

Article

Expanding the Algorithmic Information Theory Frame for Applications to Earth Observation

Daniele Cerra * and Mihai Datcu

German Aerospace Center (DLR), Remote Sensing Technology Institute, Muenchnerstr. 20, 82234 Wessling, Germany; E-Mail: mihai.datcu@dlr.de

* Author to whom correspondence should be addressed; E-Mail: daniele.cerra@dlr.de; Tel.: +49 8153 28-1496; Fax: +49 8153 28-1444.

Received: 28 November 2012; in revised form: 20 December 2012 / Accepted: 14 January 2013 / Published: 22 January 2013

Abstract: Recent years have witnessed an increased interest towards compression-based methods and their applications to remote sensing, as these have a data-driven and parameter-free approach and can be thus successfully employed in several applications, especially in image information mining. This paper expands the algorithmic information theory frame, on which these methods are based. On the one hand, algorithms originally defined in the pattern matching domain are reformulated, allowing a better understanding of the available compression-based tools for remote sensing applications. On the other hand, the use of existing compression algorithms is proposed to store satellite images with added semantic value.

Keywords: algorithmic information theory; data compression; remote sensing

1. Introduction

Information theory has provided along the years several valuable tools for remote sensing applications, especially for model selection, texture extraction, distortion evaluation and informational content characterization of satellite images [1]. For specific applications, ideas derived from Shannon's theory have been applied to Synthetic Aperture Radar (SAR) data [2,3], and to derive similarity measures for spectral signatures acquired by hyperspectral sensors [4].

While classical information theory is based on the estimation of probability density functions of random variables, algorithmic information theory (AIT) focuses on the complexity of single objects. It

is based on the concept of Kolmogorov complexity, which can be described as follows. Considering two binary strings x and y , an universal Turing machine U , and a set of self-delimiting computer programs Q which output x on U given y as input, the Kolmogorov complexity of x conditional to y is defined as:

$$K(x|y) = \min\{|q| : U(\langle y, q \rangle) = x\} \quad (1)$$

where $|q|$ is the size in bits of a program $q \in Q$. For the special case of y being the empty string ε , we rewrite the conditional complexity $K(x|\varepsilon)$ simply as the complexity $K(x)$, which can then be regarded as the size of the shortest computer program which outputs x if no auxiliary input is given to the computation [5]. Being $K(x)$ uncomputable, AIT has mostly been of theoretical interest since its definition, until in recent years the works by Li and Vitányi allowed employing it in practical applications, by approximating $K(x)$ with compression factors [6]. This resulted in the definition of similarity measures applicable to general data with a characteristic parameter-free approach: the only free parameters are the choice of the compressor, which should always be the one that delivers the most compact representation for the kind of data at hand [7], and its specific settings (for example, the size of lookup tables or buffers employed). The advantages of applying these methods to problems in Earth Observation (EO) data analysis such as clustering, classification, and artifact detection are discussed in [8]. Compression-based analysis has been proposed as a possible solution for parameter-free data mining [7], and the importance of these notions in the EO image information mining community is rising, as a recent overview on this topic confirms [9].

This paper expands the spectrum of AIT-based similarity measures and related classification methodologies. With this aim, entities originally defined in other areas such as pattern matching and data compression are reformulated in the information theory frame. This results in a better understanding of compression-based algorithms, which could represent valid tools for remote sensing applications.

The paper is structured as follows. Section 2 introduces compression-based similarity measures, while Section 3 expands the algorithmic information theory frame by exploring its relations with other methodologies, originally defined in other fields. We conclude in Section 4.

2. Preliminaries

The most widely known and used compression based similarity measure for general data is the Normalized Compression Distance (NCD), which assimilates the non-computable Kolmogorov complexity of a string x as the size of the compressed version of x [6], and is defined for any two objects x and y as:

$$NCD(x, y) = \frac{C(x, y) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}} \quad (2)$$

where $C(x)$ represents the size of x after being compressed by a general off-the-shelf compressor (such as Gzip), and $C(x, y)$ is the size of the compressed version of x appended to y . Usually C represents a lossless compressor, but for specific kinds of data such as images the use of lossy compressors has also been proposed [10]. The NCD ranges from 0 to $1 + e$, representing maximum and minimum similarity, respectively, with the e in the upper bound due to imperfections in the compression algorithms, and unlikely to be above 0.1 for most standard compressors [11]. The idea is that if x and y share common

information they will compress better together than separately, as the compressor will be able to reuse recurring patterns found in one of them to more efficiently compress the other. Such distance can be applied to diverse data types [11], and one of its main advantages is its parameter-free approach, as the NCD depends only on the compressor adopted, with performance comparisons for general compression algorithms showing this dependence to be loose [12].

