Clustering Digital Data by Compression: Applications to Biology and Medical Images

BRUNO CARPENTIERI
Dipartimento di Informatica
Università di Salerno
84084 Fisciano (SA)
ITALY
bc@di.unisa.it

Abstract: - Data compression, data prediction, data classification, learning and data mining are all strictly related as different points of views, or instances, of the same information treatment problem. Compression inspires information theoretic tools for clustering, pattern discovery and classification. For example it has been recently proposed a new, "blind", approach to clustering by compression that classifies digital objects depending on how they pair-wise compress. We will review this clustering method and we will show how this approach can be used in bio-sequences and medical images clustering.

Key-Words: - Data compression, data prediction, data classification

1 Introduction

Compression is the coding of data to minimize its representation. Data Compression is lossless if, after decompression, the reconstructed image is completely identical to the original; otherwise, it is called lossy (or irreversible or noisy) compression.

Compression is justified by the economical and logistic needs to redeem space in storage media and to preserve bandwidth in communication.

Compressed data have to be decompressed to be used, and this extra processing may be too expensive or detrimental to some applications. Hence the design of data compression algorithms involves trade-offs among various factors, including: the degree of compression, the amount of distortion introduced (if using lossy compression), and the computational resources required to compress and to decompress.

Compression is strictly related to learning and to classification and clustering.

Paul Vitányi and his Ph.D. student Rudi Cilibrasi have recently proposed a new strategy for clustering that is based on compression algorithms (see Vitányi and Cilibrasi [2] and Cilibrasi [3]). This approach leads to an interesting clustering algorithm that does not use any "semantic" information on the data to be classified but does a "blind" and effective classification of digital data that is based only on the data compressibility and not on their "meaning".

They have introduced a new distance metric, called NCD (Normalized Compression Distance). NCD is based on data compression and it can be used as a metric to cluster digital data.

In this paper we successfully apply clustering by compression in two different domains: biology and medical images.

In the next Section we review the clustering by compression approach, and the complearn software. Section 3 presents the results obtained by applying clustering by compression approach on a biological digital data. In Section 4 is devoted to the results obtained on medical images and in Section 5 we present our conclusions and outline new research directions.

2 The "Clustering by Compression" Approach to Classification

Clustering is the process of organizing objects into groups based on similarity. Clustering is performed by assigning a set of objects into homogenous groups with respect to a given distance metric.

It is an unsupervised learning problem but generally we embed some kind of information about the objects that have to be clustered cluster in the distance metric that therefore includes our knowledge of the data.

In clustering by compression this is not the case: the NCD distance metric is based only on the compressibility and does not include any explicit semantic knowledge.

ISBN: 978-960-474-350-6

Compression can be used as a distance metric; compression ratios signify a great deal of important statistical information. In fact, suppose that we have two digital files A and B. If we compress A and B with gzip or bzip or with a general-purpose, lossless, data compressor let us indicate with L(A) and L(B) the compressed lengths of A and B.

If we need to compress both A and B then we can compress A and then B and we have as resulting length of the two compressed files: L(A) +L(B). We could also append file B to file A and compress the resulting file AB by obtaining as length of the new compressed file L(AB).

Experimentally it is possible to show that, if and only if A and B are "similar", then:

L(AB) << L(A) + L(B)

Therefore if we want to cluster digital files we might be able to do it by considering how well they compress together in pairs.

Vitányi and Cilibrasi, in [2] and [3], have introduced the concept of Normalized Compression Distance (NCD) that measures how close or different, from the compression point of view, two files are one from another.

Given a compressor with length function L, the Normalized Compression Distance between two files x and y, can be defined as:

$$\mathrm{NCD}(x,y) = \frac{L(xy) - \min\{L(x), L(y)\}}{\max\{L(x), L(y)\}}$$

where L( ) is the length, in bits, of the compressed file.

The Complearn software ([3]) is a powerful software tool that takes as input a set of digital files and a general purpose data compressor and outputs a clustering of the data objects that can be visualized as an un-rooted binary tree.

