软件学报 ISSN 1000-9825, CODEN RUXUEW
[doi: 10.13328/j.cnki.jos.007568] [CSTR: 32375.14.jos.007568]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向领域自适应扰动的神经网络鲁棒性形式验证*

王麒扬, 李璟旸, 李国强

(上海交通大学 软件学院, 上海 200240)

通信作者: 李国强, E-mail: li.g@sjtu.edu.cn

摘　要: 神经网络鲁棒性验证作为保障人工智能系统安全性与可靠性的关键技术, 能够为智能决策提供形式化保证. 现有研究普遍基于数据分布各向同性的简化假设, 进行均匀 L_p 范数球邻域上的验证算法开发. 但这一理论框架在面对现实世界复杂的数据特性时显得力不从心, 例如: 数据不同特征对模型预测结果的影响及其对扰动的敏感度表现出差异性; 某些特征受物理约束限制具有不可变性; 特征间还可能存在复杂的相关性结构. 这使得基于均匀扰动域的验证算法难以准确建模现实世界对人工智能系统的鲁棒性需求. 针对上述局限, 提出基于非均匀扰动域的鲁棒性验证框架. 通过结合具体应用领域的数据分布特性, 合理构建符合领域特征的扰动域几何结构, 并对领域自适应扰动进行建模. 在此基础上, 形式化地定义 3 种新型鲁棒性概念: 椭球鲁棒性、掩码局部鲁棒性以及马氏距离鲁棒性, 并提出了相应的鲁棒性验证问题的定义与构建方法. 此外, 设计 NNV4RADAP 算法, 通过构建等价的均匀 L_p 范数球鲁棒性验证问题来将现有的验证算法推广到面向领域自适应扰动的神经网络鲁棒性验证问题上. 实验结果显示, NNV4RADAP 算法可以为神经网络提供更为精准且贴合数据分布特性的鲁棒性保证. 拓宽现有的深度神经网络的鲁棒性的形式定义, 并设计实现面向领域自适应扰动的神经网络鲁棒性的形式验证算法, 研究领域内基于数据分布的鲁棒性定义上的问题, 对于形式验证技术未来在可信人工智能技术中的落地和应用都有指导作用.

关键词: 人工智能安全; 神经网络鲁棒性; 神经网络验证; 领域自适应扰动; 非各向同性数据假设

中图法分类号: TP311

中文引用格式: 王麒扬, 李璟旸, 李国强. 面向领域自适应扰动的神经网络鲁棒性形式验证. 软件学报. http://www.jos.org.cn/1000-9825/7568.htm

英文引用格式: Wang QY, Li JY, Li GQ. Formal Verification for Neural Network Robustness Against Domain-adaptive Perturbations. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7568.htm

Formal Verification for Neural Network Robustness Against Domain-adaptive Perturbations

WANG Qi-Yang, LI Jing-Yang, LI Guo-Qiang

(School of Software, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: As a key technology for ensuring the safety and reliability of artificial intelligence (AI) systems, neural network robustness verification can provide formal guarantees for intelligent decision-making. Existing research is generally based on simplified assumptions of isotropic data distributions to develop verification algorithms based on uniform L_p-norm ball neighborhoods. However, this theoretical framework proves inadequate in the face of the real-world complex data characteristics. For instance, different data features exhibit varying influences on model predictions and sensitivities to perturbations; some features are immutable due to physical constraints; complex correlation structures may exist among features. This makes it difficult for verification algorithms based on uniform perturbation domains to accurately model the robustness requirements of the real world for AI systems. To this end, this study proposes a robustness verification framework based on non-uniform perturbation domains. By combining the data distribution characteristics of specific application domains, the study constructs geometrically structured perturbation domains aligned with domain-specific features and model domain-adaptive

* 基金项目: 国家自然科学基金 (62161146001)
收稿时间: 2025-04-08; 修改时间: 2025-07-18; 采用时间: 2025-10-29; jos 在线出版时间: 2026-03-25

perturbations. On this basis, it formally defines three novel robustness concepts, including ellipsoidal robustness, masked local robustness, and Mahalanobis distance robustness, and proposes the definition and construction methods for corresponding robustness verification problems. Furthermore, the NNV4RADAP algorithm is designed, which extends existing verification algorithms to neural network robustness verification problems for domain-adaptive perturbations by constructing equivalent uniform L_p-norm ball robustness verification problems. The experimental results demonstrate that the NNV4RADAP algorithm can provide more accurate and datadistribution-aligned robustness guarantees for neural networks. This study expands the existing formal definitions of deep neural network robustness and designs and implements a formal verification algorithm of neural network robustness for domain-adaptive perturbations. Additionally, it researches the problem in data distribution-based robustness definitions and provides guidance for the future implementation and application of formal verification techniques in trustworthy AI technologies.

Key words: artificial intelligence security; neural network robustness; neural network verification; domain-adaptive perturbation; non-isotropic data assumption

1 引 言

随着以深度神经网络为代表的人工智能技术与各产业应用深度融合, 软件 1.0 (即手动编写的代码) 与软件 2.0 (可学习的神经网络) 之间的边界越来越模糊, 如何构建可信人工智能系统已经成为学术界与工业界的共同议题. 其中, 在当前人工智能软件安全空间复杂化的背景下, 深度神经网络鲁棒性作为描述深度学习模型对抗输入扰动甚至恶意攻击能力的性质, 凸显越来越重要的作用. 具体来说, 鲁棒性是指对于输入数据中存在的微小扰动及非正常分布的数据, 神经网络模型的输出仍不受影响[1]. 传统的对抗训练[2−9]中基于对抗样本搜索的方法仅能对神经网络鲁棒性的平均性能进行评估, 而无法对最坏情况进行保证, 而这在安全攸关的领域中尤为致命. 因此, 研究者们借鉴了软件验证的理念和方法, 提出了多种技术对神经网络在一定输入域内的鲁棒性及其他性质进行证明, 并在确保神经网络安全性方面起着关键作用[10−16]. 然而, 目前的神经网络验证技术大多基于各向同性假设, 即对输入表征空间中各个维度的扰动或者攻击是等价的[10−14]. 这类假设不适用于如恶意软件检测[17]、智慧医疗[18]、垃圾邮件检测[19]等数据特征在各个维度上不具有各向同性的任务.

早期关于神经网络鲁棒性的研究主要集中在图像处理领域. 研究者通常假设数据在特征空间的所有维度上都表现出相似的分布特性及相同的鲁棒性水平, 因此扰动通常被限制在一个均匀 L_p 范数球内[2−9]. 然而在现实复杂场景中, 这一假设往往会导致不合理的变换, 这也意味着扰动域中的数据都应遵循其所在领域相关的一些约束限制. 最直接的就是考虑数据的不同特征在鲁棒性范畴上存在的异同. Tsipras 等人[20]和 Ilyas 等人[21]对特征的鲁棒性进行了定义, 并在标准训练以及对抗训练中的效应进行了分析, 发现有的特征对微小扰动是鲁棒的, 而有的则十分脆弱, 这实际上是对不同特征的鲁棒尺度的离散解释. 在手写数字识别任务中, 中心数字区域高激活像素相较于边缘像素对扰动更为敏感, 具有更严格的扰动阈值[22]. 此外, 某些特征具有不可变约束. 在网络安全领域的恶意软件检测任务中, 目标是识别一个软件是良性的还是恶意的. 机器学习模型会从可执行文件中提取多种具有语义意义的特征, 同时为了保持可执行文件功能的完整性, 某些关键特征不允许被改变, 任何对这些特征的修改都可能破坏文件的功能进而导致不现实的结果[23]. 除了特征本身的鲁棒性差异以及不可变约束限制外, 特征之间可能存在复杂的依赖关系. 例如, 在垃圾邮件检测任务中, 某些关键词或者特殊符号时常一起出现, 因此词频特征间存在一定的相关性.

在上述应用场景中, 异质化的输入空间要求鲁棒性验证不仅要确保扰动域内数据样本的安全性, 同时需保证数据分布满足特定领域的复杂规则. Liu 等人[22]基于不同特征的鲁棒性范畴的差异性, 提出了一种能求出给定的神经网络在一数据点上的椭球形非均匀扰动鲁棒区间的算法, 该算法能求出比传统算法求出的球形鲁棒区间更大的非均匀扰动鲁棒区间. Erdemir 等人[24]提出了一种在非均匀扰动鲁棒区间中的对抗训练方法, 对比起传统的对抗训练方法更好地增强了在非图像领域的网络的鲁棒性. 上述提到的研究大都集中在非均匀扰动假设对于神经网络鲁棒性的平均性能增强上, 并未考虑神经网络的形式验证算法在这一新假设下的完善. 一方面, 现有的神经网络验证技术无法形式地建模上述任务中可能面临的扰动或攻击; 另一方面, 基于各向同性假设开发的神经网络验证工具缺乏对更普适、更精确的潜在扰动或者攻击进行验证的功能支持, 是神经网络形式验证技术更广泛地应用于

安全攸关领域的阻碍.

如图 1 所示, 本文提出了一个面向领域自适应扰动的神经网络鲁棒性形式验证 (neural network verification for robustness against domain-adaptive perturbations, NNV4RADAP) 算法, 根据数据的领域特性对非均匀的扰动或攻击进行形式建模, 将现有的神经网络鲁棒性形式验证算法推广到定义在领域自适应扰动上的鲁棒性上. 我们根据鲁棒性定义的类别, 通过数据集的数据分布对数据的相关性矩阵进行计算, 并基于相关性矩阵构建面向领域自适应扰动的鲁棒性验证问题. 然后, 基于相关性矩阵进行等价验证问题构建, 将面向领域自适应扰动的鲁棒性验证问题归约到一个传统的面向均匀扰动的鲁棒性验证问题上, 并进行形式验证. NNV4RADAP 的关键优势在于它可以根据输入数据的统计属性与分布模式调整扰动空间的几何结构, 从而更加精确地为神经网络提供鲁棒性保证. 这一过程突破了以往基于均匀范数球邻域的传统验证框架的局限性, 使我们能够在更为复杂的边界条件下评估神经网络的鲁棒性.

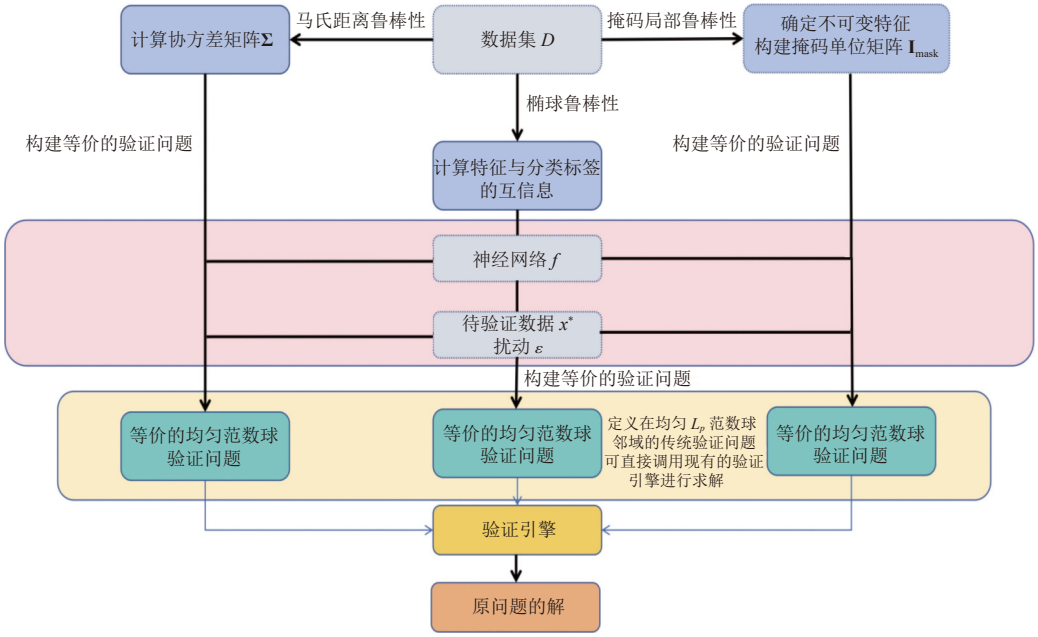

图 1　面向领域自适应扰动的神经网络鲁棒性验证算法框架

本文第 2 节对用到的预备知识进行概述, 包括神经网络的符号定义、神经网络鲁棒性的形式定义及其上的形式验证算法等. 第 3 节给出面向领域自适应扰动的神经网络鲁棒性问题, 并给出 NNV4RADAP 算法以及其正确性的理论证明. 第 4 节为案例分析, 介绍实验数据和细节, 并通过实验结果论证了 NNV4RADAP 算法的有效性以及可能存在的性能损失. 第 5 节对全文内容进行总结和概括.

2　预备知识

2.1　基础记号

我们用 $[L]$ 表示不超过 L 的正整数集合. 用正粗体符号表示矩阵, 如 $\mathbf{W}$; 用 $\mathbf{W}_{i:}$ 表示矩阵的第 i 行, 用 $\mathbf{W}_{:j}$ 表示矩阵的第 j 列, 用 $\mathbf{W}_{i,j}$ 表示矩阵的第 i 行第 j 列元素; 用 $\mathbf{I}$ 表示单位矩阵. 我们用斜粗体符号表示向量, 如 $\boldsymbol{x}$. 用下标的方式表示向量中的某个元素, 如 $\boldsymbol{x}_j$ 表示向量 $\boldsymbol{x}$ 中的第 j 个元素. 用普通符号表示标量, 如 c. 此外, 我们通过 $\|\cdot\|_p$ 来表示向量的 L_p 范数: $\|\boldsymbol{x}\|_p = \sqrt[p]{\sum_{i=1}^{n}|x_i|^p}$, 其中 $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ 为 n 维向量.

2.2　神经网络结构

一个 L 层的前馈神经网络, 它的输入层可以视为第 0 层, 最后一层被称为输出层, 第 1 至 $L-1$ 层被称为隐藏

层. 对于一个 L 层的前馈神经网络, 第 i 层 ($i\in[L]$) 有 d_i 个神经元. 我们定义神经网络的输入 $\boldsymbol{x}\in\mathbb{R}^{d_0}$, 并分别定义第 i 层的权重矩阵以及偏置向量为 $\mathbf{W}^{(i)}\in\mathbb{R}^{d_i\times d_{i-1}}$ 与 $\mathbf{b}^{(i)}\in\mathbb{R}^{d_i}(i\in[L])$. 神经网络函数 $f:\mathbb{R}^{d_0}\to\mathbb{R}^{d_L}$ 定义为 $f(\boldsymbol{x})=\boldsymbol{z}^{(L)}(\boldsymbol{x})$, 其中 $\boldsymbol{z}^{(i)}(\boldsymbol{x})=\mathbf{W}^{(i)}\hat{\boldsymbol{z}}^{(i-1)}(\boldsymbol{x})+\boldsymbol{b}^{(i)}$, $\hat{\boldsymbol{z}}^{(i)}(\boldsymbol{x})=\sigma(\boldsymbol{z}^{(i)}(\boldsymbol{x}))$, 并且 $\hat{\boldsymbol{z}}^{(0)}=\boldsymbol{x}$. σ 为激活函数, 本文中使用 ReLU 函数作为激活函数, $\boldsymbol{z}_j^{(i)}$ 与 $\hat{\boldsymbol{z}}_j^{(i)}$ 分别表示第 i 层第 j 个神经元的激活前值与激活后值. 如图 2 所示是一个简单的 3 层前馈神经网络, 网络的输出为 $f(\boldsymbol{x})=\boldsymbol{z}^{(3)}(\boldsymbol{x})$.

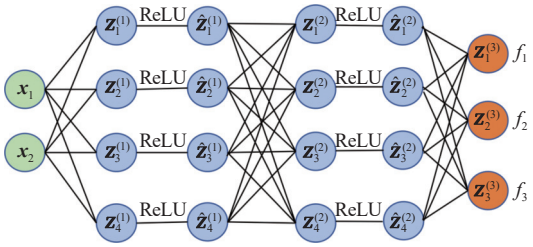

图 2 前馈神经网络的结构

2.3 神经网络验证问题的形式化定义及求解

神经网络的验证问题指的是通过形式化验证的方法论对神经网络的性质进行验证, 用数学方法去证明神经网络的某些性质, 其不存在某个缺陷或者符合某个或某些属性. 在软件 1.0 的形式化验证中, 首要的一步是用数学语言描述清楚需要证明的性质以及要解决的问题. 通过对问题建立数学模型, 根据性质限定系统在不同时刻应该以及不应该有的状态, 然后利用这些数学规则通过逻辑对这些性质进行证明. 因此, 基于第 2.2 节中的神经网络的函数定义, 同时将性质描述形式化为一个包含输入和输出的逻辑公式, 神经网络上的验证也就可以被形式化地定义, 具体如下.

定义 1. 神经网络验证. 给定两个谓词 $\mathbf{P}$ 和 $\mathbf{Q}$, 分别对应了输入约束以及待验证的性质. f 关于 $\mathbf{P}$ 和 $\mathbf{Q}$ 的验证问题就是要证明 (或证否):

$$\forall\boldsymbol{x},\mathbf{P}(\boldsymbol{x})\Rightarrow\mathbf{Q}(f(\boldsymbol{x})).$$

等价地, 我们需要去证明:

$$\mathbf{P}(\boldsymbol{x})\wedge\neg\mathbf{Q}(f(\boldsymbol{x}))$$

是不可满足的 (UNSAT). 若是在可满足 (SAT) 的情况下, 我们会找到至少一个对抗样本 $\boldsymbol{x}$ 使得 $\mathbf{P}(\boldsymbol{x})$ 为真而 $\mathbf{Q}(f(\boldsymbol{x}))$ 为假.

