GRAPH MACHINES AND THEIR APPLICATIONS TO COMPUTER-AIDED DRUG DESIGN: A NEW APPROACH TO LEARNING FROM STRUCTURED DATA

Aurélie GOULON, Arthur DUPRAT, Gérard DREYFUS
agoulon@libertysurf.fr, Arthur.Duprat@espci.fr, Gerard.Dreyfus@espci.fr

ESPCI, Laboratoire d'Électronique
10, rue Vauquelin
75005 PARIS – France
http://www.neurones.espci.fr

The recent developments of statistical learning focused on *vector machines,* i.e. on machines that learn from examples that are described by vectors of features. However, there are many fields where structured data must be handled; therefore, it would be desirable to learn from examples described by *graphs.* Obviously, a very active field where learning from graphs is highly desirable is Quantitative Structure-Activity Relationships (QSAR) and Quantitative Structure-Property relationships (QSPR), which aim at predicting activities and properties of molecules from their structures, which are naturally in the form of graphs.

The presentation will describe *graph machines*, which learn real numbers from graphs. Basically, for each input graph, a separate learning machine is built, whose structure contains the same information as the graph. If the graph is a tree, the machine is made of elementary, identical machines, which are connected together in the same way as the nodes of the graph are connected by its edges; the output of the machine is the quantity to be modeled; the input of the machine is a constant. Therefore, the basic difference between conventional learning machines and graph machines is the following: instead of training a single machine (e.g. a neural network) on several different input-output pairs, one trains several different machines on one output each. Because all graph machines are combinations of the same elementary machine, the number of parameters is kept low. Usual cross-validation or leave-one techniques extend simply to graph machines, as well as bootstrapping and bagging.

Recently developed methods handle cyclic graphs as well as trees, and accommodate features that may be different for different nodes. Typically, for QSAR applications, they allow handling molecules of arbitrary complexity, with cycles, heteroatoms, various types of bonds and of structural asymmetry.

The presentation will first describe the principles of graph machines and how they relate to the general problem of structured data representation. The second part of the talk will describe applications to real QSAR-QSPR problems.