DOI：10.7652/xjtuxb201704015

轻量级主机数据采集与实时异常事件检测方法研究

张剑[1]，童言[1]，徐明迪[2]，秦涛[3]
（1.武汉数字工程研究所系统科研部，430070，武汉；2.武汉数字工程研究所系统软件部，430070，武汉；
3.西安交通大学电子与信息工程学院，710049，西安）

摘要：针对特征值匹配方法不能检测未知异常的缺点以及常驻采集代理占用大量系统资源的问题，提出一种主机数据采集和异常检测方法。采用智能化的移动代理实现主机数据采集，大幅度降低系统中数据采集代理的数量；结合实时异常检测的需求，采用主成分分析方法对所收集的主机信息进行维度约减，并采用聚类方法对降维后的数据进行聚类分析，挖掘其中的异常点；为消除随机异常点对检测结果的影响，采用基于连续时间窗口的主机异常检测方法实现主机异常的准确检测。实验结果表明：与传统方法相比，数据规模相当的情况下，所提方法的时间复杂度减少了50%以上，检测准确率达到了95%以上，适用于主机异常的实时检测。

关键词：异常检测；移动代理；主成分分析；数据聚类
中图分类号：TP393.2 **文献标志码**：A **文章编号**：0253-987X(2017)04-0097-06

A Method for Data Collection and Real-Time Anomaly Detection of Lightweight Hosts

ZHANG Jian[1], TONG Yan[1], XU Mingdi[2], QIN Tao[3]
(1. System Research Department, Wuhan Digital Engineering Institute, Wuhan 430074, China;
2. System Software Department, Wuhan Digital Engineering Institute, Wuhan 430074, China;
3. School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: A new method for data collection and anomaly detection of hosts is proposed to focus on the problems that the methods based on signature matching cannot detect unknown anomaly and data collection agents occupy too many host resources. Intelligent mobile agents are employed to perform data collection so that the number of collection agents is greatly reduced. In order to achieve the goal of online anomaly detection, the principal component analysis method is employed to reduce the dimension of the data, and the clustering method is used to mine the abnormal events. The host anomaly detection method based on continuous time windows is adopted to eliminate the influence of random outliers. Experimental results show that the proposed method has lower computational complexity and higher detection accuracy, and for same number of records the time complexity is reduced by more than 50% and the detection accuracy is above 95%, compared with conventional method. It is concluded that the method is suitable for real-time detection of host anomaly.

Keywords: anomaly detection; mobile agent; principal component analysis; clustering method

收稿日期：2016-06-09。 作者简介：张剑(1979—)，男，高级工程师；童言(通信作者)，男，博士。 基金项目：国家自然科学基金资助项目(61502438，61672026)；陕西省自然科学基金资助项目(2016JM6040)；国防预研基金资助项目(B0820132036)。
网络出版时间：2017-03-03 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20170303.0925.006.html

计算机网络的快速发展极大地方便了人们的生活，目前全国计算机的拥有量已经超过了 1 亿台，用户数量已经超过了 6 亿，其中大部分用户的安全意识很差，导致网络攻击和病毒频发，给系统安全管理带来了巨大的挑战。主机作为病毒和攻击的直接载体，由主机层可以采集到丰富的数据源，既包括键盘、鼠标等人机交互数据，也包括文件系统及主机性能相关的参数，这些数据源为主机层的异常分析打下了良好的基础。

在过去几年，国内外学者围绕基于主机的异常检测方法做了大量的研究。基于击键数据，文献[1]设计了基于支持向量数据描述的击键生物特征身份认证系统，并将该方法应用于主机异常检测；文献[2]讨论了鼠标数据在异常行为和用户识别方面的可行性。除此之外，通过文件系统的监控也可以实现异常用户行为的检测；文献[3]通过监控文件操作过程中的系统调用来实现入侵检测；文献[4]利用系统调用子集或者定长系统调用子序列构造检测模型，针对主机层多种类型的数据进行综合分析，也可以实现异常检测；文献[5]通过分析 CPU、内存、磁盘空间、缺页率等来实现异常用户的检测；文献[6]提出一种基于多数据源的异常检测方法，从系统调用、资源使用和文件访问提取特征，结合隐马尔科夫模型实现异常检测。

综上所述，国内外学者在主机异常检测方面做了大量的工作，他们在数据采集方面多采用常驻代理形式，这会占用大量的系统资源，并且数据源维度的增多会导致检测时效下降，无法满足实时检测的需求。结合上述问题，本文提出了一种轻量级主机数据采集与异常事件检测方法。

