doi :10.3969/j.issn.1671-1122.2021.08.001

基于流量特征分类的异常 IP 识别系统的设计与实现

文伟平，胡叶舟，赵国梁，陈夏润
（北京大学软件与微电子学院，北京 100080）

摘　要：异常 IP 识别是追踪恶意主机的重要方式，是网络安全研究的热点之一。当前应用机器学习技术进行异常 IP 识别多依赖整体网络流量，在单台服务器流量下会失效，且面临标记数据成本高昂问题。针对上述问题，文章把聚类算法和遗传算法应用到对端异常 IP 主机的识别与分类技术中，利用网络流量的多维特征和单台主机上可检测的 IP 地址特征数据，使用无监督学习和半监督学习相结合的方法，实现对端异常 IP 的识别、检测，并且将方法实现为异常 IP 识别系统。系统在实验中能实现对 UNSW-NB15 数据集 9 种不同类型恶意 IP 的识别，识别精度最高可以达到 98.84%。文章方法对恶意 IP 分类工作十分有效，并且可以识别未知类型的恶意 IP，具有广泛的适用性和健壮性，已应用在国家某网络安全中心的流量识别系统中。

关键词：恶意主机；分类算法；主机识别；权重向量

中图分类号：TP309　**文献标志码：**A　**文章编号：**1671-1122（2021）08-0001-09

中文引用格式：文伟平，胡叶舟，赵国梁，等. 基于流量特征分类的异常 IP 识别系统的设计与实现 [J]. 信息网络安全，2021，21（8）：1-9.

英文引用格式：WEN Weiping, HU Yezhou, ZHAO Guoliang, et al. Design and Implementation of an Abnormal IP Identification System Based on Traffic Feature Classification[J]. Netinfo Security, 2021, 21(8): 1-9.

Design and Implementation of an Abnormal IP Identification System Based on Traffic Feature Classification

WEN Weiping, HU Yezhou, ZHAO Guoliang, CHEN Xiarun
(*School of Software and Microelectronics*, *Peking University*, *Beijing* 100080, *China*)

Abstract: Anomalous IP identification is an important way to track malicious hosts, and is one of the hot spots in network security research. Current applications of machine learning techniques for anomalous IP identification mostly rely on overall network traffic, which will fail under single server traffic and face the problem of high cost of labeled data.

收稿日期：2021-04-12
基金项目：国家自然科学基金 [61872011]
作者简介：文伟平（1976—），男，湖南，教授，博士，主要研究方向为网络攻击与防范、软件安全漏洞分析、恶意代码研究、信息系统逆向工程和可信计算技术；胡叶舟（1995—），男，河南，硕士研究生，主要研究方向为异常网络流量识别、区块链安全；赵国梁（1991—），男，山东，硕士研究生，主要研究方向为恶意代码研究、异常网络流量识别；陈夏润（1997—）男，江西，硕士研究生，主要研究方向为软件安全漏洞分析、恶意代码研究。
通信作者：文伟平 weipingwen@ss.pku.edu.cn

To address the above problems, the paper applies clustering algorithm and genetic algorithm to the identification and classification technology of end-to-end abnormal IP hosts, using the multidimensional features of network traffic and IP address feature data detectable on a single host, using a combination of unsupervised learning and semi-supervised learning to achieve the identification and detection of end-to-end abnormal IP, and implements the method as an abnormal IP identification system. The system can achieve the identification of 9 different types of malicious IP in the UNSW-NB15 dataset in the experiment, and the recognition accuracy can reach up to 98.84%. The article method is very effective for malicious IP classification work and can identify unknown types of malicious IP with wide applicability and robustness, and has been applied in the traffic identification system of a national network security center.

Key words: malicious hosts; classification algorithm; host identification; weight vector

0 引言

随着互联网和各种网络应用的迅速发展，网络的规模越来越大，传播的数据越来越多，网络与信息安全逐渐影响到国家安全。随着5G时代的到来，网络安全问题更加严重。据统计2020年上半年我国境内感染计算机恶意程序的主机数量约304万台，同比增长25.7%。恶意主机广泛分布在世界的隐秘角落。对网络安全研究者来说，找到网络上存在的恶意主机并及时地进行处置成为一个迫在眉睫的任务。

异常流量检测是发现恶意主机的有效方式。目前，中外研究人员关于异常流量检测的研究非常广泛。LAKHINA[1]等人在2005年首先使用无监督学习对流量进行自动分类，他们对Abilene和Geant骨干网络流量数据进行分析，成功检测到了网络中发生的异常，并且使用聚类分析发现了以往未被发现的异常。LEE[2]等人提出了一种评定IP地址重要程度的方法。但是文献[1]和文献[2]使用的都是网络中流通的整体流量，单纯利用单台服务器上的流量数据时，其方法会失效。

