One-class Problems and Outlier Detection

陶 卿

Qing.tao@mail.ia.ac.cn

中国科学院自动化研究所

Application-driven

- Various kinds of detection problems: unexpected conditions in engineering; abnormalities in medical data, network intrusions and etc.

- Text classification or document classification;
- Multi-classification and clustering problems;
- Intelligence and Security Informatics.

Detection and learning

- There are only some sample observations of normal behavior, and our aim is to learn a general rule which can predict normal data points and then perform outlier or anomaly detection (outlier detection or novelty detection).

- The normal behavior of the system is unknown.

Training samples

- There are too few negative samples;
- There are too many negative samples;
- In a word, the representative negative samples are difficult to obtain and the severe unbalance exists.

A theoretical problem

- How to extend the learning theory developed for binary classifications to solve one-class problems?
- Specifically, margin and support vector methods.

Outline

- Reviewing binary classifications;
- Discussion of available algorithms for one-class problems;
- Our investigation.

Binary classifications

- Assumption： (x_i, y_i) i.i.d.
- Hypothesis space： H
- Loss function： $$c(y, f(x)) = \begin{cases} 0 & if \quad f(x) = y \\ 1 & if \quad f(x) \neq y \end{cases}$$
- Objective function： $$R(f) = \int c(y, f(x)) P(x, y) dx$$

SVM

- Linearly separable: hard margin;
- Linearly inseparable: soft margin;
- Nonlinear classifier: kernel technique and soft margin algorithms in feature space;
- The entire framework: from simple to complex and in terms of hypothesis space.

Theoretical Analysis: margin

- Over the last decade, both theories and practices have pointed out that the concept of margin is central to the success of SVM and boosting.
- Generally, a large margin implies good generalization performance.

Theoretical Analysis: soft margin

- Commonly regarded as a relaxation of hard margin optimization problems.
- A specific hard margin algorithm in a feature space without changing the generalization
  [Schawe-Taylor. IEEE Trans on Information Theory. 2002]
- The maximal margin algorithm for linearly separable cases is central to the success of entire framework of SVMs

Support vector structure

- The solution is a linear combination of training samples.
- It naturally leads to dual algorithms, kernel algorithms. For example, perceptron, ridge regression.
- Also leads to sequential algorithms (objective-function driven) and geometric fast training algorithms.

Margin

- Theoretical analysis;
- Algorithms designation;
- The convergence speed of learning algorithms. For perceptron, Novikoff theorem.

Summary

- Definition: loss function;
- Margin: generalization;
- Algorithm: a complete framework and support vector structure.

- Discussion of available algorithms for one-class problems

Definition of the problems

- What is the mathematics description of one-class problem?
- Different viewpoints will lead to different algorithms.

Historical background

- Outlier detection is a well-known problem in statistics which is defined as a event in a low density level set.

- Unfortunately, the algorithms considered in those articles cannot be used since the imposed assumptions on the distribution are often tailored to specific applications and are in general unrealistic for anomalies.

Other viewpoints

- Anomalies are not concentrated.
- Estimating function with a desired region.

SVDD

- From the viewpoint of SV structure, the first paper may be

D. Tax and R. Duin. Support vector domain description. Pattern Recognition Letters. 1999, 20: 1191-1199.

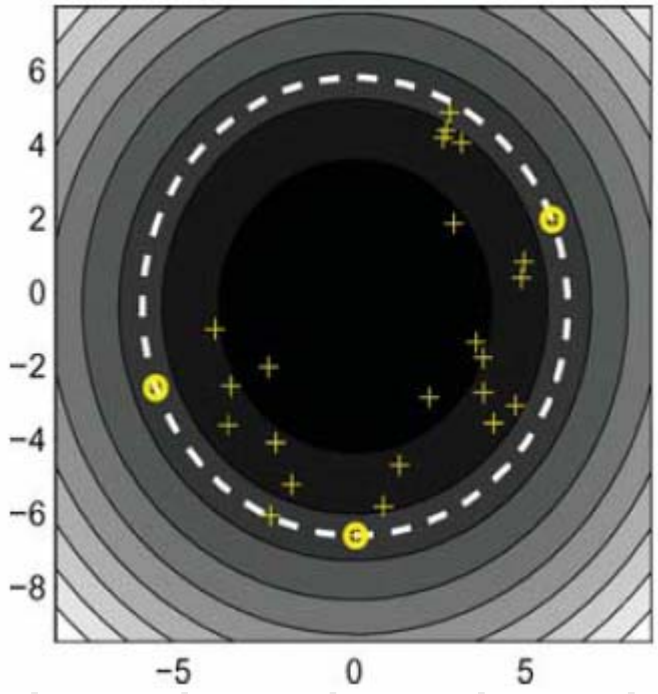
D. Tax and R. Duin. Support vector data description. Machine Learning, 2004.