The NCD and its variants have been applied to spectral signatures collected from hyperspectral sensors [13,14], SAR images [15], multispectral data and satellite image time series [8,16].

3. Expanding the Frame

This section brings in the algorithmic information theory frame notions originally defined in the pattern matching, classical information theory and data compression domains.

3.1. Pattern Recognition Based on Data Compression

The Pattern Representation based on Data Compression (PRDC) is a general classification methodology that has been successfully employed to automatically cluster airborne images [17], while its recent expansion PRDC-SSIS has been applied to the segmentation problem of satellite images in [18].

This algorithm extracts typical dictionaries with a compressor similar to LZ78 [19] from the data previously encoded into strings, with each entry in a dictionary assigning a short code to redundant sequences of symbols in the data instance. The dictionaries are later used to compress other files in order to discover similarities with them, by substituting patterns in the strings with their related codes or copy references. The distance of a string s from a class Z represented by a dictionary D_Z is:

$$PRDC(s, Z) = \frac{C_D(s|D_Z)}{|s|} \quad (3)$$

where $C_D(s|D_Z)$ represents the length of the file s , of original length $|s|$, encoded into a string in a first step and then compressed by D_Z as described. Algorithmic information theory notions are mentioned by the authors, but not considered directly connected. It is to be remarked that also the empiric relative entropy defined by Ziv and Merhav [20] is computed in a similar way.

The definitions Equations (2) and (3) can be reformulated to establish a direct link between them. For the remainder of this section, $C(x)$ is considered as the length of the compressed version of a string x when a general compressor of the LZ family such as LZW [21] is used, and the same compressor is adopted to extract a dictionary D_x from x . The prefix-closure property of LZW ensures that, if x contains recurring patterns, then D_x will contain fewer entries, as longer sequences are substituted in the string, maximizing compression. Therefore, x will be simpler than a string y of the same size generating a larger dictionary D_y , and easier to compress:

$$|D_x| < |D_y| \Rightarrow C_D(x) < C_D(y) \quad (4)$$

with $|D_x|$ the number of entries in D_x and $C_D(x) = C_D(x|D_x)$ can be assimilated to $C(x)$ under the previous assumptions. In [17] the authors consider that the total size of the dictionaries to be employed

in the analysis "... should be the smallest representative of the information source being analyzed ... also to reduce the dictionary dependency of PRDC". Therefore, if we have available n dictionaries $\{D_1, \dots, D_i, \dots, D_n\}$ for n objects of the same size belonging to a class Z , the best representative dictionary D_z with $z \in 1 \dots n$ should have the property:

$$|D_z| \leq |D_i|, \forall i \in 1 \dots n \tag{5}$$

and therefore we expect $C(z) \leq C(i), \forall i$. The correlation between dictionary size and complexity is also considered in [22], which approximates $K(x)$ with the quantity $\frac{1}{|x|}c \log_2 |x|$, where c is the number of times a new sequence of characters is found in x . The term c is equal to $|D_x|$, since an entry in D_x is created whenever a new sequence is found in x . The rule of thumb in Equation (4) comes naturally from considering two strings x and y having the same length. This is also congruous with Occam's razor principle which favours simpler models as best representations of the data [23], and agrees with Solomonoff's universal distribution for the a priori probability of a model $m(x) = 2^{-K(x)}$ [24]. Going back to NCD, if we consider two files x and y with $C(x) \leq C(y)$, we may rewrite (2) as:

$$NCD(x, y) = \frac{C(x, y) - C(x)}{C(y)} = \frac{C(y|x)}{C(y)} + O(1) \tag{6}$$

taking advantage of the property $C(x, y) = C(x|y) + C(y) + O(1) = C(y|x) + C(x) + O(1)$ [6], where $C(x|y)$ is the size of the compressed version of x , is y is given as an auxiliary input. We can now assimilate the dividend in Equation (3) to the conditional compression $C(y|x)$ in Equation (6):

$$PRDC(x, y) = \frac{C_D(y|D_x)}{|y|} \tag{7}$$

as under our previous assumptions $C(x) \leq C(y)$ and $|D(x)| \leq |D(y)|$. The differences between Equations (6) and (7), except for the additive term, are the following. Firstly, the information in y is not used to compress y in $PRDC$. Secondly, any general compressor can be used for NCD, but $PRDC$ is limited to the use of an algorithm of the LZ kind. Finally, the normalization factors are different. If we consider in Equation (7) a dictionary D_{xy} extracted from the strings x and y joint, we can insert $PRDC$ in a list of compression-based similarity measures which can be brought in a canonical form and differ by the NCD only for the normalization factor defined in [25].