在流经单台主机的流量上进行对端IP识别和分类具有实际应用意义。通过对流经单台服务器的网络流量进行整理和分析，将所有对端IP进行分类，可以有效识别提供不同服务的对端IP，可以识别出与服务器发生通讯的异常IP，定位恶意木马、病毒产生的流量，提高企事业单位的安全防护和重点服务器的安全；利用对端IP分类，对网络中出现的IP流量进行统计整理，可以为网络管理及安全响应提供IP地址流量行为的背景信息。

在恶意IP和流量的研究方面，获取流量数据相对容易，但是识别并标记样本的代价高昂，所以使用少量的标记数据来指导大量未标记数据进行半监督学习的技术得到了广泛关注[3]。

针对上述提到的对端异常IP识别与分类和标记流量数据成本过高的问题，本文把K-means聚类和层次聚类相关的算法应用到对端异常IP主机的识别与分类技术中。利用已标记样本的分类特点，识别样本中对分类影响较大的特征；对样本的各个特征属性进行重新赋值，以获取更合理的含有权重意义的特征，形成权重向量；然后，利用权重向量对未知样本的特征属性重新赋值，并进行层次聚类；最后利用已经标记的IP样本，实现了对大量未标记样本的分类和识别。

本文根据UNSW-NB15数据集进行异常IP识别测试，实验结果表明，本文方法可以实现恶意IP分类，并且可以识别未知类型的恶意IP，具有广泛的适用性和健壮性。通过对特征向量中的特征属性重新赋权，提高了分类的精度；对提高包含标签的聚类算法准确性具有参考意义。本文方案已被应用在国家某网络安全中心的流量识别系统中，通过分析对端主机特征，发现了多次国外黑客组织对我国重点服务器的恶意攻击，系统协助相关单位进行了前期的筛查和定位工作。

1 研究现状

研究显示，根据统计对象的不同，流量特征主要分为3级[4]：IP地址级、报文级、网络流级。

报文级流量的主要特征包含报文中的负载内容、通讯端口号等。MOORE[5]等人提出了一种对网络流量中的数据流进行建模，以此来识别网络中运行的服务的方法。DPI（深度报文检测）利用流量中携带的报文负载对流量进行识别和分类。

流级网络特征主要体现在报文头部信息，基于流级特征分类网络流量，主要分为直接利用流级特征和基于流级特征进行数据挖掘两种方法。SUH[6]等人用状态机的方式来对网络通讯过程进行表示，通过状态机来体现流级特征发生的不同变化。LI[7]等人在汇总他人研究的基础上，提炼出249种TCP流特征，使用朴素贝叶斯的分类方法尝试对网络流进行分类。

IP地址级流量特征通过流经特定IP地址的所有流量数据来计算统计特征。高骥翔[8]针对网络中存在的NAT网络，提出了一种使用IP地址级特征进行识别的方法。柳斌[9]等人将类似的方法应用于P2P流识别，能够正确识别超过95%的P2P流量。陈怡然[10]通过提取网络协议、上行网络流速、下行网络流速、端口号等特征，运用有监督的机器学习算法将主机进行分类，取得了很好的效果。黄思逸[4]将动物行为学模型引入IP地址流量行为模型，并对网络中所有的IP进行归纳、分类和解读。薛丽慧[11]在2019年提出了基于随机森林的恶意IP分类算法。赵艺宾[12]收集了大量APT组织中披露的恶意软件流量数据，通过对其中产生的流量特征进行统计和分析，利用深度学习中时序分析的方法，实现了恶意软件流量的检测。王勇[13]等人通过实验，提出了一种基于深度卷积神经网络的网络流量分类方法，在网络流量分类中有较好效果。

在异常流量检测方面，获取流量数据容易，但标记数据代价高昂，所以使用少量的标记数据来指导大量未标记数据进行半监督学习的技术得到了广泛关注[3]。半监督学习是监督学习与无监督学习结合的一种学习方法，所使用的数据集中既包含大量未标记数据，也包含一部分标记数据。聚类、分类和回归是3种主要的半监督学习算法。半监督聚类同时利用标记数据和无标记数据，综合了有监督学习和无监督学习的优势，改善了聚类的效果，这是基于机器学习的异常检测的一个重要方向[14]。K-means是一种半监督聚类技术，通过计算样本之间的距离将数据集中的样本分类成若干个不相交的簇，使用标记样本对簇进行标记和分类。但K-means算法容易陷入局部最优。为了解决局部最优的问题，GU[15]等人对K-means算法进行了改进，提出了一种半监督加权K-means方法，通过基于密度的初始聚类中心的选择，较好地解决了局部最优问题。

