基于逻辑斯蒂回归的恶意请求分类识别模型

陈春玲,吴 凡,余 瀚
(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要: 为了解决针对 Web 应用层的攻击,有效分类识别恶意请求,深入研究有监督的学习方法,针对请求文本内容不足、特征稀疏的缺陷,提出了一种基于非重复多 N-Gram 的 TF-IDF 分词策略和逻辑斯蒂回归方法构建的恶意请求分类模型。通过从 Secrepo 安全数据样本库等来源采集到的大量样本数据进行特征提取后对模型进行训练,以最大似然估计作为模型的优化目标,利用梯度下降的方法得到最优分类模型,并在测试集上验证模型的可靠性。实验结果表明,短文本、低语义的请求内容通过字母形式在多 N-Gram 的分词下构造的分类模型,相对于单词和单倍 N-Gram 分词的分类模型具有较高的分类准确率和得分,并且训练模型所耗时间相差不大。该方法训练出的最终模型在测试集上的准确率、召回率和 F_1 值都达到了 99%以上。

关键词: Web 请求; 逻辑斯蒂回归; 最大似然估计; TF-IDF; 分类模型
中图分类号: TP301.6 **文献标识码:** A **文章编号:** 1673-629X(2019)02-0124-05
doi: 10.3969/j.issn.1673-629X.2019.02.026

A Classification and Recognition Model of Malicious Requests Based on Logistic Regression

CHEN Chun-ling, WU Fan, YU Han
(School of Computer Science & Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In order to effectively defend the attack from Web application layer and classify and recognize the malicious requests, the supervised learning methods are researched in-depth. Aiming at the defects of insufficient content and sparse features of requests text, we propose a malicious requests classifier model based on logistic regression method and TF-IDF word segmentation with non-repetition and multi-N-Gram. The model is trained after feature extraction of a large number of sample data collected from online security database such as Secrepo. Taking the maximum likelihood estimation as the optimization goal of the model, we use the gradient descent method to obtain the optimum classification model, and its reliability is validated on the test set. The experiment shows that compared with the classification model of words and single-fold N-Gram segmentation, the classification model built by request content with short text and low semantic in letters on multi-N-Gram segmentation has higher accuracy and score. Their training time is not much different. The final model trained by this way reaches more than 99% of accuracy, recall and F_1-measure on test set.

Key words: Web requests; logistic regression; maximum likelihood estimation; TF-IDF; classification model

0 引 言

随着 Web2.0 的发展,Web 应用的数量和覆盖面得到了极大的提升,然而 Web 安全性问题却越来越突出。根据 Garter 公司的报告显示,目前有 75%的网络安全漏洞都是针对 Web 应用层的。目前常见的 Web 应用安全问题有跨站脚本攻击(XSS)、SQL 注入、远程命令执行、目录遍历、PHP 代码注入等等[1]。由于大部分 Web 应用攻击都选择避开传统防火墙,转而伪装成正常请求从 OSI 应用层[2-3]入侵,因此识别策略必须要理解请求内容,从中识别出恶意的请求并拒绝执行。目前对文本进行建模分类的常用方法是机器学习。Justin Ma 等[4]提出了一种通过对 URL 的分类检测来识别恶意站点的方案,使用静态方法基于主机来发现恶意站点,但是一旦一些知名站点被挂马,这种方

收稿日期: 2018-03-01 **修回日期:** 2018-07-05 **网络出版时间:** 2018-11-15
基金项目: 国家自然科学基金(11501302)
作者简介: 陈春玲(1961-),男,硕士,教授,研究生导师,研究方向为软件工程、分布式组件技术、网络信息安全及其应用; 吴 凡(1993-),男,硕士研究生,研究方向为云计算与物联网技术。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20181115.1047.030.html

法就无法检测出来。叶飞等[5]提出了一种基于支持向量机(SVM)分类算法的黑盒检测方法,能直接对 URL 特征进行检测,而无需了解请求的源代码,但是分类准确率还有待提升。马艳发[6]提出了一种与自学习 ML 模型相结合的新型入侵检测算法,能对 PHP 变异 Webshell 的代码进行分析,追踪程序执行中函数的执行情况,从而达到检测的目的。