Minimal covering ball

- Minimal radius

$$\min_{R \in R} \; R^2$$

$$s.t. \left| \Phi(x_i) - c \right| \le R^2, i = 1, 2, ..., l$$



SVM framework

- SV structure and dual optimization problems;

$$\max_{\alpha_i} \quad \sum_{i=1}^{m} \alpha_i k(\mathbf{x}_i, \mathbf{x}_i) - \sum_{i,j=1}^{m} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i = 1,$$

$$\mathbf{c} = \sum_{i=1}^{m} \alpha_i \varphi(\mathbf{x}_i)$$

- Soft algorithms

$$\min_{R \in \mathbb{R}} \; R^2 + \frac{1}{vl} \sum_{i=1}^{l} \xi_i$$

$$s.t. \; \left| \Phi(x_i) - c \right| \leq R^2 + \xi_i, i = 1, 2, \ldots, l$$

Existing problems in SVDD

- Only an implicit loss function, the definition of one-class problems or outliers is not discussed.
- The relation between the hard and soft algorithms.
- Generalization problems.

Support of distributions

- B. Schölkopf, et al. Estimating the Support of a High-Dimensional Distribution. Neural Computation. 2001.
- Estimating the region of high probability
- Usually called one-class SVMs.

One-class SVMs

- Estimating a region which containing a large fraction of the training data while keeping the value of some regularizer small.
- An SV-style large margin regularizer.
- Estimating a linear function $f(x) \geq \rho$: separating training data from origin.

Connection to SVDD

- If the RBF kernel is used, estimating the support of distributions in feature spaces can be reformulated as a soft minimum covering ball.

Existing problems

- The definition of one-class problems or outliers is not discussed.
- The relation between about one-class and binary classification problems is not analyzed theoretically.

DLD and outlier

- I. Steinwart et al.. A Classification Framework for Anomaly Detection. JMLR, 2005.

- Density Level Detection: finding a density level set to detect anomalous observations.

DLD and learning

- The objective function

$$S_{\mu,h,\rho}(f) := \mu\Big(\{f > 0\} \triangle \{h > \rho\}\Big),$$

- Unfortunately, there is no method known to estimate it from the training set with guaranteed accuracy and such a method may not exist.

DLD: a legitimate risk

Now let μ be a probability measure and define the risk

$$\mathcal{R}(f) := \frac{1}{1+\rho} Q(f \leq 0) + \frac{\rho}{1+\rho} \mu(f > 0).$$

$$dQ = h d\mu.$$

- Any function that minimizes R also minimizes S. Furthermore, there exists a very tight relation between R and S.

DLD-SVM

- The only difference is that the class marginal probabilities are known

$$P(\mathbf{y}=1) := 1/(1+\rho) \qquad P(\mathbf{y}=-1) := \rho/(1+\rho)$$

$$\begin{array}{ll} \min_{\psi,b,\xi} & \lambda\|\psi\|^2 + \sum_{i=1}^{n} u_i \xi_i \\ \text{s.t.} & y_i(\phi(x_i)\cdot\psi + b) \geq 1-\xi_i \\ & \xi_i \geq 0, \quad i = 1, 2, ..., n \end{array}$$

where $\lambda > 0$ and

$$u_i = \begin{cases} \frac{1}{(1+\rho)n_1}, & y_i = 1 \\ \frac{\rho}{(1+\rho)n_{-1}}, & y_i = -1 \end{cases}.$$

A classification framework

- Making the centralized viewpoints precise in statistics: Density Level Detection;
- Interpreting DLD as binary classification and their objective functions are asymptotically equivalent under mild conditions on the density.