本文采用具有移动性和智能性的移动代理实现数据采集，可以大幅度地减少系统中数据采集的代理数量。结合数据采集需求，本文设计了基于等时间间隔等待的主机资源采集方法，较好地权衡了系统资源和数据采集方面的矛盾，降低数据采集对系统性能的影响。针对异常检测的实时性需求，本文采用主成分分析方法(PCA)对所采集到的数据进行维度约减，提取数据的主要特征，利用基于密度的聚类算法(DBSCAN)实现异常点的检测。为了减少随机异常点带来的影响，设计了一种基于连续时间窗口的主机异常检测方法。实验结果表明，所提方法具有较高的检测准确率和较低的计算复杂度，可实现主机异常的实时检测。

1 异常检测方法框架

本文所提出的主机异常检测模型共分为主机数据采集、属性提取、DBSCAN 聚类分析、异常筛选以及测试验证等 4 个步骤，其流程框架如图 1 所示。

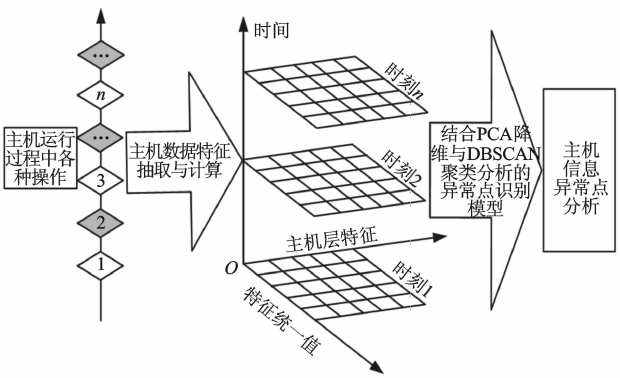


图 1 主机行为异常检测框架图

具体实施步骤分别详述如下：

步骤 1 根据相关研究和异常检测需要，采用移动代理采集主机层面的行为数据信息，包括 CPU、内存、IO、进程等，借以描述用户正常行为和异常行为在主机层面的节拍模式；

步骤 2 采用 PCA 方法提取数据的主要特征，在获得稳定性数据的同时降低了后续算法的复杂度；

步骤 3 根据所提取的主机层特征，结合 DBSCAN 聚类算法，根据获得的孤立点实现异常点的检测；

步骤 4 根据异常行为的持续性特征，设计了基于连续时间窗口的主机异常检测方法，由所搭建的实验平台采集了多台主机的相关数据，对所提方法进行了测试和验证。

2 主机层数据采集方法

2.1 主机层数据特征

为实现异常行为检测，本文采集了涵盖 CPU、内存使用、磁盘读写、数据包流量以及进程使用和开放端口等 6 类数据特征，见表 1。以上 6 类特征是根据入侵规律和异常表现而选择的[7]。例如外部入侵多表现为 CPU、内存、磁盘读写以及流量中的一个或多个显著变化。病毒会产生非法进程，具有传播功能的病毒则会增加开放端口。这些特征可为主机层异常检测提供良好的数据支持。

2.2 数据采集移动代理

移动代理技术是人工智能技术、计算机网络技术与分布式技术的结合，它为分布式应用提供了一

表 1　主机层特征定义及说明

特征	具体含义
CPU 占用率	CPU 的使用特征
内存使用率	物理内存的使用特征
磁盘读取	磁盘读取字节特征
磁盘写入	磁盘写入字节特征
接收数据包	网卡接收字节特征
发送数据包	网卡发送字节特征
进程	运行进程列表情况
端口	开放端口情况

种移动计算方式，可以提高这些应用的效率和可靠性[8]。移动代理是一类特殊的代理，它具有移动性、智能性、自主性，可以根据任务需要自主移动到有移动计算平台的计算机系统上执行相关操作。本文基于 IBM 的 Aglet 平台开发了多类数据采集移动代理[9]，提升了数据采集的效率。

2.3　基于等时间间隔轮询的主机信息采集方法

在数据采集过程中，代理需要频繁的查询相关资源，容易占用较多的系统资源。本文设计了基于等时间间隔主动轮询机制的主机信息采集方法，以降低其对系统的影响。具体的算法如下：

步骤 1　设置轮询等时间间隔 Δt 与截止采集时间 t_0；

步骤 2　获得当前系统时间 t_1，获得当前主机的 IP 和 MAC 地址以及主机的相关性能指标（例如 CPU、内存占用、磁盘读写、数据包流量以及进程和端口信息等）；

