Competence Centre on Information Extraction and Image Understanding for Earth Observation

A Novel Data Compression Technique for Remote Sensing Data Mining

Authors: Avid Roman-Gonzalez[(1)] , Miguel A. Veganzones[(2)], Manuel Graña[(2)] and Mihai Datcu[(1,3)]

[(1)]GET/Télécom Paris, 46 rue Barrault, 75013 Paris, France

[(2)]Grupo de Inteligencia Computacional, Universidad del Pais Vasco (UPV/EHU), San Sebastian, Spain

[(3)]German Aerospace Center (DLR), Oberpfaffenhofen, 82234 Weßling, Germany

Motivations / Goals

We propose a parameter-free method for Remote Sensing (RS) image databases Data Mining (DM). DM of RS images requires methodologies robust to the diversity of context found in such large datasets, as well as methodologies with low computational costs and low memory requirements. The methodology that we propose is based on the Normalized Compression Distance (NCD) over lossless compressed data. Normalized Compression Distance is a measure of similarity between two data files using the compression factor as an approximation to the Kolmogorov complexity. This approach allows to directly compare information from two images using the lossless compressed original files, and avoiding the feature extraction/selection process commonly used in pattern recognition techniques. This shortcut makes the proposed methodology suitable for DM applications in RS. We provided a classification experiment with hyperspectral data exemplifying our methodology and comparing it with common methodologies found on the literature.

Normalized Compression Distance (NCD):

Based on the Normalized Information Distance:

$$NID(x,y) = \frac{K(x,y) - \min\{K(x),K(y)\}}{\max\{K(x),K(y)\}}$$

Since the Kolmogorov Complexity *K(x)* is a non-computable function, an approximation is defined considering *K(x)* as the compressed version of *x*, and a lower limit of what can be achieved with the compressor C.

$$NCD(x,y) = \frac{C(xy) - \min[C(x),C(y)]}{\max[C(x),C(y)]}$$

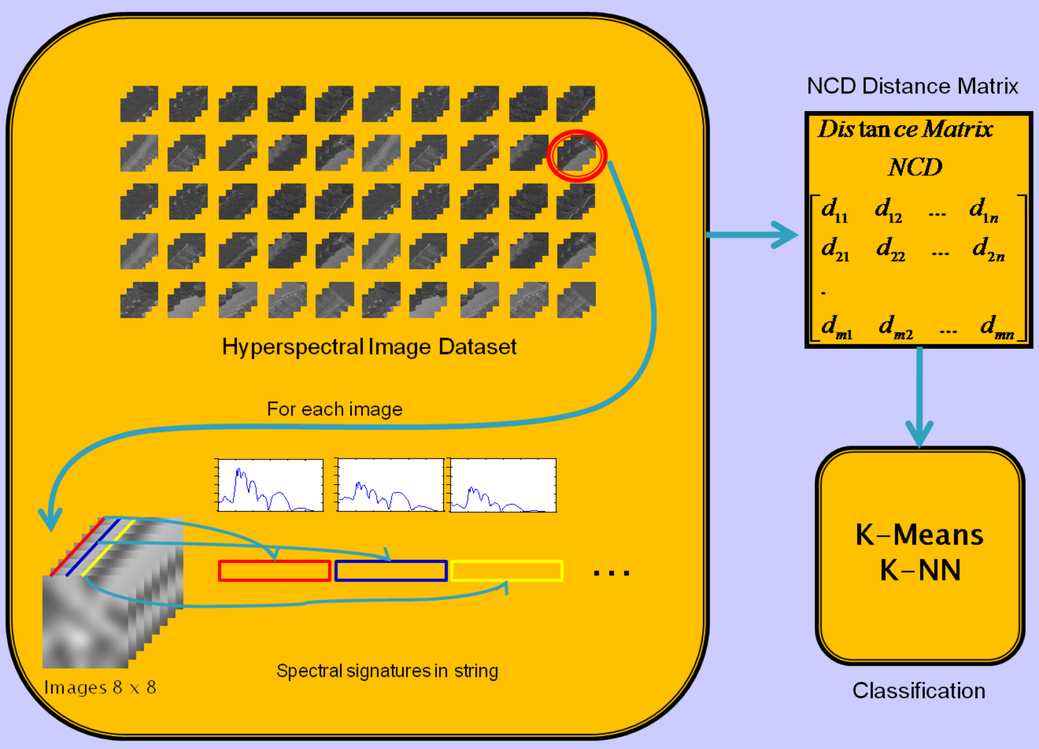
Where: *C(x,y)* represents the size of the compressed file obtained by concatenation of *x* and *y*.

