Evolving Receiver Operating Characteristics for Data Fusion

W. B. Langdon

Presented at EuroGP'2001

Computer Science, University College, London
W.Langdon@cs.ucl.ac.uk
http://www.cs.ucl.ac.uk/staff/W.Langdon

Introduction

- What is Data Fusion?
- What are Receiver Operating Characteristics
- Evolving Classifiers
- Overlapping Gaussians, Thyroid, Landsat, Drug Activity p450
- Conclusions

The Problem

- There are numerous data mining techniques

  Each tries to make sense of data

- Many can be treated as a classification/prediction E.g.:
  - Is this credit card purchase fraudulent?
  - Will this address buy double glazing?
  - Is this a planet?
  - Does this molecule block this virus?

Data Fusion

- Which classifier to use?

  Problem specific

- Can we get better results from a "fusion" of classifiers

- Can fusion be automated?

- Exponential explosion of number of ways to combine classifiers

- Use Genetic programming to find a good non-linear combination

- What is fitness of classifier?

What is Genetic Programming

- GP is randomised parallel search technique for very large search spaces

  GA $\ll$ GP, since GP explores both structure and coefficients

- Many points are tested

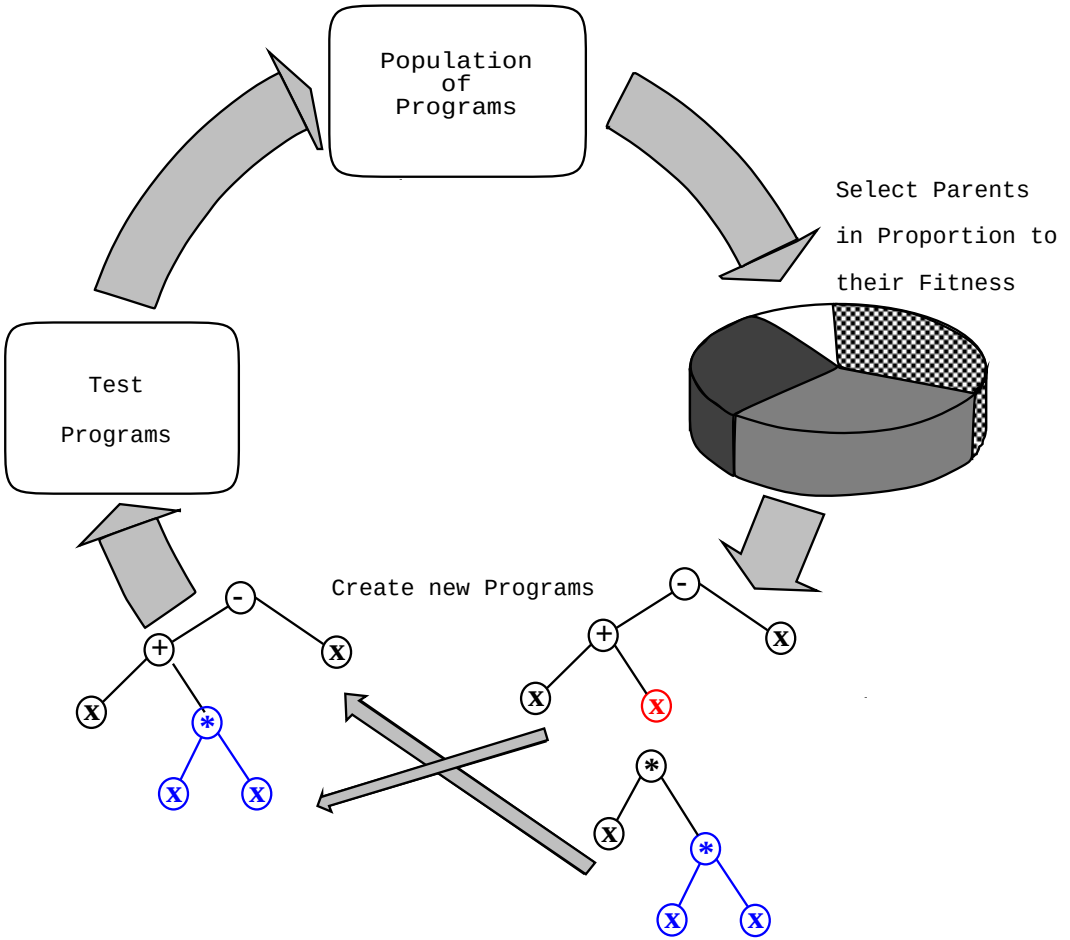
- Search continues from the better ones

- New search points are

  – neighbours of better ones

  – combinations of better ones

- Based on analogy with natural evolution, "Survival of the fittest"

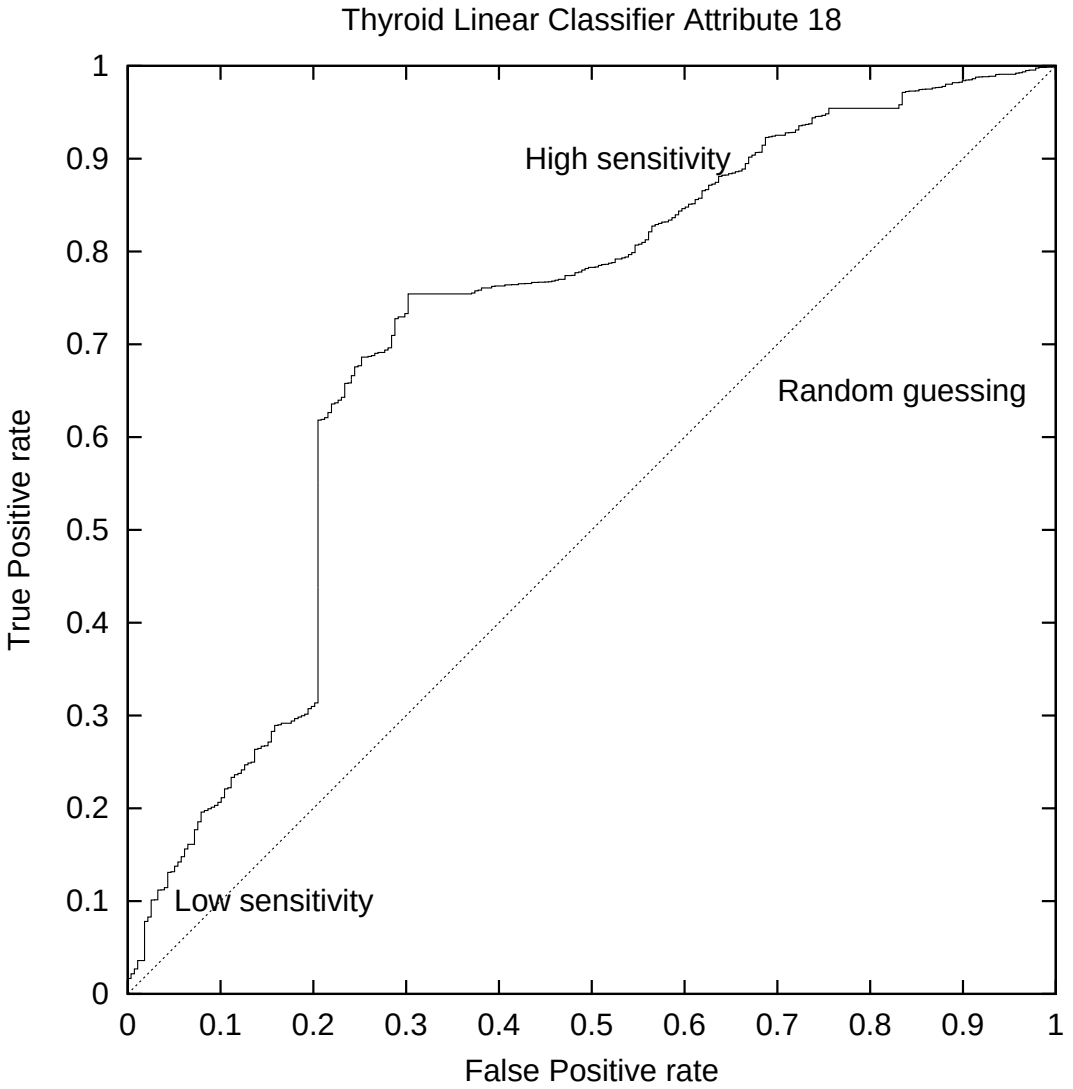


The Genetic Programming Cycle, from GP & Data Structures

Receiver Operating Characteristics

- Classifiers seldom deal with certainty
- Is this person's Thyroid OK?
- Low sensitivity, every time classifier says no
- Increase sensitivity, more yes
- Very high sensitivity, every time says yes
- ROC shows tradeoff between missing positive cases and raising false alarms

ROC of Linear Classifier on Thyroid Data



"Best" ROC = Convex Hull

- Best classifier has ROC near (0,1) (*top left corner*)

- Area under ROC, 0.5 (random), 1 (perfect)

  "Better" classifier has bigger area

- Like 0,0 to 1,1 diagonal,

  ROC of random combination of classifiers lies between them

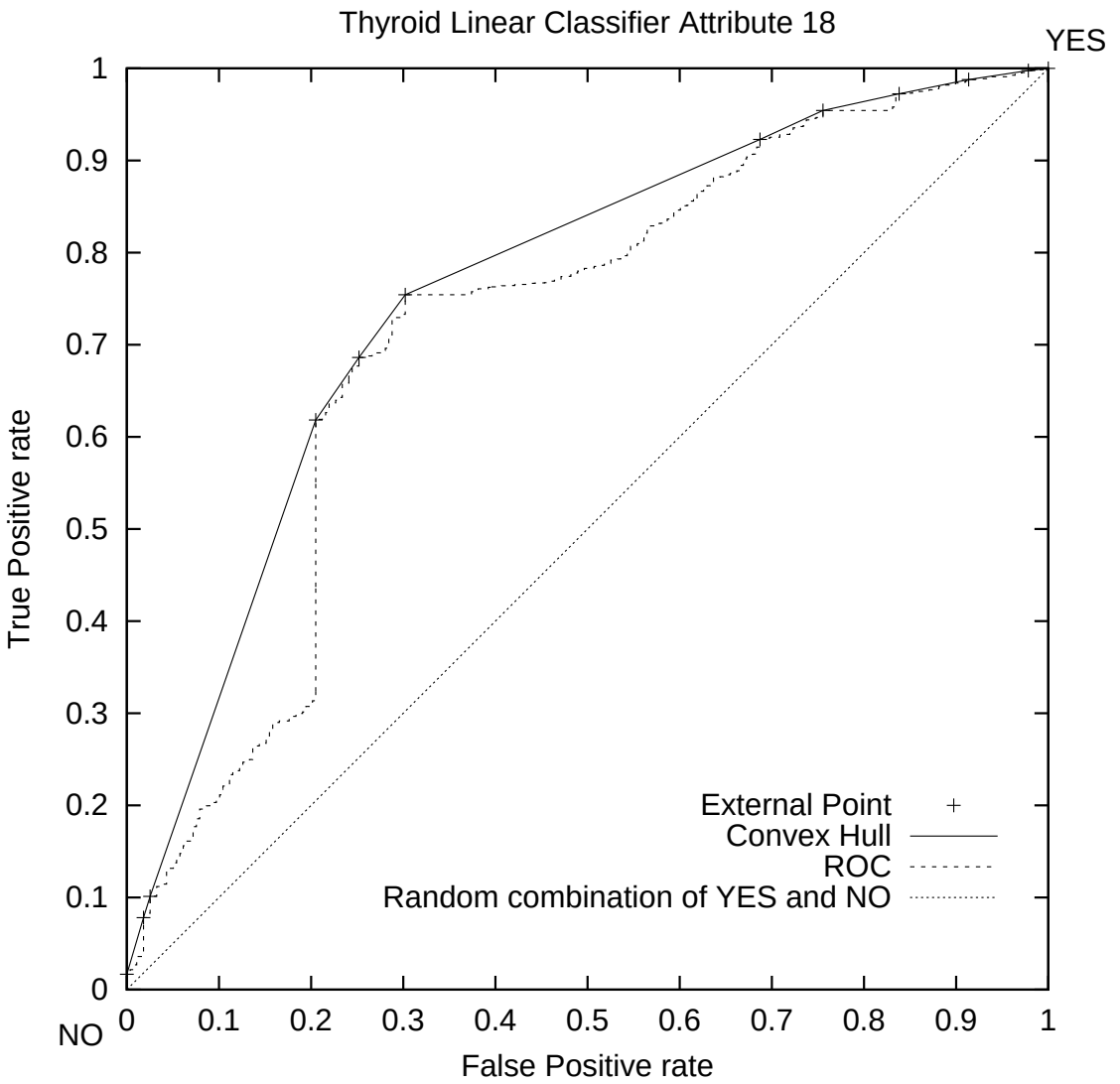
  Therefore only external points are useful

