comp4620/8620: Advanced Topics in AI
Foundations of Artificial Intelligence

Marcus Hutter
Australian National University
Canberra, ACT, 0200, Australia
http://www.hutter1.net/
6 THE UNIVERSAL SIMILARITY METRIC

- Kolmogorov Complexity
- The Universal Similarity Metric
- Tree-Based Clustering
- Genomics & Phylogeny: Mammals, SARS Virus & Others
- Classification of Different File Types
- Language Tree (Re)construction
- Classify Music w.r.t. Composer
- Further Applications
- Summary
The Similarity Metric: Abstract

The MDL method has been studied from very concrete and highly tuned practical applications to general theoretical assertions. Sequence prediction is just one application of MDL. The MDL idea has also been used to define the so called information distance or universal similarity metric, measuring the similarity between two individual objects. I will present some very impressive recent clustering applications based on standard Lempel-Ziv or bzip2 compression, including a completely automatic reconstruction (a) of the evolutionary tree of 24 mammals based on complete mtDNA, and (b) of the classification tree of 52 languages based on the declaration of human rights and (c) others.

Based on [Cilibrasi&Vitanyi’05]
Kolmogorov Complexity

Question: When is object=string $x$ similar to object=string $y$?

Universal solution: $x$ similar $y \iff x$ can be easily (re)constructed from $y$
$\iff$ Kolmogorov complexity $K(x|y) := \min \{ \ell(p) : U(p, y) = x \}$ is small

Examples:

1) $x$ is very similar to itself ($K(x|x) \doteq 0$)

2) A processed $x$ is similar to $x$ ($K(f(x)|x) \doteq 0$ if $K(f) = O(1)$).
   e.g. doubling, reverting, inverting, encrypting, partially deleting $x$.

3) A random string is with high probability not similar to any other string ($K(random|y) \equiv \text{length}(random)$).

The problem with $K(x|y)$ as similarity=distance measure is that it is neither symmetric nor normalized nor computable.
The Universal Similarity Metric

- Symmetrization and normalization leads to a/the universal metric $d$:

$$0 \leq d(x, y) := \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}} \leq 1$$

- Every effective similarity between $x$ and $y$ is detected by $d$

- Use $K(x|y) \approx K(xy) - K(y)$ (coding $T$) and $K(x) \equiv K_U(x) \approx K_T(x)$ $\implies$ computable approximation: Normalized compression distance:

$$d(x, y) \approx \frac{K_T(xy) - \min\{K_T(x), K_T(y)\}}{\max\{K_T(x), K_T(y)\}} \lesssim 1$$

- For $T$ choose Lempel-Ziv or gzip or bzip(2) (de)compressor in the applications below.

- Theory: Lempel-Ziv compresses asymptotically better than any probabilistic finite state automaton predictor/compressor.
Tree-Based Clustering

- If many objects $x_1, ..., x_n$ need to be compared, determine the similarity matrix $M_{ij} = d(x_i, x_j)$ for $1 \leq i, j \leq n$

- Now cluster similar objects.

- There are various clustering techniques.

- Tree-based clustering: Create a tree connecting similar objects,

- e.g. quartet method (for clustering)
Genomics & Phylogeny: Mammals

Let $x_1, ..., x_n$ be mitochondrial genome sequences of different mammals:

Partial distance matrix $M_{ij}$ using bzip2(?)

