Support vector data descriptions and k-means clustering: one class?

Nico Görnitz[1], Luiz Alberto Lima, Klaus-Robert Müller[1], Marius Kloft, and Shinichi Nakajima

***Abstract*—We present *ClusterSVDD*, a methodology that unifies support vector data descriptions (SVDDs) and k-means clustering into a single formulation. This allows both methods to benefit from one another, i.e. by adding flexibility using multiple spheres for SVDDs and increasing anomaly resistance and flexibility through kernels to k-means. In particular, our approach leads to a new interpretation of k-means as a regularized mode seeking algorithm. The unifying formulation further allows for deriving new algorithms by transferring knowledge from one-class learning settings to clustering settings and vice versa. As a showcase, we derive a clustering method for structured data based on a one-class learning scenario. Additionally, our formulation can be solved via a particularly simple optimization scheme. We evaluate our approach empirically to highlight some of the proposed benefits on artificially generated data, as well as on real world problems, and provide a PYTHON software package comprising various implementations of primal and dual SVDD as well as our proposed *ClusterSVDD*.**

***Index Terms*—Anomaly Detection, One-class Classification, Support Vector Data Description, Clustering, k-means**

I. INTRODUCTION

MACHINE learning methods and tools have become a vital part of research and industry these days, where raw data is cheaply available in huge amounts. Frequently, though, this comes with an absence of ground truth labels and, therefore, attention has been drawn recently to unsupervised machine learning methods.

Two of the most prominent tasks within unsupervised settings comprise *one-class classification*, the identification of common sub-structures for a given set of samples, and *clustering*, the identification of discriminative sub-structures within a given set of samples.

One-class learning, a term that was first mentioned in Moya and Hush [1], [2] in the context of neural networks, is at the core of important applications such as anomaly detection, also known as outlier detection, intrusion detection, and novelty detection (e.g. [3], [4] for an overview). The most influential methods comprise one-class support vector machines [5], [6] and support vector data descriptions [7], [8]; they have been analyzed [9], [10], extended [11], [12], [13], [14], [15], [16], [17], and refined [18], [19], [20], [21] many times since then.

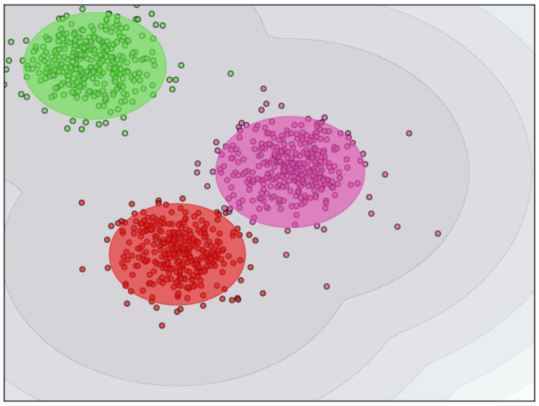

Fig. 1. Fitting multiple hyperspheres simultaneously with a pre-defined outlier fraction is the core idea of our proposed method ClusterSVDD.

Clustering, especially k-means clustering, has been around a while. In 1967, MacQueen published his classic work on multivariate classification [22], laying the foundation of one of the most successful methods for data analysis ever. It has been analyzed, extended, and adapted, and has served as a source of inspiration hundreds of times [23], [24], [25], [26], [27], [28], [23], [29], [24], [16], [30], with a still-active community today. For an extensive overview we refer to Jain et al. [31].

In this work, we fuse k-means clustering and support vector data description into a single framework, unifying one-class classification and clustering. More precisely, our contributions are:

- relating k-means clustering with one-class classification, which leads to new insights on the properties of SVDDs and k-means, i.e. identifying k-means as a regularized mode seeking algorithm,
- natural extension for k-means to kernels and outlier awareness through hypersphere formulation,
- natural extension for SVDDs for mixtures of distributions through multiple spheres.

Along these lines, we show that our solution, *ClusterSVDD*, has an especially simple form that allows for re-using existing code of k-means and SVDDs. Furthermore, we provide an open-source PYTHON software suite with implementations of dual (kernel) SVDD, fast primal (sub-gradient descent) SVDD, and our *ClusterSVDD*. A major focus of our paper is to give a rigorous review of SVDDs and its properties, e.g. deriving

[1] corresponding authors (email: nico.goernitz@tu-berlin.de, klaus-robert.mueller@tu-berlin.de)

Nico Görnitz, Klaus-Robert Müller, and Shinichi Nakajima are with the Berlin Institute of Technology, Machine Learning Group, Marchstr. 23, 10587 Berlin, Germany

Klaus-Robert Müller is also with the Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-713, Korea and with the Max Planck Institute for Informatics, Stuhlsatzenhausweg, 66123 Saarbrücken

Luiz A. Lima is with the Pontifcia Univ. Catlica do Rio de Janeiro, 22543-900 Rio de Janeiro, Brazil and Petrobras, 20031-912 Rio de Janeiro, Brazil

Marius Kloft is with the Humboldt University of Berlin, Department of Computer Science, Machine Learning Group, Rudower Chaussee 25, 12489 Berlin, Germany

new theorems where necessary and summarizing the most important properties from the works of [7], [6], [20], [19]. Finally, to leverage the link between one-class learning and clustering, we derive a clustering method for structured data based on an existing one-class scenario.

Instead of fitting cluster centers to the data, our *ClusterSVDD* minimizes hyperspheres such that the bulk of the data is within (nominal data), with a pre-defined fraction outside (anomalous data) of the spheres. From a SVDD point of view, it is a natural extension from single spheres to multiple spheres (cf. Fig. 1). Both formulations, *k*-means and SVDD, are special cases of our *ClusterSVDD*.

To clarify, the intention of this work is to allow further formal insight into the structure of two seemingly unrelated learning problems. Therefore, our focus goes beyond extensive comparison studies against all thinkable state of the art settings and methods. Rather, we point out (cf. Table V) that the novel holistic view on clustering and one-class problems permits the solution of completely novel problem classes, such as structured clustering (cf. Section V). In other words, our efforts are orthogonal to standard algorithmic improvements; the empirical evaluation serves the mere purpose of illustrating the basic functionality of our approach.

We specify the setting and introduce *k*-means clustering as well as support vector data descriptions (SVDDs) as commonly found in literature in Section II. In order to unify both methods, we first analyze and re-formulate both methods more precisely, and then we introduce our *ClusterSVDD* and corresponding optimization in Section III. Section IV presents empirical evaluations on artificially generated data as well as on real-world data. In Section V, we derive an outlier resistant clustering method as a showcase for leveraging the link between one-class classification and clustering. We conclude with Section VI, summarizing our work and outlining future research directions.

II. Preliminaries

Given a set of input instances $\mathbf{x}_1, \ldots, \mathbf{x}_\ell \in \mathcal{X}$, where $\mathcal{X}$ is an arbitrary set that is commonly assumed to be realized from a sequence of independent and identically distributed (i.i.d) random variables. Furthermore, k denotes the number of clusters and $z_i \in \{1, \ldots, k\}$ the membership of the corresponding input instance $\mathbf{x}_i$. Memberships can be expressed by partition sets $\{S_j\}_{j=1}^k$, where $i \in S_j$ if and only if $z_i = j$. It holds that $S_i \cap S_j = \emptyset$ for $i \neq j$ and $\cup_{j=1}^k S_j = \{1, \ldots, \ell\}$.

Kernel based approaches [32], [23] allow the input instances to be mapped into a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ via a feature map $\phi : \mathcal{X} \to \mathcal{H}$ for a more concise description thereof. The most important symbols are summarized in Table I.

TABLE I
Glossary of mathematical symbols and corresponding explanation

Symbol	Description
$\mathbf{c}$	Center of a (hyper-)sphere
T	Corresponding threshold (squared radius)
$\cdot^*$	Variables denoted with an asterisk are optimal
ν	Hyperparameter or regularization parameter as used in (Cluster)SVDDs and one-class SVMs
ℓ	Number of datapoints
k	Number of clusters (chosen in advance)
z_i	latent variable for instance i ($z_i \in \{1, \ldots, k\}$)
ξ_i	Slack variable for training instance i ($\xi_i \geq 0$)
$\mathsf{k}(\cdot, \cdot)$	Kernel function
$\phi(\cdot)$	Feature map: maps input instances into some (high-dimensional) feature space $\phi : \mathcal{X} \to \mathcal{H}$
$\lVert \cdot \rVert$	Euclidean norm
$\mathcal{H}$	Reproducing kernel Hilbert space (RKHS)
$\mathcal{X}$	Input space
$\lceil \cdot \rceil$	Ceiling function $\lceil x \rceil = \min\{n \in \mathbb{Z} : n \geq x\}$
$\boldsymbol{\alpha}$	Dual variable
$\lvert \cdot \rvert$	Cardinality of a set.
S_i	Partition set i: it holds that $S_i \cap S_j = \emptyset$ for $i \neq j$ and $\cup_{j=1}^k S_j = \{1, \ldots, \ell\}$

A. k-*means Clustering*

k-means clustering [22] (a recent overview is given in [31]) is usually introduced as a (non-convex) optimization problem of finding a partition $\{S_j\}_{j=1}^k$, for a pre-defined k, that minimizes the within cluster sum-of-squares (WCSS),

$$\min_{\{S_j\}_{j=1}^k} \sum_{j=1}^k \sum_{i \in S_j} \lVert \mathbf{x}_i - \mathbf{c}_j \rVert^2 , \tag{1}$$

with $\{\mathbf{c}_j \in \mathcal{X}\}_{j=1}^k$ being the means of the corresponding clusters. Solving this problem (at least locally optimal) consists of three simple steps:

(1) Initialize the cluster centers $\{\mathbf{c}_j\}_{j=1}^k$ and repeat steps (2) & (3) until no changes occur.
(2) Update the partitions $\{S_j\}_{j=1}^k$ by identifying the nearest cluster, given the intermediate cluster centers $\mathbf{c}_\cdot$, $z_i = \operatorname{argmin}_{\hat{z} \in \{1,\ldots,k\}} \lVert \mathbf{c}_{\hat{z}} - \mathbf{x}_i \rVert^2$.
(3) Update the cluster centers $\mathbf{c}_j = 1/|S_j| \sum_{i \in S_j} \mathbf{x}_i$, $\forall j = 1, \ldots, k$.

