doi: 10.12194/j.ntu.20221205003
引文格式: 张承烨, 李卓轩, 曹进德. 基于随机 k–近邻集成算法的网络流量入侵检测[J]. 南通大学学报(自然科学版), 2023, 22(3):26–32.

基于随机 k–近邻集成算法的网络流量入侵检测

张承烨[1], 李卓轩[2,3], 曹进德[2,3*]
(1. 东南大学 人工智能学院, 江苏 南京 211189;2. 东南大学 数学学院, 江苏 南京 211189;
3. 江苏省网络群体智能重点实验室, 江苏 南京 211189)

摘要:为了提高网络入侵检测模型的准确率与泛化性,提出基于随机 k–近邻集成算法的网络流量入侵检测模型。首先,该模型提出一种集成赋权距离,来提高预测精度;其次,采用一种随机策略的集成方法对 k–近邻模型进行集成,从而提高了其在异常检测过程中的全局和局部优化能力;然后,利用并行计算的方法提高了算法运行的效率;最后,构建了基于随机 k–近邻集成算法的网络入侵检测模型,并采用 KDD99 数据集进行实验。实验结果表明,基于随机 k–近邻集成算法相对于其他模型具有更好的检测效果,准确率和召回率分别达到 99.05%和 91.96%。

关键词:网络入侵检测;k–近邻模型;集成赋权距离;随机子空间;并行计算
中图分类号: TP181 **文献标志码**: A **文章编号**: 1673–2340(2023)03–0026–07

Network intrusion detection based on random k–nearest neighbor ensemble algorithm

ZHANG Chengye[1], LI Zhuoxuan[2,3], CAO Jinde[2,3*]
(1. School of Artificial Intelligence, Southeast University, Nanjing 211189, China;
2. School of Mathematics, Southeast University, Nanjing 211189, China;
3. Jiangsu Provincial Key Laboratory of Networked Collective Intelligence, Southeast University, Nanjing 211189, China)

Abstract: To improve the accuracy and generalization of network intrusion detection models, a model based on the random k–nearest neighbor (k–NN) ensemble algorithm is proposed for network flow intrusion detection. Firstly, the model introduces an ensemble weighting distance to enhance prediction accuracy. Secondly, a random strategy is employed to integrate the k–NN models, thereby improving their global and local optimization capabilities in the anomaly detection process. Furthermore, parallel computing techniques are utilized to enhance algorithm efficiency. Lastly, a network intrusion detection model based on the random k–nearest neighbor ensemble algorithm is constructed and experimented with using the KDD99 dataset. Experimental results demonstrate that the random k–nearest neighbor ensemble algorithm outperforms other models, achieving accuracy and recall rates of 99.05% and 91.96%, respectively.

Key words: network intrusion detection; k–nearest neighbor model; integrated weighted distance; random subspace; parallel computing

收稿日期: 2022–12–05 **接受日期**: 2023–03–27
基金项目: 江苏省网络集体智能重点实验室项目(BM2017002)
第一作者简介: 张承烨(2001—), 男, 本科生。
* **通信联系人**: 曹进德(1963—), 男, 教授, 博士, 博士生导师, 主要研究方向为复杂网络、群体智能算法、系统科学等。
E–mail:jdcao@seu.edu.cn

随着互联网技术的快速发展与普及，网络安全问题变得日益突出，木马程序、蠕虫病毒、DDOS（distributed denial of service）攻击等大量的威胁与安全隐患扰乱了社会正常情况的运作与经济的可持续发展。随着人工智能技术应用的发展[1-2]，网络流量异常检测方法作为网络安全系统中入侵检测技术实现的核心，受到国内外学者的广泛关注。