神经网络上的验证问题常被研究者们建模成优化问题进行求解, 这类方法通常具有良好的完备性. Katz 等人[10]通过对单纯形法的扩展来支持 ReLU 约束, 实现了 Reluplex 工具, 可以成功验证 300 个神经元的网络. 基于 Reluplex 的工作, Katz 等人[12]于 2019 年提出了 Marabou 框架, 将验证从 ReLU 激活函数扩展到了全连接和卷积神经网络上. 研究者们还通过将问题建模成混合整数规划问题来解决 SAT/LP 编码中的分支爆炸问题, 进而提升神经网络方法的可拓展性. 这方面的工作由 Lomuscio 等人[25]首次提出, 将可验证的网络规模提升到了 500 个神经元; 最具代表性的工作则是 Dutta 等人[26]提出的 Sherlock 算法, 将可验证的网络规模提升到了 6 000 个神经元. 还有工作通过网络抽象-反例制导精化的框架对现有的验证方法进行加速. Elhboer 等人[14]首次基于网络抽象技术对模型的结构进行了简化, 将每层隐藏层的神经元数量缩减到最多 4 个, 每个节点都有特定的类别及权重. 通过这样的抽象方法, 输出的上下界也就都可以通过分析抽象后的模型得到, 也就能对应用的验证方法进行加速 (如 Marabou[12]), 而其完备性也能通过反例制导精化方法进行保证.

此外, 研究者们还将验证问题建模成可达性问题进行求解, 这类方法的可拓展性较上一种往往更强, 但会以损失完备性为代价, 其本质上是对输出可达集的估计[27]. 一类方法跟抽象解释相结合, 也被称为区间传播技术. 主流的抽象解释技术[28−30]一般将神经网络的每个节点符号化并将抽象函数应用在计算过程上, 然后在具象化之后判断性质是否被满足了. Wang 等人[11]提出的 ReluVal 以及其改良后的 Neurify[31]是区间传播方法的代表工具. 另一类

具有较高可拓展性的方法是将区间算术与线性近似相结合, 其中最具代表性也是当下最先进的验证工具是 CROWN 系列的工作[13,32−35], 其中 CROWN[32]是这一系列工作的开山鼻祖, 将一个线性不等式从后往前在网络当中传播并运用线性函数对非线性激活函数进行松弛. Salman 等人[33]则证明了对 ReLU 松弛的优化使得 CROWN 可以获得跟基于线性规划的验证器同样的解. 此后, Xu 等人[34]将 CROWN 普及到了一般的通用计算图上并在[35]中通过同时优化中间层和输出层的线性界得到了比一般 LP 验证器更紧的线性界, 从而得到更优的性能.然而都无法保证验证的完备性[32−35], 因此 Wang 等人[13]通过分支界限法 (branch and bound) 将由于 ReLU 激活函数而分裂的约束统一到 CROWN 的区间边界传播算法中, 并通过 GPU 对可并行化的区间传播进行加速, 从而首次构建了一个同时具有高可拓展性和完备性的验证器.

2.4 神经网络鲁棒性的形式验证

通常而言, 神经网络背景下的鲁棒性指的是深度神经网络在面对受到干扰的输入数据时仍能保持可靠预测的能力. 最早关注神经网络鲁棒性的研究可以追溯到 Szegedy 等人[36]的发现——即便是准确率极高的神经网络也很容易被施加的某种微不可察的扰动所愚弄. 当时的研究主要集中在基于对抗样本搜索的对抗训练, 而随着时间的推移, 研究者们意识到对抗训练只能不断提高神经网络的平均鲁棒性能, 而无法为神经网络提供形式化的保证. 因此, 研究者们逐渐尝试将神经网络验证与神经网络的鲁棒性研究结合起来并提出了相应的形式化定义. 在已有的研究中, 深度神经网络上的鲁棒性往往是在局部上定义的, 一般都遵循 Huang 等人[37]对深度神经网络局部鲁棒性的定义.

定义 2. 神经网络的局部鲁棒性. 一个深度神经网络对于一个输入空间 η 是局部鲁棒的, 当且仅当该网络对 η 中所有的输入的判定结果是相同的.

在已有的研究中, 大部分研究者们都用数据点 $\boldsymbol{x}^*$ 的一个 L_p 范数球邻域来对定义 2 中的 η 进行描述.

定义 3. 神经网络在 L_p 范数球下的 ε-局部鲁棒性. 给定一个神经网络 $f:\mathbb{R}^{d_0}\to\mathbb{R}^{d_L}$, 其中 d_0 与 d_L 分别是输入向量以及输出向量的维度. 在分类问题中, 神经网络会将输入 $\boldsymbol{x}$ 分类至标签 $C_f(\boldsymbol{x})=\arg\max_{i\in[d_L]}f(\boldsymbol{x})_i$. 如果对于一个数据 $\boldsymbol{x}^*\in\mathbb{R}^{d_0}$, 以及一个正实数 ε, 在输入空间 $\eta=\{\boldsymbol{x}\,|\,\|\boldsymbol{x}-\boldsymbol{x}^*\|_p\leqslant\varepsilon\}$ 中的所有 $\boldsymbol{x}$ 在输出层中正确标签对应神经元的值恒大于其他标签的神经元值, 即满足性质 $S=\{f(\boldsymbol{x})\,|\,\arg\max_{i\in[d_L]}f(\boldsymbol{x})_i=\arg\max_{i\in[d_L]}f(\boldsymbol{x}^*)_i\}$, 则称 f 对于 $\boldsymbol{x}^*$ 是 ε-局部鲁棒的.

基于这一定义, 我们给出神经网络在 L_p 范数球下的鲁棒性验证问题.

定理 1. 神经网络在 L_p 范数球下的鲁棒性验证问题. 给定一个神经网络 $f:\mathbb{R}^{d_0}\to\mathbb{R}^{d_L}$, 输入空间 $\eta=\{\boldsymbol{x}\,|\,\|\boldsymbol{x}-\boldsymbol{x}^*\|_p\leqslant\varepsilon\}$. f 对于一个输入样本点 $\boldsymbol{x}^*\in\mathbb{R}^{d_0}$ 是 ε-局部鲁棒的, 当且仅当

$$\forall\boldsymbol{x}\in\mathbb{R}^{d_0},\ \|\boldsymbol{x}-\boldsymbol{x}^*\|_p\leqslant\varepsilon\Rightarrow(\arg\max_{i\in[d_L]}f(\boldsymbol{x})_i=\arg\max_{i\in[d_L]}f(\boldsymbol{x}^*)_i)\tag{1}$$

等价于验证 $\|\boldsymbol{x}-\boldsymbol{x}^*\|_p\leqslant\varepsilon\wedge(\arg\max_{i\in[d_L]}f(\boldsymbol{x})_i\neq\arg\max_{i\in[d_L]}f(\boldsymbol{x}^*)_i)$ 是不可满足的. 我们将这个验证问题记为 $(f,\boldsymbol{x}^*,\varepsilon)$.

3 面向领域自适应扰动的神经网络鲁棒性验证

本节首先介绍现有的神经网络鲁棒性的定义在非图像领域中的局限性, 然后对面向领域自适应扰动的神经网络鲁棒性及其对应的验证问题进行形式化定义, 并提出面向领域自适应扰动的神经网络鲁棒性验证算法 NNV4RADAP.

3.1 神经网络鲁棒性的局限性

在机器学习的安全验证领域, 均匀 L_p 范数球是一种行之有效的数据点的邻域约束, 它通过定义多维空间中具有径向对称性的超球体 (hypersphere) 作为有效邻域边界, 其优势在于无需引入领域先验知识即可建立统一的鲁棒性评估标准. 该理论框架隐含两个假设条件: (1) 特征等距假设, 即不同特征方向上允许的最大扰动幅度相同: 以 L_∞ 范数球为例, 各维度上的扰动上限被严格限定为 ε; 而对于 L_2 范数球, 则通过欧氏距离约束确保扰动向量的总能量不超过 ε^2. (2) 特征独立性假设, 即忽略特征间的语义关联性, 如自然图像中相邻像素间的空间相关性常被忽略. 在计算机视觉领域, 对抗攻击 (如 FGSM、PGD) 通常基于 L_p 范数并通过模型的梯度信息生成对抗性样本, 采

取均匀范数球作为邻域约束可确保验证覆盖这些实际攻击模式. 然而现实世界中的数据普遍表现出显著的各向异性特征, 意味着均匀的 L_p 范数球约束在许多应用领域通常会导致不合理的变换.

相较于均匀邻域约束, 我们认为非均匀邻域约束对现实任务更具普适性. 以电机故障诊断任务为例, 对高频谐波幅值, 小波变换系数等关键特征上施加微小扰动即可导致分类结果改变. 而对于次要特征, 如均值、低频能量, 即使承受较大的扰动仍能保持模型预测的稳定性. 传统上, 特征重要性评估方法 (如 Shapley 值、互信息) 倾向于强调高重要性特征对决策边界的贡献. 基于此, 我们认为若模型依靠关键特征进行决策, 那么对于这类关键特征, 我们必须设定更为严格的扰动限制; 而对于非关键特征, 则可以适当放宽限制, 保证特征的扰动预算与其重要性呈反比关系. 进一步地, 在特定领域中, 部分特征还具有物理不可变性约束. 以恶意软件检测任务为例, 攻击者的攻击方式主要可分为两类: (1) 特征空间攻击: 通过修改从二进制文件中提取的特征来生成对抗样本. (2) 问题空间攻击: 在保持软件功能完整性的前提下, 直接修改恶意软件二进制文件, 以确保修改后的对象既有效又隐蔽. 对于程序而言, 某些关键特征一旦被篡改, 将导致程序功能无法正常运行. 例如, 在 Android 恶意软件中, 权限特征 (如访问手机位置的权限) 是恶意行为的必要组件, 若这些权限被修改, 恶意软件的功能将无法正常运行. 因此, 在实际攻击中, 攻击者通常会主动避开这类特征. 基于此, 我们的方法考虑仅在可合法更改的特征空间中定义扰动域, 从而排除非法扰动的可能性. 最后, 在垃圾邮件检测任务中, 我们面临的是特征间的强关联约束. 具体而言, 在分析词频特征与符号特征时, 我们发现某些关键词和符号倾向于同时出现, 例如“click”与“here”, 以及“money”和“$”. 另一方面, 在文本结构特征上, 大写字母连续序列的平均长度、最长长度及总长度之间可能存在着较强的线性关系, 因为它们都基于同一文本属性 (大写字母在文本中的分布). 因此, 我们引入基于协方差矩阵 $\boldsymbol{\Sigma}$ 的几何约束, 确保扰动向量的方向与原始数据分布的线性相关性结构保持一致, 从而减少无意义扰动组合的发生.

为确保扰动域内输入数据 (尤其是潜在的对抗样本) 与真实数据在分布上保持高度一致, 我们有必要对数据的一致性和有效性进行明确定义. 设神经网络模型 $f:\mathbb{R}^{d_0}\to\mathbb{R}^{d_L}$ 的输入向量 $\boldsymbol{X}=(X_1,X_2,\ldots,X_{d_0})$, 其中各特征 $X_1,X_2,\ldots,X_{d_0}$ 可能存在依赖关系而非独立同分布, 这更符合实际的数据特性. 我们假设输入向量 $\boldsymbol{X}$ 服从多元联合分布 $\mathcal{F}$, 即 $\boldsymbol{X}\sim\mathcal{F}$.

定义 4. 数据的一致性. 对模型的输入 $\boldsymbol{x}^*$ 施加扰动后, 扰动域内的数据可能在给定分布下 $\mathcal{F}$ 表现为低概率密度值, 即偏离原始数据分布的区域. 设一致性阈值 $\gamma>0$, 我们称一个经扰动的数据 $\boldsymbol{x}$ 符合一致性要求当且仅当概率密度 $p_{\boldsymbol{X}\sim\mathcal{F}}(\boldsymbol{X}=\boldsymbol{x})\geqslant\gamma$.

定义 5. 数据的有效性. 尽管可以在数字层面上对不可变特征进行扰动, 但在实际应用中, 这一修改往往会导致数据的无效性. 记不可变的特征个数为 k, 其索引构成的集合为 $M=\{m_1,m_2,\ldots,m_k\}\in[d_0]$, 其中 $m_1<m_2<\ldots<m_k$. 对 $\boldsymbol{x}^*$ 施加扰动得到的数据 $\boldsymbol{x}$ 符合有效性要求, 当且仅当 $\forall i\in M,\ \boldsymbol{x}_i=\boldsymbol{x}_i^*$.

图 3(a) 揭示了基于均匀 L_p 范数球的鲁棒性验证方法中潜在的假阴性问题. 对数据施加扰动以测试模型稳定性时, 对关键特征施加过大的扰动会导致样本失真从而直接改变分类结果; 相比之下, 次要特征则能在较大的扰动范围内保持其对分类结果的影响不变, 这说明两者之间存在不同的扰动预算. 使用均匀范数球作为验证区域时, 通常依据关键特征来确定允许的最大扰动范围. 在图 3(a) 中, $\boldsymbol{x}_2$ 作为关键特征, $\boldsymbol{x}_1$ 则作为次要特征, 即使我们在鲁棒半径为 0.05 的均匀范数球内成功验证了神经网络的鲁棒性, 但这一结论可能仍具有误导性, 因为对于 $\boldsymbol{x}_1$ 而言, 0.05 的扰动预算可能过于保守, 适当放宽该维度的扰动限制值可能会揭露对抗样本的存在. 这表明了预定义的 L_p 范数球未必能准确地反映真实的鲁棒区间, 进而导致假阴性的发生: 基于敏感特征选择的扰动预算对于非敏感特征而言可能过于严格, 从而导致验证结果过于乐观. 假阴性问题可能造成神经网络完全鲁棒的假象, 这在一些安全攸关的领域中十分致命. 另一方面, 如果我们能够验证椭球区域的鲁棒性, 由于均匀范数球是被椭球所包围的, 那么在均匀范数球内网络鲁棒性也自然得到了保证. 通过将特征鲁棒性差异纳入考虑, 在椭球区域进行的神经网络鲁棒性验证, 能够为模型提供更严格的安全性保证.

图 3(b) 和图 3(c) 揭示了基于均匀 L_p 范数球的鲁棒性验证方法可能引发的假阳性问题. 当我们在验证区域中检测到对抗样本时, 这些数据可能偏离了原始数据的分布. 在图 3(b) 中, $\boldsymbol{x}_3$ 作为不可变特征, 我们在均匀范数球内

检测的对抗样本却在这一维度上发生了变化, 从而违反了领域约束条件, 导致其不符合数据的有效性假设. 图 3(c) 描绘了特征 $\boldsymbol{x}_1$ 与 $\boldsymbol{x}_2$ 之间存在正相关性的场景, 当扰动预算设置过大时, 均匀范数球内生成的对抗样本可能会破坏这一内在的相关性结构. 综上所述, 传统验证模式下发现的对抗样本可能无法满足数据的一致性或有效性要求, 导致其在实际应用场景中并不具有现实意义. 假阳性现象会引导研究者们对网络进行无意义的鲁棒性增强, 并因此耗费相当多的时间精力.

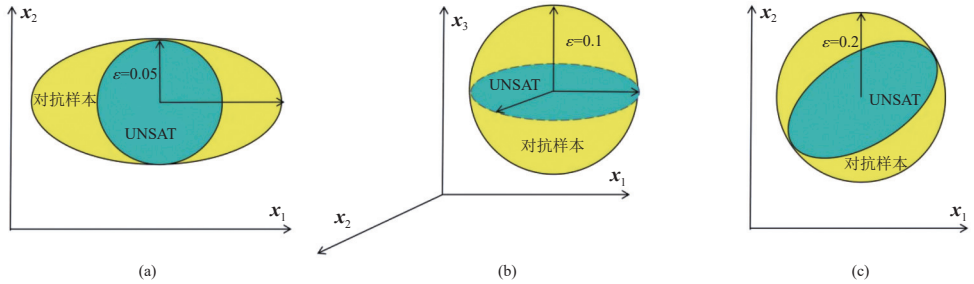

图 3 L_p 范数球鲁棒性验证中的假阳性与假阴性问题

考虑到以上问题, 在设计鲁棒性验证方法时, 我们需要充分考虑数据集中不同特征在鲁棒性范畴上的差异性及其内在关联约束, 从而构建一种面向领域自适应的扰动鲁棒性验证框架. 严格来说, 面向领域自适应扰动的鲁棒性并非一个形式化的统一定义, 而是一个包含多种性质的定义集合, 能够根据不同应用场景的需求灵活描述不同的鲁棒性特性. 基于这一理念, 我们将从这些性质中选取典型类别进行鲁棒性问题的形式化构建. 数据不同维度的特征对模型预测结果的敏感度存在差异, 基于此, 本文假设目标表示空间中各维度特征具有不同的鲁棒性范畴, 在此基础上讨论不同特征尺度下的鲁棒性, 定义并研究椭球鲁棒性. 考虑到数据某些特征具有不可变性, 我们在目标表示空间中引入特征的不可变约束, 定义并研究掩码局部鲁棒性, 通过引入掩码机制, 限制对不可变特征的扰动, 确保所发现的对抗样本符合现实场景的语义约束, 从而提升验证结果的可信度. 考虑到数据不同特征之间的依赖关系, 我们引入特征间存在线性相关性的假设, 定义并研究基于马氏距离 (Mahalanobis distance) 的鲁棒性. 通过定义以马氏距离为度量的扰动域, 从而确保扰动后的样本仍处于原始数据分布的高密度区域, 以避免识别出那些在数值上满足扰动约束、但在现实中几乎不可能出现的“假阳性”对抗样本.

