计算机科学
COMPUTER SCIENCE

局部时间序列黑盒对抗攻击

杨文博，原继东

引用本文

杨文博，原继东. 局部时间序列黑盒对抗攻击[J]. 计算机科学, 2022, 49(10): 285-290.

YANG Wen-bo, YUAN Ji-dong. Locally Black-box Adversarial Attack on Time Series[J]. Computer Science, 2022, 49(10): 285-290.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900254

局部时间序列黑盒对抗攻击

杨文博　原继东
北京交通大学计算机与信息技术学院　北京 100044
交通数据分析与挖掘北京市重点实验室(北京交通大学)　北京 100044
(weberyoung@bjtu.edu.cn)

摘　要　用于时间序列分类的深度神经网络由于其自身对于对抗攻击的脆弱性,导致模型存在潜在的安全问题。现有的时间序列攻击方法均基于梯度信息进行全局扰动,生成的对抗样本易被察觉。为此,文中提出了一种不需要梯度信息的局部黑盒攻击方法。首先,对抗攻击被描述为一个约束优化问题,并假设不能获得被攻击模型的任何内部信息;然后利用遗传算法求解该问题;最后由于时间序列 shapelets 提供了不同类别间最具辨别力的信息,因此将其设计为局部扰动区间。实验结果表明,在有潜在安全隐患的 UCR 数据集上,所提方法可以有效地攻击深度神经网络并生成对抗样本。此外,所提算法相比基准算法在保持较高攻击成功率的同时显著降低了均方误差。

关键词:黑盒对抗攻击;时间序列分类;局部扰动;遗传算法;Shapelet
中图法分类号　TP183

Locally Black-box Adversarial Attack on Time Series

YANG Wen-bo and YUAN Ji-dong
School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
Beijing Key Lab of Traffic Data Analysis and Mining(Beijing Jiaotong University), Beijing 100044, China

Abstract　Deep neural networks(DNNs) for time series classification have potential security concerns due to their vulnerability to adversarial attacks. The existing attack methods on time series performglobal perturbation based on gradient information, and the generated adversarial examples are easy to be perceived. This paper proposes a locally black-box method to attack DNNs without gradient information. First, the attack is described as a constrained optimization problem with the assumption that the method cannot get any inner information of the model, then the genetic algorithm is employed to solve it. Second, since time series shapelets provides the most discriminative information among different categories, it is designed as a local perturbation interval. Experimental results on UCR datasets that have potential security concerns indicate that the proposed method can effectively attack DNNs and generate adversarial samples. In addition, compared with the benchmark, the method significantly reduces the mean squared error while keeping a high success rate.

Keywords　Black-box adversarial attack, Time series classification, Local perturbations, Genetic algorithm, Shapelet

1　引言

现代社会中,智能设备和传感器产生了大量的时间序列数据。深度神经网络(Deep Neural Networks,DNNs)已经在计算机视觉和自然语言处理(Natural Language Processing,NLP)领域取得了极大的成功,目前它也被用来解决时间序列分类(Time Series Classification,TSC)问题。研究人员发现,在图像分类任务中,DNNs 容易受到对抗攻击(Adversarial Attack)[1]的威胁,一个典型的场景是自动驾驶汽车误识别[2]交通信号而发生事故。另外,特定的对抗样本(Adversarial Examples)可以欺骗 NLP 模型从而导致矛盾的预测[3]。在股票交易等时间序列预测领域,Dang-Nhu 等对概率自回归预测模型提出了有效的对抗攻击方法[4]。同样的安全隐患也存在于 TSC 任务中,如反窃电探测问题[5]、攻击食物频谱导致的食品安全问题[6]、对基于 DNN 的心脏诊断系统进行攻击引发的医学安全问题[7]。

人眼对颜色和曲线有着不同的感知敏感度,如人们很难注意到图片的一些像素发生了改变,但一条时间序列曲线的形状被扰动,人眼则可以轻松地识别。相比图像和文本领域的对抗攻击,时间序列对抗攻击任务最大的挑战是在保持较高攻击成功率的基础上,同时生成不易被人眼察觉的对抗样本。然而,目前图像领域的黑盒方法大都基于替代模型[8]。

到稿日期:2021-09-28　返修日期:2022-02-06
基金项目:科技创新 2030—"新一代人工智能"重大项目(2021ZD0113002);北京市自然科学基金(4214067);国家自然科学基金(61702030)
This work was supported by the National Key R & D Program of China(2021ZD0113002), Natural Science Foundation of Beijing, China(4214067) and National Natural Science Foundation of China(61702030).
通信作者:原继东(yuanjd@bjtu.edu.cn)

