DOI:10.13718/j.cnki.xsxb.2019.03.015

基于线性迭代的分类器最小扰动评估方法①

邹 瑛

四川司法警官职业学院 司法信息管理系,四川 德阳 618000

摘要:当数据集包含对抗性扰动样本时,其分类结构缺乏稳定性,传统的扰动评估方法效率低且准确率不高.针对该问题,提出一种高效准确的扰动评估方法.首先,根据样本与分类器间的物理属性,定义样本的对抗性扰动,利用线性迭代方法评估计算二类分类器的鲁棒性;然后,为了适应更加一般的情况,将该方法扩展到更加普遍的多类非线性分类器,即超平面包围的区域变为不规则多面体;最后,标记扰动样本对分类器进行优化调整,并对当前估计进行更新,以进一步提高分类器性能.不同数据集和分类器的实验结果表明:与 FGSM 方法、L-BFGS 方法和未标记方法相比,提出的方法具有更稳定高效的扰动评估性能,且可以构建鲁棒性更高的分类器.

关 键 词:多类非线性分类器;对抗性样本;扰动评估;线性迭代;鲁棒性

中图分类号:TP391 **文献标志码**:A **文章编号**:1000-5471(2019)03-0088-07

深度神经网络[1-2]因为其良好的模式识别性能,在许多领域得到广泛应用[3-6].然而,当数据集较小或包含对抗性扰动样本时,其分类性能不够稳定,从而导致与纯净样本具有特征相似性的扰动数据的误分类.对抗性攻击能够泛化到不同的模型[7-8],因此如何准确评估不同分类器对于对抗性扰动的鲁棒性,并设计鲁棒性更强的分类器是机器学习的研究热点:文献[9]通过实验说明卷积网络对于某些变换不具备不变性;文献[10]提出对抗样本的快速生成方法—快速梯度符号法(FGSM, Fast Gradient Sign Method),虽然效率较高,但只给出最优扰动向量的粗略逼近,得到次最优解;文献[11]生成对抗性扰动,在训练过程中引入平滑度惩罚参数,提高分类器的对抗性扰动鲁棒性;文献[12]对不稳定现象进行理论研究,给出一些分类器的鲁棒性上界,但这些分类器比较简单;文献[13]提出神经网络的高复杂性可能是对抗性样本鲁棒性差的原因,通过 L-BFGS(Limited-memory BFGS)算法[14]搜寻对抗样本,并利用对抗样本对模型进行规范化,提高鲁棒性,但是该方法耗时过长,难以应用于大型数据集.

为了解决上述问题,本文根据样本与分类器间的物理属性,定义样本的对抗性扰动,并利用线性迭代方法评估计算分类器的鲁棒性;通过标记对抗样本,对分类器进行优化,提高其分类性能.

1 二类分类器鲁棒性

1.1 对抗性扰动的定义

对于一个已知分类器,本文将对抗性扰动 $\rho(\boldsymbol{x};\hat{y})$ 定义为能够改变分类标签 $\hat{y}(\boldsymbol{x})$ 的最小扰动向量 $\boldsymbol{r}$:

$$\begin{aligned}&\rho(\boldsymbol{x};\hat{y})=\min_{\boldsymbol{r}}\|\boldsymbol{r}\|_2^2\\&s.t\quad \hat{y}(\boldsymbol{x}+\boldsymbol{r})\neq\hat{y}(\boldsymbol{x})\end{aligned}\tag{1}$$

其中:$\boldsymbol{x}$ 为输入样本,$\hat{y}(\boldsymbol{x})$ 为分类标签,$\rho(\boldsymbol{x};\hat{y})$ 为分类器在 x 点的鲁棒性.则分类器的整体鲁棒性 $\rho(\hat{y})$ 定义如下:

① 收稿日期:2017-12-19
基金项目:四川省自然科学基金项目(14ZJ0280);四川省科技攻关项目(2014JY0095).
作者简介:邹 瑛(1979-),女,硕士,副教授,主要从事分类器与模式识别、数字图像技术、信息安全等研究.

$$\rho(\hat{y}) = E\left(\frac{\rho(\boldsymbol{x};\ \hat{y})}{\|\boldsymbol{x}\|_2^2}\right) \tag{2}$$

其中 E 为数据分布的期望.

