中国科学: 信息科学 2021 年 第 51 卷 第 1 期: 27–39
SCIENTIA SINICA Informationis
论文

K- 近邻分类器鲁棒性验证: 从约束放松法到随机平滑法

王璐[1,2], 姜远[1,2*]

1. 南京大学计算机软件新技术国家重点实验室, 南京 210023
2. 南京大学软件新技术与产业化协同创新中心, 南京 210023
* 通信作者. E-mail: jiangy@lamda.nju.edu.cn

收稿日期: 2020–06–15; 接受日期: 2020–08–07; 网络出版日期: 2020–12–21

国家自然科学基金 (批准号: 61673201, 61921006) 和南京大学优秀博士研究生创新能力提升计划 A 资助项目

摘要 本文研究 K- 近邻分类器的鲁棒性验证问题. 形式化鲁棒性验证的目标是计算分类器在给定样本点上的最小对抗扰动的精确值或者最小对抗扰动的非平凡下界. 我们将计算 K- 近邻分类器的最小对抗扰动形式化为一组二次规划问题. 二次规划问题的数目随近邻参数 K 的增大呈指数级增长, 精确求解该组二次规划问题往往不可行. 约束放松法通过放松优化的约束项, 可以在多项式时间内求解最小对抗扰动的下界. 然而, 本文通过理论分析和实验发现, 当近邻参数 K 取值较大时, 约束放松法求得的下界往往过于宽松, 甚至会出现 K 越大下界越小的反直觉结果. 为解决这一问题, 本文提出使用随机平滑法对 K- 近邻分类器进行鲁棒性验证. 随机平滑法利用了 K- 近邻分类器对高斯 (Gauss) 白噪声鲁棒的特点, 获得了较为理想的鲁棒性验证效果. 基准数据集上的实验结果表明, 相比于最新的鲁棒神经网络, "随机平滑的" K- 近邻分类器展现出了更好的验证鲁棒性.

关键词 监督学习, 对抗机器学习, 对抗鲁棒性, 鲁棒性验证, K- 近邻分类器

1 引言

近年来的研究发现, 神经网络模型的预测结果非常容易受到对抗扰动的影响 —— 在人类看来难以察觉的微小扰动可以很容易地改变神经网络模型的预测标记 [1~3]. 对抗扰动为机器学习模型在真实场景的应用带来了安全性挑战, 同时也促使越来越多的研究者开始关注机器学习模型的形式化鲁棒性验证问题. 鲁棒性验证的目标是计算分类器在给定样本点上的可以使得预测结果发生变化的最小扰动, 即最小对抗扰动. 针对神经网络模型, 许多鲁棒性验证方法被相继提出 [4~10], 其核心思想是对非线性激活函数进行凸放松, 以求解最小对抗扰动的下界.

引用格式: 王璐, 姜远. K- 近邻分类器鲁棒性验证: 从约束放松法到随机平滑法. 中国科学: 信息科学, 2021, 51: 27–39, doi: 10.1360/SSI-2020-0172
Wang L, Jiang Y. Robustness verification of $\boldsymbol{K}$-NN classifiers via constraint relaxation and randomized smoothing (in Chinese). Sci Sin Inform, 2021, 51: 27–39, doi: 10.1360/SSI-2020-0172

本文研究 K- 近邻分类器的鲁棒性验证问题. K- 近邻分类器简单有效且应用广泛 [11~13]. 同时由于被 "认为" 有较强的鲁棒性, K- 近邻常被用作一种启发式的提升模型鲁棒性的策略 [14, 15]. 基于以上原因, 形式化地验证 K- 近邻分类器的鲁棒性就变得尤为重要.

K- 近邻分类器鲁棒性验证的困难之处在于, K- 近邻分类器在形式上是一个非连续阶梯函数, 其性质和神经网络模型有较大区别. 以往用于神经网络模型的鲁棒性验证方法往往不能直接应用到 K- 近邻分类器上. 我们的前期工作 [16] 将 K- 近邻分类器的鲁棒性验证问题形式化为一组凸二次规划问题, 并针对最近邻分类器这种特殊情况提出了一个高效算法以精确求解最小对抗扰动. 然而, 当 K 取值较大时, 情况变得更加复杂.

- 一方面, 二次规划问题的数目随近邻参数 K 的增大呈指数级增长, 精确计算最小对抗扰动往往不可行. 有鉴于此, 我们提出了约束放松法. 该方法通过放松优化问题的约束项, 可以在多项式时间内计算最小对抗扰动的下界.
- 另一方面, 当 K 值取值较大时, 约束放松法的效果并不理想. 其原因在于, 为了能够高效求解, K 越大, 约束放松法的 "放松" 程度也越大, 这就使得约束放松法求得的最小对抗扰动下界过于 "宽松". 研究者普遍认为, 在一定范围内 K 越大, 由于利用了更多近邻信息, K- 近邻分类器的鲁棒性也应越强; 但是我们发现通过约束放松法进行鲁棒性验证时, 结果却往往是 K 越大, 验证鲁棒性越弱. 这一反直觉现象促使我们设计更为准确的鲁棒性验证方法.

本文的贡献有如下几点:

- 首先, 本文在前期工作的基础上, 详尽证明了 K- 近邻分类器的约束放松法是一种严格的形式化鲁棒性验证方法, 该方法求得的是最小对抗扰动的下界.
- 其次, 本文分析了约束放松法在 K 取值较大时存在的不足, 并从实验上给出了验证.
- 最后, 本文提出了 K- 近邻分类器的随机平滑法这一鲁棒性验证方法. 该方法利用了 K- 近邻分类器对高斯 (Gauss) 白噪声较为鲁棒的特点, 取得了更好的鲁棒性验证效果. 实验表明, 相比于最新的鲁棒神经网络, "随机平滑的" K- 近邻分类器展现出了更强的验证鲁棒性.

后文组织如下: 第 2 节介绍一些关于对抗鲁棒性的背景知识; 第 3 节着重讨论我们提出的约束放松法和随机平滑法这两种 K- 近邻分类器鲁棒性验证方法; 第 4 节在基准数据集上进行实验验证, 并和鲁棒神经网络模型作对比; 第 5 节讨论一些相关工作; 第 6 节对本文进行总结, 并简述可能的未来工作.

2 对抗鲁棒性和 K- 近邻分类器

机器学习模型的对抗鲁棒性 (adversarial robustness) 为模型在真实任务场景的应用提供了必要的安全性保障. 本文将逐一介绍与之相关的对抗扰动、鲁棒性验证等概念.

