Requirements for GPC in the Real World

Anton Schwaighofer, June 2005

Fraunhofer FIRST
Intelligent Data Analysis Group

. . . Real World?

- Prediction of physiological properties of chemical compounds: Classify whether a compound interacts with 5 different proteins ("potent inhibitors")
- Highly unbalanced classes: worst case 29 positives out of 600 compounds, high cost of misclassifying the smaller class
- Probabilistic prediction required (probability that compound is a "potent inhibitor")

Bias shift, $b = \Phi^{-1}(N_2/(N_1+N_2))$ (Lawrence et al, IVM) does not seem to work reliably

Re-balancing GP Classification by Sampling

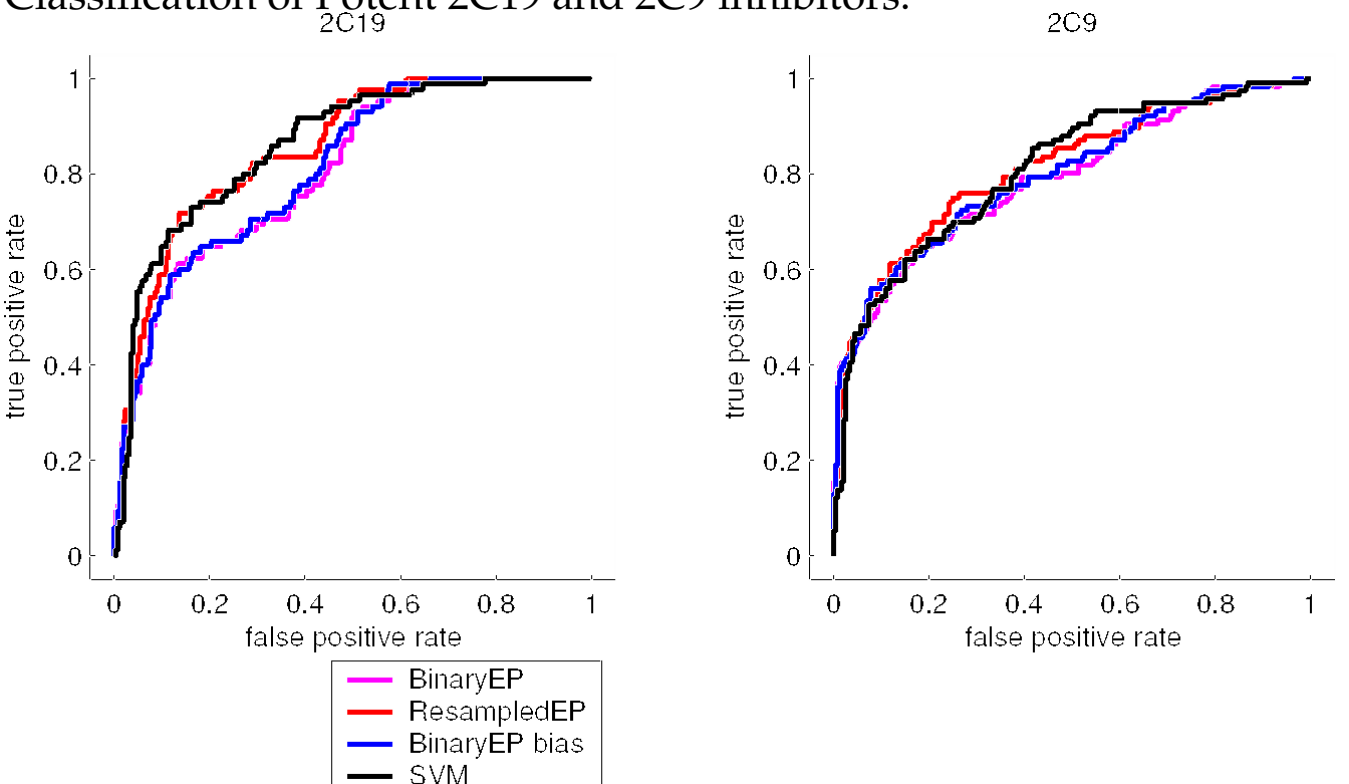
For a data set $\mathcal{D}$ with classes $\mathcal{C}_1$, $\mathcal{C}_2$, class size $N_1 > N_2$:
Build a committee of d classifiers as follows

