计算机工程
Computer Engineering
ISSN 1000-3428,CN 31-1289/TP

《计算机工程》网络首发论文

题目：　基于特征分组聚类的异常入侵检测系统研究
作者：　何发镁，马慧珍，王旭仁，冯安然
DOI：　10.19678/j.issn.1000-3428.0054476
网络首发日期：　2019-08-26
引用格式：　何发镁，马慧珍，王旭仁，冯安然．基于特征分组聚类的异常入侵检测系统研究．计算机工程. https://doi.org/10.19678/j.issn.1000-3428.0054476

计 算 机 工 程
Computer Engineering

基于特征分组聚类的异常入侵检测系统研究

何发镁[1,3]，马慧珍[2,3]，王旭仁[2,3]*，冯安然[2,3]
(1. 北京理工大学图书馆，北京 100081；2. 首都师范大学 信息工程学院，北京 100048；3. 中国科学院信息工程研究所 中国科学院网络测评技术重点实验室，北京 100093)

摘 要：利用网络安全审计数据特征可以按照连接的基本特征、内容特征、网络流量特征、主机流量特征进行分组的特点，基于 K-means 算法提出一种按照特征分组进行聚类的方法，使得网络安全审计数据特征项由原来的 41 项降为分组聚类后的 4 项，有效地进行了特征约简和数据降维。通过调整聚类参数，有效保留特征分组内的差异信息，使用决策树 C4.5 算法对降维之后的数据进行入侵分类处理，实验结果表明能够取得较好的入侵检测效果，网络数据连接的检测率为 99.73%、误检率为 0，其中刺探攻击 Probe 的检测率为 100%。

关键词：入侵检测；网络数据；K-means 算法；决策树；降维

Research on anomaly intrusion detection system based on feature grouping clustering

HE Famei[1,3], MA Huizhen[2,3], WANG Xuren[2,3]*, FENG Anran[2,3]
(1. Library, Beijing Institute of Technology, Beijing 100081, China,; 2. Information Engineering College, Capital Normal University, Beijing 100048, China; 3. Key Laboratory of Network Assessment Technology, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China)

【Abstract】According to the features of network security audit data, which can be grouped according to the basic features of connections, content features, network traffic features and host traffic features, a clustering method based on feature grouping is proposed based on K-means algorithm, which reduces the feature items of network security audit data from 41 items to 4 items after grouping clustering, and effectively performs feature reduction and data dimensionality reduction. By adjusting the clustering parameters, the difference information in the feature group is effectively preserved, and it applied decision tree C4.5 algorithm to classifying the data after dimension reduction. The experimental results show that the proposed method can achieve better detection results, and the network data connection detection rate is 99.73%, and the false detection rate is 0, and the detection rate of Probe is 100%.

【Key words】intrusion detection; network data; K-means algorithm; decision tree; data reduction
DOI:10.19678/j.issn.1000-3428.0054476

0 概述

入侵检测主要分为两类：误用检测和异常检测。误用检测通过与已知的攻击进行对比来识别异常或潜在的行为，对已知的攻击具有很好的检测率，但是检测不到新的或不常见的入侵。异常入侵检测是通过对网络传感器收集来的信息或审计日志，采用机器学习和统计学的方法进行分析处理，找到异常行为或恶意攻击时，能及时分类处理以达到保障网络空间安全目的。随着网络速度的不断提升，使得网络传输产生了大量的数据，如果不能及时处理由网络传输产生的数据，将会给网络空间安全带来不可知的威胁。而网络传感器收集或审计日志得来的数据拥有众多的属性，如何在众多的属性中提取出有用的信息提高检测率，降低误检、漏检显得尤为重要。

聚类分析是数据挖掘、模式识别等研究方向的重要研究内容之一，主要是将数据集中相似的样本

基金项目：国家自然科学基金项目资助（61872252）；中国科学院信息工程研究所中国科学院网络测评技术重点实验室开放课题资助（2018）

作者简介：何发镁（1972-），男，博士，研究方向:情报分析、数据挖掘；马慧珍，硕士研究生；王旭仁，副教授/博士；冯安然，硕士研究生.
E-mail：hefm@bit.edu.cn

