doi :10.3969/j.issn.1671-1122.2012.01.008

主成分分析在网络流量异常检测中的应用研究

张新超，董建锋，张婧

（解放军 62301 部队，北京 1000713）

摘 要：文章阐述了在网络流量异常检测中应用主成分分析的应用范畴、主成分分析的常用算法等主成分分析的方法，针对网络流量数据源特点设计出了符合大规模网络安全态势分析实际应用的异常检测算法，在对原方法进行了适当调整并建立模型的同时，着重论述了主成分分析方法的检测评估过程，并对主成分分析方法在网络流量异常检测中的应用作了比较详细的描述。

关键词：网络流量；异常；检测；主成分分析

中图分类号：TP393.08 **文献标识码：**A **文章编号：**1671-1122（2012）01-0029-03

The Principal Component Analysis in Network Traffic Anomaly Detection

ZHANG Xin-chao, DONG Jian-feng, ZHANG Jing

(*The people's Liberation Army 62301 army, Beijing* 100071, *China*)

Abstract: The article elaborated in network traffic anomaly detection using principal component analysis applied category, principal component analysis algorithms such as principal component analysis method, according to the network flow data source is designed with large-scale network security situation analysis of practical application of anomaly detection algorithm, in the original method for the appropriate adjustments and the establishment of model at the same time, discussed the principal component analysis method the test and evaluation process, and the method of principal component analysis in network traffic anomaly detection application and gives a more detailed description.

Key words: network traffic; abnormal; testing; principal component analysis

0 引言

当今社会互联网在各个领域的作用越来越重要，但由于互联网的开放性以及应用系统的复杂性所引发的安全风险也随之增多。从 CNCERT/CC 接收和监测的各类网络安全事件情况可以看出[1]，网络信息系统存在的安全漏洞和隐患层出不穷，利益驱使下的地下黑客产业继续发展，网络攻击的种类和数量成倍增长，终端用户和互联网企业成为主要的受害者，基础网络和重要信息系统均面临着严峻的安全威胁。而随着黑客攻击手法越来越具有隐蔽性，使得对这些网络犯罪行为的取证、追查和打击都变得越发困难。在这个背景下，针对网络流量特征检测的研究逐步成为热点，其分析与建模对于网络拥塞控制和资源分配优化具有重要意义和根本性的指导作用。目前，国家进一步加大了基于大规模网络安全风险评估主动防御手段的建设投入，这其中如何实现宏观网络流量异常检测和安全事件的突发检测，也就是在大规模安全事件爆发时如何进行快速、有效的监测，从而为信息网络安全防御赢得时间，成为项目成败的重中之重。

本文提出的基于主成分分析方法的网络流量异常检测模型，可以对实时网络安全数据进行及时有效的分析，以便于在互联网爆发大规模信息安全事件的情况下能够得到第一手的信息，并反馈给网络应急响应部门，帮助其决策应对办法的方向和手段。

1 主成分分析方法概述

主成分分析也称主分量分析，旨在利用降维的思想，把多指标转化为少数几个综合指标。在统计学中，主成分分析（Principal Components Analysis，PCA）是一种简化数据集的技术[2]。作为一种线性变换，主成分分析法是一种降维的统计方法，通过保留

收稿时间：2011-12-15

作者简介：张新超（1979-），男，北京，工程师，硕士，主要研究方向：信息系统工程；董建锋（1976-），男，山西，工程师，硕士，主要研究方向：信息安全；张婧（1982-），女，黑龙江，工程师，硕士，主要研究方向：网络管理。

低阶主成分，忽略高阶主成分，来达到减少数据集的维数的目的。主成分分析借助于一个正交变换，可以将数据变换至新的坐标系中，即将任何数据投影的第一大方差变换至第一个坐标（即第一主成分）上，将第二大方差变换至第二个坐标（即第二主成分）上，保持原有数据集中对方差贡献最大的特性，依次类推。这样，低阶成分能够最大限度保留原数据最重要的方面。