基于上述 3 种不同的鲁棒性概念, 我们将均匀 L_p 范数球邻域扩展为能够适应不同领域特性的非均匀邻域. 在椭球鲁棒性和马氏距离鲁棒性中, 通过构造相应的线性变换矩阵 $\boldsymbol{\Omega}$ 对均匀 L_p 范数球邻域进行线性变换, 从而刻画不同维度上的特征在扰动预算上的差异以及特征间的相关性关系; 在掩码局部鲁棒性中, 我们通过引入掩码矩阵 $\mathbf{I}_{\text{mask}}$ 确保扰动仅作用于可变特征. 上述方法允许我们在考虑特征重要性、不可变性以及相关性的基础上, 灵活地定义每个特征的最大扰动范围.

3.2 椭球鲁棒性分支

3.2.1 椭球鲁棒性的定义

在椭球鲁棒性分支上, 我们需要将特征的鲁棒性差异纳入考量, 关注数据在一个超椭球体 (hyperellipsoid) 非均匀尺度空间内的鲁棒性. 不同于球体在所有方向上半径相同的特点, 椭球体的不同轴长直接反映了对应维度上特征的扰动承受能力. 尽管研究者们期望对分类器结果影响更大的数据特征能容忍更大的扰动, 但直觉上对分类器结果影响较小的数据特征天然地拥有更高的鲁棒性. 换言之, 非关键特征能够在保持模型预测准确性的同时, 承受更多的数据变动或噪声干扰. 这是因为这些特征对于模型最终预测结果的影响微乎其微, 因此一定程度的扰动不会显著改变预测结果. 相反, 那些对分类器性能至关重要的特征, 即使是较小的扰动也可能导致预测结果的显著变化, 从而影响模型的稳定性和准确性.

我们选取了互信息 (mutual information, MI) 来对特征的重要性进行刻画. 对于特征 X 与输出标签 Y, 它们的互信息为:

$$I(X;Y)=\sum_{(x,y)\in(X\times Y)}p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{2}$$

其中, $p(x,y)$ 是 X 与 Y 的联合概率分布, $p(x)$ 与 $p(y)$ 分别是 X 与 Y 的边缘概率分布. 互信息值量化了特征与分类标签之间共享的信息量, 具体表现为该特征可以减少分类结果多少不确定性. 在多分类任务中, 互信息是评估特征对分类结果贡献度的重要指标, 互信息越高, 特征与输出的相关性越强, 对分类的贡献越大. 基于这一特性, 我们用 $mif_{i,y}$ 表示第 i 个特征与输出 y 之间的互信息值, 通过计算和比较不同特征的互信息值对其重要性进行排序.

不同特征的互信息值可能存在一定的数量级差异, 此时若以互信息最高的特征为基准分配扰动预算 ε, 其他特征可能会被施加过大的扰动值. 因此, 我们对互信息值进行幂变换 (power transformation) 操作, 在保留特征重要性顺序的前提下, 确保每个特征被分配到的扰动预算更加合理和均衡. 假设最大的互信息值为 M, 最小的互信息值为 m, 我们希望 M 和 m 间的倍数不超过某个指定值 (例如 k 倍), 可以进行如下操作将互信息值归一化到 $[1,k]$ 区间, 并记归一化后的互信息为 $\overline{mif}_{i,y}$.

首先定义最大和最小互信息值的倍数关系为 $r=\dfrac{M}{m}$, 若该倍数超过了预设的阈值, 即 $r>k$, 则执行幂变换操作, 记 $\alpha=\dfrac{\ln(k)}{\ln(r)}$, $\overline{mif}_{i,y}=\left(\dfrac{mif_{i,y}}{m}\right)^{\alpha}$.

在上述假设下, 我们给出神经网络上 L_p 范数椭球鲁棒性定义.

定义 6. 神经网络在 L_p 范数上的 ε-椭球局部鲁棒性. 给定一个神经网络 $f:\mathbb{R}^{d_0}\to\mathbb{R}^{d_L}$, 线性变换矩阵为对角矩阵 $\mathbf{D}$, 其中对角线元素记为 $\mathbf{D}_{i,i}=\overline{mif}_{i,y}, i=1,2,\ldots,d_0$, 输入空间 $\eta=\{\boldsymbol{x}\mid\|\mathbf{D}(\boldsymbol{x}-\boldsymbol{x}^*)\|_p\leqslant\varepsilon\}$. f 对于一个输入样本点 $\boldsymbol{x}^*\in\mathbb{R}^{d_0}$ 是 ε-椭球局部鲁棒的, 当且仅当

$$\forall\boldsymbol{x}\in\mathbb{R}^{d_0},\ \|\mathbf{D}(\boldsymbol{x}-\boldsymbol{x}^*)\|_p\leqslant\varepsilon\Rightarrow(\arg\max_{i\in[d_L]}f(\boldsymbol{x})_i=\arg\max_{i\in[d_L]}f(\boldsymbol{x}^*)_i) \tag{3}$$

定义 6 中神经网络在 L_p 范数下的椭球鲁棒性定义实际上将扰动值 $\boldsymbol{\delta}=\boldsymbol{x}-\boldsymbol{x}^*$ 的第 i 个特征投影到了一个以 $\dfrac{\varepsilon}{\mathbf{D}_{i,i}}$ 为半径的 L_p 范数球上. 显然, 越大的 $\overline{mif}_{i,y}$ 对应越大的 $\mathbf{D}_{i,i}$, 也意味着第 i 个特征的扰动预算越小. 值得注意的是, 若不对特征的重要性进行区分, 而是默认所有的特征具有相同的鲁棒性范畴, 则此时的线性变换矩阵 $\mathbf{D}$ 将退化为一个标量矩阵. 这意味着所有特征被赋予了相同的扰动范围, 相应地, 椭球鲁棒性也随之退化为均匀 L_p 范数球鲁棒性.

3.2.2 椭球鲁棒性验证问题的等价构建

基于均匀 L_p 范数球鲁棒性的神经网络验证工具难以直接建模上述提到的椭球鲁棒性定义. 因此, 我们尝试将椭球鲁棒性验证问题通过多项式时间算法转换到现有验证工具能解决的验证问题上. 具体而言, 对于一个神经网络 f, 我们尝试将其关于 $\boldsymbol{x}^*$ 的椭球鲁棒性验证问题转换成一个与之等价的神经网络 $\tilde{f}$ 关于 $\tilde{\boldsymbol{x}}^*$ 的均匀 L_p 范数球鲁棒性验证问题, 并提出以下定理.

定理 2. 神经网络的椭球鲁棒性验证问题的等价构建. 给定一个神经网络 $f:\mathbb{R}^{d_0}\to\mathbb{R}^{d_L}$, 正实数 $\varepsilon>0$, 数据点 $\boldsymbol{x}^*=(\boldsymbol{x}_1^*,\boldsymbol{x}_2^*,\ldots,\boldsymbol{x}_{d_0}^*)$, 以及对角矩阵 $\mathbf{D}$, 其中 $\mathbf{D}_{i,i}=\overline{mif}_{i,y}, i=1,2,\ldots,d_0$. 我们构建一个相同结构的神经网络 $\tilde{f}:\mathbb{R}^{d_0}\to\mathbb{R}^{d_L}$, 以及一个新的数据点 $\tilde{\boldsymbol{x}}^*=\mathbf{D}\boldsymbol{x}^*$, 其中 $\tilde{f}$ 中的参数满足:

$$\tilde{\mathbf{W}}^{(i)}=\mathbf{W}^{(i)},\ i=2,3,\ldots,L \tag{4}$$

$$\tilde{\mathbf{W}}^{(1)}=\mathbf{W}^{(1)}\mathbf{D}^{-1} \tag{5}$$

$$\tilde{\boldsymbol{b}}^{(i)}=\boldsymbol{b}^{(i)},\ i=1,2,\ldots,L \tag{6}$$

那么 f 在 $\eta=\{\boldsymbol{x}\mid\|\mathbf{D}(\boldsymbol{x}-x^*)\|_p\leqslant\varepsilon\}$ 上的椭球鲁棒性验证问题等价于 $\tilde{f}$ 在 $\tilde{\eta}=\{\boldsymbol{x}\mid\|\boldsymbol{x}-\tilde{\boldsymbol{x}}^*\|_p\leqslant\varepsilon\}$ 上的均匀 L_p 范数球鲁棒性验证问题.

证明: 对于 $\forall\boldsymbol{x}\in\eta$, 我们可以构造 $\tilde{\boldsymbol{x}}=\mathbf{D}\boldsymbol{x}$, 因为 $\|\mathbf{D}(\boldsymbol{x}-\boldsymbol{x}^*)\|_p\leqslant\varepsilon$, 且有 $\tilde{\boldsymbol{x}}^*=\mathbf{D}\boldsymbol{x}^*$, 故 $\|\tilde{\boldsymbol{x}}-\tilde{\boldsymbol{x}}^*\|_p\leqslant\varepsilon$, $\tilde{\boldsymbol{x}}\in\tilde{\eta}$. 当 f 以 $\boldsymbol{x}$ 作为输入, 其第 1 层神经元的激活前值 $\boldsymbol{z}^{(1)}=\mathbf{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)}$; 当 $\tilde{f}$ 以 $\tilde{\boldsymbol{x}}$ 作为输入, 其第 1 层神经元的激活前值 $\tilde{\boldsymbol{z}}^{(1)}=\tilde{\mathbf{W}}^{(1)}\tilde{\boldsymbol{x}}+\tilde{\boldsymbol{b}}^{(1)}=\mathbf{W}^{(1)}\mathbf{D}^{-1}\mathbf{D}\boldsymbol{x}+\boldsymbol{b}^{(1)}=\mathbf{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)}$, 因而有 $\boldsymbol{z}^{(1)}=\tilde{\boldsymbol{z}}^{(1)}$. 因为 f 和 $\tilde{f}$ 除了第 1 层的权重矩阵不同, 其余结构与

参数均相同, 所以有 $f(\boldsymbol{x})=\tilde{f}(\tilde{\boldsymbol{x}})$. 特别地, $f(\boldsymbol{x}^*)=\tilde{f}(\tilde{\boldsymbol{x}}^*)$.

若 $\tilde{f}$ 在 $\tilde{\eta}$ 上是鲁棒的, 则 f 在 η 上是鲁棒的. 反证法证明. 若 f 在 η 上存在对抗样本, 我们记该对抗样本为 $\boldsymbol{x}_{\text{adv}}$. 由前面的结论, 我们可以构造 $\tilde{\boldsymbol{x}}_{\text{adv}}=\mathbf{D}\boldsymbol{x}_{\text{adv}}\in\tilde{\eta}$, 且有 $f(\boldsymbol{x}_{\text{adv}})=\tilde{f}(\tilde{\boldsymbol{x}}_{\text{adv}})$. 由于 $\tilde{f}$ 在 $\tilde{\eta}$ 上是鲁棒的, 根据鲁棒性定义:

$$\arg\max_{i\in[d_L]}\tilde{f}(\tilde{\boldsymbol{x}}_{\text{adv}})_i=\arg\max_{i\in[d_L]}\tilde{f}(\tilde{\boldsymbol{x}}^*)_i \tag{7}$$

由 $f(\boldsymbol{x}_{\text{adv}})=\tilde{f}(\tilde{\boldsymbol{x}}_{\text{adv}})$ 以及 $f(\boldsymbol{x}^*)=\tilde{f}(\tilde{\boldsymbol{x}}^*)$, 我们有:

$$\arg\max_{i\in[d_L]}f(\boldsymbol{x}_{\text{adv}})_i=\arg\max_{i\in[d_L]}f(\boldsymbol{x}^*)_i \tag{8}$$

因为 $\boldsymbol{x}_{\text{adv}}$ 不会引起分类结果的改变, 所以 $\boldsymbol{x}_{\text{adv}}$ 不是对抗样本, 与假设矛盾. 所以 f 在 η 上不存在对抗样本, 即 f 在 η 上是鲁棒的.

若 $\tilde{f}$ 在 $\tilde{\eta}$ 上存在对抗样本 $\tilde{\boldsymbol{x}}_{\text{adv}}$, 则 f 在 η 上存在对抗样本 $\boldsymbol{x}_{\text{adv}}=\mathbf{D}^{-1}\tilde{\boldsymbol{x}}_{\text{adv}}$. 因为 $\|\tilde{\boldsymbol{x}}_{\text{adv}}-\tilde{\boldsymbol{x}}^*\|_p\leqslant\varepsilon$, 即 $\|\mathbf{D}(\boldsymbol{x}_{\text{adv}}-\boldsymbol{x}^*)\|_p\leqslant\varepsilon$, 故 $\boldsymbol{x}_{\text{adv}}\in\eta$, 又由鲁棒性定义, 对抗样本满足:

$$\arg\max_{i\in[d_L]}\tilde{f}(\tilde{\boldsymbol{x}}_{\text{adv}})_i\neq\arg\max_{i\in[d_L]}\tilde{f}(\tilde{\boldsymbol{x}}^*)_i \tag{9}$$

而 $f(\boldsymbol{x}_{\text{adv}})=\tilde{f}(\tilde{\boldsymbol{x}}_{\text{adv}})$ 且 $f(\boldsymbol{x}^*)=\tilde{f}(\tilde{\boldsymbol{x}}^*)$, 故 $\arg\max_{i\in[d_L]}f(\boldsymbol{x}_{\text{adv}})_i\neq\arg\max_{i\in[d_L]}f(\boldsymbol{x}^*)_i$, 所以 $\boldsymbol{x}_{\text{adv}}$ 是 η 上的对抗样本.

综上所述, 定理 2 证毕.

3.3 掩码局部鲁棒性分支

3.3.1 掩码局部鲁棒性的定义

在掩码局部鲁棒性分支中, 我们对数据分布的核心假设是: 样本具有不可变特征, 即攻击者只能在允许变动的特征子空间中对模型发起攻击. 通过专家知识, 我们可以明确哪些特征是不可变的, 哪些特征可能成为攻击的目标, 这种先验知识可以通过线性变换矩阵的形式嵌入到鲁棒性验证框架中. 具体而言, 我们引入一个掩码单位矩阵 $\mathbf{I}_{\text{mask}}$. $\mathbf{I}_{\text{mask}}$ 仅在允许变动的特征对应的对角线元素上的值为 1, 而不可变特征对应的对角线元素值为 0. 通过这种设计, 我们定义的非均匀输入空间 $\eta=\{\boldsymbol{x}\mid\|\mathbf{I}_{\text{mask}}(\boldsymbol{x}-\boldsymbol{x}^*)\|_p\leqslant\varepsilon,\ (\mathbf{I}-\mathbf{I}_{\text{mask}})x=(\mathbf{I}-\mathbf{I}_{\text{mask}})\boldsymbol{x}^*\}$ 能够严格限制数据在不可变特征上不会受到扰动. 根据上述假设, 给出神经网络上 L_p 范数掩码局部鲁棒性的定义.

定义 7. 神经网络在 L_p 范数下的 ε-掩码局部鲁棒性. 给定一个神经网络 $f:\mathbb{R}^{d_0}\to\mathbb{R}^{d_L}$, 输入数据中不可变的特征个数为 k, 其索引构成的集合记为 $M=\{m_1,m_2,\ldots,m_k\}\in[d_0]$, 其中 $m_1<m_2<\ldots<m_k$. 我们定义掩码单位矩阵 $\mathbf{I}_{\text{mask}}=diag(\{s_i\}_{i=1}^{d_0})$, 其中,

$$s_i=\begin{cases}0, & \text{if } i\in M\\ 1, & \text{otherwise}\end{cases} \tag{10}$$

输入空间 $\eta=\{\boldsymbol{x}\mid\|\mathbf{I}_{\text{mask}}(\boldsymbol{x}-\boldsymbol{x}^*)\|_p\leqslant\varepsilon,\ (\mathbf{I}-\mathbf{I}_{\text{mask}})\boldsymbol{x}=(\mathbf{I}-\mathbf{I}_{\text{mask}})\boldsymbol{x}^*\}$. f 对于输入样本点 $\boldsymbol{x}^*\in\mathbb{R}^{d_0}$ 是 ε-掩码局部鲁棒的, 当且仅当

$$\forall\boldsymbol{x}\in\mathbb{R}^{d_0},\ \|\mathbf{I}_{\text{mask}}(\boldsymbol{x}-\boldsymbol{x}^*)\|_p\leqslant\varepsilon\wedge(\mathbf{I}-\mathbf{I}_{\text{mask}})\boldsymbol{x}=(\mathbf{I}-\mathbf{I}_{\text{mask}})\boldsymbol{x}^*\Rightarrow(\arg\max_{i\in[d_L]}f(\boldsymbol{x})_i=\arg\max_{i\in[d_L]}f(\boldsymbol{x}^*)_i) \tag{11}$$

若不引入特征不可变性假设, 即认为所有特征均可被扰动, 则此时的掩码单位矩阵 $\mathbf{I}_{\text{mask}}$ 退化为单位矩阵, 相应地, 掩码局部鲁棒性也随之退化为均匀 L_p 范数球鲁棒性.

