Support Measure Data Description

A One-Class Classifier for Group Anomaly Detection

Jorge Guevara
IBM Research, Brazil
jorgegd@br.ibm.com

Stéphane Canu
Normandie Université,France
scanu@insa-rouen.fr

Roberto Hirata Jr
University of Sao Paulo, Brazil
hirata@ime.usp.br

ABSTRACT

We propose the *support measure data description* (SMDD) model which is a one-class classifier for sets of probability distributions. There are practical data mining applications where observations are better described by probability distributions rather than individual points, for instance, point-wise uncertainty, replicates measurements, clusters of points, and so on. Hence, the anomaly detection task on those datasets can be formulated as the detection of anomalous probability distributions w.r.t. the distributions with non-anomalous behavior in the data. The SMDD uses the methodology of kernel embedding of distributions and it is defined as the minimum enclosing ball of those embeddings. The SMDD does not assume anything for the probability distributions but it encodes prior knowledge of distributions by means of a kernel function. We conducted an experimental study on the group anomaly detection task on artificial and real datasets to show the effectiveness of the SMDD classifier.

CCS CONCEPTS

•Computing methodologies →Anomaly detection; Kernel methods;

KEYWORDS

Kernel embedding of distributions, group anomaly detection, kernel on probability measures, minimum volume set, support measure machine, one-class classifier

ACM Reference format:
Jorge Guevara, Stéphane Canu, and Roberto Hirata Jr. 2017. Support Measure Data Description . In *Proceedings of 23rd SIGKDD Conference on Knowledge Discovery and Data Mining, Halifax, Nova Scotia Canada, August 2017 (ACM SIGKDD 2017),* 8 pages.
DOI: 10.475/123_4

1 INTRODUCTION

One-class classifiers are widely used as data description models for datasets containing observations of the same kind. Formally, a one-class classifier is a function in some space that "encodes the behavior" of a set of points following the same probabilistic law. Such function could be explicitly given by some expert or it can be learned from data. Those classifiers are widely used for anomaly or novelty detection, clustering and classification tasks [4, 21–23, 29].

ACM SIGKDD 2017, Halifax, Nova Scotia Canada

DOI: 10.475/123_4

We propose the *support measure data description* (SMDD) one-class classifier, a novel approach to successfully describe a set of probability distributions. Although the SMDD can be used in several machine learning tasks on probability distributions (clustering for example) we study its practical application on the *anomaly detection task over a set of probability distributions*, i.e., the task of detecting anomalous probability distributions. We will show experimentally that our method has state-of-the-art performance on that task. The SMDD classifier is a kernel method that belongs to the class of support measure machines, that is, kernel machines defined on sets of probability distributions (measures)[1]. The SMDD classifier describes a dataset of probability distributions by estimating a decision function f in a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ that approximates the description of datasets comprised of probability distributions as in Lemma 2 and Theorem 1 of [18].

Thus, our main contributions are:

- We formulate the SMDD one-classifier as the minimum enclosing ball in a RKHS of the kernel embeddings of probability distributions. In this way, all the kernel embeddings of probability distributions lying within the minimum enclosing ball will characterize the non-anomalous category of probability distributions.
- We show that a SMDD classifier is an approximation of a minimum volume set of probability distributions. We show that by imposing geometrical restrictions on the kernel embeddings of those distributions it is possible to formulate different versions of the SMDD classifier. For instance, we show an scenario when a SMDD is formulated as a chance constrained program.
- We evaluate the SMDD using synthetic and real datasets as is the case of finding anomalous clusters of galaxies from the Sloan Digital Sky Survey Project. The experiments show the effectiveness of the SMDD.

Reproducibility Our code and data is available in **https://github.com/jorjasso/SMDD-group-anomaly-detection**

2 BACKGROUND AND PROBLEM DEFINITION

This section presents an overview of the Hilbert space embedding of probability distributions and the group anomaly detection task. We also present the problem definition.

2.1 Kernel embeddings of distributions

A kernel embedding of a probability measure $\mathbb{P} \in \mathcal{P}$ into a RKHS $\mathcal{H}$ is the mapping from $\mathcal{P}$ to $\mathcal{H}$ defined by $\mathbb{P} \mapsto \mu_{\mathbb{P}} = \mathbb{E}_{\mathbb{P}}[k(X, .)]$,

[1] A probability distribution is the probability measure induced by a random variable. We will use both terms interchangeably.

where $\mathbb{E}$ denotes the expectation, X is a random variable distributed according $\mathbb{P}$ and k is a real-valued positive definite kernel on $\mathbb{R}^D \times \mathbb{R}^D$. Notation $k(.,s)$ means the mapping $t \rightarrow k(t,s)$ with fixed s. The element $\mu_{\mathbb{P}}$ is called *mean map* and it is the representative function for $\mathbb{P}$ in $\mathcal{H}$ [1, 10, 12, 24, 24, 27, 28]. A sufficient condition guaranteeing the existence of $\mu_{\mathbb{P}}$ in $\mathcal{H}$ is given by assuring that $\mu_{\mathbb{P}}(X) = \mathbb{E}_{\mathbb{P}}[k(X,X)] < \infty$, and k being a measurable function [10, 24, 27]. As a consequence, the reproducing property $\langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} = \langle f, \mathbb{E}_{\mathbb{P}}[k(X,.)] \rangle_{\mathcal{H}} = \mathbb{E}_{\mathbb{P}}[f(X)]$ holds for all $f \in \mathcal{H}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product in $\mathcal{H}$.

The kernel embedding $\mu : \mathcal{P} \rightarrow \mathcal{H}$ is injective if k is *characteristic* [8, 26, 27]. Thanks to this, the embedding induces a metric on the space of probability measures. Examples of characteristic kernels are the Gaussian, Laplacian, inverse multiquadratics, B_{2n+1}-splines kernels. [27]. Moreover, empirical estimators μ_{emp} for $\mu_{\mathbb{P}}$ approximate well the true mean map, that is, the term $\|\mu_{\mathbb{P}} - \mu_{emp}\|$ is bounded by a small value [24]. As consequence, we have the following kernel on probability measures.

Proposition 2.1 (Kernel on probability measures). *A real-valued kernel on $\mathcal{P} \times \mathcal{P}$, defined by*

$$\tilde{k}(\mathbb{P}, \mathbb{Q}) = \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} = \int_{\mathbf{x} \in \mathbb{R}^D} \int_{\mathbf{x}' \in \mathbb{R}^D} k(\mathbf{x}, \mathbf{x}') d\mathbb{P}(\mathbf{x}) d\mathbb{Q}(\mathbf{x}') \quad (1)$$

is positive definite [1].

