第 68 卷 第 1 期
2022 年 2 月

武汉大学学报(理学版)
J. Wuhan Univ. (Nat. Sci. Ed.)

Vol. 68 No. 1
Feb. 2022,65~72

DOI: 10.14188/j.1671-8836.2021.0111

卡方校正算法对入侵检测特征选择的优化

刘　云,郑文凤†,张　轶

昆明理工大学 信息工程与自动化学院,云南 昆明 650500

收稿日期:2021-05-06　　† 通信联系人 E-mail:2867328528@qq.com

基金项目:国家自然科学基金(61761025),云南省重大科技专项计划项目资助(202002AD080002)

作者简介:刘　云,男,副教授,现从事数据挖掘、深度学习等方向研究。 E-mail:liuyun@kmust.edu.cn

摘　要:在入侵检测系统中互信息特征选择标准可快速选择重要特征,通常信息熵的计算偏差会降低系统的分类性能。为了减少特征选择偏差的影响,提出卡方校正算法(chi-square correction algorithm,CSCA)。首先,对所有候选特征进行离散化处理,计算互信息特征选择相关标准的偏差;然后,将偏差项添加在特征选择目标函数中,通过CSCA优化离散化水平和特征偏差;最后,在更新后的特征集中选择当前最重要的特征子集,在分类模型中检测攻击。仿真结果表明,与MIGM(mutual information gain maximize)算法和M-DFIFS(M-dynamic feature importance based feature selection)算法相比,卡方校正算法提高了入侵检测系统的精度,同时降低了系统的误报率。

关　键　词:入侵检测;特征选择;互信息;离散化;卡方检验

中图分类号:TP311　　**文献标志码**:A　　**文章编号**:1671-8836(2022)01-0065-08

Optimization of Chi-Square Correction Algorithm for Intrusion Detection Feature Selection

LIU Yun, ZHENG Wenfeng†, ZHANG Yi

Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, Yunnan, China

Abstract: Mutual information feature selection criteria can be used in intrusion detection systems to quickly select important features. Usually, the calculation deviation of information entropy will reduce the classification performance of the system. In order to reduce the influence of feature selection deviation, a chi-square correction algorithm (CSCA) is proposed. First, we discretize all candidate features and calculate the deviation of the relevant criteria for mutual information feature selection; secondly, we add the deviation term to the feature selection objective function, and optimize the discretization level and feature deviation through the CSCA; finally, in the updated feature set, the most important feature subset is selected to detect attacks in the classification model. The simulation results show that, compared with the MIGM (mutual information gain maximize) algorithm and the M-DFIFS (M-dynamic feature importance based feature selection) algorithm, the chi-square correction algorithm improves the accuracy of the intrusion detection system and reduces the false alarm rate of the system.

Key words: intrusion detection; feature selection; mutual information; discretization; chi-square test

0　引　言

数据特征选择在大数据挖掘和分析及机器学习等领域都有广泛应用,尤其在入侵检测系统(intrusion detection system,IDS)中,选择重要数据特征是识别网络攻击准确性的关键因素[1]。为了给分类模型构建高效的特征选择(feature selection,FS)算法,通常用基于互信息(mutual information,MI)的方法学习和估算

引用格式:刘云,郑文凤,张轶. 卡方校正算法对入侵检测特征选择的优化[J]. 武汉大学学报(理学版),2022,68(1):65-72. DOI:10.14188/j.1671-8836.2021.0111.

LIU Yun,ZHENG Wenfeng,ZHANG Yi. Optimization of Chi-Square Correction Algorithm for Intrusion Detection Feature Selection [J]. *J Wuhan Univ (Nat Sci Ed)*,2022,68(1):65-72. DOI:10.14188/j.1671-8836.2021.0111(Ch).

高维数据特征的重要性,该方法可以快速找到数据特征中的相关信息,选择出最大相关最小冗余(maximum relevance minimum redundancy, mRMR)的特征子集[2,3]。

Wang等[4]结合数据特征的关联性,构建了互信息增益最大化(mutual information gain maximize, MIGM)算法,通过等效分区概率对特征数据进行划分,并用最大联合互信息准则对候选特征进行评估。该方法可以快速识别有效的特征子集,比传统方法具有更好的适用性。Wei等[5]根据动态特征重要性(dynamic feature importance, DFI)指标,提出基于MI的动态特征重要性特征选择(M-dynamic feature importance based feature selection, M-DFIFS)算法,使用最大信息系数有效排除冗余特征,再通过随机森林的基尼系数选出重要特征,所选的特征数据与类要素之间有更强的相关性。但是这些方法都没有考虑信息熵的偏差校正。

