Journal of System Simulation

Volume 36 | Issue 9 Article 4

9-15-2024

Adversarial Simulation Testing Algorithm for SVM Based on Multi-objective Evolutionary Optimization

Feixing Li
Chinese Flight Test Establishment, Xi'an 710089, China

Lining Xing
School of Electronic Engineering, Xidian University, Xi'an 710071, China

Yu Zhou
School of Electronic Engineering, Xidian University, Xi'an 710071, China

Adversarial Simulation Testing Algorithm for SVM Based on Multi-objective Evolutionary Optimization

Abstract

Abstract: Machine learning typically mines underlying patterns and rules from data, making it susceptible to phenomena such as overfitting and underfitting, which in turn affects the generalization and robustness of learning models. This paper explores the potential fragility and instability of SVM from the perspective of adversarial simulation testing. The adversarial simulation strategy employed involves selectively contaminating training sample labels to simulate an attack on the SVM classifier, thereby degrading its performance and testing its dependency on training samples. To explore the ceiling of performance degradation of an SVM classifier under the combination attack of different samples, the contradictory objectives of minimum attack cost-maximum attack effectiveness are designed, and a multiobjective optimization model is constructed for SVM simulation tests. This model is fundamentally a typical multi-objective combinatorial optimization problem that can be properly solved using multiobjective evolutionary algorithms to find a set of non-dominated solutions among the objectives, facilitating the investigation of the classifier's stability under the combination attack of different samples. Comparative experimental results on simulated and real datasets show that the proposed method can identify optimal attack sample combination at varying attack levels in a single run, and achieve more severe classification performance degradation, making it more suitable and effective for investigating the stability of classifiers comprehensively.

Keywords

adversarial simulation testing, label contamination, SVM, performance degradation, multiobjective optimization, non-dominated solutions

Recommended Citation

Li Feixing, Xing Lining, Zhou Yu. Adversarial Simulation Testing Algorithm for SVM Based on Multiobjective Evolutionary Optimization[J]. Journal of System Simulation, 2024, 36(9): 2016-2031.

第 36 卷第 9 期 系统仿真学报© Vol. 36 No. 9
2024 年 9 月 Journal of System Simulation Sept. 2024

基于多目标演化优化的SVM对抗仿真测试算法

李飞行[1]，邢立宁[2]，周宇[2*]
(1. 中国飞行试验研究院，陕西 西安 710089；2. 西安电子科技大学 电子工程学院，陕西 西安 710071)

摘要：机器学习通常从数据中挖掘潜在的模式与规则，容易受到数据的影响而产生诸如过拟合、欠拟合等现象，进而影响学习模型的泛化与鲁棒性能。*从对抗仿真测试的角度考察SVM可能存在的脆弱不稳定性*，采用的对抗仿真策略是通过选择性地污染训练样本标签，模拟攻击SVM分类器使其性能退化，以测试其对训练样本的依赖性。*为探究SVM分类器在不同样本组合攻击下的性能损失上限，设计了最小攻击代价–最大攻击成效这一对矛盾目标，构建了SVM仿真测试的多目标优化模型。该模型本质上是一种典型的多目标组合优化问题，可采用适当的多目标演化算法求解目标间的一组非支配解集，揭示分类器在不同样本组合攻击下的分类性能表现。*在人工及真实数据集上的仿真对比实验结果表明：所提方法能够一次性生成不同攻击水平下的最优攻击样本组合，取得最大的分类性能损失，更能全面测试SVM分类器性能的稳定性。

关键词：对抗仿真测试；污染标签；支持向量机；性能损失；多目标优化；非支配解集
中图分类号：TP306+.2 文献标志码：A 文章编号：1004-731X(2024)09-2016-16
DOI: 10.16182/j.issn1004731x.joss.24-0101
引用格式: 李飞行, 邢立宁, 周宇. 基于多目标演化优化的SVM对抗仿真测试算法[J]. 系统仿真学报, 2024, 36(9): 2016-2031.
Reference format: Li Feixing, Xing Lining, Zhou Yu. Adversarial Simulation Testing Algorithm for SVM Based on Multi-objective Evolutionary Optimization[J]. Journal of System Simulation, 2024, 36(9): 2016-2031.

Adversarial Simulation Testing Algorithm for SVM Based on Multi-objective Evolutionary Optimization

Li Feixing[1], *Xing Lining*[2], *Zhou Yu*[2*]
(1. Chinese Flight Test Establishment, Xi'an 710089, China; 2. School of Electronic Engineering, Xidian University, Xi'an 710071, China)

Abstract: Machine learning typically mines underlying patterns and rules from data, making it susceptible to phenomena such as overfitting and underfitting, which in turn affects the generalization and robustness of learning models. *This paper explores the potential fragility and instability of SVM from the perspective of adversarial simulation testing*. The adversarial simulation strategy employed involves selectively contaminating training sample labels to simulate an attack on the SVM classifier, thereby degrading its performance and testing its dependency on training samples. *To explore the ceiling of performance degradation of an SVM classifier under the combination attack of different samples, the contradictory objectives of minimum attack cost-maximum attack effectiveness are designed, and a multi-objective optimization model is constructed for SVM simulation tests. This model is fundamentally a typical multi-objective combinatorial optimization problem that can be properly solved using multi-objective evolutionary algorithms to find a set of non-dominated solutions among the objectives,*

收稿日期：2024-01-25 修回日期：2024-05-19
基金项目：陕西省重点科技创新团队项目(2023-CX-TD-07)；陕西省重点研发计划(2024GH-ZDXM-48)
第一作者：李飞行(1978-)，男，高工，硕士，研究方向为航空智能试验设计与评估。
通讯作者：周宇(1983-)，男，研究员，博士，研究方向为智能系统测试评估。

facilitating the investigation of the classifier's stability under the combination attack of different samples. Comparative experimental results on simulated and real datasets show that the proposed method can identify optimal attack sample combination at varying attack levels in a single run, and achieve more severe classification performance degradation, making it more suitable and effective for investigating the stability of classifiers comprehensively.

Keywords: adversarial simulation testing; label contamination; SVM; performance degradation; multi-objective optimization; non-dominated solutions

0 引言

数据采集、传输与存储技术的快速发展使人类迈入了大数据时代，也对数据的分析与处理技术提出了更高的要求。如何从海量数据中挖掘提取出对既定任务有效的信息，已经成为研究人员面临的主要问题。机器学习通过研究人的学习认知过程，能够从大量数据中挖掘潜在的特征、模式、规律，因其自主、高效特性和强大的分析处理能力，已逐渐成为分析处理大规模数据的主要方法，已在交通、电网、医疗、金融、互联网、刑侦等领域成功应用，并逐步向许多安全性要求极高的场景，如自动驾驶、故障诊断等扩展。然而，机器学习的深入应用也逐渐暴露了其潜在的脆弱不稳定性，例如，在图像识别领域，通过更改图像的少量局部特征便可导致分类器的错误输出[1]。因此，如何在可控环境中通过全面高效地仿真测试，评估机器学习系统的脆弱性，明确其潜在的安全漏洞，对其后续部署应用与性能优化至关重要[2-6]。

学者们基于机器学习对数据的适应性，从对抗攻击角度出发，分别在机器学习系统的训练和推理阶段仿真模拟攻击性样本[7]，测试评估机器学习系统的数据敏感性和决策盲点，这两类测试方式分别称为“投毒”攻击和“逃逸”攻击。投毒攻击往往通过污染训练数据，从源头破坏学习系统的准确性[8-10]，例如，在训练垃圾邮件过滤器时，对手通过将合法邮件标记为垃圾邮件，即可毒害过滤器[11]。“逃逸”攻击则通过构造测试特征寻找学习系统的决策盲点或失效区域[12]，例如，对手通过在垃圾邮件中添加不相干词汇来逃避垃圾邮件过滤。在机器学习安全领域，数据的分布往往随时间不断变化，为避免遗忘并保持机器学习系统在时间维度上的推理决策精度，学者们通常采用增量学习和定期重训练的方式保障机器学习系统对新、旧数据的良好适应性，但这也为对手“投毒”攻击提供了机会[13]。通过在增量或重训练样本中投毒，对手可以有效毒害机器学习系统的推理决策性能。鉴于“投毒”攻击的显著效果，本文将站在对抗角度，将机器学习的“投毒”攻击策略转化为其性能评估的仿真测试策略，通过构建高效的样本标签“投毒”攻击方法，全面考察机器学习系统在“投毒”攻击下的性能保持能力，进一步洞察系统数据敏感性和性能稳定性。

SVM作为一种重要的机器学习模型，被广泛应用于图像与文本分类、垃圾邮件检测、入侵检测等任务中，取得了显著的效果，故也常被用作对抗攻击的仿真测试对象，以考察攻击手段的有效性。文献[14]提出了自由和受限条件下的攻击模型，并站在对抗博弈的角度针对性设计了最优SVM学习策略。文献[15-16]使用梯度上升法构建攻击样本以最大化SVM分类器的仿真测试误差，并且针对标签翻转的对抗性训练样本建立了核矩阵校正的SVM学习策略。文献[17]将寻找样本标签翻转的最优组合问题建模为有约束的二次规划问题，试图在给定预算下最大化SVM分类器的仿真决策误差。随后，他们又在有限标签翻转的约束下，通过启发式搜索方法寻找最优标签翻转组合，以最大化SVM分类器的仿真测试误差[18]。文献[19]将训练集最优攻击问题建模为双层优化问题，在隐函数上使用梯度方法求解，并在SVM等

DOI: 10.16182/j.issn1004731x.joss.24-0101

分类器上进行了仿真验证。文献[20]研究了使用批量数据流更新训练SVM分类器时，有针对性攻击对分类器性能的影响。文献[21]在高维特征空间中采用粒子群算法寻找显著性特征攻击点，再映射回原空间构建对抗样本，证明了SVM的性能脆弱性。文献[22]提供了一种在训练数据集中加入代理数据集的攻击方法，用于对手无法完全掌握训练数据知识的攻击场景。文献[23]提出了3种“投毒”攻击方法，通过有意识的聚集投毒点并引入约束来确保“投毒”点规避防御算法的检测。文献[24]构建一种用于标签翻转攻击的专家模型，并通过仿真实验验证了其在不同数据集和深度神经网络上的有效性。文献[25]通过分析现有评估方法的缺陷，提出一套新的仿真攻击方法和测试评估框架，用于考察图神经网络对标签污染攻击的鲁棒性。

