Sparse Kernel-Based Hyperspectral Anomaly Detection

Prudhvi Gurram, Member, IEEE, Heesung Kwon, Senior Member, IEEE, and Timothy Han

Abstract—In this letter, a novel ensemble-learning approach for anomaly detection is presented. The proposed technique aims to optimize an ensemble of kernel-based one-class classifiers, such as support vector data description (SVDD) classifiers, by estimating optimal sparse weights of the subclassifiers. In this method, the features of a given multivariate data set representing normalcy are first randomly subsampled into a large number of feature subspaces. An enclosing hypersphere that defines the support of the normalcy data in the reproducing kernel Hilbert space (RKHS) of each respective feature subspace is estimated using standard SVDD. The joint hypersphere in the RKHS of the combined kernel is learned by optimally combining the weighted individual kernels while imposing the $l_1$ constraint on the combining weights. The joint hypersphere representing the optimal compact support of the multivariate data in the joint RKHS is then used to test a new data point to determine if it belongs to the normalcy data or not. A performance comparison between the proposed algorithm and regular SVDD is reported using hyperspectral image data as well as general multivariate data.

Index Terms—Hyperspectral anomaly detection, sparse kernel-based ensemble learning (SKEL), support vector data description (SVDD).

I. INTRODUCTION

ENSEMBLE learning has been widely used in data and pattern classification because an ensemble decision based on exploiting a large degree of freedom available in high-dimensional multivariate data corrupted by noise and outliers can generally provide more robust generalization performance than the regular decision by a single classifier [1]. This is particularly true in the case of hyperspectral classification and detection problems [2], [3].

A novel ensemble-learning technique called sparse kernel-based ensemble learning (SKEL) [2], [3] has been recently developed by two of the current authors for hyperspectral classification problems that use support vector machines (SVMs) as subclassifiers. In SKEL, subdecision functions were first learned within the respective randomly selected feature subspaces, and an optimal sparse combination of the subdecision functions from a large number of SVMs was subsequently estimated through the $l_1$-norm-constrained optimization of the kernel weights. This is based on multiple kernel learning (MKL) algorithm developed in [5]. A significant improvement in detection performance was reported by using SKEL for hyperspectral as well as multivariate data.

In this letter, the principle of SKEL is extended to hyperspectral anomaly detection where the examples of the reference signatures of the objects of interest are no longer available in advance. The proposed technique is called sparse kernel-based anomaly detection (SKAD) where initially, a large number of one-class classifiers based on the support vector data description (SVDD) method [6] are used as subclassifiers. SVDD is one of the state-of-the-art techniques widely used for hyperspectral anomaly detection [7]. Unlike generative model-based techniques in [8] and [9], SVDD learns the support or boundary of the given normalcy data by building a minimal enclosing hypersphere containing the data. The use of nonlinear kernels allows SVDD to accurately model the nonlinear support/boundary of nontrivial multimodal distributions [6]. The kernel-based SVDD first nonlinearly maps the input data to a high-dimensional feature space called reproducing kernel Hilbert space (RKHS) and then finds the enclosing hypersphere of the data.

Similar to SKEL, in SKAD, the SVDD-based subclassifiers (one-class classifiers) are optimally designed in the respective randomly selected feature subspaces prior to optimizing the ensemble. A new RKHS associated with a kernel that is a convex combination of these individual kernels can be determined, and a hypersphere enclosing the normalcy spectra can be built. The most compact enclosing hypersphere containing the local normalcy spectra in this RKHS corresponding to the sparsely combined kernel is obtained by minimizing the radius of the hypersphere.

The optimal sparse weights of the individual kernels are estimated by using the gradient descent optimization of the kernel weights with a $l_1$ constraint applied on them. SKAD uses a two-step iterative process to do this: a) the minimization of the radius of the joint hypersphere by obtaining optimal support vectors from a combined kernel matrix that defines the boundary of the hypersphere and b) the gradient descent optimization of the $l_1$-constrained weights of the individual subclassifiers to further minimize the radius of the joint hypersphere. The optimal joint hypersphere defines the combined support of the local background data across the randomly selected spectral subspaces. The joint hypersphere provides a more compact support of the normalcy data than the individual hyperspheres built in the feature subspaces and the hypersphere built in the original input feature space using regular SVDD.