由于上述的人眼感知问题，这些方法对 TSC 任务并不适用。图 1 给出了两个通过只改变单个数据点的单像素攻击算法生成的对抗样本[9]。左侧时序曲线由于黑色区间的剧烈波动，总体形状相比原始曲线发生了巨大改变。右侧鸟的图像中，一个像素点的改变并不影响人眼对其的语义理解。

图 1 单像素攻击生成的时间序列和图像对抗样本对比

Fig. 1 Comparison of time series and image adversarial samples generated by single-pixel attacks

最近关于时间序列分类问题的对抗攻击算法都是基于人眼可见的全局扰动，而且需要获得目标模型或者替代模型的梯度信息[6,10-11]。为了生成人眼不可察觉的时间序列对抗样本和避免训练替代模型，一个直观的思想是减少扰动区间的长度和扰动幅度。时间序列 shapelet 是不同类别间最具辨别性的子序列[12]，本文利用它来限制扰动区间的长度。另外，遗传算法(Genetic Algorithm，GA)在不需要任何目标模型的内部信息的条件下，可以寻找最优的局部扰动，而且攻击者可以灵活地通过扰动因子控制扰动幅度。

图 2 给出了本文提出的 BAAT(Black-box Adversarial Attackon Time Series)算法和梯度对抗转换网络(Gradient Adversarial Transform Network，GATN)分别攻击 Italy. 数据集生成的对抗样本实例。GATN 算法生成的对抗样本在全局区间上都有剧烈的扰动，总体的曲线形状相对原始序列也发生了巨大改变。相反，BAAT 生成的样本则保持了原有曲线的整体形状，只在灰框的局部区间有小幅度扰动。

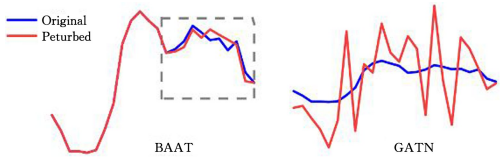


图 2 BAAT 和 GATN 生成的对抗样本的比较

Fig. 2 Comparison of two samples generated by BAAT and GATN

本文算法可以成功地攻击 24 个有潜在安全隐患的 UCR 数据集[13]，并生成与原始序列相似的对抗样本。本文的主要贡献如下：

(1)提出了一种新的黑盒时间序列对抗攻击算法(Black-box Adversarial Attackon Time Series)。本领域已有的工作主要依赖于目标或替代模型的梯度信息，而 BAAT 是不需要梯度信息的。

(2)BAAT 是第一个通过在局部 shapelet 区间添加扰动进行攻击的算法。

(3)相比基准算法，BAAT 将均方误差降低了大约两个数量级，同时保持了较高的攻击成功率，并显著提升了对抗样本的不易察觉性。

2 相关工作

2.1 时间序列分类对抗攻击

自从 Szegedy 等提出图像识别问题中对抗样本的概念[1]，学者们开始了各个领域对抗攻击的研究。目前图像领域的黑盒攻击算法主要基于替代模型和自然演化算法[14]。例如，Su 等[9]提出的单像素攻击可以仅通过改变图像的一个像素点来产生对抗样本；Papernot 等[15]则利用替代模型的可迁移性质去近似原始目标模型，进而达到黑盒攻击的目的；ANGRI 和 UPSET 模型[16]则是基于生成对抗网络的自监督训练模式。目前 TSC 问题中没有相关的黑盒攻击算法。

Oregi 等提出的方法[10]是第一项提出在 TSC 问题上实现对抗攻击的工作。尽管此方法只在模拟数据集上进行实验，但是可以成功攻击基于动态时间规整的软 k 近邻算法。Ismail 等[6]利用图像领域的白盒攻击方法去降低深度残差网络的分类正确率，例如快速梯度符号方法(Fast Gradient Sign Method，FGSM)和基本迭代方法(Basic Iterative Method，BIM)。Rathore 等[17]扩展上述两个白盒攻击为有目标(Targeted)的和普遍的对抗算法。Karim 等[11]提出了 GATN 算法来生成单维和多维时序[18]对抗样本。为了保证黑盒限制和获取梯度信息，上述方法需要通过知识蒸馏来训练一个替代模型。另外，Chen 等[7]提出了基于 DNN 的心电图诊断系统的对抗攻击算法，分析了心电图特有的性质并且提出了一种平滑度量方法来度量相邻扰动的差异。Han 等[19]也探索了对心电图信号的攻击方法，并提出了卷积平滑算法。然而，上述方法都是白盒攻击，梯度和网络结构等信息在实际场景中并不容易获取。