1.1 主成分分析的应用范畴

总的来说，主成分分析主要应用在以下四个方面：

1）降低研究对象的数据空间维数。主成分分析时利用 m 维 Y 空间代替 p 维 X 空间损失的信息很少（其中 m < p，即 Y 空间较 X 空间相对低维）。即便当仅有一个主成分 Y1 时（即 m = 1），那么这个 Y1 依然是通过全部 p 个 X 变量转换得到的。同理，在计算 Y1 的均值时，也将应用到全部 X 的均值。

2）多维数据的图形化表示方式。我们知道，空间维数大于 3 时便不能构建出任何几何图形，而多元统计研究问题基本上都包含 3 个变量以上，此时是没有办法将其通过图形表示出来的。利用主成分分析，选取前两个或者某两个主成分，参照主成分的数值，得出 n 个采样数据的二维分布情况，从而可以通过图形表示直观地量化出各采样数据在主分量中的比重。

3）筛选回归变量。合理最优化的选择回归变量具有重要的实际意义，为使建立的模型更加便于进行结构分析、控制和预报，利用主成分分析法，可以达到用较少的计算量，从原始变量子集合中筛选出最佳变量，进而获得最佳变量集合的效果。

4）构建回归模型。主成分分析可以通过将主成分作为新的自变量替换原自变量 x 的方法来进行回归分析。

1.2 主成分分析常用算法

1）指标数据采集

p 维向量 x =（X1，X2，...，Xp）T 的 n 个样品

xi =（xi1，xi2，...，xip）T （i =1，2，…，n，n > p）

2）标准化采样数据

对矩阵阵元作如下变换，构建采样数据矩阵，得到标准化阵 Z：

$$Z_{ij}=\frac{x_{ij}-\bar{x}_j}{s_j}, i=1,2,\ldots,n; j=1,2,\ldots,p$$

其中：

$$x_j=\frac{\sum_{i=1}^{n}x_{ij}}{n}, s_j^2=\frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}{n-1}$$

标准化采样数据矩阵的源代码如表 1 所示：

3）对标准化阵 Z 求相关系数矩阵

$$R=[r_{ij}]_p xp=\frac{Z^TZ}{n-1}$$

表1 采样数据矩阵标准化java源代码

```
//bufferArray[][] 为采样数据
double junZhi[]=new double[bufferArray[0].length]；// 存储均值
double fangCha[]=new double[bufferArray[0].length]；// 存储方差
junZhi=getJunZhi（bufferArray）；// 取得每个采样数据的均值存放在一维数组中，junZhi[i] 中存放的是所有样本中第 i 个采样数据的均值
fangCha=getBiaoZhunCha（bufferArray，junZhi）；// 取得每个采样数据的标准差存放在一维数组中，fangCha[i] 中存放的是所有样本中第 i 个采样数据的标准差
for（int i = 0；i<bufferArray.length；i++）// 标准化原数据
{
for（int j=0；j<bufferArray[i].length；j++）
{
bufferArray[i][j]=（bufferArray[i][j]-junZhi[j]）/Math.sqrt（fangCha[j]）；
}
}
```

其中，$r_{ij}=\frac{\sum z_{kj}\cdot z_{kj}}{n-1}, i,j=1,2,\ldots,p$。

求相关系数矩阵的 java 源代码如下：

表2 计算相关系数矩阵的java源代码

```
//XGJZ[][] 记录相关矩阵
for（int i = 0；i<XGJZ.length；i++）// 遍历每一行
{
for（int j=0；j<XGJZ[i].length；j++）// 遍历每一个元素
{
// 对每一个元素求值
for（int k=0；k<bufferArray.length；k++）
{
XGJZ[i][j]+=
bufferArray[k][i]*bufferArray[k][j]/（bufferArray.length-1）；
}
}
}
```

4）确定主成分

通过解样相关系数矩阵 R 的特征方程$|R-\lambda I_p|=0$，可以得出 p 个特征根，利用$\frac{\sum_{j=1}^{m}\lambda_j}{\sum_{j=1}^{p}\lambda_j}\geq 0.85$计算出 m 值。然后分别对每个 λj（j=1，2，...，m）求解方程组 Rb = λjb，从而计算出特征向量b_j^0。为保证信息利用率达到 85% 以上，可通过雅可比过关法[3]计算特征根和特征向量矩阵，通过有限次迭代，达到所需计算精度。

