Network Security Monitoring by Combining Semi-Supervised Learning and Active Learning

Yun Pan, Xinyang Agriculture and Forestry University, China*

ABSTRACT

In network intrusion and network security monitoring, there is massive data. When using supervised learning method directly, it will cost lots of time to collect labeled samples, which is expensive. In order to solve this issue, this paper adopts an active learning model to detect network intrusion. First, massive unlabeled samples are used to establish a weighted support vector data description model. Then, the most valuable samples are used to improve the performance of network intrusion by combining with active learning, which utilizes labeled samples and unlabeled samples to extend the weighted support data description model in a semi-supervised learning method. The experimental results show that the active learning can utilize minor labeled sample to reduce the cost of manual labeling work, which is more suitable for an actual network intrusion detection environment.

KEYWORDS

Active Learning, Feature Representation, Intrusion Detection, Network Security, Semi-Supervised Learning, Weighted Support Vector Data Description

1. INTRODUCTION

In an incremental complex network environment, network intrusions and attacks (Hong 2014, Sultana 2019) become more and more diverse and complicated, and new attack methods are emerging one after another. The anomaly detection can build a model to detect network intrusions by calculating the deviation of new visiting from the distribution of normal behavior visiting (Zhang 2015). Compared with the method based on feature rule library (Ayo 2020), the anomaly detection can identify the unknown intrusion type, which is an important part to ensure network security. In recent years, the anomaly detection based network intrusion detection has become a hot topic in the community of network security.

However, the anomaly detection methods requires to collect a large amount of labeled data as the training set to learn anomaly detection model. Additionally, the false alarm rate of anomaly detection methods are generally high. In the actual network environment, it requires expert knowledge to distinguish and label network visiting data. The process to denote and collect network visiting data is time-consuming, labor-intensive and costly (Zhang 2020). Thus, the quantity of high quality

DOI: 10.4018/IJISMD.313578

*Corresponding Author

This article published as an Open Access Article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

network visiting data is very limited. However, it is easy to obtain massive mixed data which consists of a large amount of normal network visiting data and minor abnormal network visiting data. It is urgent to utilize the impure data to improve the performance of anomaly detection for network intrusion detection. In order to solve this issue, this paper combines weighted support vector data description (Hamidzadeh 2017) and active learning (Freeman 2014) to detect potential network intrusions and attacks. First, the impure network visiting data is used to learn a weighted support vector data description model. The learnt model is used to select a small amount of high-value data to denote. Lastly, the denoted data are used to retrain a semi-supervised learning model to improve the network intrusion performance.

The rest of this paper is organized as follows. The related work is introduced in Section 2. Section 3 adopts active weighted support vector data description to detect network intrusions and attacks. Section 5 is the experiments and simulations. The discussion and conclusion is provided in the last section.

2. RELATED WORK

Network intrusion and attack detection is an important and difficult task in the community of network security. Many researchers have conducted a lot of efforts and proposed many anomaly detection methods for network intrusion detection. The methods include data mining based anomaly detection (Wang 2018), fractal time series based anomaly detection (Radivilova 2019), information fusion based anomaly detection (Zhang 2008), principal component analysis based anomaly detection (Salman 2018), wavelet analysis based anomaly detection (Lu 2009), and fractal feature parameters based anomaly detection (Ya-min 2009). These methods performs feature analysis from different aspects to establish anomaly detection model and have been well applied in practice. These methods focus on how to extract features to train anomaly detection model, which can achieve a high detection accuracy. However, it requires massive labeled samples which are difficult to collect in actual network environment.

One way to avoid collecting massive labeled samples is unsupervised anomaly detection. However, the unsupervised methods have high false error rate. Another way is to learn a one-class classification model by using existing data. In general, it cannot guarantee that the existing network visiting data is normal. It must control the influence of the mixed abnormal visiting data in the training set. In order to solve this issue, robust one class classification methods are proposed (Zhu 2016). In order to further improve the performance of one-class classification model for network intrusion and attack detection, some high quality labeled visiting samples are necessary. In order to solve the acquisition of labeled data usually depends on expert knowledge and is time-consuming, active learning is adopted to select the samples that are most conducive to improve the performance of the machine learning to submit to experts for annotating. Then, the annotated samples are used to train supervised learning model to improve the performance of the machine learning model. Recently, the application of active learning has been attracted the attention of the researchers in the community of network security.

