doi:10.3969/j.issn.1000-436x.2014.05.003

基于 *K* 均值多重主成分分析的 App-DDoS 检测方法

杨宏宇，常媛
(中国民航大学 计算机科学与技术学院，天津 300300)

摘 要：针对应用层分布式拒绝服务攻击，利用 Web 日志的数据挖掘方法提出一种 *K* 均值多重主成分分析算法和基于该算法的 App-DDoS 检测方法。首先，通过分析正常用户和攻击者的访问行为区别，给出提取统计属性特征的方法；其次，根据主成分分析法的数据降维特性并利用最大距离划分法，提出一种 *K* 均值多重主成分分析算法，构建基于该算法的检测模型。最后，采用 CTI-DATA 数据集及模拟攻击获取的数据集，进行与模糊综合评判、隐半马尔科夫模型、D-S 证据理论 3 种检测方法的 App-DDoS 攻击检测对比实验，实验结果证明 KMPCAA 检测算法具有较好的检测性能。

关键词：应用层；网络攻击；主成分分析；均值聚类；日志
中图分类号：TP309,TP393.08 **文献标识码**：A **文章编号**：1000-436X(2014)05-0016-09

App-DDoS detection method based on *K*-means multiple principal component analysis

YANG Hong-yu, CHANG Yuan
(School of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China)

Abstract: Aiming at the application layer distributed deny of service(App-DDoS) attacks, a *K*-means multiple principal component analysis algorithm(KMPCAA) utilizing the Web log mining was proposed, then an App-DDoS detection method based on KMPCAA was presented. Firstly, a statistical properties feature extracting method was designed by analyzing the difference between normal users' and attackers' access behavior. Secondly, a *k*-means multiple principal component analysis algorithm was proposed by using the maximum distance classification method according to the data dimension reduction property of the principal component analysis, and then the testing model based on the algorithm was established. Finally, an App-DDoS attack detection experiment on the CTI-DATA dataset and the simulated attack dataset was conducted. In this experiment, the proposed method was compared with the fuzzy synthetical evaluation (FSE) algorithm, the hidden semi-Markov model (HsMM) detection algorithm and the dempster-shafer evidence theory (D-S) algorithm. Experimental results demonstrate that the KMPCAA detection algorithm has better detection performance.
Key words: application layer; network attack; principal component analysis; means clustering; log

1 引言

随着互联网技术的大范围普及和迅速发展，互联网领域的安全问题日益突出。在所有的网络入侵行为中，分布式拒绝服务攻击(DDoS, distributed denial of service)作为一种常见且有效的攻击方式，被认为是安全领域中最难解决的问题之一[1~3]。尽管传统 DDoS 攻击检测防御机制已较为完善，但应

收稿日期：2013-08-24；**修回日期**：2013-12-31
基金项目：国家科技重大专项基金资助项目(2012ZX03002002)；国家自然科学基金资助项目(60776807, 61179045)；国家高技术研究发展计划（“863”计划）重点基金资助项目(2006AA12A106)；天津市科技计划重点基金资助项目(09JCZDJC16800)；中国民航科技基金资助项目(MHRD201009, MHRD201205)
Foundation Items: The National Science and Technology Major Project (2012ZX03002002); The National Natural Science Foundation of China (60776807, 61179045); The National High Technology Research and Development Program of China (863 Program) (2006AA12A106); The Tianjin Key Project of Science and Technology Support Program (09JCZDJC16800); The Science & Technology Project of CAAC (MHRD201009, MHRD201205)

用层 DDoS(App-DDoS, application layer DDoS)攻击作为一种新型攻击方式，完全模拟正常用户的访问行为，因此很难在请求发出的时候被检测到。目前现有的检测算法针对性单一，并且检测率偏低，应对 App-DDoS 的检测效果不理想。

2　相关工作

目前，国内外对于 App-DDoS 攻击的检测方法研究主要基于 3 个方向：流量监测控制、验证机制、用户访问特征。

文献[4]采用计算可信度的方法，依据用户请求速率和请求负载 2 个特征，分析了正常用户的 HTTP 数据请求分布规律，并将其作为计算可信度的标准，从而对 DDoS 攻击进行检测防御。

文献[5]提出一种轻量级验证机制，该机制通过在客户端连接目标服务器时加入验证码，利用客户端计算，过滤恶意攻击，从而达到防御目的，但该方法会影响正常用户的使用心理，给用户带来一定压力。

文献[6]提出一种采用图灵测试和 Zipf 法则分别建立黑白名单，利用权重队列调度来限制访问速率的早期随机丢弃算法，达到检测防御的目的。

文献[7]通过分析用户浏览行为，提出一种采用基于 AR 模型和卡尔曼滤波的自回归模型，该模型以学习和预测正常用户访问的方式来判断和定位异常访问源，并通过前端路由器对其限流或过滤该访问源，从而达到防御目的。

文献[8]认为当流量拥堵阻塞现象出现的时候，大量的 IP 地址簇会重复出现，然而在 DDoS 攻击中会出现大量新的 IP 簇。一个合法的用户会在流量拥堵的时候降低访问速率，但是一个攻击者会持续增加访问速率。还有一个重要区别就是文件的关联度。当流量拥堵现象产生时，文件访问会遵循 Zipf 规则分布，而在 DDoS 攻击中，一些新的文件访问会很频繁。但上述规律只在简单情形下存在，对于复杂的 DDoS 攻击手段则无法检测。

