Automatic Feature Selection for Operational Scenarios of Satellite Image Understanding using Measures of Mutual Information

Dragos Bratasanu[1], Ion Nedelcu[1] and Mihai Datcu[2]

[1] Romanian Space Agency ROSA, Mendeleev 21-25, 010362, Bucharest, Romania
[2] DLR German Aerospace Center, Oberpfaffenhoffen D-82234 Wessling, Germany

Abstract. The Earth Observation processing tools operating in the recent scenario need to be tailored to the new products offered by the sub-meter spatial resolution imaging sensors. The new methods should provide the image analysts the essential automatic support to discover relevant information and identify significant elements in the image. We advocate an automatic technique to select the optimum number features used in classification, object detection and analysis of optical satellite images. Using measures of mutual information between the target classes and the available features, we investigate the criterions of maximum-relevance and maximum-relevance-minimumredundancy for automatic feature selection at very-low cost. Following a comprehensive set of experiments on multiple sensors, applications and classifiers, the results demonstrate the possible operational use of the method in future scenarios of human-machine interactions in support of Earth Observation technologies.

1 Introduction

The methods and ways users operate the Earth Observation (EO) satellite data in the present scenario are beginning to change the paradigms of classical image analysis. If in the past the existent automatic classification and segmentation tools provided good results for mapping of decameters resolution images, nowadays these tools fail to offer the users the necessary support in discovering relevant information in the image.

The old methods for knowledge-based image understanding were operating on two distinct levels: pixel level (e.g. classification techniques in which each pixel is assigned with a label) and region level (e.g. segmentation techniques in which homogeneous image regions are assigned with labels). The resolution of new optical sensors has reached values of centimeters (e.g. *GeoEye*-1 0.41m, *Quickbird* 0.6m and *World-View*-2 0.50m) and outran the capabilities of standard information mining tools to infer knowledge using spectral and spatial information. Future tools need to look at the way analysts understand the data and how the current manual operations are performed. We introduce an approach based on patch-level analysis, capturing *contextual information* – sub-meter resolution image areas interconnect complex structures covering many pixels with high diversity of spectral information. By using this approach we line up to the way users create cartographic products for multiple applications (e.g.

Bratasanu D., Nedelcu I. and Datcu M..
Automatic Feature Selection for Operational Scenarios of Satellite Image Understanding using Measures of Mutual Information.
DOI: 10.5220/0003143101090117
In *Proceedings of the International Workshop on Semantic Sensor Web* (SSW-2010), pages 109-117
ISBN: 978-989-8425-33-1

maps for emergency response, geo-intelligence, forensics). Maybe the most important step in all automated procedures is to identify the optimum set of attributes – *feature selection* - to minimize the classification error. Using the minimum-redundancy-maximum-relevance (mRMR) criterion [1] based on mutual information, we introduce a method to select an ideal set of features from the available set. Fig.1 shows the workflow of our procedure.

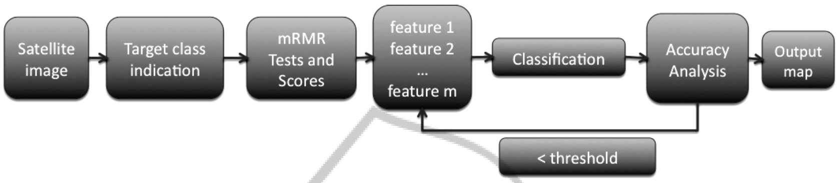


Fig. 1. Workflow for selecting the optimum features for target class classification.

We test the resulting attributes using four classifiers (Minimum Distance, Maximum Likelihood, Mahalanobis Distance and Latent Dirichlet Allocation (LDA) [2]). The minimum of the classification error yields the optimum feature set as a function of the operating classifier. The results confirm that the mRMR method applied to sub-meter resolution data improves classification accuracy for applications based on the new sensors.

