doi :10.3969/j.issn.1671-1122.2016.11.008

基于流量行为特征的异常流量检测

胡洋瑞，陈兴蜀，王俊峰，叶晓鸣
（四川大学计算机学院，四川成都 610065）

摘　要：针对真实网络流量缺乏标记数据集的问题，文章提出了一种无监督异常流量检测方法。通过对四川大学网络出口流量行为的分析和研究，构建了用户行为特征集，利用改进的 *k*–means++ 余弦聚类方法建立正常流量行为模型，通过度量流量行为与正常行为模型之间的偏离距离以识别异常流量。文章通过 Spark 大数据处理平台实现了特征抽取、*k*–means 改进算法和异常检测的研发，通过实验验证了该方法的可行性和有效性，实验结果表明文章提出的方法对异常流量行为检测具有较高的准确性和敏感性。

关键词：大数据；异常流量检测；*k*–means
中图分类号：TP391　　**文献标识码：**A　　**文章编号：**1671–1122（2016）11–0045–07

中文引用格式：胡洋瑞，陈兴蜀，王俊峰，等 . 基于流量行为特征的异常流量检测 [J]. 信息网络安全，2016（11）：45–51.

英文引用格式：HU Yangrui, CHEN Xingshu, WANG Junfeng, et al. Anomalous Traffic Detection Based on Traffic Behavior Characteristics [J]. Netinfo Security, 2016 (11) : 45–51.

Anomalous Traffic Detection Based on Traffic Behavior Characteristics

HU Yangrui, CHEN Xingshu, WANG Junfeng, YE Xiaoming
(*College of Computer Science of Sichuan University, Chengdu Sichuan* 610065, *China*)

Abstract: Real network environment lack of labeled data set, so traditional anomaly traffic detection method based on labeled data set is unsuitable for actual large-scale network. To resolve this, the paper proposes an improved *k*-means anomaly traffic detection method for unlabeled data sets. Firstly, collect the Sichuan University network outlet flow and store in the distributed file system; secondly, construct user behavior feature set on the basis of network flow analysis, and extract relevant characteristics by Spark big data processing platform. Referenced principles of group to define the normal behavior of clusters in the actual flow, construct normal flow behavior model on improved K-means++ cosine clustering method; Finally, the cosine distance between the normal behavior model and user actual flow behavior is calculated to detected anomaly flow behavior. The feasibility and validity of the method are verified by attacking experiment. The experimental results show that the normal flow behavior model for anomaly flow detection has higher accuracy.

Key words: big data; anomaly traffic detection; *k*-means

收稿日期：2016–07–01
基金项目：国家自然科学基金 [61272447]

作者简介：胡洋瑞（1991—），男，四川，硕士研究生，主要研究方向为信息安全、数据分析；陈兴蜀（1968—），女，四川，教授，博士，主要研究方向为大数据、云安全与网络安全；王俊峰（1976—），男，四川，教授，博士，主要研究方向为空间信息网络、智能交通；叶晓鸣（1981—），女，四川，博士研究生，主要研究方向为基于大数据安全的网络流量分析。
通信作者：陈兴蜀 chenxsh@scu.edu.cn

0 引言

随着信息时代的到来，网络中异常流量激增，网络拥塞问题日益突显，通过异常流检测减小网络流量负载，缓解网络拥塞问题是当前面临的迫切问题。异常流量检测是一种通过对保护系统信息收集和分析，从而发现异常的技术。它主要是通过对计算机系统和网络进行实时监控，发现和识别网络流量中的异常流量，给出异常流量警报。可将异常流检测看作是区分“正常”还是“异常”的二分类问题[1]。对异常流检测系统的要求是正确性和实时性，这样才能及时处理网络中传输的海量数据，不会因为速度慢而丢失信息、造成漏警，更能及时采取措施，将异常流量带来的损失降到最低。在有关文献中，HE[2] 等人基于相似流，在相似数据预测的思想上，提出了Grubbs-KNN算法，并将其应用到实时网络流量检测中，实现了大数据异常流量的实时检测。HUANG[3] 等人对无监督学习上的各种网络异常流量检测方法做了一个归纳总结。王苏南[4] 提出了一种新的流量模型TSTM，该模型提出了一种可用于高速复杂网络环境异常流量检测的URCA无监督异常流量检测算法。许晓东[5] 等人基于 k-means方法，在网络流量的多维特征下进行了无监督网络流量异常检测。FORREST[6] 等人把异常检测看作是区分“自我”(也就是“正常”)和“非自我”(也就是“异常”)的过程，提出了基于免疫模型的异常检测技术。