3. NETWORK SECURITY MONITORING VIA ACTIVE WEIGHTED SUPPORT VECTOR DATA DESCRIPTION

The idea of this method is summarized as follows. It first trains a weighted support vector data description by using existing network visiting samples which have not annotated; then uses the active learning method to select a small number of samples to request labeling; lastly combines these labeled data to retrain the model in a semi-supervised manner, in which the training set contains some labeled samples and massive unlabeled samples. The sample selection and model training are performed again. The process is repeated until the termination condition is satisfied.

3.1 Weighted Support Vector Data Description

Let $X = \left\{x_1, \ldots, x_l\right\}$ represent the training set. In X, most samples are normal. The aim of classical support vector data description is to find a hyper-sphere which can enclose most of the training sample with the volume as small as possible. An illustration is shown in Figure 1.

Let $f : X \rightarrow Y$ represent the evaluation function. For a sample x_i, function $f(x_i)$ returns a label y_i. According to the principle of structural risk minimization, the optimization object can be rewritten as following equation:

$$\begin{aligned} \min_{R,c} \quad & R^2 + C\sum_{i=1}^{l} \xi_i \\ \text{s.t.}: \quad & \varphi\left(x_i\right) - c^2 \leq R^2 + \xi_i, \quad i = 1, \ldots, l \\ & \xi_i \geq 0, \quad i = 1, \ldots, l \end{aligned} \tag{1}$$

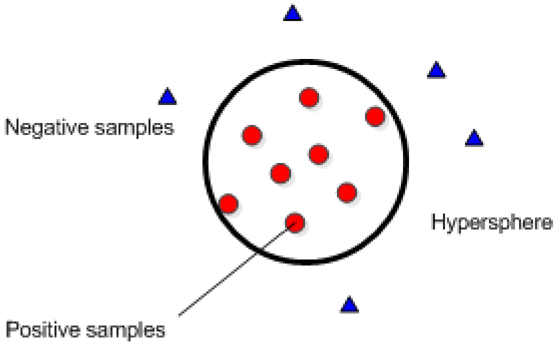
In the equation (1), R is the radius of hypersphere, while c is the center of the hypersphere; $\phi(x_i)$ is the mapping of sample x_i in the reproducing kernel Hilbert space, ξ_i is the associated slack variable of sample x_i, and C is penalty factor to balance the empirical risk and expected risk.

In classical support vector data description, it assumes that all samples in training set have been perfectly labeled. However, it is unavoidable that the network visiting samples contains minor abnormal ones which may make the support vector data description deteriorate seriously. In order to solve this issue, weighted support vector data description is proposed, in which each sample is assigned with a weight to denote the probability of this sample is normal. Similar to weighted one-class support vector machine (Zhu 2016), the training set in weighted support vector data description is reorganized as $\{x_i, \eta_i\}_{i=1}^{l}$ $\left(0 \leq \eta_i \leq 1\right)$. Then, the optimal programming is reformulated as follows:

$$\begin{aligned} \min_{R,c} \quad & R^2 + C\sum_{i=1}^{l} \eta_i \xi_i \\ \text{s.t.}: \quad & \varphi\left(x_i\right) - c^2 \leq R^2 + \xi_i, \quad i = 1, \ldots, l \\ & \xi_i \geq 0, \quad i = 1, \ldots, l \end{aligned} \tag{2}$$

The equation (2) is still a convex optimal programming problem, which can be solved by Lagrange multipliers. The associated Lagrange function is rewritten as follows:

Figure 1. The illustration of support vector data description



$$L = R^2 + C\sum_{i=1}^{l}\eta_i\xi_i - \sum_{i=1}^{l}\alpha_i\left(R^2 + \xi - \varphi\left(x_i\right) - c^2\right) - \sum_{i=1}^{l}\beta_i\xi_i \tag{3}$$

