CESVM: Centered Hyperellipsoidal Support Vector Machine Based Anomaly Detection

Sutharshan Rajasegarar¹, Christopher Leckie², Marimuthu Palaniswami¹

¹ARC Special Research Center for Ultra-Broadband Information Networks (CUBIN) Department of Electrical and Electronic Engineering
²NICTA Victoria Research Laboratory, Department of Computer Science and Software Engineering The University of Melbourne, Australia

Email: {r.sutharshan, swami}@ee.unimelb.edu.au, caleckie@csse.unimelb.edu.au

Abstract—A challenge in using machine learning for tasks such as network intrusion detection and fault diagnosis is the difficulty in obtaining clean data for training in order to model the normal behavior of the system. Unsupervised anomaly detection techniques such as one class support vector machines (SVMs) have been introduced to overcome this difficulty. One class support vector machines model the normal or target data using non-linear surfaces in the input space while ignoring the anomalous data. Our approach to this problem is based on fitting a hyperellipsoid with a minimal effective radius, centered at the origin, around a majority of the data vectors in a higher dimensional space. We formulate this as a linear optimisation problem, which is advantageous in terms of its computational complexity. We demonstrate using real data from the Great Duck Island Project that our approach achieves better detection performance and flexibility in terms of parameter selection, compared to an earlier detection scheme using a quarter sphere SVM.

I. INTRODUCTION

Network management tasks, such as intrusion detection and fault diagnosis, are increasingly making use of data mining and machine learning techniques to detect abnormal activities in networks. This involves building a model of normal behavior from the data collected from the network. However in order to build such a normal model, machine learning techniques require cleaned data or labeled data for training. One of the challenges in these tasks is how to obtain such error free or labeled data for training. This requires techniques that can model the normal profile of the monitored data while automatically ignoring the errors or anomalies present in the data. Unsupervised anomaly detection techniques such as one class support vector machines (SVMs) aim to address this challenge. An anomaly or outlier in a set of data is defined as an observation that appears to be inconsistent with the remainder of the data set [1]. One class support vector machines model normal behavior using a non-linear surface, so that anomalies can be identified as those data points that deviate from the normal model.

Recently, several one-class SVM approaches have been proposed for anomaly detection. The general approach of these one-class SVMs is to first map the data vectors (measurements) from the input space (the space where the measurements are collected) to a higher dimensional space called the feature space by means of a non-linear function. The mapped vectors in the feature space are called the image vectors. Then a smooth surface or boundary is found in the feature space that separates the image vectors into normal and anomalous measurements. By using Mercer kernels [2], the data vectors can be implicitly mapped to a higher dimensional inner product space. The use of Mercer kernels allows the use of data vectors in the input space for performing inner product computations in the feature space without the knowledge of the mapping function. As a result, the boundaries found in the feature space will yield a non-linear boundary in the input space [3]. Below we discuss several examples of such one-class SVM formulations in detail.

Scholkopf et al. [4] have proposed a hyperplane-based one-class SVM for anomaly detection. In this approach, image vectors in the feature space are separated from the origin by a hyperplane with the largest possible margin. The vectors that fall towards the origin of the hyperplane are identified as anomalous. This scheme involves a quadratic optimisation process to arrive at the hyperplane. Campbell et al. [5] have formulated a linear programming approach for the above hyperplane based one-class SVM when used with a radial basis function (Rbf) kernel. This formulation is based on attracting the hyperplane towards the average of the image data distribution, rather than minimising the maximum norm distance to the bounding hyperplane from the origin as in [4].

Tax et al. [6] have formulated the one-class SVM using a hypersphere. In this approach, a minimal radius hypersphere is fixed around the majority of the image vectors in the feature space. The data that falls outside the hypersphere are identified as anomalous. This hypersphere formulation involves solving a quadratic programming optimisation. Wang et al. [7] have formulated the one-class SVM using hyperellipsoids with minimum effective radii around a majority of the image vectors in the feature space. This hyperellipsoidal formulation involves two phases. First the image vectors are partitioned into a number of distinct clusters using Ward’s linkage algorithm [8]. Second, the image vectors in each cluster are fixed with a hyperellipsoid that encapsulates a majority of the image vectors in that cluster. The image vectors that do not fall within any of the hyperellipsoids are identified as anomalous. This
problem is formulated as a second order cone programming optimisation problem.

