One-Class Support Tensor Machines with Bounded Hinge Loss Function for Anomaly Detection

Imran Razzak, Tariq Mahmood Khan
Department of Information Technology
Deakin University, Geelong, Australia
{imran.razzak, tariq.khan}@deakin.edu.au

***Abstract*—Traditional one class support tensor machine (OCSTM) is a popular classifier that is widely adopted for one class classification, however, outliers in the data negatively affects its performance. To improve the robustness of OCSTM against outliers, in this paper, we present OCSTM with bounded loss function rather than finding optimized support vectors with unbounded loss function. To solve the corresponding optimization problem, we have presented half quadratic optimization to drive the problem to traditional OCSTM, followed by solving a typical OCSTM optimization problem iteratively. We further demonstrate our algorithms through experiments on eight real-world benchmark datasets. Experimental results show that the proposed approach separates well most of the samples of interested class from origin even in the presence of outliers.**

***Index Terms*—Support Tensor Machine, Bounded hinge loss, Outliers, STM, OCSTM, One-Class**

I. Introduction

Rare and inconsistent patterns in data that do not conform to expected behavior are known as anomalies. Anomaly detection is an interesting area of data mining and machine learning where the main task is to identify these instances and reduce their deleterious effect on data interpretation tasks. It has been shown that classification of data consisting of outliers most likely results in incorrect categorization [14]. Anomalies or outliers are rarely occurring events, however, their affect can be significant, i.e. anomalous credit card transactions can indicate stolen credit card or unusual traffic pattern in network may show unauthorized access to network respectively. Anomaly detection has been widely applied in several application domains such as healthcare [29], fault diagnosis [9], [38] intrusion detection [16], [21], industrial damage [20], fraud detection [18], robot behavior [26], sensor networks [32] and astronomical data [31]. Synonymously, it has been termed as anomaly detection, novelty detection, deviation detection, and exception mining.

Traditional one class support vector machine (OCSVM) is widely adopted in one class classification, however, results show that one-class support tensor machine (OCSTM) is negatively influenced by the outliers. To improve the robustness of (OCSVM), many variants have been proposed. These variants can be categorized into two groups, weighting samples and improving the loss function. In the first approach, each sample is assigned a weight that indicates the importance of that sample within the training dataset. Thus, outliers are assigned smaller weights in comparison to reduce their influence [3]. Distance to the center of the samples [37], adaptive weight strategy based on the range of distance [36], weight based on KNN [40] have been used to assign a weight to each sample. In the second category, the focus is on improving the hinge loss function limit the loss due to the outliers. Xiao et al. utilized a non-convex ramp loss function into OCSVM optimization to reduce the effect of outliers [33]. Similar to [33], Yingjie et.al. presented a robust and sparse anomaly detection approach by replacing the hinge loss with non-convex ramp loss function to make a robust and sparse semi-supervised algorithm and used the concave-convex procedure to solve the model that is a non-differentiable non-convex optimization problem [30]. Recently, to improve the robustness of traditional OCSVM against outliers, Xing et.al. replaced the hinge loss with rescaled hinge loss function [34]. Experimental results showed that these methods can effectively decrease the impact of outliers to some extent, however, they are computationally complex.

OCSVM has proven to be an effective classifier for unsupervised anomaly detection. However, the classical anomaly detection methods such as OCSVM or Kernel Density Estimation often fail for high-dimensional data. For example, the data has to be reshaped into vectors which could in-turn destroy the structural information embedded within. A tensor can preserve the high order correlation among modes of the data. The utilization of structural information of the original features are very important for anomaly detection of high dimensional data where strong correlation exits between data points [25]. Furthermore, tensor to vector representation also results in the curse of dimensionality. An alternative is the OCSTM. Since the parameters to be solved in tensor based methods are far less than those in vector-based methods, thus, OCSTM is especially suitable for small sample size problems. Experiment results have shown that the accuracy of OCSTM is superior to that of the traditional OCSVM. However, OCSTM is still sensitive to outliers and impractical for large datasets.

To improve the robustness of OCSTM against outliers and computational efficiency, recently, many researchers

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

TABLE I
NOTATIONS AND THEIR DESCRIPTION

Symbols	Description
x	Lowercase letter represents a scalar
$\boldsymbol{x}$	Boldface lowercase letter represents a vector
$\boldsymbol{X}$	Boldface uppercase letter represents a matrix
$\mathcal{X}$	Calligraphic letter represents a tensor
$\mathcal{R}$	Rank of tensor
y_i	$y_i \in \{1, -1\}$ are the corresponding class labels
$[1 : M]$	Set of integers in the range of 1 to M inclusively
$vec(\cdot)$	Denotes the column stacking operation
$\langle \cdot, \cdot \rangle$	Denotes the inner product of tensors
$\otimes$	Denotes product of tensors
δ	Denotes delta function
$\mathcal{K}(\cdot, \cdot)$	Denotes kernel function

have focused on the improvement of the loss function and kernel methods respectively. He et al. presented a structure-preserving kernel for nonlinear tensor learning by deriving the kernel based on structure-preserving feature mapping [15]. Erfani et al presented a randomized kernel support tensor machine based on nonlinear randomized projection, however, it is sensitive to outliers [11]. Anaissi et al. presented sparse and smooth representations by replacing with ℓ_1 regularized tensor decomposition to overcome the sensitivity of OCSTM against outliers [1]. Yanyan et al. developed Linear Support Tensor Domain Description (LSTDD) based on a linear tensor-based algorithm to find a closed hypersphere with the minimal volume in the tensor space [8]. Traditional support tensor machine is not robust to outliers as unboundedness of the loss function results in larger loss due to outliers and the decision boundary may deviate from the optimal hyperplane [22], [24]. Several non-convex and bounded loss functions have been presented to substitute the hinge loss function in order to suppress the effect of outliers and improve the robustness for support vector machines. It is well known that methods based on tensors are better in term of both computational complexity as well as accuracy [7], [11], [23]. However, according to our knowledge, no work has been done so for on the improvement of one-class tensor machines [23]. Extensive experimental analysis shows that proposed bounded one-class support tensor machines considerably improves the robustness against outliers and significantly reduces the computational complexity as compared to state of the art anomaly detection methods.