文献[9]根据用户浏览特征描述，提出通过计算单位时间内请求页面的熵来检测攻击的算法，其主要依据正常用户与攻击者的请求模式不同，因此对页面产生的熵值也不同，依此来检测攻击的发生与否，该方法的不足在于对低速率攻击，熵值变化不大，检测效果不佳。

本文提出一种基于 K 均值多重主成分分析算法的检测方法。通过分析正常用户和攻击者的访问行为区别，给出了提取统计属性特征的方法，对改进 K 均值聚类算法的初始中心点选取方式进行了优化。在此基础上，结合主成分分析法的数据降维特性，提出一种 K 均值多重主成分分析算法，并构建了基于该算法的检测模型。

3　K 均值多重主成分分析算法

3.1　属性特征提取

为了能够最大程度表征用户浏览行为，本文分析了攻击行为特征与 Web 日志记录属性的关联，并依据用户浏览行为的属性特征和 Web 日志记录，构建了数据统计特征向量。

假定有 M 个用户访问一个 Web 服务器，产生了 N 个 session 会话，服务器总共有 D 个 Web 页面。设 s_i 为这个会话的页面请求序列，定义变量 L_i 和 $\overline{L}$

$$L_i=\sum_{d-1}^{D}\eta_{d,i}\ ,\quad \overline{L}=\sum_{i-1}^{N}L\ /N \tag{1}$$

其中，L_i 表示会话 i 的请求序列中总共的请求次数，$\overline{L}$ 表示 N 个会话 session 的平均请求次数。

定义属性 1　为了观察用户网页浏览量与平均量的关系，设变量 τ_i 为会话 i 总请求次数与每个会话的平均请求次数的比值，计算式为

$$\tau_i=L_i/\overline{L} \tag{2}$$

定义属性 2　为了了解用户的兴趣广泛程度，定义变量 α_i，表示会话 i 所访问的总页面数占服务器总页面数的百分比，计算式为

$$\alpha_i=\sum_{d=1}^{D}b_{d,\ i}/D\begin{cases}b_{d,\ i}=1, & \eta_{d,\ i}>0\\ b_{d,\ i}=0, & \eta_{d,\ i}\leqslant 0\end{cases} \tag{3}$$

定义属性 3　为了了解用户对偏好页面的喜爱程度，设 q_i 表示会话 i 访问次数最多的页面，即该用户对此类页面主题兴趣较高，q_i 计算式为

$$q_i=\arg\max_d\{\eta_{d,\ i}\} \tag{4}$$

γ_i 定义为用户对 q_i 的访问次数与对所有页面的访问次数之比，计算式为

$$\gamma_i=\eta_{q_i,\ i}/L_i \tag{5}$$

定义属性 4~8　为了细粒度了解用户的行为特征，从页面角度加以考虑，对页面的属性特征进行分析，从中提取用户行为特征，构造适合会话 session 的属性特征。

定义变量 o_i 表示会话 i 访问的一级页面次数与

总访问次数 L_i 的比值；定义变量 t_i 表示会话 i 访问的二级页面次数与总访问次数 L_i 的比值；定义变量 th_i 表示会话 i 访问的三级及更深层级页面次数与总访问次数 L_i 的比值，这 3 个变量的计算式参考式(1)。

根据动态页面与静态页面的区别，定义变量 dy_i 表示会话 i 访问的动态页面的次数与总访问次数 L_i 的比值；定义变量 st_i 表示会话 i 访问静态页面的次数与总访问次数 L_i 的比值。由于页面只有动态和静态 2 种情况，因此 dy_i 与 st_i 2 个属性具有较强的相关性。

定义属性 9 用户误操作后，服务器通常会跳转到错误页面，而若是攻击者产生错误，访问错误页面的次数会增加，因此定义变量 e_i，表示会话 i 访问错误页面次数与总访问次数 L_i 之比，日志记录中状态码属性表示的是请求响应后的服务器状态。

定义属性 10 攻击者的访问时间与普通用户有很大区别，攻击者阅读时间短，发送请求频率高，而普通用户阅读时间长，发送请求频率低。过高速率访问表示攻击者在频繁发送请求，过低速率访问则表示攻击者一直占用服务器连接，或者一直请求访问大文件页面。定义变量 sp_i 表示请求访问速率，用以区分攻击者与普通用户，用会话 i 总访问次数与整个会话的持续时间 T_i 的比值来表示

$$sp_i = L_i / T_i \tag{6}$$

综上所述，每一条 session 会话记录都可以用式(7)的 10 个维度属性向量表示

$$w_i = [\alpha_i, \tau_i, \gamma_i, o_i, t_i, th_i, dy_i, st_i, e_i, sp_i] \tag{7}$$

Web 日志中的所有数据记录，可以表示为会话属性特征矩阵 $\boldsymbol{W}=[w_1 \cdots w_n]^{\mathrm{T}}$，行表示 session 会话记录，列表示属性特征。

3.2 主成分分析法

主成分分析法[10]是一种可以降低数据维度的数学变换方法。计算思路是先将给定的一组相关变量经过线性变换转成另一组不相关的变量，然后按照新变量方差依次递减的顺序排列，从而选取出与研究问题相关性较大的前几项影响指标，作为问题的主成分。其主要的计算过程如下。

