Robust Ensemble Learning for Data Mining

The field of machine learning is expanding at a rapid pace, especially in medicine. This pace is being driven by an information avalanche that is unprecedented. To handle these data, specialized research databases and metadatabases have been established in many domains. Machine learning technology applied to many of these databases has the potential of revolutionizing scientific knowledge. In the area of bioinformatics, for example, many large scale sequencing projects have produced a tremendous amount of data on protein sequences. This has created a huge gap between the number of identified sequences and the number of identified protein structures. Machine learning methods capable of fast and accurate prediction of protein structures hold out the promise of not only reducing this gap but also of increasing our understanding of protein heterogeneity, protein-protein interactions, and protein-peptide interactions, which in turn would lead to better diagnostic tools and methods for predicting protein/drug interactions.

What are needed to handle the problems of today are not yesteryear's solutions, which were typically based on training a single classifier on a set of descriptors extracted from a single source of data. It is generally acknowledged that ensembles are superior to single classifiers, and much recent work in machine learning has focused on methods for building ensembles. In protein prediction some powerful ensembles have recently been proposed that utilize the combined information available in multiple descriptors extracted from protein representations. Particularly interesting, however, are ensemble systems combining multiple descriptors extracted from many protein representations that are trained across many databases. In this address, I shall describe the ensemble research that I am involved in with Drs. Loris Nanni and Alessandra Lumini. My focus will be on describing the methods our research group uses for building ensemble systems that work extremely well across multiple databases. Powerful general purpose ensembles are of value to both the general practitioner and expert alike. Such classifier systems can serve as a base for building systems optimized for a given problem. Moreover, general purpose ensembles can further our general understanding of the classification problems to which they are applied.