GHOSH[7] 等人利用神经网络来提取特征和分类。LEE[8] 等人从数据挖掘技术的角度探讨了异常检测的实现问题。此外，常用的异常检测方法[9] 有：基于特征选择的异常检测方法[10,11]，基于贝叶斯推理[12]、贝叶斯网络[13] 的异常检测方法，基于模式预测的异常检测方法[14]，基于贝叶斯聚类的异常检测方法[15]。

基于统计的异常检测方法对所观测到的行为进行识别，主要使用的是有监督的学习识别，通过研究用户行为建立用户行为轮廓。在实际的网络环境下，随着用户规模的扩大产生海量的数据，数据量累计增长，人工标记的方法在大数据面前显得无能为力。通过小规模标记数据集训练得到的模型检测在海量数据中的异常检测效果并不理想。本文拟采用无监督网络异常检测技术解决大数据环境下数据标记难和训练模型检测技术精确率低等问题。本文在前人工作的基础上，根据文献[16]网络中异常流量相对于正常流量是小众行为的思想定义了主流行为簇，在实际流量中根据实际情况变更特征可信值域，解决复杂环境中特征可信值域随时间的变化特点。本文直接在实际网络的数据中进行数据提取，训练正常行为模型。同时，由于用户的相同行为在时间上的累积会对提取的流行为特征集数据在欧式距离上产生影响，导致聚类的模型不收敛或模型中的流行为相似度的问题，因此本文基于流行为特征向量，提出了 k-means++ 余弦聚类算法，采用余弦聚类来避免数据在时间累积上的影响导致的模型误差，完成模型建立，实现对异常流的检测。

1 流行为描述与特征提取

1.1 网络行为流

网络行为是伴随现代网络技术出现的，可定义为：行为主体为实现某种特定的目标，采用基于计算机系统的电子网络作为手段和方法而进行的有意识的活动。网络用户行为可以从多个角度来描述，本文立足于流的角度来描述用户的网络行为，称其为网络用户行为流。为构建网络用户行为特征，本文使用华为公司的Netstream流统计工具，其采集的流有丰富的用户行为流属性，本文此后的工作都是在有Netstream统计流的基础之上进行的，表1给出了本文立足于流角度，构建网络行为流使用到的Netstream流的字段描述。

表1 NetStream 字段描述

名称	描述
Srcaddr	源IP地址
Dstaddr	目标IP地址
Packets	流中的封包数量
dOctets	流中第3层字节数
Srcport	TCP/UDP源端口
Sstport	TCP/UDP目标端口，ICMP类型和代码
tcp_flags	流中所有TCP标志进行“或”运算的结果
Prot	IP协议

1.2 流行为分析及特征提取

网络用户行为的研究与心理学、社会学、社会心理学、人类学等学科有关。网络用户行为的分析就是运用多学科知识研究和分析网络用户的构成、特点及其在网络应用过程中行为活动上所表现出来的规律。

网络用户行为可以用某些特征量的统计特征[17] 或特

征量的关联关系定量或定性地表示。网络用户行为常见的表示有四元组（源 IP、目的 IP、统计参数、统计参数值）[18]。其中统计参数的选取可以根据研究的目的而定。流行为是网络用户行为在流上的体现，本文立足于网络用户行为在流量上的特征，通过对流量数据进行采样，分析正常用户上网行为的流量特征，借鉴文献 [19] 中对网络用户行为特征描述的成果，在 Netstream 流数据的基础上构建了用户行为特征集。最后，得出如表 2 所示的流行为特征表。

