基于改进聚类分析的网络流量异常检测方法

李洪成[1]，吴晓平[1]，姜洪海[2]
（1. 海军工程大学 信息安全系，湖北 武汉 430033；
2. 海军北海舰队 司令部，山东 青岛 266071）

摘 要：针对传统基于聚类分析的网络流量异常检测方法准确性较低的问题，提出了一种基于改进 *k*-means 聚类的流量异常检测方法。通过对各类流量特征数据的预处理，使 *k*-means 算法能适用于枚举型数据检测，进而给出一种基于数值分布分析法的高维数据特征筛选方法，有效解决了维数过高导致的距离失效问题，并运用二分法优化 *K* 个聚簇的划分，减少了初始聚类中心选择对 *k*-means 算法结果的影响，进一步提高了算法的检测率。最后通过仿真实验验证了所提出算法的有效性。

关键词：网络安全；流量异常检测；聚类分析；*k*-均值算法
中图分类号：TP309.7
文献标识码：A
doi: 10.11959/j.issn.2096-109x.2015.00009

Traffic anomaly detection method in networks based on improved clustering algorithm

LI Hong-cheng[1], WU Xiao-ping[1], JIANG Hong-hai[2]
(1. Information Security Department, Naval University of Engineering, Wuhan 430033, China;
2. Headquarters, Command of Naval North-Sea Fleet, Qingdao 266071, China)

Abstract: To solve the problem that traditional traffic abnormal detection methods were not accurate enough, a traffic anomaly detection method based on improved *k*-means was proposed. All kinds of network traffic data were preprocessed to make *k*-means algorithm can apply to enumeration data detection. Then a features selection method was proposed with the analysis of the distribution of network traffic data to avoid the distance useless caused by too much features. Furthermore, the clustering process of *K* clusters was optimized based on dichotomy, aiming to reduce the effects of initial clusters centers selection. Simulation results demonstrate the effectiveness of the algorithm.

Key words: network security; traffic abnormal detection; clustering analysis; *k*-means algorithm

1 引言

网络流量异常检测是指以网络流数据为输入，通过统计分析、数据挖掘和机器学习等方法，发现异常的网络数据分组和异常网络交互等信息。网络流量异常检测过程首先需要使用 sniffer、NetFlow、fprobe 和 flow-tools 等数据流抓取工具来采集海量的网络数据流信息，然后从数据中提取和选择出可用于检测异常的数据属性，通过对数据属性的分析得出该数据记录为正常或异常的

收稿日期：2015-09-15；**修回日期**：2015-10-08。**通信作者**：李洪成，ytztybzb@163.com
基金项目：国家自然科学基金资助项目（61100042）；中国博士后基金资助项目（2014M552656）；湖北省自然科学基金资助项目（2015CFC867）。
Foundation Items: The National Natural Science Foundation of China (61100042); Postdoctoral Science Foundation of China (2014M552656); The Natural Science Foundation of Hubei Province (2015CFC867)

结论。

网络流量异常检测方法主要包括基于无监督学习、监督学习和半监督学习的方法。其中，唐成华等[1]和 Xu 等[2]利用无监督的聚类方法自动提取网络流量异常模式，该类方法的缺点是算法复杂度高、检测率低[3]；武小年等[4]和郑黎明等[5]分别利用支持向量机和流量熵理论来对网络流量数据进行监督学习，该类方法在时效性和准确性上较优，但需要大量经过标记的样本[6-8]；陆悠等[9]利用半监督学习将监督学习和无监督学习结合，在准确性和标记成本之间取得了很好的折中。

由于在实时网络流量异常检测中无法得到大量带标记的样本记录，所以应采用无监督的聚类检测算法[10]。另外，由于网络流量数据体量巨大，因此采用快速聚类算法 *k*-means 较为适合。基本 *k*-means 算法的流程如图 1 所示。