3.3.2 掩码局部鲁棒性验证问题的等价构建

类似地, 对于一个神经网络 f, 我们尝试将其关于 $\boldsymbol{x}^*$ 的掩码局部距离鲁棒性验证问题转换成一个与之等价的神经网络 $\tilde{f}$ 关于 $\tilde{\boldsymbol{x}}^*$ 的 L_p 范数球鲁棒性验证问题, 并提出以下定理.

定理 3. 神经网络的掩码局部鲁棒性验证问题的等价构建. 给定一个神经网络 $f:\mathbb{R}^{d_0}\to\mathbb{R}^{d_L}$, 正实数 $\varepsilon>0$, 数据点 $\boldsymbol{x}^*=(\boldsymbol{x}_1^*,\boldsymbol{x}_2^*,\ldots,\boldsymbol{x}_{d_0}^*)$, 其中不可变的特征个数为 k, 其索引构成的集合为 $M=\{m_1,m_2,\ldots,m_k\}\in[d_0]$, 其中 $m_1<m_2<\ldots<m_k$; 允许改变的特征个数为 d_0-k, 其索引构成的集合为 $N=\{n_1,n_2,\ldots,n_{d_0-k}\}\in[d_0]$, 其中 $n_1<n_2<\ldots<n_{d_0-k}$. 构建掩码对角矩阵 $\mathbf{I}_{\text{mask}}$, 其中 $\mathbf{I}_{\text{mask}}$ 仅在允许变动的特征对应的对角线元素上的值为 1, 不可变特征对应的对角线元素

值为 0. 同时我们构建 $\tilde{\boldsymbol{x}}^* = (\boldsymbol{x}^*_{n_1}, \boldsymbol{x}^*_{n_2}, \ldots, \boldsymbol{x}^*_{n_{d_0-k}})$ 对 $\boldsymbol{x}^*$ 进行压缩仅保留允许改变的特征, 并仅在这些特征上施加扰动. 我们定义压缩矩阵:

$$\mathbf{V} = \begin{pmatrix} \boldsymbol{e}^{\mathrm{T}}_{n_1} \\ \boldsymbol{e}^{\mathrm{T}}_{n_2} \\ \vdots \\ \boldsymbol{e}^{\mathrm{T}}_{n_{d_0-k}} \end{pmatrix} \in \mathbb{R}^{(d_0-k)\times d_0} \tag{12}$$

其中, $\boldsymbol{e}_i$ 为标准基向量, 仅在第 i 维度上为 1, 其余维度为 0. 压缩操作通过 $\tilde{\boldsymbol{x}}^* = \mathbf{V}\boldsymbol{x}^*$ 实现, 且易证得等式关系 $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_{\mathrm{mask}}$. 对于不可变的特征, 它们在网络中进行前向传播会形成常量增量, 体现在第 1 层是 $\sum_{i=1}^{k} \mathbf{W}^{(1)}_{j,m_i} \boldsymbol{x}^*_{m_i}$, $j = 1,2,\ldots,d_1$. 综上所述, 我们构建一个新的神经网络 $\tilde{f}: \mathbb{R}^{d_0-k} \to \mathbb{R}^{d_L}$, 其中网络 $\tilde{f}$ 中参数满足:

$$\tilde{\mathbf{W}}^{(i)} = \mathbf{W}^{(i)},\ i = 2,3,\ldots,L \tag{13}$$

$$\tilde{\mathbf{W}}^{(1)} = \mathbf{W}^{(1)}\mathbf{V}^{\mathrm{T}} \tag{14}$$

$$\tilde{\boldsymbol{b}}^{(i)} = \boldsymbol{b}^{(i)},\ i = 2,3,\ldots,L \tag{15}$$

$$\tilde{\boldsymbol{b}}^{(1)} = \boldsymbol{b}^{(1)} + \mathbf{W}^{(1)}(\mathbf{I} - \mathbf{I}_{\mathrm{mask}})\mathbf{x}^* \tag{16}$$

那么 f 在 $\eta = \{\boldsymbol{x} \mid \|\mathbf{I}_{\mathrm{mask}}(\boldsymbol{x} - \boldsymbol{x}^*)\|_p \leqslant \varepsilon,\ (\mathbf{I} - \mathbf{I}_{\mathrm{mask}})\boldsymbol{x} = (\mathbf{I} - \mathbf{I}_{\mathrm{mask}})\boldsymbol{x}^*\}$ 上的掩码局部鲁棒性验证问题等价于 $\tilde{f}$ 在 $\tilde{\eta} = \{\boldsymbol{x} \mid \|\boldsymbol{x} - \tilde{\boldsymbol{x}}^*\|_p \leqslant \varepsilon\}$ 上的均匀 L_p 范数球鲁棒性验证问题.

证明: 对于 $\forall \boldsymbol{x} \in \eta$, 我们构造 $\tilde{\boldsymbol{x}} = \mathbf{V}\boldsymbol{x}$, 由于 $\tilde{\boldsymbol{x}}^* = \mathbf{V}\boldsymbol{x}^*$, 所以有等式关系 $\|\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}^*\|_p = \|\mathbf{V}(\boldsymbol{x} - \boldsymbol{x}^*)\|_p$ 成立. 因为输入约束 $\|\mathbf{I}_{\mathrm{mask}}(\boldsymbol{x} - \boldsymbol{x}^*)\|_p \leqslant \varepsilon$, 以及 $\|\mathbf{I}_{\mathrm{mask}}(\boldsymbol{x} - \boldsymbol{x}^*)\|_p = \|\mathbf{V}(\boldsymbol{x} - \boldsymbol{x}^*)\|_p$, 可得 $\|\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}^*\|_p \leqslant \varepsilon$, 即 $\tilde{\boldsymbol{x}} \in \tilde{\eta}$. 当 f 以 $\boldsymbol{x}$ 作为输入, 其第 1 层神经元的激活前值为 $\boldsymbol{z}^{(1)} = \mathbf{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$; 当 $\tilde{f}$ 以 $\tilde{\boldsymbol{x}}$ 作为输入, 其第 1 层神经元的激活前值为 $\tilde{\boldsymbol{z}}^{(1)} = \tilde{\mathbf{W}}^{(1)}\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{b}}^{(1)} = \mathbf{W}^{(1)}\mathbf{V}^{\mathrm{T}}\mathbf{V}\boldsymbol{x} + \boldsymbol{b}^{(1)} + \mathbf{W}^{(1)}(\mathbf{I} - \mathbf{I}_{\mathrm{mask}})\boldsymbol{x}^* = \mathbf{W}^{(1)}\mathbf{I}_{\mathrm{mask}}\boldsymbol{x} + \mathbf{W}^{(1)}(\mathbf{I} - \mathbf{I}_{\mathrm{mask}})\boldsymbol{x}^* + \boldsymbol{b}^{(1)}$, 由输入约束 $(\mathbf{I} - \mathbf{I}_{\mathrm{mask}})\boldsymbol{x} = (\mathbf{I} - \mathbf{I}_{\mathrm{mask}})\boldsymbol{x}^*$, 可得 $\tilde{\boldsymbol{z}}^{(1)} = \mathbf{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$, 所以有 $\boldsymbol{z}^{(1)} = \tilde{\boldsymbol{z}}^{(1)}$ 成立. f 和 $\tilde{f}$ 除了第 1 层的权重矩阵以及偏置向量不同, 其余结构以及参数均相同, 所以有 $f(\boldsymbol{x}) = \tilde{f}(\tilde{\boldsymbol{x}})$. 特别地, $f(\boldsymbol{x}^*) = \tilde{f}(\tilde{\boldsymbol{x}}^*)$.

若 $\tilde{f}$ 在 $\tilde{\eta}$ 上是鲁棒的, 则 f 在 η 上也是鲁棒的. 用反证法证明. 假设 f 在 η 上存在对抗样本, 我们记该对抗样本为 $\boldsymbol{x}_{\mathrm{adv}}$. 由前面的结论, 我们构造 $\tilde{\boldsymbol{x}}_{\mathrm{adv}} = \mathbf{V}\boldsymbol{x}_{\mathrm{adv}} \in \tilde{\eta}$, 且满足 $f(\boldsymbol{x}_{\mathrm{adv}}) = \tilde{f}(\tilde{\boldsymbol{x}}_{\mathrm{adv}})$. 因为 $\tilde{f}$ 在 $\tilde{\eta}$ 上是鲁棒的, 根据鲁棒性定义有:

$$\arg\max_{i\in[d_L]} \tilde{f}(\tilde{\boldsymbol{x}}_{\mathrm{adv}})_i = \arg\max_{i\in[d_L]} \tilde{f}(\tilde{\boldsymbol{x}}^*)_i \tag{17}$$

又因为 $f(\boldsymbol{x}_{\mathrm{adv}}) = \tilde{f}(\tilde{\boldsymbol{x}}_{\mathrm{adv}})$ 以及 $f(\boldsymbol{x}^*) = \tilde{f}(\tilde{\boldsymbol{x}}^*)$, 故

$$\arg\max_{i\in[d_L]} f(\boldsymbol{x}_{\mathrm{adv}})_i = \arg\max_{i\in[d_L]} f(\boldsymbol{x}^*)_i \tag{18}$$

由公式 (18) 知 $\boldsymbol{x}_{\mathrm{adv}}$ 不会引起分类结果的改变, 因此 $\boldsymbol{x}_{\mathrm{adv}}$ 不是对抗样本, 与假设矛盾, 所以 f 在 η 上不存在对抗样本, 即 f 在 η 上是鲁棒的.

若 $\tilde{f}$ 在 $\tilde{\eta}$ 上存在对抗样本 $\tilde{\boldsymbol{x}}_{\mathrm{adv}}$, 则 f 在 η 上能找到对抗样本 $\boldsymbol{x}_{\mathrm{adv}}$. 我们将 $\boldsymbol{x}_{\mathrm{adv}}$ 和 $\tilde{\boldsymbol{x}}_{\mathrm{adv}}$ 第 j 个元素分别记为 $\boldsymbol{x}_{\mathrm{adv}\_j}$ 以及 $\tilde{\boldsymbol{x}}_{\mathrm{adv}\_j}$, 构造 $\boldsymbol{x}_{\mathrm{adv}}$:

$$\boldsymbol{x}_{\mathrm{adv}\_n_j} = \tilde{\boldsymbol{x}}_{\mathrm{adv}\_j},\ j = 1,2,\ldots,d_0 - k \tag{19}$$

$$\boldsymbol{x}_{\mathrm{adv}\_m_j} = \boldsymbol{x}^*_{m_j},\ j = 1,2,\ldots,k \tag{20}$$

这是压缩的逆过程, 通过对 $\tilde{\boldsymbol{x}}_{\mathrm{adv}}$ 在不允许扰动的维度上补充 $\boldsymbol{x}^*$ 对应的元素值, 从而将 $\tilde{\boldsymbol{x}}_{\mathrm{adv}}$ 还原成 $\boldsymbol{x}_{\mathrm{adv}}$, 显然有 $\tilde{\boldsymbol{x}}_{\mathrm{adv}} = \mathbf{V}\boldsymbol{x}_{\mathrm{adv}}$. $\tilde{\eta}$ 上的输入约束保证了 $\|\tilde{\boldsymbol{x}}_{\mathrm{adv}} - \tilde{\boldsymbol{x}}^*\|_p = \|\mathbf{V}(\boldsymbol{x}_{\mathrm{adv}} - \boldsymbol{x}^*)\|_p \leqslant \varepsilon$, 因 $\|\mathbf{I}_{\mathrm{mask}}(\boldsymbol{x}_{\mathrm{adv}} - \boldsymbol{x}^*)\|_p = \|\mathbf{V}(\boldsymbol{x}_{\mathrm{adv}} - \boldsymbol{x}^*)\|_p$, 所以有:

$$\|\mathbf{I}_{\mathrm{mask}}(\boldsymbol{x}_{\mathrm{adv}} - \boldsymbol{x}^*)\|_p \leqslant \varepsilon \tag{21}$$

维度扩展操作保证了 $\boldsymbol{x}_{\mathrm{adv}}$ 在不可变维度的特征的值和 $\boldsymbol{x}^*$ 是一样的, 即:

$$(\mathbf{I}-\mathbf{I}_{\text{mask}})\boldsymbol{x}_{\text{adv}}=(\mathbf{I}-\mathbf{I}_{\text{mask}})\boldsymbol{x}^{*} \tag{22}$$

根据公式 (21)、(22), $\boldsymbol{x}_{\text{adv}}\in\eta$. 由鲁棒性定义, 对抗样本满足:

$$\arg\max_{i\in[d_L]}\tilde{f}(\tilde{\boldsymbol{x}}_{\text{adv}})_i\neq\arg\max_{i\in[d_L]}\tilde{f}(\tilde{\boldsymbol{x}}^{*})_i \tag{23}$$

因为 $f(\boldsymbol{x}_{\text{adv}})=\tilde{f}(\tilde{\boldsymbol{x}}_{\text{adv}})$ 且 $f(\boldsymbol{x}^{*})=\tilde{f}(\tilde{\boldsymbol{x}}^{*})$, 故 $\arg\max_{i\in[d_L]}f(\boldsymbol{x}_{\text{adv}})_i\neq\arg\max_{i\in[d_L]}f(\boldsymbol{x}^{*})_i$, 所以 $\boldsymbol{x}_{\text{adv}}$ 是 η 上的对抗样本. 综上所述, 定理 3 证毕.

3.4 马氏距离鲁棒性分支

3.4.1 马氏距离鲁棒性的定义

在某些数据集上, 不同特征之间往往并非相互独立, 而是存在一定的相关性. 对于这样的数据分布, 椭球鲁棒性和掩码局部鲁棒性较难建模其上的鲁棒性. 相比之下, 马氏距离能够同时刻画特征的缩放特性及其维度间的相关性, 从而更准确地描述这类数据的分布特性. 数据集 X 的协方差矩阵记为 $\boldsymbol{\Sigma}$, X 上的向量 $\boldsymbol{x}$, $\boldsymbol{x}^{*}\in\mathbb{R}^{d}$ 之间的马氏距离可以用以下表达式描述:

$$d_{\text{M}}(\boldsymbol{x},\boldsymbol{x}^{*})=\sqrt{(\boldsymbol{x}-\boldsymbol{x}^{*})^{\text{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{x}^{*})} \tag{24}$$

如果向量满足各向同性假设 (即特征间独立且同分布), 则协方差矩阵将退化为一个标量矩阵, 其对角线元素均为特征的方差 (在同分布条件下, 所有特征具有相同的方差). 在这种情况下, 马氏距离与欧氏距离将产生相同的度量效果. 基于这一特性, 我们进一步形式化地定义了神经网络上的马氏距离鲁棒性.

定义 8. 神经网络在 L_p 范数下的 ε-马氏距离局部鲁棒性. 给定一个神经网络 $f:\mathbb{R}^{d_0}\to\mathbb{R}^{d_L}$, 数据分布上的协方差矩阵 $\boldsymbol{\Sigma}$, 我们将协方差矩阵的逆矩阵进行 Cholesky 分解, $\boldsymbol{\Sigma}^{-1}=\mathbf{U}^{\text{T}}\mathbf{U}$, 输入空间为 $\eta=\{\boldsymbol{x}\mid\|\mathbf{U}(\mathbf{x}-\mathbf{x}^{*})\|_p\leqslant\varepsilon\}$. f 对于输入 $\boldsymbol{x}^{*}\in\mathbb{R}^{d_0}$ 是 ε-马氏距离局部鲁棒的, 当且仅当:

$$\forall\boldsymbol{x}\in\mathbb{R}^{d_0},\ \|\mathbf{U}(\boldsymbol{x}-\boldsymbol{x}^{*})\|_p\leqslant\varepsilon\Rightarrow(\arg\max_{i\in[d_L]}f(\boldsymbol{x})_i=\arg\max_{i\in[d_L]}f(\boldsymbol{x}^{*})_i) \tag{25}$$

马氏距离在距离度量计算时使用了平方根, 为了与这一特性保持一致, 我们一般仅考虑 L_2 范数下马氏距离鲁棒性. 如果不引入特征间的线性相关性假设, 并假设所有的特征方差相同且彼此不相关时, 此时协方差矩阵为标量矩阵, 由协方差矩阵的逆进行 Cholesky 分解得到的线性变换矩阵 $\mathbf{U}$ 为标量矩阵. 在这一情形下, 马氏距离鲁棒性也随之退化为均匀 L_p 范数球鲁棒性.

3.4.2 马氏距离鲁棒性与数据的一致性

根据第 3.1 节中的数据一致性定义, 设一致性阈值 $\gamma>0$, 一个经扰动的数据 $\boldsymbol{x}$ 符合一致性要求, 当且仅当 $p_{X\sim\mathcal{F}}(X=\boldsymbol{x})\geqslant\gamma$.

定理 4. 当数据服从均值为 $\boldsymbol{\mu}$, 协方差为 $\boldsymbol{\Sigma}$ 的高斯分布时, 其概率密度函数记为 $p(\boldsymbol{x})$. 在输入空间 $\eta=\{\boldsymbol{x}\mid\|\mathbf{U}(\boldsymbol{x}-\boldsymbol{x}^{*})\|_2\leqslant\varepsilon\}$ 内, 数据的概率密度函数存在一个下界:

$$\ln p(\boldsymbol{x})\geqslant C-\frac{1}{2}\varepsilon^{2}-\varepsilon\left\|\left(\mathbf{U}^{-1}\right)^{\text{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^{*}-\boldsymbol{\mu})\right\|_2 \tag{26}$$

其中, $C=-\ln\left[(2\pi)^{\frac{d_0}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}\right]-\frac{1}{2}(\boldsymbol{x}^{*}-\boldsymbol{\mu})^{\text{T}}\sum^{-1}(\boldsymbol{x}^{*}-\boldsymbol{\mu})$, d_0 是输入向量 $\boldsymbol{x}$ 的维度.