As the NCD is a relation between compression factors, while the $PRDC$ is basically a compression factor in itself, we expect the former to be more reliable than the latter, which fails at normalizing according to the single complexity of each object the similarity indices obtained: for example, if y is very simple, the similarity with another string x should be high only if y is very well compressed by the dictionary D_x . Performing such normalization of $PRDC$ results in the definition of a Normalized $PRDC$ ($NPRDC$), which is symmetric:

$$NPRDC(x, y) = \frac{C_D(xy|D_{xy}) - \min\{C_D(x|D_x), C_D(y|D_y)\}}{\max\{C_D(x|D_x), C_D(y|D_y)\}} \tag{8}$$

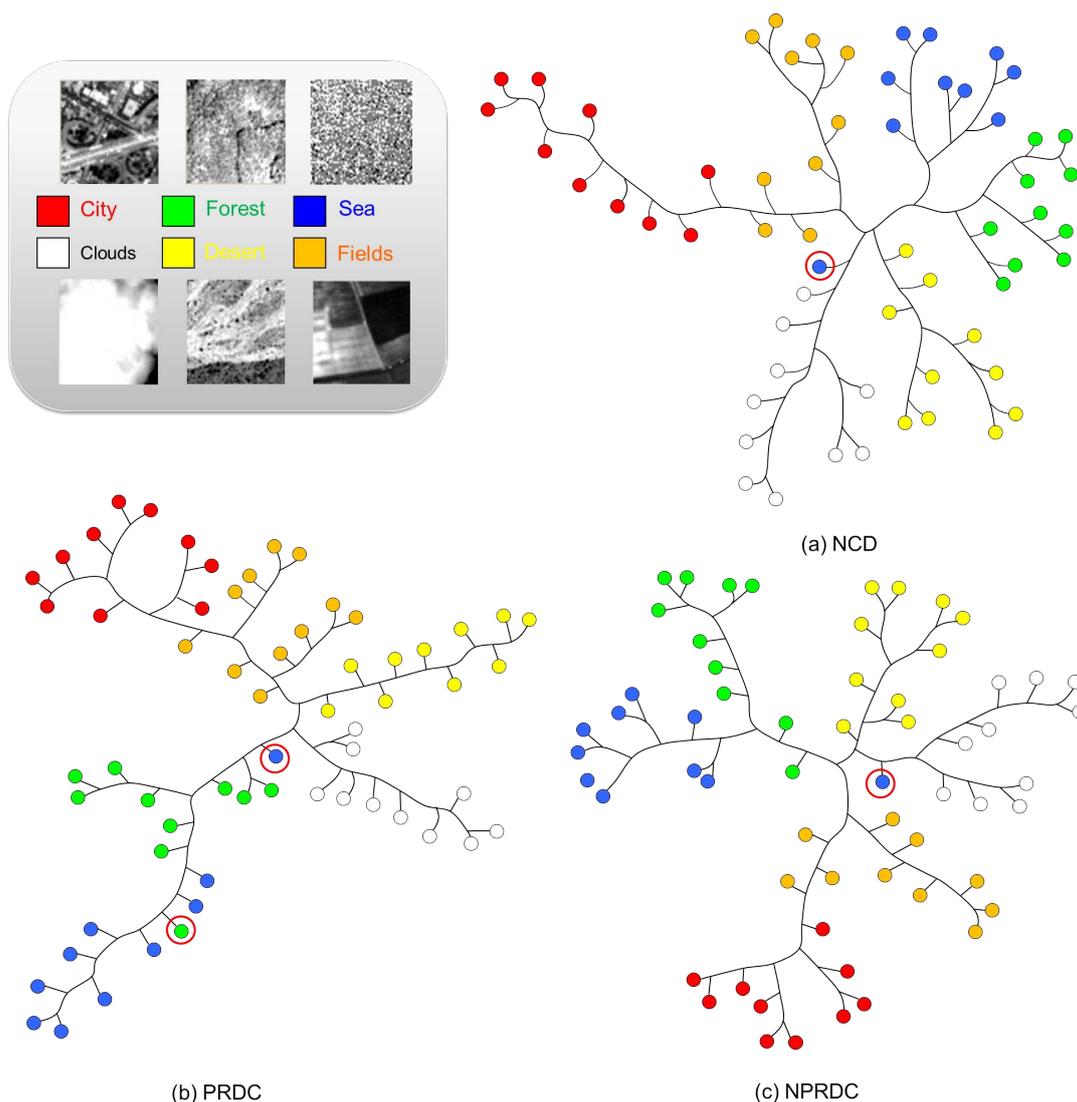
where D_{xy} is the dictionary extracted via the LZW algorithm from x and y merged.