上述对抗攻击测试方法所能提供的攻击代价和攻击收益信息均存在一定的单一性，无法提供攻击代价—攻击成效之间的演变信息，不利于掌握SVM分类器在不同攻击水平下的性能表现。为了在不同的攻击代价下均取得最大的攻击收益，以全面评估SVM分类器在不同攻击强度下的性能稳定性，本文提出最小攻击代价－最大攻击成效这一矛盾的优化目标，构建SVM性能稳定性测试的多目标优化模型。通过设计高效的演化求解算法，可一次性获得目标间的一组非支配解集及对应的目标向量，分别表征不同攻击水平下的最优攻击样本组合和最大分类器性能损失，为分类器的稳定性研究提供更全面的测试评估信息。

1 SVM分类器与对抗仿真测试

1.1 SVM分类器

SVM作为一种重要的机器学习模型，常被用作算法仿真测试的对象。它的基本优化模型为

$$\begin{aligned}&\arg\min_{\boldsymbol{w}}\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}\\&\text{s.t }y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\geqslant 1,\ \forall i=1,2,\cdots,N\end{aligned}\tag{1}$$

式中：$\{\boldsymbol{x}_i,y_i\}$为训练样本及对应标签；N为训练样本数；$\boldsymbol{w}$、b为模型参数。考虑到实际应用中由于噪声等原因造成的数据线性不可分情况，可以在支持向量(即满足$y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)=1$的正、负样本)中引入松弛变量$\xi_i$，将式(1)转化为

$$\begin{aligned}&\min_{\boldsymbol{w},\,b}\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+C\sum_{i=1}^{N}\xi_i\\&\text{s.t. }y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\geqslant 1-\xi_i,\\&\quad\xi_i\geqslant 0,\ \forall i=1,2,\cdots,N\end{aligned}\tag{2}$$

式中：$C\geqslant 0$为控制因子，表示对松弛变量的容忍程度，C越大表示对松弛变量ξ_i的容忍度越低。模型(2)的解由式(3)给出:

$$\begin{aligned}&\boldsymbol{w}=\sum_{i=1}^{N}\alpha_i y_i\boldsymbol{x}_i\\&\text{s.t. }\sum_{i=1}^{N}\alpha_i y_i=0,\\&\quad 0\leqslant\alpha_i\leqslant C,\ i=1,2,\cdots,N\end{aligned}\tag{3}$$

式中：参数α_i为拉格朗日对偶问题的解。

对于给定的测试样本$\boldsymbol{x}$，它的类别可由判别函数式(4)给出：

$$F(\boldsymbol{x})=\operatorname{sign}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)=\operatorname{sign}\left(\sum_{i=1}^{N}\alpha_i y_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}+b\right)\tag{4}$$

当训练样本集在原空间中线性不可分时，通常采用适当的非线性函数$\phi(\cdot)$将其映射到高维特征空间中以增加可分行。假设$\{\phi(\boldsymbol{x}_i),\ i=1,2,\cdots,N\}$在高维空间中线性可分，则其SVM分类器遵循上述推导过程，测试样本$\boldsymbol{x}$的类别由判别函数给出：

$$F(\boldsymbol{x})=\operatorname{sign}\left(\sum_{i=1}^{N}\alpha_i y_i[\phi(\boldsymbol{x}_i)]^{\mathrm{T}}\phi(\boldsymbol{x})+b\right)\tag{5}$$

式(5)中存在映射函数的内积运算，因此可以直接构造核函数$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=[\phi(\boldsymbol{x}_i)]^{\mathrm{T}}\phi(\boldsymbol{x}_j)$，将式(5)转化为

$$F(\boldsymbol{x})=\operatorname{sign}\left(\sum_{i=1}^{N}\alpha_i y_i\kappa(\boldsymbol{x}_i,\boldsymbol{x})+b\right)\tag{6}$$

即可在原空间中判断给定样本的所属类别。

1.2 对抗仿真测试

本文聚焦于研究高效的样本标签“投毒”攻

击方法，通过仿真测试SVM分类器在该类攻击下的推理差异，评估其数据敏感性和性能稳定性。因此，需要率先建立标签翻转攻击前后的推理差异度量方法。标签翻转攻击通过篡改部分训练样本的标签，从而在训练阶段毒害分类器，从源头上破坏分类器的准确性。具体地，假定训练数据集$\{\boldsymbol{x}_i,y_i\}_{i=1}^N\in X\{-1,1\}$，其标签只能在−1和1中取值，对手通过将标签从−1篡改为1或者从1篡改为−1来“毒害”训练样本。这一过程可以通过构造0-1向量$\boldsymbol{\varepsilon}=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_N)$, $\varepsilon_i\in\{0,1\}$来实现，翻转后的标签为$\tilde{y}_i=y_i(1-2\varepsilon_i)$，当$\varepsilon_i=0$时，样本$\boldsymbol{x}_i$的标签$y_i$不翻转，当$\varepsilon_i=1$时，样本$\boldsymbol{x}_i$的标签$y_i$翻转为$-y_i$。

根据1.1中SVM分类器的推导过程可得，由篡改后的训练样本集$\{\boldsymbol{x}_i,\tilde{y}_i\}_{i=1}^N$所确定的SVM分类器，其判别函数为

$$\tilde{F}(\boldsymbol{x})=\mathrm{sign}\left(\sum_{i=1}^{N}\alpha_i y_i(1-2\varepsilon_i)\kappa(\boldsymbol{x}_i,\boldsymbol{x})+b\right) \tag{7}$$

假定测试样本集包含M个由样本$\boldsymbol{z}_i$及标签r_i构成的样本对，即$\{\boldsymbol{z}_i,r_i\}_{i=1}^M$，通过比较式(6)和式(7)，可得标签翻转攻击前后的分类变化率为

$$R_z=\frac{\sum_{i=1}^{M}\mathrm{Is}(\tilde{F}(\boldsymbol{z}_i)\neq F(\boldsymbol{z}_i))}{M} \tag{8}$$

式中：指示函数$\mathrm{Is}(\cdot)=1$，当且仅当$\tilde{F}(\boldsymbol{z}_i)\neq F(\boldsymbol{z}_i)$。式(8)计算了受攻击前后SVM分类器的推理结果差异，可用于度量攻击对分类器的影响程度。

2 多目标演化建模

单一攻击代价下的SVM分类性能表现并不足以反映分类器的抗攻击能力。为了全面评估SVM的性能稳定性，需通过仿真测试其在不同样本组合攻击下的分类性能变化，掌握不同攻击强度下SVM分类器的抗攻击效果。显然，攻击强度和攻击效果是一对相关量，当攻击强度较小时，SVM的推理性能难以发生较大的转变，攻击效果不明显；当攻击效果明显，SVM推理性能发生剧烈变化时，所需要的攻击强度往往较大。因此，通过改变样本攻击的强度，测试分类器在不同攻击强度下的推理性能变化，对全面评估分类器的性能稳定性具有重要意义。

2.1 最小攻击代价-最大攻击成效建模

由于样本标签翻转攻击具有鲜明的二值特性，其攻击强度的高低可以直接采用翻转标签的数目来表征，因此可以构造目标函数来刻画标签翻转攻击的代价最小化问题：

$$\min_{\boldsymbol{\varepsilon}} f_1=\|\boldsymbol{\varepsilon}\|_1 \tag{9}$$

对于攻击效果的强弱，本文从攻击前后学习系统的推理差异性方面构造目标函数。由于标签翻转攻击的根本目的是从源头上毒害SVM分类器，使之产生有别于原分类器的推理差异。因此，攻击前后的SVM分类器在未知样本集上的推理差异越大，说明攻击方案产生的攻击效果越好。假定样本集为$\{\boldsymbol{c}_i\}_{i=1}^M$，且每个样本的标签未知，则标签翻转攻击前后SVM分类器的推理差异最大化问题可表征为

$$\max_{\boldsymbol{\varepsilon}} R_c=\frac{\sum_{i=1}^{M}\mathrm{Is}(\tilde{F}(\boldsymbol{c}_i)\neq F(\boldsymbol{c}_i))}{M} \tag{10}$$

为了保障机器学习系统的任务针对性，式(10)中的样本集$\{\boldsymbol{c}_i\}_{i=1}^M$不能随意选取，避免脱离任务而实施无实意的攻击。因此，样本集$\{\boldsymbol{c}_i\}_{i=1}^M$应当尽可能与训练数据$\{\boldsymbol{x}_i\}_{i=1}^N$有相似的分布，这样搜索出的攻击方案才能在原任务上产生尽可能大的攻击效果。问题(10)中，R_c取值在[0,1]之间，可通过构造函数$f_2=1-R_c$将其转化为等价目标最小化问题。

为了实现在不同的攻击代价下均产生最大的攻击效果，可将式(9)和式(10)的等价最小化问题联立，得到SVM攻击测试的多目标优化模型

$$\min_{\boldsymbol{\varepsilon}} \boldsymbol{F}(\boldsymbol{\varepsilon})=\left\{\|\boldsymbol{\varepsilon}\|_1,1-\frac{\sum_{i=1}^{M}\mathrm{Is}(\tilde{F}(\boldsymbol{c}_i)\neq F(\boldsymbol{c}_i))}{M}\right\} \tag{11}$$

该模型致力于搜索代价最低且成效最高的攻击测试方案。根据多目标优化理论，在攻击成本和成效之间应存在一组非支配解。

https://dc-china-simulation.researchcommons.org/journal/vol36/iss9/4
DOI: 10.16182/j.issn1004731x.joss.24-0101