步骤 3　将步骤 2 所获得的信息按适当格式存入数据库；

步骤 4　如果 $t_1 > t_0$，则结束采集，否则延迟时间间隔 Δt，转至步骤 2。

采用上述方法在 2015 年 12 月至 2016 年 3 月分两次采集了 10 台主机的相关数据，见表 2。其中磁盘读写特征包括磁盘读和磁盘写。网络数据特征包括网卡接收和发送的数据包。

表 2　主机数据样本

样本类型	样本数量	样本含主机数
CPU 使用率	39 561	10
内存使用率	39 561	10
磁盘读写	39 561	10
网络数据包	39 561	10
运行进程	1 086	10
开放端口	1 086	10

3　主机数据 PCA 降维分析

由于所采集的主机层特征数据维度较高，提升了后续聚类检测算法的计算复杂度，本文首先采用主成分分析方法（PCA）进行主机信息维度约减。主成分分析利用线性变换把数据变换到一个新的坐标系统中，使得数据投影的方差按大小一次投影在各个主成分上，在减少数据维度的同时也可保持数据集的显著特征[10]。我们对所采集的数据（6 000 个连续时刻的数据集）进行了降维分析，相关系数矩阵的特征根和贡献率见表 3，结果可以看出，前两个最大的特征根的贡献率达到了 94.9%，即当数据维度约减为 2 时，依然可以保留原始数据近 95%的特征。

表 3　样本相关系数矩阵的特征根及其贡献率

特征根	贡献率/%	累计贡献率/%
5.622	80.3	80.3
1.020	14.6	94.9
0.244	3.5	98.4
0.101	1.4	99.8
0.013	0.2	100
0.000	0.0	100

4　基于时间窗口和 DBSCAN 的主机异常检测

4.1　基于 DBSCAN 的主机异常点检测方法

DBSCAN 是基于密度的聚类算法，与划分和层次聚类算法不同，它将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在噪声的数据库中发现任意形状的聚类[11]，其中算法有 MinPoints 和 Eps 两个重要的参数，在本文中分别用 M 和 ε 表示。不同的取值对结果影响很大，本文在实验的基础上实现上述参数的选择。首先，利用 DBSCAN 算法对降维后的主机数据进行聚类分析，并根据聚类得到的孤立点进行异常事件检测。设 $U=\{p_1, p_2, \cdots, p_n\}$ 为测试数据集，其中 p_i 为数据点。经过 DBSCAN 聚类，得到若干个簇 C_1，$C_2, \cdots, C_l$ 和离群点 $m_1, m_2, \cdots, m_n$。将每个离群点作为一个单独的簇，按照簇中的元素数降序排序，使得所得到的簇满足下式

$$|C_1| \geqslant |C_2| \geqslant \cdots \geqslant |C_k| \geqslant |\{m_1\}| \geqslant |\{m_2\}| \geqslant \cdots \geqslant |\{m_n\}| \quad (1)$$

根据网络管理经验，必有大多数数据点都集中于几个较大的簇，正常簇和异常簇的点数有很大差

别，其中较大的簇为正常数据点，即主机在大多数情况下都在正常运行。上述网络管理经验可以表示为

$$\left.\begin{aligned}&\frac{|C_1|+|C_2|+\cdots+|C_l|}{\sum|C_i|+\sum|\{m_i\}|}\geqslant\alpha\\&\frac{|C_l|}{|C_{l+1}|}\geqslant\beta\end{aligned}\right\}\tag{2}$$

式中：α 表示正常主机点在主机总数中所占的比值；β 表示由异常点所构成的簇要远小于正常簇中点的数量，可以根据实验结果实现这两个阈值的选取。

4.2 基于连续滑动时间窗口的主机异常检测模型

由于用户在正常使用计算机或操作系统进行资源调度时也将导致 CPU、内存等数据特征偶尔发生异常变化，单个时间点的异常不能准确判定主机是否感染了病毒或者受到了攻击，而主机受到入侵、病毒感染时都会产生一个时间窗口以上的持续性异常，不只是单点异常。据此，本文提出了基于时间窗口的主机异常检测算法，如图 2 所示，其中 t_i 为时间点，而时间窗口包括了多个时间点。

