doi:10.3969/j.issn.1001-4616.2025.03.009

一种混合特征选择的朴素贝叶斯网络入侵检测算法

郑锦波,王慧玲

(伊犁师范大学网络安全与信息技术学院,新疆 伊宁 835000)

[**摘要**] 在入侵检测应用中,机器学习算法发挥着至关重要的作用,特征选择作为关键的数据预处理步骤,可以有效提升分类器的分类效果. 而现有的特征选择算法未考虑数据分布不均匀时特征间存在的伪相关性,影响了分类器的泛化能力. 针对此问题,本文提出了一种混合特征选择的朴素贝叶斯网络入侵检测算法,将相关性度量准则引入特征提取阶段,避免特征间存在的伪相关性,更好地满足朴素贝叶斯算法的强假设,使模型检测性能有效提升. 该方法采用了两步特征选择策略:第一步筛选数据集中和类变量相关性较强特征;第二步去除冗余特征,筛选出相互条件独立的特征作为特征子集,并将此特征子集送入朴素贝叶斯算法进行检测. 实验结果表明,提议的方法在检测率和泛化性能上都优于参与对比的 6 个传统机器学习算法,并且在一定程度上克服了数据分布不平衡导致的精度低的问题,与近期提出的两个深度学习算法相比较,在准确率和精确率上优于两个对比深度学习算法.

[**关键词**] 网络入侵检测,条件独立性,特征选择,条件互信息,pearson 相关系数
[**中图分类号**]TP309　[**文献标志码**]A　[**文章编号**]1001-4616(2025)03-0073-11

A Naive Bayes Network Intrusion Detection Algorithm with Mixed Feature Selection

Zheng Jinbo, Wang Huiling

(School of Network Security and Information Technology, Yili Normal University, Yining 835000, China)

Abstract: In intrusion detection applications, machine learning algorithms play a crucial role. Feature selection, as a key data preprocessing step, can effectively improve the classification performance of classifiers. However, existing feature selection algorithms do not consider the existence of pseudo-correlations between features when the data distribution is imbalanced, which affects the generalization ability of classifiers. To address this issue, a hybrid feature selection naive Bayes network intrusion detection algorithm is proposed, which introduces correlation measurement criteria into the feature extraction stage to avoid the pseudo-correlations between features and better satisfy the strong assumption of the naive Bayes algorithm, thereby improving the detection performance of the model. This method adopts a two-step feature selection strategy. In the first step, features that are strongly correlated with the class variable are selected from the dataset. In the second step, redundant features are removed to select a subset of mutually conditionally independent features, which are then fed into the naive Bayes algorithm for detection. Experimental results show that the proposed method outperforms 6 traditional machine learning algorithms in terms of detection rate and generalization performance, and it partially overcomes the problem of low accuracy caused by imbalanced data distribution. Compared with two recently proposed deep learning algorithms, it performs better in terms of accuracy and precision.

Key words: network intrusion detection, conditional independence, feature selection, conditional mutual information, pearson correlation coefficient

近年来,随着网络的迅速发展,网络入侵攻击行为不断增多,网络安全受到很大威胁[1]. 根据中国信息通信研究院 2020 年网络安全威胁信息研究报告[2]所述,在 2020 年全球新型冠状病毒感染爆发后,全球约有 6 431 498 台主机受到网络攻击,国内约有 880 607 台,约占全球的 1/8. 入侵检测(intrusion detection)通过收集和分析被保护系统的数据能够有效检测出入侵行为. 常用入侵检测技术包括统计学习、传统机

收稿日期:2024-12-15.
基金项目:新疆维吾尔自治区自然科学基金项目(2022D01C337)、伊犁师范学院重点学科开放课题(XJZDXKphy202302).
通讯作者:王慧玲,高级实验师,研究方向:计算机视觉,图像内容理解. E-mail:37993034@qq.com

器学习、深度学习、强化学习和可视化等[3-7],其中机器学习理论成熟,具有简单易实现、精度高等优点,已成为网络入侵检测中的应用研究热点[8-10]. 由于入侵检测数据中可能含有冗余、噪声和不相关的数据,所以增加了检测算法的计算代价,泛化性能较低. 针对此问题,学者提出了基于特征选择的检测方法,在数据集全特征中选择最优的特征子集,实现特征降维,去除冗余数据[11]. 文献[12]结合改进的粒子群算法和粗糙集理论提出了一种基于粒子群的朴素贝叶斯入侵检测方法. 文献[13]使用 KDDCup99 数据集,首先将其进行归一化、离散化处理,并进行 PCA 特征提取,后利用朴素贝叶斯分类器进行分类,检测率达到了 87%~97%. 文献[14]在 KDDCup99 数据集上使用了最小依赖和最大重要性算法(minimum dependence maximum significance,MDMS)进行特征选取,利用 KNN 来预测网络攻击,所用方法较好地识别了拒绝服务攻击和探针式攻击. 文献[15]使用 CFS、CONS 和 INTERACT 算法进行特征选择,提出了一种基于特征选择的网络入侵检测方法,文献[16]使用了基于并行 QPSO 的特征选择方法,利用朴素贝叶斯算法(NB)进行分类. 但上述方法未考虑数据分布不均匀时存在伪相关性,影响了泛化性能.

为解决上述问题,本文提出了一种混合特征选择的朴素贝叶斯网络入侵检测算法. 该方法借助于相关性度量准则,利用 pearson 相关系数和条件互信息获取数据集中与类变量强相关且彼此满足条件独立的特征子集,将 NB 算法应用到该特征子集上,适应其特征属性之间相互条件独立的假设,提升算法的检测率. 为验证混合特征选择的朴素贝叶斯网络入侵检测算法的有效性,通过使用 Friedman 检验[17]与 Nemenyi 后续检验[18]验证特征选择能够有效提升 NB 算法的检测性能,在检测率等 4 项重要指标上,NB 在本研究方法筛选出的特征子集上的分类效果较其他经典的机器学习算法有显著提升,并且在部分指标上优于目前较为先进的深度学习算法.