表2 流行为特征

源IP统计属性		目的IP统计属性	
S1	数据包字节数	D1	数据包字节数
S2	数据包包数	D2	数据包包数
S3	源端口个数	D3	源端口个数
S4	目的端口个数	D4	目的端口个数
S5	目的IP个数	D5	源IP个数
S6	前N个协议流量占比	D6	前N个协议流量占比
S7	前N个源端口流量占比	D7	前N个源端口流量占比
S8	前N个目的端口流量占比	D8	前N个目的端口流量占比

1.3 流特征提取技术

本文的研究工作是基于 Spark 技术实现的，Spark 是 UC Berkeley AMP lab 所开发的开源的类 Hadoop MapReduce 的通用的并行计算框架。Spark 基于 MapReduce 算法实现的分布式计算，拥有 MapReduce/Hadoop 具有的优点；但不同于 MapReduce 的是 Job 中间输出和结果可以保存在内存中，从而不再需要读写 HDFS，因此 Spark 能更好地适用于数据挖掘与机器学习等需要迭代的 MapReduce 的算法。Spark 是新一代的大数据处理工具，在各方面的性能上都大幅度地优于 Hadoop-Hive[20]。在大型网络中，进行异常流量检测首要解决的是海量数据的存储与处理，为在全网流量的背景下进行数据挖掘与统计提供条件。

2 基于聚类的的异常检测算法

2.1 余弦聚类

k-means[21] 余弦聚类方法要先确定聚类数 K，C 为样例集 $C=\{C_1,C_2,\cdots,C_i,\cdots,C_k\}$，其中 C_i 为其中的某个类，$i=1,2,\cdots,k$。$x_{ij}\in C_i$，$j=1,2,\cdots,|C_i|$，中心 u_i 代表对属于同一个类的样本中心点的猜测。首先随机选取 k 个中心点；然后对每一个样本 x_{ij} 计算其到 k 个中心的余弦距离；其次选取余弦距离最大的中心作为样例 x_{ij} 所属类别，经过这一步后每一个样例都有了所属的类；最后，对于每一个类，重新计算它的中心 u_i（同类所有样例坐标加和的平均值）。重复迭代第二步和第三步直到中心不变或者变化很小，具体步骤方程如下：

Step1: 随机选取 k 个聚类中心点 $u_1,u_2,\cdots,u_k\in R^n$。

Step2: 对每一个样例 x，计算其应该属于的类

$$C_i=\max_{i=1}^{k}(\cos(u_i,x)) \quad \cdots\cdots(1)$$

Step3: 对于每一个类 i，重新计算该类的中心

$$u_i=\frac{\sum_{j=1}^{|C_i|}x_{ij}}{|C_i|} \quad \cdots\cdots(2)$$

Step4: 重复以上 Step1 到 Step3 直到中心不变或变化很小为结束。

2.2 类数的确定

分类类别数的确定在聚类算法中一直是一个难题，大多数是靠经验来确定最后的类别数，而靠经验来确定最后的类别数，存在主观因素的成分。同时，靠经验来确定类别数很大程度上存在不准确性，在数据量较大时存在一定的困难性 [22-24]。本文采用 TIBSHIRANI 提出的聚类数确定算法间隙统计（Gapstatistic）[25] 来确定类别数。

间隙统计方法中的基本定义数据集中每个观测点用 x_i 表示，$i=1,2,\cdots,n$。共有 n 个相互独立的观测点．观测点用 x_i 可以表示成向量 $(x_{i1},x_{i2},\cdots,x_{ij})$，$j=1,2,\cdots,p$，即每个观测点具有 p 个特征．令 $d_{ii'}$ 表示观测点 i 到观测点 i' 的距离，在此选择的是应用广泛的欧几里德距离：

$$d_{ii'}=\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 \quad \cdots\cdots(3)$$

