·人工智能及识别技术· 文章编号:1000-3428(2017)05-0204-06 文献标志码:A 中图分类号:TP309

基于词袋模型聚类的异常流量识别方法

马林进,万 良,马绍菊,杨 婷

(贵州大学 计算机科学与技术学院,贵阳 550025)

摘 要: 针对现有异常流量检测方法的识别准确率低且快速识别需要确定阈值等问题,基于词袋模型聚类,提出一种改进的网络异常流量识别方法。通过对已有的异常流量和正常流量进行 K-means 均值聚类,得到网络流量中的流量关键点,将网络流量转化映射到相应流量关键点后建立直方图,并采用半监督学习方式对异常流量进行检测。实验结果表明,与基于朴素贝叶斯、支持向量机等的识别方法相比,该方法具有更好的异常流量识别效果。

关键词: 词袋模型;机器学习;聚类;数据挖掘;异常流量识别

中文引用格式: 马林进,万 良,马绍菊,等. 基于词袋模型聚类的异常流量识别方法[J]. 计算机工程,2017,43(5):204-209.

英文引用格式: Ma Linjin,Wan Liang,Ma Shaoju,et al. Abnormal Traffic Identification Method Based on Bag of Words Model Clustering[J]. Computer Engineering,2017,43(5):204-209.

Abnormal Traffic Identification Method Based on Bag of Words Model Clustering

MA Linjin,WAN Liang,MA Shaoju,YANG Ting

(College of Computer Science and Technology,Guizhou University,Guiyang 550025,China)

【Abstract】 In view of the problem that the accuracy of abnormal traffic identification is low and fast identification is dependent on the threshold,an abnormal traffic identification method based on BoW(Bag of Words) model clustering is proposed. By means of K-means mean clustering for existing abnormal traffic and normal traffic,it finds the key points of network traffic. The original traffic is tranformed and mapped to the corresponding traffic critical points and then histogram is established. Abnormal traffic is detected by using semi-supervised learning. The experimental results show that this method has better recognition effect of abnormal traffic compared with identification method based on Naive Bayes(NB),Support Vector Machine(SVM) and others.

【Key words】 Bag of Words(BoW) model;machine learning;clustering;data mining;abnormal traffic identification

DOI: 10.3969/j.issn.1000-3428.2017.05.033

0 概述

随着网络技术的快速发展,互联网中流量迅速增加,充斥各种协议的网络环境变得越来越复杂,而网络异常流量分析在网络管理中至关重要。在大量涌入的流量中,若能够准确、及时地检测网络流量的异常行为,减轻异常流量对网络及其承载业务的危害,保证网络有效运行,则对提高网络的可用性和可靠性具有非常重要的意义,同时也是学术界和工业界共同关注的前沿课题之一。由于网络攻击的复杂化、自动化和大规模化,依靠传统的人工响应方式已经远不能满足异常流量检测和分析需求,因此将机器学习、数据挖掘等方法应用到大容量的网络流量数据的检测识别中,已成为研究的热点之一。为此,本文基于机器学习方法,提出一种基于流量关键点(Stream Point,SP)的词袋聚类异常流量识别方法。对归一化后的网络流量进行特征提取,将提取的特征向量进行 K-means 均值聚类计算聚类中心,聚类结果为流量关键点。使训练集特征向量映射到最近流量关键点上计算关键点直方图,经均衡处理得到词袋关键点训练模型。将测试集映射到流量关键点中获得测试集关键点直方图,通过分类器识别异常流量。