2.1 对抗扰动

在测试样本上施加扰动, 使得分类器在扰动后的样本上的预测结果异于原测试样本的真实标记, 这样的扰动即为对抗扰动 (adversarial perturbation). 最小对抗扰动 (minimal adversarial perturbation) 尤为重要 —— 在小于最小对抗扰动的范围内, 在测试样本上施加的任何扰动都不会使分类器的预测结果异于 "真实" 标记.

在上述定义中, 如果分类器对原始的测试样本的预测是错误的, 那么最小对抗扰动为 0. 本文以 ℓ_2 范数衡量对抗扰动的大小, 并以 $\|\cdot\|$ 指代 ℓ_2 范数.

2.2 对抗攻击

计算最小对抗扰动的可行解的过程即为对抗攻击 (adversarial attack). 注意到每一个可行解都对应于最小对抗扰动的上界. 对抗攻击通常会对扰动的大小做一定的限制. 如果在给定范围内找到可行解, 则称为攻击成功; 否则攻击失败. 攻击成功的扰动样本往往称为对抗样本 (adversarial example). 在许多图片分类任务中, 对抗样本和原始样本的差别常常难以被肉眼察觉. 需要强调的是, 一个对抗攻击方法在给定扰动范围内攻击失败, 并不能保证这一范围内不存在对抗样本. 这是因为对抗攻击方法往往并不能保证在对抗扰动存在的情况下找到可行解. 可以保证精确计算出最小对抗扰动的对抗攻击被称为最优对抗攻击.

形式化地, 给定分类器 $c:\mathbb{X}\to\mathbb{Y}$ 和测试样例 $(\boldsymbol{x},y)\in\mathbb{X}\times\mathbb{Y}$, 最优对抗攻击的目标是精确求解如下优化问题:

$$\min_{\boldsymbol{\delta}}\ \|\boldsymbol{\delta}\| \quad \text{s.t. } c(\boldsymbol{x}+\boldsymbol{\delta})\neq y. \tag{1}$$

显然, 该问题的每一个可行解即为对抗扰动.

2.3 鲁棒性验证

计算最小对抗扰动的下界的过程即为鲁棒性验证 (robustness verification). 之所以被称为 “验证”, 是因为在小于对抗扰动下界的范围内, 无论在测试样本上施加什么样的扰动, 都不会使分类器的预测结果发生变化, 这就相当于给分类器的预测结果做了 “认证”. 如前文所述, 最小对抗扰动同样满足该性质, 且是满足该性质的最大值. 最优鲁棒性验证即精确计算最小对抗扰动的大小.

形式化地, 给定分类器 $c:\mathbb{X}\to\mathbb{Y}$ 和测试样例 $(\boldsymbol{x},y)\in\mathbb{X}\times\mathbb{Y}$, 最优鲁棒性验证的目标是求解如下优化问题:

$$\max_{\epsilon}\ \epsilon \quad \text{s.t. } c(\boldsymbol{x}+\boldsymbol{\delta})=y,\ \forall\boldsymbol{\delta}\in\mathbb{X},\ \|\boldsymbol{\delta}\|\leqslant\epsilon. \tag{2}$$

容易看出, 最优对抗攻击和最优鲁棒性验证满足如下关系:

$$\|\boldsymbol{\delta}^*\|=\epsilon^*. \tag{3}$$

对于很多模型来说, 精确求解优化问题 (2) 非常困难. 在这些情况下, 鲁棒性验证往往仅能给出优化问题 (2) 的非平凡可行解, 即计算最小对抗扰动的下界.

2.4 验证鲁棒错误率和验证模型

鲁棒性验证是针对每一个测试样本进行的. 问题是, 如何评估模型整体的对抗鲁棒性呢? 验证鲁棒错误率 (certified/verified robust error) 用于评估模型的 (验证) 对抗鲁棒性. 针对特定的模型和鲁棒性验证方法, 验证鲁棒错误率指的是最小对抗扰动下界 (通过给定的鲁棒性验证方法求得) 小于或等于目标半径的比例. 形式化地, 验证鲁棒错误率定义如下:

$$\mathrm{cre}(\epsilon)=\mathbb{E}_{(\boldsymbol{x},y)}\left[\mathbf{1}\{\underline{\epsilon}(\boldsymbol{x},y)\leqslant\epsilon\}\right], \tag{4}$$

其中 $\underline{\epsilon}(\boldsymbol{x},y)$ 为通过给定的鲁棒性验证方法求得的对抗扰动下界. 显然, 倘若鲁棒性验证方法总能给出非平凡解, 那么如果目标半径 ϵ 选定为 0, 则验证鲁棒错误率即为分类错误率.

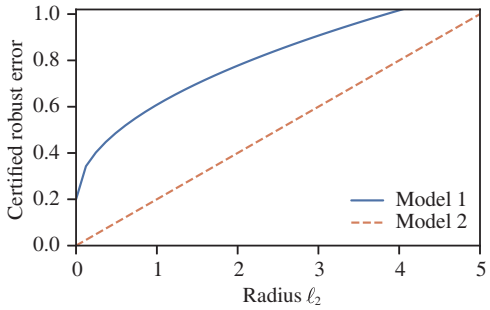

图 1　(网络版彩图) 验证鲁棒错误率曲线

Figure 1　(Color online) Curves of certified robust errors

需要特别强调的是, 验证鲁棒错误率不仅和模型有关, 还和鲁棒性验证方法有关. 验证鲁棒错误率的下界对应于最优鲁棒性验证. 所以, 为了量化模型的对抗鲁棒性, 鲁棒性验证方法是必不可少的. 我们将模型和相应的鲁棒性验证方法合称为验证模型 (verified model).

由于验证鲁棒错误率是关于目标半径的函数, 仅仅用某个半径取值上的验证鲁棒错误率并不能准确地衡量模型的对抗鲁棒性. 为了进行直观且全面的比较, 通常会画出其函数图像, 即验证鲁棒错误率曲线. 例如, 图 1 中曲线 2 处处在曲线 1 的下方, 所以曲线 2 所对应的验证模型体现了较好的验证对抗鲁棒性.

2.5　K- 近邻分类器

K- 近邻分类器将测试样本在训练集中 K 个近邻样本标记的多数投票结果作为最终的预测结果. 形式化地, 记 $\mathbb{S}=\{(\boldsymbol{x}_1,y_1),\dots,(\boldsymbol{x}_n,y_n)\}$ 为训练集, 其中对于任意 $i\in[n]$, 有 $(\boldsymbol{x}_i,y_i)\in\mathbb{X}\times\mathbb{Y}$. 对于任意样本 $\boldsymbol{x}\in\mathbb{X}$ 和正整数 $k\leqslant n$, 函数 $\pi(k,\boldsymbol{x};\mathbb{S})$ 返回训练集中第 k 近邻样本的下标 (本文取欧式距离). K- 近邻分类器的预测函数为

$$c(\boldsymbol{x})=\mathrm{mode}(y_{\pi(1,\boldsymbol{x};\mathbb{S})},\dots,y_{\pi(K,\boldsymbol{x};\mathbb{S})}),\tag{5}$$

其中 mode 表示取众数.