1.2　鲁棒性计算

本文首先提出用于二类分类器的算法. 假定 $\hat{y}(\boldsymbol{x}) = \text{sign}(f(\boldsymbol{x}))$, $f(\boldsymbol{x})$ 为线性分类函数 $f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{b}$, $\Gamma = \{\boldsymbol{x}:\ f(\boldsymbol{x}) = \boldsymbol{0}\}$ 表示 f 的零水平集，即分类面. 从图 1 可以看出，f 在点 $\boldsymbol{x}_0$ 处的鲁棒性 $\rho(\boldsymbol{x}_0;\ f)$ 与从 $\boldsymbol{x}_0$ 到分类面 $\Gamma = \{\boldsymbol{x}:\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{b} = \boldsymbol{0}\}$ 的距离相等. 影响分类器决策的最小扰动 $\boldsymbol{r}_*(\boldsymbol{x}_0)$ 对应于 $\boldsymbol{x}_0$ 在 Γ 上的正交投影，如式(3) 所示.

$$\begin{aligned} &r_*(\boldsymbol{x}_0) = \min_{\boldsymbol{r}} \|\boldsymbol{r}\|_2^2 \\ &s.t \quad \text{sign}(f(\boldsymbol{x}_0 + \boldsymbol{r})) \neq \text{sign}(f(\boldsymbol{x}_0)) \end{aligned} \tag{3}$$

其中 $\text{sign}(f(\boldsymbol{x}_0)) = -\dfrac{f(\boldsymbol{x}_0)}{\|\boldsymbol{w}\|_2^2}\boldsymbol{w}$.

若 f 为非线性两类分类器，本文利用迭代过程计算其鲁棒性 $\rho(\boldsymbol{x}_0;\ f)$. 每一次迭代时，在当前点 $\boldsymbol{x}_i$ 周围对 f 进行线性化处理，该线性化分类器的最小扰动如式(4) 所示.

$$\begin{aligned} &\min_{\boldsymbol{r}_i} \|\boldsymbol{r}_i\|_2^2 \\ &s.t \quad f(\boldsymbol{x}_i) + \nabla f(\boldsymbol{x}_i)\boldsymbol{r}_i = \boldsymbol{0} \end{aligned} \tag{4}$$

第 i 步的扰动 $\boldsymbol{r}_i$ 由式(3) 求得，并对下一次迭代 $\boldsymbol{x}_{i+1}$ 进行更新. 当 $\boldsymbol{x}_{i+1}$ 改变分类器的分类标签时，算法停止. 具体过程如算法 1 所示.

实际中，该算法能够收敛至分类面 Γ 上的一个点. 为了能够达到分类面的另一边，将最终扰动向量 $\hat{\boldsymbol{r}}$ 乘以常数 $1+\eta$，其中 $\eta \ll 1$. 本文取 $\eta = 0.02$.

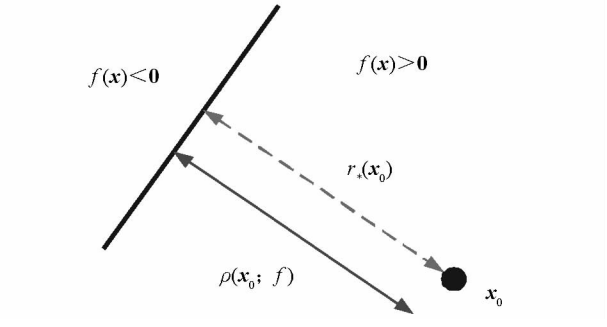

图 1　线性二类分类器对抗性样本

算法 1

输入：样本 $\boldsymbol{x}$，分类器 f

输出：扰动向量 $\hat{\boldsymbol{r}}$

初始化 $\boldsymbol{x}_0 \leftarrow \boldsymbol{x}$, $i \leftarrow 0$

while $\text{sign}(f(\boldsymbol{x}_i)) = \text{sign}(f(\boldsymbol{x}_0))$ do

$\boldsymbol{r}_i \leftarrow \dfrac{f(\boldsymbol{x}_i)}{\|\nabla f(\boldsymbol{x}_i)\|_2^2}\nabla f(\boldsymbol{x}_i)$

$\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{x}_i + \boldsymbol{r}_i$

$i \leftarrow i + 1$

end while

return $\hat{\boldsymbol{r}} = \sum_i \boldsymbol{r}_i$

2　多类分类器鲁棒性

下面将抖动计算扩展到多类情况. 假设分类器有 c 个输出，即类别数量为 c，通过式(5) 的映射关系进行分类.

$$\hat{y}(\boldsymbol{x}) = \arg\max_i f_i(\boldsymbol{x}),\ i = 1,2,\cdots,c \tag{5}$$

其中 $f_i(\boldsymbol{x})$ 为分类器的第 i 个类别的输出. 与二类分类器一样，首先对应于线性情况，然后扩展到非线性情况.