For our purposes, we need an alternative formulation of *k*-means. Instead of stating that the cluster centers $\{\mathbf{c}_j\}_{j=1}^k$ should be the means of input instances corresponding to the cluster, we can re-write OP (1) more concisely as

$$\min_{\{S_j\}_{j=1}^k} \sum_{j=1}^k \min_{\mathbf{c}_j} \sum_{i \in S_j} \lVert \mathbf{c}_j - \mathbf{x}_i \rVert^2 .$$

This yields the same solution, since it is a convex problem w.r.t. $\mathbf{c}_j$ (fixing the partitions), and we can analytically derive the optimal solution by $\partial \left(\sum_{i \in S_j} \lVert \mathbf{c}_j - \mathbf{x}_i \rVert^2 \right) / \partial \mathbf{c}_j = 0$, therefore $\mathbf{c}_j^* = 1/|S_j| \sum_{i \in S_j} \mathbf{x}_i$. We can now define an equivalent constrained formulation of OP (1).

*Definition 1 (*k*-means Constrained Problem):* The constrained optimization problem for *k*-means is given by

$$\sum_{j=1}^k \min_{\mathbf{c}_j} \sum_{i \in S_j} \lVert \mathbf{c}_j - \mathbf{x}_i \rVert^2 \tag{2}$$

$$\text{subject to } z_i = \operatorname*{argmin}_{\hat{z} \in \{1,\ldots,k\}} \lVert \mathbf{c}_{\hat{z}} - \mathbf{x}_i \rVert^2 ,$$

where $i \in S_j$, if and only if $z_i = j$, $\forall\ i = 1, \ldots, \ell$ and $\forall\, j = 1, \ldots, k$.

Since its introduction, efforts have been made to increase the flexibility of the description [23], [24], [25], e.g. through the use of kernels ([32], [23], [33] for an introduction to kernel methods), and increasing the robustness of the method [26], [27], [28] against outliers and the curse of dimensionality. In this work, we tackle all of the above mentioned into a single framework.

Another line of research, which we do not investigate further in this work, deals with the inference of the correct number of partitions k [29], [34], [35], [36], [37].

B. Support Vector Data Description (SVDD)

One-class SVMs [6], [5] and support vector data descriptions (SVDD)[7] are among the most prominent methods for *one-class classification*.

The aim of one-class classification is to fit a model of *normality*, that is, to find a set containing the most typical instances and reject instances that deviate significantly from this model.

The task can be formally phrased within the framework of density level set estimation as follows, where input instances are drawn i.i.d. according to a probability distribution P. Let $p(x)$ denote the density of P for $x \in \mathcal{X}$ so that $p(x) = P(X = x)$ for some random variable X taking values on $\mathcal{X}$ and $b_\nu \in [0, 1]$ a threshold on the fraction of anomalies.

Denoting by $\mathcal{X}$ another i.i.d. copy according to P, the theoretically optimal nominal set is $L_\nu := \{x \in \mathcal{X} : p(x) \geq b_\nu\}$ for $\nu \in [0, 1]$ and b_ν, such that $P(X \notin L_\nu) = \nu$, which is called the ν *density level set* and can be interpreted as follows: L_ν contains the most likely inputs under the density p, while rare or untypical data are modeled to lie outside of L_ν. The parameter ν indicates the fraction of outliers in the model.

The aim is to compute, based on the data $x_1, \ldots, x_\ell \in \mathcal{X}$, a good approximation of L_ν, that is, to determine a function $f : \mathcal{X} \to \mathbb{R}$ giving rise to an estimated density level set $\hat{L}_\nu := \{x \in \mathcal{X} : f(x) \leq 0\}$. It is desirable that $\hat{L}_\nu$ closely approximates the true density level set L_ν, i.e., $\hat{L}_\nu$ converges to L_ν in probability, that is,

$$P(\hat{L}_\nu \backslash L_\nu \cup L_\nu \backslash \hat{L}_\nu) \to 0 \text{ for } \ell \to \infty.$$

This implies that $\hat{L}_\nu$ has asymptotically probability mass ν, that is, $P(X \notin \hat{L}_\nu) \to \nu$ for $\ell \to \infty$.

Density level set estimation is closely related to minimum volume set estimation [38], excess mass estimation, and density estimation [9]. It has been—implicitly or explicitly—applied to anomaly detection [3], [39], outlier detection, novelty detection, and change detection.

Classic kernel-based [32] approaches include the one-class support vector machine [6], [5] (OC-SVM) and the support vector data description [8], [7] (SVDD).

The one-class SVM learns a hyperplane that separates the bulk of the data from the origin with maximum margin. The primal optimization problem reads:

$$\begin{aligned} &\min_{\mathbf{w}, \rho, \boldsymbol{\xi} \geq 0} \frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} \xi_i \\ &\text{subject to} \quad \langle \mathbf{w}, \phi(\mathbf{x}_i)\rangle \geq \rho - \xi_i\ , \quad \forall\, i = 1, \ldots, \ell \quad , \end{aligned} \tag{3}$$

where $\rho \in \mathbb{R}$ is a threshold, $\mathbf{w} \in \mathcal{H}$ the hyperplane parameter and $0 < \nu \leq 1$ controls the number of support vectors. Moreover, one-class SVMs with RBF kernels have been shown to be consistent density level set estimators [10].

The SVDD employs a quadratic model $f_{\mathbf{w},R}(x) = \|\mathbf{c} - \phi(x)\|^2 - R^2$. It subsequently encloses a fraction of $1 - \nu$ many inputs within a hypersphere, with center $\mathbf{c}$ and radius R.

The Primal SVDD Problem according to [8] is given by,

$$\begin{aligned} &\min_{\mathbf{c}, R, \boldsymbol{\xi} \geq 0} R^2 + C\sum_{i=1}^{\ell} \xi_i \\ &\text{subject to} \quad \|\mathbf{c} - \phi(\mathbf{x}_i)\|^2 \leq R^2 + \xi_i\ , \quad \forall\, i = 1, \ldots, \ell \end{aligned} \tag{4}$$

for $C > 0$. It can be shown that both methods are equal for specific kernels [12], i.e. translation-invariant kernels. In Section III, we present an equivalent result for the primal formulations and utilize it in Section V.

Extensions to semi-supervised and LPUE (learning with positive and unlabeled examples) settings comprise [11], [12], [13]. Using standard SVDD for clustering was first attempted by [16]. Besides the obvious extension to multiple one-class learning problems, attempts to integrate multiple spheres were done in [14], [15], [40], [17], [18], [41]. Our work is in the spirit of these attempts, but goes further by directly relating *k*-means to SVDDs, being more extensive by reviewing details and stating more precisely the problem, adding flexibility through kernels, and also being more *simple*.

III. ClusterSVDD

In order to unify both worlds of *k*-means clustering and support vector data description, we need to re-formulate and analyze more precisely the underlying optimization problems before we introduce our *ClusterSVDD* and corresponding optimization.

A. Revisiting SVDD

As noted in the literature [20], [18], [19], there are some issues with the original formulation of the SVDD as defined in Section II-B. First, the formulation is not convex due to R^2 in the constraints and second, the primal-dual relation breaks down for $0 < C < 1/\ell$. However, this can be fixed, and we derive here a rigorous formulation of the SVDD based on the work of Chang et al. [20].

Definition 2 (Primal Constrained Problem): The primal SVDD optimization problem as a quadratically constrained linear program (QCLP) is given by:

$$\begin{aligned} &\min_{\mathbf{c}, T \geq 0, \boldsymbol{\xi} \geq 0} T + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} \xi_i \\ &\text{subject to} \quad \|\mathbf{c} - \phi(\mathbf{x}_i)\|^2 \leq T + \xi_i \quad \forall\, i = 1, \ldots, \ell \end{aligned} \tag{5}$$

for all $0 < \nu < 1$ the constraint $T \geq 0$ is dispensable (cf. Lemma (2)). We will denote the OP (5) as $\text{Svdd}(\nu, \{\mathbf{x}_i\}_{i=1}^{\ell})$.

Note that ξ_i in OP (5) can be substituted, which allows for an unconstrained formulation of the SVDD.

Definition 3 (Primal Unconstrained Problem): The primal convex, non-smooth, and unconstrained SVDD optimization problem is given by:

$$\min_{\mathbf{c}, T \geq 0} T + \frac{1}{\ell\nu} \sum_{i=1}^{\ell} \max(0, \|\mathbf{c} - \phi(\mathbf{x}_i)\|^2 - T) \ . \qquad (6)$$

This definition is important for solving SVDDs practically, using sub-gradient based solvers (cf. Section III-E).

Deriving a linearly constrained quadratic program (QP) allows us to pin the relation between SVDDs and OC-SVMs, which will become important later in Section V.

Theorem 1 (Quadratic Program Formulation and Equivalence to One-class SVM): The SVDD primal optimization problem, given by OP (5), can be transformed into the following equivalent linearly constrained quadratic program (QP):

$$\min_{\mathbf{w}, \rho, \boldsymbol{\xi} \geq 0} \frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\ell\nu} \sum_{i=1}^{\ell} \xi_i \qquad (7)$$

$$\text{subject to} \quad \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle \geq \rho + \frac{1}{2}\|\phi(\mathbf{x}_i)\|^2 - \xi_i \ , \ \forall\, i = 1, \ldots, \ell \ ,$$

i.e. for L_2-normalized feature vectors $\|\phi(\mathbf{x})\| = const$, the above formulation reduces to the one-class SVM formulation as given in OP (3).