基金项目: 贵州省科学技术基金(贵黔合 LH 字[2014]7634 号,黔科合 J 字[2012]2328 号)。

作者简介: 马林进(1991—),男,硕士研究生,主研方向为网络安全监测;万 良,教授;马绍菊、杨 婷,硕士研究生。

收稿日期: 2016-10-26 **修回日期:** 2016-12-06 **E-mail:** majin78@126.com

1　异常流量检测

传统的网络流量特征如五元组(协议,源 IP 地址,目的 IP 地址,目的端口以及目的 IP 地址)、分组数、字节数、流计数、熵值、审计记录数据等均是常用统计量。文献[1]利用信息熵值的稳定性为依据,以IP、端口和活跃 IP 数量为维度的空间信息结构建模,将流量异常检测转化为基于 SVM 的分类决策问题。文献[2]使用相对领域信息熵重新定义离群度,提出一种基于直推式网络的异常检测算法 TCM-RNE,能有效降低网络噪声数据对检测的影响。文献[3]用小波融合和 Hilbert-Huang 变换提出了自相似参数估计算法。文献[4]设计一种异常流量动态自适应检测方法,采用小波分析估计 Hurst 参数,根据网络自相似程度自适应地调整检测阈值。文献[5]引入有监督的朴素贝叶斯(Naive Bayes,NB)机器学习方法进行流量分类与应用识别,使用核密度估计对朴素贝叶斯方法进行改进。文献[6]提出基于 C4.5 决策树分类器的有监督络流量分类方法,讨论特征选择和 Boosting 增强方法 2 种改进策略。实验结果表明,C4.5 分类器的训练复杂度适中、准确率高且分类速度快。文献[7]利用加权自相似参数,提出一种实时基于经验模态分解的异常流量检测方法,降低计算成本。文献[8]改进之前数据挖掘只追踪源目的地(OD)流,提出 Defeat 方法,通过随机聚合的 IP 流(草图)来准确追踪异常 IP 流。文献[9]通过比较交叉链路的流量来反映流量在多个层次的联系,使用非参数估计和多重假设检验方法对这种联系进行分析,能够有效识别异常行为。文献[10]考虑流量矩阵的时空相关性,利用小波变换和 PCA 提出了一种在线的 MSPCA 全网络异常检测方法。文献[11]基于 BP 神经网络模型和线性神经网络模型,分别提出了2 种无线传感器网络异常数据检测方法,结合网络中的流量特征信息设计不同的自动学习算法,根据网络的流量情况自动构建检测模型,分析系统网络的异常行为。文献[12]把流量数据投影到多维 Hash 直方图上选用无监督 SVDD 构建检测向量,提出基于 TRW(Threshold Random Walk)的多窗口关联检测与样本优化选择算法,提高检测精度。文献[13]通过快速分布式特征提取框架从原始网络中提取显著特征,提出基于异常点的多步骤流量异常检测方法。

基于统计的异常检测方法通常很难确定阈值,为了增加检出率需要降低阈值,容易造成大量误报。统计方法通常需要精确的统计分布,不是所有的异常情况都可以通过完全的统计方法表示,多数基于统计的异常检测技术需要假定该网络流量是一个似稳态的过程,因此不能应用于所有异常流量检测。基于数据挖掘的异常流量算法利用计算机进行数据挖掘操作,期间需要用到大量数据,其产生的中间结果和其中的偏差通常不可干预,其建立模型过于复杂,不适用于小型机房异常流量检测。基于机器学习的异常检测算法通过以往经验在已有的训练集上建立模型,通常针对特定的一组任务或模型,能依据先前结果提高性能和检测率,具有较好的泛化识别能力和多种可选择的算法。

2　词袋模型

词袋(Bag of Words,BoW)[14]模型最早出现在自然语言处理领域,是信息检索和文本分类中非常重要的模型,在基于图像信息的图像分类中 BoW 也被称为特征词(Bag of Feature,BoF)[15]。图像中局部区域特征可以被看作构成图像的词汇[16],其中相近的区域特征可以视为一个词,这样就能把文本分类的方法应用到图像分类及图像检索[17]中。随着互联网的发展,各种网络交织形成复杂网络,其产生的数据流量非常大,尤其当发生大规模流量攻击时,人们无法通过频繁更新规则库和签名来应对,词袋模型通过机器学习的方法能很好地解决该问题。