2.2 遗传算法

遗传算法[20]是基于达尔文自然选择思想建立的优化算法。与传统的搜索算法不同，遗传算法通过一组随机初始种群，通过选择、变异和交叉运算生成下一代种群。在每一代中，适应性函数计算出个体适应度，适应度更高的个体将更有可能存活到下一代。类似于生物繁殖，交叉是通过多个父代个体生成子代个体的过程。变异是为了增加个体多样性，并提供更大的搜索空间。遗传算法在解决组合优化问题时表现较好[21]，很适合寻找最优的局部扰动并生成不易被察觉的对抗样本。

2.3 时间序列 shapelets

时间序列 shapelets 是 Ye 等[12]提出的一个概念，它是不同类别之间最具辨别性的子序列。时间序列可以通过最好的 k 个 shapelets 被转换到另一个特征空间，因此各种分类器可以直接用于 TSC[22]任务。因为时间序列 shapelets 特有的辨别性质，所以本文假设它可以提供比其他子序列更有效的扰动区间。本文中，BAAT 利用最佳的 shapelet 区间作为对抗攻击的扰动区间。图 3 给出了时间序列及其最佳的 shapelet，其中加粗的序列片段就是 ECGFiveDays 数据集上最佳的 shapelet。

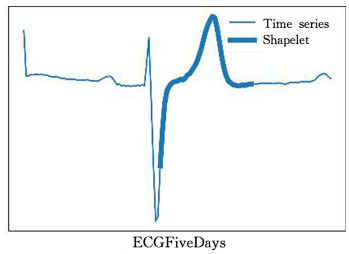


图 3 时间序列 shapelet

Fig. 3 Time series shapelet

2.4 全卷积网络

全卷积网络(Fully Convolutional Network，FCN)是 TSC 领域中一个强大的基准模型[23]，其网络结构如图 4 所示。它包含 3 个卷积层，过滤器个数分别为 128，256 和 128，卷积层之后进行批归一化操作[24]；在 ReLU[25] 激活函数之后，全局池化层在时间维度上进行池化操作；softmax 层最后输出概率向量，其中神经元数目和数据集类标个数相同。为了与基准算法进行对比，本文将 FCN 作为实验的目标模型。

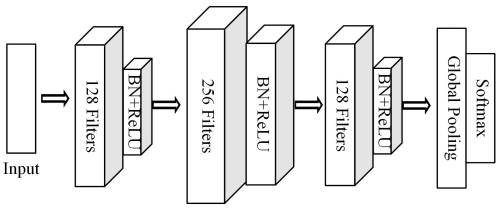


图 4 全卷积网络

Fig. 4 Fully convolutional network

3 BAAT 算法

3.1 黑盒限制

本文假定攻击者以黑盒方式访问被攻击模型，只需要获得模型预测的类标和其对应的概率(置信度)。攻击者不知目标模型的内部信息，包括参数权重、网络结构和训练数据[8]，因此，实验采用 UCR[13] 数据集的测试集 D 作为被攻击的数据集，目标模型从未见过测试集 D。此外，shapelet 作为先验知识，它应该从一个独立的数据集中得到，因此 D 被划分成两个类标平衡的数据集 D_{shap} 和 D_{test}，分别用来计算 shapelet 和作为被攻击数据集。

3.2 问题描述

基于 DNN 的时间序列分类模型的对抗攻击过程可以形式化为一个约束优化问题。令 f 表示目标模型，$\boldsymbol{x}=(x_1,\cdots,x_n)$ 表示长度为 n 的时间序列，其中每一个元素代表相应时间步的数值。如果原始序列属于类标 ori，则模型输出的概率为 $f_{\text{ori}}(x)$。类似地，施加在 shapelet 区间 l 的扰动可以表示为向量 $\boldsymbol{\varepsilon}(x)$。扰动幅度 σ 通过扰动因子 β 控制，如式(4)所示：

$$\sigma=\beta\cdot\left(\frac{1}{T}\sum_{i=1}^{T}(x_i^{\max}-x_i^{\min})\right) \tag{1}$$