Proof: Starting from the formulation of the primal SVDD in OP (5), we first extend the constraints from $\|\mathbf{c} - \phi(\mathbf{x}_i)\|^2 \leq T + \xi_i$ to $\|\mathbf{c}\|^2 - 2\langle \mathbf{c}, \phi(\mathbf{x}_i) \rangle + \|\phi(\mathbf{x}_i)\|^2 \leq T + \xi_i$. Second, we re-arrange terms and arrive at $\frac{\|\mathbf{c}\|^2 - T}{2} + \frac{\|\phi(\mathbf{x}_i)\|^2}{2} - \frac{\xi_i}{2} \leq \langle \mathbf{c}, \phi(\mathbf{x}_i) \rangle$. In a third step, we substitute $\rho = \frac{\|\mathbf{c}\|^2 - T}{2} \in \mathbb{R}$, $\zeta_i = \frac{\xi_i}{2} \in \mathbb{R}^+$, and $\mathbf{c} = \mathbf{w} \in \mathcal{H}$, which changes the objective function $T + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} \xi_i$ towards $\|\mathbf{w}\|^2 - 2\rho + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} 2\zeta_i$. Without changing the minimizer, we can multiply the objective by $\frac{1}{2}$ and arrive at the one-class SVM objective $\frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} \zeta_i$ with corresponding constraints $\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle \geq \rho + \frac{1}{2}\|\phi(\mathbf{x}_i)\|^2 - \zeta_i$. This proves the first part of the theorem. For the second part, a simple substitution $\hat{\rho} = \rho + \frac{1}{2}\|\phi(\mathbf{x}_i)\|^2 = \rho + \frac{const^2}{2} \in \mathbb{R}$ leads to the desired outcome. ■

The following lemmas are important for understanding the various solutions to the SVDD problem. These lemmas help establishing the link between *k*-means and our proposed method *ClusterSVDD* (cf. Section III-B)

Lemma 1: Assume $\nu \leq 1/\ell$ is given, then OP (5) reduces to the minimum enclosing ball (MEB) problem, i.e. it holds that $\{\xi_i\}_{i=1}^{\ell} = 0$ (hard margin).

Proof: Assume an optimal solution of OP (5) is given by $(T^*, \mathbf{c}^*, \{\xi_i^*\}_{i=1}^{\ell})$. Assume another solution $(T^* + \xi_m^*, \mathbf{c}^*, \{0\}_{i=1}^{\ell})$, where $\xi_m^* = \max_{i \in \{1,\ldots,\ell\}} \xi_i^*$, which is a feasible solution. Therefore,

$$(T^* + \xi_m^*) + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} 0 = T^* + \xi_m^* \leq T^* + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} \xi_i^*$$

$$\Rightarrow \quad \nu\xi_m^* \leq \frac{1}{\ell}\sum_{i=1}^{\ell} \xi_i^* = \frac{1}{\ell}\left(\sum_{i \backslash m}^{\ell} \xi_i^* + \xi_m^*\right) ,$$

is strictly fulfilled for $\nu < 1/\ell$ and hence, any optimal solution must include $\{\xi_i^*\}_{i=1}^{\ell} = \{0\}_{i=1}^{\ell}$ and for $\nu = 1/\ell$, the set of optimal solutions does include $\{\xi_i^*\}_{i=1}^{\ell} = \{0\}_{i=1}^{\ell}$. ■

Lemma 2: Assume $0 < \nu < 1$ is given, then the non-negativity constraint in OP $(5), T \geq 0$, can be omitted.

Proof: (According to [20], Theorem 3, Proof in Appendix A) Assume an optimal solution of OP (5) is given by $(T^*, \mathbf{c}^*, \{\xi_i^*\}_{i=1}^{\ell})$. Further, assume that $T^* = -|T^*|$ and another feasible solution that does not change the constraints is given by $(0, \mathbf{c}^*, \{\xi_i^* - |T^*|\}_{i=1}^{\ell})$, i.e. $0 \leq \|\mathbf{c}^* - \phi(\mathbf{x}_i)\|^2 \leq -|T^*| + \xi_i^*$. It holds that

$$-|T^*| + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} \xi_i^* \geq 0 + \frac{1}{\ell\nu}\sum_{i=1}^{\ell}(\xi_i^* - |T^*|)$$

$$= \frac{-|T^*|}{\nu} + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} \xi_i^*$$

is true for $\nu < 1$ and hence, is a contradiction to the assumption that $-|T^*| = T^*$. ■

Lemma 3: Assume $\nu \geq 1$ is given, then due to the non-negativity constraint in OP $(5), T \geq 0$, the optimal solution must have $T^* = 0$.

Proof: (According to [20], Theorem 3, Proof in Appendix A) Assume an optimal solution of OP (5) is given by $(T^*, \mathbf{c}^*, \{\xi_i^*\}_{i=1}^{\ell})$ and another feasible solution, that does not change the constraints, is given by $(0, \mathbf{c}^*, \{\xi_i^* + T^*\}_{i=1}^{\ell})$. It holds that

$$T^* + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} \xi_i^* \geq 0 + \frac{1}{\ell\nu}\sum_{i=1}^{\ell}(\xi_i^* + T^*) = \frac{T^*}{\nu} + \frac{1}{\ell\nu}\sum_{i=1}^{\ell} \xi_i^* \ ,$$

is true for $\nu \geq 1$ and hence, the optimal solution must have $T^* = 0$. ■

Therefore, we can now state precise primal and dual optimization problems. Furthermore, we can derive a closed form solution to a special case of OP (5).

Theorem 2 (Primal Problem and Solution for $\nu \geq 1$): If $\nu \geq 1$ the primal optimization problem reduces to

$$\min_{\mathbf{c}} \sum_{i=1}^{\ell} \|\mathbf{c} - \phi(\mathbf{x}_i)\|^2, \qquad (8)$$

and the optimal solution is given by $\mathbf{c} = 1/\ell \sum_{i=1}^{\ell} \phi(\mathbf{x}_i)$.

Proof: According to Lemma 3, $T = 0$ and $\frac{1}{\ell\nu} > 0$ can be discarded, hence we arrive at

$$\min_{\mathbf{c}, \boldsymbol{\xi} \geq 0} \sum_{i=1}^{\ell} \xi_i$$

$$\text{subject to} \quad \|\mathbf{c} - \phi(\mathbf{x}_i)\|^2 \leq \xi_i \quad \forall\, i = 1, \ldots, \ell.$$

Further, $\xi_i \geq 0$ is due to the 2-norm always fulfilled and minimization yields the smallest possible $\xi_i = \|\mathbf{c} - \phi(\mathbf{x}_i)\|^2$, which reads unconstrained

$$\min_{\mathbf{c}} L(\mathbf{c}) = \sum_{i=1}^{\ell} \|\mathbf{c} - \phi(\mathbf{x}_i)\|^2 \ .$$

This quadratic form has a unique optimum at $\partial L(\mathbf{c})/\partial \mathbf{c} = 0$, which is $\mathbf{c} = 1/\ell \sum_{i=1}^{\ell} \phi(\mathbf{x}_i)$.

For $\nu \geq 1$, the dual problem can be solved analytically by $\boldsymbol{\alpha} = 1/\ell$. ■

To employ kernelized versions of the SVDD (i.e. OP (5)), we need to derive the corresponding dual problem.

Theorem 3 (Dual Problem): For $0 < \nu \leq 1$ and appropriately defined Mercer-kernel $\mathsf{k} : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}, \mathsf{k}(\mathbf{x}, \mathbf{y}) \mapsto \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$, the dual problem is given by

$$\max_{0 \leq \boldsymbol{\alpha} \leq \frac{1}{\ell\nu}} \sum_{i=1}^{\ell} \alpha_i \mathsf{k}(\mathbf{x}_i, \mathbf{x}_i) - \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j \mathsf{k}(\mathbf{x}_i, \mathbf{x}_j) \qquad (9)$$

$$\text{subject to} \quad \sum_{i=1}^{\ell} \alpha_i = 1$$

with expansions $\mathbf{c} = \sum_{i=1}^{\ell} \alpha_i \phi(\mathbf{x}_i)$.

Proof: Due to Lemma 2, we can skip the non-negativity constraint $T \geq 0$ of the convex OP (5). The resulting Lagrangian arrives at

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{c}, T, \boldsymbol{\xi}) = T + \frac{1}{\ell\nu} \sum_{i=1}^{\ell} \xi_i + \sum_{i=1}^{\ell} \alpha_i (\|\mathbf{c} - \phi(\mathbf{x}_i)\|^2 - T - \xi_i) - \sum_{i=1}^{\ell} \beta_i \xi_i \ ,$$

and solving for the Lagrange dual function $g(\boldsymbol{\alpha}, \boldsymbol{\beta})$ (with $\boldsymbol{\alpha} \geq 0, \boldsymbol{\beta} \geq 0$ and $g(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\mathbf{c}, T, \boldsymbol{\xi}} L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{c}, T, \boldsymbol{\xi})$) yields

(1) $\frac{1}{\ell\nu} - \beta_i - \alpha_i = 0$ and hence, $0 \leq \boldsymbol{\alpha} \leq \frac{1}{\ell\nu}$
(2) the expansion $\mathbf{c} = \sum_{i=1}^{\ell} \alpha_i \phi(\mathbf{x}_i)$
(3) the equality constraint $\sum_{i=1}^{\ell} \alpha_i = 1$.

Substitution and re-arrangement then gives us the dual optimization problem in OP (9). In order for strong duality to hold, some constraint qualifications, such as *Slater's condition*, must be fulfilled (which holds trivially, cf. [20] Section 3.1). For any primal $(\mathbf{c}^*, T^*, \boldsymbol{\xi}^*)$ and dual optimal solution $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$, the complementary slackness constraints are given by $\alpha_i^*(\|\mathbf{c}^* - \phi(\mathbf{x}_i)\|^2 - T^* - \xi_i^*) = 0$ and $\beta_i^* \xi_i^* = 0$. ■

Interestingly, the above formulation reduces to the dual one-class SVM optimization problem [6], if $\mathsf{k}(\mathbf{x}, \mathbf{y})$ is a constant for $\mathbf{x} = \mathbf{y}$. Also, the dual formulation allows for a neat interpretation of the ν parameter.

Finally, we show that the hyper-parameter ν relates to the fraction of outliers, which makes the SVDD an suitable method for anomaly detection.

Theorem 4: Given $0 < \nu \leq 1$, then $\lceil \ell\nu \rceil$ is a lower bound on the number of support vectors and an upper bound on the number of outliers.

Proof: Due to the complementary slackness constraints (Thm. 3, cf. [20], Eq. (12,17)), we know that constraints in Eq. (5) that are not strictly fulfilled yield $\alpha^* = 1/\ell\nu$ ($\xi_i^* > 0 \Rightarrow \beta_i^* = 0$ and $\frac{1}{\ell\nu} - \beta_i^* - \alpha^* = 0$ must hold), whereas constraints that are strictly fulfilled receive $\alpha^* = 0$ ($\xi_i^* = 0$, $\|\mathbf{c}^* - \phi(\mathbf{x}_i)\|^2 < T^*$ and complementary slackness must hold). For data points lying exactly on the border, it holds that $0 \leq \alpha^* \leq 1/\ell\nu$ ($\xi_i^* = 0$ and $\|\mathbf{c}^* - \phi(\mathbf{x}_i)\|^2 = T^*$). Therefore, in order to fulfill the equality constraint in Problem (9), at most $\lceil \ell\nu \rceil$ data points can strictly lie outside and there must be at least that many support vectors. ■

Therefore, it makes sense to restrict ν to be in range $]0, 1]$.