Laskov et al. [9] have observed that the “typical distribution of features used in intrusion detection systems (IDS) is one-sided on $\mathbb{R}^+_0$. Further, the data with numerical attributes can be normalised to zero mean and unit variance [9]. This results in a greater mass of the data lying in the vicinity of the origin. A geometric construction that takes into account this one-sidedness of the data distribution can be obtained by extending the hypersphere-based one-class SVM approach. Laskov et al. have extended such an approach into a special type of SVM called a one-class quarter-sphere SVM, which finds a minimal radius hypersphere centered at the origin, encompassing a majority of the image vectors in the feature space [9]. Data vectors that fall outside the quarter sphere are classified as anomalies. This problem is formulated as a linear programming problem.

In this paper we use the hyperellipsoidal SVM of Wang et al. [7] to extend this quarter-sphere SVM scheme (QSSVM) proposed by Laskov et al. [9] into a centered-hyperellipsoidal SVM based anomaly detection scheme (CESVM). We formulate this problem as a linear optimisation problem. Below we give the mathematical formulation of the problem in detail.

II. CENTERED HYPERELLIPSOIDAL SVM BASED ANOMALY DETECTION (CESVM)

Consider a data set $X = \{x_i : i = 1 \ldots n\}$ of $d$ variate data vectors $x_i = (x^1, x^2, \ldots, x^d)$, $i = 1, 2, \ldots, n$ in the input space, where $n$ is the number of data vectors. Data vectors $X$ are mapped to a feature space via a non linear function $\phi(.) : \mathbb{R}^d \rightarrow \mathbb{R}^p$, resulting in image vectors $X = \{\phi(x_i) : i = 1 \ldots n\}$, where $p$ is the dimension of the feature space. Our aim is to fit a hyperellipsoid in the feature space with minimum effective radius $R$ ($> 0$), centered at origin, encompassing a majority of the image vectors $X$. This problem can be formulated as an optimisation problem, similar to the structured one-class classifier formulation given in [7], as follows:

$$\min_{R \in \mathbb{R}, \xi_i \in \mathbb{R}^n} R^2 + \frac{1}{\nu n} \sum_{i=1}^{n} \xi_i$$

subject to:

$$\phi(x_i)\Sigma^{-1}\phi(x_i)^T \leq R^2 + \xi_i, \quad \xi_i \geq 0, \quad i = 1 \ldots n$$

(1)

where $\{\xi_i : i = 1 \ldots n\}$ are the slack variables that allow some of the data vectors to lie outside the hyperellipsoid. The number of data vectors is denoted by $n$. In contrast to [7], we use the parameter $\nu \in (0, 1)$, as the regularisation parameter that controls the fraction of data vectors that lie outside the hyperellipsoid, i.e., the fraction of data vectors that can be outliers or anomalies. This parameter is similar to the one proposed for the $\nu$-Support Vector Classifier ($\nu$-SVC) in [3], which gives more intuition behind the selection of a value for that parameter $\nu$. Note that $\Sigma^{-1}$ is the inverse of the covariance matrix $\Sigma$ of the image vectors $X$, which is given as follows:

$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} (\phi(x_i) - \mu)(\phi(x_i) - \mu)^T, \quad \mu = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)$$

(2)

For centered image data $\mu = 0$. Using Mercer kernels, the dot product computations of the image vectors in the feature space can be computed in the input space without any knowledge of the non-linear function $\phi(.)$ [2]. Our aim is to formulate the optimisation problem (1) in terms of a Mercer kernel function $k(y, z) = \phi(y) \phi(z)^T$. Let $K \in \mathbb{R}^{n \times n}$ be the kernel matrix of the data vectors $X$. The image vectors $X$ are centered in the feature space by subtracting the mean $\mu$ of the image vectors from $X$, resulting in a centered kernel matrix $K_c$ [10]. The $K_c$ can be expressed in terms of the kernel matrix $K$ using $K_c = K - I_nK - K_1n + nK_1n$, where $I_n$ is the $n \times n$ matrix with all entries equal to $1/n$.