1 预备知识

1.1 pearson 相关系数

pearson 相关系数是由协方差演变而来的. 假设两组数据分别是 $X=\{x_1,x_2,\cdots,x_n\}$ 和 $Y=\{y_1,y_2,\cdots,y_n\}$,那么 X、Y 的期望分别为:

$$E(X)=\frac{\sum_{i=1}^{n}x_i}{n},E(Y)=\frac{\sum_{i=1}^{n}y_i}{n}, \tag{1}$$

总体协方差是:

$$\mathrm{cov}(X,Y)=\frac{\sum_{i=1}^{n}(x_i-E(X))(y_i-E(Y))}{n}. \tag{2}$$

协方差得正值时表明两个变量正相关,负值表明负相关,零值时表明相互独立,绝对值越小,两个变量相互之间的影响越小,反之越大. 根据协方差公式可知,两个变量间的相关程度是在同一物理量纲下度量的,而对于在不同量纲下的两个变量,协方差会产生较大差异,因此演变出了 pearson 相关系数. 假设有两组数据 $X=\{x_1,x_2,\cdots,x_n\}$ 和 $Y=\{y_1,y_2,\cdots,y_n\}$,那么总体期望为公式(1)所示,总体协方差为公式(2)所示,总体 pearson 相关系数定义为:

$$r_{XY}=\frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{\sum_{i=1}^{n}\frac{(x_i-E(X))}{\sigma_X}\frac{(y_i-E(Y))}{\sigma_Y}}{n}, \tag{3}$$

式中,σ_x、σ_y 分别是 X 和 Y 的标准差:

$$\sigma_X=\sqrt{\frac{\sum_{i=1}^{n}(x_i-E(X))^2}{n}},\sigma_Y=\sqrt{\frac{\sum_{i=1}^{n}(y_i-E(Y))^2}{n}}. \tag{4}$$

pearson 相关系数是两个变量进行标准化计算后得到的协方差,不受变量量纲的影响,反应了两个变量之间的线性相关程度,结合 r_{XY}的取值,当$|r_{XY}|>0.8$ 时,表示 X 和 Y 强相关;当 $0.3\leqslant|r_{XY}|\leqslant0.8$ 时,表示 X 和 Y 弱相关;当$|r_{XY}|<0.3$ 时,表示 X 和 Y 没有相关性[19].

1.2 条件互信息

相对熵用于度量两个随机分布间的距离,概率密度函数 $p(X)$ 和 $p(Y)$ 的相对熵定义为:

$$D(p||q)=\sum_{x\in X}p(x)\log\frac{p(x)}{q(x)}=E_p\log\frac{p(X)}{q(X)},\tag{5}$$

由式(5)观察可得,$p(X)$ 和 $q(X)$ 两个分布越接近,其相对熵的值就越小,当两个分布完全相同时,相对熵的值等于 0. 由此可知,可以利用相对熵度量两个分布之间的相似性.

互信息可以用于衡量两个随机变量之间的独立程度. 对于随机变量 X 和 Y,$p(X,Y)$ 表示它们的联合分布,边缘分布分别为 $p(X)$ 和 $p(Y)$. 互信息 $I(X;Y)$ 是联合分布 $p(X,Y)$ 与边缘分布 $p(X)p(Y)$ 的相对熵,即

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}=D(p(x,y)||p(x)p(y)),\tag{6}$$

随机变量 Y 的熵定义为 $H(Y)$. 在给定变量 X 的条件下,Y 的条件熵定义为 $H(Y|X)$,则互信息与熵的关系如下式:

$$I(X;Y)=H(Y)-H(Y|X).\tag{7}$$

由式(7)可知,在给定 X 的条件下,互信息可以看成是 Y 的不确定程度的消减量. 如果 X 和 Y 相互独立,则 $H(Y|X)=H(Y)$,那么互信息 $I(X;Y)=0$.

条件互信息涉及 3 个变量间的关系,计算量较互信息大,但在信息的度量上更为准确,在给定条件 Z 的情况下,条件互信息定义如下:

$$I(X;Y|Z)=\sum_{z\in Z}p(z)\sum_{x\in X}\sum_{y\in Y}p(x,y|z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)},\tag{8}$$

条件互信息和条件熵关系如下:

$$I(X;Y|Z)=H(X|Z)-H(X|Y,Z).\tag{9}$$

1.3 朴素贝叶斯分类器

朴素贝叶斯分类器结构简单,计算效率高,被应用于各领域中,但其需要“属性条件独立性假设”支持,要求各个属性单独影响分类结果. 假设数据集 D 包括含 n 个属性,记作 $A=\{A_1,A_2,\cdots,A_n\}$,实例 $d\in D$ 表示为 $d=(a_1,a_2,\cdots,a_n)$. 其中,a_i 是属性 A_i 的取值. 类别用 C 表示. 则贝叶斯公式如下:

$$P(C|X)=\frac{P(X|C)P(C)}{P(X)},\tag{10}$$

式中,类先验概率记为 $P(C)$;$P(X|C)$(或称为“拟然”)是样本 X 相对于类标记 C 的条件概率;$P(X)$ 是用于归一化的“证据”因子,根据朴素贝叶斯的条件独立性假设,可得下式:

$$P(C|X)=\frac{P(X|C)P(C)}{P(X)}=\frac{P(C)}{P(X)}\prod_{i=1}^{d}P(x_i|C),\tag{11}$$

式中,$P(x_i|C)$ 可以从训练数据集中求得. 证据因子 $P(X)$ 在给定样本 X 时与类标记无关,朴素贝叶斯分类器将 X 预测为具有最大后验概率 $P(C|X)$ 的类. 由此,朴素贝叶斯分类器描述为:

$$h_{nb}(x)=\arg\max_{c\in y}P(C)\prod_{i=1}^{d}P(x_i|C).\tag{12}$$

2 PCMI-NB

当特征数量过多、超过一定限度时,分类器的性能会受到影响[15]. 以 KDDCup99 数据集为例,通过在 10%的 KDDCup99 训练数据集上生成特征与标记以及特征和特征之间的相关性热力图,其中,$a_1,a_2,\cdots,a_{41}$ 表示了 41 个特征,类标用 label 表示. 如图 1 所示,和 label 相关性强的特征,它们之间也存在强相关性,比如 a_2,a_3. 从因果性角度解释,这些特征和 label 间可能存在伪相关性,其包含的信息会降低分类效果.