为了在入侵检测系统中减少信息熵偏差对分类性能的影响,构建实时可靠的IDS,本文提出了卡方校正算法(chi-square correction algorithm, CSCA)。首先对所有候选特征进行离散化处理,根据特征的潜在概率分布估计特征选择相关标准的偏差;然后,通过考虑偏差的目标函数得到特征选择的临界值;最后,通过χ^2检验优化离散化水平和特征偏差,更新特征子集。

1 入侵检测特征选择

1.1 入侵检测特征选择模型

选择重要相关的特征并及时做出数据更新对IDS防御攻击非常重要。特征选择的目的就是在所有候选特征集F中选出最优特征子集$\boldsymbol{f}=\{f_1,f_2,\cdots,f_s\}$,为分类模型提供在类之间具有最大区分力的特征数据,并降低数据维度[6]。针对高维数据流中存在的特征变化,基于MI的特征选择方法可减少无关特征和冗余特征对模型的影响。

为了简化特征数据,通常在特征选择之前把连续值的特征转换为离散特征。在图1入侵检测模型中,将数据流中提取的特征和训练后的标准特征动态离散化为n个区间,则离散化间隔为$\{[d_0,d_1],\cdots,[d_{n-1},d_n]\}$,特征$f_i$的最小值为$d_0$,最大值为$d_n$,$d_i$s是间隔点。特征离散化可能存在信息丢失,但可以减少数据噪声,增加模型的稳定性和分类准确性[7]。

数据特征离散化后,通过特征选择的标准为分类器选择最优特征子集,并将该特征子集更新为下一次检测的候选特征。最后使用支持向量机(support vector machines, SVM)分类模型检测攻击,SVM有效减少了数据维度的影响,有较高的分类精度[8]。

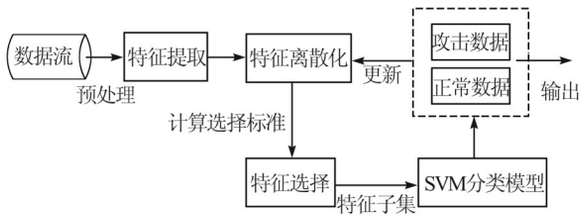

图1 入侵检测的特征选择模型

Fig. 1 Feature selection model for intrusion detection

为了独立于分类算法并降低计算成本,一般构建基于MI过滤的特征选择方法评估特征。MI方法能够测量随机变量之间的非线性依赖关系,评估复杂分类任务中数据特征的信息内容[9]。两个变量的MI通常用熵表示,熵为信息不确定性的度量,计算形式如下

$$I(X;Y)=H(X)+H(Y)-H(X,Y) \quad (1)$$

其中,$H(X)$和$H(Y)$为边际熵;$H(X,Y)$为联合熵,为变量X和Y平均所需要的信息量。若X是连续型随机变量,则$H(X)=-\int_{-\infty}^{+\infty}f(x)\log f(x)\mathrm{d}x$,$f(x)$表示$X$的概率分布函数。若用$\mathcal{N}$个等长$\Delta x$分隔$X$,则第$i$个间隔中观测样本的概率值$p_i=\int f(x)\mathrm{d}x\approx f(x_i)\Delta x$,假设$f(x)$在一个区间内近似恒定,熵的近似值如下

$$\hat{H}(X)=-\sum_i\left(\frac{n_i}{N}\log\frac{n_i}{N}\right)+\log(\Delta x)=$$
$$-\frac{1}{\ln 2}\sum_i\left(\frac{n_i}{N}\ln\frac{n_i}{N}\right)+\log(\Delta x) \quad (2)$$

其中,N表示总的样本数,n_i为第i个间隔的观测样本,$i\in(0,\mathcal{N})$且$\frac{n_i}{N}=p_i$。

1.2 特征选择标准

假设一组数据集中有K个特征数据和M个数据类别,用互信息$I(\boldsymbol{f};C)$表示特征数据中包含的关于目标类别C的信息量,从最大化下式

$$I(\boldsymbol{f};C)=$$
$$\sum_{f_1,\cdots,f_s}\sum_C P(f_1,\cdots,f_s;C)\log\frac{P(f_1,\cdots,f_s;C)}{P(f_1,\cdots,f_s)P(C)} \quad (3)$$

可以得到特征子集$\boldsymbol{f}$的最小集合。

为了减少$I(\boldsymbol{f};C)$的计算成本,通过联合互信息

(joint mutual information,JMI)方法[10]选择第i个特征f_i的模型如下式

$$J(f_i)=R+\frac{1}{|f|}\sum_{f_k\in f}(\mathrm{CI}-r) \tag{4}$$

其中,$|f|$为所选特征子集f的特征数,$R=I(f_i;C)$表示f_i和C之间的相关性;$r=I(f_i;f_k)$表示数据特征f_i和f_k之间的冗余;$\mathrm{CI}=I(f_i;f_k|C)$表示在给定的类别C下f_i和f_k之间的互补信息[11]。