2.1　线性分类器

该种情况下，分类器函数 $f(\boldsymbol{x}) = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{b}$ 最小扰动的计算如式(6) 所示.

$$\begin{aligned}&\arg\min_{\boldsymbol{r}} \|\boldsymbol{r}\|_2^2\\ &s.t\ \ \exists k:\boldsymbol{w}_i^T(\boldsymbol{x}_0+\boldsymbol{r})+\boldsymbol{b}_i\geqslant \boldsymbol{w}_{\hat{y}(\boldsymbol{x}_0)}^{\mathrm{T}}(\boldsymbol{x}_0+\boldsymbol{r})+\boldsymbol{b}_{\hat{y}(\boldsymbol{x}_0)}\end{aligned}\tag{6}$$

其中 $\boldsymbol{w}_i$ 为矩阵 $\boldsymbol{W}$ 的第 i 列向量. 几何学上，该问题为求解 $\boldsymbol{x}_0$ 与凸多面体 P 的补集之间的距离，

$$P=\bigcap_{i=1}^{c}\{\boldsymbol{x}:f_{\hat{y}(\boldsymbol{x}_0)}(x)\geqslant f_i(\boldsymbol{x})\}\tag{7}$$

其中 $\boldsymbol{x}_0$ 位于 P 内，本文将该距离表示为 $D(\boldsymbol{x}_0,P^c)$. 多面体 P 定义为分类函数 f 输出标签 $\hat{y}(\boldsymbol{x}_0)$ 的空间的区域，3条线围成的区域即为 P(图2). 定义 $\hat{l}(\boldsymbol{x}_0)$ 为最接近 P 的界限的超平面，图2中 $\hat{l}(\boldsymbol{x}_0)=3$. 则 $\hat{l}(\boldsymbol{x}_0)$ 计算如式(8)所示：

$$\hat{l}(\boldsymbol{x}_0)=\arg\min_{i\neq\hat{y}(\boldsymbol{x}_0)}\frac{|f_i(\boldsymbol{x}_0)-f_{\hat{y}(\boldsymbol{x}_0)}(\boldsymbol{x}_0)|}{\|\boldsymbol{w}_i-\boldsymbol{w}_{\hat{y}(\boldsymbol{x}_0)}\|_2^2}\tag{8}$$

最小扰动 $r_*(\boldsymbol{x}_0)$ 为将 $\boldsymbol{x}_0$ 投影到 $\hat{l}(\boldsymbol{x}_0)$ 超平面上的向量，即 $\boldsymbol{x}_0$ 到 P 表面的最短距离：

$$r_*(\boldsymbol{x}_0)=\frac{|f_{\hat{l}(\boldsymbol{x}_0)}(\boldsymbol{x}_0)-f_{\hat{y}(\boldsymbol{x}_0)}(\boldsymbol{x}_0)|}{\|\boldsymbol{w}_{\hat{l}(\boldsymbol{x}_0)}-\boldsymbol{w}_{\hat{y}(\boldsymbol{x}_0)}\|_2^2}(\boldsymbol{w}_{\hat{l}(\boldsymbol{x}_0)}-\boldsymbol{w}_{\hat{y}(\boldsymbol{x}_0)})\tag{9}$$