Hilbert space embedding of measures was introduced in [12], and later by [1, 28], and by [24] when the measures are probability measures. Some applications in machine learning include, dimensionality reduction [7], measuring independence of random variables [11], two-sample test [10], embeddings of Hidden Markov Models into RKHS [25], among others [26, 27]. The kernel on probability measures can be estimated using (2) without requiring fitting some probabilistic models to the observations. Other related kernels on distributions which assume probabilistic models for observations are the Fisher kernel [13], the kernel based on the symmetrized Kullback-Leibler (KL) divergence on distributions [17], the Bhattacharyya kernel [16], and the probability product kernel [14].

2.2 Group anomaly detection

This kind of anomaly detection is defined as the process of finding out anomalous groups of points (observations) from datasets of the form:

$$\mathcal{T} = \{s_i \mid s_i \subset \mathbb{R}^D, 1 \leq i \leq N\} \quad (2)$$

where $N \in \mathbb{N}$ is the number of *observations* (groups of points) and each observation s_i is a non-empty set of points in $\mathbb{R}^D$. Group anomalies can be categorized in two types: [30] *point-based* anomaly, i.e., the aggregation of anomalous points, or a *distribution-based* anomaly, i.e., the anomalous aggregation of non-anomalous points. Due the nature of this task, it is very important to incorporate all the information provided by all the points within each s_i into the machine learning model. Figure 1 shows how simple features, as the mean statistic per each s_i can not work on this scenario.

Previous works on group anomaly detection do not include information of the probability distributions of each s_i into the classifier, instead some features are extracted from each group s_i to further apply an anomaly detector on the induced feature space [3, 15]. Another works ignore the fact that group anomalies can be distribution-based [5]. State-of-the-art techniques for the group anomaly detection task are given by the hierarchical probabilistic model [30, 31] and the one-class support measure machine [19].

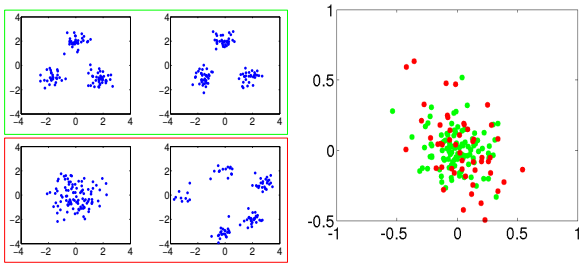


Figure 1: a) Red box: Two group anomalies. Green box: Two non-anomalous groups. b) From several anomalous and non-anomalous groups similar to the ones from a) we compute the mean statistic per group. Red points are the means of anomalous groups. Blue points are the means of non-anomalous groups. There is an overlapping between the statistical means of anomalous and non-anomalous groups which turns hard the detection of this type of anomalies.

2.3 Problem definition

The group anomaly detection can be posed as the task of detecting anomalies from a set of probability distributions. To that end, let $\{\mathbb{P}_1, \mathbb{P}_2, \ldots, i, \ldots, \mathbb{P}_N\} \subset \mathcal{P}$ denote a sample of unknown probability distributions, such that each $s_i \in \mathcal{T}$ is completely characterized by $\mathbb{P}_i$, i.e., $s_i \in \mathcal{T}$ contains the outcomes of a random variable $X_i \sim \mathbb{P}_i$. We define our problem as follows:

Problem Definition 2.2. *Design a one-class classifier for a set of probability distributions with the following properties: a) it does not have to assume any form for each $\mathbb{P}_i$, b) it can encode prior knowledge about $\mathbb{P}_i$ by mean of a kernel function, c) the description obtained by this classifier is robust in the sense that it can be used to detect anomalous probability distributions.*

We will see in the next section that the SMDD classifier satisfy those requirements. The SMDD does not assume any form for $\mathbb{P}_i$, but it can incorporate prior knowledge about probability distributions by means of a kernel on probability measures. For the SMDD, detecting group anomalies equals to finding out anomalous distributions by estimating the kernel embeddings outside of the minimum enclosing ball in the RKHS.

3 THE SMDD CLASSIFIER

The quantile function and Minimum-Volume (MV) set are primal concepts used to define one-class classifiers on $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^D$ [21–23]. We use them to derive the SMDD classifier for the i.d.d sample $\{\mathbb{P}_1, \mathbb{P}_2, \ldots, \mathbb{P}_N\} \subset \mathcal{P}$. Let $(\mathcal{P}, \mathcal{A}, \mathcal{E})$ be a probability space where $\mathcal{P}$ is the space of all probability measures $\mathbb{P}$ on $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$,

the set $\mathcal{A}$ is some suitable σ-algebra of $\mathcal{P}$ and the function $\mathcal{E}$ is a probability measure on $(\mathcal{P}, \mathcal{A})$. A MV-set is the set in $\mathcal{A}$ satisfying:

$$G_\gamma^* = \underset{G\in\mathcal{A}}{\operatorname{argmin}}\{\rho(G) \mid \mathcal{E}(G) \geq \gamma, G \in \mathcal{A}\},\ \gamma \in [0,1], \tag{3}$$

where ρ is real-valued function on $\mathcal{A}$. The SMDD classifier assumes that the class $\mathcal{A}$ is formed by sets of the form :

$$G(R,c) = \{\mathbb{P} \in \mathcal{P} \mid \|\mu_{\mathbb{P}} - c\|_{\mathcal{H}}^2 \leq R^2\}, \tag{4}$$

where the set $G \in \mathcal{A}$ is explicitly parametrized by the hypersphere parameters: $R \in \mathbb{R}^+$ and $c \in \mathcal{H}$. Function $\mu_{\mathbb{P}} \in \mathcal{H}$ is the mean map of $\mathbb{P}$. The SMDD does not try to find a MV-set in the input space $\mathcal{P}$ instead for a specific finite sample $\{\mathbb{P}_1, \mathbb{P}_2, \ldots, \mathbb{P}_N\} \subset \mathcal{P}$ of size N a MV-set $G_\gamma^*(R^*, c^*)$ associated to that sample is estimated by optimizing over $R \in \mathbb{R}^+$ and $c \in \mathcal{H}$. Therefore, the function $\rho(G)$ can be regarded as measure of an enclosing ball in terms of R and c. The SMDD estimates the minimum enclosing ball (R, c) in $\mathcal{H}$ using the set of mean maps $\{\mu_{\mathbb{P}_1}, \mu_{\mathbb{P}_2}, \ldots, \mu_{\mathbb{P}_N}\} \subset \mathcal{H}$ from the sample $\{\mathbb{P}_1, \mathbb{P}_2, \ldots, \mathbb{P}_N\} \subset \mathcal{P}$ as follows:

Problem 1.

$$\begin{aligned} \underset{c\in\mathcal{H}, R\in\mathbb{R}^+, \xi\in\mathbb{R}^N}{min} \quad & R^2 + \lambda \sum_{i=1}^N \xi_i \\ \textit{subject to} \quad & \|\mu_{\mathbb{P}_i} - c\|_{\mathcal{H}}^2 \leq R^2 + \xi_i, i = 1, \ldots, N \\ & \xi_i \geq 0,\ i = 1, \ldots, N, \end{aligned}$$

where ξ contains the slack variables ξ_i and λ is a non-negative regularization parameter.

