Geological Units Classification of Multispectral Images by Using Support Vector Machines

Miloš Kovačević, Branislav Bajat
Faculty of Civil Engineering
University of Belgrade
Belgrade, Serbia
e-mail: milos@grf.bg.ac.rs bajat@grf.bg.ac.rs

Branislav Trivić, Radmila Pavlović
Faculty of Mining and Geology
University of Belgrade
Belgrade, Serbia
e-mail: trivic@rgf.bg.ac.rs prada@rgf.bg.ac.rs

Abstract—Quantitative techniques for spatial prediction and classification in geological survey are developing rapidly. The recent applications of machine learning techniques confirm possibilities of their application in this field of research. The paper introduces Support Vector Machines, a method derived from recent achievements in the statistical learning theory, in classification of geological units based on the source of the Landsat multispectral images. The initial experiments suggest the usefulness of the proposed classification approach.

Keywords—image classification; Landsat; multispectral images; support vector machines

I. INTRODUCTION

Satellite images offer important advantages when compared to other methods of data gathering. Therefore, they are used in a wide range of applications in geosciences. The availability of such kind of data in digital form and the development of computer technology and image analysis software supports the integration of remote sensing data and geoinformation systems (GIS). At the same time, the improved computational capabilities and efficiency resulted in the increased usage of sophisticated statistical and machine learning methods, in a wide variety of environmental sciences.

The paper introduces Support Vector Machines (SVM), a recent method in statistical learning theory, used to recognize and classify geological units based on the Landsat multispectral imagery. Nearly all of researches related to SVM in remote sensing technologies are focused on hyperspectral image classification [1], [2].

Many applications in environmental studies use also SVM jointly with other contemporary techniques, like cellular automata in the spatial simulating of land use changes [3] or fuzzy K-means in the image pattern recognition in precise farming [4]. Most machine learning methods which are used in geoinformatics and environmental sciences proved to act as a support to contemporary acquisition technologies like remote sensing [5], [6]. All those applications are characterized with a huge number of available input data.

II. MATERIAL AND METHODS

A. Problem Statement

Let \( S \) be a set of all possible samples (pixels) covering some geographical area given in the following form:

\[
S = \{ x \mid x \in \mathbb{R}^n \}
\]

Each sample is represented as an \( n \)-dimensional real vector and every particular coordinate \( x_i \) represents a pixel digital number for separate band of a satellite image. First, we define the task of classification of pixels into geological units. Let \( C = \{ c_1, c_2, \ldots, c_l \} \) be the set of \( l \) classes that correspond to some predefined geological units such as Quaternary or Triass. The function \( f_c : S \rightarrow C \), is called a classification if for each \( x_i \in S \) it holds that

\[
f_c(x_i) = c_j \quad \text{if } x \text{ belongs to the class } c_j.
\]

In practice, one only has a limited set of \( m \) labeled examples \( (x_i, y_i) \), \( x_i \in \mathbb{R}^n, y_i \in C, \quad i = 1, \ldots, m \). Labeled examples form a training set for the classification problem at hand. The machine learning approach tries to find the function \( f_c \), which is a good approximation of the real, unknown function \( f_c \), using only the examples from the training set and a specific learning method such as Artificial Neural Networks (ANN) or Decision Trees (DT) [7].

B. Brief overview of SVM classification

Support Vector Machines method is a recent approach in pattern classification and it deals with binary classification model [8]. Binary model assumes that a pixel belongs to one class only and that there are just two classes \( (C = \{ c_1, c_2 \}) \). Each classification task with \( n \) classes can be modeled as a sequence of \( \binom{n}{2} \) binary tasks using the one-vs-one approach in which one trains \( n^*(n-1)/2 \) binary classifiers, one for each pair of classes. The final decision is made by voting i.e. the most frequently predicted class is selected as the output. Let \( (x_i, y_i), x_i \in \mathbb{R}^n, y_i \in \{-1,1\}, \quad i = 1, \ldots, m \) be the training...
set (-1 stands for class $c_1$ and 1 for $c_2$). Fig. 1 is used to illustrate the basic idea of SVM classification. White and grey squares represent samples from a training set comprised of two distinct classes.

