Feature Selection Methods for Remote Sensing Images Classification

E. Goncharova[1], A. Gaidel[1,2]

[1]Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia
[2]Image Processing Systems Institute – Branch of the Federal Scientific Research Centre "Crystallography and Photonics" of Russian Academy of Sciences, 151 Molodogvardeyskaya st., 443001, Samara, Russia

Abstract

Different methods of feature selection are used to improve the performance of remote sensing images classification. In this work two methods of feature selection are examined. The first one is based on the discriminant analysis, and the second one rests on building the regression model. Histogram and textural features are considered as characteristics of an image. The experiments on the remote sensing dataset UC Merced Land Use show the effectiveness of these methods. As the result, the largest fraction of correctly classified images accounts for the 95%. Dimension of the initial feature space consisting of 18 features has been reduced to 3 features.

Keywords: Feature selection; classification; remote sensing images; discriminant analysis; regression analysis

1. Introduction

Remote sensing images are a huge storage of data, which have become readily available lately. The analysis of such images allows us not only to enrich human's knowledge about the Earth but also to solve large number of applied problems. For example, to control the cultivation of croplands, trace the spread of crop pests, prevent forest fires, etc. To solve the outlined problems the high-level and effective methods of image processing should be developed.

The dimension reduction, or feature selection, is a crucial step in performing the classification task. This fact may be explained by the following reasons.

1. An image is described by various features, however their extraction requires large amount of resources. The more features are extracted, the more challenging the task is. Therefore, choosing the most informative features makes the classification cheaper and faster.
2. Each feature influences the object discrimination differently. Moreover, the classifier is not ideal, therefore it includes some error, which depends on the quality of feature space. Thus, uninformative and noise descriptors may complicate the process of building a prediction model.

There are a large number of feature extraction methods, which guarantee good performance. For instance, in [1] the combination of various descriptors was used to divide images into 19 classes. The mean portion of the correctly classified objects was 93.6%, in some classes it peaked at 100%. The problem of reducing the number of features for the purpose of pattern recognition was investigated in [2]. The feature space included several hundred thousand characteristics (pixels of the initial images), and its dimension was reduced to several dozens of features.

Various approaches for feature selection are widely used in the analysis of biomedical images. In [3] the group of 5 significant features was extracted from the set of 169 properties, which characterize the progress of the chronic obstructive pulmonary disease (COPD). The classification error rate of 0.11 was obtained using this reduced feature space.

In this work it is proposed to examine the histogram and textural features. The images for classification were received from the available UC Merced Land Use dataset, including aerial optical images, belonging to different classes (agricultural field, forest, beach, etc.). The two approaches of feature selection were proposed. The former was based on the discriminant analysis, the latter – on the regression model. To assess the performance of the proposed methods the nearest neighbor algorithm of object classification was applied.

2. The object of the study

The object of the study is the set of features, characterizing an image, and methods of selection the most informative subset of features, which has the strongest discriminatory power.

The histogram and textural image characteristics and a degree of their influence on the performance of dividing images into two classes are analyzed.

The first method of feature selection is based on the maximization of the discriminant analysis criterion and a greedy strategy of adding a feature to the informative subset. In the second method we propose to assess the importance of a feature according to its coefficient in the regression model. The greedy strategy of removing a feature with the minimal coefficient from the informative subset is used in the implementation of this method.

The set of image characteristics that should be considered to get accurate classification results was extracted via the use of these two methods. The k-nearest neighbors algorithm was implemented to perform the classification task.

3. Methods

3.1. Feature extraction

An image is represented by its intensity matrix $I^{(M\times N)}$, where $M \times N$ is an image size. The intensity of each pixel of image (RGB color space) is defined as follows:

$$I(m,n) = \frac{R(m,n) + G(m,n) + B(m,n)}{3},\ m = \overline{1,M},\ n = \overline{1,N},$$

where R, G, B is an intensity of red, green, and blue component of the image resolution cell having coordinates (m,n) respectively.

$I(m,n)$ ranges in value from 0 to $L-1$, where L is a maximum gray level.

