KDD Cup 2002:

Single Class SVM for Yeast Gene Regulation Prediction

Adam Kowalczyk
Bhavani Raskutti

Telstra Research Laboratories
Australia

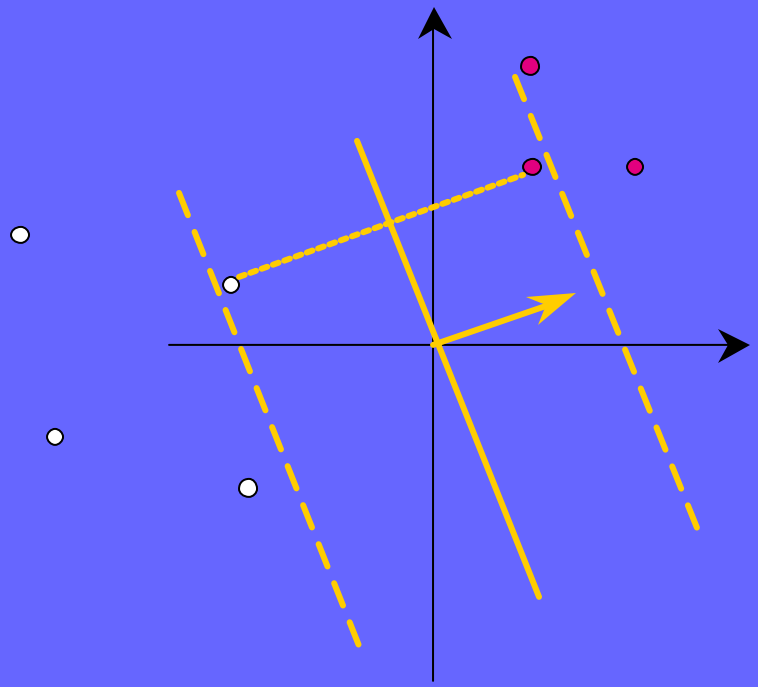
Overview

- Objective
  - Prediction of yeast gene regulation
- Data
  - Training - 3018
    - 38 positive in narrow partition
    - 84 positive in broad partition
  - Test - 1489
- Challenge
  - Data representation, Missing values, High dimensionality
- Solution
  - Single Class Support Vector Machines

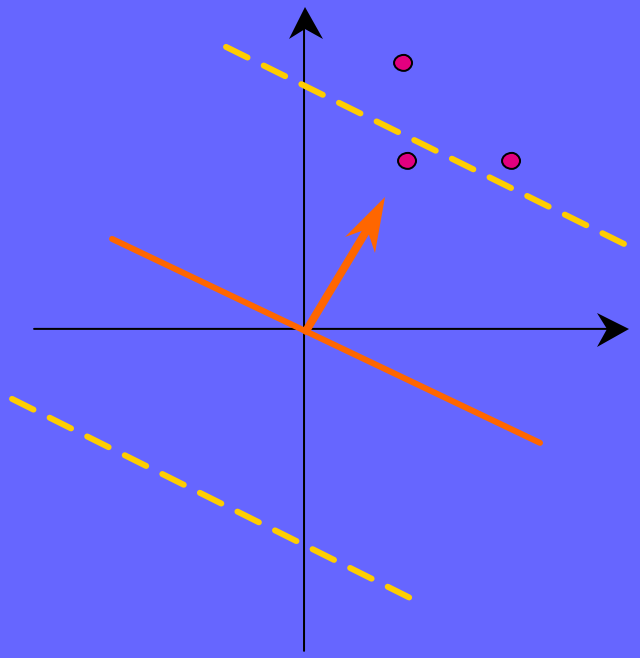
Data Representation

- Two kinds of data used by our approach
  - gene abstracts from MEDLINE database
    - Attributes: all words corresponding to abstract for training genes
    - 48829 words (excluding standard stoplist words)
      - Reduced to 12480 by deleting most frequent and least frequent words
  - data from the MIPS comprehensive yeast genome database
    - localization, protein classes, function
      - Information represents hierarchy, e.g., chromosome structure | nucleus
      - Attributes: each unique term for each data type (409 features)
      - Binary vector with 1 at every level of hierarchy
    - Gene Interactions
      - Attributes: all genes interacting with the training genes (1447 features)

Homogeneous SVM: Geometric View



Two class SVM



Single class SVM

Homogeneous SVM : Formal Definition

$$(x_i, y_i) ? \; ? \; ^{m} ? \{?1\}, \quad i ? 1,..., 3018$$

$$f(x) ? \; w ?x ? b ? (w, b) ?(x, 1)$$

$$? \; \tilde{w} ?\tilde{x}$$

$$\tilde{w} ? \arg\min \left\|\tilde{w}\right\|^2 ? \sum_i C_{y_i} [\max(0, \; 1 ? \; y_i \; \tilde{w} \tilde{x}_i)]^p$$

