Data Compression and Clustering: a "Blind" Approach to Classification

BRUNO CARPENTIERI
Dipartimento di Informatica
Università di Salerno
Via S. Allende – 84081 Fisciano (SA)
ITALY
bc@dia.unisa.it http://www.dia.unisa.it/professori/bc

Abstract: Data Compression is today essential for a wide range of applications: for example Internet and the World Wide Web infrastructures benefits from compression. New general compression methods are always being developed, in particular those that allow indexing over compressed data or error resilience.
Compression also inspires information theoretic tools for pattern discovery and classification, in particular it is possible to use data compression as a metric for clustering. This leads to a powerful clustering strategy that does not use any "semantic" information on the data to be classified but does a "blind" and effective classification that is based only on the compressibility of digital data and not on its "meaning". Here we experiment with this strategy and show its effectiveness.

Key-Words: - Data Compression, Clustering, Dictionary based compression, Classification.

1 Introduction

The theoretical background of data compression dates back to the seminal work of Shannon who, more than half a century ago, gave precise limits on the performance of any compression algorithm (see Shannon and Weaver [1]).

Data Compression is today essential for a wide range of applications: for example Internet and the World Wide Web infrastructures benefits from compression and compression inspires information theoretic tools for pattern discovery and classification, especially for biosequences. Additionally, new general compression methods are always being developed, in particular those that allow indexing over compressed data or error resilience.

Today we know that data compression, data prediction, data classification, learning and data mining are facets of the same (multidimensional) coin. In particular it is possible to use data compression as a metric for clustering.

In 2005, Paul Vitányi and his Ph.D. student Rudi Cilibrasi proposed a new idea for clustering, based on compression algorithms (see Vitányi and Cilibrasi [2] and Cilibrasi [3]).

This idea leads to a powerful clustering strategy that does not use any "semantic" information on the data to be classified but does a "blind" and effective classification that is based only on the compressibility of digital data, and not on its "meaning".

Cilibrasi and Vitányi introduced a new distance metric, called *NCD* (Normalized Compression Distance), that is based on data compression and showed how to cluster digital data by using this NCD metric.

In this paper we review recent work on clustering by compression and we experiment with different data sets to test the effectiveness of this new approach.

In the next Section we review the strict relationship between compression and clustering, the Normalized Compression Distance, and the *complearn* software. Section 3 presents the results obtained in testing this clustering by compression approach on a wide variety of digital data and in Section 4 we present our conclusions and outline new research directions.

2 Clustering by Compression

Clustering is the job of assigning a set of objects into clusters, i.e. into homogenous groups, so that the objects in the same cluster are more similar with each other than to those in other clusters with respect to a given distance metric.

Generally we need to "know" about the objects to cluster a set and this knowledge is made explicit in the distance metric that includes our knowledge on the data.

ISBN: 978-1-61804-126-5

In clustering by compression this is not the case, the distance metric is based on the compressibility of the data and does not include any explicit semantic knowledge about the objects.

To intuitively understand why compression can be used as a distance metric, let us suppose that we have two digital files *A* and *B*. If we compress *A* and *B* with a general-purpose, lossless, data compressor (for example *gzip* or *bzip*) we can indicate with *L(A)* and *L(B)* the compressed lengths (in bits) of A and B.

If we need to compress together A and B then we can first compress A and then B and we have as resulting length of the two compressed files: *L(A) +L(B)*. Another option we have is to append file *B* to file *A* and compress the resulting file AB. The resulting length of the new compressed file shall be *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)

This is because compression ratios signify a great deal of important statistical information. This observation gives us a hint that if we want to cluster a set of digital files we might be able to do it by considering how well they compress together in pairs.

In [2] and [3], Vitányi and Cilibrasi have introduced the concept of Normalized compression Distance (NCD). NCD measures how different two files are one from another. NCD depends on a particular compressor and may give different results for the same pair of objects when used with different compressors.

For a given compressor with length function L, the Normalized Compression Distance between two digital objects *x* and *y*, can be formally defined as:

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

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

In 2007 Cilibrasi finished his dissertation and the implementation of the *Complearn* software ([3]). *Complearn* is a powerful software tool that takes as input a set of digital objects and a data compressor and produces a clustering of the data objects visualized on the computer screen as an un-rooted binary tree. It is freely available from complearn.org. It works by building a *distance matrix* composed by the pairwise Normalized Compression Distances between the objects in the data set that we want to cluster. This matrix is the input to a classification algorithm based on the quartet method: the output will be an un-rooted binary tree where each digital object is now represented at a leaf.