There are a large number of different features, which can characterize an image. In this work we use the histogram features that describe the spatial distribution of gray values. If the discrete image is considered as a two-dimensional stochastic process, we can estimate its spatial distribution of gray values and, therefore, raw (2) and central moments (3).

$$\nu_k = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} I^k(i,j). \quad (2)$$

$$\mu_k = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(I(i,j) - \nu_1\right)^k. \quad (3)$$

The calculated features are:

- mean intensity:

  $\overline{I} = \nu_1$, and also (I_R, I_G, I_B – mean intensity of red, green, and blue component respectively);

- second raw moment (mean energy):

  $s = \nu_2$;

- standard deviation:

  $\sigma = \sqrt{\mu_2}$;

- skewness:

  $\gamma_1 = \frac{\mu_3}{\sigma^3}$;

- kurtosis (a measure of the "tailedness" of the probability distribution):

  $\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$.

The autocorrelation matrix (4) describes dependence among the pixels of an image [4].

$$R(m,n) = \frac{\frac{1}{(M-|m|)(N-|n|)}\sum_i\sum_j I(i,j)I(i+m,j+n)}{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} I^2(i,j)}. \quad (4)$$

Two textural features are presented by the average of four values of the function (4) for two distances:

- $r_1 = \frac{1}{4}\left(R(0,-1) + R(0,1) + R(1,0) + R(-1,0)\right)$;
- $r_5 = \frac{1}{4}\left(R(0,-5) + R(0,5) + R(5,0) + R(-5,0)\right)$.

Another type of textural characteristics is the widely known Haralick's features. Let $P_{d_1\theta}(i,j)$ be a frequency with which two pixels of image, separated by distance d_1 in direction θ, occur on the image with the intensity i and j respectively. Then the gray-level spatial dependence matrix can be build according to the following rule [5]:

$$P_{d_1 d_2}(i,j) = \{(m,n) \in \{1,2,...,M\} \times \{1,2,...,N\} \mid I(m,n) = i, I(m+d_1, n+d_2) = j\},\ i,j = \overline{0, L-1}.$$

Textural features are extracted from the spatial dependence matrices, which are calculated for eight different distances (d_1, d_2): $(1,0)$, $(0,1)$, $(1,\pm1)$, $(2,0)$, $(0,2)$, $(2,\pm2)$. To get the invariant under rotation features, they are extracted from the average matrices. Thus, eight more textural features can be defined as follows:

- angular second moment:
$$f_1 = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} \left(\frac{|P(i,j)|}{R} \right)^2;$$
- contrast:
$$f_2 = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} (i-j)^2 \frac{|P(i,j)|}{R};$$
- entropy:
$$f_3 = -\sum_{i=0}^{L-1} \sum_{j=0}^{L-1} \frac{|P(i,j)|}{R} \log_2 \left(\frac{|P(i,j)|}{R} \right);$$
- correlation:
$$f_4 = \frac{\sum_{i=0}^{L-1} \sum_{j=0}^{L-1} ij \frac{|P(i,j)|}{R} - M_X M_Y}{\sqrt{D_X D_Y}},$$

where $P(i,j)$ – an element of averaged over the four dimensions $(1,0),(0,1),(1,\pm1)$ and $((2,0),(0,2),(2,\pm2))$.

R – a number of neighboring pixel pairs;

M_X, M_Y – the row and column means;

D_X, D_Y – the row and column variance.

3.2. Feature selection methods

Let $\mathbf{\Omega}$ be a set of objects for recognition. In this work a feature vector $\boldsymbol{x}_k \subseteq \boldsymbol{R}^K$, where K is a number of features, is considered as the element of this set. The set is divided into two classes $\mathbf{\Delta} = \{\mathbf{\Omega}_j\}_{j=1}^2$ with the following properties:

1) $\mathbf{\Omega}_0 \cup \mathbf{\Omega}_1 = \mathbf{\Omega}$;
2) $\mathbf{\Omega}_0 \cap \mathbf{\Omega}_1 = \varnothing$.

