UNIVERSITÀ DEGLI STUDI DI FIRENZE
Facoltà di Ingegneria – Dipartimento di Sistemi e Informatica

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

Learning Preference and Structured Data: Theory and Applications

Sauro Menchetti

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

Ph.D. Candidate
Sauro Menchetti

Ph.D. Coordinator
Prof. Edoardo Mosca

Advisor
Prof. Paolo Frasconi

Anno Accademico 2004–2005

Abstract

This dissertation deals with the theory and applications to natural language processing and computational molecular biology of learning preference and structured data.

From a theoretical point of view, a new and unpublished interpretation in the dual space of the voted perceptron algorithm is provided, including an on–line update rule and an upper bound for dual variables. Accordingly, a novel formulation of regularization theory for this algorithm is devised.

A further new theoretical analysis based on a partial order model of preference and ranking problems, explains why a setwise loss function which directly tackles the problem exhibits a better performance of a pairwise loss function based on an utility function. In the context of preference learning, we report applications to two large scale problems involving learning a preference function that selects the best alternative in a set of competitors: reranking parse trees generated by a statistical parser and the prediction of first pass attachment under strong incrementality hypothesis. We compare convolution kernels and recursive neural networks, two effective approaches to solve the investigated problems, finding that the choice of the loss function plays an essential role.

A novel, general and computationally efficient family of kernels on discrete data structures called weighted decomposition kernels is developed within the general class of decomposition kernels. We report experimental evidence that the proposed family of kernel is highly competitive with respect to more complex and computationally demanding state–of–the–art methods on a set of practical problems in bioinformatics, involving protein sequence and molecule graph classification.

Finally, we tackle the prediction task of the zinc binding sites and proteins that is still a little widespread problem in the machine learning community. We propose an ad–hoc remedy to the autocorrelation problem between residues close in sequence. This approach lead to a significant improvement in the prediction performance by modelling the linkage between examples in such a way that sequentially close pairs of candidate residues are classified as being jointly involved in the coordination of a zinc ion. We develop a kernel for this particular type of data that can handle variable length gaps between candidate coordinating residues.
Keywords: Structured Data, Preference Learning, Kernel Machines.