The spectral signatures that lie outside the joint hypersphere are considered outliers. To apply SKAD on hyperspectral image data sets for anomaly detection, for each test pixel of a hyperspectral image, a sliding dual rectangular window [9] is used, where the spectra between the inner and the outer...
windows represent the normalcy data. The size of the inner window is set to enclose the potential objects of interest, and the size of the outer window is determined to include an enough number of spectral signatures to estimate a compact hypersphere surrounding the local normalcy data in the RKHS. The spectral signature of the test pixel from the center of the inner window is tested against the joint optimal hypersphere to find whether it belongs to the background or is an outlier and, hence, a target pixel. Both qualitative and quantitative performance comparisons between SKAD and regular SVDD are provided using hyperspectral image data sets as well as multivariate data sets.

The rest of this letter is organized as follows. In Section II, the concept of SVDD is introduced. Section III describes the SKEL using SVDD anomaly detectors for hyperspectral target detection and explains the implementation issues. Simulation results are presented in Section IV. Finally, Section V concludes this letter with some remarks about the proposed method.

II. SVDD

SVDD, introduced in [6], characterizes the background data set by containing only the relevant data and excluding the superfluous space around the data set. The boundary of the data set is defined by the vectors or samples in the background data, which are called support vectors. The pixels that lie outside this boundary are detected as outliers or anomalies. Consider a data set containing samples represented as \( \{x_i\} \), where \( x_i \in \mathbb{R}^d \) is a \( d \)-dimensional feature vector of each data sample \( i \). Let \( \Phi \) be a function that transforms the input feature vector to a high-dimensional (possibly infinite) RKHS associated with the kernel function \( k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle \). The kernel-based SVDD algorithm tries to find the smallest hypersphere in the RKHS that encloses the given background data set in order to exclude the superfluous space around the background data set as much as possible. This sphere is defined by its center \( a \) and radius \( R \). So, the functional \( R^2 \) is minimized with a constraint that the hypersphere contains all the background data points. If there is a possibility of outliers existing in the background data, then slack variables are used to allow for the outliers. It can be expressed as

\[
\begin{align*}
\min L(a, R) &= R^2 + C \sum_i \xi_i \\
\text{subject to } &\|\Phi(x_i) - a\|^2 \leq R^2 + \xi_i \quad \forall i = 1, 2, \ldots, N, \\
\xi_i &\geq 0 \quad \forall i = 1, 2, \ldots, N, 
\end{align*}
\]  

(1)

where parameter \( C \) controls the tradeoff between the volume of the hypersphere and the errors. Applying the Lagrange multipliers \( \{\alpha_i, i = 1, 2, \ldots, N\} \) and Karush–Kuhn–Tucker (KKT) conditions (see [7] for details), the dual problem can be written as

\[
\begin{align*}
\max L(\alpha_i) &= \sum \alpha_i k(x_i, x_i) - \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) \\
\text{subject to } &0 \leq \alpha_i \leq C \quad \forall i = 1, 2, \ldots, N, \quad \sum \alpha_i = 1. 
\end{align*}
\]  

(2)

This is a convex quadratic programming problem and can be easily solved to obtain the optimal Lagrangian multipliers \( \{\alpha_i^*\} \). The center of the hypersphere (although cannot be determined explicitly) after solving (2) is given by

\[
a = \sum_i \alpha_i^* \Phi(x_i). 
\]  

(3)

The vectors with \( \alpha_i^* = 0 \) lie inside the hypersphere and are considered to be a part of the background data. The vectors with the corresponding Lagrange multipliers \( 0 < \alpha_i^* < C \) are the support vectors that actually lie on the boundary of the hypersphere. The vectors with the corresponding Lagrange multipliers \( \alpha_i^* = C \) are the outliers (still support vectors) that are allowed by the introduction of slack variables. These vectors lie outside the hypersphere. Then, the radius of the hypersphere is given by

\[
R^2 = \frac{1}{N_b} \sum_{k=1}^{N_b} \{ \|\Phi(x_k) - a\|^2 \} \\
= \frac{1}{N_b} \sum_{k=1}^{N_b} \left\{ k(x_k, x_k) - 2 \sum_i \alpha_i^* k(x_k, x_i) + \sum_{i,j} \alpha_i^* \alpha_j^* k(x_i, x_j) \right\} 
\]  

(4)

where \( \Phi(x_k), k = 1, 2, \ldots, N_b \), denotes the support vectors that lie on the boundary of the background data set, and \( N_b \) is the total number of such boundary support vectors.

