doi: 10.7690/bgzdh.2022.04.011

基于改进的 OCSVM 算法的工控网络异常检测算法

徐　园，梅　勇，龚　俊，孙梧雨
（中国兵器装备集团自动化研究所有限公司特种计算机事业部，四川 绵阳 621000）

摘要：为提高工控系统异常流量检测能力，设计一种结合孤立森林（isolation forest，iForest）和单类支持向量机（one-class support vector machine，OCSVM）的混合算法。采用孤立森林算法检测训练数据中的离群点，将离群点剔除以降低其对单类支持向量机决策函数的影响；基于正常数据训练单类支持向量机模型，结合特征选取和参数优化进一步提高异常检测模型的检测率。实验结果表明：在燃气管道数据集上，该算法模型的检测率提高至 92.51%，特别是对异常行为的召回率和查准率上升，优化了异常检测模型的性能，满足可靠性要求。

关键词：工业控制网络；异常检测；单类支持向量机；孤立森林
中图分类号：TP391　　**文献标志码：**A

Industrial Control Network Anomaly Detection Algorithm Based on Improved OCSVM Algorithm

Xu Yuan, Mei Yong, Gong Jun, Sun Wuyu
(*Department of Special Computer*, *Automation Research Institute Co*., *Ltd. of China South Industries Group Corporation*, *Mianyang* 621000, *China*)

Abstract: In order to improve the ability of anomaly traffic detection in industrial control system, a hybrid algorithm combining isolated forest (iForest) and one-class support vector machine (OCSVM) is designed. The isolated forest algorithm is used to detect outliers in the training data, and the outliers are eliminated to reduce their impact on the one-class support vector machine decision function.The OCSVM model is trained based on normal data, and the detection rate of the anomaly detection model is further improved by combining feature selection and parameter optimization. The experimental results show that the detection rate of the algorithm model is improved to 92. 51% on the gas pipeline data set, especially the recall rate and precision rate of abnormal behavior are improved, which optimizes the performance of the anomaly detection model and meets the reliability requirements.

Keywords: industrial control network; anomaly detection; OCSVM; isolated forest

0　引言

随着信息化和工业化的深度融合，大部分关键的基础设施和工业系统，比如电力、水利、交通运输等行业走向了传统控制和信息智能化控制相结合的道路。传统工控网络缺乏相应的安全防护机制，工控系统网络的变化不可避免地带来一些网络安全问题，工控系统网络智能化的同时也暴露出传统工控系统存在的安全漏洞。工控系统网络安全与国家关键基础设施安全、国家安全息息相关[1-5]。

实际工业控制系统的异常样本难以获得且数量占比很小，存在样本不平衡的问题，导致常见的 2 类检测模型难以建立且异常检测精度低。Schölkopf 等[6]提出的单类支持向量机（OCSVM）是一种只需单类样本即正常数据就可以训练检测模型的方法，对噪声具有鲁棒性，可以用来建立较准确的异常检测模型。A.Terai 等[7]利用数据包的间隔和长度特征构建的 OCSVM 判别模型，提高了异常检测能力，具有一定实用价值。刘万军等[8]提出一种结合 DBSCAN 和 K-means 方法的 OCSVM 算法，消除离群点和内部异常点对 OCSVM 训练模型精度的影响，提高了在气体管道数据集上的总体检测率，达到 91.81%。尚文利等[9]通过粒子群优化算法对 OCSVM 的参数进行优化，减少误报率。

1　单类支持向量机算法

传统机器学习的方法大多是针对二分类或者多分类数据，通过不同的方法来寻找最优的分界面，但当面对单类数据或者样本不均衡的情况时，如工控网络异常检测，普通的机器学习方法检测精度低，并不能直接适用。

考虑到实际场景中恶意样本的缺失，决定仅采

收稿日期：2021-12-20；修回日期：2022-01-28
作者简介：徐　园（1996—），女，江苏人，硕士，从事工控网络安全研究。E-mail: 15851856605@163.com。

用正常样本来构建网络通信数据的分类模型。单类支持向量机是基于 SVM 算法发展起来的，只需要单类数据样本来训练模型的算法。利用单类数据，并将该类数据映射到尽可能小的区域内。以坐标原点为参考，表现在特征空间上就是获得一个能够将所有样本与原点隔开并且距离原点尽可能远的超平面[10-11]。

设训练样本 $D=\{x_1, x_2, \cdots, x_l\}$，求解以下的二次规划问题：

$$\min_{\omega,\varepsilon,\rho}\frac{\|\omega\|^2}{2}+\frac{1}{\nu l}\sum_{i=1}^{l}\xi_i-\rho;$$

$$\text{s.t.}\ (\omega\bullet\Phi(x_i))\geqslant\rho-\xi_i,\quad \xi_i\geqslant 0, i=1,2,\cdots,l。\tag{1}$$