The eigen decomposition of the symmetric positive semidefinite matrix $K_c$ is given by $K_c = P \Lambda P^T$, where $\Lambda$ is a diagonal matrix with positive eigenvalues as the diagonal elements, and $P$ is the eigenvector matrix corresponding to the positive eigenvalues [11]. Hence the covariance matrix $\Sigma$ can be expressed as $\Sigma = QGQ^T$, where $Q = XT \Lambda^{-\frac{1}{2}}$ and $G = \frac{1}{n} \Lambda$. Refer to Appendix III of [7] for a derivation. Therefore, the pseudo inverse $\Sigma^+ \Sigma^{-1}$ can be expressed [11] as $\Sigma^+ \Sigma^{-1} QG^T = nX^T \Lambda^{-\frac{1}{2}} P^T X$. Further, $\phi(x_i)\Sigma^{-1}\phi(x_i)^T$ can be expressed as follows:

$$\phi(x_i)\Sigma^{-1}\phi(x_i)^T = \phi(x_i)\Sigma^+\phi(x_i)^T = \phi(x_i)nX^T \Lambda^{-\frac{1}{2}} P^T X \phi(x_i)^T$$

$$= \|\sqrt{n}\Lambda^{-\frac{1}{2}} P^T K_c^i\|^2$$

(3)

where $K_c^i$ is the $i^{th}$ column of the kernel matrix $K_c$. Hence, the optimisation problem (1) can now be expressed as follows:

$$\min_{R \in \mathbb{R}, \xi_i \in \mathbb{R}^n} R^2 + \frac{1}{\nu n} \sum_{i=1}^{n} \xi_i$$

subject to:

$$\|\sqrt{n}\Lambda^{-\frac{1}{2}} P^T K_c^i\|^2 \leq R^2 + \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, \ldots, n$$

(4)

We form the dual for this primal optimisation problem (4) using the Lagrange technique [3]. The Lagrange function for
this optimisation will be as follows:

\[
L(R^2, \xi_i, \alpha_i, \gamma_i) = R^2 + \frac{1}{\nu n} \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \gamma_i \xi_i
- \sum_{i=1}^{n} \alpha_i (R^2 - \|\sqrt{\Lambda}^{-1} P^T K_c \|_2^2 + \xi_i)
\]

(5)

where \(\alpha_i \geq 0, \gamma_i \geq 0, \forall i\) are the Lagrange multipliers. Equating the partial derivatives of \(L\) with respect to \(R^2\) and \(\xi_i\) to zero yields:

\[
\frac{\partial L}{\partial R^2} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i = 1
\]

(6)

\[
\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \gamma_i = \frac{1}{\nu n} - \alpha_i
\]

(7)

From (7) and using \(\alpha_i \geq 0, \gamma_i \geq 0\), we can obtain \(0 \leq \alpha_i \leq \frac{1}{\nu n}\). Substituting (6) and (7) in (5) results in \(L = \sum_{i=1}^{n} \alpha_i \|\sqrt{\Lambda}^{-1} P^T K_c \|_2^2\). Therefore, the dual problem is as follows:

\[
\min_{\alpha \in \mathbb{R}^n} - \sum_{i=1}^{n} \alpha_i \|\sqrt{\Lambda}^{-1} P^T K_c \|_2^2
\]

subject to:

\[
\sum_{i=1}^{n} \alpha_i = 1, \quad 0 \leq \alpha_i \leq \frac{1}{\nu n} \quad (8)
\]

where \(i = 1...n\). This dual problem (8) is a linear optimisation problem. The \(\{\alpha_i\}\) can be obtained using widely available linear optimisation techniques such as the simplex method or interior point method [12], [13]. This linear optimisation formulation is advantageous in terms of its computational complexity.

From the solution of (8) for \(\{\alpha_i\}\), the data vectors can be classified as follows (refer to Figure 1(a)). The data vectors with \(\alpha_i = 0\) will fall inside the hypersphere. The data vectors with \(\alpha_i > 0\) are called the support vectors. Support vectors with \(\alpha_i = \frac{1}{\nu n}\) are termed as outliers, which fall outside the hypersphere. Support vectors with \(0 < \alpha_i < \frac{1}{\nu n}\) will reside on the surface of the hypersphere, and hence are called the border support vectors. Moreover, the effective radius of the hypersphere \(R\) can be obtained using \(R = \|\sqrt{\Lambda}^{-1} P^T K_c \|_2\) for any border support vector \(x_b \in X\).