The **key contributions** of this work are:

- We present novel support tensor machines with bounded hinge loss which is monotonic, bounded and non-convex, thus robust to outliers by limiting the loss due to outliers.
- To solve the non-convex objective function, we devised an iterative approach using the half quadratic optimization.
- Extensive experimental analysis on ten real-time datasets evident of substantial improvement in anomaly detection.
- We provide theoretical analysis and analyze the robustness of OCSTM-BH against outliers.

II. One-Class STM with Bounded Hinge Loss

While a one-class support tensor machine has been proven an effective approach for anomaly detection, their ability to model large corrupted datasets is limited as the traditional loss function is unbounded which results in larger loss caused by outliers. Thus is not able to efficiently identify anomalies. Bounding the hinge function could in turn help to reduce the effect of outliers. Thus, the aim of this work is to improve the anti-outliers ability and design a robust one class support tensor machine for anomaly detection for corrupted datasets. To overcome the aforementioned challenge for the corrupted dataset, in this section, we present a novel anomaly detection approach by replacing the traditional hinge loss with a bounded loss function. This results in improving the performance against outliers significantly. In the following discussion, we first present support tensor machines with bounded loss function for tensor data followed by algorithm optimization using half quadratic, convergence analysis, computational complexity, and theoretical analysis.

A. Bounding Loss Function

One-class support tensor machines for anomaly detection tries to find an optimal hyperplane in high dimensional data that best separates the data from anomalies with maximum margin. However, the hinge loss of traditional one-class support vector machines is unbounded, which results in larger loss caused by outliers affecting its performance for anomaly detection. Bounding hinge loss function results in significantly overcoming the influence of outliers by reducing the impact of samples that are far from their labels.

$$\min_{\mathcal{W},p,\zeta} \quad \frac{1}{2}||\mathcal{W}||_F^2 + \frac{1}{Nv}\sum_{i=1}^{N}\zeta_i - p \tag{1}$$

To this end, we can modify the optimization problem of OCSTM (given in Eq. 1) as

$$\max_{\mathcal{W},p} J(\mathcal{W},p) = \frac{1}{2}||\mathcal{W}||_F^2 - \frac{1}{vN}\sum_{i=1}^{N}\aleph_i - p \tag{2}$$

$$\text{subject to} \quad \langle \mathcal{W}.\phi(\mathcal{X}_i)\rangle \geq p - \aleph_i$$

$$\aleph_i \geq 0 \quad \forall i = 1, \ldots, N$$

where $\aleph_i = max\{0, p - \mathcal{Z}_i\}$ is the hinge loss function with $\mathcal{Z}_i = \mathcal{W}\phi(\mathcal{X}_i)$.

Notice that the hinge loss in Eq. 2 is unbounded which results in larger loss due to the outliers which in turn affects the performance of anomaly detection. To overcome the aforementioned challenge, we present the following objective function (Eq.3) with bounded loss function (Eq.4).

$$\max_{\mathcal{W},p} J(\mathcal{W},p) = \frac{1}{2}||\mathcal{W}||_F^2 - p + \frac{1}{vN}\sum_{i=1}^{N}\wp_i \tag{3}$$

$$\text{subject to} \quad \langle \mathcal{W}.\phi(\mathcal{X}_i)\rangle \geq p - \wp_i$$

$$\wp_i \geq 0 \quad \forall i = 1, \ldots, N$$

$$\wp_i = \beta[1 - e^{-\eta\aleph_i}] \tag{4}$$

where $\beta = \frac{1}{1-e^{-\eta}}$ is the normalization constant and $\eta \geq 0$ is the scale constant.

The normalization constant β ensures that $\wp_i = 1$. Here, the scale constant η controls the upper bound. For $\eta = 0$ the bounded loss function ($\wp$) degenerates to traditional hinge loss ($\aleph$), thus the traditional hinge loss function (Eq.2) is a special case of bounded loss function (3).

Eq.2 shows that similar to the traditional one-class support tensor machines, the bounded loss function is also monotonic, bounded however non-convex. By simplifying the Eq. 2 and Eq. 3, We can rewrite the objective function as

$$\max_{\mathcal{W},p} J(\mathcal{W}, p) = \frac{\beta}{vN}\sum_{i=1}^{N} e^{-\eta\aleph_i} + p - \frac{1}{2}||\mathcal{W}||_2^2 \tag{5}$$