通过最大化(4)式得到特征选择标准的最优值,即最大特征的相关性R、最小的特征冗余r和最大特征的互补信息CI,根据这三个标准可以选择当前最重要的特征。用潜在的概率分布估计特征熵时会引入偏差,假设样本n_i是多项式独立分布的,根据概率函数在$\frac{n_i}{N}$处的泰勒级数,样本期望为

$$E\{\hat{H}(X)\}=\frac{1}{\ln 2}\sum_i\left(-\frac{\bar{n}_i}{N}\ln\frac{\bar{n}_i}{N}-\left(\frac{1}{N}+\frac{1}{N}\ln\frac{\bar{n}_i}{N}\right)\cdot E\{(n_i-\bar{n}_i)\}-\frac{E\{(n_i-\bar{n}_i)^2\}}{2N\bar{n}_i}+E\{R_i^3(n_i)\}\right)+\log(\Delta x) \tag{5}$$

取$\frac{n_i}{N}=\frac{\bar{n}_i}{N}$,(5)式第二项为0。第三项简化为

$$\frac{1}{\ln 2}\sum_i\frac{E\{(n_i-\bar{n}_i)^2\}}{2N\bar{n}_i}=\frac{1}{2N\ln 2}(\mathcal{N}-1) \tag{6}$$

$\mathcal{N}$是变量X中的离散间隔数,观测样本的期望和方差如(7)式

$$E\{n_i\}=\bar{n}_i=Np_i;D\{n_i\}=Np_i(1-p_i) \tag{7}$$

(8)式表示忽略高阶项的$E\{\hat{H}(X)\}$。

$$E\{\hat{H}(X)\}\approx\sum_i\left(-\frac{\bar{n}_i}{N}\log\frac{\bar{n}_i}{N}\right)-\frac{(\mathcal{N}-1)}{2N\ln 2}+\log(\Delta x)\approx H(X)-\frac{(\mathcal{N}-1)}{2N\ln 2} \tag{8}$$

将变量Y分为$\mathcal{J}$个间隔,$E\{\hat{H}(Y)\}$可用下式表示

$$E\{\hat{H}(Y)\}\approx H(Y)-\frac{(\mathcal{J}-1)}{2N\ln 2} \tag{9}$$

若x-y平面被分为$\mathcal{N}\times\mathcal{J}$个大小相等($\Delta x\times\Delta y$)的单元,则变量$X$和$Y$联合熵的期望为

$$E\{\hat{H}(X,Y)\}=H(X,Y)-\frac{(\mathcal{N}\mathcal{J}-1)}{2N\ln 2} \tag{10}$$

结合式(1)可以得到

$$E\{\hat{I}(X;Y)\}=I(X;Y)-\frac{(\mathcal{N}-1)(\mathcal{J}-1)}{2N\ln 2} \tag{11}$$

根据两个连续随机变量熵的期望,可以看到互信息$I(X;Y)$的偏差是$\frac{(\mathcal{N}-1)(\mathcal{J}-1)}{2N\ln 2}$。

2　卡方校正算法(CSCA)

2.1　偏差校正

根据MI计算信息熵中存在的偏差会对特征选择产生影响,因此需要进行偏差校正。设存在有$\mathcal{N}$个间隔的连续随机变量X和离散的类变量C,$\mathcal{K}$表示类别C的个数,X和C的MI用信息熵表示为

$$I(X;C)=H(X)-H(X|C)=H(X)-\sum_{c=1}^{\mathcal{K}}p(c)H(X|C=c) \tag{12}$$

结合(8)式,假设给定X的信息熵是一个固定值,用$\hat{H}(X)$代替$E\{\hat{H}(X)\}$,用$\frac{n_c}{N}$估计$p(c)$,则(12)式可写为

$$I(X;C)=\hat{H}(X)+\frac{(\mathcal{N}-1)}{2N\ln 2}-\sum_{c=1}^{\mathcal{K}}p(c)\left(\hat{H}(X|C=c)+\frac{(\mathcal{N}-1)}{2n_c\ln 2}\right)=\hat{I}(X;C)+\frac{(\mathcal{N}-1)}{2N\ln 2}-\sum_{c=1}^{\mathcal{K}}\frac{n_c}{N}\frac{(\mathcal{N}-1)}{2n_c\ln 2}=\hat{I}(X;C)-\frac{(\mathcal{N}-1)(\mathcal{K}-1)}{2N\ln 2} \tag{13}$$