证明: 因为数据服从均值为 $\boldsymbol{\mu}$, 协方差为 $\boldsymbol{\Sigma}$ 的高斯分布, 所以概率密度的对数为:

$$\ln p(\boldsymbol{x})=-\ln\left[(2\pi)^{\frac{d_0}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}\right]-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\text{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \tag{27}$$

因为 $\boldsymbol{x}\in\eta=\{\boldsymbol{x}\mid\|\mathbf{U}(\boldsymbol{x}-\boldsymbol{x}^{*})\|_2\leqslant\varepsilon\}$, 我们记 $\boldsymbol{\delta}=\boldsymbol{x}-\boldsymbol{x}^{*}$, 因而有 $\|\mathbf{U}\boldsymbol{\delta}\|_2\leqslant\varepsilon$, 且有:

$$\begin{aligned}\ln p(\boldsymbol{x})&=-\ln\left[(2\pi)^{\frac{d_0}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}\right]-\frac{1}{2}(\boldsymbol{x}^{*}-\boldsymbol{\mu}+\boldsymbol{\delta})^{\text{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^{*}-\boldsymbol{\mu}+\boldsymbol{\delta})\\&=-\ln\left[(2\pi)^{\frac{d_0}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}\right]-\frac{1}{2}(\boldsymbol{x}^{*}-\boldsymbol{\mu})^{\text{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^{*}-\boldsymbol{\mu})-\frac{1}{2}\boldsymbol{\delta}^{\text{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}-(\boldsymbol{x}^{*}-\boldsymbol{\mu})^{\text{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}\end{aligned} \tag{28}$$

我们记 $C=-\ln\left[(2\pi)^{\frac{d_0}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}\right]-\frac{1}{2}(\boldsymbol{x}^{*}-\boldsymbol{\mu})^{\text{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^{*}-\boldsymbol{\mu})$, 则:

$$\inf_{x\in\eta}\ln p(\boldsymbol{x})=C-\frac{1}{2}\sup_{\|\mathbf{U}\boldsymbol{\delta}\|_2\leqslant\varepsilon}\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}-\sup_{\|\mathbf{U}\boldsymbol{\delta}\|_2\leqslant\varepsilon}(\boldsymbol{x}^*-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} \tag{29}$$

因为 $\boldsymbol{\delta}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U}\boldsymbol{\delta}=\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}\leqslant\varepsilon^2$, 故 $\sup_{\|\mathbf{U}\boldsymbol{\delta}\|_2\leqslant\varepsilon}\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}=\varepsilon^2$.

引理 1. 柯西-施瓦茨不等式. 设 $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^n$, $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}\leqslant\sum_{i=1}^{n}|\boldsymbol{a}_i\boldsymbol{b}_i|\leqslant\|\boldsymbol{a}\|_2\|\boldsymbol{b}\|_2$. 等号成立的条件当且仅当 $\boldsymbol{a}=\lambda\boldsymbol{b}$, $\lambda\geqslant 0$.

根据引理 1, 我们令 $\boldsymbol{a}=\varepsilon(\mathbf{U}^{-1})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^*-\boldsymbol{\mu})$, $\boldsymbol{b}=\dfrac{\mathbf{U}\boldsymbol{\delta}}{\varepsilon}$, 则有:

$$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}=(\boldsymbol{x}^*-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}\leqslant\left\|\left(\mathbf{U}^{-1}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^*-\boldsymbol{\mu})\right\|_2\|\mathbf{U}\boldsymbol{\delta}\|_2\leqslant\varepsilon\left\|\left(\mathbf{U}^{-1}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^*-\boldsymbol{\mu})\right\|_2 \tag{30}$$

即 $\sup_{\|\mathbf{U}\boldsymbol{\delta}\|_2\leqslant\varepsilon}(\boldsymbol{x}^*-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}=\varepsilon\left\|(\mathbf{U}^{-1})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^*-\boldsymbol{\mu})\right\|_2$, 所以:

$$\inf_{x\in\eta}\ln p(\boldsymbol{x})=C-\frac{1}{2}\varepsilon^2-\varepsilon\left\|\left(\mathbf{U}^{-1}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^*-\boldsymbol{\mu})\right\|_2 \tag{31}$$

定理 4 证毕.

概率密度函数的下界是关于 ε 的二次函数, 并随着 ε 的增大而减小, 这意味着在马氏距离鲁棒性所定义的输入空间中, 我们可以通过调整 ε 的大小来使其中数据的概率密度大于一致性阈值 γ, 进而保证数据一致性.

3.4.3 马氏距离鲁棒性验证问题的等价构建

对于一个神经网络 f, 我们尝试将其关于 $\boldsymbol{x}^*$ 的马氏距离鲁棒性验证问题转换成一个与之等价的神经网络 $\tilde{f}$ 关于 $\tilde{\boldsymbol{x}}^*$ 的 L_p 范数球鲁棒性验证问题, 并提出以下定理.

定理 5. 神经网络的马氏距离鲁棒性验证问题的等价构建. 给定一个神经网络 $f:\mathbb{R}^{d_0}\to\mathbb{R}^{d_L}$, 根据分布上的协方差矩阵 $\boldsymbol{\Sigma}$, 正实数 $\varepsilon>0$, 数据点 $\boldsymbol{x}^*=(\boldsymbol{x}_1^*,\boldsymbol{x}_2^*,\ldots,\boldsymbol{x}_{d_0}^*)$, 以及矩阵 $\mathbf{U}$ 满足 $\boldsymbol{\Sigma}^{-1}=\mathbf{U}^{\mathrm{T}}\mathbf{U}$. 我们构建新的数据点 $\tilde{\boldsymbol{x}}^*=\mathbf{U}\boldsymbol{x}^*$, 以及与 f 拥有相同结构的神经网络 $\tilde{f}:\mathbb{R}^{d_0}\to\mathbb{R}^{d_L}$, $\tilde{f}$ 中的参数满足:

$$\tilde{\mathbf{W}}^{(i)}=\mathbf{W}^{(i)},\ i=2,3,\ldots,L \tag{32}$$

$$\tilde{\mathbf{W}}^{(1)}=\mathbf{W}^{(1)}\mathbf{U}^{-1} \tag{33}$$

$$\tilde{\boldsymbol{b}}^{(i)}=\boldsymbol{b}^{(i)},\ i=1,2,\ldots,L \tag{34}$$

f 在 $\eta=\{\boldsymbol{x}\mid\|\mathbf{U}(\boldsymbol{x}-\boldsymbol{x}^*)\|_p\leqslant\varepsilon\}$ 上的马氏距离鲁棒性验证问题等价于 $\tilde{f}$ 在 $\tilde{\eta}=\{\boldsymbol{x}\mid\|\boldsymbol{x}-\tilde{\boldsymbol{x}}^*\|_p\leqslant\varepsilon\}$ 上的均匀 L_p 范数球鲁棒性验证问题.

证明: 对于 $\forall\boldsymbol{x}\in\eta$, 我们可以构造 $\tilde{\boldsymbol{x}}=\mathbf{U}\boldsymbol{x}$, 因为 $\|\mathbf{U}(\boldsymbol{x}-\boldsymbol{x}^*)\|_p\leqslant\varepsilon$, 且有 $\tilde{\boldsymbol{x}}^*=\mathbf{U}\boldsymbol{x}^*$, 故 $\|\tilde{\boldsymbol{x}}-\tilde{\boldsymbol{x}}^*\|_p\leqslant\varepsilon$, $\tilde{\boldsymbol{x}}\in\tilde{\eta}$. 当 f 以 $\boldsymbol{x}$ 作为输入, 其第 1 层神经元的激活前值 $\boldsymbol{z}^{(1)}=\mathbf{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)}$; 当 $\tilde{f}$ 以 $\tilde{\boldsymbol{x}}$ 作为输入, 其第 1 层神经元的激活前值 $\tilde{\boldsymbol{z}}^{(1)}=\tilde{\mathbf{W}}^{(1)}\tilde{\boldsymbol{x}}+\tilde{\boldsymbol{b}}^{(1)}=\mathbf{W}^{(1)}\mathbf{U}^{-1}\mathbf{U}\boldsymbol{x}+\boldsymbol{b}^{(1)}=\mathbf{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)}$, 因而有 $\boldsymbol{z}^{(1)}=\tilde{\boldsymbol{z}}^{(1)}$. 因为 f 和 $\tilde{f}$ 除了第 1 层的权重矩阵不同, 其余结构与参数均相同, 所以有 $f(\boldsymbol{x})=\tilde{f}(\tilde{\boldsymbol{x}})$. 特别地, $f(\boldsymbol{x}^*)=\tilde{f}(\tilde{\boldsymbol{x}}^*)$.

若 $\tilde{f}$ 在 $\tilde{\eta}$ 上是鲁棒的, 则 f 在 η 上是鲁棒的. 用反证法证明. 若 f 在 η 上存在对抗样本, 我们记该对抗样本为 $\boldsymbol{x}_{\mathrm{adv}}$. 由前面的结论, 我们可以构造 $\tilde{\boldsymbol{x}}_{\mathrm{adv}}=\mathbf{U}\boldsymbol{x}_{\mathrm{adv}}\in\tilde{\eta}$, 且有 $f(\boldsymbol{x}_{\mathrm{adv}})=\tilde{f}(\tilde{\boldsymbol{x}}_{\mathrm{adv}})$. 由于 $\tilde{f}$ 在 $\tilde{\eta}$ 上是鲁棒的, 根据鲁棒性定义:

$$\arg\max_{i\in[d_L]}\tilde{f}(\tilde{\boldsymbol{x}}_{\mathrm{adv}})_i=\arg\max_{i\in[d_L]}\tilde{f}(\tilde{\boldsymbol{x}}^*)_i \tag{35}$$

由 $f(\boldsymbol{x}_{\mathrm{adv}})=\tilde{f}(\tilde{\boldsymbol{x}}_{\mathrm{adv}})$ 以及 $f(\boldsymbol{x}^*)=\tilde{f}(\tilde{\boldsymbol{x}}^*)$, 我们有:

$$\arg\max_{i\in[d_L]}f(\boldsymbol{x}_{\mathrm{adv}})_i=\arg\max_{i\in[d_L]}f(\boldsymbol{x}^*)_i \tag{36}$$

因为 $\boldsymbol{x}_{\mathrm{adv}}$ 不会引起分类结果的改变, 所以 $\boldsymbol{x}_{\mathrm{adv}}$ 不是对抗样本, 与假设矛盾. 所以 f 在 η 上不存在对抗样本, 即 f 在 η 上是鲁棒的.

若 $\tilde{f}$ 在 $\tilde{\eta}$ 上存在对抗样本 $\tilde{\boldsymbol{x}}_{\mathrm{adv}}$, 则 f 在 η 上存在对抗样本 $\boldsymbol{x}_{\mathrm{adv}}=\mathbf{U}^{-1}\tilde{\boldsymbol{x}}_{\mathrm{adv}}$. 因为 $\|\tilde{\boldsymbol{x}}_{\mathrm{adv}}-\tilde{\boldsymbol{x}}^*\|_p\leqslant\varepsilon$, 即 $\|\mathbf{U}(\boldsymbol{x}_{\mathrm{adv}}-\boldsymbol{x}^*)\|_p\leqslant\varepsilon$, 故 $\boldsymbol{x}_{\mathrm{adv}}\in\eta$, 又由鲁棒性定义, 对抗样本满足:

$$\arg\max_{i\in[d_L]}\tilde{f}(\tilde{\boldsymbol{x}}_{\text{adv}})_i \neq \arg\max_{i\in[d_L]}\tilde{f}(\tilde{\boldsymbol{x}}^*)_i \tag{37}$$

而 $f(\boldsymbol{x}_{\text{adv}})=\tilde{f}(\tilde{\boldsymbol{x}}_{\text{adv}})$ 且 $f(\boldsymbol{x}^*)=\tilde{f}(\tilde{\boldsymbol{x}}^*)$, 故 $\arg\max_{i\in[d_L]}f(\boldsymbol{x}_{\text{adv}})_i \neq \arg\max_{i\in[d_L]}f(\boldsymbol{x}^*)_i$, 所以 $\boldsymbol{x}_{\text{adv}}$ 是 η 上的对抗样本.

综上所述, 定理 5 证毕.

3.5 NNV4RADAP 算法

我们给出 NNV4RADAP 算法, 如算法 1 所示. NNV4RADAP 算法的输入包括: 神经网络 f、扰动约束 ε、数据样本 x^* 以及训练数据集 *Dataset*, 该数据集用于椭球鲁棒性和马氏距离鲁棒性验证中的线性变换矩阵的计算. 算法的核心思想是: 通过对神经网络第 1 层的权重与偏置参数, 以及待验证的数据样本进行线性变换, 将前文所述 3 种非均匀扰动域下的鲁棒性验证问题, 等价地归约为传统的均匀 L_p 范数球邻域下的验证问题. 对于椭球鲁棒性验证问题, 我们首先基于训练数据集 *Dataset* 计算各特征与输出标签之间的互信息值 $\overline{mif}_{i,y}$, $i=1,2,\ldots,d_0$, 并据此构建对角矩阵 $\mathbf{D}$, 其中 $\mathbf{D}_{i,i}=\overline{mif}_{i,y}$, $i=1,2,\ldots,d_0$, 其对角线元素表示各特征的重要性, 并根据定理 2 对原网络的参数和待验证样本进行线性变换; 对于掩码局部鲁棒性验证问题, 我们首先确定哪些特征不允许改变, 之后依据定理 3 构建掩码单位矩阵 $\mathbf{I}_{\text{mask}}$ 和压缩矩阵 $\mathbf{V}$, 并对原网络的参数和待验证样本进行线性变换; 对于马氏距离鲁棒性验证问题, 我们首先计算训练数据集 *Dataset* 上的协方差矩阵 $\boldsymbol{\Sigma}$, 并对其逆矩阵进行 Cholesky 分解, 获得线性变换矩阵 $\mathbf{U}$, 之后依据定理 5, 对原网络的参数和待验证样本进行线性变换. 对于 3 种非均匀鲁棒性验证问题, 在上述操作后我们能得到变换后的网络 $\tilde{f}$ 以及数据 $\tilde{\boldsymbol{x}}^*$, 根据定理 2、定理 3 与定理 5, 原网络在 3 种非均匀扰动域下的验证问题等价于变换后网络 $\tilde{f}$ 在 $\tilde{\eta}=\{\boldsymbol{x}\,|\,\|\boldsymbol{x}-\tilde{\boldsymbol{x}}^*\|_p\leqslant\varepsilon\}$ 上的均匀 L_p 范数球鲁棒性验证问题. 因此, 我们可直接调用现有的验证引擎, 对变换后的验证问题进行求解, 从而输出最终的验证结果.

算法 1. NNV4RADAP 算法.

输入: 待验证的神经网络 f, 训练数据集 *Dataset*, 扰动预算 ε, 待验证数据点 $\boldsymbol{x}^*$;
输出: 验证结果 (SAT, UNSAT, UNK).

1. If 验证椭球鲁棒性 then
2. 计算 *Dataset* 上的特征与分类标签的互信息值 $\overline{mif}_{i,y}$, $i=1,2,\ldots,d_0$
3. 构建对角矩阵 $\mathbf{D}$, 其中 $\mathbf{D}_{i,i}=\overline{mif}_{i,y}$, $i=1,2,\ldots,d_0$
4. 构造数据点 $\tilde{\boldsymbol{x}}^*=\mathbf{D}\boldsymbol{x}^*$
5. 构建具有相同结构的神经网络 $\tilde{f}$, 参数赋值为:
6. $$\begin{cases}\tilde{\mathbf{W}}^{(i)}=\mathbf{W}^{(i)},\ i=2,3,\ldots,L\\ \tilde{\mathbf{W}}^{(1)}=\mathbf{W}^{(1)}\mathbf{D}^{-1}\\ \tilde{\boldsymbol{b}}^{(i)}=\boldsymbol{b}^{(i)},\ i=1,2,\ldots,L\end{cases}$$
7. Elif 验证掩码局部鲁棒性 then
8. 根据不可变的特征计算掩码单位矩阵 $\mathbf{I}_{\text{mask}}$ 以及压缩矩阵 $\mathbf{V}$
9. 构造数据点 $\tilde{\boldsymbol{x}}^*=\mathbf{V}\boldsymbol{x}^*$
10. 构建新的神经网络 $\tilde{f}$, 参数赋值为:
11. $$\begin{cases}\tilde{\mathbf{W}}^{(i)}=\mathbf{W}^{(i)},\ i=2,3,\ldots,L\\ \tilde{\mathbf{W}}^{(1)}=\mathbf{W}^{(1)}\mathbf{V}^{\mathrm{T}}\\ \tilde{\boldsymbol{b}}^{(i)}=\boldsymbol{b}^{(i)},\ i=2,3,\ldots,L\\ \tilde{\boldsymbol{b}}^{(1)}=\boldsymbol{b}^{(1)}+\mathbf{W}^{(1)}(\mathbf{I}-\mathbf{I}_{\text{mask}})\boldsymbol{x}^*\end{cases}$$
12. Else
13. 计算 *Dataset* 上的协方差矩阵 $\boldsymbol{\Sigma}$
14. 计算 $\mathbf{U}$ 满足 $\boldsymbol{\Sigma}^{-1}=\mathbf{U}^{\mathrm{T}}\mathbf{U}$

15. 构造数据点 $\tilde{\boldsymbol{x}}^* = \mathbf{U}\boldsymbol{x}^*$
16. 构建具有相同结构的神经网络 $\tilde{f}$, 参数赋值为:
17. $\begin{cases} \tilde{\mathbf{W}}^{(i)} = \mathbf{W}^{(i)},\ i = 2,3,\ldots,L \\ \tilde{\mathbf{W}}^{(1)} = \mathbf{W}^{(1)}\mathbf{U}^{-1} \\ \tilde{\boldsymbol{b}}^{(i)} = \boldsymbol{b}^{(i)},\ i = 1,2,\ldots,L \end{cases}$
18. 调用验证引擎对等价验证问题 $(\tilde{f}, \tilde{\boldsymbol{x}}^*, \varepsilon)$ 进行验证
19. 输出验证结果

4 实验与分析

本研究基于 Python 实现了 NNV4RADAP 算法. 首先我们对网络的训练集进行了数据相关性分析, 并据此构建了特征矩阵. 接着, 利用特征矩阵对网络参数进行了重构并调用通用的验证引擎对等价问题进行验证. 值得注意的是, 我们的算法可以与任一面向 L_p 范数球的深度网络验证算法结合.