B. ClusterSVDD

In this section, we introduce our unifying formulation *ClusterSVDD* and prove that *k*-means and SVDD can be recovered as special cases.

Definition 4 (Primal Problem): Primal non-convex ClusterSVDD optimization problem (again $0 < \nu \leq 1$):

$$\min_{\{\mathbf{c}_j\}_{j=1}^k, \mathbf{T} \geq 0, \boldsymbol{\xi} \geq 0} \sum_{j=1}^{k} T_j + \sum_{j=1}^{k} \sum_{i=1}^{\ell} \frac{\mathbf{1}[z_i = j]}{\sum_l \mathbf{1}[z_l = j]\nu} \xi_i \qquad (10)$$

$$\text{subject to} \quad \|\mathbf{c}_{z_i} - \phi(\mathbf{x}_i)\|_2^2 \leq T_{z_i} + \xi_i, \ \forall\, i = 1, \ldots, \ell$$

$$\text{with} \quad z_i = \operatorname*{argmin}_{\hat{z} \in \{1, \ldots, k\}} \|\mathbf{c}_{\hat{z}} - \phi(\mathbf{x}_i)\|^2 - T_{\hat{z}}$$

Theorem 5 (Decomposability): The Problem (10) is decomposable into k sub-problems with k disjunct sets of hypershpere constraints and ℓ global cluster membership constraints.

Proof: Notice that the data can be partitioned, that is, each datum x_i can only belong to a single set S_j at any given time, where $i \in S_j$ for $j \in 1, \ldots, k$ iff $z_i = j$. It follows that $S_i \cap S_j = \emptyset$ for $i \neq j$ and $\cup_{j=1}^{k} S_j = \{1, \ldots, \ell\}$. Re-writing $\sum_{i=1}^{\ell} \frac{\mathbf{1}[z_i = j]}{\sum_l \mathbf{1}[z_l = j]\nu} \xi_i = \frac{1}{|S_j|\nu} \sum_{i \in S_j} \xi_i$ (in Problem (10)) and arranging terms accordingly achieves

$$\min_{\{\mathbf{c}_j\}_{j=1}^k, \mathbf{T} \geq 0, \boldsymbol{\xi} \geq 0} \sum_{j=1}^{k} \left(T_j + \frac{1}{|S_j|\nu} \sum_{i \in S_j} \xi_i \right)$$

$$= \sum_{j=1}^{k} \min_{\mathbf{c}_j, T_j \geq 0, \boldsymbol{\xi} \geq 0} T_j + \frac{1}{|S_j|\nu} \sum_{i \in S_j} \xi_i \qquad (11)$$

$$\begin{aligned} \text{subject to} \quad & \|\mathbf{c}_j - \phi(\mathbf{x}_i)\|^2 \leq T_j + \xi_i, \ \forall\, i \in S_j, j = 1 \\ & \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ & \|\mathbf{c}_j - \phi(\mathbf{x}_i)\|^2 \leq T_j + \xi_i, \ \forall\, i \in S_j, j = k \\ \text{with} \quad & z_i = \operatorname*{argmin}_{\hat{z} \in \{1, \ldots, k\}} \|\mathbf{c}_{\hat{z}} - \phi(\mathbf{x}_i)\|^2 - T_{\hat{z}} \ , \ \forall\, i = 1, \ldots, \ell. \end{aligned}$$

The above optimization problem is now decomposed into k distinct SVDD optimization problems that are coupled solely through the global cluster assignment constraint. Hence, by applying the notation introduced in Def. 2, the *ClusterSvdd* optimization problem OP (11) can be written as

$$\sum_{j=1}^{k} \text{Svdd}(\nu, \{\mathbf{x}_i\}_{i \in S_j}) \qquad (12)$$

$$\text{subject to} \quad z_i = \operatorname*{argmin}_{\hat{z} \in \{1, \ldots, k\}} \|\mathbf{c}_{\hat{z}} - \phi(\mathbf{x}_i)\|^2 - T_{\hat{z}} \ , \ \forall\, i = 1, \ldots, \ell.$$

■

This is also an interesting result for the optimization in Section III-D. Notably, given the partitions $\{S_j\}_{j=1}^k$, OP (12) is just a sum of convex optimization problems, which itself is a convex optimization problem [42]. Because the problem decomposes neatly into SVDD sub-problems (with exact primal-dual relations where strong duality holds), using kernels is straightforward by simply solving the dual SVDD Problem (9) instead of the primal version. We now proceed further and show the equivalence to k-means when $\nu \geq 1$.

Theorem 6 (Equivalence I): Assume $\nu \geq 1$ given and $\phi : \mathcal{X} \rightarrow \mathcal{X},\ \mathbf{x} \mapsto \mathbf{x}$ being the identity function $id_{\mathcal{X}}$, then the *ClusterSVDD* optimization problem is identical to the k-means optimization problem: OP (10) = OP (2).

Proof: Since the OP (5) can be decomposed into OP (12) and Thm. 2 holds for each sub-SVDD,

$$\sum_{j=1}^{k} \text{Svdd}(\nu \geq 1, \{\mathbf{x}_i\}_{i \in S_j}) = \sum_{j=1}^{k} \min_{\mathbf{c}_j} \sum_{i \in S_j} \|\mathbf{c}_j - \phi(\mathbf{x}_i)\|^2$$

$$\text{subject to } z_i = \underset{\hat{z} \in \{1,\ldots,k\}}{\operatorname{argmin}} \ \|\mathbf{c}_{\hat{z}} - \phi(\mathbf{x}_i)\|^2\ ,\ \forall\, i = 1, \ldots, \ell$$

which is identical to the k-means OP (2). ■

Theorem 7 (Equivalence II): Assume $k = 1$ given, then the *ClusterSVDD* optimization problem is identical to the SVDD optimization problem: OP (10) = OP (5).

Proof: Since the OP (5) can be decomposed into OP (12), the sum can be omitted, as well as the cluster membership constraints, as they always deliver $1 = z_i = \operatorname{argmin}_{\hat{z} \in \{1,\ldots,k\}} \|\mathbf{c}_{\hat{z}} - \phi(\mathbf{x}_i)\|^2\ ,\ \forall\, i = 1, \ldots, \ell$. The resulting optimization problem is SVDD$(\nu, \{\mathbf{x}_i\}_{i=1}^{\ell})$, which is in fact the original SVDD formulation as defined in Def. 2. ■

C. Relation to Kernel k-means and Spectral Clustering

From the decomposability theorem (Thm. 5) a kernelized version of our *ClusterSVDD* can be derived using the dual of the SVDD, as given in Thm. 3. Due to the expansion of $\mathbf{c} = \sum_{i=1}^{\ell} \alpha_i \phi(\mathbf{x}_i)$ of a single SVDD, we can equivalently rewrite the global cluster membership constraint of OP (10) as

$$z_i := \underset{j \in \{1,\ldots,k\}}{\operatorname{argmin}} \sum_{m,n \in S_j} \alpha_m \alpha_n \mathsf{k}_j(\mathbf{x}_m, \mathbf{x}_n)$$
$$- 2 \sum_{m \in S_j} \alpha_m \mathsf{k}_j(\mathbf{x}_m, \mathbf{x}_i) + \mathsf{k}_j(\mathbf{x}_i, \mathbf{x}_i) - T_j \quad .$$

Moreover, a proper dual version of k-means can be derived as a special case due to Thm. 6, which ensures the equivalence to kernel k-means [25]. Interestingly, Dhillon et al. [24] showed that an explicit theoretical connection between kernel k-means and spectral clustering [43] can be drawn under certain conditions. In return, there is also a connection between our *ClusterSVDD* and spectral clustering, with kernel k-means being the link.

D. Optimization

Following the ideas of CCCP [44] (concave-convex procedure), a variant of DC-programming [45] (difference of convex functions), which itself is a special instance of MM (majorization-minimization), we separate the problem into two sub-problems:

(1) inferring the partition, and
(2) calculating the new hypersphere centers and radii.

This approach does not guarantee the globally optimal solution (except for $k = 1$), but will provide locally optimal solutions. Due to Theorem (5), the optimization is similar to the original k-means optimization, where the first step also considers kernels, and the second step can be solved using existing SVDD implementations. The resulting optimization algorithm is described in Algorithm 1 (a kernelized version is given in Appendix A). Despite the non-convex nature of the optimization problem, we found in our experiments that the algorithm tends to converge fast.

Algorithm 1 ClusterSVDD

```
input data $x_1, \ldots, x_\ell$ and outlier fraction $\nu > 0$
put $t = 0$
choose $z_i \in \{1, \ldots, k\}\ \forall i \in \{1, \ldots, \ell\}$ (e.g. randomly)
let $(\mathbf{c}_j^t, T_j^t)$ be the optimal arguments when solving the SVDD optimization problem OP (5) with subset $\mathbf{x}_i,\ i \in S_j,\ \forall\, j = 1, \ldots, k$.
repeat
  t:=t+1
  for $i = 1, \ldots, \ell$ do
    $z_i^t := \operatorname{argmin}_{j \in \{1,\ldots,k\}} ||\mathbf{c}_j^{t-1} - \phi(\mathbf{x}_i)||^2 - T_j^{t-1}$
  end for
  let $(\mathbf{c}_j^t, T_j^t)$ be the optimal arguments when solving the SVDD optimization problem OP (5) with subset $\mathbf{x}_i,\ i \in S_j,\ \forall\, j = 1, \ldots, k$.
until $\forall\, i = 1, \ldots, \ell :\ z_i^t = z_i^{t-1}$
Return optimal model parameters $\mathbf{c}_\cdot := \mathbf{c}_\cdot^t$, $T_\cdot = T_\cdot^t$, the cluster memberships $z_i := z_i^t \quad \forall\, i = 1, \ldots, \ell$, and the anomaly scores $s_i := ||\mathbf{c}_{z_i}^t - \phi(\mathbf{x}_i)||^2 - T_{z_i}^t$
```

E. Implementation & Practical Considerations

We provide a PYTHON software package that includes a primal subgradient descent solver and a dual quadratic program solver for the SVDD, as well as an implementation of our unifying method *ClusterSVDD*. The package can be installed conveniently using *pip install git+https://github.com/nicococo/ClusterSvdd.git*.