通过有效的特征选择,可以去除这些冗余特征或噪声特征,从而提高检测的准确性和泛化能力,降低训练时间.

基于上述问题,本文提出 PCMI 特征选择算法进行特征选择,选取与 label 真相关的特征,进而实现基

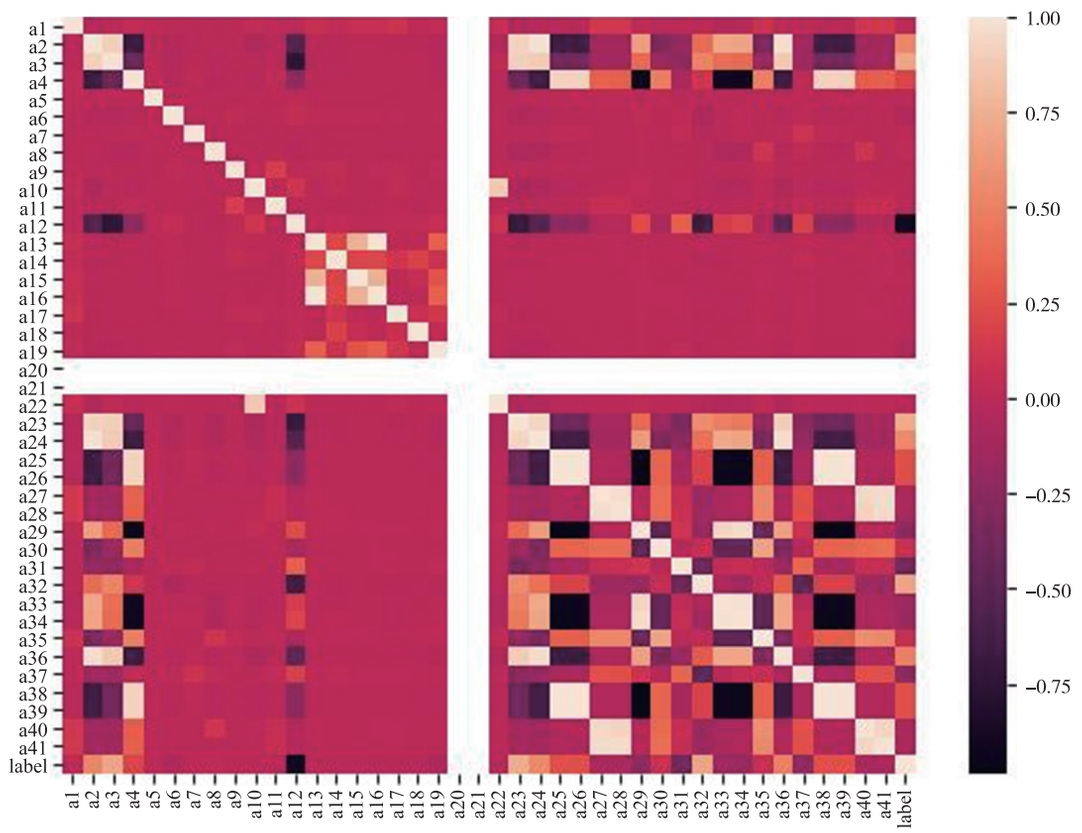


图 1　pearson 相关系数结果热力图

Fig. 1　Pearson correlation coefficient result heatmap

于特征选择的朴素贝叶斯网络入侵检测方法.

定义 1　类变量相关集

如果特征 X_i 和类变量 C 的相关性大于指定阈值,则称特征 X_i 是与类变量 C 相关,把这样的特征 X_i 构成的集合称为类变量相关集.

本文中使用了 person 相关系数作为度量变量之间是否相关的度量函数. 得到如下集合:

$$P=\{X\mid\ |r_{XC}|>0.5\}. \tag{13}$$

定义 2　条件于类变量 C 的相关集

给定类变量 C,如果特征 X 和 Y 的条件互信息小于指定阈值,把这样的特征 X 和 Y 构成的集合称为条件于类变量 C 的相关集.

对于该定义,本文使用了条件互信息作为度量特征之间相关性的函数. 得到如下集合:

$$PM=\{X,Y|I(X;Y|C)_{X,Y\in P}<\text{average}\}, \tag{14}$$

式中,$\text{average}=\dfrac{\sum_{X,Y\in P}I(X;Y\mid\ C)}{\text{len}(P)}$.

基于定义 1 和定义 2,会得到与类变量相关且相互条件独立的特征集 PM. $P(x_i|C)$ 可以从训练集 PM 中求得,证据因子 $P(X)$ 在给定样本 X 时与类标记无关,朴素贝叶斯分类器将 X 预测为具有最大后验概率 $P(C|X)$ 的类. 由此,朴素贝叶斯分类器描述为公式(15):

$$h_{nb}(x)=\arg\max_{c\in y}P(C)\prod_{x_i\in PM}P(x_i|C). \tag{15}$$

显然,从公式(15)中可以看出,用于分类模型的特征是和 label 相关且条件独立的特征.

基于公式(15),提议的网络入侵检测方法总体框架如图 2 所示,其中,PCMI 特征选择算法包括两部分:第一部分进行 person 相关系数计算,保留与类相关系数绝对值大于 0.5 的特征,记为 D_P;第二部分输入 D_P,进行条件互信息计算,选取条件互信息数值小于平均值,即相对独立的特征作为最终的特征子集 D_{PM},对特征

子集 D_{PM}归一化处理后,送入朴素贝叶斯分类器进行分类. 本文提出了两种条件互信息计算方法,并进行测试比较.