熊钢等[3]针对网络空间安全中面临的威胁进行了综述，将入侵检测系统功能分为两大类：异常检测软件系统和签名检测系统。异常检测软件系统在检测未知攻击方面表现更理想，然而会产生很高的误报率。因此，提高入侵检测系统的检测精度和学习速度仍然是一项艰巨的任务。Papamartzivanos 等[4]提出了名为 Dendron 的入侵检测系统，通过遗传算法对决策算法进行优化，旨在生成准确且无偏的决策树，能够检测常见和罕见的侵入性攻击。生龙等[5]为了提高网络入侵检测模型的准确率与泛化性，采用核主成分分析提取入侵性攻击的特征，通过改进的引力搜索算法对混合核极限学习机进行优化。徐国天[6]提出一种基于裁剪树方法优化的 k–近邻（k–nearest neighbor，k–NN）的入侵检测模型，维持较高查询效率。王月等[7]提出名为 imTk–NN 的网络入侵检测模型，采用改进的三元组网络和 k–近邻算法进行组合预测，精度和效率都有大幅提升。

综上，本文提出一种新的 k–NN 集成算法——随机 k–NN（random k–nearest neighbor，Rk–NN）算法，应用于网络异常流量检测。主要包含以下部分：1）提出一种基于熵权法与变异系数法的集成赋权欧氏距离；2）采用一种随机策略的集成方法，首先采用 bootstrap 再抽样技术，对抽样后的数据集随机选取特征构成随机子空间，然后生成多个组件 k–NN 分类器，最后对整个集成方法进行并行优化；3）对每一个子 k–NN 分类器按照均匀分布随机生成 k 值。

1 相关理论基础

1.1 k–NN 算法

k–NN 分类算法[8]简单、易于实现，是机器学习领域的一个经典机器学习算法，引起了人们的广泛关注[9]。k–NN 算法模型已广泛应用于文本挖掘、模式识别等领域，它是一种基于实例的学习法，可以根据被预测对象与已知对象之间的相似度进行归类，并不能显式地构建模型，k–NN 算法模型过于依赖待分类具体实例与训练集中最近的 k 个邻居。

对于训练集 $D=\{(x_i,y_i)\mid i=1,2,\cdots,n\}$，待分类实例 $z=(x,y)$，表 1 给出 k–NN 的分类过程[1]。

表 1 k–NN 算法

Tab. 1 k–NN algorithm

算法 1 k–NN 算法
输入：训练集 D，待分类实例 z，近邻值 k；
输出：分类结果 $h_{k-\mathrm{NN}}$。
计算 z 与训练集 D 中每个数据之间的距离
根据 k 值确定出 k 近邻集 D_z
计算类别 $h_{k-\mathrm{NN}}=\mathrm{argmax}_l\sum_{(x_i,y_i)\in D_z}V(l=y_i)$
return $h_{k-\mathrm{NN}}$

算法 1 中：l 为类标签；y_i 是第 i 个近邻的类标签；$V(\cdot)$为一个指示函数，当条件为真时返回 1，当条件为假时返回 0。

与神经网络和决策树分类器有许多参数不同，k–NN 分类器只有两个参数，即用于计算给定测试样本与训练样本的距离度量和邻居 k 的数量，这使得 k–NN 的优化研究具有挑战性。k–NN 可以根据与该实例最近的 k 个实例进行归类，然而 k 的优与劣将完全决定着 k–NN 分类的使用效果。诚然，k–NN 算法的关键之一在于如何采纳一个合理的 k 值[10]。如果 k 值取得过小，则算法易受噪声的影响，使分类结果不稳定；k 值取得过大，则近邻集中含有太多其他类别的实例，导致分类错误，同时，过大的 k 值也会增加算法的时间总开销。理论上可以通过采用枚举法，循环遍历所有可能的 k 值，最后从中选出最好的 k 值作为近邻数。但在实际应用中，枚举法带来的时间总开销一般说来是根本无法接受的。

1.2 bagging 集成方法

1996 年，Breiman[11]提出 bagging 集成方法，它是最早也最简单的集成算法之一。借助 bootstrap 抽样从原始训练集得到多个同等规模的训练集合的子集，接着使用相同学习算法模型在这些训练集上训练出多个基分类算法，最终以多数投票方式组合所