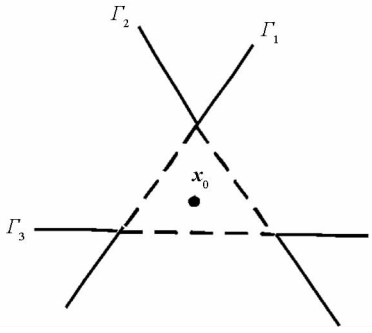

图2 多类线性分类器

2.2 非线性分类器

下面将扰动算法扩展到多类非线性分类器. 对于多类非线性分类器，式(7)中输出标签 $\hat{y}(\boldsymbol{x}_0)$ 所围区域集合 P 不再为规则多面体. 同二类非线性分类器相似，本文通过一个规则多面体 $\tilde{P}_i$ 对第 i 次的集合 P 进行逼近，如图3所示. 各超平面包围的区域为 P，虚线包围的区域为 $\tilde{P}$，计算公式如下：

$$\tilde{P}_i=\bigcap_{y=1}^{c}\{\boldsymbol{x}:f_y(\boldsymbol{x}_i)-f_{\hat{y}(\boldsymbol{x}_0)}(\boldsymbol{x}_i)+\nabla f_k(\boldsymbol{x}_i)\mathrm{T}\boldsymbol{x}-\nabla f_{\hat{y}(\boldsymbol{x}_0)}(\boldsymbol{x}_i)\mathrm{T}\boldsymbol{x}\leqslant 0\}\tag{10}$$

本文通过 $D(\boldsymbol{x}_i,\tilde{P}_i^c)$ 逼近 $D(\boldsymbol{x}_i,P^c)$ 的方法计算第 i 次迭代中 $\boldsymbol{x}_i$ 与 P 补集之间的距离，即每次迭代计算 $\boldsymbol{x}_i$ 到多面体 $\tilde{P}_i$ 边界的扰动向量，并对当前估计进行更新，具体如算法2所示. 二类非线性分类器的算法1为具有自适应步长的梯度下降算法，每次迭代会自动选择步长. 多类非线性分类器的算法2类似连续凸优化算法，每步迭代进行线性化约束.

算法2

输入：样本 $\boldsymbol{x}$，分类器 f

输出：扰动 $\hat{\boldsymbol{r}}$

初始化 $\boldsymbol{x}_0\leftarrow\boldsymbol{x}$，$i\leftarrow 0$.

while $\hat{y}(\boldsymbol{x}_i)=\hat{y}(\boldsymbol{x}_0)$ do

for $y\neq\hat{y}(\boldsymbol{x}_0)$ do

$\boldsymbol{w}_y'\leftarrow\nabla f_y(\boldsymbol{x}_i)-\nabla f_{\hat{y}(\boldsymbol{x}_0)}(\boldsymbol{x}_i)$

$f_y'\leftarrow f_y(\boldsymbol{x}_i)-f_{\hat{y}(\boldsymbol{x}_0)}(\boldsymbol{x}_i)$

end for

$\hat{l}\leftarrow\arg\min_{y\neq\hat{y}(\boldsymbol{x}_0)}\frac{|f_y'|}{\|\boldsymbol{w}_y'\|_2^2}$

$\boldsymbol{r}_i\leftarrow\frac{|f_{\hat{l}}'|}{\|\boldsymbol{w}_{\hat{l}}'\|}\boldsymbol{w}_{\hat{l}}'$

$\boldsymbol{x}_{i+1}\leftarrow\boldsymbol{x}_i+\boldsymbol{r}_i$

$i\leftarrow i+1$

end while

return $\hat{\boldsymbol{r}}=\sum_i\boldsymbol{r}_i$

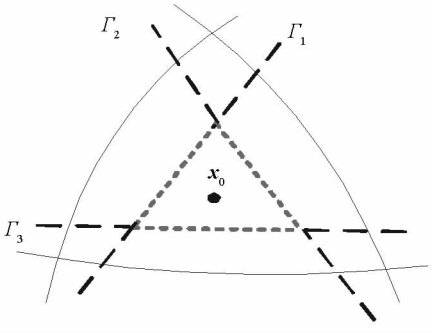

图3 多类非线性分类器

2.3　扩展 l_p 范数

上述算法基于 l_2 范数对扰动进行度量，下面扩展到 l_p 范数($p \in [1, +\infty]$). 为实现该扩展，需要将算法 2 中的 $\hat{l}$，$\boldsymbol{r}_i$ 进行如下修改：

$$\hat{l} \leftarrow \arg\min_{y \neq \hat{y}(x_0)} \frac{|f'_y|}{\|\boldsymbol{w}'_y\|_q} \tag{11}$$

$$\boldsymbol{r}_i \leftarrow \frac{|f'_{\hat{l}}|}{\|\boldsymbol{w}'_{\hat{l}}\|_q^q} |\boldsymbol{w}'_{\hat{l}}|^{q-1} \odot \mathrm{sign}(\boldsymbol{w}'_{\hat{l}}) \tag{12}$$

其中：$\odot$ 为逐点乘积，$q = \frac{p}{p-1}$，当 p 趋于无穷大时，$q = 1$.