尽可能划分为相同的簇，而把相异的样本尽可能划分为不同的簇。在入侵检测研究领域 K-means 聚类有着广泛的应用[1~3]。k 的取值与初始聚类中心的选择对聚类结果有着重大的影响。罗等人[4]提出通过设置簇半径 L 来解决 k 的取值和初始聚类中心选择问题，但是 L 的取值成了又一新的问题。夏等人[5]提出通过改进 Seeded K-means 算法中种子集初始聚类中心的选择来解决由此产生的随机性和盲目性，但并未从根本上解决该问题，因为种子集本身存在随机性和盲目性。文献[6]通过改进 K-means 算法解决聚类初始中心问题，但其相当于运行两次 K-means 算法，效率不高。文献[7-8]通过采用 K-means 聚类与其他算法组合的方式来解决 k 值与初始聚类中心选择导致的聚类结果不稳定问题。Gaddam 等人[9]提出 K-means 与 ID3 决策树组合的方法，有效的避免 K-means 聚类存在的强制分配与类优势问题。Wang 等人[10]提出使用 Self-Organizing Map 对数据进行初步的聚类，得到聚类中心作为 K-means 的初始聚类中心，重新定义簇的边界。Muda 等人[11]和 Yassin 等人[12]提出使用 K-means 将数据聚为 k 个簇，然后使用 Naive Bayes 分别对每一个簇进行分类处理。Fachkha 等人[13]采用 EM（expectation maximization）算法获得簇数目 k，然后使用 K-means 将数据聚成 k 个簇，该方法解决了 k 过大过小的问题。Chandrasekhar 等人[14]提出使用 K-means 将数据聚成 k 个簇，对每一个簇的数据使用 Fuzzy Neural Network 进行处理，然后每条数据用(k+1)维的向量表示，其中第(k+1)个数值由簇中数据的隶属度表示，从而达到降维的作用，随后采用 SVM 分类处理。Yaseen 等人[15]提出 M-LHSVM-ELM（Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine）算法，将数据聚成 k 个簇，用 k 个聚类中心点代替原来的数据，对新得到的数据采用多层 SVM 和 ELM 混合的方法分类处理，降低了训练模型的时间。He[16]提出了神经模糊分类（Neural fuzzy classifier，NFC），该方法融合了神经网络和模糊理论。

以上工作通过改进 K-means 算法以及组合算法来避免由于 k 值和初始聚类中心选择导致的聚类结果不稳定问题，但是由于数据分布不均匀，在数据处理过程中未能识别出数据中存在的微小差异，导致对数据量少的攻击类检测率不是很高。文献[17]中通过特征取值分组来更好的描述同类表情中不同特征取值作用的差异。本文提出基于 K-means 的特征分组聚类来描述网络数据中存在的差异，同时可以降低数据维度，然后采用 C4.5 分类处理，实验使用 kddcup99 数据集。初始聚类中心选择不稳定导致的聚类效果不佳和依赖问题由 C4.5 来弱化，从而提高异常检测率和数据量少的攻击类检测率，尤其是 Probe 的检测率得到极大提高。

本文组织如下：第一节介绍本文所涉及到的相关算法，第二节介绍本文的入侵检测模型、特征分组策略、特征分组的聚类算法研究，第三节介绍实验数据特征分组及实验评估标准，第四节介绍实验结果及讨论其在不同数据上的性能，第五节总结本文方法的优缺点。