条件互信息计算方法 1:如算法 1 所示,通过计算训练集 D 各特征与类的 person 相关系数,获得与类相关性较强的特征子集 D_P,将特征子集 D_P 作为输入,计算各特征以类标为条件的条件互信息,预留最小值,形成最小条件互信息集合 CMI,并计算最小条件互信息平均值,在 CMI 中选取小于最小互信息平均值的特征作为最终特征子集 D_{PM}.

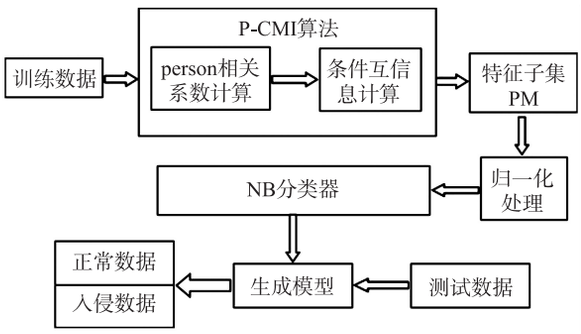


图 2　总体框架

Fig. 2　Overall framework

算法 1　结合条件互信息计算方法 1 的 PCMI 特征选择算法

输入:训练集 D
输出:特征子集 D_{PM}
Begin
PCMI( ){
$D_{PM}=\varnothing$/ * 初始化特征子集 * /
$D_P=\varnothing$
输入训练数据集
计算训练集 D 各特征与类的 person 相关系数
For $i=1:m$/ * m 为特征维数,遍历特征与类之间的相关系数 * /
　If $|P(i)|\geqslant 0.5$/ * 相关系数绝对值大于 0.5 * /
　　$D_P=D_P\leftarrow D_i$
End For
输入 D_P,
For $n=1:h$/ * 唯一选定 X_i,h 是 D_P 的维数 * /
　For $r=(n+1):h$/ * 选取特征 X_k * /
　　计算条件互信息 $I(X_i;X_k|C)$
　CMI←最小条件互信息
End For
average←求取 CMI 平均值
For $s=1:h$/ * h 是 D_P 的维数 * /
　If $DP(s)<$average
　$D_{PM}\leftarrow Ds(s)$
　End For
返回特征子集 D_{PM}}

条件互信息计算方法 2:如算法 2 所示,通过计算训练集 D 各特征与类的 person 相关系数,获得与类相关性较强的特征子集 D_P,将特征子集 D_P 作为输入,计算各特征以类标为条件的条件互信息和其平均值,预留小于平均值的特征,形成特征子集 D_{PM}.

算法 2　结合条件互信息计算方法 2 的 PCMI 特征选择算法

输入:训练集 D
输出:特征子集 D_{PM}
Begin
PCMI( ){
$D_{PM}=\varnothing$/ * 初始化特征子集 * /
$D_P=\varnothing$
输入训练数据集
计算训练集 D 各特征与类的 pearson 相关系数

7 For $i=1:m$/ * m 为特征维数,遍历特征与类之间的相关系数 * /
8 If $|P(i)|\geqslant 0.5$/ * 相关系数绝对值大于 0.5 * /
9 $D_P=D_P\leftarrow D_i$
10 End For
11 输入 D_P,
12 For $n=1:h$/ * 唯一选定 X_i,h 是 D_P 的维数 * /
13 For $r=1:h$/ * 选取特征 X_k * /
14 If $r==n$:continue
15 计算条件互信息 $I(X_i;X_k|C)$
16 $X_j\leftarrow$条件互信息数值小于平均值的特征
17 End For
18 $D_{PM}\leftarrow$计算全部小于平均值特征的并集
19 返回特征子集 D_{PM}}

3 实验及结果

3.1 数据集

本文实验选用 KDDCup99 和 UNSW-NB15 数据集验证所提网络入侵检测方法的性能. KDDCup99 数据集是网络入侵检测领域的标准数据集,它包括 41 个网络连接特征(详见表 1)和 1 个类标签,包括基本网络连接特征、网络连接内容特征和网络连接流量特征,数据集中共 500 万余条记录. UNSW-NB15 数据集在 2015 年发布,广泛应用在入侵检测领域,数据集共 49 个网络连接特征(详见表 1),254 万余条记录,为了便于验证本文所提算法,本次验证实验中使用的是 10%的 KDDCup99 数据集和 UNSW-NB15 数据集的 10%抽样数据[20].

表 1 数据集特征编号及对应的特征名

Table 1 Dataset feature numbers and corresponding feature names

数据集	编号	特征名
KDDCup99	$a_1,a_2,a_3,\cdots,a_{41}$	"duration","protocol_type","service","flag","src_bytes","dst_bytes","land","wrong_fragment","urgent","hot","num_failed_logins","logged_in","num_compromised","root_shell","su_attempted","num_root","num_file_creations","num_shells","num_access_files","num_outbound_cmds","is_host_login","is_guest_login","count","srv_count","serror_rate","srv_serror_rate","rerror_rate","srv_rerror_rate","same_srv_rate","diff_srv_rate","srv_diff_host_rate","dst_host_count","dst_host_srv_count","dst_host_same_srv_rate","dst_host_diff_srv_rate","dst_host_same_src_port_rate","dst_host_srv_diff_host_rate","dst_host_serror_rate","dst_host_srv_serror_rate","dst_host_rerror_rate","dst_host_srv_rerror_rate"
UNSW-NB15	$b_1,b_2,b_3,\cdots,b_{49}$	"srcip","sport","dstip","dsport","proto","state","dur","sbytes","dbytes","sttl","dttl","sloss","dloss","service","sload","dload","spkts","dpkts","swin","dwin","stcpb","dtcpb","smeansz","dmeansz","trans_depth","res_bdy_len","sjit","djit","stime","ltime","sintpkt","dintpkt","tcprtt","synack","ackdat","is_sm_ips_ports","ct_state_ttl","ct_flw_http_mthd","is_ftp_login","ct_ftp_cmd","ct_srv_src","ct_srv_dst","ct_dst_ltm","ct_src_ltm","ct_src_dport_ltm","ct_dst_sport_ltm","ct_dst_src_ltm","attack_cat","Label"

3.2 特征选择

特征选择的主要目的是筛选出影响分类效果的重要特征,针对朴素贝叶斯算法,与类标相关性较强且各特征之间相关性较小的特征将对其分类效果影响更大. 以 10%的 KDDCup99 数据集为例,计算 KDDCup99 训练集 D 中 41 个特征之间及 41 个特征和类标的 pearson 相关系数,选取 41 个特征中与类标相关系数绝对值大于等于 0.5 的特征,获得与类相关性较强的特征子集 $D_P=\{a_2,a_3,a_{12},a_{23},a_{24},a_{32},a_{36}\}$,选取的 7 个特征相较其他特征和类标相关性更强,对分类效果将影响更大. 在此基础上计算特征子集 D_P 中各特征以类为条件的条件互信息,结果如图 3 所示. 可以发现存在给定 label,特征之间仍然存在强相关性,如 $a2$,$a3$. 这违背了朴素贝叶斯的强假设.