The NCD is a positive result *0 ≤ NCD ≤ 1+e*, with *e* as a representation of imperfections of the compression algorithms.

The NCD works with different classes, in different environments and for different applications.

Methodology:

The goal is to calculate a distance between two hyperspectral images without the need of selecting/extracting features, not tuning any parameter neither. First, hyperspectral images are converted to strings, by concatenating the spectral response of the pixels of each image one after another. To compare two hyperspectral images, H_1 and H_2, the string representations of each image, x_1 and x_2, are compressed by a lossless compressor *C*, and their individual compression factors, $C(x_1)$ and $C(x_2)$, are calculated. Then, x_1 and x_2 are concatenated and compressed to calculate $C(x_1, x_2)$. Now, we can measure the distance between the two images, H_1 and H_2, using the Normalized Compression Distance, $NCD(x_1, x_2)$.



Experimental Results:

We realized some experiments to show the use of the proposed methodology for hyperspectral classification, and we compared the results to other methodologies found on the literature. The hyperspectral data taken by the HyMap sensor have been provided by the German Aerospace Center (DLR).

The sensed scene corresponds to the facilities of the DLR center in Oberpfaffenhofen and its surroundings, mostly fields, forests and small towns. Figure 2 shows some subscenes of the hyperspectral image used on the experiments. The data cube has 2878 lines, 512 samples and 125 bands; and the pixel values are represented by 2-bytes signed integers.

We compared the results obtained with the proposed NCD-based methodology to results obtained using the average patch radiance and the induced endmembers characterization:

NCD distances	Buildings	Fields	Forests	Total
Buildings	30	0	0	30
Fields	5	32	13	50
Forests	0	13	37	50

Overall accuracy: 76.15%. *KHAT*: 74.81%.

Average radiance	Buildings	Fields	Forests	Total
Buildings	26	4	0	30
Fields	12	38	0	50
Forests	0	0	50	50

Overall accuracy: 87.69%. *KHAT*: 87.14%.

Endmembers	Buildings	Fields	Forests	Total
Buildings	28	2	0	30
Fields	12	38	0	50
Forests	0	0	50	50

Overall accuracy: 89.23%. *KHAT*: 88.69%.

Table 1. Results using the unsupervised K-Means algorithm

NCD distances	Buildings	Fields	Forests	Total
Buildings	30	0	0	30
Fields	0	47	3	50
Forests	0	8	42	50

Overall accuracy: 91.54%. *KHAT*: 91.06%.

Average radiance	Buildings	Fields	Forests	Total
Buildings	26	4	0	30
Fields	2	48	0	50
Forests	0	0	50	50

Overall accuracy: 95.38%. *KHAT*: 95.18%.

Endmembers	Buildings	Fields	Forests	Total
Buildings	22	7	1	30
Fields	2	48	0	50
Forests	1	0	49	50

Overall accuracy: 91.53%. *KHAT*: 91.28%.

Table 2. Results using the supervised K-NN algorithm

All the experiments have been run using a K-Fold resampling with 10 folds. Experiments with the induced endmember characterization of the patches have been done using the EIHA endmember induction algorithm and the Spectral Image Distance function to calculate an endmember distance matrix analogous to the NCD distance matrix used as input to the classifiers. Tables 1 and 2 show the confusion matrix, the overall accuracy and the KHAT index for the K-Means and K-NN algorithms respectively.

Conclusions

In this work we proposed the Normalized Compression Distance (NCD) as the base for a novel data compression technique suitable for Remote Sensing data mining. We provided a methodology for the mining of hyperspectral datasets based on this technique. This methodology is easily modifiable for its use with any Remote Sensing data. The applicability of this methodology was tested by a classification and clustering experiments over a dataset of hyperspectral images. The results of the experiments show that the proposed methodology performance is similar to other methodologies found on the literature, while presents some advantages (e.g. no need for a feature extraction/selection process, parameter-free, adaptability to any image size) that makes it suitable for RS data mining. The presented methods can be applied not only to the artifacts detection in optical satellite imagery, but to the anomaly detection in hyperspectral images, SAR images, image fakery, steganalysis and also in medical images. Finally, we can say that these methods can be used for data analysis throughout the area of multimedia.

Acknowledgement

HyMap data was made available from HyVista Corp. and DLRs optical Airbone Remote Sensing and Calibration Facility service (http://www.OpAiRS.aero). The authors very much acknowledge the support of Dr. Martin Bachmann from DLR.