Code2Vect: an efficient representation and classification technique for heterogeneous data

Clara Argerich Martin*, Rubén Ibanez Pinillo[†] and Francisco Chinesta**

*ENSAM ParisTech
151 Boulevard de l'Hôpital, F-75013 Paris, France.
Clara.argerich_martin@ensam.eu
** ESI GROUP Chair @ ENSAM ParisTech
151 Boulevard de l'Hôpital, F-75013 Paris, France.
[†] IcI Ecole Centrale de Nantes
1, rue de la Noe, 44321 Nantes Cedex 3, France.

ABSTRACT

Nowadays there are many Big Data techniques such as neural networks, regression trees, t-sne [1,2,3] that are widely used to classify data and obtain this data behaviour. One of the newest works on the field was developed by Google [4] in order to codify large texts coming with a representation of a word as a neighbour-weighted vector. Based on this algorithm we present in our work a new classification technique for heterogeneous, non-linear and high-dimensional data.

In this work we develop a mathematical tool that will enable the user to represent a given set of data according to a chosen objective or target. For that purpose we picture a parametric space, figure 1, left, where ideally, each parameter defines an axis. The first problem given this type of conventional representation: usually there will be more parameters/axes that what humans can visualize. We will call the data in this space Inputs, represented by letter **y** [dy].

The second problem that arises is the comparison of parameters whose nature is completely different. In order to circumvent this issues we aim to build a non-linear mapping operator, **W(y)** that will enable the connection between the input space and the vector space, depicted in figure 1, right-side, and defined as **x** [dx].

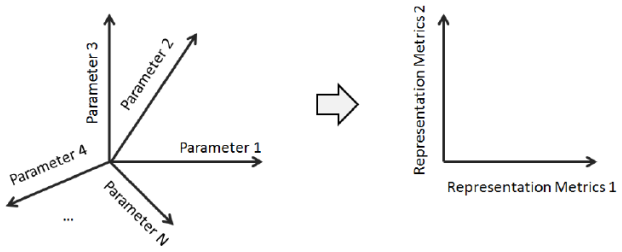

Figure 1. Input Space (left) and representation space (right).

Moreover, this vector space will be built according to a given output or target, so metrics at this space will be constraint by the output.
The operator **W(y)** is a matrix of dimension [dx × dy] with non-linear entries depending on **y** and it is built by enforcing metric constraints that derive from the target or output. Indeed, data appearing in the input space will be ordered and clustered in the representation space according to the metrics defined in the target space. The representation of data according to a requirement will be achieved. In order to illustrate the applications of the algorithm it will be tested for two examples regarding a manufacturing engineering framework, one highly non-linear data as an inputs and a second one regarding the representation of high-dimensional data.

REFERENCES

[1] S. Roweis and L. Saul . "Nonlinear dimensionality reduction by locally linear embedding". *Science* v.290 no.5500, pp.2323--2326, 2000..

[2] L. Vans Der Maaten and G. Hinton . "Visualizing Data using t-SNE". *Journal of Machine arning Research* 9, 2579-2605, 2008.

[3] Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin. "A Neural Probabilistic Language Model" *urnal of Machine Learning Research 3*, 1137-1155, 2003.

[4] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean. "Distributed Representation of Words and Phrases an their Compositionality". *ProceedingNIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems*, Volume 2 2013