1 相关算法

1.1 K-means 算法

设训练数据集 S 由 m 个数据对象组成的集合，每个对象由 n 个属性组成。将 S 分成 k 个不相交的簇 $C_1, C_2, \ldots, C_k$，其中 c_i, i=1, 2, …, k, 为簇 C_i 数据的平均值，即聚类中心。在本文中，K-means 算法对于数据对象之间的相似度度量采用欧氏距离，距离越小，相似度越大，反之，距离越大，相似度越小。 欧氏距离的计算方法如下：

$$d=\sqrt{|x_{i1}-c_{j1}|^2+|x_{i2}-c_{j2}|^2+...+|x_{in}-c_{jn}|^2} \quad (1)$$

K-means 算法常采用误差和准则函数作为聚类准则函数，其定义如下：

$$J=\sum_{i=1}^{k}\sum_{j=1}^{m_i}d_{ij}(x_j,c_i) \quad (2)$$

其中 m_i 是簇 C_i 中的数据对象个数，x_j 是簇 C_i 中的数据对象，J 是所有数据对象的误差和。聚类准则函数的作用是尽可能使簇内数据相似度最大，簇间相似度最小，从而使聚类算法获得最佳的聚类结果。

K-means 算法的处理流程描述如算法 1。

算法 1：K-means
输入：簇的数目 k；包含 m 个对象的数据集 S。
输出：k 个簇的集合。
从 S 中任意选择 k 个对象作为初始簇中心；
repeat
　根据公式(1)计算 d，将每个对象分配到最相似的簇；
　更新簇均值，重新计算每个簇的均值 c_i；
until 准则函数收敛。

1.2 决策树

决策树在分类问题中构建的模型以树形结构呈现。1986 年 Quinlan 将信息论中的熵引入决策树研

究领域，利用熵计算数据对象各特征对类标签的影响程度，以此寻找最优特征构建决策树[18]。C4.5 使用信息增益比作为构建决策树过程中子节点选择时最优属性划分的评估方法[19]。

设训练数据集为 S，m 为总的训练样本个数，$F=\{f_1, f_2, \ldots, f_n\}$为特征集合，每条数据由 n 个特征组成。假设有 p 个类，第 i 个类包含 m_i 个训练样本。计算训练样本集 S 的经验熵：

$$H(m_1,m_2,...,m_p)=-\sum_{i=1}^{p}\frac{m_i}{m}\log_2\frac{m_i}{m} \quad (3)$$

特征 f_k,k=1,2,..,n,有 v 个不同的取值$\{\rho_1, \rho_2,\ldots, \rho_v\}$，根据特征 f_k 的取值将样本集 S 分为 v 个子集$\{S_1, S_2,\ldots, S_v\}$，S_j 是特征 f_k 取值为 ρ_j 的样本集合，其中 m_{ij} 是 S_j 第 i 类的样本个数，特征 f_k 的经验条件熵：

$$E(f_k)=\sum_{j=1}^{v}\frac{m_{1j}+m_{2j}+\ldots+m_{pj}}{m}\times H(m_{1j},m_{2j},\ldots,m_{pj}) \quad (4)$$

特征 f_k 的信息增益：

$$Gain(f_k)=H(m_1,m_2,\ldots,m_p)-E(f_k) \quad (5)$$

特征 f_k 的信息增益比：

$$GainRatio(f_k)=\frac{Gain(f_k)}{H(m_1,m_2,\ldots,m_p)} \quad (6)$$

2 基于特征分组聚类思想的入侵检测系统

2.1 入侵检测模型

本文提出基于特征分组聚类的 K-means 算法和决策树 C4.5 算法复合的方法对数据进行分析处理：首先对数据进行预处理；按照一定标准对特征进行分组；使用 K-means 算法对于每个分组内的数据进行聚类，则分组内的所有特征由聚类后的标签所替代，实现了低层高维（数据向高层抽象数据的转化，这样可以更好的区分出数据相似特征中存在的微小的相异性，同时可以降低维度，方便数据进一步处理；对高层抽象数据使用 C4.5 训练分类模型；最后进行测试，验证模型的有效性，具体流程如图 1 所示。