### III. QUARTER SPHERE SVM BASED ANOMALY DETECTION (QSSVM)

The hyperrhosphoidal formulation in (1) is the general form of the hypersphere based scheme used in [6], [9], [14]. To obtain the quarter sphere based scheme from (1), the covariance matrix \(\Sigma\) can be replaced by a unit matrix \(I\), where \(I\) is a diagonal matrix with all diagonal elements set to 1. Hence, the problem formulation of finding a hypersphere of minimal radius \(R > 0\), centered at the origin, encompassing a majority of the image vectors in the feature space will become as follows:

\[
\min_{R \in \mathbb{R}, \xi \in \mathbb{R}^n} R^2 + \frac{1}{\nu n} \sum_{i=1}^{n} \xi_i
\]

subject to:

\[
k(x_i, x_i) \leq R^2 + \xi_i, \quad \xi_i \geq 0
\]

(9)

where \(i = 1...n\). The dual formulation of (9) will become

\[
\min_{\alpha \in \mathbb{R}^n} - \sum_{i=1}^{n} \alpha_i k(x_i, x_i)
\]

subject to:

\[
\sum_{i=1}^{n} \alpha_i = 1, \quad 0 \leq \alpha_i \leq \frac{1}{\nu n}
\]

(10)

where \(i = 1...n\). The solution for this linear programming problem and the classification of data vectors will be similar to the hyperrhosphoidal case explained above in Section III (refer to Figure 1(b)). In the following section, we compare the performance of this hypersphere based scheme with our hyperrhosphoidal based scheme.

### IV. EVALUATION

The purpose of the evaluation is to analyse and compare the performance of the proposed centered hyperrhosphoidal SVM based anomaly detection scheme (CESVM) with the quarter sphere SVM based anomaly detection scheme (QSSVM). Our evaluation data are a set of real sensor measurements gathered from a deployment of wireless sensors in the Great Duck Island, Maine, USA during 2003 [15]. The network was deployed for monitoring the habitat of a sea bird called the Leach’s Storm Petrel [15]. They recorded light, temperature, pressure and humidity measurements at 5 minutes intervals.

We consider a data set formed from the combined data measurements collected by seven sensor nodes, namely the nodes 101, 109, 111, 116, 118, 122 and 123. A 24 hour period of data recorded on 1st July 2003 was used in our evaluation. We used three attributes: humidity, temperature and pressure measurements for each data vector. The data is cleaned manually by removing erroneous and spurious measurements with the help of scatter plots. The cleaned data is labeled as Normal for use in our evaluation. A randomly generated set of anomalous data was introduced into the above normal data measurements. The anomalous data for each attribute comprised a set of data vectors drawn from a uniform distribution over the normal measurements of each attribute. The number of introduced anomalous data are 10% of the normal measurements and were labeled as Anomalies. The data is normalised to the range \([0,1]\) in order to alleviate the effect of different ranges of values for attributes causing a bias in kernel computations.

The CESVM and QSSVM algorithms are implemented in Matlab. We considered three kernel functions in our evaluations: (1) Rbf: A distance based radial basis function kernel \(k_{Rbf} = \exp(-\|y - z\|^2 / \sigma^2)\), where \(\sigma\) is the width parameter of the kernel function and \(\{y, z\}\) are the data vectors; (2) Poly: A polynomial kernel function \(k_{Poly} = (y.z + 1)^r\), where \(r\) is the degree of the polynomial; (3) Linear: A linear kernel function \(k_{Linear} = (y.z)\). Kernel matrices obtained using the above kernel functions are centered and used in our evaluations.

In each simulation, we recorded the false positives, which occur when a normal measurement is identified as anomalous by the detector, and the true positives, which occur when an actual anomalous measurement is correctly identified by the
detector. The false positive rate (FPR) is computed as the percentage ratio between the false positives and the actual normal measurements, and the detection rate (DR) is computed as the percentage ratio between the true positives and the actual anomalous measurements. For comparison of the CESVM and QSSVM schemes, Receiver Operating Characteristic (ROC) curves are obtained for each anomaly detection scheme, i.e., DR vs FPR plots. The larger the area under the ROC curve, the better the performance of the corresponding classifier. The effective radius \( R \) for the CESVM (and radius \( r \) for QSSVM) is computed either using any border support vectors or using the mean of the support vectors should such border support vectors not exist.

We have examined the effect of two parameters for the two anomaly detection schemes using the \( Rbf, Poly \) and \( Linear \) kernels. The regularisation parameter \( \nu \) is varied in the range from 0.01 to 0.25 in intervals of 0.01. The kernel parameters are varied as follows depending on the kernel function used. For the \( Rbf \) kernel, \( kpm = \sigma \) is varied in the range from 0.0025 to 1310.72 in exponential intervals of 2, and for the \( Poly \) kernel \( kpm = r \) is varied in the range from 1 to 20 in intervals of 1.