III. SPARSE KERNEL-BASED SVDD ENSEMBLE LEARNING

Recently, it has been shown that SKEL provides a better generalization performance over a single classifier in the case of binary classification [2], [3]. In SKEL, the final kernel \( k(x_i, x_j) \) is considered as a convex combination of basis kernels \( k_m(x_i, x_j), m = 1, 2, \ldots, M \). Each of the basis kernels \( k_m \) is associated with its corresponding RKHS. Each kernel and RKHS is generated from a randomly selected feature subspace of the input data. Each feature subspace is obtained by the random selection of features using a uniform distribution. Mathematically, it can be written as

\[
k(x_i, x_j) = \sum_{m=1}^{M} d_m k_m(P^m x_i, P^m x_j) \\
\text{subject to } d_m \geq 0 \quad \forall m = 1, 2, \ldots, M, \\
\sum_{m=1}^{M} d_m = 1 
\]  

(5)

where \( M \) is the total number of kernels and \( d_m \) denotes the weights of the individual kernels. \( P^m \) is an \( F_m \times d \) projection matrix which defines the input features randomly selected to form the \( m \)th subspace. It is defined by the elements \( P_{f,n}^m = 1 \) if the \( n \)th input feature is selected as the \( f \)th feature of the \( m \)th subspace and zero, otherwise, for all \( n = 1, 2, \ldots, d \) and \( f = 1, 2, \ldots, F_m \). The \( l_1 \) norm constraint is applied on the weights of the basis kernels to promote sparsity among them and to select only the best feature subsets that help in improving the generalization performance of the classifier. In this letter, we apply the same principle to SKAD using SVDD-based one-class classifiers as the subclassifiers. For this purpose, the primal SVDD problem from (1) can be rewritten as

\[
\begin{align*}
\min L(a, R) &= R^2 + C \sum_i \xi_i \\
\text{subject to } &\|a\|^2 - 2 \langle a, \Phi(x_i) \rangle \leq R^2 - 1 + \xi_i \\
\xi_i &\geq 0 \quad \forall i = 1, 2, \ldots, N. 
\end{align*}
\]  

(6)
Following the framework of the MKL presented in [5], the primal SVDD-based ensemble problem can be written as

\[
\begin{align*}
\min_{\mathbf{d}} \quad & L(\mathbf{d}, R) = R^2 + C \sum \xi_i \\
\text{subject to} \quad & \sum_{m=1}^{M} \frac{1}{d_m} \|\mathbf{a}_m\|^2 - 2 \left( \sum_{m=1}^{M} \mathbf{a}_m, \Phi_m(\mathbf{P}^m \mathbf{x}_i) \right) \\
& \leq R^2 + 1 - \xi_i \quad \forall i = 1, 2, \ldots, N, \\
& \xi_i \geq 0 \quad \forall i = 1, 2, \ldots, N, \\
& d_m \geq 0 \quad \forall m = 1, 2, \ldots, M, \\
& \sum_{m=1}^{M} d_m = 1
\end{align*}
\]

(7)

where each \(d_m\) controls the effect of the hypersphere in the RKHS associated with each kernel \(k_m\) on the joint hypersphere in the RKHS associated with the combined kernel \(k\). \(\Phi_m\) is the nonlinear function that maps the input feature vectors into the RKHS associated with the kernel \(k_m\). One can observe that the \(l_1\) constraint applied on these weights \(d_m\) leads to the selection of the best feature subsets based on the fact that the joint enclosing hypersphere in the RKHS associated with the combined kernel from these feature subsets has a minimum radius. Introducing the Lagrange multipliers \(\alpha_i \geq 0\), \(\gamma_i \geq 0\), \(\eta_m \geq 0\), and \(\lambda \geq 0\), the Lagrangian of (7) is obtained as

\[
L_p = R^2 + C \sum \xi_i + \sum \alpha_i \\
\times \left\{ \sum_{m=1}^{M} \frac{1}{d_m} \|\mathbf{a}_m\|^2 - 2 \sum_{m=1}^{M} \mathbf{a}_m, \Phi_m(\mathbf{P}^m \mathbf{x}_i) \right\} + 1 - R^2 - \xi_i \\
- \sum_i \gamma_i \xi_i + \lambda \left( \sum_{m=1}^{M} d_m - 1 \right)
\]

