基于模糊神经网络的异常网络数据挖掘算法

许 磊 王建新
(北京林业大学信息学院 北京 100083)

摘 要 异常网络数据受到聚类中心的模糊加权扰动的影响,导致数据挖掘的聚类性不好。文中提出一种基于模糊神经网络的异常网络数据挖掘算法,该算法根据异常网络数据的混合分类属性进行相似度分析,提取异常网络数据的数值属性特征和分类属性特征,采用联合关联规则分析方法进行异常网络数据的模糊融合处理,采用基于模糊质心相异性的度量方法构建异常网络数据的分类模糊集,并在模糊数据集中进行异常网络数据混合加权和自适应分块匹配,进而提取异常网络数据的弱关联化特征量,最后将提取的特征量输入到模糊神经网络分类器中进行数据分类识别,完成异常网络数据的优化挖掘。仿真结果表明,采用所提方法进行异常网络数据挖掘的数据聚类性较好,挖掘过程的收敛性和抗干扰性较强。

关键词 模糊神经网络,异常网络数据,挖掘,特征提取

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.04.011

Data Mining Algorithm of Abnormal Network Based on Fuzzy Neural Network

XU Lei WANG Jian-xin
(School of Information Science & Technology, Beijing Forestry University, Beijing 100083, China)

Abstract Abnormal network data are affected by the fuzzy weighted disturbance of the clustering center, which leads to poor clustering of data mining. This paper proposed a data mining algorithm of anomaly network based on fuzzy neural network. Similarity analysis is performed according to the mixed classification attributes of abnormal network data, the numerical and classification attributes of abnormal network data are extracted, and the fuzzy fusion processing of the abnormal network data is carried out by using the joint association rule analysis method. The classification fuzzy set of abnormal network data is constructed based on fuzzy centroid heterogeneity measurement method. In the fuzzy data set, abnormal network data are weighted by mixing and adaptive block matching, and the weak correlation characteristic quantity of abnormal network data is extracted. The extracted features are input into the fuzzy neural network classifier for data classification and recognition, and the optimized mining of abnormal network data is completed. The simulation results show that the proposed method has good data clustering ability for anomaly network data mining. The mining process has strong convergence and anti-interference.

Keywords Fuzzy neural network, Abnormal network data, Mining, Feature extraction

随着大数据信息处理技术的发展,各种应用型大数据分布在网络中,并以TB级别的日志数据猛增,数据集的规模、容量以及增长速度大幅攀升。大数据信息处理在金融业、通信、网络、信息管理、智慧城市、人工智能等各个行业中表现出巨大的优势[1]。在异构网络中,大数据通过无线网络实现远程传输和控制,网络中数据的安全性受到人们的极大关注。本文通过对异常网络数据的准确挖掘,提取异常网络数据的关联规则特征量,采用信息融合和数据聚类方法实现异常网络数据的挖掘,提高对网络异常的诊断能力和抗攻击能力。研究异常网络数据挖掘方法在实现网络故障分析和网络安全检测方面具有重要意义,数据挖掘算法的相关研究受到人们的极大关注[2]。

挖掘异常网络数据的本质是进行异常网络数据的时间序列分析,结合异常网络入侵检测和网络拥塞控制方法进行异常网络流量分析,根据网络传输的流量特征进行异常数据挖掘,采用相关的信息处理和数据检测方法来提高异常数据挖掘的准确性和抗干扰能力[3]。对异常网络数据进行挖掘的传统算法主要采用分集检测和谱分析方法,利用自相关特征谱分解方法对异常网络数据的输出流量进行自适应学习[4],建立异常网络数据输出预测模型,采用功率谱分析和分类器进行异常网络数据的特征提取和数据挖掘,结合模糊数值分析和簇聚类方法实现异常数据的挖掘。根据上述原理,相关学

到稿日期:2018-03-28 返修日期:2018-06-09 本文受国家自然科学基金(61170268)资助。

许 磊(1992-),男,硕士生,主要研究方向为数据挖掘、大数据;王建新(1972-),男,博士,教授,博士生导师,主要研究方向为应用数学、软件工程、数据挖掘,E-mail:wangjx@bjfu.edu.cn(通信作者)。

