F-052

Clustering the Normalized Compression Distance for Biological Data

Kimihito Ito[†] Thomas Zeugmann[‡] Yu Zhu [§]

Abstract. Our recently results supporting the usefulness of the normalized compression distance for the task to classify genome sequences of virus data are reported in this paper. Specifically, the problem to cluster the hemagglutinin(HA) sequences of influenza virus data for the HA gene in dependence 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 studies.

1. Introduction

The similarity between objects is of fundamental importance in everyday life. Quite frequently, domain knowledge is used to define a suitable domain-specific distance measure. So many data mining 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 be expensive, error prune and time consuming to arrive at a suitable tuning.

Recently, the approach of parameter-free data mining has emerged. This is aiming at scenarios where we are not interested in a certain similarity measure but in the similarity between the objects themselves. The most promising approach to this paradigm uses Kolmogorov complexity theory [7] as its basis. The key ingredient to this approach 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., [1, 3]).

More formally the *normalized information distance* between two strings x and y is defined as

$$NID(x,y) = \frac{\max\{K(x|y),\ K(y|x)\}}{\max\{K(x),\ K(y)\}}\ , \qquad (1)$$

where $K(x|y)$ is the length of the shortest program that outputs x on input y, and $K(x)$ is the length of the shortest program that outputs x on the empty input. We refer the reader to [8].

[†]Research Center for Zoonosis Control, Hokkaido University.
[‡]Division of Computer Science,Graduate School of Information Science and Technology, Hokkaido University.
[§]Division of Computer Science,Graduate School of Information Science and Technology, Hokkaido University.

Since its definition involves the Kolmogorov complexity $K(\cdot)$, the NID cannot be computed. Therefore, to apply this idea to real-world data mining tasks, standard compression algorithms, such as `gzip`, `bzip2`, or `PPMZ`, have been used as approximations of the Kolmogorov complexity.

To apply this idea to data mining tasks, standard compression algorithm have to be invoked to approximate the Kolmogorov complexity K. This yield the *normalized compression distance*(NCD) as approximation of the NID (cf. Definition 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

As explained in the Introduction,the definition of the NID depends on the function 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 [8].

Definition 1. *The* normalized compression distance *between two strings* x *and* y *is defined as*

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

where C *is any given data compressor.*

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

Definition 2 ([8]). *A compressor* C *is said to be* normal *if it satisfies the following axioms for all strings* x, y, z *and the empty string* λ.

(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`, `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 Tookit (cf. [2]) to computer the distance matrix $D = (d^{ncd}(x,y))_{x,y\in X}$, 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. [4]). For a detailed description of the algorithms applied, we refer the reader to [5].

3. Experiments and Results

3.1 Clustering the NCD for Influenza Viruses

Our first group of experiments dealt with influenza viruses. 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`.

Here, we shortly describe experiments dealing with influenza viruses hosted by duck and human. Usually birds can pass avian influenza viruses to swines, where the viruses have to mutate for being transmissible between swines. If one consider 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 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, in the following 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.

3.2 Clustering the NCD for Dengue Virus Data

Dengue viruses 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. 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 shock syndrome, occur in patients who have secondary infections by a different serotype from previous infections. So, it is 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. For a complete description of the data used, please see

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

Moreover, we repeated these experiments with a non-balannced 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.

Then we computed the distance matrix as described above by appllying the standard compressors `bzlib`, `ppmd` and `zlib`. Our hierarchical clustering was perfect for the compressors `ppmd` and `zlib`, while spectral clustering delivered correct results in all three cases. Details can be found in [6].

To summarize, our results are very promising and show that one can obtain an (almost) perfect clustering for all the problems studied.

References

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

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

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

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

[5] K. Ito, T. Zegumann, and Y. Zhu. Clustering the normalized compression distance for virus data. In T. Elomaa, H. Mannila, and P. Orponen, editors, *Algorithms and Applications*, volume 6060 of *Lecture Notes in Computer Science*, pages 130–146. Springer, Heidelberg, 2010.

[6] K. Ito, T. Zegumann, and Y. Zhu. Rencent experiences in parameter-free data mining. *In proceedings of 25th International Symposium on Computer and Information Sciences*, 2010.

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

[8] 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.