DLDSVM: generating samples

- Justifying the well-known heuristic:

  to generate a labeled data set by assigning one label to the original unlabeled data and another label to a set of artificially generated data.

Summary

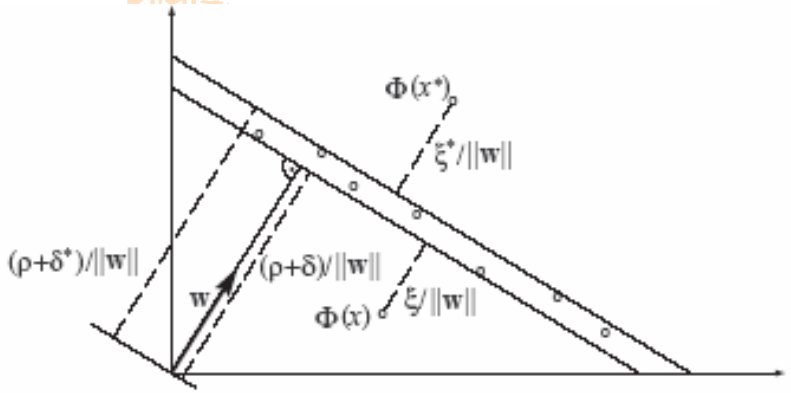
- SVDD: a SVM-like framework for one-class algorithms based on intuition.
- One-class SVMs: a binary classification framework for one-class problem.
- DLD-SVM: a classification framework based on making the viewpoint precise.

Restricting the optimal region

- Estimating function with a desired region.
- Slab-SVM: B. Schölkopf, et al. Kernel Methods for Implicit Surface Modeling. NIPS, 2004.
- Slab-one-class: Q. Tao, et al. A new maximum margin algorithm for one-class problems and its boosting implementation. Pattern Recognition. 2005.

Slab-SVM

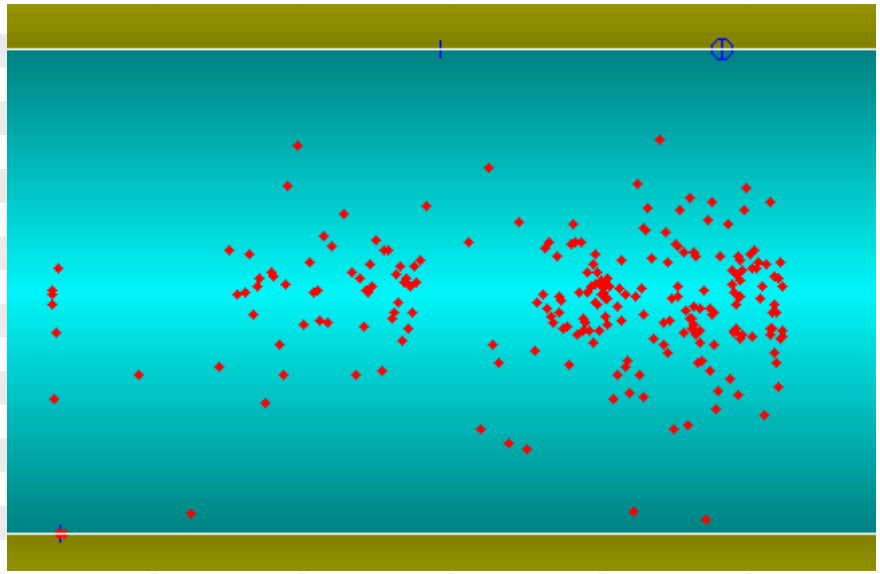
- Estimate a region which is a slab in the RKHS, i.e., further restricting the optimal region.



$$\begin{aligned} \underset{\mathbf{w}\in\mathcal{H},\boldsymbol{\xi}^{(*)}\in\mathbb{R}^m,\rho\in\mathbb{R}}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu m}\sum_i(\xi_i+\xi_i^*) - \rho \\ \text{subject to} \quad & \delta - \xi_i \le \langle \mathbf{w}, \Phi(x_i)\rangle - \rho \le \delta^* + \xi_i^* \\ \text{and} \quad & \xi_i^{(*)} \ge 0. \end{aligned}$$

Slab one-class

- Loss function;
- Margin.