In the equation (3), $\alpha_i \geq 0$, $\beta_i \geq 0$. The following equation can be obtained by the partial derivatives L for R, c and ξ_i:

$$\sum_{i=1}^{l}\alpha_i = 1 \tag{4}$$

$$c = \sum_{i=1}^{l}\alpha_i\varphi\left(x_i\right) \tag{5}$$

$$\alpha_i = C\eta_i - \beta_i \tag{6}$$

By substituting equations (4), (5) and (6) into (2), we can obtain the following equation:

$$\begin{aligned} \underset{\alpha}{min} \quad & \sum_{i,j=1}^{l}\alpha_i\alpha_j K\left(x_i, x_j\right) - \sum_{i=1}^{l}\alpha_i K\left(x_i, x_i\right) \\ \text{s.t.}: \quad & 0 \leq \alpha_i \leq \eta_i C, \quad i = 1,\ldots,l \\ & \sum_{i=1}^{l}\alpha_i = 1 \end{aligned} \tag{7}$$

After obtaining α_i, and c, the decision function $f(x)$ is represented as follows:

$$f\left(x\right) = \varphi\left(x\right) - c^2 = K\left(x,x\right) - 2\sum_{i=1}^{l}\alpha_i K\left(x, x_i\right) + \sum_{i,j=1}^{l}\alpha_i\alpha_j K\left(x_i, x_j\right) \tag{8}$$

In the equation (8), $K(x_i, x_j) =< \phi(x_i), \phi(x_j) >$. When $f(x) < R^2$, x is located within the hypersphere and is a normal sample; when $f(x) > R^2$, x is located outside the hypersphere and is an abnormal sample. Different from classical support vector description, the weighted support vector data description is robust to the noises in the training set. In weighted support vector data description, the requirement of the quality of training set is not as strict as that in classical support vector data description.

3.2 Semi-Supervised Model Based on Weighted Support Vector Data Description

After obtaining the labels for available samples, the semi-supervised learning is used to extend and optimize the weighted support vector data description. The labeled samples are obtained by active learning. In semi-supervised learning, the training set is represented as $X = \{x_1, \ldots, x_l, x_{l+1}, \ldots, x_{l+n}\}$. The former l samples are the unlabeled samples, while the latter m samples are labeled +1 or -1. When it is denoted as +1, it is a positive sample; otherwise, it is a negative sample. Then, the objective of semi-supervised weighted support vector data description is represented as follows:

$$\begin{array}{ll} \min\limits_{R,c} & R^2 + C_1\sum_{i=1}^{l}\eta_i\xi_i + C_2\sum_{i=l+1}^{l+n_1}\xi_i + C_3\sum_{i=l+n_1+1}^{l+n}\xi_i \\ \text{s.t.}: & \varphi\left(x_i\right) - c^2 \le R^2 + \xi_i, \xi_i \ge 0, \quad i = 1,\ldots,l \\ & \varphi\left(x_i\right) - c^2 \le R^2 + \xi_i - \gamma, \xi_i \ge 0, \quad i = l+1,\ldots,l+n_1 \\ & \varphi\left(x_i\right) - c^2 \ge R^2 + \xi_i - \gamma, \xi_i \ge 0, \quad i = l+n_1+1,\ldots,l+n \end{array} \tag{9}$$

In the equation (9), it is assumed that the former n_1 labeled samples in the training set are positive, while the latter n_2 labeled samples in the training set are negative. Variable γ represents the expected minimum distance between positive samples and negative samples. The constants C_1, C_2, and C_3 are penalty factors. Constant C_1 reflects the importance of the unlabeled samples for semi-supervised weighted support vector data description model, while C_2 and C_3 reflects the importance of labeled samples for semi-supervised weighted support vector data description. When C_1 is close to 0, the effect of labeled samples will be weaken. When C_1 is set as 0, the semi-supervised weighted support vector data description is degenerated as classical support vector data description. Constants C_2 and C_3 depends on the requirements of false alarm rate and false negative rate. Since the cost of misjudgment of abnormal data is higher than that of normal data, constant C_2 is generally set smaller than constant C_3. In general, the constants are set as $C_1 < C_2 < C_3$.

