A Comparison of ILP and Feature-based Approaches

Prediction of Mutagenicity within Nitro-Aromatic Compounds

General Approach

- View mutagenicity as a biological activity, and attempt to find structure-activity relationships (SAR analysis).
- Earliest such approach used linear regression (as with earliest approaches to SAR in general).

Data

- 230 nitro-aromatic compounds (many found in car exhaust, for example).
- 138 of these are mutagenic. (They have been found to cause DNA mutations in rodents. Therefore there exists a reasonable chance they will be carcinogenic in humans.)

Data (Continued)

- Feature-based algorithms will use those features that were found helpful in linear regression.
- ILP will use atom-and-bond structure as well as definitions of chemical groups as background knowledge. It also will use the features provided to the feature-based algorithms.

Features Used

- *logP*: measure of hydrophobicity.
- *eLUMO*: energy level of the lowest unoccupied molecular orbital.
- *I1*: Boolean variable indicating presence of 3 fused rings.
- *Ia*: Boolean variable indicating whether the compound belongs to the acenthrylene family.

Algorithms

- Linear regression
- Neural network (backpropagation)
- Decision tree (Ind implementation of CART algorithm)
- Prolog implementation of PROGOL (successor is ALEPH).

Methodology

- Use an earlier split of the data into 42 "regression-unfriendly" compounds and the other 188 compounds.
- Use leave-one-out cross-validation on the 42, and use 10-fold cross-validation on the 188.
- Apply each algorithm to each set. Use McNemar's Test to evaluate significance.

Methodology (Continued)

- With ILP, use two types of background knowledge. S1: atom and bond information only. S2: S1 + definitions of groups of interest.
- Question: Why not use groups of interest as features for feature-based techniques?
- Question: Since S1 can capture everything in S2, what is added value of S2?

Results

- Progol did *not* significantly outperform IndCART, and it significantly underperformed with S1 only.
- With S2, Progol performed as well as the best of the other algorithms (IndCART).
- One of Progol's rules was interesting to domain experts---a possible new structural alert for mutagenicity? *Comprehensibility*.