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

\[(x_i, y_i) \in \{\pm 1\}, \quad i = 1, \ldots, 3018\]

\[f(x) = \langle w, x \rangle + b \Rightarrow (w, b) \in (x, 1)\]

\[\tilde{w} \in \arg \min \|\tilde{w}\|^2 \quad \text{subject to} \quad C \int \max(0, 1 - y_i \tilde{w} \tilde{x}_i) \]

\[p = 1 \text{ or } 2,\]

\[C_{\pm 1} = \frac{C}{n_{\pm 1}}, \quad C_{\pm 2} = \frac{C}{n_{\pm 1}} B\]

\[C = 0,\]

\[0 \leq B \leq 1 \quad \text{ Balance Factor}\]

\[n_{\pm 1} = 38 \text{ or } 84,\]

\[n_{\pm 1} = 2980 \text{ or } 2934\]
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 behaviour
- best break-even at balance factor = 1e-2
- best ROC at balance factor = 0
- consistent behaviour across 3 classes
Average ROC

• Hard Balance:
  - Vary amount of ‘+1’ or ‘-1’ examples

• Behaviour for Reuters is normal
  - Best ROC with some ‘+1’ and ‘-1’ examples
  - Single class ROC better than random

• Surprising behaviour with Yeast Gene data
  - Best ROC with only positive examples

(a) Area under ROC for Reuters Dataset

(b) Area under ROC for Yeast Gene Dataset
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:
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?