由此可以看到$I(X;C)$和$I(X;Y)$产生的偏差是相似的,同理可得$I(f_i;C)$,$I(f_i;f_j)$的偏差。因此,

$$I(X;Y|C)=H(X|C)+H(Y|C)-H(X,Y|C)=\hat{I}(X;Y|C)-\frac{(\mathcal{N}-1)(\mathcal{J}-1)\mathcal{K}}{2N\ln 2} \tag{14}$$

若$\mathcal{N}$和$\mathcal{J}$是特征f_i和f_j中的间隔数,可以得到$I(f_i;f_j|C)$的偏差是$\frac{(\mathcal{N}-1)(\mathcal{J}-1)\mathcal{K}}{2N\ln 2}$。此时,在(4)式的目标函数中添加偏差项为

$$J_{\mathrm{CSCA}}(f_i)=I(f_i;C)-\frac{(\mathcal{N}-1)(\mathcal{K}-1)}{2N\ln 2}+\frac{1}{|f|}\sum_{f_k\in f}\left(I(f_i;f_k|C)-\frac{(\mathcal{N}-1)(\mathcal{J}-1)\mathcal{K}}{2N\ln 2}-I(f_i;f_k)+\frac{(\mathcal{N}-1)(\mathcal{J}-1)}{2N\ln 2}\right) \tag{15}$$

添加偏差后的目标函数有助于准确计算JMI,分别对R,r和CI进行偏差校正,选出最优的特征子集。

由(15)式可以发现,当f_i和C统计独立时,相关

性$I(f_i;C)$服从自由度为$(\mathcal{N}-1)(\mathcal{K}-1)$的$\chi^2$分布。当$f_i$和$f_j$统计独立时,冗余$I(f_i;f_j)$服从自由度为$(\mathcal{N}-1)(\mathcal{J}-1)$的$\chi^2$分布。临界值$\chi^2_C(R)$和$\chi^2_C(r)$分别如下

$$\chi^2_C(R)=2N(\ln 2)\times I(f_i;C)$$
$$\chi^2_C(r)=2N(\ln 2)\times I(f_i;f_j) \tag{16}$$

同理若给定变量Z,两个独立随机变量X和Y的条件互信息为

$$I(X;Y|Z)=\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z}p(x,y,z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)}=\frac{1}{\ln 2}\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z}p(x,y,z)\ln\frac{p(x,y,z)}{p(x|z)p(y|z)P(z)} \tag{17}$$

其中,$p(x,y,z)$是联合概率,$p(x|z)$和$p(y|z)$是条件概率。假设$q(x,y,z)\equiv p(x|z)p(y|z)p(z)$,则有

$$I(X;Y|Z)=\frac{1}{\ln 2}\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z}p(x,y,z)\ln\frac{p(x,y,z)}{q(x,y,z)} \tag{18}$$

令$p_1=p(x,y,z)$, $q_1=q(x,y,z)$, 则$f(p_1)=p_1\ln\frac{p_1}{q_1}$,$f(p_1)$在$p_1=q_1$时泰勒展开,展开式第三项简化为$\frac{(p_1-q_1)^2}{2q_1}$并忽略高阶项,得到下式

$$I(X;Y|Z)\approx\frac{1}{\ln 2}\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z}\frac{(p(x,y,z)-q(x,y,z))^2}{2q(x,y,z)} \tag{19}$$

将(19)式化为标准表达式后,$I(X;Y|Z)$服从自由度为$(|X|-1)(|Y|-1)Z$的χ^2分布。因此,对于给定类别C且在f_i和C统计独立的情况下,$I(f_i;f_j|C)$服从自由度为$(\mathcal{N}-1)(\mathcal{J}-1)\mathcal{K}$的$\chi^2$分布。$\chi^2_C(\mathrm{CI})$的计算如下

$$\chi^2_C(\mathrm{CI})=2N\ln(2)\times I(f_i;f_j|C) \tag{20}$$

由于特征数据的R,r和CI都服从χ^2分布,将χ^2分布同时用于检验离散化和特征选择,可以减少离散化的信息丢失,同时优化偏差得到最优目标函数[12]。因此,通过比较(15)式、(16)式和(20)式的临界值,可以做出特征选择和离散化水平决策。

2.2 算法实现及评估指标

首先,CSCA算法使用类信息C计算其MI,根据其相关性分别计算每个特征的离散度

$$J_R(f_i)=I(f_i^{d_i};C) \tag{21}$$

其中,$f_i^{d_i}$表示特征f_i的离散化水平为d_i,算法用χ^2分布检验离散化可以选择最小的d_i,即$J_R(f_i)>\chi^2_C(R)$。根据特征对类别C的依赖,每个特征中不同值的总数可以确定最大离散间隔d。因此,不同特征值有不同的离散间隔。但是,为了在最大的d内实现特征数据最佳的离散化水平,对所有特征使用相同的间隔。离散化水平设置和特征选择的主要步骤如算法1。