The constraint conditions in the equation (9) are expressed in the form of risk function. Then, it is converted as an unconstrained optimization problem which is written as follows:

$$\begin{array}{l} J = R^2 + C_1\sum_{i=1}^{l}\eta_i\left(R^2 - \varphi\left(x_i\right) - c^2\right) + C_2\sum_{i=l+1}^{l+n_1}\left(R^2 - \varphi\left(x_i\right) - c^2 - \gamma\right) \\ +C_3\sum_{i=l+n_1+1}^{l+n}\left(\varphi\left(x_i\right) - c^2 + \gamma - R^2\right) \end{array} \tag{10}$$

In the equation (10), the center c is represented as follows:

$$c = \sum_{i=1}^{l}\alpha_i\varphi\left(x_i\right) + \sum_{i=l+1}^{l+n}\alpha_i y_i\varphi\left(x_i\right) \tag{11}$$

In the equation (11), it can be found that the center is decided by both labeled samples and unlabeled samples. The equation (10) can be solve by gradient method.

3.3 Active Learning for Selecting Samples to Denote

In the above semi-supervised learning model, the process of obtaining labeled samples is cumbersome and expensive. The cost to collect labeled samples should be minimized. This paper adopts active learning to denote samples for semi-supervised learning. The keys of active learning are the selection strategy and termination conditions.

According to the ways of acquiring samples through active learning, the selection strategy can be classified into three types: membership query comprehensive method, flow-based selective sampling method and pool-based selective sampling method. Among them, the pool-based selective sampling method has been thoroughly studied. In pool-based selective sampling method, it first utilizes unlabeled samples to compose a sample pool with relatively fixed distribution and characteristics, then proceeds sample evaluation and selection according to a certain strategy. According to the selection strategies, the pool-based selective sampling is classified into uncertainty reduction based method, version space reduction based method, and generalization error reduction based method. The uncertainty reduction based method requests for labelling the samples with the most ambiguous classification information. The version space reduction based method requests for labelling the samples that can minimize the version space. The generation error reduction based method reduces the classification error to improve the classification ability of the classifier.

In general, the reasons for high false alarm in anomaly detection include the purity of the training set and the completeness of the training set. For the former, the anomaly detector would degrade if the training set contains abnormal samples. For the latter, if the training set cannot depict whole characteristics of the normal samples, it may increase the false alarm rate. Thus, when selecting samples for semi-supervised weighted support vector data description, we need to first select high-confidence samples to improve the purity of the training set, and then select representative samples to cover all characteristics of the training set as much as possible.

In anomaly detection, the samples near the decision boundary are usually selected for labelling. These samples usually have the largest uncertainty and can provide more information for optimizing the model. When there is no labeled sample in the training set, this strategy is stopped.

Merely selecting near boundary samples, the anomalies may be selected when the boundary samples pass through a sparse areas. However, these anomalies cannot represent the characteristics of the normal samples, which are not inductive to improve the performance of the semi-supervised learning model. When the boundary samples pass through a dense area, a large number of samples in this area will be requested to be labeled. These samples usually have the same characteristics, which would increase the cost for labeling. Therefore, it is hard to completely describe the characteristics of the training set merely using limited labeled samples. In order to solve this issue, this paper adopts adjacent metric to select samples for labeling.

In active learning, it needs to set termination condition to control the learning process. In general, the learning process stops when it achieves a certain condition, such as the limited iterations, a certain performance indicator. This paper adopts the following termination condition:

$$con = MSE\left(f\left(x\right) - y\right) + \operatorname{var}\left(f\left(x\right) - y\right) \tag{12}$$

In the equation (12), the first term represents the error of the labeled sample between predicted label and ground truth; the second term represents the ratio of the difference between the predicted values for all unlabeled samples. The whole procedure for semi-supervised weighted support vector data description by using active learning is summarized as shown in Algorithm 1.

4. EXPERIMENTS AND SIMULATIONS

In this section, we will use the KDD Cup 99 to evaluate the proposed network intrusion detection method. In order to verify the effectiveness of the proposed labeling method and the semi-supervised

Algorithm 1. Semi-supervised weighted support vector data description by using active learning

```
Input: Training set $X$ which contains minor anomaly samples
Output: Anomaly detector
Step 1: Using original training set to learn a weighted support vector data description model;
Step 2: Carry out active learning to select samples for labeling;
Step 3: Combining labeled samples and unlabeled samples to learn semi-supervised model;
Step 4: Using active learning to select samples for labeling according to semi-supervised model;
Step 5: Using new labeled samples to learn semi-supervised model again;
Step 6: If reaching termination condition, algorithm stops; otherwise, it goes to Step 4.
```

method for anomaly detection, we compare the proposed method with classical support vector data description (SVDD) (Tax 2004), one-class support vector machine (OC-SVM) (Zhu 2016), weighted support vector data description (WSVDD) (Cha 2014), and support vector data description with random labeling samples (SVDD (Ran)). The proposed method is short for Semi-SVDD.