(8)

After applying the KKT conditions by setting the derivatives of \(L_p\) with respect to \(R^2\), \(\mathbf{a}_m\), \(\xi_i\), and \(d_m\) to zero and substituting these conditions in (8), the dual problem of (7) is obtained as

\[
\begin{align*}
\max_{\alpha} \quad & L_d = 1 - \lambda \\
\text{subject to} \quad & \sum \alpha_i = 1, \ 0 \leq \alpha_i \leq C \quad \forall i = 1, 2, \ldots, N, \\
& \sum \alpha_i \alpha_j k_m(\mathbf{P}^m \mathbf{x}_i, \mathbf{P}^m \mathbf{x}_j) \leq \lambda \quad \forall m = 1, 2, \ldots, M.
\end{align*}
\]

(9)

This dual problem is similar to the dual problem of MKL using the SVM for binary classification, which is explained in [5] and is very difficult to optimize due to the last constraint in (9). Instead, similar to [5], an alternate form of the constrained optimization problem is considered here

\[
\begin{align*}
\min_{\mathbf{d}} \quad & J(d) \\
\text{subject to} \quad & \sum_{m=1}^{M} d_m = 1, \\
& d_m \geq 0 \quad \forall m = 1, 2, \ldots, M.
\end{align*}
\]

(10)

where \(J(d)\) is shown at the bottom of the page. The problem in (11) is the primal regular SVDD problem expressed in (1) and (6). Hence, it can be solved by using a standard SVDD quadratic programming solver. The dual problem for the primal problem expressed in (10) and (11) can be formed by substituting the combined kernel (5) into the dual problem of the standard SVDD (2) as follows:

\[
\begin{align*}
\max_{\alpha} \quad & J(d, \alpha) = \sum \alpha_i \sum_{m=1}^{M} d_m k_m(\mathbf{P}^m \mathbf{x}_i, \mathbf{P}^m \mathbf{x}_j) \\
\text{subject to} \quad & 0 \leq \alpha_i \leq C \quad \forall i = 1, 2, \ldots, N, \sum \alpha_i = 1.
\end{align*}
\]

(12)

Using the combined kernel \(k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{M} d_m k_m(\mathbf{P}^m \mathbf{x}_i, \mathbf{P}^m \mathbf{x}_j)\), (13) can be solved as a standard SVDD problem. Once the optimal Lagrange multipliers \(\alpha_i^*\) are obtained, the objective value of (13) is going to be

\[
J(d) = 1 - \sum \alpha_i^* \sum_{m=1}^{M} d_m k_m(\mathbf{P}^m \mathbf{x}_i, \mathbf{P}^m \mathbf{x}_j).
\]

(14)

This is solved using the gradient descent algorithm, where the gradient of \(J(d)\) with respect to each weight \(d_m\) is given by

\[
\frac{\partial J}{\partial d_m} = - \sum \alpha_i^* \alpha_j^* k_m(\mathbf{P}^m \mathbf{x}_i, \mathbf{P}^m \mathbf{x}_j) \quad \forall m.
\]

(15)

Once the gradient of \(J(d)\) is computed, \(d\) is updated by using a descent direction calculated using a reduced gradient method described in [5]. The reduced gradient method ensures that the weights \(d\) are updated in such a way that the equality constraint and the nonnegativity constraints on \(d\) [shown in (10)] are satisfied. Optimal weights of the subclassifiers \(d_m^*\) are obtained when the gradient descent algorithm converges, and the optimal Lagrange multipliers \(\alpha_i^*\) are obtained for the final combined kernel-based SVDD. Now, the radius of the joint hypersphere is given by

\[
R^2 = \frac{1}{N_b} \sum_{k=1}^{N_b} \left\{ 1 - \sum \alpha_i^* \sum_{m}^{} d_m^* k_m(\mathbf{P}^m \mathbf{x}_k, \mathbf{P}^m \mathbf{x}_i) \\
+ \sum \alpha_i^* \alpha_j^* \sum_{m}^{} d_m^* k_m(\mathbf{P}^m \mathbf{x}_i, \mathbf{P}^m \mathbf{x}_j) \right\}
\]