其中：l 为集合 D 的大小；ξ_i 为非零松弛变量，是为了避免函数过拟合而加入的惩罚项；ν 为正则化参数，通常 $\nu\in(0,1)$。如果 ω 和 ρ 是该问题的解，构建决策函数为

$$f(x)=\text{sgn}((\omega\bullet\Phi(x))-\rho)。\tag{2}$$

决策函数 $f(x)$为正，x 的检测结果是正常，反之检测出 x 是一条异常数据。

和基于 SVM 的有监督学习算法相比，可以发现基于 OCSVM 的异常检测算法性能表现还有一定差距，主要体现在最优召回率上。但在实际的工控网络安全场景中，不依靠恶意样本标记的算法才有实际应用价值，基于 OCSVM 的异常检测算法是一种可行的方案。

2 孤立森林算法

孤立森林(iForest)也是一种可以用于异常检测的无监督学习算法，常用于离群点检测和奇异值检测。笔者把 iForest 作为一种剔除训练数据中离群点的方法。

iForest 是由南京大学周志华等[12-13]提出的基于集成学习快速离群点检测方法。与其他异常检测算法通过距离、密度等量化指标刻画样本间的疏离程度不同，iForest 通过样本点的孤立来检测异常值。

iForest 利用一种名为孤立树(iTree)的二叉树搜索结构来孤立样本，iForest 由 t 个 iTree 组成。数据集中的异常数据的样本密度低，在二叉树结构中会较早地被孤立出来，离 iTree 的根节点距离近；相比而言，正常样本就会离 iTree 的根节点较远。$h(x)$是从根节点遍历 iTree 到外部节点的路径长度，$c(n)$是给定 n 的 $h(x)$的平均值，用来归一化 $h(x)$。异常值分数 $s(x, n)$定义为：

$$s(x,n)=2^{-E(h(x))/c(n)}。\tag{3}$$

其中：

$$c(n)=\begin{cases}2H(n-1)-2(n-1)/n, & n>2\\ 1, & n=2\\ 0, & \text{其他}\end{cases}。\tag{4}$$

$H(i)$是可以通过 $\ln(i)+0.577\ 215\ 664\ 9$(欧拉常数)来估算的调和数。当 $E(h(x))\to c(n)$时，$s\to 0.5$；当 $E(h(x))\to 0$ 时，$s\to 1$；当 $E(h(x))\to n-1$ 时，$s\to 0$。

如果 $s\to 1$，那么 x 是绝对异常；如果 $s\to 0$，则 x 被认为是正常数据；如果所有的 $s\to 0.5$，那么整个样本基本正常。

3 OCSVM 算法的优化

OCSVM 算法广泛应用于异常检测中，但是该算法对离群点很敏感，所谓敏感就是在模型训练时会把离群点划到决策边界内，从而对决策函数产生影响，使得检测精度下降。为降低离群点对模型训练的负面影响，在用 OCSVM 算法训练模型前应先剔除离群点。

笔者选用 iForest 算法剔除离群点，但是 iForest 不适用于特别高维的数据集，噪声以及维度信息的低效利用，会影响算法的可靠性；因此，在利用 iForest 算法剔除离群点前应先进行特征选择，删去无关维度。

笔者作了如下改进：

1) 为消除无关或次相关属性对 OCSVM 检测模型检测率的影响以及对 iForest 算法可靠性的影响，删去无关属性。首先用 VarinanceThreshold 特征选择算法剔除数据集中特征值完全相同的一列，然后通过控制变量法对剩余特征属性逐一删减，观察该属性对检测精度是否存在负影响，适当删减属性条目，提高检测率。

2) 为避免离群点对训练模型的影响，采用 iForest 算法剔除训练集中的异常离群点，降低其对决策函数产生的影响，再进行参数优化，改善模型的训练效果。

iFOCSVM 算法构建过程如图 1 所示。

4 算法验证

4.1 仿真环境

仿真平台：Intel(R) Core(TM) i5-9300H CPU @ 2.40 GHz，8 GB 内存，Windows10 操作系统，选用 Python 编程语言，实现 OCSVM 算法主要来自机器学习库 sklearn 中的 SVM 模块，选用径向基函数