2.2 演化搜索算法

模型(11)本质上是一个多目标组合优化问题，其决策变量$\boldsymbol{\varepsilon}$为0-1向量，故适合采用多目标演化算法搜索目标间的一组非支配解集，得到SVM分类器在不同攻击成本下的最大推理差异。演化算法作为一种高鲁棒性和强适用性的全局优化算法，具有自组织、自适应、自学习、隐并行的特点，已经广泛应用于资源调度[26-28]、任务规划[29-31]等复杂问题。作为演化多目标优化的代表性算法，基于分解的演化多目标优化(multiobjective evolutionary algorithm based on decomposition, MOEA/D)方法可以将多目标优化问题分解成一系列对多个目标具有不同侧重的单目标优化子问题，利用邻居子问题间的协同进化机制，搜索原问题的一组非支配解，并通过调整子问题的加权参数改善解的分布均匀性。因此，本文采用MOEA/D框架[32]求解模型(11)，其结构如算法1所示。算法1中种群$\{\boldsymbol{I}_i\}_{i=1}^{P}$的每个个体$\boldsymbol{I}_i$代表着决策变量$\boldsymbol{\varepsilon}$的一个可行解，MOEA/D算法最终可以得到决策变量$\boldsymbol{\varepsilon}$的一组非支配解集，以及这组解集对应的帕累托前沿。

算法1：MOEA/D算法框架

输入：种群大小P，邻居子问题数目T，样本集$\{\boldsymbol{x}_i,y_i\}_{i=1}^{N}$，$\{\boldsymbol{c}_i\}_{i=1}^{M}$

输出：非支配解集S_{pareto}及其对应目标向量集F_{pareto}

在(0, 1)和(1, 0)之间均匀产生P个权重向量$\{\boldsymbol{\lambda}_i\}_{i=1}^{P}$

用边界交叉法将式(11)聚合为一组单目标子问题$\{S_{\text{pb}}^{i}\}_{i=1}^{P}$

基于欧式距离查找每个子问题S_{pb}^{i}的T个邻居子问题并记录其索引$B\{i\}$

初始化种群$\{\boldsymbol{I}_i\}_{i=1}^{P}$、集合$S_{\text{pareto}}=F_{\text{pareto}}=\varnothing$和理想点$\boldsymbol{z}^*=\min\{\boldsymbol{F}(\boldsymbol{I}_i)|i=1,2,\cdots,P\}$

while 不满足终止条件 do

for $i=1\to P$ do

查找第i个子问题S_{pb}^{i}的邻居索引$B\{i\}$

从邻居子问题的解中选择父代个体，采用多点交叉和单点变异产生子代个体$\hat{\boldsymbol{I}}_i$

更新$\boldsymbol{z}^*=\min\{\boldsymbol{z}^*,\boldsymbol{F}(\boldsymbol{I}_i)\}$；

若$g(\hat{\boldsymbol{I}}_i|\boldsymbol{\lambda}_j,\boldsymbol{z}^*)\leqslant g(\boldsymbol{I}_j|\boldsymbol{\lambda}_j,\boldsymbol{z}^*),j\in B\{i\}\cup i$，

则$\boldsymbol{I}_j=\hat{\boldsymbol{I}}_i,\ \boldsymbol{F}(\boldsymbol{I}_i)=\boldsymbol{F}(\hat{\boldsymbol{I}}_i)$

end

end while

从$\{\boldsymbol{I}_i\}_{i=1}^{P}$和$\{\boldsymbol{F}(\boldsymbol{I}_i)\}_{i=1}^{P}$中分别提取$S_{\text{pareto}}$，$F_{\text{pareto}}$。

在(0,1)和(1, 0)之间均匀产生权重向量$\{\boldsymbol{\lambda}_i\}_{i=1}^{P}$之后，采用边界交叉法[32]将模型(11)转化为一系列子问题，并通过子问题解的演化来搜索模型(11)的非支配解集。具体子问题$\{S_{\text{pb}}^{i}\}_{i=1}^{P}$的构建方式和优化原理为

$$\begin{aligned}&S_{\text{pb}}^{i}:\ \min g(\boldsymbol{I}|\boldsymbol{\lambda}_i,\boldsymbol{z}^*)=d_1+\theta d_2\\&\quad\text{s.t. }\boldsymbol{I}\in\{0,1\}^N\end{aligned}\tag{12}$$

式中：$\boldsymbol{z}^*$为理想点，是用于引导模型(11)中目标优化的理想位置，令$\boldsymbol{z}^*$等于种群已搜索到的目标函数的最小值；θ为罚参，用于控制对d_2的重视程度。

$$d_1=\|(\boldsymbol{F}(\boldsymbol{I})-\boldsymbol{z}^*)^{\mathrm{T}}\boldsymbol{\lambda}_i\|/\|\boldsymbol{\lambda}_i\|\tag{13}$$

$$d_2=\|\boldsymbol{F}(\boldsymbol{I})-(\boldsymbol{z}^*+d_1\boldsymbol{\lambda}_i/\|\boldsymbol{\lambda}_i\|)\|\tag{14}$$

如图1所示，d_1、d_2分别为$\boldsymbol{F}(\boldsymbol{I})$在$\boldsymbol{\lambda}_i$方向和$\boldsymbol{\lambda}_i$正交方向上与理想点$\boldsymbol{z}^*$的距离。因此，模型(12)的目的是在目标空间中沿着$\boldsymbol{\lambda}_i$的反方向搜索子问题S_{pb}^{i}的解，并通过θ控制搜索方向在$\boldsymbol{\lambda}_i$正交方向上的偏移程度。如此，通过一系列均匀产生权重向量$\{\boldsymbol{\lambda}_i\}_{i=1}^{P}$，模型(12)可以在目标空间中沿着指向$\boldsymbol{z}^*$的不同方向搜索到模型(11)的一组非支配解集。

非支配解集的演化搜索是在子问题的邻域内进行的，具体搜索过程如下：对于每个子问题S_{pb}^{i}，根据其对应权重$\boldsymbol{\lambda}_i$的欧式距离查找最近的T个邻域子问题，并将其索引存储在集合$B\{i\}$中；对搜索种群、非支配解集、目标向量集与理想点进行初始化，为了生成含有不同标签翻转个数的初始种群，针对性产生含有不同1元素数目的P个N维0-1向量对种群赋初始值；采用遗传操作对种群进行迭代，并对相关变量进行更新，直至满足程序终止条件；从末代种群及其目标函数集中提取出非支配解集S_{pareto}及对应的帕累托前沿F_{pareto}即可。

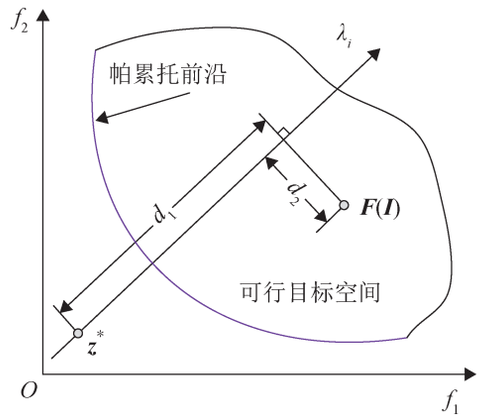

图1　边界交叉法构造子问题原理

Fig. 1　Illustration on penalty-based boundary intersection method for constructing subproblems

算法1的详细遗传操作过程如下。

首先，从子问题S_{pb}^i的邻居索引中选定2个邻居子问题$S_{\text{pb}}^j, S_{\text{pb}}^k (j,k \in B\{i\})$，将两者的解$\boldsymbol{I}_j$与$\boldsymbol{I}_k$作为父代个体，通过多点交叉算子产生新个体：

$$\bar{I}_{i,t}=\begin{cases} I_{j,t}, & r^t \leqslant 0.5 \\ I_{k,t}, & \text{其他} \end{cases} \tag{15}$$

式中：$\bar{I}_{i,t}, I_{j,t}, I_{k,t}, r_t$分别为$\bar{\boldsymbol{I}}_i$、$\boldsymbol{I}_j$、$\boldsymbol{I}_k$、$\boldsymbol{r}$的第$t$个元素，$\boldsymbol{r}$是与个体等长且元素值在$(0,1)$内的随机向量。

算子(15)可以令$\boldsymbol{I}_j$与$\boldsymbol{I}_k$进行充分的基因交换以产生$\bar{\boldsymbol{I}}_i$。对于$\bar{\boldsymbol{I}}_i$，以微小的概率对其基因位点进行变异，以产生进化的原始材料，具体公式为

$$\hat{I}_{i,t}=1-\bar{I}_{i,t}; \text{If rand}(0,\ 1) \leqslant\ 1/N \tag{16}$$

$\bar{\boldsymbol{I}}_i$基于式(16)的期望变异基因数目为1。

在子代个体$\hat{\boldsymbol{I}}_i$产生之后，需要先更新理想点$\boldsymbol{z}^*$，以保障其存储当前搜索到的最小目标值，从而发挥其导向作用。然后对于子问题S_{pb}^i及其邻居子问题，评估子代个体$\hat{\boldsymbol{I}}_i$相对于上一代个体的优劣，若子代个体的目标函数值更小，则用其取代上一代个体，同时更新对应的目标向量。在完成种群迭代之后，需从末代种群$\{\boldsymbol{I}_i\}_{i=1}^P$中筛选出非支配解集$S_{\text{pareto}}$，并从$\{\boldsymbol{F}(\boldsymbol{I}_i)\}_{i=1}^P$中筛选出对应的目标向量集$F_{\text{pareto}}$。

$$\begin{cases} S_{\text{pareto}}=\{\boldsymbol{I}_l | \neg \exists \boldsymbol{I}_i,\ \boldsymbol{I}_i \prec \boldsymbol{I}_l,\ \ i,l=1,2,\cdots,P\} \\ F_{\text{pareto}}=\{\boldsymbol{F}(\boldsymbol{I}_l) | \boldsymbol{I}_l \in S_{\text{pareto}}\} \end{cases} \tag{17}$$

当且仅当

$$\begin{aligned} &\forall\, d=1,2\ \ F_d(\boldsymbol{I}_i) \leqslant F_d(\boldsymbol{I}_l) \text{且} \\ &\exists\, d \in \{1,2\}\, F_d(\boldsymbol{I}_i) < F_d(\boldsymbol{I}_l) \end{aligned} \tag{18}$$

