分类号：　　密级：
U D C ：　　学号：T2016018

南 昌 大 学 同 等 学 历 申 请 硕 士 学 位 研 究 生

学 位 论 文

基于 UEBA 的企业用户网络异常行为检测研究

Research on Network Abnormal Behavior Detection of Enterprise Users based on UEBA

苏文英

培养单位（院、系）：数学与计算机学院

指导教师姓名、职称：邱桃荣 教授

申请学位的学科门类：工 学

学科专业名称：计算机科学与技术

论文答辩日期： 2022 年 5 月 22 日

答辩委员会主席：甘登文

评阅人：

2022 年 5 月 22 日

摘　要

随着物联网、网格计算、移动互联网等信息化技术的发展，企业内部的管理、生产、经营等活动产生了大量且重要的各类数据（包括普通数据，也包括敏感数据和机密信息等）。针对企业内部核心数据面临日益严峻的安全问题，本文致力于解决企业用户网络异常行为检测问题，采用用户和实体行为分析方法进行企业员工日常访问行为记录的异常行为分析，构建了三个行为基线，提出基于K均值聚类-孤立森林的企业用户网络异常行为检测方法，通过在公开数据集上进行对比测试，验证了所提出的检测方法的有效性。论文主要研究内容如下。

（1）面向企业用户网络异常行为的检测需要，基于用户和实体行为分析方法进行数据处理，整合满足构建三个行为基线需要的数据。本文利用用户和实体行为分析方法对企业员工日常访问行为记录进行异常行为检测，识别异常的上网行为。首先对企业上网日志数据进行分析，通过时间序列处理法、统计值构造法和组合特征法构造90个特征作为基础数据。其次，通过方差选择方法和皮尔逊相关系数法对基础数据进行特征约简，整合得到能有效反映三个行为基线的17个特征。最后，基于17个基础特征，建立了企业上网行为基线、部门上网行为基线、个人上网行为基线。

（2）基于行为基线开展了企业用户网络异常行为检测方法研究，结合K均值聚类和孤立森林算法各自优势，提出使用K均值聚类-孤立森林算法的企业用户网络异常行为检测模型。该模型首先通过K-Means算法进行企业用户网络异常行为检测，设定异常行为值为1，正常行为值为0；其次，使用孤立森林算法对整体数据，不同的部门数据，不同的账号数据分别建模并获取行为异常分数；最后，将两种算法的异常检测结果进行结合分析。由此模型可以自动分析得到企业员工日常访问行为的异常分数。

通过计算模型得出的行为异常分数与人工标注的行为异常分数之间的均方根误差值和平均绝对误差值来分析模型的精准性，即误差值越小表示模型越精准。实验结果显示通过K均值聚类-孤立森林5折交叉验证法对整体数据建模，模型更优（在测试集上的均方根误差值为0.225347，平均绝对误差值为0.171624）。该方法和现有的孤立森林5折交叉验证算法相比，在数据集相同的情况下，均方

根误差值和平均绝对误差值更小，异常检测模型更精准。

关键词：用户和实体行为分析；异常行为检测；K 均值聚类；孤立森林

ABSTRACT

With the development of information technology such as the Internet of Things, grid computing, and mobile Internet, a large number of significant data (including ordinary data, sensitive data and confidential information, etc.) have been generated within the internal management, production and operation of enterprises. For enterprises, internal core data are facing increasingly severe security issues, this thesis devoted to solving the problem of the enterprise user network abnormal behavior detection, user and entity behavior analysis method is used for the enterprise users to access behavior daily record abnormal behavior analysis, and three behavioral baselines are built. This thesis proposed an enterprise user network abnormal behavior detection method based on K-Means Clustering - Isolated Forest algorithm, the effectiveness of the proposed method is verified by comparative tests on public data sets. The main research contents of this thesis are as follows.

(1) To meet the detection needs of abnormal network behaviors of the enterprise user, data processing is performed based on user and entity behavior analysis method, and data that meet the needs of three behavioral baselines are constructed. This thesis uses user and entity behavior analysis method to detect abnormal behaviors of enterprise users' daily access behavior records and identify abnormal online behaviors. Firstly, the log data of enterprises are analyzed, and 90 features are constructed as basic data through the time series processing method, the statistical value construction method and the combined feature method. Secondly, through the variance threshold and the pearson correlation coefficient are used to reduce the features of the basic data, and 17 features which can effectively reflect the three behavioral baselines are constructed Finally, based on 17 basic features, the baseline of enterprise online behavior, the baseline of departmental online behavior, and the baseline of personal online behavior are established.