2.1　K 均值算法

K 均值算法以欧式距离作为相似度测度,需要提前制定 K 值进行聚类,是半监督聚类算法。K-means 算法能迅速将输入数据集划分为 k 个集群,每个集群都由一个质心表示。其主要思想是:划分 N 个输入数据 $x_1,x_2,\cdots,x_N$ 到 k 个不相交的子集 C_i $(i=1,2,\cdots,k)$ 中,对于每个 n_i 数据集,$0<C_i<N$,能使式(1)的均方差最小:

$$J_{\mathrm{MSE}}=\sum_{i=1}^{k}\sum_{x_t\in C_i}\|x_t-c_i\|^2 \tag{1}$$

其中,x_t 表示子集 C_i 中第 t 个数据;c_i 是 C_i 的几何中心,K-means 算法旨在使得目标函数得到最小值 $\min J_{\mathrm{MSE}}$;$\|x_t-c_i\|^2$ 表示 x_t 与 c_i 之间的距离度量。

当数据集成员使得函数 $I(x_t,i)=1$ 时,K-means 算法划分输入数据 x_t 到第 i 个集群中。

$$I(x_t,i)=\begin{cases}1, i=\mathrm{argmin}(\|x_t-c_j\|^2)\\ 0,\text{其他}\end{cases} \tag{2}$$

其中,$j=1,2,\cdots,k$;$c_1,c_2,\cdots,c_k$ 是聚类中心,通过以下步骤实现收敛:

步骤 1　初始化 k 个随机指定的聚类中心 $c_1,c_2,\cdots,c_k$,对于每个输入数据 x_t 和 k 个集群,重复步骤 2、步骤 3 直到收敛。

步骤 2　通过式(2)计算 $I(x_t,i)$,决定第 k 个集群最靠近哪个中心。

步骤 3　对于 k 个聚类中心,令 c_i 为集群 C_i 的质量中心。

对于给定一组观察值的序列 $x_1,x_2,\cdots,x_N$，每一个观察值都是一个 d 维的实值向量。K 均值聚类的目标是划分这 N 个训练数据到 k 个聚类中心，最后得到 k 个 d 维的训练聚类中心。

2.2 直方图表示

直方图模型可用于 K-means 聚类结果表示，对于 k 个 d 维的训练聚类中心用词频直方图进行表示，如式(3)所示。

$$h(i)=\sum_{i=1}^{C_i}I(x_t,i),i=1,2,\cdots,k \tag{3}$$

通常最终结果可以用频率直方图进行表示：

$$\bar{h}(i)=\frac{h(i)}{\sum_{i=1}^{C_i}h(i)},i=1,2,\cdots,k \tag{4}$$

其中，$\bar{h}(i)$可以用来表示原始网络流量特征提取后的网络流量特征。

3 SP-BoW 算法

单个网络数据包本身具有的结构可以被视为文档对象，网络数据包中自带的特征可以被看作构成某一种异常流量的特有特征，由此对不同网络包进行训练建模，可以得到词袋模型的流量代码本，称作流量关键点(SP)，由此可以将词袋模型应用到异常流量检测中。本文提出了一种流量关键点词袋模型(Stream Point Bag of Word，SP-BoW)算法，通过训练词袋模型建立流量关键点，将直方图代换原始流量样本，对于训练样本使用流量关键点模型进行表示。

3.1 SP-BoW 训练过程

SP-BoW 训练过程具体如下：

1)对于训练组数据流量，首先提取特征向量，对于训练集，K 个类别有 M 个数据，共 $K\times M$ 个训练集数据，为消除不同的特征数据度量对聚类造成影响，对网络流量进行如下变换：

$$x_i'=\frac{x_i-\bar{x}}{S},i=1,2,\cdots,N \tag{5}$$

其中，x_i'是变换后的数据，数值在 0～1 之间；$\bar{x}$ 是特征数据的数值平均值，$\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i$；$S$ 是样本的特征标准差，$S=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2}$；$N=K\times M$ 是数据样本数量。