1) 设研究对象 $\boldsymbol{x}$ 的维度为 p，其 p 维向量为 $\boldsymbol{x}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$，将 n 个 $\boldsymbol{x}$ 样本 $\boldsymbol{x}=(X_{i1},X_{i2},\cdots,x_{ip})^{\mathrm{T}}$，$i=1,2,\cdots,n$，$n>p$，构造样本阵，并将矩阵元素进行标准化变换

$$Z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j},\ i = 1,2,\cdots,n; j = 1,2,\cdots,p \tag{8}$$

其中，

$$\overline{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n},\ s_j^{\,2} = \frac{\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2}{n-1} \tag{9}$$

得到标准化矩阵 $\boldsymbol{Z}$。

2) 对标准化矩阵 $\boldsymbol{Z}$ 求相关系数矩阵 $\boldsymbol{R}$

$$\boldsymbol{R} = [r_{ij}]_p = \frac{Z^{\mathrm{T}} Z}{n-1} \tag{10}$$

其中，

$$r_{ij} = \frac{\sum_{i=1}^{n} z_{ij} z_{ij}}{n-1},\ i,j = 1,2,\cdots,p \tag{11}$$

3) 利用奇异值分解法，求解样本相关系数矩阵 $\boldsymbol{R}$ 的特征方程：$|\boldsymbol{R}-\lambda I_p|=0$，得 p 个特征根，根据特征根的方差累计贡献度确定主成分个数 m，表示 m 个主成分使信息的利用率达 85%以上，对每个特征根 λ_j，$j=1,2,\cdots,m$，解方程组 $Rb=\lambda_j b$ 得单位特征向量 b_j。

$$\frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \geqslant 0.85 \tag{12}$$

4) 将标准化后的指标变量转换为主成分

$$U_{ij} = z_i^{\mathrm{T}} b_j, j = 1,2,\cdots,m \tag{13}$$

其中，U_1 称为第一主成分，U_2 称为第二主成分，……，U_p 称为第 p 主成分。

3.3 改进 K 均值聚类算法

传统 K 均值聚类算法存在两点主要缺陷，分别是对中心点选取敏感和受孤立点影响较大。因此，在借鉴已有研究成果的基础上[11]，本文对其初始中心值的选取方式进行了改进，依据极远邻二点 K 初始值选取算法(Farest2-K-means)[11]，提出最大距离划分法。

最大距离划分法的计算思路为：计算数据对象间的欧式距离后找出距离最大的数据对象 x_1、x_2，形成各自集合 X_1、X_2，并将其从集合 W 内删除，然后计算 X_1 类的数据对象 x_1 与 W 内各样本的距离，找出符合条件的数据对象并将其放入 X_1 内，同时从集合 W 中删除，X_2 做如 X_1 的相同处理，直至找不出符合条件的数据对象为止。接着再从集合 W 内找到最

大距离的数据对象 x_3、x_4 形成 X_3、X_4，重复上面的搜索过程，直至 W 集合内数据全部划分完。最后对各对象集合进行算术平均，形成 K 个初始中心点。

3.4 *K* 均值多重主成分分析算法

为了尽可能还原真实的网络，本文提出采用多重模式的主成分分析算法来分析 Web 日志记录，即利用第一重主成分分析降低聚类操作的复杂性，完成聚类后，再在每一个聚类内部，采用第二重主成分表示数据的主要特征，同时也代表了该聚类的特征。利用该特征，判断新会话数据的聚类归属程度，就能够识别攻击数据与正常数据。本文定义此算法为 K 均值多重主成分分析算法(KMPCAA, *K*-means multiple principal component analysis algorithm)

根据 3.1 节中对会话对象的特征分析，结合 Web 日志记录特点，本文首先采用主成分分析法对会话进行分析，用新的主成分表示数据。

依据 3.2 节中主成分分析算法的介绍，定义研究对象 w_i 为会话对象 i 的向量表示，w_i 表示如式(7)，定义均值向量 μ_0 和相关系数矩阵 $\boldsymbol{C}$ 为

$$\mu_0 = \left(\sum_{i=1}^{N} w_i\right)\Big/N\ ,\quad \boldsymbol{C} = XX^{\mathrm{T}}/N\ ,$$
$$\boldsymbol{X} = [x_i \cdots x_N],\ i = 1,\cdots,N \tag{14}$$

其中，$\boldsymbol{X}$ 表示由 W=[w_i,⋯,w_N]构建的标准化矩阵，通过奇异值分解相关系数矩阵 $\boldsymbol{C}$ 来计算特征值和特征向量。

设 u_j 为通过相关系数矩阵 $\boldsymbol{C}$ 的第 j 个特征根 λ_j 解出的相应特征向量，则所有特征向量可表示为

$$\tilde{U} = \left[u_1 \cdots u_p\right] \tag{15}$$

其中，P (<10)表示依据累计方差贡献率得到的主成分数量。将 w_i 通过 $\tilde{U}$ 表示为新的维度空间 a_i，如式(16)所示，其中，a_i=[U_1,⋯,U_p]是以 P 维主成分向量的形式表示 w_i

$$a_i = \tilde{U}^{\mathrm{T}} x_i,\ \ i = 1,\cdots,N \tag{16}$$

依据 3.3 节中提出的改进 K 均值聚类算法将给定的 N 个数据样本[w_i,⋯,w_N]划分为 k 个聚类。为了避免属性特征矩阵有很高的稀疏度，利用 a_i 值取代 w_i 值来进行计算，a_i 值是低维度非稀疏值。