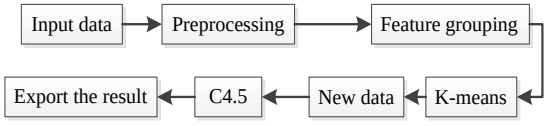

图 1 KBFG-C4.5 入侵检测系统模型

该模型简称为 KBFG-C4.5（K-means Based on Feature Grouping Plus C4.5）模型。

2.2 特征分组策略

如何将相似的特征分组到一起？分组恰当，会更好地提取高层抽象数据，提高分类模型的准确率；反之不合适的特征分组会加重训练模型的时间复杂度，降低模型的准确率。本文提出两种特征分组方式。

基于先验知识的特征分组。例如将图片的特征按照颜色、形状、光亮度进行分组。

均匀分组策略。例如一个 n 维的特征分成 l 组，那么每组包含[n/l]个特征。

2.3 基于特征分组的聚类算法研究

假设数据集 $S=\{x_1, x_2, \ldots, x_m\}$，其特征集合为 $F=\{f_1, f_2, \ldots, f_n\}$，每条数据由 n 个特征组成。现将特征集合 F 依据某种标准进行分组，分离为 l 个属性集，$l<n$。即 $F=F_1\cup F_2\cup\ldots\cup F_l$，$F_1\cap F_2\cap\ldots\cap F_l=\varnothing$，$F_1=\{f_1, f_2,\ldots, f_i\},\ldots, F_l=\{f_j, f_{j+1}, \ldots, f_n\}$，$1<i<j<n$。依据特征分组 F_1，F_2，…，F_l 对应得到 l 个数据集 S_1，S_2，……，S_l。假设数据集 S_1，S_2，……，S_l 聚类数目分别为 $k_1, k_2,\ldots, k_l$，则 KBFG 算法流程如算法 2 所示。

算法 2：KBFG
输入：样本集 S，特征分组数据集 S_t 对应的聚类数目 k_t，$1\le t\le l$。
输出：特征分组数据集簇集合 $C=\{C_1, \ldots, C_t\}$，S_t 对应的簇集合为 C_t，$1\le t\le l$。
for i=1: l
　从 S_i 中任意选择 k_i 个样本作为初始簇聚类中心$\{c_{i,1}, \ldots, c_{i,k_i}\}$；
end
for i=1: l
　repeat
　　(1) 用 n_i 表示第 i 个分组的特征数目，将 x_j 分组数据$(x_{j,1}, \ldots, x_{j,n_i})$根据公式(1)计算 d
　　(2) 将$(x_{j,1}, \ldots, x_{j,n_i})$标记为最小的 d 所对应的类别 λ，更新簇集合 $C_{i,\lambda}=C_{i,\lambda}\cup\{(x_{j,1}, \ldots, x_{j,n_i})\}$
　　(3) 重新计算簇 $C_{i,\lambda}$ 的聚类中心

$$c_{i,\lambda}=\frac{1}{|C_{i,\lambda}|}\sum_{x\in C_{i,\lambda}}x$$

　until 准则函数收敛。
end

使用 KBFG 算法对于每个样本数据按照分组进行聚类，执行完毕后，将样本集 S 内的样本数据 x，依次将 x 中第 i 个分组的 n_i 个特征值由该分组聚类后的标签 λ 所替代，$1\le i\le l$。这样样本 x 的特征值由原来的 n 个变为 l 个，实现了低层高维数据向高层抽象数据的转化。讨论几种特殊情况：

当 l=1 时，KBFG 算法退化为 K-means 算法，KBFG-C4.5 模型不适用；

当 l=n 时，KBFG 算法对样本集 S 内的样本数

据 x 的所有特征进行了聚类处理，虽然没有降低特征的维数，但是每个特征的特征值进行了聚类抽象，实现了特征值从低层到高层数据的转换。KBFG-C4.5 模型适用；

当 $l<<n$，使用 2.2 节讨论的方法进行确定。即达到特征高层抽象、数据量缩减（甚至是数量级的减少）的目的，又能够实现高效的入侵检测。

在 KBFG 算法中，特征分组数据集 S_t 对应的聚类数目 k_t（$1\le t\le l$）的值需要通过反复实验，根据 KBFG-C4.5 模型检测评估效果确定。

3 实验数据特征分组及实验评估标准

3.1 数据集及特征分组选择

本文使用 kddcup99 数据集[20]。kddcup99 数据集的每条数据由 41 个属性和 1 个类标签组成，数据记录格式如下例子所示：

0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,1,0,0,255,254,1,0.01,0,0,0,0,0,0,Normal.

