doi: 10.12068/j.issn.1005-3026.2019.01.006

基于流时间影响域的网络流量异常检测

徐久强, 周洋洋, 王进法, 赵 海
(东北大学 计算机科学与工程学院, 辽宁 沈阳 110169)

摘 要: 针对如何提高网络流量异常行为检测准确率的问题,提出基于网络流时间影响域(TID)的网络流量检测模型. 通过分析正常和异常情况下流量网络模型平均度的变化,构建了基于复杂网络平均度指标的网络流量异常检测算法. 实验结果表明,基于网络流时间影响域的流量网络模型能合理地描述网络流量间的依赖关系,具有良好的检测性能,同时该网络模型仅需时间戳、源IP、目的IP三维网络特征即可实现,检测方法适用于绝大多数网络类型,检测效率优于其他网络流量异常检测方法,具有较高的普适性.

关 键 词: 网络流量;异常检测;流时间影响域;流量网络模型;网络平均度
中图分类号: TP 393 **文献标志码**: A **文章编号**: 1005-3026(2019)01-0026-06

Anomaly Detection of Network Traffic Based on Flow Time Influence Domain

XU Jiu-qiang, ZHOU Yang-yang, WANG Jin-fa, ZHAO Hai
(School of Computer Science & Engineering, Northeastern University, Shenyang 110169, China. Corresponding author: WANG Jin-fa, E-mail: jinfa.wong@gamil.com)

Abstract: Aiming at improving the accuracy rate of anomaly network traffic detection, a network traffic detection model was proposed based on the time influence domain(TID) of network flow. By analyzing the changes of average degree of traffic network model under the normal and abnormal conditions, an anomaly detection algorithm of network traffic based on the average degree metric of complex network was developed to detect the abnormal traffic. Experimental results show that based on the flow time influence domain, the anomaly detection model of traffic network can reasonably describe the inter-dependency relationship between network traffic. The proposed method has a better detection performance, meanwhile only three network features, i.e. timestamp, source IP and destination IP, are needed to implement the above model. Detection efficiency is better than other methods. The method proposed meets most network types and has a better ubiquity.

Key words: network traffic; anomaly detection; flow time influence domain; traffic network model; network average degree

目前互联网所承载的数据,基本形态为网络流量. 为保证业务的可靠运行,构建可信的网络环境,减少各类异常事件对通信网络及其承载业务的危害,网络流量异常行为的检测变得更加重要. 针对网络流量异常检测技术,程艳云等提出一种基于大数据的全新时间序列异常点检测方法[1]. 赵海等为了研究地震网络的动力学行为,提出基于时空影响域的地震网络[2]. 贺涛提出基于网络数据流依赖关系的拟阵构造[3]. 由此可知,网络间的数据流之间存在依赖关系,并且可以进行量化. 大多数异常攻击的通信行为在时间上是分阶段进行的,在异常通信行为中,通信行为在一个合理的时间影响域内存在依赖关系[4].

基于网络流量特征具有自相似特性[5],本文提出基于流时间影响域的流量网络模型(TID网络模型)并用于流量异常检测,提出基于网络平均度的网络流的异常检测方法,并通过实验数据进行验证.

收稿日期: 2017-10-18
基金项目: 中央高校基本科研业务费重大科技创新项目(N161608001).
作者简介: 徐久强(1966-),男,辽宁北镇人,东北大学教授; 赵 海(1959-),男,辽宁沈阳人,东北大学教授,博士生导师.

1　TID 网络模型构建

异常行为是连续行为，即当在一个源 IP 的流中发现了异常行为，那么在下一阶段中网络行为仍是异常的是个大概率事件，因此可知网络通信行为具有时间局部性特征，在一个时间影响域(time-influence-domain，TID)内的网络数据流之间存在依赖关系是一个大概率事件，因此基于流时间影响域的网络流量异常检测是可行的.

通过对网络数据流中的每条数据流提取三维网络特征 x(时间戳 t、源 IP、目的 IP)的网络流量数据集 $D=\{x_1,x_2,\cdots,x_n\}$，$n\geqslant 2$，构造基于时间影响域的流量网络来表示数据流之间的依赖关系，并通过分析其特征值的变化来检测网络异常行为发生的时间. 首先将网络数据流量按照时间演进顺序分解为更小的尺度分量，每个小分量为一个采样窗口，窗口长度为 Δw，然后基于流时间影响域内网络数据流间的依赖关系构造 TID 网络模型，最后计算网络特征值，分析可能的网络异常时间段，进行网络流量异常检测[6-10].