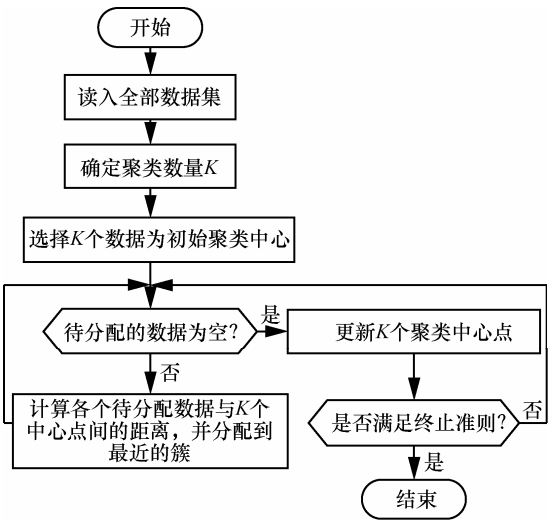

图 1　基本 *k*-means 聚类算法流程

将 *k*-means 聚类方法直接应用于网络流量异常检测工作会面临以下几点问题：①只能处理连续型数据；②对初始聚类中心的选择过于敏感；③当特征数量过多时，空间中的点会变得稀疏，从而使距离失效[11,12]。本文针对 *k*-means 聚类算法在检测中存在的问题，拟提出基于改进聚类算法的异常检测方法，进而提高聚类检测的准确率。

2　基于改进 *k*-means 的网络流量异常检测

2.1　网络流量特征数据的预处理

假设网络流量数据集中的记录总数为 N，各记录的特征个数为 P，第 i 条数据记录的第 p 个特征的数值表示为 $a_{ip}(1\leqslant i\leqslant N,\ 1\leqslant p\leqslant P)$。

典型网络流量数据包含以下几种特征类型：①无序枚举型特征，如协议类型、连接状态和目标主机的网络服务类型等；②有序枚举型特征，如访问系统敏感文件的次数、过去两秒内与当前连接有相同目标主机的连接数等；③{0,1}型特征，如是否成功登录、是否为 guest 用户等；④有序连续型特征，如在固定时间内，与当前连接在某特征处取值相同的连接数占总连接数的百分比等。对于不同的特征类型，应采用不同的数据预处理和距离度量方法。

考虑到网络流量数据体量十分巨大，因此在数据的预处理过程中应尽量简化运算过程，提高预处理的时效性。本文在数据预处理的过程中，对于有序枚举型和有序连续型特征，利用如下公式进行数据预处理

$$b_{ip}=\frac{a_{ip}}{R_p} \tag{1}$$

其中，b_{ip} 为 a_{ip} 预处理后的数值，R_p 为数据集中第 p 维特征数值取值范围的上限。

以上预处理方法运算过程简单，可以有效提高数据预处理的效率。同时，对于网络流量数据，特征数值取值范围的上限可以一定程度上反映特征数值的大小，预处理过程可以理解为归一化过程，因此利用式(1)进行有序枚举型和有序连续型特征的预处理可以在时效性和合理性之间取得很好的折中。

对于无序枚举型特征，为便于存入数组并进行下一步操作，用整数数字代表特征的不同数值；对于{0,1}型特征，不进行数据预处理，直接代入下一步运算。

2.2　网络流量数据特征优化选择

在网络流量异常检测过程中，流量数据往往具有很多特征，如果将这些特征全部参与聚类过程，那么会出现维数过高的难题。维数过高会导致空间中的点变得稀疏，从而使距离失效。此外，网络流量各维特征在异常检测中的重要程度往往差异较大，因此有必要对高维网络流量进行特征选择，提高异常检测的准确性。本文提出一种基于特征数值分布分析的特征优化选择方法。该方