β-CROWN[13]是一款基于高效线性边界传播 (bound propagation) 与分支定界 (branch and bound) 技术的神经网络形式化验证引擎. 我们选择了 β-CROWN 作为通用验证引擎, 并结合 PGD 攻击来寻找对抗样本. 在椭球鲁棒性和掩码局部鲁棒性分支中, 我们研究基于 L_∞ 范数下的验证区域; 在马氏距离鲁棒性分支中, 我们研究基于 L_2 范数下的验证区域. 验证结果分为以下几种情况.

(1) 当验证结果显示 UNSAT 时, 表明网络对于当前输入是鲁棒的.

(2) 返回 SAT 表明存在对抗样本.

(3) 返回 UNK 表明因为超过了验证器设定的最大分支限制或最大验证时间限制, 此时无法判断网络是鲁棒的还是不安全的.

为评估 NNV4RADAP 算法的有效性, 我们在几个典型应用领域 (如电机故障诊断、恶意软件检测和垃圾邮件检测) 进行了实验验证, 并与传统的均匀扰动验证模式进行了对比分析, 实验结果将在第 4.1–4.4 节详细展示.

我们的研究旨在回答以下几个问题.

Q1: NNV4RADAP 算法是否是有实际意义的? 具体来说, 对于一个神经网络、一个数据点 $\boldsymbol{x}^*$ 以及一个正实数 ε, 如果中心为 $\boldsymbol{x}^*$, 半径为 ε 的均匀 L_p 范数球被证明是 SAT (UNSAT) 的, NNV4RADAP 是否能证明一个更符合实际的以 $\boldsymbol{x}^*$ 为中心的 L_p 范数邻域是 UNSAT (SAT) 的?

Q2: NNV4RADAP 算法在性能上是否存在损耗? 具体来说, 对于给定的一个验证问题, NNV4RADAP 算法的性能相比原始神经网络验证算法如何?

因此, 我们的实验涉及以一个基于边界传播的验证工具 β-CROWN 为验证内核的 NNV4RADAP 算法实现以及与该验证工具的原型之间的对比. 实验在一台运行 Ubuntu 18.04 的 Linux 服务器上进行, 服务器配备有 Intel Xeon CPU E5-2678 v3, 拥有 12 个 CPU 核心.

为了更直观体现对比实验结果, 我们定义了以下指标.

定义 9. 非一致率. 给定 n 组对比验证实验结果 $(A_1, B_1), \ldots, (A_n, B_n)$, 其中 $A_i, B_i \in \{\text{SAT, UNSAT, UNK}\}$ 我们用非一致率 (non-consistency rate, *NCR*) 描述在对照实验中对照结果均明确 (不为 UNK) 的情况下结果不一致的比例. 非一致率定义如下:

$$NCR = \frac{card(i \mid A_i \neq \text{UNK} \wedge B_i \neq \text{UNK} \wedge A_i \neq B_i)}{card(i \mid A_i \neq \text{UNK} \wedge B_i \neq \text{UNK})} \times 100\% \tag{38}$$

其中, $card(X)$ 表示集合 X 中的元素数量.

4.1 椭球鲁棒性验证

4.1.1 数据集介绍

我们选择了源自 UCI 的 SDD (sensorless drive diagnosis) 数据集[38]进行实验, 以评估椭球鲁棒性验证方法相

较于传统的均匀范数球验证方法的有效性及两者在性能上的差异. 该数据集主要用于电机故障诊断, 目标是区分电机的不同工作状态 (包括正常状态和多种故障状态). SDD 数据集来源于对电机驱动器电流信号的监测, 旨在通过传感器数据识别设备的不同故障模式. 数据集中每个样本对应驱动器的运行状态, 包含 48 个连续特征, 采样自电流信号的多维度分析, 例如频域特征以及时域统计量 (如均值、方差等), 均为浮点类型. 此外, 每个样本还包含一个类别标签, 用于表示 11 种可能的故障模式: 标签 0 表示正常运作状态, 标签 1–10 分别表示不同的故障状态, 如轴承故障、转子不平衡等.

该数据集一共包含了 58 509 条记录, 类别分布均匀. 我们按 8:2 划分训练集和测试集并分别训练了两个神经网络用于后续实验, 分别为 SDD-S 和 SDD-L. 两个模型的具体神经网络结构分别是 48×100×100×100×11 和 48×300×300×300×11. 其中, 两个模型的隐藏层维度分别为 100 和 300, 输入、输出维度均为 48、11.

4.1.2 阴性非一致率

为了更直观地展示 L_p 范数球鲁棒性验证中的假阴性问题, 我们对非一致率这一指标进行了细分.

定义 10. 阴性非一致率. 给定 n 组对比验证实验结果 $(A_1,B_1),\ldots,(A_n,B_n)$, 其中 A_i 是对照组, B_i 是实验组, 且 $A_i,B_i\in\{\text{SAT, UNSAT, UNK}\}$, 我们用阴性非一致率 (UNSAT non-consistency rate, *US-NCR*) 描述在对照实验中对照组结果为 UNSAT 且实验组结果不为 UNK 的情况下实验组结果为 SAT 的比例. 阴性非一致率定义如下:

$$US\text{-}NCR=\frac{card(i\mid A_i=\text{UNSAT}\wedge B_i=\text{SAT})}{card(i\mid A_i=\text{UNSAT}\wedge B_i\neq\text{UNK})}\times 100\% \tag{39}$$

阴性非一致率这一指标揭示了在椭球扰动域内被认定为不安全的样本, 却在均匀范数球验证中被视为安全的现象. 具体而言, 这一比率衡量了椭球鲁棒性验证模式下识别潜在对抗样本的能力, 而这些对抗样本在基于均匀范数球的方法中未能被识别. 当给定一个合理的扰动预算时, 较高的阴性非一致率表明均匀范数球验证模式可能对模型的安全性评估过于乐观.

我们使用 NNV4RADAP 进行椭球鲁棒性验证, 为了回答 Q1, 基于 SDD-S 和 SDD-L 进行了验证实验.

首先, 我们通过 sklearn 库中的 mutual_info_classif 函数计算各个特征与分类标签之间的互信息值. 在对互信息进行归一化处理中我们选择了 $k=2$ 保证最小互信息值与最大互信息值的倍数不超过两倍. 最终, 我们根据归一化后的互信息值构建了互信息对角矩阵 $\mathbf{D}$.

针对每个待验证的数据点 $\boldsymbol{x}^*$ 以及给定的扰动 ε, 我们首先构建了基于均匀 L_p 范数球的鲁棒性验证问题作为对照组. 在实验组设置上, 如图 3(a), 我们通过识别互信息矩阵 $\mathbf{D}$ 中的最大对角线元素所对应的特征 (对分类结果贡献最大的特征), 并将该特征对应的扰动预算设定为 ε, 构造出最短轴长度为 ε 的椭球验证区域, 该椭球区域包含了均匀范数球区域.

针对每种电机故障模式, 我们从测试集中各随机选取了 100 个数据点, 总共选取了 1 100 个数据点, 并针对 L_p 范数球鲁棒性验证中假阴性的情况进行了实验, 实验结果如表 1 所示.

表 1 均匀范数球扰动和椭球扰动下的验证结果 (SDD 数据集)

模型	扰动预算 (ε)	扰动类型	UNSAT	SAT	UNK	平均验证时间 (s)			*US-NCR* (%)
						返回UNSAT	返回SAT	返回UNSAT或UNK	
SDD-S	0.001	均匀范数球扰动	1 071	28	1	0.126	0.008 1	0.127	0.56
		椭球扰动	1 065	35	0	0.128	0.010 0	0.128	
	0.003	均匀范数球扰动	960	137	3	0.146	0.012 6	0.149	6.78
		椭球扰动	893	202	5	0.154	0.010 6	0.158	
	0.005	均匀范数球扰动	749	349	2	0.205	0.009 2	0.210	14.61
		椭球扰动	631	457	12	0.234	0.008 7	0.277	
	0.007	均匀范数球扰动	537	558	5	0.396	0.008 0	0.404	21.09
		椭球扰动	419	670	11	0.351	0.008 1	0.483	
	0.01	均匀范数球扰动	321	778	1	0.350	0.008 0	0.407	23.75
		椭球扰动	244	854	2	0.444	0.007 7	0.477	

表 1 均匀范数球扰动和椭球扰动下的验证结果 (SDD 数据集)(续)

模型	扰动预算 (ε)	扰动类型	UNSAT	SAT	UNK	平均验证时间 (s)			*US-NCR* (%)
						返回UNSAT	返回SAT	返回UNSAT或UNK	
SDD-S	0.02	均匀范数球扰动	144	954	2	0.395	0.0074	0.395	9.86
		椭球扰动	128	970	2	0.662	0.0074	0.731	
	0.03	均匀范数球扰动	103	995	2	1.584	0.0073	2.227	42.16
		椭球扰动	59	1038	3	2.541	0.0076	3.044	
	0.04	均匀范数球扰动	45	1055	0	0.874	0.0073	0.874	31.11
		椭球扰动	31	1069	0	1.034	0.0077	1.034	
	0.05	均匀范数球扰动	28	1072	0	1.268	0.0071	1.268	57.14
		椭球扰动	12	1088	0	2.807	0.0079	2.807	
	0.06	均匀范数球扰动	12	1088	0	1.955	0.0074	1.955	100.00
		椭球扰动	0	1100	0	0.000	0.0082	0.000	
SDD-L	0.001	均匀范数球扰动	1089	11	0	0.130	0.0083	0.130	0.09
		椭球扰动	1088	12	0	0.128	0.0081	0.128	
	0.003	均匀范数球扰动	1056	44	0	0.130	0.0220	0.130	12.32
		椭球扰动	1042	57	1	0.141	0.0195	0.143	
	0.005	均匀范数球扰动	960	140	0	0.182	0.0120	0.182	9.07
		椭球扰动	872	227	1	0.242	0.0107	0.249	
	0.007	均匀范数球扰动	774	323	3	0.285	0.0093	0.309	17.7
		椭球扰动	637	461	2	0.388	0.0091	0.428	
	0.01	均匀范数球扰动	495	604	1	0.529	0.0079	0.565	23.84
		椭球扰动	377	720	3	0.661	0.0081	0.905	
	0.02	均匀范数球扰动	155	945	0	1.167	0.0772	1.167	27.10
		椭球扰动	113	987	0	1.555	0.0073	1.555	
	0.03	均匀范数球扰动	89	1011	0	0.754	0.0073	0.754	7.87
		椭球扰动	82	1018	0	5.326	0.0073	5.326	
	0.04	均匀范数球扰动	60	1039	1	2.732	0.0071	2.945	36.67
		椭球扰动	38	1060	2	11.338	0.0072	22.790	
	0.05	均匀范数球扰动	27	1073	0	4.953	0.0072	4.953	74.07
		椭球扰动	7	1093	0	5.766	0.0071	5.766	
	0.06	均匀范数球扰动	7	1093	0	6.170	0.0073	6.170	100.00
		椭球扰动	0	1100	0	0.000	0.0071	0.000	

在扰动预算 ε 设定在较低的区间 ($\varepsilon \leqslant 0.003$) 时, 两个模型上验证结果返回 UNSAT (安全) 的样本占比均比较高 (达到 87.3% 以上), 同时阴性非一致率均维持在相对较低的水平, 尤其在 $\varepsilon = 0.001$ 时, 该比率不足 1%. 在如此小的扰动范围内, 无论是均匀范数球邻域还是椭球邻域, 其体积都非常有限, 几乎不包含对抗样本, 但是仍旧存在假阴性的现象, 这也揭示了传统球形验证邻域方法的固有局限性. 随着扰动预算的增加, 阴性非一致率总体上呈现非线性增长的趋势. 当 $\varepsilon = 0.007$ 时, SDD-S 上的验证结果为 UNSAT (安全) 的数据占比为 48.81%, 这意味着接近一半的数据通过了安全性验证, 然而此时阴性非一致率达到 21.09%, 这表明约 1/5 的所谓“安全”的样本实际上存在漏检风险. 类似地, SDD-L 在 $\varepsilon = 0.01$ 时表现出更显著的验证偏差, 虽然在球形邻域中的验证结果返回 UNSAT 的比例为 45%, 但对应的阴性非一致率已达到 23.84%. 与球形扰动域相比, 椭球扰动域在处理非关键特征时扩大了扰动半径, 从而扩展了验证范围, 也更容易发现潜在的对抗样本.

随着扰动预算的进一步增长, 当 $\varepsilon = 0.05$ 时, SDD-S 与 SDD-L 上的阴性非一致率分别攀升至 57.14% 与 74.07%. 尽管此时假阴性率非常高, 但值得注意的是即使是在球形扰动域内, 验证结果为 UNSAT 的数据比例也已不足 3%, 而在椭球邻域内这一比例则更低. 当 $\varepsilon = 0.06$ 时, SDD-S 与 SDD-L 在球形邻域内几乎不存在被验证结果

为安全的数据点, 而椭球邻域内所有数据点均返回 SAT (不安全) 的验证结果, 导致此时阴性非一致率达到 100%. 我们认为随着扰动预算的增长, 对抗样本出现的可能性增加, 同时导致原本的 UNSAT 安全保证变得越来越容易被额外扩展的验证区域所破坏.

4.1.3 S^{∞} 鲁棒区域

对于一个数据点 $\boldsymbol{x}^* \in \mathbb{R}^{d_0}$, 我们定义在无穷范数下的集合 $S_{\varepsilon}^{p}(\boldsymbol{x}^*) = \{\boldsymbol{x}^* + \boldsymbol{\delta} \mid \|\boldsymbol{\Omega}\boldsymbol{\delta}\|_p \leqslant \varepsilon\}$. 在给定扰动 ε、线性变换矩阵 $\boldsymbol{\Omega}$ 以及范数 $p(0,1,2,\infty)$ 的情况下, $S_{\varepsilon}^{p}(x^*)$ 是围绕着 $\boldsymbol{x}^*$ 的一个邻域. 当 $\boldsymbol{\Omega} = \mathbf{I}$ 时, 这是一个均匀 L_p 范数球邻域; 当 $\boldsymbol{\Omega} = \mathbf{D}$ 时, 则形成一个椭球邻域. 设定矩阵 $\boldsymbol{\Omega}$ 的前提下, 我们定义 S^{∞} 鲁棒区域为验证安全 (无对抗样本存在) 的邻域 $S_{\varepsilon}^{\infty}(\boldsymbol{x}^*)$. 给定扰动预算 ε 的情况下, 均匀范数球区域的体积记为 $\sqrt[d_0]{\varepsilon^{d_0}} = \varepsilon$; 对于椭球区域, 给定扰动预算 ε, 特征会根据其互信息值被分配到不同的扰动值, 如第 3.2.1 节中所述, 对于特征 $\boldsymbol{x}_i$ 它被施加的扰动值实际上是 $\varepsilon_i = \dfrac{\varepsilon}{\mathbf{D}_{i,i}}$, 此时椭球区域的体积可以表示为 $\sqrt[d_0]{\prod_{i=1}^{d_0}\varepsilon_i}$. 我们的目标是找到具有最大体积的 S^{∞} 鲁棒区域.

SDD-S 以及 SDD-L 均为归一化训练的网络, 采用下述方法计算均匀扰动和椭球扰动下最大 S^{∞} 鲁棒区域的体积: 选择 $\varepsilon_1 = \dfrac{1}{2}$ 为初始值, 在第 i 次迭代的过程中, 调用验证引擎验证网络关于 ε_i 的扰动邻域是否安全, 如果返回 UNSAT, 则将 $\varepsilon_{i+1} = \varepsilon_i + \left(\dfrac{1}{2}\right)^{i+1}$ 作为下一个取值; 否则, 若返回 SAT 或者 UNK, 则将 $\varepsilon_{i+1} = \varepsilon_i - \left(\dfrac{1}{2}\right)^{i+1}$ 作为下一次的取值. 重复上述的迭代过程直到找到满足精度要求的 ε_i, 此时我们便能计算出最大的鲁棒区域的体积.