Loss function

$$L(f(x),\eta)=\begin{cases}0 & if \ \ |f(x)|\le\eta\\ 1 & else\end{cases}$$

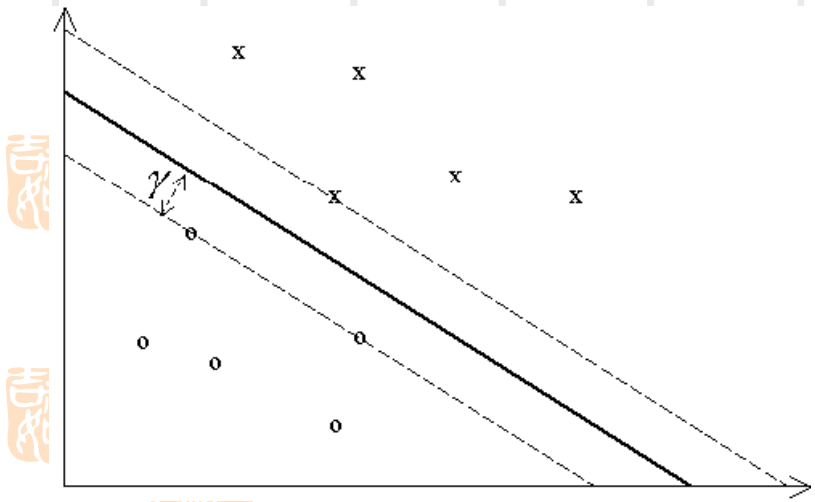
$$R(f,\eta)=\int L(f(x),\eta)\,p(x)dx$$

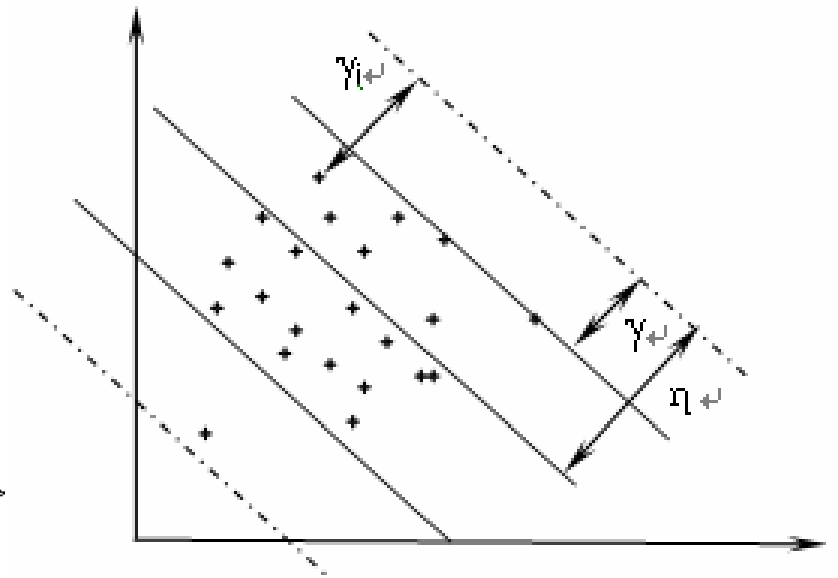
- The expected risk is the total probability that a point is outside the slab. It reflects the averaged error of all the samples.

Introducing η

- $\|w\|_p = 1 \quad \gamma_i = m(f, x_i, \eta) = \eta - |w \cdot x_i + b|$

$$\gamma = m(f, S, \eta) = \min\left(m(f, x_i, \eta)\right), \; i = 1, 2, \ldots, l$$





Maximum margin algorithms

- To obtain a minimal width zone that contains all the samples.

$$\begin{cases} \max\limits_{w,b,\gamma} \ \gamma \\ \|w\|_p = 1 \quad w \cdot x_i + b + \eta \geq \gamma \quad w \cdot x_i + b - \eta \leq -\gamma \quad i = 1, 2, ..., l \end{cases}$$

- Let $\rho = \eta - \gamma$

$$\begin{cases} \min\limits_{w,b,\rho} \ \rho \\ \|w\|_p = 1 \quad w \cdot x_i + b \geq -\rho \quad w \cdot x_i + b \leq \rho \quad i = 1, 2, ..., l \end{cases}$$

Theoretical analysis

- We can analyze the maximum margin optimization problem using the methodology developed for support vector machines.