5）主成分转换

通过以下算法，将标准化的数据矩阵转换为主成分：

$$U_{ij}=z_i^Tb_j^0, j=1,2,\ldots,m$$

U1 即第一主成分，U2 即第二主成分，…，Up 即第 p 主成分。

2 主成分分析方法应用于网络流量异常检测

2.1 构建数据矩阵

这里将骨干网核心路由器 NetFlow 数据作为流量异常检测的数据来源，采用主成分分析的方法可以有效检测出主动扫描、大规模僵尸网络爆发，以及 DDoS 攻击等引起网络流量异常的网络安全事件。

Netflow 由 Cisco 创造，提供网络流量的会话级视图，记录下每个 TCP/IP 事务的信息。流量（Traffic）在逻辑上等价

于呼叫或者连接。流（Flow）指的是两个节点之间数据包的单向序列（也就是说，对每一个连接会话都有两个流：从服务器到客户端的和从客户端到服务器的），用起止时间分隔开，可以用下面 7 个关键域进行标识：源 EP 地址、目的 EP 地址、源端口、目的端口、协议类型、服务类型、路由器输入接口等。无论什么时候路由器收到数据包，都要查找这 7 个域后再做决定；若该数据包属于已经存在的“流”，则相应流的流量值增加；反之，将创建一个新的流。与“流”相关的属性值（如源 / 目的地址、包数、字节数等）反映了在起止时间内发生的事件。一个流的起止时间是固定的，终止时间随着流的延续而增长。通常我们关心的流记录分类依据由一个五元组（即源地址、目的地址、源端口、目的端口和协议类型）所组成。

表3 流记录内容

源地址	目的地址	源端口	目的端口	协议类型	分组数	字节数	其他

Netflow 系统主要有三个组成部分：探测器、采集器、报告系统。探测器监听网络数据；采集器收集探测器传来的数据；报告系统用来从采集器收集到的数据中生成易读的报告 [4]。NetFlow 数据分为广州、上海、北京 3 个国际出入口数据，每个出入口分为 24 小时，每小时一个文件。其数据输出要求先在路由器和交换机上定制 NetFlow 流输出，并选择输出流的缓冲区大小、版本、个数等，配置相应 NetFlow 流量收集器（FlowCollector）的端口、IP 地址等信息。此时路由器或交换机可以用户数据报协议（UDP）的方式向外发送流信息，之后在 NetFlow FlowCollector 端配置接收端口号，设置汇聚、过滤策略和流量文件存放格式、目录等。

一般来说，NetFlow FlowCollector 都选用 Unix 工作站来收集数据，收集到的数据将存放在本地磁盘中（路径由用户定义）。同时，它也可以通过网关以 Socket 方式发送信息到其他网管分析软件，或直接读取存放在 NetFlow FlowCollector 工作站中的数据文件进行分析处理。本文将国际互联网出入口产生的 NetFlow 数据导入 Oracle 10g 数据库作为流量数据源。

针对以上数据源的特点，我们采用表 4 所示的 14 个内容作为采样矩阵的行向量，设定观测窗口为 6 小时，每 5 分钟产生一个行向量，即可得到一个 72*14 的数据矩阵。

表4 流量异常检测采样矩阵行向量

源IP数量	目的IP数量	TCP协议字节数	TCP协议报文数	UDP协议字节数	UDP协议报文数	其他流量字节数	其他流量报文数	WEB流量字节数	WEB流量报文数	前10个源IP在总字节中所占比例	前10个源IP在总报文数所占比例	前1个目的IP在总字节数所占比例	前10个目的IP在总报文数所占比例

2.2 建立检测模型

以采样数据的 72 个样本当作建模空间，通过滑动时间窗口建立模型，这些样本的采样数据构成数据矩阵 X，且矩阵的行向量分别由 14 个元素构成。

我们知道，主成分可划分为正常和异常两种，分别代表网络的正常流量和异常流量，两者区别主要体现于变化趋势上。正常主成分随时间变化的幅度较平缓，存在明显的周期性；异常主成分随时间变化的幅度较大，具有较强的突发性。以下是如何判断正常主成分的算法：