2)给定 K 值，使用半监督的 K-means 聚类算法对训练组进行聚类，得到向量聚类中心。对于给定一组观察值的序列$(x_1,x_2,\cdots,x_N)^{\mathrm{T}}$，每一个观察值都是一个 d 维的实值向量。训练结果 K-means 流量的兴趣点称为流量关键点。对于训练集，K 个类别有 M 个数据，共 $K\times M$ 个训练集数据，指定 4 个聚类中心，从而得到 4 个向量聚类中心 $\boldsymbol{w}_1,\boldsymbol{w}_2,\boldsymbol{w}_3,\boldsymbol{w}_4$，即有 4 个流量关键点。

3)对于每一个训练集数据，重新计算其与各个流量关键点 SP 的距离，得到与之最接近的 SP。因此，可以将该训练集映射到第 k 个 SP 上，如式(6)所示。

$$T_{i,k}=\min_{i,k}\left\{\sqrt{\sum_{j=1}^{KK}(x_i-S_j)^2}\right\},k\in KK \tag{6}$$

其中，$i=1,2,\cdots,N$；KK 为指定的聚类中心个数。

3.2 SP-BoW 测试样本处理过程

对于测试样本，同样经过式(5)的数据变化，提取得到特征向量，将特征向量通过式(6)映射到训练得到的流量关键点中。

3.3 SP-BoW 识别过程

对于已经标记的 K 个类别，每个类别有 M 个数据，共 $K\times M$ 个数据已经映射到流量关键点的训练集数据。给定一个聚类数目 $C=200$，使得 $M \bmod C=0$，可以将每一类别分为 M/C 个分片，训练集可以表示为 $C=\{C_1,C_2,\cdots,C_{K\times M/C}\}$，同一类别的训练集划分到 M/C 个分片内。对每一个分片，统计分片内的流量关键点直方图特征，具体如下：

$$H(k)=\sum C_{i+1},C_{i+2},\cdots,C_{i+M/C} \tag{7}$$

其中，$i=0,\frac{M}{C},2\times\frac{M}{C},\cdots,(K-1)\times\frac{M}{C}$。

对于训练集的每一个类别分片，都能得到相应的流量关键点直方图。对于训练集得到的不同类别的词典直方图，还需要进行均衡处理，如式(8)所示，避免训练集样本个数对最终识别结果造成的影响。

$$x_i''=x_i/\sum_{i=1}^{n}x_i,i=1,2,\cdots,n \tag{8}$$

训练识别流程如图 1 所示。

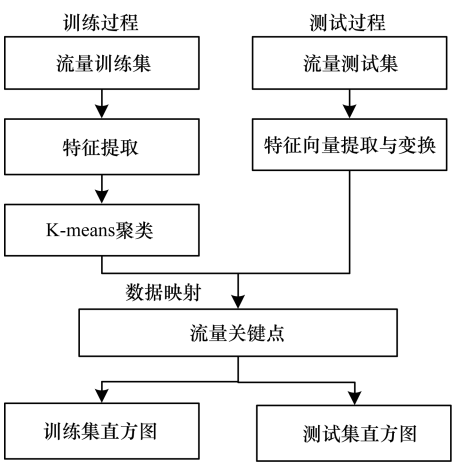

图 1 训练识别流程

采用欧式距离(式(9))对归一化后的训练集直方图与测试集直方图进行计算，将得到的分类结果与类

别标签进行比较,计算最终测试集样本的识别率。

$$Distance(C_i, C_j) = \sqrt{\sum_{k=1}^{n}(C_i - C_j)^2} \quad (9)$$

4　实验与结果分析

本文实验使用美国麻省理工学院林肯实验室 DARPA 数据集 KDD-99 进行实验测试。该数据集由 MIT LL 采集、哥伦比亚大学 IDS 实验室整理形成的数据集 KDD CUP99,是公认的异常流量检测数据集。KDD 每行数据共有 42 项(最后一项是标签),可分为 4 类特征:基本特征,内容特征,基于时间的流量统计特征,基于主机的统计特征。