Complearn, at the moment of writing, is freely available from complearn.org. The software works by building a distance matrix composed by the computing the NCDs between each pair of files in the data set that we want to cluster. This matrix is then given as input to a classification algorithm based on the quartet method. The final output is an un-rooted binary tree where each digital object is now represented at a leaf.

Complearn requires no background knowledge about data. There are no domain-specific parameters, and only a few general settings, to set.

The clustering results depend on the choice of the compressor: different compressors lead to different clustering trees.

Complearn has been successfully tested in many domains, see for example [1], [4], and [5],

3 Clustering Biological Data

The computational analysis of biological data is today a challenge of considerable interest

Complearn has not been designed specifically to cluster biological data. In particular the main obstacles for Complearn are the large amount of data (for example the size of the human DNA consists of about three billion elements) and the complex informational content.

Our objective is to verify the behavior of Complearn in identifying relationships between protein sequences represented by digital files, by clustering them in appropriate groups.

The content of the test dataset consists of one hundred and one protein sequence files. The files are semantically homogeneous; in fact they contain information related to proteins of the same type, i.e. that represents transmembrane receptors, namely integral membrane proteins localized mainly at the level of the cytoplasmic membrane.

The transmembrane proteins differ by the membrane proteins because, unlike membrane, they stretch in the hydrophobic interior of the lipid bilayer, while the membrane proteins remain adherent to only one of the faces of the membrane. Tying with one specific molecule, defined ligand, the transmembrane receptors mediate an intracellular biochemical response, acting as a fundamental role in the process of signal transduction.

The tests were carried out by considering all the hundred and one files that are named with scientific nomenclature of the species to which they refer.

We have begun to perform our biological test by using the standard BZLIB compressor that is included in the Complearn suite.

Fig. 1 shows the un-rooted clustering tree obtained by Complearn. In this tree we notice for example that files that refer to mammals, as for instance Homo Sapiens, Mus Musculus, Rattus Norvegicus, etc., are close.

Fig 2 zooms on three particulars of Fig. 1 to show how organisms that are close in nature are still close in the clustering.

The quality of the hierarchical clustering tree can be measured by the normalized tree benefit score S(T) that is associated with the clustering.

ISBN: 978-960-474-350-6

For this clustering tree the value of S(T) is 0.915592 (where 1 is the maximum) so the clustering tree is good.

4 Clustering Medical Images

Medical images are an important source of digital data. There is an increasing interest in this data because of new medical applications, such as telemedicine, tele-radiology, real time tele-consultation, PACS (Picture Archiving and Communication Systems), etc..

Some of these digital imaging technologies, such as magnetic resonance (MR) and computed tomography (CT), produce three-dimensional images. In the case of MR and CT images each examination produces multiple slices.

A slice is the graphical representation of a cross section of the part of the human body that is currently analyzed.

The collection of all these slices composes a 3-D image. From the compression point of view, the 3-D medical images show a strong correlation among consecutive slices (inter-slice) and a high relation in the spatial context (intra-slice).

We have experimented the Complearn clustering strategy on several CT and MR images by using the same test set generally used in literature for compression testing on 3-d medical images.

The test data set is described in Table I.

Each slice has 256 columns, 256 lines and 8-bit per sample.

Figure 3 shows the clustering tree obtained. The result is almost optimal. In fact the images belonging to the type CT (computed tomography) are grouped together while those belonging to the type MR (magnetic resonance imaging) are grouped together.

5 Conclusion

Compression inspires information theoretic tools for clustering, pattern discovery and classification. Complearn is a powerful tool for clustering by compression. It is a blind clustering but it can be a valid classification method in many domains.

Type	***History*** **(Age / Sex / # of Slices)**	Image
CT	*Tripod fracture* (16 / M / 192)	CT_skull
	Healing scaphoid dissection (20 / M / 176)	CT_wrist
	Internal carotid dissection (41 / F / 64)	CT_carotid
	Apert's syndrome (2 / M / 96)	CT_Aperts
MR	*Normal* (38 / F / 48)	MR_liver_t1
	Normal (38 / F / 48)	MR_liver_t2e1
	Left exophthalmos (42 / M / 48)	MR_sag_head
	Congenital heart disease (1 / M / 64)	MR_ped_chest

Table 1: Medical Images

Here we have successfully experimented this approach on biological data and medical images.