Let $\Phi(\boldsymbol{x}_k) : \mathbf{\Omega} \to \mathbf{\Delta}$ be the ideal operator that puts an object in correspondence with its class. As long as the ideal operator is unknown, another operator $\Phi(\boldsymbol{x}_k) : \mathbf{\Omega} \to \mathbf{\Delta}$ can be created. $\Phi(\boldsymbol{x}_k)$ tries to predict a class of input object, according to the information got from a training set of data $U \subseteq \mathbf{\Omega}$, in which the outcome of object is observable.

As the features can be measured in varied units, firstly, they should be standardized to get zero mean and unit variance. For this purpose the expected value:

$$M(i) = \frac{1}{|U|} \sum_{k=1}^{|U|} x_k(i),\ i = \overline{1,K},\ M \in \mathbf{R}^K$$

and variance:

$$R(i,i) = \frac{1}{|U|} \sum_{k=1}^{|U|} (x_k(i) - M(i))^2,\ i = \overline{1,K},\ R \in \mathbf{R}^{K \times K}$$

should be estimated for each feature.

Therefore, the feature vectors can be standardized by applying the formula (5).

$$x_k(i) = \frac{x_k(i) - M(i)}{\sqrt{R(i,i)}},\ k = \overline{1,|U|},\ i = \overline{1,K}. \quad (5)$$

To extract the subset of informative features two methods were examined. The former belongs to the discriminant analysis theory. According to this method, we choose the set of features that provides the largest value of the criterion $J(Q)$ [6]:

$$J(Q) = \frac{\operatorname{tr} R}{\sum_{j=1}^{2} \mathrm{P}(\Omega_j) \operatorname{tr} R_j},$$

where Q – current set of features;

R – mixture covariance matrix;

R_j – within-class covariance matrix;

$\mathrm{P}(\boldsymbol{\Omega}_j)$ – prior probability of class $\boldsymbol{\Omega}_j$, there $\mathrm{P}(\boldsymbol{\Omega}_j) = \frac{1}{2}$.

Thus, the stronger the scattering between two classes exceeds the average within-class scattering, the better selected set of features is.

To form the set of the most informative descriptors a greedy strategy of adding a feature was applied. Let the initial feature set be empty – $Q_{(0)} = \varnothing$. In step i we consider all the sets, like $Q_{(i,j)} = Q_{(i-1)} \cup \{j\}$, and calculate the criterion $J_{i,j} = J\left(Q_{(i,j)}\right)$. Then choose the set that maximizes the criterion:

$$Q_{(i)} = Q_{(i-1)} \cup \left\{ \underset{j \in [1;K] \cap \mathbf{Z} \setminus \mathbf{Q}_{(i-1)}}{\arg\max} J_{i,j} \right\} = Q_{(i-1)} \cup \left\{ \underset{j \in [1;K] \cap \mathbf{Z} \setminus \mathbf{Q}_{(i-1)}}{\arg\max} J\left(Q_{(i-1)} \cup \{j\}\right) \right\}.$$

These steps are iterated until a required number of features are obtained.

The second approach is based on the regression analysis. The regression analysis estimates the relationships among the dependent variable and one, or more, independent variables.

We propose that the number of class, which $\boldsymbol{x}_k$ can belongs to, is an independent variable $y(\boldsymbol{x}_k)$. This implies that the feature vector $\boldsymbol{x}_k$ influences $y(\boldsymbol{x}_k)$, and the regression model (6) can be built as follows:

$$y = X\theta + \xi, \tag{6}$$

where $y = (y_1 \quad y_2 \quad \ldots \quad y_n)^T$ – output vector;

X – feature matrix;

$\theta = \left(\theta_0 \quad \theta_1 \quad \ldots \quad \theta_{|Q|}\right)^T$ – regression weights;

$\xi = (\xi_1 \quad \xi_2 \quad \ldots \quad \xi_n)^T$ – error vector.

The unknown coefficients belonging to the vector θ are determined from the training set data via the ordinary least squares method:

$$(y - X\theta)^T (y - X\theta) \to \min_{\theta}.$$

The value of each feature is directly related to its weight in the regression equation (6). According to this proposal, the greedy strategy of removing a feature can be applied to forming the set of the informative descriptors.