3　实验结果与分析

本部分实验以 Matlab2016b 为平台，软件仿真界面如图 4 所示. 处理器为英特尔 Core i7 4800M 2.9 GHz；CUP 内存为 16 GB；操作系统为 64 位 Windows 10. 使用 MNIST 和 CIFAR-10 数据集分别训练 Alexnet 网络和 ResNet 网络分类器，对本文算法性能进行检验，同时与文献[10]的 FGSM 和文献[13]的 L-BFGS 两种算法进行比较. 利用式(2)计算分类器对扰动的鲁棒性.

图像加入微小扰动后被误分类的示例如图 5 所示. 所用图像是 CIFAR-10 数据集中两幅图像. 其中，第一行源图像是“鲸鱼”，加入一些扰动后，变成人眼很难区别的另一幅图像，但会被分类器错误分类为“海龟”. 第二行源图像是“熊猫”，加入一些扰动后，分类器误分类为“长臂猿”.

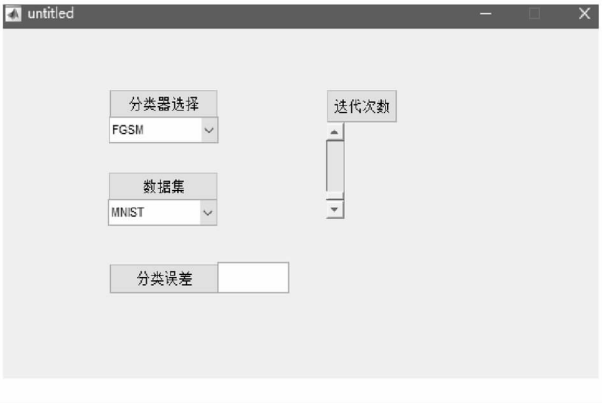

图 4　分类器最小扰动评估界面

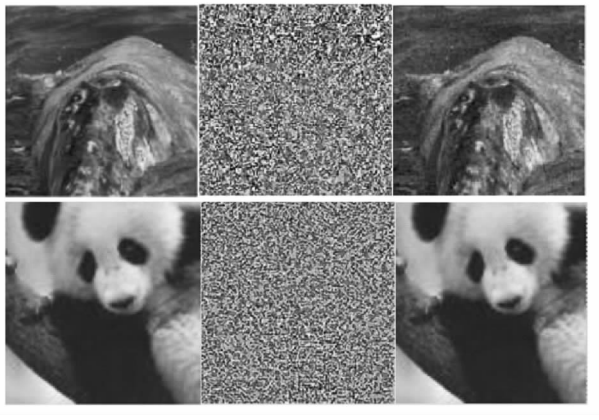

图 5　图像加入微小扰动被误分类的示例

3.1　鲁棒性评估

不同方法对于 MNIST 和 CIFAR-10 数据集的结果，分别如表 1 和表 2 所示.

由表 1 和表 2 知，对于不同分类器，本文方法估计出的鲁棒性 ρ 最小，其更接近式(1)定义的最小对抗扰动. 本文方法估计的鲁棒性与 L-BFGS 方法相近，但略小于后者；同时比 FGSM 方法小一个量级. 说明本文方法在检测分类器网络扰动对抗方面具有更高的准确性. 而 FGSM 方法和 L-BFGS 方法通过分类器的最小几何变换，给出分类器对几何变换鲁棒性的量化度量，且只能提供最优扰动向量的一个粗略逼近. 同时由表 1 和表 2 知本文方法的运算时间远低于 L-BFGS 方法，略高于 FGSM 方法. 因为 L-BFGS 方法涉及一系列目标函数最优化处理，时间复杂度较高. 而本文方法经过较少的迭代次数就能够收敛到稳定的抖动向量结果，一般迭代次数为 3 即可. 说明本文方法在提高扰动评估性能的同时，能够保持快速高效的计算效率，可有效应用于要求更深层复杂神经网络的大数据集情况. 对于本文方法，每次迭代中影响分类器决策的最小扰动、对应于 x_0 在 Γ 上的正交投影值如表 3 所示，可以看出，在 3 次迭代后，其投影值基本处于稳定状态.