基于先验知识将特征分组分为 4 个组：TCP 连接基本特征(第 1~9 个特征,其中 3 个标称属性)，TCP 连接的内容特征（第 10~22 个特征），基于时间的网络流量统计特征（第 23～31 个特征），基于主机的网络流量统计特征（第 32～41 个特征）。

3.2 数据预处理

将数据集中的标称属性值数值化[21]。假设一个标称属性包含 m 个值，则将该标称属性值变为 m 维向量。例如，数据集的“protocol type”的特征值包括 tcp、udp、icmp，则将特征值变为(1,0,0)。

数据中各特征的取值范围不同，其中取值较大的特征将会降低取值较小的特征在计算过程中所起的作用，即大数吃小数现象。因此对数据进行归一化处理。本文将特征的取值转化为 0 到 1 区间内的值，如公式（7）所示：

$$newValue=\frac{oldValue-\min}{\max-\min} \tag{7}$$

其中 min 和 max 分别为特征值的最小值和最大值。

3.3 数据集划分

kddcup99 数据集包括训练（train data）与测试数据（test data）集，总共包含 38 种攻击，被分为以下 4 种主要类型：拒绝服务攻击(Dos)、远程攻击(R2L)、本地用户非法提升权限的攻击(U2R)、网络刺探(Probe)。数据集的数据分布如表 1 所示。

本文实验使用 10%的训练数据，共 494,021 个连接数据，其中包含 97,278 个正常连接，攻击连接数据包含 22 种攻击类型；10%的测试集包含 311,029 个连接数据，正常连接有 60593 个，攻击连接数据包含 14 种在训练数据集中不存在的新的攻击类型。数据集中攻击数据分布如表 2 所示。

表 1 实验数据分布

	Normal	Dos	Probe	R2L	U2R	total
train	97278	391458	4107	1126	52	494021
test	60593	229853	4166	16189	228	311029

3.4 评估方法

为了衡量 KBFG-C4.5 的分类效果，采用 DR(detection rate)，FPR(false positive rate)作为评估措施。将实验数据分为正常类和攻击类。一个性能好的入侵检测具有高的 DR，低的 FPR 值。

$$\mathrm{DR}=\frac{TP}{TP+FN} \tag{8}$$

$$\mathrm{FPR}=\frac{FP}{FP+TN} \tag{9}$$

其中：

TP—将攻击类预测为攻击类数；

FN—将攻击类预测为正常类数；

FP—将正常类预测为攻击类数；

TN—将正常类预测为正常类数。

表 2 攻击数据分布

类别	train 攻击类型	test 攻击类型
Normal	normal	normal
Dos	back、land、neptune、pod、smurf、teardrop	back、land、neptune、pod、smurf、teardrop 、apache2、mailbomb、processtable、udpstorm
Probe	Ipsweep 、nmap 、portsweep、satan	Ipsweep 、nmap 、portsweep、satan、mscan、saint
R2L	ftp_write 、guess_passwd、imap、multihop、phf、spy、warezclient 、warezmaster	ftp_write、guess_passwd、imap 、multihop 、phf 、warezmaster 、named 、sendmail、snmpgetattack、snmpguess、worm、xlock、xsnoop
U2R	buffer_overflow 、loadmodule 、perl 、rootkit	buffer_overflow 、loadmodule、perl、rootkit、httptunnel、ps、sqlattack、xterm