根据提议的 PCMI 特征选择算法 1 和算法 2,得到了如表 2 和表 3 所示的 KDDCup99 特征选择结果,其中带有下划线的特征选择的特征,和图 3 的热力图结果相比较,不难发现,选择的特征没有违背朴素贝叶斯的强条件假设. 结合 PCMI 特征选择算法,在 UNSW-NB15 数据集上对 49 个特征进行特征选择,选择

结果如表 4 和表 5 所示,选择特征分别为(b_4,b_{20},b_{23},b_{34},b_{35})和(b_4,b_9,b_{10},b_{20},b_{23},b_{28},b_{32},b_{34},b_{35},b_{36}).

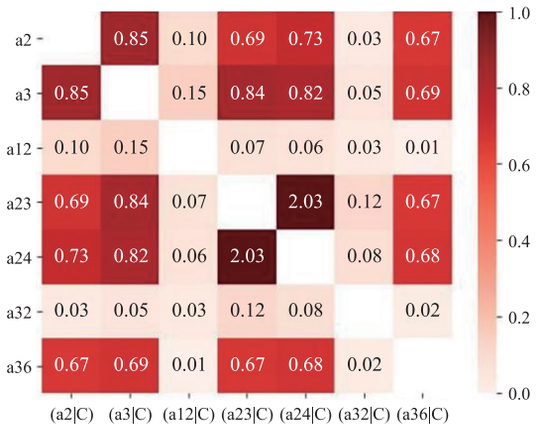


图 3　KDDCup99 特征子集 D_P 中各特征以类标为条件的条件互信息计算结果

Fig. 3　The conditional mutual information calculation results of each feature in the KDDCup99 feature subset D_P based on class labels

表 2　算法 1 KDDCup99 数据集特征选取结果

Table 2　Feature selection results of algorithm 1 KDDCu99 dataset

特征序号	a2	a3	a12	a23	a24	a32	a36	average
最小条件互信息	0.03	0.05	0.01	0.12	0.08	0.02	—	0.05

表 3　算法 2 KDDCup99 数据集特征选取结果

Table 3　Feature selection results of algorithm 2 KDDCup99 dataset

特征序号	a2	a3	a12	a23	a24	a32	a36	average
a2	—	0.85	0.10	0.69	0.73	0.03	0.67	0.48
a3	0.85	—	0.15	0.84	0.82	0.05	0.69	0.56
a12	0.10	0.15	—	0.07	0.06	0.03	0.01	0.08
a23	0.69	0.84	0.07	—	2.03	0.12	0.67	0.75
a24	0.73	0.82	0.06	2.03	—	0.08	0.68	0.74
a32	0.03	0.05	0.03	0.12	0.08	—	0.04	0.05
a36	0.67	0.69	0.01	0.67	0.68	0.04	—	0.54

表 4　算法 1 UNSW-NB15 数据集特征选取结果

Table 4　Feature selection results of algorithm 1 UNSW-NB15 dataset

特征序号	b4	b9	b10	b13	b20	b23	b28	b32	b34	b35	b36	average
最小条件互信息	0.02	0.05	0.03	0.11	0.01	0	0.03	0.04	0.01	0.01	—	0.02

表 5　算法 2 UNSW-NB15 数据集特征选取结果

Table 5　Feature selection results of algorithm 2 UNSW-NB15 dataset

特征序号	b4	b9	b10	b13	b20	b23	b28	b32	b34	b35	b36	average
b4	—	0.26	0.03	0.17	0.17	0.04	0.03	0.16	0.1	0.02	0.03	0.101
b9	0.26	—	0.15	0.41	0.41	0.17	0.09	0.39	0.32	0.43	0.05	0.268
b10	0.03	0.15	—	0.5	0.49	0.16	0.1	0.45	0.4	0.46	0.1	0.284
b13	0.17	0.41	0.5	—	0.49	0.2	0.11	0.51	0.43	0.57	0.11	0.35
b20	0.17	0.41	0.49	0.49	—	0.05	0.01	0.04	0.01	0.08	0.01	0.176
b23	0.04	0.17	0.16	0.2	0.05	—	0.01	0.04	0.01	0.06	0	0.074
b28	0.03	0.09	0.1	0.11	0.01	0.01	—	0.37	0.31	0.4	0.06	0.149
b32	0.16	0.39	0.45	0.51	0.04	0.04	0.37	—	0.35	0.44	0.04	0.279
b34	0.1	0.32	0.4	0.43	0.01	0.01	0.31	0.35	—	0.08	0.01	0.202
b35	0.02	0.43	0.46	0.57	0.08	0.06	0.4	0.44	0.08	—	0.01	0.255
b36	0.03	0.05	0.1	0.11	0.01	0	0.06	0.04	0.01	0.01	—	0.042

3.3　混合特征选择的朴素贝叶斯网络入侵检测性能对比分析

为评估本文提出的网络入侵检测方法的有效性,使用 Friedman 检验与 Nemenyi 后续检验对比传统机器学习算法在数据集上的性能. 根据表 6 混淆矩阵,使用评估指标:检测率(Recall=TP/(TP+FN))、准确率(Accuracy=(TP+TN)/(TP+TN+FP+FN))、精确率(Precision=(TP/(TP+FP))和 F1-score(F1-score=2×Precision×Recall/(Precision+Recal))来评估 NB 在特征子集上和其他机器学习算法在全特征集上的分类性能,使用 ROC 曲线和 AUC 评估泛化能力.