Future work will include a wider testing on images and remote sensing data and improvements in the visualization of the clustering trees.

Acknowledgments

We wish to thank Marco Pastena, Annarita Leone, Raffaele Pizzolante and Giovanni Murano for performing some of the experiments described in this paper.

References:

[1] B. Carpentieri, "A "Blind" Approach to Clustering Through Data Compression". *International Journal of Matematics and Computers in Simulation*, Vol.7, No.2, pp. 162-170, 2013.

[2] R. Cilibrasi and P. Vitányi. "Clustering by Compression". *IEEE Transactions on Information Theory*, 51(4):1523-1545, 2005.

[3] R. Cilibrasi, Statistical Inference through Data Compression. Ph.D. Dissertation, University of Amsterdam, 2007.

[4] R. Cilibrasi, P. Vitányi and R Wolf, "Algorithmic clustering of music". Computer Music Journal, Vol. 28, pp. 49-67, 2004.

[5] P. Ferragina, R. Giancarlo, V. Greco, G. Manzini, G. Valiente, "Compression-based classification of biological sequences and structures via the Universal Similarity Metric: experimental assessment". BMC Bioinformatics. 14(252), 2007.

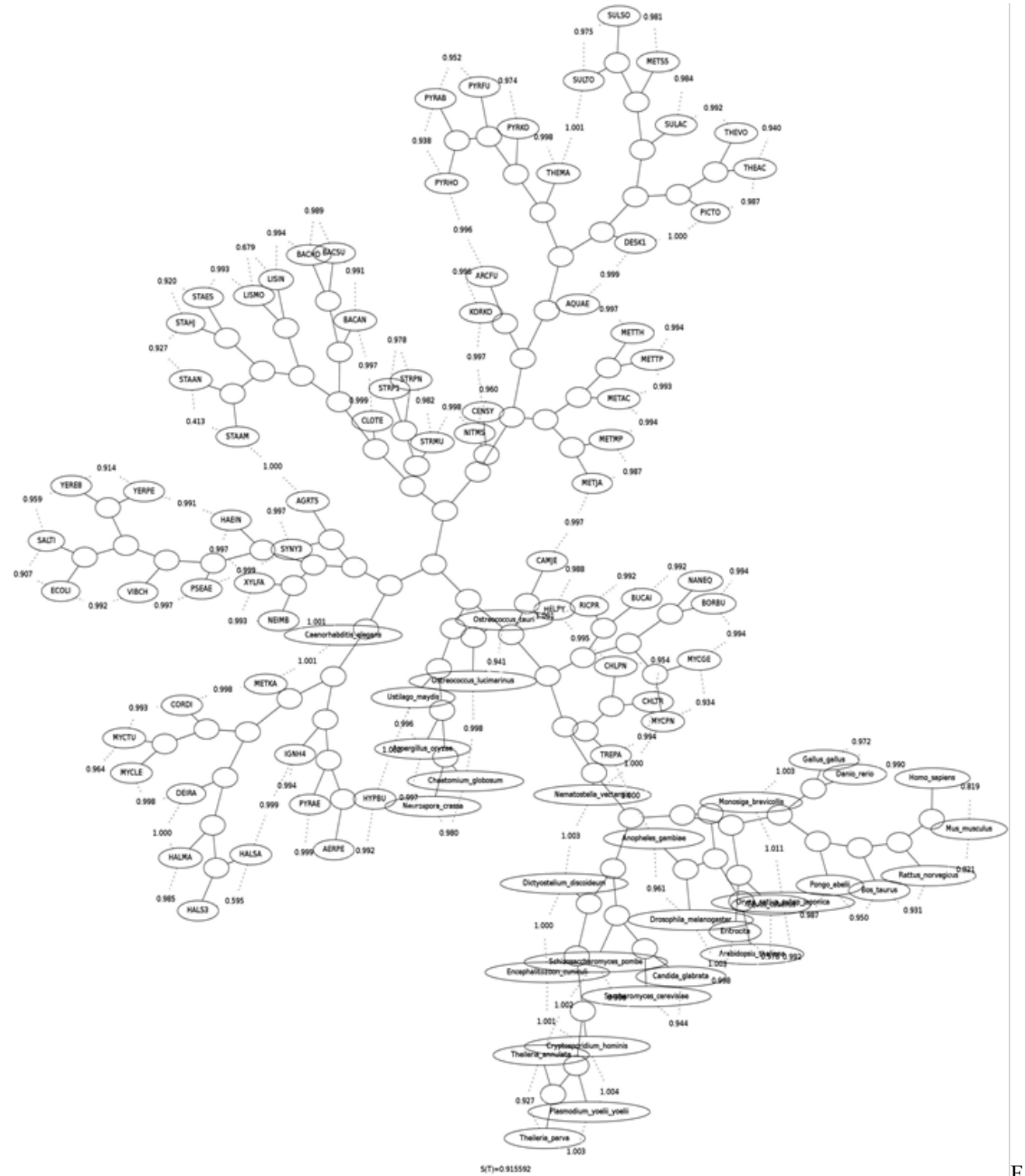

Fig. 1: Clustering obtained on the 101 transmembrane receptors by the BZLIB compressor