式中：$F_d(\cdot)$为$\boldsymbol{F}(\cdot)$的第d个目标。

3　仿真实验设计与结果分析

通过仿真与真实数据对比实验验证所提SVM性能攻击测试方法的有效性。首先，对方法的相关参数进行设置。其次，介绍标签翻转攻击的代表性算法，用于对比实验。最后，在选定数据集上依次开展实验，全面验证所提方法的攻击测试性能。由于实验考察各算法的攻击效果，因此，采用分类器的推理误差作为攻击结果的评估指标：

$$\delta_{idx}=\sum_{i=1}^{M}(\tilde{F}(\boldsymbol{t}_i) \neq \text{label}i)/M \tag{19}$$

式中：$\sum_{i=1}^{M}(\tilde{F}(\boldsymbol{t}_i) \neq \text{label}i)$为分类器在测试集$\{\boldsymbol{t}_i\}_{i=1}^M$上推理结果与实际标签不一致的数目。为了综合衡量模型在二分类问题上的准确性，采用F1分数作为补充性参考指标。F1分数是查准率和召回率的调和平均数，用于衡量模型对正类样本的推理准确性和全面性，具体计算方式见文献[33]，其数值越小表明分类器性能越差。

3.1　参数设置

本文方法的参数均与MOEA/D算法有关，种群大小为200、邻居子问题数目为20、交叉概率为1、变异概率为0.95、罚参θ的值设置为1，并采用控制变量法微调其中的单一参数，发现实验结果对参数并不敏感。

3.2　标签翻转攻击类对比算法

实验采用几种代表性标签翻转方法攻击SVM分类器，以验证所提算法在标签翻转类攻击方法中的优越性，突出其攻击测试能力：①随机标签翻转攻击[17](random label flip attack, RLFA)，从训练样本集中随机选择样本进行标签翻转，即随机

DOI: 10.16182/j.issn1004731x.joss.24-0101

引入标签噪声；②就近优先翻转攻击[17](near-first label flip attack, NLFA)，从训练样本集中优先挑选与原SVM分类超平面距离近的样本，进行标签翻转；③就远优先翻转攻击[17] (far-first label flip attack, FLFA)，从训练样本集中优先挑选与原SVM分类超平面距离远的样本，进行标签翻转；④对抗标签翻转攻击[18] (adversarial label flip attack, ALFA)：从对抗的角度寻找满足攻击预算的最大标签翻转子集，以最大化SVM分类器在原训练集上的经验损失[13];⑤连续标签松弛ALFA[18] (ALFA with continuous label relaxation, ALFA-CR)，将对抗标签翻转攻击的子集选择问题转化为连续变量优化问题，并采用梯度上升法求解[14]；⑥超平面倾斜ALFA[18] (ALFA based on hyperplane tilting, ALFA-Tilt)，从对抗角度寻找最优的标签翻转组合，以最大化受攻击前后SVM分类超平面的夹角[14]；⑦关联聚类ALFA[18] (ALFA based on correlated clusters, ALFA-CC)，随机产生标签翻转聚类，并随机对聚类进行变异，采用贪心算法寻找能产生最大经验损失的聚类作为标签翻转组合[14]。

以上算法包含了规则类算法、交替迭代算法、梯度算法和启发式贪心算法，能够较为充分地对比验证所提方法的攻击测试性能。其中，ALFA、ALFA-CR、ALFA-Tilt、ALFA-CC的具体参数设置与文献出处保持一致。

3.3 数值仿真实验与结果分析

3.3.1 应用仿真案例

基于一组线性可分的二维合成数据集展开，该数据集包含400个训练样本、600个测试样本，主要用于阐述所提方法的具体应用仿真实施过程。为了方便描述，将本文的攻击方法记为ALFA-MO(ALFA with multiobjective optimization)。由于ALFA-MO算法在优化过程中需要与训练样本分布类似的样本集，以评估种群中个体的适应度。所以，可以直接将测试样本集的标签隐去，用于计算受攻击前后SVM分类器的推理性能差异。算法得到模型(11)的帕累托前沿如图2所示。从图中可以看出，ALFA-MO算法所得帕累托前沿对应目标间的一组非支配解集，提供了不同攻击成本下对SVM分类器的最大攻击效果，且随着攻击成本的提升，攻击前后SVM分类器的推理差异$(1-F_2(\varepsilon))$在逐渐增大，故该帕累托前沿表征了SVM分类性在不同攻击强度下的抗攻击能力。

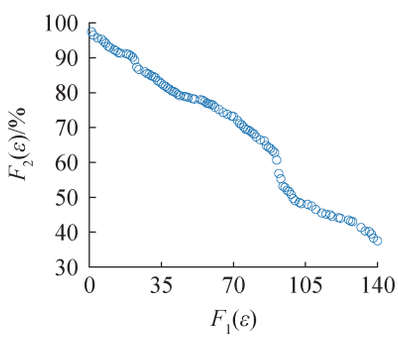

图2 ALFA-MO所得模型(11)的帕累托前沿
Fig. 2 Pareto front of model (11) obtained by ALFA-MO

图3为SVM分类器测试误差随ALFA-MO算法攻击强度的变化情况。其中，图3(b)为SVM分类器受攻击前的分类边界，其测试误差为2%，图3(c)~(f)分别为攻击样本数等于10、20、30、40时的最佳攻击样本组合以及攻击后的分类边界，对应测试误差逐渐增加，分别为7%、9.33%、15.17%、19.5%。

3.3.2 人工数据仿真对比实验

本实验将在线性可分模式、抛物线可分模式和环形可分模式的人工数据上对所提算法进行对比实验验证。由于采用不同核函数的SVM分类器具有不同的数据适应性和分类性能，为了充分衡量所提算法的攻击效果，实验将对采用不同核函数的SVM分类器进行攻击，包括线性核函数SVM、径向基核函数SVM、多项式核函数SVM，以全面考察攻击算法对SVM分类器的攻击能力。由于不同攻击水平下不同算法的攻击效果难以全部展

示，本实验仅展示攻击样本数目在训练样本中占比 10% 时，各算法对不同核函数 SVM 的攻击效果。

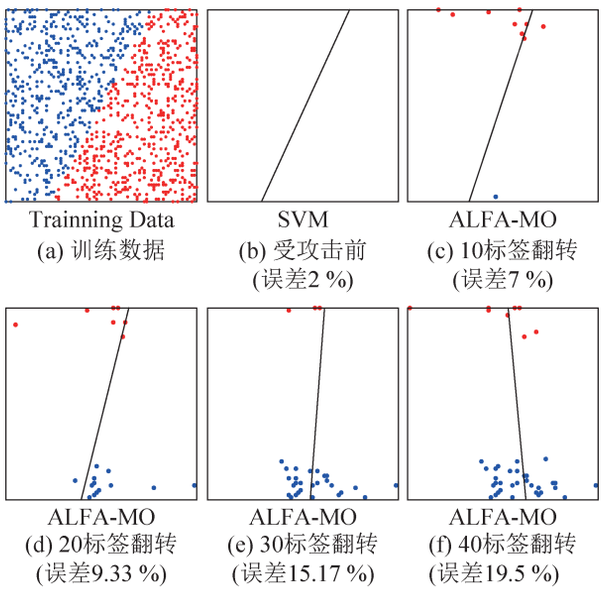

图3　ALFA-MO 算法攻击效果随攻击样本数目的变化
Fig. 3　Attack effect of ALFA-MO changes with number of label flips

第一组实验仍然在线性可分模式的数据集上展开，该数据集和第 3.3.1 节中的数据集相同。图 4 为各算法在该数据集上对采用线性核 SVM 的攻击效果。其中，图 4(c)~(f)分别为各攻击算法在攻击样本数占比 10% 时的最佳攻击样本组合以及攻击后的分类边界(图 5~11 的(c)~(f)同此含义)。从图中可知，受攻击前 SVM 分类器在测试数据集上的分类误差为 2%，在受到不同算法的攻击之后，SVM 的分类误差均有增加。其中，ALFA、ALFA-Tilt、FLFA 的攻击效果较为明显，导致攻击后的 SVM 分类误差均超过 10%。相较之下，ALFA-MO 算法的攻击能力最强，导致攻击后的 SVM 分类误差达到 19%，比攻击效果最好的对比算法 ALFA-Tilt 提升了 2.67 个百分点。此外，从表 1 中亦可以看出，ALFA-MO 算法在线性可分模式的数据集上对采用线性核 SVM 的分类器攻击效果最好，因为受其攻击后的分类器所得 F1 分数较对比算法更低。

图 5 为各算法在该数据集上对采用径向基核函数 SVM 的攻击效果。从图中可知，受攻击之前 SVM 分类器在测试数据集上的分类误差为 1.67%，在受到不同算法的攻击之后，SVM 的分类误差既有增加亦有降低。其中，ALFA、ALFA-CR、ALFA-Tilt 的攻击效果较为明显，导致攻击后的 SVM 分类误差均超过 10%，ALFA-MO 算法的攻击能力最强，导致攻击后的 SVM 分类误差达到 21%。其他对比算法的攻击效果则比较差，FLFA 算法攻击后的分类器精度甚至还有提升，说明其攻击无效。此外，从表 1 中 F1 分数亦可以看出，ALFA-MO 算法在线性可分模式的数据集上对采用径向基核 SVM 的分类器攻击效果最好。