Proposition 3.1 (Dual form). *The dual form of the previous problem is:*

Problem 2.

$$\begin{aligned} \underset{\alpha\in\mathbb{R}^N}{max} \quad & \sum_{i=1}^N \alpha_i \tilde{k}(\mathbb{P}_i, \mathbb{P}_i) - \sum_{i,j=1}^N \alpha_i \alpha_j \tilde{k}(\mathbb{P}_i, \mathbb{P}_j) \\ \textit{subject to} \quad & 0 \leq \alpha_i \leq \lambda,\ i = 1, \ldots, N \\ & \sum_{i=1}^N \alpha_i = 1 \end{aligned}$$

where $\boldsymbol{\alpha}$ is a Lagrange multiplier vector with non negative components α_i and $\tilde{k}$ is the kernel defined by (1).

3.1 Representer theorem

The representer theorem of kernel methods, specifically of support measure machines (Theorem 1, [18]), states that the solution of those classifiers is characterized by a linear combination of mean maps. From the KKT conditions of Problem 1, we have that the center c can be rewritten as:

$$c = \sum_i \alpha_i \mu_{\mathbb{P}_i}, \quad i \in \{i \in \mathcal{I} \mid 0 < \alpha_i \leq \lambda\},$$

where $\mathcal{I} = \{1, 2, \ldots, N\}$. Moreover, if we define the index sets: $\mathcal{I}_0 = \{i \in \mathcal{I} \mid \alpha_i = 0\}$, $\mathcal{I}_< = \{i \in \mathcal{I} \mid 0 < \alpha_i < \lambda\}$ and $\mathcal{I}_\alpha = \{i \in \mathcal{I} \mid \alpha_i = \lambda\}$ then the sets $\{\mathbb{P}_i \mid i \in \mathcal{I}_0\}$ and $\{\mathbb{P}_i \mid i \in \mathcal{I}_<\}$ are within the description estimated for the SMDD, because their mean maps are inside the hypersphere (R, c). The set $\{\mathbb{P}_i \mid i \in \mathcal{I}_<\}$ is the *support measure set* because their correspondent mean maps are *support vectors* in RKHS. Finally, the set $\{\mathbb{P}_i \mid i \in \mathcal{I}_\alpha\}$ induces mean maps outside the hypersphere because they are affected for the regularization parameter λ. The radius R can be estimated using following result:

Algorithm 1: Training a SMDD.

Input: Training set $\mathcal{T} = \{s_i \mid s_i \subset \mathbb{R}^D,\ 1 \leq i \leq N\}$.
Input: A positive definite kernel k.
Output: The radius and the norm of the center of a minimum enclosing hypersphere: $(R, \|c\|_{\mathcal{H}}^2)$
for *each* $s_i, s_j \in \mathcal{T},\ 1 \leq i, j \leq N$ **do**
  $L_i = |s_i|,\ L_j = |s_j|$;
  $K_{i,j} = \tilde{k}(\mathbb{P}_i, \mathbb{P}_j) \approx \frac{1}{L_i L_j} \sum_{l=1}^{L_i} \sum_{l'=1}^{L_j} k(\mathbf{x}_l^{(i)}, \mathbf{x}_{l'}^{(j)}) \quad \mathbf{x}_l^{(i)} \in s_i, \mathbf{x}_{l'}^{(j)} \in s_j$
end
$\boldsymbol{\alpha}$=solveSMDD(K) (Problem 2);
$\|c\|_{\mathcal{H}}^2 = \boldsymbol{\alpha}^\top K \boldsymbol{\alpha}$;
$R = -\eta + \|c\|_{\mathcal{H}}^2$ (Proposition 3.2);

Proposition 3.2. *Let $\mathcal{L}$ be Lagrangian of Problem 2. If η is the Lagrange multiplier of the constraint $\sum_{i=1}^N \alpha_i = 1$, then $R^2 = -\eta + \|c\|_{\mathcal{H}}^2$.*

From those results the decision function is a function from $\mathcal{H}$ to $\{-1, 1\}$ given by $f(\mu_{\mathbb{P}_t}; R^*, c^*) = \text{sign}(R^2 - \|\mu_{\mathbb{P}_t} - c\|_{\mathcal{H}}^2)$, where (R^*, c^*) are the minimizers of Problem 1. Thanks to the representer theorem of kernel methods the decision function can be written as a function from $\mathcal{P}$ to $\{-1, 1\}$ by the following expression

$$f(\mathbb{P}_t; R^*, c^*) = \text{sign}\left(R^2 - \tilde{k}(\mathbb{P}_t, \mathbb{P}_t) + 2\sum_i \alpha_i \tilde{k}(\mathbb{P}_i \mathbb{P}_t) - \|c\|_{\mathcal{H}}^2\right),$$

where we used $\tilde{k}(\mathbb{P}_i, \mathbb{P}_j) = \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j} \rangle_{\mathcal{H}}$ and $c = \sum_i \alpha_i \mu_{\mathbb{P}_i}$.

Algorithm 1 shows a training procedure for a SMDD classifier on datasets given by Equation 2. The kernel matrix K is computed using an empirical estimator for $\tilde{k}$. The output of the procedure is the radius R and the value $\|c\|^2$ that are used to define the decision function.

3.2 SMDD with stationary kernels

Feature maps $k_I(\mathbf{x}, .)$ under positive definite stationary kernels: $k_I(\mathbf{x}, \mathbf{x}') = f(\mathbf{x} - \mathbf{x}')$ have constant norm [9], i..e., $\|k_I(\mathbf{x}, .)\|_{\mathcal{H}} = \sqrt{\epsilon}$, where ϵ is a constant. However, mean maps under stationary kernels do not have constant norm because: $\|\mu_{\mathbb{P}}\|_{\mathcal{H}} = \|\mathbb{E}_{\mathbb{P}}[k_I(X.,)]\|_{\mathcal{H}} \leq \mathbb{E}_{\mathbb{P}}[\|k_I(X.,)\|_{\mathcal{H}}] = \sqrt{|\epsilon|}$. One way to force those mean maps to have constant norm is by using the following normalization procedure:

$$\tilde{\tilde{k}}(\mathbb{P}_i, \mathbb{P}_j) = \frac{\tilde{k}(\mathbb{P}_i, \mathbb{P}_j)}{\sqrt{\tilde{k}(\mathbb{P}_i, \mathbb{P}_i)\tilde{k}(\mathbb{P}_j, \mathbb{P}_j)}} = \frac{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}}{\sqrt{\langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}}}, \tag{5}$$

which preserves the positivity of the kernel and the injectivity of the embeeding [19]. From the MV-set perspective, the class $\mathcal{A}$ induced by the kernel $\tilde{\tilde{k}}(\mathbb{P}_i, \mathbb{P}_j)$ contains sets of the form:

$$G(R,c) = \{\mathbb{P} \in \mathcal{P} \mid \|\mu_{\mathbb{P}} - c\|_{\mathcal{H}}^2 \leq R^2,\ \|\mu_{\mathbb{P}}\|_{\mathcal{H}}^2 = 1\}. \tag{6}$$