  Scott [BMVC'98]'s "Maximum Realisable ROC", MRROC.

  MRROC = convex hull

- Fitness = area under convex hull of ROC

Convex Hull of Receiver Operating Characteristics



GP to Evolve Classifiers

- Function set: traditional + classifiers

  All classifiers unary functions, current test case + threshold

  return classification of test case and "confidence"

- Terminal set: traditional + threshold

- 5 Trees: sum value returned by each tree (i.e. weighted vote)

  Wrapper: sum $< 0 \Rightarrow$ negative class

- Fitness: Run on test set for threshold 0, 0.1, 0.2 ... 1.0

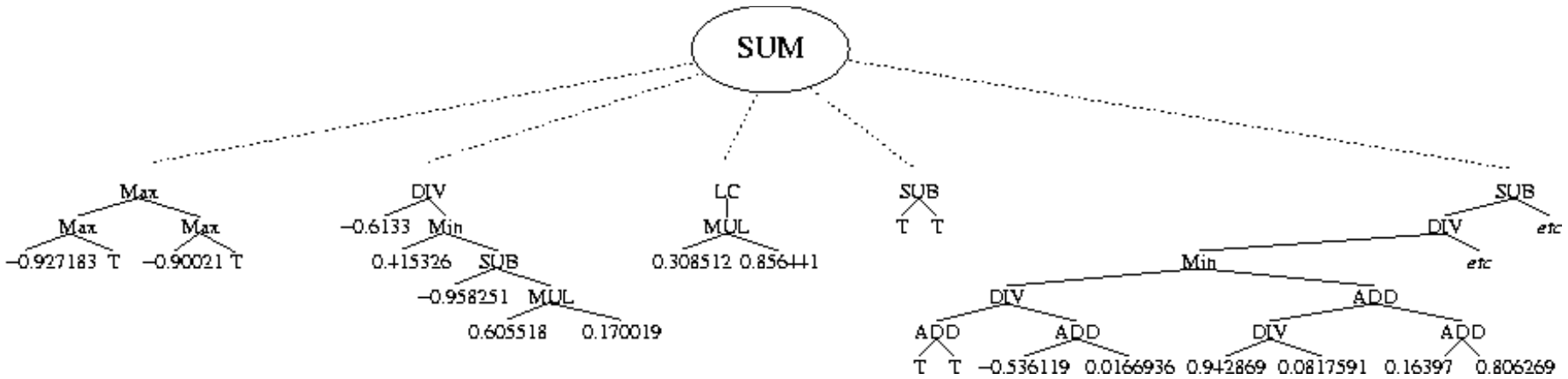
  Calculate true positive and false positive rate for each threshold

  Fitness = area under convex hull of 13 points (also 0,0 and 1,1)

- Population 500. 50 generations

Evolved Classifier

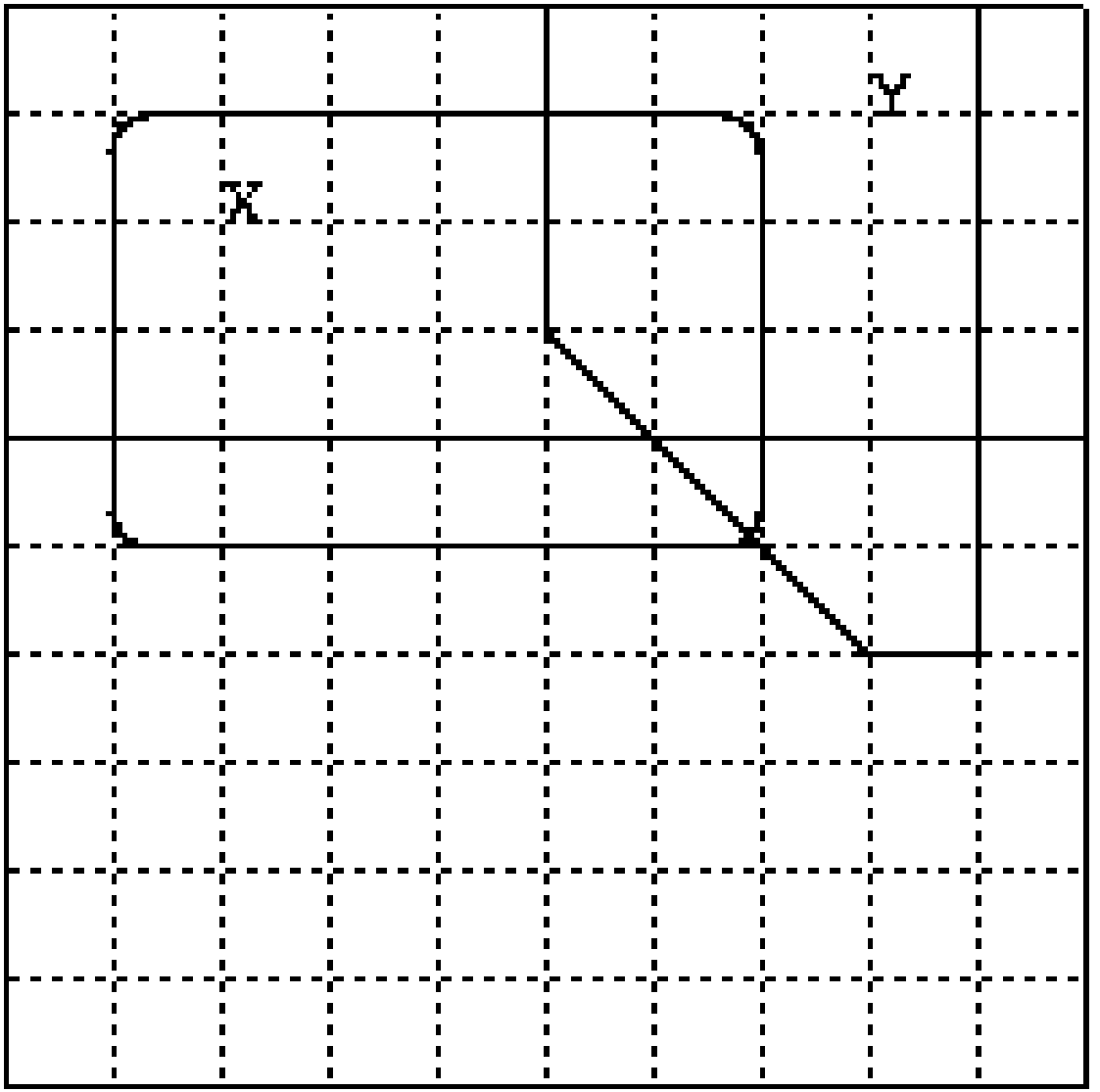
Answer = Sum of five GP trees



GP Parameters

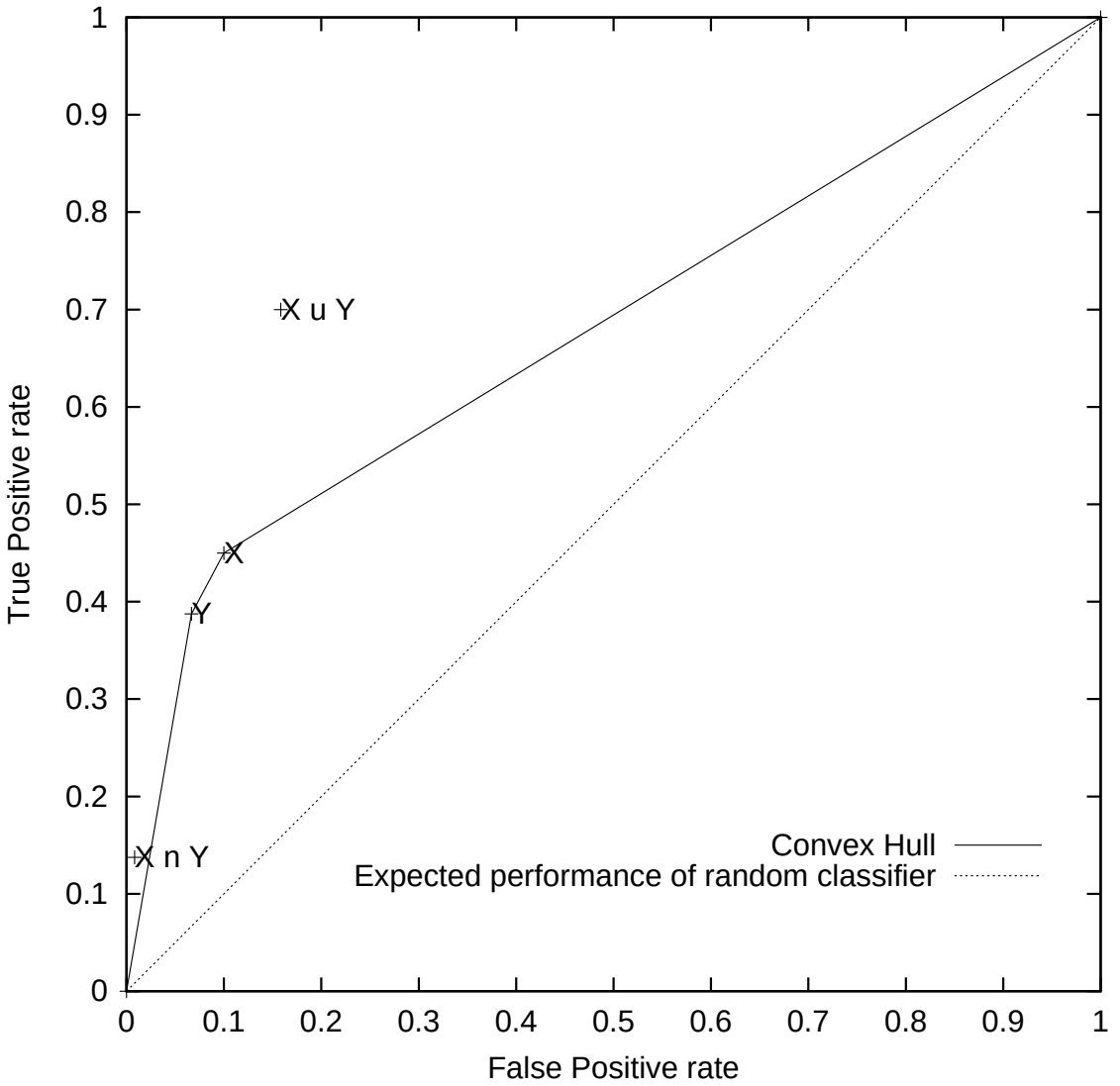
Objective:	Evolve a function with Maximum Convex Hull Area
Function set:	INT FRAC Max Min MaxA MinA MUL ADD DIV SUB IFLTE LC
Terminal set:	T, 0, 1, 200 unique constants randomly chosen in $-1 \ldots +1$
Fitness:	5000 randomly chosen test points Area under convex hull of 11 ROC points.
Selection:	generational (non elitist), tournament size 7
Wrapper:	$\geq 0 \Rightarrow$ positive, negative otherwise
Pop Size:	500
No size or depth limits	
Initial pop:	ramped half-and-half (2:6) (half terminals are constants)
Parameters:	50% size fair crossover 50% mutation (point 22.5%, constants 22.5%, shrink 2.5% subtree 2.5%)
Termination:	generation 50