$p ? 1 \; or \; 2,$

$$C_{?1} ? \frac{C}{n_{?1}}, \quad C_{?1} ? \frac{C}{n_{?1}} B$$

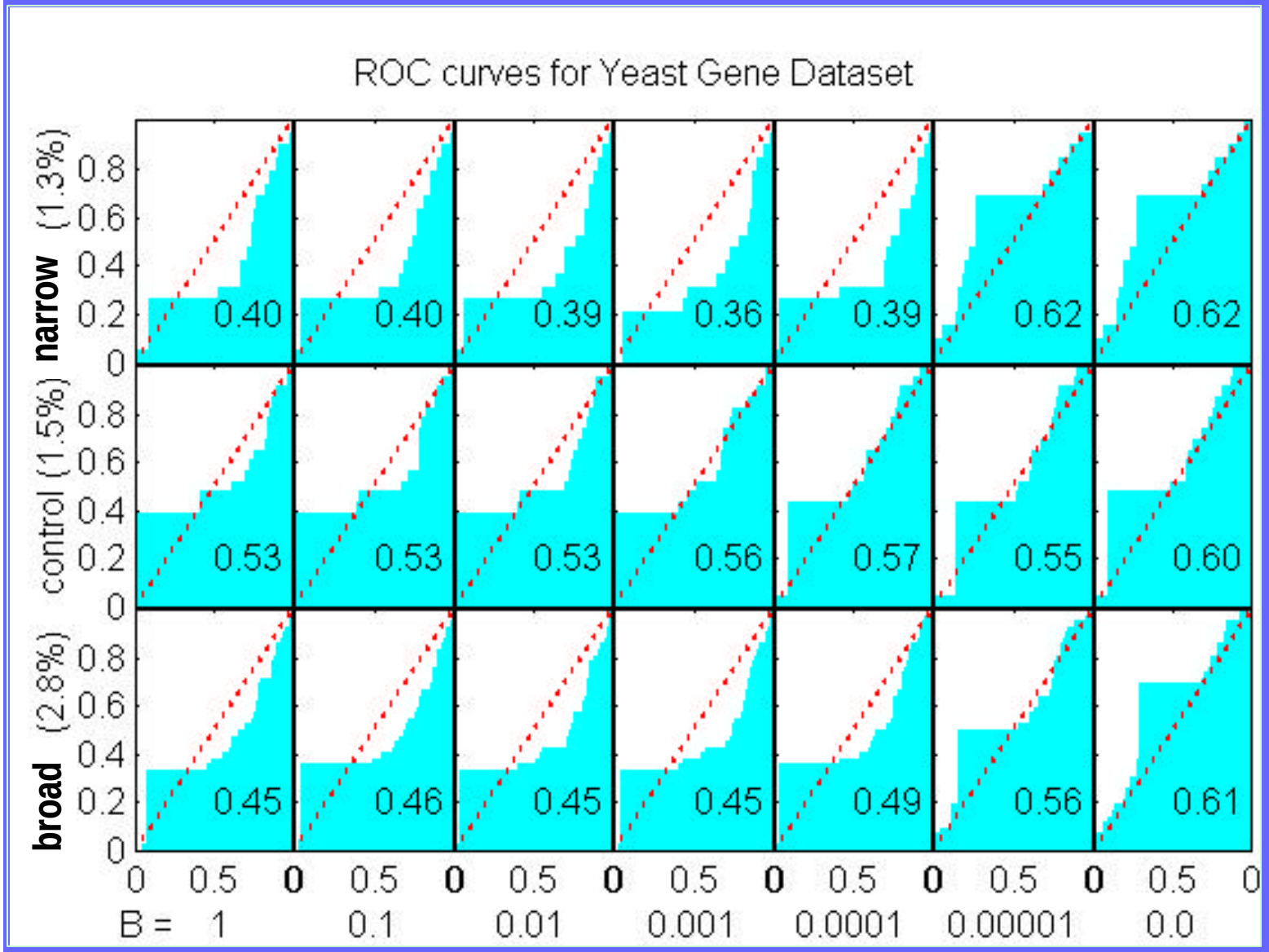
$C ? 0,$