现有文献都是基于对请求的内容进行分析建模,设计出对应的分类识别模型。但是存在适用范围小、分类准确率不高、训练时间长等问题。因此,文中提出一种基于逻辑斯蒂回归[7-8]的有监督学习模型,通过 Python 编程实现,在 Secrepo 安全数据样本库和 GitHub 代码仓库的大量真实样本上进行训练,找到似然估计最高的参数模型作为最终分类模型,并通过在测试集上进行实验来验证其分类准确率。

1　基础知识

1.1　逻辑斯蒂回归模型

设 X 是连续随机变量,X 服从逻辑斯蒂分布是指 X 具有下列的分布函数和密度函数:

$$F(x)=P(X\leqslant x)=\frac{1}{1+e^{-\frac{(x-\mu)}{\gamma}}} \tag{1}$$

$$f(x)=F'(X\leqslant x)=\frac{e^{-(x-\mu)/\gamma}}{\gamma(1+e^{-(x-\mu)/\gamma})^2} \tag{2}$$

其中,μ 表示位置参数; γ 表示形状参数。通常令 μ 为 0,γ 为 1/2,即 $F(x)=\frac{1}{1+e^{-\frac{x}{2}}}$,该函数称之为 sigmod 函数。

逻辑斯蒂回归是一种二分类模型,由条件概率分布 $P(Y|X)$ 表示,形式就是参数化的逻辑斯蒂分布。其中自变量 X 取值为实数,因变量 Y 为 0 或者 1。二项逻辑斯蒂回归的条件概率公式为:

$$P(Y=1|x)=\frac{e^{\omega\cdot x}}{1+e^{\omega\cdot x}} \tag{3}$$

$$P(Y=0|x)=\frac{1}{1+e^{\omega\cdot x}} \tag{4}$$

ω 即为要求解的模型参数,通常采用最大似然估计。即找到一组参数,使得在这组参数下数据的似然度达到最大。

设: $P(Y=1|x)=\pi(x)$,$P(Y=0|x)=1-\pi(x)$,则似然函数为:

$$L(\omega)=\prod[\pi(x_i)]^{y_i}[1-\pi(x_i)]^{(1-y_i)} \tag{5}$$

式 5 的对数似然函数为:

$$\ln L(\omega)=\sum[y_i(\omega\cdot x_i)-\ln(1+e^{\omega\cdot x_i})] \tag{6}$$

对数似然损失在单个数据点上的定义为:

$$-y\ln p(y|x)-(1-y)\ln[1-p(y|x)]=-[y_i\ln\pi(x_i)+(1-y_i)\ln(1-\pi(x_i))] \tag{7}$$

则整个数据集上的平均对数似然损失为:

$$J(\omega)=-\frac{1}{N}\ln L(\omega) \tag{8}$$

因此求最小化对数似然损失函数和最大化对数似然函数是等价的。文中通过梯度下降法求解 ω 的估计值。梯度下降算法迭代步骤为:

(1) 取初始值 $\omega_0\in R^n$,令 $k=0$。

(2) 计算 $J(\omega_k)$。

$$J(\omega_k)=-\frac{1}{N}\ln L(\omega_k)\Rightarrow-\ln L(\omega_k)=\sum[y_i(\omega_k\cdot x_i)-\ln(1+e^{\omega_k\cdot x_i})]$$

(3) 计算梯度 $g_k=g(\omega_k)=\nabla J(\omega)\in g(\omega_k)=\sum[x_i\cdot y_i-\frac{x_i\cdot e^{\omega_k\cdot x_i}}{1+e^{\omega_k\cdot x_i}}]=\sum[x_i\cdot y_i-\pi(x_i)]$。

若 $\|g_k\|<\varepsilon$,$\omega^*=\omega_k$,转步骤 5。

否则,令 $p_k=-g(\omega_k)$,求 λ_k,使得 $J(\omega_k+\lambda_k p_k)=\min[J(\omega_k+\lambda_k p_k)]$。

(4) $\omega_{k+1}=\omega_k+\lambda_k p_k$,计算 $J(\omega_{k+1})$。当 $\|J(\omega_{k+1})-J(\omega_k)\|<\omega$ 或 $\|\omega_{k+1}-\omega_k\|<\omega$,$\omega^*=\omega_{k+1}$,转步骤 5,否则,令 $k=k+1$,转步骤 3。