(2) Based on the behavioral baselines, the research into abnormal network behavior detection methods of enterprise user are carried out. Combining the respective advantages of K-Means Clustering Algorithm with Isolated Forest Algorithm, a detection model of abnormal network behavior of enterprise user using

K-Means Clustering-Isolated Forest Algorithm is proposed. Firstly, the model detects the abnormal network behavior of enterprise user through K-Means Algorithm, and sets abnormal behavior value to 1 and normal behavior value to 0. Secondly, the Isolated Forest Algorithm is used to model the overall data, different departments' data, and different accounts' data respectively and obtain abnormal behavior scores. Finally, the abnormal detection results of the two algorithms are combined and analyzed. This model can automatically analyze and obtain the abnormal scores of enterprise users' daily access behaviors.

The accuracy of the model is analyzed by calculating the root mean square error value and mean absolute error value between the abnormal behavior scores are obtained by the model and the manually labeled abnormal behavior scores, that is, the smaller the error value is, the more accurate the model is. The experimental results show that the model is better by K-Means Clustering-Isolated Forest five-fold cross-validation (the root mean square error value on the test set is 0.225347, the mean absolute error value on the test set is 0.171624). Compared with the existing Isolated Forest five-fold cross-validation algorithm, this algorithm has a smaller root mean square error value and mean absolute error value, and a more accurate abnormal behavior detection model under the same data set.

Key Words: UEBA; Abnormal Behavior Detection; K-Means Clustering; Isolation Forest

目 录

第 1 章 绪论 1
1.1 研究背景和意义 1
1.2 国内外研究现状 2
1.2.1 网络异常行为检测国内外研究现状 2
1.2.2 UEBA 的国内外研究现状 3
1.3 本文主要工作 5
1.4 本论文结构 5
第 2 章 相关理论与技术概述 7
2.1 企业用户异常行为 7
2.1.1 企业用户异常行为概述 7
2.1.2 企业用户异常行为分类 7
2.1.3 企业用户网络异常行为检测 8
2.2 UEBA 10
2.2.1 UEBA 概述 10
2.2.2 UEBA 适用场景 10
2.2.3 UEBA 算法 11
2.3 K-Means 算法 12
2.3.1 聚类 12
2.3.2 算法概述 12
2.3.3 算法实现 13
2.3.4 算法优缺点 13
2.4 孤立森林 14
2.4.1 异常 14
2.4.2 算法概述 14
2.4.3 孤立树 16
2.4.4 异常检测 17

2.4.5 算法优缺点 18
2.5 本章小结 18
第 3 章 特征分析与行为基线构造 19
3.1 数据集介绍 19
3.2 面向企业用户网络异常行为检测需要的上网行为基线构造 21
3.3 特征构造 21
3.3.1 特征构造常见方法 22
3.3.2 企业用户上网行为的特征构造 22
3.4 特征选择 27
3.4.1 特征选择常见方法 27
3.4.2 方差选择 28
3.4.3 皮尔逊相关系数 28
3.5 本章小结 31
第 4 章 结合 K 均值聚类和孤立森林算法的异常行为检测模型 32
4.1 结合 K 均值聚类-孤立森林算法的异常行为检测模型图 32
4.2 基于 K-Means 聚类的异常行为检测 35
4.2.1 数据归一化 35
4.2.2 K 值确认 36
4.2.3 数据聚类 38
4.2.4 企业用户网络异常行为检测 38
4.3 基于孤立森林 5 折交叉验证的异常行为分数获取 39
4.3.1 5 折交叉验证概述 39
4.3.2 基于整体数据的企业用户网络异常行为检测模型 39
4.3.3 基于部门数据的企业用户网络异常行为检测模型 40
4.3.4 基于账号数据的企业用户网络异常行为检测模型 41
4.4 结合 K-Means 和孤立森林算法的异常行为结果融合分析 42
4.5 本章小结 42
第 5 章 实验与结果分析 43
5.1 模型评价指标 43