聚类划分完成后，在每一个聚类 $S^{(k)}$ 上建立一个主成分分析模型。$w_i^{(k)}$ 表示会话 i 属于聚类 $S^{(k)}$ 集合，对于每一个聚类，首先将属性向量标准化

$$x_i^{(k)} = (w_i^{(k)} - \mu^{(k)})\big/(\sigma^{(k)})^2 \tag{17}$$

其中，$\mu^{(k)}$ 表示均值，$(\sigma^{(k)})^2$ 表示方差。

然后依据 3.2 节中公式计算聚类 $S^{(k)}$ 的 P 维主成分向量 $U^{(k)}$，并依据 $U^{(k)}$ 重新构建原始数据向量 $\hat{x}_i^{(k)}$，并计算误差值 ε_i

$$\hat{x}_i^{(k)} = (U^{(k)})(U^{(k)})^T x_i^{(k)} \tag{18}$$

$$\varepsilon_i = \left\| x_i^{(k)} - \hat{x}_i^{(k)} \right\|^2 \tag{19}$$

为构建误差值的统计偏差，定义聚类 $S^{(k)}$ 的阈值 $\delta^{(k)}$ 来判断给定的数据样本是否正常，$\delta^{(k)}$ 的计算式为

$$\delta^{(k)} = \mathrm{E}[\varepsilon] + \beta\sqrt{\mathrm{E}\left[\varepsilon - \mathrm{E}[\varepsilon]\right]^2} \tag{20}$$

其中，E[ε]和 E[ε−E[ε]]^2 分别表示聚类 $S^{(k)}$ 重构误差的均值 μ 和方差 σ^2，β 值是决定偏差范围的关键因素。根据检测偏差研究[18]，偏差范围应该偏离均值 2 到 3 个标准绝对误差。

当有新会话 t 请求访问服务器时，首先构建属性特征 w_t，然后计算适合请该求数据的聚类 π

$$\pi = \arg\min_k \left\{\tilde{U}^{\mathrm{T}}(w_t - \mu_0) - m_k\right\} \tag{21}$$

其中，k=1,⋯,K。K 表示聚类的总个数，m_k 表示聚类 $S^{(k)}$ 的中心点。在确定了最合适聚类 π 后，依式(17)对 w_t 进行标准化，然后根据式(19)计算重构误差值 ε_t，与式(20)计算出的阈值 $\delta^{(\pi)}$ 作比较，若 ε_t>$\delta^{(\pi)}$，判定当前用户为攻击者。

基于上述算法设计思路，本文设计了 KMPCCA 的处理流程（如图 1 所示）。该算法的流程如下。

step1　提取属性特征。对原始数据进行预处理，并依据式(1)~式(7)提取所需属性特征向量组成数据集合[w_i,⋯,w_N]。

step2　主成分分析。将数据集进行主成分分析。

1) 依据式(2)和式(3)计算标准化矩阵 $\boldsymbol{X}$。

2) 依据式(4)和式(5)计算均值向量 μ_0 和相关系数矩阵 $\boldsymbol{C}$。

3) 采用奇异值分解的形式，解样本相关系数矩阵 $\boldsymbol{C}$ 的特征方程|$\boldsymbol{C}$−λI_p| = 0，得到 p 个特征根及相应的特征向量 $\tilde{U} = \left[u_1 \cdots u_p\right]$。

4) 依据式(13)，将 w_i 通过 $\tilde{U}$ 表示为新的维度空间 a_i。

step3　K 均值聚类。将 N 个数据样本[w_i,⋯,w_N]替代为[a_i,⋯,a_N]进行聚类分析。

1) 初始化迭代次数 m，原始数据集 dataSet，中心数据集 center，加权平均平方距离和 J_l，依据

3.2 节计算聚类初始中心值并存入中心数据集中，并初始化 K 值。

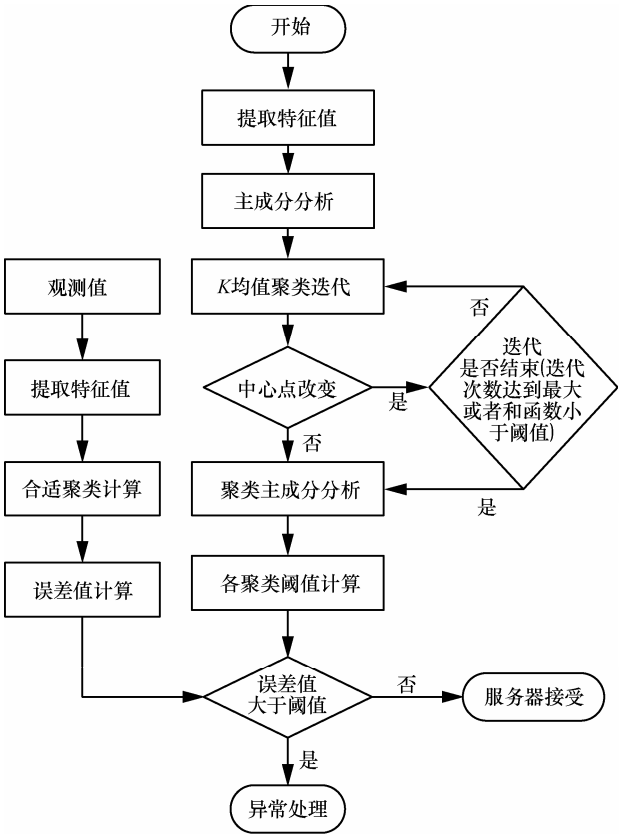

图 1　KMPCCA 流程

2) 迭代开始，计算中心距并进行归类。

3) 中心点是否改变，若未改变，收敛已完成，迭代结束，转至执行 step4；若中心点改变，判断 m 是否为 0，若为 0，转至执行 step4，若不为 0，继续执行 4)。