B. Optimization

As discussed earlier, the objective function in Eq. 5 is non-convex due to non-convexity of hinge loss function, thus traditional optimization can not be applied directly. Eq. 5 can be solved using the half quadratic optimization by defining a convex function as

$$g(u) = -ulog(-u) + u, \quad s.t. \quad u < 0 \tag{6}$$

where $u = [u_1, \ldots u_N]^T \in \mathbb{R}^N$ with its element $u_i < 0$. By applying the conjugate function theory, we get

$$e^{-\eta\aleph} = \sup_{u<0} \eta\aleph u - g(u) \tag{7}$$

We can obtain the supermum of $e^{-\eta\aleph}$ at $u = -e^{-\eta\aleph} < 0$. Now, we can rewrite the Eq. 5 as

$$\max_{\mathcal{W},p} J(\mathcal{W}, p) = \frac{\beta}{vN}\sum_{i=1}^{N} \sup_{u_i<0}\{\eta\aleph_i u_i - g(u_i)\} + p - \frac{1}{2}||\mathcal{W}||_F^2 \tag{8}$$

$$\max_{\mathcal{W},p} J(\mathcal{W}, p) = \frac{\beta}{vN}\sup_{u<0}\Big\{\sum_{i=1}^{N} \eta\aleph_i u_i - g(u_i)\Big\} + p - \frac{1}{2}||W||_F^2 \tag{9}$$

$$\max_{\mathcal{W},p} J(\mathcal{W}, p) = \sup_{u<0}\Big\{\frac{\beta}{vN}\sum_{i=1}^{N} \eta\aleph_i u_i - g(u_i) + p - \frac{1}{2}||W||_F^2\Big\} \tag{10}$$

The above Eq. 9 can be simplified as

$$\max_{\mathcal{W},u,p} J(\mathcal{W}, u, p) = \frac{\beta}{vN}\sum_{i=1}^{N} \eta\aleph_i u_i - g(u_i)\Big\} + p - \frac{1}{2}||\mathcal{W}||_F^2 \tag{11}$$

By using the alternating method to solve the above equation iteratively and compute $\mathcal{W}$, u and p respectively, we get

$$\max_{\mathcal{W},p} J(\mathcal{W}, p) = \frac{\beta}{vN}\sum_{i=1}^{N} \eta\aleph_i u_i + p - \frac{1}{2}||\mathcal{W}||_F^2 \tag{12}$$

We can rewrite the above Eq. 12 as

$$\min_{\mathcal{W},p} J_o(\mathcal{W}, p) = \frac{1}{2}||\mathcal{W}||_F^2 + \frac{\beta}{vN}\sum_{i=1}^{N} \eta\aleph_i u_i - p \tag{13}$$

The above optimization problem in Eq. 13 can be solved by applying Lagrange multiplier as

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j \mathcal{K}(\mathcal{X}_i, \mathcal{X}_j) \tag{14}$$

$s.t.$ $\sum_{i=1}^{N} \alpha_i = 1$ and $0 \geq \alpha_i \leq \frac{1}{vN}s_i$ for $i = 1, \ldots, N$
where $\alpha = [\alpha_1, \ldots, \alpha_N]^T$ is the vector of Lagrange multipliers, k is the kernel matrix.

After solving the dual optimization problem 14, the weight tensor $\mathcal{W}$ can be calculated as

$$\mathcal{W} = \sum_{i=1}^{N} \alpha_i \phi(\mathcal{X}_i) \tag{15}$$

Finally, the decision function is defined as

$$f(x) = sgn\Big(w\phi(x) - p\Big) \tag{16}$$

$$f(x) = sgn\Big(\sum_{i=1}^{N} \alpha_i(x_i, x) - p\Big) \tag{17}$$

The solution to the above quadratic problem in Eq. 17 is characterized by parameter v that sets an upper and lower bound on the fraction of anomalies and the number of training samples used as support vectors respectively.

C. Convergence

Theorem 1: The sequence

$$\max_{\mathcal{W},u,p} J(\mathcal{W}^t, u^t, p^t) = \frac{\beta}{vN}\sum_{i=1}^{N} \eta\aleph_i u_i^t - g(u_i^t)\Big\} + p^t - \frac{1}{2}||\mathcal{W}^t||_F^2$$

converges.

Proof: Comparing Eq.10 and Eq.11, we can conclude that

$$p + \frac{\beta}{v} \geq \sup_{u<0}\Big\{\frac{\beta}{vN}\sum_{i=1}^{N} \eta\aleph_i u_i - g(u_i) + p - \frac{1}{2}||W||_F^2\Big\} \geq \frac{\beta}{vN}\sum_{i=1}^{N} \eta\aleph_i u_i - g(u_i)\Big\} + p - \frac{1}{2}||\mathcal{W}||_F^2 \tag{18}$$

TABLE II
ALGORITHMIC PROCEDURE OF OCSTM-BH

```
Input: : Training dataset: $\mathcal{X}_{i}{}_{i=1}^{N}$ where $X_j \in \mathbb{R}^{m \times n}$ for $j = 1, ..., N$, kernel function $\mathcal{K}(\mathcal{X}_i, \mathcal{X}_j$ trade-off parameter $\tau$, scale constant $\eta$, $T_{max}$

Output: Lagrange multiplier $\alpha$ and margin parameter p,

 Step-I: Parameter Initialization: Auxiliary variable $u \in \mathbb{R}^M$ such that $u_i < 0$, Number of iteration T=0,
While T $\leq T_{max}$ do
 Step-II: Compute $\alpha^{T+1}$ and margin parameter p by solving Eq. 14,
 Step-III: Compute $u^{T+1} = -e^{-\eta\aleph}$.
 Step-IV: Increment T by 1 and repeat the step II-III until converges.
end while
 Step-VI: Return $\alpha$ and p
```

Thus, we can say Eq.11 is upper bounded.
Furthermore, we can deduce that,

$$\begin{aligned} J_0(\mathcal{W}^t, u^t, p^t) &\leq J_0(\mathcal{W}^{t+1}, u^{t+1}, p^t) \\ &\leq J_0(\mathcal{W}^{t+1}, u^{t+1}, p^{t+1}) \quad (19) \end{aligned}$$

Therefore, $J(\mathcal{W}^t, u^t, p^t)$ is non decreasing for $t = 1, 2, 3, ...$ We can conclude that the objective function in Eq. 11 $\max_{\mathcal{W},u,p} J(\mathcal{W}^t, u^t, p^t)$ converges.