表 6　混淆矩阵

Table 6　Confusion matrix

真实情况	预测结果	
	正例	反例
正例	真正例(TP)	假反例(FN)
反例	假正例(FP)	真反例(TN)

3.3.1　混合特征选择对各分类算法的影响

为验证混合特征选择方法对朴素贝叶斯的影响,使用 Friedman 检验与 Nemenyi 后续检验方法在 KDDCup99 数据集和 UNSW-NB15 数据集上分别对传统机器学习算法,包括朴素贝叶斯(NB)、贝叶斯网络(BN)、逻辑回归(LR)、决策树(DT)、自适应增强(AdaBoost)、随机森林(RF)和 k 邻近(KNN)进行性能对比,如图 4 和图 5 所示,在未作特征选择的前提下,7 个机器学习算法性能没有明显差异,其中 NB 算法性能序值排名分别为 6 和 4,主要原因在于数据集中部分特征之间存在相关性,影响了 NB 算法性能. 结合 PCMI 两个特征选择方法,将两个特征子集作为两个数据集,送入 Friedman 检验与 Nemenyi 后续检验,得到的检验结果如图 6 和图 7 所示,NB 算法在两个特征子集上平均序值排名为 1,在两个数据集上,其性能相较其他参与比较的机器学习算法最好,由此可以推断,通过 PCMI 两个特征选择方法提取的特征对 NB 算法影响较大,去除了具有相关性的特征,降低了特征属性之间的相关性,更好地满足了 NB 算法的特征属性之间相互条件独立的假设.

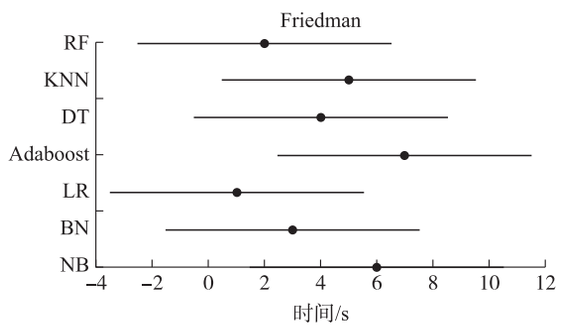


图 4　KDDCup99 数据集全特征下检验图

Fig. 4　KDDCup99 dataset full feature inspection map

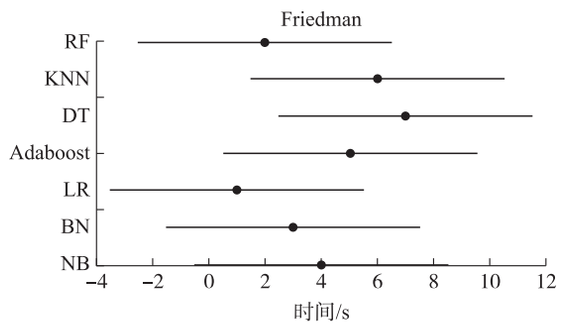


图 5　UNSW-NB15 数据集全特征下检验图

Fig. 5　UNSW-NB15 dataset full feature test plot

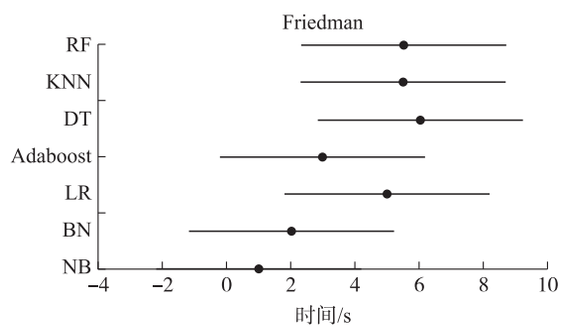


图 6　KDDCup99 数据集特征选择后检验图

Fig. 6　KDDCup99 dataset feature selection validation map

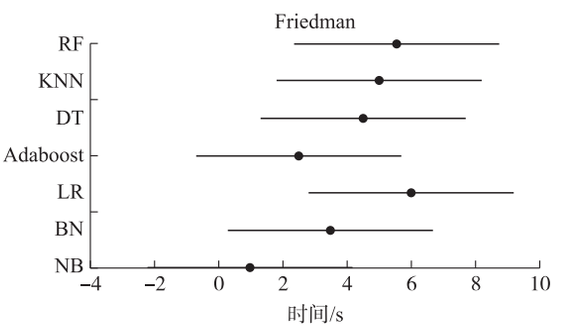


图 7　UNSW-NB15 数据集特征选择后检验图

Fig. 7　UNSW-NB15 dataset feature selection post validation plot

3.3.2　消融实验

为评估混合特征选择的朴素贝叶斯网络入侵检测方法的性能,特设置消融实验. 在相同条件下,分别使用未特征选择的 NB 算法和混合特征选择的 NB 算法在 KDDCup99 数据集和 UNSW-NB15 数据集上进行对比,检测结果如表 7 和表 8 所示,特征选择后,NB 算法的检出率等 4 项评估指标都有所提升,其中

PCMI 两个特征选择方法中,方法 1 的效果最好. 由此验证了特征选择算法在 NB 分类器下的有效性. 另外,特征选择后,输入的特征较少,NB 分类器的计算时间也比无特征选择的计算时间有所减少. 无特征选择和特征选择下 NB 入侵检测的 ROC 曲线和 AUC 如图 8 和图 9 所示,无特征选的情况下,特征维数较多,噪声过多,致使 NB 分类器过拟合,造成泛化能力下降,特征选择后,选取的特征和类标相关性较强且特征间相关性较小,使 NB 分类器能够更好地拟合数据,有更好的泛化能力,其中通过方法 1 的特征选择后,NB 泛化能力最好.