4) 重新计算中心距，并将原始数据重新归类，重复执行 3)。

step4　聚类主成分分析。对每个聚类 $S^{(k)}$建立主成分分析模型。

1) 依据式(14)将聚类 $S^{(k)}$内原始向量 $w_i^{(k)}$标准化。

2) 依据 3.1 节统计属性的定义计算聚类 $S^{(k)}$的 P 维主成分向量 $U^{(k)}$。

3) 依据式(18)和式(19)重新构建原始数据向量 $\hat{x}_i^{(k)}$，并计算误差值 ε_i。

4) 计算误差值 ε_i的期望与方差，并依据式(20)计算聚类 $S^{(k)}$的阈值 $\delta^{(K)}$。

step5　计算观测值。提取属性向量 w_t，依据式(21)计算合适聚类 π，依式(17)对 w_t进行标准化，并计算重构误差值 ε_t。

step6　判定。若 $\varepsilon_t>\delta(\pi)$，则判定为攻击者，进行异常处理；否则判定为正常会话请求。

3.5　基于 *K* 均值多重主成分分析的检测模型

根据公共入侵检测框架(CIDF, common intrusion detection framework)规范中所给出的入侵检测系统的通用模型，将 KMPCAA 算法引入 App-DDoS 攻击检测中，提出一个基于 K 均值多重主成分分析算法的 App-DDoS 攻击检测模型（如图 2 所示）。

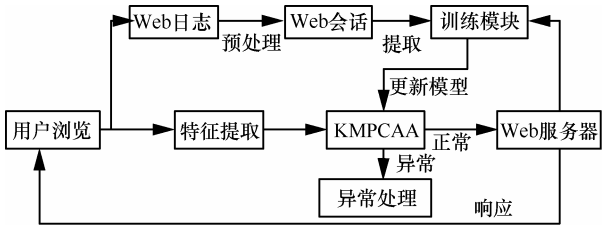

图 2　基于 KMPCCA 的检测流程

该检测模型共包含 5 个核心模块，分别为 Web 会话模块、KMPCAA 训练模块、特征提取模块、KMPCAA 检测模块和异常处理模块。

4　仿真实验与结果分析

本节对 KMPCAA 算法进行的仿真实验与算法检测验证。参考相关研究方法[8,9,12~15]，选择模糊综合评判（FSE, fuzzy synthetical evaluation）算法[14]、隐半马尔科夫模型(HsMM, hidden semi-Markov model)检测算法[15]、D-S 证据理论(Dempster-Shafer evidence theory) 检测算法[16]3 种主流检测方法，与 KMPCAA 算法进行 App-DDoS 攻击检测对比实验，以验证 KMPCAA 算法的检测性能。

4.1　实验设计

为验证 KMPCAA 算法对 App-DDoS 攻击的检测性能，将整个实验过程分 2 个部分 4 个步骤，第 1 部分是实验数据处理，第 2 部分是实验验证。实验设计思路如下。

1) 实验数据预处理

从 CTI-DATA 预处理数据集中，提取用户行为特征；将提取特征后的数据集划分为训练数据集 dataSet_train 和测试数据集 dataSet_normal_test。

2) 代码实现

编制 MATLAB 代码实现 KMPCAA 算法检测，源代码主要包括数据集预处理模块，主成分分析模块和 K 均值聚类分析模块 3 个模块。

3) 攻击实验

设置不同的攻击参数，采用 CC 攻击软件，模拟 App-DDoS 攻击过程，获得攻击数据集 dataSet_attack_test。

4) 对比实验

对 KMPCAA 和 FSE、HsMM、D-S 3 种主流检测方法进行应用层 DDoS 攻击的检测性能对比实验。在该实验中，首先设置上述 4 种检测算法的初始参数，然后进行检测实验。

4.2 实验数据与处理

实验采用 DePaul University 大学教育网站服务器的 Web 日志 CTI-DATA 数据集，其中包括学生频繁浏览访问的教育信息板块页面。它提供了预处理后的数据样本分别为 cti_cod.txt、cti_nav.txt、cti_std.txt、cti_tra.txt、cti_stats.txt 5 个文本，数据样本内容如表 1 所示。

表 1　　数据样本介绍

文件名称	内容
cti_cod.txt	所有 Web 页面
cti_nav.txt	会话预处理信息
cti_std.txt	会话—页面时长矩阵
cti_tra.txt	会话访问无重复序列
cti_stats.txt	原始数据记录的统计信息

数据集中的会话记录均为正常用户访问产生，不包含攻击数据。依据时间记录选取前 224 个小时的数据定义为 dataSet_train，作为训练数据集，用来对检测算法进行训练，将剩余的 112 个小时数据定义为 dataSet_normal_test，作为算法的正常检测数据集，如表 2 所示。

表 2　　实验数据集的组成

数据集名称	请求数目	会话数	页面数
dataSet_train	58 361	9 674	411
dataSet_normal_test	20 146	4 071	233

训练样本集中包括了 411 条 Web 页面 URL 及 9 674 条会话记录，正常测试样本数据集包括了 233 条 Web 页面 URL 及 4 071 条会话记录。