<table>
<thead>
<tr>
<th></th>
<th>Cat</th>
<th>Echidna</th>
<th>Gorilla</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>BrownBear</td>
<td>0.002</td>
<td>0.943</td>
<td>0.887</td>
<td>0.935</td>
</tr>
<tr>
<td>Carp</td>
<td>0.943</td>
<td>0.006</td>
<td>0.946</td>
<td>0.954</td>
</tr>
<tr>
<td>Cat</td>
<td>0.887</td>
<td>0.946</td>
<td>0.003</td>
<td>0.926</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>0.935</td>
<td>0.954</td>
<td>0.926</td>
<td>0.006</td>
</tr>
<tr>
<td>Cow</td>
<td>0.906</td>
<td>0.947</td>
<td>0.897</td>
<td>0.926</td>
</tr>
<tr>
<td>Echidna</td>
<td>0.944</td>
<td>0.955</td>
<td>0.942</td>
<td>0.948</td>
</tr>
<tr>
<td>FinbackWhale</td>
<td>0.915</td>
<td>0.952</td>
<td>0.905</td>
<td>0.926</td>
</tr>
<tr>
<td>Gibbon</td>
<td>0.939</td>
<td>0.951</td>
<td>0.928</td>
<td>0.849</td>
</tr>
<tr>
<td>Gorilla</td>
<td>0.940</td>
<td>0.957</td>
<td>0.931</td>
<td>0.731</td>
</tr>
<tr>
<td>HouseMouse</td>
<td>0.934</td>
<td>0.956</td>
<td>0.919</td>
<td>0.943</td>
</tr>
<tr>
<td>Human</td>
<td>0.930</td>
<td>0.946</td>
<td>0.922</td>
<td>0.667</td>
</tr>
</tbody>
</table>
Genomics & Phylogeny: Mammals
Evolutionary tree built from complete mammalian mtDNA of 24 species:

Carp
Cow
BlueWhale
FinbackWhale
Cat
BrownBear
PolarBear
GreySeal
HarborSeal
Horse
WhiteRhino
Gibbon
Gorilla
Human
Chimpanzee
PygmyChimp
Orangutan
SumatranOrangutan
HouseMouse
Rat
Opossum
Wallaroo
Echidna
Platypus

Ferungulates
Eutheria
Primates
Eutheria - Rodents
Metatheria
Prototheria
Genomics & Phylogeny: SARS Virus and Others

- Clustering of SARS virus in relation to potential similar virii based on complete sequenced genome(s) using bzip2:

- The relations are very similar to the definitive tree based on medical-macrobio-genomics analysis from biologists.
Genomics & Phylogeny: SARS Virus and Others
Classification of Different File Types

Classification of files based on markedly different file types using bzip2

- Four mitochondrial **gene** sequences
- Four excerpts from the **novel** “The Zeppelin’s Passenger”
- Four **MIDI** files without further processing
- Two Linux x86 ELF executables (the **cp** and **rm** commands)
- Two compiled **Java** class files

No features of any specific domain of application are used!
Classification of Different File Types

Perfect classification!
Language Tree (Re)construction

- Let $x_1, ..., x_n$ be the “The Universal Declaration of Human Rights” in various languages $1, ..., n$.

- Distance matrix $M_{ij}$ based on gzip. Language tree constructed from $M_{ij}$ by the Fitch-Margoliash method. [Li 03]

- All main linguistic groups can be recognized (next slide)
Classify Music w.r.t. Composer

Let $m_1, \ldots, m_n$ be pieces of music in MIDI format.

Preprocessing the MIDI files:

- Delete identifying information (composer, title, ...), instrument indicators, MIDI control signals, tempo variations, ...
- Keep only note-on and note-off information.
- A note, $k \in \mathbb{Z}$ half-tones above the average note is coded as a signed byte with value $k$.
- The whole piece is quantized in $0.05$ second intervals.
- Tracks are sorted according to decreasing average volume, and then output in succession.

Processed files $x_1, \ldots, x_n$ still sounded like the original.
Classify Music w.r.t. Composer

12 pieces of music: $4 \times \text{Bach} + 4 \times \text{Chopin} + 4 \times \text{Debussy}$. Class. by bzip2

Perfect grouping of processed MIDI files w.r.t. composers.
Further Applications

• Classification of Fungi
• Optical character recognition
• Classification of Galaxies
• Clustering of novels w.r.t. authors
• Larger data sets

See [Cilibrasi&Vitanyi’05]
The Clustering Method: Summary

- based on the universal similarity metric,
- based on Kolmogorov complexity,
- approximated by bzip2,
- with the similarity matrix represented by tree,
- approximated by the quartet method
- leads to excellent classification in many domains.
Exercises

1. [C20] Prove that $d(x, y) := \frac{\max\{K(x|y), K(y|x)\} - 1}{\max\{K(x), K(y)\}}$ is a metric.

Literature