5.2 实验环境 43
5.3 不同异常行为检测模型的实验与结果分析 44
5.3.1 基于整体数据的异常行为检测模型实验与结果分析 44
5.3.2 基于部门数据的异常行为检测模型实验与结果分析 44
5.3.3 基于账号数据的异常行为检测模型实验及结果分析 46
5.4 两种算法构建的不同模型的对比实验与结果分析 47
5.5 本章小结 48
第 6 章 总结与展望 49
6.1 总结 49
6.2 展望 49
致 谢 51
参考文献 52
攻读学位期间的研究成果 55

第 1 章 绪论

1.1 研究背景和意义

随着物联网、网格计算、移动互联网等信息化技术的发展，给我们的生活带来了很大的便利，如：足不出户即可购买各种所需用品，即使相隔很远，也能通过一个电话或视频诉说彼此的思念等。但与此同时，大量数据的融合、分析与应用给用户带来前所未有的隐私泄露威胁。2020年4月中央电视台曝光了一起特大电信诈骗案：贾某通过网络发布贷款广告，收集了40多万条网贷用户的信息，并将这些信息贩卖给境外诈骗团伙，用来对这些用户实施精准诈骗，其中有1万多人被骗，被骗总额超两亿；2021年1月份新闻消息：李某某、范某某和其他30多位犯罪嫌疑人，贩卖6亿多条公民的个人信息，非法获得利益800多万元……网络安全已成为国家安全的重要组成部分，也是保障信息社会和信息技术可持续发展的核心基础。习近平总书记说过“没有网络安全就没有国家安全”。

与此同时，企业的信息化水平也在不断提升。原来企业内部的管理生产经营等活动都是通过纸质文件记录，现各企业基本都有 OA、客户信息管理、业务等系统，员工、客户的隐私信息以及工作商业的机密信息查找比较方便，且数据比较完整，所以企业的数据安全也是至关重要。经过各相关机构对各种网络安全事件的分析发现：影响最大、破坏力最强的安全威胁是来自可以访问企业敏感数据的内部员工。《2020 Securonix 内部威胁报告》中提到：一些准备离职的内部员工里面，有 80%的员工可能在离职的时候带走部分公司敏感数据，这些员工中涉及 60% 的内部网络安全和数据泄露事件。Fortinet 的《2019 年内部威胁报告》中提到：53%的网络安全专业人员确认在过去一年里，他们的组织遭受过内部的攻击。2019 年巴基斯坦人行贿可以进入 AT&T 公司系统的内部员工，让其在 AT&T 系统上安装恶意软件，使它们能够按要求自动解锁大量价格昂贵的 iPhone 手机，解锁后的手机可以在 AT&T 网络之外使用。为此，美国电信供应商 AT&T 公司损失了价值 2 亿美元的用户，涉及近 200 万部手机被解锁。2020 年 5 月新闻报道：银行内部员工丁某，将公民的身份证、电话号码、银行

卡及交易等信息，以每条 80-110 元的价格，卖给犯罪人员，涉及个人信息 50000 多条，该犯罪团伙诈骗金额超 2000 万元……类似安全事件数不胜数。

内部员工的安全威胁主要有 2 大类原因。一类是主观有意而为之的安全威胁，包括：员工对公司有诸多的意见，为了报复企业，对企业敏感数据进行主动攻击或恶意破坏系统等行为；或内部员工被人用金钱或权利等利益贿赂，对企业敏感数据进行窃取。另一类是客观无意识的安全威胁，包括：内部员工的电脑被外部攻击或中了木马病毒等原因频繁的访问企业的重要信息并将其泄露出去的行为；或员工虽然不是恶意破坏也没有被外部攻击，但其操作不规范也会导致企业的敏感数据泄露等问题。虽然现在绝大多数企业围绕敏感数据保护都出台了相关管理办法和操作行为准则，但仍常有企业涉密事件。所以，及时、准确、有效的识别或预测企业用户网络异常行为是非常重要且有意义的课题。

本文的主要研究内容是基于用户和实体行为分析（User and Entity Behavior Analytics，UEBA），对企业内部用户的网络操作行为进行定义，并通过机器学习算法构建行为异常检测模型，以确定哪些内部用户的行为是正常行为，哪些内部用户的行为是异常行为，从用户行为层面管理内部网络和数据的安全问题。该技术基于海量数据对内部用户的异常行为或内部威胁进行预测，主动出击，在数据泄漏之前进行阻止，并为安全分析人员提供可靠的依据。实现对企业数据资产使用情况的事前预测、事中阻断、事后溯源全闭环管控。

