Protein Classification

Real Time Learning in Intelligent Systems

Joaquim Leitão

· U C ·

Objectives

- Protein function identification from genetic sequence
  - 55 families – 2 classifiers per family
- Incremental vs Non-Incremental Approaches
  - LASVM vs SVM
- SCOP40_Minidatabse

Methodology

1. Pre-processing
2. Data Representation
3. Classifier Training
4. Results Assessment

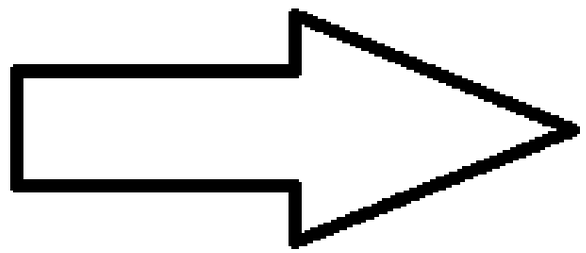
Pre-Processing

- SCOP40mini.fasta
  - Genetic coding sequences for several protein sequences
- SCOP40mini_sequence_minidatabase_19.cast
  - Map between sequences to be used in training and testing for each family

Data Representation

- Count number of occurrences of each nucleotide
- Exclude “x” nucleotides – Missing Data

vdaeavvqqkcis



V	3
D	1
E	1
A	1
Q	2
K	1
I	1
C	1
S	1

Support Vector Machines

- Radial Basis Kernel for high-dimensional mapping
  - $\exp(-\gamma * |x_i - x_j|^2)$
- Misclassification penalty – Cost parameter, C
- Tune parameters C and γ

Support Vector Machines - Tuning

- "A practical guide to support vector classification", Hsu *et al.* (2003):
    - Grid-search on exponentially growing sequences of C and γ

- $C = 2^{-5}, 2^{-4}, \ldots, 2^{15}$
- $\gamma = 2^{-15}, 2^{-14}, \ldots, 2^{15}$

Support Vector Machines - Results

- Average AUC of 0.74 in test dataset
- Best AUC = 0.99 for family "b.29.1._b.29.1.2." (C = 256 ; γ = 3.9e-03)
- Overfitting in families "c.26.2._c.26.2.1." and "c.47.1._c.47.1.10." (Train AUC = 1; Test AUC around 0.5)
- Families where no classifier could be learned (Train AUC = Test AUC around 0.5)

LASVM

- Online version of SVM featuring a support vector removal step
- Radial Basis Kernel + Misclassification penalty
- Tune parameters C and γ similar to SVM case
  - $C = 2^{-5}, 2^{-4}, \ldots, 2^{15}$
  - $\gamma = 2^{-15}, 2^{-14}, \ldots, 2^{15}$

LASVM - Results

- Average AUC of 0.81 in test dataset
- Best AUC = 0.99 in families “b.29.1._b.29.1.11.” (C = 32 ; γ = 1.22e-04); “b.29.1._b.29.1.2.” (C = 4 ; γ = 9.7e-04); and “c.67.1._c.67.1.4.” (C = 512 ; γ = 2.44e-04)
- Does not appear to suffer from overfitting
- Families where no classifier could be learned (Train AUC = Test AUC around 0.5)

Conclusions

- LASVM produce better results than SVM and does not appear to suffer from overfitting
- Best results in test dataset do not correspond to high performances in the training dataset (AUC = 0.7-0.8)
- Average results lower than SCOP40 benchmark

Future Work

- Extend parameter fine tuning
- Missing data imputation techniques
- More powerful representation techniques
- Other machine learning techniques, such as neural networks or random forests

Questions?