本机测试环境为 CPU Intel I5,主频 3.2 GB,内存 4 GB,操作系统是 Win7 64 位。对于现有的机器识别方法,采用 Weka 实验平台进行测试。对于本文提出方法,通过 C ++ 编程进行测试。从数据集中选择 11 个类别,每个类别包含 2 000 条数据。选取 1 000 条数据样本作为训练样本,其余 1 000 条数据样本作为测试样本,共计 22 000 条数据。各类别数据情况如表 1 所示。

表 1　各标签类别数据

名称	类别	训练集	测试集	总计
D1	ipsweep	1 000	1 000	2 000
D2	nmap	1 000	1 000	2 000
D3	portsweep	1 000	1 000	2 000
D4	satan	1 000	1 000	2 000
D5	back	1 000	1 000	2 000
D6	neptune	1 000	1 000	2 000
D7	smurf	1 000	1 000	2 000
D8	guess_passwd	1 000	1 000	2 000
D9	snmpgetattack	1 000	1 000	2 000
D10	snmpguess	1 000	1 000	2 000
D11	normal	1 000	1 000	2 000

首先需要对网络流量进行归一化处理,如式(5)所示,然后将流量数据分为训练集和测试集 2 类,对训练集进行特征提取与变换,得到的特征向量进行 K-means 聚类,计算流量关键点。将训练集的每个标记标签的类别进行分片,对于每个分片,计算相应的流量关键点直方图,过程如图 2 所示。对测试集进行特征提取变换,与训练集中的流量关键点进行计算,得到最近关键点进行映射,将映射得到的关键点按分片大小统计关键点直方图,用式(9)进行识别。总的分类过程如图 2 所示。

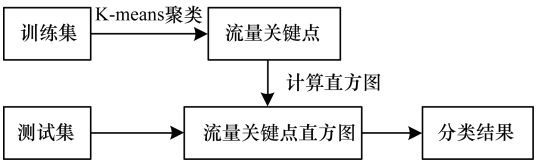

图 2　流量关键点词袋模型分类过程

为对比词袋聚类效果,使用 SVM、C4.5 决策树、NB 分类器、OneR 的简单 1-R 分类法、LMT、Logistic 回归模型、KNN 最近邻(最近邻参数为 5)、LogitBoost、Jrip 规则学习方法进行比较,采用正样本(True Positive,TP)和负样本(False Positive,FP)数作为评价指标。样本识别率结果通过式(10)表示:

$$P_{\text{Presicion}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \times 100\% \quad (10)$$

实验结果如表 2、表 3 所示,将聚类中心设置为 600,分片大小设置为 50,平均识别率为 90.9% 。对于数据 D3(数据类别为 portsweep),本文方法识别率偏低;对于 D10 识别率为 95% ,本文方法识别率稍低于 KNN,C4.5,LogitBoost,NB 等识别方法;对于其余类别,本文方法识别率均达到 100% 。从平均值来看,本文方法平均识别率为 90.9% ,KNN 为 88.5% 、SVM 为 82.1% 、NB 为 81.1% 、C4.5 仅为 54.6% ,可见本文方法的识别效果要优于对比方法。全部测试数据如表 2 所示。