1.2 国内外研究现状

1.2.1 网络异常行为检测国内外研究现状

随着新的通信技术和服务的发展，以及互联网络设备、网络用户、服务和应用程序数量的增加，使得计算机网络变得越来越复杂，网络安全事件频繁发生。自 19 世纪早期以来，越来越多的研究人员致力于网络异常行为检测。

Swarnkar 和 Hubballi[1]通过使用基于载荷异常检测的一类朴素贝叶斯分类器，将短序列的频率信息与一类朴素贝叶斯分类器相结合，准确地检测出网络数据包中的可疑载荷内容。Wang 等人[2]基于增强特征的支持向量机(Support Vector Machines，SVM)创建了一种有效的入侵检测系统（Intrusion Detection System，IDS）框架，集成了 SVM 和对数边际密度比转换(Logarithm Marginal Ensity Ratios

Transformation，LMDRT)，将数据集转换为新数据集，新数据集被用来训练 SVM 分类器，以提高系统的检测能力。通过使用常用的 NSL-KDD 数据集（美国空军局域网上采集的流量进行特征提取得到的数据集的改进版）对该框架进行评估，获得较快的训练速度、较高的准确率。Subba 等人[3]通过人工神经网络（Artificial Neural Network，ANN）模型，引入智能代理来检测审计数据的异常性，实验表明该方法具有较高的性能和准确性。Saeed 等人[4]在物联网环境中使用循环神经网络(Recurrent Neural Network，RNN) 来实现智能安全架构，提出了一种多层有效的入侵检测方法和低功耗物联网系统的预防机制，通过在无线传感器节点的物联网系统中进行测试，验证了该方案对入侵检测的有效性。David 等人[5]提出了一种通过快速熵法和基于流的分析对拒绝服务攻击进行增强检测的方法。该方法首先将观察到的流聚合为单个流，然后计算每个连接的流量计数的快速熵，最后基于快速熵和流量计数的均值和标准差生成自适应阈值。宁亚飞等人[6]提出了通过时空卷积自编码方法进行异常行为检测，首先将连续三帧正常的视频灰度化，以便对输入图像进行时空块处理，然后将灰度化的图像输入到卷积自动编码器中进行训练，以获得视频中正常行为的时空特征，最后基于阈值规则并结合编码器的重构误差和相应的规则评分来判断异常行为。刘良鑫等人[7]通过 3D 双流卷积神经网络技术，对视频进行异常行为检测，首先对输入的视频进行预处理，利用 I3D（基于 Inception-V1 的 3D 卷积）提取视频文件，每 16 帧提取一次，然后处理视频特征并将数据进行保存，最后通过异常检测模型对已处理好的特征数据进行评估，并给出概率估计。

为了更好的检测网络异常行为，更主动和更准确地发现安全问题，UEBA 用户和实体行为分析技术被提出，它可以识别基于日志的行为异常。基于 UEBA 技术，安全专业团队可以通过识别的异常，主动管理网络信息安全，是对安全信息和事件管理的有效补充[8]。

1.2.2 UEBA 的国内外研究现状

2014 年美国 Gartner 研究机构发布了用户行为分析(User Behavior Analytics，UBA)市场定义，当时对 UBA 的技术定位是在信息窃取和利用窃取来的信息实现非法牟利上，该技术可以帮助企业或机构检测内部安全威胁、攻击和诈骗。但是后来随着数据泄露事件的增多，Gartner 把这部分从诈骗检测的技术中分离

出来，在 2015 年正式将 UBA 更改为用户和实体行为分析(UEBA)。并在第二年被列入十大安全技术，2017 年虽然没有入榜，但 2018 年以检测与响应项目中的一个分支重新登入榜单[9]。随着 UEBA 的发展，越来越多的学者研究基于 UEBA 的异常行为检测。