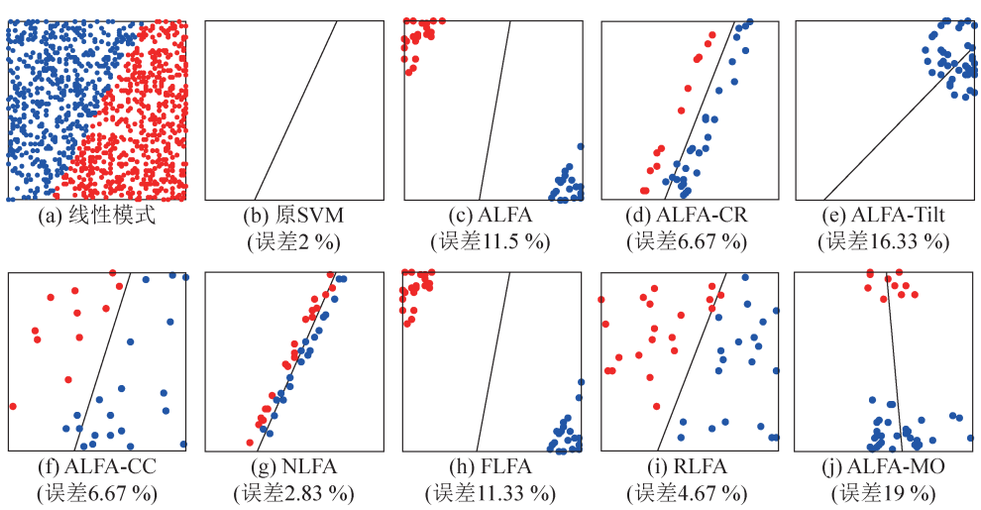

图4　各算法在线性可分数据上对线性核 SVM 的攻击效果
Fig. 4　Attack effect of all algorithms on SVM with linear kernel based on linear separable data

https://dc-china-simulation.researchcommons.org/journal/vol36/iss9/4
DOI: 10.16182/j.issn1004731x.joss.24-0101

第 36 卷第 9 期
2024 年 9 月
系统仿真学报
Journal of System Simulation
Vol. 36 No. 9
Sept. 2024

图 6 为各算法在该数据集上对采用多项式核函数 SVM 的攻击效果。从图中可知，受攻击之前 SVM 分类器在测试数据集上的分类误差为 2.33%，除 NLFA 之外，在受到不同算法的攻击之后，SVM 的分类误差均有增加。其中，ALFA、ALFA-Tilt、ALFA-CC 和 FLFA 的攻击效果较为明显，导致攻击后的 SVM 分类误差均超过 10%。ALFA-MO 算法的攻击能力最强，导致攻击后的 SVM 分类误差达到 24.33%，相较于 ALFA-Tilt 提升了 5%。此外，从表 1 中 F1 分数亦可以看出，ALFA-MO 算法在线性可分模式的数据集上对采用多项式核 SVM 的分类器攻击效果最好。

综上可知，ALFA-MO 在线性可分数据上对采用几种常见核函数的 SVM 分类器均能产生较对比算法更优的攻击效果。相对而言，部分对抗性和非对抗性对比方法对分类器的攻击效果并不稳定，如 ALFA-CC、FLFA。这表明在线性可分数据上，ALFA-MO 算法更适合于 SVM 分类器的抗攻击性能测试，以全面评估其在不同攻击强度下的性能稳定性。

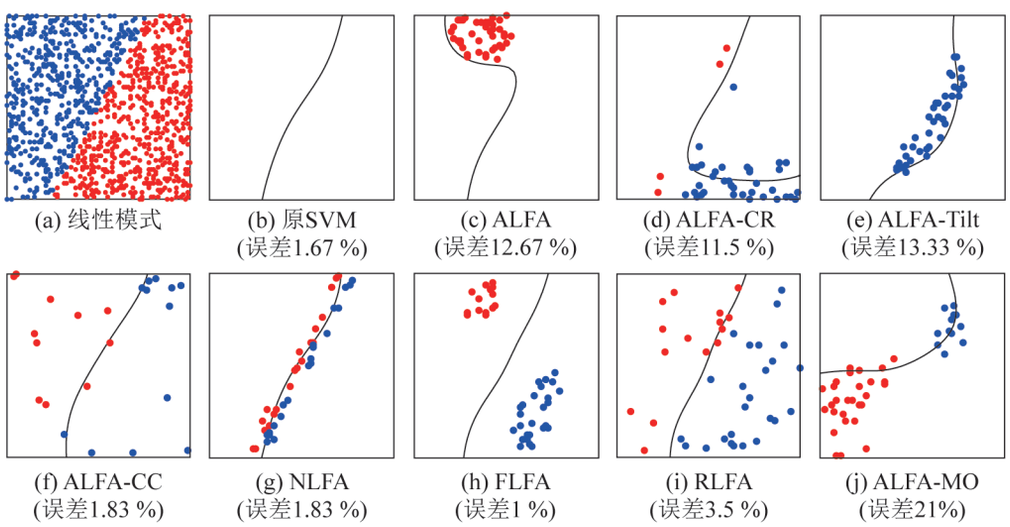

图 5　各算法在线性可分数据集上对径向基核 SVM 的攻击效果

Fig. 5　Attack effect of all algorithms on SVM with RBF kernel based on linear separable data

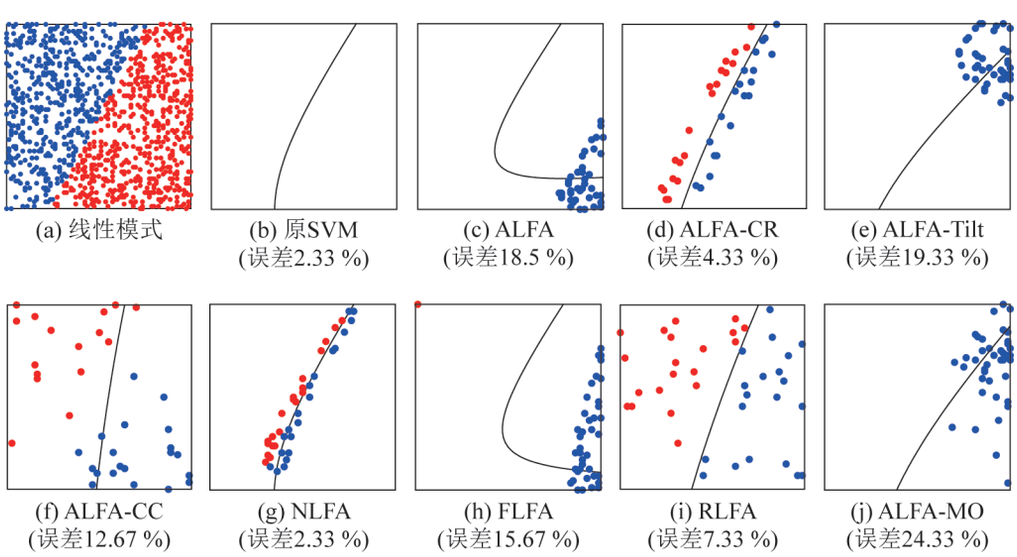

图 6　各算法在线性可分数据上对多项式核 SVM 的攻击效果

Fig. 6　Attack effect of all algorithms on SVM with polynomial kernel based on linear separable data

表1　不同算法在不同数据集上攻击采用不同核函数SVM分类器所得F1分数

Table 1　F1 scores obtained by different methods attacking SVM with different kernels on different data sets

模式	核函数	原模型	ALFA	ALFA-CR	ALFA-Tilt	ALFA-CC	NLFA	FLFA	RLFA	ALFA-MO
线性可分模式	线性核	0.987 8	0.950 1	0.928 8	0.891 6	0.928 0	0.975 0	0.915 1	0.955 5	0.822 3
	径向基核	0.976 1	0.890 5	0.952 4	0.838 5	0.942 1	0.952 1	0.966 7	0.974 4	0.820 2
	多项式核	0.979 5	0.788 2	0.934 7	0.847 8	0.902 2	0.974 8	0.808 9	0.937 9	0.737 8
抛物可分模式	线性核	0.760 1	0.656 7	0.762 1	0.656 3	0.761 4	0.760 6	0.755 7	0.760 8	0.640 3
	径向基核	0.972 9	0.899 4	0.953 3	0.924 6	0.924 6	0.956 9	0.912 2	0.965 1	0.823 3
	多项式核	0.871 3	0.618 5	0.820 4	0.762 7	0.810 2	0.871 7	0.616 5	0.835 5	0.603 0
环形可分模式	径向基核	0.956 7	0.932 3	0.952 4	0.911 9	0.887 1	0.943 9	0.939 9	0.948 0	0.785 4
	多项式核	0.763 8	0.397 2	0.589 2	0.469 4	0.727 6	0.745 1	0.397 2	0.721 1	0.357 1

第 2 组实验将在抛物线可分数据集上展开，为了便于展示，数据集的构造仍然在二维空间中进行，该数据集同样包含 400 个训练样本、600 个测试样本。图 7 为各算法在该数据集上对采用线性核 SVM 的攻击效果。从图中可知，受攻击之前 SVM 分类器在测试数据集上的分类误差为 23.67%，说明采用线性核函数的 SVM 分类器并不适合此数据集。尽管如此，在受到攻击后，能够引起 SVM 分类误差明显增加的算法有 ALFA、ALFA-Tilt、ALFA-MO，而 ALFA-MO 对 SVM 的攻击能力明显强于前 2 种对比算法，比攻击效果最好的对比算法 ALFA-Tilt 提升了 17.24 个百分点。从图中甚至能看出，ALFA-MO 攻击后的 SVM 分类器分类边界已经脱离数据的分布区域。此外，从表 1 中的 F1 分数亦可以看出，ALFA-MO 算法在抛物线可分模式的数据集上对采用线性核 SVM 的分类器攻击效果最好。

图 8 为各算法在该数据集上对采用径向基核 SVM 的攻击效果。从图中可知，受攻击之前 SVM 分类器在测试数据集上的分类误差为 3.67%，说明采用径向基核函数的 SVM 分类器适用于此数据集。在受到不同算法的攻击之后，SVM 的分类误差均有增加。其中，ALFA 和 ALFA-MO 算法的攻击效果较为明显，导致攻击后的 SVM 分类误差均超过 10%，ALFA-MO 较 ALFA 攻击后产生的分类误差提升了 11.5 个百分点，显著领先各对比算法。此外，从表 1 中 F1 分数亦可以看出，ALFA-MO 算法在抛物线可分模式的数据集上对采用径向基核 SVM 的分类器攻击效果最好。