表1 MNIST数据集结果

分类器	误差/%	算法	鲁棒性	时间/s
Alexnet	1.4	FGSM	1.51	0.02
		L-BFGS	0.22	4.80
		本文	0.19	0.13
ResNet	2.3	FGSM	0.47	0.01
		L-BFGS	0.17	2.80
		本文	0.15	0.04

表2 CIFAR-10数据集结果

分类器	误差/%	算法	鲁棒性	时间/s
Alexnet	9.5	FGSM	0.140	0.15
		L-BFGS	0.027	48.00
		本文	0.025	1.30
ResNet	19.6	FGSM	0.138	0.47
		L-BFGS	0.042	65.00
		本文	0.040	0.18

表3 未优化时每次迭代的投影值

迭代次数/次	MNIST数据集		CIFAR数据集	
	Alexnet分类器	ResNet分类器	Alexnet分类器	ResNet分类器
1	0.70	0.67	0.76	0.71
2	0.52	0.57	0.65	0.60
3	0.36	0.40	0.45	0.47
4	0.31	0.41	0.46	0.48
5	0.32	0.40	0.46	0.47
6	0.31	0.41	0.45	0.48

3.3 分类器优化

扰动性分析只是初步对分类器网络的性能进行评估，最终的目的是提高分类器网络的分类准确度. 为了进一步分析本文方法的性能，本节实验分别利用FGSM方法、L-BFGS方法和本文方法标记对抗性样本，然后利用该对抗性样本对分类器性能进行微调优化，提高鲁棒性. 本节实验在3.2节训练网络的基础上增加6次迭代，并且仅在扰动训练集上将学习率降低50%. 分类器优化后，统一使用本文方法评估其鲁棒性. 对MNIST数据集和CIFAR-10数据集采用未添加对抗样本、FGSM方法标记对抗样本、L-BFGS方法标记对抗样本、本文方法标记对抗样本4种方法进行微调，6次迭代的分类器鲁棒性分别如图6和图7所示. 对于图6的MNIST数据集，可以看出，本文方法比其他方法的曲线更早达到平稳的态势，即更快达到收敛状态. 对于图7的CIFAR-10数据集，除了本文方法，其他方法表现出没有规律的鲁棒性变化，而本文方法变化平稳.