Martin Alejandro G 等人[10]提出利用 UEBA 来提高联邦身份管理(Federated Identity Management，FIM)解决方案的安全性。该方法基于用户或实体的静态和动态行为数据建立会话指纹，并应用异常检测技术来检测由于凭证盗窃或会话劫持而引起的冒充行为。Madhu Shashanka Niara 等人[11]提出了一种基于马氏距离的 UEBA 研究，采用基于奇异值分解(Singular Value Decomposition，SVD)的机器学习算法自动检测异常行为。这种潜在恶意活动的异常行为会向分析人员发出警报，并提供相关的背景信息，以便进一步调查和采取行动。美创[12]推出的新一代数据安全管理平台中，使用到了 UEBA 技术对安全事件进行分析和预警。 胡绍勇[13]提出了基于 UEBA 技术进行数据泄露风险的分析，介绍了观安信息技术股份有限公司的 UEBA 解决方案。UEBA 在数据泄漏分析上可以动态构建个人、部门、资产等多维度基线，相比传统基于固定规则的检测，UEBA 结合每个人的历史操作行为、登录 IP 和登录时间等再对当下操作行为做判断，大大提升了检测的准确度。同时该技术使用机器学习算法，大大减少规则制定的工作量，而且容易发现未知的数据泄露威胁。吴宏胜[14]将 UEBA 运用到了智慧政务系统中，通过抽象出各类实体，对用户进行画像分析，建立动态行为基线和数据标准，检测和抵御安全威胁与攻击，提升智慧政务系统的安全性。徐飞[15]将 UEBA 应用在智慧公路上，通过获取 ETC 收费日志、日常运维管理日志建立用户行为基线，将用户画像与行为基线做对比，检测攻击行为。该方法可以将智慧公路信息系统日志、身份验证日志、防御日志等多源数据进行融合，统一分析，使得分析更全面，准确性更高，同时减少人工核对单个报警的工作，提升工作效率。莫凡等人[16]提出 UEBA 在账号异常检测的应用，通过采集大量的用户行为数据，包括身份信息、行为数据、及其关联数据，通过核密度估计算法确认上班时间，建立行为基线。该研究使用孤立森林算法、一类支持向量机（One Class SVM）以及局部异常因子算法建立异常检测模型，计算异常得分，并通过人工核查，验证了异常检测的准确性。

随着数字化经济的发展，UEBA 会被广泛应用到各种安全产品中。

1.3 本文主要工作

首先，本文利用用户和实体行为分析（UEBA）方法对企业员工日常访问行为记录进行异常行为检测，识别企业用户异常的访问行为，即通过对数据集进行特征构造和特征选择确认得到用于企业用户网络异常行为检测需要的核心特征，依托这些核心特征建立上网行为基线。其次，本文结合K均值聚类算法和孤立森林算法对整体数据、不同的部门数据、不同的账号数据构建企业用户网络异常行为检测模型。最后，本文将所提出的企业用户网络异常行为检测模型与基于孤立森林算法构建的企业用户网络异常行为检测模型进行对比实验，在相同数据集上，验证了所提出的基于K均值聚类-孤立森林法构建的企业用户网络异常行为检测模型更精准。

本文的主要工作安排如下：

第一，对企业员工日常访问行为数据进行分析。通过时间序列处理法、统计值构造法和组合特征法构造扩展特征，通过方差选择方法和皮尔逊相关系数法选择最终需要的特征，依托这些特征分别建立企业上网行为基线、部门上网行为基线、个人上网行为基线。

第二，使用K均值聚类-孤立森林5折交叉验证法对整体数据，不同的部门数据，不同的账号数据分别构建企业用户网络异常行为检测模型。第一步，通过K-Means算法进行企业用户网络异常行为检测，并记录检测结果。首先对数据进行StandardScale归一化，然后通过肘部法则和轮廓系数确认K值，使用K-Means聚类算法对数据做聚类分析，最后计算每个数据点与其聚簇中心之间的距离，设置异常值的阈值，通过阈值来判断行为是否异常，异常行为值为1，正常行为值为0。第二步，使用孤立森林5折交叉验证法对整体数据，不同的部门数据，不同的账号数据分别建模，获取异常得分。第三步，将两种异常检测结果进行结合分析，得到最终的异常分数。

第三，使用均方根误差（Root Mean Square Error，RMSE）和平均绝对误差（Mean Absolute Error，MAE）分别对整体数据，不同的部门数据，不同的账号数据建立的企业用户网络异常行为检测模型进行模型评估。