2 Theoretical Approaches

2.1 Feature Selection using Statistical Measures of Mutual Information

In all information-mining applications, feature selection (characterizing attributes for a given class) is a critical step in minimizing the classification error. Having available a data set *D* described by *M* features $X = \{x_i, i = 1...M\}$ and the target class *C,* the problem is to discover a subspace of *m* features R^m in the feature space R^M that characterizes *C.* Of course, the score for each possible feature set needs to be related to an operational classification algorithm. The question that rises is '*what subset* R^m Ì R^M *is the optimum for the problem at hands?'* Since there are countless combinations of the existent attributes, we use an incremental method explained in the following paragraphs.

The optimal characterization condition most of the times implies minimal classification error for the target class, which in turn requires the maximal statistical dependency of the target class *C* on the data distribution in the subspace R^m. In literature, this condition is known as *maximal dependency.*

The maximal dependency approach is widely debated in information theory publications and the most familiar way to obtain it is to use *maximum relevance* (MR) criterion – selecting the features *m* with the highest relevance to the target class *C.* The relevance of features R^m to the features in *C* is defined in terms of mutual information. Given two random variables *x* and *y* with probability density functions *p*(*x*) and *p*(*y*), the following formula gives their measure of mutual information:

$$I(x,y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \quad (1)$$

The selected features x_i in R^m need to yield the largest mutual information $I(x_i, C)$ with the target class *C*. In the incremental search, the first *m* maximum scores of mutual information $I(x_i, C)$ yield the selection of the best *m* features.

The MR criterion gives a good start in the feature selection problem but it does not answer two critical questions: *'what is the optimum number of features m to select from the mutual-information $I(x_i, C)$ scores'* and '*are all these selected attributes useful to minimize the classification error?'* In information theory it is widely accepted that the combination of individually good features may lead to confusion in classification results if they have a high level of mutual redundancy. One approach to reduce redundancy amongst characteristics is the *minimum redundancy* (mR) criterion.

In [3] the authors introduce a *minimum-redundancy-maximal-relevance* (mRMR) framework to select the optimum number of features and minimize redundancy amongst them. We will describe this method in the following paragraphs.

2.2 Category Discovery using Maximum relevance Minimum Redundancy

When considering mutual information the reasoning for discovering the optimum set of attributes, the goal is to find a feature set *S* with *m* features $\{x_i\}$, which jointly have the largest dependency, on the target class *C* (2).

$$MaxDependency(S,C) = \max(I(x_i, i = 1...m), C) \quad (2)$$

Because maximum dependency is often hard to implement even for discrete random variables, maximum relevance criterion (MR) has been proposed as alternative in publications. The MR criterion approximates the dependency between multiple random variables by selecting the features approximating $\max(I(x_i, i = 1...m), C)$ with the mean value of all mutual information values between individual features x_i and class *C* (3).

$$\max D(S,C) = \max\frac{1}{|S|}\sum_{x_i,x_j \in S} I(x_i, C) \quad (3)$$

It is well known in remote sensing literature that usually features (e.g. spectral bands) may present high redundancy for a specific target class. When two or more variables have a rich content of mutual information, discriminating between the target class and the rest does not change if one feature is removed. In order to select mutually exclusive features, the criterion of minimum redundancy (mR) may be used following the MR. (4)

$$\min(R,(S)) = \min\left(\frac{1}{|S|^2}\sum_{x_i,x_j \in S} I(x_i, x_j)\right) \quad (4)$$

The criterion that combines the MR and the mR is called minimum-redundancy-max-relevance (mRMR) [3]. We choose the form (5) to combine the above formulas, where D is the max-relevance and R is the min-redundancy:

$$mRMR = \max(D - R) \tag{5}$$

In operational application we used an incremental version for selecting the feature set. If we have *m*-1 selected attributes, the task is to select the *m*-th feature that maximizes the mRMR condition (6):

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j, C) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j, x_i) \right] \tag{6}$$

2.3 Choosing the Optimal Feature Set

After determining the score for each feature with the mRMR tests, the remaining issue is to determine the optimum number of attributes *m.* To discover the feature set $S_m \subset S$ we follow the workflow in figure 2.