3　K- 近邻鲁棒性验证

本节探讨两种鲁棒性验证方法 —— 约束放松法和随机平滑法, 并分别给出 K- 近邻分类器的解决方案, 进而分析其优劣.

3.1　约束放松法

约束放松法是一种从优化角度直接对原始分类器进行鲁棒性验证的方法. 其基本思想是, 放松最优对抗攻击的约束项, 使得优化问题易于求解, 进而得到最小对抗扰动的下界.

如前所述, 优化问题 (1) 的最优点即为最小对抗扰动. 如果我们对 (1) 的约束项进行放松, 也即增大可行域, 那么新的优化问题的最优值即是最小对抗扰动的下界. 从而, 通过求解放松后的优化问题, 即实现了鲁棒性验证.

此处的关键问题在于, 我们要如何放松约束项, 放松到何种程度才能使得优化问题易于求解, 同时得到尽可能紧致的最小对抗扰动下界. 一般来说, 我们必须要在 “更易于求解” 和 “更紧致的最小对

抗扰动下界” 之间做折中. 对于许多复杂的模型, 为了使得优化问题易于求解, 约束项常被设置得过于宽松. 在这种情况下, 约束放松法求解得到的鲁棒性验证结果往往也不尽如人意.

下面我们介绍一种以约束放松法为基础的 K- 近邻分类器的鲁棒性验证方法.

首先考虑一个简单的三元组问题: 给定 $\boldsymbol{x}$, $\boldsymbol{x}^+$, $\boldsymbol{x}^-\in\mathbb{X}$ 三个样本, 对 $\boldsymbol{x}$ 添加扰动 $\boldsymbol{\delta}$ 使得 $\boldsymbol{x}+\boldsymbol{\delta}$ 到 $\boldsymbol{x}^-$ 的距离小于或等于到 $\boldsymbol{x}^+$ 的距离, 目标是扰动 $\boldsymbol{\delta}$ 尽可能小. 这一问题可以形式化为如下优化问题:

$$\min_{\boldsymbol{\delta}}\ \|\boldsymbol{\delta}\|\quad \text{s.t. } \|\boldsymbol{x}+\boldsymbol{\delta}-\boldsymbol{x}^-\|\leqslant\|\boldsymbol{x}+\boldsymbol{\delta}-\boldsymbol{x}^+\|. \tag{6}$$

该问题的最优值为

$$\frac{\left[\|\boldsymbol{x}^--\boldsymbol{x}\|^2-\|\boldsymbol{x}^+-\boldsymbol{x}\|^2\right]_+}{2\|\boldsymbol{x}^--\boldsymbol{x}^+\|}. \tag{7}$$

其中 $[\cdot]_+$ 表示 $\max(\cdot,0)$. 该最优值即为 $\boldsymbol{x}$ 到 $\boldsymbol{x}^+$ 和 $\boldsymbol{x}^-$ 之间中垂线的距离 (当然, 前提是 $\boldsymbol{x}$ 距离 $\boldsymbol{x}^+$ 更近, 否则上式取值为 0). 下面我们简要给出式 (7) 的证明.

证明 式 (6) 等价于如下优化问题:

$$\min_{\boldsymbol{\delta}}\ \boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\delta} \tag{8}$$

$$\text{s.t. }\ \boldsymbol{a}^{\mathrm{T}}\boldsymbol{\delta}\leqslant b, \tag{9}$$

其中

$$\boldsymbol{a}=(\boldsymbol{x}^+-\boldsymbol{x}^-),$$

$$b=\frac{1}{2}\left(\|\boldsymbol{x}-\boldsymbol{x}^+\|^2-\|\boldsymbol{x}-\boldsymbol{x}^-\|^2\right).$$

对偶函数为

$$g(\lambda)=\inf_{\boldsymbol{\delta}}\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\delta}+\lambda(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\delta}-b) \tag{10}$$

$$=-\frac{1}{4}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}\lambda^2-b\lambda, \tag{11}$$

这里 inf 在 $\boldsymbol{\delta}=-\lambda\boldsymbol{a}/2$ 处取得. 于是, 对偶问题为

$$\max_{\lambda\geqslant0}\ -\frac{1}{4}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}\lambda^2-b\lambda, \tag{12}$$

其最优点为

$$\left[-\frac{2b}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}}\right]_+, \tag{13}$$

最优值为

$$\begin{cases}0, & \text{如果 } b\geqslant0,\\ \dfrac{b^2}{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}}, & \text{否则.}\end{cases} \tag{14}$$

根据 Slater 条件, 如果 $\boldsymbol{x}^+$ 和 $\boldsymbol{x}^-$ 不重合, 则强对偶性满足 (容易验证, 该问题下, 即使 $\boldsymbol{x}^+=\boldsymbol{x}^-$, 下述最优值依然成立). 所以式 (6) 的最优值为

$$\left[\frac{-b}{\sqrt{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}}}\right]_+=\frac{\left[\|\boldsymbol{x}^--\boldsymbol{x}\|^2-\|\boldsymbol{x}^+-\boldsymbol{x}\|^2\right]_+}{2\|\boldsymbol{x}^--\boldsymbol{x}^+\|}. \tag{15}$$

从而得证.