法通过计算不同网络流量数据记录在某特征处取值的均方差，并对其进行归一化处理，得到该特征的权重。计算公式如下

$$\overline{b_p}=\frac{1}{N}\sum_{i=1}^{N}b_{ip} \tag{2}$$

$$\sigma_p=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(b_{ip}-\overline{b_p}\right)^2} \tag{3}$$

$$w_p=\frac{\sigma_p}{\sum_{p=1}^{P}\sigma_p} \tag{4}$$

其中，$\overline{b_p}$ 表示第 p 个特征分布的平均值，σ_p 表示第 p 个特征分布的均方差，w_p 表示第 p 个特征的权重。

在计算特征权重之后，选取权重较大的特征进行下一步运算，并在数据记录距离度量中代入特征权重进行计算，进一步提高聚类检测结果的准确性。

2.3 网络流量数据记录的距离度量

网络流量数据记录的距离度量是 *k*-means 聚类的重要依据，对于不同类型的特征，其距离度量过程必须区别对待才能保证聚类结果的科学性。

对于{0,1}型特征、有序枚举型特征和有序连续型特征，在特征数据预处理之后，即可代入距离度量函数进行计算，典型的距离度量函数有欧几里得距离和曼哈顿距离，其计算公式如下

$$d_{\mathrm{E}}(i,j)=\sqrt{\sum_{p=1}^{P}\left(b_{ip}-b_{jp}\right)^2} \tag{5}$$

$$d_{\mathrm{M}}(i,j)=\sum_{p=1}^{P}\left|b_{ip}-b_{jp}\right| \tag{6}$$

其中，$d_{\mathrm{E}}(i,j)$ 为第 i 个记录和第 j 个记录之间的欧几里得距离，$d_{\mathrm{M}}(i,j)$ 为第 i 个记录和第 j 个记录之间的曼哈顿距离。考虑到聚类检测效率的因素，选择复杂度较小的曼哈顿距离进行计算，并通过引入特征权重对其进行改进，具体计算公式如下

$$d_{\mathrm{A}}(i,j)=\sum_{p=1}^{P}\left(w_p\times\left|b_{ip}-b_{jp}\right|\right) \tag{7}$$

其中，$d_{\mathrm{A}}(i,j)$ 为第 i 个记录和第 j 个记录之间的{0,1}型特征、有序枚举型特征和有序连续型特征的距离。

对于无序枚举型特征，如果 $b_{ip}=b_{jp}$，则距离 $d_p(i,j)=0$；若 $b_{ip}\neq b_{jp}$，则距离 $d_p(i,j)=1$。无序枚举型特征距离的计算公式如下

$$d_{\mathrm{B}}(i,j)=\sum_{p=1}^{P}\left[w_p\times d_p(i,j)\right] \tag{8}$$

其中，$d_{\mathrm{B}}(i,j)$ 为第 i 个记录和第 j 个记录之间的无序枚举型特征的距离。

综上，网络流量数据记录的距离为以上两部分之和，公式为

$$d(i,j)=d_{\mathrm{A}}(i,j)+d_{\mathrm{B}}(i,j) \tag{9}$$

2.4 基于二分法的 *k*-means 聚类检测

在利用 *k*-means 聚类方法进行网络流量异常检测的过程中，存在的较大问题是初始聚类中心难以确定。对于基于聚类的网络流量异常检测，确定聚类个数 K 和 K 个初始聚类中心点的位置是检测的必要前提，但是网络流量数据是连续动态产生的，初始聚类中心点的位置很难一次性确定，而 *k*-means 聚类方法对于初始聚类中心点位置的选择非常敏感，导致检测率难以保证。为了更加科学合理地确定初始聚类中心点位置，提高聚类检测率，本文提出一种基于二分法的 *k*-means 聚类检测方法。

基于二分法的 *k*-means 聚类分析思想是：设最终需要获得的聚簇个数是 K，首先采用随机选择初始聚类中心的方法将数据一分为二，然后在这 2 个簇中选择较大的一个簇，重复以上过程，并以得到的聚簇数量达到 K 为终止准则。其算法步骤如下。