D. Theoretical Analysis

In this section, we provide the analysis and reason for robustness of proposed OCSTM-BH. Similar to [12], we have investigated the robustness from weight viewpoint. Consider $max\{0, p - \mathcal{W}^T\phi(\mathcal{X}_i\} = [p - \mathcal{W}^T\phi(\mathcal{X}_i)]_+ = [p - \mathcal{Z}_i]_+$ can be written as $u_i^t = -e^{\eta[p^t - \mathcal{Z}_i^t}$ for $i = 1, \ldots, N$. The weight $\mathcal{S}_i^t$ for t^{th} iteration can be computed as $\mathcal{S}_{t = \beta\eta(-u_i^t)}$. The decision function for OCSTM-BH can be written as $f(x) = \mathcal{W}^T\phi(\mathcal{X} - p = \sum_{i-1}^{N}\mathcal{K}(X_i, y) - p$. For incorrect classification, the value of decision function is $f(x) < 0$ and for correct classification with its true label the value of decision function is +1. For non target samples with its true label, the value -1, the decision function value is $f(x) > 0$. Thus, we can say, larger the values of $f(x)$, the value could be outliers. As we know that the value of $\mathcal{S}_i^t$ ($\mathcal{S}_i^t = \beta\eta(-u_i^t)$) decreases with an increase in $f(x)$ for incorrect classification. Thus, we can conclude that the bounded hinge loss reduces the impact of samples that are far from their labels.

III. EXPERIMENTS AND ANALYSIS

In this section, we investigate the performance of proposed support tensor machines in multiple experiments on eight real-world datasets downloaded from UCI machine learning repository. We have compared the performance of OCSTM-BH with state of the art tensor methods (such as OCSTM [7], R1STM [11], LOCSTM [7]), vector methods (OCSVM [5], LOCSVM [6], R1SVM [10]) and deep one class classification methods (One-Class Deep SVDD [27], Soft Bounded Deep SVDD (SB Deep SVDD) [27]). Since, RBF kernel showed better performance in comparison to other kernels [19], thus in this work we have used RBF kernel function. The accuracy and AUC (area under curve) are the evaluation measure commonly used for one class classifier [7], thus in this work we have considered both test accuracy and AUC as evaluation metrics.

To validate the robustness of OCSTM-BH against outliers, we have contaminated the datasets with anomalies by introducing label noise. In order to estimate the effectiveness of bounded loss function over traditional loss function in the construction of the projection matrices, we repeated our cross-validation experiments ten times for all datasets. We have randomly selected 30% of the training data to form a validation dataset that we have used to tune parameters. The size of the training dataset is very important for efficient anomaly detection, thus, we have performed several experiments with the variable size of the training data.

A. Dataset

In order to validate the performance of the proposed approach, in this paper, we have conducted several experiments on eight different types of dataset (healthcare, handwritten text and face). We have download publicly available datasets from UCI machine learning repository that are Breast Cancer [28], SONAR [13], USPS [35], Daily and Sport Activity (DSA) [2], University of Southern California Human Activity Dataset (USC-HAD) [39], ORL and MNIST [17] . Table III describes the detail of datasets. As some of the above datasets are originally vector based thus, we have transformed these datasets to tensor form. For the vector dataset, we have generated tensor data by transforming the vector data [4] and select the tensor size based on [7].

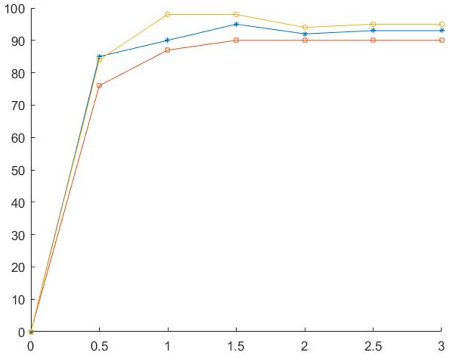

Fig. 1. Performance of OCSTM-BH on MNIST Subject (4,5,6) with different value of bounding factor η

B. Parameter Setting

We performed several experiments with different values of parameters. There are four parameters (scale constant η , width

TABLE III
DESCRIPTION OF DATASETS

		Tensor Representation		
	Dataset	Feature Dimension	Objects	Target Classes
2nd Order Tensor	Breast-Cancer	13×4	683	2
	USPS	16×16	1005	1
	SONAR	8×8	203	1
	MNIST	28×28	70,000	10
	ORL	112 × 92	400	40
3rd Order Tensor	DSA	125×45×60	152	19
	USC-HAD	6×600×5	168	10

TABLE IV
AVERAGED ACCURACY(%) AND AUC (%) WITH STANDARD DEVIATIONS OF DIFFERENT METHODS (OCSVM, LOCSVM, SOFT-BOUNDED DEEP SVD, ONE-CASS DEEP SVD, OCSTM, LOCSTM AND PROPOSED OCSTM-BH) ON BREAST-CANCER DATASET FOR VARIOUS TRAINING DATA SIZE