数据集 D 以采样窗口 Δw 进行切割后得到流量子集为

$U_i=\{x_j \mid x_j\in D, t_{i\cdot\Delta w}<t_{x_j}<t_{(i+1)\cdot\Delta w}\}$，$x_j$ 表示在 $i\cdot\Delta w$ 到 $(i+1)\cdot\Delta w$ 采样窗口下的网络数据流.

1.1　网络节点

考虑到通信行为的方向问题，本文构造两种不同节点的 TID 网络模型，节点的认定方式如下：

SrcIP→DstIP：如果存在从源 IP 到目的 IP 的通信数据流就认定为一个节点；

SrcIP↔DstIP：如果存在从源 IP 到目的 IP 或从目的 IP 到源 IP 的通信数据流就认定为一个节点.

1.2　连接关系

在网络攻击行为中，在网络流的时间影响域内，每一条网络流(TID 网络模型的节点)与在该条流之后该影响域内的每条网络流之间均具有相关性. 因此在该时间影响域内可以建立连接关系. TID 网络模型中的连接关系代表网络设备间通信行为的依赖关系.

网络数据流是按照时间来进行数据采集的，按照采样窗口长度 Δw 进行数据切割后，得到若干个采样窗口，每个采样窗口均造一个 TID 网络，得到若干个 TID 网络，记为 $T=\{t_1,t_2,\cdots,t_n\}$，$n\geqslant 2$.

1.3　TID 网络构建实例

以图 1 为例，图 1a 中数据描述如下：

时间戳(0 ~ 8)表示当前网络采样窗口下数据流时间戳；数据流(A ~ D)表示当前网络采样窗口下存在的数据流，A ~ D 分别对应具有不同的源 IP 与目的 IP 组合的数据流；矩形条带表示该网络流的影响范围，即流时间影响域.

根据当前采样窗口的网络数据流，以每条网络数据流为初始节点，通过时间影响域判断对应节点在该网络数据流的时间影响域中是否和其他网络节点形成连接关系，形成 TID 网络模型，如图 1b 所示.

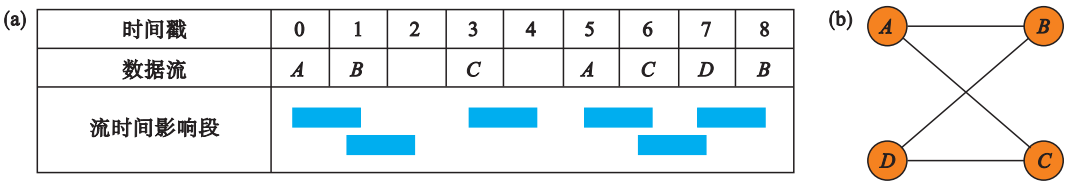

图 1　对应采样窗口 t_i 的 TID 网络模型构建示例

Fig. 1　An example for TID network model in time-window t_i

(a)—网络拓扑时序图；(b)—对应 TID 网络 w_i.

2　流量异常检测方法研究

近年来的研究发现，数据网络中的业务量呈自相似特性[11]，这种网络的自相似性成为网络流量异常检测的理论基础. 在异常情况下，网络拓扑结构发生变化，网络拓扑特征值也相应发生变化.

当网络数据流发生变化时会导致在当前时间影响域下的网络平均度均会偏离，故可通过分析网络平均度分布的变化，推断可能的网络异常行为时间段.

本文构造 TID 网络模型来表示网络数据流之间的依赖关系，通过分析网络拓扑结构的网络平均度分布情况来判断可能的异常网络时间段.

异常网络行为在时间上是有序的，大多数的异常行为都有一个共同的行为特征，它们发生时间较短，并且攻击行为集中. 这种异常行为会导致网络特征分布发生变化. 图 2 所示为不同的攻击

行为导致的网络异常流量特征分布情况.

