A “Blind” Approach to Clustering Through Data Compression

Bruno Carpentieri

Abstract—Data compression, data prediction, data classification, learning and data mining are all facets of the same (multidimensional) coin. In particular it is possible to use data compression as a metric for clustering. In this paper we test a clustering method that does not rely on any knowledge or theoretical analysis of the problem domain, but it relies only on general-purpose compression techniques. Our experiments, on different kinds of digital data, show that the results obtained are impressive: the system is versatile and, under appropriate conditions, robust. The experimental results are presented for clustering of digital data representing heterogeneous data, text in different languages, drugs, cereals, and music.

Keywords— Data Compression, Clustering, Dictionary based compression, Classification.

I. Introduction

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

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

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]).

Current research on lossless compression focuses on developing compressors for new types of digital data or on improving, often only by small amounts, existing compressors for a specific class of data (see for example Carpentieri [4], [5], [6], [9]). Even a small improvement is an important achievement, given the economical impact it can have on data transmission and storage (see for example Ansalone and others [7], Matola and others [8], Rizzo and others [10]).

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.

II. 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 of objects and this knowledge is made explicit in the distance metric that includes our knowledge on the data.

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.

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

Prof. Bruno Carpentieri is with the Dipartimento di Informatica, Università di Salerno, Via S. Allende – 84081 Fisciano (SA), ITALY (phone: +39 089969500; fax: +39 089969600; e-mail: bc@dia.unisa.it).

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 file 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:

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

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.

III. 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 digital data of the clustering by compression approach.

A. *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

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.

B. *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 three 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.

C. 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.

Also in this case the classification obtained fulfills our expectations. Figure 3 shows the un-rooted binary tree resulting by the clustering algorithm. The drugs are grouped mainly according 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.

D. Cereals

This test covers nineteen varieties of cereals, for each of which we collected information related to its taxonomic classification (specifically information about class, order, family, genus and species).

The nineteen cereals examined are: *Avena* (in english Oats), *Cece* (Chickpea), *Cicerchia* (Chickling), *Fagiolo* (Bean), *Farro* (Spelt), *Fava* (Broad Bean), *Grano saraceno* (Buckwheat), *Grano duro* (Durum wheat), *Grano tenero* (Soft wheat), *Lenticchia* (lentil), *Lupino* (Lupine), *Miglio* (Millet), *Mais* (Corn), *Orzo* (Barley), *Pisello* (Pea), *Riso* (Rice), *Segale* (Rye), *Sorgo* (Sorghum), *Triticale* (Triticale).

For example the "Fagiolo" file is:

Classe: Dicotyledonae
Ordine: Leguminosae
Famiglia: Papilionaceae
Tribù: Genisteae
Specie: Phaseolus vulgaris L.

The "Riso" file is:

Classe: Monocotyledones
Ordine: Glumiflorae
Famiglia: Graminaceae (Gramineae o Poaceae)
Tribù: Orizeae
Specie: Oryza sativa L.
Al genere Oryza appartiene anche la specie O. glaberrima Steud., coltivata solo in piccole zone dell'Africa tropicale occidentale.

The "Grano saraceno" file is:

Classe: Dicotyledones
Famiglia: Polygonaceae
Specie: Fagopyrum esculentum Moench.
Sinonimo: Polygonum fagopyrum L.

Figure 4 shows the clustering obtained. The results meet perfectly as expected, with stronger links grouping closer different species with a common genus.

In the figure it has been highlighted by using two colors (green for cereals, red for legumes), that complearn has been able to clearly separate the different varieties of cereals on one side and of legumes on the other, with buckwheat to act as a point of contact due its description file (given with less details).

E. Music

This domain is particularly interesting when you consider the exponential growth of websites that offer musical contents and that, therefore, need efficient methods to organize their contents. Currently, much of the analysis and classification needed is carried out by real people, or it is simply based on the buying pattern of customers, and there is an increasing interest in techniques that may automate the extraction procedures from audio files of common characteristics such as hue, rhythm, harmony, etc..

