The mean vector $x$ of an object $X$ is calculated:

$$x = \frac{1}{k} \sum_{l=1}^{k} x^l .$$

Discriminant function of the class ( $\omega_i$ ) is

$$g_i(x) = \ln(p(\omega_i)) - 0.5 \ln(|B_i|) - \\ - 0.5k(x - m_i)^T B_i^{-1} (x - m_i),$$

where $p(\omega_i)$ – a priori probability of the class $\omega_i$ . Let's designate $T_i$ based on the distribution $\chi^2$ threshold value for rejected vectors of the class $\omega_i$ :

$$T_i = \ln(p(\omega_i)) - 0.5A(N,Q) - 0.5\ln(|B_i|),$$

where $A(N,Q)-$ the critical value of the level $Q$ of the distribution $\chi^2$. Let $t_i$- the variable, the value of which depends on classifier parameter $thr$: $t_i = -\infty$, $thr = 1$, $t_i = T_i$, $thr = 2$, $t_i = \min_{l=1}^{m} T_l$, $thr = 3$, $t_i = \max_{l=1}^{m} T_l$, $thr = 4$, $t_i = (\sum_{l=1}^{m} T_l)/m$, $thr = 5$. Then, the decision rule for this classifiers is as follows: the central element of the object $X$ is entered to the class $\omega_i$ if $g_i(x) > g_j(x)$ for all $j \neq i$ and $g_i(x) > t_i$. On the contrary, the central object element is entered to the class of rejected vectors.

Statistical characteristics, which are necessary for the construction of discriminant functions – mean vectors, covariance matrices, coefficients of spatial correlation between the coordinate values of neighbourhood vectors in horizontal and vertical directions - are defined on the base of the vectors from training fields. All classifiers can be used in two modes – test and operational. According to the results of the work of the classifiers in test mode error matrix is formed. While analyzing the latest, we can estimate the quality of training. The result of the work of the classifiers in the operational mode is oneband (byte) image, which pixels values are the numbers of the classes. This image is colored into preliminary defined colors, which can be replaced in interactive mode with the colors defined by a user. Besides, to this image we can apply one of the two functions of post classification to delete isolated pixels.

The system of supervised classification has the following characteristics: the number of training images – up to 9, the number of classes – up to 15, the number of training and control fields in a class – up to 10, the size of each field – up to 50*50, object size – from 1*1 up to 11*11, the dimension of the data vectors isn't limited.

*Cluster analysis* is represented by two algorithms – the method of *K*-means and the method of modes analysis of multidimensional histogram [1].

The first approach is based on iterative procedure of referring of features vectors to the clusters according to the criterion of distance minimum from a vector to the cluster center. The optimal splitting of input vectors to the clusters is considered the one when the deviation inside the cluster can't be decreased while transferring any vector from one cluster to another. Algorithm consists of the following steps:

1. On the base of specified ratio of pure and mix vectors the dividing the vectors to pure and mix ones takes place. With this purpose at first the gradient image is calculated for the original set of vectors of measurements with the use of multidimensional Roberts operator while gradient histogram is being built. Proceeding from the specified percentage of the number of mix vectors, the threshold is defined according to histogram, which divides the vectors to pure and mix.

2. The merging of pure vectors to connective components. On this step all pure vectors merge into connective components, which are numerated sequentially. Corresponding algorithm, which is conceptually close to the algorithm of filling of areas with arbitrary boundary according to the criterion of connectivity [3], can detect and label any number of multi-connective areas simultaneously without any limits of form and width of their boundaries. For each connective component the vector of means is calculated.

3. Iterative clusterization of means vectors. The initial centers of clusters are defined according the following scheme. As the first two centers the pair of vectors is taken, which are the farthest from each other. Then the whole sample is divided to clusters according to the criterion of closeness to the chosen centers. In each cluster the vector, which is the farthest from center is found. For all such vectors the sum distance to all centers is calculated. As a new center the vector is taken, for which the sum distance is maximal and the procedure of vector distribution to the clusters is repeated

4. Connective components distribution to clusters. On this step connective components are labeled with numbers. A new number is labeled to a component according to a number of the cluster to which the mean vector of this component has fell.