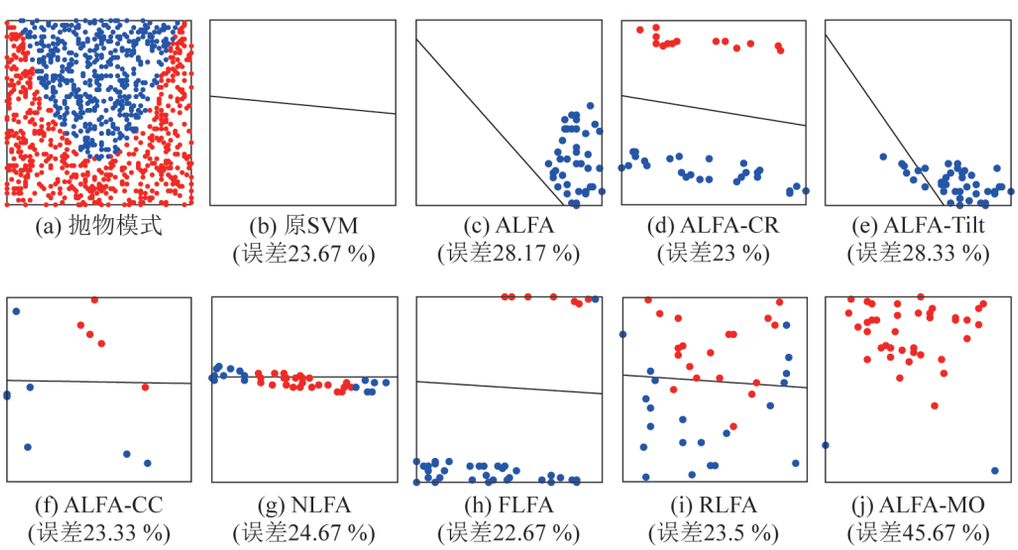

图7　各算法在抛物线可分数据上对线性核SVM的攻击效果

Fig.7　Attack effect of all algorithms on SVM with linear kernel based on parabolic separable data

https://dc-china-simulation.researchcommons.org/journal/vol36/iss9/4
DOI: 10.16182/j.issn1004731x.joss.24-0101

第 36 卷第 9 期
2024 年 9 月
系统仿真学报
Journal of System Simulation
Vol. 36 No. 9
Sept. 2024

图 9 为各算法在该数据集上对采用多项式核 SVM 的攻击效果。从图中可知，受攻击之前 SVM 分类器在测试数据集上的分类误差为 14.33%，说明采用多项式核函数的 SVM 分类器同样不太适合于此数据集。在受到不同算法的攻击之后，SVM 的分类误差均有增加。其中，ALFA、ALFA-Tilt、ALFA-CC、FLFA 和 ALFA-MO 算法的攻击效果均比较为明显，导致攻击后的 SVM 分类误差均超过 20%，ALFA-MO 攻击后的分类器分类误差高达 39.83%，比攻击效果最好的对比算法 FLFA 提升了 3 个百分点。此外，从表 1 中 F1 分数亦可以看出，ALFA-MO 算法在抛物线可分模式的数据集上对采用多项式核 SVM 的分类器攻击效果最好。

综上可知，ALFA-MO 在抛物线可分数据上对采用几种常见核函数的 SVM 分类器均能产生较对比算法更优的攻击效果。即便是对分类效果差强人意的 SVM 分类器，ALFA-MO 仍能产生较大的攻击效能。这表明在抛物线可分数据上，ALFA-MO 算法更适合于 SVM 分类器的抗攻击性能测试，以全面评估其在不同攻击强度下的性能稳定性。

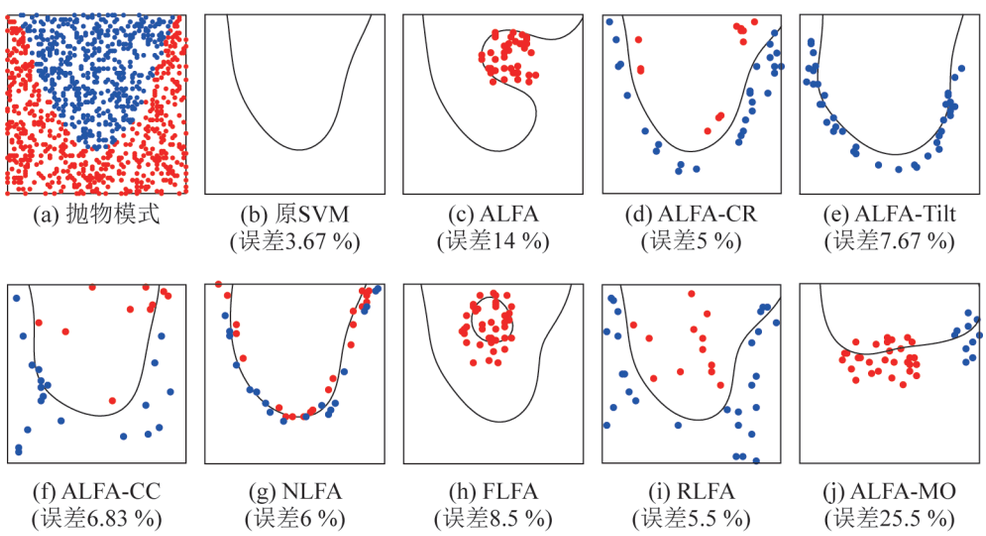

图 8　各算法在抛物线可分数据上对径向基核 SVM 的攻击效果

Fig. 8　Attack effect of all algorithms on SVM with RBF kernel based on parabolic separable data

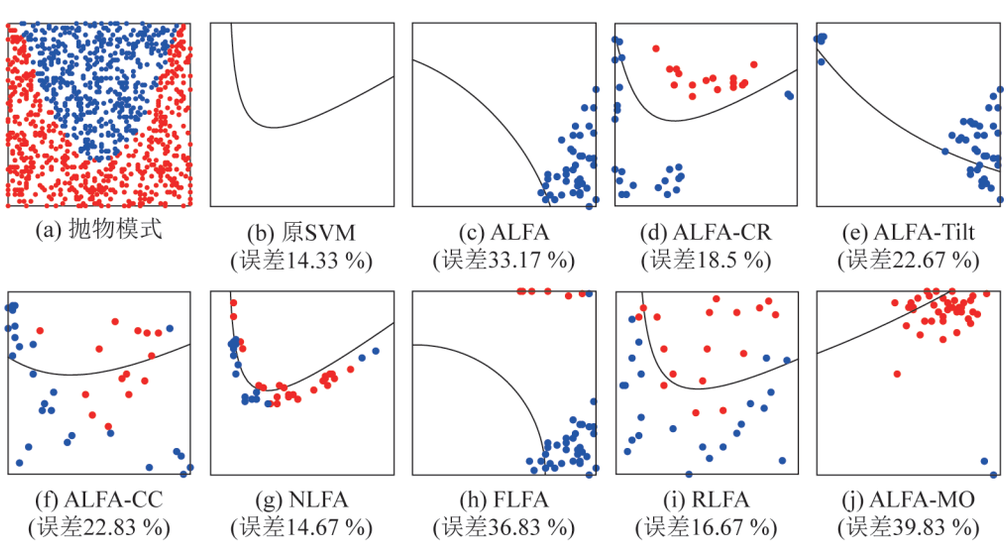

图 9　各算法在抛物线可分数据上对多项式核 SVM 的攻击效果

Fig. 9　Attack effect of all algorithms on SVM with polynomial kernel based on parabolic separable data

第 3 组实验将在环型可分数据集上展开，为了便于展示，数据集的构造仍然在二维空间中进行，该数据集同样包含 400 个训练样本、600 个测试样本。由于采用线性核的 SVM 在该数据集上的分类误差达到 50% 左右，已不具备基本的分类能力，本实验不再对线性核 SVM 分类器实施攻击。图 10 为各算法在该数据集上对采用径向基核 SVM 的攻击效果。从图中可知，受攻击之前 SVM 分类器在测试数据集上的分类误差为 4.67%，说明采用径向基核函数的 SVM 分类器适用于此数据集。在受到不同算法的攻击之后，SVM 的分类误差均有增加。其中，ALFA-CC 和 ALFA-MO 算法的攻击效果较为明显，导致攻击后的 SVM 分类误差均超过 10%， ALFA-MO 较 ALFA-CC 攻击后产生的分类误差提升了 3.83 个百分点，显著领先各对比算法。ALFA-CC 和 ALFA-MO 的攻击在 SVM 的分类边界中成功引入了一个缺口，这是造成分类误差明显增加的原因。此外，从表 1 中 F1 分数亦可以看出，ALFA-MO 算法在环形可分模式的数据集上对采用径向基核 SVM 的分类器攻击效果最好。

图 11 为各算法在该数据集上对采用多项式核 SVM 的攻击效果。从图中可知，受攻击之前 SVM 分类器在测试数据集上的分类误差为 23.5%，说明采用径向基核函数的 SVM 分类器也不太适用于此数据集。在受到不同算法的攻击之后，SVM 的分类误差均有增加。其中，ALFA、FLFA 和 ALFA-MO 算法的攻击效果均较为明显，致使攻击后的 SVM 分类误差均超过 40%，ALFA-MO 攻击后的分类器分类误差高达 52.17%，比攻击效果最好的对比算法 ALFA 和 FLFA 提升了 9.67 个百分点，致使被攻击后的 SVM 分类器分类结果不再具有参考意义，表明 ALFA-MO 具有最优的攻击测试能力。从表 1 中 F1 分数亦可以看出，ALFA-MO 算法在环形可分模式的数据集上对采用多项式核 SVM 的分类器攻击效果最好。

综上可知，ALFA-MO 在环形可分数据上对采用径向基核与多项式核函数的 SVM 分类器均能产生较对比算法更优的攻击效果。即便是对分类效果相对较差的 SVM 分类器，ALFA-MO 仍能产生较大的攻击收益。这表明在环形可分数据上，ALFA-MO 算法仍然更适合于 SVM 分类器的抗攻击性能测试，以全面评估其在不同攻击强度下的性能稳定性。

以上 3 组采用不同分布模式的数据对比实验验证了 ALFA-MO 算法在攻击各类 SVM 分类器时的优越性，表明其在 SVM 性能攻击测试方面的出色能力，是一种稳定可靠的性能攻击测试方法。