Let us assume for a moment that classes are linearly separable, and neglect the circled examples in Fig. 1. During the learning phase one seeks the separating hyper-plane which best separates the examples of two classes. Let $h_1: w \cdot x + b = 1$ (where "." denotes the dot product) and $h_{-1}: w \cdot x + b = -1$, $w, x \in \mathbb{R}^n$, $b \in \mathbb{R}$, be possible hyper-planes, with all the white examples lying above $h_1$ ($y_i = 1$) and all the grey examples lying below $h_{-1}$. ($y_i = -1$). Hence for all training examples $(x_i, y_i)$ it follows that:

$$y_i(w \cdot x_i + b) \geq 1, \quad i = 1, 2, \ldots, m$$

(1)

One chooses $h: w^* \cdot x + b^* = 0$ to be the best separating hyper-plane lying in the middle between the already fixed hyper-planes $h_1$ and $h_{-1}$. The notion of the best separation can be formulated to find the maximum margin $M$ that separates the data from both classes. Since the margin is equal to $\|w\|_2$, maximizing the margin is equal to minimizing the $\|w\|_2$. The best separating hyper-plane can now be found by solving the following nonlinear convex programming problem (for solving of optimization problem see [9]): find $w, b$ so that:

$$\min_{w, b} \frac{1}{2} \|w\|^2$$

w.r.t. $1 - y_i (w \cdot x_i + b) \leq 0, \quad i = 1, 2, \ldots, m$\n
(2)

In practical classification problems, examples are usually not linearly separable (circled examples from Fig. 1). Therefore, some additional positive slack variables $\epsilon_i$ are introduced, representing the distances of points on the wrong side of the separating hyper-plane (circled squares). The nonlinear convex program (2) now becomes:

$$\min_{w, b} \frac{1}{2} \|w\|^2 + C \sum_i \epsilon_i$$

w.r.t. $1 - y_i (w \cdot x_i + b) \leq 0, \quad -\epsilon_i \leq 0, \quad i = 1, 2, \ldots, m$\n
(3)

The parameter $C$ models the penalty for misclassified points in a training set. One wants to find a hyper-plane to minimize misclassification errors while maximizing the margin between classes. The optimization problem (3) is usually solved in its dual form and the solution is:

$$w^* = \sum_{i=1}^m \alpha_i y_i x_i, \quad C \geq \alpha_i \geq 0, \quad i = 1, \ldots, m$$

(4)

The solution $w^*$ for optimal hyper-plane is a linear combination of training examples. However, it can be shown that $w^*$ represents a linear combination of those vectors $x_i$ (support vectors) for which the corresponding $\alpha_i$ is a non-zero value. Support vectors for which $C > \alpha_i > 0$ holds belong either to $h_1$ or $h_{-1}$ (depending on $y_i$). Let $x_a$ and $x_b$ be two support vectors ($C > \alpha_a, \alpha_b > 0$) for which holds $y_a = 1$ and $y_b = -1$. Now $b^* = -\frac{1}{2}w^* \cdot (x_a + x_b)$ and finally the classification function becomes:

$$f(x) = \text{sgn}\left[\sum_{i=1}^m \alpha_i y_i (x_i \cdot x) + b^*\right]$$

(5)

In order to deal with non-linearity of the classification problem, the SVM approach goes one step further. One can define mapping of examples to a so-called feature space of very high dimension: $\phi: \mathbb{R}^n \rightarrow \mathbb{R}^d$, $n<<d$ i.e. $x \rightarrow \phi(x)$. The basic idea of this mapping into high dimensional space is to transform the non-linear case into the linear one as illustrated in Fig. 2, and then to use the above explained linear algorithm. In such a space, dot-product from (5) transforms into $\phi(x_i) \cdot \phi(x)$. It is recognized that there exists a class of functions called kernels [10] for which holds $k(x_i, x) = \phi(x_i) \cdot \phi(x)$. These functions represent the dot-products in some high dimensional spaces, but can be easily computed in the input space. Using kernels equation, (5) becomes:

$$f(x) = \text{sgn}\left[\sum_{i=1}^m \alpha_i y_i k(x_i \cdot x) + b^*\right]$$

(6)

In this paper, Gaussian $k(x, y) = \exp\left(-\gamma \|x - y\|^2\right)$ and linear kernel [no mapping, equation (5)] are used.
If removing of all training data which are not the support vector points but retraining the classifier, the same solution is reached once again. Hence, support vectors represent examples from the training set that best describe the classes. The ability to distinguish between the support vectors and the noisy data points enables SVM to increase its generalization capacity in the learning process. For detailed review of SVM for pattern classification please refer to [10].

C. Case study area

Within the Project “Realisation de la cart des ressources en eau souterraine du Nord de l’Algérie” most of the Northern Algeria was subjected to digital geological mapping. The investigation covers the area of approximately 300,000 km² in North Africa, between the Mediterranean Sea in the north, state borders with Tunisia and Morocco in the east and west respectively, and South Atlas Fault zone in the south. In geological sense, generally, investigation covers three different regions, namely, from north to south: Tell Atlas (l’Atlas tellien), High Plateau (Hauts-plateaux) and Saharan Atlas (l’Atlas saharien). The collected and analyzed geological documentation was validated and updated by application of remote sensing method, which provided creation of digital geological maps for the whole investigated area. The set of multispectral Landsat 7 ETM+ satellite images was used for this purpose.