4.3 验证实验

将预处理后的数据集依据 3.1 节中属性定义 1~10，编制 M 文件 data_extract.m 对原始属性数据进行统计属性提取。该特征提取函数以预处理数据集为输入，通过输入 Web 日志预处理后的数据集合：统计量 data.txt、会话序列矩阵 basic.txt、一二三级页面序号向量 ott.txt、错误页面序号 error.txt 以及访问时间向量 time.txt，计算得出会话统计特征矩阵 data_result。

根据 3.4 节的算法设计，采用数学建模工具 MATLAB 编制了 KMPCAA 算法源代码。由于目前在 App-DDoS 攻击领域，还没有通用的攻击类数据集可以提供测试研究，在研究并分析了相关研究方法和文献[8,9,13~16]后，本文采用网页攻击工具 CC 攻击软件，模拟 CC 攻击过程，获取攻击数据。

本次实验过程选择的攻击目标是一台部署了基于 J2EE 架构的信息管理平台系统的服务器主机。经过 JVM 内存及 Web 服务器内存参数设置，将该系统作为被攻击对象，并监测 Web 日志数据记录情况。实验限定了总持续攻击时间为 12 min，共产生完整有效日志记录数 3 万多条，从中随机筛选出 5 000 条记录数作为攻击测试数据集 dataSet_attack _test，如表 3 所示。

表 3　　dataSet_attack_test 数据集的组成

数据集名称	请求数目	会话数	页面数
dataSet_attack_test	29 418	5 000	53

为了验证 KMPCAA 算法的检测性能，本文针对 FSE 方法、HsMM 方法和 D-S 方法 3 种检测算法进行了检测对比实验。

模糊综合评判法的观察特征主要为访问请求速率、CPU 负载情况 2 个特征。该方法不需要前期学习，参照文献[14]中所设计的各评价因素的隶属函数和权重因素以及最大隶属度原则，设定初始参数如表 4 所示。

表 4　　FSE 算法参数设定

重要参数	设定值
隶属度函数边界值$[a, c]$	0~7
权重矩阵一致性指标 CR	0.1
权重矩阵元素标度范围	1 ~ 9

HsMM 算法的观测特征主要为访问对象的序列和请求时间间隔 2 个特征。该模型算法与本文提出的 KMPCAA 算法一样，均需要前期学习，参照文献[15]中设定的模型构建方法，依据训练数据集 dataSet_train 构建 HsMM 模型，模型参数、或然概率和状态序列分别由前向—后向算法、极大或然概率估计和多观测序列模型参数估计方法计算得到。

初始参数设定如表 5 所示。

表 5 HsMM 算法参数设定

重要参数	设定值
序列长度阈值 $T0$	50
正常度范围	−3 ~ −7
阈值判断的左右阈值	−12.3 ~ −2.2

D-S 证据理论法的观测特征为页面浏览时间、访问网页先后顺序和请求页面的资源量 3 个特征。该算法不需要前期学习，参考文献[16]中的实验过程，分别计算主 mass 函数 m_1 和 m_2，其中 m_2 函数的构建依据 HsMM 模型，并使用前向后向算法确定模型参数。初始参数设置如表 6 所示。

表 6 D-S 算法参数设定

重要参数	设定值
平均请求时间间隔参数 x	0.9
置信度区间	0.068 9 ~ 0.147 8

依据上述 3 种算法的初始参数设置，对数据集 dataSet_attack_test 进行攻击行为检测，并与 KMPCAA 检测结果进行对比，如表 7 所示。

由表 7 可见，针对具有不同参数的相同类型攻击行为，4 种检测算法在 120 线程/1 秒时的检测率与 80 线程/4 秒时相比整体上升，同时针对 4 种不同攻击类型呈现出一致的检测规律。针对主页面类型的攻击，KMPCAA 检测率普遍高于其他算法，这是由于主页面攻击表现出的高频率、高兴趣度、低广泛度特征与正常用户行为偏差过大，因此检测率偏高；针对随机页面类型的攻击，HsMM 检测算法的检测率要高于其他 3 种算法，这是由于 HsMM 算法在构建检测模型时加入了序列特征属性，因此对于序列特征混乱的随机页面攻击能够轻易检测出；主流页面攻击的攻击频率不高，该攻击主要以动态页面为目标，且经常包含对大容量文件的请求，因此对于主要检测请求负载率的 FSE 与 D-S 算法，能够被较快地检测出；针对重复序列攻击，该类攻击是重复发送正常用户访问序列，对于检测用户跳转序列的 HsMM 算法，攻击特征过于明显，因此能够较快地被 HsMM 算法检测出。

由表 7 还可得出，针对具有相同参数的不同类型攻击行为，KMPCAA 算法针对 4 种类型攻击的检测率浮动范围很小，具有稳定的检测率；而 FSE 算法针对不同攻击的检测率浮动范围偏大，在攻击频率达到 120 线程/间隔 1 秒时，对主页面攻击和主流页面攻击检测率较高，这说明其偏重于检测高请求速率、高负载的攻击；HsMM 算法也较稳定，在攻击频率达到 120 线程/间隔 1 秒时，对随机页面攻击的检测率达到了 89.9%，对重复序列攻击的检测率达到了 90.6%，说明其更偏向于检测访问特征混乱的随机页面攻击与重复序列攻击；同样 D-S 算法的检测率浮动明显，这是由于其检测特征的构建过于薄弱，只能够针对特征明显的攻击进行检测。