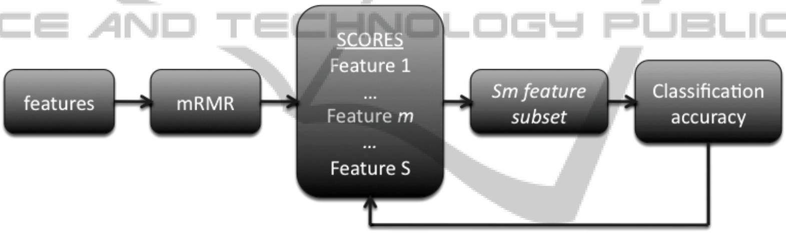


Fig. 2. Selecting the optimum feature set.

The process of selecting the optimum characteristics for a specific target class includes the following steps:

1. Determine the available features for the satellite image – spectral bands, textural information, etc
2. Choose a patch from the image representing the target class
3. Determine the score of the mRMR tests for each feature
4. Begin with the top 3 features S_3 and determine the classification accuracy for each adjacent features sets $S_3 \subset S_4 \subset S_5 ... \subset S_m$. The feature set that yields the lowest classification error $e = \min(e_k)$ gets selected to be the optimum one.

2.4 Multiple Classifiers

The mRMR feature selection scheme does not imply the use of a specific classifier. We have tested this approach on multiple supervised classifiers with *the same training set* and discovered that optimum attributes differ not only with respect to the target class but also with respect to the operational classifier. We used in our case stu-

dies the Minimum Distance, Mahalanobis Distance, Maximum Likelihood and Latent Dirichlet Allocation described in [2].
The *Minimum Distance* classifier discovers the classes of interest by the following rules [4]. Suppose m_i are the means for the *M classes* determined from the training data and *x* is the position of the pixel to be classified. Classification is performed on the basis of:

$$x\,\text{Î}\ w_i \ \text{ if } d(x,m_i)^2 < d(x,m_j)^2 \ \text{ for all } j\ ^1\ \ i \tag{7}$$

Maximum Likelihood classification works on the following principles [4]. Let the spectral classes for an image be represented by $w_i, i = 1...M$ with *M* the total number of classes. A pixel *x* is assigned to class w_i if:

$$p(w_i \mid x) > p(w_j \mid x) \ \text{ for all } j\ ^1\ \ i \tag{8}$$

$$p(w_i \mid x) = p(w_j \mid x) p(w_i) / p(x) \tag{9}$$

$P(\omega_i)$ is the *a priori* probability of class w_i to occur in the image.

The *Mahalanobis Distance* classifier assigns each pixel in the image to one of the training classes based on the following distance measure [4]:

$$d(x,m_i)^2 = (x - m_i)^t\, \text{å}^{-1} (x - m_i) \tag{10}$$

The *Latent Dirichlet Allocation based* classifier is thoroughly described in [2] and works by assigning each pixel and each patch in the image to a specific ‘latent’ topic generated from the training patches. For all classifiers, experiments and results are presented in the following section.

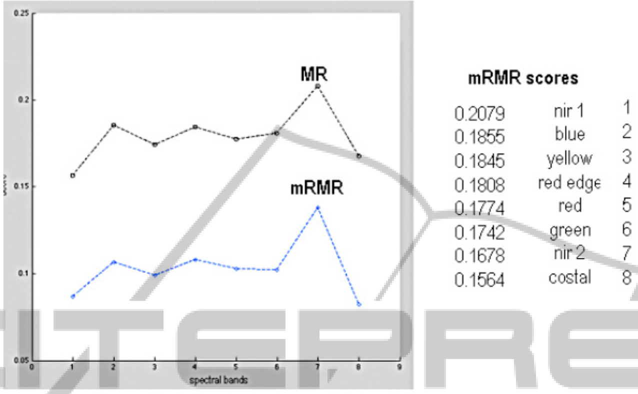
3 Operational Case Studies and Results

3.1 Urban Areas Extraction using WorldView-2 Satellite Image

In the first case study we aim to discover the optimum feature set and extract the urban areas from a WorldView-2 image (resolution of 0.5 meters / pixel). Figure 3 shows the test image and the training patch used to indicate the target class.

