支持向量机的异常检测应用研究

肖 铮

(工商职业技术学院,四川 成都 611830)

摘要: 机器学习应用于入侵检测的领域已经有很长一段时间。采用支持向量机的入侵检测手段有助于应对日益更新的攻击方法。从数据来源和分析的角度,可以将入侵检测模型分为基于工业流量,工业过程,用户行为,恶意文件四大类。在实验中,使用 Matlab 作为接口来完成与双容水箱系统的通信,通过区域之间状态的推导,就能准确地判断出遭受到攻击的区域。

关键词: 机器学习; 支持向量机; 异常检测

中图分类号: TP301.6 **文献标识码:** A **文章编号:** 2096 - 790X(2020)08 - 0067 - 03

DOI: 10.19576/j.issn.2096 - 790X.2020.08.016

Research on the Application of Support Vector Machine in Anomaly Detection

Xiao Zheng

(*Sichuan Technology & Business College*, *Chengdu City*, *Sichuan Province* 611830)

Abstract: Machine learning has been applied to intrusion detection for a long time. The intrusion detection method based on support vector machine is helpful to deal with the increasingly updated attack methods. From the perspective of data source and analysis, intrusion detection model can be divided into four categories: industrial traffic, industrial process, user behavior and malicious files. In the experiment, Matlab is used as the interface to complete the communication with the dual tank system. Through the derivation of the state between the areas, the area that is attacked can be accurately determined.

Key words: machine learning; support vector machine; anomaly detection

0 支持向量机(SVM)

支持向量机(Support vector machines, SVM)是一种基于统计学习理论的有效监督分类方法。该方法自 20 世纪 90 年代中期发展起来,由 Cortes 和 Vapnik 提出。支持向量机在许多问题中表现出了独有的优势,比如小样本、非线性及高维模式问题[1]。支持向量机能够适应基于有限样本信息的复杂模型。支持向量机算法能够在有限训练样本里找到最佳识别能力使其达到学习精度与样本的正确分类,具有良好的泛化能力。

支持向量机是基于定义在特征空间上的间隔最大线性分类器的一种二分类模型。因为间隔最大这一特征,支持向量机与感知机不同。由于该技巧的原因,支持向量机在实质上可以叫做线性分类器。支持向量机通过间隔最大化这一学习策略,将模型化为凸二次规划问题[2]。

(1)线性可分支持向量机

支持向量机(SVM)的目的是为了解决线性可分性问题而提出的,假设训练数据的样本量为 l,该样本可以表示为$\{(x_i, y_i), i = 1, 2, 3, \cdots, l\}$,训练数据样本集分为两个类别,如果 x_i 为第一类,则 $y_i = 1$;如果 x_i 为第二类,则 $y_i = -1$。

假设有一个超平面的分类,如公式(1)所示:

$$\omega x + b = 0 \tag{1}$$

数据样品可准确地分为两种类型,把不同类型

收稿日期: 2020 - 03 - 09

基金项目: 教育部科技发展中心产学研创新基金(2018A03007);四川省高等教育人才培养质量和教学改革项目(JG2018 - 1168);2019 年中国轻工业联合会教育工作分会立项课题(QGJY2019020);四川工商职业技术学院校级课题(20190016)

作者简介: 肖铮(1983 -),男,辽宁黑山人,副教授,硕士,主要研究方向为图形图像识别、机器学习。

的数据样品均打入超平面的两侧，使此数据标本集线可分离。即对于$(x_i, y_i) \in D$，若$y_i = +1$，则有$\omega^T x_i + b > 0$；若$y_i = -1$，则有$\omega^T x_i + b < 0$。即公式(2)所示。

$$\begin{cases} \omega^T x_i + b \geq +1, y_i = +1 \\ \omega^T x_i + b \leq -1, y_i = +1 \end{cases} \quad (2)$$

将样本空间中任意点x到超平面(ω, b)的距离记为公式(3)所示。

$$r = \frac{|\omega^T x_i + b|}{||\omega||} \quad (3)$$

从图1中可以看出，这个白色和黑色的圆分别代表两个样品，而H则区分平面。支持向量是SVM的分类决策中的重要部分。如果$|\omega^T x_i + b| = 1$，那么两类数据样本点之间的距离是$2\frac{|\omega^T x_i + b|}{||\omega||} = \frac{2}{||\omega||}$，当满足公式(2)的条件下求得的平面由公式(4)所示。使$\frac{2}{||\omega||}$最大，也就是$\frac{||\omega||^2}{2}$最小。即需要求解如下最小值问题:

$$\begin{cases} \min \frac{1}{2}||w||^2 \\ s.t.\ y_i(\omega^T x_i + b) \geq 1, i = 1, 2, \cdots, m. \end{cases} \quad (4)$$

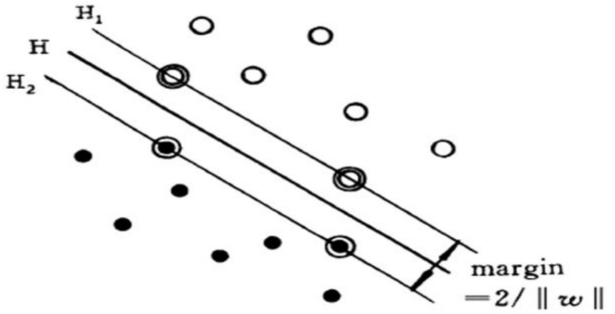

图1 支持向量机分类

(2)非线性支持向量机

在实际应用中，由于大部分问题不能用线性分开，所以可能被线性分开的支持向量机无法解决这些问题。使用非线性支持向量机可以提供解决方法。当$\phi: R^n \to H$，原输入空间的样品将比照到高维特征空间H中，在高维特征空间里找到最优的平面。