将数据分为 k 类 $C_1,C_2,\cdots,C_k$，在此 C_r 表示观测点属于第 r 类，$n_r=|C_r|$ 为属于类 C_r 的观测点个数。公式（4）定义第 r 类中任意两点的距离和：

$$D_r=\sum_{i,i'\in C_i}d_{ii'} \quad \cdots\cdots(4)$$

公式（5）定义了类中的平方和均值 W_k：

$$W_k=\sum_{r=1}^{k}\frac{1}{2n_r}D_r \quad \cdots\cdots(5)$$

公式（6）定义了 $Gap_n(k)$，即观测数据集的期望值与参考数据集的间隙距离：

$$Gap_n(k)=E_n^*(\log(W_k)-\log(W_k)) \quad \cdots\cdots(6)$$

E_n^* 表示对参考数据集的期望值。间隙统计算法的基本思路是比较观测数据集 $\log(W_k)$ 与参考数据集的期望值 $E_n^*(\log(W_k))$，寻找 $\log(W_k)$ 下降最快的 k 值为最优聚类数。关于参考数据集的生成，本文使用文献 [25] 中提出的两种

参考数据集生成方法中的第一种。即先确定观测数据集所有特征表达值的取值区间，每个参考数据集的特征表达值平均分布在观测数据集相应特征表达值的取值区间内。为使参考数据集尽可能均匀分布，需要生成多个参考数据集，数据集的个数用 B 表示。

间隙统计值的计算可分为以下三个步骤：

Step1: 对已知观测数据集聚类，将观测数据聚成不同的类个数进行观测，聚类数 k 取 k=1,2,⋯,K，计算 W_k。

Step2: 生成 B 个参考数据集，并计算每个参考数据集的 W^*_{kb}，b=1,2,⋯,B，k=1,2,⋯,K。计算 Gap 值，如公式（7）：

$$Gap(k)=(1/B)\sum_b \log(W^*_{kb})-\log(W_k) \quad \cdots\cdots (7)$$

Step3: 通过公式（8）选择满足条件的最小 k 值，S_{k+1} 的值由公式（9）、公式（10）、公式（11）获得：

$$Gap(k) \geqslant Gap(k+1)-s_{k+1} \quad \cdots\cdots (8)$$

$$\bar{l}=(1/B)\sum_b \log(W^*_{kb}) \quad \cdots\cdots (9)$$

$$sd_k=[(1/B)\sum_b \log(W^*_{kb})-\bar{l}] \quad \cdots\cdots (10)$$

$$s_k=sd_k\sqrt{1+1/B} \quad \cdots\cdots (11)$$

2.3 初始聚类中心的确定

k–means 聚类中，初始的聚类中心是随机选择，因此结果不稳定，很多时候得不到最好的聚类，容易陷入局部最优，针对此问题 ARTHUR 等人提出了 k–means++[26] 算法，k–means++ 算法选择初始 seeds 的基本思想就是：初始的聚类中心之间的相互距离要尽可能的远。从输入的数据点集合中随机选择一个点作为第一个聚类中心，步骤如下：

Step1: 对于数据集中的每一个点 x，计算它与最近聚类中心（指已选择的聚类中心）的距离 $D(x)$。

Step2: 选择一个新的数据点作为新的聚类中心，选择的原则是：$D(x)$ 较大的点，被选取作为聚类中心的概率较大。

Step3: 重复步骤 2 和步骤 3 直到 k 个聚类中心被选出来

Step4: 利用这 k 个初始的聚类中心来运行标准的 k–means 算法。

相较于 k–means 算法，使用 k–means++ 的聚类结果更加稳定。

2.4 余弦聚类的改进

翟东海等人在文献 [27] 中提出了将文本的余弦相似度转化为文本距离的思路，改进了 k–means 余弦聚类算法，在文本聚类上与原始 k–means、雷小锋 [28] 等人的工作进行了对比，获得了良好的聚类效果。