表 7　在 KDDCup99 数据集上消融实验检测结果

Table 7　Detection results of ablation experiment on the KDDCup99 dataset

特征数量	Recall/%	Accuracy/%	Precison/%	F1-score/%
41	92.32	92.94	98.84	95.47
6	97.14	97.35	99.55	98.33
3	98.53	97.52	99.90	99.21

表 8　在 UNSW-NB15 数据集上消融实验检测结果

Table 8　Detection results of ablation experiments on the UNSW-NB15 dataset

特征数量	Recall/%	Accuracy/%	Precison/%	F1-score/%
49	67.54	75.92	74.04	70.64
10	76.00	80.26	79.61	77.76
5	79.01	82.72	83.04	80.97

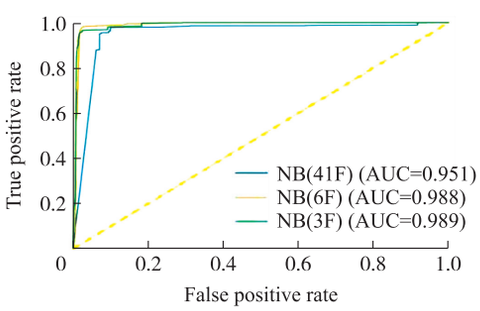


图 8　在 KDDCup99 数据集上无特征选择和特征选择下 NB 入侵检测 ROC 曲线和 AUC 比较

Fig. 8　Comparison of ROC curve and AUC of NB intrusion detection on the KDDCup99 dataset with and without feature selection

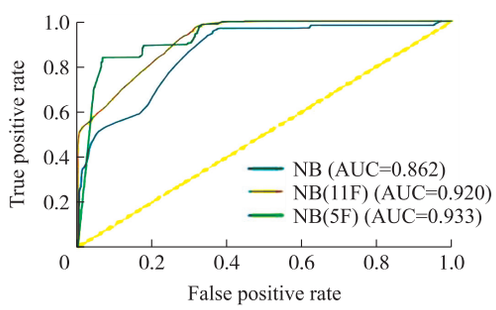


图 9　在 UNSW-NB15 数据集上无特征选择和特征选择下 NB 入侵检测 ROC 曲线和 AUC 比较

Fig. 9　Comparison of ROC curve and AUC of NB intrusion detection on the UNSW-NB15 dataset with and without feature selection

3.3.3　与基于机器学习和基于深度学习的入侵检测方法对比试验

为评估混合特征选择的朴素贝叶斯检测方法与传统的机器学习算法的泛化能力,在 KDDCup99 数据集和 UNSW-NB15 数据集上,将使用方法 1 进行特征选择后的 NB 和在全特征下 6 种分类器进行对比. 如表 9 和表 10 所示,可以看出,由于存在特征冗余,对分类器产生了影响,6 种分类器分类效果整体较低,对比 ROC 曲线和 AUC,如图 10 和图 11 所示,特征选择后的 NB 分类器的泛化能力优于全特征下的其他 6 类分类器,AUC 值分别达到了 0.989 和 0.933.

表 9　在 KDDCup99 数据集上特征选择后的 NB 和其他分类器的入侵检测结果对比

Table 9　Comparison of intrusion detection results between feature selected NB and other classifiers on the KDDCup99 dataset

指标	NB(3F)	BN	LR	DT	AdaBoost	RF	KNN
Recall/%	98.53	97.54	81.85	91.2	92.70	91	91.4
Accuracy/%	97.51	92.30	97.31	93.72	93.75	92.61	93.83
Precison/%	99.90	93.21	98.84	98.6	99.40	98.7	98.80
F1-score/%	99.21	95.33	89.55	95.2	96.00	95.2	95.4

表 10　在 UNSW-NB15 数据集上特征选择后的 NB 和其他分类器的入侵检测结果对比

Table 10　Comparison of intrusion detection results between NB and other classifiers after feature selection on the UNSWNB15 dataset

指标	NB(5F)	BN	LR	DT	AdaBoost	RF	KNN
Recall/%	79.01	70.1	68.34	61.89	69.64	66.93	69.43
Accuracy/%	82.72	76.12	71.7	64.37	76.22	71.82	77.43
Precison/%	83.04	76.29	67.22	59.33	77.33	70.02	75.55
F1-score/%	80.97	73.06	68.31	60.58	73.17	68.44	72.36

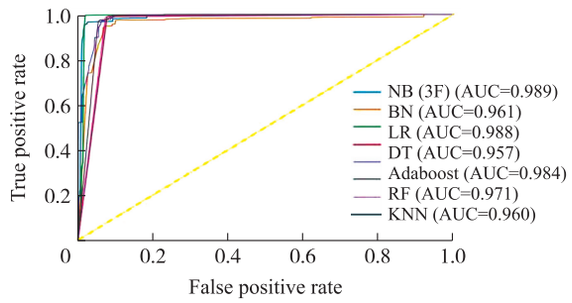


图 10　KDDCup99 数据集下特征选择后的 NB 和全特征下其他分类器 ROC 曲线和 AUC 比较

Fig. 10　Comparison of ROC curves and AUC of NB after feature selection and other classifiers under full feature in the KDDCup99 dataset

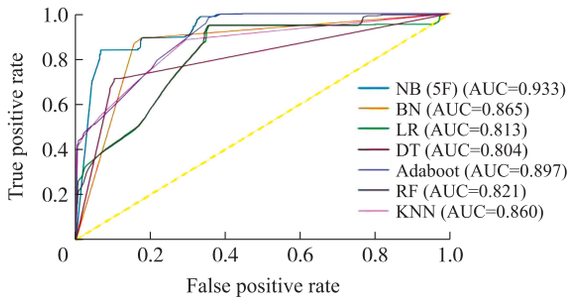


图 11　在 UNSW-NB15 数据集下特征选择后的 NB 和全特征下其他分类器 ROC 曲线和 AUC 比较

Fig. 11　Comparison of ROC curves and AUC of NB after feature selection and other classifiers under full feature in the UNSW-NB15 dataset

为验证本文所提算法的有效性,在 UNSW-NB15 数据集上与近期提出的深度学习算法 GRU-RNN[21] 和 CNN-BiLSTM[22] 进行比较,比较结果如表 11 所示,对比发现,本文所提算法在准确率和精确率上优于两个对比深度学习算法,但在召回率和 F1-score 上性能稍弱,主要原因是通过特征选择,有效地拉开了正常样本和攻击样本之间的类间距,辅助 NB 分类器取得了更好的检测性能,而深度学习模型参数较多,存在过拟合的现象.

