# Recent Experiences in Parameter-Free Data Mining

Kimihito Ito[1], Thomas Zeugmann[2⋆] and Yu Zhu[2]

[1] Research Center for Zoonosis Control
Hokkaido University, N-20, W-10 Kita-ku, Sapporo 001-0020, Japan
itok@czc.hokudai.ac.jp
[2] Division of Computer Science
Hokkaido University, N-14, W-9, Sapporo 060-0814, Japan
{thomas,yuar}@mx-alg.ist.hokudai.ac.jp

**Abstract.** Recent results supporting the usefulness of the normalized compression distance for the task to classify genome sequences of virus data are reported. Specifically, the problem to cluster the hemagglutinin (HA) sequences of influenza virus data for the HA gene in dependence on the host and subtype of the virus, and the classification of dengue virus genome data with respect to their four serotypes are studied.
A comparison is made with respect to hierarchical and spectral clustering via the kLine algorithm by Fischer and Poland (2004), respectively, and with respect to the standard compressors `bzlip`, `ppmd`, and `zlib`.
Our results are very promising and show that one can obtain an (almost) perfect clustering for all the problems studied.

## 1 Introduction

In many data mining applications the similarity between objects is of fundamental importance. Quite frequently, domain knowledge is used to define a suitable domain-specific distance measure. As a consequence, many of the resulting algorithms tend to have many parameters which have to be tuned. This is not only difficult but also including the risk of being biased. Furthermore, it may make it hard to verify the results obtained.

Recently, as a radically different approach, the paradigm of parameter-free data mining has emerged (cf. Keogh *et al.* [12]). The main idea of parameter-free data mining is the design of algorithms that have no parameters and that are universally applicable in all areas. At first glance this may seem impossible. How can an algorithm perform well if it is not based on extracting the important features of the data and if we are not allowed to adjust these parameters? As pointed out by Vitányi *et al.* [17], parameter free data mining is aiming at scenarios where we are not interested in a certain similarity measure but in *the* similarity between the objects themselves.

⋆ Supported by MEXT Grant-in-Aid for Scientific Research on Priority Areas under Grant No. 21013001.

The most promising approach to this paradigm uses Kolmogorov complexity theory [14] as its basis. The key ingredient is the so-called *normalized information distance* (*NID*) which was developed by various researchers during the past decade in a series of steps (cf., e.g., [2, 13, 8]). The intuitive idea behind it is as follows. If two objects are similar then there should be a simple description of how to transform each one of them into the other one. And conversely, if all descriptions for transforming each one of them into the other one are complex, then the objects should be dissimilar. Then, the *normalized information distance* between two strings $\mathtt{x}$ and $\mathtt{y}$ is defined as

$$NID(\mathtt{x},\mathtt{y}) \;=\; \frac{\max\{\mathsf{K}(\mathtt{x}|\mathtt{y}),\, \mathsf{K}(\mathtt{y}|\mathtt{x})\}}{\max\{\mathsf{K}(\mathtt{x}),\, \mathsf{K}(\mathtt{y})\}}\;, \tag{1}$$

where $\mathsf{K}(\mathtt{x}|\mathtt{y})$ is the length of the shortest program that outputs $\mathtt{x}$ on input $\mathtt{y}$, and $\mathsf{K}(\mathtt{x})$ is the length of the shortest program that outputs $\mathtt{x}$ on the empty input. For the technical details of the *NID*, we refer the reader to Vitányi *et al.* [17].

To apply this idea to data mining tasks, standard compression algorithms have to be invoked to approximate the Kolmogorov complexity $\mathsf{K}$. This yields the *normalized compression distance* (*NCD*) as approximation of the *NID* (cf. Definition 1). The *NCD* has been successfully applied to a variety of data mining problems (cf., e.g., [8, 12, 5, 6, 1]).

In this paper, we report the usefulness of the *NCD* for three classification problems for virus data. One task is to cluster the hemagglutinin (HA) sequences of influenza virus data for the HA gene in dependence on the subtype, where all data originate from the same host. The second task is the same classification but in dependence on the subtype *and* host of the virus. The third problem deals with the classification of dengue virus genome data with respect to their four serotypes.