算法1 卡方校正算法(CSCA)

输入: 特征数据集$F=\{f_1,f_2,\cdots,f_n\}$,类信息C,最大间隔d

输出: 特征子集$\boldsymbol{f}$,离散化水平$D=\{d_1,d_2,\cdots,d_k\}$

1 Begin
2 设定特征集$F_C\leftarrow\varnothing$
3 初始化$j=2$,用间隔j离散化f_i
4 如果$J_R(f_i)>\chi^2_C(R)$,则
$D_C\leftarrow D_C\cup j;J_C\leftarrow J_C\cup J_R(f_i);F_C\leftarrow F_C\cup f_i$
5 $j=j+1$
6 当j达到最大间隔d时,得到初始r个特征集F_C,离散化水平D_C和相关性J_C
7 F_C和D_C根据对应的J_C值降序排序,将最大J_C值对应的特征和间隔定义为f_1和d_1
$\boldsymbol{f}\leftarrow f_1;D\leftarrow d_1;F_C\leftarrow F_C\backslash f_1$
8 初始化$i=2$,T表示临界点,此时$j=d_i-\delta$
9 用间隔j离散化特征f_i,如果
$J_{\mathrm{CSCA}}(f_i)>\chi^2_C(R),\chi^2_C(r),\chi^2_C(\mathrm{CI})$
则
$d_i\leftarrow j;J_i\leftarrow J_{\mathrm{CSCA}}(f_i)$;
$T\leftarrow\chi^2_C(R),\chi^2_C(r),\chi^2_C(\mathrm{CI})$
10 迭代计算直到$j=d_i+\delta$,得到J_i,T和d_i
11 如果$J_{\mathrm{CSCA}}>T$,则
$\boldsymbol{f}\leftarrow\boldsymbol{f}\cup f_i;D\leftarrow D\cup d_i$
12 计算$F_C\leftarrow F_C\backslash f_i$
13 迭代训练$i=3,4,\cdots,r$,得到特征子集$\boldsymbol{f}$和相应的离散度D
14 End

在算法1中,训练得到的特征按照相关性降序排序,选出相关性高的特征向量。为了使选择的特征和类变量C的联合性更加相关,通过偏移少量的δ改变新特征的离散化水平,若偏移过大则使特征数据将变得不相关。算法在J_{CSCA}大于特征选择标准的临界值时,数据特征被选择。

$$J_{\mathrm{CSCA}}(f_i)=I(f_i^{d_i};C)-\frac{(\mathcal{N}-1)(\mathcal{K}-1)}{2N\ln 2}+$$

$$\frac{1}{|f|}\sum_{f_k\in f}\left(I(f_i^{d_j};f_k|C)-\frac{(\mathcal{N}-1)(\mathcal{J}-1)\mathcal{K}}{2N\ln 2}-I(f_i^{d_j};f_k)+\frac{(\mathcal{N}-1)(\mathcal{J}-1)}{2N\ln 2}\right) \tag{22}$$

通过算法1检验所有特征的R,r和CI来评估特征的重要程度,若某个特征的r高,则丢弃,若R和CI有意义,则选择该特征,最终可得到适当离散程度的最优特征子集。

CSCA在应对高维数据特征时,可以筛选出重要相关的约简特征,提高IDS的检测准确性。为了评估算法的性能,使用准确率(A)、召回率(R)、查准率(P)和误报率来衡量IDS的优劣。

准确率为正确分类的样本数占总样本的比例,计算如下

$$A=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{23}$$

召回率表示被正确分类的样本中正常数据所占的比例,计算如下

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{24}$$

查准率表示预测为正确的样本数中,正确分类数据的占比,计算如下

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{25}$$

以上式子中,TP表示准确分类正常数据的样本数,FP表示正常的数据中存在误报的样本数,TN表示准确分类攻击数据的样本数,FN表示攻击数据中被漏报的样本数[5,13]。

实际数据中R和P指标不会都达到理想值,需要加权调和平均$F1$值进行综合考虑,计算如下

$$F1=\frac{2\cdot R\cdot P}{R+P} \tag{26}$$

误报率是将正常的样本数据预测为攻击类别所占的比例,计算公式如下

$$\mathrm{FAR}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \tag{27}$$

高性能的IDS一定是低误报率。

3　仿真分析

3.1　数据集及仿真环境

为了研究数据特征选择的影响,选用两个常用的基准数据库NSL-KDD数据集[14]和UNSW-NB15数据集[15]进行实验仿真。NSL-KDD数据集有含125 973个训练样本和22 543个测试,有4个攻击类型,提取了41个分类特征,分为符号特征,0-1类型特征和百分比类型特征。UNSW-NB15数据集有175 341个训练数据和82 332个测试数据,包含9个攻击类,提取了44个数据特征,主要分为:时间特征,内容特征,流特征,基本特征,标记特征和其他原始特征。具体的数据特征分布如表1。