者对数据挖掘算法进行了研究。文献[5]提出一种基于简化梯度算法的网络通信中异常数据的挖掘模型,其采用匹配滤波器进行网络通信数据的干扰滤波,结合自适应的信道均衡控制方法进行网络输出信道的均衡设计,提高了数据挖掘的抗干扰能力;但该方法进行异常数据挖掘时存在带宽受限和维数较大等问题。文献[6]提出一种基于模糊指向性聚类的异常网络数据挖掘方法,其采用模糊 K 质心方法进行异常网络数据的模糊加权,在保留异常网络数据集内在的不确定性的条件下实现数据的优化聚类,提高了数据挖掘时模糊决策性;但该方法在进行大规模的异常网络数据挖掘的存在计算开销较大和复杂度较高的问题。文献[7]提出基于自适应分数间隔采样异常网络数据的挖掘方法,其采用分数间隔进行数据替代处理,结合自适应级联匹配检测方法进行异常数据的特征检测和滤波分析,提取高阶谱特征量,并在 BP 分类器中进行数据的分类识别,提高了数据挖掘的精度;但该方法在对随机性较强的异常数据进行挖掘时准确性不高。

针对上述问题,本文提出一种基于模糊神经网络的异常网络数据挖掘算法。首先构建异常数据挖掘的总体模型,并进行数据特征分析;然后提取异常网络数据的特征量,采用模糊神经网络进行异常网络数据的挖掘和分类识别;最后通过实验测试展示了本文方法在提高异常数据挖掘准确性方面的优越性能。

1 异常网络数据的特征分析与提取

1.1 数据相似度分析

为了实现对异常网络数据的优化挖掘,首先分析异常网络数据的相似度信息,采用 C 分类的决策树模型进行异常网络数据的相似度分解[7]。

对异常网络数据进行混合特征识别和数据分类,构造异常网络数据的混合属性模糊分类模型,并根据数据的混合分类属性进行相似度分析,对模糊信息的分段属性集 X 进行奇异值(SVD)分解:

$$X=\boldsymbol{UDV}^{\mathrm{T}} \tag{1}$$

其中,$\boldsymbol{U}\in R^{m\times m}$ 为本体映射的特征矩阵,$\boldsymbol{V}\in R^{M\times M}$ 表示异常网络数据的类间闭频繁项矩阵,且 $\boldsymbol{U}^{\mathrm{T}}=\boldsymbol{U}^{-1}$,$\boldsymbol{V}^{\mathrm{T}}=\boldsymbol{V}^{-1}$;$\boldsymbol{D}\in R^{m\times M}$,且满足 $\boldsymbol{D}=[\Sigma\ \ 0]$,在本体映射下,异常网络数据分布式特征量的加权值为 $\Sigma=\mathrm{diag}(\sqrt{\lambda_1},\sqrt{\lambda_2},\cdots,\sqrt{\lambda_m})$,$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m$,求得异常网络数据的语义概念集的分布矩阵为 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$。取非零特征值作为训练子集,进行数据信息流模型的重构,采用混合相似度特征分析方法对异常网络数据进行相空间重构[8],相空间重构后输出的平均互信息特征表达式为:

$$I(\boldsymbol{Q},S)=\sum_i\sum_j p_{sq}(s_i,q_j)\log_2[p_{sq}(s_i,q_j)/p_s(s_i)] \tag{2}$$

其中,$p_{sq}(s_i,q_j)$ 表示异常网络数据的语义本体概念集 s_i 和冗余数据概念集 q_j 的联合分布概率。定义异常网络数据的簇中的信息分布模型为$[s,q]=[x(t),x(t+\tau)]$,得到模糊信息的闭频繁项。在第 l 个模糊质心,用 s 代表采样的异常网络数据信息流 $x(t)$ 的随机序列样本,用 q 表示延迟时间序列样本,输出的延迟序列为 $x(t+\tau)$,得到的 $I(\boldsymbol{Q},S)$ 为以 τ 为自变量的模糊决策函数。

假设每个分类属性值的初始码元为 $C_0=C_{N/2}=0$,$C_{N-n}=C_n^*$,$n=0,1,2,\cdots,N/2-1$,异常网络数据的挖掘对象和簇中心分布的关系模型为:

$$P_r=\frac{P_t}{(4\pi)^2(\frac{d}{\lambda})^{\gamma}}[1+\alpha^2+2\alpha\cos(\frac{4\pi h^2}{d\lambda})] \tag{3}$$