Thus, the normalized kernel $\tilde{\tilde{k}}(\mathbb{P}_i, \mathbb{P}_j)$ implies a constant value $\sum_{i=1}^{N} \alpha_i \tilde{\tilde{k}}(\mathbb{P}_i, \mathbb{P}_i)$ in the objective function of Problem 2. Consequently, Problem 2 can be written as

Problem 3.

$$\begin{aligned} \max_{\boldsymbol{\alpha} \in \mathbb{R}^N} \quad & -\sum_{i,j=1}^{N} \alpha_i \alpha_j \tilde{\tilde{k}}(\mathbb{P}_i, \mathbb{P}_j) \\ \textit{subject to} \quad & 0 \leq \alpha_i \leq \lambda, \ i = 1, \ldots, N \\ & \sum_{i=1}^{N} \alpha_i = 1. \end{aligned}$$

4 SMDD WITH CHANCE CONSTRAINTS

In this section we show how it is possible to create another versions of SMDD classifiers by imposing restrictions to the sets of probability measures G in the class of sets $\mathcal{A}$. For example if we consider event sets of the form:[2] $A(X, R, c) = \{\omega \mid \|k(X(\omega), .) - c\|^2_{\mathcal{H}} \leq R^2\}$ where $X \sim \mathbb{P}$ and $\omega \in \mathbb{R}^D$ and, moreover, if we bound the probability measure of A by an arbitrary value $\rho \in [0, 1]$, that is: $\mathbb{P}(A(X, R, c)) \geq \rho$, we end up with a SMDD classifier that assumes that the class $\mathcal{A}$ is formed by sets of the form:

$$\begin{aligned} G(R, c) &= \{\mathbb{P} \in \mathcal{P} \mid \mathbb{P}(A(X, R, c)) \geq \rho\} \\ &\equiv \{\mathbb{P} \in \mathcal{P} \mid \mathbb{P}(\|k(X, .) - c\|^2_{\mathcal{H}} \leq R^2) \geq 1 - \kappa\}, \ \kappa = 1 - \rho \end{aligned} \tag{7}$$

If we keep the value of ρ close to one and optimize over (R, c) such $\mathbb{P}(A(X, R, c)) \geq \rho$ is satisfied then we restrict most of the realizations of $k(X, .)$ to be within the hypersphere (R, c). Thus, given a set $\{\kappa_1, \ldots, \kappa_N\} \subset [0, 1]^N$ (bounding values) and a sample $\{\mathbb{P}_1, \mathbb{P}_2, \ldots, \mathbb{P}_N\} \subset \mathcal{P}$, a SMDD model is defined by the following chance constrained optimization problem:

Problem 4.

$$\begin{aligned} \min_{c \in \mathcal{H}, R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N} \quad & R^2 + \lambda \sum_{i=1}^{N} \xi_i \\ \textit{subject to} \quad & \mathbb{P}_i(\|k(X_i, .) - c(.)\|^2_{\mathcal{H}} \leq R^2 + \xi_i) \geq 1 - \kappa_i, \\ & \xi_i \geq 0, . \end{aligned}$$

for $i = 1, \ldots, N$.

Instead of taking into account every possible outcome of $X \sim \mathbb{P}_i$ we use the Markov's inequality to bound the probabilistic constraints. Markov's inequality states that $\mathbb{P}(X \geq t)$ is bounded by $\mathbb{E}_{\mathbb{P}}[X]/t$, only if $X \sim \mathbb{P}$ is a nonnegative random variable and $t > 0$. Using Markov's inequality, and noticing that each chance constraint can be rewritten as $\mathbb{P}_i(\|k(X_i, .) - c(.)\|^2_{\mathcal{H}} \geq R^2 + \xi_i) \leq \kappa_i$, the following expression

$$\mathbb{P}_i(\|k(X_i, .) - c(.)\|^2_{\mathcal{H}} \geq R^2 + \xi_i) \leq \frac{\mathbb{E}_{\mathbb{P}_i}[\|k(X_i, .) - c(.)\|^2_{\mathcal{H}}]}{R^2 + \xi_i}, \tag{8}$$

holds, for all $i = 1, 2, \ldots, N$. Moreover, we impose κ_i as being an upper bound for the i constraint:

$$\mathbb{E}_{\mathbb{P}}[\|k(X_i, .) - c(.)\|^2_{\mathcal{H}}]/(R^2 + \xi_i) \leq \kappa_i \tag{9}$$

We have the following result.

Proposition 4.1.

$$\mathbb{E}_{\mathbb{P}}[\|k(X, .) - c(.)\|^2_{\mathcal{H}}] = tr(\Sigma^{\mathcal{H}}) + \|\mu_{\mathbb{P}} - c(.)\|^2_{\mathcal{H}},$$

where[3]

$$tr(\Sigma^{\mathcal{H}}) = \mathbb{E}_{\mathbb{P}}[k(X, X)] - \tilde{k}(\mathbb{P}, \mathbb{P}). \tag{10}$$

Thus, by replacing the result given in Proposition 4.1 into (9) and then into the constrains of Problem 4 we get a SMDD with non probabilistic constraints expressed as the following optimization problem

Problem 5.

$$\begin{aligned} \min_{c \in \mathcal{H}, R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N} \quad & R^2 + \lambda \sum_{i=1}^{N} \xi_i \\ \textit{subject to} \quad & \|\mu_{\mathbb{P}_i} - c(.)\|^2_{\mathcal{H}} \leq (R^2 + \xi_i)\kappa_i - tr(\Sigma^{\mathcal{H}}_i), \\ & \xi_i \geq 0, \quad i = 1, \ldots, N \end{aligned}$$

where $tr(\Sigma^{\mathcal{H}}_i)$ is given by (10).

The dual form of this problem is presented in the next proposition.