其中，$x_i^{\max}$ 和 $x_i^{\min}$ 分别代表每个时间序列的最大值和最小值，并假设数据集包含 T 条时间序列。BAAT 的目标是寻找下面约束优化问题的最优解 $\boldsymbol{\varepsilon}(x)^*$。

$$\begin{aligned}&\underset{\boldsymbol{\varepsilon}(x)^*}{\text{minimize}}\ f_{\text{ori}}(x+\boldsymbol{\varepsilon}(x))\\&\text{s.t.}\ \ \|\boldsymbol{\varepsilon}(x)\|_0\leqslant l\\&\qquad\|\boldsymbol{\varepsilon}(x)\|_\infty\leqslant\sigma\end{aligned} \tag{2}$$

其中，$\|\boldsymbol{\varepsilon}(x)\|_0$ 和 $\|\boldsymbol{\varepsilon}(x)\|_\infty$ 分别代表扰动向量 $\boldsymbol{\varepsilon}(x)$ 的 L_0 和 L_∞ 范数。式(2)表示 BAAT 仅在 shapelet 区间施加幅度不大于 σ 的扰动，进而使模型输出原始类标的概率下降，最终达到对抗攻击的效果。为了与基准算法进行对比，本文只讨论无目标攻击(Untargeted Attack)的情况。

3.3 攻击过程

BAAT 的目的是成功生产对抗样本并同时满足上述的约束。因此，本文提出利用 GA 和 shapelet 区间的黑盒攻击策略去解决此约束优化问题。图 5 给出了 BAAT 算法的整体流程图，其包含两个阶段，分别如算法 1 和算法 2 所示。

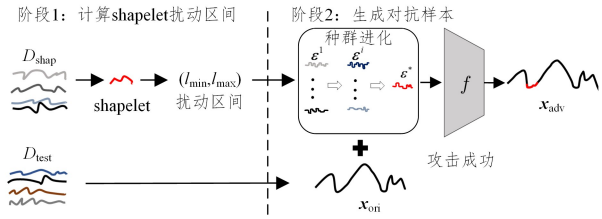


图 5 BAAT 算法的流程图

Fig. 5 Process of BAAT

算法 1 计算 shapelet 区间

输入：数据集 D_{shap}；滑动窗口大小 w

输出：shapelet 区间 l

```
1. candidates←GenCandidates(D_shap, w)
2. shapelets←[]
3. for S in candidates do
4.    quality←checkCandidate(S, D_shap)
5.    shapelets.add(S, D_shap)/*加入列表*/;
6. sortByQuality(shapelets)/*排序*/;
7. removeSelfSimilar(shapelets)
8. bestShapelet←shapelets[0]/*最佳 shapelet*/;
9. l←getInterval(bestShapelet)/*扰动区间*/;
10. return l.
```

在攻击模型前，首先需要获得 shapelets 来作为扰动区间。算法 1 是计算 shapelet 区间的伪代码。首先根据用户定义的滑动窗口长度 w 得到所有的候选子序列(见算法 1 中的第 1 行)。然后计算每个子序列的质量并将其加入到 shapelet 列表中(见算法 1 中的第 3－5 行)。根据质量对 shapelets 排序后，自相似的 shapelets 将被删除(见算法 1 中的第 6－7 行)。最终得到最佳 shapelet 区间的起始位置和结束位置(见算法 1 中的第 8－10 行)。该算法的详细过程请参考文献[12，22]。

算法 2 描述了对抗样本生成的伪代码。首先根据算法 1 获得扰动区间(见算法 2 中的第 1 行)，然后根据式(1)计算扰动幅度(见算法 2 中的第 2 行)。在 GA 优化过程中，施加的扰动被编码为一个实数向量(候选解)并被随机初始化(见算法 2 中的第 4－6 行)，其长度等于扰动区间 l 的长度，其数值范围小于或等于扰动幅度，即式(2)的约束条件。目标模型 f 的预测概率为个体的适应度(见算法 2 中的第 10 行)。如果最优个体的预测类标和原始类标不同，那么此样本攻击成功(见算法 2 中的第 12－15 行)。否则，种群继续进化直到最大进化代数 G。父代个体通过适应度归一化概率抽样得到(见算法 2 中的第 17－21 行)，并通过交叉和变异运算产生下一代个体(见算法 2 中的第 22－24 行)。尽管攻击过程被描述为一个最优化问题，但是在程序实际运行中，一旦攻击成功，程序就会停止并输出相应的对抗样本。