(5) 算法结束,输出 ω^*。

正则化: 当模型的参数过多时,很容易遇到过拟合的问题[9]。而正则化是结构风险最小化的一种实现方式[10-11],通过在经验风险上加一个正则化项,来惩罚过大的参数以防止过拟合。即:

$$J(\omega)=>J(\omega)+\lambda\|\omega\|_p \tag{9}$$

$p=1$ 或者 $p=2$,表示 L_1 范数或 L_2 范数。L_1 范数是指向量中各个元素的绝对值之和; L_2 范数是指向量各元素的平方和的平方根。式 9 中 λ 的作用为 λ 权衡拟合能力和泛化能力对整个模型的影响,λ 越大,对参数值惩罚越大,泛化能力越好[12]。文中使用 L_2 范数对模型参数进行正则化。

1.2　TF-IDF 方法

TF-IDF[13-14](term frequency-inverse document frequency)是一种用于资讯检索与资讯探勘的加权技术,是一种统计方法,用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数呈正比增加,但同时会随着它在语料库中出现的频率呈反比下降。

词频(term frequency,TF)指的是某一个给定的词语在该文件中出现的频率。对于在特定文档 j 中的词汇 i 来说,它出现的频率可表示为:

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_{k=0}^{n} n_{k,j}} \tag{10}$$

其中，$n_{i,j}$ 为词 i 在文档 j 中的出现次数，分母表示在文档 j 中所有字词的出现次数之和。

逆向文件频率（inverse document frequency，IDF）是一个词语普遍重要性的度量。某一特定词语的 IDF，可以由总文件数目除以包含该词语的文件数目，再将得到的商取对数得到：

$$\mathrm{idf}_i = \log\left(\frac{|D|}{|\{j: t_i \in d_j\}|}\right) \tag{11}$$

其中，$|D|$ 是语料库中的文件总数；$|\{j: t_i \in d_j\}|$ 是包含词语 t_i 的文件数目（即 $n_{i,j} \neq 0$ 的文件数目）。如果该词语不在语料库中，就会导致被除数为零，因此一般情况下使用 $1 + |\{j: t_i \in d_j\}|$，然后计算出单词 i 在文件 j 中的 TF-IDF 值：

$$\mathrm{fidf}_{i,j} = \mathrm{tf}_{i,j} * \mathrm{idf}_i \tag{12}$$

2 逻辑斯蒂回归分类模型的构造

逻辑斯蒂回归分类模型通过构造一个二分类逻辑斯蒂回归方程，对未知的请求进行分类。回归方程构造过程分为取样、特征选取、参数拟合、测试。首先训练样本的丰富性和可靠性是有监督学习中非常重要的基础，为了保证实验结果的可靠性，使构造的逻辑斯蒂回归方程准确率更高，选取了来自 Secrepo 安全数据样本库和 GitHub 代码仓库中的数据集合，且采集使用的样本量较大，以便覆盖多种恶意请求。然后基于给定的样本数据通过合适的分词策略后得到相应的特征矩阵，分词策略的选择应当充分考虑分类准确率和训练时消耗的时间，选取最合适的方案。最后以最大似然估计为约束拟合出分类准确率最高的回归方程系数组合，将其带入原方程，作为最终的分类模型对测试数据进行分类，见图 1。

数据特征的选取方式对于逻辑斯蒂回归模型是至关重要的。文中采用了 TF-IDF 方法构造样本数据的特征向量。TF-IDF 是一种用来衡量一个关键词对一个词库中一份文档的重要程度的统计方法，关键词的重要程度随着它在整个词库中出现的频率呈反比下降。TF-IDF 的优势在于可以过滤掉一些对相似性检测毫无用处的词，即所谓的停用词。因为在式 11 中，一旦词库中的文档总数 $|D|$ 与包含关键词 t_i 的文档数相等，就会导致 idf 为 0，这表明该类关键词与文档主题几乎没有关系。去掉这些干扰词后可以突出重要的关键词，以提高逻辑斯蒂分类回归模型的准确率。但是由于请求字符串文本具有长度极短、单词字符间关联度较低、语义不明确等特点，常规分词处理后的特征向量字典将会非常庞大，而单条请求的有效特征又非常少，结果将导致特征矩阵非常稀疏，不利于之后特征权重向量的构造。