为叙述方便, 我们将该值记为关于三个样本的函数 $\tilde{\epsilon}:\mathbb{X}\times\mathbb{X}\times\mathbb{X}\to\mathbb{R}^+$, 即

$$\tilde{\epsilon}(\boldsymbol{x},\boldsymbol{x}^+,\boldsymbol{x}^-)=\frac{\left[\|\boldsymbol{x}^- -\boldsymbol{x}\|^2-\|\boldsymbol{x}^+-\boldsymbol{x}\|^2\right]_+}{2\|\boldsymbol{x}^- -\boldsymbol{x}^+\|}. \tag{16}$$

接下来形式化描述 K- 近邻的对抗鲁棒性问题. 简单起见令 K 为奇数. 记测试样例为 $(\boldsymbol{x},y)$, 训练集中与该样本标记相同的样本构成集合 $\mathbb{S}^+=\{\boldsymbol{x}_1^+,\ldots,\boldsymbol{x}_{n^+}^+\}$, 与该样本标记不同的样本构成集合 $\mathbb{S}^-=\{\boldsymbol{x}_1^-,\ldots,\boldsymbol{x}_{n^-}^-\}$. 考虑二元组 $(\mathbb{I},\mathbb{J})$, 其中 $\mathbb{I}\subseteq[n^+]$, $|\mathbb{I}|=(K-1)/2$, $\mathbb{J}\subseteq[n^-]$, $|\mathbb{J}|=(K+1)/2$. 我们将优化问题定义如下: 在 $\boldsymbol{x}$ 加上扰动 $\boldsymbol{\delta}$, 使得 $\boldsymbol{x}+\boldsymbol{\delta}$ 到任意 $\boldsymbol{x}_j^-$ 的距离小于或等于到任意 $\boldsymbol{x}_i^+$ 的距离, 其中 $j\in\mathbb{J}$, $i\in[n^+]-\mathbb{I}$, 目标是扰动 $\boldsymbol{\delta}$ 尽可能小. 该优化问题的约束项保证了以 $\mathbb{J}$ 中元素为下标的样本全部在 $\boldsymbol{x}+\boldsymbol{\delta}$ 的 K- 近邻中. 形式化地, 该优化问题定义如下:

$$\min_{\boldsymbol{\delta}}\ \|\boldsymbol{\delta}\|\ \ \text{s.t.}\ \ \ \|\boldsymbol{x}+\boldsymbol{\delta}-\boldsymbol{x}_j^-\|\leqslant\|\boldsymbol{x}+\boldsymbol{\delta}-\boldsymbol{x}_i^+\|,\ \ \forall i\in[n^+]-\mathbb{I},\ \forall j\in\mathbb{J}. \tag{17}$$

容易证明, 上式是一个凸二次规划问题. 我们将式 (17) 的最优值记为 $\epsilon^{(\mathbb{I},\mathbb{J})}$, 则容易验证最小对抗扰动满足

$$\epsilon^*\geqslant\min_{(\mathbb{I},\mathbb{J})}\epsilon^{(\mathbb{I},\mathbb{J})}. \tag{18}$$

特别地, 如果分类问题是二分类问题, 即 $|\mathbb{Y}|=2$, 那么上式等号成立. 可以看到, 通过遍历 $(\mathbb{I},\mathbb{J})$ 二元组并求解每一个凸二次规划问题, 我们可以得到最小对抗扰动的下界; 如果是二分类问题, 我们甚至得到了最小对抗扰动的精确值, 即做到了最优鲁棒性验证. 但是, 需要注意的是, 优化问题的总数, 也即二元组 $(\mathbb{I},\mathbb{J})$ 的数目是

$$\binom{n^-}{\frac{K+1}{2}}\binom{n^+}{\frac{K-1}{2}}. \tag{19}$$

根据斯特灵公式 (Stirling's formula), 其规模 $\Omega((\frac{n}{K})^K)$ 随 K 的增大呈指数级增长. 所以, 直接遍历二元组并求解所有优化问题的做法很难应用到实际问题中. 为此, 我们考虑继续对该下界进行放松, 并得到如下定理.

定理1 (K- 近邻分类器的最小对抗扰动下界) K- 近邻分类器的最小对抗扰动满足

$$\epsilon^*\geqslant s\,\text{thmin}_{j\in[n^-]}\ s\,\text{thmax}_{i\in[n^+]}\ \tilde{\epsilon}(\boldsymbol{x},\boldsymbol{x}_i^+,\boldsymbol{x}_j^-), \tag{20}$$

其中 $s=(K+1)/2$, $s\,\text{thmin}$ 表示取第 s 小的元素, $s\,\text{thmax}$ 表示取第 s 大的元素.

该定理最早在我们的前期工作 [16] 中作为最近邻对抗鲁棒性分析的扩展提出, 这里我们给出完整的证明, 并在后文做更多的对比分析.

证明 通过比较式 (6) 和 (17) 的约束项, 我们有

$$\epsilon^{(\mathbb{I},\mathbb{J})}\geqslant\max_{i\in[n^+]-\mathbb{I},\ j\in\mathbb{J}}\tilde{\epsilon}(\boldsymbol{x},\boldsymbol{x}_i^+,\boldsymbol{x}_j^-). \tag{21}$$

代入式 (18), 有

$$\epsilon^*\geqslant\min_{\mathbb{I},\mathbb{J}}\epsilon^{(\mathbb{I},\mathbb{J})} \tag{22}$$

$$\geqslant\min_{\mathbb{I},\mathbb{J}}\ \max_{i\in[n^+]-\mathbb{I},\ j\in\mathbb{J}}\ \tilde{\epsilon}(\boldsymbol{x},\boldsymbol{x}_i^+,\boldsymbol{x}_j^-) \tag{23}$$

$$\geqslant \min_{\mathbb{I},\mathbb{J}} \max_{j\in\mathbb{J}} \max_{i\in[n^+]-\mathbb{I}} \tilde{\epsilon}(\boldsymbol{x}, \boldsymbol{x}_i^+, \boldsymbol{x}_j^-) \tag{24}$$

$$\geqslant \min_{\mathbb{I},\mathbb{J}} \max_{j\in\mathbb{J}} s\text{thmax}_{i\in[n^+]} \tilde{\epsilon}(\boldsymbol{x}, \boldsymbol{x}_i^+, \boldsymbol{x}_j^-) \tag{25}$$

$$\geqslant \min_{\mathbb{I},\mathbb{J}} s\text{thmin}_{j\in[n^-]} s\text{thmax}_{i\in[n^+]} \tilde{\epsilon}(\boldsymbol{x}, \boldsymbol{x}_i^+, \boldsymbol{x}_j^-) \tag{26}$$

$$= s\text{thmin}_{j\in[n^-]} s\text{thmax}_{i\in[n^+]} \tilde{\epsilon}(\boldsymbol{x}, \boldsymbol{x}_i^+, \boldsymbol{x}_j^-). \tag{27}$$

从而得证.

于是, 我们只需要求解 n^+n^- 个有闭式解的优化问题, 即可得到 K- 近邻分类器的最小对抗扰动的下界, 也即实现了鲁棒性验证.