Simple Example

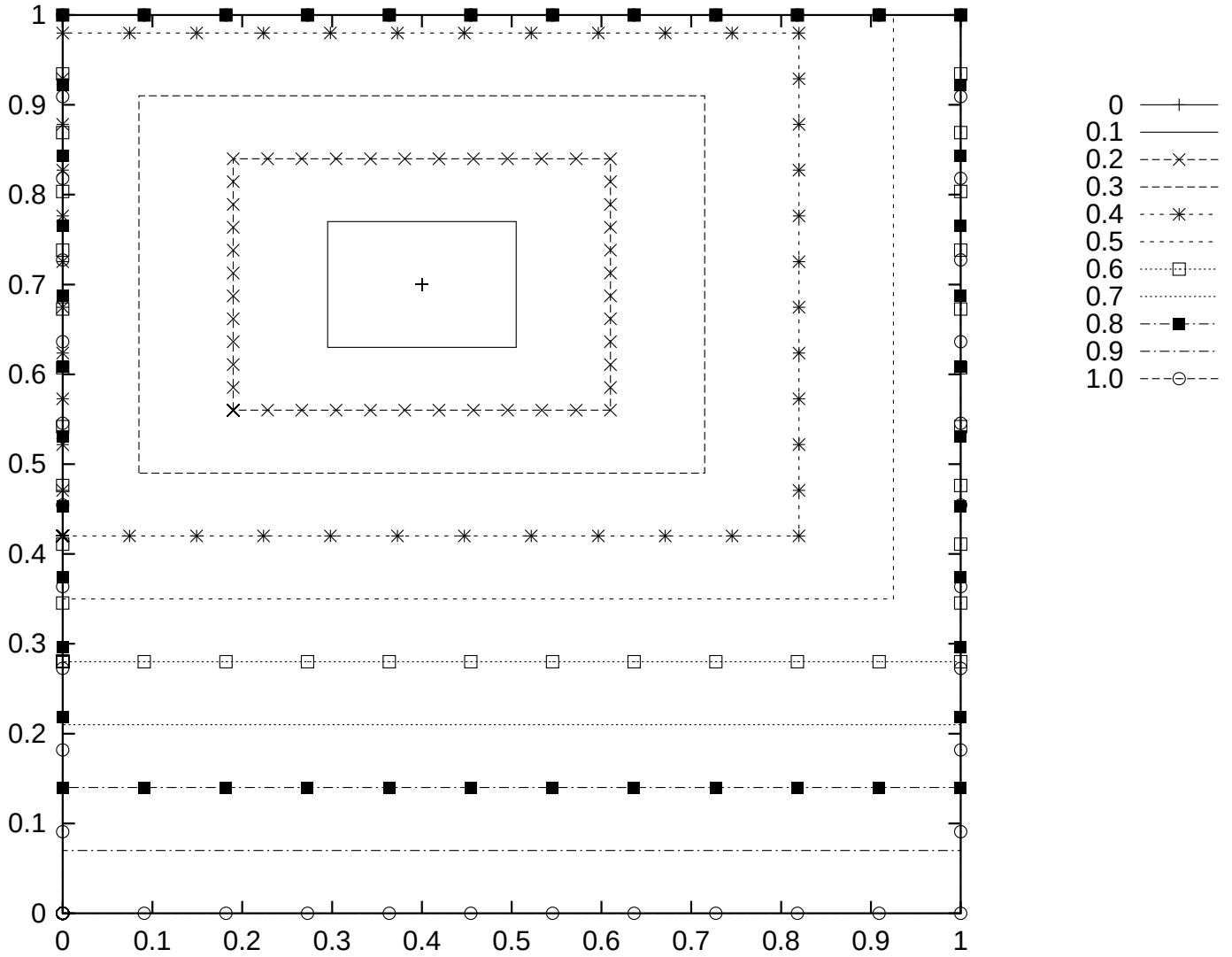


Data points above horizontal line are in the class

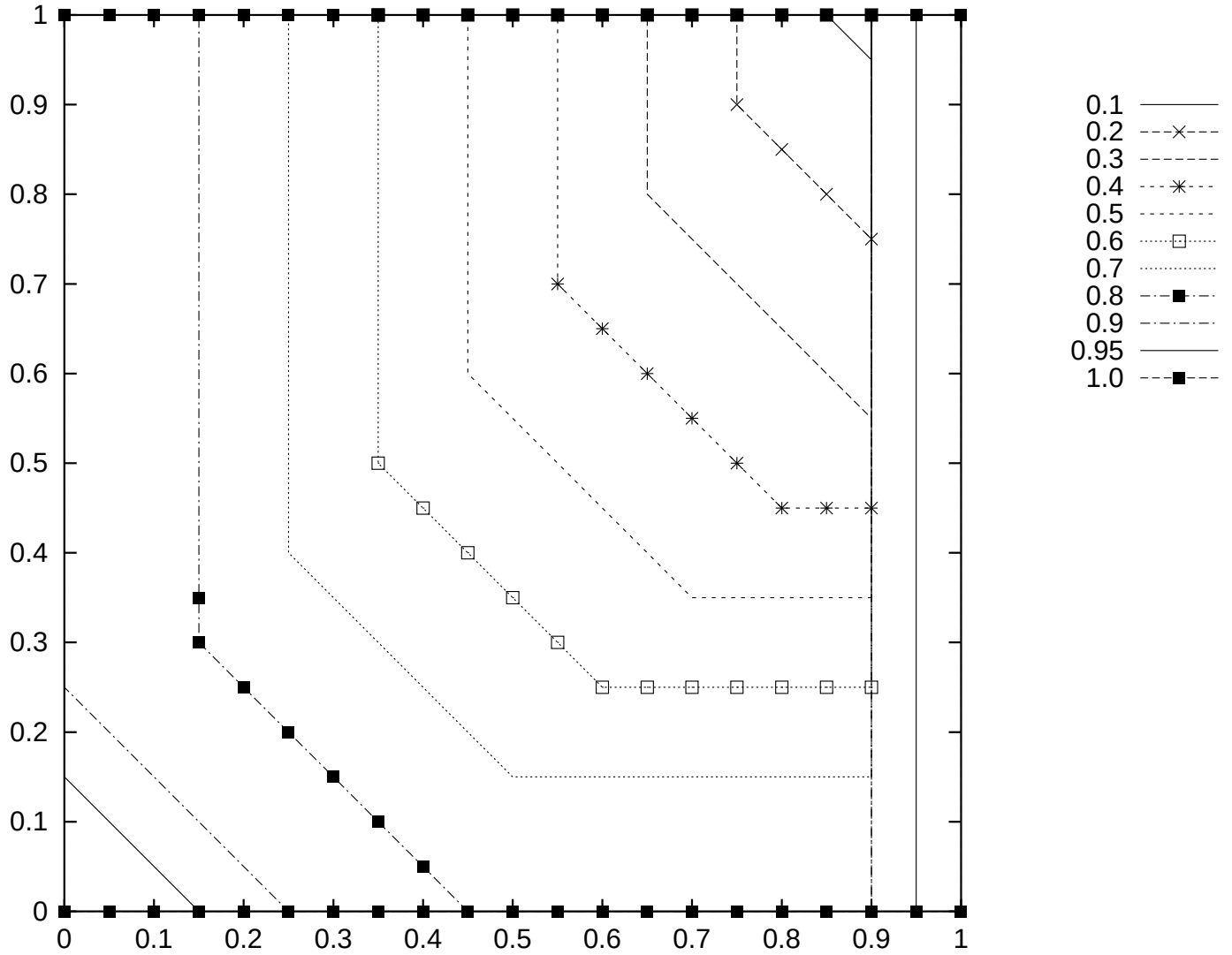
Combined Classifier better than Convex Hull