Figure 2 shows the ROC curves obtained for the CESVM and QSSVM schemes using a \( Linear \) kernel function. The CESVM scheme provides better detection performance compared to the QSSVM scheme. The reason lies in the difference in the geometry of the boundary produced by the two detection schemes. CESVM produces a hyperellipsoidal boundary around the data in the feature space, while QSSVM produces a hypersphere boundary. The volume of space occupied by the hyperellipsoid in CESVM is smaller than the volume occupied by the hypersphere in QSSVM. Therefore, more of the uniformly distributed anomalous data will be captured inside the hypersphere than the hyperellipsoid, resulting in the reduced performance for QSSVM than CESVM. This behavior is also observed in ROC curves obtained using polynomial kernel functions (refer to Figures 3(c) and 3(d)) and \( Rbf \) kernel function (refer to Figures 3(a) and 3(b)).

Figures 3(a) and 3(b) shows the ROC curves obtained for the CESVM and QSSVM schemes using the \( Rbf \) kernel respectively. A selected set of curves are shown in the figures for clarity. For kernel parameter values \( kpm < 0.64 \), in the case of the CESVM, and \( kpm < 0.04 \), in the case of the QSSVM, the detection rate and the false positive rates are observed to be very low (these are not shown in the plots). This is because for lower values of \( kpm \), a ragged boundary around the data is produced in the input space, causing the anomalous data to be detected as normal. Further, for larger values of \( kpm \), the detection performance again starts to diminish. This can be observed for \( kpm > 1311 \) in the case of the CESVM and for \( kpm > 0.64 \) in the case of the QSSVM (refer to Figure 3(b)). This is because as the \( kpm \) increases, the effective radius of the hyperellipsoid (and radius of the hypersphere) become larger. This causes more anomalous data to fall inside the hyperellipsoid (or hypersphere), and hence reduces the detection rate. This shows there exists a window of kernel parameter values \( kpm \) within which the CESVM and QSSVM schemes provide higher detection accuracy. The selection of appropriate parameter settings is application dependent, and can be done using training data and cross-validation [16].

Furthermore, Figure 3(a) shows that the CESVM performs better for a wide range of kernel parameter \( kpm \) values, while QSSVM shows greater sensitivity to the kernel parameter values (refer to Figure 3(b)). For example, a detection rate in excess of 80% and a false positive rate below 10% is attainable for CESVM as \( kpm \) varies from 0.64 to 1311. In contrast, QSSVM can only achieve this accuracy over the \( kpm \) range of 0.04 to 0.32. This shows that the CESVM is less sensitive to the choice of parameter values than the QSSVM.

Figures 3(c) and 3(d) compare the ROC curves obtained using the polynomial kernel function for the CESVM and QSSVM schemes. These plots demonstrate there is a window of kernel parameters (\( kpm = r \)) where each method achieves its best performance. Overall, CESVM achieves better detection accuracy than QSSVM, while being less sensitive to the choice of kernel parameter. When compared with the plots obtained using the \( Rbf \) kernel function (refer to Figures 3(a) and 3(b)), the use of the polynomial function shows lower detection performance. These results are consistent with the observations made by Tax et al. [6]. The performance of the polynomial kernel is influenced by the norms of the data vectors, while the distance based kernels, such as the \( Rbf \) kernel, are independent of the norms of the vectors. Larger norms result in a larger effective radius (or radius in the case of QSSVM), and hence yield a lower detection rate.

\[ V. \text{ CONCLUSION} \]

In this paper, we have proposed a centered hyperellipsoidal support vector machine based anomaly detection scheme (CESVM). Here the data vectors in the input space are first mapped to a higher dimensional space and centered by means of kernel functions. Then, a hyperellipsoid, centered at the origin, with minimal effective radius is fitted around the majority of the data vectors in the higher dimensional space. The vectors that fall on or inside the hyperellipsoid are normal and those falling outside the hyperellipsoid are classified.
as anomalous. We formulate this as a linear optimisation problem, which is advantageous in terms of computational complexity. We evaluated our approach using real data gathered from a sensor network deployment in the Great Duck Island. We compared the performance of our scheme with the quarter sphere support vector machine based anomaly detection scheme and demonstrated that the CESVM scheme achieves better detection accuracy, and greater flexibility in terms of parameter selection.

**ACKNOWLEDGMENT**

We thank the ARC Research Network on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP). This work was supported by the Australian Research Council.

**REFERENCES**