通过聚类算法，可以得到一些显著的离群点（离群点又称孤立点，其分布状况与其他正常样本有很大差异）。通过提取流量数据中的特征，对其进行离群点检测，可以发现不同于正常请求的恶意网络攻击。AHMAD[16]等人做过数据流离群点检测的研究，但由于网络流量数据的标记成本高昂，获取大量标记的数据集较为困难，因此在网络流量异常检测方面有监督的深度检测算法应用很少。无监督的检测方法利用样本内部属性来识别离群点，然而分析出数据集中样本的内在属性难度很大，导致检测效果不佳。利用类似深度自编码器[17]的半监督学习方法，可以达到很好的检测效果。当有足够多的样本时，深度自编码器可以获得较高的准确率。

2 相关技术

本文提出的基于流量特征分类的异常IP识别方法使用了聚类算法和遗传算法，其中，恶意IP检测使用K-means聚类算法，将已经标记的样本根据特征进行聚类，并对各个簇进行识别；未知异常IP检测使用聚类算法识别未知类型的簇。此外，不同特征对分类结果的影响不同，本文采用遗传算法找到合适的权重向量，以衡量各个特征对于分类效果的影响。本节介绍使用到的聚类算法和遗传算法。

2.1 聚类算法

聚类是一种将数据集中的样本按照相似度划分成

多个不相交的簇的过程，聚类完成后，各个簇之间的数据样本相似度比较低，而每个簇内的数据样本有很高的相似度。聚类方法在各个领域都有着广泛的应用，如模式识别、图像分析、数据挖掘等。

本文中对对端IP的聚类方法主要用到的是基于划分的K-means聚类[18]和基于层次的AGNES聚类[19]算法。

2.1.1 基于划分的 K–means 聚类

对于数据集D中包含n个样本数据，通过指定簇数k，基于划分的聚类算法把样本数据划分成为k（$k \leqslant n$）个不相交的簇。这些簇形成一组对整体数据样本的分类描述，在同一个簇中的样本是“相似”的，而不同簇之中的样本是“不相似”的。

优化簇内变差是一个NP-困难问题，可以使用贪心算法对该问题求局部最优解。K-means就是一种常用的贪心方法，简单且常用。首先随机选择k个样本数据作为算法初始的聚类的簇中心；然后计算各个种子数据与各簇中心的距离，之后将样本分配给距离最近的聚类中心对应的簇；将样本分配给簇中心后，会引起当前簇的样本集合发生变化，因此需要对该簇的中心进行重新计算。不断重复这个过程，直到满足预先设定的终止条件。

2.1.2 基于层次的聚类

层次聚类将样本构建成一棵具有层次嵌套的树，根据层次结构的构建方向不同，层次聚类可以分为自顶向下的分裂层次聚类（Divisive Hierarchical Clustering）和自底向上的凝聚层次聚类（Agglomerative Hierarchical Clustering）两种构建方式。凝聚的层次聚类方法使用自底向上的策略。开始时，把每个样本作为一个簇，并且迭代地把相邻的簇合并成一个更大的簇，直到达到某个预设的终止条件。在合并过程中，依据样本属性的相似度综合考虑，找到最近的簇，并将它们合并成一个更大的簇。直到所有样本数据都合并到同一个簇中，或达到预先设定的终止条件。

本文使用的AGNES（AGglomerative NESting）算法就是一种凝聚层次聚类算法。AGNES算法将每个样本作为一个初始的聚类簇，在算法运行过程中，找到距离最近的两个簇进行合并，不断重复这个过程，直至满足终止条件。

2.2 遗传算法

遗传算法是诸多进化算法中的一种。模拟了自然界中生物自然淘汰的进化过程,学习不仅可以通过单个生物个体的适应来完成,还可以通过种群的进化来实现,并将其运用到计算机模型之中[20]。具体流程如图1所示。

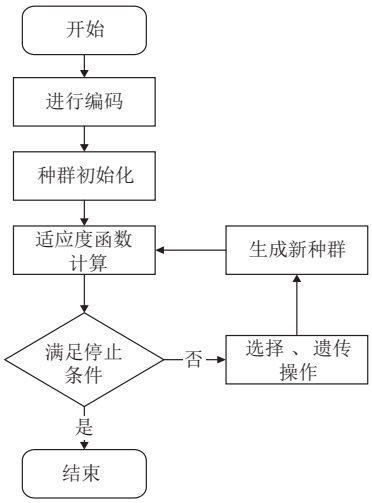

图 1 遗传算法的一般流程

1）编码。一个种群由经过基因编码的一定数目的个体组成，因此遗传算法首先是对原始输入进行编码，从而生成初始种群[21]。

2）适应度函数。在生成初始种群后，需根据适应度函数分别计算每个个体的适应度，以区分个体的优劣[21]。

3）选择策略。遗传算法在多次迭代中，从父代种群中选择哪些个体作为下一代个体的策略。

4）遗传操作。遗传操作分为交叉计算和变异计算。交叉计算是将两个父代种子进行组合，形成新的种子的过程。通过对比种群中所有个体的个体适应度，将适应度比较高的种子作为父代种子，通过这种方法可以使“优良”基因更容易被传递到下一代种子中。变异计算是父代两个种子交叉后产生的子代种子，其基因位可能以很小的概率发生转变，这种变异增强了种