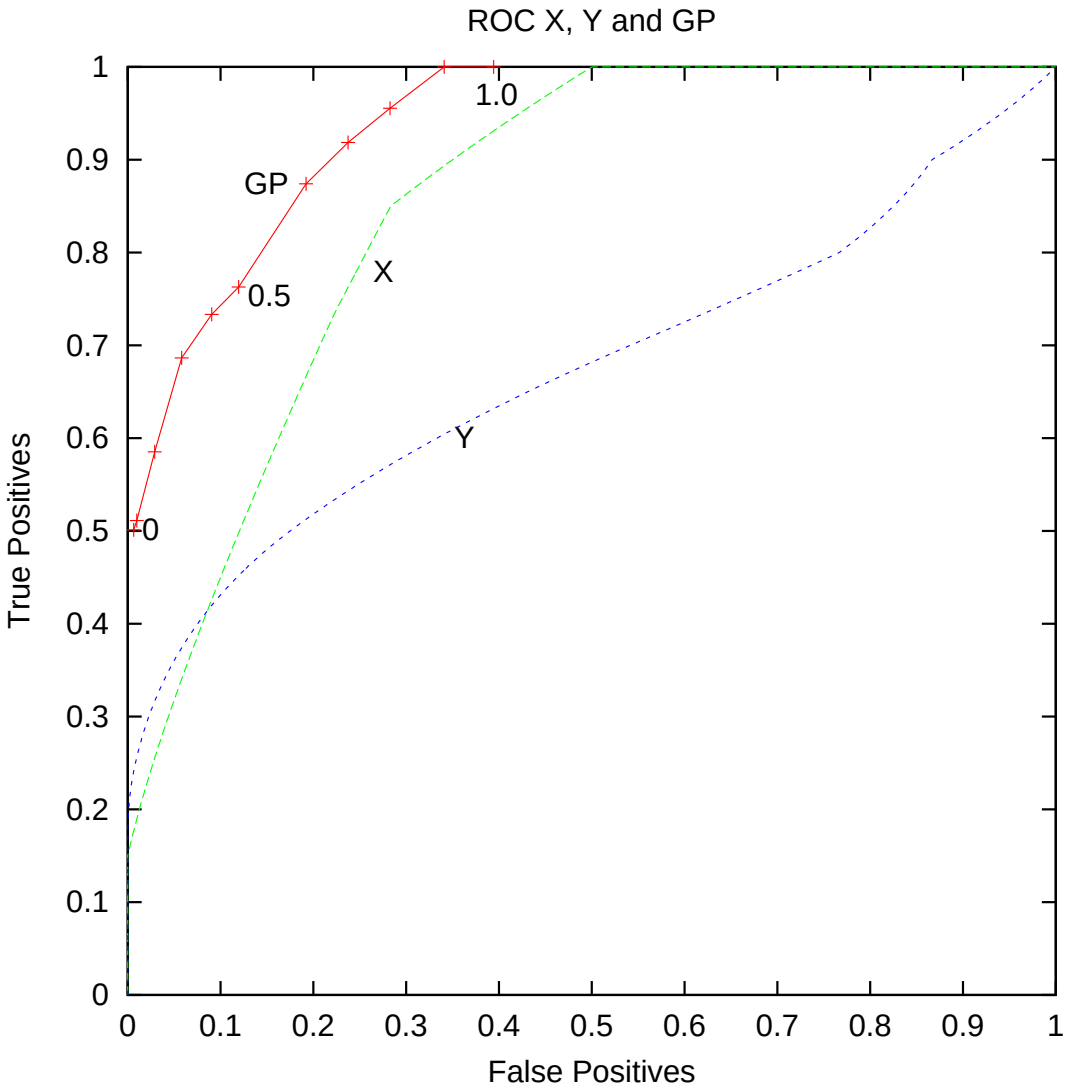
Sensitivity makes X rectangle grow or shrink



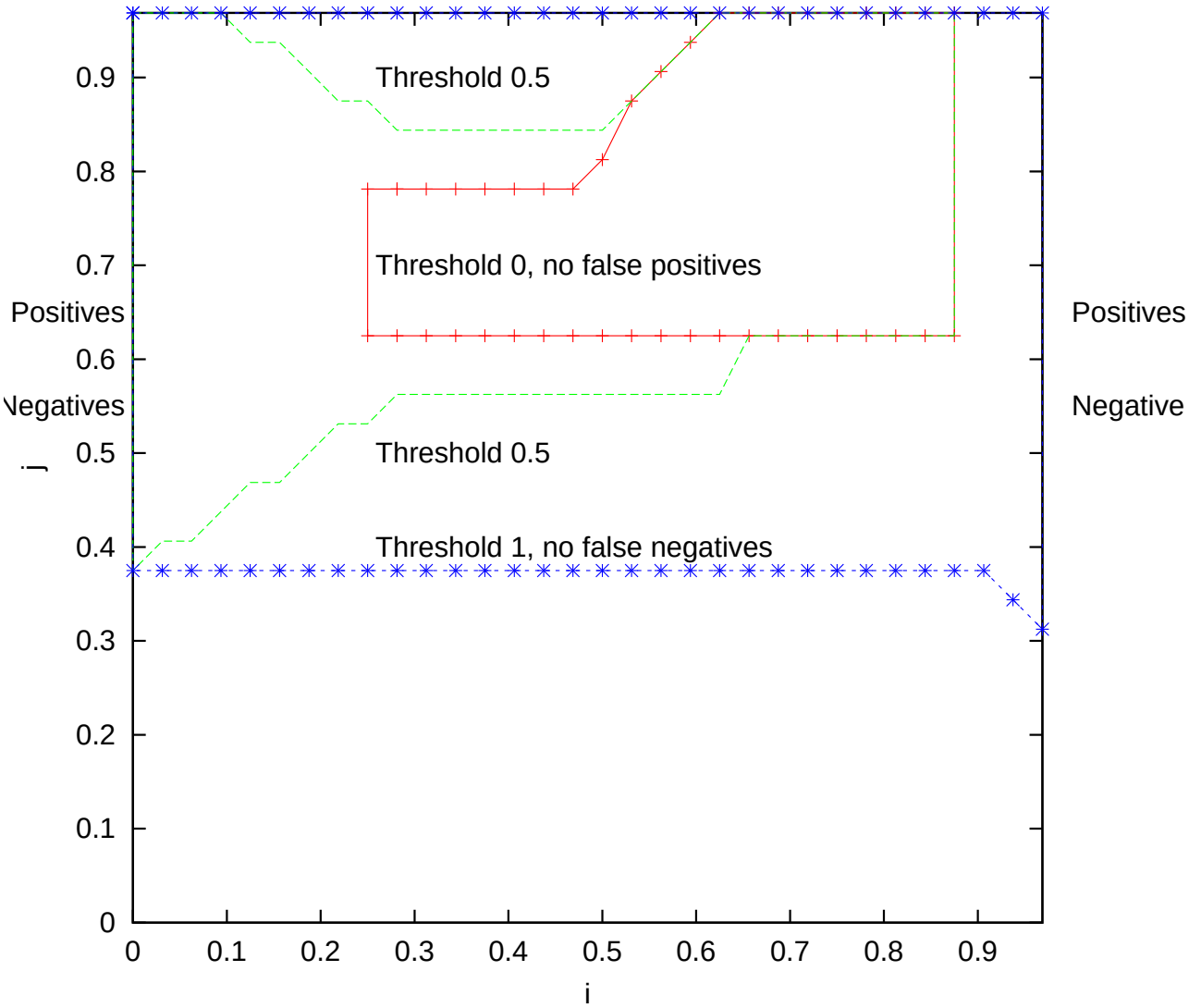
Sensitivity makes Y "lozenge" grow or shrink



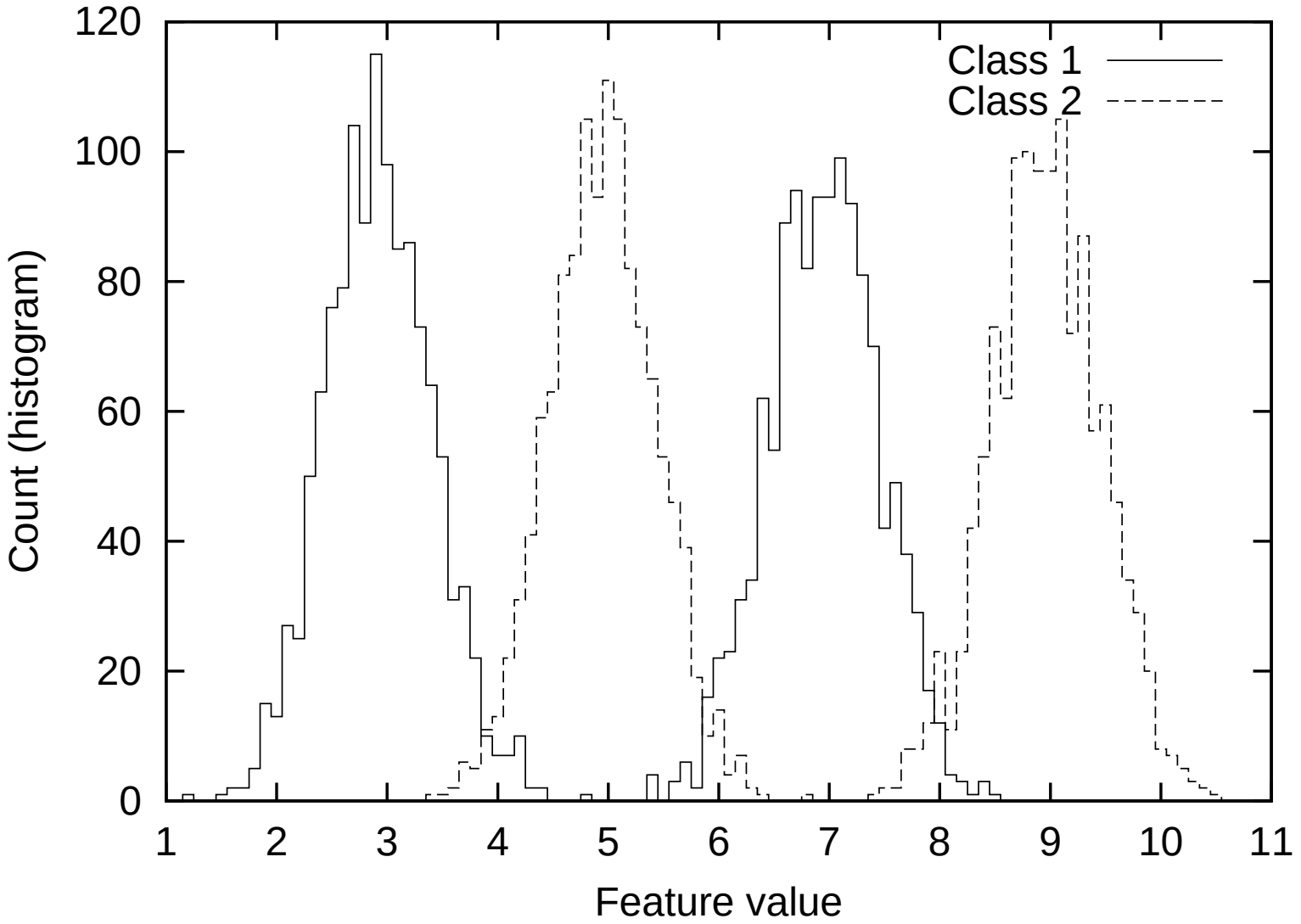
Receiver Operating Characteristics of X, Y and Evolved Combination



