CMSC435 Group 9 First Predictions - Model Masters

Github Repo: https://github.com/SonLee17/ModelMasters

Features Generated:

- Amino Acid Frequency (ACDEFGHIKLMNPQRSTVWY): 20 features
- Positive
- Polarity
- Nonpolarity
- Cyclic
- Aromaticity
- Basicity
- Hydrophobicity
- Hydrophilicity
- Sulfur_content
- Small
- Large

Total Features: 31

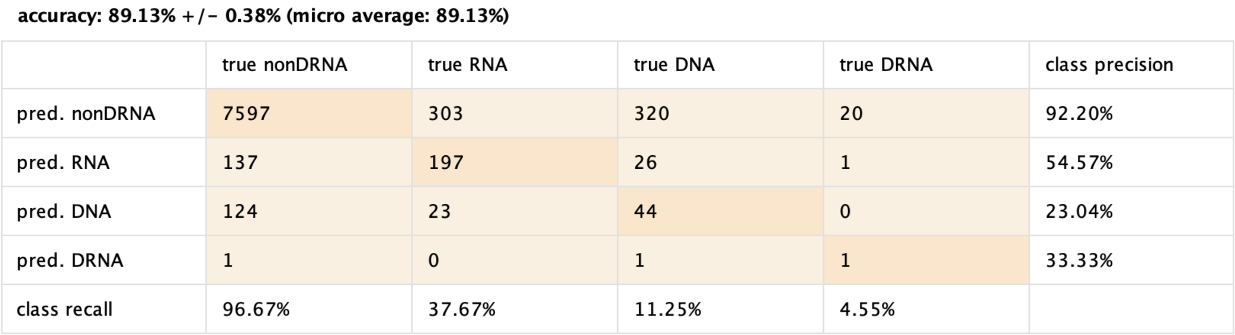
Data Processing Steps:

- Created python script to measure Amino Acid Frequency in each protein sequence in sequences_training.txt, generated Amino Acid Frequency CSV.
- Used Pfeature to encode sequences into features, generated 11 physio-chemical features from the protein sequences, generated Pfeature CSV.
- Created python scripts to combine Amino Acid Frequency feature CSV with Pfeature features CSV.
    - This is our training dataset file: processed_features.csv

Classification Algorithm Used:

- Decision Trees
    - Information gain criteria
    - Maximal depth = 10
    - Pruning & pre pruning
    - Confidence = 0.1
    - Minimal gain = 0.01
    - Minimal leaf size = 2
    - Minimal size for split = 4
    - Number of pre pruning alternatives = 3

4x4 Confusion Matrix & The Four MCC Values



accuracy: 89.13% +/- 0.38% (micro average: 89.13%)

	true nonDRNA	true RNA	true DNA	true DRNA	class precision
pred. nonDRNA	7597	303	320	20	92.20%
pred. RNA	137	197	26	1	54.57%
pred. DNA	124	23	44	0	23.04%
pred. DRNA	1	0	1	1	33.33%
class recall	96.67%	37.67%	11.25%	4.55%	

MCC nonDRNA = 0.355
MCC RNA = 0.425
MCC DNA = 0.134
MCC DRNA = 0.122