4 实验结果及讨论

本文的仿真实验使用一台处理器为 Intel® Core™ i7-4790 CPU @ 3.60GHz 、内存 8.00GB 的

64 位 Windows 7 旗舰版的台式机，weka 工具辅助。

首先对实验数据进行预处理，预处理包含两个部分：标称属性值数值化、所有数据归一化处理；然后依据数据的特征类型将数据分离成 4 个新的数据集，即 l=4，示例如图 2 所示；最后对处理后的数据使用算法 2 进行降维，将数据转化为图 2 中用簇号表示的一条数据，其图中 $C_{1,6}$ 表示该样本第一个特征分组数据经过 KBFG 算法被聚到簇 6 中，以此类推如图中所示。最后使用 C4.5 对高层抽象之后的数据进行分类，输出结果。

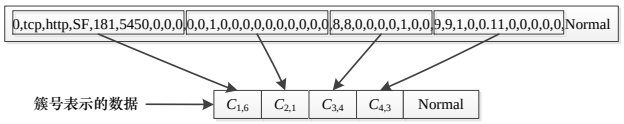

图 2 数据属性分组示例

4.1 分组数据集 S_t 对应的聚类数目 k_t ($1 \leqslant t \leqslant l$)的确定

在 KBFG 算法中，各分组的聚类数目 k_t ($1 \leq t \leq l$) 的取值由人工指定。通过反复实验，约定所有分组聚类数目相等，即 k_1 = ... =k_t=... =k_l =k。

当 k=5,6,7,8,9,10 时，图 3 显示的是 KBFG-C4.5 模型对攻击类连接数据的检测率的变化：当 k=10 时 KBFG-C4.5 模型检测率 DR=0.9973 对比 k=9 时降低 0.003。

当 k=5,6,7,8,9,10 时，图 4 显示的是 KBFG-C4.5 模型对五种连接数据检测准确率的变化；当 k=10 时 R2L 的 DR=0.3309、Dos 的 DR=0.9993，对比 k=9 时分别降低约 0.03、提高约 0.03。Normal、U2R、Probe 的 DR 在 k=10、k= 9 均未有明显变化。

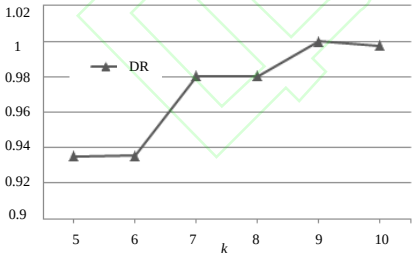

图 3 基于 KBFG-C4.5 的检测率

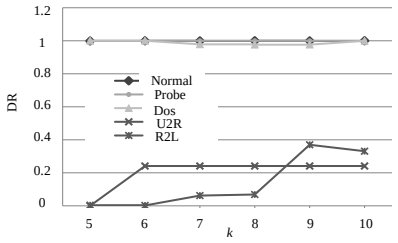

图 4 基于 KBFG-C4.5 的各类别的检测率

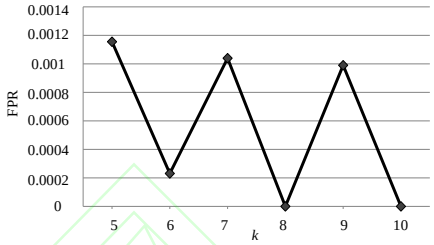

图 5 基于 KBFG-C4.5 的误检率

当 k=5,6,7,8,9,10 时，图 5 显示 KBFG-C4.5 模型对正常类 Normal 连接数据误检率的变化：当 k=10 时误检率 FPR=0。综合上面的实验讨论，本文取 k=10 为最终参数值。

当 k=10 时，KBFG 降维前后数据量对比如表 3 所示。训练集数据量约缩减 87.9%，测试集数据量约缩减 85.8%。KBFG 方法能够缩减网络数据的数据量，约在 85%~87%之间。