Let the initial feature set $Q_{(0)} = Q$ contain all the analyzed features. In each step i the linear regression model $y_{(i)} = X_{(i)}\theta_{(i)}$ is built in the corresponding feature space. Then a feature with the minimal coefficient is removed from the set according to the following rule:

$$Q_{(i+1)} = Q_{(i)} \setminus \left\{ \underset{j \in [1;K] \cap \mathbf{Z} \cap Q_{(i)}}{\arg\min} \left|\theta_{(i)}(j)\right| \right\}.$$

As in the previous case these steps are iterated until a required number of features are obtained.

To estimate the classification power of the obtained feature subsets the nearest-neighbor classification is carried out. The Euclidean distance in feature space is defined as follows:

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{K} \left(x(i) - y(i)\right)^2}.$$

The classifier assigns the class of the vector $\boldsymbol{x}$ to the class of its closest point in the training set. In terms of the computational complexity, this method is rather simple in comparison with others. Since this classifier is memory-based, if the number of objects in the training set becomes large, this computational requirement may become excessive. The nearest-neighbor misclassification rate is no more than twice larger than the Bayes error rate [7].

The nearest-neighbor error rate is assessed as follows:

$$\varepsilon = \frac{\left|\left\{\boldsymbol{x}_k \in \mathbf{U} \mid \Phi\left(\boldsymbol{x}_k\right) \neq \Phi\left(\boldsymbol{x}_k\right)\right\}\right|}{\left|\mathbf{U}\right|}, \; k = \overline{1, \left|\mathbf{U}\right|},$$

where $\left|\mathbf{U}\right|$ – test set.

4. Results and Discussion

To assess the performance of the proposed approaches two image sets from the remote-sensing UC Merced Land Use dataset were used. This dataset includes aerial optical images, belonging to different classes (agricultural field, forest, beach, etc.), 100 for each class. Each image measures 256×256 pixels (RGB color space). There are two classes of images (agricultural fields and forest) being examined in this work. Figure 1 illustrates sample images belonging to the two classes.

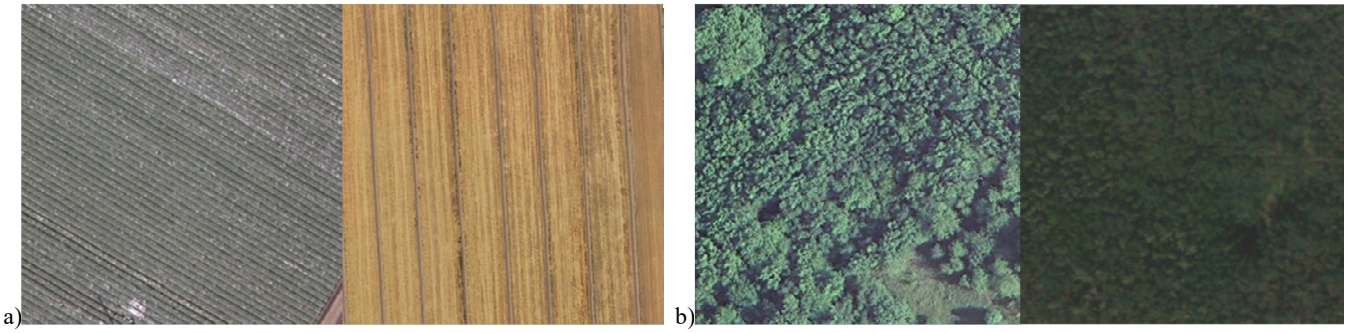

Fig.1. Sample images from UC Merced Land Use dataset (a – agricultural field, b - forest).

To carry out the experiments we used 5-fold cross-validation. The results obtained with the discriminant and regression analysis methods are shown in tables 1 and 2 respectively.

Table 1. Groups of the first 8 informative features selected with the discriminant analysis.

Features	ε
I_R	0.5
$I_R, \overline{I}$	0.075
$I_R, \overline{I}, s$	0.05
$I_R, \overline{I}, s, I_G$	0.075
$I_R, \overline{I}, s, I_G, I_B$	0.225
$I_R, \overline{I}, s, I_G, I_B, r_1$	0.175
$I_R, \overline{I}, s, I_G, I_B, r_1, r_5$	0.175
$I_R, \overline{I}, s, I_G, I_B, r_1, r_5, \gamma_1$	0.2

Table 2. Groups of the first 8 informative features selected with the regression analysis.