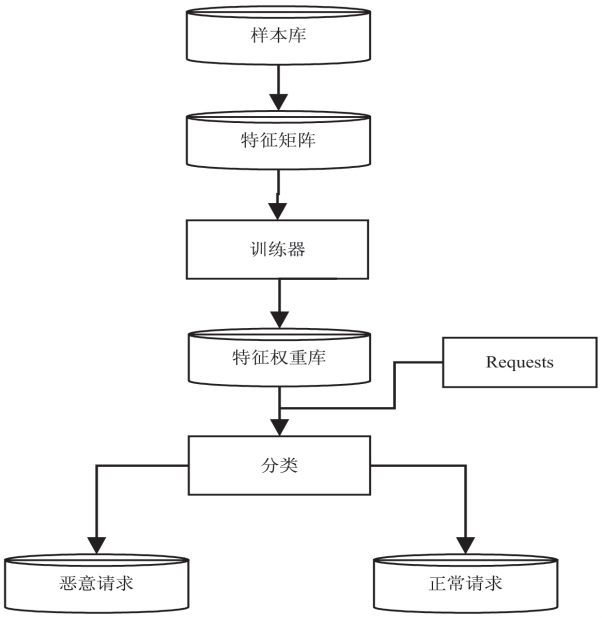

图 1 逻辑斯蒂回归分类模型

为了减少请求文本的这种特性对构造最终模型可能产生的影响，在构造特征字典时引入了非重复的 N-Gram[15-16] 分词模型，表示为：$|G_N(s)| + |G_N(t)| - 2 * |G_N(s) \cap G_N(t)|$。其中，$|G_N(s)|$ 是字符串 s 的 N-Gram 集合。通过非重复的 N-Gram 分词，相邻的词会组合成新的分词，成为特征词典的一部分，既能隐性地增加文本长度，提高有效特征比例，又能将请求中某些暗含的固定字词联系发掘出来，成为潜在的重要特征。一般 N-Gram 分词中 N 取 2 或 3，数值太高将导致衍生词过多，同样会导致特征矩阵系数，降低分类准确率。

特征矩阵构造完成后，将其带入二分类逻辑斯蒂回归模型中，求解模型的损失函数，通过梯度下降法递归地求解出最优拟合参数 ω，将 ω 带入之前的模型中，得到该样本下的最优逻辑斯蒂回归分类模型。对于待分类的样本，只需要先通过特征字典获取样本对应的特征列表 x，然后计算 $g(x) = \mathrm{simgod}(\sum \omega \cdot x) - 0.5$ 的结果即可，若 $g(x) \geqslant 0$，则样本为正例，否则为负例。文中所指的正例即判断为恶意请求的样本。

3 实验结果与分析

3.1 特征选取

使用 Python3.6 语言编程环境，选取 Secrepo 安全数据样本库和 GitHub 代码仓库中的数据进行实验。选取的数据集包含了 8 万条正常请求以及 4 万条恶意请求。恶意请求包括 XSS、SQL 注入、远程命令执行、

目录遍历、PHP 代码注入等多种类型,比较全面地覆盖了常见的恶意请求类型。

首先对采集到的样本数据进行预处理。文中使用的是 Python 的 Scikit-Learn 包中的 TfidfVectorizer 类,该类实现了基于 TF-IDF 的特征词典的构造和特征向量的计算。通过设定该方法的 ngram_range 参数以及 analyzer 参数为 char,可实现在构造特征字典时引入非重复的 N-gram 分词且分词基于字母而非单词。最终得到的特征字典大小为 73 405。以/rss.php? page [path] =XXpathXX? &cmd=ls 为例,该文本在特征矩阵中非零参数有 85 个,这意味着 TF-IDF 分词算法将该词条分割成了 85 个非重复特征单元,并计算了每个特征对于本词条在全部样本中的影响力。由于请求字段长度短、非语义化等特点,如果采用基于单词的分词方法进行特征选取和 TF-IDF 的计算,容易出现单个样本特征稀少而特征词典过于庞大,导致特征矩阵过于稀疏的情况,不利于提高模型分类的准确性。在以单词为单位的特征字典中,案例文本仅能获得 17 个非零参数,而特征字典大小达到了惊人的 260 095。