我们从测试数据集中随机选取了 3 300 个数据点, 确保类别标签均匀分布, 并分别计算了这些数据在 SDD-S 和 SDD-L 网络中, 均匀扰动和椭球扰动条件下最大鲁棒区域体积的平均值. 为了提高计算效率, 我们选择了 CROWN[32]作为迭代过程中的验证工具. CROWN 采用线性松弛方法对网络输出的可达集进行上近似处理, 其结果仅能返回 UNSAT(安全) 或 UNK, 这意味着它只能保证验证结果的可靠性而非完备性. 因此, 我们所计算出的邻域体积通常小于实际的最大鲁棒区域体积.

实验的结果如表 2 所示. 如我们在第 3.2.1 节中讨论的, 椭球鲁棒性中的参数 k 作为归一化系数被引入, 其作用在于确保对不同特征所施加的扰动的最大值和最小值之比不超过 k 倍. 当 $k = 1$ 时, 验证区域实质上等同于一个均匀范数球邻域, 通过调整 k 的值, 我们构造了 4 种不同形状的椭球区域. 实验结果显示, 在 SDD-S 和 SDD-L 网络中, 4 种取值下的 k 均能界定出相比均匀范数球有更大体积的安全区域. 在 SDD-S 中, 当 $k = 2.5$ 时, 鲁棒区域体积达到最大; 在 SDD-L 中, 当 $k = 2$ 时, 鲁棒区域体积达到最大. 因此我们认为, 与均匀范数球鲁棒性相比, 椭球鲁棒性因为考虑到了各特征间鲁棒性的差异, 从而能够提供更为严格的安全保障.

表 2 均匀范数球扰动和椭球扰动下所能界定的最大鲁棒区域的平均体积

模型	扰动类型	平均体积	比值
SDD-S	均匀范数球扰动	0.007 215	1.000
	椭球扰动 (k=1.5)	0.008 110	1.124
	椭球扰动 (k=2)	0.008 231	1.141
	椭球扰动 (k=2.5)	0.008 374	1.161
	椭球扰动 (k=3)	0.008 346	1.157
SDD-L	均匀范数球扰动	0.007 702	1.000
	椭球扰动 (k=1.5)	0.008 396	1.090
	椭球扰动 (k=2)	0.009 059	1.176
	椭球扰动 (k=2.5)	0.008 606	1.117
	椭球扰动 (k=3)	0.008 169	1.061

4.2 掩码局部鲁棒性验证

4.2.1 数据集介绍

在掩码局部鲁棒性分支上, 我们选择 EMBER (endgame malware benchmark for research) 数据集[39]进行实验. EMBER 是一个专为 Windows 恶意软件检测研究设计的开源数据集, 它从 Windows PE (portable executable) 文件

中提取了 2 831 维结构化特征, 具体可分为两类.

(1) 解析特征

a) 一般文件信息: 文件的虚拟大小, 导入/导出函数的个数, 文件是否包含调试段、线程本地存储、资源、重定位或签名以及符号数量.

b) 头信息: 头中的时间戳、目标机器、图像特征列表、目标子系统、DLL 特征、文件的幻数标志、系统和子系统的版本、映像版本、链接器版本以及代码、标头和提交大小.

c) 导入函数: 导入地址表提取的函数.

d) 导出函数: 导出函数列表.

e) 节/段信息: 每个部分的属性, 包括名称、大小、熵、虚拟大小以及表示节特性的字符串列表.

(2) 格式无关特征

a) 字节直方图: 字节直方图包含 256 个整数值, 表示文件中每个字节值的计数.

b) 字节熵直方图: 熵 H 和字节值 X 的联合分布 $p(H, X)$ 的量化和归一化.

c) 字符串信息: 字符串的数量及其平均长度; 这些字符串中可打印字符的直方图; 所有可打印字符串中字符的熵.

此外, 每个样本还包含一个二元目标变量, 用于判定可移植的可执行文件是良性软件还是恶意软件.

在数据规模上, EMBER 数据集包含了 60 万条有效的带标注的训练样本以及 20 万个测试样本. 我们根据训练样本进行网络的标准化训练形成模型 EMBER-L 并用于后续实验, 网络具体结构为 2831×500×500×500×1.

4.2.2 阳性非一致率

为了更直观展示 L_p 范数球鲁棒性验证中的假阳性问题, 我们也对阳性非一致率进行了定义.

定义 11. 阳性非一致率. 给定 n 组对比验证实验结果 $(A_1, B_1), \ldots, (A_n, B_n)$, 其中 A_i 是对照组, B_i 是实验组, 且 $A_i, B_i \in \{\text{SAT, UNSAT, UNK}\}$, 我们用阳性非一致率 (SAT non-consistency rate, *S-NCR*) 描述在对照实验中对照组结果为 SAT 且实验组结果不为 UNK 的情况下实验组结果为 UNSAT 的比例. 阳性非一致率定义如下:

$$S\text{-}NCR = \frac{card(i \mid A_i = \text{SAT} \wedge B_i = \text{UNSAT})}{card(i \mid A_i = \text{SAT} \wedge B_i \neq \text{UNK})} \times 100\% \tag{40}$$

面对一个恶意软件样本, 攻击者的核心目标是操纵深度神经网络, 使其错误地将该样本识别为良性软件. 为了实现这一目标, 攻击者通常采用问题空间攻击策略, 即直接对恶意软件的二进制文件进行修改, 而非针对提取的特征进行操作. 这些在问题空间实施的攻击手段, 在保持恶意软件核心功能不变的前提下, 仅向二进制文件中添加各种字节信息. 因此, 这些修改仅对提取特征中的字节直方图、字节熵直方图和节信息这 3 个特征组产生影响[24]. 基于这一专家知识, 我们确定了仅这 3 个特征组是可修改的, 其中这 3 个特征组的特征总数为 767 维, 而其他特征则需要保持不变. 这一设定使我们能够构建相应的掩码局部鲁棒性验证问题.

与假阴性问题的处理方式相似, 对于每个数据点 $\boldsymbol{x}^*$, 我们构建了基于 L_p 范数球的鲁棒性验证问题作为对照组. 在实验组的设置上, 我们仅在允许变动的 3 个特征组上施加扰动, 构建掩码局部鲁棒性验证问题作为实验组.

鉴于攻击者的主要目的是使恶意软件逃避检测, 我们在数据集中随机抽取了 100 个标签为恶意软件的样本, 而未选取良性软件样本, 并针对 L_p 范数球鲁棒性验证中假阳性的情况进行了实验, 实验结果如表 3 所示.

根据表 3 所示, 随着扰动幅度的增加, 阳性非一致率也呈现出增长趋势, 在扰动值达到 0.02 时达到峰值 88.68%. 我们认为此现象主要归因于: 在均匀扰动下, 随着扰动预算 ε 的增大, 对抗样本出现的概率显著提升; 而在掩码扰动条件下, 尽管概率同样上升, 但由于扰动仅发生在约 1/3 的维度上, 其增长速度相对缓慢. 从表 3 可以观察到, 在扰动预算 ε 从 0.01 增加到 0.014 的过程中, 均匀扰动下验证为 SAT (不安全) 的样本增加了 15 个, 但在掩码扰动下未见增长. 同样地, 在均匀扰动条件下, 当扰动预算 ε 由 0.006 逐渐增加至 0.024 时, 对所有 2 831 维特征进行扰动, 被验证为 SAT 的样本数从 5 显著上升至 61; 相反, 在采用掩码扰动策略时, 即仅对实际可能被修改的 767 维特征进行扰动, 被验证为 SAT 的样本数量仅从 0 略微增至 7. 这表明在相同的扰动水平下, 对全部的 2 831 维特征施加扰动比仅对 767 维特征施加扰动更容易揭示对抗样本的存在, 但这些样本大多数是不符合现实情况的, 从而导致较高的阳性非一致率. 在 $\varepsilon = 0.024$ 的情况下, 存在 52 个数据点在均匀扰动模式下能发现对抗样本,

但是在掩码扰动模式下, 它们事实上是被验证为 UNSAT (安全) 的, 这也进一步证实了均匀范数球验证模式下的假阳性问题. 此外, 我们注意到使用掩码扰动方法时, 由于输入维度的减少, 导致验证结果为 UNK (验证超时) 的比例大幅下降.

表 3　当 ε 不同时, EMBER-L 模型在均匀范数球扰动和掩码扰动下的验证结果 (EMBER 数据集)

扰动预算 (ε)	扰动类型	UNSAT	SAT	UNK	平均验证时间 (s)			*S-NCR* (%)
					返回UNSAT	返回SAT	返回UNSAT或UNK	
0.006	均匀范数球扰动	94	5	1	1.242	0.687	4.558	40.00
	掩码扰动	97	3	0	1.255	0.548	1.252	
0.008	均匀范数球扰动	86	10	4	3.825	0.628	19.674	60.00
	掩码扰动	96	4	0	1.269	0.545	1.269	
0.01	均匀范数球扰动	75	16	9	60.712	0.605	92.818	68.75
	掩码扰动	95	5	0	1.254	0.566	1.254	
0.012	均匀范数球扰动	51	19	30	1.997	0.595	134.752	73.68
	掩码扰动	95	5	0	1.256	0.570	1.256	
0.014	均匀范数球扰动	42	31	27	5.442	0.587	144.363	83.87
	掩码扰动	95	5	0	1.473	0.586	1.473	
0.016	均匀范数球扰动	30	43	27	12.771	0.549	177.458	86.05
	掩码扰动	94	6	0	1.344	0.613	1.344	
0.018	均匀范数球扰动	18	47	35	15.459	0.551	243.313	87.23
	掩码扰动	94	6	0	1.513	0.607	1.513	
0.02	均匀范数球扰动	15	53	32	17.692	0.553	278.664	88.68
	掩码扰动	92	6	2	3.321	0.591	10.910	
0.022	均匀范数球扰动	11	54	35	17.664	0.553	278.644	87.37
	掩码扰动	86	7	7	1.942	0.524	28.891	
0.024	均匀范数球扰动	1	61	38	1.269	0.545	351.421	85.25
	掩码扰动	84	9	7	3.318	0.532	30.098	

4.3　马氏距离鲁棒性验证

4.3.1　数据集介绍

在马氏距离鲁棒性分支上, 我们选择源自 UCI 的 Spambase 数据集[38]进行实验. 该数据集用于检测一封邮件属于垃圾邮件还是正常邮件. Spambase 的输入特征共有 57 维, 均为连续数值类型, 主要分为 3 类.

(1) 词频统计特征: 统计特定的单词 (例如 make, address) 在邮件中出现的频率, 总共包含了 48 个特征.

(2) 统计量特征: 包括连续大写字母序列的平均长度, 最长大写字母的序列长度, 大写字母总数量, 总共包含了 3 个特征.

(3) 特殊符号特征: 统计邮件中特殊字符出现的频率 (例如!, $, #).

此外, 每个样本还包含一个二元目标变量, 用于判定该邮件属于正常邮件还是垃圾邮件. 该数据集共有 4 601 条数据, 其中垃圾邮件 1 813 条, 占比 39.4%, 我们按 8:2 的比例划分了训练集和测试集, 并按照 57×100×100×100×1 的网络结构进行标准化训练, 得到模型 Spam-S.

4.3.2　马氏距离鲁棒性验证

在第 3.1 节中我们定义了数据的一致性, 用于量化对抗样本与原始数据分布的偏离程度. 为实现精确评估, 首先需计算数据分布的概率密度函数 $p(\boldsymbol{x})$, 我们采用标准化流 (normalization flow)[40]方法对概率密度进行建模. 一致性阈值 γ 的选择基于训练集上的预分析确定: 计算全体训练样本的概率密度后, 以下 1% 分位点作为一致性阈值 γ. 这意味着训练集中 99% 样本被视为正常数据, 而概率密度低于阈值的 1% 样本则标记为低概率异常样本. 在对抗样本的一致性检测中, 若其概率密度高于一致性阈值则判定为正常样本, 反之则视为分布偏离实例.

与掩码局部鲁棒性验证直接排除涉及不可变特征修改的对抗样本不同, 基于马氏距离的鲁棒性验证需结合概

率密度阈值进行进一步判定. 为准确评估验证系统的可靠性, 我们引入概率密度约束条件对第 4.2.2 节的阳性非一致率定义进行重构. 具体而言, 当对照组 A_i 返回 SAT 的验证结果时, 记所得对抗样本为 $\boldsymbol{a}_i$, 若其概率密度值 $p(\boldsymbol{a}_i)<\gamma$, 则表明该对抗样本偏离了正常数据分布. 在此条件下, 若实验组 B_i 返回 UNSAT 的验证结果, 即可判定对照组 A_i 产生了假阳性误报. 基于上述的讨论, 我们给出马氏距离鲁棒性下阳性非一致率的定义:

$$S\text{-}NCR_{\mathrm{M}}=\frac{card(i\mid A_i=\mathrm{SAT}\wedge B_i=\mathrm{UNSAT}\wedge p(\boldsymbol{a}_i)<\gamma)}{card(i\mid A_i=\mathrm{SAT}\wedge B_i\neq\mathrm{UNK})}\times 100\% \tag{41}$$

对于每个数据点 $\boldsymbol{x}^*$, 我们构建了基于 L_p 范数球的鲁棒性验证问题作为对照组. 在实验组的设置上, 如图 3(c) 所示, 我们构建的马氏距离扰动域 $\eta=\{\boldsymbol{x}\mid\|\mathbf{U}(\boldsymbol{x}-\boldsymbol{x}^*)\|_2\leqslant\varepsilon\}$ 本质上是一个以 $\boldsymbol{x}^*$ 为中心的经过仿射变换生成的高维超椭球体, 其各主轴长度可量化为 $\frac{\varepsilon}{\sigma_i}$, 其中 σ_i 表示矩阵 $\mathbf{U}$ 的奇异值. 为确保实验组与对照组具有可比性, 我们通过参数约束条件使椭球体的最大半轴长度与 L_p 范数球的半径严格等长, 即保证实验组中构建的马氏距离扰动域完全内切于对照组中的 L_p 范数球.

随机选取测试集 400 个样本 (正常邮件与垃圾邮件各 200 个) 进行鲁棒性验证. 本文构建的马氏距离扰动域是一个完全内切于 L_p 范数球的椭球区域, 在确定了扰动预算 ε (L_p 范数球的半径) 后, 对验证过程中返回的两类对抗样本进行分类统计.

(1) 内蕴对抗样本: 处于马氏距离鲁棒性所定义的椭球体内的对抗样本.

(2) 外延对抗样本: 处于 L_p 范数球内但在椭球体外的对抗样本.

表 4 呈现了不同扰动预算下两类对抗样本违反数据一致性原则的统计结果. 当扰动预算较小时, 两类对抗样本的非一致性比例均维持在相对较低水平, 表明此时扰动域内的对抗样本能较好保持原始数据分布特性. 随着扰动预算的增加, 两个区域内的对抗样本不符合一致性原则的比例均呈现上升趋势, 这表明无论是内蕴对抗样本还是外延对抗样本都在逐渐偏离真实的数据分布. 然而, 外延对抗样本的非一致性比例始终高于内蕴对抗样本, 这揭示了外延对抗样本中存在更为显著的分布偏移问题.

表 4　内蕴对抗样本和外延对抗样本的数据非一致性的比例 (%)

扰动预算 (ε)	0.05	0.07	0.1	0.15	0.2	0.25	0.3	0.35	0.4
内蕴对抗样本	0	0	0	29.4	53.7	59.2	67.3	63.8	69.1
外延对抗样本	0	41.2	57.6	65.1	82.5	81.5	85.7	94	95.8

本研究还针对 L_p 范数球鲁棒性验证框架中的假阳性情况进行了实验, 其结果汇总于表 5. 当扰动预算 $\varepsilon\geqslant 0.1$ 时, 实验结果呈现出较高的阳性非一致率, 这表明传统 L_p 范数球验证模式会将不符合数据分布特征的对抗样本错误归类为有效对抗样本, 进而导致对神经网络安全性的误判.