有分类结果。

相对于新的未见实例,由训练阶段得到的个体分类算法分别对其所属类别投票,总票数最大的类别即是最终的决策结果。所提出的 bagging 集成算法具体直观有效, 尤其适用于数据集很小的情景。表 2 给出了提出 bagging 算法描述[2]。

表 2 bagging 算法

Tab. 2 bagging algorithm

算法 2 bagging 算法
输入:训练集 D,组件分类器 L,组件个数 M;
输出:分类结果 H。
for $m = 0,1,2,3,\cdots,M$ do
D_m = bootstrap-sample(D)
$h_m = L_m(D_m)$
end for
return $H(x) = \text{argmax}_y \sum_{m=1}^{M}(h_m(x) = y)$

虽然 bagging 集成方法在决策树(decision tree, DT)[12]和神经网络(neural network,NN)[13]上取得了巨大的成功,但它很难在 k-NN 分类器上很好地工作[11]。因为 k-NN 是一个稳定的分类器,而 bagging 使用 bootstrap 再抽样技术来生成准确但多样的组件分类器,这对不稳定的机器学习方法如决策树和类感知机算法的工作是更有效的。

2 Rk-NN 算法

2.1 集成赋权距离

为了更合理地分析各个数据对象之间的差异程度,本文利用一种信息熵与变异系数结合的集成赋权法来计算各数据对象的赋权欧式距离。变异系数法通过利用各项特征反映的信息计算得到特征的权重,它反映的是特征值的差异程度及数据的分布情况,假设数据集 X 有 m 个对象和 n 个特征,变异系数权重计算方法为

$$v_i = \delta_i / \bar{x}_i,$$

$$w_i^v = v_i / \sum_{i=1}^{n} v_i,$$

其中:δ_i 为第 i 项指标的标准差; $\bar{x}_i$ 为第 i 项指标的平均数;w_i^v 为变异系数法确定的第 i 个特征权重。

在信息论中,熵是对数据不确定性的一种度量和描述,熵值越大,数据不确定性越小;反之亦然,熵值越小,不确定性则越大。一个方案的无序性和随机性的程度可以由熵特性来确定。此方法可以消除单位或平均值不同对多种数据变异程度的影响。假设数据集 X 有 m 个对象和 n 维属性,设标准化后的数据为 x_{ij},熵权计算方法为

$$P_{ij} = x_{ij} / \sum_{i=1}^{n} x_{ij},$$

$$e_j = -k \sum_{i=1}^{n} P_{ij} \ln(P_{ij}),$$

$$g_i = 1 - e_i,$$

$$w_j^e = g_j / \sum_{j=1}^{n} g_j,$$

其中:e_j 为第 j 个特征的熵值,$k > 0$,$e_j > 0$,k 与样本量 m 有关,一般令 $k = 1/\ln(m)$,则 $0 \leqslant e_j \leqslant 1$;对于第 j 项指标,指标值 x_{ij} 差异越大,熵值越小,g_i 越大,在计算距离时占比越大;权重 w_j^e 为熵权法确定第 j 个特征的权重。

本文通过线性加权组合的方法,综合熵值法与变异系数法的权重确定结果,增加权重确定的科学性。其计算公式为

$$w_j^c = w_j^e w_j^v / (\sum_{j=1}^{n} w_j^e w_j^v),$$

其中 w_j^c 为组合权重。

最终通过集成赋权法,确定赋权欧氏距离为

$$d(x_q, x_p) = \sqrt{\sum_{j=1}^{n} w_j^c (x_{pj} - x_{qj})^2}。$$

加权欧氏距离可以用来判断不同数据对象之间的相似程度,可以提高分类精度。

2.2 随机策略优化

为了提高 k-NN 的预测精度, 首先进行随机特征(输入变量)选取,在引入特征随机性的同时,减小相关系数而保持强度不变。本文对训练集的输入属性进行随机扰动,从最初训练集的属性集中抽取出若干个属性子集,基于每个子集训练一个 k-NN 组件分类器,不同的“子空间”提供了观察数据的不同视角。本文通过设定特征抽取比例 α,对训练集的原始输入特征进行按比例随机选取,再进行 bootstrap 抽样,最终生成随机特征空间,过程如表 3 所示。显然,上述方案能解决 bagging 算法对于k-NN这种稳定分类器优化效果差的问题。