根据数据的不同属性在聚类中的差异性,对异常数据进行特征识别,得到准确的概率密度函数为:

$$P_S=p_{2D}^k(1-p_{2D})^{N-1-k}\sum_{i=1}^{\infty}\lambda_S^i=\frac{\lambda_S}{1-\lambda_S} \tag{4}$$

其中,λ_S 为在采样时刻进行数据采集的相似度系数,p_{2D} 为簇中的信息分布概率密度。异常网络数据簇中心之间的相异度计算公式为:

$$DisSim(A,B)=1-\left|\frac{SameDis(A)-SameDis(B)}{Dis(A)+Dis(B)}\right| \tag{5}$$

其中,$Dis(A)$ 表示聚类过程中的扩展损失,$Dis(B)$ 表示属性数据集。

1.2 异常网络数据的属性特征提取

在对异常网络数据进行相似度分析的基础上,提取异常网络数据的数值属性特征和分类属性特征。假设 X 为具有 m 个属性的分类属性的异常网络数据集,第 i 个属性值的异常数据 $y(k)$ 和分类训练数据集 $\varphi(k)$ 可以表示为:

$$y(k)=s_1(k)+n_1(k),\varphi(k)=s_2(k)+n_2(k) \tag{6}$$

$$s_1(k)=AA_H\mathrm{e}^{\mathrm{j}(\Omega k+\theta_H)},s_2(k)=AA_{H_B}\mathrm{e}^{\mathrm{j}(\Omega k+\theta_{H_B})} \tag{7}$$

其中,A_H,A_{H_B} 和 θ_H,θ_{H_B} 分别是前 p 个元素的数值属性值、系统函数 $H(z)$ 和 $H_B(z)$ 的离散化数值属性特征量和响应幅值。结合最小化目标成本方法进行自适应寻优[9],得到数值属性特征和分类属性特征为:

$$R_\beta X=U\{E\in U/R\,|\,c(E,X)\leqslant\beta\} \tag{8}$$

$$R_\beta X=U\{E\in U/R\,|\,c(E,X)\leqslant 1-\beta\} \tag{9}$$

对于第 i 个分类属性的两个数据块 m_i 和 m_j,对数据对象 $m_{i,j}(1\leqslant i\leqslant n,1\leqslant j\leqslant k)$ 进行混合特征分解,异常数据的聚类特征系数为$\{\lambda_i:1\leqslant i\leqslant S\}$,判别准则为$\{R_j:1\leqslant j\leqslant L\}$。根据异常网络数据分类属性的属性值之间的差异性,得到训练函数 f 和基 d_{γ_0} 之间的模糊概念集为:

$$\lambda^n(d_{\gamma_0})=\int_{-\infty}^{+\infty}f(t)d_{\gamma_0}^*(t)\mathrm{d}t \tag{10}$$

采用联合关联规则分析方法进行异常网络数据的模糊融合处理[10],求得异常网络数据在共同出现的属性值下的自相关特征分块函数:

$$S_b=\sum_{i=1}^{e}p(\omega_i)(u_i-u)(u_i-u)^{\mathrm{T}} \tag{11}$$

$$S_\omega=\sum_{i=1}^{e}p(\omega_i)E\left[\frac{(u_i-u)(u_i-u)^{\mathrm{T}}}{\omega_i}\right] \tag{12}$$

$$S_i=S_b+S_\omega \tag{13}$$

其中,$p(\omega_i)$ 为指定离散区间内的规则向量集,$\mu=E(x)$ 为离散区间数。对异常网络数据的联合关联规则模型 $X(t)$ 进行归一化处理,得到新的聚类模态函数满足:

$$X'(t)=\frac{X(t)}{\|X(t)\|} \tag{14}$$

其中,$\|X(t)\|$ 表示对 $X(t)$ 取模。通过对异常网络数据的模

糊融合处理，可以提高数据挖掘的分类识别能力。

2　数据挖掘算法的优化

2.1　异常网络数据的混合加权分块匹配

在根据异常网络数据的混合分类属性进行相似度分析，提取异常网络数据的数值属性特征和分类属性特征以及数据融合处理的基础上，对异常网络数据优化挖掘算法进行改进设计。本文提出一种基于模糊神经网络的异常网络数据挖掘算法，在混合属性条件下，异常网络数据挖掘的加权滤波函数式为：

$$X(u)=\sqrt{\frac{1-\mathrm{j}\cot\alpha}{2\pi}}\int_{-\infty}^{+\infty}x(t)\exp\left[\mathrm{j}\frac{t^2+u^2}{2}\cot\alpha-\mathrm{j}\csc\alpha\right]\mathrm{d}t \tag{15}$$