- For $i = 1 \ldots d$: Build
  $\mathcal{D}^i = \{\text{all examples from } \mathcal{C}_2\} \cup \{N_2 \text{ examples chosen randomly from } \mathcal{C}_1\}$
- Train GP classifier i on $\mathcal{D}^i$

At prediction stage: Average predictions of all d GP classifiers

Real World (again)

Classification of Potent 2C19 and 2C9 inhibitors:



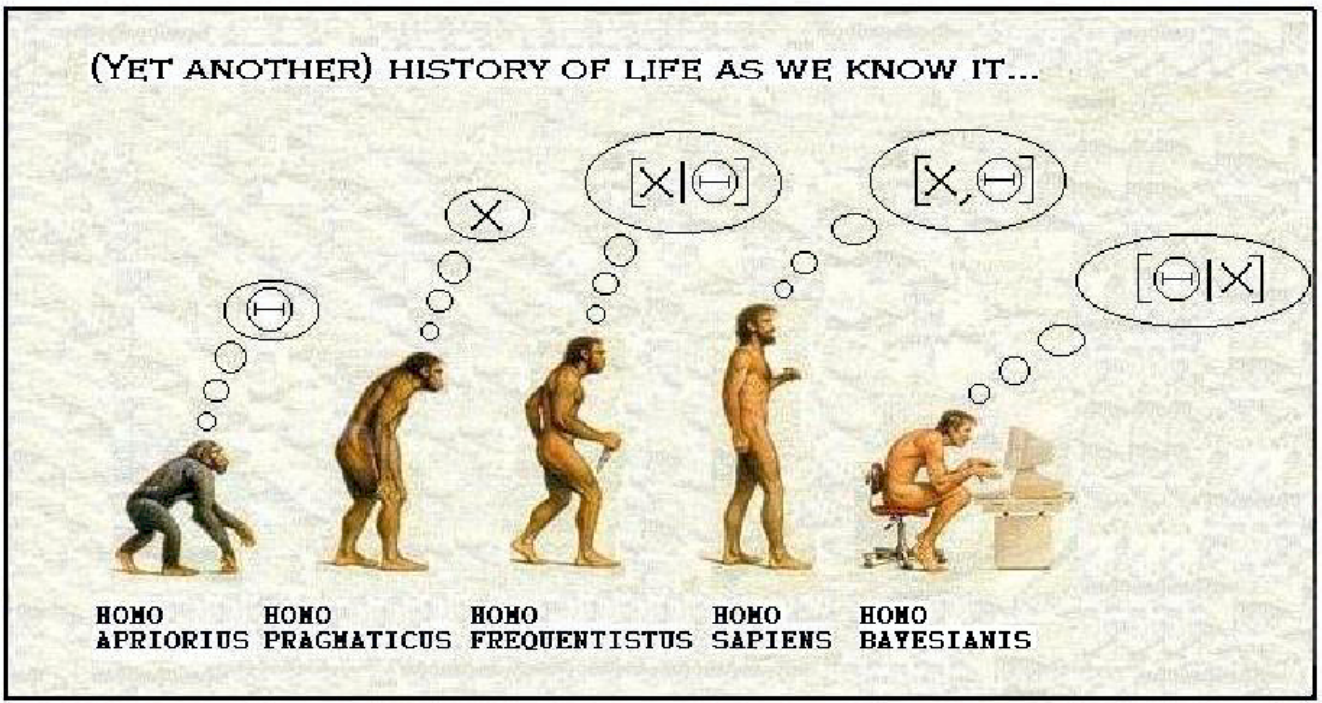
Why I Like it

- It's simple (good for computer scientists)
- It solves the problem of unbalanced classes
- It (could) help dealing with large data sets

Open issues

Unbalanced data sets are extremely common real-world applications, with high cost for misclassification of the smaller class

What are the possible GPC solutions for this problem?


(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...
HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS

(c) Mike West