算法 2 无目标攻击情况下生成对抗样本

输入：数据集 D_{shap} 和 D_{test}；扰动因子 β；真实类标 t_{ori}

输出：对抗样本列表 adv_list

```
1. interval←FindingShapeletInterval(D_shap)
2. σ←GetMagnitude(D_test, β)/*扰动幅度*/;
3. adv_list←[];
4. for x_ori in D_test do
```

```
for i=1,2,…,M in population do
  ε_i^1 ← init(interval,σ) /* 初始化种群 */;
for g=2,…,G in generation do
  for i=1,2,…,M in population do
    p_i^{g-1} ← f_ori(x_ori + ε_i^{g-1});
  ε* ← ε_{argmin_j p_j}^{g-1} /* 最优候选解 */;
  if argmax_c f_c(x_ori + ε*) ≠ t_ori
    x_adv ← x_ori + ε* /* 攻击成功 */;
    adv_list.append(x_adv);
    break;
  else
    prob ← Normalize(p^{g-1})
    ε_1^g ← ε* /* 保留上代最优 */;
    for i=1,2,…,M in population do
      par1 ← sample(ε^{g-1}, prob);
      par2 ← sample(ε^{g-1}, prob);
      child ← Crossover(par1, par2);
      child_mut ← Mutation(child);
      ε_i^g ← child_mut;
return adv_list.
```

3.4 评估方法

为了准确计算数据集D_{test}上成功攻击的样本数，实验采取两步验证的方法。第一步首先确定攻击前模型预测的类标是否与真实类标一致。如果一致，那么第二步判断攻击后模型预测的类标与原始类标是否一致。两步验证法避免了那些本来就不能被模型正确分类的样本导致的计算误差。实验使用 UCR 数据集官方提供的类标。

4 实验和结果分析

如表 1 所列，本文选择 UCR 数据集中心电图(Electrocardiogram，E)和传感器(Sensor，S)这两类有潜在安全隐患的数据集[11,13]进行实验。实验对比的基准算法是 GATN，它是目前黑盒条件下的最佳算法。算法的评估指标如下。

表 1 数据集及其类型

Table 1 Name and type of datasets

Name	Type	Name	Type	Name	Type
Car	S	FordB	S	Phoneme	S
Chlorine.	S	InsectWngSnd	S	Plane	S
CinCECG.	S	Italy.	S	SonyAIBOSurf1	S
Earthquakes	S	Lightning2	S	SonyAIBOSurf2	S
ECG200	E	Lightning7	S	StarLightCurves	S
ECG5000	E	MoteStrain	S	Trace	S
ECGFiveDays	E	NonIFECGTho1	E	TwoLeadECG	E
FordA	S	NonIFECGTho2	E	Wafer	S

(1)攻击成功率(Success Rate，SR)。SR 计算被攻击成功的样本占D_{test}的比例。

(2)均方误差(Mean Square Error，MSE)。MSE 度量攻击成功需要的扰动幅度，MSE 越小说明算法越有效。实验依据式(3)，使用所有对抗样本的平均 MSE 描述所需要的平均扰动幅度。

$$\frac{1}{Tn}\sum_{j=1}^{T}\sum_{i=1}^{n}(x_{j,ori}^{i}-x_{j,adv}^{i})^2 \tag{3}$$

其中，$x_{j,ori}^{i}$($x_{j,adv}^{i}$)代表第 j 个原始(对抗)样本中的第 i 个元素值。

(3)平均迭代次数(Average Numberof Iterations，ANI)。ANI 描述了每个数据集被攻击成功所需的 GA 的平均进化代数。同一数据集的 ANI 越低表明算法越有效。

4.1 实验设置和细节

实验选择 FCN 作为目标模型 f 并使用 Adam 优化器和默认的超参数配置进行训练[23]。目标模型和训练的超参数均与对比的 GATN 算法一致[11]。表 2 列出了具体的参数设置情况。计算 shapelet 扰动区间的滑动窗口大小时通过百分比系数 w 来控制，例如 0.1 表示窗口大小为总长度的 10%。GA 的最大可进化代数 G 设为 50，种群数量 M 设置为 60。在实际运行中，为了加速每个样本的攻击过程，程序设置了提前放弃的策略。如果种群进化了一定代数后，模型预测的概率依旧很高，此样本的攻击过程将会立即停止，这些样本被认为是很难被攻击成功的，但是这可能会降低攻击成功率。扰动因子 β 依据 4.2 节中的实验设为 3 个值：0.01，0.02 和 0.04。