Fig. 3. WorldView-2 image and the target class *Urban areas.*

The first set of features contained only the 8 spectral bands provided by the sensor (costal, blue, green, yellow, red, red-edge, nir-1, nir-2). We computed the MR and mRMR values between all features in the data set and the target class and drew the score table as shown in figure 4.



mRMR scores

0.2079	nir 1	1
0.1855	blue	2
0.1845	yellow	3
0.1808	red edge	4
0.1774	red	5
0.1742	green	6
0.1678	nir 2	7
0.1564	costal	8

Fig. 4. The scores for the features in the data set and the results of mRMR criterion.

After evaluating the mRMR scores, we train the test classifiers with patches representing the target class and classes labeled as ‘others’. The first three top features (nir-1, blue, yellow) are tested first and then each feature in the scores table in added to the classifier. Figure 5 shows how each classifier operates better on a different set of input features, yielding different accuracy values (the y-axis) as function of the input features (the x-axis). The Maximum Likelihood classifier gives the best accuracy value (93%) with the first three features (nir-1, blue, yellow) and the target class mask is shown in figure 6.

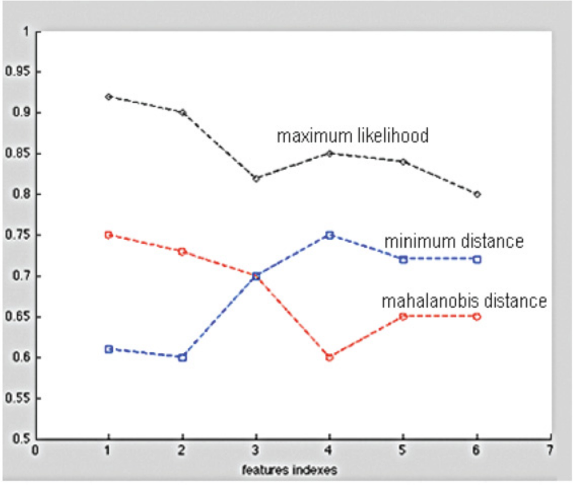


Fig. 5. Classification accuracy as a function of the input features and the classifier used.

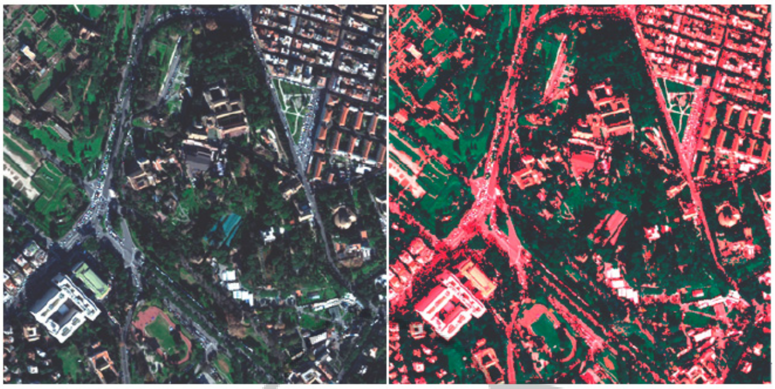
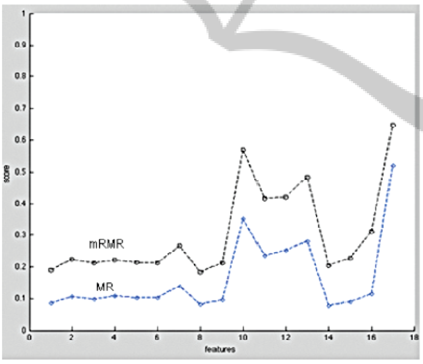


Fig. 6. WorldView-2 image and the target class extracted with Max Likelihood.