为验证 KMPCAA 算法在保持高检测率时，误报率能否也保持在较低范围内，对数据集 dataSet_normal_test 进行误报率检测。首先将数据集 dataSet_normal_test 按照会话中访问数多少划分为高、中、低 3 种类型，单个会话的平均访问数为 5，设定高访问量为大于 10 的会话，低访问量为小于 5 的会话，如表 8 所示。

表 8 dataSet_normal_test 数据集的组成

数据集名称	请求数目 x	会话数	页面数
高访问量	x>10	1 203	53
中访问量	5≤x<10	2 355	181
低访问量	x <5	713	26

依据 FSE、HsMM、D-S 3 种检测算法的初始参数设置，对数据集 dataSet_normal_test 进行正常访问行为误报率检测，并与 KMPCAA 检测结果进

表 7 检测率对比

攻击类型	40 线程/间隔 7 秒				80 线程/间隔 4 秒				120 线程/间隔 1 秒			
	KMPCAA	FSE	HsMM	D-S	KMPCAA	FSE	HsMM	D-S	KMPCAA	FSE	HsMM	D-S
主页面攻击	58.4	55.4	55.5	70.8	87.6	83.4	81.7	78.1	85.2	83.4	87.1	80.8
随机页面攻击	55.5	50.2	59.3	68.3	84.7	68.2	86.5	75.6	90.3	70.4	89.9	82.3
主流页面攻击	57.1	58	53.4	62.1	86.3	83	79.6	69.4	91.9	88.0	85.0	76.1
重复序列发送	54.2	55.2	59.0	55.0	83.4	74.2	85.2	72.3	89.6	75.2	90.6	79
平均检测率	56.3	57.2	57.1	64.6	85.5	75.2	83.3	73.9	89.1	82.2	87.7	80.6

行对比，如表 9 所示。

表 9　　误报率对比

误报率	检测算法			
	KMPCAA	FSE	HsMM	D-S
高访问量	5.1	7.1	5.3	8.8
中访问量	3.5	5.2	3.9	4.7
低访问量	1.1	1.4	0.9	2.2
平均	4.1	5.6	4.3	6.5

通过分析表 9 中 4 种检测算法的误报率，可以看出随着访问量的增加，4 类检测算法的误报率均上升。高访问量状态下的网络流量与发生 DDoS 攻击时的流量状态相似，因此误报率偏高。本文提出的 KMPCAA 算法与其他算法相比，平均误报率偏低，证明该算法的误报率能够被控制在一定范围内，在检测 App-DDoS 攻击时性能显著。

为了更加全面地评价 KMPCAA 算法的检测性能，采用 ROC(receiver operating characteristic)曲线分析检测算法的动态检测性能，ROC 曲线描述了检测算法在不同检测阈值条件下检测率与误报率之间的折中关系（如图 3 所示）。

由图 3 可以看出，本文提出的 KMPCAA 算法综合检测率偏高，ROC 曲线下面积最大，能够实现低误报率和高检测率，同其他 3 种算法相比，在相同的误报率下，能够实现更高的检测率。由此可以得出本研究设计的 *K* 均值多重主成分分析算法是非常有效可行的，值得进一步研究，从而提升其性能。

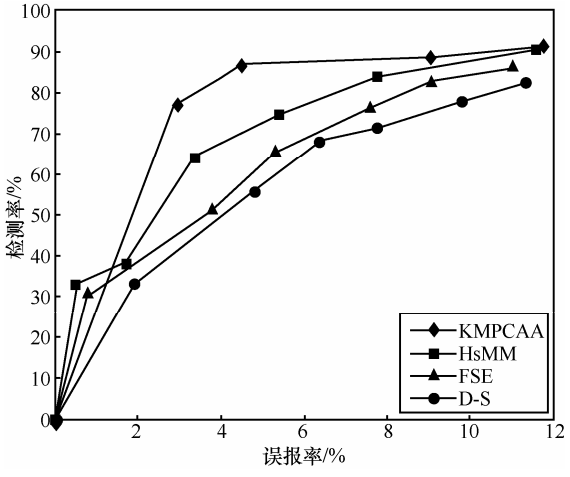

图 3　ROC 曲线对比

5　结束语

本文基于用户访问行为模式，分析了攻击者与正常用户的访问行为区别，提取出适合表征用户访问特征的统计属性。采用最大距离划分法对传统 *K* 均值聚类算法进行了改进，利用递归原理将分布距离最远的中心点逐步筛选到中心点集合中，实现了聚类 *K* 和中心点的自动设定；结合主成分分析法的数据降维特性，提出了一种 *K* 均值多重主成分分析算法。使用经过 MATLAB 编程实现的 KMPCAA 算法同 FSE、HsMM、D-S 3 种检测算法进行检测对比实验，结果证明了 KMPCAA 算法具有较好的检测性能。

由于目前攻击样本种类的局限，未来工作的重点是考虑采用更多的攻击样本进行训练，检验并改进 KMPCAA 算法对多种类型网站系统 App-DDoS 攻击的检测性能。

参考文献：

[1] 第 30 次中国互联网络发展状况统计报告[R]. 中国互联网络信息中心, 2012.
The 30th China Internet Network Development State Statistic Report[R]. China Internet Network Information Center, 2012.

[2] 张鹏. Arbor Pravail APS:专注抵御应用层 DDoS 攻击[J]. 通信世界, 2011:41.
ZHANG P. Arbor Pravail APS: focus on the application layer DDoS attack[J]. Communications World Weekly, 2011:41.