We have conducted some experiments, first by considering music in the MP3 digital format but the results were quite disappointing. To explain this failure, we must consider that the mp3 files are already highly compressed, therefore if we try to apply further compression in order to measure the compression distance and to isolate the characteristics common to different pieces of music this cannot work.

We therefore decided to perform the music analysis by using files in the MIDI format. Each MIDI file has been subjected to a step of preprocessing to increase their uniformity by eliminating the header fields containing title, names of the author/composer, the name of the program used to create the file, and essentially all the information not strictly necessary and not related to the musical content, in fact this preprocessing does not alter the musical content of the MIDI file.

In our test we have considered twelve pieces of classical music, twelve rock songs and twelve jazz pieces.

Table 1 shows the 11 classical pieces we used for our test.

COMPOSER	MUSIC
J.S. Bach	*Wohltemperierte Klavier II: Preludes 1,2 and fugues 1,2*
Chopin	*Préludes op. 28: 1, 15, 22, 24*
Debussy	*Suite bergamasque, 4 movements*

Table 1: Classical Music

Tab. 2 shows the 11 jazz pieces we used for our test.

COMPOSER	MUSIC
John Coltrane	*Blue Trane* *Giant Steps* *Lazy Birt* *Impressions*
Miles Davis	*Milestone* *Seven Steps To Heaven* *Solar* *So What*
George Gershwin	*Summertime*
Dizzy Gillespie	*Night in Tunisia*
Thelonious Monk	*Round Midnight*
Charlie Parker	*Yardbird Suite*

Table 2: Jazz

Tab 3. shows the 11 rock/pop pieces we used for our test.

COMPOSER	MUSIC
The Beatles	*Eleanor Rigby* *Michelle*
Eric Clapton	*Cocaine* *Layla*
Dire Straits	*Money for Nothing*
Led Zeppelin	*Stairway to Heaven*
Metallica	*One*
Jimi Hendrix	*Hey Joe* *Voodoo Child*
The Police	*Every Breath You Take* *Message in a Bottle*
Rush	*Yyz*

Table 3: Rock/Pop

Figure 5 shows the clustering obtained.

The clustering seems good but not quite perfect. The top right of the figure contains ten of the twelve jazz pieces, but also the Prelude no.5 of Chopin and the Bach Prelude, while the other two jazz pieces are placed in other positions, according to some hidden relationship not easily evident at first glance. S

imilarly, nine of the twelve rock pieces are placed in the bottom left of the tree, with the other instead accommodated elsewhere.

Most classical pieces occupy the central part and the angle at the bottom right of the tree. The movements of Debussy and three of the four Preludes by Chopin are, as expected, close to each other, while it is rather strange to note that two of the Bach's pieces are grossly out of place.

The reason why this happens is not clear and may be due to an error in the measurement of the compression distance that we have chosen to adopt, but on the other hand the fact that they have been arranged in the vicinity of some pieces rather than others may reflect some similarities not so obvious at first glance.

One of the purposes of this new clustering methodology is in fact to obtain a process of data-mining that could be able to discover similarities between characteristics both known and unknown, so a deeper musicologic study of the graph seems to be needed.

IV. Conclusion and Future work

The clustering method that we 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 experimental work and of a deeper analysis of the appropriate compressors to be used 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 experimenting the use of this clustering strategy for biological data and also for band ordering in the compression of hyperspectral images.

Moreover we are 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.

Acknowledgment

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.

[4] Carpentieri, B, "Interactive compression of digital data", (2010). Algorithms, 3 (1), pp. 63-75.

[5] Carpentieri, B., "Interactive compression of books", (2010). WSEAS Transactions on Computers, 9 (3), pp. 278-287.

[6] Carpentieri, B., "Image compression via textual substitution", (2009). WSEAS Transactions on Information Science and Applications, 6 (5), pp. 768-777.

[7] Ansalone, A., Carpentieri, B., "How to set "don't care" pixels when lossless compressing layered documents", (2007). WSEAS Transactions on Information Science and Applications, 4 (1), pp. 220-225.