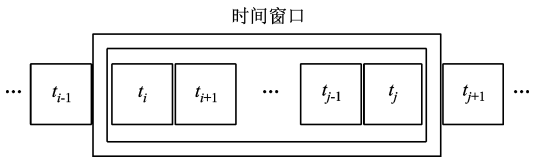


图 2 基于时间窗口的主机异常检测算法

在时间窗口 $\Delta t=t_j-t_i+1$ 内，使用 DBSCAN 聚类算法获得当前时间窗口内的主机数据异常点，用 A_i 表示窗口内检测到的异常点。当异常点的数量 ω 超过一定阈值时，认定该时间窗口内产生了主机异常，如下式所示

$$\left.\begin{aligned}&\omega=\sum_{t-\Delta t}^{t}1_{\{A_i\}}\\&A(t)=\begin{cases}1,\ \omega(t)\geqslant\theta\\0,\ \omega(t)\geqslant\theta\end{cases}\end{aligned}\right\}\tag{3}$$

式中：阈值 θ 可以根据实验结果选取。该算法根据时间窗口内的异常点的数量判断该时间窗口是否产生异常，能规避单一时间点异常所造成的偶然性。

5 算法测试与分析

5.1 测试环境

为了对所提算法进行测试与验证，搭建了实验平台并进行主机信息收集。实验平台中的正常主机和异常主机具有相同的配置。测试主机配有 Interl (R) E1280 CPU，8.00 GB 内存，100 Mb/s 带宽网卡。在软件方面，测试主机安装有 Windows 7 64 bit 操作系统以及 Aglets 2.0.2 运行平台。同时，这些测试主机中安装有常用软件，具备正常主机的状态。根据前文所设计的主机数据采集方法，连续采集接近 40 000 次，共计超过 600 h 的主机数据(所有主机合计)。选取其中一台主机作为异常主机，在该机器上利用 IBS 套件模拟合成消耗主机信息资源的病毒和 Worm 病毒分时段感染该主机。将感染病毒时间段内主机的状态定义为异常状态，将未感染病毒的主机时间段定义为正常状态来实现测试数据集的构建。

5.2 维度约减效率分析

利用 5.1 节所收集到的数据，经 PCA 将数据约减到不同的维度，在聚类参数 M 取值为 6、ε 取值为核心对象与其 M 个最近邻对象的距离平均值时的聚类结果见表 4。结果表明，将高维度数据约减到合适维度后，聚类仍能得到与原始高维度数据集相一致的结果，说明维度约减后的数据保留了原始高维数据的特征。计算数据约减到不同维度时 DBSCAN 聚类算法所耗费的时间，如图 3 所示，经 PCA 将原始数据约减到二维时可以大幅度降低聚类算法的计算复杂度，在数据集大小为 1 000 时，约提高 128.8%的时效。

根据表 3、表 4 以及图 3 的结果可见，将原始数据维度约减到 2 时既能保持原始数据的特征，也能有效降低聚类算法的时间复杂度。

表 4 PCA 降维数据与原始数据集聚类结果对比

数据类别	聚类数	离群点数	聚类结果相似度/%
7 维数据	2	11	100
3 维数据	2	11	100
2 维数据	2	11	100
1 维数据	2	1	99

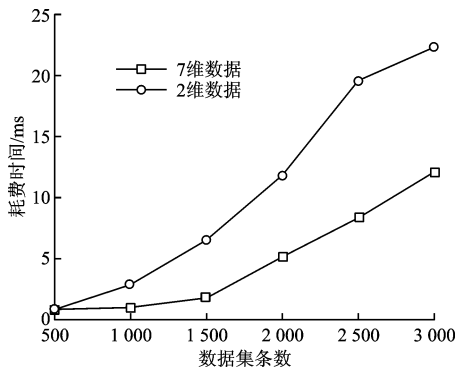
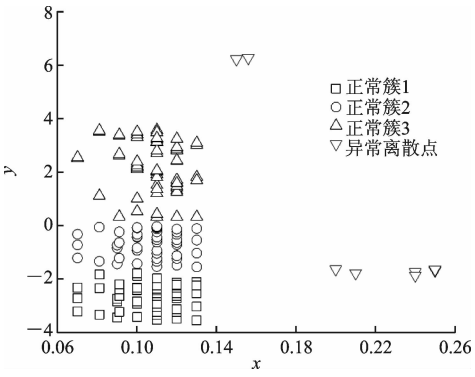