群的多样性，将搜索范围跳出局部最优。

3 系统架构

本文提出的基于流量特征分类的异常IP识别方法主要包括特征工程、特征权重计算、恶意IP检测和未知异常IP检测4个步骤。特征工程主要用于提取对端IP的数据特征，并对部分特征进行重构；特征权重计算主要使用遗传算法，找到一个权重向量来衡量各个特征对分类效果的影响；恶意IP检测主要是使用聚类算法，将未知样本根据特征数据进行聚类，并对各个簇进行识别；未知异常IP检测主要是用于识别聚类算法中形成的未知类型的簇。

异常IP识别方法首先提取对端IP的数据特征，并对部分特征进行重构，然后使用遗传算法，找到一个权重向量来衡量各个特征对分类效果的影响，使用K-means算法对已标记的样本集进行聚类处理，聚类完成后，使用已知的标签标记每个簇，使用无监督的层次聚类方法，将未标记的样本集根据流量特征进行聚类训练，聚类完成后，使用K-means算法的聚类结果指导层次聚类结果的标记。最后，处理待检测样本得到特征数据，与各个簇的聚类中心进行比较，从而实现对异常IP的识别。

本文基于异常IP识别方法进行了系统实现，系统分为5个功能模块，分别是：流量预处理模块、特征提取模块、模型训练模块、恶意检测模块和检测报告生成模块，如图2所示。

流量预处理模块用来处理主机与各个对端IP产生的网络流量数据。特征提取模块用来提取和汇总各对端IP特征数据，主要的特征包括源端口、目的端口、发送包、接收包、服务协议、持续时间、状态、丢包率、TTL和流量速度等。模型训练模块分为K-means模型训练和层次聚类训练。恶意检测模块使用模型训练的结果，对未知IP进行检测和分类。检测结果主要分为正常IP、各类已知恶意IP和未知异常IP。报告生成模块用来为用户提供对端IP整体态势分析和恶意IP分析报告。

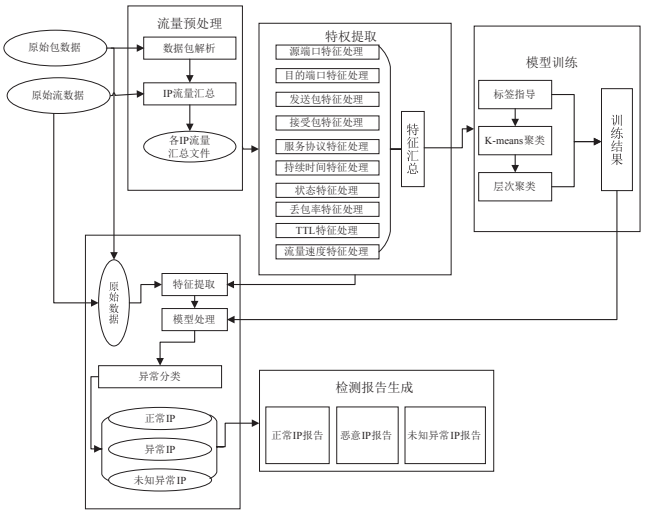

图 2 基于流量特征分类的异常 IP 识别系统

4 系统功能实现

4.1 流量预处理

流量预处理是为了处理和统计与本地主机发生网络通讯的各个对端IP产生的网络流量数据，并将其存储到各自的流量汇总文件中。当主机与网络上的远端计算机发生通讯时，统计IP数据包，得到数据包的基本信息，如源IP、目的IP、源端口号等。系统要做的工作就是将这些数据包根据对端IP进行分类，分类后可以得到与每个对端IP的通讯情况，将主机与每个对端IP发生的通讯数据包数据分别写入对应IP的数据包汇总文件中，为特征提取做准备，如图3所示。

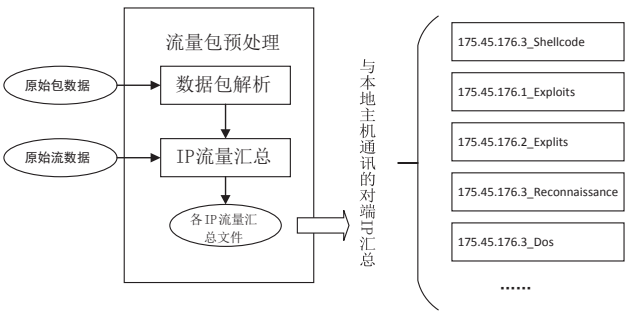

图 3 数据预处理示意图

4.2 特征提取

特征提取模块用来提取和汇总各对端IP特征数据。在源端口号、目的端口号方面，提取的特征为小于等于1024的端口熵值、大于1024的端口熵值和小于等于

1024的端口的占比。在发送包、接收包方面，提取的特征为包的平均大小、包大小的方差以及发送包速度。在协议方面，提取的特征为各个协议的占比向量。在流持续时长方面，提取的特征为均值和方差。在状态方面,提取的特征为各状态的占比向量。在丢包率方面，提取的特征为丢包率的均值和方差。在TTL方面，提取的特征为TTL的均值和方差。在流量速度方面，提取的特征为双向流量速度的均值和方差，训练特征统计如表1所示。