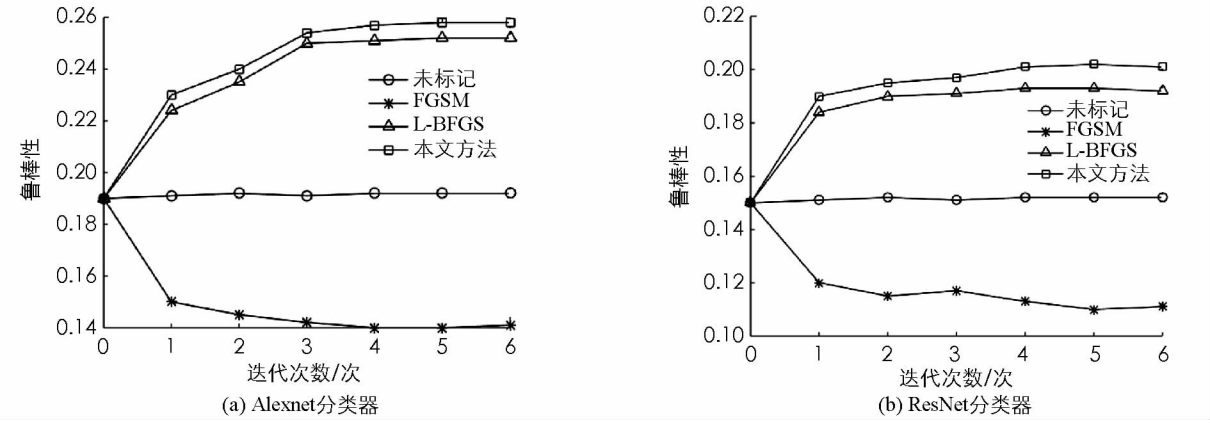

图6 MNIST数据集

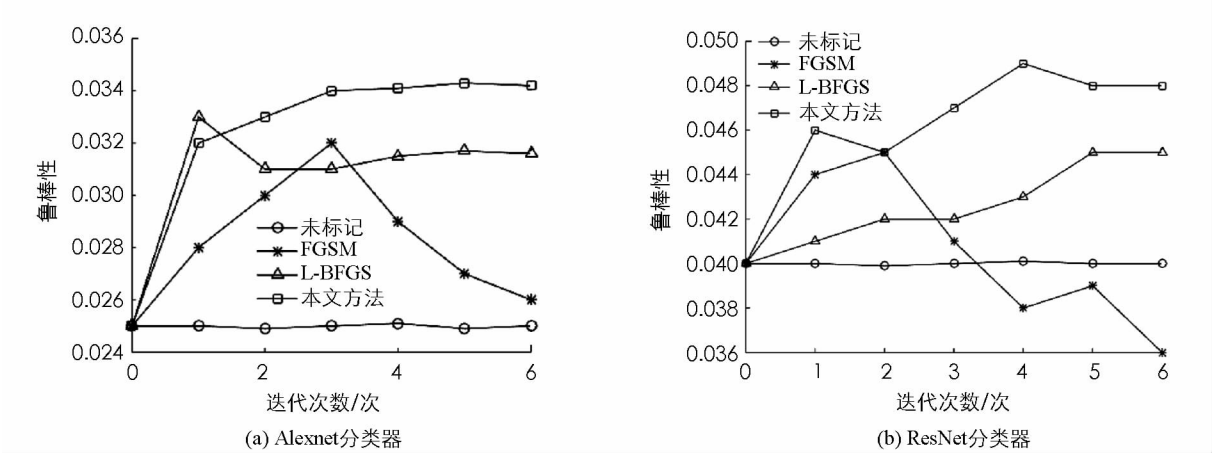

图 7　CIFAR-10 数据集

由图 5 和图 6 可以看出，对分类器进行优化后，本文方法所获得的鲁棒性明显提高，优于其他方法，在增加 3 次迭代后，分类器性能趋于稳定. L-BFGS 方法对分类器进行优化后，效果较本文稍差一些，但明显优于 FGSM 方法. FGSM 方法的效果较差，这是因为 FGSM 输出含有严重受扰动的图像，并导向那些不会出现在测试数据中的图像. 由此降低了该方法的性能. L-BFGS 拥有一个不代表原始数据分布的正则项，稍微弱于本文方法. 对于图 6 的 MNIST 数据集，分类器优化后鲁棒性反而随着迭代次数的增加逐渐降低；对于图 7 的 CIFAR-10 数据集，分类器优化后鲁棒性不稳定. 出现该情况是由于 FGSM 方法估计出的扰动远大于本文方法的最小扰动，使用扰动性过大的样本对网络进行调整，降低了该网络对扰动的鲁棒性.

为了进一步说明本文方法的良好性能，对各方法优化后分类器的分类误差进行比较. 分类误差的定义如下：

$$\eta = sl - \frac{t}{S} \tag{13}$$

其中：sl 表示理论基准分类率，t 表示本次实验准确分类的样本数，S 表示总样本数.

不同方法的优化结果如表 4 所示. 可以看出，FGSM 方法由于标记出的抖动样本不准确，致使优化之后分类器的性能反而有所下降，分类误差上升. L-BFGS 方法和本文方法通过正确标记抖动样本，优化后的分类误差明显更低，这主要得益于多类非线性分类器的性能，使得最优扰动向量的逼近值更佳. 其中，本文优化后的每次迭代投影值如表 5 所示，与表 3 相比，投影值趋于稳定的速度更快，这也从另一个侧面说明优化后的鲁棒性更佳.

表 4　不同方法分类器优化后的分类误差结果

数据集	分类器	分类误差/%			
		未标记	FGSM	L-BFGS	本文
MNIST	Alexnet	1.4	2.2	1.0	0.7
	ResNet	2.3	3.1	1.7	1.3
CIFAR-10	Alexnet	9.5	9.2	5.3	4.8
	ResNet	19.6	22.1	15.4	10.1

表 5　优化后每次迭代的投影值

迭代次数/次	MNIST 数据集		CIFAR 数据集	
	Alexnet 分类器	ResNet 分类器	Alexnet 分类器	ResNet 分类器
1	0.69	0.65	0.75	0.69
2	0.49	0.51	0.60	0.61
3	0.35	0.39	0.51	0.53
4	0.31	0.41	0.45	0.48
5	0.31	0.41	0.45	0.48
6	0.31	0.41	0.45	0.48