Research area described in this paper covers small part of the terrain of rectangular shape (N 32°54’, E 0°02’-South West corner, N 33°24’, E 1°05’-Nort East corner) in the south part of Saharan Atlas, southwest of Brezina. Various kind of geological maps and documentation were available. Some parts of the terrain were covered by detailed geological maps of the scale 1:50000, while for other parts only simplified maps of the scale from 1:200000 up to 1:500000 were available. In general, available geological maps had distinctive scientific quality, as well as various levels of details in defined geological units and have been published in distinct period during the last century. In addition, maps of the same scales had different conceptual approaches (like stratigraphic and tectonic) or they had just the lithological approach. Different level of geological details induces many practical cartographic problems. In some cases logical trends of geological units in vicinity sheets, in areas directly

connected, were not recorded as the same unit. On the contrary, in certain locations units with different geological compounds were notified and presented on adjacent sheets, as a unique one.

This is why the application of remote sensing techniques provided integral geological approach, not only for the investigated area but, for the whole area included in this project. Cartographic discrepancies and spreading of geological units were resolved, as much as possible, on the prepared maps by the application of the adopted methodology. Based on the satellite image visual interpretation, as well as the analysis of image processing results, contact between the units and their spatial position was determined, whereas chrono-stratigraphy (age) was taken into account from the existing geological maps.

III. Experiment

The multispectral Landsat 7 ETM+ satellite images data used in this project was acquired on February, 2004. Although seven bands were available, only 5 spectral bands as well as panchromatic images were used in this study. The standard resolution of multispectral images is 30 m, and of the panchromatic ones is 15 m. The images are delivered in separated files that are in digital TIFF format in the UTM coordinate system, with eliminated radiometric errors. The natural color RGB composite model with standard deviation stretch type is presented in Fig. 3.

On digital multispectral satellite image data standard image processing was performed. Different procedures of spectral enhancement techniques were used [11], [12], like clay minerals, iron mineral and mineral composite indexes. At last, principal component analysis and correlation stretching gave very good basement for digital mineral mapping of this area [13].

Taking into account all these results, based on visual analysis of images and comparison with geological maps of various contents, very qualitative geological interpretation is provided.

Results are shown in digital geology map of the scale of 1:200000 which covers the whole researched area. The small part to the south of Saharan Atlas that is approximately about 5000 km² was covered with two case studies. Simplified geological map of this area is shown in Fig. 4.

The oldest rock in the selected area is Triassic in age. Mainly, it is build of evaporates and clay with intercalated dolomites. The rocks are overlain by middle and upper Jurassic sediments, dominantly, silty clay with intercalations of sandstone, limestone and marls. Mesozoic stratigraphic succession is ended with cretaceous limestone and dolomite. They are associated with marl, gypsum and arenites. In lower parts of Cenozoic formation, clays and conglomerates Mio-Pliocene in age and different type of Quaternary deposits are designated. In structural-tectonic sense Mesozoic sediments are gently folded with subhorizontal axes trending from the northeast to the southwest.
In the paper, two case study areas are presented (Fig. 4). The smaller (C1) is located in the left central part of the map. It is chosen because of relatively simple geological composition of the couple conformable units with dip direction to the southeast. In the second case in the central part of the map (C2), more geologically complicated relationship is presented.

A. Results

Input features for the SVM classification are given as pixel digital numbers of 5 bands (thermal infrared channel and panchromatic band are skipped), as well as, XY coordinates of pixels in the UTM projection. Label values of geological units are obtained from the digital thematic map (Fig. 4.). Different geological units are presented according to the symbology enclosed in map legend. The map (in the shape format) is transformed to the raster format with grid
cell size same as the image pixel dimensions of 28.5 m. Classes of geological units were assigned to each grid cell.

In order to test the quality of SVM classification methods we used standard performance measures: accuracy and kappa index. They are given with:

\[
acc = \frac{ncc}{N} \quad (7)
\]

\[
\kappa = \frac{(N \sum_{i} x_{ij} - \sum_{i} x_{i.} x_{.j}) / (N^2 - \sum_{i} x_{i.} x_{.j})}{N - \sum_{i} x_{i.} x_{.j}} \quad (8)
\]

In (7) \(N\) represents the total number and \(ncc\) the number of correctly classified test pixels. Kappa (\(\kappa\)) is considered to be an improvement over \(acc\). It is very popular in remote sensing community and is applied to raster maps. The best way to compute \(\kappa\) is to derive it from the confusion matrix, a cross table with \(n\) rows and \(n\) columns (\(n\) - number of classes) in which \(x_{ij}\) represents the number of pixels from the actual class \(j\) that are classified by SVM as the class \(i\).

