Learning Minimum Volume Sets with Support Vector Machines

Mark A. Davenport, Richard G. Baraniuk
Rice University
Department of Electrical and
Computer Engineering

Clayton D. Scott
University of Michigan
Department of Electrical Engineering
and Computer Science

Overview
Use support vector machines to estimate
minimum volume sets (MV-sets)

• anomaly detection
• clustering

Key idea: reduce MV-set estimation to
Neyman-Pearson classification

• treat MV-set estimation (one-class problem)
as a two-class problem like classification
• draw second class from uniform distribution

Support Vector Machines
Method for learning classifiers from training data
• Use "kernel-trick"
• Maximize the "margin"

\[
\min_{w, b, \nu} \frac{1}{2}||w||^2 - \nu - \frac{1}{n} \sum_{i=1}^{n} \xi_i \\
\text{s.t. } (\langle w, x_i \rangle + b) y_i \geq \rho - \xi_i
\]

\[\nu \in [0, 1]\]

Neyman-Pearson Classifiers
Consider cost-sensitive SVM
• Introduce class-specific weights
• Adjust weights to achieve desired error rates

Relies on accurate error estimation
• cross-validation

\[
\min_{w, \lambda, \nu} \frac{1}{2}||w||^2 - \nu - \frac{\nu_+}{n+} \sum_{i \in F_+} \xi_i + \frac{\nu_-}{n-} \sum_{i \in F_-} \xi_i \\
\text{s.t. } \langle w, x_i \rangle + b Y_i \geq \rho - \xi_i
\]

\[\nu_+, \nu_- \in [0, 1]^2\]

Minimum Volume Sets
Given
• Probability measure \(P\)
• Reference measure \(\mu\) (typically Lebesgue)
• Target mass \(\beta\)

The minimum volume set is

\[G_\beta = \arg \min\{\mu(G) : P(G) \geq \beta, G \text{measurable}\}\]

Neyman-Pearson Classification
Given
• Probability measures \(Q_+\) and \(Q_-\)
• Target power \(\alpha\)

Let

\[P_\alpha(f) = Q_+(\{x : f(x) = +1\})
\]
\[P_\beta(f) = Q_+(\{x : f(x) = -1\})\]

The Neyman-Pearson classifier is

\[f_\alpha = \arg \min\{P_\beta(f) : P_\alpha(f) \leq \alpha\}\]

Uniform Data: Thinning
In high dimensions we must confront the
"curse of dimensionality"

One option is thinning the data to ensure a large
distance between any pair of points
• results in an approximate "packing set"

Results: MV-set Estimation
Compare with one-class SVM
Modified LIBSVM software

Highlights:
• manifold sampling performs best
• two-class methods more reliable
• impact of discrete data

Results: Anomaly Detection
Test validity of uniform prior
Compare
• MV-set (one class)
• NP-classifier (both classes)

Conclusions
Minimum volume sets are an effective way to
approach anomaly detection

We can accurately estimate minimum volume sets using Neyman-Pearson SVMs

The procedure used for generating "uniform" samples can significantly impact performance

Our approach tends to perform
• better than the one-class SVM
• often nearly as well the NP classifier trained using both classes

dsp.rice.edu