After this normalization, we consider a test set of 20 single-band 64×64 satellite image subsets, and compute the NCD and $NPRDC$ between all the elements, choosing LZW as compressor for the former.

Subsequently, we build a vector v with the absolute differences between each couple of elements: the mean $\mu(v) = -1.84 \times 10^{-3}$ and variance $\sigma^2(v) = 2.45 \times 10^{-4}$ of v suggest that these two distances behave almost identically.

The results of an unsupervised clustering using the three described methods on a dataset comprising 60 SPOT image subsets are reported in Figure 1. The images have been first encoded into strings by traversing them in raster order before applying PRDC and NPRDC. The tool maketree [26] has been used to cluster the results generating the best-fitting binary trees related to the three distance matrices obtained. Results are assessed by visually inspecting if images belonging to the same class are correctly clustered in some branch of the tree, *i.e.*, by checking how much each class can be isolated by "cutting" the tree at convenient points. All clusterings present the same misplacement of an image belonging to the class sea. When switching from PRDC to NPRDC, a misplacement of a forest image in the class *sea* is avoided.

Figure 1. Visual description and hierarchical clustering of NCD (a), PRDC (b) and Normalized PRDC (c) distances between 60 satellite images of size 64×64 belonging to 6 classes. Misplacements are circled in red.



Both PRDC and NCD can be employed in remote sensing applications: the former should be preferred if execution speed is favoured over results accuracy; otherwise, the latter would be a better choice.

3.2. Relative Entropy

In one of the works that pioneered practical applications of compression-based similarity measures, Benedetto *et al.* defined the relative entropy of a string x related to a string y as:

$$H_r(x||y) = \frac{C(x + \Delta y) - C(x) - (C(y + \Delta y) - C(y))}{|\Delta y|} \quad (9)$$

where C is a general compressor, and Δy represents a small fraction of a string y [27]. Equation (9) quantifies how well a small fraction of y can be represented using the information contained in x , with respect to the full information contained in y . This intuition, derived from information theoretical notions, arose great interest within the community along with some controversies [28,29]. Eventual relations with Kolmogorov complexity were only hinted throughout the paper.

Later on, the relative complexity and its computable approximation based on data compression was defined in the algorithmic information theory frame in [16], and applied to classify satellite images. This section establishes a connection between these two notions. Consider the relative complexity as defined in [16] with Δy and x as its arguments:

$$C(\Delta y||x) = \frac{C(\Delta y \oplus x) - C(\Delta y)}{|\Delta y| - C(\Delta y)} \quad (10)$$

where $C(\Delta y \oplus x)$ represents the cross-complexity of Δy related to x , which is the size of the compressed version of Δy using a dictionary D_x extracted from x simultaneously to the compression (therefore, only a part of D_x will be available at any given step) [16]. This term is intuitively close to $C(x + \Delta y) - C(x)$ in Equation (9), as both aim at expressing a small fraction of y only in terms of x .

Secondly, the term $C(y + \Delta y) - C(y)$ in Equation (9) is intuitively close to $C(\Delta y)$ in Equation (10), where in the former a representative dictionary extracted from y is used to code the fraction Δy , while the latter discards any limitation regarding the size of the analysed objects and considers the full string y . This solves a problem raised in [30], which investigates the optimal size for Δy in Equation (9), which does not represent y well enough if set too small, while it uses too much information from y itself in the compression step if set too large.

