Machine Learning and Applications in Biology

Track 11. Bioinformatics
Session 2. Extracting Biology from Bioinformatics Algorithms and Databases [*]

Hyunjung (Helen) Shin
Dept. of Industrial & Information Systems Engineering
Ajou University
shin@ajou.ac.kr

14.Sep.2007

Abstract

The emergence of the fields of computational biology and bioinformatics has alleviated the burden of solving many biological problems, saving the time and cost required for experiments and also providing predictions that guide new experiments. Within computational biology, machine learning algorithms have played a central role in dealing with the flood of biological data. The goal of this tutorial is to raise awareness and comprehension of machine learning so that biologists can properly match the task at hand to the corresponding analytical approach. We start by categorizing biological problem settings and introduce the general machine learning schemes that fit best to each or these categories. We then explore representative models in further detail, from traditional statistical models to recent kernel models, presenting several up-to-date research projects in bioinfomatics to exemplify how biological questions can benefit from a machine learning approach. Finally, we discuss how cooperation between biologists and machine learners might be made smoother.

[*] BioKorea 2007, vol. 3, pp. 337–366, COEX, Seoul, Korea

Content

1. Tasks:
   - Prediction
     - Classification (batch or incremental)
     - Regression (batch or incremental)
   - Description
     - Clustering
     - Feature Description
   - Feature Selection (Extraction)
   - Sample Selection
   - Data Integration
2. Learning Schemes:
   - Supervised
   - Unsupervised
   - Semi-Supervised
3. Models
   - Traditional Statistical Models: Regression, *Logistic Regression*, *Discriminant Analysis*, *Principal Component Analysis (PCA)*, Canonical Correlation Analysis (CCA), Factor Analysis (FA), ANalysis Of VAriance (ANOVA), *k-Means Clustering*, Agglomorative Clustering, Stratified Sampling, etc.
   - Neural Networks: Feed-Forward Network, Self-Organized Map, etc.
   - Decision Trees: CART, CHAID, *C4.5*, QUEST, FOREST, etc.
   - Kernel Methods: *Support Vector Classifier/Regressor (SVC/SVR)*, *kPCA*, kCCA, Independent Component Analysis (ICA), etc.
   - Ensemble Methods: Bagging, Boosting, Arcing, etc.
   - Semi-Supervised Learning (SSL) and Transductive Inference Methods: *Graph-based Semi-Supervised Learning*, etc.
   - Generative (Probabilistic) Methods: Bayesian Models, Näive Bayes, Position Weighted Matrix (PWM), Gaussian Process, etc.
4. Evaluation Measurement
   - Error Rate
   - Receiver Operating Characteristic (ROC) score
   - Sensitivity/Specificity, True/False Alarm

5. Examples:
    - Alternative Splicing with SVM – supervised
    - Protein Class Classification with SSL – semi-supervised and data integration
    - Microarray Analysis with Bi-clustering – unsupervised
6. How to Communicate with Machine Learners?
    - Objective of Task
    - Data Availability and Format
        - Number and Type of Features
        - Number of Labelled Observations
        - Number of Unlabelled Observations
    - Prior Knowledge based on Previous Research in Biology
    - Model Building
    - Biological Interpretation/Implication on the Result