- N. Cristianini and J. Schawe-Taylor. An Introduction to Support Vector Machines. Cambridge: Cambridge Univ Press. 2000.

Soft and nonlinear algorithms

- Soft algorithms:

$$\begin{cases} \min\limits_{w,\rho,\xi,\xi^*} \ \nu\rho + \sum\limits_{i=1}^{l}\left(\xi_i + \xi_i^*\right) \\ \Phi\left(x_i\right) \geq -\rho - \xi_i^*, \ \Phi\left(x_i\right) \leq \rho + \xi_i, \ \ i = 1, 2, ..., l \\ \left\|w\right\|_1 = 1, \quad \xi_i \geq 0, \ \ \xi_i^* \geq 0, \ \ w \in R_+^N, \quad i = 1, 2, ..., l \end{cases}$$

- By restricting the margin to be in terms of L1 norm, the boosting techniques can be employed.

Interpretation of ν

$$\sum_{i=1}^{l} I\left(\left|f_{w_0}(x_i)\right| > \rho_0\right) \leq \nu \leq \sum_{i=1}^{l} I\left(\left|f_{w_0}(x_i)\right| \geq \rho_0\right)$$

- Similar to that in ε-insensitive ν-support vector regression.

Existing problems

- The optimization is not convex.
- No SV structure in terms of L2 norm.
- Even in L1 norm cases, the direction of the weight vector is imposed on some restriction.

Our investigation

- Reconsidering our previous paper.
- New findings on SVDD.

Philosophy in unsupervised learning

- Without imposing some restrictive conditions on the model, it is hard to obtain any meaningful theoretical results;

- To allow theoretical analysis, these conditions may be so that the model does not exactly refer to any specific practical problem.

Consistency problems

- The assumption may be not reasonable, or, the objective function may be very specified;
- The proposed algorithm should coincide with the objective function in terms of learning.

Unsupervised learning algorithms

- Every unsupervised learning algorithm must come with a set of parameters that can be used to adjust the algorithm to a particular problem or according to a particular principle.
- Such parameters should be as small as possible, and their effect on the behavior of the algorithm should be as clear as possible.

Reconsidering problems

- The defined loss function is still reasonable.
- Non-convex and pre-defined parameters: the same phenomena exists in regression problems.

New findings

- The maximum margin algorithm specializes to SVDD for the case of pointwise hypothesis space.
- The soft SVDD optimization problem which allows margin violation is essentially a specific maximum margin algorithm using an auxiliary function space.

Margin for MEB

- Changing the hypothesis space

$$H = \{f : \ f \in R^N\} \text{ and define } \triangle(x_i, f) = \|x_i - f\|^2.$$

$$\gamma = min\{\eta - \|x_i - x_0\|^2, i = 1, 2, \ldots, l.\}.$$

$$\begin{cases} max_{\{\gamma,\ x_0 \in R^n\}} \ \gamma \\ \eta - \|x_i - x_0\|^2 \geq \gamma, \quad i = 1, 2, \ldots, l. \ \ \gamma \geq 0. \end{cases}$$

Consistency about soft margin

$$\begin{cases} min_{\{R,x_0,\xi_i\}} \quad R^2 + \dfrac{1}{\nu l}\displaystyle\sum_{i=1}^{l}\xi_i \\ \|x_i - x_0\|^2 \le R^2 + \xi_i, \ \xi_i \ge 0, \quad i = 1, 2, \ldots, l. \end{cases}$$

$$\begin{cases} max_{\{\gamma,x_0,\xi_i\}} \quad \gamma - \dfrac{1}{\nu l}\displaystyle\sum_{i=1}^{l}\xi_i \\ \eta - \|x_i - x_0\|^2 \ge \gamma - \xi_i, \ \xi_i \ge 0, \quad i = 1, 2, \ldots, l. \quad \eta \ge \gamma \ge 0. \end{cases}$$

Conclusion

- Giving a strong justification for the SVDD formulations from the point of view of our defined loss function and risk.

- η-one-class is a reasonable framework for designing and analyzing one-class learning algorithms.

Summary and reconsidering

- From the point of view of unsupervised learning:
- We restrict information preservation by making some a-priori assumptions on the loss function and dimension reduction by specifying the hypothesis space.

- **Thanks!**