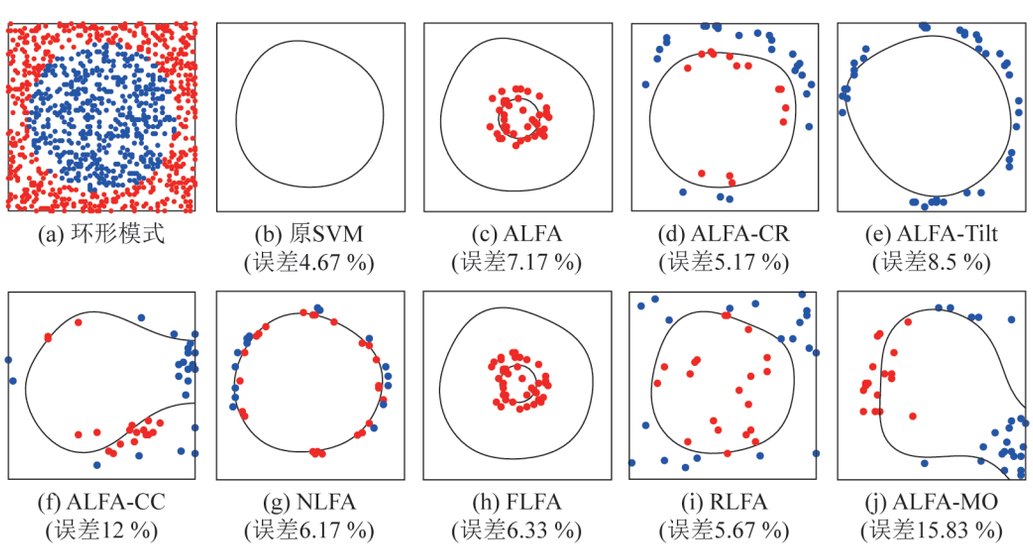

图 10　各算法在环形可分数据上对径向基核 SVM 的攻击效果

Fig.10　Attack effect of all algorithms on SVM with RBF kernel based on circle separable data

https://dc-china-simulation.researchcommons.org/journal/vol36/iss9/4
DOI: 10.16182/j.issn1004731x.joss.24-0101

第 36 卷第 9 期
2024 年 9 月
系统仿真学报
Journal of System Simulation
Vol. 36 No. 9
Sept. 2024

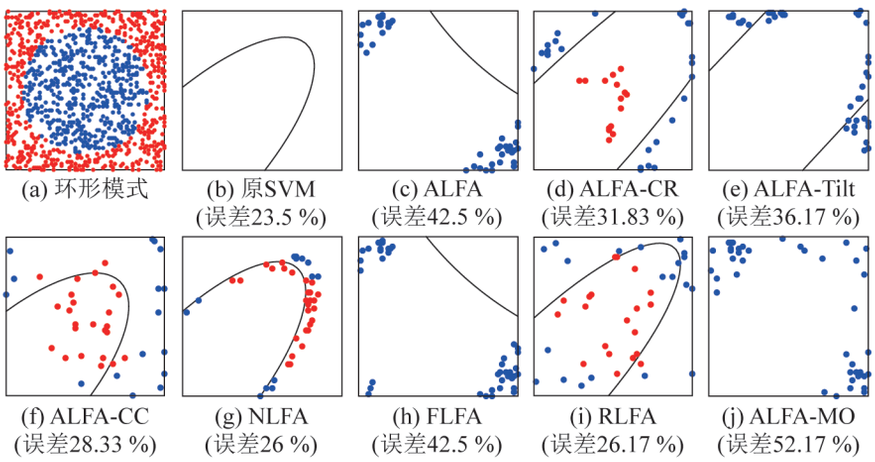

图 11　为各算法在环形可分数据上对多项式核 SVM 的攻击效果

Fig. 11　Attack effect of all algorithms on SVM with polynomial kernel based on circle separable data

3.3.3 真实数据仿真对比实验

本实验将在 LIBSVM 网站下载的 5 个真实数据集上开展 SVM 性能攻击测试实验。每个数据集均包含 1 000 个样本，从中随机选择 400 个样本用于训练 SVM 分类器，600 个样本用于测试分类器。为了全面评估 SVM 的分类性能稳定性，本实验测试了 ALFA-MO 与各对比算法在不同攻击成本下(即样本标签翻转数量)的攻击效果。为保证实验结果的可信性，每个实验指标均为 20 次独立重复实验的平均值，同时给出了独立重复实验中指标的最大值与最小值，以评估攻击效果的波动性。图 12~13 分别为 ALFA-MO 与各对比算法在不同攻击成本下，对采用径向基核 SVM 和多项式核 SVM 的攻击效果变化情况。

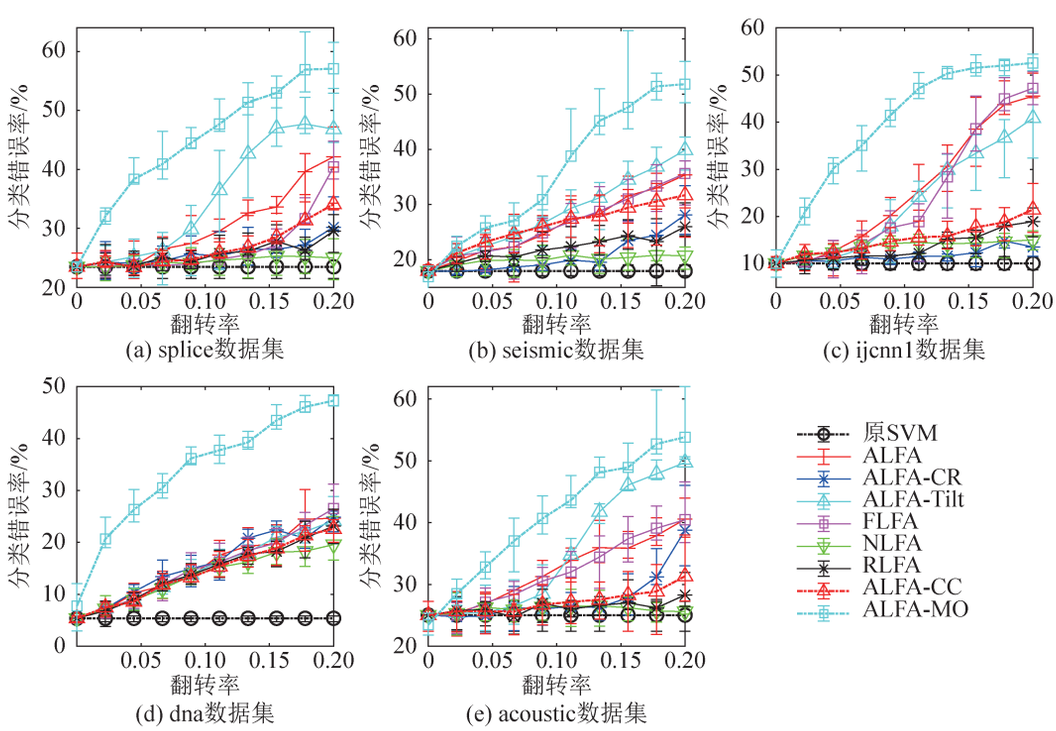

图 12　各算法在多个真实数据集上对径向基核 SVM 的攻击效果随标签翻转规模的变化

Fig. 12　Attack effect of all algorithms on SVM with RBF kernel against label flip scale with diverse real datasets

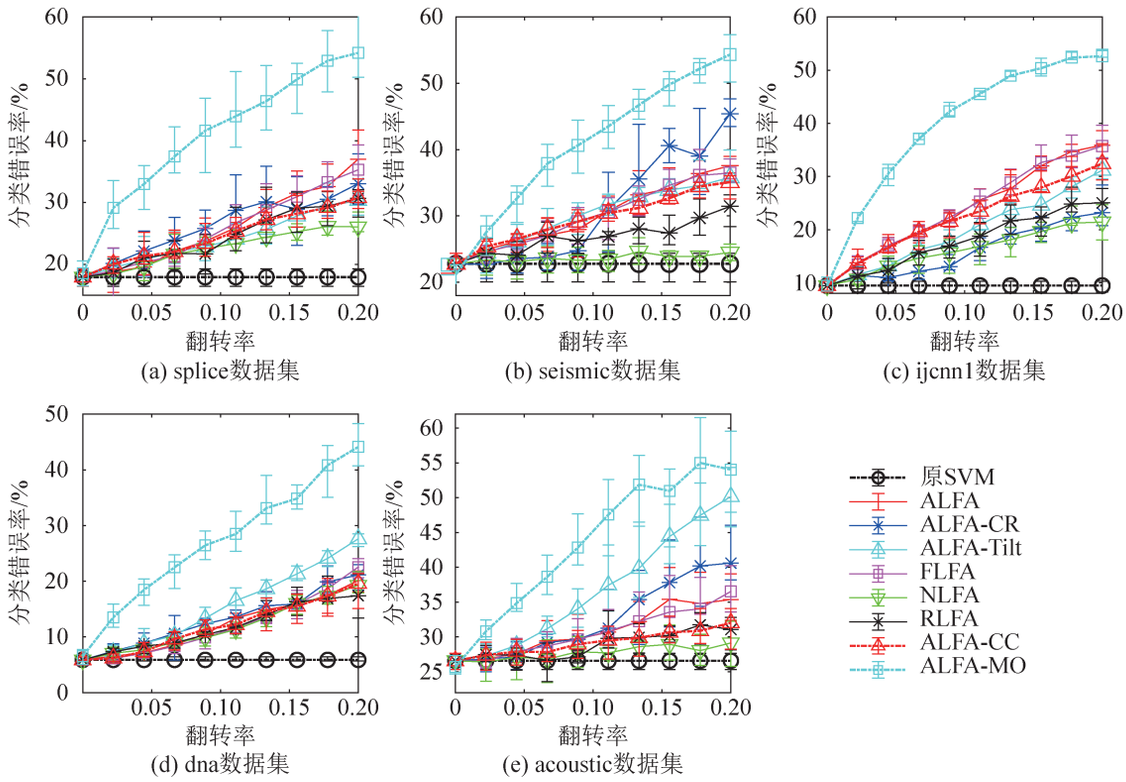

图 13　各算法在多个真实数据集上对多项式核 SVM 的攻击效果随标签翻转规模的变化

Fig. 13　Attack effect of all algorithms on SVM with polynomial kernel against label flip scale with diverse real datasets