Proposition 4.2 (Dual form). *The dual form of Prob. 5 is given by the following fractional programming problem*[4]*:*

Problem 6.

$$\begin{aligned} \max_{\boldsymbol{\alpha} \in \mathbb{R}^N} \quad & \sum_{i=1}^{N} \alpha_i \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_i} \rangle_{\mathcal{H}} - \frac{\sum_{i,j=1}^{N} \alpha_i \alpha_j \langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j} \rangle_{\mathcal{H}}}{\sum_{i=1}^{N} \alpha_i} + \sum_{i=1}^{N} \alpha_i tr(\Sigma^{\mathcal{H}}_i) \\ \textit{subject to} \quad & 0 \leq \alpha_i \kappa_i \leq \lambda, \quad i = 1, \ldots, N \\ & \sum_{i=1}^{N} \alpha_i \kappa_i = 1, \end{aligned}$$

where $\langle \mu_{\mathbb{P}_i}, \mu_{\mathbb{P}_j} \rangle_{\mathcal{H}}$ is computed by $\tilde{k}(\mathbb{P}_i, \mathbb{P}_j)$, $\boldsymbol{\alpha}$ is a Lagrange multiplier vector with non negative components α_i and $tr(\Sigma^{\mathcal{H}}_i)$ is given by (10).

4.1 Representer theorem

The representer theorem is given by the following proposition:

Proposition 4.3 (Representer theorem). *Let be the index sets: $\mathcal{I} = \{1, 2, \ldots, N\}$, $\mathcal{I}_0 = \{i \in \mathcal{I} \mid \alpha_i = 0\}$, $\mathcal{I}_< = \{i \in \mathcal{I} \mid 0 < \alpha_i \kappa_i < \lambda\}$ and $\mathcal{I}_\alpha = \{i \in \mathcal{I} \mid \alpha_i \kappa_i = \lambda\}$. Then,*

$$c(.) = \frac{\sum_i \alpha_i \mu_{\mathbb{P}_i}}{\sum_i \alpha_i}, \ i \in \mathcal{I}_< \cup \mathcal{I}_\alpha, \tag{11}$$

Furthermore, the sets $\{\mathbb{P}_i, \mid i \in \mathcal{I}_0\}$ and $\{\mathbb{P}_i, \mid i \in \mathcal{I}_<\}$ are within the description (the MV-set) estimated by the SMDD. The set $\{\mathbb{P}_i, \mid i \in \mathcal{I}_\alpha\}$ induce mean maps outside of the hypersphere in the RKHS. This set depend on the regularization parameter λ.

The radius R can be estimated with the following result

[2] We consider a random vector defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as a Borel measurable map: $X : \Omega \to \mathbb{R}^D$ where $\Omega = \mathbb{R}^D$ and $\mathcal{F} = \mathcal{B}(\mathbb{R}^D)$. If X satisfy $X(\omega) = \omega$, $\forall \omega \in \Omega$, i.e, X is an identity map, then for $B \in \mathcal{B}(\mathbb{R}^D)$ the probability measure (probability distribution) induced by X on $\mathbb{R}^D$ given by $\mathbb{P}_X(B) = \mathbb{P}\{\omega : X(\omega) \in B\}$ equals to the probability measure $\mathbb{P}(B)$, i.e., $\mathbb{P}_X = \mathbb{P}$.

[3] Formally, $\Sigma^{\mathcal{H}}$ is the covariance operator on $\mathcal{H}$ (Definition ??).

[4] A reference for this kind of optimization problem is [6].

Proposition 4.4. *Let η be the Lagrange multiplier of the constraint $\sum_{i=1}^{N} \alpha_i \kappa_i = 1$ of the Lagrangian of Problem 6, then $R^2 = -\eta$.*

The decision function is the map from $\mathcal{H}$ to $\{-1, 1\}$ given by $f(\mu_{\mathbb{P}_t}; R, c) = \text{sign}\left(R^2 - \|\mu_{\mathbb{P}_t} - c\|^2_{\mathcal{H}} - tr(\Sigma_t^{\mathcal{H}})\right)$ which can be rewritten as the following map from $\mathcal{P}$ to $\{-1, 1\}$:

$$\text{sign}\left(R^2 - \tilde{k}(\mathbb{P}_t, \mathbb{P}_t) + 2\sum_i \alpha_i \tilde{k}(\mathbb{P}_i, \mathbb{P}_t) - \sum_{i,j} \alpha_i \alpha_j \tilde{k}(\mathbb{P}_i, \mathbb{P}_j) - tr(\Sigma_t^{\mathcal{H}})\right) \tag{12}$$

5 RELATED WORK

The closest machine learning model to the SMDD is the One-class support measure machine (OCSMM) [19, 22]. Making a parallel, OCSMM considers that the class $\mathcal{A}$ consist of sets of the form: $G(f, b) = \{\mathbb{P} \in \mathcal{P} \mid \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} \geq b\}$. SMDD models and OCSMM are not equivalent, but they are in the following scenario.

Proposition 5.1. *SMDD given by Problem 2 and Problem 3 and OCSMM are equivalent if the kernel used for all of them is the one given by* (5)*. Moreover, the deterministic version of the SMDD with chance constraints given by Problem 6 is equivalent to OCSMM with kernel* (5) *if it uses joint constraints: $\kappa_1 = \kappa_2 = \cdots = \kappa_N$, the same covariance operator $\Sigma_1^{\mathcal{H}} = \Sigma_2^{\mathcal{H}} = \cdots = \Sigma_N^{\mathcal{H}}$ and a kernel given by* (5).

A connection between kernel density estimation and OCSMM was described in [19, 22] for the case of training sets of gaussian distributions.

6 EXPERIMENTS

This section perform an experimental study of performance of the SMDD models on the task of group anomaly detection. To that end, we used an artificial dataset with four types of group anomalies. Moreover, we used astronomical data from the *Sloan Digital Sky Survey* with the aim to find out anomalous galaxy clusters (Section 6.3).

6.1 Experimental setting

We used the AUC measure as performance metric estimated by a nested cross validation procedure. We experimented with several SMDD models, the Support vector data description method and the OCSMM. The kernels on probability measures were estimated using empirical estimators. All the experiments described in this section can be reproduced using the code in https://github.com/jorjasso/SMDD-group-anomaly-detection. Next, we describe in detail such procedures in the sections below.