为了提高模型的训练效率,在预处理中对特征数据进行归一化处理,将非数值特征数据转换为数值,使所有的特征值处于相同的数量级[16],并采用10折交叉验证进行数据仿真,取10次交叉仿真结果的均值作为对算法性能的估计。

3.2　检测精度分析

IDS检测数据的准确率是分析算法精度的重要

表1　NSL-KDD和UNSW-NB15数据集的特征分布

Table 1　Feature distribution of NSL-KDD and UNSW-NB15 datasets

NSL-KDD数据集的特征	UNSW-NB15数据集的特征
protocol_type, service, flag, land, logged_in, root_shell, duration, su_attempted, is_host_login, diff_srv_rate, is_guest_login,src_bytes, dst_bytes, wrong_fragment, urgent, hot, num_root,num_failed_logins, num_compromised, num_file_creations, num_shells, dst_host_count, num_access_files, num_outbound_cmds, count, serror_rate, srv_count, srv_serror_rate, rerror_rate, srv_rerror_rate, same_srv_rate, srv_diff_host_rate, 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	dur, proto, service, state, spkts, dpkts, sbytes,dbytes, rate, sttl, dtll, sload, dload, sloss,dloss,sinpkt, dinpkt, sjit, djit, swin, stcpb,dtcpb, dwin,tcprtt, synack, ackdat, smean,dmean, trans_depth, response_body_len,ct_srv_src, ct_state_ttl, ct_dst_ltm,ct_src_dport_ltm, ct_dst_sport_ltm, ct_dst_src_ltm, is_ftp_login, ct_ftp_cmd, ct_flw_http_mthd,ct_src_ltm, ct_srv_dst, is_sm_ips_ports, attack_cat, label
合计　41个	44个

指标,可以评估不同特征选择方法的有效性。在两个不同的数据集中分别选择几组特征子集进行训练,可以提高算法的可靠性和稳定性。MIGM,M-DFIFS和CSCA在SVM分类器中的检测准确率结果如图2所示。

根据图2所示,算法的检测准确率随着特征子集的增加而增加,达到稳定值后变化不明显。CSCA检测的准确率明显更高,平均准确率比另外两个算法提高了约5%,达到95.2%。MIGM算法受冗余特征的影响较大,在准确率达到峰值后,检测准确率随着特征的增加有明显下降。但CSCA达到收敛时选择的特征数更少,说明该算法降低了候选特征的冗余性。

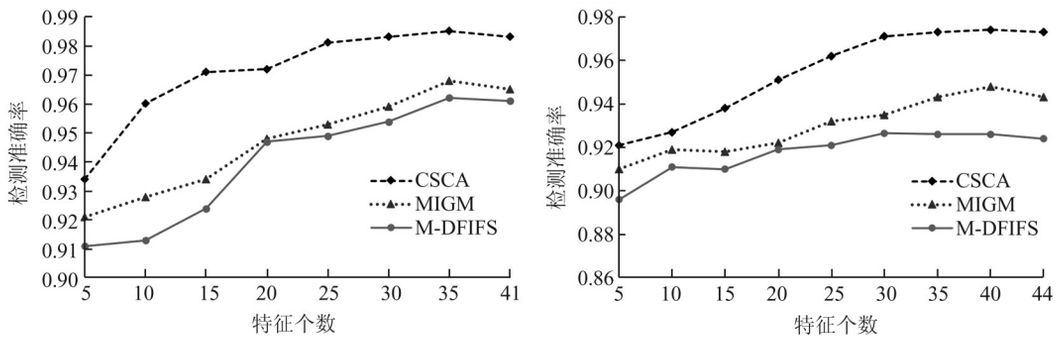

图2 不同数据集中算法的检测准确率分析

Fig. 2 Analysis of detection accuracy of algorithms in different data sets

算法收敛时,IDS的召回率和查准率也能直接反映特征数据选择算法的检测精度。表2和表3记录了不同数据集中算法仿真的平均召回率、查准率和$F1$值。

在算法的仿真结果中,通过观察不同特征数在多次10折交叉验证中的综合指标的均值,可以对比不同算法检测正常数据的精度。从表2和表3中可以看到CSCA在两个数据集中的检测结果都比较

表2 NSL-KDD数据集中算法的平均综合指标分析

Table 2 Analysis of the average comprehensive index of the algorithm in the NSL-KDD dataset