We added textural information to the first set of features and recomputed the MR and mRMR values between all features in the data set and the target class. The table score is presented in figure 7. We compute the classification accuracy for the first three top features and then add an extra attribute each step and re-perform the experiments. From the graphics in figure 8 we can easily understand that the best accuracy is obtained by using the first 7 features in the score table with Maximum Likelihood.



0.64	correlation	1
0.56	mean	2
0.48	contrast	3
0.41	homogeneity	4
0.41	variance	5
0.30	second moment	6
0.26	nir-1	7
0.22	entropy	8
0.22	blue	9
0.22	yellow	10
0.21	red	11
0.21	pan	12
0.21	green	13
0.21	red edge	14
0.20	dissimilarity	15
0.18	costal	16
0.18	nir-2	17

Fig. 7. MR and mRMR scores for the features in test.

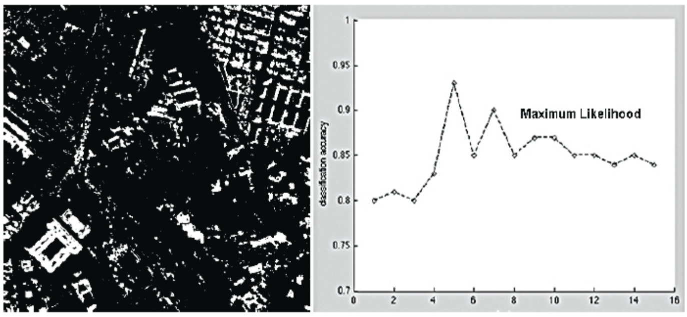


Fig. 8. Target class and the classification accuracy as a function of the features in test.

3.2 Urban Areas Extraction using GeoEye-1 Satellite Data

In the second case study we aim to discover the optimum feature set and extract the urban areas from a Geoeye-1 image (resolution of 0.41 meters / pixel). Figure 9 shows the test image and the training patch used to indicate the target class.

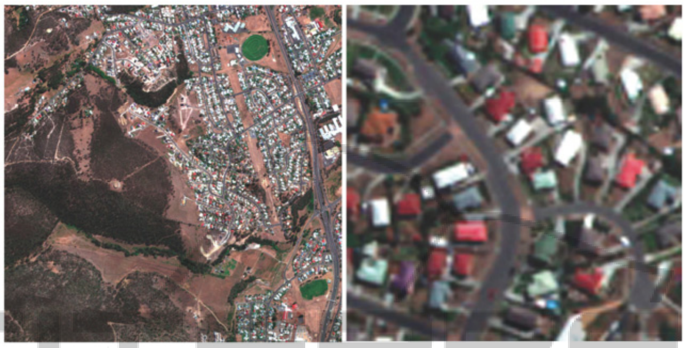
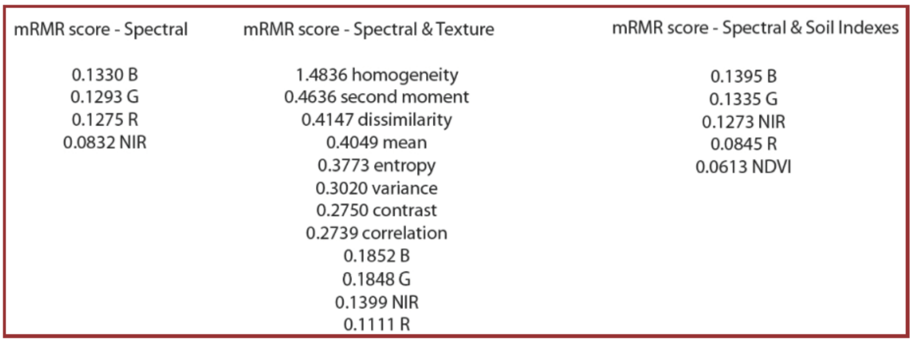


Fig. 9. GeoEye-1 image and the target class *Urban areas.*

On this image we performed three types of experiments, with different input feature sets. The first set consists only of spectral information, the second contains the spectral bands and textural information and the third has the spectral bands and the vegetation index NDVI. Each experiment gave different scores for the mRMR tests and the top is presented in figure 10. Figure 11 shows the classification accuracy of the target class as function of the input features and the classifier used. The graphics show that the lowest error is obtained by the Minimum Distance classifier applied to the first four textural features in the score table.