The implementation of the primal SGD SVDD is based on Def. 3 and hence, does not allow for kernelization. Instead, it is a lightweight implementation that solely depends on the PYTHON-NUMPY library. This is in contrast to the dual QP SVDD, as defined in OP 9, where choosing an appropriate kernel is mandatory and the implementation depends on third-party frameworks like CVXOPT, which itself can work as a wrapper for commercial heavy-load solvers, i.e. MOSEK.

Subgradient descent methods usually do not exhibit the best convergence behavior. Luckily, SVDDs come with simple, always feasible (and often near-optimal) warm-start solutions, namely, the center-of-mass method with some appropriate radius $0 < T < \max_{i \in \{1,\ldots,\ell\}} \|\mathbf{c} - \phi(\mathbf{x}_i)\|^2$.

Unlike the primal case, when optimizing the dual QP SVDD the radius T needs to be estimated after optimization. Choosing the correct radius is not straightforward. Since SVs lying on the boundary are in the range $0 \leq \alpha \leq 1/\ell\nu$, there are cases where all SVs are $\alpha = 1/\ell\nu$. In our framework, we therefore choose $T = \min_{i \in SV} \|\mathbf{c}^* - \phi(\mathbf{x}_i)\|^2$ as an upper bound on the true radius. This particular problem is discussed in Wang et al. [19].

IV. Experiments

To sketch the benefits of our *ClusterSVDD*, we conducted experiments on artificially generated data and on real-world data. First, we would like to show that multiple spheres can be beneficial for anomaly detection tasks. Second, we will attempt to show that our *ClusterSVDD* can be beneficial for cluster membership identification. We emphasize that the experiments are of a descriptive nature, showing that improvements *can*, for anomaly detection and clustering, be achieved by using *ClusterSVDD* under comparable conditions.

To get a rough idea how clustering and anomaly detection benefits from our *ClusterSVDD* formulation, we show results for the two settings in Fig. 2 and Fig. 3 respectively. In both cases, we generated data from four Gaussians and added uniformly generated anomalies on top. The plots show the ground truth, the solution for our *ClusterSVDD* as well as the corresponding baseline method (*k*-means for clustering and SVDD for anomaly detection). Further, results are shown for linear cases and for RBF kernel cases.

For each of the experiments, we report area under the ROC curve (AUROC) for anomaly detection accuracy and adjusted Rand index (ARI) for cluster membership accuracy. AUROC is defined as the integral of the curve given by the true positive rate (y-axis) and false positive rate (x-axis) in the interval $[0, 1]$. AUROC can be seen as a ranking measure of how well positive labeled data is separated from the negative data, which makes it also applicable in unbalanced label settings. The adjusted Rand index (ARI) is an extension of the Rand index (RI) and measures the accuracy of the predicted assignment when compared to the ground truth assignment (e.g. regardless of the permutations of cluster numbering) adjusted for chance level. The ARI is a natural choice for assessing cluster method performance in the presence of ground truth label data.

A. Concise Descriptions for Anomaly Detection

We draw 1.000 training instances and 2.000 test instances from three isotropic Gaussian distributions in two dimensions. Experiments were repeated 50 times, and we report means and standard errors. Two out of three Gaussians are labeled as normal, and one will serve as an anomaly source. The fraction of outliers is fixed to 5%, and we report AUROC scores for our *ClusterSVDD* with $k = 2, 3, 4$ as well as standard SVDD using the primal formulation (cf. Fig. 4, left) and RBF kernels (cf. Fig. 4, right). For both experiments, we varied the regularization parameter ν and chose additional parameters (i.e. RBF variance $\sigma^2 \in \{0.1, 0.25, 0.5, 1.0, 2.0\}$) using a cross validation approach where training data was further split into training and validation data. Results for primal formulation, in Fig. 4 (left), suggest that our multiple spheres approach is beneficial when the description is not overly rich (e.g. $k > 2$). Otherwise, the multiple spheres tend to enclose the outliers as well as the normal data, making them harder to distinguish. Due to the rich description in the case of RBF kernels (cf. Fig. 4, right), we assume not much difference in accuracy for both methods. And, indeed, despite a small advantage for *ClusterSVDD*, we observe similar maximum accuracy. However, *ClusterSVDD* shows an overall more stable behavior for the whole range of regularization parameters.

B. Robustness for Cluster Identification

Again, we draw 1.000 training instances and 2.000 test instances from three isotropic Gaussian distributions in two dimensions. Experiments were repeated 50 times, and we report means and standard errors. Here, we test cluster identification accuracy in adjusted Rand index (ARI) for a varying number of $k = 2, 3, 4$ for *ClusterSVDD* and *k*-means. Fig. 5 (left) shows the results for cluster identification against *k*-means (red, $\nu = 1$). It can be seen that, even in the isotropic Gaussian case, which is the Gold standard for *k*-means, improvements can be achieved when using our *ClusterSVDD*. Interestingly, below a certain threshold ($\nu < 0.225$), the accuracy drops drastically. This can be explained by a significant change of the cluster means towards uncommon data points, which need to be embedded within the hypershperes boundaries. Fig. 5 (right) shows the same setting, but with the kernelized version of our method, which especially includes solutions for kernel *k*-means. Here, a similar behavior is visible, with our method achieving highest accuracy, although it seems to behave less stably in the region around $\nu \in [0.6, 0.9]$.

C. Results on Real-world Data Sets

In the previous sections IV-A and IV-B, we showed very well on artificially generated data that our approach can indeed be useful for improving cluster membership identification accuracy and anomaly detection accuracy. In this section, however, we apply our method to real world data sets.

We selected two multi-class data sets from the libSVM data set library[1] where features are normalized between $[-1, +1]$ (details shown in Table II). Half of the data was used for training, and the other half for testing. Validation data was further split from training data for model selection. Experiments are repeated 10 times, and means and standard deviations are reported. We injected 0%, 2%, 5%, 10% and 15% of uniformly generated random data points as outliers (which we refer to as *noise level*), and tested various k (cf. Table II) and $\nu \in \{1.0, 0.95, 0.9, 0.5, 0.1, 0.01\}$. Results are shown in Table III and Table IV.

In both experiments, a data description with multiple spheres proved beneficial for anomaly detection, when compared to the traditional single sphere description of the SVDD. This comes as no surprise, as the data consists of multiple clusters and anomalies emerge as a convex combination of normal data.

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

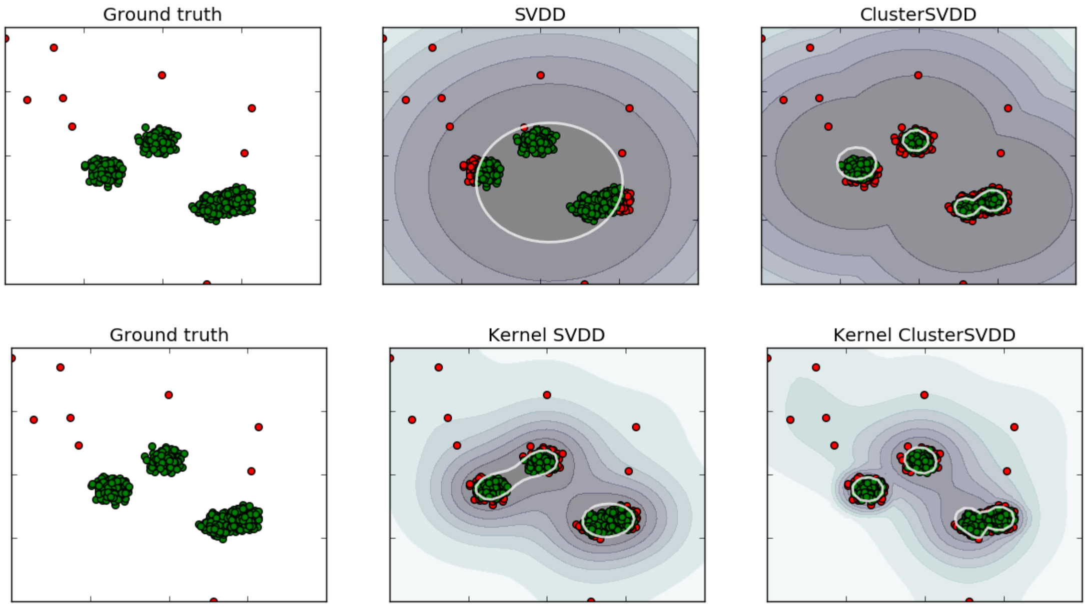

Fig. 2. Results for anomaly detection settings for our *ClusterSVDD* (right column) and SVDD (center column) for linear settings (top row) and RBF kernel settings (bottom row) by assumption of 10% outlier ($\nu = 0.1$). Ground truth (left column) shows nominal data in green color and outliers in red. Note that the description learned by our *ClusterSVDD* is much more concise than for SVDD for the linear and the kernel case.

TABLE II
Overview of used data sets and tested number of clusters.

Name	Features	Instances	Classes	k
Segment	19	2,310	7	1, 5, 7, 10, 14
SatImage	36	4,435	6	1, 3, 6, 9

Interestingly, cluster membership accuracy did only improve slightly (cf. Table III), or not at all (cf. Table IV). A possible explanation for this discovery led us to a re-interpretation of k-means clustering. Traditionally, k-means is supposed to be an un-regularized, non-convex optimization problem that assumes all data to be normal but from Thm (7) and Thm (6) we can directly relate k-means to SVDDs with $\nu = 1$. The regularization parameter ν itself has an interpretation as an outlier fraction (cf. Thm (4)) and as a density level set (cf. Section II). Applied to k-means, this means k-means does indeed assume a maximum amount of outliers in the data, or, in other words, it is aiming for the *mode* of the density.

Furthermore, SVDDs and related one-class SVMs (Thm (1)) are properly regularized risk minimization problems. This makes k-means also a properly regularized risk minimization problem.

In light of these results, together with the theoretical framework derived in Section III, we therefore should re-interpret k-means clustering:

k-*means clustering is a regularized*
mode seeking algorithm.

TABLE III
Results on Segment data set.

	AUROC		ARI	
Noise level	SVDD	*ClusterSVDD*	k-means	*ClusterSVDD*
0%	-/-	-/-	0.52/0.04	0.52/0.05
2%	0.98/0.01	1.00/0.00	0.50/0.03	0.50/0.01
5%	0.98/0.00	1.00/0.00	0.52/0.02	0.53/0.03
10%	0.97/0.00	1.00/0.00	0.49/0.05	0.52/0.02
15%	0.97/0.01	1.00/0.00	0.50/0.05	0.51/0.03

TABLE IV
Results on SatImage data set.