输入 训练数据集 D，二分次数 M，目标聚类簇数 K。

输出 K 个聚簇集合。

步骤 1 设当前的所有记录组成的簇集为 S，当前的簇数为 k'，令 $k'=1$。

步骤 2 在簇集 S 里选出一个包含记录数量最多的簇。

步骤 3 采用随机选择初始聚类中心的 *k*-means 方法将数据一分为二，共选择 M 次初始聚类中心，聚类 M 次。

步骤 4 在上述每次二分聚类之后，算出各次聚类中子簇对的误差平方和 *Judgement*，把 *Judgement* 最小的子簇放入簇集 S，$k'=k'+1$。

步骤 5 以 $k'=K$ 为终止准则，若 $k'\neq K$，则

重复进行步骤 2~步骤 5。

由于上述步骤使用二分聚类方法逐步产生 K 个聚类，所以该方法检测结果不易受初始聚类中心点位置影响。二分 k-means 聚类中的评价函数是误差平方和 *Judgement*，各步骤中需要找出误差平方和最小的一对子簇。子簇误差平方和的计算公式为

$$Judgement = \sum_{b_i \in U_1} d(i, m_1) + \sum_{b_i \in U_2} d(i, m_2) \quad (10)$$

其中，b_i 为预处理后的第 i 条网络流量数据记录，U_1 和 U_2 分别为本次二分聚类中 2 个簇，m_1 和 m_2 分别为 U_1 和 U_2 的簇中心标识。误差平方和是二分 k-means 算法中判断二分质量优劣的依据。二分 k-means 聚类方法不是随机选择初始聚类中心，而是在每一轮迭代中，选择 m 个子簇对中 *Judgement* 最小的子簇对，进而反复进行二分聚类。

算法的基本流程如图 2 所示。

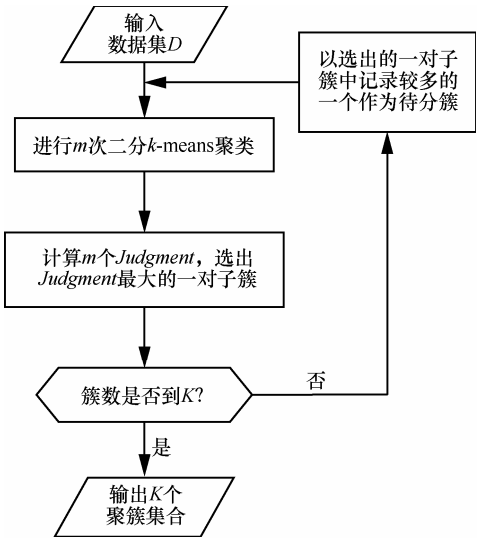

图 2　二分 k-means 聚类基本流程

以上算法运行完成后，可输出最优的聚类方案，进而实现了对网络流量数据记录的划分，根据划分结果可以将各时间段的网络行为辨别为正常行为或各类网络攻击行为。

3　实验设计及结果分析

本文仿真实验使用 KDD99 网络流量数据集。该数据集是对 1998 年的 DARPA 网络环境原始数据经过统计分析得出的，KDD99 数据集分为训练数据集和测试数据集两部分。

在 KDD99 数据集的各条数据记录中，最后一个特征是攻击类型，这些攻击类型分为 4 个大类 39 个小类，其中 17 个小类出现在测试集（Corrected 数据集）中。另外，KDD99 还提供一个 10%的训练子集，KDD99 测试集和 10%训练子集包含攻击大类及其所对应的具体攻击类型如表 1 所示。

在 KDD99 数据集的 10%训练集和测试集中，各类型数据的统计情况如表 2 所示。

实验使用的计算机配置为 Inter Core(TM) i3-3220/ 3.29 GHz/ 2.99 GB RAM，利用 MATLAB 7.1 软件编程。

首先，利用 2.1 节的方法处理各维数据。然后，利用 2.2 节提出的方法，进行特征优化选择，选出 13 维特征参与检测，这 13 个特征分别为 KDD99 数据集的第 2、3、4、10、12、22、23、28、32、33、34、37、39 维特征，这些特征代表的意义如表 3 所示。

