A NOVEL DATA COMPRESSION TECHNIQUE FOR REMOTE SENSING DATA MINING

Avid Román-González, Miguel A. Veganzones, Manuel Graña and Mihai Datcu

TELECOM ParisTech, Paris, France
Grupo de Inteligencia Computacional, Universidad del País Vasco (UPV/EHU), San Sebastián, Spain
German Aerospace Center - DLR, Oberpfaffenhofen, Germany

ABSTRACT

In this article 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 exemplarizing our methodology and comparing it with common methodologies found on the literature.

1. INTRODUCTION

Earth Observation (EO) data have increased significantly over the last decades with orbital and suborbital sensors collecting and transmitting several terabytes of data a day. Image information mining (IIM) is a new field of study that has arisen to seek solutions to automating the mining (extracting) of information from EO archives that can lead to knowledge discovery and the creation of actionable intelligence (exploiting) [1].

Hyperspectral sensors densely measure the spectral response of the elements on the sensed scene, forming images with rich information for thematic mapping, identification of materials (even at subpixel resolution) or anomalies detection. However, hyperspectral image analysis poses several issues [2] as the Hughes phenomenon, due to the small ratio between the number of samples and the number of spectral bands; a high spatial correlation, violating the sampling independence assumption; and the non-stationary behavior of the spectral signatures in the spatial domain. Also, the lack of reference (labelled) samples with respect to the complexity of the problem and the lack of inherent superiority of any predictive learning classifier as well as data clustering algorithm, make mapping assesment and comparison a really hard problem [3]. This issues have been faced with new promissing techniques as such implementing the Structural Risk Minimization (SRM) inductive principle [4], semisupervised learning methods or unsupervised thematic map quality assesments.

However, Pattern Recognition techniques have to cope with new problems when information have to be extracted from large databases of remote sensing images. Feature extraction/selection techniques and classification algorithms are usually chosen and their parameters tuned to solve the mapping task for an specific scene and fail to be applied to different contexts or similar scenes taken in different conditions. Moreover, Data Mining imposes another restrictions such as computational efficiency or memory allocation capacity. In this work, we propose a new parameter-free methodology suitable for data mining of large hyperspectral image databases. The proposed methodology uses lossless compression techniques to compress the hyperspectral data, and defines a Normalized Compression Distance (NCD) over the compressed data as a meassure of similarity between two data files, using the compression factor as an approximation to the Kolmogorov complexity.

We provide a classification experiment over a hyperspectral dataset, showing an example of applicability of the proposed methodology, and comparing it to other methodologies found on the literature. We show in our experiment that the proposed parameter-free NCD-based methodology performance is similar to other well-known approaches, while skips the feature extraction/selection process.

The paper is structured as follow. Section 2 presents the proposed methodology and the theorical basis on which it is based. Section 3 shows a practical application in classification of buildings, forest and fields over a hyperspectral dataset. Finally, section 4 reports our conclusions.

2. PROPOSED METHODOLOGY

In this section we present the theoretical bases for the methodology such as the Kolmogorov complexity and the normalized compression distance; and finally we present the methodology itself.

2.1. Kolmogorov Complexity and Normalized Information Distance (NID)

The Kolmogorov Complexity $K(x)$ of an object x is defined as the minimum amount of computational resources, q, needed to represent x:

$$K(x) = \min_{q \in Q_x} |q| \tag{1}$$

where Q_x is the set of instantaneous codes that give x as output.

It is true that some dependence of the size on the descriptive language we use exists, but it is not very worrying as it is reduced to some constant, i.e., given two languages L_1 and L_2, and any string of symbols x, $|K_1(x) - K_2(x)| < k$. This may seem surprising, but it is not, because all we need to move from a description in L_1 to another in L_2 is a program interpreter of L_1 in L_2 writing. The interpreter may be more or less long, but it's fixed, so that its size is a constant (and the corresponding program size in L_2 is the size on L_1 program plus the interpreter). Another interesting but less surprising result is that there exists a constant k (depending on the language) such that for any string x, $K(x) < |x| + k$. This is easy to see if we think of the worst case, a program containing the string itself as internal constant.