其中，$x(t)$表示网络异常数据元素是分类属性值，假设 α 为时间窗函数。结合描述性统计分析，得到数据的模糊划分的块匹配函数为：

$$\rho_{SRm}=\frac{\lambda_{SRm}}{\mu_{SRm}}=\sum_{i=1}^{M}\frac{\lambda_i p_{im}}{\mu_{im}} \tag{16}$$

当数据满足宽平稳条件时，得到数据分类的模糊质心系数 $u\in L_{t,x}^{d+1}(K\times IR^d)$。采用相关变量分析方法，得到数据的相异性度量$(u,u_t)\in C_t(K,\dot{H}_x^{s_c}\times\dot{H}_x^{s_c-1})$。结合模糊加权方法，得到簇中心的数值属性加权变量满足：

$$\begin{aligned}0\leqslant&[y^{\mathrm{T}}(t)\Sigma^{\mathrm{T}}T\Sigma y(t)-f^{\mathrm{T}}(y(t))Tf(y(t))]+[-y^{\mathrm{T}}(t)\\&U\Sigma_1 y(t)+2y^{\mathrm{T}}(t)U\Sigma_2 f(y(t))-f^{\mathrm{T}}(y(t))Uf(y(t))]+\\&[-y^{\mathrm{T}}(t-\sigma)V\Sigma_1 y(t-\sigma)+2y^{\mathrm{T}}(t-\sigma t)V\Sigma_2 f(y(t-\sigma))-\\&f^{\mathrm{T}}(y(t-\sigma))Vf(y(t-\sigma))]\end{aligned} \tag{17}$$

采用联合关联规则分析方法进行异常网络数据的模糊融合处理，利用基于模糊质心相异性度量方法构建异常网络数据的分类模糊集，为：

$$\alpha_{\mathrm{desira}}^{i}=\alpha_1-\frac{Density_i}{\sum_i Density_i}+\alpha_2\frac{AP_i}{AP_{\mathrm{init}}} \tag{18}$$

异常网络数据的模糊融合加权系数满足：

$$\begin{cases}\alpha_1+\alpha_2=1,\alpha_1,\alpha_2\in[0,1]\\ \alpha_2=\dfrac{\max\limits_i(AP_i)-\min\limits_i(AP_i)}{AP_{\mathrm{init}}}\end{cases} \tag{19}$$

在相邻的目标数据挖掘节点 $node_i$ 中引入一个调制参量，设数据挖掘的分配规则 $A\subset V$，$B\subset V$ 且 $A\cap B=\varnothing$，在有限数据集下，得到异常网络数据混合加权分块匹配集为：

$$\begin{aligned}x_{id}(t+1)=&wx_{id}(t)+c_1r_1[r_3^{t_0>T_0}p_{id}-x_{id}(t)]+\\&c_2r_2[r_4^{t_g>T_g}p_{gd}-x_{id}(t)]\end{aligned} \tag{20}$$

其中，t_0 和 t_g 分别表示目标数据挖掘的成本函数和统计特征，T_0 和 T_g 分别表示数值属性和分类属性。由此实现对异常网络数据的混合加权分块匹配。

2.2　数据挖掘的模糊神经网络分类

在模糊数据集中进行异常网络数据混合加权和自适应分块匹配，在此基础上提取异常网络数据的弱关联化特征量，并将提取的特征量输入到模糊神经网络分类器中进行数据分类识别[11-16]，从而得到一维数据矢量 X_n。提取的特征量为：

$$x_k=\sum_{n=0}^{N-1}C_n-\mathrm{e}^{\mathrm{j}2\pi kn/N},k=0,1,\cdots,N-1 \tag{21}$$