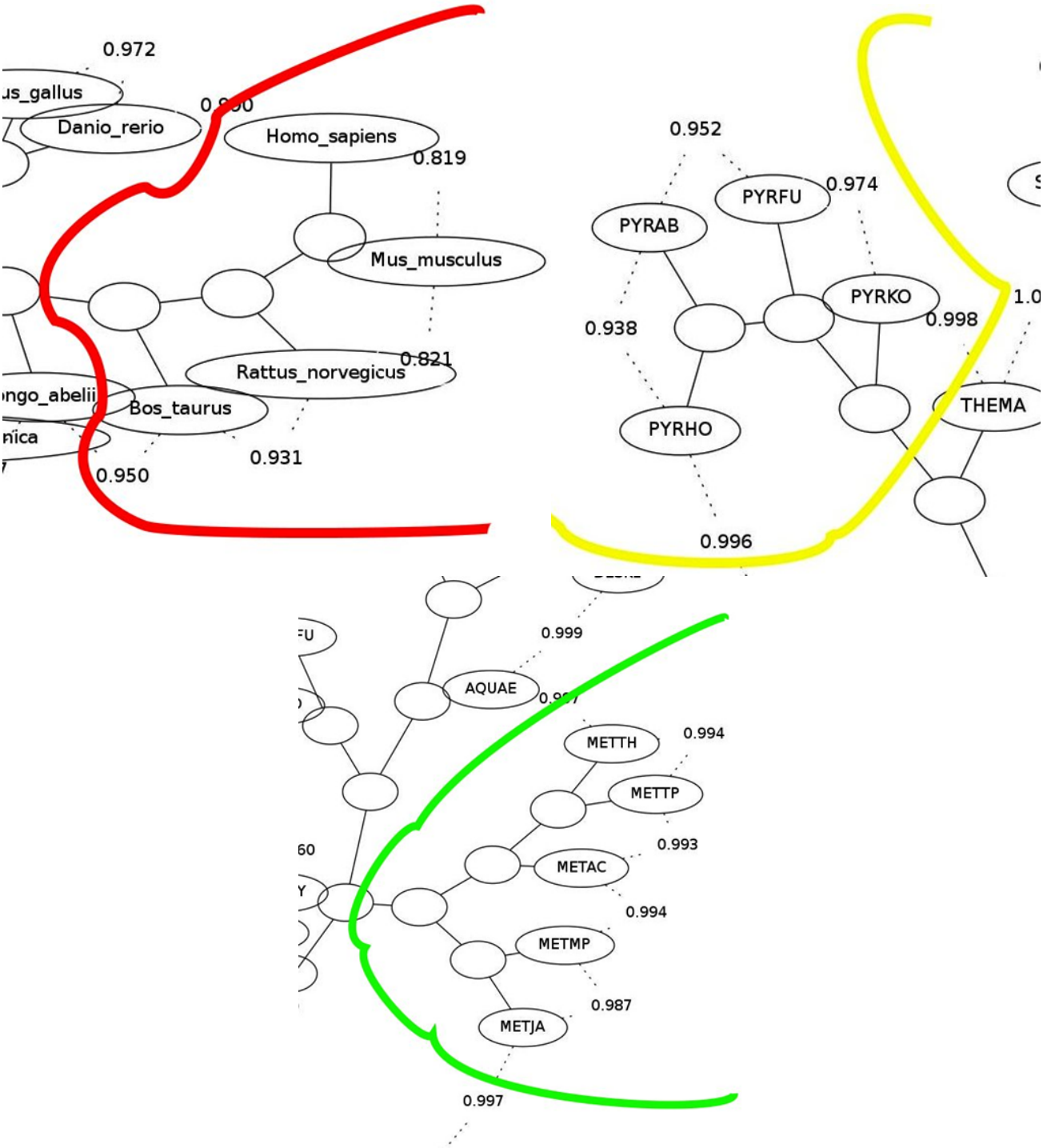

Fig. 2: Three zooms of the clustering in Fig.1.