3.2　实验结果

将样本按指定的分词策略进行分词后得到的特征矩阵按 4∶1 分成两部分,分别为训练集 A 和测试集 B,其中 A 作为训练样本带入逻辑斯蒂回归方程进行训练,将训练完成后得到的最终模型用于测试集 B,检验该模型的分类准确性。通常评估模型好坏的指标为准确率(accuracy)、精确率(precision)、召回率(recall)和 F_1-Measure。假设 TP 为正类中判定为正类的数量,FP 为负类中判定为正类的数量,FN 为正类中判定为负类的数量,TN 为负类中判定为负类的数量,则可以定义为准确率=(TP+TN)/总样本数,精确率=TP/(TP+FP),召回率=TP/(TP+FN),F_1=2TP/(2TP+FP+FN)。

测试结果如表 1 所示,其中前四条数据是不同参数的 TF-IDF 分词方法通过逻辑斯蒂回归得到的结果,第五条数据为在分词标准 char、N-Gram 系数为 3 的情况下通过 SVM 训练得到的结果。

表 1　不同分词对应模型的分类结果

序号	分词标准	N-Gram 系数	准确率	精确率	召回率	F_1 值	训练时间/s
1	word	1	0.978	0.975	0.960	0.967	0.624
2	word	3	0.981	0.983	0.960	0.971	1.336
3	char	1	0.945	0.883	0.967	0.923	1.031
4	char	3	0.994	0.988	0.995	0.992	2.036
5	char	3	0.982	0.991	0.956	0.973	491

图 2 为不同分词情况下模型的分类结果的对比。

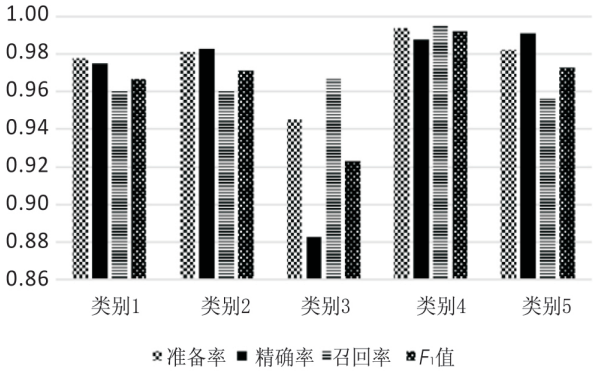

图 2　不同分词情况下模型的分类结果对比

3.3　结果分析

由实验结果可知,同为逻辑斯蒂回归模型的四种情况中,类别四分词策略取得了最好的效果,四种指标都超过了 99%,其中召回率的数据超出其他策略很多,而分类模型中召回率越高,正类被漏判的概率就越低,这也意味着对正常用户请求的误伤概率降到了最低。既保障了最高的分类检出率,也不会过于影响用户体验。对比 SVM 的训练模型指标,类别四也有一些优势。

而在训练用时方面,由于使用字母和较高的 N-Gram 系数分词策略,使得每条样本的非零特征增多,在训练时会增加一些时间。从实验结果来看,增加的时间在可忍受范围内,但是提高了模型的分类准确率,因此是可取的。

从表 1 中能发现 SVM 模型的训练用时相对于逻辑斯蒂回归模型多了很多,这可能是由于 SVM 会将全部特征进行空间映射,从而找出一个线性可分的超平面作为分类的标准,即所谓的支持向量,然而文中特征选取的策略会导致特征空间非常庞大,导致 SVM 模型的训练时间很长,而逻辑斯蒂回归模型在这种情况下效率仍很高。总之,实验结果表明,类别四的分词策略配合逻辑斯蒂回归模型对于文中的请求分类具有最好的结果。

4　结束语

提出的二分类逻辑斯蒂回归分类模型通过使用基于 TF-IDF 的非重复 N-Gram 分词,有效避免了对请求文本进行分词时遇到的诸多不利因素,使训练出的模型能有效分类新遇到的请求,识别其中的恶意请求,

从而弥补了传统安全防御模式在应用层上的不足，提高了服务器端的安全水平。该模型以样本的最大似然估计为训练目标，使用 L_2 范数达到最大泛化效果，避免模型出现过拟合现象。在从 Secrepo 安全数据样本库和 GitHub 代码仓库采集的数据集下进行实验，结果表明该分类模型的分类准确率、精确率、召回率和 F_1 值都较高，且训练时间开销不大，可在极低的误判下有效识别出恶意请求。但是该模型基于有监督学习，面对形式新颖、变化较大的恶意请求攻击模式可能无法做到有效识别，因此还需要对机器学习算法、恶意请求形式进行更深入的研究。

参考文献:

[1] RAFIQUE S，HUMAYUN M.Systematic review of web application security vulnerabilities detection methods [J] .Journal of Computer and Communications，2015，3：28-40.