$0 ? B ? 1 \quad ? \; \text{BalanceFactor}$

$n_{?1} ? 38 \text{ or } 84,$

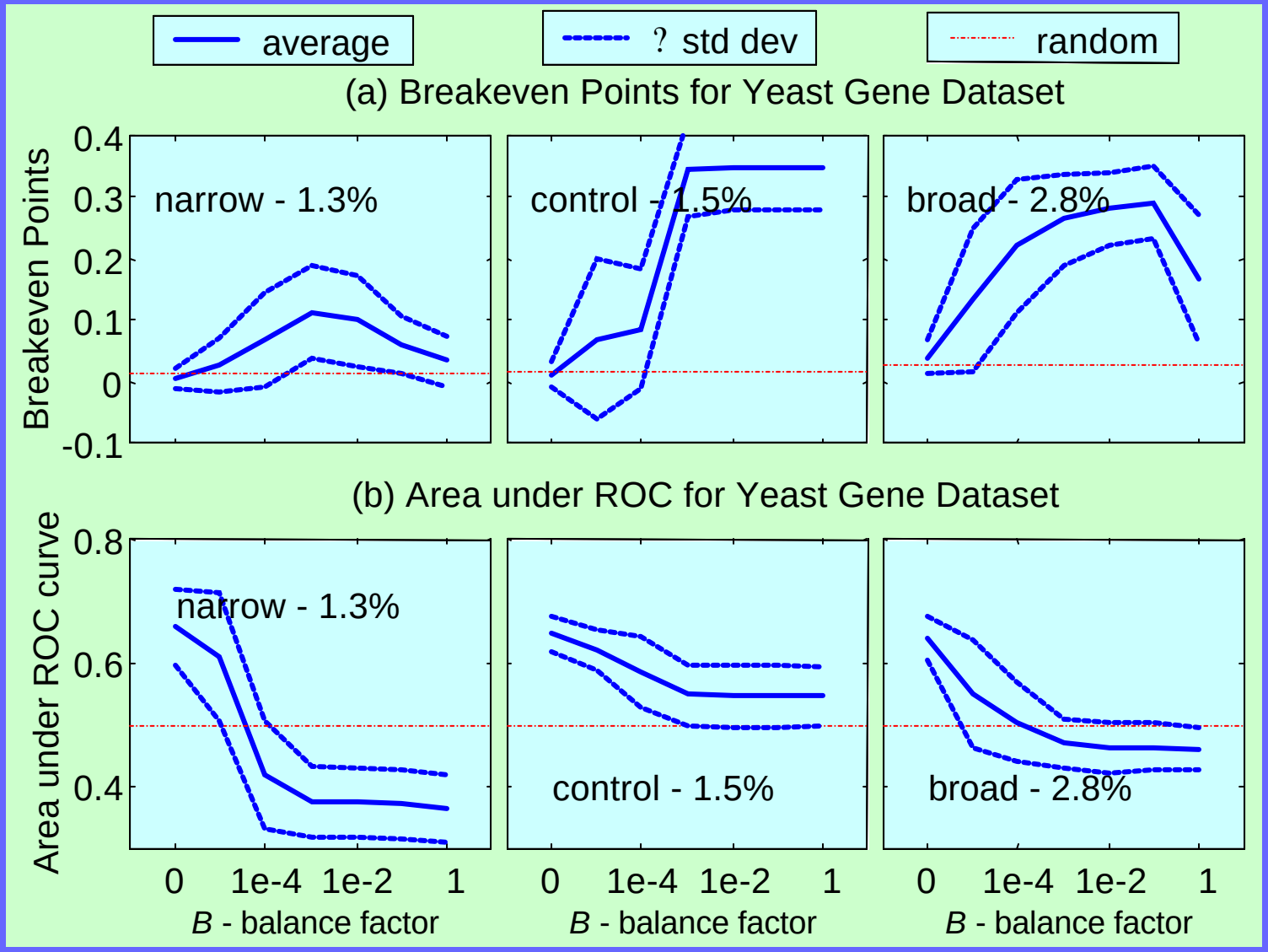
$n_{?1} ? 2980 \text{ or } 2934$

Why single class SVM?