(radical basis function，RBF)作为核函数。

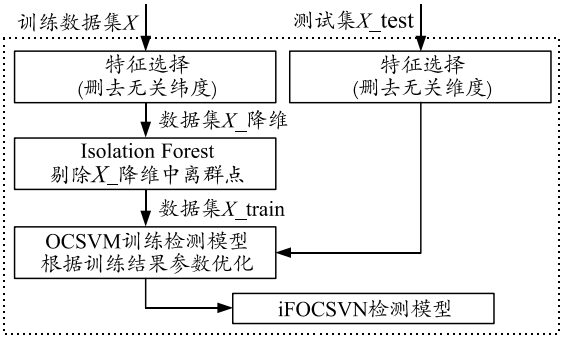

图 1　iFOCSVM 检测模型构建过程

4.2　方案设计

1) 训练集和测试集准备：先对燃气管道数据集进行预处理，选取合适的特征参数，再进行数据集划分，训练集与测试集比例为 3:1。用 iForest 算法删去训练集中的离群点，再提取训练集中所有的正常数据作为单类支持向量机模型的训练数据，组成单类支持向量机的训练集。

2) 模型训练：用 iForest 算法剔除离群点后，基于训练集训练单分类检测模型，再利用测试集测试训练模型的检测效果。根据评价指标进行参数优化，注意避免过拟合。在多次参数优化、结果对比后选择最优的模型参数。具体训练流程如图 2 所示。

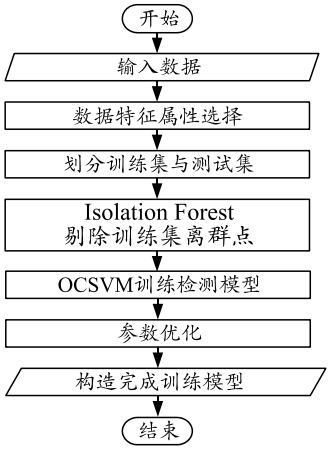

图 2　iFOCSVM 算法检测模型训练流程

3) 对比实验设计：为确定算法的可行性以及算法的训练效果，选择几组用于异常检测的算法进行对比实验。笔者选择的对比算法是同样应用在异常检测方向的几种算法，包括孤立森林 iForest、单分类器 OCSVM、局部异常因子 LOF 等。用同样的数据集训练几种算法模型，再用相同的测试集检验训练效果，将改进后的算法 iFOCSVM 与已有异常检测算法的训练结果作对比，讨论改进的 iFOCSVM 算法在工控系统异常检测中的整体效果。

4) 评价指标：笔者通过以下几种指标来度量检测模型效果：检测率(accurancy)、正样本召回率(positive recall rate，PRR)、正样本查准率(positive precision rate，PPR)、负样本召回率(negative recall rate，NRR)、负样本查准率(negative precision rate，NPR)，各指标计算公式如下：

$$accuracy=(TP+TN)/(TP+TN+FP+FN); \quad (5)$$

$$PRR=TP/(TP+FN); \quad (6)$$

$$PRR=TP/(TP+FP); \quad (7)$$

$$FRR=TN/(TN+FP); \quad (8)$$

$$NPR=TN/(TN+FN)。 \quad (9)$$

其中：TP、FN、FP、TN 定义如表 1 所示。

表 1　各指标公式中数值定义说明

实际	预测	
	Positive	Negative
Positive	True Positive(TP)	False Negative(FN)
Negative	False Positive(FP)	True Negatve(TN)

4.3　数据集

实验采用的数据集是来自密西西比亚州立大学公开的数据集，选择实验室规模的天然气管道数据集。该数据集中有 100 000 多条数据，共 27 个参数，分成 7 类，具体分类如表 2 所示。

表 2　数据类型

数据类型	标签
正常	0
幼稚恶意响应注入	1
复杂恶意响应注入	2
恶意状态命令注入	3
恶意参数命令注入	4
恶意函数代码注入	5
拒绝服务攻击	6
侦察	7

由于实验采用的是无监督方法，所以训练检测模型的时候会删去该列参数，该参数仅用来计算检测模型的检测率等模型评价指标。

原始数据集包含 27 个参数，但实际训练模型中只选取其中 10 个参数特征。在实验过程中发现：删去相关性低的参数，可以提高检测模型的检测率、召回率等，并且留下的每个参数和异常行为有很强的相关性，可以作为异常行为数据的分类依据。

在选取数据特征时，考虑到数据集庞大，为提高效率，选取 10%的子数据集。通过控制变量逐一删去参数，在 OCSVM 模型上观察某参数对于检测结果的影响，如果超出 0.1%的影响范围，则认为该参数重要并且保留，否则删去。最终保留 10 个特征参数，测试集检测率由原来的 87.53%提高到

88.85%，（参数设置为 kernel='rbf'，gamma=0.07，nu=0.2）。

4.4 结果分析

首先选取 10%的子数据集(约 1 万多条数据)用以训练模型，模型的训练结果如表 3 所示。OCSVM 算法和 LOF 算法在燃气管道数据集上的训练结果要优于 iForest 算法，其中：OCSVM 对异常行为的召回率达到 98.96%，LOF 对异常行为的召回率达到 99.27%，而 iForest 只有 72.22%，并且前两者的检测率都达到 88%。从数据上看，OCSVM 算法和 LOF 算法在异常检测上比 iForest 算法更优，但是 OCSVM 算法和 LOF 算法的异常查准率都只有 70%多，在实际应用中存在可靠性问题。改进后的 OCSVM 算法(iFOCSVM)的检测率达到 93.15%，异常行为的召回率为 99.06%，查准率提升至 84.62%，降低误报率，性能明显改善。