表 1 训练特征统计表

	信息熵	比例	均值	方差	速度	占比向量
源端口	√	√				
目的端口	√	√				
发送包			√	√	√	
接收包			√	√	√	
协议						√
持续时长			√	√		
状态						√
丢包率			√	√		
TTL			√	√		
流量速度			√	√		

4.3 模型训练

模型训练模块使用特征提取模块提取的特征，对样本数据进行分类，并根据已有的标签，将各个簇进行标记。模型训练模块分为K-means模型训练和层次聚类训练。主体思路是首先将有标签的数据注入K-means模型进行无监督聚类训练，聚类完成后根据已有标签的样本对各个簇进行识别并标记；之后把无标签的数据，注入层次聚类进行无监督聚类，聚类完成后，使用K-means聚类的各个簇依据相似度来标注层次聚类的聚类结果，如果没有对应的相似簇，则归为“未知异常类”，如图4所示。

模型训练模块利用提取的特征进行训练，过程主要有四步，计算不同特征的特征权重向量；给定的标记数据进行K-means聚类；将无标记数据进行层次聚类后；利用K-means聚类的结果进行分类。

4.3.1 计算特征权重向量

为了避免聚类时发生不同标签的样本被聚类到同一个簇的问题，首先要找到一个权重向量，当样本的特征值按照权重重新赋值时，标签相同的样本聚类到同一个簇。

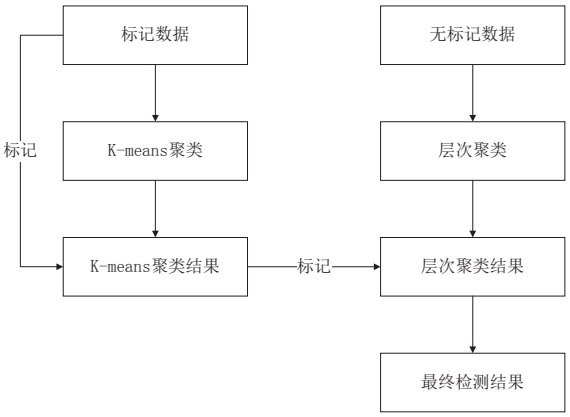

图 4 模型训练模块示意图

本文通过遗传算法获得不同属性样本的权重向量。编码方面，将权重向量$\boldsymbol{w}$采用实数编码的方式编码为$wi(a_1,a_2,\cdots,a_d)$，$a_i\in[0,10]$。采用随机生成的方法初始化种群为$P=\{w_1,w_2,\cdots,w_{count}\}$。选择距离自己的聚类中心最近的样本所占的比例（后面称为覆盖率）作为适应度函数，表示为$F(w)=\frac{Count_{correct}}{Count_{t_x}}$。使用最常用的正比选择策略，表示为$P_i=\frac{F_k}{\sum_{k=1}^{size}F_k}$。定义交叉计算为子代权重向量中的每个权重，都是随机从两个父代权重的相应位置随机取得的。即$w_{child}=(random(a_1^{parent1},a_1^{parent2}),\cdots,random(a_d^{parent1},a_d^{parent2}))$。突变操作为在种子权重向量中随机挑选一个权重，将其赋值为一个0到1之间的随机值。通过设置迭代最大次数*MaxLoop*的方式来使遗传算法停止。

算法1最优权重向量$\boldsymbol{w}$生成遗传算法

Input: 带标签的 IP 样本列表 SampleList

Output: 最优化的权重向量 $\boldsymbol{w}$

Begin

1）按照前述的编码方法，随机生成 *count* 个种子权重，作为初始种群 *P*。

2）对 *P* 中的每个个体进行适应度计算，得到 Flist。

3）使用正交选择策略，按照概率 P_i 选择 *count* 个候选个体 candidateList。

4）将 candidateList 中的种子，两两随机交叉运算，得到包

含 *count* 个新种子的种群。

5）对 *P* 中的每个个体进行适应度计算，得到 Flist。

6）执行 *MaxLoop* 次迭代。

输出最后种群中适应度最高的 ***w***。

End

4.3.2 标签样本 K–means 聚类

获取最优的权重向量 ***w*** 之后，进行K-means聚类，本文中初始化的 *k* 个样本选用加权后 *k* 个带标签的样本中心，作为 *k* 个初始化的簇中心，如算法2所示：