约束放松法的局限性. 从我们的证明中可以看出, 上述约束放松法有一重要局限性: K 越大, 约束放松程度越大, 求得的最小对抗扰动下界越宽松, 进而 K- 近邻分类器的验证对抗鲁棒性越弱. 这显然与我们的直观认识相悖. 通常认为 K 越大, 分类器利用的近邻信息越多, K- 近邻分类器的对抗鲁棒性应越强. 这里的关键是要区分对抗鲁棒性和验证对抗鲁棒性. 对抗鲁棒性本身和验证方法无关, 而验证对抗鲁棒性依赖于验证方法. 之所以会出现上述矛盾, 是因为约束放松法 "放松过度" 导致鲁棒性验证结果不理想.

约束放松法的适用范围. 由于上述局限性, 约束放松法通常只适用于 K 值较小的情况; 当 K 值较大时, 约束放松法通常难以达到令人满意的验证效果. 我们将在实验部分对此做进一步探讨.

3.2 随机平滑法

随机平滑法 [17] 是一种通用的鲁棒性验证方法, "理论" 上可以适用于任意分类器. 在实际应用中, 随机平滑法通常仅被看作一种针对神经网络的验证防御方法, 其目标是提升神经网络模型的验证对抗鲁棒性. 其基本思想是, 先将基分类器转化为 "平滑" 分类器, 然后验证平滑分类器的对抗鲁棒性. 随机平滑法在难以对原始分类器进行直接的对抗鲁棒性验证的情况下, 提供了一种替代方案. 需要注意的是, 随机平滑法验证的是平滑分类器, 而不是基分类器本身.

平滑分类器. 平滑分类器的预测过程依赖于基分类器: 对给定样本注入高斯白噪声, 将基分类器在噪声样本上的输出概率最大的类别作为最终的预测结果. 形式化地, 记 $c:\mathbb{X}\to\mathbb{Y}$ 为基分类器, $\boldsymbol{x}\in\mathbb{X}$ 为测试样本, 则平滑分类器 $c':\mathbb{X}\to\mathbb{Y}$ 在 $\boldsymbol{x}$ 上的预测结果为

$$c'(\boldsymbol{x}) = \arg\max_{y\in\mathbb{Y}} \Pr_{\boldsymbol{\delta}}[c(\boldsymbol{x}+\boldsymbol{\delta}) = y \mid \boldsymbol{\delta}\sim\mathcal{N}(0,\sigma^2\boldsymbol{I})], \tag{28}$$

其中超参数噪声强度 σ 用于在鲁棒性和准确性之间进行折中.

平滑分类器的鲁棒性验证. 记 y_A 为基分类器输出概率最大的类别 (即平滑分类器的输出), p_A 为输出 y_A 的概率; y_B 为基分类器输出概率第二大的类别, p_B 为输出 y_B 的概率. 记

$$\epsilon_{\text{smooth}} = \frac{\sigma}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right), \tag{29}$$

其中, $\Phi^{-1}(\cdot)$ 为标准正态分布的累积分布函数的反函数 (即分位数函数, 函数图像见图 2). 那么对于任意 $\|\boldsymbol{\delta}\|\leqslant\epsilon_{\text{smooth}}$, 都有 $c'(\boldsymbol{x}) = y_A$. 换句话说, ϵ_{smooth} 是平滑分类器 c' 的最小对抗扰动的下界, 计算 ϵ_{smooth} 的过程即是平滑分类器的一个鲁棒性验证方法. 注意到, $\Phi^{-1}(\cdot)$ 是一个单调递增函数, 从而将 p_A 替换为其下界 $\underline{p_A}$, 将 p_B 替换为其上界 $\overline{p_B}$ 后, 我们得到的仍然是平滑分类器的最小对抗扰动的下

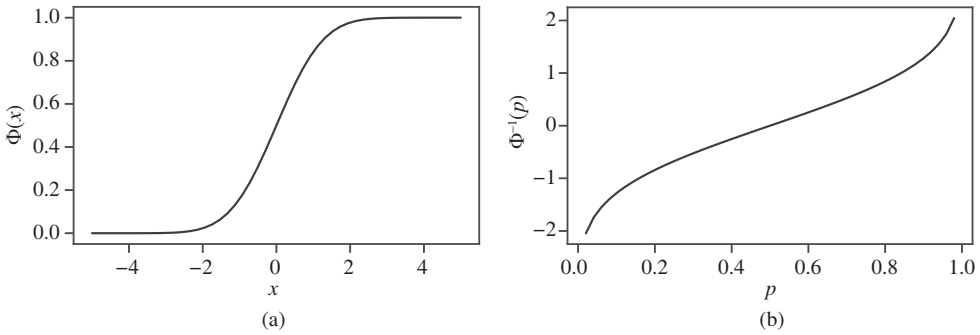

图 2　标准正态分布的 (a) 累积分布函数及 (b) 分位数函数

Figure 2　(a) Cumulative distribution function and (b) percent point function of the standard normal distribution

界 (相比于使用 p_A 和 p_B, 新的下界更宽松了), 即

$$\epsilon'_{\text{smooth}} = \frac{\sigma}{2}\left(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})\right) \leqslant \epsilon_{\text{smooth}}. \tag{30}$$

另外, 在实际应用中, 对于许多分类器 (神经网络模型, K- 近邻分类器等), 即使是 $\underline{p_A}$ 和 $\overline{p_B}$, 我们也很难计算出其精确值. 为了解决这一问题, 随机平滑法采用蒙特卡洛 (Monte Carlo) 方法估计 $\underline{p_A}$ 和 $\overline{p_B}$, 使得以任意高的概率 (概率为 $1-\alpha$, 这里 α 为超参数, 通常取 0.001), 满足 $\underline{p_A} \leqslant p_A$ 和 $\overline{p_B} \geqslant p_B$.

随机平滑法的适用范围. 虽然上述讨论对任意基分类器都成立. 但是, 在实际应用中, 随机平滑法所适用的基分类器并不是任意的. 这里的关键是, 为了得到有意义的鲁棒性验证结果, 基分类器必须对高斯白噪声具有一定的鲁棒性. 例如, 如果基分类器是线性分类器, 那么容易验证随机平滑后的分类器依然是线性分类器本身. 再比如, 普通的神经网络模型由于对高斯白噪声并不鲁棒, 往往并不能利用随机平滑法实现有意义的鲁棒性验证. 所以, 在对神经网络模型应用随机平滑法时, 往往需要利用高斯白噪声数据增广法重新训练模型使其对高斯白噪声具有一定的鲁棒性. 需要强调的是, 由于随机平滑法的特殊性, 除了线性模型和神经网络, 目前尚没有工作将随机平滑法应用到其他分类器上.