表 3 KBFG 降维前后数据量对比

	train/MB	test/MB
降维前	71.4	45
降维后	8.63	6.39

4.2 KBFG-C4.5 模型对新类型数据的检测能力测试

以上实验证实当 k=10 时 KBFG-C4.5 模型具有较好的检测效果。为了更进一步的证明 KBFG-C4.5 模型的有效性，从训练集与测试集抽取部分数据验证该方法能够检测出新攻击类型，其实验数据详情见表 4、5，在表中可以看到训练集与测试集中的攻击类型仅有“ipsweep”一类相同，其余均不同。

表 4 抽样训练数据分布

Normal	Dos		Probe		R2L		U2R	
normal	back	pod	ipsweep	satan	guess_passwd	warezclient	rootkit	buffer_overflow
2000	900	264	900	900	53	1020	10	30

表 5 抽样测试数据分布

Normal	Dos		Probe			R2L		U2R
normal	apache2	mailbomb	ipsweep	mscan	saint	named	snmpguess	Httptunnel
1000	200	56	16	200	200	17	200	50

此次抽样实验将实验数据分为五大类，而实验参数 k=10，本次的抽样实验结果如表 6 所示。由表 6 可知 Normal 的检测准确率 DR 达到 0.997，Dos、Probe、R2L、U2R 检测准确率达到 100%，说明 KBFG-C4.5 模型可以检测出新攻击类且正确分类。

4.3 实验对比

通过以上实验可知当 k=10 时，实验效果最佳。表 7 与表 8 是 k=10 与已有工作的对比试验。在表 7 与表 8 中的 C4.5 为未降维数据使用 C4.5 的实验结果。从表 7 可知 KBFG-C4.5 模型的 DR 为 99.72%，比 M-LHSVM-ELM 高出 4.56 个百分点、NFC 高出 4.53 个百分点、C4.5 高出 8.54 个百分点，而 KMs-C4.5 的 FPR=0，其它三种方法的 FPR 虽然很低，却仍然是高于 KMs-C4.5。

从表 8 可知，各类数据的检测率对比，KBFG-C4.5 模型对于 Normal、Dos、U2R、Probe 的检测率最优，分别为 100%、99.93%、24.12%、100%，均高于其它三种方法。因此通过以上分析可知 KBFG-C4.5 模型无论在检测率方面，还是对数据各类的检测率均优于其他方法，所以本文提出的 KBFG-C4.5 模型，由于使用了分组聚类思想，能够在入侵检测方面取得较好的效果。

表 6 抽样实验混淆矩阵

		预测				
		Normal	Dos	Probe	R2L	U2R
原始	Normal	997	0	0	0	3
	Dos	0	256	0	0	0
	Probe	0	0	416	0	0
	R2L	0	0	0	217	0
	U2R	0	0	0	0	50
	DR	0.997	1	1	1	1

表 7 与已有工作检测率对比

	M-LHSVM-ELM[14]	NFC[16]	KBFG-C4.5	C4.5
DR(%)	95.17	95.2	**99.73**	91.19
FPR(%)	1.87	1.9	0	0.5

表 8 与已有工作各类检测率对比 /%

方法	Normal	Dos	U2R	R2L	Probe
M-LHSVM-ELM[14]	98.13	99.54	21.93	31.39	87.22
NFC[16]	98.2	99.5	14.1	31.5	84.1
KMs-C4.5	**100**	**99.93**	**24.12**	**33.09**	**100**
C4.5	99.49	97.31	3.95	5.84	74.7

5 结束语

本文提出的基于 KBFG 算法的 KBFG-C4.5 模型对网络数据进行检测分析，基于分组聚类思想的 KBFG 算法弱化了 K-means 算法初始聚类中心选择对异常检测的影响，同时极大限度的降低了数据的维度，同时提高了入侵检测率，其缺点对于 U2R、R2L 的检测率没有做到很大的提高，下一步工作将讨论多种特征分组策略对入侵检测模型的影响，并构建一个更合适的决策函数，提高 U2R、R2L 的检测率。

参考文献

[1] Zhang C, Zhang G, Sun S. A mixed unsupervised clustering based intrusion detection model[C]// in Proc. 3rd International Conference on Genetic and Evolutionary Computing, USA: IEEE, 2009: 426–428.