## 2 Background and Theory

The definition of the *NID* depends on the function $\mathsf{K}$ which is *uncomputable*. Thus, the *NID* is *uncomputable*, too. Using a real-word compressor, one can approximate the *NID* by the *NCD* (cf. Definition 1). Again, we omit details and refer the reader to [17].

**Definition 1.** *The* normalized compression distance *between two strings* $\mathtt{x}$ *and* $\mathtt{y}$ *is defined as*

$$NCD(\mathtt{x},\mathtt{y}) \;=\; \frac{\mathsf{C}(\mathtt{x}\mathtt{y}) - \min\{\mathsf{C}(\mathtt{x}),\, \mathsf{C}(\mathtt{y})\}}{\max\{\mathsf{C}(\mathtt{x}),\, \mathsf{C}(\mathtt{y})\}}\;,$$

*where* $\mathsf{C}$ *is any given data compressor.*

Common data compressors are `bzlib`, `ppmd`, `zlib`, etc[3]. Note that the compressor $\mathsf{C}$ has to be computable and *normal* in order to make the *NCD* a useful approximation. This can be stated as follows.

[3] We use here the same naming convention as in the CompLearn Toolkit [4]. Essentially, these compressors coincide with `bzip2`, `ppmz`, and `gzip`.

**Definition 2 ([17]).** *A compressor* $C$ *is said to be* normal *if it satisfies the following axioms for all strings* $x$, $y$, $z$ *and the empty string* $\lambda$.

(1) $C(xx) = C(x)$ *and* $C(\lambda) = 0$; (identity)
(2) $C(xy) \geqslant C(x)$; (monotonicity)
(3) $C(xy) = C(yx)$; (symmetry)
(4) $C(xy) + C(z) \leqslant C(xz) + C(yz)$; (distributivity)

*up to an additive* $O(\log n)$ *term, with* $n$ *the maximal binary length of a string involved in the (in)equality concerned.*

Good real-world compressors like `bzlib`, `ppmd`, and `zlib` turned out to be normal for our data, and we used these compressors for our experiments. We used the `ncd` function from the CompLearn Toolkit (cf. [4]) to compute the distance matrix $D = \left(d^{ncd}(x, y)\right)_{x,y \in X}$, where $X = (x_1, \ldots, x_n)$ is the relevant data list.

To cluster the data we used hierarchical clustering and spectral clustering via kLines (cf. Fischer and Poland [9]). For a detailed description of the algorithms applied, we refer the reader to our paper [11].

## 3 Clustering Virus Data – Experiments and Results

The first paper using the *NCD* to analyze virus data was Cilibrasi and Vitányi [7]. In this paper the authors used the SARS TOR2 draft genome assembly 120403 from Canada's Michael Smith Genome Sciences Centre and compared it to other viruses by using the *NCD*and the `bzlib` compressor. After applying their quartet tree heuristic for hierarchical clustering, they obtained a ternary tree showing relations very similar to those shown in the definitive tree based on medical-macrobiological genomics analysis which was obtained later (see [7] for details).

Our first group of experiments dealt with influenza viruses, too. We have been interested in learning whether or not specific gene data for the hemagglutinin of influenza viruses are *correctly* classifiable by using the concept of the *NCD*. For any relevant background concerning the biological aspects of the influenza viruses we refer the reader to Palese and Shaw [16] and Wright *et al.* [18].

The family of *Orthomyxoviridae* is defined by viruses that have a negative-sense, single-stranded, and segmented RNA genome. There are five different genera in the family of *Orthomyxoviridae*: the influenza viruses A, B and C; *Thogotovirus*; and *Isavirus*. Influenza A viruses have a complex structure and possess a lipid membrane derived from the host cell.

We were only interested in their HA gene, since HA is the major target of antibodies that neutralize viral infectivity, and responsible for binding the virus to the cell it infects. In [11] we considered all 16 subtypes of the HA and collected a data set from the National Center for Biotechnology Information (NCBI) [15] containing a total of 106 sequences (all taken from viruses hosted by their the natural host) which could be (almost) successfully clustered into the relevant 16 subtypes of the HA. So, the HA subtype is *the* similarity between the different sequences.

