Devising and experimenting correlation-based metrics for evaluating the effectiveness of input encoding techniques in prediction tasks

G. Armano, E. Tamponi✉

Department of Electrical and Electronic Engineering, University of Cagliari, Italy

Motivations

Defining an optimal encoding for input data is fundamental to achieve high performances in prediction tasks. Its main responsibility is to transform input data to a format suitable for the classification algorithm. The selection of the best encoding is typically done by resorting to the knowledge of a human expert, entrusted with extracting the features that s/he deems useful for the task. In some cases, the evaluation of an encoding can be performed by heavyweight tests, where most of the computational effort is spent to train classifiers. Furthermore, these tests may introduce a bias due to many factors, which depend on the adopted learning technique rather than on the encoding under analysis. To overcome these problems, we propose to investigate the correlation between the input variables (whose actual values depend on the adopted encoding technique) and the output variables (which encode the labels associated with each input). According to this insight, we devised and experimented some ""correlation-based metrics"" aimed at evaluating the encodings used for prediction tasks. The availability of efficient metrics would be particularly useful in domains where the selection of the right encoding is crucial, such as in the prediction of biomedical data. For example, in the prediction of protein secondary structure, different encodings greatly affect the overall performance of a system. In particular, multiple alignments of amino acid sequences bring an exceptional performance increase with respect to one-hot encoding.

Methods

We started our research by finding suitable methods for evaluating the input-output correlation. An obvious starting point was the Pearson product-moment correlation coefficient, but we found many more algorithms that resulted more suitable for our needs. In particular, we tested the novel distance correlation coefficient and the generalized correlation ratio. A fundamental part of the metrics is to extract a ""synthetic value"" from the correlation matrices obtained during the input-output correlation analysis. To evaluate our system, we selected a particularly difficult domain, the prediction of protein secondary structure. We measured the performance of 30 input encodings using our metrics. For each encoding, we compared the performance predicted by our metrics with the effective results obtained by actual prediction systems. To implement and test the system, we used the GAME framework.

Results

The tests showed that our metrics predicted the performance of actual prediction systems with high accuracy. In particular, the metrics based on the generalized correlation ratio resulted very effective and fast (more than 1000 times faster than traditional performance tests).

Availability

http://iasc2.diee.unica.it/GAME/