表 5　均匀范数球扰动和马氏距离扰动下的验证结果 (Spambase 数据集)

扰动预算 (ε)	扰动类型	UNSAT	SAT	UNK	平均验证时间 (s)			$S\text{-}NCR_{\mathrm{M}}$ (%)
					返回UNSAT	返回SAT	返回UNSAT或UNK	
0.05	均匀范数球扰动	392	8	0	0.026	0.0089	0.026	0
	马氏距离扰动	399	1	0	0.018	0.0089	0.018	
0.1	均匀范数球扰动	360	40	0	0.039	0.0088	0.039	47.50
	马氏距离扰动	392	7	1	0.026	0.0065	0.027	
0.15	均匀范数球扰动	339	60	1	0.135	0.0093	0.224	46.67
	马氏距离扰动	382	17	1	0.038	0.0090	0.047	
0.2	均匀范数球扰动	318	81	1	0.310	0.0097	0.404	40.74
	马氏距离扰动	359	41	0	0.026	0.0087	0.026	
0.25	均匀范数球扰动	284	114	2	0.351	0.0094	0.557	46.49
	马氏距离扰动	351	49	0	0.033	0.0090	0.033	
0.3	均匀范数球扰动	259	139	2	0.492	0.0081	0.718	51.80
	马氏距离扰动	345	55	0	0.082	0.0093	0.082	

表 5　均匀范数球扰动和马氏距离扰动下的验证结果 (Spambase 数据集)(续)

扰动预算 (ε)	扰动类型	UNSAT	SAT	UNK	平均验证时间 (s)			$S\text{-}NCR_{\mathrm{M}}$ (%)
					返回UNSAT	返回SAT	返回UNSAT或UNK	
0.35	均匀范数球扰动	243	153	4	1.093	0.0089	1.561	51.63
	马氏距离扰动	330	69	1	0.166	0.0091	0.172	
0.4	均匀范数球扰动	219	176	5	0.875	0.0091	1.525	51.70
	马氏距离扰动	319	81	0	0.116	0.0092	0.116	

距离数据集中心 (均值) 的马氏距离能够有效衡量数据点偏离整体分布的程度, 其值越大表明该数据与分布的偏离程度越高. 设训练数据在 d_0 维特征空间的均值向量 $\boldsymbol{\mu} \in \mathbb{R}^{d_0}$ 为分布中心, 定义为 $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$, 其中 n 为训练样本数量, $\boldsymbol{x}_i$ 为第 i 个训练样本的向量表征. 我们统计了 L_p 均匀范数球扰动域以及马氏距离扰动域内的对抗样本到均值 $\boldsymbol{\mu}$ 的平均马氏距离, 结果呈现在表 6 和图 4 中. 随着扰动预算的增加, 对抗样本的马氏距离也随之增大. 在相同扰动预算下, 马氏距离扰动域中的对抗样本的平均马氏距离显著低于 L_p 范数球扰动域内的对抗样本. 这进一步表明, 基于马氏距离扰动域验证得到的对抗样本具有更高的数据一致性, 更好地保持了原始数据的分布特性.

表 6　不同扰动域内对抗样本的平均马氏距离

扰动预算 (ε)	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6
马氏距离扰动域	3.77	3.80	3.90	3.99	3.95	3.97	3.85	3.95	4.03	4.12	4.12
均匀范数球扰动域	4.15	4.11	4.12	4.32	4.42	4.53	4.64	4.67	4.77	4.94	4.93

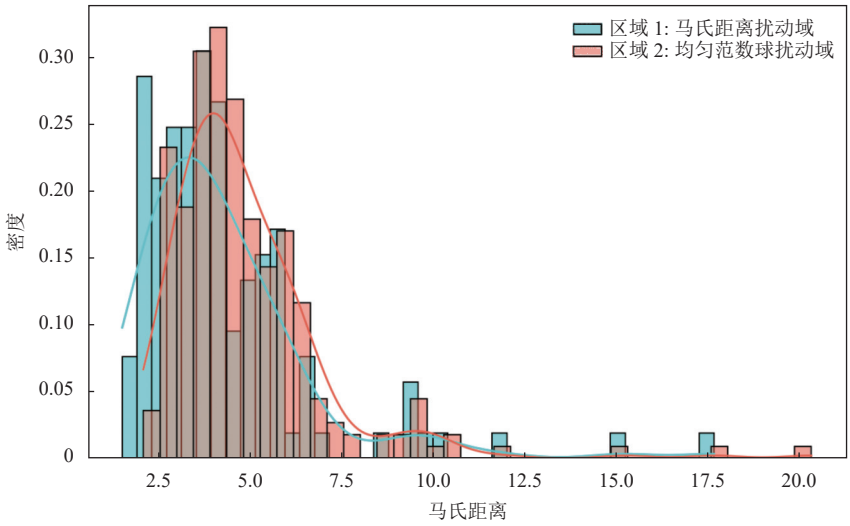

图 4　扰动预算 $\varepsilon = 0.6$ 下, 马氏距离扰动域和均匀范数球扰动域内对抗样本的平均马氏距离的密度直方图

4.4　NNV4RADAP 算法的性能

为了回答 Q2, 我们分别在椭球鲁棒性验证 (SDD-S/SDD-L)、掩码局部鲁棒性验证 (EMBER-L) 和马氏距离鲁棒性验证 (Spam-S) 这 4 个模型上, 将实验组与对照组中基准的均匀 L_p 范数球验证方法进行了时间性能对比实验. 表 1、表 3、表 5 中的平均验证时间数据表明, 在通过 PGD 攻击寻找对抗样本的时间开销方面, 实验组与对照组基本持平, 未产生显著差异. 因此我们主要对比在限定时间内验证器返回 UNSAT 或 UNK 结果时的平均验证耗时.

图 5(a)(b) 显示, 在椭球鲁棒性验证场景中, 实验组的平均验证时间较对照组有一定的增加. 这主要归因于椭球验证域相较 L_p 范数球具有更大的空间覆盖范围, 从而在分支定界过程中产生更多子问题, 最终导致了验证时间的增加. 在掩码扰动场景中, 如图 5(c) 显示, 实验组的平均验证时间较基准方法大幅降低, 这得益于验证特征空间

的维度压缩——输入特征从原始 2 831 维降至 767 维, 这使得在排除非法扰动的同时, 也使得分支定界过程中需要处理的决策变量减少进而提升了验证效率. 类似地, 图 5(d) 中的马氏距离验证展现出一定的时间效率提升, 因其验证域本质上是原始 L_p 范数球的一个紧致内切椭球子空间.

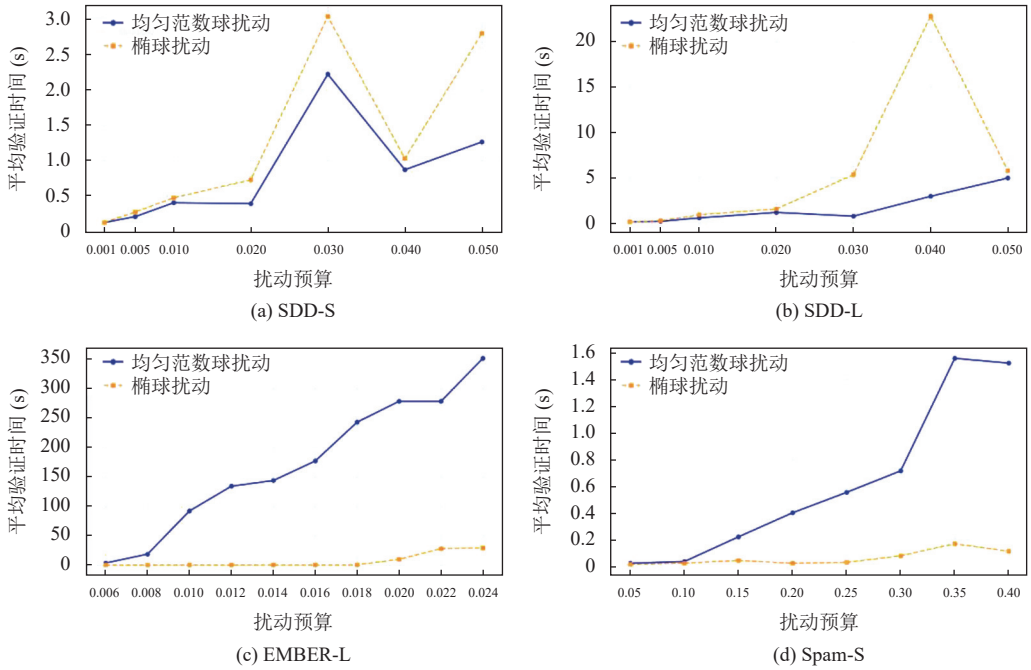

图 5 实验组与对照组验证结果返回 UNSAT/UNK 所用平均验证时间对比

从实验结果上分析, 相对于均匀 L_p 范数球鲁棒性, 椭球鲁棒性的验证时间呈现可控范围内的适度增长, 而掩码局部鲁棒性与马氏距离鲁棒性在验证效率上则表现出显著的提升. NNV4RADAP 算法在处理等价验证问题时, 并未改变网络原有的隐藏层与输出层结构, 因此我们认为验证时间的变化主要归因于待验证输入空间的变化. 具体而言, 在椭球鲁棒性的验证过程中, 由于我们扩展了输入空间, 使其从一个均匀范数球扩展为包含该范数球的椭球, 这直接导致了验证所需时间的增加. 相反, 在掩码局部鲁棒性实验中, 仅对允许变更的特征施加扰动, 以及在马氏距离鲁棒性实验中, 选择内切于原范数球的椭球作为验证空间, 这两种方法均有效地减小了实际需要考虑的输入空间大小, 从而显著降低了验证的时间成本.

5 总结与未来工作

本文关注面向领域自适应扰动的神经网络鲁棒性验证问题, 提出了 3 种基于领域数据特性建模的神经网络鲁棒性定义及验证算法 NNV4RADAP, 基于 3 种数据分布假设对领域特性矩阵进行计算, 将面向领域自适应扰动的神经网络鲁棒性验证转换为等价的 L_p 范数球鲁棒性验证问题. 同时, 基于实验证明了 NNV4RADAP 算法的有效性, 及其较低的时间性能损耗.

我们未来希望能对领域自适应扰动的定义及验证算法进行更深的拓展, 尝试建模数据特征之间的非线性相关性, 同时对非数据特征维度上的鲁棒尺度进行刻画, 基于更广义的数据分布假设在一些更具实际意义的子问题上进行鲁棒性的定义建模与验证算法的开发.

References

[1] Huber PJ. Robust Statistics. New York: Wiley, 1981. [doi: 10.1002/0471725250]

[2] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the 3rd Int'l Conf. on Learning

Representations. San Diego, 2015. 1–11.

[3] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018. 1–23.

[4] Song Y, Shu R, Kushman N, Ermon S. Constructing unrestricted adversarial examples with generative models. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 8322–8333.

[5] Zhang JF, Xu XL, Han B, Niu G, Cui LZ, Sugiyama M, Kankanhalli MS. Attacks which do not kill training make adversarial learning stronger. In: Proc. of the 37th Int'l Conf. on Machine Learning. JMLR.org, 2020. 1046.

[6] Shafahi A, Najibi M, Ghiasi A, Xu Z, Dickerson J, Studer C, Davis LS, Taylor G, Goldstein T. Adversarial training for free! In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 302.

[7] Zhang DH, Zhang TY, Lu YP, Zhu ZX, Dong B. You only propagate once: Accelerating adversarial training via maximal principle. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 21.

[8] Zhang HY, Yu YD, Jiao JT, Xing E, El Ghaoui L, Jordan M. Theoretically principled trade-off between robustness and accuracy. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 7472–7482.

[9] Wang YS, Zou DF, Yi JF, Bailey J, Ma XJ, Gu QQ. Improving adversarial robustness requires revisiting misclassified examples. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020. 1–14.

[10] Katz G, Barrett C, Dill DL, Julian K, Kochenderfer MJ. Reluplex: An efficient SMT solver for verifying deep neural networks. In: Proc. of the 29th Int'l Conf. on Computer Aided Verification. Heidelberg: Springer, 2017. 97–117. [doi: 10.1007/978-3-319-63387-9_5]

[11] Wang SQ, Pei KX, Whitehouse J, Yang JF, Jana S. Formal security analysis of neural networks using symbolic intervals. In: Proc. of the 27th USENIX Security Symp. Baltimore: USENIX, 2018. 1599–1614.

[12] Katz G, Huang DA, Ibeling D, Julian K, Lazarus C, Lim R, Shah P, Thakoor S, Wu HZ, Zeljić A, Dill DL, Kochenderfer MJ, Barrett CW. The Marabou framework for verification and analysis of deep neural networks. In: Proc. of the 31st Int'l Conf. on Computer Aided Verification. New York: Springer, 2019. 443–452. [doi: 10.1007/978-3-030-25540-4_26]

[13] Wang SQ, Zhang H, Xu KD, Lin X, Jana S, Hsieh CJ, Kolter Z. Beta-CROWN: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. Curran Associates Inc., 2021. 2289.

[14] Elboher YY, Gottschlich J, Katz G. An abstraction-based framework for neural network verification. In: Proc. of the 31st Int'l Conf. on Computer Aided Verification. Los Angeles: Springer, 2019. 43–65. [doi: 10.1007/978-3-030-53288-8_3]

[15] Jin P, Tian JX, Zhi DP, Wen XJ, Zhang M. TRAINIFY: A CEGAR-driven training and verification framework for safe deep reinforcement learning. In: Proc. of the 34th Int'l Conf. on Computer Aided Verification. Haifa: Springer, 2022. 193–218. [doi: 10.1007/978-3-031-13185-1_10]

[16] Liu WW, Song F, Zhang THR, Wang J. Verifying ReLU neural networks from a model checking perspective. Journal of Computer Science and Technology, 2020, 35(6): 1365–1381. [doi: 10.1007/s11390-020-0546-7]

[17] Ferrag MA, Maglaras L, Moschoyiannis S, Janicke H. Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. Journal of Information Security and Applications, 2020, 50: 102419. [doi: 10.1016/j.jisa.2019.102419]

[18] Zhang L, Wang XS, Yang D, Sanford T, Harmon S, Turkbey B, Wood BJ, Roth H, Myronenko A, Xu DG, Xu ZY. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Trans. on Medical Imaging, 2020, 39(7): 2531–2540. [doi: 10.1109/TMI.2020.2973595]

[19] Liu NH, Yang HX, Hu X. Adversarial detection with model interpretation. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. London: ACM, 2018. 1803–1811. [doi: 10.1145/3219819.3220027]

[20] Tsipras D, Santurkar S, Engstrom L, Turner A, Madry A. There is no free lunch in adversarial robustness (but there are unexpected benefits). arXiv:1805.12152, 2018.

[21] Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B, Madry A. Adversarial examples are not bugs, they are features. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 12.

[22] Liu C, Tomioka R, Cevher V. On certifying non-uniform bounds against adversarial attacks. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 4072–4081.

[23] Li DQ, Li QM. Adversarial deep ensemble: Evasion attacks and defenses for malware detection. IEEE Trans. on Information Forensics and Security, 2020, 15: 3886–3900. [doi: 10.1109/TIFS.2020.3003571]

[24] Erdemir E, Bickford J, Melis L, Aydöre S. Adversarial robustness with non-uniform perturbations. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. Curran Associates Inc., 2021. 1464.

[25] Lomuscio A, Maganti L. An approach to reachability analysis for feed-forward ReLU neural networks. arXiv:1706.07351, 2017.

[26] Dutta S, Jha S, Sankaranarayanan S, Tiwari A. Output range analysis for deep feedforward neural networks. In: Proc. of the 10th Int'l Symp. on NASA Formal Methods. Newport News: Springer, 2018. 121–138. [doi: 10.1007/978-3-319-77935-5_9]

[27] Xiang WM, Tran HD, Johnson TT. Output reachable set estimation and verification for multilayer neural networks. IEEE Trans. on Neural Networks and Learning Systems, 2018, 29(11): 7777–5783. [doi: 10.1109/TNNLS.2018.2808470]

[28] Gehr T, Mirman M, Drachsler-Cohen D, Tsankov P, Chaudhuri S, Vechev M. AI2: Safety and robustness certification of neural networks with abstract interpretation. In: Proc. of 2018 IEEE Symp. on Security and Privacy. San Francisco: IEEE, 2018. 3–18. [doi: 10.1109/SP.2018.00058]

[29] Mirman M, Gehr T, Vechev M. Differentiable abstract interpretation for provably robust neural networks. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 3575–3583.

[30] Singh G, Gehr T, Püschel M, Vechev M. An abstract domain for certifying neural networks. Proc. of the ACM on Programming Languages, 2019, 3(POPL): 41. [doi: 10.1145/3290354]

[31] Wang SQ, Pei KX, Whitehouse J, Yang JF, Jana S. Efficient formal safety analysis of neural networks. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 6369–6379.

[32] Zhang H, Weng TS, Chen PY, Hsieh CJ, Daniel L. Efficient neural network robustness certification with general activation functions. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 4944–4953.

[33] Salman H, Yang G, Zhang H, Hsieh CJ, Zhang PC. A convex relaxation barrier to tight robustness verification of neural networks. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2019. 882.

[34] Xu KD, Shi ZX, Zhang H, Wang YH, Chang KW, Huang ML, Kailkhura B, Lin X, Hsieh CJ. Automatic perturbation analysis for scalable certified robustness and beyond. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 96.

[35] Xu KD, Zhang H, Wang SQ, Wang YH, Jana S, Lin X, Hsieh CJ. Fast and complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021. 1–15.

[36] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ, Fergus R. Intriguing properties of neural networks. In: Proc. of the 2nd Int'l Conf. on Learning Representations. Banff: OpenReview.net, 2014. 1–10.

[37] Huang XW, Kwiatkowska M, Wang S, Wu M. Safety verification of deep neural networks. In: Proc. of the 29th Int'l Conf. on Computer Aided Verification. Heidelberg: Springer, 2017. 3–29. [doi: 10.1007/978-3-319-63387-9_1]

[38] Merz CJ, Murphy P. UCI repository of machine learning databases. 1996. https://archive.ics.uci.edu/datasets

[39] Anderson HS, Roth P. EMBER: An open dataset for training static pe malware machine learning models. arXiv:1804.04637, 2018.

[40] Rezende D, Mohamed S. Variational inference with normalizing flows. In: Proc. of the 32nd Int'l Conf. on Machine Learning. Lille: PMLR, 2015. 1530–1538.

作者简介

王麒扬, 硕士生, 主要研究领域为人工智能安全, 神经网络验证.

李璟旸, 博士生, CCF 学生会员, 主要研究领域为人工智能安全, 神经网络验证.

李国强, 博士, 副教授, CCF 高级会员, 主要研究领域为形式化验证, 程序语言理论.