表 2 参数设置

Table 2 Parameter settings

Parameters	Values	Description
w	0.1,0.2,0.3,0.4	滑动窗口比例
G	50	最大进化代数
M	60	GA 种群数量
β	0.01,0.02,0.04	扰动因子

4.2 参数分析

如图 6 所示，当 β 取[0.01，0.02，0.04，0.06，0.08，0.1]时，图中的 5 个数据集 SR 都与 β 成正相关。随着 β 的增大，施加的扰动幅度也越大，这违背了对抗样本不易察觉的基本特征。因此，扰动因子的取值是样本的不易察觉性和 SR 权衡的结果。根据实验可知，如果 β 超过 0.04，生成的对抗样本可以轻易地被人眼识别，失去了本身的对抗意义，因此本文之后的实验将 β 的取值限制为[0.01，0.02，0.04]。

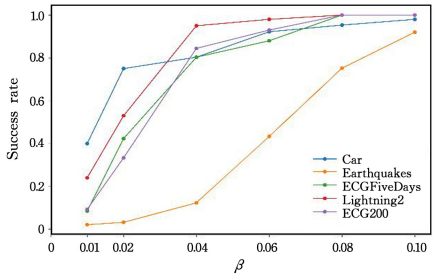


图 6 5 个数据集上的成功率变化趋势

Fig. 6 Trend of SR on five datasets

由于空间限制，本文只展示在 24 个 UCR 数据集上实验的统计结果。表 3 列出了 BAAT 和 GATN 在 24 个数据集上的指标平均统计结果(见表 3 中的第 1—3 行)。总体来看，随着 β 的减小，SR 也随之降低。同时，MSE 却变得非常小，从而可以生成人眼不易察觉的对抗样本。表 3 还列出了各 ANI 范围的数据集个数(见表 3 中的第 4—6 行)，对于大部分数据集，BAAT 可以在 10 代以内攻击成功。不同数据集攻击的难易程度是不一样的，例如当 β 为 0.04 时，StarLightCurves 数据集的 SR 为 97.41%，而 Trace 数据集只有 6.32%。

表 3 不同 β 下的 BAAT 和 GATN 的统计结果

Table 3 Statistical results of BAAT and GATN with different β

	β=0.01	β=0.02	β=0.04	GATN
Avg_SR	19.01%	32.05%	57.64%	46.34%
Avg_MSE	0.0005	0.0020	0.0084	0.1273
Avg_ANI	8.90	9.32	9.12	—
#*ANI*<=2	2	3	5	—
#2<*ANI*<10	11	15	9	—
#*ANI*>10	7	5	8	—

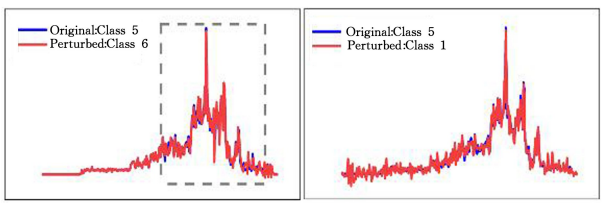
4.3 BAAT 与 GATN 算法的对比

如表 3 所列，BAAT 生成的对抗样本的平均 MSE 相比 GATN 降低了大约两个数量级，这极大地提升了对抗样本的不易察觉性。当 β 等于 0.04 时，BAAT 的平均 *SR* 高于 GATN；当 β 等于 0.02 和 0.01 时，GATN 的平均 *SR* 高于 BAAT。如表 4 所列，本文对两种算法也进行了 Wilcoxson 秩和检验[26]。如果 p 值小于 0.05，那么可以得出二者有显著性差异的结论。当 β 等于 0.04 和 0.02 时，BAAT 和 GATN 在 *SR* 上没有显著性差异；当 β 等于 0.01 时，GATN 的 *SR* 显著高于 BAAT；BAAT 的 *MSE* 结果在 3 个 β 取值下均显著优于 GATN。基于 shapelet 区间的局部扰动思想，BAAT 大大降低了 *MSE* 的结果，但同时局部扰动也适当地削弱了对抗攻击的威力，导致部分情况下的 *SR* 低于基准算法。另外，图 7 给出了两种算法在 Lightning7 数据集上生成的对抗样本的直观差异（蓝色为原始曲线，红色为扰动后的曲线），图 7(a) 为 BAAT 生成的样本，其中灰色框为 shapelet 区间，图 7(b) 为 GATN 生成的样本。BAAT 生成的对抗样本和原始样本相比，只在图中灰框的局部 shapelet 扰动区间中有微小幅度的改变，人眼几乎察觉不到。相反地，GATN 算法生成的对抗样本在全局区间上都有剧烈扰动，人眼可以轻易识别其为假样本。实验结果证明，BAAT 显著提升了对抗样本的不易察觉性。