In KDD Cup 99, it contains nearly 5 million records. Each record contains 41 features to describe network connecting status. The dataset contains normal data, and four types of abnormal data. The details are reported in Table 1.

Most of the training set is normal samples. However, it is unavoidable to contain minor abnormal samples in the training set which we do not know. The ratio of abnormal samples in test set is relatively high. In SVDD based methods and one-class support vector machine, the Gaussian function is adopted as kernel function:

$$K\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|}{\sigma^2}\right)$$

The width of the Gaussian function is set as $\sigma^2 = 1.25$ directly. The parameter C is tuned by grid search to ensure highest accuracy.

In anomaly detection, the dataset is usually unbalanced, in which most samples are normal. The misclassification of abnormal samples will induce more serious consequence. In order to better evaluate the anomaly detection methods, the experimental results are reported in terms of accuracy, recall, and precision. Let TP represent true positive, FP represent false positive, FN represent false negative, and TN represent true negative, which are illustrated in Figure 2.

Table 1. The description of the KDD cup dataset

	Training	Ratio	Test	Ratio
norml	21,000	91.81%	7,835	43.71%
buffer_overflow	1,365	5.97%	2,426	13.54%
xss	509	2.22%	1,438	8.02%
code_injection	0	0%	2,115	11.8%
other	0	0%	4,109	22.93%
total	22,874	100%	17,923	100%

Figure 2. The illustration of TP, FP, FN, and TN

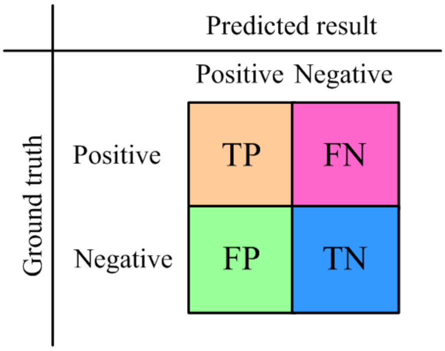


Table 2. The result of network intrusion and attack detection for KDD cup dataset

	Labeled samples			Performance		
	Pos.	**Neg.**	**Sum**	**Accuracy**	**Recall**	**Precision**
SVDD	0	0	0	79.83%	83.23%	81.51%
OC-SVM	0	0	0	80.03%	82.46%	81.62%
WSVDD	0	0	0	81.17%	83.87%	82.53%
SVDD (Ran)	9.67%	0.98%	10.65%	83.72%	86.34%	85.18%
Semi WSVDD	6.13%	0.87%	7.00%	88.64%	91.37%	88.03%

The accuracy is defined as $\frac{TP+TN}{TP+FP+FN+TN}$. The recall is defined as $\frac{TN}{FP+TN}$. The precision is defined as $\frac{TN}{FN+TN}$. The accuracy reflects the overall performance of the learning model. The recall reflects the sensitivity of the learning model to anomalies. The details of the experimental results are reported in Table 2.

From the result in Table 1, it can be found that the accuracy, recall, and precision of SVDD achieve 79.83%, 83.23%, and 81.51%, respectively; the accuracy, recall, and precision of OC-SVM achieve 80.03%, 82.46%, and 81.62%, respectively; the accuracy, recall, and precision of WSVDD achieve 81.17%, 83.87%, and 82.53%, respectively; the accuracy, recall, and precision of SVDD (Ran) achieve 88.64%, 91.37%, and 88.03%, respectively. Obviously, when considering semi-supervised learning, the one class classifiers performs better for network intrusion problem. Semi WSVDD labels few samples than SVDD (Ran), however the accuracy, recall, and precision of Semi WSVDD are all higher than that of SVDD (Ran). Compared with randomly selecting samples to label, active learning can need few samples to be label and these samples contains more representation information.

