VIRTUAL LEAVE-ONE-OUT ESTIMATION OF GENERALIZATION ERROR FOR GRAPH MACHINES AND RECURSIVE NETWORKS

Aurélie GOULON, Arthur DUPRAT, Gérard DREYFUS
Aurelie.Goulon@espci.fr, Arthur.Duprat@espci.fr, Gerard.Dreyfus@espci.fr
ESPCI-Paristech, Laboratoire d'Électronique
10, rue Vauquelin
75005 PARIS – France
http://www.neurones.espci.fr

Virtual leave-one-out is an attractive alternative to cross-validation for estimating the prediction error of models, especially nonlinear ones: training is performed on the available data, and an estimate of the prediction error that would have been incurred on each example, if it had been withdrawn from the training set, is computed. For linear models, the virtual leave-one-out score reduces to the PRESS statistic. The computation of the leave-one-out score involves the computation of the *leverage* of each example, which indicates the influence of each example on the model[1].

There is a growing interest in graph machines and recursive networks, which learn from structured data, i.e. examples that are described by graphs instead of vectors[2]. Graph machines encode the structure of the graphs and simultaneously provide a prediction of the properties of interest. Therefore the representation of the structured data is learnt together with the learning of the task, which exempts the model designer from finding a vector representation for the data.

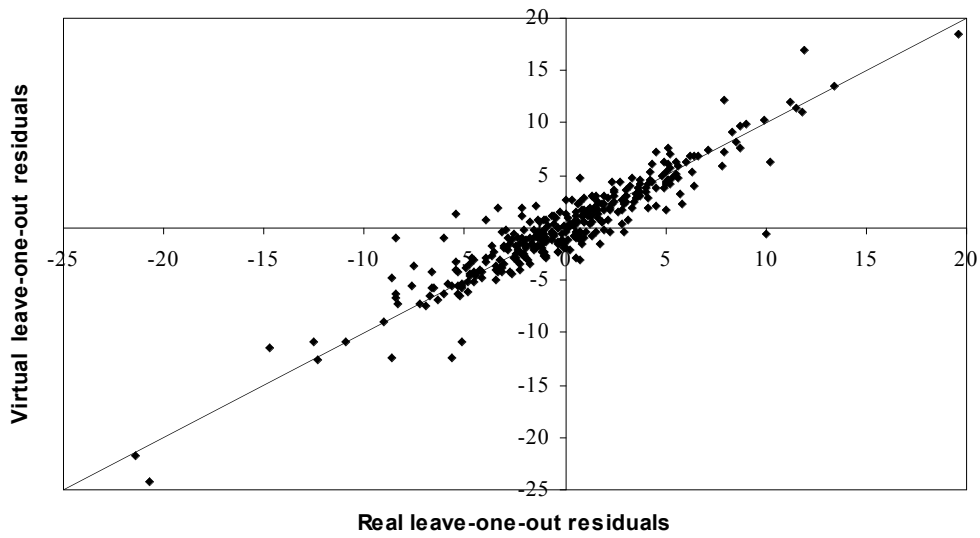

Figure 1: Comparison of the real and virtual leave-one-out residuals for the prediction of the boiling point of 330 haloalcanes

[1] Bates, D.M., Watts D.G.: Nonlinear Regression Analysis and its Applications. John Wiley and Sons (1998).

[2] Goulon, A., Duprat, A., Dreyfus, G.: "Graph machines and their applications to computer-aided drug design: a new approach to learning from structured data", Unconventional Computation 2006, Lecture Notes in Computer Science, 4135, 1-19 (2006).

We show how the computation of leverages, hence of the virtual leave-one-out score, can be extended to graph machines and recursive networks. We compare the real and virtual leave-one-out scores on several data sets (Figure 1). We describe examples of graph machine selection by virtual leave-one-out, and we show that, in addition, virtual leave-one-out can provide insight into the design of the predictors, i.e. the encoding of the input data into directed acyclic graphs. We illustrate these topics on several regression tasks, e.g. the estimation of the Gibbs free energy of solvation of molecules[3], the toxicity of halogenated aliphatic compounds[4], or the agonist activities of ecdysteroids[5].

Topics: estimation, prediction, and sequence modeling
Preference: poster

---

[3] Bernazzani, L., Duce, C., Micheli, A., Mollica, V., Sperduti, A., Starita, A., Tiné, M.R.: "Predicting physical-chemical properties of compounds from molecular structures by recursive neural networks", Journal of Chemical Information and Modeling, 46, 2030-2042 (2006).

[4] Öberg, T.: "A QSAR for baseline toxicity: validation, domain of application, and prediction", Chemical Research in Toxicology, 17, 1630-1637 (2004).

[5] Ravi, M., Hopfinger, A.J., Hormann, R.E., Dinan, L.: "4D-QSAR Analysis of a set of ecdysteroïds and a comparison to CoMFA modelling", Journal of Chemical Information and Computer Sciences, 41, 1587-1604 (2001).