表 3　*RS* 算法

Tab. 3　*RS* algorithm

算法 3　*RS* 算法
输入：训练集 D，特征抽取比例 α；
输出：随机训练集 $\mathscr{F}$。
按照特征抽取比例 α，对训练集 D 进行特征抽取，生成子集 D_α
对 D_α 进行 bootstarp 抽样，得到随机训练集 $\mathscr{F}$
return $\mathscr{F}$

通常选取超参 k 值进行 k–NN 算法的优化[14]，为数据集找到最优 k，并依此进行分类。因为没有先验知识，直接指定 k 值具有很大的盲目性，同时，穷举法又由于时间开销过大而不可取。本文提出一种对多个 k–NN 组件分类器 k 值的选取方法，按照区间 $\Omega = [N, 2N+1]$ 均匀分布生成单一 k–NN 组件分类器的 k 值，为

$k_i \sim U(\Omega), i = 1, 2, \cdots, M,$

其中：i 为 k–NN 组件分类器编号；M 为 k–NN 组件分类器总个数。算法过程如表 4 所示。

表 4　Rk–NN 算法

Tab. 4　Rk–NN algorithm

算法 4　Rk–NN 算法
输入：训练集 D，k–NN 组件分类器 L，组件个数 M，特征抽取比例 α，待分类实例 $\mathscr{Z}$；
输出：分类结果 H。
$K = \{k_i \mid k_i \sim U(\Omega), i = 1, 2, \cdots, M\}$
for $m = 0, 1, 2, 3, \cdots, M$ do
$k = k_m \in K$
$\mathscr{F}_m = \mathscr{RS}(D, \alpha)$
$h_m = L_m(\mathscr{F}_m, k, z)$
end for
计算类别：$H = \operatorname{argmax}_l \sum_{m=1}^{M} V(h_m = y)$
return H

由于算法有良好的并行性，本文对 Rk–NN 算法分别实现数据处理的并行化，实现模型训练的并行化。

2.3　基于 Rk–NN 算法的网络入侵检测模型

基于以上内容，本文提出了一种新型的网络入侵检测模型——Rk–NN 算法的网络入侵检测模型，结构如图 1 所示[5]。

图 1　基于 Rk–NN 算法的网络入侵检测模型[5]

Fig. 1　Network intrusion detection model based on Rk–NN algorithm[5]

该模型的具体步骤如下：

Step1　输入数据集，初始化 Rk–NN 算法；

Step2　对数据集进行划分，拆分成训练集和测试集；

Step3　使用预处理的训练数据集生成历史数据搜索空间；

Step4　使用测试数据集评估获得的最佳 Rk–NN 模型的性能；

Step5　输出模型检测结果的评估指标。

3　实验与分析

3.1　实验数据

为了验证 Rk–NN 算法的有效性，选取 6 组 UCI 数据集中的公开测试数据（见表 5）和 KDD99 数据集进行实验。

表 5　UCI 数据集[6]

Tab. 5　UCI dataset[6]

数据集	类别数量	样本数量	数据维度
wine	3	178	14
zoo	7	101	18
sonar	2	208	61
ionosphere	2	351	35
diabites	2	768	9
balance–scale	3	625	5