5. CONCLUSION

In order to solve the issue that it is difficult to obtain labeled data in network anomaly detection, this paper proposes an anomaly detection framework by combining semi-supervised weighted support vector data description with active learning. First, a weighted support vector data description model is learnt by using available network visiting data which mainly consists of normal visiting data. Then, active learning is adopted to select representative samples to be labeled. The labeled samples and remaining unlabeled samples are used to train semi-supervised weighted support data description model. The experimental results demonstrate that compared with previous works, the proposed method only need to labeled fewer samples to achieve better accuracy, recall, and precision.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Ayo, F.E., Folorunso, S., Abayomi-Alli, A., Adekunle, A.O., & Awotunde, J.B. (2020). Network intrusion detection based on deep learning model optimized with rule-based hybrid feature selection. *Information Security Journal: A Global Perspective, 29*, 267 - 283.

Cha, M., Kim, J., & Baek, J. (2014). Density weighted support vector data description. *Expert Systems with Applications*, *41*(7), 3343–3350. doi:10.1016/j.eswa.2013.11.025

Freeman, S., Eddy, S. L., McDonough, M., Smith, M., Okoroafor, N., Jordt, H., & Wenderoth, M. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8410–8415. doi:10.1073/pnas.1319030111 PMID:24821756

Hamidzadeh, J., Sadeghi, R., & Namaei, N. (2017). Weighted support vector data description based on chaotic bat algorithm. *Applied Soft Computing*, *60*, 540–551. doi:10.1016/j.asoc.2017.07.038

Hong, J., Liu, C., & Manimaran, G. (2014). Detection of cyber intrusions using network-based multicast messages for substation automation. *ISGT*, *2014*, 1–5. doi:10.1109/ISGT.2014.6816375

Lu, W., & Ghorbani, A. (2009). Network Anomaly Detection Based on Wavelet Analysis. *EURASIP Journal on Advances in Signal Processing*, *2009*, 1–16.

Radivilova, T., Kirichenko, L., Ageyev, D., Tawalbeh, M., Bulakh, V., & Zinchenko, P. (2019). Intrusion Detection Based on Machine Learning Using Fractal Properties of Traffic Realizations. *2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT)*, 218-221. doi:10.1109/ATIT49449.2019.9030452

Salman, M., Husna, D., Apriliani, S.G., & Pinem, J.G. (2018). Anomaly based Detection Analysis for Intrusion Detection System using Big Data Technique with Learning Vector Quantization (LVQ) and Principal Component Analysis (PCA). *AIVR 2018.*

Sultana, N., Chilamkurti, N., Peng, W., & Alhadad, R. (2019). Survey on SDN based network intrusion detection system using machine learning approaches. *Peer-to-Peer Networking and Applications*, *12*(2), 493–501. doi:10.1007/s12083-017-0630-0

Tax, D., & Duin, R. P. (2004). Support Vector Data Description. *Machine Learning*, *54*(1), 45–66. doi:10.1023/B:MACH.0000008084.60811.49

Wang, W., & Yin, C. (2018). Research on the Method of Network Intrusion Detection Based on Data Mining Technology. *CSE 2018.*

Ya-min, S. (2009). Anomaly detection algorithm based on fractal characteristics of large-scale network traffic. *Journal of Communication.*

Zhang, F., Geng, J., Qin, Z., & Zhang, J. (2008). Using data fusion for awareness of intrusion in large-scale network. *2008 International Conference on Communications, Circuits and Systems*, 519-523. doi:10.1109/ICCCAS.2008.4657827

Zhang, M., Xu, B., & Gong, J. (2015). An Anomaly Detection Model Based on One-Class SVM to Detect Network Intrusions. *2015 11th International Conference on Mobile Ad-hoc and Sensor Networks (MSN),* 102-107.

Zhang, S., & Du, C. (2020). Semi-Supervised Deep Learning based Network Intrusion Detection. *2020 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC),* 35-40.

Zhu, F., Yang, J., Gao, C., Xu, S., Ye, N., & Yin, T. (2016). A weighted one-class support vector machine. *Neurocomputing*, *189*, 1–10. doi:10.1016/j.neucom.2015.10.097

Zhu, F., Yang, J., Xu, S., Gao, C., Ye, N., & Yin, T. (2016). Relative density degree induced boundary detection for one-class SVM. *Soft Computing*, *20*(11), 4473–4485. doi:10.1007/s00500-015-1757-7