5. Mix vectors clusterization. On finish step non iterative clusterization of mix vectors is developed according to minimum distance from clusters center $C_1,...,C_k$. Mix vector $z$ would be referred to the nearest cluster $\omega_i$, if $\|C_i - z\| < 0.5\max\|C_i - C_l\|,\ i,l = 1,...,k,\ i \neq l.$

The figures 1 – 3 illustrate the results of clusterization by method $K$-means. On fig. 1 presents the fragment of 3-band image received from the satellite Meteor-3M (scanner MSU-E, resolution 40m) 26.07.2003. Figure 2 illustrates the result of the algorithm for the following input data: the number of clusters for detection is 10, the relation of the numbers of mix and pure vectors is 0.35. Figure 3 shows the detection of the same number of clusters, but the relation of the number of mix and pure vectors is 0.25. Substantial differences in figures 2 and 3 show that reliability of the results of clusterization can be evaluated comparing some variants of processing.

The base of the second approach is the assumption that original data are the sample from distribution multimodes law, and the vectors, being in agreement with the separate mode, form the cluster. So the problem is reduced to the analysis of multidimensional histograms modes.

We have still assumed that the dimension of vectors is equal to $N$. The histogram is generated by sequential survey of data vectors and the comparison of each vector with current vectors list. As the result of generation, the corresponding frequency value changes or a vector is added to the list. To compute vectors addresses in the list hash coding is used. The key is calculated on the base of ordering in defined succession of bits of binary representation of the vector features $X$.

The first step of modal analysis is the search for the nearest neighborhoods of the vector from the list among other vectors.

As determined, the vector $X$ is the nearest neighborhood of the vector $Y$, if $|X_i - Y_i| \leq 1$ for $i = 1,\ldots,N$. Each of the possible nearest neighborhoods of the vector can be received from the given vector plus some vector of the shift, which components take the values from the set $\{-1,0,1\}$.