	AUROC		ARI	
Noise level	SVDD	*ClusterSVDD*	k-means	*ClusterSVDD*
0%	-/-	-/-	0.54/0.01	0.54/0.01
2%	0.94/0.01	1.00/0.00	0.55/0.04	0.55/0.04
5%	0.93/0.01	1.00/0.00	0.52/0.03	0.52/0.03
10%	0.93/0.01	1.00/0.00	0.55/0.04	0.55/0.04
15%	0.93/0.00	1.00/0.00	0.53/0.02	0.53/0.02

V. Discussion: Deriving new Algorithms

The previous section shed some light on the properties of our *ClusterSVDD* and delivered insights into SVDD and k-means when used in anomaly detection and clustering tasks on artificial and real data. In this section, however, we show how new algorithms can be derived by leveraging the link between clustering and one-class learning. Specifically, we derive a clustering method for structured data by transferring knowledge from some one-class setting, namely joint kernel support estimation (JKSE) [46], to clustering. Therefore, we are blending the ideas of k-means, kernel k-means, SVDD, One-class SVM, and JKSE into a single framework for clustering structured data (cf. Table V). At the same time, we extend JKSE with a multiple spheres mixture model and evaluate the structured output accuracy.

Joint kernel support estimate [46] is an unsupervised method to estimate the support of a joint probability density

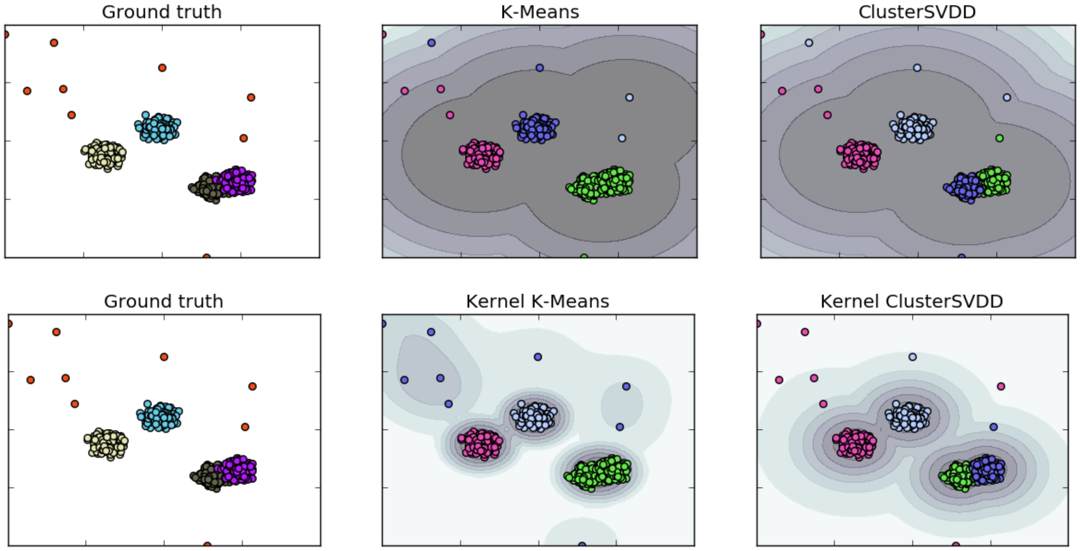

Fig. 3. Clustering results for our *ClusterSVDD* and k-means for linear settings (top row) and RBF kernel settings (bottom row). The ground truth (left column) was generated from four Gaussians with similar variance plus some uniformly generated anomalies (red dots). As can be seen, (kernel) k-means assigns one cluster center to fit the anomalies and is therefore not able to reveal the four Gaussian clusters, whereas our *ClusterSVDD* concisely finds the four Gaussians.

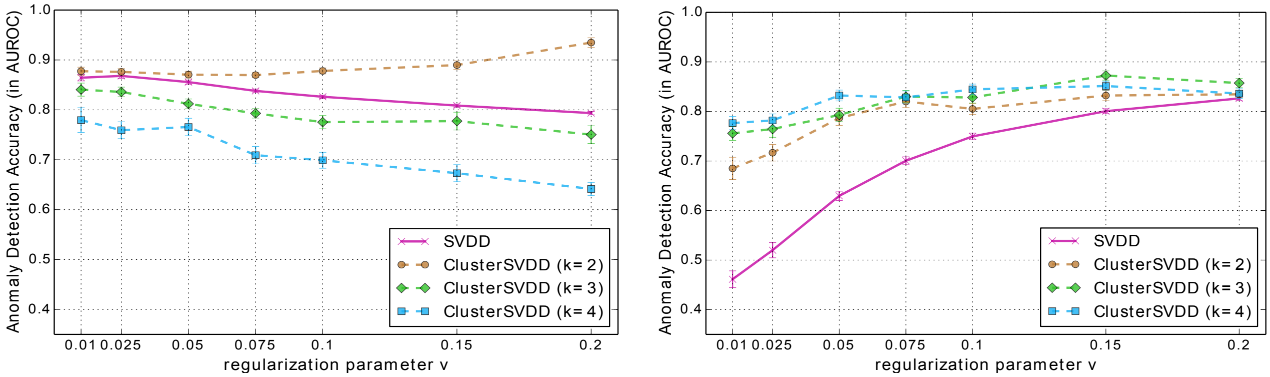

Fig. 4. Anomaly detection accuracy (in AUROC) of standard SVDD against our *ClusterSVDD* with $k = 2, 3, 4$ for varying regularization parameter ν and a fraction of 5% of outlier in the data set. Left: linear version. Right: kernelized version (RBF kernel).

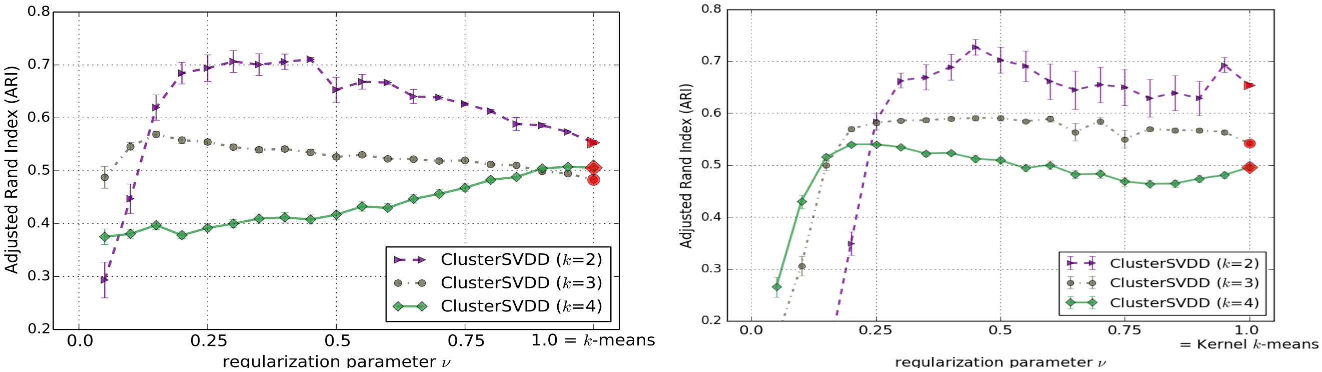

Fig. 5. Cluster membership identification accuracy (in adjusted Rand index, ARI) of our *ClusterSVDD* including standard k-means clustering for varying regularization parameter ν. The original k-means solution is recovered at $\nu = 1.0$ and plotted in red color. Left: linear version. Right: kernelized version (RBF kernel).

TABLE V
SUMMARY OF METHODS AND THEIR RESPECTIVE PROPERTIES.

	Citation	Clustering	Outlier-detection	Kernels	Mixture-models	Latent State Inference	Structured Output
k-means	[22]	x	-	-	x	x	-
Kernel k-means	[23]	x	-	x	x	x	-
SVDD	[8], [7]	-	x	x	-	-	-
One-class SVM	[6]	-	x	x	-	-	-
JKSE	[46]	-	x	x	-	-	x
ClusterSVDD for structured data	Section V	x	x	x	x	x	x

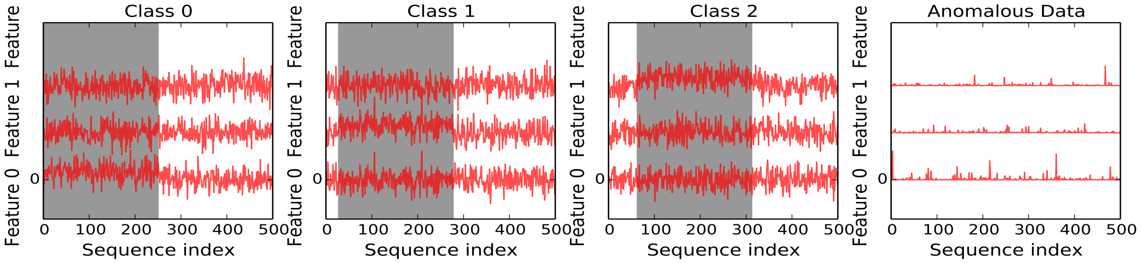

Fig. 6. The generated structured data consists of multivariate (3 dimensions: Feature 0, Feature 1, and Feature 2) Gaussian sequences of length 500 (red) and corresponding 2-state label sequences (white and gray areas) of three classes and a fraction of anomalies. Notice that the differences between the classes are subtle, rendering the overall problem non-trivial.

$p(\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{\ell})$ of observations $\mathbf{x}_i \in \mathcal{X}$ and corresponding output structures $\mathbf{y}_i \in \mathcal{Y}$. For that, JKSE employs the one-class SVM with an adaptation of the feature vector to joint feature maps of both observations and labels $\phi(\mathbf{x}_i) := \Psi(\mathbf{x}_i, \mathbf{y}_i)$. Given the estimated solution $\mathbf{w}^*$ of the one-class SVM, the latent state structure $\mathbf{y}$ is obtained by solving the *maximum-a-posteriori* (MAP) inference problem:

$$\mathbf{y} = \underset{\hat{\mathbf{y}} \in \mathcal{Y}}{\operatorname{argmax}} \langle \mathbf{w}^*, \Psi(\mathbf{x}, \hat{\mathbf{y}}) \rangle \ . \tag{13}$$

In the case of sequences with corresponding latent states (hidden Markov models), this can be efficiently solved by dynamic programming (i.e. Viterbi [47]).