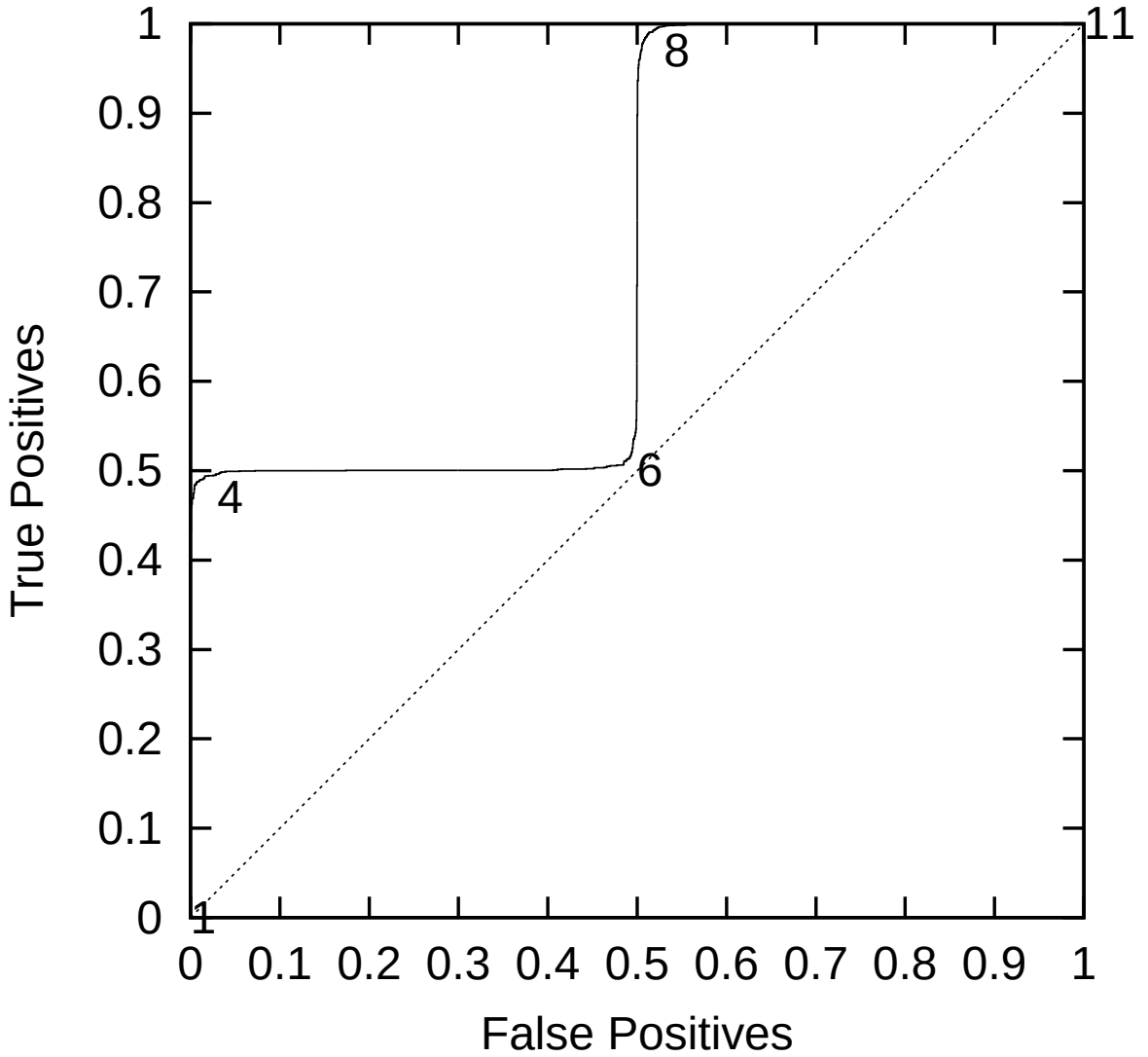
Decision Boundary of Evolved Classifier



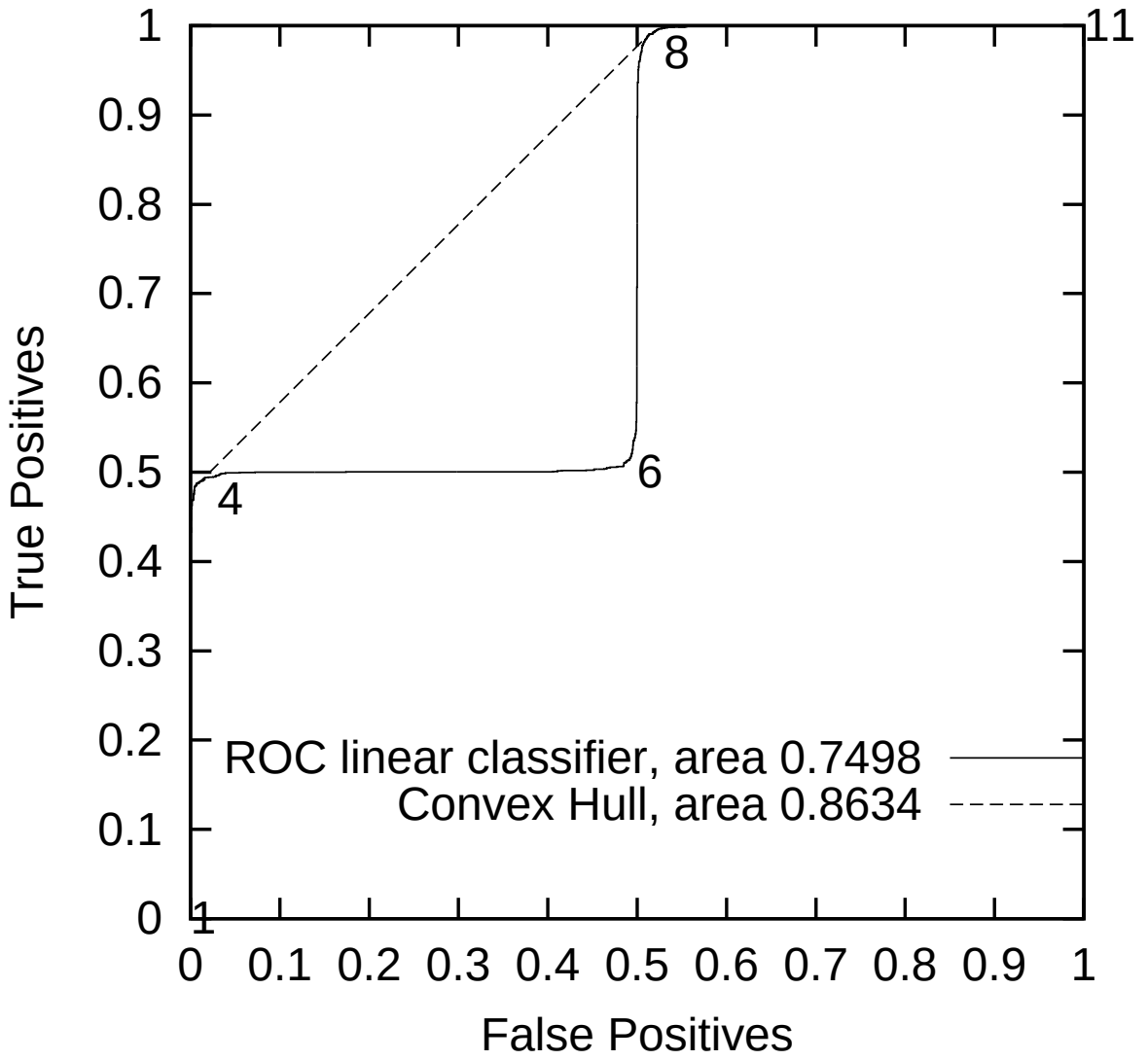
Overlapping Gaussians



Receiver Operating Characteristics of Linear Classifier on Overlapping Gaussians

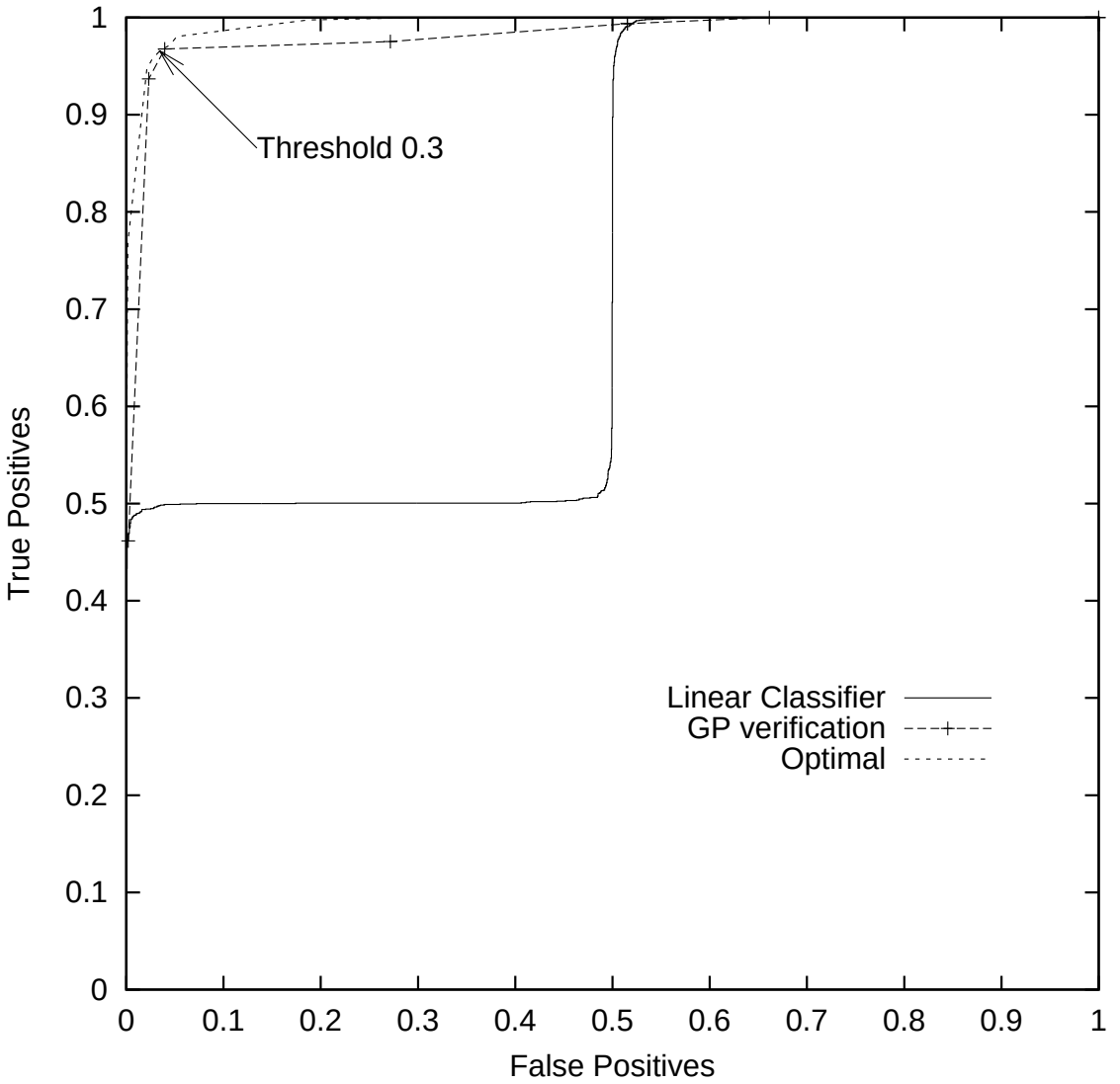


Convex Hull of ROC on Overlapping Gaussians

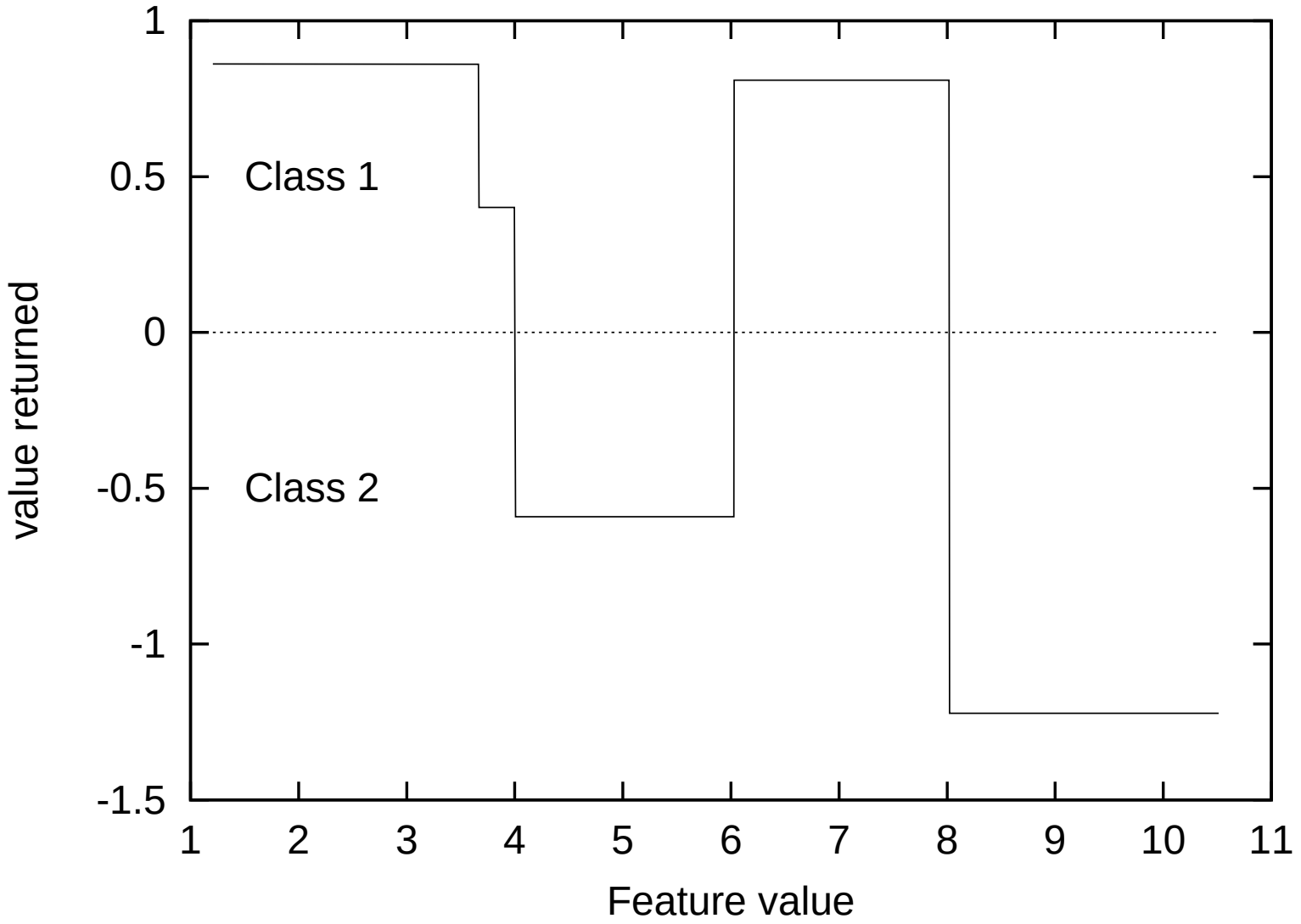


Note random combination of classifiers 4.0 and 8.0 is better than classifier with threshold 6.0