[8] Matola, L., Carpentieri, B., "Color re-indexing of palette-based images", (2006). WSEAS Transactions on Information Science and Applications, 3 (2), pp. 455-461.

[9] Carpentieri, B., "Sending compressed messages to a learned receiver on a bidirectional line", (2002). Information Processing Letters, 83 (2), pp. 63-70.

[10] Rizzo, F., Storer, J.A., Carpentieri, B., "LZ-based image compression", (2001). Information Sciences, 135 (1-2), pp. 107-122.

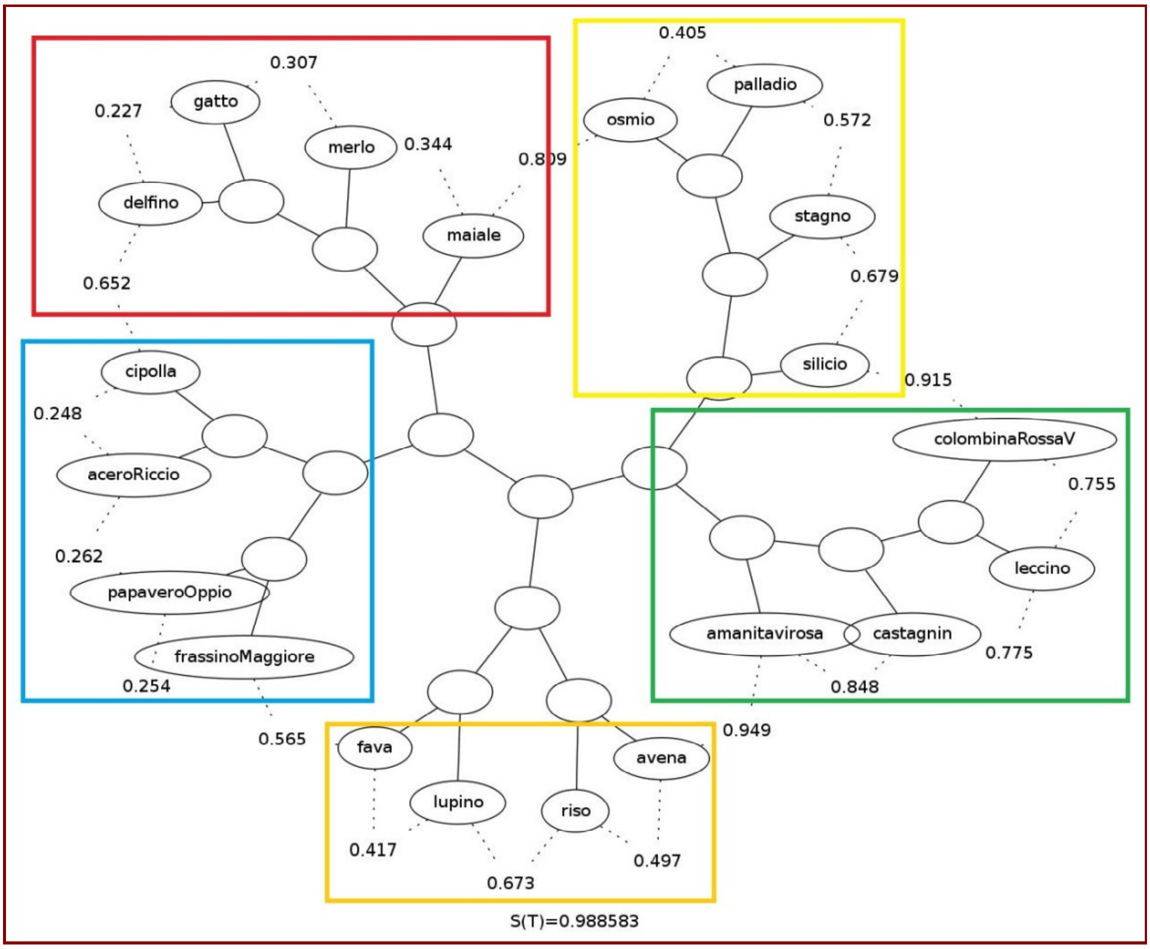

Figure 1. Heterogeneous Data

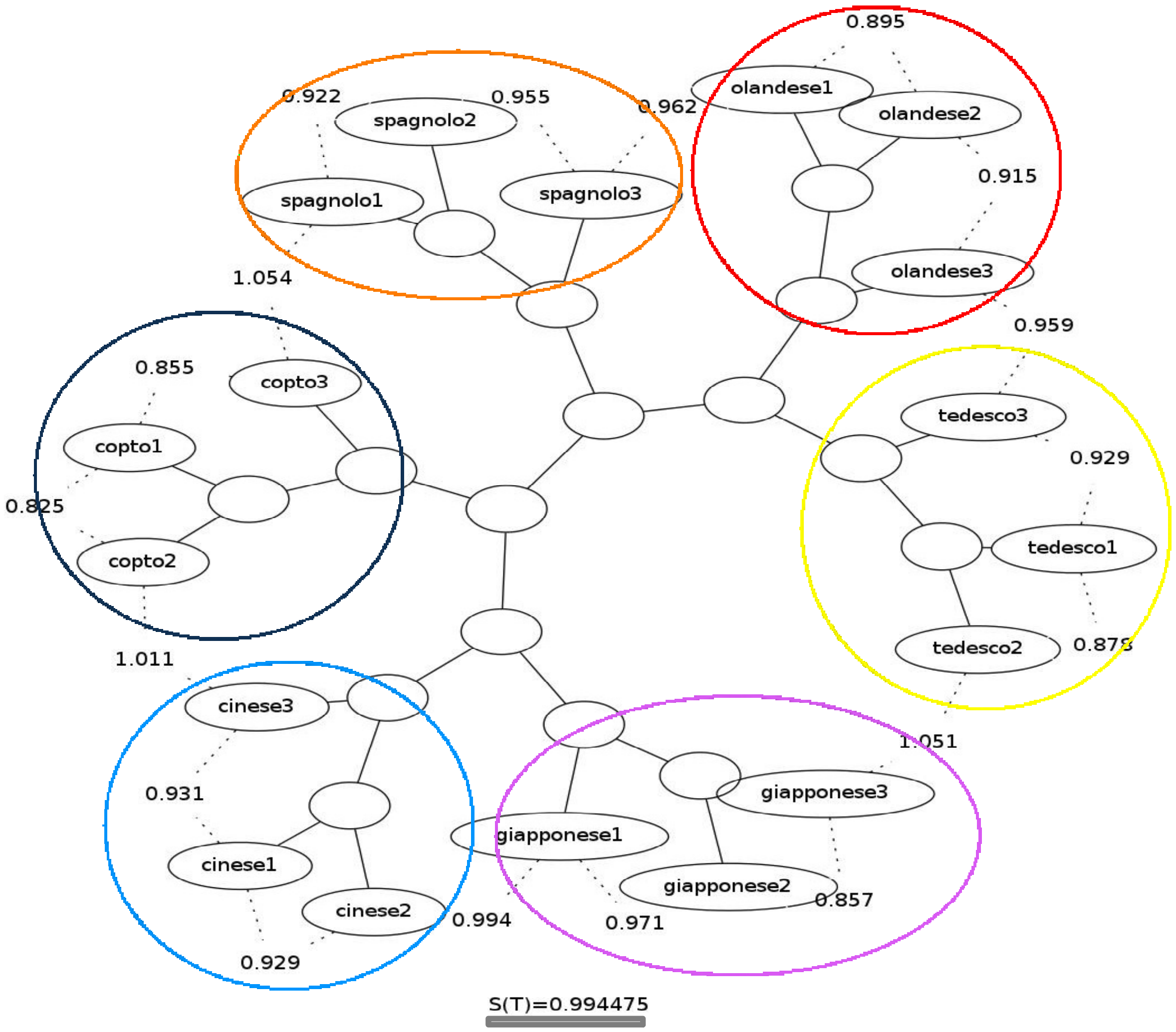

Figure 2. Text in Different Languages