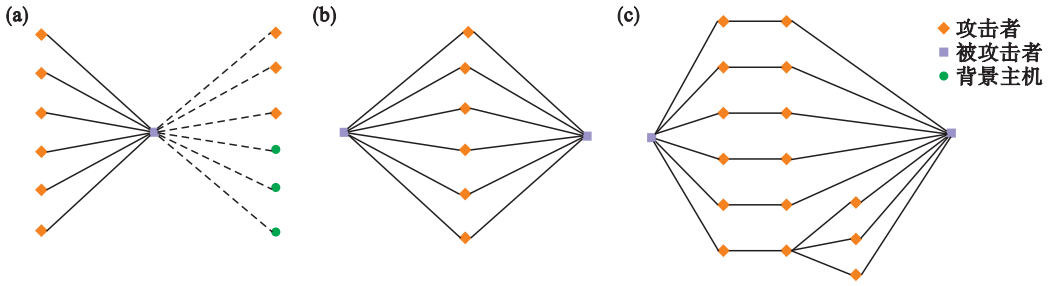

图 2 基于时间顺序的异常流量特征分布

Fig. 2 Distribution of abnormal traffic characteristics based on chronological order

(a)—分布式拒绝服务攻击;(b)—端口扫描攻击;(c)—蠕虫攻击.

基于以上分析可知,网络发生异常行为时,会使网络节点的节点度分布发生变化,因此可知,根据度分布来分析异常网络行为是可行的. 复杂网络中,节点的度及网络的平均度 $<k>$ 定义如下:

定义 1 节点的度. 在网络中,节点 v_i 的邻边数 k_i 称为该节点 v_i 的度.

定义 2 网络平均度. 在网络中对所有节点的度求平均值,可得到网络的平均度 $<k>$:

$$<k> = \frac{1}{N}\sum_{i=1}^{N} k_i. \tag{1}$$

在 TID 网络模型中,当其样本数量变大,随着时间顺序连续变化时,其网络行为是连续行为,因此网络平均度分布符合中心极限定理[12]. 互联网具有自相似性,因此当实验数据结果超过某个阈值时,可认为该时间段为可能的异常时间段.

在本文中,异常判定的规则如下定义:

$$v(U_i) = \begin{cases} 1, & D_c(t_i, U_i) \leqslant \eta; \\ 0, & \text{其他}. \end{cases} \tag{2}$$

其中:$D_c(t_i, U_i)$ 为 $i\cdot\Delta w$ 到 $(i+1)\cdot\Delta w$ 采样窗口下的 TID 网络模型的网络平均度;η 为网络拓扑的网络平均度偏差;0 表示该采样窗口对应的时间段可能为异常网络时间段;1 表示该采样窗口对应的时间段可能为非异常网络时间段.

该规则表明,当在第 i 个采样窗口中,当前 TID 网络模型的网络平均度的偏差超过 η 时,认为采样窗口对应的时间段为异常时段. η 的取值,本文通过对其他场景下正常网络流量的数据集进行训练获得[13]. 本文中,μ 为 TID 网络模型的网络平均度均值,σ 为标准差,L 为给定对应置信区间的正态分布的分位数,置信区间由正常流量训练可得. 阈值 η 的计算公式为

$$\eta = \mu + L\cdot\sigma\ . \tag{3}$$

3 实验结果分析

3.1 实验数据集

本文使用数据集包括背景流量与异常流量所构成的真实僵尸网络流量,选取 3 个僵尸网络场景,并用标签表示该条数据流是否为异常流量[14]. 僵尸网络场景描述如表 1,表 2 所示.

表 1 僵尸网络场景特点[5]

Table 1 The characteristics of the botnet scene

场景	IRC	SPAM	CF	PS	DDoS	US
1	有	无	无	无	有	有
2	有	无	无	有	无	无
3	有	无	无	无	有	有

注:IRC(IRC botnet)为基于 IRC 协议控制的僵尸网络;SPAM 为垃圾邮件;CF(click fraud)为点击欺诈;PS(port scan)为端口扫描;DDoS 为分布式拒绝服务攻击;US(compiled and controlled by us)为对恶意软件尝试编译并可控.

