UNIVERSITÀ DEGLI STUDI DI FIRENZE
Dipartimento di Sistemi e Informatica

Dottorato di Ricerca in
Ingegneria Informatica e dell'Automazione
XVI Ciclo

Kernel Methods, Multiclass Classification and Applications to Computational Molecular Biology

Andrea Passerini

Dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
in Computer and Control Engineering

Ph.D. Coordinator
Prof. Edoardo Mosca

Advisors
Prof. Paolo Frasconi

Prof. Giovanni Soda

Anno Accademico 2003-2004

Abstract

Support Vector Machines for pattern recognition were initially conceived for the binary classification case. A common approach to address multi-classification problems with binary classifiers is that of reducing the multiclass problem to a set of binary sub-problems, and combine their predictions in order to obtain a multiclass prediction. Reduction schemes can be represented by Error Correcting Output Codes, while binary predictions can be combined using a decoding function which outputs a score for each possible class. We propose a novel decoding function which computes the conditional probability of the class given the binary predictions, and present an extensive set of experiments showing that it outperforms all the decoding functions commonly used in practice. An alternative approach for solving multiclass problems is that of directly extending binary classification algorithms to the multiclass case. Various multicategory extensions for SVM have been proposed so far. We review most of them showing their similarities and differences, as well as the connections to the ECOC schemes. We report a series of experiments comparing different multiclass methods under various conditions, showing that while they perform similarly at the optimum, interesting differences emerge when forced to produce degenerate solutions. Moreover, we present a novel bound on the leave-one-out error of ECOC of kernel machines, showing that it can be successfully employed for model selection.

In the second part of this thesis, we present applications of kernel machines to problems in computational molecular biology. We address the problem of predicting disulphide bridges between cysteines, which represent strong constraints on proteins 3D structure, and whose correct location can significantly help the overall folding prediction. The problem of disulphide bridges prediction can be divided in two successive steps: firstly, for each cysteine in a given protein, predict whether it is involved or not in a disulphide bond; secondly, given the subset of disulphide bonded cysteines in the protein, predict their connectivity pattern, by coupling each cysteine with the correct partner. We focus on the first step, and develop state-of-the-art learning algorithms combining kernel machines and connectionist models. Disulphide bridges are not the only type of binding a cysteine can be involved in, and many cysteines actually bind different types of ligands, usually including metal ions, forming complexes which play very important roles in biological systems. We employed kernel machine algorithms to learn to discriminate between ligand bound and disulphide bound cysteines, in order to get deeper insights into the role of each cysteine in a given protein. We developed ad-hoc kernels able to exploit information on residues similarity, showing that they obtain performances similar to standard kernels with much simpler models.