Num	Class	Metrics	OCSVM	LOCSVM	SB-Deep SVDD	OC Deep SVDD	OCSTM	LOCSTM	OCSTM-BH
2	Class 1	Accuracy	43.92 ± 9.82	68.43 ± 13.42	70.21±10.21	71.24±12.2	63.64 ± 15.08	73.74 ± 14.21	74.43±10.13
		AUC	99.48 ± 0.13	99.48 ± 0.05	99.02±0.04	99.11±0.03	99.32 ± 0.16	99.51 ± 0.03	99.63±0.03
	Class 2	Accuracy	65.20 ± 0.00	65.83 ± 8.75	67.11±10.11	68.41±9.23	69.40 ± 5.17	65.85 ± 16.75	70.22±8.77
		AUC	80.93 ± 28.50	76.64 ± 31.88	80.23±12.12	80.21±21.10	84.62 ± 23.68	77.55 ± 30.05	85.11±19.21
4	Class 1	Accuracy	59.80 ± 13.51	79.58 ± 10.81	83.22±12.19	84.23±13.70	75.67 ± 13.02	84.16 ± 10.49	85.87±8.74
		AUC	99.43 ± 0.29	99.46 ± 0.06	98.950±0.0	99.21±0.01	98.52 ± 1.92	99.48 ± 0.11	98.51±0.0
	Class 2	Accuracy	70.59 ± 4.89	63.92 ± 18.67	76.43±6.7	78.21±9.43	78.57 ± 7.83	64.80 ± 24.57	79.69±11.43
		AUC	89.59 ± 15.31	71.35 ± 31.82	86.32±8.7	90.43±10.93	92.13 ± 10.49	70.19 ± 30.92	93.21±14.21
6	Class 1	Accuracy	71.68 ± 13.16	84.47 ± 9.38	88.32±12.43	89.21±10.08	82.47 ± 10.39	87.03 ± 10.61	88.45±12.98
		AUC	99.16 ± 1.10	99.27 ± 0.98	99.10±0.01	99.43±0.06	98.31 ± 1.84	99.02 ± 3.41	99.45±0.11
	Class 2	Accuracy	78.16 ± 5.51	68.13 ± 18.7	82.47±4.40	83.86±5.32	83.88 ± 5.95	65.88 ± 26.02	84.01±2.4
		AUC	93.82 ± 6.84	75.58 ± 26.79	94.71±3.2	94.45±7.2	92.96 ± 9.81	71.55 ± 29.03	94.21±2.7
8	Class 1	Accuracy	76.45 ± 11.65	86.00 ± 8.64	90.21±10.34	90.65±12.65	83.26 ± 10.26	89.02 ± 8.17	90.18±14.51
		AUC	99.33 ± 0.73	99.30 ± 0.62	99.51±0.09	99.50±0.11	98.50 ± 1.82	99.47 ± 0.12	99.54±0.08
	Class 2	Accuracy	80.90 ± 5.73	70.34 ± 21.13	85.91±16.67	86.21±19.21	84.96 ± 6.65	70.89 ± 23.37	86.43±15.12
		AUC	93.81 ± 7.86	75.91 ± 26.61	94.45±3.66	94.51±4.32	92.21 ± 10.69	75.93 ± 25.09	93.43±14.55

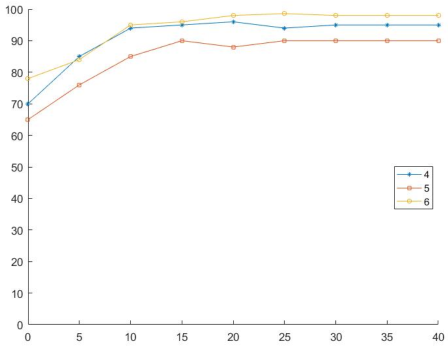

Fig. 2. Performance of OCSTM-BH on MNIST Subject (4,5,6) with different value of σ

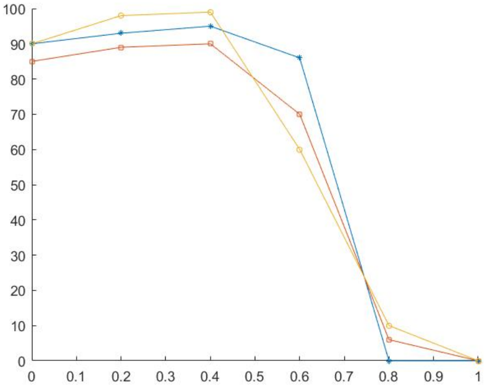

Fig. 3. Performance of OCSTM-BH on MNIST Subject (4,5,6) with different value of v

parameter σ and trade-off parameter v. Inappropriate selection of these parameters may result in poor anomaly detection, thus the value of these parameters should be selected carefully. Figure 3 shows the results of proposed OCSTM-BH on three subjects (4,5,6) of the MNIST dataset with a different value of scale constant, width parameter and trade-off parameter. We noticed that the best performance we have achieved at value in interval $5 \geq \sigma \leq 25$, $0.75 \geq \eta \leq 1.50$ and $0.30 \geq v \leq 0.60$ for all datasets. The performance of OCSTM-BH starts to degrade significantly outsider this interval. We performed a grid search strategy for each approach with 3 fold validation to confirm the optimal range of parameters for each dataset. Once, we have an optimal range of parameters, we have performed 10 fold validation on each training datasets.

C. Experimental Results

To evaluate and compare the performance of the proposed approach with state of the art anomaly detection methods, we have divided the experiment into two phases. In our first phase, we have performed several experiments on real datasets and in the second experiment, we have corrupted the each dataset with outliers by randomly adding 5% of the opposite class samples into the target training dataset. The average AUC values on the training and test dataset together with their corresponding standard deviation on the optimal value of parameters (scaling constant η , width parameter σ and trade-off parameter υ) are shown in table IV, table V, table VII and figure 4 on real datasets. Similarly, table VI and figure 5 describes the results on outliers contaminated datasets.