表 2　不同方法识别率比较

数据集	KNN	SVM	Logistic	NB	LogitBoost	C4.5	Jrip	OneR	LMT	本文方法
D1	0.982	1.000	0.983	1.000	0.050	0.000	0.000	0.000	0.000	1.000
D2	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.000	0.040	1.000
D3	0.993	0.997	0.985	0.924	0.999	0.991	0.999	0.000	0.998	0.050
D4	1.000	1.000	1.000	1.000	1.000	0.995	0.995	0.373	0.995	1.000
D5	0.996	0.995	0.001	0.844	0.999	1.000	1.000	0.016	0.000	1.000
D6	1.000	1.000	1.000	0.992	1.000	1.000	1.000	0.986	1.000	1.000
D7	1.000	1.000	1.000	1.000	1.000	0.000	0.000	0.914	1.000	1.000
D8	0.995	1.000	0.989	0.855	0.999	0.000	1.000	0.000	0.000	1.000
D9	0.738	0.007	0.995	0.272	0.995	0.995	0.004	0.727	0.000	1.000
D10	0.995	0.997	0.996	0.997	0.984	0.984	0.000	0.693	0.000	0.950
D11	0.999	0.999	0.980	0.993	0.400	0.001	0.000	0.438	0.001	1.000
平均值	0.885	0.821	0.815	0.811	0.770	0.546	0.458	0.377	0.367	0.909

不同识别方法消耗时间如图 3 所示,表 3 为具体数据。图 3 中 OneR 方法训练与识别总耗费时间为 0.3 s,但其识别率平均为 37.7% ,KDD 训练与识别累计时间为 24.89 s,准确率为 88.5% ,本文方法训练所用时间为 0.39 s,测试时间为 6.85 s,仅次于 SVM 的 5.83 s,但识别准确率高达90.9% ,比 SVM 高 8.8% 。识别所耗时间还取决于算法优化,NB,OneR,SVM 等算法优化程度较好,运行所耗费时间短,而本文算法复杂度还可以进一步优化。

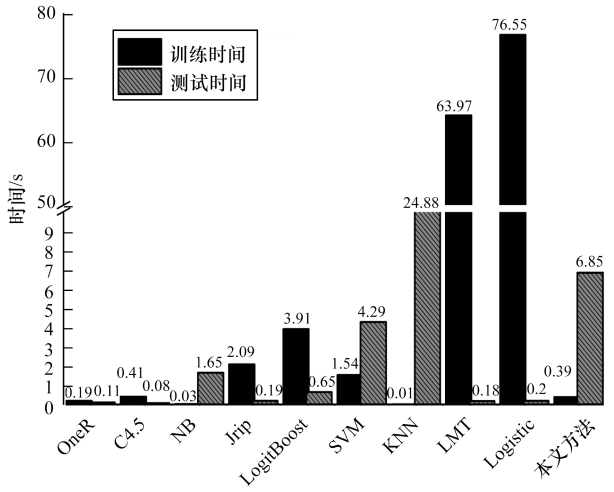

图 3 不同识别方法耗费时间对比

表 3 不同识别方法耗费时间的具体数据 s

识别方法	训练时间	测试时间	总时间
OneR	0.19	0.11	0.30
C4.5	0.41	0.08	0.49
NB	0.03	1.65	1.68
Jrip	2.09	0.19	2.28
LogitBoost	3.91	0.65	4.56
SVM	1.54	4.29	5.83
KNN	0.01	24.88	24.89
LMT	63.97	0.18	64.15
Logistic	76.55	0.20	76.75
本文方法	0.39	6.85	7.24

为进一步测试词袋聚类算法的识别性能,本文在不同聚类中心、分片大小下做了进一步实验测试,如图 4 所示。当聚类个数在 100 ~900 之间时,随着聚类个数的增加,识别率不断上升。当聚类个数为600,900 时,识别率较高,分别为90.9% ,91.8% 。随着聚类个数的增加,样本识别率迅速上并并趋于收敛。