上文中分析了 k–means 余弦聚类的具体思路，此处针对其思路中的不足之处进行分析，进而提出改进方法。k–means 算法是采用欧式距离来计算类的中心。在余弦聚类中使用欧式距离来计算中心造成了距离度量标准不一致，欧氏距离反映数据间的距离，余弦距离反映数据间的相似度，因此传统的余弦聚类算法，由于距离度量的不一致性会出现得不到最优解、得不到最大相似度、收敛速度慢、甚至不收敛等问题。为解决余弦聚类算法收敛慢、不收敛、得不到最优解等问题。本文改进了基于余弦聚类算法，在中心迭代过程中基于余弦距离最大化的思想更新迭代中心。由此统一了距离度量标准，改进后的余弦聚类的中心迭代算法如公式（12）所示：

$$u_i=\max_{j=1}^{|C_i|}\left(\sum_{l=1}^{|C_i|}\cos(x_{ij},x_{il})\right) \quad \cdots\cdots (12)$$

改变中心迭代方法后，实验结果表明改进后的算法可以少量迭代次数达到收敛，其最终的目标函数（13）得到了全局最优解。

$$\max\sum_{i=1}^{K}\sum_{x\in C_i}\cos(c_i,x) \quad \cdots\cdots (13)$$

图 1 比较了 k–means 余弦聚类与本文中改进的 k–means++ 余弦聚类：

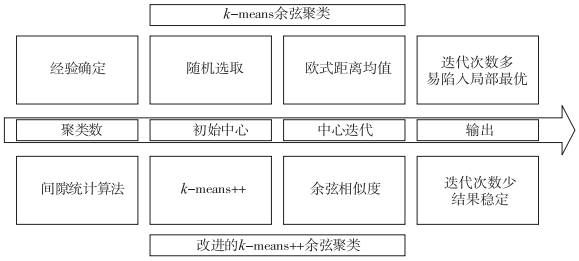

图1 算法对比

为了验证本文提出的余弦聚类中心迭代改进的有效性，本文提取了 7000 条用户流量行为数据，比较了原始 k–means 余弦聚类、文献 [27] 算法、本文算法。表 3 展示了三种算法在聚类数为 12 的情况下的聚类效果，表 3 中的模型距离定义见公式（15），模型距离本质为整个模型的行为相似程度，所得相似度越高，建立的模型越有效。

表3 聚类效果对比

算法	迭代次数	最优解（模型距离）
原始 k–means 余弦	52	0.83544301
文献[27]算法	41	0.85135648
本文算法	11	0.93719251

从表 3 中可以看出，在聚类中本文提出的改进算法在迭代 11 次的情况下获得了收敛，并且获得了模型距离的最优解。

在图 2 中，本文对比了在不同聚类数时，三种算法的迭代次数。由图 2 可以看出本文提出的改进方法，在不同聚类数下以较少的迭代次数完成了收敛。

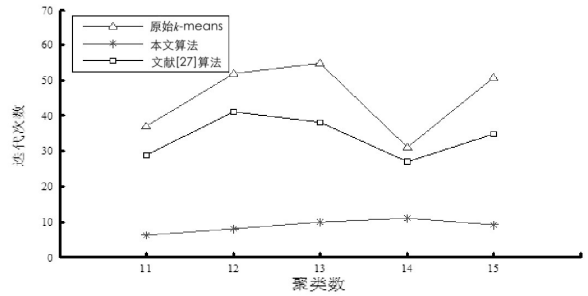

图2 迭代数对比

2.5 正常行为模型建立

由于一般的网络环境内，正常行为是主流，而入侵则表现为个别现象，因此获得的正常实例的规模远大于入侵行为数目，因而可采用异常检测的思想来发现数据集中的异常模式[16]。该方法假设入侵行为相对于正常行为 (主流的行为模式) 是些孤立的异常数据,且主流行为具有“抱团”的性质。基于以上研究成果，本文定义校园网中的主流行为簇：

定义 1 主流行为簇 假设 $C=\{C_1,C_2,\cdots,C_k\}$ 为聚类算法获得簇的集合，如果簇满足 $|C_m|=\max_{1\le j\le k}(|C_j|)$，其中，$C_j$ 表示簇的大小，那么 $|C_m|$ 就是数据集的主流行为簇。