ROC of Evolved Classifier on Overlapping Gaussians



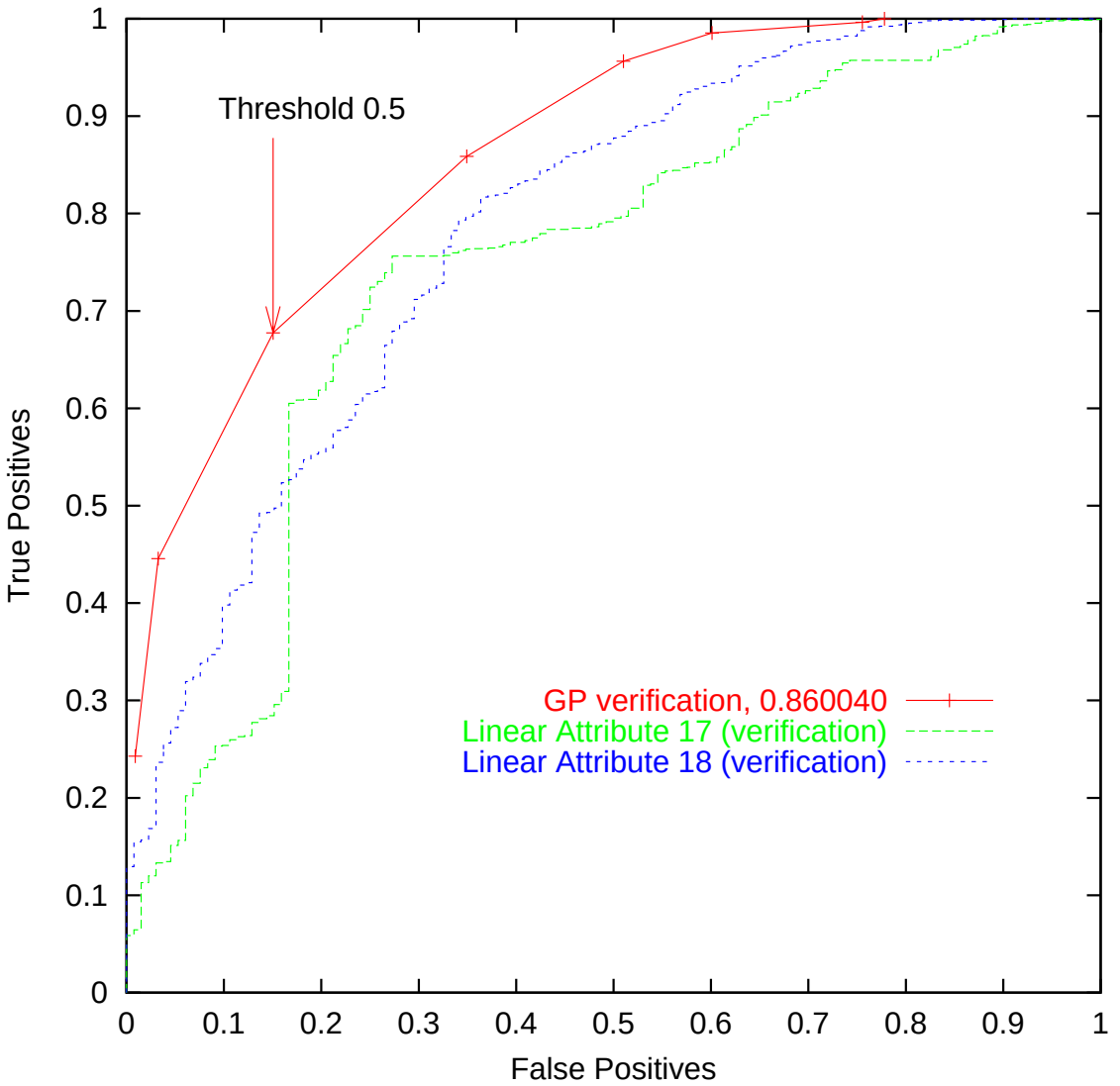
Output at Threshold 0.3 of Evolved Classifier on Overlapping Gaussians



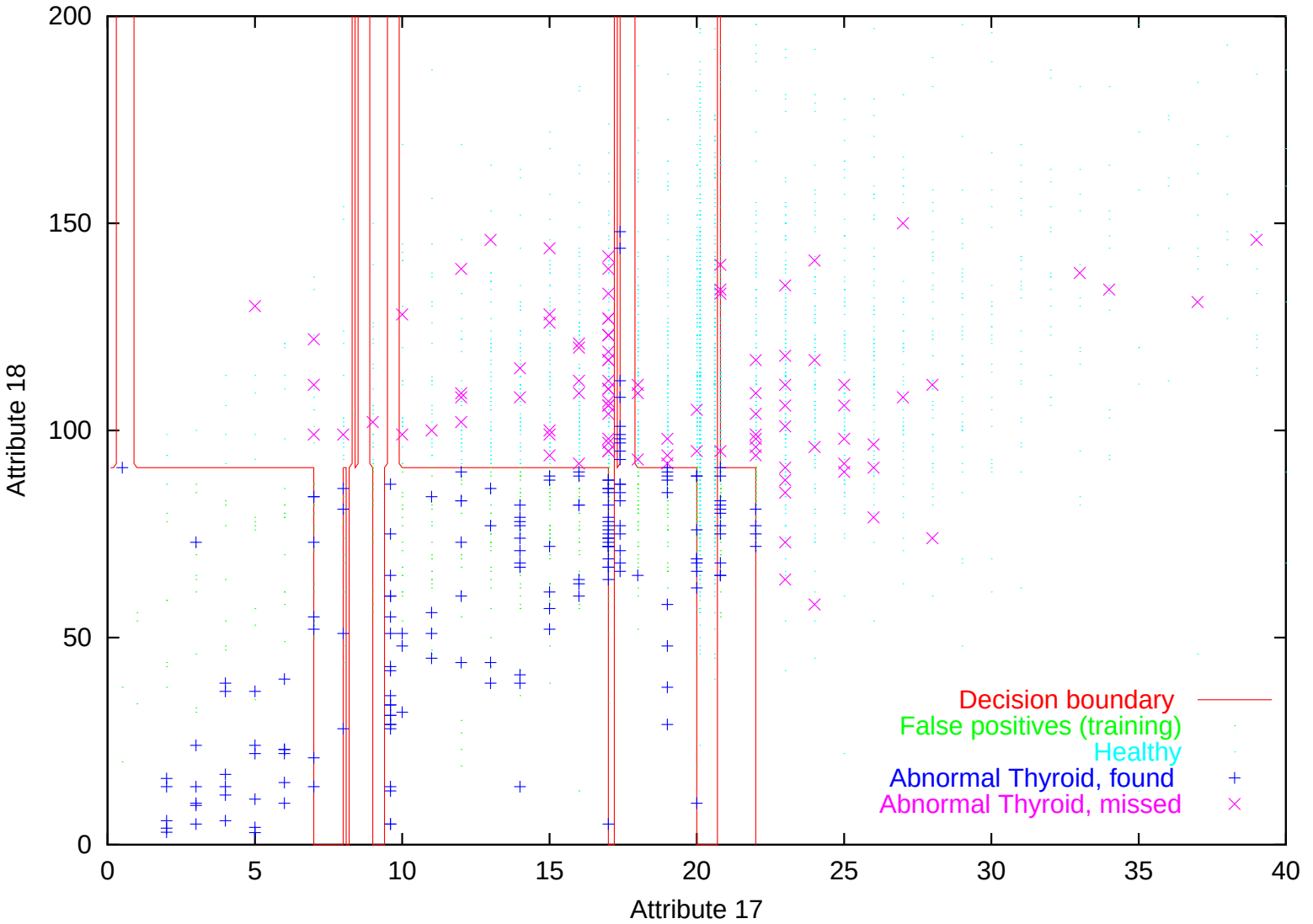
Thyroid

- Second of Scott's [BMVC,1998] bench marks
- Real data
- Linear classifiers on two attributes (15 binary 6 continuous)
- ROC of evolved combination better than either
- GP forms non-linear combinations