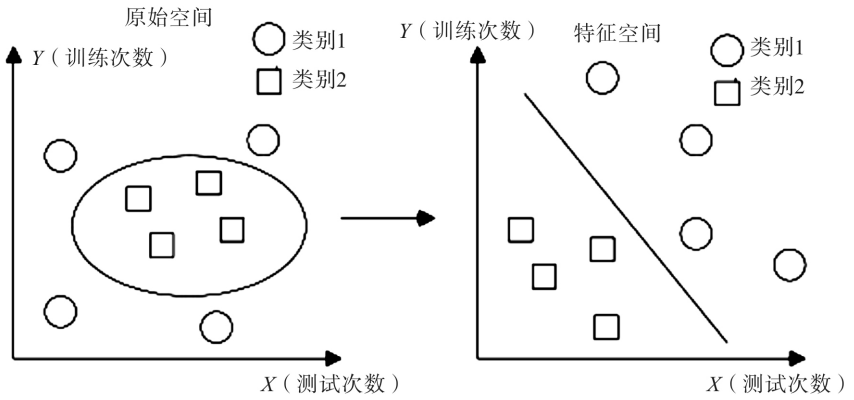

图2 非线性支持向量机模型

1 网络异常应用验证

1.1 实验原理

对于工业网络入侵检测的研究，其实就是对工业网络特征数据进行预测和分类的模式识别问题。根据数据样本的特点，使用标准C－SVM分析和判别数据，得到入侵检测结果。

1.2 实验流程

本实验采用的线性核函数，从数据集中选取800组正常数据和200组异常数据。每一组数据包括3个水箱的液位("level1"、"level2"、"level3")，两个阀门开度("valve1"、"valve2")，流量("flow")，压力("pressure")。正常数据样本标记为0类，异常数据样本标记为1类。测试数据中有140组异常数据，943组正常数据。惩罚参数C[3]指的是惩罚松弛变量，当其接近1时，该松弛变量对误分类的惩罚增大，这样调整参数可以提高测试训练集的准确率，缺点是泛化能力会较弱。当C值小的时候，对误分类的惩罚会减小，容错程度增加，泛化能力会增强。经过多次尝试和调整，本实验的取值是0.6。

使用学习过后的支持向量机检测测试数据，下图是运行的结果。可以看出，正常的数据集的检测结果全为正常数据。异常数据集的检测结果是，从

43 个数据开始,系统进入异常状态。说明发动攻击40s 以上后,系统才开始出现异常。该方法可以准确地检测出系统何时出现异常,判断系统是处在正常运行状态还是异常运行状态。

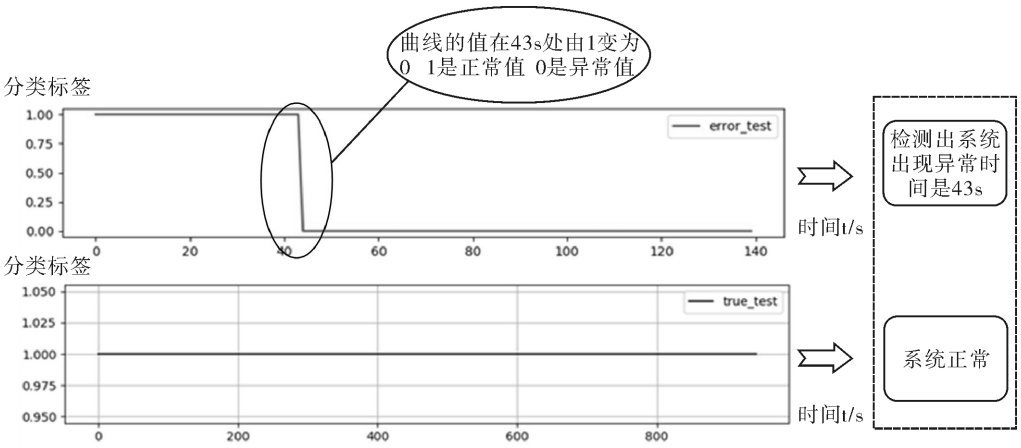

图 3 入侵检测响应曲线

2 结束语

使用支持向量机分析系统运行过程中的数据,对工业控制系统运行状态进行二分类的处理,该方法的优点是简洁、响应速度快、训练时间较短,可以迅速判断工控系统是否出现了异常。然而,这种处理方式不能够判断系统遭受了什么样的攻击,系统的哪个部分出现故障。工控系统数据的特点是其关联性较强、维度高,各种数据之间存在一定的关联。因此需要考虑其他的方式对系统的异常进行更深层的分析。

参考文献:

[1] 黄炜,刘坤. 面向信息特征模式识别的核方法研究综述[J]. 现代情报,2014,34(3):168-176.
[2] 孙晶,孙旭. 基于数据挖掘的电子商务推荐系统[J]. 仪器仪表用户,2011,18(6):88-90.
[3] 饶刚. 支持向量机(SVM)算法的进一步研究[D]. 重庆:重庆大学,2012.