随机平滑法和 K- 近邻分类器. 相比于神经网络, K- 近邻分类器对高斯白噪声往往具有更强的鲁棒性. 基于此我们发现, K- 近邻作为基分类器的时候, 可以得到较好的鲁棒性验证结果. 这一现象在 K 值较大的时候表现得尤为明显 —— 利用了更多的近邻信息的分类器有更强的对抗鲁棒性. 这一特点恰恰和约束放松法形成互补.

随机平滑法的局限性. 随机平滑法的局限性有三方面.

- 首先, 随机平滑法验证的对象不是基分类器, 而是基于基分类器构造的平滑分类器. 所以严格来说, 随机平滑法并不是针对基分类器的形式化鲁棒性验证方法. 已有研究表明, 平滑分类器相比于基分类器, 往往会损失一定的 (正常样本) 分类正确率以换取验证鲁棒性.
- 其次, 应用平滑分类器时, 需要调用多次基分类器. 例如, 平滑分类器做单次预测时往往需要调用超过 100 次基分类器; 而平滑分类器做单次验证时, 甚至需要调用超过 100000 次基分类器.
- 最后, 对于多数基分类器来说, 随机平滑法的鲁棒性验证结果都是基于概率的, 而并非确定性的. 当然, 应用蒙特卡洛方法时, 我们可以使结果以任意高的概率成立. 然而, 目标概率越大, 对基分类器的访问次数就越多, 这又进一步加剧了运行开销问题.

基于以上考虑, 我们将随机平滑法视作约束放松法的替代折中方案, 用于处理约束放松法验证结果太过宽松的情况. 对应到 K- 近邻分类器, 随机平滑法可以一定程度上弥补约束放松法在 K 值较大时的不足.

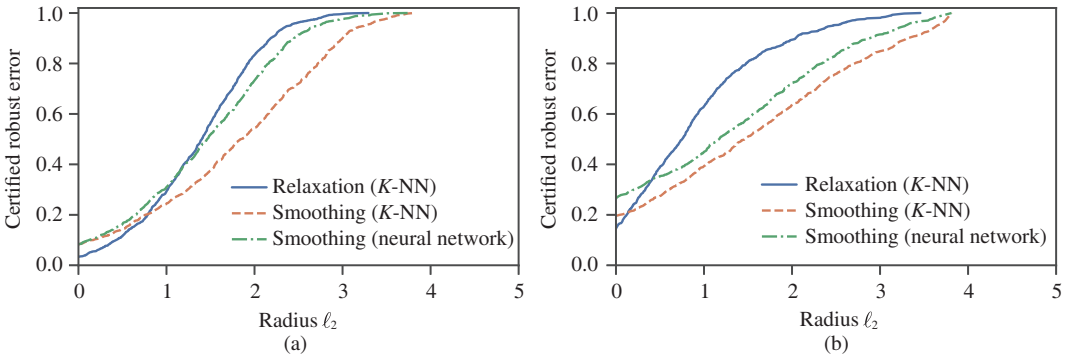

图 3 (网络版彩图) K- 近邻和神经网络鲁棒性验证对比

Figure 3 (Color online) Comparison between K-NN and neural networks for robustness verification. (a) MNIST; (b) Fashion-MNIST

4 实验

我们以约束放松法和随机平滑法分别对 K- 近邻分类器做鲁棒性验证, 比较二者的验证结果.

我们在 MNIST [18] 和 Fashion-MNIST [19] 上进行实验, 二者各有 60000 个训练样本和 10000 个测试样本, 属性向量维度为 784, 类别个数为 10. 选用这两个数据集主要是考虑到两个因素: 一是 K- 近邻分类器在 MNIST 和 Fashion-MNIST 上的分类错误率较低, 从而可以更准确地比较验证鲁棒错误率 (降低分类错误率的影响); 二是 MNIST 和 Fashion-MNIST 广泛应用于鲁棒性验证问题中, 从而我们可以方便地与最新的鲁棒神经网络的性能做对比. 我们从测试集上随机选取 1000 个样本对模型 (K- 近邻或者神经网络) 进行鲁棒性验证, 并绘制出相应的验证鲁棒错误率曲线.

4.1 主要结果

K- 近邻约束放松法和随机平滑法的验证鲁棒错误率曲线的典型结果见图 3. 为了做更为精确的对比, 我们将图中半径取值为 0, 1, 2, 3 上的验证鲁棒错误率记在表 1 和 2 上.

这里, 约束放松法作用于 $K = 1$ 的情况; 随机平滑法作用于 $K = 51$ 的情况, 噪声超参数 $\sigma = 1$. 参数的取值问题我们将在后面的实验中展开讨论. 作为对比, 我们使用高斯白噪声数据增广法训练深度神经网络 [17], 并利用随机平滑法对其进行鲁棒性验证, 噪声参数同样取 $\sigma = 1$. 从图表中我们可以看出:

- 平滑分类器在分类错误率 (半径为 0 处的验证鲁棒错误率) 上有一定劣势 (见表 1 和 2). 这验证了随机平滑法会损失一定的分类准确率的论断. 我们将在 4.3 小节中进一步讨论应用随机平滑法时面临的分类错误率和验证鲁棒错误率之间的权衡.
- 随机平滑法作用在 K- 近邻分类器 (平滑 K- 近邻分类器) 上时的结果显著优于神经网络. 例如, 在 Fashion-MNIST 半径为 0, 1, 2, 3 上的实验数据中, 平滑 K- 近邻分类器的验证鲁棒错误率比平滑神经网络分别低 6.8%, 5.7%, 8.4% 和 6.7%. 这体现了 (平滑) K- 近邻分类器更好的验证鲁棒性.

4.2 近邻参数 K 对验证结果的影响

如前讨论, 近邻参数 K 对约束放松法和随机平滑法的影响程度有较大差异, 因此我们分别对两种方法进行实验探究.