图 3 不同维度数据聚类算法耗时分析

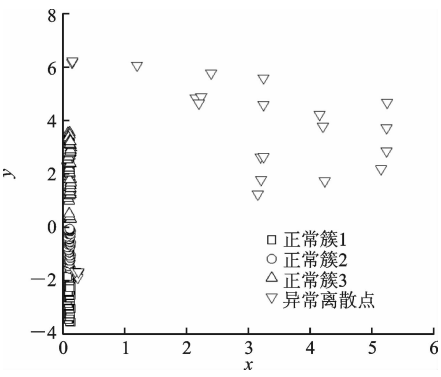
5.3　基于 DBSCAN 的主机异常点识别算法测试

将收集到的正常数据和异常数据混合作为测试数据集，并将数据集分为多个子数据集，每个测试子数据集的规模在 1 000 个元素以上，正常数据的元素个数占据 98%，异常元素个数占据 2%。聚类分析结果如图 4 所示。正常状态一般表现为两种模式：静默模式是在用户不参与的情况下系统所表现出来的状态；用户正常行为模式是用户正常的行为活动参与表现出来的模式。病毒入侵后计算机的行为模式发生改变，多表现为较小的簇或离散点，这些离散点和正常簇之间的差别较大，如图 4b 中相同的坐标下，正常点聚集在一起，而异常点分布较为散乱。

在基于主机数据的异常点检测中，参数 β 的取值范围对结果影响不大。将 β 取为聚类结果中最小簇元素的个数与正常数据集中异常点数最大值的比值，经计算其值为 5。α 的取值十分重要，将聚类结果中两个最大簇中元素的总数与数据集元素总数的比值作为 α 的取值，经多次计算得到 $\alpha=0.95$。



(a)正常运行时聚类结果



(b)异常运行时聚类结果

图 4　正常数据和含异常数据聚类结果对比

5.4　基于时间窗口的主机异常检测算法测试分析

在连续时间窗口异常检测模型中，为了保证检测算法的实时性，本文将窗口大小设定为 10 个计数值范围，因此 ω 的取值范围为 $0\leqslant\omega\leqslant10$。将 DBSCAN 算法检测结果应用于时间窗口的主机异常检测算法，获得了不同阈值 θ 下的检测结果，见表 5。由表 5 可以看出，随着阈值的增大误报率逐渐降低，但漏报率会大幅增加。在异常检测与控制中，通常会优先降低误报率，而当阈值在[3　4]之间时，可在有效减少误报的同时得到较高的识别率(95%左右)，因此本文选择阈值为 3。

本文所提算法无须训练，因此其可以实现未知异常的检测，这对于弥补由于特征匹配等常见的入侵检测方法无法检测未知异常的缺点至关重要。我们也将算法结果与仅使用 DBSCAN 算法在最优情况下的误报率和漏报率进行了对比，结果见表 6，实验结果表明本文提出的方法更为有效，大幅度降低了误报率。

表 5　异常阈值与检测主机异常精度

异常阈值	检测率/%	误报率/%	漏报率/%
1	98.5	3.2	1.5
2	97.2	1.0	2.8
3	95.5	0.7	4.5
4	94.1	0.0	5.9
5	91.4	0.0	8.6
6	86.5	0.0	13.5
7	77.9	0.0	22.1
8	62.1	0.0	37.9
9	39.8	0.0	60.2
10	14.6	0.0	85.4

表 6　DBSCAN-TW 算法与 DBSCAN 算法结果对比

算法	漏报率/%	误报率/%
EDBSCAN-TW($\theta=3$)	4.5	0.7
DBSCAN	3.1	42.9

6　结　论

本文分别从主机数据采集、维度约减和异常检测算法 3 个方面详细介绍了主机数据采集与异常事件检测方法。首先，采用智能化的移动代理实现主机信息的采集，并根据异常检测的需要，采用 PCA 方法对主机信息进行了降维分析，实验结果表明在将数据约减为 2 维时依旧可以保持原始数据近 95%的特征。结合异常事件的特点，提出了聚类分析和连续滑动时间窗口相结合的主机异常检测算法。最后，基于 Aglet 搭建了实验平台，并采集了大量数据对算法进行了测试和验证。分析结果表明，所提出的方法具有较高的检测准确率和较低的计算

复杂度,可用于主机异常的实时检测。由于本文所提的方法仅适合于快速传播异常的检测,在下一步工作中,我们将重点分析时间窗口的选择和动态聚类方法对检测结果的影响,实现隐蔽性攻击的检测。

参考文献:

[1] 倪桂强,李佳桢,潘志松,等. 基于支持向量数据描述的击键生物特征认证 [J]. 模式识别与人工智能,2008,21(5):704-708.
NI Guiqiang, LI Jiazhen, PAN Zhisong, et al. Verification based on keystroke biologic characteristics using support vector data description [J]. Pattern Recognition and Artificial Intelligence, 2008, 21(5): 704-708.