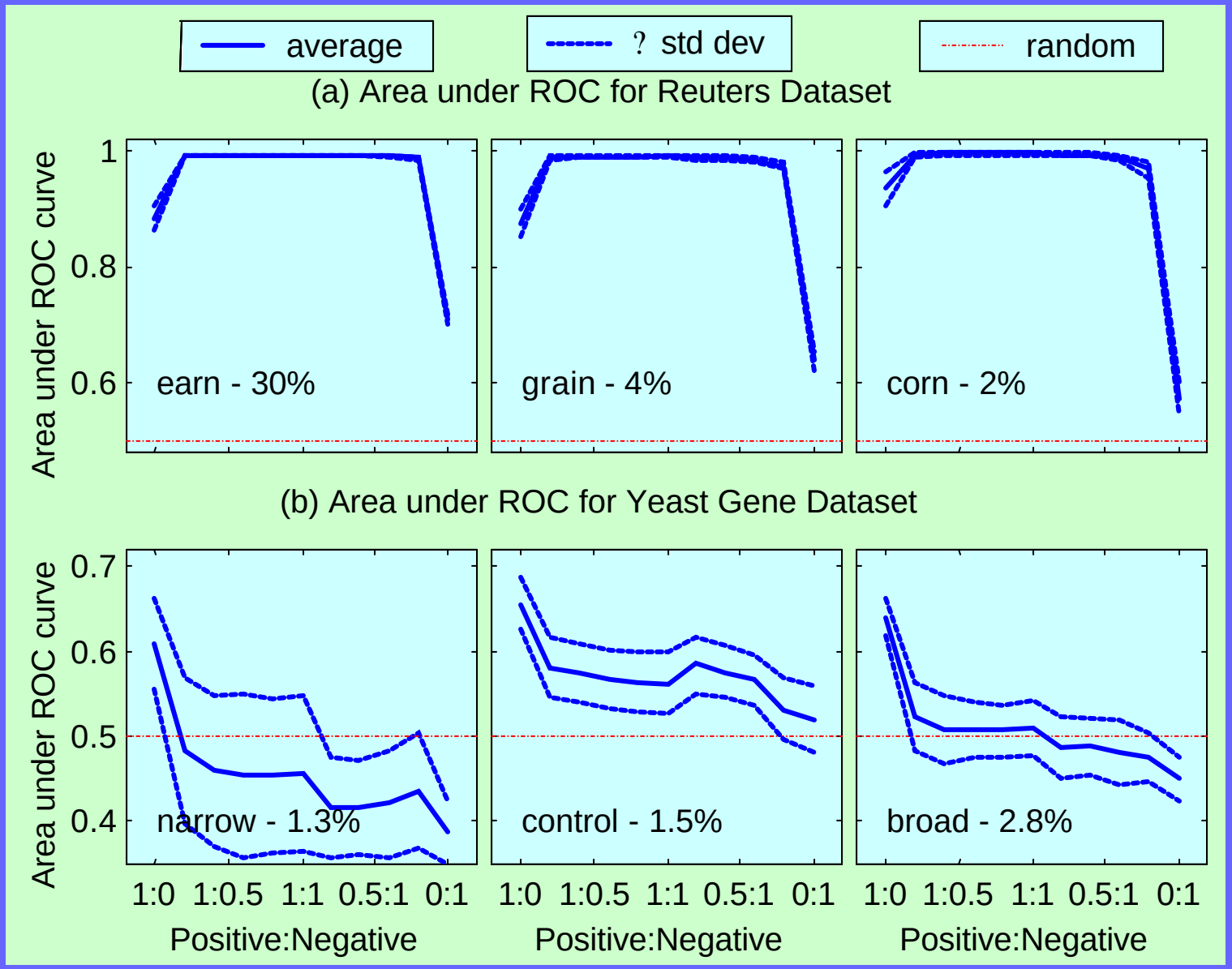
- even balance factor = 0.00001 is worse than 0
- ignoring negative examples gets best ROC!!
- is this true for other validation splits?

ROC and Breakeven Points



- different metrics
  - Different behavi
- best break-even a balance factor = 1e
- best ROC at balance factor = 0
- consistent behavio across 3 classes

Average ROC



- Hard Balance:
  - Vary amount of '+1' or '-1' example
- Behaviour for Reuters is normal
  - Best ROC with some '+1' and '-1' examples
  - Single class ROC better than random
- Surprising behaviou with Yeast Gene da
  - Best ROC with or positive examples

Winning model

- Single class (B = 0)
  - trained on 38 (narrow) and 84 (broad) out of 3018 examples
- "Hard margin" (C = 10000)
- Quadratic penalty (p = 2)
- 'All features'
- Lessons:
  - Discrimination is not always the best method
  - Explore single-class learning when negative class is noisy
- More info:
  - B. Raskutti and A. Kowalczyk, A Case when Supervised Learning Works Well or Better with Knowledge of a Single Class Only, submitted to NIPS 2002.

Open question: Why single class model does so well?

- Fluke?
- Strange data representation?
- Extraordinary data set?
- A feature of the (yeast) genetic code?

- How much the result can be improved if single class SVM is combined with different data representation and feature selection?