特征个数	R			P			$F1$值		
	CSCA	MIGM	M-DFIFS	CSCA	MIGM	M-DFIFS	CSCA	MIGM	M-DFIFS
10	0.953	0.930	0.919	0.949	0.926	0.911	0.951	0.928	0.915
20	0.960	0.942	0.934	0.963	0.944	0.932	0.961	0.943	0.933
30	0.984	0.958	0.957	0.980	0.957	0.951	0.982	0.956	0.954
40	0.879	0.753	0.679	0.977	0.945	0.947	0.925	0.838	0.971
均值	0.944	0.896	0.872	0.967	0.943	0.935	0.955	0.917	0.898

表3 UNSW-NB15数据集中算法的平均综合指标分析

Table 3 Analysis of the average comprehensive index of the algorithm in the UNSW-NB15 dataset

特征个数	R			P			$F1$值		
	CSCA	MIGM	M-DFIFS	CSCA	MIGM	M-DFIFS	CSCA	MIGM	M-DFIFS
10	0.931	0.917	0.912	0.933	0.916	0.911	0.932	0.917	0.912
20	0.954	0.921	0.919	0.958	0.924	0.920	0.956	0.922	0.920
30	0.978	0.963	0.946	0.976	0.949	0.945	0.977	0.956	0.946
40	0.905	0.872	0.783	0.967	0.926	0.920	0.935	0.898	0.846
均值	0.942	0.918	0.890	0.959	0.929	0.924	0.950	0.923	0.906

稳定,相比于另外两个算法,该方法R和P的结果表现更好,且综合调整的$F1$值分别达到0.955和0.950,明显提高了模型的检测性能。

3.3　误报率分析

误报率是衡量IDS系统性能的关键指标,误报过高会导致系统检测性能下降。随着特征选择个数的变化,NSL-KDD数据集和UNSW-NB15数据集的误报率结果如图3所示。

从图3可以看出,选择的特征过少时,特征子集不足以代表所有的类别信息,算法检测数据的误报率较高。随着特征个数的增多,算法的误报率整体呈下降趋势,但在某些区域会出现增长的情况,这主要是受到冗余特征的影响。CSCA通过偏差校正后减少了冗余特征,模型的平均误报率均比另外两个算法低3%左右,并且收敛性更快,训练选择最优的特征子集后的误报率都接近0.01,可以在数据动态变化时减少分类错误。

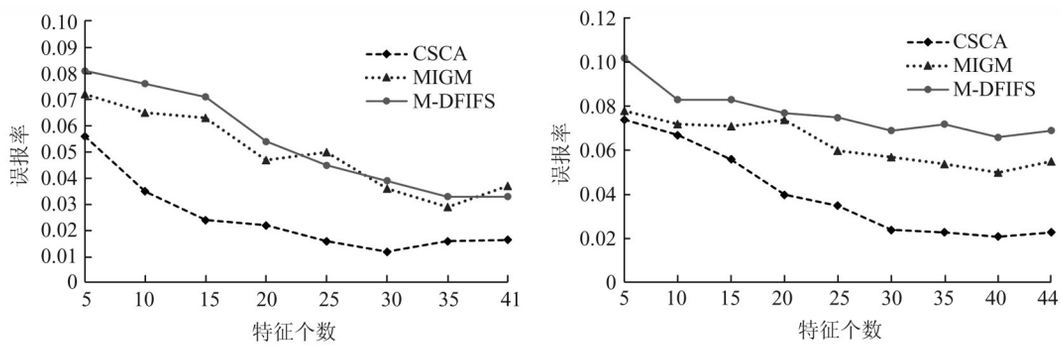

图3　不同数据集中算法的误报率

Fig. 3　False alarm rate of algorithms in different data sets

4　结　语

在面临高维复杂的网络入侵数据时,传统MI特征选择算法的性能达不到理想的效果。基于动态离散化的CSCA能更好地减少特征冗余,选择最优的特征子集。利用离散化方法对所有候选特征预处理,再根据MI的特征选择标准计算相关偏差;然后通过卡方检验校正偏差目标函数,最后选出当前最重要的特征子集。仿真结果表明,CSCA精确提取重要的数据特征,使IDS具有更高的检测精度和更低的误报率。为了提高IDS对未知攻击的检测度,下一步将深入研究如何应对数据流中的漂移变化。

参考文献:

[1] NANCY P, MUTHURAJKUMAR S, GANAPATHY S, *et al*. Intrusion detection using dynamic feature selection and fuzzy temporal decision tree classification for wireless sensor networks[J]. *IET Communications*, 2020, **14**(5): 888-895. DOI:10.1049/iet-com.2019.0172.