算法2 改进的K-means聚类算法

Input: 样本中不同的种类数目 *k*，包含 *n* 个样本的数据集 *X*
Output: *k* 个簇的集合
Begin
将 *k* 个带标签的样本中心作为 *k* 个初始簇中心。
使用最优权重向量 ***w***，对 *X* 中的 *n* 个样本重新赋权，得到新的样本集 *D*
repeat
根据簇中对象的均值，将每个对象分配到最相似的簇。
更新簇均值，即重新计算每个簇中对象的均值。
until 不再发生变化
End

4.3.3 无标签样本层次聚类

首先使用遗传算法的结果权重向量 ***w*** 修正原始特征向量。赋权后，使用层次聚类算法AGNES算法进行聚类。通过AGNES聚类，得到了多个无标记的簇，如算法3所示。

算法3 AGNES聚类算法

Input: 包含 *n* 个样本的数据集 *X*
Output: 满足终止条件的若干个簇
Begin
repeat
分别计算两个聚类簇中心之间的距离，获取两个相距最近的簇;
把两个距离最近的簇合并为一个簇;
until 终止条件得到满足
End

终止条件说明：为保证聚类形成的簇数量合适，簇内类型清晰，将层次聚类的终止条件设置为聚类的结果使得簇间聚类最大，而簇内距离最小。

4.3.4 聚类簇标记

K-means聚类后的簇是有标签的，而层次聚类后的簇是没有标签的，使用K-means聚类得到的簇来标记未标记的簇的方法是将各个层次聚类簇的中心与K-means聚类的中心比较，将层次聚类簇的中心标记为距离最近的K-means聚类簇的标签；如果层次聚类簇的中心与所有K-means聚类的中心的距离超过最大阈值，则标记为未知异常类。

4.4 恶意检测

恶意检测模块使用模型训练的结果，对新输入的未知IP样本进行检测和分类。检测结果主要分为正常IP、恶意IP和未知异常IP。恶意检测模块将提取到的特征集合与模型训练得到的各个聚类中心进行比较，将其归入最近的簇中。如果该未知IP与最近的聚类中心的距离超过阈值，则归入未知异常类，如图5所示。

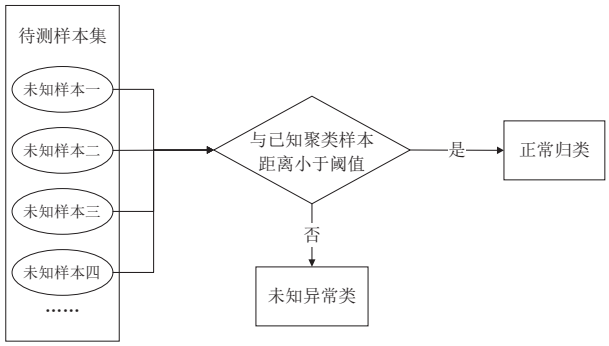

图 5 恶意检测模块示意图

5 实验及分析

为了验证模型及方法的有效性，选取UNSW-NB15数据集作为测试对象。选取2015年1月22日的16个小时和2015年2月17日的15个小时，总计100GB的实验数据作为实验样本，运用本文提出的系统进行实验。实验平台使用AMD Ryzen5、8核、2.00GHz主频CPU，8GB内存，操作系统为Windows 10的PC机。

1）恶意IP分类检测有效性实验

本文使用遗传算法计算特征权重向量，通过获取的权重向量指导层次聚类，并使用有标签的样本簇为层次聚类结果进行分类识别，实验结果如图6所示。

图6中Precision为检测精确率，Recall为召回率，ACC为准确率。从实验结果可以看出，所有分类的ACC指

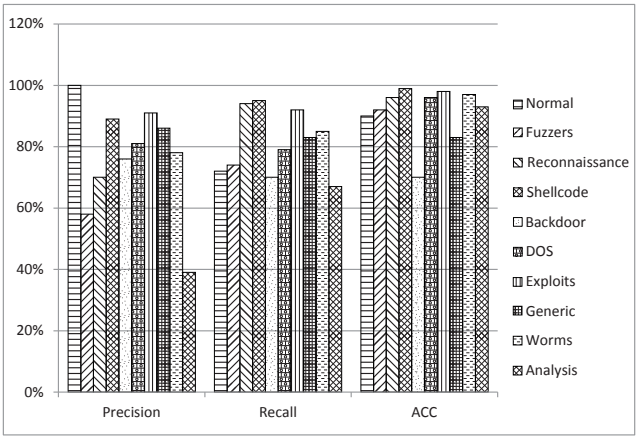

图 6 恶意 IP 分类检测有效性实验

标都处于较高水平。

本文给出不同种类的样本被分类到各个标签的数量统计，可以更直观精确地了解实验结果，如表2所示。（其中各个字母对应的标签分别为：Normal-N，Fuzzers -F，Reconnaissance -R，Shellcode -S，Backdoor -B，DOS -D，Exploits -E，Generic -G，Worms -W，Analysis -A）