Next, we shortly describe experiments dealing with influenza viruses hosted by duck and human. Note that H1N1 is a subtype of influenza A and the most common cause of influenza in humans. In June 2009, the World Health Organization declared that a new strain of `swine` origin H1N1 was responsible for the 2009 flu pandemic. Usually birds can pass avian influenza viruses to `swines`, where the viruses have to mutate so that they can circulate in the swine population. Then a new strain emerges which can be passed to humans or to other hosts. Of course, in order to become pandemic, the viruses may mutate again.

If one considers sequences for the HA gene originating from different hosts, it is only natural to ask which property is more "similar," the *host* or the *subtype*. For answering this question we chose 32 sequences having different HA subtypes that originated from both the duck and human host (again from NCBI). For a complete list of the data description we refer the reader to

`http://www-alg.ist.hokudai.ac.jp/nhuman_vs_duck.html` .

For the ease of presentation, below we use the following abbreviation for the data entries. Instead of giving the full description, e.g.,

>gi|218664152|gb|CY036815| /Human/4 (HA)/H2N2/South Korea/1968/// Influenza A virus (A/Korea/426/1968(H2N2)) segment 4, complete sequence

we refer to this datum as hH2N2CY036815 for short. The h stands for human here, and we use d if the host is the duck.

Each datum consists of a sequence of roughly 1800 letters from the alphabet {A, T, G, C}, e.g., looking such as

AAAAGCAGGGGAATTTCACAATTAAA . . . TGTATATAATTAGCAAA.

The results obtained by using the `zlib` and `bzlib` compressor and then applying hierarchical clustering are shown in Figure 1 and 2, respectively.

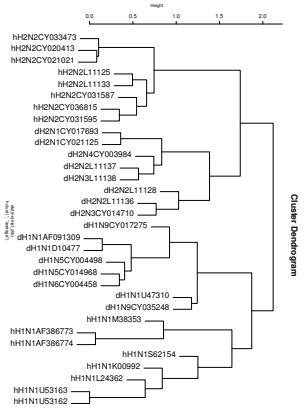


**Fig. 1.** Classification of HA sequences hosted by `human` and `duck`; compr.: `zlib`

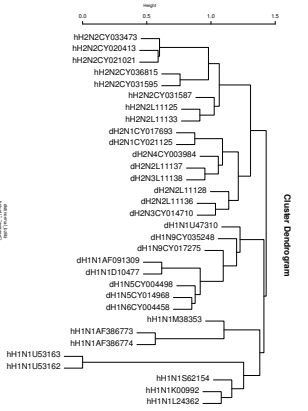


**Fig. 2.** Classification of HA sequences hosted by `human` and `duck`; compr.: `bzip`

As these clustering results show, for this data set the similarity between subtypes is stronger than the similarity between the hosts. We could confirm this outcome by using spectral clustering, where we used two clusters.

### 3.1 Clustering the *NCD* for Dengue Virus Data

Dengue virus is an RNA virus that causes dengue fever, one of the most important emerging diseases, infecting 100 million people annually in more than one hundred countries around the world [3]. The genome of dengue virus consists of nucleotides approximately 11 KB long, and 10 viral proteins are encoded in the genome. Dengue virus exhibits extensive genetic diversity, and there exist four antigenically distinct serologic types (1 through 4). It is known that severe cases, called dengue hemorrhagic fever / dengue shock syndrome, occur in patients who have secondary infections by a different serotype from previous infections [10]. Around 250,000 cases of dengue hemorrhagic fever / dengue shock syndrome are annually reported. Nucleotide sequences of all four dengue virus groups have been determined, and the rapid development of molecular biology over the last two decades is accelerating the accumulation of genomic data on the pathogen.

So, it is only natural to ask whether or not we can correctly cluster dengue virus genome data with respect to their four serotypes. To answer this question, we used 80 sequences (20 for each serotype) from NCBI ([15]). For a complete description of the data used, please see

`http://www-alg.ist.hokudai.ac.jp/Dengue-Data.html` .