根据主成分及采样数据计算求得第一主成分变量，分析第一主成分变量 72 个数值的均值 μ1 和方差 σ1，找出其中与均值偏离最大的元素，观察偏离程度是否超过 3σ1。若最大偏离值超过阈值，那么第一主成分应当为正常主成分，其他主成分即为异常主成分，则主成分转换矩阵为：U=[L1]；若最大偏离值没有超过阈值，则需转入判断下一个主成分，直到时最后取得 U=[L1……Li–1]。如上所述，作为第一主成分应当表现出较强的周期性，而其它主成分的周期性应当逐渐减弱，同时其突发性应当逐渐增强，从而反映出网络的正常流量与异常流量之间的差别。计算转换矩阵 U 的 java 源代码为：

表5 计算转换矩阵的java源代码

```
for ( int i=0 ; i<14 ; i++) // 找出正常主成分变换矩阵
{
// 找出第 i 主成分变量所对应的 72 个数据放进 temp[j]
for ( int j=0 ; j<72 ; j++) {temp[j]=bufferArray[j][getByOrder ( i,
TZZ) ] ; }
// 如果第 i 主成分对应的数据中偏离均值最大的值不超过 3 倍标准差，
则再选一个主成分为正常，直至所有都正常
if ( getMaxPianLi ( getJunZhi ( bufferArray) [getByOrder ( i, TZZ) ],
temp)
<3*getBiaoZhunCha( bufferArray, getJunZhi( bufferArray))[getByOrder( i,
TZZ) ])
{
for ( int k=0 ; k<14 ; k++) {temp1[k]=TZXL[k][getByOrder ( i, TZZ) ] ;
}
if ( U.getNumRows ( ) ==1)
{
U=new Matrix ( temp1.length, 1) ;
U.setData ( temp1) ;
}
else
U.appendCol ( temp1) ;
break ;
}
}
U=U.multiply ( U.transpose ( )) ;
```

在求得转换矩阵 U 后，将每一个采样数据 Sk=（Xk1，Xk2……Xkp）的主成分分别投影至 P 维空间重建，重建后向量的计算公式为：

$$T_k = (S_k - \overline{X})^T (UU^T)$$

然后，计算每一个采样数据重建前后向量间的欧氏距离，也就是我们所说的残差：

$$d_k = \|S_k - T_k\|$$

同理，分别得出当前时刻前的 72 次采样数据的残差，再

下转第 35 页

拦截，即使拦截也不能通过验证，大大提高验证过程的安全性。（责编 张岩）

参考文献：

[1] Banday M. Tariq and N. A. Shah. A Study of CAPTCHAs for Securing Web Services[J].IJSDIA International Journal of Secure Digital Information Age, Vol. 1. No.2, December 2009.

[2] G Mori and J Malik. Recognizing objects in adversarial clutter: breaking a visual CAPTCHA[C].IEEE Conference on Computer Vision & Pattern Recognition (CVPR), 2003.

[3] Yan J and A Salah El Ahmad. Low-cost automated attacks on Yahoo CAPTCHAs[R]. TECHNICAL REPORT SERIES No. CS-TR-1127,November, 2008.

[4] Yan J and A S, El Ahmad. A Low-Cost Attack on a Microsoft CAPTCHA[C].Proc. 15th ACM Conf. Computer and Communications Security (CCS 08), ACM Press, 543-554,2008.

[5] 吉治钢 . 基于验证码破解的 HTTP 攻击原理与防范 [J]. 计算机工程，2006，20：1501-1504.

[6] K Chellapilla and K Larson, P Simard and M Czerwinski. Building Segmentation Based Humanfriendly Human Interaction Proofs[C] .2nd Int'l Workshop on Human Interaction Proofs, Springer-Verlag, LNCS 3517, 2005.

[7] Athanasopoulos E and S. Antonatos. Enhanced captchas: Using animation to tell humans and computers apart[C]. In Proceedings of the 10th IFIP Open Conference on Communications and Multimedia Security, October 2006.