本文在主流行为簇的基础上建立正常行为模型，基于实验结果采用最大的主流行簇作为正常行为簇，并定义了下列几个衡量参数：

假设待检测行为向量为 $\boldsymbol{w}$，正常行为模型簇为 C。$\boldsymbol{w}$ 与 C 的夹角为 $\cos\theta$，正常行为簇的大小为 $|C|$，簇中的行为向量用 $\boldsymbol{x}_i$ 表示，i=1,2,…,$|C|$，公式 (14) 定义行为向量到行为模型的余弦距离：

$$\cos\theta=\min_{i=1}^{|C|}(\cos(w,x_i)) \quad (14)$$

其中模型的余弦距离定义为：

$$D^*=\min(\min_{j=1}^{|C|}(\cos_{l=1}^{|C|}(x_{ij},x_{il}))) \quad (15)$$

相对余弦距离 D 定义为公式 (16)：

$$D=1-\cos\theta \quad (16)$$

距离度 R：

$$R=\frac{D-D^*}{D^*} \quad (17)$$

在检测中使用距离度 R 度量行为异常的程度。

正常流量行为模型建立的流程图如图 3 所示。

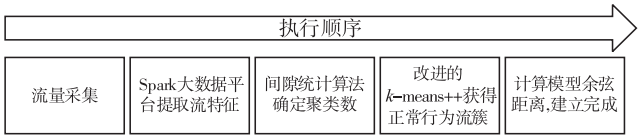

图3 模型建立流程图

3 实验与结果分析

本文采集 2015 年 12 月 22 日至 2015 年 12 月 30 日某高校校园网真实流量的数据包头信息，在 Spark 平台下对用户流量数据进行处理，通过间隙统计算法确定用户流量行为数据集分为 13 个类别。

本文的实验由 3 个实验组成。

实验 1：对已知的异常流行为进行检测，抓取实验主机攻击网络上的服务器的流量数据与正常访问时的数据流，通过正常流行为模型检测，评估模型对异常流的检测效果。

实验 2：检测不同规模的已知异常流量，通过检测不同规模的异常流来评估模型对异常流的的敏感程度。

实验 3：通过检测 1000 次被实验人员的正常上网的行为流量来评估模型对正常行为流的误报率。

3.1 已知异常流量行为检测

本文首先通过采集实验主机在进行异常行为时的流量进行特征提取，与本文训练的正常流量行为模型计算距离度，判断异常行为与正常行为模型的距离度，基于距离度的阀值来检测异常行为。最后评估对异常行为的检测效果来对本模型进行评估。本文采集了实验主机的几种攻击行为：TCP Flood、PortScan、UDP Flood，一种正常行为流量数据 Normal，在此基础上抽取出其行为特征，然后，基于本文所建立的正常流量行为模型 (Model) 进行距离度检测,得到表 4。

表4 异常流量检测

	cos θ	D	R
TCP Flood	0.81805	0.18195	1.89
Port Scan	0.83131	0.16869	1.68
UDP Flood	0.7498	0.2502	2.98
Normal	0.94782	0.05218	–0.16
Model	0.93719	0.06281	0

由表 4 可以看出各类异常行为的距离度，定义距离度阈值，当距离度阈值达到一定程度就判定为异常，距离度的选择要根据实际情况根据自己模型的余弦距离与模型距离，并结合模型的敏感程度定义，本文中针对本模型选取距离度为 1 为异常行为的阈值。由此可以看出，表 4 中的几种异常行为都被检测为异常行为，正常流量统计的距离度小于 0，表示其行为被包含在了正常行为模型范围内。由此可以看出本模型对在流量上有变化的异常行为有良好效果。

