Multi-classifier systems for Bioinformatics problems

Vasile Palade
Oxford University, Computing Laboratory,
Parks Road, OX1 3QD, Oxford, United Kingdom.
Telephone: +44-01865 - 283606
Fax: +44-01865 - 273839
E-mail: Vasile.Palade@comlab.ox.ac.uk

Abstract:

Multi-Classifier Systems have fast been gaining popularity among researchers working in machine learning and applications for their ability to fuse together multiple classification outputs for better accuracy and classification. This talk is concerned with current issues in the design of multi-classifier systems and presents our multi-classifier developments for several Bioinformatics problems.

The talk first presents some important issues in the design of multi-classifier systems, with a focus on the diversity and combination of the outputs of individual classifiers. Few diversification and combination schemes are presented. Then, a neural network based multi-classifier system for the identification of Escherichia Coli (E.Coli) promoter sequences in strings of DNA is presented. A data set containing known E.Coli promoter and non-promoter sequences were encoded using different encoding methods. The encoded sequences were then used to train multiple neural networks. Optimal neural network configurations and encoding methods are determined using genetic algorithms. We then move onto recognizing DNA sequences for the Human DNA data set (http://www.fruitfly.org/sequence/human-datasets.html). We introduce a genetic algorithm based optimization of the standard majority voting combiner. A collection of novel functions for feature extraction from DNA data are used for the construction of data sets used for training and testing the classifiers. In a later extension, a selection of the classifiers to be considered in the combiner part of the multi-classifier system is done using genetic algorithms. The talk also presents guidelines for the selection of different training paradigms and performance metrics based on the properties and distribution of the genomic data.

The presentation will then proceed with introducing an ensemble of neuro-fuzzy networks for micro-array cancer gene expression data classification. Neuro-fuzzy ensemble approach not only provides good classification results, but the behaviour of neuro-fuzzy models can be explained and interpreted in human understandable terms. At the end of the talk, an analysis on the performance of several classifier fusion techniques in a protein secondary structure prediction problem will be provided. The presented approaches and results prove that ensembles of classifiers can be used as effective computational tools in solving difficult Bioinformatics problems.