(16)

where \(\Phi(\mathbf{x}_k), \ k = 1, 2, \ldots, N_b\), denotes the support vectors that lie on the boundary of the background data set in the
combined RKHS, and \( N_0 \) is the total number of such boundary support vectors.

### A. Algorithm Implementation

The algorithm is initiated by selecting random feature subspaces from the input data to form weak classifiers. The weights of all the weak classifiers or kernels are set to the same value, i.e., \( 1/M \). The final kernel is obtained by combining all the weighted individual kernels. Then, the optimization problem in (13) is solved to obtain the best solution of \( \alpha^*_i \). They are plugged into (15) to be used in the gradient descent algorithm for updating the weights of the individual kernels. These two steps are continued until the algorithm convergence criterion is met. In this letter, the total change in the weights is used as the convergence criterion. The algorithm terminates when the change in the weights of the individual kernels is below a certain threshold. The final optimized sparse weights are used to combine the kernels and obtain the hypersphere in the RKHS associated with the combined kernel as shown in (16).

In this letter, the dual-window technique described in [9] and [7] is used to obtain the local background information for every pixel in the hyperspectral image. The center of the inner window represents the test pixel \( x_T \). The pixels in between the inner window and the outer window constitute the background data set for that particular test pixel. The test statistic that can be used to determine if the test pixel is an anomaly or not using SKAD is given by

\[
F_{SKAD}(x_T) = 1 - 2 \sum \alpha^*_i \sum_m d^* m k_m(P^m x_T, P^m x_i) + \sum \alpha^*_i \alpha^*_j \sum_m d^* m k_m(P^m x_i, P^m x_j) \geq R^2. \quad (17)
\]

This statistic provides a measure of similarity between the test pixel and its background. Since the radii of the hyperspheres enclosing the local background data sets at different test pixels in a hyperspectral image are not equal, a ROC curve cannot be generated by directly applying varying thresholds on the statistics in (17). So, we use the normalized SKAD statistic (similar to [8]) to generate ROC curves and quantify the performance of the detector. The normalized SKAD test statistic is expressed in (18), shown at the bottom of the page.

### IV. Simulation Results

#### A. Hyperspectral Image Data

In this section, we apply the kernel-RX (KRX) [9], regular SVDD, and SKAD algorithms on two hyperspectral digital imagery collection experiment (HYDICE) images—the Forest Radiance I (FR-I) and the Desert Radiance II (DR-II) images.

The sizes of the inner window and the outer window used for the dual-window technique in both the test images are \( 5 \times 5 \) and \( 15 \times 15 \), respectively. The size of the inner window was set to enclose the targets whose approximate size was determined based on a prior knowledge of the range of the scene and field of view of the hyperspectral sensor. The outer window size was determined to provide an enough number of spectral samples to model the support of the local background. The value of \( C \) for SVDD and SKAD algorithms is set to make the lower bound on the number of support vectors or the upper bound on the number of outliers allowed to be 20% of the total number of local background pixels being used from the dual-window technique [10]. The Gaussian RBF kernel parameter \( \sigma \) for SVDD and SKAD is determined by implementing the approximate minimax technique on the randomly selected ten regions of the image to represent the background as done in [8]. The same value is used over all the test pixels in the image. The kernel parameter for KRX algorithm from [9] is used in this letter as well. Each test pixel is then tested if it falls outside or inside of the hypersphere associated with the local background region to determine if it is an anomaly or a background pixel. Since there is randomness involved in the selection of the feature subspaces for each of the kernels, we have run the SKAD algorithm five times and plotted error bars on the ROC curves for all the data sets. However, each run generates different values of probability of detection and false alarm rate. In order to generate the error bars, all the ROC curves are interpolated onto the same scale of false alarm rate.