Leveraging the relations between one-class SVMs and SVDDs (Thm (1)), SVDDs and *ClusterSVDD* (Thm (6)), and finally *ClusterSVDD* and k-means clustering (Thm (7)), we only need to apply JKSE using our *ClusterSVDD* and increase $k = 1$ for a clustering setting.

In order to conform to the properties that tie one-class SVMs and SVDDs together, we need to ensure that only translation-invariant kernels are used, or, that the (joint) feature maps are L_2-normalized. Then, the latent state sequence can be inferred by OP (13) with $\mathbf{c}_z^* =: \mathbf{w}^*$ for given $z \in \{1, \ldots, k\}$:

$$\begin{aligned} \mathbf{y} &= \underset{\hat{\mathbf{y}} \in \mathcal{Y}}{\operatorname{argmin}} \, \|\mathbf{c}_z^* - \Psi(\mathbf{x}, \hat{\mathbf{y}})\|^2 - T_z^* \\ &= \underset{\hat{\mathbf{y}} \in \mathcal{Y}}{\operatorname{argmin}} \, \|\mathbf{c}_z^*\|^2 - 2\langle \mathbf{c}_z^*, \Psi(\mathbf{x}, \hat{\mathbf{y}}) \rangle + \|\Psi(\mathbf{x}, \hat{\mathbf{y}})\|^2 - T_z^* \\ &= \underset{\hat{\mathbf{y}} \in \mathcal{Y}}{\operatorname{argmax}} \langle \mathbf{c}_z^*, \Psi(\mathbf{x}, \hat{\mathbf{y}}) \rangle \quad . \end{aligned}$$

Finally, cluster membership z of a given input instance $\mathbf{x}$ can be inferred by

$$z := \underset{\hat{z} \in \{1, \ldots, k\}}{\operatorname{argmin}} \, \|\mathbf{c}_{\hat{z}}^* - \Psi(\mathbf{x}, \underset{\mathbf{y}}{\operatorname{argmax}} \langle \mathbf{c}_{\hat{z}}^*, \Psi(\mathbf{x}, \mathbf{y}) \rangle)\|^2 - T_{\hat{z}}^* \ .$$

Here, we summarize the steps that lead us from JKSE to structured clustering using our methodology:

1. JKSE uses one-class SVMs to train a generative model of examples with dependency structure using the notion of joint feature maps [46], [48].
2. One-class SVMs are equivalent to SVDDs if (a) translation invariant kernels (e.g. RBF) are used, or, (b) if input instances are L_2-normalized (Thm (1))
3. *ClusterSVDD* is equivalent to SVDD for $k = 1$ (Thm (7))
4. If $k > 1$ and $\nu = 1$ *ClusterSVDD* is equivalent to k-means (or kernel k-means respectively, Thm (6))

For our example, we generated 2500 sequences of length 500 from multivariate Gaussians where the mean depends on a corresponding latent 2-state label sequence. While for state 1, all sequences stem from a zero-mean Gaussian, for the second state, one out of three features is generated from a Gaussian with mean 0.5, depending on the corresponding class. This simulates complex structured labels, where single state sequences can be generated by different processes (cf. Fig. 6). A real-world example that exhibits this behavior is the problem of gene finding, where label sequences consist of 'genic' and 'intergenic' states, but both have biases depending on position on the DNA, sequencing technique, etc. We would like to emphasize that no changes to the algorithm need to be made if the sequences have varying lengths. The total length of all training sequences was 1,000,000, and for testing, 250,000.

We measure the structured output loss in Hamming-distance of the true to the predicted latent state sequence (denoted as Δ_{Hamming}) and normalize it to be in the range $[0, 1]$ (where 0 is perfect reconstruction and 1 is the inverse reconstruction) and, as usual, the cluster membership identification accuracy in ARI. The experiment was repeated 10 times, and means and standard deviations are reported. Results are shown in Table VI for a range of regularization parameter ν and show a near-perfect cluster membership identification for $\nu = 0.1$

TABLE VI
RESULTS FOR CLUSTERING ON STRUCTURED DATA.

	Δ_{Hamming}-loss		ARI	
Regularizer ν	SVDD	ClusterSVDD	k-means	ClusterSVDD
1.00	0.33/0.00	0.30/0.02	0.84/0.20	0.84/0.20
0.90	0.34/0.00	0.30/0.02	-/-	0.75/0.20
0.50	0.33/0.00	0.29/0.01	-/-	0.84/0.24
0.10	0.33/0.00	0.28/0.01	-/-	0.95/0.14
0.01	0.33/0.00	0.31/0.02	-/-	0.53/0.35

with an ARI of 0.95. Furthermore, a multiple spheres approach decreases the structured output loss by 5%.

VI. Conclusions

In this work, we introduced *ClusterSVDD*, blending the ideas of clustering using *k*-means and one-class classification using support vector data descriptions (SVDDs) into a single framework. We rigorously reviewed their properties and showed empirically that the new methodology performs better in clustering and anomaly detection settings. The revealed relation between *k*-means clustering and one-class classification enabled us to identify *k*-means as a mode seeking algorithm solving a regularized risk minimization problem. With our new method, we were able to turn JKSE, a structured prediction method, into a clustering method for structured data. For the practitioner, we provide an easy-to-use PYTHON software package, which can be conveniently installed using *pip install git+https://github.com/nicococo/ClusterSvdd.git*.

Future lines of research will focus on more sophisticated optimization schemes [49], [50], representation learning such as multiple kernel learning (MKL) [51], [52], [53], [54], and imputation of missing information as well as handling of additional (e.g. label) information [12].

VII. Acknowledgments

NG was supported by BMBF ALICE II grant 01IB15001B. MK acknowledges support by the German Research Foundation through the grant KL 2698/1-1 and KL 2698/2-1. We also acknowledges the support by the German Research Foundation through the grant DFG MU 987/6-1 and RA 1894/1-1. KRM thanks for partial funding by the National Research Foundation of Korea funded by the Ministry of Education, Science, and Technology in the BK21 program. SN was supported by the German Ministry for Education and Research as Berlin Big Data Center BBDC, funding mark 01IS14013A. LAL acknowledges the support of Petrobras.

Appendix

A. Kernel ClusterSVDD Algorithm

A kernelized version for Algorithm 1 is given in Algorithm 2.

References

[1] M. M. Moya, "A constrained second-order network with mean square error minimization and boundary size minimization for one-class classification," Ph.D. dissertation, The University of New Mexico, USA, 1991.

[2] M. M. Moya and D. R. Hush, "Network constraints and multi-objective optimization for one-class classification," *Neural networks*, vol. 9, no. 3, pp. 463–474, 1996.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[4] C. C. Aggarwal, *Outlier Analysis*. Heidelberg, Germany: Springer, 2013.

[5] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, a. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," Microsoft Research, Redmond, USA, Tech. Rep., 1999.

[6] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, a. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution." *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[7] D. Tax and R. Duin, "Support Vector Data Description," *Machine Learning*, pp. 45–66, 2004.

[8] D. Tax and R. Duin, "Data domain description using support vectors," in *Proceedings of the European Symposium on Artificial Neural Networks*, vol. 256, 1999, pp. 251–256.

[9] A. Munoz and J. M. Moguerza, "One-class support vector machines and density estimation: The precise relation," in *Iberoamerican Congress on Pattern Recognition (CIARP)*, vol. 9. Springer, 2004, pp. 216–223.

[10] R. Vert and J.-P. Vert, "Consistency and Convergence Rates of One-Class SVMs and Related Algorithms," *JMLR*, 2006.

[11] N. Görnitz, M. Kloft, and U. Brefeld, "Active and semi-supervised data domain description," in *ECML*. Springer, 2009, pp. 407–422.

[12] N. Görnitz, M. Kloft, K. Rieck, and U. Brefeld, "Toward Supervised Anomaly Detection." *Journal of Artificial Intelligence Research (JAIR)*, vol. 46, pp. 235–262, 2013.

[13] X.-l. Li and B. Liu, "Learning from Positive and Unlabeled Examples with Different Data Distributions," in *ECML*, 2005.

[14] N. Görnitz, A. K. Porbadnigk, A. Binder, C. Sanelli, M. Braun, K.-R. M̈ller, and M. Kloft, "Learning and Evaluation in Presence of Non-i . i . d . Label Noise," in *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 33, 2014.

[15] N. Görnitz, M. Braun, and M. Kloft, "Hidden Markov Anomaly Detection," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[16] A. Ben-hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "Support Vector Clustering," *JMLR*, vol. 2, pp. 125–137, 2001.

[17] T. Le, D. Tran, W. Ma, and D. Sharma, "A Theoretical Framework for Multi-sphere Support Vector Data Description," in *ICONIP*, 2010, pp. 132–142.

[18] C.-c. Chang and H.-c. Tsai, "A Minimum Enclosing Balls Labeling Method for Support Vector Clustering," National Taiwan University of Science and Technology, Taiwan, Tech. Rep., 2007.

[19] X. Wang, F.-l. Chung, and S. Wang, "Theoretical analysis for solution of support vector data description," *Neural networks*, vol. 24, pp. 360–369, 2011.

[20] W.-c. Chang, C.-p. Lee, and C.-j. Lin, "A Revisit to Support Vector Data Description (SVDD)," 2010.

[21] G. Rätsch, S. Mika, B. Schölkopf, and K.-R. Müller, "Constructing boosting algorithms from SVMs: an application to one-class classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1184–1199, 2002.

[22] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. The Regents of the University of California, 1967.

[23] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, jul 1998.

[24] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means, Spectral Clustering and Normalized Cuts," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, New York, USA: ACM Press, aug 2004, p. 551.

[25] M. Girolami, "Mercer kernel-based clustering in feature space," *IEEE Transactions on Neural Networks and Learning (TNNLS)*, vol. 13, no. 3, pp. 780–4, jan 2002.

[26] P. A. Forero, V. Kekatos, and G. B. Giannakis, "Robust Clustering Using Outlier-Sparsity Regularization," *IEEE Transactions on Signal Processing*, apr 2012.

[27] Y. Kondo, "Robustification of the sparse K-means clustering algorithm," Master's thesis, The University Of British Columbia, Vancouver, Canada, 2011.