表 1　MNIST 验证鲁棒错误率 (%)

Table 1　Certified robust errors on MNIST

Radius ℓ_2	Constraint relaxation (K-NN)	Random smoothing (K-NN)	Random smoothing (neural network)
0	**3.3**	8.2	8.3
1	29.3	**24.4**	30.9
2	83.3	**54.4**	73.2
3	99.6	**89.9**	97.6

表 2　Fashion-MNIST 验证鲁棒错误率 (%)

Table 2　Certified robust errors on Fashion-MNIST

Radius ℓ_2	Constraint relaxation (K-NN)	Random smoothing (K-NN)	Random smoothing (neural network)
0	**14.5**	19.6	26.4
1	63.0	**39.2**	44.9
2	89.3	**63.6**	72.1
3	98.0	**84.8**	91.5

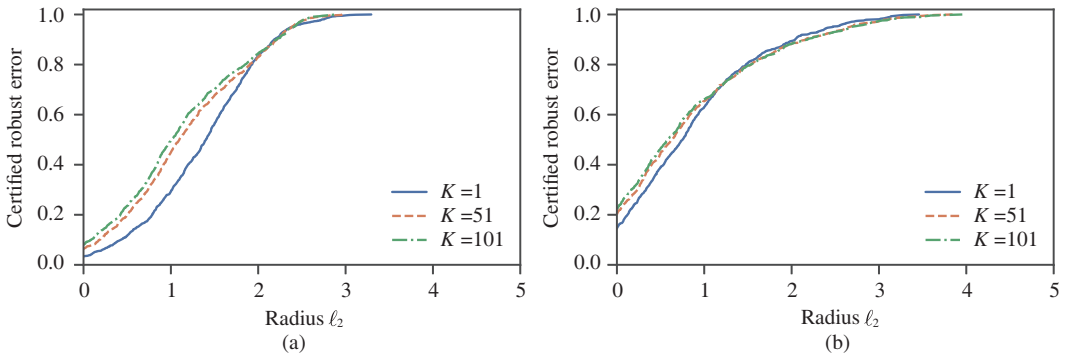

图 4　(网络版彩图) K- 近邻约束放松法随 K 的变化

Figure 4　(Color online) K-NN constraint relaxation method for different K. (a) MNIST; (b) Fashion-MNIST

约束放松法的结果见图 4. 可以看到, 随着 K 的增大, 验证鲁棒错误率整体呈上升趋势. 这是由于在 K 较大时, 约束放松法求得的对抗扰动下界太过宽松, 难以反映 K- 近邻分类器真实的对抗鲁棒性. 所以, 约束放松法通常只适用于 K 值较小的情况.

随机平滑法的结果见图 5. 不同于约束放松法, 随机平滑法求得的验证对抗鲁棒性随着 K 的增大而增强. 这一现象符合预期: 利用更多近邻信息的分类器有更好的对抗鲁棒性. 所以, 当 K 值较大时, 随机平滑法可以更为准确地评估 K- 近邻分类器的对抗鲁棒性 (确切地说, 评估的是平滑 K- 近邻分类器).

4.3　噪声参数对随机平滑法的影响

随机平滑法需要选择合适的噪声参数 σ. 不同 σ 对应的平滑 K- 近邻分类器的验证结果见图 6. 可以看到, 与平滑神经网络类似 [17], σ 起到折中准确性和鲁棒性的作用: 较小的 σ 对应于较大的分类准确率; 较大的 σ 对应于较好的验证对抗鲁棒性.

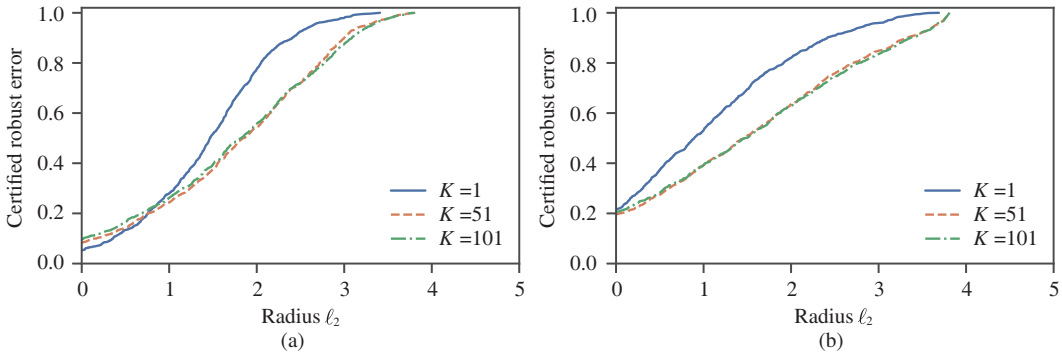

图 5 (网络版彩图) K- 近邻随机平滑法随 K 的变化

Figure 5 (Color online) K-NN randomized smoothing method for different K. (a) MNIST; (b) Fashion-MNIST

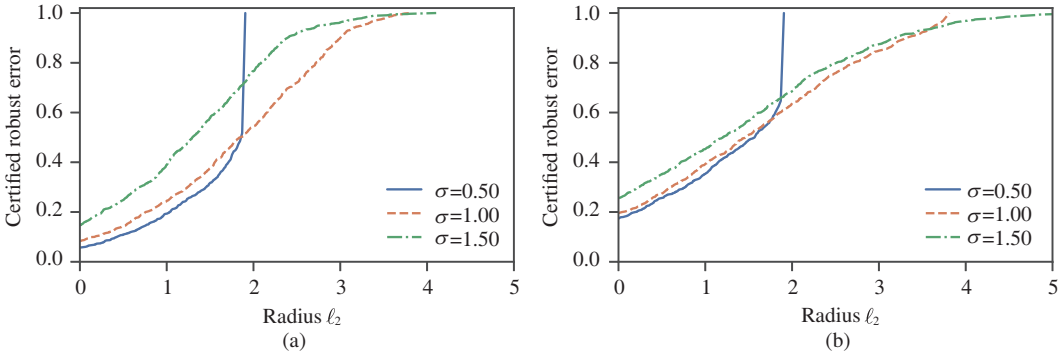

图 6 (网络版彩图)K- 近邻随机平滑法随 σ 的变化

Figure 6 (Color online) K-NN randomized smoothing method for different σ. (a) MNIST; (b) Fashion-MNIST

5 相关工作

目前关于 K- 近邻分类器对抗鲁棒性的研究大多还集中在对抗攻击上. 例如, 为了应对 K- 近邻分类器不连续的特点, 有研究提出攻击 K- 近邻分类器的可微替代模型 [20,21]; 另有工作提出启发式攻击方法, 例如在测试样本和训练样本的连线上搜索对抗样本 [21,22]; 最近的一些研究工作 (包括我们的前期工作) 更进一步, 考虑利用 K- 近邻空间划分的特点设计对抗攻击方法 [16,23]. 但是, 正如我们在第 2 节中讨论的, 对抗攻击计算的是最小对抗扰动的上界, 而鲁棒性验证计算的是最小对抗扰动的下界. 只有下界才能为模型提供严格的形式化鲁棒性保证. 虽然神经网络模型的鲁棒性验证问题在最近几年里受到了越来越多的关注 [4~10], 现阶段关于 K- 近邻分类器的鲁棒性验证问题的研究还相对较少 [16]. 与之相对的是, 尽管缺少明确的研究结果的支撑, 研究者普遍认为 K- 近邻分类器有较强的对抗鲁棒性, 一些工作将最近邻模型看作一种启发式的经验防御方法, 并部署到实际应用中 [14,15]. 如前文所述, 缺少验证鲁棒性保证的机器学习模型会带来潜在的安全性风险. K- 近邻分类器的对抗鲁棒性, 特别是验证鲁棒性问题需要得到更多的关注.