It is well known that it is hard to solve the classification problem with the small number of sample size especially in case of high dimensional data. To further verify the effectiveness of OCSTM-BH for small sample size, we performed experiments with different sample sizes. Results on real dataset showed that proposed OCSTM-BH considerable performed better in comparison to OCSVM, OCSVM-SVDD OCSTM and R1OCSTM for all datasets. We can observe that one-class deep SVDD [27], soft bounded Deep SVDD [27]) performed marginally better for datasets with large number of samples as shown in table V and figure 4. Notice that OCSTM-BH performed better for small sample size (2,4) whereas soft bounded Deep SVDD showed partially better or similar performance for sample size (6,8), however, it is computationally complex. Figure 4 and figure 5 show that OCSTM-BH is scalable and can be used for high dimensional data consisting of small number of samples.

The main goal of this work is to improve the robustness of anomaly detection. Thus, to elucidate the performance against corrupted datasets, we haven contaminated the datasets with 5% anomalies. Table V and figure 5 show that OCSTM-BH showed significantly better performance in comparison to OCSVM, LOCSVM, R1TVM, OCSTM, LOCSTM, Deep SVDD, Soft bounded Deep SVDD on all contaminated datasets. This validate that OCSTM-BH has better anti-outliers ability especially compared to Deep SVDD and Soft bounded Deep SVDD.

Comprehensive evaluation indicates that proposed OCSTM-BH showed better performance in comparison to state of the art method. From figure 5 and table VI, we can observe that the gain in performance on corrupted dataset is significantly better in comparison to other methods. This shows the robustness of OCSTM-BH against outliers. We observe that the proposed OCSTM-BH degenerates to traditional one-class support tensor machines for scale constant $\eta = 0$. Thus, we can say that OSTM-BH is a special case of OCSTM. We further observe that bounding hinge loss function results in significantly overcoming the influence of outliers by reducing the impact of samples that are far from their labels.

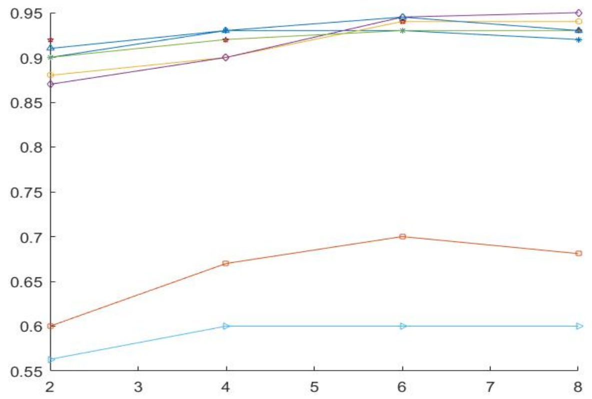

Fig. 4. Averaged performance (AUC) of the 40 target classes on ORL dataset (real) with respect to the training sample sizes (2,4,6,8)

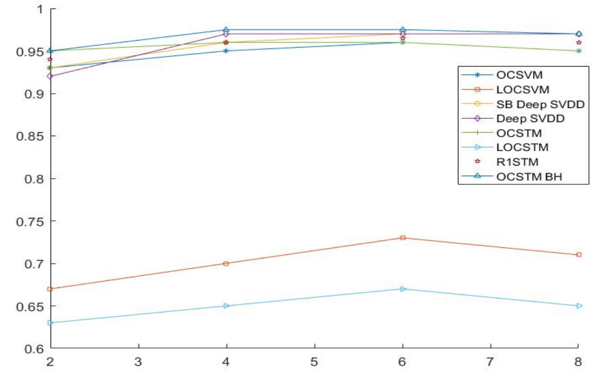

Fig. 5. Averaged performance (AUC) of the 40 target classes on ORL dataset (contaminated) with respect to the training sample sizes (2,4,6,8)

IV. Conclusion

In this work, we presented a novel one class support tensor machines based on correntropy-induced loss function. We replaced the traditional hinge loss function with bounded loss function which is monotonic and non-convex, thus, robust against outliers by decreasing the loss caused by outliers. To solve the non-convex optimization, We have presented half quadratic optimization based on alternating optimization method and drive the problem to traditional OCSTM which then can be solved using dual optimization. Extensive evaluation on eight publicly available real world datasets showed that OCSTM-BH provided better performance especially in the presence of outliers which shows that OCSTM-BH is effective in dealing with outliers.

.

TABLE V
AVERAGE AUC(%) WITH STANDARD DEVIATIONS ON MNIST DATASET

Class	OCSVM-SVDD	SB Deep SVDD	OC-Deep-SVDD	OCSTM	R1OCSTM	OCSTM-BH
0	96.75± 0.5	97.8± 0.7	**98.5± 0.7**	97.87± 0.9	97.90+1.1	98.1± 0.5
1	99.15± 0.4	99.6± 0.1	**99.7± 0.08**	99.65± 0.6	99.60± 0.7	99.6± 0.07
2	79.4± 0.9	89.5± 0.2	91.7± 0.8	90.20± 0.4	90.43± 0.9	**92.1± 0.5**
3	86.1± 0.6	90.3± 0.1	91.9± 0.5	91.1± 0.3	91.00± 0.7	**92.1± 0.4**
4	94.21± 0.3	93.8± 0.5	**95.32± 0.8**	93.21± 0.1	92.8± 0.54	95.2± 0.9
5	73.1± 0.8	85.8± 0.5	**89.23± 0.9**	86.22± 0.4	97.1± 0.3	89.12± 0.9
6	95.5± 0.2	98.0± 0.4	98.3± 0.5	98.0± 0.7	98.10± 0.1	**98.60± 0.2**
7	92.16± 0.1	92.7± 0.4	94.6± 0.9	92.7± 0.3	93.85± 0.7	**95.00± 0.5**
8	89.09± 0.4	**94.2± 0.4**	93.9± 0.6	93.6± 0.2	92.96± 0.4	94.1± 0.4
9	92.71± 0.2	94.9± 0.6	96.35± 0.3	95.7± 0.3	95.71± 0.2	**96.5± 0.5**