表 3 不同算法在 10%燃气管道数据集上的训练效果 %

	检测率	召回率(正)	查准率(正)	召回率(负)	查准率(负)
OCSVM	88.85	83.12	99.29	98.96	76.88
iForest	80.11	84.59	84.29	72.22	72.67
LOF	88.32	82.11	99.50	99.27	75.89
iFOCSVM	93.15	89.79	99.41	99.06	84.62

在原始数据集(约 10 万多条数据)上训练的结果如表 4 所示。

表 4 不同算法在燃气管道数据集上的训练效果 %

	检测率	召回率(正)	查准率(正)	召回率(负)	查准率(负)
OCSVM	86.07	78.89	98.91	98.49	72.96
iForest	79.57	82.14	85.10	75.14	70.87
LOF	83.28	92.47	83.06	67.38	83.81
iFOCSVM	92.51	88.99	99.09	98.59	83.82

对比表 3 和 4，在数据量变大的情况下，模型的训练指标有所下降。其中，对 LOF 算法的影响最大，从表 4 中可以观察到：LOF 模型对正常行为的召回率提高但查准率降低，对异常行为的查准率降低但召回率降至 67.38%，说明在实际工程数据量大的情况下，该算法不适用。数据量的增加基本上没有影响 OCSVM、iForest 和 iFOCSVM 的对比结果，表 4 中 iFOCSVM 算法的训练结果要明显优于其他对比算法，整体检测率达到 92.51%，对于异常行为的召回率达到 98.59%，查准率 83.82%，提高了可靠性，有较高的实用价值。

5 结束语

笔者由工控系统异常检测的研究背景及战略意义，引出 OCSVM 算法在工控网络异常检测中的应用前景。针对其对异常点的敏感提出一种结合 iForest 算法的改进单类支持向量机算法，即 iFOCSVM 算法，提高了单类支持向量机模型的检测率、召回率等。通过仿真对比分析验证了该算法的可行性，仿真结果表明：该算法在燃气管道数据集上的检测率达到 92.51%，满足工控系统可靠性的要求，有较高的实用价值。

参考文献：

[1] BHATTACHARYYA D K, KALITA J K. Network Anomaly Detection: A Machine Learning Perspective[M]. Chapman and Hall/CRC, 2013: 4-6.

[2] 李琳, 尚文利, 姚俊, 等. 单类支持向量机在工业控制系统入侵检测中的应用研究综述[J]. 计算机应用研究, 2016, 33(1): 7-11.

[3] 梁海洋, 张瀚铭, 孙科星. 基于工业互联网的高危产品装配生产线智能管控平台设计[J]. 兵工自动化, 2021, 40(12): 24-28.

[4] 罗耀锋. 面向工业控制系统的入侵检测方法的研究与设计[D]. 杭州: 浙江大学, 2013.

[5] 糜旗. 网络安全态势感知平台架构设计[J]. 兵工自动化, 2021, 40(1): 17-21.

[6] SCHÖLKOPF, BERNHARD, PLATT, et al. Estimating the Support of a High-Dimensional Distribution[J]. Neural Computation, 2001, 13(7): 1443-1471.

[7] TERAI A, ABE S, KOJIMA S, et al. Cyber-Attack Detection for Industrial Control System Monitoring with Support Vector Machine Based on Communication Profile[C]// IEEE European Symposium on Security &Privacy Workshops. IEEE, 2017.

[8] 刘万军, 秦济韬, 曲海成. 基于改进单类支持向量机的工业控制网络入侵检测方法[J]. 计算机应用, 2018, 38(5): 1360-1365.

[9] 尚文利, 李琳, 万明, 等. 基于优化单类支持向量机的工业控制系统入侵检测算法[J]. 信息与控制, 2015, 44(6): 678-684.

[10] CHEN L P. Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar: Foundations of machine learning, second edition[J]. Statistical Papers, 2019, 60(5): 1793-1795.

[11] MAGLARAS L A, JIANG J. A novel intrusion detection method based on OCSVM and K-means recursive clustering[J]. Eai Endorsed Transactions on Security & Safety, 2015, 2(3): 1-10.

[12] FEI T L, KAI M T, ZHOU Z H. Isolation Forest[C]// IEEE International Conference on Data Mining. IEEE, 2008.

[13] 杨杰, 张东月, 周丽华, 等. 基于网格耦合的数据流异常检测[J]. 计算机工程与科学, 2020(1): 25-35.