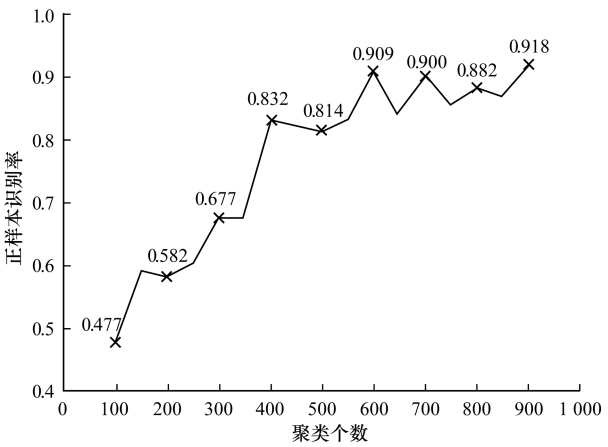

图 4 不同聚类个数下本文方法的正样本识别率 1

图 5 为聚类个数小于 10 000 时测试样本的正样本识别率,前期随着聚类中心的增加,识别率增加并收敛;后期(测试样本大于 2 000)随着聚类个数的不断增加,聚类中心开始影响最终识别率,总体呈现周期性的上下波动的态势。当聚类中心大于9 000 时,识别率又重新达到最大的 90.9% 。通过曲线拟合得到曲线 1:$y = a \times x \wedge b$,$a = 0.613\ 4$,$b = 0.038\ 74$,直线 2:$y = a + b \wedge x$,$a = 0.805$,$b = 7.951\mathrm{E}-6$。总体来说,聚类中心的个数增加对最终识别率有正面效果。

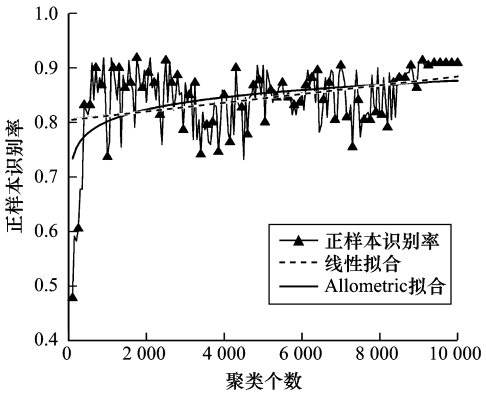

图 5 不同聚类个数下本文方法的正样本识别率 2

5 结束语

本文提出基于词袋聚类模型的网络异常流量识别方法,利用词袋模型进行聚类训练,依据关键点直方图进行异常流量识别。实验结果表明,该方法比现有的 SVM,Logistic,NB 等方法识别率更高,训练时间更短,为复杂网络环境下准确识别异常流量提供了一种可行的方法。下一步将通过调整聚类中心个数以提高算法计算效率。

参考文献

[1] 朱应武,杨家海,张金祥. 基于流量信息结构的异常检测[J]. 软件学报,2010,21(10):2573-2583.

[2] 李向军,张华薇,郑思维,等. 基于相对邻域熵的直推式网络异常检测算法[J]. 计算机工程,2015,41(8):132-139.

[3] Cheng Xiaorong, Kun Xie, Dong Wang. Network Traffic Anomaly Detection Based on Self-similarity Using HHT and Wavelet Transform[C]//Proceedings of the 5th International Conference on Information Assurance and Security. Washington D. C., USA: IEEE Press, 2009: 710-713.

[4] 夏正敏,陆松年,李建华,等. 基于自相似的异常流量自适应检测方法[J]. 计算机工程,2010,36(5):23-25.

[5] Moore A W, Zuev D. Internet Traffic Classification Using Bayesian Analysis Techniques[J]. ACM SIGMETRICS Performance Evaluation Review, 2005, 33(1):50-60.

[6] Wang Yu, Yu Shunzheng. Internet Traffic Classification Based on Decision Tree[J]. Journal of Chinese Computer Systems, 2009, 30(11):2150-2156.

[7] Han Jieying, Zhang J Z. Network Traffic Anomaly Detection Using Weighted Self-similarity Based on EMD[C]//Proceedings of 4th IEEE International Conference on Software Engineering and Service Science. Washington D. C., USA: IEEE Press, 2013: 23-25.