表 2 分类实验中样本识别平均值

类别	N	F	R	S	B	D	E	G	W	A
N	**409.3**	**85.4**	**33.8**	**6.5**	**6.5**	**4.8**	**5.1**	**1.1**	**6**	**6.5**
F	0	118.7	26.1	0.9	2.9	0.3	0	0.1	2.5	5.5
R	0	0.9	153.1	0	2	0	0	0	0	4
S	0	0	0	154.6	2.8	0	0	0	0	1.6
B	0	0.2	0.6	8.8	102	4.8	0	0.6	7	19
D	0	0.2	0	0	5.9	126.6	5.2	17.2	0.6	4.3
E	0	0	0	0	2.6	5.7	149.4	0.1	0	2.2
G	0	1	0	0.5	2.9	11.7	1.3	132	4.8	5.8
W	0	2.1	0.9	0	2.3	0.3	0	1.1	90.6	7.7
A	0	1.1	2.1	0	5.4	2.7	0.4	0	4.7	33.6

实验结果表明，本文提出的方法，对ShellCode、Exploits和Worms类恶意IP识别效果非常好，可达到90%以上；Reconnaissance、Backdoor、DOS类恶意IP的识别效果次之，可达到70%以上；对Fuzzers和Analysis的分类效果较差。

2）未知异常IP检测有效性实验

本文在有标签的聚类实验中，故意将某种恶意对端IP剔除，而在无监督的层次聚类中加入此类恶意IP，来检查对未知异常类型的发现能力，实验结果如图7所示。

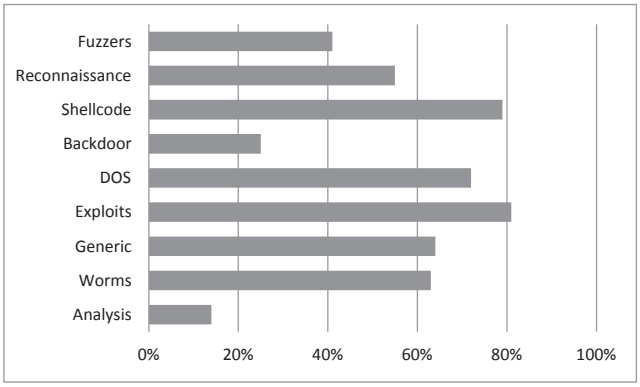

图 7 未知恶意 IP 分类检测有效性实验

实验证明，本文系统可以发现未知类型的恶意IP簇。在没有相应类型的样本参考时，本文方法可以识别80%左右的Exploits和ShellCode恶意IP；Worms、Generic、DOS、Reconnaissance的识别率可以达到50%以上。

6 结束语

本文设计并实现了基于流量特征分类的异常IP识别系统，首先对Pcap网络流量文件进行整合与特征提取，获得对端IP与本地主机发生通讯的流量中源端口、目的端口、发送包、接收包、服务协议、持续时间、状态、丢包率、TTL和流量速度的相关特征。然后利用有标签的样本数据，针对各个特征的重要程度计算权重向量。进一步利用特征重要性的权重向量指导无标签样本的聚类工作，最后利用有标签的样本簇对无标签的聚类簇进行标记，完成各类恶意IP的识别工作。

实验结果表明，本文提出的基于聚类的检测方法，对于ShellCode、Exploits和Worms类型的恶意主机的检测效果最佳，Reconnaissance、Backdoor、DOS和Worms类型的恶意主机的检测效果次之，Fuzzers和Analysis类型主机检测效果最差，究其原因是由于前者呈现的特征相似度高，容易形成良好的聚类效果；而Fuzzers和Analysis类型恶意主机攻击方式多种多样，不能形成明显的聚类导致的。

本文方案已经应用在国家某网络安全中心的流量

识别系统中，发现了多次国外黑客组织对我国重点服务器的恶意攻击。未来方案将研究通讯内容与流量特征相结合的方式进行流量识别，并在只含有少量标记或无标记流量分类和识别方面进行更深入研究。

参考文献：

[1] LAKHINA A, CROVELLA M, DIOT C. Mining Anomalies Using Traffic Feature Distributions[J]. ACM SIGCOMM Computer Communication Review, 2005, 35(4): 217–228.

[2] LEE D J, BROWNLEE N. A Methodology for Finding Significant Network Hosts[C].//IEEE. 32nd IEEE Conference on Local Computer Networks (LCN 2007). October 15–18, 2007, Dublin, Ireland. Piscataway: IEEE, 2007: 981–988.

[3] ZIMBA A, CHEN Hongsong, WANG Zhaoshun, et al. Modeling and Detection of the Multi–stages of Advanced Persistent Threats Attacks Based on Semi–supervised Learning and Complex Networks Characteristics[J]. Future Generation Computer Systems, 2020, 106 (5) : 501–517.