[28] Y. Kondo, M. Salibian-Barrera, and R. Zamar, "A robust and sparse K-means clustering algorithm," *Arxiv*, 2012.

Algorithm 2 Kernel ClusterSVDD

```
input data $x_1, \dots, x_\ell$ and outlier fraction $\nu > 0$
put $t = 0$
choose $z_i \in \{1, \dots, k\}\ \forall i \in \{1, \dots, \ell\}$ (e.g. randomly) and therefore fixing $S_j^t, \forall j = 1, \dots, k$
let $(\boldsymbol{\alpha}_j^t, T_j^t)$ be the optimal arguments when solving the SVDD dual optimization problem OP (9) with subset $\mathbf{x}_i,\ i \in S_j^t$, $\forall\, j = 1, \dots, k$ for the corresponding kernel $\mathsf{k}_j$.
repeat
  t:=t+1
  for $i = 1, \dots, \ell$ do
    $z_i^t := \text{argmin}_{j \in \{1,\dots,k\}} \sum_{m,n \in S_j^{t-1}} \alpha_m^{t-1} \alpha_n^{t-1} \mathsf{k}_j(\mathbf{x}_m, \mathbf{x}_n) - 2 \sum_{m \in S_j^{t-1}} \alpha_m^{t-1} \mathsf{k}_j(\mathbf{x}_m, \mathbf{x}_i) + \mathsf{k}_j(\mathbf{x}_i, \mathbf{x}_i) - T_j^{t-1}$
    Note: $z_i^t, \forall i = 1, \dots, \ell$ implies $S_j^t, \forall\, j = 1, \dots, k$
  end for
  let $(\boldsymbol{\alpha}_j^t, T_j^t)$ be the optimal arguments when solving the SVDD dual optimization problem OP (9) with subset $\mathbf{x}_i,\ i \in S_j^t$, $\forall\, j = 1, \dots, k$ for the corresponding kernel $\mathsf{k}_j$.
until $\forall\, i = 1, \dots, \ell:\ z_i^t = z_i^{t-1}$
Return optimal model parameters $\boldsymbol{\alpha}_\cdot := \boldsymbol{\alpha}_\cdot^t$, $T_\cdot = T_\cdot^t$, the cluster memberships $z_i := z_i^t \quad \forall\, i = 1, \dots, \ell$, and the anomaly scores $s_i := \sum_{m,n \in S_{z_i}^t} \alpha_m^t \alpha_n^t \mathsf{k}_j(\mathbf{x}_m, \mathbf{x}_n) - 2 \sum_{m \in S_{z_i}^t} \alpha_m^t \mathsf{k}_j(\mathbf{x}_m, \mathbf{x}_i) + \mathsf{k}_j(\mathbf{x}_i, \mathbf{x}_i) - T_j^t$
```

[29] G. Hamerly and C. Elkan, “Learning the k in k-means,” in *NIPS*, 2004.

[30] Q. Gu and J. Han, “Clustered Support Vector Machines,” in *AISTATS*, vol. 31, 2013, pp. 307–315.

[31] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, jun 2010.

[32] K. R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, “An introduction to kernel-based learning algorithms.” *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, jan 2001.

[33] B. Schölkopf, S. Mika, C. J. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. J. Smola, “Input Space Versus Feature Space in Kernel-Based Methods,” *IEEE Transactions on Neural Networks*, 1999.

[34] Jinwen Ma and Taijun Wang, “A cost-function approach to rival penalized competitive learning (RPCL),” *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 4, pp. 722–737, 2006.

[35] D. Bacciu and A. Starita, “Competitive Repetition Suppression (CoRe) Clustering: A Biologically Inspired Learning Model With Application to Robust Clustering,” *IEEE Transactions on Neural Networks*, vol. 19, no. 11, pp. 1922–1941, 2008.

[36] Yiu-ming Cheung, “On rival penalization controlled competitive learning for clustering with automatic cluster number selection,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 11, pp. 1583–1588, 2005.

[37] H. Jia, Y.-M. Cheung, and J. Liu, “Cooperative and penalized competitive learning with application to kernel-based clustering,” *Pattern Recognition*, vol. 47, pp. 3060–3069, 2014.

[38] C. D. Scott and R. D. Nowak, “Learning Minimum Volume Sets,” *JMLR*, vol. 7, pp. 665–704, 2006.

[39] A. B. Tsybakov, “On nonparametric estimation of density level sets,” *Annals of Statistics*, vol. 25, no. 3, pp. 948–969, 1997.

[40] Y. Xiao, B. Liu, L. Cao, X. Wu, C. Zhang, Z. Hao, F. Yang, and J. Cao, “Multi-sphere Support Vector Data Description for Outliers Detection on Multi-distribution Data,” in *2009 IEEE International Conference on Data Mining Workshops*, 2009, pp. 82–87.

[41] A. K. Porbadnigk, N. Görnitz, C. Sannelli, A. Binder, M. Braun, M. Kloft, and K.-R. Müller, “Extracting latent brain states Towards true labels in cognitive neuroscience experiments,” *NeuroImage*, vol. 120, pp. 225–253, 2015.

[42] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambrigde University Press, 2004.

[43] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On Spectral Clustering: Analysis and an algorithm,” in *NIPS*, 2002.

[44] A. L. Yuille and A. Rangarajan, “The concave-convex procedure,” *Neural computation*, vol. 15, no. 4, pp. 915–36, apr 2003.

[45] R. Horst and N. V. Thoai, “DC Programming : Overview,” *Journal of Optimization Theory and Applications*, vol. 103, no. 1, pp. 1–43, 1999.

[46] C. H. Lampert and M. B. Blaschko, “Structured prediction by joint kernel support estimation,” *Machine Learning*, vol. 77, no. 2-3, pp. 249–269, apr 2009.

[47] L. R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, 1989.

[48] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, “Large Margin Methods for Structured and Interdependent Output Variables,” *Journal of Machine Learning Research*, vol. 6, pp. 1453–1484, 2005.

[49] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers,” *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[50] N. Parikh and S. Boyd, “Proximal Algorithms,” *Foundations and Trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[51] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, “SimpleMKL,” *JMLR*, vol. 9, pp. 2491–2521, 2008.

[52] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien, “lp-Norm Multiple Kernel Learning,” *JMLR*, vol. 12, pp. 953–997, 2011.

[53] M. Kloft, U. Brefeld, P. Laskov, K.-R. Müller, A. Zien, and S. Sonnenburg, “Efficient and Accurate Lp-Norm Multiple Kernel Learning,” in *Advances in Neural Information Processing Systems*, 2009, pp. 997–1005.

[54] J. A. Nasir, N. Görnitz, and U. Brefeld, “An Off-the-shelf Approach to Authorship Attribution,” in *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, 2014.

Nico Görnitz is a research associate in the machine learning group at TU Berlin (Berlin Institute of Technology, Germany), headed by Klaus-Robert Müller. Before an internship with the eScience Group, led by David Heckerman (Microsoft Research, Los Angeles, US) in 2014, he was employed as a research associate at TU Berlin (2010-2014) and during 2010-2012 also affiliated with the Friedrich Miescher Laboratory of the Max Planck Society in Tübingen, Germany, where he was co-advised by Gunnar Rätsch. Nico is interested in machine learning in general and in one-class learning based anomaly detection for data with dependency structure, large-margin structured output learning and corresponding optimization techniques in specific. Applications that Nico has been working on cover computational biology, computer security, computational sustainability, brain-computer-interfaces, natural language processing and porosity estimation for geosciences.

Luiz Alberto Lima is a senior consultant at the Exploration Division of Petrobras - Petrleo Brasileiro S.A. in Brazil. He received his Ph.D. degree in Electrical Engineering with emphasis on Decision Support Systems from PUC-Rio in Brazil in 2017. During his Ph.D., he worked for two years with the machine learning group at the Berlin Institute of Technology (TU Berlin), Germany, under the supervision of Prof. Klaus-Robert Müller. His interests are machine learning, in particular structured output learning, and computer graphics, with focus on Geosciences classification and prediction problems.

Klaus-Robert Müller studied physics at University of Karlsruhe, Germany, from 1984 to 1989 and received the Ph.D. degree in computer science from University of Karlsruhe in 1992. He has been a Professor of computer science at Technische Universität Berlin, Berlin, Germany, since 2006. At the same time he has been the Director of the Bernstein Focus on Neurotechnology Berlin until 2013; from 2014 he has been Co-director of the Berlin Big Data Center. After completing a postdoctoral position at GMD FIRST in Berlin, he was a Research Fellow at the University of Tokyo from 1994 to 1995. In 1995, he founded the Intelligent Data Analysis group at GMD-FIRST (later Fraunhofer FIRST) and directed it until 2008. From 1999 to 2006, he was a Professor at the University of Potsdam. Dr. Müller was awarded the 1999 Olympus Prize by the German Pattern Recognition Society, DAGM, and, in 2006, he received the SEL Alcatel Communication Award. In 2014, he received the Berliner Wissenschaftspreis des regierenden Bürgermeisters. In 2012, he was elected to be a member of the German National Academy of Sciences-Leopoldina. His research interests are intelligent data analysis, machine learning, signal processing, and brain-computer interfaces.

Marius Kloft is a junior professor of machine learning at the Department of Computer Science of Humboldt University of Berlin. At the same time, he is leading since 2015 the Emmy-Noether research group on statistical learning from dependent data. Prior to joining HU Berlin he was a joint postdoctoral fellow at the Courant Institute of Mathematical Sciences and Memorial Sloan-Kettering Cancer Center, New York, working with Mehryar Mohri, Corinna Cortes, and Gunnar Rätsch. From 2007-2011, he was a PhD student in the machine learning program of TU Berlin, headed by Klaus-Robert Müller. He was co-advised by Gilles Blanchard and Peter L. Bartlett, whose learning theory group at UC Berkeley he visited from 10/2009 to 10/2010. In 2006, he received a diploma (MSc equivalent) in mathematics from the University of Marburg with a thesis in algebraic geometry.

Shinichi Nakajima is a senior researcher in the Berlin Big Data Center, Machine Learning Group, Technische Universität Berlin. He received the master degree on physics in 1995 from Kobe university, and worked with Nikon Corporation until September 2014 on statistical analysis, image processing, and machine learning. He received the doctoral degree on computer science in 2006 from Tokyo Institute of Technology. His research interest is in theory and applications of machine learning, in particular, Bayesian learning theory, computer vision, and data mining.