Then, we computed the distance matrix as described above by applying the standard compressors `bzlib`, `ppmd`, and `zlib`. It should be noted that the dengue virus genome data are much larger than the influenza virus data, i.e., 10.6 KB versus 1.7 KB. Our hierarchical clustering was perfect for the compressors `ppmd`, and `zlib` (see Figure 3 for an example), but not for `bzlib`. Hierarchically clustering the distance matrix computed via the `bzlib` compressor gave 11 errors. On the other hand, spectral clustering delivered correct results in all three cases.

Moreover, we repeated these experiments with a non-balanced data set, see

`http://www-alg.ist.hokudai.ac.jp/imbalanced-dengue.html` ,

where we used 44 sequences of type 1 and 20 sequences of type 2, 3, and 4.

The results have been almost the same, i.e., hierarchical clustering and spectral clustering have been correct for the compressors `ppmd`, and `zlib`.

Using the `bzlib` compressor and spectral clustering as described in [11] produced two errors. However, by using a different kernel width for transforming the distance matrix in a similarity matrix (i.e., 1.23), the clustering was again perfect. Moreover, in contrast to the experiments performed with the influenza virus data, the kernel width was much less influential.

To summarize, our results are very promising and show that one can obtain an (almost) perfect clustering for all the problems studied. Note that we do not have reported the running time here, since it was in the range of several seconds. The clustering algorithms used in our experiments will nicely scale up to the amount of data for for which we can efficiently compute the distance matrix.

## References


[1] D. Benedetto, E. Caglioti, and V. Loreto. Language trees and zipping. *Phys. Rev. Lett.*, 88(4):048702–1–048702–4, 2002.

[2] C. H. Bennett, P. Gács, M. Li, P. M. B. Vitányi, and W. H. Zurek. Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423, 1998.

[3] D. S. Burke, G. Kuno, and T. P. Monath. Flaviviruses. In D. M. Knipe and P. M. Howley et al., editors, *Fields' Virology*, pages 1153–1252. Lippincott Williams & Wilkins, Philadelphia, fifth edition, 2007.

[4] R. Cilibrasi. The CompLearn Toolkit, 2003-. http://www.complearn.org/.

[5] R. Cilibrasi and P. Vitányi. Automatic meaning discovery using Google. Manuscript, CWI, Amsterdam, 2006.

[6] R. Cilibrasi and P. Vitanyi. Similarity of objects and the meaning of words. In *Theory and Applications of Models of Computation, Third International Conference, TAMC 2006, Beijing, China, May 2006, Proceedings*, volume 3959 of *Lecture Notes in Computer Science*, pages 21–45, Berlin, 2006. Springer.

[7] R. Cilibrasi and P. M. Vitányi. A new quartet tree heuristic for hierarchical clustering. In D. V. Arnold, T. Jansen, M. D. Vose, and J. E. Rowe, editors, *Theory of Evolutionary Algorithms*, number 06061 in Dagstuhl Seminar Proceedings. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany, 2006.

[8] R. Cilibrasi and P. M. B. Vitányi. Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545, 2005.

[9] I. Fischer and J. Poland. New methods for spectral clustering. Technical Report IDSIA-12-04, IDSIA / USI-SUPSI, Manno, Switzerland, 2004.

[10] S. B. Halstead. Pathogenesis of dengue: Challenges to molecular biology. *Science*, 239 (4839):476–481, 1988.

[11] K. Ito, T. Zeugmann, and Y. Zhu. Clustering the normalized compression distance for influenza virus data. In T. Elomaa, H. Mannila, and P. Orponen, editors, *Algorithms and Applications, Essays Dedicated to Esko Ukkonen on the Occasion of His 60th Birthday*, volume 6060 of *Lecture Notes in Computer Science*, pages 130–146. Springer, Heidelberg, 2010.

[12] E. Keogh, S. Lonardi, and C. A. Ratanamahatana. Towards parameter-free data mining. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 206–215. ACM Press, 2004.

[13] M. Li, X. Chen, X. Li, B. Ma, and P. M. Vitányi. The similarity metric. *IEEE Transactions on Information Theory*, 50(12):3250–3264, 2004.

