One-Class Support Measure Machines for Group Anomaly Detection

Krikamol Muandet **Bernhard Schölkopf**

Empirical Inference Department
Max Planck Institute for Intelligent Systems, Tübingen, Germany

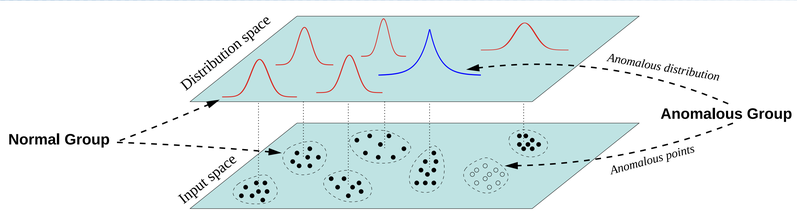
MAX-PLANCK-GESELLSCHAFT

uai2013

Abstract

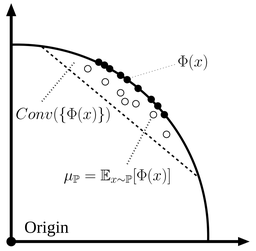
We propose one-class support measure machines (OCSMMs) for group anomaly detection. Unlike traditional anomaly detection, OCSMMs aim at recognizing anomalous aggregate behaviors of data points. The OCSMMs generalize well-known one-class support vector machines (OCSVMs) to a space of probability measures. By formulating the problem as quantile estimation on distributions, we can establish interesting connections to the OCSVMs and variable kernel density estimators (VKDEs) over the input space on which the distributions are defined, bridging the gap between large-margin methods and kernel density estimators. In particular, we show that various types of VKDEs can be considered as solutions to a class of regularization problems studied in this paper. Experiments on Sloan Digital Sky Survey dataset and High Energy Particle Physics dataset demonstrate the benefits of the proposed framework in real-world applications.

Group Anomaly Detection



Geometric Interpretation of OCSMMs

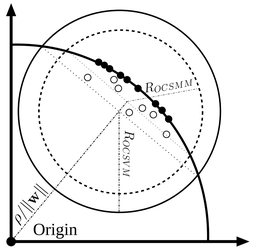
Kernel Mean Embedding in RKHS



The two dimensional representation of the RKHS of Gaussian RBF kernels. Since the kernels depend only on $x - x'$, $k(x,x)$ is constant. Therefore, all feature maps $\Phi(x)$ (black dots) lie on a sphere in feature space. Hence, for any probability distribution $\mathbb{P}$, its mean embedding $\mu_{\mathbb{P}}$ always lies in the convex hull of the feature maps, which in this case, forms a segment of the sphere.

Theorem 1. *There exists a unique separating hyperplane $\boldsymbol{w}$ as a solution to (2) that separates $\mu_{\mathbb{P}_1}, \mu_{\mathbb{P}_2}, \ldots, \mu_{\mathbb{P}_\ell}$ from the origin.*

OCSMM and Minimum Enclosing Sphere

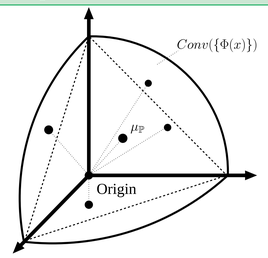


In general, the solution of OCSMM is different from the minimum enclosing sphere [TD99; TD04]. However, we can apply the spherical normalization

$$\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \longmapsto \frac{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}}{\sqrt{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}}}.$$

After the normalization, $\|\mu_{\mathbb{P}}\|_{\mathcal{H}} = 1$ for all $\mathbb{P} \in \mathfrak{P}_{\mathcal{X}}$. That is, all mean embeddings lie on the unit sphere in the feature space.

Spherical Normalization



Three dimensional sphere in the feature space. For the Gaussian RBF kernel, the kernel mean embeddings of all distributions always lie inside the segment of the sphere. In addition, the angle between any pair of mean embeddings is always greater than zero. Consequently, the mean embeddings can be scaled, e.g., to lie on the sphere, and the map is still injective.