[2] 刘云, 肖雪. 基于邻域维护准则的特征选择算法优化研究[J]. 重庆大学学报, 2019, **42**(3): 58-64. DOI: 10.11835/j.issn.1000-582X.2019.03.006.
LIU Y, XIAO X. Research on optimization of feature selection algorithm based on neighborhood preservation criterion [J]. *Journal of Chongqing University* (*Natural Science Edition*), 2019, **42**(3): 58-64. DOI:10.11835/j.issn.1000-582X.2019.03.006 (Ch).

[3] ZHANG X, MEI C L, CHEN D G, *et al*. Active incremental feature selection using a fuzzy-rough-set-based information entropy[J]. *IEEE Transactions on Fuzzy Systems*, 2020, **28**(5): 901-915. DOI: 10.1109/TFUZZ.2019.2959995.

[4] WANG X Z, GUO B, SHEN Y, *et al*. Input feature selection method based on feature set equivalence and mutual information gain maximization [J]. *IEEE Access*, 2019, **7**: 151525-151538. DOI:10.1109/ACCESS.2019.2948095.

[5] WEI G F, ZHAO J, FENG Y L, *et al*. A novel hybrid feature selection method based on dynamic feature importance[J]. *Applied Soft Computing*, 2020, **93**: 106337. DOI:10.1016/j.asoc.2020.106337.

[6] FAHY C, YANG S X. Dynamic feature selection for clustering high dimensional data streams[J]. *IEEE Access*, 2019, **7**: 127128-127140. DOI:10.1109/ACCESS.2019.2932308.

[7] TRAN B, XUE B, ZHANG M J. A new representation in PSO for discretization-based feature selection [J]. *IEEE Transactions on Cybernetics*, 2018, **48**(6): 1733-1746. DOI:10.1109/TCYB.2017.2714145.

[8] ATIENCIA R S, WEBER R. Dynamic rough-fuzzy sup-

port vector domain description for outlier detection [C]// 2018 *IEEE International Conference on Fuzzy Systems* (*FUZZ-IEEE*). New York: IEEE Press, 2018: 1-8. DOI:10.1109/FUZZ-IEEE.2018.8491618.

[9] 李郅琴, 杜建强, 聂斌, 等. 特征选择方法综述[J]. 计算机工程与应用, 2019, **55**(24): 10-19. DOI:10.3778/j.issn.1002-8331.1909-0066.
LI Z Q, DU J Q, NIE B, *et al*. Summary of feature selection methods [J]. *Computer Engineering and Applications*, 2019, **55**(24): 10-19. DOI:10.3778/j.issn.1002-8331.1909-0066(Ch).

[10] YAN B H, HAN G D. Effective feature extraction via stacked sparse autoencoder to improve intrusion detection system [J]. *IEEE Access*, 2018, **6**: 41238-41248. DOI: 10.1109/ACCESS.2018.2858277.

[11] LUO C, LI T R, YI Z. An incremental feature selection approach based on information entropy for incomplete data [C]//2019 *IEEE International Conference on Dependable, Autonomic and Secure Computing*. New York: IEEE Press, 2019: 483-488. DOI:10.1109/dasc/picom/cbdcom/cyberscitech.2019.00097.

[12] 陈谌, 梁雪春. 基于基尼指标和卡方检验的特征选择方法[J]. 计算机工程与设计, 2019, **40**(8): 2342-2345. DOI:10.16208/j.issn1000-7024.2019.08.040.
CHEN C, LIANG X C. Feature selection method based on Gini index and chi-square test [J]. *Computer Engineering and Design*, 2019, **40**(8): 2342-2345. DOI:10.16208/j.issn1000-7024.2019.08.040(Ch).

[13] GAO X W, SHAN C, HU C Z, *et al*. An adaptive ensemble machine learning model for intrusion detection [J]. *IEEE Access*, 2019, **7**: 82512-82521. DOI:10.1109/ACCESS.2019.2923640.

[14] NGUYEN M T, KIM K. Genetic convolutional neural network for intrusion detection systems [J]. *Future Generation Computer Systems*, 2020, **113**: 418-427. DOI:10.1016/j.future.2020.07.042.

[15] MOUSTAFA N, SLAY J. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set) [C]//2015 *Military Communications and Information Systems Conference* (*MilCIS*). New York: IEEE Press, 2015: 1-6. DOI:10.1109/MilCIS.2015.7348942.

[16] 宋勇, 蔡志平. 大数据环境下基于信息论的入侵检测数据归一化方法[J]. 武汉大学学报(理学版), 2018, **64**(2): 121-126. DOI:10.14188/j.1671-8836.2018.02.004.
SONG Y, CAI Z P. Normalized method of intrusion detection data based on information theory in big data environment [J]. *Journal of Wuhan University* (*Natural Science Edition*), 2018, **64**(2): 121-126. DOI:10.14188/j.1671-8836.2018.02.004(Ch).

□