表 2 僵尸网络场景标签分配[5]

Table 2 Labeling of botnet scene

场景	背景流量		异常行为流量		正常流量	
	数量	占比/%	数量	占比/%	数量	占比/%
1	3 895 469	94.60	6 466	0.15	33 610	7.93
2	6 881 228	90.22	383 215	5.02	362 594	4.75
3	4 535 493	87.54	323 441	6.24	321 917	6.21

3.2　实验结果及评价

3.2.1　实验数据结果

本文中，采样窗口长度 $\Delta w = 2.5$ min，流时间影响域确定为 $\Delta t = 75$ ms. 抽取背景流量，即为无攻击行为的网络数据流量. 由此，可将实验分为两部分：背景流实验和攻击流实验. 将两种不同的网络流量均按照 TID 网络模型进行处理. 背景流实验与正常流实验结果的对比过程仅用于对 TID 网络模型的正确性验证，并不作为可能的异常时间段的确定. 在实验过程中，发现以 SrcIP↔DstIP 和以 SrcIP→DstIP 为节点的 TID 网络模型的网络平均度呈现同样的变化趋势，故在本实验中仅表示以 SrcIP↔DstIP 为节点的 TID 网络模型实验结果.

图 3 为这 3 个场景的实验结果及可能的异常网络时间段判别结果. 其中，横坐标表示采样窗口 w_i，为了方便表示，将其映射为对应的时间，纵坐标表示 TID 网络模型的网络平均度，检测结果表示根据攻击流实验结果检测的可能的异常行为时间段.

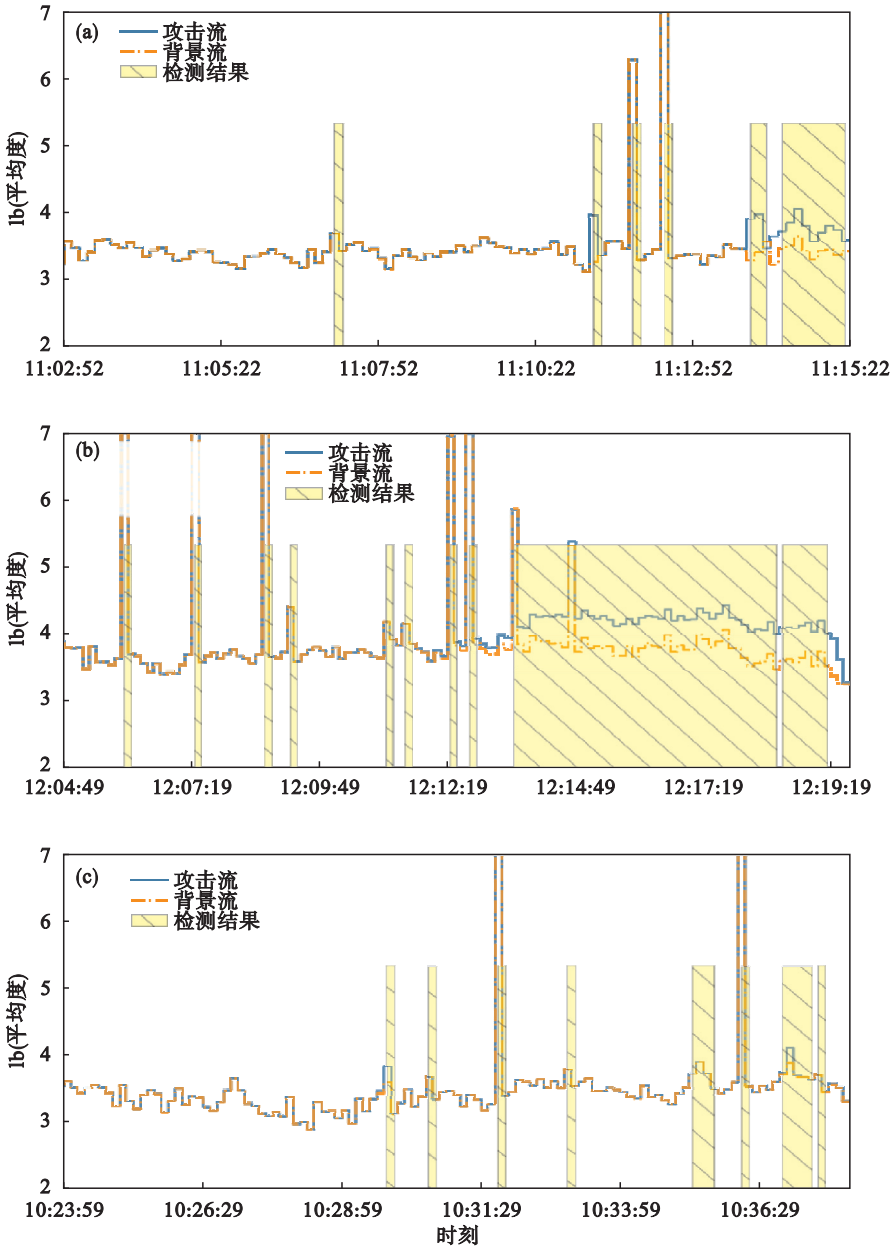

图 3　攻击流实验与背景流实验对比及检测结果

Fig. 3　Comparison between attack flow and background flow and detection results

(a)—场景 1；(b)—场景 2；(c)—场景 3.