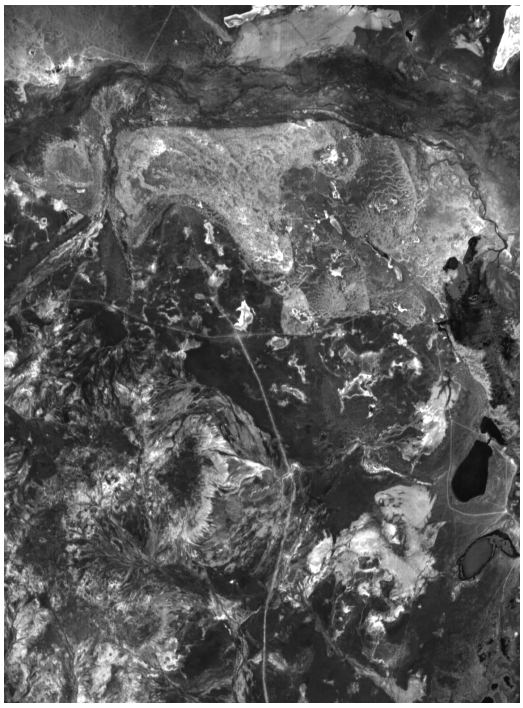

Fig. 1

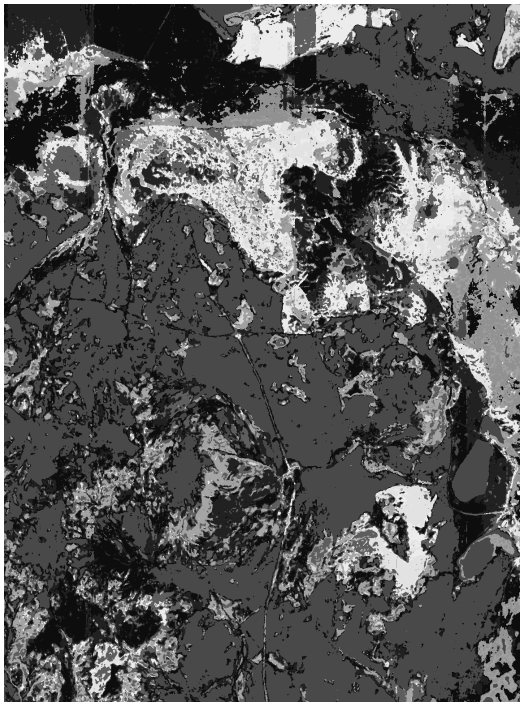

Fig. 2

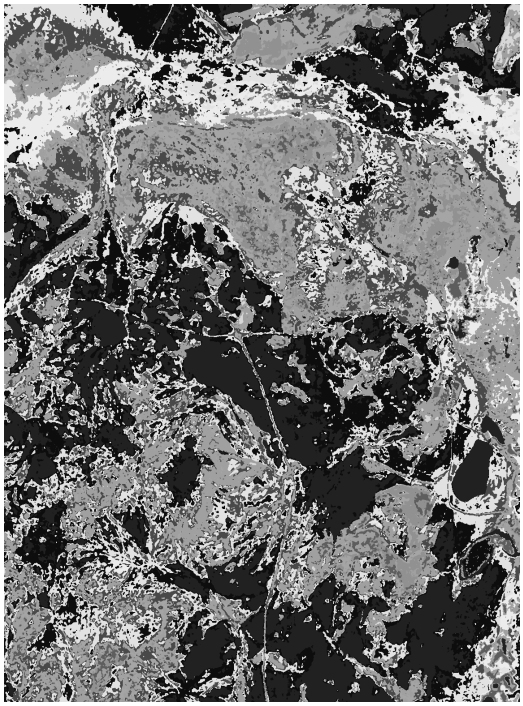

Fig. 3

Algorithmically $i$-th vector of the shift, $i = 1,2,...,3^N - 1,$ can be received if we decrease by 1 any of the coefficients of decomposition of the number $i$ in triple notation. As in real histogram not all the nearest neighbourhoods are present, for the efficient search the vectors are preliminary ordered into multidimensional binary trees. In this case the time of searching for the nearest neighbourhoods of the specified vector is proportional to the numbers of the neighbourhoods, which are really present. The process of building of multidimensional binary tree is as follows. The vectors $X$ are considered as $N$-mention keys. First the dispersions are calculated by all the coordinates and the coordinate $j$ is chosen which has the maximal dispersion. Median value of the sample by the given coordinate is used as a key to divide the vectors to two subset: the first subset – the vectors which value by the coordinate $j$ is not larger than the value of the threshold, the second – the vectors which value by the coordinate $j$ is larger than the threshold. Each received subset is divided further in this way.

On the second step of histogram analysis its smoothing is carried out in the way of substitution of the frequency $h(X)$ of the given vector to the mean value of the frequencies of its nearest neighborhoods or in the way of decreasing of data vectors digits (shift to the right of data bites to the corresponding number of digits). The smoothing is carried out only for the vectors, whose frequencies are smaller than the mean frequent.

Further the localization of histogram modes is carried out. In the beginning each vector on the base of analysis of its nearest neighbourhoods is put in correspondence the gradient. The number of vector with maximum value of the gradient is assigned to the vector. If the gradient is less than zero, that is the vector coordinates are the center of mode and vector own number is assigned to it. Finally, each histogram mode

corresponds oriented graph, which root corresponds to the point of mode. If the number of generated clusters (the number of histogram local maximums) is larger than the specified number, the histogram smoothing is carried out. At final step the oriented graph is colored with one color, i.e. all graph vertices are assigned the value applied to its root.

Let's note that the system of recognition and cluster analysis described here is the part of joint developed software on ERS data processing of ICMMG SB RAS and SRC "Planeta". This software is the single technological environment, having practically full set of operations proposed in [4] for ERS data processing.

The work was carried out partly with financial sponsoring of Integration project RAS (project N 13.14).

## References

[1]. V.V. Asmus. Software hardware complex of satellite data processing and its application for solving the hydrology and environment monitoring problems. Thesis for the Doktor Nauk degree . Moscow – 2002. (in Russian).
[2]. Methods of computer image processing. /Edited by Soifer V.A. Moscow, Fizmatlit, 2001 (in Russian).
[3]. Theo Pavlidis. Algorithms for graphics and image processing. 1982 Computer Science Press, Inc.
[4]. Remote Sensing: The Quantitative Approach. Edited by P.H. Swain and S.M. Davis. USA, McGraw-Hill, Inc., 396 p.

## About Authors

**Asmus V.V**. is the Head of Scientific Research Center of SpaceHydrometeorology "Planeta", ROSHYDROMET. Phone: (8095)255-69-14, e-mail: asmus@planet.ittp.ru

**Buchnev A.A.** is Senior Researcher at the Institute of Computational Mathematics and Mathematical Geophysics of SB RAS. Phone: (83832)307-332, e-mail: baa@ooi.sscc.ru

**Pyatkin V.P.** is the Head of laboratory at the Institute of Computational Mathematics and Mathematical Geophysics of SB RAS. Phone: (83832)307-332, e-mail: pvp@ooi.sscc.ru