Theorem 2. *Assume that k is characteristic and the samples are linearly independent in the feature space $\mathcal{H}$. Then, the spherical normalization preserves the injectivity of the mapping $\mu : \mathfrak{P}_{\mathcal{X}} \to \mathcal{H}$.*

Experimental Results

We compare the following group anomaly detection algorithms:

1. k-nearest neighbor with NP-L_2 divergence (KNN-L2)
2. k-nearest neighbor with NP-Renyi divergence (KNN-Renyi)
3. Multinomial genre model (MGM)
4. One-class support vector machine (OCSVM)
5. One-class support measure machine (OCSMM)

Noisy Data

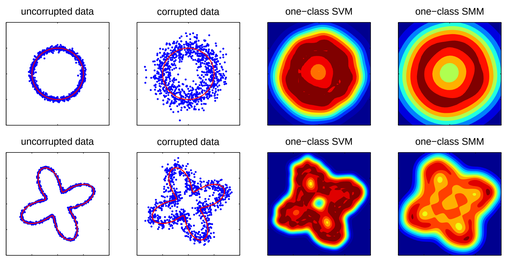


Figure 1: The density functions estimated by the OCSVM and the OCSMM using the corrupted data.

Synthetic Data

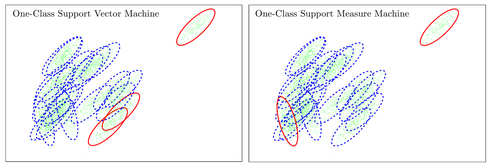


Figure 2: The results of group anomaly detection on synthetic data obtained from the OCSVM and the OCSMM.

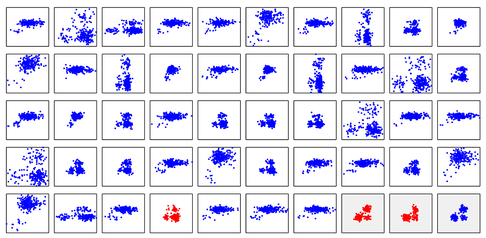


Figure 3: The results of the OCSMM on the synthetic data of the mixture of Gaussian. The shaded boxes represent the anomalous groups that have different mixing proportion.

High Energy Particle Physics

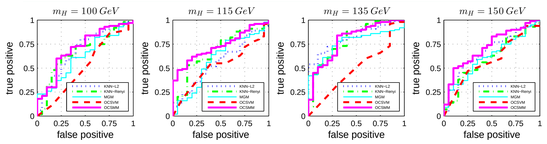


Figure 4: The ROC of different group anomaly detection algorithms on Higgs boson datasets with various Higgs masses.

Table 1: The associated AUC scores for different settings

m_H	KNN-L2	KNN-Renyi	MGM	OCSVM	OCSMM
100 GeV	0.6835	0.6655	0.6350	0.5125	**0.7085**
115 GeV	0.5645	0.6783	0.5860	0.5263	**0.7305**
135 GeV	**0.8190**	0.7925	0.7630	0.4958	0.7950
150 GeV	0.6713	0.6027	0.6165	0.5862	**0.7200**

Sloan Digital Sky Survey

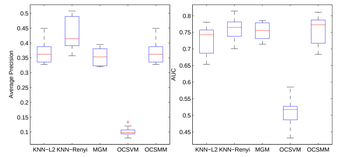


Figure 5: The average precision (AP) and area under the ROC curve (AUC) of different group anomaly detection algorithms.

One-Class Support Measure Machines (OCSMMs)

Kernel Mean Embedding: The kernel mean map from a space of distributions $\mathfrak{P}_{\mathcal{X}}$ into an RKHS $\mathcal{H}$ is defined as

$$\mu : \mathfrak{P}_{\mathcal{X}} \to \mathcal{H}, \;\; \mathbb{P} \longmapsto \int_{\mathcal{X}} k(x, \cdot)\, \mathrm{d}\mathbb{P}(x) \;. \tag{1}$$

OCSMM Formulation: Using (1), the primal optimization problem for one-class SMM can be subsequently formulated in an analogous way to the one-class SVM [Sch+01] as follow:

$$\underset{\mathbf{w}, b, \xi, \rho}{\text{minimize}} \quad \frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle_{\mathcal{H}} - \rho + \frac{1}{\nu\ell}\sum_{i=1}^{\ell} \xi_i \qquad \text{subject to} \quad \langle \mathbf{w}, \mu_{\mathbb{P}_i} \rangle_{\mathcal{H}} \geq \rho - \xi_i, \xi_i \geq 0. \tag{2}$$