ISBN: 978-960-474-350-6

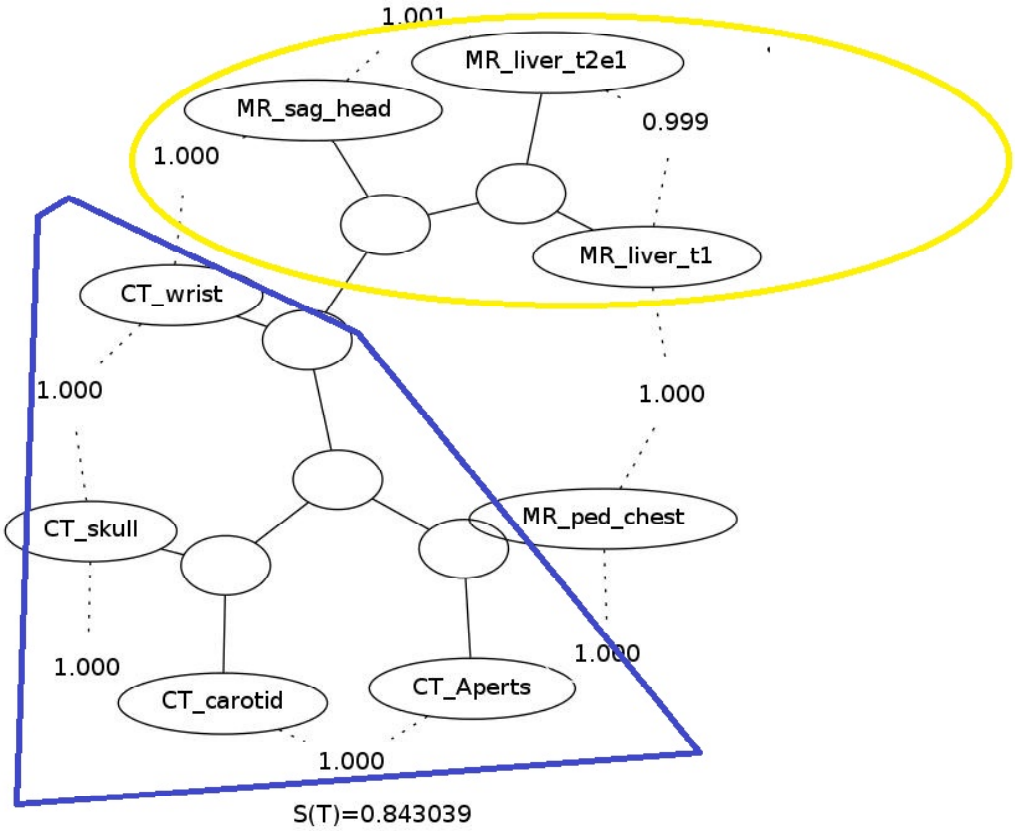

Fig. 3: Clustering obtained on medical images