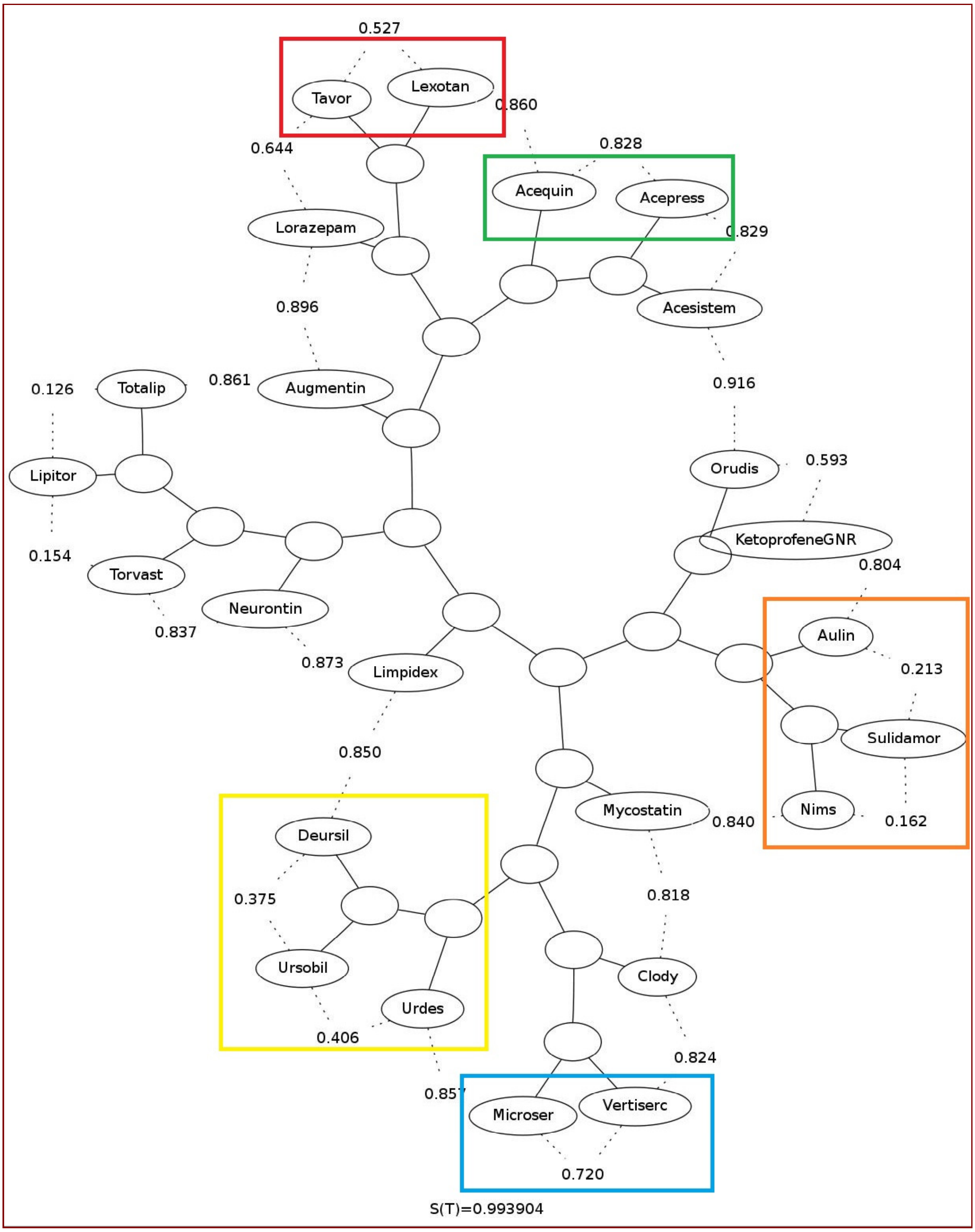

Figure 3. Drugs

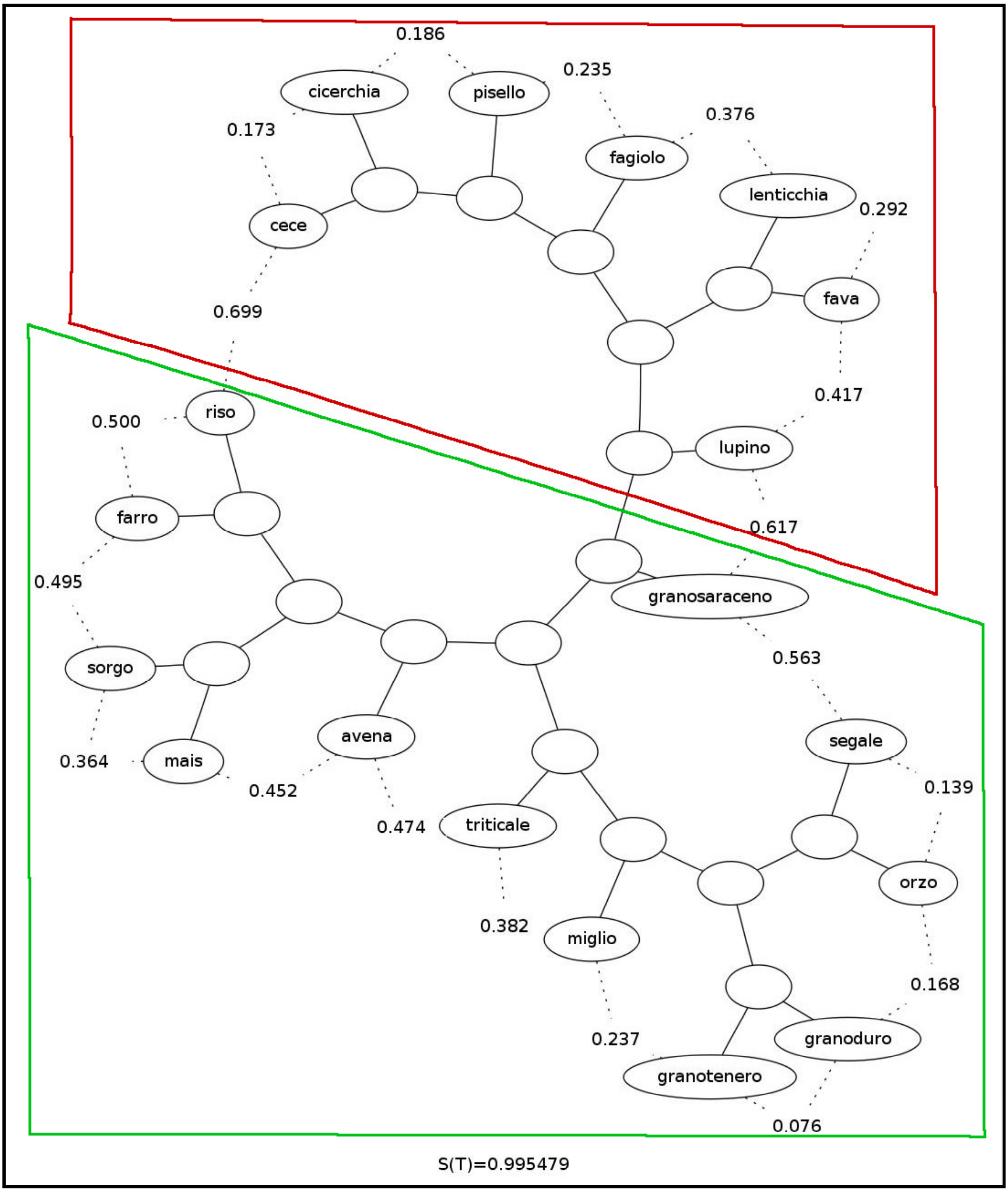

Figure 4. Cereals

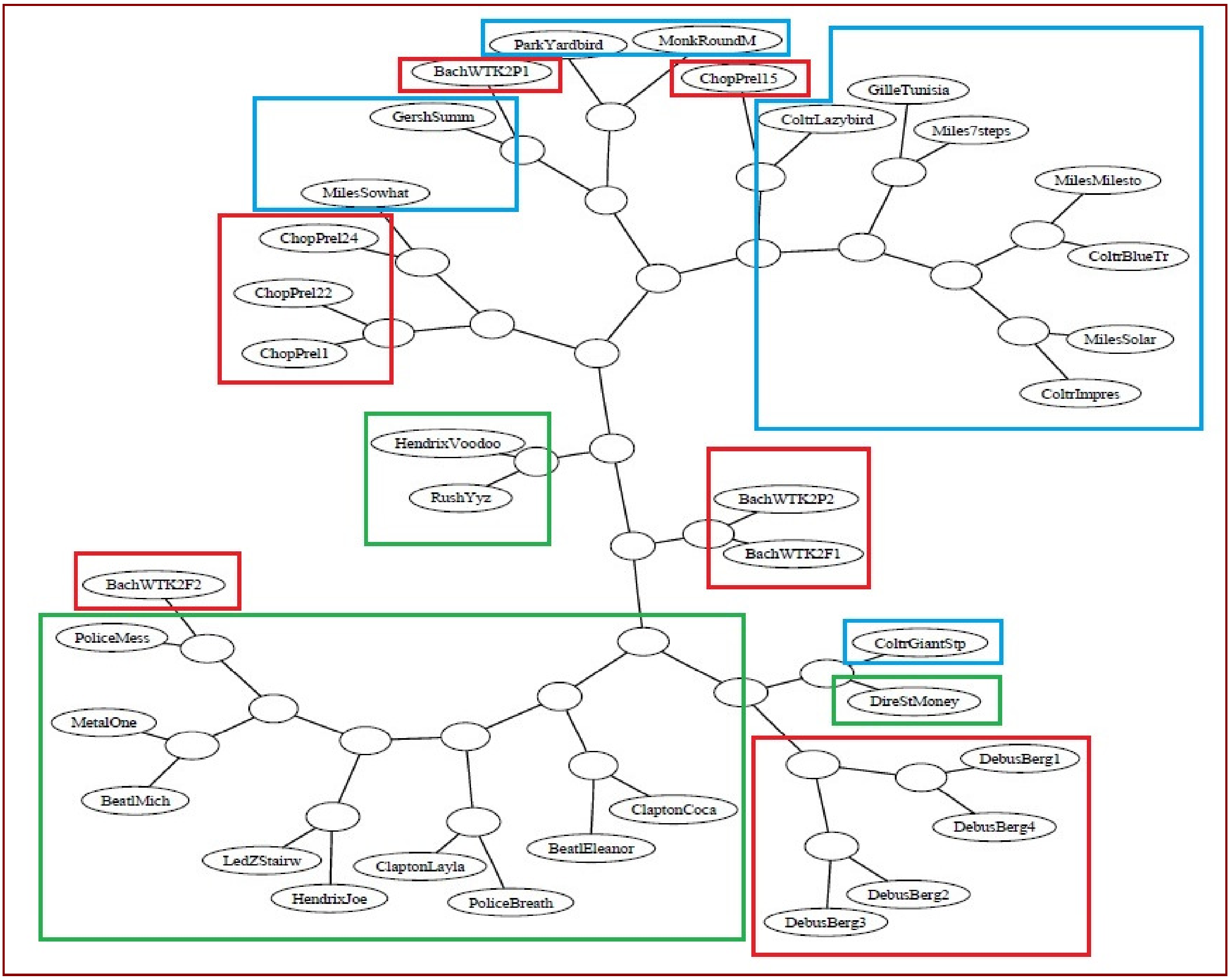

Figure 5. Music