[14] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer, 3rd edition, 2008.

[15] National Center for Biotechnology Information. Influenza Virus Resource, information, search and analysis. http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html.

[16] P. Palese and M. L. Shaw. Orthomyxoviridae: The viruses and their replication. In D. M. Knipe and P. M. Howley et al., editors, *Fields' Virology*, pages 1647–1689. Lippincott Williams & Wilkins, Philadelphia, fifth edition, 2007.

[17] P. M. B. Vitányi, F. J. Balbach, R. L. Cilibrasi, and M. Li. Normalized information distance. In *Information Theory and Statistical Learning*, pages 45–82. Springer, New York, 2008.

[18] P. F. Wright, G. Neumann, and Y. Kawaoka. Orthomyxoviruses. In D. M. Knipe and P. M. Howley et al., editors, *Fields' Virology*, pages 1691–1740. Lippincott Williams & Wilkins, Philadelphia, fifth edition, 2007.

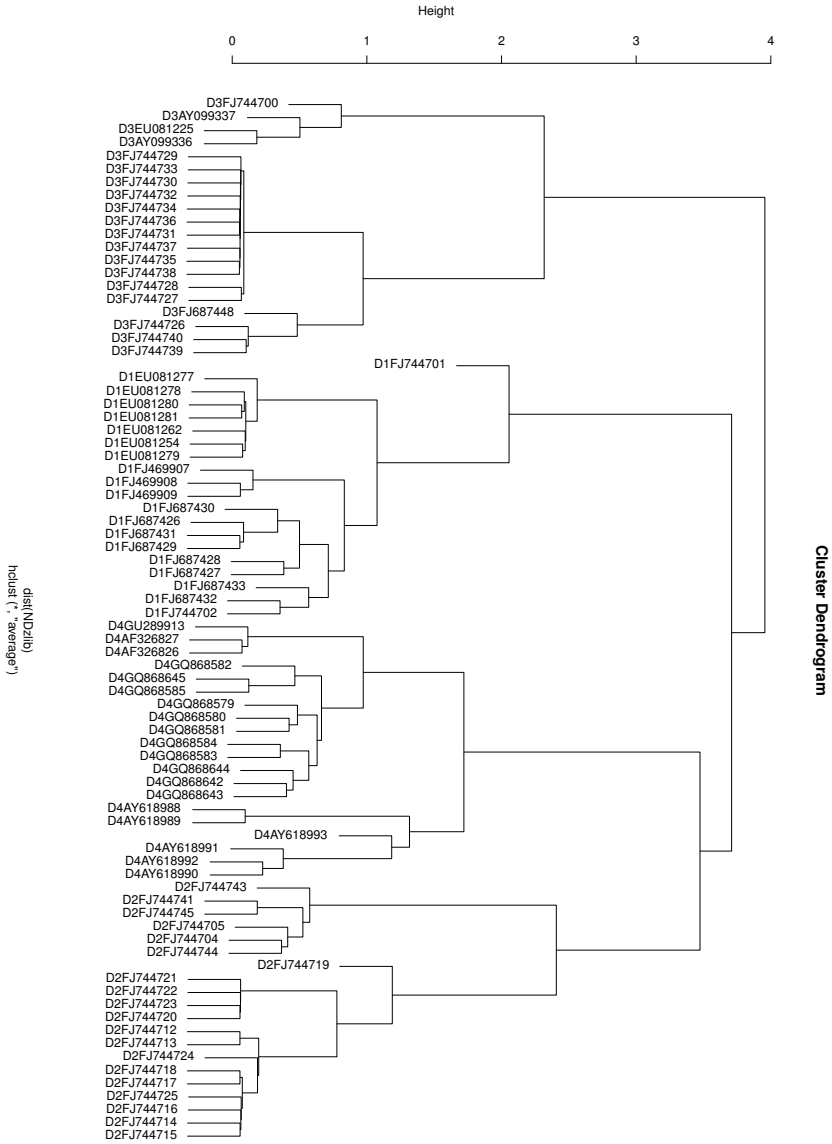


**Fig. 3.** Classification of dengue genome sequences; compr. `zlib`