6.1.1 Kernel and covariance estimation. We used a Gaussian kernel $k(x, y) = \exp -\gamma \|x - y\|^2, \gamma > 0$ as base kernel for the kernel on probability measures $\tilde{k}$ given by Equation (1). Using a dataset as the one given by (2) we approximated $\tilde{k}$ by the empirical estimator:

$$\tilde{k}(\mathbb{P}_i, \mathbb{P}_j) \approx \frac{1}{L_i L_j} \sum_{l=1}^{L_i} \sum_{l'=1}^{L_j} k(x_l^{(i)}, y_{l'}^{(j)}), \tag{13}$$

where $x^{(i)}_{1 \leq l \leq L_i}$ and $y^{(j)}_{1 \leq l' \leq L_j}$ are elements in the sets s_i and s_j, respectively, and L_i and L_j are the cardinality of that sets. We also approximated the trace of the covariance operator in a RKHS given by Equation (10) by the empirical estimator:

$$tr(\Sigma_i^{\mathcal{H}}) \approx \frac{1}{L_i - 1} \sum_{l=1}^{L_i} k(x_l^{(i)}, x_l^{(i)}) - \frac{1}{L_i(L_i - 1)} \sum_{l=1}^{L_i} \sum_{l=1}^{L_i} k(x_l^{(i)}, x_l^{(i)}). \tag{14}$$

6.1.2 Classifiers. Table 1 shows the models used in the experiments. We remark that the SVDD classifier under this experimental setting, i.e., using a Gaussian kernel, is equivalent to the One-class support vector machine [22]. We trained the SVDD model using the empirical mean per group. The OCSMM is the one-class support vector machine with kernel $\tilde{k}$ [19]. The SMDD is the classifier given by Problem 2 with kernel $\tilde{k}$. The SMDD.N is the classifier given by Problem 2 but with the normalized $\tilde{\tilde{k}}$, we point out that by Proposition 5.1 the SMDD.N is equivalent to a OCSMM classifier with the normalized $\tilde{\tilde{k}}$. The SMDD.C.k.1 and SMDD.C.k.1.N classifiers are the SMDD's from Problem 6 with kernels $\tilde{k}$ and $\tilde{\tilde{k}}$, respectively. For similar experiments with generative models vs the OCSMM classifier in a the same task we refer [19].

Model	Problem/Ref.	kernel
SVDD	[29]	Gaussian
OCSMM	[19]	Eq (1)
SMDD	2	Eq (1)
SMDD.C.k.1	6	Eq (1)
SMDD.N	3	Eq (5)
SMDD.C.k.1.N	6	Eq (5)

Table 1: One-class classifiers used in the experiments

6.1.3 Nested cross validation experiments. The main difficulty in the estimation of a description of datasets by one-class classifiers is that it is hard to perform model selection on unlabeled data and moreover it is unusual to have a set of anomalies beforehand. To overcome that, we artificially introduced anomalies into the original unlabeled data and we added a label to identify whether or not an observation is a group anomaly. That resulted in a dataset with two classes: *anomalous* and *non-anomalous*, although this setting resembles to a binary classification task we emphasize that one-class classifiers as the ones presented in this paper do not take account the labels of observations in the estimation of the description of the data. With that in mind, we performed the experiments using a nested cross-validation procedure (it is know that the results using this type of cross validation are less biased [2]) which uses an internal loop to perform model selection and an outer loop to access the model performance. Notice that each time we trained the classifier either within the inner loop for model selection or the outer loop we only used the observations from the non-anomalous class, however we validated it (in the inner loop) or test it (in

the outer loop) using both labels. We used the Area under the ROC curve (AUC) as a target metric either to access the model performance in the outer loop or for optimize the hyper-parameters in the model selection within the inner loop. The model selection (internal loop) was done by estimating the hyper-parameters with largest cross validation AUC value over a grid of hyper-parameters. The hyper-parameters were given by the regularization parameter λ and the kernel parameter γ.

6.2 Group Anomaly Detection on a Gaussian Mixture Distribution dataset

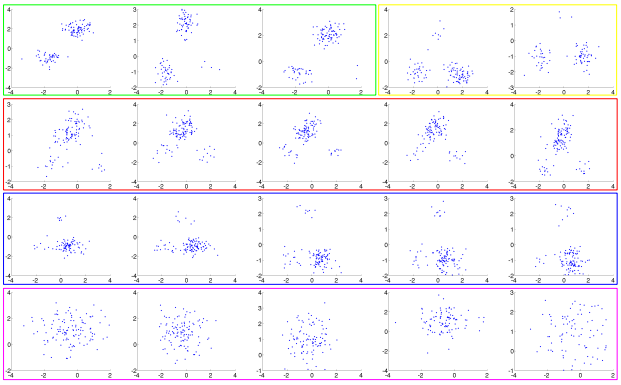


Figure 2: Group anomaly detection dataset. Green and yellow boxes contain non-anomalous groups of points. Red, blue, and magenta boxes contain anomalous groups of points.

We generated a dataset of 300 non-anomalous groups. Each non-anomalous group is a sample drawn from either of two different Gaussian mixture distributions, the probability of chosen either one distribution or the another was 0.5 The parameters for the first Gaussian mixture distribution were mixture weights: $(0.33, 0.64, 0.03)$; means: $(-1.7, -1)$, $(1.7, -1)$, $(0, 2)$; and $0.2 * I_2$ as the sharing covariance matrix, where I_2 denotes the 2×2 identity matrix. The second Gaussian mixture distribution has the same parameters but the mixture weights were given by $(0.33, 0.03, 0.64)$. The number of points per group was given by a value drawn from a Poisson distribution with parameter $\beta = 100$. The green box in Figure 2 shows three non-anomalous groups for the first Gaussian mixture distribution and the yellow box shows two non-anomalous groups for the another one.

We experimented with four types of group anomalies, we describe them as follows

6.2.1 First type of group anomalies. We generated 30 anomalous groups. Each group is a sample drawn from a normal distribution with parameters: mean $(-0.4, 1)$ and covariance matrix given by an 2×2 identity matrix. The number of points per anomalous group is a value taken from a Poisson distribution with parameter $\beta = 100$. The magenta box in Figure 2 shows five of those groups.

6.2.2 Second type of group anomalies. We generated 30 anomalous groups. Each group is a sample drawn from a Gaussian mixture distribution with parameters weights: $(0.1, 0.08, 0.07, 0.75)$; means: $(-1.7, -1)$, $(1.7, -1)$, $(0, 2)$, $(0.6, -1)$; and a sharing covariance matrix given by $0.2 * I_2$. The number of points per anomalous group was the same as the First type of group anomalies. Blue box in Figure 2 shows five of those groups.

6.2.3 Third type of group anomalies. We generated 30 anomalous groups. Each group is a sample drawn from a Gaussian mixture distribution with parameters weights: $(0.14, 0.1, 0.28, 0.48)$; means: $(-1.7, -1)$, $(1.7, -1)$, $(0, 2)$, $(-0.5, 1)$; and $0.2 * I_2$ as the sharing covariance matrix. The number of points per anomalous group was the same as the First type of group anomalies. Red box in Figure 2 shows five of those groups.

