The Model based Similarity Metric

Lionel Gueguen[+], Mihai Datcu[*+]
[+]GET-Télécom Paris, 46 rue Barrault, 75013 Paris, France
[*]German Aerospace Center DLR, Oberpfaffenhofen, D-82234 Wessling, Germany

This paper addresses the problem of building a compact representation of objects which enables to define a similarity measure. The key idea is to provide a mean to extract and represent the relevant information of an object. By integrating the Minimum Length Description to an informational similarity metric based on Kolmogorov complexity [1], we propose a new informational measure based on models. First, we make the strong assumption that for any object x there exists a model $\mathcal{M}_x$ such that $K(x) = K(x \mid \mathcal{M}_x) + K(\mathcal{M}_x)$. So, the Li's similarity metric between two objects x and y is reformulated, in the case when $K(x) \leq K(y)$, by:

$$d(x, y) = \alpha \frac{K(x, y \mid \mathcal{M}_{x,y}) - K(x \mid \mathcal{M}_x)}{K(y \mid \mathcal{M}_y)} + (1 - \alpha) \frac{K(\mathcal{M}_{x,y}) - K(\mathcal{M}_x)}{K(\mathcal{M}_y)} \tag{1}$$

$$\alpha = \frac{K(y \mid \mathcal{M}_y)}{K(y \mid \mathcal{M}_y) + K(\mathcal{M}_y)} = \frac{K(y \mid \mathcal{M}_y)}{K(y)} \tag{2}$$

One can notice that the similarity metric is divided in two parts as in the MDL principle. The right part measures the similarity between objects expressed in their models, while the second one measures the similarity between models. Moreover, when α goes to 0, the similarity metric is more based on the models complexity. On the contrary, when α tends to 1, the similarity metric takes into account mostly the similarity between the randomness of objects. The object y is highly random when $\alpha \approx 1$, while it is highly deterministic when $\alpha \approx 0$. Consequently, it appears natural that the similarity metric takes into account the randomness of an object to compare it to others. Secondly, we make the assumption that an object y is independent of the randomness of an object x such that $K(y \mid x) = K(y \mid \{x \mid \mathcal{M}_x, \mathcal{M}_x\}) = K(y \mid \mathcal{M}_x)$. Thus, if we consider $y \mid M_x$ as a new object z with its associated model M_z, the similarity metric becomes:

$$\delta(x, y) = \alpha \frac{K(z \mid \mathcal{M}_z)}{K(y \mid \mathcal{M}_y)} + (1 - \alpha) \frac{K(\mathcal{M}_z)}{K(\mathcal{M}_y)} \tag{3}$$

In conclusion, representing an object by its model enables to compare it to another using the proposed similarity. In addition, the model constitutes a compact representation of objects.

References

[1] Ming Li, Xin Chen, Xin Li, Bin Ma, and P.M.B Vitanyi, "The Similarity Metric," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3250–3264, Dec. 2004.