KDD99 数据集中共包含 23 种攻击类型，可以分为四大攻击类别：DoS、PRB、U2R 和 R2L[15]。由于利用 KDD99 完整的数据集在进行机器学习算法训练时很复杂，因此大多数研究人员都使用了 KDD99 数据集中 10%子集作为实验数据。表 6 给出了KDD99 数据集的详细信息，其中训练数据集和两个测试数据集分别表示为 T_0、T_1 和 T_2。训练数据集 T_0 与测试数据集 T_1 一起用来验证 Rk–NN 算法的有效性。此外，本文还利用测试数据集 T_2 进行Rk–NN 算法与其他研究模型的性能比较。

表 6 KDD 数据集划分[7]

Tab. 6 KDD dataset division[7]

类别	T_0	T_1	T_2
Normal	200	1 000	10 000
Dos	60	500	40 000
PRB	40	500	400
U2R	30	52	52
R2L	60	1 000	100

3.2 实验结果评价指标

为了评估 Rk-NN 算法的性能,采用准确率(accuracy,Acc)和 F-score 值,对算法在 UCI 数据集上的表现进行评价。为了评估 Rk-NN 算法在网络入侵检测中的效果,采用精准率(precision)和 F-score 值作为每类攻击的评估指标;准确率(accuracy,Acc)、召回率(recall)、平均 F-score(mean F-score,MF)作为数据集整体的评估指标。在数据集中的网络连接可以分为正常(标记为 0)、攻击(标记为 1,2,3,4)两大类。计算方式如表 7 所示。

表 7 评价指标

Tab. 7 Evaluation index

名称	计算公式
准确率	$(TP+TN)/(TP+TN+FP+FN)$
精准率	$TP/(TP+FP)$
召回率	$TP/(TP+FN)$
F-score	$(1+\beta)^2\times precision\times recall/(\beta^2\times(precision+recall))$
平均 F-score	$(\sum_{i=1}^{n}F-score_i)/n$

其中:TP(真阳性)为正样本被正确预测为正样本;TP(假阳性)为负样本被错误预测为正样本;TN(真阴性)为负样本被正确预测为负样本;FN(假阴性)为正样本被错误预测为负样本,在 F-score 中令 $\beta=1$。

3.3 实验结果分析对比

3.3.1 Rk-NN 算法在 UCI 数据集上的表现

在 UCI 数据集上将 Rk-NN 算法与同类变种算法进行对比(见表 8),对比算法分别为:k-NN 算法、bagging 集成的 k-NN 算法和改进距离的 k-NN 算法,其中 k-NN 算法,bagging 集成的 k-NN 算法采用欧氏距离,Rk-NN 算法采用集成赋权距离。

由表 8 可看出,Rk-NN 算法具有更好的分类精度。相比欧式距离,Rk-NN 算法采用集成赋权距离能有效挖掘数据之间的联系,即 Rk-NN 通过权重计算能更有效描述不同样本之间的联系,具有更好的分类性能;不仅保持 bg k-NN 集成算法的优点,能有效描述样本之间的联系,而且通过随机集成策略构建随机子空间,Rk-NN 算法能有效利用数据集上的信息形成分类器间的差异,强化不同分类器中信息的重要性,提高集成的效果,具有更高的分类性能。Rk-NN 算法相较 k-NN 算法在准确率上平均提升了 13.00%,在平均 F-score 指标上平均提升了 15.20%。

3.3.2 Rk-NN 算法在 KDD99 数据集上的表现

本文利用训练数据集 T_0 和测试数据集 T_1 对 Rk-NN 算法性能进行分析,对于不同攻击类别进行比较,如图 2 所示,性能如表 9 所示。

如表 9 所示,Rk-NN 算法整体达到了一个较高水平,针对 Normal 类别,召回率达到 99.40%,但精准率只有 90.86%,存在误报的情况;对于 Dos 攻击能够精准识别,精准率达到 99.60%,同时 F-score和召回率也分别达到了 99.70%和 99.80%;对 R2L,U2R 和 PRB 的识别,本文算法也达到了一个