[4] HUANG Siyi. Ip Address Characterizing Based on Netflow[D]. Nanjing: Southeast University, 2017.
黄思逸 . 基于流记录的 IP 地址角色挖掘 [D]. 南京：东南大学，2017.

[5] MOORE A, ZUEV D, CROGAN M. Discriminators for Use in Flow–based Classification[R]. London: Queen Mary University of London, RR–05–13, 2013.

[6] SUH K, FIGUEIREDO D R, KUROSE J F, et al. Characterizing and Detecting Skype–relayed Traffic[C]//INFOCOM. The 25th Conference on Computer Communications. April 23–29, 2006, Barcelona, Spain. New York: IEEE, 2006: 2706–2717.

[7] LI Wei, CANINI M, MOORE A W, et al. Efficient Application Identification and the Temporal and Spatial Stability of Classification Schema[J]. Computer Networks, 2009, 53(6): 790–809.

[8] GAO Jixiang. NAT Recognition Method Based on Network Traffic Features[D]. Chengdu: University of Electronic Science and Technology of China, 2012.
高骥翔 . 基于网络流量特征的 NAT 识别方法 [D]. 成都：电子科技大学，2012.

[9] LIU Bin, LI Zhitang, LI Jia. A New Method on P2P Traffic Identification Based on Flow[J]. Journal of Xiamen University(Natural Science), 2007, 2046(2): 132–135.
柳斌，李之棠，李佳 . 一种基于流特征的 P2P 流量实时识别方法 [J]. 厦门大学学报 (自然科学版) . 2007，2046 (2): 132–135.

[10] CHEN Yiran. Study of the Host Behavior Classification Method Based on The Network Features of Flow and Connection[D]. Chengdu: University of Electronic Science and Technology of China, 2016.
陈怡然 . 基于网络流和连接特征的端主机分类 [D]. 成都：电子科技大学，2016.

[11] XUE Lihui. Research on Malicious IP Classification Algorithm Based on Big Data Platform[D]. Beijing: Beijing Jiaotong University, 2019.
薛丽慧 . 基于大数据平台的恶意 IP 分类算法研究 [D]. 北京：北京交通大学，2019.

[12] ZHAO Yibin. Research on APT Malware Traffic Detection Method Based on Association Rules and Timing[D]. Chengdu: University of Electronic Science and Technology of China, 2020.
赵艺宾 . 关联规则与时序特征结合的 APT 恶意软件流量检测方法研究 [D]. 成都：电子科技大学，2020.

[13] WANG Yong, ZHOU Huiyi, FENG Hao, et al. Network Traffic Classification Method Basing on CNN[J]. Journal on Communications, 2018, 39(1): 14–23.
王勇，周慧怡，俸皓，等 . 基于深度卷积神经网络的网络流量分类方法 [J]. 通信学报，2018，39 (1): 14–23.

[14] IDHAMMAD M, AFDEL K, BELOUCH M. Semi–supervised Machine Learning Approach for DDoS Detection[J]. Applied Intelligence, 2018, 48(10): 3193–3208.

[15] GU Yonghao, LI Kaiyue, GUO Zhenyang, et al. Semi–supervised K–means DDoS Detection Method Using Hybrid Feature Selection Algorithm[J]. IEEE Access, 2019, 2019(7): 64351–64365.

[16] AHMAD S, LAVIN A, PURDY S, et al. Unsupervised Real–time Anomaly Detection for Streaming Data[J]. Neurocomputing, 2017, 262(1): 134–147.

[17] XIAO Yawen, WU Jun, LIN Zongli, et al. A Semi–supervised Deep Learning Method Based on Stacked Sparse Auto–encoder for Cancer Prediction Using RNA–seq Data[J]. Computer Methods and Programs in Biomedicine, 2018, 166(11): 99–105.

[18] LU Yi, LU Shiyong, FOTOUHI F, et al. FGKA: A Fast Genetic K–means Clustering Algorithm[C].//ACM. Proceedings of the 2004 ACM Symposium on Applied Computing. March 14，2004, Nicosia Cyprus. New York: ACM, 2004: 622–623.

[19] LIU Anan, SU Yuting, NIE Weizhi, et al. Hierarchical Clustering Multi–task Learning for Joint Human Action Grouping and Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(1): 102–114.

[20] WANG Yinnian. The Research and Application of Genetic Algorithm[D]. Wuxi: Jiangnan University, 2009.
王银年 . 遗传算法的研究与应用 [D]. 无锡：江南大学，2009.

[21] WEI Zaoyu. Research on Blockchain Smart Contract Vulnerability Detection Based on Taint Analysis and Genetic Algorithm[D]. Beijing: Beijing University of Posts and Telecommunications, 2020.
韦早裕 . 基于污点分析和遗传算法的区块链智能合约漏洞检测技术研究 [D]. 北京：北京邮电大学，2020.