In (8) \(x_{ij}\) and \(x_{i.}\) are the total number of observations in the row \(i\) and the column \(i\) of the cross-table, respectively. The idea of \(\kappa\) is to remove the effect of the random agreement between the two experts (here between SVM and the expert map). Based on [14] Kappa index values falling in range 0.61-0.81 are categorized as substantial and values higher than 0.81 are consider as almost perfect.

The input data for each case study area are divided in 3 data sets:

a) With all input features (pixel digital numbers of 5 bands + XY coordinates of pixels).

b) Data sets only with bands.

c) Only with XY positional coordinates.

Linear and Gaussian kernels are applied for all data sets. Various values of Gaussian kernel parameter \(\gamma\) and \(C\) are tested. The best results are obtained for \(\gamma = 100\) and \(C = 2\). Overall accuracy results of 2-fold cross validation (50% of pixels for training the SVM classifier and other 50% of pixels for testing) for both case study areas C1 (5 class units) and C2 (14 class units) are shown in tables I and II.

In all experiments the training set consists of pixels uniformly distributed across the case study area. The experiment with Gaussian kernel is repeated for the case study area C2, with the varied partitions of training pixels (0.1\%, 1\% and 5\% of the total number of available pixels) and the corresponding partitions of test pixels (99.9\%, 99\% and 95\% of the total number of available pixels).

Gaussian kernel was also applied on the overall area (20 geological units – classes) with 0.1\% of training pixels. The obtained results with \(\gamma = 1000\) and \(C = 2\) are shown in table IV.

### TABLE I. ACCURACY ESTIMATION OF THE CASE STUDY AREA C1

<table>
<thead>
<tr>
<th></th>
<th>Linear kernel</th>
<th>Gaussian kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bands</td>
<td>63%</td>
<td>68%</td>
</tr>
<tr>
<td>Coordinates</td>
<td>76%</td>
<td>97%</td>
</tr>
<tr>
<td>Coord.+bands</td>
<td>88%</td>
<td>97%</td>
</tr>
</tbody>
</table>

### TABLE II. ACCURACY ESTIMATION OF THE CASE STUDY AREA C2

<table>
<thead>
<tr>
<th></th>
<th>Linear kernel</th>
<th>Gaussian kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bands</td>
<td>45%</td>
<td>50%</td>
</tr>
<tr>
<td>Coordinates</td>
<td>42%</td>
<td>93%</td>
</tr>
<tr>
<td>Coord.+bands</td>
<td>84%</td>
<td>94%</td>
</tr>
</tbody>
</table>

### TABLE III. ACCURACY ESTIMATION AND KAPPA INDEXES OF THE CASE STUDY AREA C2

<table>
<thead>
<tr>
<th>Partitions of training pixels</th>
<th>0.1%</th>
<th>1%</th>
<th>5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bands</td>
<td>39%</td>
<td>48%</td>
<td>49%</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.28</td>
<td>0.38</td>
<td>0.39</td>
</tr>
<tr>
<td>Coordinates</td>
<td>68%</td>
<td>84%</td>
<td>89%</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.62</td>
<td>0.82</td>
<td>0.87</td>
</tr>
<tr>
<td>Coord.+bands</td>
<td>69%</td>
<td>84%</td>
<td>90%</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.6</td>
<td>0.81</td>
<td>0.89</td>
</tr>
</tbody>
</table>

### TABLE IV. ACCURACY ESTIMATION AND KAPPA INDEXES OF THE OVERALL AREA

<table>
<thead>
<tr>
<th></th>
<th>accuracy</th>
<th>Kappa index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bands</td>
<td>39%</td>
<td>0.25</td>
</tr>
<tr>
<td>Coordinates</td>
<td>48%</td>
<td>0.38</td>
</tr>
<tr>
<td>Coord.+bands</td>
<td>61%</td>
<td>0.54</td>
</tr>
</tbody>
</table>

### IV. DISCUSSIONS AND CONCLUSIONS

The results presented above, demonstrate high performance of the SVM technique for the geological units classification based on the source of multispectral Landsat images. Gaussian kernel performed better than linear in both case studies suggesting that the problem at hand is not linear by its nature. Results based just on the knowledge of locations (XY coordinates) showed that the distribution of the geological units can be predicted sufficiently well if one has a good quality training set. Based on the accuracy estimated by 2-fold cross validation, the results obtained by Gaussian kernel pointed out that the knowledge of pixel location has more influence on SVM classification than pixel digital numbers from bands. With reducing the number of examples in the training sets, both the accuracy and the Kappa index still remain high while preserving the significance of coordinates over the band data. Please note that the class labels from the training set are obtained using other expert methods that utilize the information from the bands.

However, when applying the method on the whole terrain, the information based on band values becomes more important.

Despite the small number of input features that could be provided by multispectral imagery, this study shows that SVM classification could be used for the classification of geological units based on the source of Landsat images.

### REFERENCES