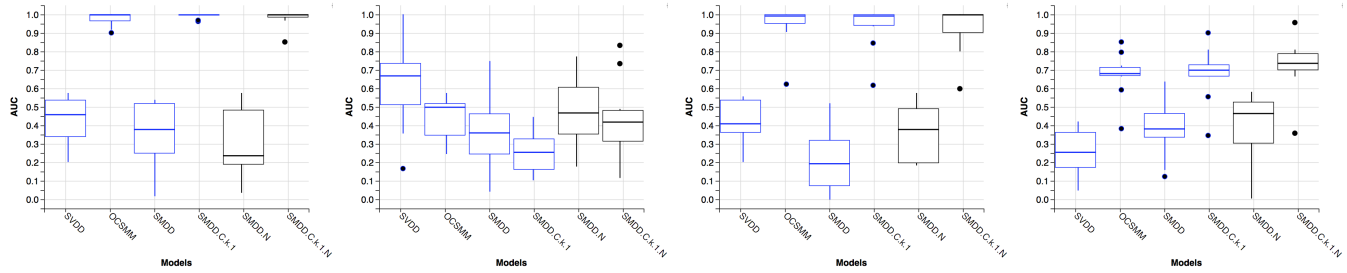
6.2.4 Fourth type of group anomalies. We generated 30 anomalous groups by combining 10 groups of the first type of group anomalies, 10 groups from the second type and 10 groups from the third type. The number of points per anomalous group was the same as the First type of group anomalies.

6.2.5 Results. Figure 3 shows the results. Each box plot contain information of the AUC measures from the outer loop of the nested cross validation procedure. For the first type of anomalies we observed that the OCSMM, SMDD.C.k.1 and SMDD.C.k.1.N outperform the other classifiers. The SMDD.N classifier (which is equivalent to a OCSMM with normalized kernel), the SMDD and the SVDD, have a poor performance in terms of the AUC measure. For the second type of anomalies, all the support measure machines perform badly, and a simple SVDD outperforms all the other classifiers. It seems that for this particular setting a simple statistic per group, such as the mean, it is enough to describe the non-anomalous groups. For the third and fourth type of anomalies we observed that the OCSMM, SMDD.C.k.1 and SMDD.C.k.1.N can detect more anomalies than the other models.

6.3 Group Anomaly Detection on Astronomical Data

The aim of this experiment was to test the SMDD classifiers using real data. To this end, we used data from *The Sloan Digital Sky Survey*[5] *(SDSS)* project. This data contains massive spectroscopic surveys of the Milky Way galaxy and extra solar planetary systems. We used the same setting described in [19, 20, 31] to construct the non-anomalous groups. That is, initially, the dataset contains information of 7530 galaxies, each galaxy is represented by 4000 values of spectral information. Then, this dataset is processed by down-sampling each observation to get only 500 values of spectral information. Finally, it is formed 505 groups of galaxies using a clustering procedure. Thus, we used a dataset of 505 non-anomalous groups of galaxies, where each group of galaxies contain between $10 - 15$ galaxies. In order to reduce even more the dimensionality of the data, we applied a Principal Component Analysis (PCA) procedure. We noticed that the first four PCA components preserved 85% of the variance. Then the dataset of non-anomalous groups is formed by 505 groups, each of them containing the four dimensional PCA vectors. Figure 4 summarize this procedure.

[5] http://www.sdss3.org/



(a) Results first type of anomalies (b) Results second type of anomalies (c) Results third type of anomalies (d) Results fourth type of anomalies

Figure 3: Experimental results for a group anomaly detection task over a Gaussian mixture distribution dataset.

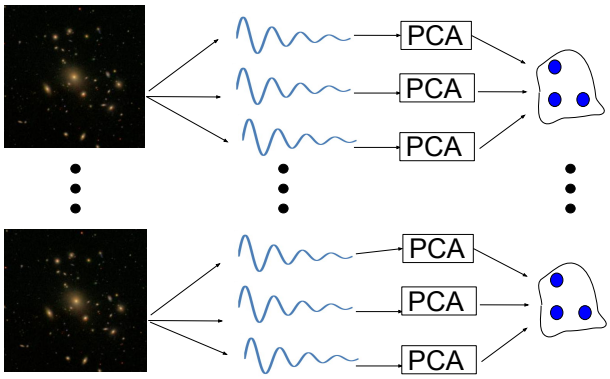


Figure 4: Feature extraction pipeline for the group anomaly experiment on astronomical data .

In order to verify the performance of the SMDD classifiers we injected three types of group anomalies. We describe them as follows.

6.3.1 First type of group anomalies. We injected 50 groups anomalies. Each group anomaly is formed by randomly selecting n galaxies from all the dataset of galaxies, where n is a value distributed according to a Poisson distribution with parameter $\beta = 15$.

6.3.2 Second type of group anomalies. We injected 50 groups anomalies where each group anomaly is a sample from a Gaussian mixture distribution with the following parameters: weights = $\{1/3, 1/3, 1/3\}$, means =$\{mean(A_1), mean(A_2), mean(A_3)\}$, where A_1, A_2, A_3 are three random subsets of galaxies from the original set of galaxies, and covariance matrix given by the $0.5 * \Sigma$, where Σ is the mean of the empirical covariance matrices of non-anomalous groups.

6.3.3 Third type of group anomalies. We use the same setting as the second type of group anomalies but we set the covariance matrix of the Gaussian mixture distribution to be $1.0 * \Sigma$

6.3.4 Results. Figure 5 shows the results in terms of the AUC metric for the models in Table 1. We observed that for the first type of anomalies all the SMDD models outperform the other methods and the OCSMM has the worst performance. For the second type of anomalies the OCSMM, SMDD.C.K.1 and SMDD.C.K.1.N have better performance and the SMDD and SMDD.N (or OCSMM with normalized kernel) have the worst performance. For the third type of anomalies all the measure machines performs equivalent and the SVDD has the best performance.

6.4 Discussion

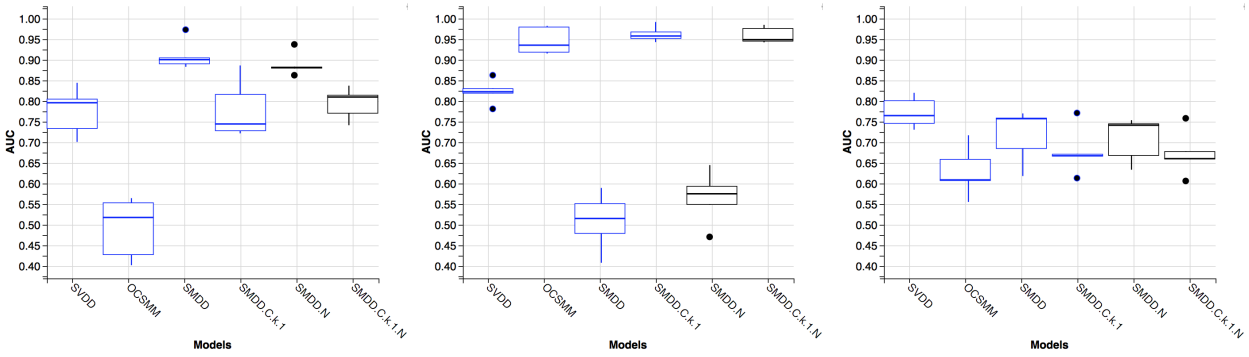
From the last experiments it is possible to observe that either the SMDD.C.K.1 or the SMDD.C.K.1.N are the classifiers with best performance across all the experiments: they have small variance of the AUC metric per experiment, moreover, the variance of the AUC metric across all the experiments is small as well. We observe also that the performance of the other measures machines and the SVDD will depend on the type of group anomalies. For example, a SVDD (or a one class support vector machine) would have a good performance if a statistic per group is a discriminant feature.