Thyroid Receiver Operating Characteristics



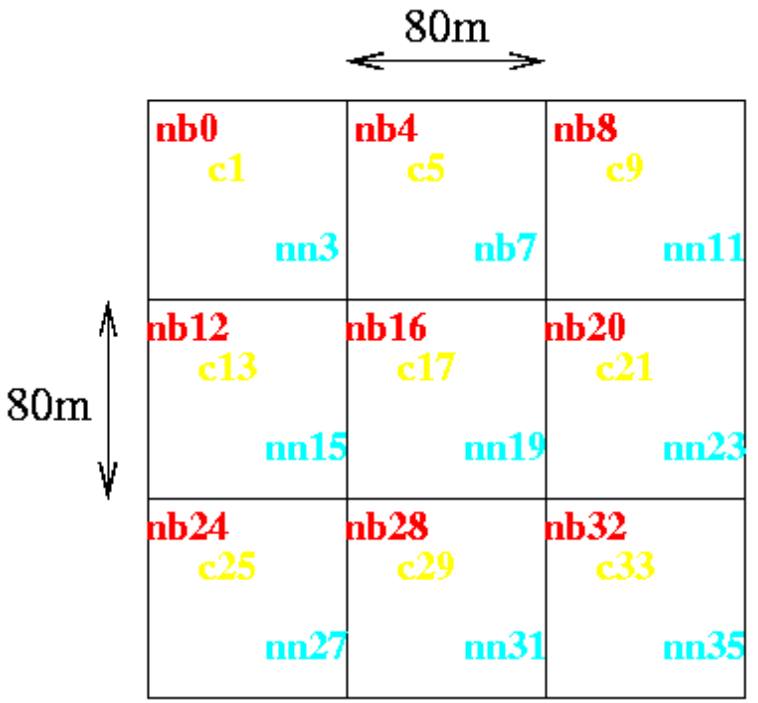
Thyroid Evolved Classifier (Threshold 0.5)



Landsat (Naive Bayes)

- Last of Scott's Bench marks
- Binary classification of soil type from 4 wave band Landsat data
- Best of Scott's classifiers given to GP

Landsat, Nine Pixels × 4 Bands

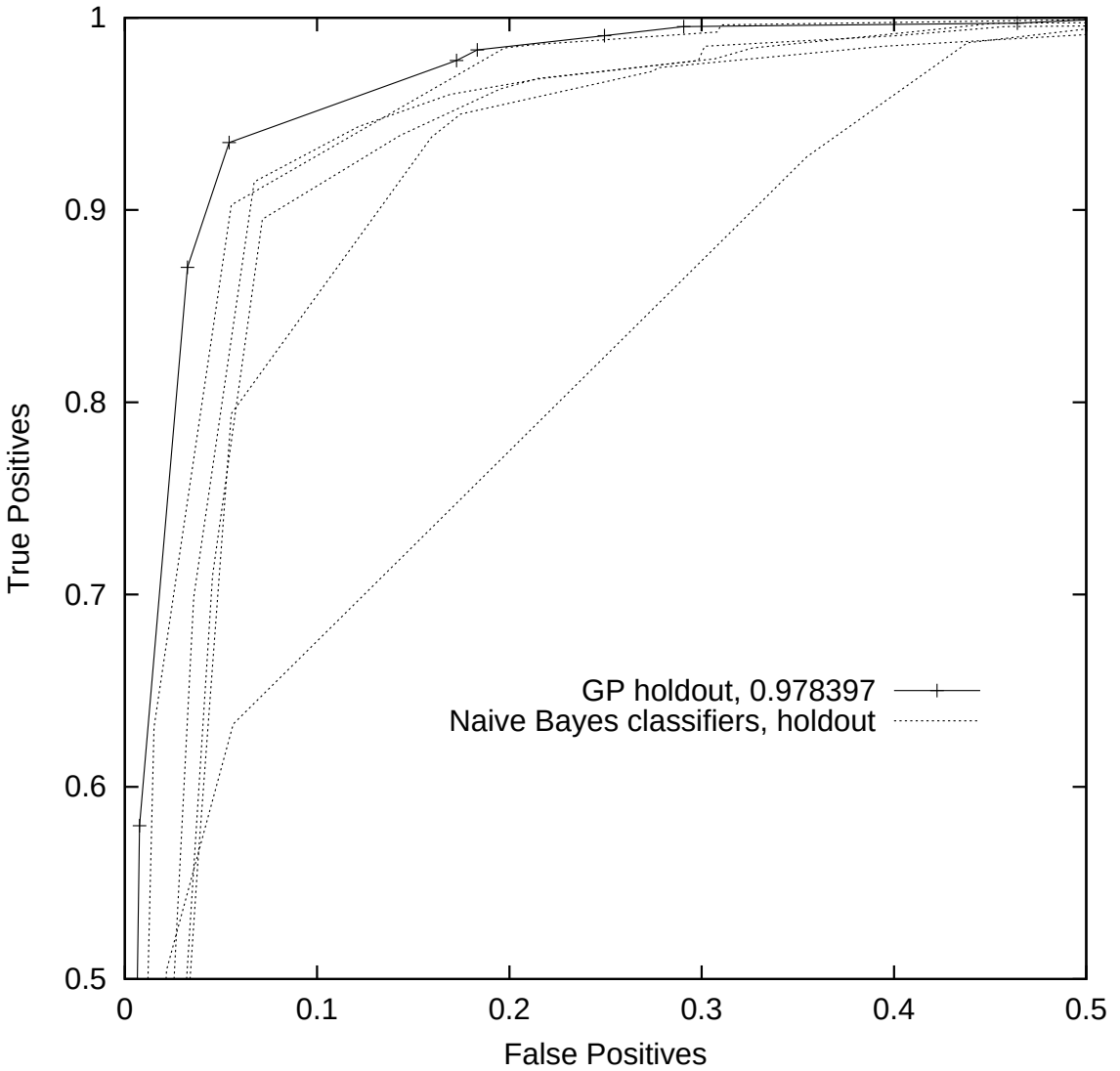


Each record contains data from nine adjacent Landsat pixels

nb16, nb16,23 nb16,23,24 nb23,24 and nb8,23,24 together use four attributes

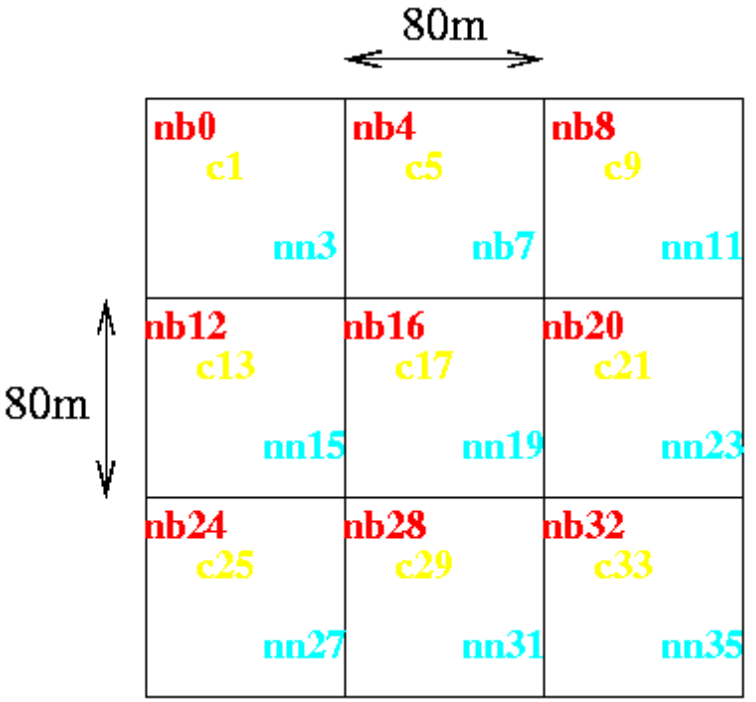
Three (8, 16, 24) use spectral band 0 and the other (23) uses band 3

Landsat Receiver Operating Characteristics



(Note X and Y-Ranges)

Landsat, Nine Pixels × 4 Bands (3 used)



Each record contains data from nine adjacent Landsat pixels

Two near infrared, two visible bands

Naive Bayes trained on Band 0

C4.5 Trained on Band 1

Artificial Neural Network trained on Band 3 (Band 2 poor)

Landsat (Naive Bayes, C4.5, Neural Network)

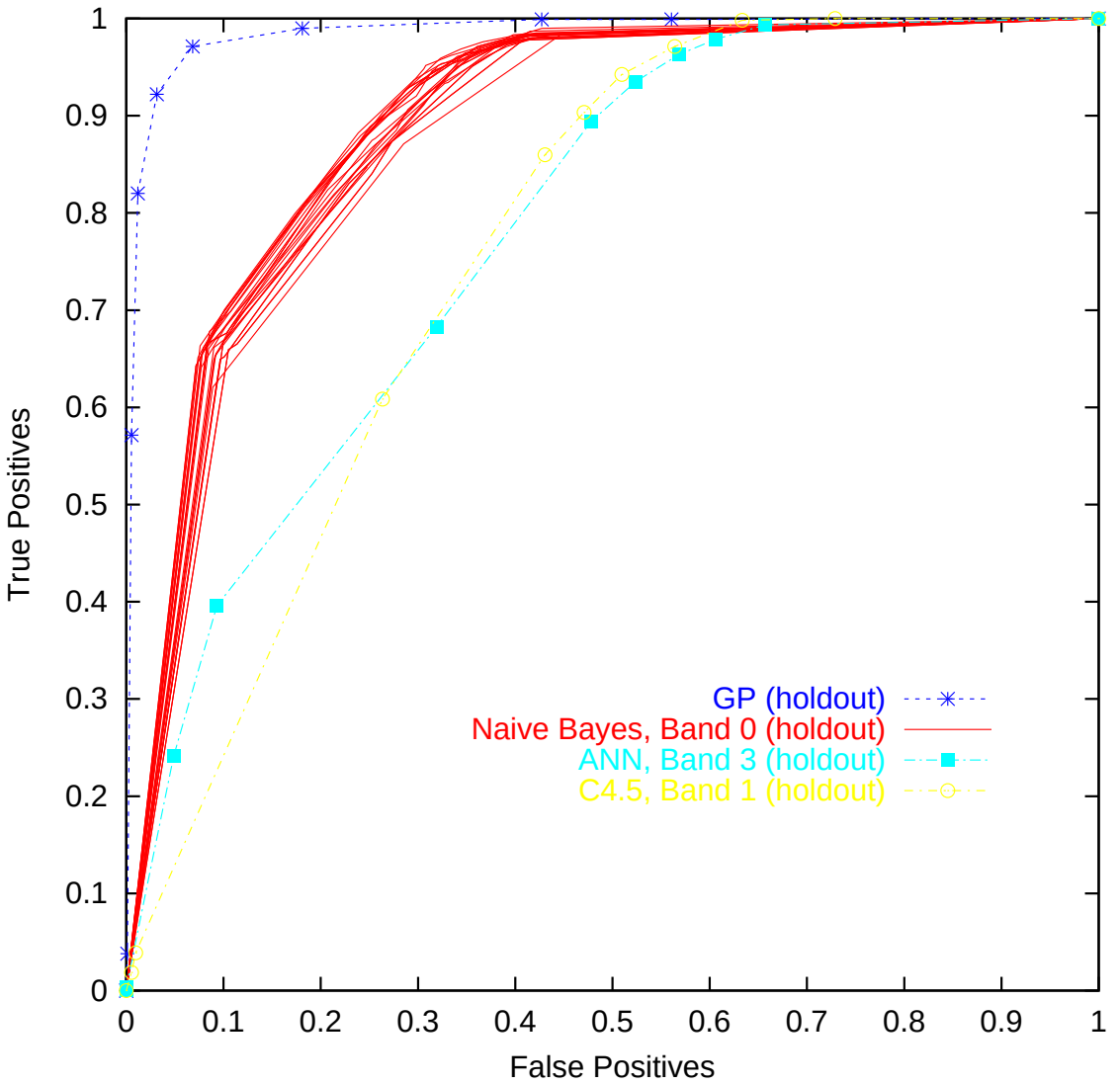
- Landsat data but 3 radically different classifiers
- 21 Naive Bayes classifiers trained on Band 0