表 4 $p<0.05$ 时 Wilcoxson 秩和检验结果

Table 4 Results of Wilcoxson signed-rank test when $p<0.05$

	β=0.01 vs. GATN	β=0.02 vs. GATN	β=0.04 vs. GATN
SR	$9.17*10^{-4}$	$1.28*10^{-1}$	$2.19*10^{-1}$
MSE	$1.32*10^{-4}$	$2.70*10^{-5}$	$1.82*10^{-5}$



(a)BAAT (b)GATN

图 7 Lightning7 数据集上两种算法生成的对抗样本（电子版为彩图）

Fig. 7 Adversarial examples of the two algorithms on Lightning7

4.4 验证 shapelet 区间的有效性

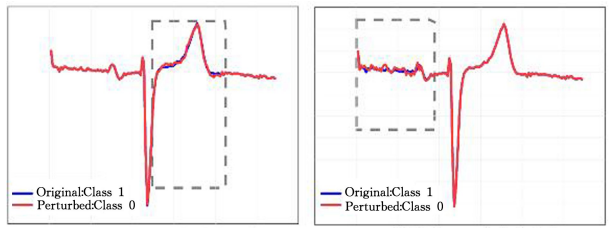
为了验证 shapelet 区间的有效性，本文在每个数据集上随机选取与 shapelet 长度相等的区间并将其作为扰动区间进行攻击。如表 5 所列，加粗数据表示胜出值，标星的数据表明其显著劣于加粗数据，无标记则表示无统计意义的显著性差异。表 5 中的数据表明，以 shapelet 区间为扰动区间的攻击成功率显著高于随机区间；攻击成功所需要的 *ANI* 也显著少于随机区间；对于 *MSE* 则没有太大差别。

表 5 随机区间和 shapelet 区间的平均结果

Table 5 Average results of random and shapelet intervals

	SR/%	MSE	ANI
random_interval	40.04*	0.0086	9.97*
shapelet_interval	**57.64**	**0.0084**	**9.12**

图 8 给出了数据集 ECGFiveDays 的一个具体例子，分别在 shapelet 区间和普通区间施加扰动，模型的预测类标虽然经过攻击都能从 1 变为 0，但是 GA 进化代数不同。图 8(a) 中的 shapelet 区间（灰色虚线框）是心电图中的“T 波”，代表了心脏的舒张和下一次收缩的准备阶段，有着清晰的实际意义。基于普通扰动区间的样本需要 69 代被攻击成功，远远多于 shaplelet 区间的 17 代。实验结果证明，shapelet 区间有显著的优势。



(a)shapelet 区间：17 代攻击成功 (b)普通区间：69 代攻击成功

图 8 shapelet 区间和普通区间的实例（电子版为彩图）

Fig. 8 Examples of shapelet interval and common intervals

4.5 时间复杂度分析

表 6 列出了同等实验条件下，GATN 和 BAAT 在各自前期阶段和生成对抗样本阶段消耗的平均时间。GATN 算法主要在前期的训练替代模型过程中消耗了较长时间，生成单个样本的时间短于 1s。BAAT 前期计算 shapelet 区间消耗的时间较短，但是由于 GA 的迭代进化特性，其生成单个对抗样本的时间效率较低。

表 6 GATN 和 BAAT 消耗的平均时间

Table 6 Average time consumption of GATN and BAAT

	前期过程/min	生成对抗样本
GATN（训练替代模型）	41.7	小于 1 s
BAAT（计算 shapelet）	16.3	3.2 min

另外，查询次数是对抗攻击领域常用的时间复杂度衡量指标，一次查询代表目标模型对样本的一次评估。在本文的实验中，每个样本的查询次数等于 GA 种群数量乘以攻击成功所需的迭代次数。假设一个数据集包含 T 个样本，平均迭代次数 *ANI* 等于 a，种群数量为 M，那么 BAAT 的时间复杂度为 $O(TMa)$。

结束语 本文提出了一种基于局部 shapelet 区间和遗传算法的时间序列黑盒攻击方法。相比本领域中已存在的方法，本文方法有两个优势：1）方法简单，BAAT 直接对目标模型发起攻击，不需要训练替代模型来满足黑盒限制；2）对抗

样本不易被人眼察觉。

实验结果表明，BAAT 可以有效、灵活地攻击用于时间序列分类的 DNN 模型，并显著提升对抗样本的不易察觉性。尽管 GA 可以在进化若干代之后攻击成功，但是其本身计算复杂度较高并且不能应用于实时场景。未来的工作会考虑提升算法的执行效率。最后，希望研究人员更多地关注时间序列分类领域的潜在安全问题。

参 考 文 献

[1] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [J]. arXiv: 1312.6199, 2013.