[8] Li Xin. Detection and Identification of Network Anomalies Using Sketch Subspaces[C]//Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement. New York, USA: ACM Press, 2006: 147-152.

[9] Chhabra P. Distributed Spatial Anomaly Detection[J]. Pattern Analysis & Applications, 2008, 14(4):329-348.

[10] 钱叶魁,陈　鸣,叶立新,等. 基于多尺度主成分分析的全网络异常检测方法[J]. 软件学报,2012,23(2):361-377.

[11] 胡　石,李光辉,卢文伟,等. 基于神经网络的无线传感器网络异常数据检测方法[J]. 计算机科学,2014,41(11):208-211.

[12] 郑黎明,邹　鹏,贾　焰,等. 网络流量异常检测中分类器的提取与训练方法研究[J]. 计算机学报,2012,35(4):719-730.

[13] Bhuyan M H, Bhattacharyya D K, Kalita J K. A Multi-step Outlier-based Anomaly Detection Approach to Network-wide Traffic[J]. Information Sciences, 2016, 348:243-271.

[14] Zhao Chunhui, Li Xiaocui, Cang Yan. Bisecting K-means Clustering Based Face Recognition Using Block-based Bag of Words Model[J]. International Journal for Light and Electron Optics, 2015, 126(19):1761-1766.

[15] Qiang Qiu, Cao Qixin, Adachi M. Filtering out Background Features from BoF Representation by Generating Fuzzy Signatures[C]//Proceedings of 2014 International Conference on Fuzzy Theory & Its Applications. Washington D. C., USA: IEEE Press, 2014:26-28.

[16] Ma Linjin, Wang Hangjun. A New Method for Wood Recognition Based on Blocked HLAC[C]//Proceedings of the 8th International Conference on Natural Computation. Washington D. C., USA: IEEE Press, 2012.

[17] Philbin J. Object Retrieval with Large Vocabularies and Fast Spatial Matching[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2007:1-8.

编辑　陆燕菲

(上接第 203 页)

[15] 汪锦岭,金蓓弘,李　京. 一种高效的 RDF 图模式匹配算法[J]. 计算机研究与发展, 2005, 42(10):1763-1770.

[16] 王　颖,刘　群,王慧强,等. 一种基于 RDF 图的本体匹配方法[J]. 计算机应用,2008,28(2):460-462.

[17] 吴江宁,刘巧凤. 基于最大公共子图文本相似度算法研究[J]. 情报学报,2010,29(5):785-791.

[18] 赖兴瑞. 基于最大公共子图的中文 Web 文本分类研究[D]. 厦门:厦门大学,2011.

[19] Bunke H, Shearer K. A Graph Distance Metric Based on the Maximal Common Subgraph[J]. Pattern Recognition Letters, 1998, 19(3-4):255-259.

[20] Conte D, Sansone C. A Comparison of Three Maximum Common Subgraph Algorithms on a Large Database of Labeled Graphs[C]//Proceedings of the 4th International Conference on Graph Based Representations in Pattern Recognition. Berlin, Germany: Springer, 2003:130-141.

[21] Xiao Yanghua, Dong Hua, Wu Wentao, et al. Structure-based Graph Distance Measures of High Degree of Precision[J]. Pattern Recognition, 2008, 41(12):3547-3561.

[22] Buke H. On a Relation Between Graph Edit Distance and Maximum Common Subgraph[J]. Pattern Recognition Letters, 1997, 18(9):689-694.

[23] Yan Xifeng, Zhu Feida, Han Jiawei. Feature-based Similarity Search in Graph Structures[J]. ACM Transactions on Database System, 2006, 31(4):1418-1453.

[24] Hochbaum D. Approximation Algorithms for NP-hard Problems[J]. ACM Sigact News, 1996, 28(2):40-52.

编辑　顾逸斐