Finally, the normalization term in the two equations is different: the upper bound for Equation (9) is smaller than 1 in the case of x and y being algorithmically independent (*i.e.*, x and y share no common information): we have $|\Delta y| > \max\{C(x + \Delta y) - C(x) - (C(y + \Delta y) - C(y))\}$, as $|\Delta y| > |\Delta y| - \min\{C(y + \Delta y) - C(y)\}$, due to the monotonicity property of C , which ensures that the quantity $C(y + \Delta y) - C(y)$ is strictly positive, $\forall y$. Therefore, the maximum distance in Equation (9) also depends on the complexity of Δy , while it should in principle be independent.

This strong link between the two distances, both based on data compression but defined in the frames of classical and algorithmic information theory respectively, solves some controversies [28,29] on Benedetto's seminal work, as it repositions it in the consistent theoretical background linking compression to complexity theory, later defined in the works of Li and Vitányi [5,6,11]. The relative

entropy can be added to the list of practical tools derived from information theory notions, and that can be successfully applied to EO data analysis.

3.3. Delta Encoding as Conditional Compression

The conditional compression $C(x|y)$ is the size of the compressed version of a string x , if y is given as an auxiliary input. This quantity is directly used in Equation 6, and can be computed in several ways (see [10] for an example). Yet, delta encoding (or differential file compression) [31] is an existing algorithm in literature which could be used for such computation. Traditional algorithms based on Delta encoding represent x as a compressed file having size $C_{\Delta}(x|y)$, which contains the information to fully recover x if y is available. Common strings between x and y are found through the Longest Common Sub-sequence (LCS) method or by edit-distance methods, which estimate the shortest sequence of edits to convert one string into another. The copy references for the common strings or the sequences of edits are there replaced in the original string.

In the last years the volume of acquired EO data has been growing exponentially [9], and repeated acquisitions of the same areas on ground are always more frequent: delta encoding could then represent an important tool for handling and interpreting remotely sensed images. If we have an available compressed image in an archive, and we wish to store a second one acquired over the same area in a different time, we could use delta encoding to store only the differences between the two, in a compressed format. This would enable a direct comparison between scenes, as the computation of the conditional compression $C(x|y)$ would be implicit, and therefore available for later use with no computational overhead.

4. Conclusions

This paper expands the frame of algorithmic information theory (AIT), by including classification methodologies and general concepts originally defined in other areas of interest. The value of the established connections is different in each case. The reformulation of the Pattern Recognition based on Data Compression (PRDC) methodology [17], coming from the pattern recognition domain, reveals the similarities and differences with the most popular compression-based similarity measure, the Normalized Compression Distance (NCD) [6]: as a result, indications are given regarding the preferred algorithm to employ in practical remote sensing applications. This correspondence can be inserted in the overview contained in [32]. Subsequently, the reformulation in the AIT frame of the concept of relative entropy [27] allows solving some problems under investigation related to this similarity measure, and at the same time some controversies about its theoretical soundness. Finally, the differential file compression [31] is regarded as a possible solution for storing EO data with the added value of computing a similarity between the archived scenes: this index obtained with no computational overhead could have a great value for change detection and image information mining applications.

Acknowledgements

The authors would like to thank Toshinori Watanabe for fruitful discussions.

