C++ implementation, optimization and application of the focusing Belief Propagation algorithm

Daniele Dall'Olio[1], Nico Curti[1,2], Gastone Castellani[1,2], Armando Bazzani[1,2], and Daniel Remondini[1,2]

[1]Department of Physics and Astronomy, University of Bologna
[2]INFN Bologna

Abstract

Nowadays the machine learning researchers' community is mostly focused on the application of algorithms which are trained on large datasets. The most common tendency is to collect a huge dataset and to supply the chosen learning algorithm with a large fraction of this dataset, in the attempt of teaching an underlying nature of the data.

The idea of learning from a small portion of data appears infeasible, but Belief Propagation based algorithms introduced in the last decade [1, 2] may be exploited to question this largely accepted opinion. We mined one of these new algorithms called replicated focusing Belief Propagation (rfBP). The rfBP aims both to partly explain the reason why pure heuristic algorithms perform optimally in both time and outcomes, and to show an alternative way of designing effective algorithms [3].

We introduce a new C++ library implementing the replicated focusing Belief Propagation and its associated Python wrapper. This implementation is optimized for parallel computing and is endowed with a newly written C++ library called *scorer*, which is able to compute a large number of statistical measurements based on a hierarchical graph scheme. With this optimized implementation we believe we can encourage researchers to approach these alternative algorithms and to use them more frequently in real context.

We tested the rfBP on EU's COMPARE project data. The dataset was composed of 210 Salmonella enterica genome sequences, 8189 bases long, living inside animals. Our early goal was to discriminate those bacteria living in pigs (159 samples) with respect to all the others animals (51 samples). We first set the training set size to 25% of the total (leaving all rest to the testing set) and then we performed a stratified K-fold.

		Actual label	
		P	**N**
Predicted label	**P**	111.35	6.55
	N	8.65	31.45
Total		120	38

Despite the unusual sizes exchange between training and testing set, the fBP performs on average with an accuracy of about 90% and a Matthews Correlation Coefficient of about 0.75. This early result on daily analyzed data suggests that the learning of data's underlying nature can be feasible even from small datasets. This motivates us to explore further applications of this algorithm in multiple fields.

References

[1] Braunstein A. and Zecchina R. Learning by message passing in networks of discrete synapses. 2006.

[2] Baldassi C. et al. Efficient supervised learning in networks with binary synapses. 2007.

[3] Baldassi C. et al. Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes. 2016.