表 8 在 UCI 数据集上的不同算法比较

Tab. 8 Comparison of different algorithms on UCI dataset %

数据集	Rk-NN		bg$k$-NN		改进距离的 k-NN		k-NN	
	Acc	MF	Acc	MF	Acc	MF	Acc	MF
wine	98.15	98.33	74.07	73.31	75.93	75.17	74.07	72.61
zoo	90.32	71.43	83.87	67.53	87.10	70.67	83.87	62.60
sonar	88.89	88.31	74.60	73.47	76.19	75.97	74.60	73.47
ionosphere	93.40	92.78	84.91	81.76	85.85	83.63	83.96	80.44
diabites	73.61	69.81	71.43	68.97	71.43	67.42	71.00	68.24
balance-scale	89.36	62.18	87.23	60.84	86.17	60.63	85.64	59.88

较高的水平。

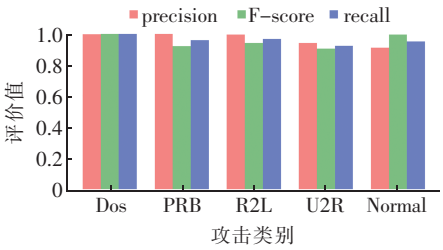

图 2　Rk-NN 算法检测结果

Fig. 2　Rk-NN algorithm detection results

表 9　不同攻击的检测性能比较

Tab. 9　Comparison of detection performance of different attacks　%

攻击	精准率	F-score	召回率
Normal	90.86	94.94	99.40
Dos	99.60	99.70	99.80
PRB	99.78	95.73	92.00
U2R	94.00	92.16	90.38
R2L	99.36	96.61	94.00

3.3.3　Rk-NN 算法与其他算法的对比

对于在测试数据集 T_2 上，将本文提出的基于 Rk-NN 算法的网络入侵检测模型与其他常用的机器学习算法如：支持向量机(support vector machine，SVM)算法、k-近邻(k-nearest neighbor，k-NN)随机森林(random forests，RF)模型、决策树(decision tree，DT)算法、极限梯度树(extrme gradient boosting，XGBoost)模型算法分别进行测试，结果如表 10 所示。

表 10　本文算法与其他算法比较结果

Tab. 10　Compare with other algorithms　%

算法	Acc	MF	召回率
本文算法	**99.05**	97.44	**91.96**
SVM	97.67	95.49	64.33
k-NN	97.87	65.51	86.59
RF	98.68	**97.56**	81.99
DT	96.59	96.20	60.80
XGBoost	97.87	97.29	72.04

Rk-NN 模型检测结果的评估指标准确率、平均 F-score 和召回率的值分别为 99.05%、97.44%和 91.96%。本文算法相较 SVM 算法在准确率、平均 F-score、召回率分别提升了 1.41%、2.04%和42.95%；相较 DT 算法在准确率、平均 F-score、召回率分别提升了 2.55%、1.29%和 51.25%；相较XGBoost 算法在准确率、平均 F-score、召回率分别提升了1.21%、0.15%和 27.65%；相较 RF 算法在平均 F-score 没有提升，在准确率、召回率分别提升了 0.37%、12.16%。可以看出，与其他方法相比，本文算法在流量入侵检测上具有更高的分类精度并且有更高的召回率。这充分说明 Rk-NN 模型在 KDD99 数据集上具有更好的性能。

4　结论

本文所提的 Rk-NN 模型在计算距离时引入一种集成赋权的欧式距离，采用熵权法和变异系数法计算不同特征之间的权值差异性，使得能够有效识别出关键特征。加入集成赋权欧式和随机策略优化方法，提升各个子分类器差异程度，提高了在网络入侵检测中的识别精度。Rk-NN 模型有良好的并行性，本文通过并行优化，提升算法运行效率。下一步将通过对 Rk-NN 模型进行模型剪枝，提升模型精度和算法效率，建立更加实时高效的检测模型，完善网络流量入侵的检测体系。