Fig. 1(a) and (b) show the anomaly detection results of KRX, regular SVDD, and the proposed sparse kernel-based SVDD (SKAD) on the FR-I image and the DR-II image. From Fig. 1(a) and (b), one can see that the performance of SKAD is generally better than that of regular SVDD and KRX over the entire range of false alarms for both the images FR-I and DR-II. The dimensionality of the feature subspaces is set to 50 for both the images. The dimensionality of 50 was determined based on the performance analysis of SKEL, which shares the similar ensemble concept with SKAD. It turned out that this number provided a near-optimal performance for SKEL. SKAD has been applied on FR-I and DR-II images with different dimensionalities of feature subspaces, and the ROC curves are plotted in Fig. 2. It can be observed that SKAD with any dimensionality of the subspace outperforms SVDD in terms of anomaly detection. The algorithm is initialized with 100 kernels. The average number of kernels that is selected (with nonzero weights) after running the SKAD algorithm on the DR-II images is 72. The average number of nonzero weight kernels selected on the FR-I image is 71.

\[
F_{SKADN}(x_T) = \frac{1 - 2 \sum \alpha^*_i \sum_m d^* m k_m(P^m x_T, P^m x_i) + \sum \alpha^*_i \alpha^*_j \sum_m d^* m k_m(P^m x_i, P^m x_j)}{R^2} \quad (18)
\]
In this letter, a SKAD has been proposed for hyperspectral images as well as some multivariate data sets. SKAD is premised on the principle of ensemble learning that an optimal combination of independently estimated subdecision models for the respective feature subspaces can significantly improve the prediction performance of a multivariate vector for detection/classification. In this method, the weights of individual SVDD-based one-class classifiers built in randomly selected feature subspaces are optimized by minimizing the radius of the hypersphere enclosing the normalcy/background data in the joint RKHS. Moreover, the weights are sparse in nature due to the $l_1$ constraint applied on them. The sparsity of the weights allows the algorithm to choose the best subsets of features that generate a joint RKHS, in which the enclosing hypersphere has the most compact support.

It is well known that ensemble-learning techniques have to use much more computational resources than regular algorithms to perform multiple learning as opposed to single learning, and hence, computational time required for SKAD is much longer than that required for SVDD. However, this is true only in the case of hyperspectral images because of the dual-sliding-window technique used. In the case of multivariate data sets, the speed of SKAD is comparable to that of SVDD. The future work involves reducing the computational complexity by optimizing the implementation of the dual-sliding-window SVDD used in the proposed work, which eventually leads to a computationally simplified SKAD.

### V. Conclusion

The KRX, regular SVDD, and SKAD algorithms are also compared based on their performance on multivariate data sets. In this letter, two multivariate one-class data sets from [11] have been used. The first data set is the sonar mines data set, in which the sonar signals that bounced off of cylindrical metal mines represent the normalcy and the sonar signals that bounced off of roughly cylindrical stones represent the outliers. The second data set is the ionosphere data set which includes good radar returns with some structure as the normalcy data and the returns with no evidence of structure as the outliers. For both the data sets, 50% of the normalcy data points are randomly selected and used as the training set of the normalcy class. The other 50% of the normalcy data points along with the outliers are used in test sets to generate the ROC curves presented in Fig. 3(a) and (b). The initial number of kernels used is 50 for both the data sets. The sonar mines data set has 60 features, and the dimensionality of the randomly selected feature subspaces is three. The ionosphere data set has 34 features, and 5-D feature subspaces have been used. The dimensionality of the feature subspaces for the two data sets was determined by choosing the one that provides the best improvement in the detection performance. The value of $C$ is set to make the lower bound on the number of support vectors or the upper bound on the number of outliers allowed to be 10% of the total number of training samples. The final number of kernels selected after the optimization for the two data sets are 42% and 12%, respectively. One can, again, observe from the ROC curves that SKAD provides a significantly improved probability of detection at a low false alarm rate over regular SVDD and KRX for sonar mines data set and over KRX for ionosphere data set. The confidence intervals of the SKAD ROC curves for the multivariate data sets are larger than those for the hyperspectral data sets due to the small number of samples in the test set.

### References