[8] Egele M., L. Bilge, E. Kirda, and C. Kruegel, CAPTCHA Smuggling: Hijacking Web Browsing Sessions to Create CAPTCHA Farms[C]. In The 25th Symposium On Applied Computing (SAC), pages 1865-1870. ACM, March 2010.

[9] Mithal Anant Kartik , Sarah A. Douglas. Differences in movement microstructure of the mouse and the finger-controlled isometric joystick[C], Proceedings of the SIGCHI conference on Human factors in computing systems: common ground, p.300-307, April 13-18, 1996.

[10] Plamondon R, Feng C, Woch A. kinematic theory of rapid human movement[M]. Biological cybemetics，2003，89(2)：126-138.

[11] 桑应朋 . 基于计算机击键动力学的用户身份鉴别 [D]. 成都：西南交通大学信息科学与技术学院，2004.

上接第 31 页

计算出其标准差 σd 和均值 μd。至此，转换矩阵 U、残差标准差 σd 和残差均值 μd 就构成了网络流量模型，也就是进行网络流量异常检测的基础。

2.3 检测算法

建立网络流量模型后，用与之前类似的方法对新观测向量 N 进行中心化，即：

$$N_d = N - \overline{X}$$

然后，将中心化后的观测向量 N 投影至 P 维空间进行重建，即：

$$T_d = UU^T N_d^T$$

计算出残差值 d，即

$$d = \|N_d - T_d\|$$

最后，通过以下算法量化残差 d，即：

$$q(d) = \frac{d - \mu_d}{\sigma_d}$$

这里，若该观测值正常，重建前后的向量应该非常相似，则计算出的残差值 d 应该很小；若观测值反应出的流量与建立模型时差距很大，则计算出的残差值 d 应该会较大。

2.4 评估方法

参考观测向量的量化值 q（d），可以判断出当前网络流量是否正常。经验表明，若$|q(d)|$<5，则网络基本正常；若 5 ≤ $|q(d)|$<10，则网络轻度异常；若 10 ≤$|q(d)|$，则网络重度异常。根据数据源的特点和用户需求，在原有算法的基础上，可以把采样矩阵分为两种，一种是每分钟采样一次，每个矩阵维持 72 分钟数据，一种是把每个小时的数据进行汇总，作为一个行向量，数据矩阵维持 72 小时的数据。然后再对两个数据矩阵分别进行主成分分析。这种算法具有比较好的性能，这里采用通过随机数产生随机矩阵对本算法进行测试，如图 1 所示。在对随机产生的 n*14 测试矩阵进行主成分分析时，得出的运算时间随矩阵行数 n 的增加平稳增长，说明使用此算法可以检测更长时间的数据。

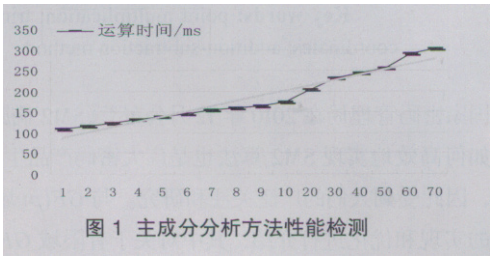

图 1 主成分分析方法性能检测

3 结束语

本文从应用范畴和常用算法两方面对网络流量主成分分析方法进行了探究，又对应用于网络流量异常检测的主成分分析方法进行了详细的阐述。从而比较全面地论述了网络流量主成分分析这一网络监控技术手段，对于网络安全管理人员在获取、分析网络流量信息，实施网络应急管理和保持网络业务连续性等方面有一定指导意义。下一步，我们还将对网络流量主成分的细化和分析方法展开进一步研究，以便获得更为有效的网络流量监控和网络安全管理方案。（责编 杨晨）

参考文献：

[1] 国家计算机网络应急技术处理协调中心 . CNCERT/CC 2007 年网络安全工作报告 [R]. 北京，2008.

[2] 何晓群 . 多元统计分析 [M]. 北京：中国人民大学出版社，2004.3-53.

[3] 周长发 . Java 数值计算算法编程 [M]. 北京：电子工业出版社，2007.5-96.

[4] Cisco System. NetFlow Services and Appilications[R]. White Paper(S)，2003.