在实验中，聚类个数与攻击大类（包括正常行为）保持一致，取 K=5。聚类检测的评价指标采用检测率和误报率，其定义是：检测率为正常检测出的数据记录数和数据记录总数的比值；误报率为被错误标识的数据记录数和数据记录总数的比值。

利用表 2 所示的测试集，分别运用基本 k-means 算法和本文算法进行聚类检测，将聚类检测的结果与测试集数据记录的标记信息对照，得到的检测率和误报率如表 4 所示。

表 1　KDD99 的 10%训练集和测试集所含的攻击类型

所属攻击大类	在 10%训练集中出现的攻击类型	只在测试集中出现的攻击类型
DoS	back、land、neptune、pod、smurf、teardrop	apache2、mailbomb、processtable、udpstorm
R2L	ftp_write、guess_passwd、imap、multihop、phf、warezmaster、spy、warezclient	named、sendmail、snmpgetattack、snmpguess、worm、xlock、xsnoop
U2R	buffer_overflow、loadmodule、perl、rootkit	httptunnel、ps、sqlattack、xterm
Probing	ipsweep、nmap、portsweep、satan	mscan、saint

表 2 KDD99 数据集的统计情况

类型	在 10%训练集中的数量/个	在 10%训练集中占的比例/%	在测试集中的数量/个	在测试集中占的比例/%
DoS	391 458	79.24	229 851	73.90
R2L	1 126	0.23	16 311	5.24
U2R	52	0.01	104	0.03
Probing	4 107	0.83	4 166	1.34
Normal	97 278	19.69	60 593	19.48
Overall	494 021	100.00	311 025	100.00

表 3 选择出的特征名称及其含义

序号	特征名称	特征含义
2	protocol_type	网络协议的类型
3	service	网络流目的主机的服务类别
4	flag	连接正常则为 1，错误为 0
10	hot	网络流访问密级较高资源的次数
12	logged_in	是否成功登录的状态
22	is_guest_login	是否为 guest 登录的状态
23	count	两秒时间里，与该网络连接目的主机一致的连接数目
28	srv_rerror_rate	两秒时间里，标记“REJ”错误的网络连接占与该网络连接服务类型一致的网络连接的比例
32	dst_host_count	在此前 100 个网络连接里与该网络连接目的主机一致的网络连接数目
33	dst_host_srv_count	在此前 100 个网络连接里，与该网络连接目的主机和服务类型均一致的网络连接数目
34	dst_host_same_srv_rate	在此前 100 个网络连接里，与该网络连接目的主机和服务类型均一致的网络连接占的比例
37	dst_host_srv_diff_host_rate	在此前 100 个网络连接里，在与该网络连接目的主机和服务类型均一致的网络连接中，与该网络连接源主机不一致的网络连接所占比例
39	dst_host_srv_serror_rate	在此前 100 个网络连接里,在与该网络连接目的主机和服务类型均一致的网络连接中,标记“SYN”错误的网络连接所占比例

表 4 算法的检测率和误报率

类型	基本 *k*-means 的检测率/%	基本 *k*-means 的误报率/%	本文算法的检测率/%	本文算法的误报率/%
DoS	82.08	17.92	87.01	12.99
R2L	83.44	16.56	88.45	11.55
U2R	85.07	14.93	90.17	9.83
Probing	80.20	19.80	85.01	14.99

由表 4 可以看出，相比于基本的 *k*-means 聚类检测方法，本文采用的改进检测方法检测率较高、误报率较低，这是因为本文的数据预处理方法使得网络流量特征数据更加科学合理地参与检测运算，二分法的使用解决了初始聚类中心选择对检测结果影响过大的问题。

4 结束语

基于数据挖掘的网络流量异常检测工作虽然可以较好地检测出入侵行为，但在应用中仍存在一些问题。例如，对高安全等级网络流量进行数据挖掘工作，会泄露网络流量特征的隐私信息，