mRMR score - Spectral	mRMR score - Spectral & Texture	mRMR score - Spectral & Soil Indexes
0.1330 B	1.4836 homogeneity	0.1395 B
0.1293 G	0.4636 second moment	0.1335 G
0.1275 R	0.4147 dissimilarity	0.1273 NIR
0.0832 NIR	0.4049 mean	0.0845 R
	0.3773 entropy	0.0613 NDVI
	0.3020 variance	
	0.2750 contrast	
	0.2739 correlation	
	0.1852 B	
	0.1848 G	
	0.1399 NIR	
	0.1111 R	

Fig. 10. mRMR score for three different feature sets of GeoEye-1 image.

4 The Sensor Data-knowledge Continuum

The recent and future scenarios of sensorics arise with the promise of new capabilities for collecting and distributing information about the world. However, there is still a gap between the raw data coming from the sensor in form of numbers (measures of

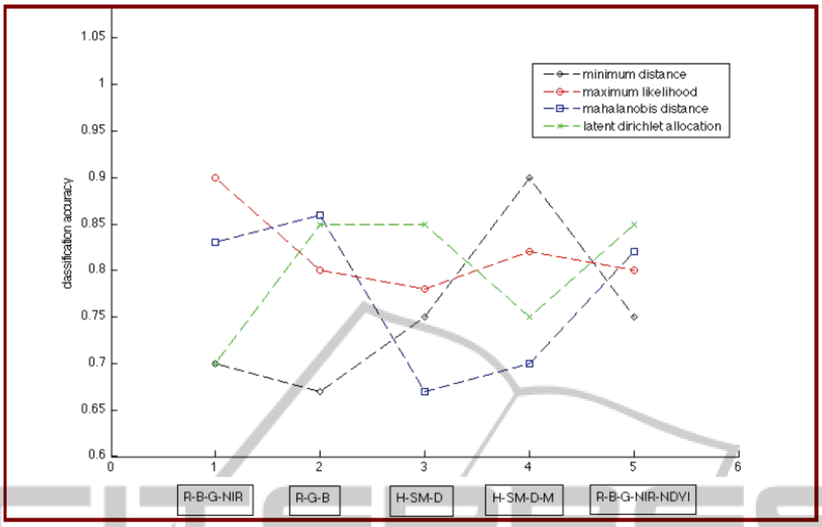


Fig. 11. Classification accuracy scores for GeoEye-1 image, target class *Urban areas*.

physical phenomena) and applied user-oriented knowledge (in form of meaning).

When the human mind understands an object or an event, it classifies it automatically into an acknowledged category with sense and implicit semantic denotation.

We introduce a procedure to describe an important layer of automatic data processing i.e. feature selection with the aim of supporting the users to find what they are looking for in the collection of data. In order to add human-oriented conceptual meaning to the abstract representation of sensor information, semantics may be introduced at different steps of the workflow.

The mRMR technique yields highly accurate results when used with a layered classifier – inferring classes one at a time. The user indicates the target category; adds a descriptive caption to it and detects the optimum features simultaneously. This approach ensures that the class of interest is extracted with maximum accuracy and in the same time it is represented in the concepts domain by the wording chosen by the user. Thus, the added semantics provide an inter-connection between layers of information in the sensor data-knowledge continuum.

References

1. Peng, H., Long, F., Ding, C.: Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 27. No.8. 2005. Pp 1226-1238
2. Bratasanu, D., Nedelcu, I., Datcu, M.: Bridging the gap for satellite image annotation and automatic mapping applications. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, accepted paper 2010
3. Ding, C., Peng, H.: Minimum Redundancy Feature Selection from Microarray Gene Expression Data. Proc. Second IEEE Computational Systems Bioinformatics Conference. 2003. Pp 523-528
4. Richards, J., Jia, X.: Remote Sensing Digital Image Analysis. 4th Edition. Springer, 2006.