TABLE VI
AVERAGE AUC(%) WITH STANDARD DEVIATIONS ON ANOMALIZED MNIST DATASET

Class	OCSVM-SVDD	SB Deep SVDD	OC-Deep-SVDD	OCSTM	R1OCSTM	OCSTM-BH
0	91.75± 0.22	92.11± 0.16	93.32± 1.1	93.05± 1.22	93.55+1.32	**95.22± 1.28**
1	92.45± 0.9	93.46± 1.4	93.32± 0.34	93.05± 1.4	92.02± 0.5	**93.87± 0.22**
2	72.43± 1.45	78.32± 1.2	82.54± 1.45	83.08± 1.21	82.01± 1.43	**84.20± 1.32**
3	78.2± 1.54	84.34± 1.56	84.19± 0.4	82.43± 1.06	84.41.± 1.6	**85.11± 1.1**
4	84.43± 2.1	86.47± 0.9	85.32± 1.44	84.61± 1.05	84.78± 1.00	**85.13± 1.22**
5	65.43.1± 3.1	77.98± 1.3	80.11± 0.9	79.11± 3.6	83.79± 1.22	**85.00± 1.54**
6	80.89± 1.70	86.76± 2.2	88.54± 2.5	86.54± 1.21	87.76± 2.10	**88.67± 1.43**
7	80.22± 0.1	83.57± 0.84	85.55± 2.44	86.27± 1.23	85.43± 1.27	**86.81± 0.76**
8	78.65± 2.4	86.4± 1.54	84.43± 2.60	82.46± 2.23	84.76± 2.43	**87.43± 0.80**
9	81.89± 1.6	85.55± 1.96	84.45±1.75	80.54± 4.5	84.79± 1.29	**86.76± 1.1**

TABLE VII
COMPARISON OF AVERAGE AUC(%) ON VARIOUS DATASETS

	AUC							
Dataset	OCSVM	LOCSVM	SB-Deep SVDD	OC Deep SVDD	OCSTM	LOCSTM	R1STM	OCSTM-BH
Breast	90.17	87.65	92.2	99.22	98.29	89.74	99.02	**99.25**
SONAR	58.43	66.21	72.13	**72.23**	61.88	67.87	69.43	72.11
USPS	99.43	99.61	**99.91**	99.85	99.75	97.81	99.87	**99.91**
USCHAD	83.42	89.41	98.67	**99.13**	95.12	97.11	98.47	99.06
ORL	96.12	73.87	97.21	97.50	96.43	69.43	96.89	**97.58**
DSA	79.43	83.47	98.57	99.12	98.24	98.12	99.17	**99.20**

REFERENCES

[1] Ali Anaissi, Young Lee, and Mohamad Naji. Regularized tensor learning with adaptive one-class support vector machines. In *International Conference on Neural Information Processing*, pages 612–624. Springer, 2018.

[2] Billur Barshan and Murat Cihan Yüksek. Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. *The Computer Journal*, 57(11):1649–1667, 2014.

[3] Manuele Bicego and Mario AT Figueiredo. Soft clustering using weighted one-class support vector machines. *Pattern Recognition*, 42(1):27–32, 2009.

[4] Deng Cai, Xiaofei He, and Jiawei Han. Learning with tensor representation. Technical report, 2006.

[5] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[6] Yanyan Chen, Liyun Lu, and Ping Zhong. One-class support higher order tensor machine classifier. *Applied Intelligence*, 47(4):1022–1030, 2017.

[7] Yanyan Chen, Kuaini Wang, and Ping Zhong. One-class support tensor machine. *Knowledge-Based Systems*, 96:14–28, 2016.

[8] Yanyan Chen, Kuaini Wang, and Ping Zhong. A linear support higher order tensor domain description for one-class classification. *Journal of Intelligent & Fuzzy Systems*, (Preprint):1–11, 2018.

[9] LI Dong, LIU Shulin, and Hongli Zhang. A method of anomaly detection and fault diagnosis with online adaptive learning under small training samples. *Pattern Recognition*, 64:374–385, 2017.

[10] Sarah Erfani, Mahsa Baktashmotlagh, Sutharshan Rajasegarar, Shanika Karunasekera, and Chris Leckie. R1svm: a randomised nonlinear approach to large-scale anomaly detection. 2015.

[11] Sarah M Erfani, Mahsa Baktashmotlagh, Sutharshan Rajasegarad, Vinh Nguyen, Christopher Leckie, James Bailey, and Kotagiri Ramamohanarao. R1stm: One-class support tensor machine with randomised kernel. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 198–206. SIAM, 2016.

[12] Yunlong Feng, Yuning Yang, Xiaolin Huang, Siamak Mehrkanoon, and Johan AK Suykens. Robust support vector machines for classification with nonconvex and smooth losses. *Neural computation*, 28(6):1217–1247, 2016.