同时会为攻击者有针对性地实施攻击和设计特定的攻击手段提供依据。这就要求网络流量数据挖掘工作必须使攻击者不能利用已掌握的流量统计信息推测出其想要获取的流量信息，进而保护网络流量数据隐私。

参考文献：

[1] 唐成华, 刘鹏程, 汤申生, 等. 基于特征选择的模糊聚类异常入侵行为检测[J]. 计算机研究与发展, 2015, 52(3): 718-728.
TANG C H, LIU P C, TANG S S, et al. Fuzzy clustering abnormal behavior detection based on feature selection[J]. Computer Research and Development, 2015, 52(3): 718-728.

[2] XU J H, LIU H. Web user clustering analysis based on k-means algorithm[C]//Proceeding of the 2010 International Conference on Information, Networking and Automation (ICINA), New York. 2010: 6-9.

[3] LI P, LIU L, GAO D, et al. On challenges in evaluating malware clustering[C]. Recent Advances in Intrusion Detection, Ottawa. 2010: 238-255.

[4] 武小年, 彭小金, 杨宇洋, 等. 入侵检测中基于 SVM 的两级特征选择方法[J]. 通信学报, 2015, 36(4): 19-26.
WU X N, PENG X J, YANG Y Y, et al. Two levels of feature selection methods based on SVM in intrusion detection[J]. Journal on Communication, 2015, 36(4): 19-26.

[5] 郑黎明, 邹鹏, 韩伟红, 等. 基于多维熵值分类的骨干网流量异常检测研究[J]. 计算机研究与发展, 2012, 49(9): 1972-1981.
ZHENG L M, ZOU P, HAN W H, et al. Backone network traffic anomaly detection study based on multidimensional entropy classification[J]. Computer Research and Development, 2012, 49(9): 1972-1981.

[6] SHINGO M, CHEN C, LU N N. Intrusion-detection model based on fuzzy class-association-rule mining using genetic programming network[J]. IEEE Transaction on Systems, Man, and Cybernetics, 2011, 41(1): 130-139.

[7] DASH S K, REDDY K S. Adaptive naive Bayes method for masquerade detection[J]. Security and Communications Networks, 2011, 4(4): 410-417.

[8] 张玲, 白中英, 罗守山, 等. 基于粗糙集和人工免疫的集成入侵检测模型[J]. 通信学报, 2013, 34(9): 166-176.
ZHANG L, BAI Z Y, LUO S S, et al. Ensemble intrusion detection model based on rough set and artificial immunity[J]. Journal on Communication, 2013, 34(9): 166-176.

[9] 陆悠, 李伟, 罗军舟, 等. 一种基于选择性协同学习的网络用户异常行为检测方法[J]. 计算机学报, 2014, 37(1): 28-40.
LU Y, LI W, LUO J Z, et al. A method of network users abnormal behavior based on selective collaborative learning[J]. Chinese Journal of Computer, 2014, 37(1): 28-40.

[10] FLAVIO C, NILESH D, RAVI K. Correlation clustering in MapReduce[C]//Proceeding of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2014), New York. 2014: 641-650.

[11] 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.
QIAN Y K, CHEN M, YE L X, et al. Methods of full network anomaly detection based on multi-scale principle component analysis[J]. Journal of Software, 2012, 23(2): 361-377.

[12] RUBINSTEIN B, NELSON B, HUANG L, et al. Stealthy poisoning attacks on PCA-based anomaly detectors[C]//Proceeding of the ACM SIGMETRICS, New York. 2009.

作者简介：

李洪成（1991-），男，河南商丘人，海军工程大学博士生，主要研究方向为信息安全、数据挖掘。

吴晓平（1961-），男，山西新绛人，博士，海军工程大学教授、博士生导师，主要研究方向为信息安全、密码学。

姜洪海（1972-），男，山东乳山人，海军北海舰队司令部工程师，主要研究方向为信息安全。