[2] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Robust physical-world attacks on deep learning visual classification [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2018.

[3] ZHANG W E, SHENG Q Z, ALHAZMI A, et al. Adversarial attacks on deep-learning models in natural language processing: A survey [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2020, 11(3): 1-41.

[4] DANG-NHU R, SINGH G, BIELIK P, et al. Adversarial attacks on probabilistic autoregressive forecasting models[C]// Proceedings of the International Conference on Machine Learning. PMLR, 2020.

[5] ZHENG Z, YANG Y, NIU X, et al. Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids [J]. IEEE Transactions on Industrial Informatics, 2017, 14(4): 1606-1615.

[6] FAWAZ H I, FORESTIER G, WEBER J, et al. Adversarial attacks on deep neural networks for time series classification [C]// Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019.

[7] CHEN H, HUANG C, HUANG Q, et al. Ecgadv: Generating adversarial electrocardiogram to misguide arrhythmia classification system[C]// Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, 2020.

[8] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]// Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. ACM, 2017.

[9] SU J, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks [J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.

[10] OREGI I, DEL SER J, PEREZ A, et al. Adversarial sample crafting for time series classification with elastic similarity measures [C]// Proceedings of the International Symposium on Intelligent and Distributed Computing. Springer, 2018.

[11] KARIM F, MAJUMDAR S, DARABI H. Adversarial attacks on time series [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(10): 3309-3320.

[12] YE L, KEOGH E. Time series shapelets: a new primitive for data mining[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2009.

[13] DAU H A, BAGNALL A, KAMGAR K, et al. The UCR time series archive [J]. IEEE/CAA Journal of Automatica Sinica, 2019, 6(6): 1293-1305.

[14] PAN W W, WANG X Y, SONG M L, et al. Survey on Generating Adversarial Examples [J]. Journal of Software, 2020, 31(1): 67-81.

[15] PAPERNOT N, MCDANIEL P, GOODFELLOW I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples [J]. arXiv: 1605.07277, 2016.

[16] SARKAR S, BANSAL A, MAHBUB U, et al. UPSET and ANGRI: Breaking high performance image classifiers [J]. arXiv: 1707.01159, 2017.

[17] RATHORE P, BASAK A, NISTALA S H, et al. Untargeted, Targeted and Universal Adversarial Attacks and Defenses on Time Series[C]// Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020.

[18] HARFORD S, KARIM F, DARABI H. Adversarial attacks on multivariate time series [J]. arXiv: 2004.00410, 2020.

[19] HAN X, HU Y, FOSCHINI L, et al. Deep learning models for electrocardiograms are susceptible to adversarial attack [J]. Nature Medicine, 2020, 26(3): 360-363.

[20] JI G L. Survey on genetic algorithm [J]. Computer Applications and Software, 2004, 21(2): 69-73.

[21] ANDERSON E J, FERRIS M C. Genetic algorithms for combinatorial optimization: the assemble line balancing problem [J]. ORSA Journal on Computing, 1994, 6(2): 161-173.

[22] YAN W H, LI G L. Research on time series classification based on shapelet [J]. Computer Science, 2019, 46(1): 29-35.

[23] WANG Z, YAN W, OATES T. Time series classification from scratch with deep neural networks: A strong baseline[C]// Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN). IEEE, 2017.

[24] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]// Proceedings of the International Conference on Machine Learning. PMLR, 2015.

[25] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [J]. Advances in Neural Information Processing Systems, 2012, 25: 1097-1105.

[26] DEMŠAR J. Statistical comparisons of classifiers over multiple data sets [J]. The Journal of Machine Learning Research, 2006, 7: 1-30.

YANG Wen-bo, born in 1997, postgraduate. His main research interests include artificial intelligence and time series classification.

YUAN Ji-dong, born in 1989, doctor, associate professor. His main research interests include data mining and pattern recognition.

（责任编辑：喻藜）