1.4 本论文结构

本文使用用户和实体行为方法对企业用户网络异常行为检测进行研究，论文主要分为六个章节，每个章节的内容如下：

第1章为绪论，首先介绍本文的研究背景与研究意义，并对国内外相关的研究情况进行阐述，然后对本文研究工作进行梳理，最后介绍了论文的组织结构。

第2章为相关理论与技术概述，首先介绍了企业用户异常行为的概述、分类及检测方法，接着介绍了UEBA的概述、适用场景和算法，最后分别介绍K-Means算法和孤立森林算法。

第3章为特征分析与行为基线构造，首先介绍了数据的来源及每个字段的含义，然后介绍了面向企业用户网络异常行为检测需要的上网行为基线构造，接着介绍了特征构造常见的方法并通过时间序列处理法、统计值构造法和组合特征法构造上网行为特征，最后，介绍了特征选择常见方法并通过方差选择和皮尔逊相关系数选择法对基础数据进行特征约简，整合得到能有效反映三个行为基线的17个特征。

第4章为结合K均值聚类和孤立森林算法的异常行为检测模型，首先通过K均值聚类算法进行企业用户网络异常行为检测；然后通过孤立森林5折交叉验证法对企业整体数据、不同的部门数据、不同的账号数据分别构建企业用户网络异常行为检测模型，并计算行为异常分数。将两种算法的异常检测结果进行结合分析。由此模型可以自动分析得到企业员工日常访问行为的异常分数。

第5章为实验与结果分析，首先介绍了模型评价指标：均方根误差和平均绝对误差，然后介绍了实验环境，接着分别呈现了基于K均值聚类-孤立森林5折交叉验证和孤立森林5折交叉验证对整体数据、不同的部门数据、不同的账号数据建模的实验结果，最后对本文提出的K均值聚类-孤立森林与孤立森林算法在相同数据集上的实验结果进行对比分析。

第6章为总结与展望，主要对本文的研究工作进行总结，分析本文的贡献和不足之处，并提出未来的研究方向。

第 2 章 相关理论与技术概述

2.1 企业用户异常行为

2.1.1 企业用户异常行为概述

企业用户是指企业或单位为维持正常经营活动而聘用的员工。企业用户行为是指企业用户表现出来的各种外表活动，例如登入系统、查看数据、下载数据、分享数据等。企业用户异常行为是指企业用户对系统和数据的滥用行为，是与常规行为不同的用户行为，对企业与他人造成一定影响的行为。

2.1.2 企业用户异常行为分类

根据异常的性质进行异常分类可分为：点异常（Point Anomalies）、上下文异常（Contextual Anomalies）和集合异常（Collective Anomalies）。

点异常：是某一次行为与通常的行为模式的偏差，比如某人通常是9:00-17:00上网，突然某天是22:00上网，那么22:00就是点异常。

集体异常：是一组相似的数据实例相对于整个数据集出现异常的行为。例如：一个人无法登入系统不算异常，一个部门都无法登入系统才算是异常。

上下文异常：具有差异的异常，该异常指向基于上下文的知识，当缺少上下文信息时，可能无法识别这种异常。

根据因果关系可将异常分为：操作/配置错误/故障事件、合法但非正常使用、网络滥用异常/恶意攻击。

操作/配置错误/故障事件：它发生在网络系统中，主要由硬件故障、软件bug或人为错误引起。服务器崩溃、断电、配置错误、流量拥塞、非恶意的大文件传输、资源配置不足、或者由于施加速率限制或添加新设备而导致的网络行为的显著变化。

合法但非正常使用：短时间大量用户集中访问特定的网络资源时，就会导致服务器负载急剧增加。这通常是热事件相互反应的结果，但远远大于系统可以处理的负载。如，当竞赛结果在官网上公布时，或者当电子商务网站宣布大减价时，或者甚至是由于软件发布时，网络拥塞就会出现。虽然不是恶意行为，但如果没有足够的时间来响应并提供必要的资源来处理过载需求，这些事件可