[2] SHIROSHITA T.A data processing performance model for the OSI application layer protocols [J] .ACM SIGCOMM Computer Communication Review，1990，20(4)：60-68.

[3] KOŁODZIEJCZYK M，OGIELA M. Applying of security mechanisms to middle and high layers of OSI/ISO network model [J] .Theoretical and Applied Informatics，2012，24(1)：95-106.

[4] MA J，SAUL L K，SAVAGE S，et al.Identifying suspicious URLs：an application of large-scale online learning [C] //Proceedings of the 26th annual international conference on machine learning. [s.l.]：[s.n.]，2009：681-688.

[5] 叶 飞，龚 俭，杨 望.基于支持向量机的 Webshell 黑盒检测 [J] .南京航空航天大学学报，2015，47(6)：924-930.

[6] 马艳发.基于 WAF 入侵检测和变异 WebShell 检测算法的 Web 安全研究 [D] .天津：天津理工大学，2015.

[7] LEE E T.A computer program for linear logistic regression analysis [J] .Computer Programs in Biomedicine，1974，4(2)：80-92.

[8] SABERI H，RAHAI A，HATAMI F. A fast and efficient clustering based fuzzy time series algorithm (FEFTS) for regression and classification [J] .Applied Soft Computing，2017，61：1088-1097.

[9] KIM K I，SIMON R.Overfitting，generalization，and MSE in class probability estimation with high-dimensional data [J] .Biometrical Journal，2014，56(2)：256-269.

[10] 朱劲夫，刘明哲，赵成强，等.正则化在逻辑回归与神经网络中的应用研究 [J] .信息技术，2016，40(7)：1-5.

[11] XU Chen，PENG Zhiming，JING Wenfeng.Sparse kernel logistic regression based on L1/2 regularization [J] .Science China：Information Sciences，2013，56(4)：75-90.

[12] LU Shuai，PEREVERZEV S V，SHAO Yuanyuan，et al.Discrepancy curves for multi-parameter regularization [J] .Journal of Inverse and Ill-Posed Problems，2010，18(6)：655-676.

[13] DIVYA K S，SUBHA R，PALANISWAMI S.Similar words identification using naive and TF-IDF method [J] .International Journal of Information Technology and Computer Science，2014，6(11)：42-47.

[14] 王小林，杨 林，王 东，等.改进的 TF-IDF 关键词提取方法 [J] .计算机科学与应用，2013，3(1)：64-68.

[15] 张家旺，李燕伟.基于 N-gram 算法的恶意程序检测系统研究与设计 [J] .信息网络安全，2016(8)：74-80.

[16] 徐建平.基于改进的 N-gram 恶意 PDF 文档静态检测技术研究 [D] .上海：华东理工大学，2017.

(上接第 123 页)

A，2014，378(7-8)：635-640.

[8] LI C H.Dynamics of a network-based SIS epidemic model with nonmonotone incidence rate [J] .Physica A：Statisical Mechanics and Its Applications，2015，427：234-243.

[9] CHEN Lijuan，SUN Jitao.Global stability and optimal control of an SIRS epidemic model on heterogeneous networks [J] .Physic A：Statistical Mechanics and Its Applications，2014，410：196-204.

[10] GRAHAM M，HOUSE T.Dynamics of stochastic epidemis on heterogeneous networks [J] .Journal of Mathematical Biology，2014，68(7)：453-485.

[11] PASTOR-SATORRAS R，VESPIGNANI A.Epidemics and immunization in scale-free networks [M] //Handbook of graphs and networks. [s.l.]：[s.n.]，2003.

[12] 郭进利.复杂网络和人类行为动力学演化模型 [M] .北京：科学出版社，2013.

[13] 巩永旺，宋玉蓉，蒋国平.移动环境下网络病毒传播模型及其稳定性研究 [J] .物理学报，2012，61(11)：110205.

[14] 吕 剑，宋玉蓉，蒋国平.自适应网络异步元胞自动机病毒传播模型 [J] .计算机技术与发展，2012，22(7)：132-135.

[15] SAHNEH F D，SCOGLIO C.Epidemic spread in human networks [C] //Proceedings of CDC-ECC. [s.l.]：IEEE，2011：3008-3013.