表 11　与基于深度学习算法的入侵检测方法对比结果

Table 11　Comparison results with intrusion detection methods based on deep learning algorithms

特征数量	Recall/%	Accuracy/%	Precison/%	F1-score/%
GRU_RNN	81.99	82.59	79.75	81.14
CNN_BiLSTM	82.08	82.68	80	81.32
NB(5F)	79.01	82.72	83.04	80.97

4　结论

本文提出了一种混合特征选择的朴素贝叶斯网络入侵检测算法,使用 PCMI 算法筛选出与类标线性相关且相对独立的特征子集,利用 NB 分类器进行入侵检测分类. 在 10%的 KDDCup99 数据集和 UNSW-NB15 数据集的 10%抽样数据上进行实验结果验证,从横向和纵向对比了特征选择后的 NB 分类器和未特征选取的 6 类常用机器学习算法的分类效果. 同时,在 UNSW-NB15 数据集上与近期提出的深度学习算法进行了比较. 实验结果表明,特征选择后 NB 的分类效果和泛化能力最优,并且在一定程度上克服了样本分布不平衡的问题. 本文工作主要研究了入侵检测中的特征选择,采用了传统分类器,特征选择主要考虑的是特征与类标线性相关性和特征间的独立性,忽略了特征间及特征与类标间其他方面的相关性,这也是未来研究工作的一个重点.

[参考文献]

[1] WU J X,LI J H,JI X S. Security for cyberspace:Challenges and opportunities[J]. Frontiers of information technology & electronic engineering,2018,19(12):1459-1461.

[2] 中国信息通信研究院安全研究所. 2020 年网络安全威胁信息研究报告[EB/OL]. [2021-12-03]. http://www.caict.ac.cn/kxyj/qwfb/ztbg/202112/P020211203576374176759.pdf.

[3] MOHAMED M B,MEDDEB-MAKHLOUF A,FAKHFAKH A. Intrusion cancellation for anomaly detection in healthcare applications[C]//International Wireless Communications & Mobile Computing Conference(IWCMC). Split Croatia:IEEE,2019:313-318.

[4] GHOSH K,NEOGY S,DAS P K,et al. Intrusion detection at international borders and large military barracks with multi-sink wireless sensor networks:An energy efficient solution[J]. Wireless pearsonal communications,2018,98(1):1083-1101.

[5] GAO L L,LI F,XU X,et al. Intrusion detection system using SOEKS and deep learning for in-vehicle security[J]. Cluster computing,2019,22(S6):14721-14729.

[6] HU Y,YANG A,LI H,et al. A survey of intrusion detection on industrial control systems[J]. International journal of distributed sensor networks,2018,14(8):155014771879461.

[7] HALDER S,GHOSAL A,CONTI M. Efficient physical intrusion detection in internet of things:A node deployment approach [J]. Computer networks,2019,154:28-46.
[8] SENTHILNAYAKI B,VENKATALAKSHMI K,KANNAN A. Intrusion detection system using fuzzy rough set feature selection and modified KNN classifier[J]. The internation arab journal of information technology,2019,16(4):746-753.
[9] AUNG Y Y,MIN M M. Hybrid intrusion detection system using K-Means and K-Nearest neighbors algorithms[C]//ACIS International Conference on Computer and Information Science(ICIS). Osaka,Japan:IEEE,2018:34-38.
[10] SAHU S K,KATIYAR A,KUMARI K M,et al. An SVM-based ensemble approach for intrusion detection[J]. International journal of information technology and Web engineering,2019,14(1):66-84.
[11] KAUR A, GULERIA K, TRIVEDI N K. Feature selection in machine learning: methods and comparison [C]//2021 International Conference on Advance Computing and Innovative Technologies in Engineering(ICACITE). Greater Noida, India:IEEE,2021:789-795.
[12] 任晓奎,缴文斌,周丹. 基于粒子群的加权朴素贝叶斯入侵检测模型[J]. 计算机工程与应用,2016,52(7):122-126.
[13] YU N. A novel selection method of network intrusion optimal route detection based on Naive Bayesian[J]. International journal of applied decision sciences,2018,11(1):1
[14] LIANG J,MA M,SADIQ M,et al. A filter model for intrusion detection system in vehicle Ad Hoc Networks:A hidden Markov Methodology[J]. Knowledge-based systems,2019,163:611-623.
[15] 戴远飞,陈星,陈宏,等. 基于特征选择的网络入侵检测方法[J]. 计算机应用研究,2017,34(8):2429-2433.
[16] 戴敏. 基于并行特征选择和分类的网络入侵检测方法[J]. 计算机工程与设计,2019,40(3):654-661.
[17] DEMSAR J. Statistical comparisons of classifiers over multiple data sets[J]. The journal of machine learning research,2006,7:1-30
[18] NEMENYI P B. Distribution-free multiple comparisons[M]. USA:Princeton University,1963.
[19] 赵洪斌. 多时间尺度下基于相关系数的光伏电站出力特性分析[J]. 青海大学学报,2020,38(5):60-65.
[20] KOCHER G,GUISHAN K A. Analysis of machine learning algo-rithms with feature selection for intrusion detection using UNSW-NB15 dataset[J]. International journal of network security & its applications,2021,13(1):21-31.
[21] 李俊,夏松竹,兰海燕,等. 基于 GRU-RNN 的网络入侵检测方法[J]. 哈尔滨工程大学学报,2021,42(6):879-884.
[22] JAYSM,MANOLLAS. Efficient deep CNN-BiLSTM model for network intrusion detection[C]//Proceedings of the 3rd International Conference on Artificial Intelligence and Pattern Recognition. Xiamen:Semantic Scholoar,2020.

[责任编辑:黄　敏]