Within the Information Theory we can say that the Kolmogorov complexity or algorithmic complexity (as was top lines) is the amount of information needed to recover x. It is important to note that $K(x)$ is a non-calculable function. The conditional complexity $K(x, y)$ of x related to y is defined as the length of the shortest program with which we can obtain an output x from y. An important application of this notion is to estimate the shared information between two objects: The Normalized Information Distance (NID) [5]. The NID is proportional to the length of the shortest program that can calculate x given y. The distance calculated from these considerations is then normalized as follows:

$$NID(x, y) = \frac{K(x, y) - \min\{K(x), K(y)\}}{\max\{K(x), K(y)\}} \tag{2}$$

The NID result is a positive value r, $0 \leq r \leq 1$, being $r = 0$ if the objects are identical and $r = 1$ for the maximum distance between them.

If we have a string x and a program P smaller than x which describes it, then this program can be interpreted as a compression (lossless information) of x. Therefore, if $K(x) < |x|$ we say that x is compressible.

2.2. Normalized Compression Distance (NCD)

Since the Kolmogorov Complexity, $K(x)$, is a non-computable function, the work in [5] defines the Normalized Compression Distance (NCD) as an approximation to the Normalized Information Distance (NID) considering $K(x)$ as the compressed version of x, and a lower limit of what can be achieved with the compressor C. That is, to approximate $K(x)$ with $C(x) = K(x) + k$, the length of the compressed version of x obtained by a lossless compressor C plus an unknown constant k. The presence of k is necessary because it is not possible to estimate how close of $K(x)$ this approach is. To clarify this concept, we take two strings b and p having the same length n, where the first is a random output of a Bernoulli process and the second represents the first n digits of π. The quantity $K(p)$ would be smaller tan $K(b)$ because there is a natural language program of length $K(p) \ll n$ which output is the number π, while a program that has as output a random sequence of bits would have a close to n length, so $K(p) \ll K(b)$. Thus, equation (2) could be estimated by the Normalized Compression Distance (NCD):

$$NCD(x, y) = \frac{C(x, y) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}} \tag{3}$$

where $C(x, y)$ represents the size of the compressed file obtained by the concatenation of x and y.

The NCD can be calculated explicitly between two strings or two files x and y, and this represents how different are these files, facilitating the use of this result into various applications with different data into a parameter-free approach [6, 7, 8, 9]. The NCD is a positive result $0 \leq NCD \leq 1 + e$, being e a representation of the imperfections of the compression algorithms. It is necessary to remark that the $K(x)$ approximation by $C(x)$ depends on the data with which to work. Knowing that common compressors are built based on different hypotheses, some are more efficient than others with a specific type of data.

2.3. Methodology

Figure 1 shows a block diagram of the proposed 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 concatened and compressed to calculate $C(x_1, x_2)$. Now, we can meassure the distance between the two images, H_1 and H_2, using the Normalized Compression Distance, $NCD(x_1, x_2)$ as it was defined in (3).

It is possible to use this methodology to calculate the matrix of NCD distances between pairs of hyperspectral images from a hyperspectral dataset. That is,

$$D = \{d_{ij}\}; i = 1, .., N; j = 1, .., N \tag{4}$$

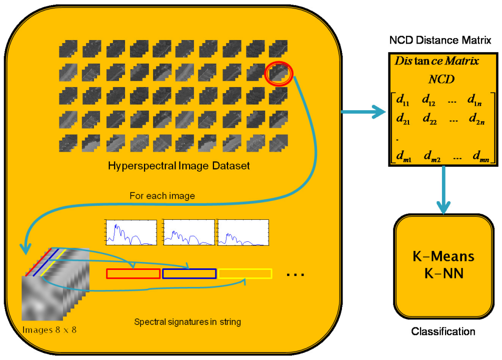
Hyperspectral Image Dataset
For each image
Images 8 x 8
Spectral signatures in string
NCD Distance Matrix
Distance Matrix
NCD
d_{11} d_{12} ... d_{1n}
d_{21} d_{22} ... d_{2n}
d_{m1} d_{m2} ... d_{mn}
K-Means
K-NN
Classification

Fig. 1. Block diagram of the proposed methodology

Fig. 2. Subscenes of the hyperspectral image showing the DLR facilities (Band 7).

where N is the number of images in the dataset, and d_{ij} is the NCD distance between images H_i and H_j. Then, the matrix D of NCD distances can be used as the input to a classifier. Although we make emphasis here in hyperspectral imagery, this methodology is easily modifiable for the analysis of other RS data. It is also to be noted that the analysis of the RS images by this methodology allows to use the image as a whole, independently of their size.