[13] R Paul Gorman and Terrence J Sejnowski. Analysis of hidden units in a layered network trained to classify sonar targets. *Neural networks*, 1(1):75–89, 1988.

[14] Riyaz Ahamed Ariyaluran Habeeb, Fariza Nasaruddin, Abdullah Gani, Ibrahim Abaker Targio Hashem, Ejaz Ahmed, and Muhammad Imran. Real-time big data processing for anomaly detection: A survey. *International Journal of Information Management*, 2018.

[15] Lifang He, Xiangnan Kong, Philip S Yu, Xiaowei Yang, Ann B Ragin, and Zhifeng Hao. Dusk: A dual structure-preserving kernel for supervised tensor learning with applications to neuroimages. In *Proceedings*

of the 2014 SIAM International Conference on Data Mining, pages 127–135. SIAM, 2014.

[16] Amin Karami and Manel Guerrero-Zapata. A fuzzy anomaly detection system based on hybrid pso-kmeans algorithm in content-centric networks. *Neurocomputing*, 149:1253–1269, 2015.

[17] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[18] Hao Lin, Guannan Liu, Junjie Wu, Yuan Zuo, Xin Wan, and Hong Li. Fraud detection in dynamic interaction network. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[19] Subhransu Maji, Alexander C Berg, and Jitendra Malik. Efficient classification for additive kernel svms. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):66–77, 2013.

[20] Nour Moustafa, Jiankun Hu, and Jill Slay. A holistic review of network anomaly detection systems: A comprehensive survey. *Journal of Network and Computer Applications*, 128:33–55, 2019.

[21] Nour Moustafa and Jill Slay. The evaluation of network anomaly detection systems: Statistical analysis of the unsw-nb15 data set and the comparison with the kdd99 data set. *Information Security Journal: A Global Perspective*, 25(1-3):18–31, 2016.

[22] Imran Razzak. Cooperative evolution multiclass support matrix machines. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[23] Imran Razzak, Michael Blumenstein, and Guandong Xu. Multi-class support matrix machines by maximizing the inter-class margin for single trial eeg classification. *IEEE Transaction of Neural System and Rehabilitation Engineering*.

[24] Imran Razzak, Michael Blumenstein, and Guandong Xu. Robust support matrix machine. *Pattern Recognition*.

[25] Imran Razzak, Raghib Abu Saris, Michael Blumenstein, and Guandong Xu. Integrating joint feature selection into subspace learning: A formulation of 2dpca for outliers robust feature selection. *Neural Networks*, 121:441–451, 2020.

[26] Oliver Rettig, Silvan Müller, Marcus Strand, and Darko Katic. Which deep artifical neural network architecture to use for anomaly detection in mobile robots kinematic data? In *Machine Learning for Cyber Physical Systems*, pages 58–65. Springer, 2019.

[27] Lukas Ruff, Nico Görnitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Robert Vandermeulen, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International Conference on Machine Learning*, pages 4390–4399, 2018.

[28] W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, pages 861–871. International Society for Optics and Photonics, 1993.

[29] Yu-Xing Tang, You-Bao Tang, Mei Han, Jing Xiao, and Ronald M Summers. Deep adversarial one-class learning for normal and abnormal chest radiograph classification. In *Medical Imaging 2019: Computer-Aided Diagnosis*, volume 10950, page 1095018. International Society for Optics and Photonics, 2019.

[30] Yingjie Tian, Mahboubeh Mirzabagheri, Seyed Mojtaba Hosseini Bamakan, Huadong Wang, and Qiang Qu. Ramp loss one-class support vector machine; a robust and effective approach to anomaly detection problems. *Neurocomputing*, 310:223–235, 2018.

[31] Niall Twomey, Haoyan Chen, Tom Diethe, and Peter Flach. An application of hierarchical gaussian processes to the detection of anomalies in star light curves. *Neurocomputing*, 2019.

[32] Franco van Wyk, Yiyang Wang, Anahita Khojandi, and Neda Masoud. Real-time sensor anomaly detection and identification in automated vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[33] Yingchao Xiao, Huangang Wang, and Wenli Xu. Ramp loss based robust one-class svm. *Pattern Recognition Letters*, 85:15–20, 2017.

[34] Hong-Jie Xing and Man Ji. Robust one-class support vector machine with rescaled hinge loss function. *Pattern Recognition*, 84:152–164, 2018.

[35] Lei Xu, Adam Krzyzak, and Ching Y Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE transactions on systems, man, and cybernetics*, 22(3):418–435, 1992.

[36] Jinhong Yang, Tingquan Deng, and Ran Sui. An adaptive weighted one-class svm for robust outlier detection. In *Proceedings of the 2015 Chinese Intelligent Systems Conference*, pages 475–484. Springer, 2016.

[37] Shen Yin, Xiangping Zhu, and Chen Jing. Fault detection based on a robust one class support vector machine. *Neurocomputing*, 145:263–268, 2014.

[38] Liangwei Zhang, Jing Lin, and Ramin Karim. An angle-based subspace anomaly detection approach to high-dimensional data: With an application to industrial fault detection. *Reliability Engineering & System Safety*, 142:482–497, 2015.

[39] Mi Zhang and Alexander A Sawchuk. Usc-had: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 1036–1043. ACM, 2012.

[40] Fa Zhu, Jian Yang, Cong Gao, Sheng Xu, Ning Ye, and Tongming Yin. A weighted one-class support vector machine. *Neurocomputing*, 189:1–10, 2016.