[3] 2011 年中国互联网网络安全态势综述[R]. 国家计算机网络应用技术处理协调中心, 2012.
In 2011 China's Internet Network Security Situation Were Reviewed[R]. National Computer Network Application Technology Processing Coordination Center, 2012.

[4] 龙士工, 赵梦龙. 基于可信度的 App-DDoS 攻击的分布式流量控制模型[J]. 信息安全, 2009, 25(3-3):75-76, 302.
LONG S G, ZHAO M L. Distributed flow control model of the App - DDoS attacks based on the credibility[J]. Information Security, 2009, 25(3-3):75-76, 302.

[5] 魏兵, 徐震. 基于验证机制的应用层 DDoS 攻击防御方法[J]. 计算机工程与设计, 2010, (31)2:231-234.
WEI B, XU Z. Defense approach against application level DDoS attacks based on authentication mechanism[J]. Computer Engineering and Design, 2010, (31)2:231-234.

[6] 张苒. 基于权重队列的 HTTP DDoS 防范技术研究[J]. 辽宁师专学报, 2007, 9(4):40-42.
ZHANG R. Based on the weight of queue HTTP DDoS prevention technology research[J]. Journal of Liaoning Teachers College, 2007, 9(4):40-42.

[7] 赵国锋, 喻守成, 文晟. 基于用户行为分析的应用层 DDoS 攻击检测方法[J]. 计算机应用研究, 2011, 28(2):717-719.
ZHAO G F, YU S C, WEN C. Detecting application-layer DDoS attack based on analysis of users'behaviors[J]. Application Research of Computers, 2011, 28(2):717-719.

[8] 郁继锋. 基于数据挖掘的 Web 应用入侵异常检测研究[D]. 武汉:华中科技大学, 2011.

YU J F. Research on Anomaly Intrusion Detection of Web Application Based on Data Mining[D]. Wuhan: Huazhong University of Science and Technology, 2011.

[9] 张烜. 基于应用层的 DDoS 攻击检测防御技术研究[D]. 北京:北京邮电大学, 2009.
ZHANG H. Based on Application Layer DDoS Attack Detection Defense Technology Research[D]. Beijing: Beijing University of Posts and Telecommunications, 2009.

[10] 毛丽玮. 基于 BP 神经网络的产能建设单井效益评价研究[D]. 青岛:中国石油大学(华东), 2012.
MAO L W. Research into the Evaluation of Oil Well Performance in Productivity Construction Based on the BP Neural Network[D]. Qingdao: China university of Petroleum (East China), 2012.

[11] 王秀芳, 王岩. 优化 *K* 均值随机初始中点的改进算法[J]. 化工自动化及仪表, 2012, 39(10):1302-1304.
WANG X F, WANG Y. Optimize *K*-means random initial midpoint algorithm[J]. Chemical industry automation and instrumentation, 2012, 39(10):1302-1304.

[12] 谢逸, 余顺争. 基于 Web 用户浏览行为的统计异常检测[J]. 软件学报, 2007, 18(4):967-977.
XIE Y, YU S Z. A large-scale hidden semi-Markov model for anomaly detection on user browsing behaviors[J]. IEEE/ACM Trans Netw, 2009,17(1):54-65.

[13] KHATTAB S, GOBRIEL S, MELHEM R, *et al*. Live baiting for service-level DoS attackers[A]. Proc of the Infocom[C]. 2008.682-690.

[14] 徐鹏. 通用应用层 DDoS 检测防护模型的研究[D]. 南京:南京理工大学, 2008
XV P. General Protective Model of Application Layer DDoS Detection[D]. Nanjing: Nanjing University of Science and Technology, 2008.

[15] 谢亚. 基于模糊综合评判的应用层 DDOS 攻击检测方法研究[D]. 成都:西南交通大学, 2009.
XIE Y. Research on Anomaly Intrusion Detection of Web Application Based on Data Mining[D]. Chengdu: Southwest Jiaotong University, 2009.

[16] 张伟, 范宽, 张梦媛. 基于D-S证据理论应用层DDOS攻击检测[J]. 江苏科技大学学报(自然科学版), 2012, 26(3):295-299.
ZHANG W, FAN K, ZHANG M Y. An application layer DDoS attack detection method based on D-S evidence theory[J]. Journal of Jiangsu University of Science and Technology(Natural Science Edition), 2012, 26(3):295-299.

[17] LEE S, SUNG J, KIM D. Incremental update of linear appearance models and its application to AAM: incremental AAM[A]. Lecture Notes Computer Science[C]. 2007.538-547.

[18] YATAGAI T, ISOHARA T, SASASE I. Detection of HTTP-GET flood attack based on analysis of page access behavior[A]. Proceedings IEEE Pacific RIM Conference on Communications. Computers and Signal Processing[C]. Victoria B C, Canada, 2007.232-235.

[19] XUAN Y, SHIN I, THAIMT E T A L. Detecting application denial-of-service attacks: a group-testing-based approach[J]. IEEE Trans on Parallel and Distributed Systems, 2010, 21(8):1203-1216.

作者简介：

杨宏宇（1969-），男，吉林长春人，中国民航大学教授、博士生导师，主要研究方向为网络信息安全。

常媛（1992-），女，河北石家庄人，中国民航大学硕士生，主要研究方向为网络信息安全。