

PROBABILISTIC MULTI-CLASS MULTI-KERNEL LEARNING: ON PROTEIN FOLD RECOGNITION AND REMOTE HOMOLOGY DETECTION

THEODOROS DAMOULAS
INFERENCE GROUP
DEPARTMENT OF COMPUTING SCIENCE
UNIVERSITY OF GLASGOW

ABSTRACT. The problems of protein fold recognition and remote homology detection have recently attracted a great deal of interest as they represent challenging multi-feature multi-class problems for which modern pattern recognition methods achieve only modest levels of performance. As with many pattern recognition problems, there are multiple feature spaces or groups of attributes available, such as global characteristics like the amino-acid composition (C), predicted secondary structure (S), hydrophobicity (H), van der Waals volume (V), polarity (P), polarizability (Z), as well as attributes derived from local sequence alignment such as the Smith-Waterman scores. This raises the need for a classification method that is able to assess the contribution of these potentially heterogeneous object descriptors while utilizing such information to improve predictive performance. To that end, we offer a single multi-class kernel machine that informatively combines the available feature groups and, as is demonstrated in this paper, is able to provide the state-of-the-art in performance accuracy on the fold recognition problem. Furthermore, the proposed approach provides some insight by assessing the significance of recently introduced protein features and string kernels. The proposed method is well-founded within a Bayesian hierarchical framework and a variational Bayes approximation is derived which allows for efficient CPU processing times.

RESULTS:

The best performance which we report on the SCOP PDB-40D benchmark data-set is a 70% accuracy by combining all the available feature groups from global protein characteristics but also including sequence-alignment features. We offer an 8% improvement on the best reported performance that combines binary SVM classifiers while at the same time reducing computational costs and assessing the predictive power of the various available features. Furthermore, we examine the performance of our methodology on the SCOP 1.53 benchmark data-set that simulates remote homology detection and examine the combination of various state-of-the-art string kernels that have recently been proposed.