Features	ε
I_R	0.5
I_R, I_G	0.075
$I_R, I_G, \overline{I}$	0.2
$I_R, I_G, \overline{I}, I_B$	0.175
$I_R, I_G, \overline{I}, I_B, r_5$	0.075
$I_R, I_G, \overline{I}, I_B, r_5, r_1$	0.1
$I_R, I_G, \overline{I}, I_B, r_5, r_1, s$	0.1
$I_R, I_G, \overline{I}, I_B, r_5, r_1, s, f_{22}$	0.275

Table 3 shows a so called confusion matrix for the group of three features, extracted by the discriminant analysis method and performed best on this task. Table rows show the real classes of objects, while the columns indicate the predicted ones. The fraction of objects that were predicted correctly is represented by the diagonal cells.

Having analyzed the results, we can conclude that the discriminant analysis method performed best on this classification task. The lowest classification error rate of 0.05 was achieved in three-dimensional feature space, consisting of $I_R, \overline{I}, s$. The studied textural features have no significant effect on the quality of this classification. The inclusion of more textural characteristics, considering the correlation of features on various distances, may provide a better performance of this feature group.

Table 3. Confusion matrix.

	Predicted class		
True class	agricultural	forest	
agricultural	100%	0%	
forest	10%	90%	
			95%

5. Conclusion

Thus, for the task of the remote sensing images classification the subset of informative features was extracted. On the images from the UC Merced Land Use dataset, the histogram features produced the best outcome. It should be mentioned that the images were represented in RGB color space; hence the mean intensity of these three components appeared to have considerable impact on the discriminatory power.

The feature vector, selected with the discriminant analysis method, produced the best classification performance (using the nearest-neighbor classification method) on the images from the UC Merced Land Use dataset. The minimal classification error rate made up 0.05, therefore the proportion of the correctly classified images was 95%. This rate was achieved in the reduced three-dimensional feature space, consisting of the descriptors $I_R, \overline{I}, s$.

Thus, applying the feature selection methods leads to improving the image classification performance. In this study, the combination of three of the 18 initial descriptors appeared to be informative, while the other features increased the misclassification rate.

The method based on the discriminant analysis criterion provided good results and can be applied to fulfill the task of feature selection. Overall, in the future work we are interested in considering more features, which can characterize an image, and multiclass classification that can enable us to get more universal results.

Acknowledgements

The work was partially supported by the Russian Foundation of Basic Research (grant 16-41-630761 p_a), the Russian Federation Ministry of Education and Science as a part of Samara University's competitiveness enhancement program in 2013-2020 and the RAS based research program “Bioinformatics, modern information technologies and mathematical methods in medicine”.

References

[1] Guofeng Sheng, Wen Yang, Tao Xu, Hong Sun. Guofeng Sheng. High-resolutionsatellite scene classification using a sparse coding based multiple featurecombination. International Journal of Remote Sensing 2012; 33(8): 2395–2412.
[2] Glumov NI, Myasnikov EV. Method of the informative features selection on the digital images. Computer Optics 2007; 31(3): 73–76. (in Russian)
[3] Gaidel AV, Zelter PM, Kapishnikov AV, Khramov AG. Computed tomography texture analysis capabilities in diagnosing a chronic obstructive pulmonary disease. Computer Optics 2014; 38(4): 843–850.
[4] Gaidel AV, Pervushkin SS. Research of the textural features for the bony tissue diseases diagnostics using the roentgenograms. Computer Optics 2013; 37(1): 113–119. (in Russian)
[5] Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics 1973; 3: 610–621.
[6] Goncharova EF, Gaidel AV, Khramov AG. Statistical study of the factors affecting the cardiovascular disease. Information Technology and Nanotechnology 2016;1020–1025. (in Russian)
[7] Fukunaga K. Introduction to statistical pattern recognition . San Diego: Academic Press, 1990; 592 p.