3. 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 took the original image and divided it into several patches of 8×8 pixels, and we selected by visual inspection 130 among them corresponding to three different classes: buildings (30 patches), fields (50 patches) and forests (50 patches). We transformed these patches into strings, x_i, as it has been described in section 2, and we calculated the matrix of distances D (4), between all image pairs using the NCD function (3). Then, we used the distance matrix D as the input to well known classification algorithms: K-Means (unsupervised clustering) and K-NN (supervised classification).

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

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

- The average patch radiance $\bar{r_i}$ for a patch p_i is given by $\bar{r_i} = \frac{1}{N}\sum_{j=1}^{N} p_i^{(j)}$, where $p_i^{(j)}$ is the j-th pixel belonging to patch i, and N is the total number of pixels in the patch ($N = 64$ in our case).
- The induced endmember characterization $E_i = (\mathbf{e_{i1}}, \ldots, \mathbf{e_{im_i}})$ for a patch p_i is a set of endmembers, $\mathbf{e_{ij}}$, $j = 1, \ldots, m_i$, where m_i is the number of endmembers for the patch p_i, obtained by applying an Endmember Induction Algorithm [10, 11] to the patch.

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 [12] and the endmembers distance function in [13] 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.

4. CONCLUSIONS

In this work we proposed the Normalized Compression Distance (NCD) as the base for a novel data compression tech-

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

nique 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 advantadges (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.

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.

5. REFERENCES

[1] M. Datcu, S. D'Elia, R. L. King, and L. Bruzzone, "Introduction to the special section on image information mining for earth observation data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45, no. 4, pp. 795–798, 2007.

[2] L. Bruzzone and C. Persello, "A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 47, no. 9, pp. 3180–3191, 2009.

[3] A. Baraldi, L. Bruzzone, and P. Blonda, "Quality assessment of classification and cluster maps without ground truth knowledge," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 43, no. 4, pp. 857–873, 2005.

[4] Vladimir N. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, Sept. 1998.

[5] Ming Li, Xin Chen, Xin Li, Bin Ma, and P.M.B. Vitanyi, "The similarity metric," *Information Theory, IEEE Transactions on*, vol. 50, no. 12, pp. 3250–3264, 2004.

[6] Eamonn Keogh, Stefano Lonardi, and Chotirat Ann Ratanamahatana, "Towards parameter-free data mining," Seattle, WA, USA, 2004, pp. 206–215, ACM.

[7] T. Watanabe, K. Sugawara, and H. Sugihara, "A new pattern representation scheme using data compression," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 579–590, 2002.

[8] R. Cilibrasi and P.M.B. Vitanyi, "Clustering by compression," *Information Theory, IEEE Transactions on*, vol. 51, no. 4, pp. 1523–1545, 2005.

[9] D. Cerra, A. Mallet, L. Gueguen, and M. Datcu, "Algorithmic information Theory-Based analysis of earth observation images: An assessment," *Geoscience and Remote Sensing Letters, IEEE*, vol. 7, no. 1, pp. 8–12, 2010.

[10] A. Plaza, P. Martinez, R. Perez, and J. Plaza, "A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 42, no. 3, pp. 650–663, 2004.

[11] Miguel A. Veganzones and Manuel Grana, "Endmember extraction methods: A short review," in *Knowledge-Based Intelligent Information and Engineering Systems, 12th International Conference, KES 2008, Zagreb, Croatia, September 3-5, 2008, Proceedings, Part III*, vol. 5179 of *Lecture Notes in Computer Science*, pp. 400–407. Springer, 2008.

[12] Manuel Grana, Ivan Villaverde, Jose O. Maldonado, and Carmen Hernandez, "Two lattice computing approaches for the unsupervised segmentation of hyperspectral images," *Neurocomput.*, vol. 72, no. 10-12, pp. 2111–2120, 2009.

[13] J.O. Maldonado, D. Vicente, M.A. Veganzones, and M. Grana, "Spectral indexing for hyperspectral image CBIR," Torrejon air base, Madrid (Spain), 2006.