3.2 已知异常流量不同规模的检测

为检测模型在异常流量行为上的敏感程度，本文主要通过对服务器实施常见不同攻击规模的网络攻击，实现对模型敏感程度的效果评估。表 5 是模型对 UDP Flood、TCP Flood、Port Scan 三种常见网络攻击的检测效果。针对不同的攻击类型，实验分别选取了 3 种不同的攻击规模（每个时间窗口内的探测次数），共实施了 9 组实验，每组实验实施 30 次攻击，其攻击检测结果如表 5 所示。

表5 网络攻击的检测结果

攻击类型	攻击规模	攻击次数	检测次数	平均距离度	检测率(%)
TCP Flood	800	30	23	2.23	76.6
TCP Flood	1000	30	27	2.66	90.0
TCP Flood	1200	30	30	2.78	100
UDP Flood	1200	30	22	2.28	73.3
UDP Flood	1500	30	28	2.35	93.3
UDP Flood	1800	30	30	2.54	100
Port Scan	1000	30	24	2.83	80
Port Scan	1500	30	27	2.94	90
Port Scan	2000	30	30	3.13	100

结果显示模型对攻击流量行为具有良好的灵敏度，Port Scan、TCP Flood 的攻击规模到达 1000 就能达到 90% 的检测率，UDP Flood 的攻击规模到达 1500 模型能达到 93.3% 的检测率。随着攻击规模的增大，检测率和距离度也在不断增大。由实验结果可知，模型对能够影响网络流量结构的常见网络攻击具有较好的检测效果。

3.3 正常行为流量检测

为评估模型对正常行为流量的误报率，本文在 Spark 大数据平台上连续采集了特定研究人员一周的正常上网的正常行为流量数据（Normal），基于每 5 分钟为一个时间窗口，提取其流量行为特征，并抽取出其中 1000 次行为来评估模型的误报率，其实验结果如表 6 所示。

表6 正常行为检测结果

行为类型	行为次数	误报次数	平均距离度	误报率(%)
Normal	1000	45	0.14	4.5

4 结束语

本文基于网络行为流量结构的稳定性，基于用户的正常网络行为流量进行建模，根据用户的网络行为流量结构与正常网络行为在流量上的距离度进行异常流量检测。本文从正常流量固有稳定性及业务特性、用户特性等表现的流量稳定性两方面，对网络行为流量进行了研究，同时，直接从实际数据集入手，定义了正常簇，采用无监督学习真实数据集训练正常行为模型。实验结果表明，本文所建立的正常行为模型对常见网络攻击及未知网络流量异常都具有较好检测效果。（责编 吴晶）

参考文献：

[1]BHUYAN M H, BHATTACHAYYA D K, KALITA J K. Network Anomaly Detection: Methods, Systems and Tools[J]. Communications Surveys & Tutorials, 2014, 16(1): 303-336.

[2]HE G, TAN C, YU D, et al. A Real-Time Network Traffic Anomaly Detection System Based on Storm[C]//IEEE. IEEE 7th IHMSC Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), August 26 - 27, 2015, Hangzhou, China. New York:IEEE, 2015, 1: 153-156.

[3]HUANG H, ALAZZAWI H, BRANI H. Network Traffic Anomaly Detection[EB/OL]. https://arxiv.org/pdf/1402.0856.pdf，2016-6-1.

[4] 王苏南 . 高速复杂网络环境下异常流量检测技术研究 [D]. 郑州：解放军信息工程大学，2012.

[5] 许晓东，杨燕，李刚 . 基于 K-means 聚类的网络流量异常检测 [J]. 无线通信技术，2013，4（1）：21-26.

[6]FORREST S, PERELSON A S, ALLEN L, et al. Self-nonself Discrimination in a Computer[EB/OL]. http://www.cs.unm.edu/~immsec/publications/virus.pdf, 2016-6-1.

[7]GHOSH A K, MICHAEL C, SCHATZ M. A Real-time Intrusion Detection System Based on Learning Program Behavior[EB/OL]. http://link.springer.com/chapter/10.1007%2F3-540-39945-3_7#page-1, 2016-6-1.