参考文献：

[1] 李卓轩，赵璇，曹进德，等. 政务服务中群众留言答复意见评价模型[J]. 南京信息工程大学学报(自然科学版)，2022，14(2)：178-185.
LI Z X，ZHAO X，CAO J D，et al. Evaluation of replies to public consultations in government service[J]. Journal of Nanjing University of Information Science and Technology (Natural Science Edition)，2022，14(2)：178-185. (in Chinese)

[2] 李卓轩，林凯迪，郭建华，等. 基于车联网数据的运输车辆安全评价模型[J]. 南通大学学报(自然科学版)，2020，19(1)：26-32.
LI Z X，LIN K D，GUO J H，et al. Transportation vehicle safety evaluation model based on vehicle network data[J]. Journal of Nantong University (Natural Science Edition)，2020，19(1)：26-32. (in Chinese)

[3] 熊钢，葛雨玮，褚衍杰，等. 基于跨域协同的网络空间威胁预警模式[J]. 网络与信息安全学报，2020，6(6)：88-96.
XIONG G，GE Y W，CHU Y J，et al. Model of cy-

berspace threat early warning based on cross-domain and collaboration[J]. Chinese Journal of Network and Information Security, 2020, 6(6):88-96. (in Chinese)

[4] PAPAMARTZIVANOS D, GÓMEZ MÁRMOL F, KAMBOURAKIS G. Dendron:genetic trees driven rule induction for network intrusion detection systems[J]. Future Generation Computer Systems, 2018, 79:558-574.

[5] 生龙, 袁丽娜, 武南南, 等. 基于 GSA 与DE优化混合核 ELM 的网络异常检测模型[J]. 计算机工程, 2022, 48(6):146-153.
SHENG L, YUAN L N, WU N N, et al. Network anomaly detection model based on GSA and DE optimizing hybrid kernel ELM[J]. Computer Engineering, 2022, 48(6):146-153. (in Chinese)

[6] 徐国天. 网络入侵检测中 K 近邻高速匹配算法研究[J]. 信息网络安全, 2020, 20(8):71-80.
XU G T. Research on K-nearest neighbor high speed matching algorithm in network intrusion detection[J]. Netinfo Security, 2020, 20(8):71-80. (in Chinese)

[7] 王月, 江逸茗, 兰巨龙. 基于改进三元组网络和 K 近邻算法的入侵检测[J]. 计算机应用, 2021, 41(7):1996-2002.
WANG Y, JIANG Y M, LAN J L. Intrusion detection based on improved triplet network and K-nearest neighbor algorithm[J]. Journal of Computer Applications, 2021, 41(7):1996-2002. (in Chinese)

[8] COVER T, HART P. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13(1):21-27.

[9] WU X D, KUMAR V, QUINLAN J R, et al. Top 10 algorithms in data mining[J]. Knowledge and Information Systems, 2008, 14(1):1-37.

[10] GATES G. The reduced nearest neighbor rule(corresp.)[J]. IEEE Transactions on Information Theory, 1972, 18(3):431-433.

[11] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24(2):123-140.

[12] KOCEV D, VENS C, STRUYF J, et al. Tree ensembles for predicting structured outputs[J]. Pattern Recognition, 2013, 46(3):817-833.

[13] TIAN J, LI M Q, CHEN F Z, et al. Coevolutionary learning of neural network ensemble for complex classification tasks[J]. Pattern Recognition, 2012, 45(4):1373-1385.

[14] 杜磊, 杜星, 宋擒豹. 一种 k-NN 分类器 k 值自动选取方法[J]. 控制与决策, 2013, 28(7):1073-1077.
DU L, DU X, SONG Q B. An automatic selection method of k in k-NN classifier[J]. Control and Decision, 2013, 28(7):1073-1077. (in Chinese)

[15] TAVALLAEE M, BAGHERI E, LU W, et al. A detailed analysis of the KDD CUP 99 data set[C]//Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, July 8-10, 2009, Ottawa, ON, Canada. New York:IEEE Xplore, 2009:1-6.

(责任编辑:仇慧)