从图 3 中可以得出，在无攻击情况下攻击流的实验结果与背景流的实验结果相同；当发生攻击行为时实验结果不同，在数据结果不同的时间段为异常行为发生的时间. 本文中背景流实验与

攻击流实验数据结果不同时,所得的异常时间段同捷克理工大学判定的异常时间段相吻合. 由此可知,TID 模型对异常行为检测是合理且有效的. 接下来将给出仅依据攻击流实验可能出现的异常时间段的判定结果.

3.2.2 依据攻击流实验确定可能的异常时间段

实验过程中针对场景 1、场景 2、场景 3 实验数据集分别构造了 TID 网络,依据网络平均度分布确定阈值,获取可能的异常网络时间段. 场景 1、场景 2、场景 3 依据 TID 模型进行可能的异常网络时间判别结果见表 3.

其中场景 2 与其他网络流量异常检测方法的实验对比结果见表 4. 本实验评价指标采用准确率(accuracy,ACC)、精确率(precision,PRE)、召回率(recall,REC)和综合评价指标($F_{b=1}$). 其中 TP(true positive)表示模型预测为正常流量且模型预测正确的样本数量,TN(true negative)表示模型预测为异常流量且模型预测正确的样本数量,FP(false positive)表示模型预测为正常流量但模型预测错误的样本数量,FN(false negative)表示模型预测为异常流量的样本且模型预测错误的样本数量.

表 3 TID 检测异常网络时间判别结果

Table 3 TID detection results for abnormal network time

场景	TP	TN	FP	FN	ACC/ %	PRE /%	REC /%	$F_{b=1}$/%
1	10	91	4	3	93.52	71.43	76.92	74.07
2	50	61	19	7	82.22	74.63	87.72	80.65
3	4	107	3	10	89.52	57.14	28.57	38.09

表 4 场景 2 实验对比结果

Table 4 Experimental comparison results for scenes 2

模型	TP	TN	FP	FN	ACC/%	PRE /%	REC /%	$F_{b=1}$/%
TID	50	61	19	7	82.22	74.63	87.72	80.65
捷克理工	46	57	26	8	75.18	63.89	85.19	73.02

实验结果表明,TID 模型对网络流量的异常检测的准确率最高达 93.52%,精确度最高达 74.63%,通过实验对比可知,TID 网络模型的检测结果优于异常检测结果. 对比 3 个场景的实验结果,场景 3 的异常发生的持续时间较突然且短暂,因此导致异常网络流量数据中的数据依赖程度较低,且通过分析数据集可知,在场景 3 中存在用户数据集激增的情况,因实验数据是基于捷克理工大学的真实校园网络的背景流量,而背景流量存在不可控及不可预知性,其网络平均度远高于正常情形下的网络平均度,推测原因为在此采样窗口下,校园用户激增,故导致在场景 3 中的 ACC 较高,但 PRE 较低.

4 结 论

本文提出流时间影响域的概念,同时基于流时间影响域,首次提出基于流时间影响域的网络流量异常检测模型,用于网络流量异常时间段的检测. 以网络平均度作为衡量指标,通过分析其分布特点,判断可能的异常网络时间段,并通过实验加以验证. 实验结果表明,TID 网络模型,在网络流量异常检测分类中取得了良好的效果,其平均准确率达到 88.42%,可以较为准确地判断出可能的异常网络时间段,且更加适用于异常行为连续发生的情况. 本文为异常网络流量检测的研究提供了新的思路和方法,该方法仅需获取网络数据流中的时间戳、源 IP、目的 IP,可以满足绝大部分网络,具有普适性.

参考文献:

[1] 程艳云,张守超,杨杨. 基于大数据的时间序列异常点检测研究[J]. 计算机技术与发展,2016,26(5):139 - 144.
(Cheng Yan-yun, Zhang Shou-chao, Yang Yang. Research on time series outlier detection based on big data. [J]. *Computer Technology and Development*, 2016, 26(5): 139 - 144.)