References

1. Datcu, M.; Seidel, K.; Walessa, M. Spatial information retrieval from remote sensing images: Part A. information theoretical perspective. *IEEE Trans. Geosci. Remote Sens.* **1998**, *36*, 1431–1445.
2. Cloude, S.; Pottier, E. An entropy based classification scheme for land applications of polarimetric SAR. *Geosci. Remote Sens. IEEE Trans.* **1997**, *35*, 68–78.
3. Hegarat-Masclé, S.L.; Vidal-Madjar, D.; Taconet, O.; Zribi, M. Application of Shannon information theory to a comparison between L- and C-band SIR-C polarimetric data versus incidence angle. *Remote Sens. Environ.* **1997**, *60*, 121–130.
4. Du, H.; Chang, C.; Ren, H.; Chang, C.; Jensen, J.; D'Amico, F. New hyperspectral discrimination measure for spectral characterization. *Opt. Eng.* **2004**, *43*, 1777–1786.
5. Li, M.; Vitányi, P. *An Introduction to Kolmogorov Complexity and Its Applications*; Springer-Verlag: New York, NY, USA, 2008.
6. Li, M.; Chen, X.; Li, X.; Ma, B.; Vitányi, P.M.B. The similarity metric. *IEEE Trans. Inf. Theory* **2004**, *50*, 3250–3264.
7. Keogh, E.; Lonardi, S.; Ratanamahatana, C. Towards Parameter-free Data Mining. In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004; p. 215.
8. Cerra, D.; Mallet, A.; Gueguen, L.; Datcu, M. Algorithmic information theory-based analysis of earth observation images: An assessment. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 8–12.
9. Quartulli, M.; Olaizola, I.G. A review of EO image information mining. *ISPRS J. Photogr. Remote Sens.* **2013**, *75*, 11–28.
10. Campana, B.J.L.; Keogh, E.J. A compression-based distance measure for texture. *Stat. Anal. Data Min.* **2010**, *3*, 381–398.
11. Cilibrasi, R.; Vitányi, P.M.B. Clustering by compression. *IEEE Trans. Inf. Theory* **2005**, *51*, 1523–1545.
12. Granados, A.; Cebrian, M.; Camacho, D.; Rodriguez, F. Evaluating the impact of information distortion on normalized compression distance. *Coding Theory Appl.* **2008**, *5228*, 69–79.
13. Veganzones, M.A.; Datcu, M.; Graña, M. Dictionary based Hyperspectral Image Retrieval. In Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods, Vilamoura, Algarve, Portugal, 2012; pp. 426–432.
14. Cerra, D.; Bieniarz, J.; Avbelj, J.; Reinartz, P.; Mueller, R. Compression-based unsupervised clustering of spectral signatures. In Proceedings of the Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2011 3rd Workshop on, Lisbon, Portugal, 6–9 June 2011; pp. 1–4.
15. Cerra, D.; Datcu, M. Compression-based hierarchical clustering of SAR images. *Remote Sens. Lett.* **2010**, *1*, 141–147.
16. Cerra, D.; Datcu, M. Algorithmic relative complexity. *Entropy* **2011**, *13*, 902–914.
17. Watanabe, T.; Sugawara, K.; Sugihara, H. A new pattern representation scheme using data compression. *IEEE Trans. Patt. Anal. Mach. Intell.* **2002**, *24*, 579–590.

18. Nakajima, M.; Watanabe, T.; Koga, H. Compression-based Semantic-Sensitive Image Segmentation: PRDC-SSIS. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS'12), Munich, Germany, 22-27 July 2012.
19. Ziv, J.; Lempel, A. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inf. Theory* **1978**, *24*, 530–536.
20. Ziv, J.; Merhav, N. A measure of relative entropy between individual sequences with application to universal classification. *IEEE Trans. Inf. Theory* **1993**, *39*, 1270–1279.
21. Welch, T. Technique for high-performance data compression. *Computer* **1984**, *17*, 8–19.
22. Kaspar, F.; Schuster, H. Easily calculable measure for the complexity of spatiotemporal patterns. *Phys. Rev. A* **1987**, *36*, 842–848.
23. Soklakov, A. Occam's razor as a formal basis for a physical theory. *Found. Phys. Lett.* **2002**, *15*, 107–135.
24. Solomonoff, R. The universal distribution and machine learning. *Comput. J.* **2003**, *46*, 598.
25. Sculley, D.; Brodley, C. Compression and Machine Learning: A New Perspective on Feature Space Vectors. In Proceedings of the Data Compression Conference, Snowbird, UT, USA, 28–30 March 2006; pp. 332–341.
26. Cilibrasi, R.; Cruz, A.; de Rooij, S.; Keijzer, M. CompLearn. 2002. Available online: <http://www.complearn.org> (accessed on 20 November 2012).
27. Benedetto, D.; Caglioti, E.; Loreto, V. Language trees and zipping. *Phys. Rev. Lett.* **2002**, *88*, 48702.
28. Goodman, J. Extended comment on language trees and zipping. **2002**, arXiv:cond-mat/020238.
29. Benedetto, D.; Caglioti, E.; Loreto, V. On J. Goodman's comment to "Language Trees and Zipping". **2002**, arxiv: cond-mat/0203275.
30. Puglisi, A.; Benedetto, D.; Caglioti, E.; Loreto, V.; Vulpiani, A. Data compression and learning in time sequences analysis. *Phys. D* **2003**, *180*, 92–107.
31. Shapira, D.; Storer, J. In place differential file compression. *Comput. J.* **2005**, *48*, 677.
32. Wyner, A.; Ziv, J.; Wyner, A. On the role of pattern matching in information theory. *IEEE Trans. Inf. Theory* **1998**, *44*, 2045–2056.