6 结束语

本文提出并讨论了两种 K- 近邻分类器的鲁棒性验证方法: 约束放松法和随机平滑法. 约束放松

法是一种严格的形式化验证方法, 它可以直接验证 K- 近邻分类器的对抗鲁棒性. 然而在 K 值较大的情况下, 约束放松法的验证效果并不理想, 和最小对抗扰动相差较大. 随机平滑法利用了 K- 近邻分类器对高斯白噪声不敏感的特点, 在 K 值较大的情况下可以取得非常理想的验证效果, 这在很大程度上弥补了约束放松法的不足. 随机平滑法同时也存在着一些缺点. 其一是验证对象是平滑分类器, 而不是基分类器本身; 其二是应用平滑分类器做预测和验证时需要较多额外运行开销; 其三是随机平滑法是基于概率的, 严格来说不是形式化验证方法.

K- 近邻鲁棒性验证的工作才刚刚开始. 约束放松法的优势恰恰是随机平滑法的劣势, 而随机平滑法的优势同时又是约束放松法的劣势. 结合二者优势, 克服二者劣势的鲁棒性验证方法将是一个值得研究的方向.

参考文献

1 Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. In: Proceedings of International Conference on Learning Representations, 2013

2 Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of International Conference on Learning Representations, 2015

3 Biggio B, Roli F. Wild patterns: ten years after the rise of adversarial machine learning. Pattern Recogn, 2018, 84: 317–331

4 Weng T, Zhang H, Chen P, et al. Evaluating the robustness of neural networks: an extreme value theory approach. In: Proceedings of International Conference on Learning Representations, 2018

5 Wong E, Kolter J Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In: Proceedings of International Conference on Machine Learning, 2018. 5283–5292

6 Weng T W, Zhang H, Chen H, et al. Towards fast computation of certified robustness for relu networks. In: Proceedings of International Conference on Machine Learning, 2018. 5273–5282

7 Gehr T, Mirman M, Drachsler-Cohen D, et al. Ai2: safety and robustness certification of neural networks with abstract interpretation. In: Proceedings of IEEE Symposium on Security and Privacy, 2018. 3–18

8 Wang S, Pei K, Whitehouse J, et al. Efficient formal safety analysis of neural networks. In: Proceedings of Advances in Neural Information Processing Systems, 2018. 6367–6377

9 Zhang H, Weng T W, Chen P Y, et al. Efficient neural network robustness certification with general activation functions. In: Proceedings of Advances in Neural Information Processing Systems, 2018. 4939–4948

10 Zhang H, Zhang P, Hsieh C J. RecurJac: an efficient recursive algorithm for bounding Jacobian matrix of neural networks and its applications. In: Proceedings of AAAI Conference on Artificial Intelligence, 2019. 5757–5764

11 Boiman O, Shechtman E, Irani M. In defense of nearest-neighbor based image classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008. 1–8

12 Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res, 2009, 10: 207–244

13 Xi X, Keogh E, Shelton C, et al. Fast time series classification using numerosity reduction. In: Proceedings of International Conference on Machine Learning, 2006. 1033–1040

14 Dubey A, van der Maaten L, Yalniz Z, et al. Defense against adversarial images using web-scale nearest-neighbor search. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 8767–8776

15 Papernot N, McDaniel P. Deep k-nearest neighbors: towards confident, interpretable and robust deep learning. 2018. ArXiv: 1803.04765

16 Wang L, Liu X, Yi J, et al. Evaluating the robustness of nearest neighbor classifiers: a primal-dual perspective. 2019. ArXiv: 1906.03972

17 Cohen J M, Rosenfeld E, Kolter J Z. Certified adversarial robustness via randomized smoothing. In: Proceedings of International Conference on Machine Learning, 2019. 1310–1320

18 Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE, 1998, 86: 2278–2324

19 Xiao H, Rasul K, Vollgraf R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. 2017. ArXiv: 1708.07747
20 Papernot N, McDaniel P D, Goodfellow I J. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. 2016. ArXiv: 1605.07277
21 Sitawarin C, Wagner D. On the robustness of deep k-nearest neighbors. 2019. ArXiv: 1903.08333
22 Amsaleg L, Bailey J, Barbe D, et al. The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality. In: Proceedings of IEEE Workshop on Information Forensics and Security, 2017. 1–6
23 Yang Y Y, Rashtchian C, Wang Y, et al. Robustness for non-parametric classification: a generic attack and defense. In: Proceedings of International Conference on Artificial Intelligence and Statistics, 2020

Robustness verification of K-NN classifiers via constraint relaxation and randomized smoothing

Lu WANG[1,2] & Yuan JIANG[1,2*]

1. *National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210023*, China*;
2. *Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing University, Nanjing* 210023*, China*
* Corresponding author. E-mail: jiangy@lamda.nju.edu.cn

Abstract We study the robustness verification problem for K-NN classifiers. The objective of formal robustness verification is to find the exact minimal adversarial perturbation or a guaranteed lower bound of the perturbation. We find that the robustness verification of K-NN classifiers could be formalized as a series of quadratic programming problems. Solving these quadratic programming problems is not possible in general because the number of problems grows exponentially with respect to K. The constraint relaxation method is proposed to compute the lower bound of the minimal adversarial perturbation in polynomial time. However, we find that the resulting lower bound tends to be extremely loose when K is large; hence, K-NN with a large K being less robust is counterintuitive. To tackle this issue, we propose to employ the randomized smoothing method to verify the robustness of K-NN classifiers. By exploiting the resistance of K-NN to random Gaussian noise, the randomized smoothing method achieves high performance in verification. Our experiments on benchmark datasets show that the smoothed K-NN classifier is more verifiably robust than state-of-the-art robust neural networks.

Keywords supervised learning, adversarial machine learning, adversarial robustness, robustness verification, K-NN classifier

Lu WANG was born in 1992. He received his B.S. degree in Computer Science and Technology from Nanjing University, Nanjing, China, in 2014. Currently, he is a Ph.D. candidate in Nanjing University. His research interests mainly include machine learning and data mining.

Yuan JIANG was born in 1976. She received her Ph.D. degree in Computer Science from Nanjing University, China, in 2004. Currently, she is a professor at Nanjing University. Her main research interests include artificial intelligence, machine learning, and data mining.