能会使服务器宕机。

网络滥用异常/恶意攻击：在非授权的情况下，试图对信息进行存取、处理或者破坏以使系统不可靠、不可用的故意行为[17]。

2.1.3 企业用户网络异常行为检测

企业用户网络异常行为检测是根据企业用户平时的行为或资源平时使用的情况，来预测当前行为是否偏离了正常的行为基准。首先使用各类日志数据建立正常行为基线，然后通过计算实际目标行为模式与正常行为模式之间的偏差值识别事件的操作是正常还是异常。如果发现是异常行为系统将对其进行告警。异常检测的核心问题是正常使用行为基线的建立以及如何利用该基线对当前的系统或用户行为进行对比，判断当前行为与正常行为的偏离程度。由于任何不符合正常基线的行为都被认为是入侵行为，所以能够发现未知的攻击[18]。

网络异常行为检测的技术和方法主要有基于统计分析的网络异常行为检测、基于聚类的网络异常行为检测、基于有限状态机的网络异常行为检测、基于分类算法的网络异常行为检测、基于进化论的网络异常行为检测[19]。

基于统计分析的网络异常行为检测：这种方法通常是基于与训练数据相关的概率模型来跟踪网络行为。异常与网络数据的突然变化有关，大多数情况下，这些突变是通过建模硬阈值来检测的。统计技术的主要挑战是找到减少硬阈值引起的虚拟告警的方法。基于统计的方法检测网络异常的内在能力比其他任何方法都强，能够了解流量(网络系统)的预期行为，这些方法不需要任何关于系统的先验知识作为输入。但是某些类型的攻击可能是训练数据集的常规部分，可能被纳入正常行为中，导致被认为是正常行为。由于阈值的局限性和静态特性，在某些现实情况下阈值的使用可能不可靠。

基于聚类的网络异常行为检测：聚类分析的目的是将一组对象分成相似对象的类。这些类或组(称为集群)及其对象彼此相似(在某种程度上)，但与其他集群中的类或组不同。基于集群的流程可以适应变化，并帮助筛选出区分不同组的有用特征。聚类技术可以用于离群点检测，识别离群点太远的值，或者作为其他算法/方法的预处理步骤。

基于有限状态机的网络异常行为检测：有限状态机(Finite State Machine，FSM)，也称为有限自动机，是一种路由状态、转换和动作组成的数学行为模型，用于表示计算机问题或逻辑电路。它的每个状态都存储关于过去的信息，这些

信息是自进入状态以来从系统开始到现在发生的更改。这种类型的机器一次只能处于一种状态，转换指示状态变化，必须达到相应条件才能发生转换。有限状态机的动作是对必须在特定时间执行的活动的描述。此外，有限状态机具有强大的分析能力、很强的鲁棒性和灵活性。

基于分类算法的网络异常行为检测：分类是异常检测领域广泛应用的一种方法。基于分类进行网络异常行为检测的主要思想可以概括为两个步骤。首先，在训练阶段，使用标记的训练数据建立(学习)分类器。然后，使用该分类器将实例分类为正常或异常(测试阶段)。根据每个可用的标记数据进行训练，基于分类的异常检测技术可以是多类的或者单类。针对单类数据，则意味所有训练数据只有一个正常的类标签。针对多类数据，则训练实例有多个正常的类标签，在这种情况下，将构建一个分类器来区分普通类和不属于任何类的实例(异常)。常用的分类器有朴素贝叶斯、支持向量机、人工神经网络。朴素贝叶斯是一种简单的概率分类器，常用来解决网络入侵检测问题。它结合了先验信息和样本信息，并将其应用于统计推断中，用概率来表示各种形式的不确定性。它的原则建立在所有输入属性都有条件地相互独立的假设之上。支持向量机（Support Vector Machines,SVM）是一个有监督的学习概念，其特征是使用特征向量/核函数(如径向基核函数-RBF，Radial Basis Function)，同时具有稳健性和稀疏性，通过 SVM 得到的分类器具有良好的泛化能力。人工神经网络被认为是一种仿生模型，但它们在异常检测领域主要用作分类器。基于分类的方法简单、有效，而且测试和训练很灵活，对已确认的攻击检测率高。但是这种方法资源消耗高，没有相关的训练信息，无法检测到未知的异常。神经网络的使用可能导致过度拟合，在某些情况下，实时性能很难获得。

基于信息论的网络异常行为检测：这种方法主要依赖于指定流量特征的互信息或熵值演算，以便识别异常分布，该方法可扩展性强。由于它采用了时间序列的统计特性，因此这种方法可能会导致检测结果不准确，只有当数据集中存在显著的异常时，才有可能检测到异常，而且它很难将异常分数和案件联系起来。