[8]LEE W, STOLFO S J, MOK K W. A Data Mining Framework for Building Intrusion Detection Models[C]// IEEE. IEEE Conference on Security and Privacy, May 9-12, 1999. Oakland, California, USA. New York:IEEE, 1999: 120-132.

[9] 隋新，杨喜权，陈棉书，等 . 入侵检测系统的研究 [J]. 科学技术与工程，2012，12（33）：8971-8979.
[10]HAI T N, PETROVIC S, FRANKE K. A Comparison of Feature-Selection Methods for Intrusion Detection[J]. Department of Computer & Information Science, 2010, 6258(19) :242-255.
[11] 边肇祺 . 模式识别 [M]. 北京：清华大学出版社，2000.
[12]LUNT T T, TAMARU A, GILLHAM F. A Real-time Intrusion-detection Expert System (IDES)[EB/OL]. https://pdfs.semanticscholar.org/82e1/95f39a355fdb5dd2436c145cd1f2c5ee27e9.pdf，2016-6-1.
[13]VALDES A, SKINNER K. Adaptive, Model-based Monitoring for Cyber Attack Detection[EB/OL]. http://www.csl.sri.com/papers/adaptbn/adaptbn.pdf, 2016-6-1.
[14]TENG H S, CHEN K, LU S C. Adaptive Real-time Anomaly Detection Using Inductively Generated Sequential Patterns[EB/OL]. http://www.cs.unc.edu/~jeffay/courses/nidsS05/ai/Teng-AdaptiveRTAnomaly-SnP90.pdf, 2016-6-1.
[15]KUMAR S. Classification and Detection of Computer Intrusions[D]. USA：Purdue University, 1995.
[16] 黄学宇，魏娜，陶建锋 . 基于人工免疫聚类的异常检测算法 [J]. 计算机工程，2010，36（1）：166-169.
[17]QIN T, GUAN X, LONG Y, et al. Users' Behavior Character Analysis and Classification Approaches in Enterprise Networks[C]// IEEE. Eighth IEEE/ACIS International Conference on Computer and Information Science, June 1-3, 2009, Shanghai, China. New York: IEEE, 2009: 323-328.
[18] 董富强 . 网络用户行为分析研究及其应用 [D]. 西安：西安电子科技大学，2005.
[19] 杨铮 . 基于流量识别的网络用户行为分析 [D]. 重庆：重庆大学，2009.
[20]GU L, LI H. Memory or Time: Performance Evaluation for Iterative Operation on Hadoop and Spark[C]//IEEE. High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC), November 13 - 15, 2013, Zhangjiajie, China. New York: IEEE, 2013: 721 - 727.
[21] 吴夙慧，成颖，郑彦宁，等 . K-means 算法研究综述 [J]. 现代图书情报技术，2011，27（5）：28-35.
[22] 穆筝，吴进，许书娟 . 高速网络下 P2P 流量识别研究 [J]. 信息网络安全，2015（5）：69-76.
[23] 沈昌祥，张焕国，冯登，等 . 信息安全综述 [J]. 中国科学：E 辑，2007（2）：129-150.
[24] 肖梅，辛阳 . 基于朴素贝叶斯算法的 VoIP 流量识别技术研究 [J]. 信息网络安全，2015（10）：74-79.
[25]TIBSHIRANI R, WALTHER G, HASTIE T. Estimating the Number of Clusters in a Data Set via the Gap Statistic[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2001, 63(2): 411-423.
[26]ARTHUR D, VASSILVITSKII S. k-means++: The Advantages of Careful Seeding[C]//ACM. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, January 7-9, 2007, New Orleans, Louisiana, USA. Philadelphia: Society for Industrial and Applied Mathematics, 2007: 1027-1035.
[27] 翟东海，鱼江，高飞，等 . 最大距离法选取初始簇中心的 K-means 文本聚类算法的研究 [J]. 计算机应用研究，2014，31（3）：713-715.
[28] 雷小锋，谢昆青，林帆，等 . 一种基于 K-Means 局部最优性的高效聚类算法 [J]. 软件学报，2008，19（7）：1683-1692.