4　结　　论

本文提出一种新的扰动评估方法，其基于分类器的线性迭代评估能够改变分类标签的最小扰动，同时

将该方法扩展到多类非线性分类器. 对不同数据集和分类器进行的比较实验表明，本文方法在计算对抗性扰动方面更加准确高效. 同时本文利用该方法生成扰动样本，对分类器进行优化调整，进一步提高分类器的性能. 因此，本文提出的方法能够有效用于对最小扰动向量的准确估计，并构建鲁棒性更高的分类器.

参考文献：

[1] 尹宝才，王文通，王立春. 深度学习研究综述 [J]. 北京工业大学学报，2015，41(1)：48-59.
[2] 刘建伟，刘 媛，罗雄麟. 深度学习研究进展 [J]. 计算机应用研究，2014，31(7)：1921-1930.
[3] 王 军. 基于深度神经网络的中期电力负荷预测 [J]. 重庆工商大学学报(自然科学版)，2018，35(6)：17-21.
[4] 李文洁，张晴晴，张鹏远，等. 基于维特比算法的深度神经网络语音端点检测 [J]. 重庆邮电大学学报(自然科学版)，2018，30(2)：210-215.
[5] 卢宏涛，张秦川. 深度卷积神经网络在计算机视觉中的应用研究综述 [J]. 数据采集与处理，2016，31(1)：1-17.
[6] 赵 军，赵 艳，杨 勇，等. 基于降维的堆积降噪自动编码机的表情识别方法 [J]. 重庆邮电大学学报(自然科学版)，2016，28(6)：844-848.
[7] GU S，RIGAZIO L. Towards Deep Neural Network Architectures Robust to Adversarial Examples [J]. Computer Science，201423(5)：67-79.
[8] MOOSAVIDEZFOOLI S. On the Robustness of Deep Networks for Images Classification [J]. General Information，2015，7(34)：435-447.
[9] 张 晖，苏 红，张学良，等. 基于卷积神经网络的鲁棒性基音检测方法 [J]. 自动化学报，2016，42(6)：959-964.
[10] GOODFELLOW I J，SHLENS J，SZEGEDY C. Explaining and Harnessing Adversarial Examples [J]. Computer Science，2014，37(8)：48-61.
[11] 常清泉. 容错神经网络学习算法设计与收敛性研究 [D]. 兰州：兰州大学，2015.
[12] FAWZI A，MOOSAVIDEFOOLI S M，FROSSARD P. Robustness of Classifiers：from Adversarial to Random Noise [J]. Computer Science，2016，8(31)：765-785.
[13] SZEGEDY C，ZARAMBA W，SUTSKEVER I，et al. Intriguing Properties of Neural Networks [J]. Computer Science，2014，7(12)：1-10.
[14] 张永义. 基于深度神经网络的场景分类方法研究 [D]. 南昌：南昌航空大学，2016.

Minimal Perturbation Evaluation Approach for Classifier Based on Linear Iteration

ZOU Ying

Department of Judicial Information Management，Sichuan Vocational College of Judicial Police，Deyang Sichuan 618000，China

Abstract：When data sets contain adversarial perturbation samples，their classification structure lacks stability，and traditional perturbation evaluation methods are complex，inefficient and accurate. To solve this problem，a perturbation evaluation method with efficiency and accuracy has beenproposed. Firstly，according to physical properties of sample and antagonism between the classifiers，sample antagonistic perturbations are defined，and the linear iterative method is used to evaluate the two classes robustness of classifiers. Secondly，in order to adapt to more general cases，the proposed method is extended to multiple class classification with more general nonlinear，which means that hyperplane encircled region becomes an irregular polyhedron. And finally，perturbation samples are tagged to optimize classifier，and updates the current estimate，so that the classifier performance gets further improvement. Through experiments for different data-sets and classifiers，experimental results show that the proposed method could get more stable and efficient perturbation evaluation performance in comparison with traditional methods，which makes classifiers more robust.

Key words：multi-class nonlinear classifier；adversarial samples；perturbation evaluation；linear iteration；robust

责任编辑 张 枸