[2] Meng Jianliang, Shang Haikun, Bian Ling. The application on intrusion detection based on k-means cluster algorithm[C]// International Forum on Information Technology and Application, USA: IEEE, 2009: 150-152.

[3] Pathak V, Anathanarayana V S. A novel Multi-Threaded K-Means clustering approach for intrusion detection[C]// Proceedings of 2012 IEEE 3rd International Conference on Software Engineering and Service Science, USA:IEEE, 2012: 757-760.

[4] 罗敏，王丽娜，张焕国. 基于无监督聚类的入侵检测方法[J]. 电子学报, 2003, 31(11): 1713-1716.

[5] 夏战国，万玲，蔡世玉，等. 一种面向入侵检测的半监督聚类算法[J]. 山东大学学报(工业版)，2012，42(6): 1-7.

[6] 陈光平，王文鹏，黄俊. 一种改进初始聚类中心选择的 K-means 算法[J]. 小型微型计算机系统，2012，33(6): 1320-1323.

[7] 肖苗苗，魏本征，尹义龙. 基于 BFOA 和 K-means 的复合入侵检测算法[J]. 山东大学学报(工业版)，2018，48(3): 115-119.

[8] 薛卫，杨荣丽，赵南，等. 空间密度相似性度量 K-means 算法[J]. 小型微型计算机系统，2018,39(1): 53-57.

[9] Gaddam S. R., Phoha V. V. , Balagani K. S. . K-Means+ID3: A novel method for supervised anomaly detection by cascading K-Means clustering and ID3 decision tree learning methods[J]. IEEE Trans. Knowl. Data Eng. 2007, 19(3): 345-354.

[10] Wang Huaibin, Yang Hongliang, Xu Zhijian, et al. A clustering algorithm use SOM and K-Means in intrusion detection[C]// Proceedings of the International Conference on E-Business and E-Government, USA: IEEE, 2010: 1281-1284.

[11] Muda Z, Yassin W, Sulaiman M N, et al. Intrusion detection based on K-Means clustering and Naïve Bayes classification[C]// International Conference on Information Technology in Asia, USA: IEEE, 2011: 1-6.

[12] Yassin W, Udzir N, Muda Z, et al. Anomaly-based intrusion detection through K-means clustering and naives bayes classification[C]// Proceedings of the 4 th International Conference on Computing and Informatics,Sarawak, Malaysia: Universiti Utara Malaysia, 2013: 298-303.

[13] Fachkha C, Harb E B, Debbabi M. Inferring distributed reflection denial of service attacks from darknet[J]. Computer Communications, 2015, 62 (1/2): 59-71.

[14] Chandrasekhar A, Raghuveer K. Intrusion detection technique by using k-means, fuzzy neural network and SVM classifiers[C]// International Conference on

Computer Communication and Informatics, USA: IEEE, 2013 :1-7.
[15] Yaseen W A, Othman Z, Nazri M. Multi-Level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system[J]. Expert Systems with Applications, 2017, 6(7): 296-303.
[16] He Liang. An improved intrusion detection based on neural network and fuzzy algorithm[J]. Journal of Networks, 2014, 9(5): 1274–1280.
[17] 武宇文，刘宏，查红彬. 基于特征分组加权聚类的表情识别[J]. 计算机辅助设计与图形学报，2005，17(11): 2394-2401.
[18] Quinlan JR. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.
[19] Quinlan J R. C4.5: programs for machine learning. Morgan Kaufmann Publisher, San Mateo, CA, 1993: 27-48.
[20] Hettich S, Bay S D. KDD cup 1999 data[EB/OL]. 1999. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[21] Bouzida Y, Cuppens F, Boulahia N C, et al. Efficient intrusion detection using principal component analysis[C]// In 3eme Conference surla Securite et Architectures Reseaux (SAR), La Londe: France, 2004.