[2] CAI Zhongmin, SHEN Chao, GUAN Xiaohong. Mitigating behavioral variability for mouse dynamics: a dimensionality reduction method [J]. IEEE Transactions on Human-Machine Systems, 2014, 44(2): 244-255.

[3] PENNINGTON A, STUNK J, GRIFFIN J, et al. Storage based intrusion detection: watching storage activity for suspicious behavior [C]// Proceedings of the 12th USENIX Security Symposium. New York, USA: ACM, 2003: 137-151.

[4] FORREST S, HOFMEYR A, SOMAYAJI A, et al. A sense of self for UNIX processes [C]// Proceedings of the 1996 IEEE Symposium on Security and Privacy. Piscataway, NJ, USA: IEEE Communication Society, 1996: 120-128.

[5] 李涵,包立辉. 基于聚类算法的异常入侵检测模型的研究与实现 [J]. 计算机应用与软件,2006,23(10):126-127.
LI Han, BAO Lihui. Research and implementation of an anomaly intrusion detection system model based on cluster analysis [J]. Computer Applications and Software, 2006, 23(10): 126-127.

[6] HAN Sangjun, CHO Sungbae. Rule-based integration of multiple measure-models for effective intrusion detection systems [C]// Proceedings of the 2003 IEEE International Conference on Man and Cybernetics. Piscataway, NJ, USA: IEEE System, Man and Cybernetics Society, 2003: 120-125.

[7] 陶敬,马小博,赵娟,等. 基于资源可用性的主机异常检测 [J]. 电子科技大学学报,2007,36(S3):1449-1452.
TAO Jing, MA Xiaobo, ZHAO Juan, et al. A method for host abnormal detection based on resource availability [J]. Journal of University of Electronic Science and Technology of China, 2007, 36(S3): 1449-1452.

[8] DALMEIJER M, HAMMER D, AERTS A. Mobile software agents [J]. Computers in Industry, 2000, 41(3): 251-260.

[9] IBM. IBM aglets software development kit-home page [EB/OL]. (2003-06-23) [2014-05-10]. http://web.media.mit.edu/~stefanm/ibm/AgletsHome Page/index_new4.html.

[10] 赵蔷. 主成分分析方法综述 [J]. 软件工程,2016,6(19):1-3.
ZHAO Qiang. A review of principal component analysis [J]. Software Engineering, 2016, 6(19): 1-3.

[11] CAMPELLO R, MOULAVI D, SANDER J. Density based clustering based on hierarchical density estimates [C]// Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery in Databases. Berlin, Germany: Springer, 2013: 160-172.

(编辑 武红江)

(上接第 5 页)

LI Changhe, YUAN Suoxian, XIU Shichao. Automatic balancing technology of grinding wheel in superhigh speed grinding [J]. New Technology & New Process, 2004(5): 29-31.

[3] DIONYS H. Apparatus for compensating for unbalance of a rotary body: USA 4050195[P]. 1977-09-27.

[4] 章云,梅雪松,胡振邦,等. 注液式高速切削主轴动平衡装置设计及其性能研究 [J]. 西安交通大学学报,2013,47(3):13-17,23.
ZHANG Yun, MEI Xuesong, HU Zhenbang, et al. Design and performance analysis of hydrojet-typed balancing device for high-speed machine tool spindle [J]. Journal of Xi'an Jiaotong University, 2013, 47(3): 13-17, 23.

[5] 吴海琦,潘鑫,高晖. 气压液体式转子在线自动平衡器: CN103874868A[P]. 2014-06-18.

[6] 贺世正. 释放液体式自动平衡头的研究 [J]. 浙江大学学报(工学版),2001,35(4):70-74.
HE Shizheng. Study of liquid release auto-balancing head [J]. Journal of Zhejiang University(Engineering Science), 2001, 35(4): 70-74.

[7] 王积伟,章宏甲,黄谊. 液压与气压传动 [M]. 2 版. 北京:机械工业出版社,2005:45-48.

[8] 杨继宏. 激光打微孔作用机理及工艺研究 [D]. 天津:天津大学,2008:10-11.

(编辑 杜秀杰)