[2] 赵海,张娅,何璇,等. 基于时空影响域的地震网络动力学演化特征分析[J]. 东北大学学报(自然科学版),2015,36(9):1232 - 1236.
(Zhao Hai, Zhang Ya, He Xuan, et al. Dynamic evolution analysis of earthquake network based on the time-space influence domain [J]. *Journal of Northeastern University (Natural Science)*, 2015, 36(9): 1232 - 1236.)

[3] 贺涛. 基于网络数据流依赖关系的拟阵构造[D]. 上海:复旦大学,2009.
(He Tao. Matroid contruction based on data streams dependent relationship in network [D]. Shanghai: Fudan University, 2009.)

[4] 程光,龚俭,丁伟. 基于抽样测量的高速网络实时异常检测模型[J]. 软件学报,2003,14(3):594 - 599.

(Cheng Guang, Gong Jian, Ding Wei. A real-time anomaly detection model based on sampling measurement in a high-speed network [J]. *Journal of Software*, 2003, 14(3): 594 – 599.)

[5] Grill M, Stiborek J, Zunino A. An empirical comparison of botnet detection methods [J]. *Computers & Security*, 2014, 45: 100 – 123.

[6] Akoglu L, Tong H, Koutra D. Graph based anomaly detection and description: a survey [J]. *Data Mining & Knowledge Discovery*, 2014, 29(3): 626 – 688.

[7] Oz L E, Eilertson E, Lazarevic A, et al. MINDS-minnesota intrusion detection system[J]. *Cd Technology*, 2007, 31(5): 151 – 153.

[8] Ahmed M, Mahmood A N, Hu J. A survey of network anomaly detection techniques [J]. *Journal of Network & Computer Applications*, 2016, 60: 19 – 31.

[9] Sridharan N A, Ye T, Bhattacharyya N S. Connectionless port scan detection on the backbone [C]// IEEE International Performance Computing and Communications Conference. Phoenix, 2006: 567 – 576.

[10] Tavallaee M, Stakhanova N, Ghorbani A A. Toward credible evaluation of anomaly-based intrusion-detection methods[J]. *IEEE Transactions on Systems Man & Cybernetics: Part C*, 2010, 40(5): 516 – 524.

[11] Xu K, Zhang Z L, Bhattacharyya S. Profiling internet backbone traffic: behavior models and applications[J]. *ACM SIGCOMM Computer Communication Review*, 2005, 35(4): 169 – 180.

[12] Barbara D, Wu N, Jajodia S. Detecting novel network intrusions using bayes [C]// Siam Conference on Data Mining. Philadelphia, 2001: 308 – 317.

[13] Sengar H, Wang X, Wang H, et al. Online detection of network traffic anomalies using behavioral distance [C]// International Workshop on Quality of Service. Charleston: IEEE, 2009: 1 – 9.

[14] Hua L, Joe H. Strength of tail dependence based on conditional tail expectation [J]. *Journal of Multivariate Analysis*, 2014, 123(1): 143 – 159.

(上接第 5 页)

[13] Hesse W, Moller E, Arnold M, et al. The use of time-variant EEG Granger causality for inspecting directed interdependencies of neural assemblies [J]. *Journal of Neuroscience Methods*, 2003, 124(1): 27 – 44.

[14] Afshari S, Jalili M. Directed functional networks in alzheimer's disease: disruption of Glabal and local connectivity measuers [J]. *IEEE Journal of Biomedical and Health Informatics*, 2016, 21(4): 949 – 955.

[15] Rubinov M, Sporns O. Complex network measures of brain connectivity: uses and interpretations [J]. *NeuroImage*, 2009, 52(3): 1059 – 1069.

[16] Huang D, Ren A, Shang J, et al. Combining partial directed coherence and graph theory to analyse effective brain networks of different mental tasks [J]. *Frontiers in Human Neuroscience*, 2016, 10(5): 1 – 11.

[17] 黄璐, 王宏. 基于约束独立分量分析的脑电特征提取[J]. 东北大学学报(自然科学版), 2014, 35(3): 419 – 423.
(Huang Lu, Wang Hong. EEG feature extraction based on constrained ICA [J]. *Journal of Northeastern University (Natural Science)*, 2014, 35(3): 419 – 423.)