将上述特征量输入到模糊神经网络分类器中。

假设输入层有 $2n$ 个相同的神经元，对于 n 组异常网数据的弱关联规则特征量$(x_{i1},x_{i2},\cdots,x_{i,m-1},y_i)$，$i=1,2,\cdots,n$，采用模糊神经网络分类进行自适应训练，训练式为：

$$\begin{cases}net_{s1}(k)=r_s(k)\\ net_{s2}(k)=y_s(k)\end{cases} \tag{22}$$

在模糊神经网络的输入层，根据数值属性得到神经元状态为：

$$u_{si}(k)=net_{si}(k) \tag{23}$$

在神经网络的中间层，神经元的输出为：

$$x_{si}(k)=\begin{cases}1, & u_{si}(k)>1\\ u_{si}(k), & -1\leqslant u_{si}(k)\leqslant 1\\ 1, & u_{si}(k)<-1\end{cases} \tag{24}$$

其中，$u_{si}(k)$表示比例神经元和积分神经元的自适应加权边向量。根据神经元的输出测量误差进行自适应修正，得到异常网络数据挖掘的输出误差为：

$$\begin{cases}e_1=\varphi_a-\varphi_{ad}\\ e_2=\dot{\varphi}_a-\dot{\varphi}_{ad}\end{cases} \tag{25}$$

最后得到数据挖掘输出的误差收敛于：

$$\lambda_i=\lim_{N\to\infty}\frac{1}{N}\sum_{j=0}^{N-1}\log(\boldsymbol{R}_j)_{ii}=0 \tag{26}$$

可见，采用本文方法进行异常网络数据的挖掘时，挖掘过程是稳态收敛的，误差趋于零。

3　仿真实验与性能分析

本节通过仿真实验来测试本文方法在实现异常网络数据优化挖掘时的性能，实验采用 Matlab 进行设计。选择 Zoo 数据集作为测试数据集，数据集的初始规模大小为 2000 个对象数据包，训练数据集的规模为 100 个数据包，测试数据集的分类属性为 24。数据的分组属性维数为 3，由 16 个分类属性和 1 个数值属性描述。模糊神经网络的输入层神经元个数为 12，模型参数为 0.24。关联特征的主旁瓣高度比为 30 dB，异常网络数据挖掘的干扰强度为 −10 dB，对异常网络数据的初始采样间隔为 0.01。根据上述仿真参量的设定进行异常网络数据挖掘仿真实验。首先进行网络传输数据的原始采样，得到网络传输数据的采样时域波形，如图 1 所示。

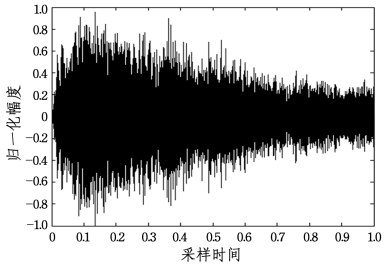
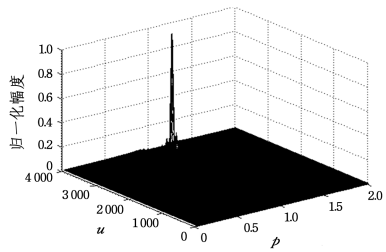


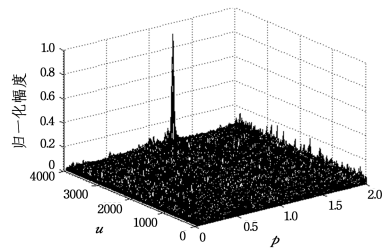
图 1　网络传输数据的原始采样时域波形

Fig. 1　Original sampling time domain waveform of network transmission data

以图 1 的数据为研究对象进行异常网络数据挖掘，提取异常网络数据的数值属性特征和分类属性特征，并根据信息融合结果提取异常网络数据的弱关联化特征量，从而得到特征提取结果，如图 2 所示。图 2 同时给出了采用传统的谱分析方法进行异常网络数据特征提取的输出结果。



(a)本文方法



(b) 谱分析方法

图 2 异常网络数据的特征提取结果

Fig. 2 Feature extraction results of abnormal network data

分析图 2 得知,本文方法在提取异常网络数据特征时具有很好的抗旁瓣干扰能力,输出异常网络数据的弱关联规则性特征强度较大。对具有 10 个分类属性的异常网络数据包进行挖掘,并测试挖掘准确率,得到的结果如表 1 所列。分析表 1 得知,本文方法对异常网络数据进行挖掘的准确率较高,比传统的谱分析方法和 FCM 方法分别高出 19.6% 和 21.8%。

表 1 挖掘准确率测试结果

Table 1 Mining accuracy test

SNR/dB	准确率/%		
	本文方法	FCM	谱分析
−10	93.2	67.2	71.2
−5	98.8	78.3	82.3
0	99.7	81.9	87.5
5	99.9	89.3	92.3
10	100.0	92.1	95.4