This method requires no background knowledge about any particular classification. There are no domain-specific parameters to set and only a few general settings.

3 Clustering Real Data.

In the Compression laboratory of the University of Salerno we have extensively worked in testing and improving the Complearn approach.

Here we present the results of some meaningful tests on real life data of the clustering by compression approach.

3.1 Heterogeneous Data.

For this test we decided to check the behavior of the clustering algorithm on data collected from different domains. Specifically we selected twenty files describing elements evenly distributed between animals, plants, grains, mushrooms and metals.

The four text files regarding animals are: "gatto", "delfino", "merlo", "maiale" (in English: "cat", "dolphin", "blackbird", "pig").

For example the "delfino" file is:

Regno:	Animalia
Sottoregno:	Eumetazoa
Superphylum:	Deuterostomia
Phylum:	Chordata
Subphylum:	Vertebrata
Superclasse:	Tetrapoda
Classe:	Mammalia
Sottoclasse:	Theria
Ordine:	Cetacea
Sottordine:	Odontoceti
Famiglia:	Delphinidae
Genere:	Delphinus
Specie:	D. delphis

The four text files regarding plants are: "aceroRiccio", "papaveroOppio", "cipolla", "frassinoMaggiore" (in English: "maple", "poppy", "onion", "ash").

For example the "aceroRiccio" file is:

Regno:	Plantae
Divisione:	Magnoliophyta
Classe:	Magnoliopsida
Ordine:	Sapindales
Famiglia:	Aceraceae
Genere:	Acer
Specie:	A. platanoides

ISBN: 978-1-61804-126-5

The four text files regarding grains are: “fava”, “lupino”, “riso”, “avena” (in English: “stone”, “lupine”, “rice”, “oatmeal”). The four text files regarding mushrooms are: “colombinaRossa”, “leccino”, “amanitavirosa”, “castagnin”.

The four text files regarding metals are: “osmio”, “palladio”, “stagno”, “silicio” (in English: “osmium”, “palladium”, “tin”, “silicon”).

These files represent elements readily identifiable and classifiable by a human user in order to facilitate the subsequent analysis.

Figure 1 shows the un-rooted binary tree resulting by the clustering algorithm. It is pretty clear (and the colored lines we have drawn show just that) that the results obtained in this test are optimal. The factors used in the analysis were correctly classified accordingly to their most important features, in a way that every branch of the graph represents one of the five domains examined.

3.2 Text in Different Languages

For this test we decided to check the potentiality of the clustering algorithm in identifying and clustering texts depending on their languages.

To do this we have tested the complearn approach on eighteen text files, three of which were in Dutch (“olandese”), three in Copt (“copto”), three in Japanese (“giapponese”), three in Spanish (“spagnolo”) and three in German (“tedesco”).

The test is fully successful, and the text files are grouped by language. Moreover complearn is able to identify also the relationships between the different languages. As can be seen from Figure 2 the three texts in Chinese and the three in Japanese are very close in the resulting clustering tree and the Dutch files are grouped next to the German files.

This experiment was repeated by increasing the number of text files to fifty-five files in eleven different languages, by including texts in Korean, Greek, Arabic, Portuguese and Danish and the clustering obtained was successful in this case too.

3.3 Drugs

This test involves twenty-four different drugs, each described by a text file containing the same explanations that you can find on the drug’s leaflet.

The description files provide information on the following sections: the active ingredient, excipients, indications, contraindications, side effects, precautions for use, various uses, dosage, overdose.

In this case too the classification obtained fulfills our expectations. Figure 3 shows the un-rooted binary tree resulting by the clustering algorithm. The drugs are grouped mainly accordingly to their common characteristics and in particular with respect to their active ingredients.

The colored lines we have drawn emphasize this result, showing that the drugs that have common molecules have been properly inserted in the graph into contiguous locations.

4 Conclusion and Future work

Today we know that data compression, data prediction, data classification, learning and data mining are facets of the same (multidimensional) coin. In particular it is possible to use data compression as a metric for clustering.