  All single attributes (9)

  All pairs of best 6 attributes (15)
- C4.5 trained only on band 1
- ANN trained only on band 3
- All 7 combinations of single band, pairs and triple combinations
- In all 7 cases GP higher ROC area than best input classifier

Landsat 3 Classifiers

Receiver Operating Characteristics



Drug Activity GlaxoSmithKline p450 Data

- 1500 compounds, 300 positive 1200 negative

  Especially noisy data discarded. "Similar" compounds removed

- 699 attributes

  GSK problem knowledge used to calculate attributes from primary molecular structure

  Ten types of attribute. 3 small group together, biggest 2 split $\Rightarrow$ 15 attribute groups

- 4 balanced training sets (same 300 pos, 300 different inactives)

- One Clementine neural network trained on each group of attributes and each training set

  $15 \times 4 = 60$ neural networks

"Blind Trial" p450 Data

- Only neural networks and classification

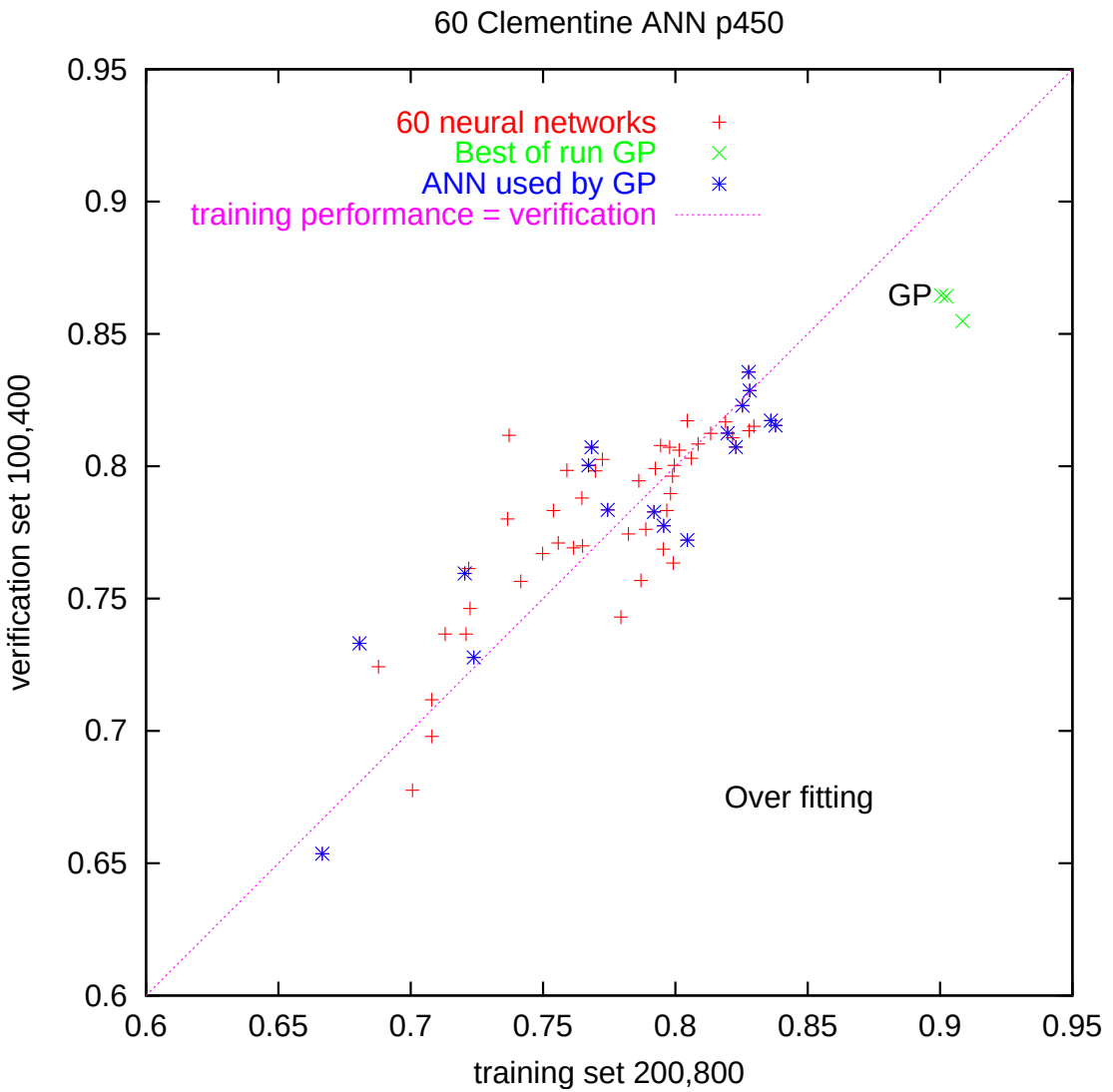
  No access to private data (699 attributes)

- GlaxoSmithKline keeping 1500 different records for validation

- 1500 randomly split 2:1 training:verification

- Training set used by GP

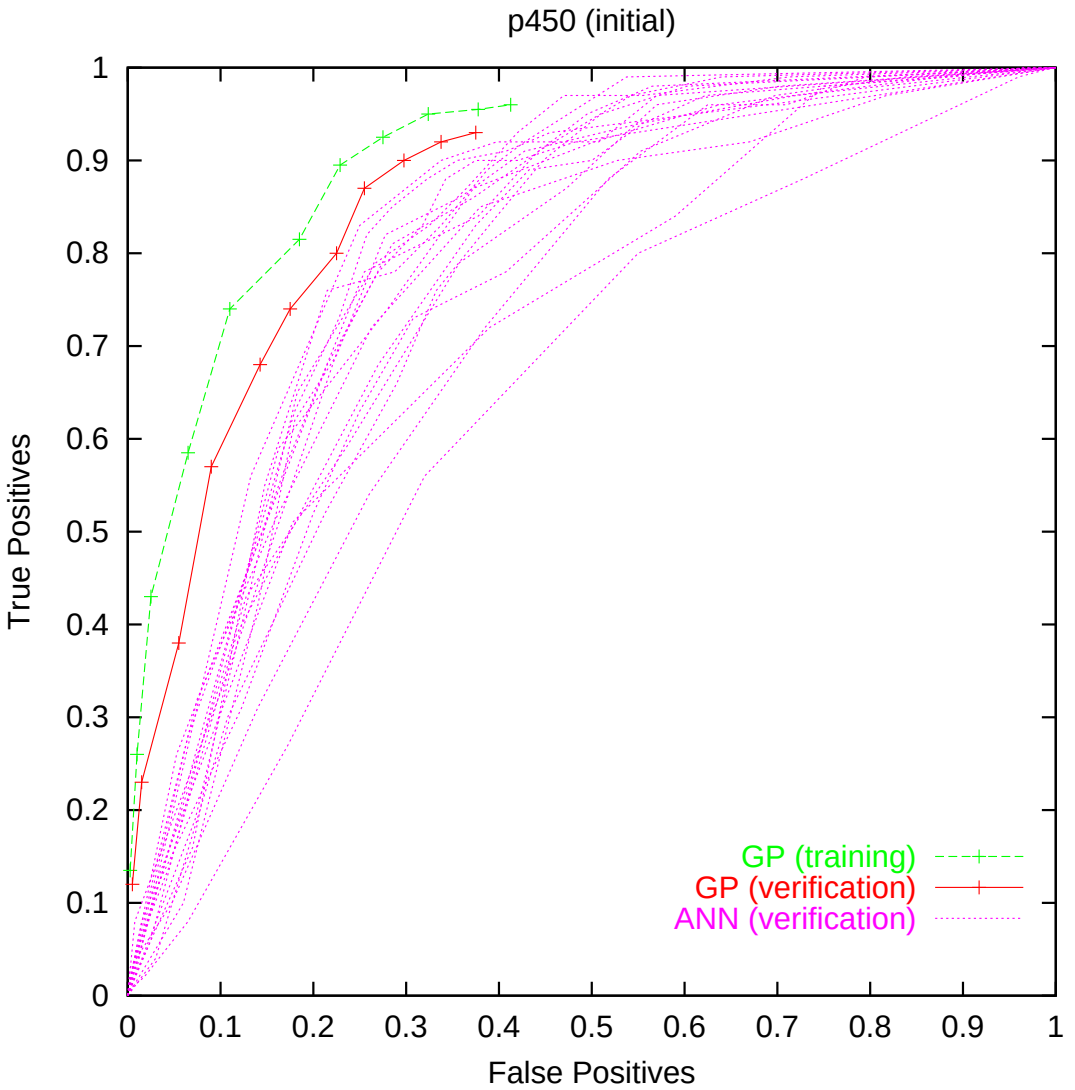
p450 ROC Area ANN and GP



GP marginally better (significant?) best of neural networks

GP over fitting?

p450 ROC ANN and GP



When Will GP-ROC Work?

- We may hope for improvement when
  - Have both aggressive (say positive when can) and conservative (only when sure) classifiers
  - Classifiers which are good at different parts of the feature space
  - Small number of significant features interacting in a complicated way
- Future work
  - Complete p450 experiments
  - Evolve (and combine) specialist classifiers
  - Boosting?

Conclusions

- Scott's "Maximum Realisable ROC" not always the best
- GP has done better on Scott's benchmarks and real problems
- We can automatically get better results from a "fusion" of classifiers. Demonstrated on:

  Same classifier (overlapping Gaussians)

  Classifiers of same type (Thyroid)

  Classifiers of different types, trained on different data (Landsat)

- GP-ROC technique not specific to a domain
- Size fair crossover and mutation effective against bloat

Evolution of Size