结束语 为了对异常网络数据进行准确挖掘,提取异常网络数据的关联规则特征量,采用信息融合和数据聚类方法实现异常网络数据的挖掘,提高网络异常的诊断能力和抗攻击能力,本文提出一种基于模糊神经网络的异常网络数据挖掘算法。该算法采用基于模糊质心相异性度量方法构建异常网络数据的分类模糊集,进行异常网络数据混合加权和自适应分块匹配,提取异常网络数据的弱关联化特征量,采用模糊神经网络进行数据分类挖掘。研究表明,本文方法进行异常网络数据挖掘的准确率较高,抗干扰能力较强。

参 考 文 献

[1] HUANG C. Cloud computing environment of huge amounts of optical fiber communication fault data mining algorithms[J]. Laser Journal, 2017, 38(1): 96-100.

[2] ZHANG S, ZHANG L. W-PAM restricted clustering algorithm in data mining[J]. Computer Science, 2016, 43(S2): 447-450. (in Chinese)
张松,张琳. 一种数据挖掘中的 W-PAM 限制聚类算法[J]. 计算机科学,2016,43(S2):447-450.

[3] PANG T J, LIANG J Y. A large standard hybrid data clustering algorithm based on sampling[J]. Computer Science, 2016, 43(9): 209-212. (in Chinese)
庞天杰,梁吉业. 一种基于抽样的大规范混合数据聚类集成算法[J]. 计算机科学,2016,43(9):209-212.

[4] MORADI M, KEYVANPOUR M R. An analytical review of XML association rules mining[J]. Artificial Intelligence Review, 2015, 43(2): 277-300.

[5] DONG G L, RYU K S, BASHIR M, et al. Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction[J]. Journal of Medical Systems, 2013, 37(2): 1-10.

[6] KHALILI A, SAMI A. SysDetect: a systematic approach to critical state determination for industrial intrusion detection systems using Apriori algorithm[J]. Journal of Process Control, 2015, 2776: 154-160.

[7] MERNIK M, LIU S H, KARABOGA M D, et al. On clarifying misconceptions when comparing variants of the Artificial Bee Colony Algorithm by offering a new implementation[J]. Information Sciences, 2015, 291(10): 115-127.

[8] ZHAO X J, SUN Z X, YUAN Y. An efficient association rule mining algorithm based on predetermined screening[J]. Journal of Electronics & Infromation Technology, 2015, 37(7): 1620-1625.

[9] XU K Y, GONG X R, CHENG M C. An audit log association rule mining based on improved Apriori algorithm[J]. Journal of Computer Applications, 2016, 36(7): 1847-1851.

[10] LI M D, ZHAO H, WENG X W, et al. Differential evolution algorithm based on optimal gaussian random walk and individual selection strategy[J]. Control and Decision, 2016, 31(8): 1379-1386.

[11] BI S, HO C K, ZHANG R. Wireless powered communication: opportunities and challenges[J]. IEEE Communications Magazine, 2015, 53(4): 117-125.

[12] SUN L J, CHEN X D, HAN C, et al. New Fuzzy-Clustering Algorithm for Data Stream[J]. Journal of Electronics & Infromation Technology, 2015, 37(7): 1620-1625.

[13] BI A Q, DONG A M, WANG S T. A dynamic data stream clustering algorithm based on probability and exemplar[J]. Journal of Computer Research and Development, 2016, 53(5): 1029-1042.

[14] XING C Z, LIU J. Evolutionary data stream clustering algorithm based on integration of affinity propagation and density[J]. Journal of Computer Applications, 2015, 35(7): 1927-1932.

[15] CHEN H, WAN G X, XIAO Z J. Intrusion detection method of deep belief network model based on optimization of data processing[J]. Journal of Computer Applications, 2017, 37(6): 1636-1643.

[16] ZHANG X C, SANG R T, ZHOU Z H, et al. A Short Text Classification Method Based on Two-Channel Convolutional Neural Network [J]. Journal of Chongqing University of Technology (Natural Science), 2019, 33(1): 45-52. (in Chinese)
张小川,桑瑞婷,周泽红,等. 一种基于双通道卷积神经网络的短文本分类方法法[J]. 重庆理工大学学报(自然科学),2019,33(1):45-52.