By introducing Lagrange multipliers $\boldsymbol{\alpha}$, we have $\mathbf{w} = \sum_{i=1}^{\ell} \alpha_i \mu_{\mathbb{P}_i} = \sum_{i=1}^{\ell} \alpha_i \mathbb{E}_{\mathbb{P}_i}[k(x, \cdot)]$ and the dual form of (2) can be written as

$$\underset{\boldsymbol{\alpha}}{\text{minimize}} \quad \frac{1}{2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell} \alpha_i \alpha_j \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j} \rangle_{\mathcal{H}} \qquad \text{subject to} \quad 0 \leq \alpha_i \leq \frac{1}{\nu\ell}, \; \sum_{i=1}^{\ell} \alpha_i = 1\,. \tag{3}$$

Kernel on Probability Distributions: From (3), we can see that $\mu_{\mathbb{P}}$ is a feature map associated with the kernel $K : \mathfrak{P}_{\mathcal{X}} \times \mathfrak{P}_{\mathcal{X}} \to \mathbb{R}$, defined as $K(\mathbb{P}_i, \mathbb{P}_j) = \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j} \rangle_{\mathcal{H}}$. It follows from Fubini's theorem and reproducing property of $\mathcal{H}$ that

$$\langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j} \rangle_{\mathcal{H}} = \iint k(x, y)\, \mathrm{d}\mathbb{P}_i(x)\, \mathrm{d}\mathbb{P}_j(y) \qquad \Longrightarrow \qquad K(\widehat{\mathbb{P}}_i, \widehat{\mathbb{P}}_j) = \frac{1}{n_i \cdot n_j}\sum_{k=1}^{n_i}\sum_{l=1}^{n_j} k(x_k^{(i)}, x_l^{(j)})$$

where $x_k^{(i)} \in S_i$, $x_l^{(j)} \in S_j$, and n_i is the number of samples in S_i for $i = 1, \ldots, \ell$.

OCSMM and Kernel Density Estimation

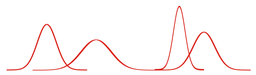
The KDE of an unknown density f from an i.i.d. sample $x_1, x_2, \ldots, x_n$ is

$$\hat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{y - x_i}{h}\right)$$

Identical Bandwidth

If $\sigma_i = \sigma_j$ for all $1 \leq i, j \leq n$, the OCSMM is equivalent to the OCSVM on the training samples $m_1, m_2, \ldots, m_n$ with Gaussian RBF kernel $k_{\sigma^2+\sigma_i^2}$. Hence, the OCSMM corresponds to the OCSVM on the means of the distributions with kernel of larger bandwidth.

Variable Bandwidth



If $\sigma_i \neq \sigma_j$ for some $1 \leq i, j \leq n$, the OCSMM is equivalent to the OCSVM on the training samples $m_1, m_2, \ldots, m_n$ with Gaussian RBF kernel $k_{\sigma^2+\sigma_i^2}$. The kernel bandwidth may be different at each training samples, i.e., OCSMM $\equiv$ OCSVM with variable bandwidth parameters.

Conclusions

- We propose a simple and efficient algorithm for detecting group anomalies called one-class support measure machine (OCSMM).
- To handle aggregate behaviors, groups are represented as probability distributions which account for higher-order information arising from those behaviors. The set of distributions are represented as mean functions in the RKHS via the kernel mean embedding.
- We also extend the relationship between the OCSVM and the KDE to the OCSMM in the context of variable kernel density estimation, bridging the gap between large-margin approach and kernel density estimation.

References

[Mua+12] Krikamol Muandet et al. "Learning from Distributions via Support Measure Machines". In: *Advances in Neural Information Processing Systems (NIPS)*. 2012, pp. 10–18.

[Sch+01] Bernhard Schölkopf et al. "Estimating the Support of a High-Dimensional Distribution". In: *Neural Computation* 13 (7 2001), pp. 1443–1471.

[TD04] David M. J. Tax and Robert P. W. Duin. "Support Vector Data Description". In: *Machine Learning* 54.1 (2004), pp. 45–66.

[TD99] David M. J. Tax and Robert P. W. Duin. "Support vector domain description". In: *Pattern Recognition Letters* 20 (1999), pp. 1191–1199.