7 CONCLUSIONS

In this paper, we propose the SMDD one-class classifier as a tool to estimate the description of a set of probability distributions. The SMDD is a kernel method, specifically, it is a support measure machine whose solution is a function that depend on a subset of the kernel embeddings of probability distributions, and hence a subset of the training set: *the support measures*. The SMDD model does not assume any form for each probability distribution, however it can include prior knowledge of the distribution via a kernel function. We experimentally show the robustness of the estimated description of a set of probability distribution by the SMDD through a set of experiments on the group anomaly detection task.

ACKNOWLEDGMENTS

The authors are thankful with FAPESP grant # 2011/50761-2,FAPESP 2015/01587-0, CNPq, CAPES, NAP eScience - PRP - USP for their financial support.



(a) Results Experiment I - Sloan dataset (b) Results Experiment II - Sloan dataset (c) Results Experiment III- Sloan dataset

Figure 5: The results of the experiment for the group anomaly detection task over a SDSS III dataset.

REFERENCES

[1] Alain Berlinet and Christine Thomas-Agnan. 2004. *Reproducing kernel Hilbert spaces in probability and statistics*. Vol. 3. Kluwer Academic Boston.

[2] Gavin C Cawley and Nicola LC Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* 11, Jul (2010), 2079–2107.

[3] P.K. Chan and M.V. Mahoney. 2005. Modeling multiple time series for anomaly detection. In *Data Mining, Fifth IEEE International Conference on*. 8 pp.–. DOI: http://dx.doi.org/10.1109/ICDM.2005.101

[4] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly Detection: A Survey. *ACM Comput. Surv.* 41, 3, Article 15 (July 2009), 58 pages.

[5] Kaustav Das, Jeff Schneider, and Daniel B. Neill. 2008. Anomaly Pattern Detection in Categorical Datasets. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*. ACM, New York, NY, USA, 169–176. DOI: http://dx.doi.org/10.1145/1401890.1401915

[6] Christodoulos A Floudas and Panos M Pardalos. 2008. *Encyclopedia of optimization*. Vol. 1. Springer Science & Business Media.

[7] Kenji Fukumizu, Francis R Bach, and Michael I Jordan. 2004. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *The Journal of Machine Learning Research* 5 (2004), 73–99.

[8] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. 2008. Kernel Measures of Conditional Dependence. In *Advances in Neural Information Processing Systems 20*, J.C. Platt, D. Koller, Y. Singer, and S. Roweis (Eds.). MIT Press, Cambridge, MA, 489–496.

[9] Marc G. Genton. 2002. Classes of Kernels for Machine Learning: A Statistics Perspective. *J. Mach. Learn. Res.* 2 (March 2002), 299–312.

[10] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research* 13 (2012), 723–773.

[11] Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. 2005. Kernel Methods for Measuring Independence. *J. Mach. Learn. Res.* 6 (Dec. 2005), 2075–2129.

[12] C Guilbart. 1979. Produits scalaires sur l'espace des mesures. In *Annales de l'institut Henri Poincaré (B) Probabilités et Statistiques*, Vol. 15. Gauthier-Villars, 333–354.

[13] Tommi Jaakkola, David Haussler, and others. 1999. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems* (1999), 487–493.

[14] Tony Jebara and Risi Kondor. 2003. Bhattacharyya and expected likelihood kernels. In *Learning Theory and Kernel Machines*. Springer, 57–71.

[15] E. Keogh, J. Lin, and A. Fu. 2005. HOT SAX: efficiently finding the most unusual time series subsequence. In *Data Mining, Fifth IEEE International Conference on*. 8 pp.–. DOI: http://dx.doi.org/10.1109/ICDM.2005.79

[16] Risi Kondor and Tony Jebara. 2003. A kernel between sets of vectors. In *ICML*. 361–368.

[17] Pedro J Moreno, Purdy P Ho, and Nuno Vasconcelos. 2003. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In *Advances in neural information processing systems*. None.

[18] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schlkopf. 2012. Learning from Distributions via Support Measure Machines. In *Advances in Neural Information Processing Systems 25*, P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger (Eds.). 10–18.

[19] Krikamol Muandet and Bernhard Schoelkopf. 2013. One-Class Support Measure Machines for Group Anomaly Detection. In *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*. AUAI Press, Corvallis, Oregon, 449–458.

[20] Barnabás Póczos, Liang Xiong, and Jeff G. Schneider. 2012. Nonparametric Divergence Estimation with Applications to Machine Learning on Distributions. *CoRR* abs/1202.3758 (2012).

[21] Wolfgang Polonik. 1997. Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications* 69, 1 (1997), 1 – 24.

[22] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural computation* 13, 7 (2001), 1443–1471.

[23] Clayton Scott and Robert D. Nowak. 2006. Learning Minimum Volume Sets. *Journal of Machine Learning Research* 7 (2006), 665–704.

[24] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. 2007. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory*. Springer, 13–31.

[25] Le Song, Byron Boots, Sajid M Siddiqi, Geoffrey J Gordon, and Alex J Smola. 2010. Hilbert space embeddings of hidden Markov models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 991–998.

[26] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert Lanckriet, and Bernhard Schlkopf. 2008. Injective hilbert space embeddings of probability measures. In *In COLT*.

[27] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. 2010. Hilbert space embeddings and metrics on probability measures. *JMLR* 99 (2010), 1517–1561.

[28] Charles Suquet and others. 1995. Distances euclidiennes sur les mesures signees et applications a des theoremes de Berry-Esseen. *Bulletin of the Belgian Mathematical Society Simon Stevin* 2, 2 (1995), 161–182.

[29] David MJ Tax and Robert PW Duin. 2004. Support vector data description. *Machine learning* 54, 1 (2004), 45–66.

[30] Liang Xiong, Barnabás Póczos, and Jeff G. Schneider. 2011. Group Anomaly Detection using Flexible Genre Models. In *NIPS*. 1071–1079.

[31] Liang Xiong, Barnabás Póczos, Jeff G. Schneider, Andrew J. Connolly, and Jake VanderPlas. 2011. Hierarchical Probabilistic Models for Group Anomaly Detection. In *AISTATS*. 789–797.