从图 12~13 可以看出，无论是在哪个数据集上，ALFA-MO 对径向基核 SVM 和多项式核 SVM 的攻击效果，在相同的攻击成本下均比对比算法更优，而且优势极为明显。例如，在 splice 数据集上，ALFA-MO 算法对径向基核 SVM 的攻击效果明显强于对比算法中表现最好的 ALFA-Tilt 算法，其对多项式核 SVM 的攻击效果也明显强于众对比算法，只是在攻击性能上有一定的波动。尽管如此，ALFA-MO 算法在产生最差的攻击效果时，也仍然优于各对比算法的攻击效果。

4　结论

本文从对抗仿真测试的角度，模拟了通过篡改训练样本标签以最大程度毒害 SVM 分类器性能的攻击过程。这种攻击不需要访问数据本身的特征，是一种隐蔽高效的攻击策略，既适用于二分类模型，也适用于多分类模型。通过在训练集中模拟不同程度地篡改样本标签数量，可以有效测试评估 SVM 分类器在不同攻击水平下的性能损失。

为了仿真测试出 SVM 分类器在不同样本组合攻击下的性能退化强度，全面掌握分类器的数据敏感性与性能稳定性，本文提出了最小攻击代价-最大攻击成效的多目标优化模型。通过设计高效的演化求解算法，可以得到目标间的一组非支配解集，表征不同攻击成本下的分类器性能退化上限。在人工和真实数据集上的仿真对比实验结果表明，在不同攻击预算下，本文所提 ALFA-MO 算法对使用不同核函数 SVM 分类器的攻击效果均优于一众对比算法，说明 ALFA-MO 算法能够有效攻击对分类器具有重要影响的样本组合，来产生尽可能大的攻击效果，从而全面测试评估分类器在不同攻击强度下的性能稳定性。然而，ALFA-MO 算法仅能搜索高效的标签翻转组合攻击测试方案，尚无法在训练样本的特征层面构建有效的攻击测试样本。这是由于训练样本的特征空

https://dc-china-simulation.researchcommons.org/journal/vol36/iss9/4
DOI: 10.16182/j.issn1004731x.joss.24-0101

第 36 卷第 9 期 系统仿真学报 Vol. 36 No. 9
2024 年 9 月 Journal of System Simulation Sept. 2024

间维度过高，对其特征进行直接攻击需要解决攻击方式的设计和攻击成本的度量等问题。后续，将结合分类模型的学习与推理机制，探究全局搜索与局部微调相结合的样本攻击测试方法，拓展最小攻击代价-最大攻击成效的思路在更广泛的对抗测试问题中应用。

参考文献：

[1] Battista Biggio, Luca Didaci, Giorgio Fumera, et al. Poisoning Attacks to Compromise Face Templates[C]// 2013 International Conference on Biometrics (ICB). Piscataway: IEEE, 2013: 1-7.

[2] Rosenfeld E, Winston E, Ravikumar P, et al. Certified Robustness to Label-flipping Attacks via Randomized Smoothing[C]//Proceedings of the 37th International Conference on Machine Learning. Chia Laguna Resort: PMLR, 2020: 8230-8241.

[3] Battista Biggio, Giorgio Fumera, Fabio Roli. Pattern Recognition Systems Under Attack: Design Issues and Research Challenges[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2014, 28(7): 1460002.

[4] Patrick P K Chan, Luo Fengzhi, Chen Zitong, et al. Transfer Learning Based Countermeasure Against Label Flipping Poisoning Attack[J]. Information Sciences, 2021, 548: 450-460.

[5] Lü Zhuo, Cao Hongbo, Zhang Feng, et al. AWFC: Preventing Label Flipping Attacks Towards Federated Learning for Intelligent IoT[J]. The Computer Journal, 2022, 65(11): 2849-2859.

[6] Najeeb Moharram Jebreel, Josep Domingo-Ferrer, David Sánchez, et al. LFighter: Defending Against the Label-flipping Attack in Federated Learning[J]. Neural Networks, 2024, 170: 111-126.

[7] Fahri Anıl Yerlikaya, Serif Bahtiyar. Data Poisoning Attacks Against Machine Learning Algorithms[J]. Expert Systems with Applications, 2022, 208: 118101.

[8] Battista Biggio, Igino Corona, Blaine Nelson, et al. Security Evaluation of Support Vector Machines in Adversarial Environments[M]//Ma Yunqian, Guo Guodong. Support Vector Machines Applications. Cham: Springer International Publishing, 2014: 105-153.

[9] Xu Qianqian, Yang Zhiyong, Zhao Yunrui, et al. Rethinking Label Flipping Attack: From Sample Masking to Sample Thresholding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45 (6): 7668-7685.

[10] Konda Reddy Mopuri, Vaisakh Shaj, R Venkatesh Babu. Adversarial Fooling Beyond "Flipping the Label"[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2020: 3374-3382.

[11] Zhang Hongpo, Cheng Ning, Zhang Yang, et al. Label Flipping Attacks Against Naive Bayes on Spam Filtering Systems[J]. Applied Intelligence, 2021, 51(7): 4503-4514.

[12] Xiao Han, Thomas Stibor, Claudia Eckert. Evasion Attack of Multi-class Linear Classifiers[C]//Proceedings of the 16th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Berlin: Springer Berlin Heidelberg, 2012: 207-218.

[13] Barreno M, Nelson B, Joseph A D, et al. The Security of Machine Learning[J]. Machine Learning, 2010, 81(2): 121-148.

[14] Zhou Yan, Kantarcioglu M, Thuraisingham B, et al. Adversarial Support Vector Machine Learning[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2012: 1059-1067.

[15] Battista Biggio, Blaine Nelson, Pavel Laskov. Poisoning Attacks Against Support Vector Machines[C]// Proceedings of the 29th International Coference on Machine Learning. Madison: Omnipress, 2012: 1467-1474.

[16] Battista Biggio, Blaine Nelson, Pavel Laskov. Support Vector Machines Under Adversarial Label Noise[C]// Proceedings of the Asian Conference on Machine Learning. Chia Laguna Resort: PMLR, 2011: 97-112.

[17] Xiao Han, Xiao Huang, Claudia Eckert. Adversarial Label Flips Attack on Support Vector Machines[C]// Proceedings of the 20th European Conference on Artificial Intelligence. NLD: IOS Press, 2012: 870-875.

[18] Xiao Huang, Battista Biggio, Blaine Nelson, et al. Support Vector Machines Under Adversarial Label Contamination[J]. Neurocomputing, 2015, 160: 53-62.

[19] Mei Shike, Zhu Xiaojin. Using Machine Teaching to Identify Optimal Training-set Attacks on Machine Learners[C]//Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2015: 2871-2877.

[20] Burkard C, Lagesse B. Analysis of Causative Attacks Against SVMs Learning from Data Streams[C]// Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics. New York: Association for Computing Machinery, 2017: 31-36.

[21] 钱亚冠, 卢红波, 纪守领, 等. 基于粒子群优化的对抗样本生成算法[J]. 电子与信息学报, 2019, 41(7): 1658-1665.
Qian Yaguan, Lu Hongbo, Ji Shouling, et al. Adversarial Example Generation Based on Particle Swarm Optimization[J]. Journal of Electronics & Information Technology, 2019, 41(7): 1658-1665.

[22] Anirban Chakraborty, Manaar Alam, Dey V, et al. A Survey on Adversarial Attacks and Defences[J]. CAAI Transactions on Intelligence Technology, 2021, 6(1): 25-45.

[23] Koh P W, Steinhardt J, Liang P. Stronger Data Poisoning Attacks Break Data Sanitization Defenses[J]. Machine Learning, 2022, 111(1): 1-47.

[24] Jha R D, Hayase J, Oh S. Label Poisoning Is All You Need[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2023: 71029-71052.

[25] Vijay Lingam, Mohammad Sadegh Akhondzadeh, Aleksandar Bojchevski. Rethinking Label Poisoning for Gnns: Pitfalls and Attacks[C]//ICLR 2024. New York: ICLR, 2024: 1-29.

[26] Yao Feng, Du Yonghao, Li Lei, et al. General Modeling and Optimization Technique for Real-world Earth Observation Satellite Scheduling[J]. Frontiers of Engineering Management, 2023, 10(4): 695-709.

[27] Wang Yuting, Han Yuyan, Gong Dunwei, et al. A Review of Intelligent Optimization for Group Scheduling Problems in Cellular Manufacturing[J]. Frontiers of Engineering Management, 2023, 10(3): 406-426.

[28] He Yongming, Xing Lining, Chen Yingwu, et al. A Generic Markov Decision Process Model and Reinforcement Learning Method for Scheduling Agile Earth Observation Satellites[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(3): 1463-1474.

[29] Xing Lining, Rohlfshagen P, Chen Yingwu, et al. An Evolutionary Approach to the Multidepot Capacitated Arc Routing Problem[J]. IEEE Transactions on Evolutionary Computation, 2010, 14(3): 356-374.

[30] Xing Lining, Rohlfshagen P, Chen Yingwu, et al. A Hybrid Ant Colony Optimization Algorithm for the Extended Capacitated Arc Routing Problem[J]. IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics, 2011, 41(4): 1110-1123.

[31] Wang Xuewu, Hua Yi, Gao Jin, et al. Digital Twin Implementation of Autonomous Planning Arc Welding Robot System[J]. Complex System Modeling and Simulation, 2023, 3(3): 236-251.

[32] Zhang Qingfu, Li Hui. MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition[J]. IEEE Transactions on Evolutionary Computation, 2007, 11(6): 712-731.

[33] Ni Wayan Surya Wardhani, Masithoh Yessi Rochayani, Atiek Iriany, et al. Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data[C]//2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA). Piscataway: IEEE, 2019: 14-18.

DOI: 10.16182/j.issn1004731x.joss.24-0101