The clustering method that we have tested in this paper does not rely on any knowledge or theoretical analysis of the problem domain, but it relies only on general-purpose compression techniques. On this basis the results are certainly impressive: the system is especially versatile and, under appropriate conditions, robust.

On the other side, other experiments have shown that when the number n of objects that we want to cluster increases than the clustering results degrade rapidly. This confirms the need of more experiments to search for a solution to this problem.

However, the potential of the method is surprising and there are many possible interesting developments. For example we are also experimenting the usage of this approach to design a data analysis engine that could be able to discover automatically the unknown characteristics that make different objects similar.

Acknowledgements

I would like to thank my students Vito Cuozzo, Raffaele Pizzolante, and Danilo Coppola for carrying out some of the experimental work described in this paper.

References:

[1] C. S. Shannon and W. Weaver, The mathematical theory of communication. University of Illinois Press, Urbana, IL., 1949.

[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.

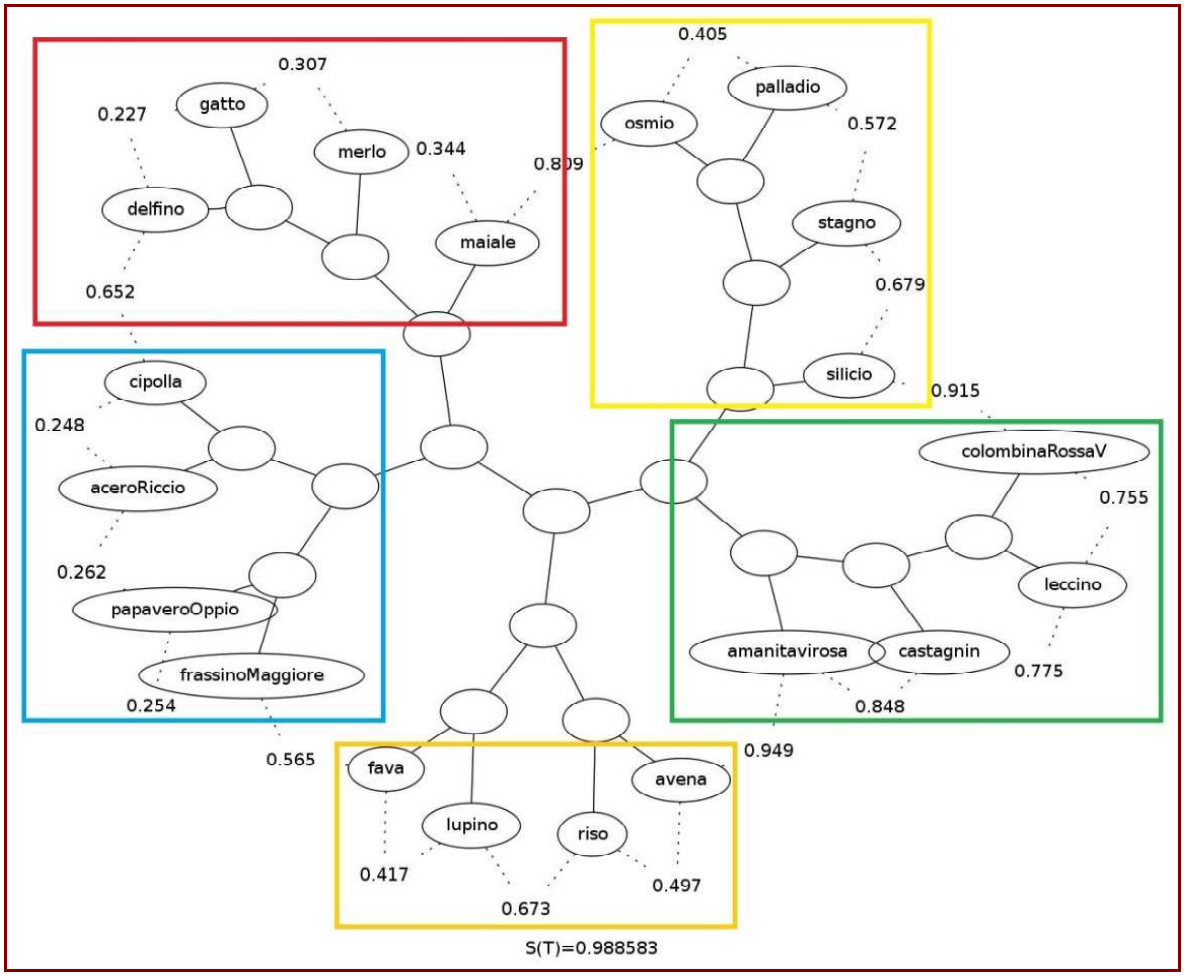

Figure 1. Heterogeneous Data

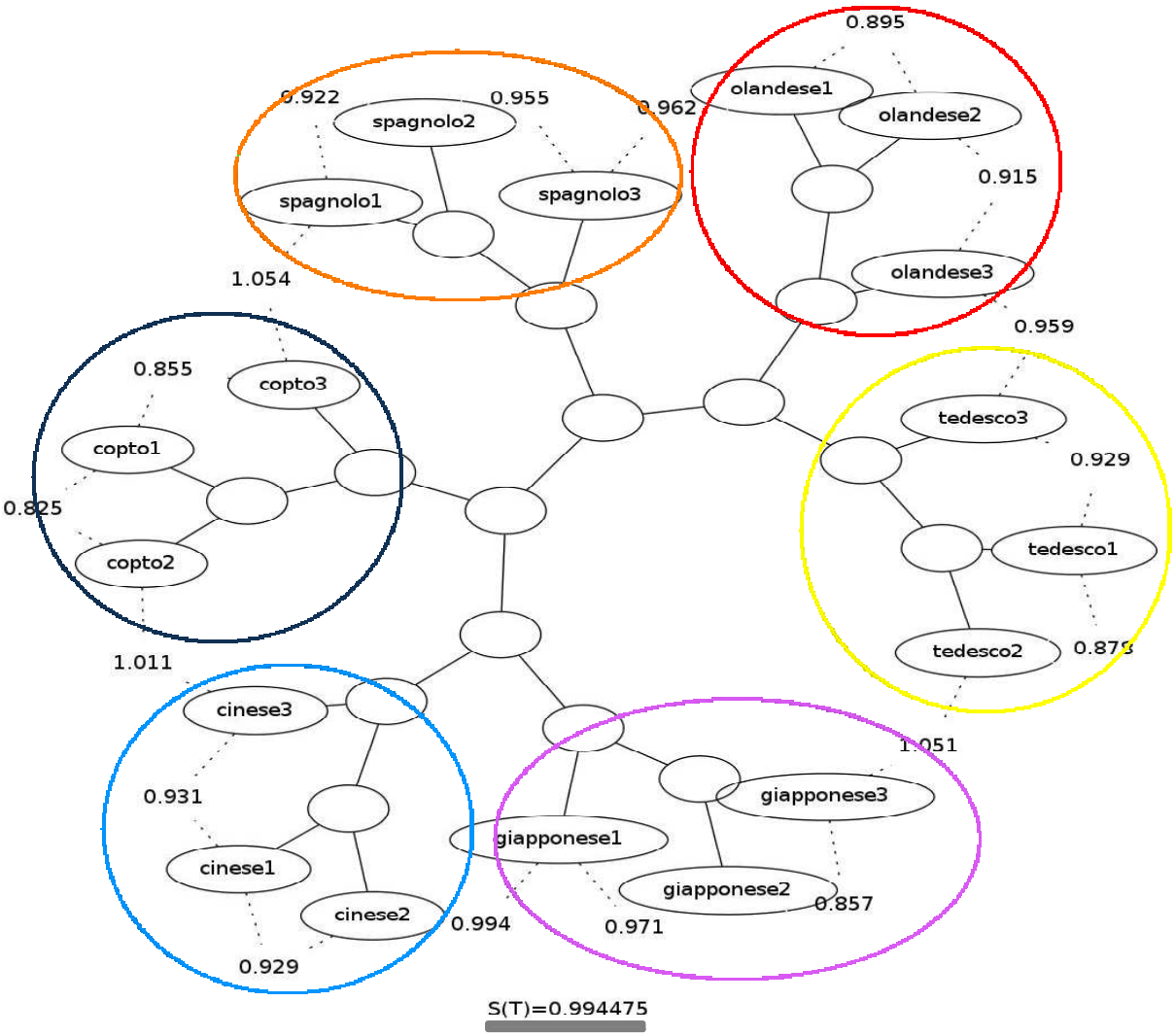

Figure 2. Text in Different Languages

ISBN: 978-1-61804-126-5

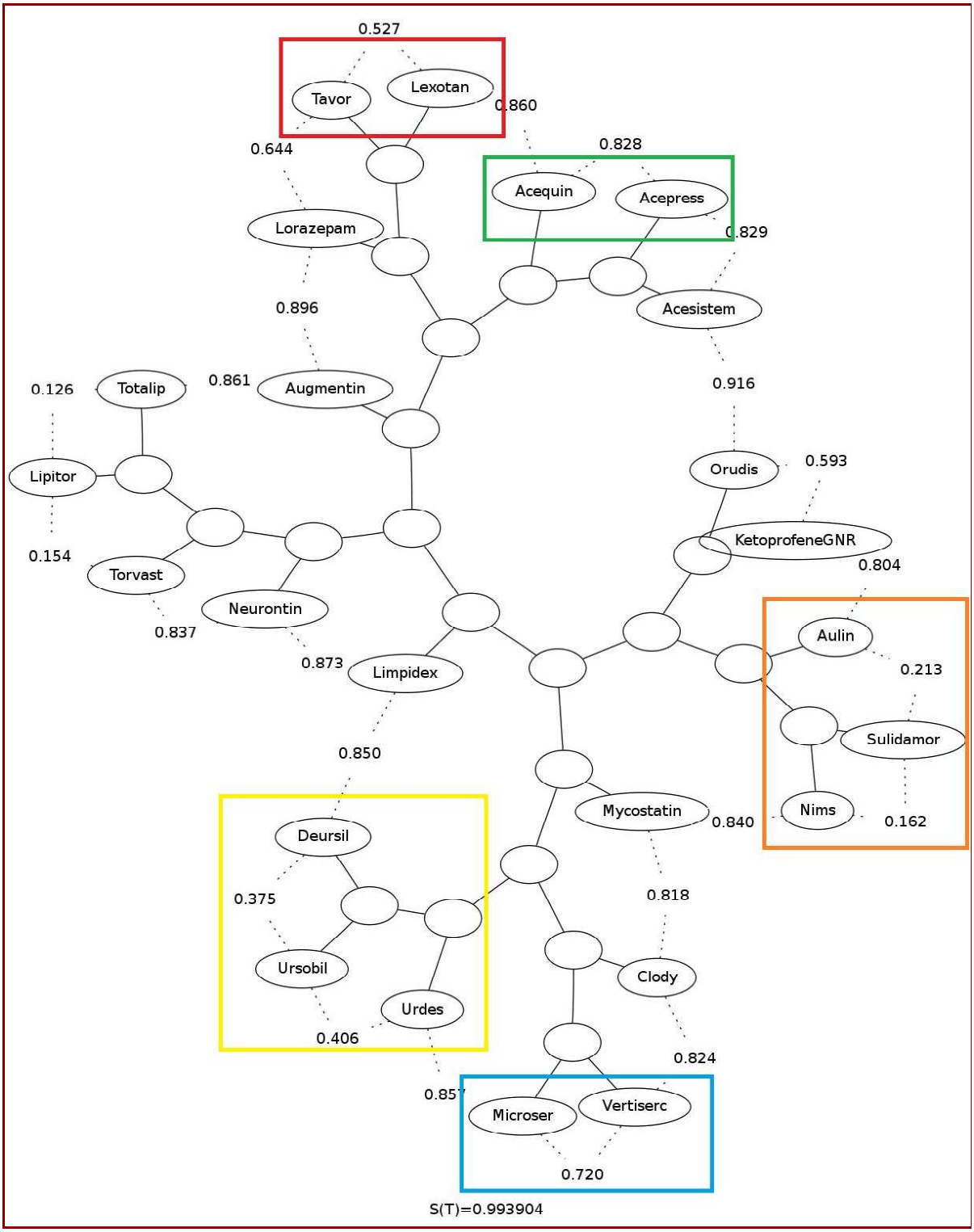

Figure 3. Drugs