基于进化论的网络异常行为检测：进化计算领域，也被称为生物启发计算，是一套受自然进化启发的智能算法，能够像生物有机体一样学习和适应。由于进化计算方法的智能算法可以像真实的生物体一样学习和适应，因此能够为许多复杂网络问题提供潜在的解决方案，但是适应度函数很难找到，选择最优参

数也非常困难，有时，用生物学方法来解决问题可能是一项复杂的任务。

2.2 UEBA

2.2.1 UEBA 概述

UEBA(User and Entity Behavior Analytics)称为用户和实体行为分析技术，主要是以用户和实体为对象，结合规则以及机器学习模型，对用户行为进行分析和异常检测，尽可能快速地感知内部用户的可疑非法行为。在安全领域已经有了较为深入的应用，在 2016 年 Gartner 就已经将其作为当年十大信息安全技术之一。

UEBA 相对来说具有洞察力和可扩展性，简单说 UEBA 是大数据驱动，且采用机器学习方法进行安全分析，能够检测高级、隐藏和内部威胁的行为分析技术，不需要使用签名或规则。该方法在杀伤链上能关联数据，能进行有针对性地发现异常行为。UEBA 方法中使用的分析技术包括机器学习、行为建模、分类、对等组分析、统计模型和图形分析。企业结合 UEBA 分析和人工设置的评分机制，对比有嫌疑用户的活动，最终实现对异常行为和威胁行为的检测。同时，UEBA 还包括进行威胁可视化分析，以可视化的方式进行跨越杀伤链分析。但 UEBA 在进行用户异常行为分析之前需要采集和处理多个数据源的大量行为类数据。由于这些广泛用户行为类数据源格式不同，在使用 UEBA 进行用户异常行为分析之前需要对这些数据进行处理，将这些可能是结构化也可能是非结构化的多源数据中提取能用于开展异常行为分析的有价值信息。因此，高质量多种数据源是 UEBA 的核心。

UEBA 检测到的是“异常”，异常是说和平时的行为不一样，但有时不一样不一定是威胁，例如双 11 活动，访问的数据量、购买量以及退货量会和平时不一样。异常表示需要引起关注，评估后给出威胁判断，威胁指标则代表了关注度的逐级上升。比如通过检测发现了 50 个异常，进一步聚合为 10 个威胁特征，再次产生了 1-2 个威胁指标，这种数据扩展的方式让 UEBA 能够进行异常和威胁检测。

2.2.2 UEBA 适用场景

UEBA 主要适用于分析内部恶意者、失陷账号、安全事件上下文关联、实体

分析以及数据泄露防护。

分析内部恶意者：包括有权限进入 IT 系统的内部员工和合同商，他们可能会对企业进行网络攻击。因为内部员工拥有企业数据资产的访问权限，所以一般很难通过日志文件或常规安全事件来衡量内部恶意者的恶意程度或发现内部恶意者，但是使用 UEBA 技术，可以通过建立用户典型行为的基线并检测异常活动来有效发现存在异常行为的用户[20]。

失陷账号：攻击者通常会渗透到组织并破坏网络上的特权账户或可信主机，并从那里继续攻击。如果当前不了解攻击模式或杀死链（例如零日攻击），或者如果攻击通过更改凭据、IP 地址或机器间横向移动传播等进行，那么传统安全工具很难检测到被黑的内部人员，且难以跟随攻击者的技术升级，但 UEBA 技术可以根据资产的行为与已建立的不同基线来快速检测和分析攻击者通过受感染账户进行的不良活动。

安全事件上下文关联：安全信息和事件管理从多个安全工具和关键系统那里收集事件和日志，并生成必须由安全人员调查的大量警报，从而导致警报疲劳。UEBA 通过将安全事件进行逻辑关联，可以帮助区分哪些事件在组织环境中特别异常、可疑或存在潜在危险，可以通过添加有关组织结构的数据（例如，资产的关键性、特定组织功能的角色和访问级别）来构建基线和威胁模型。

实体分析：UEBA 在处理物联网（Internet of Things，IoT）安全风险方面的作用尤为重要。企业部署大量连接设备，通常只做很少的安全措施，甚至不做安全措施。攻击者可能会破坏物联网设备，使用它们窃取数据或获取对其他 IT 系统的访问权限，或者更糟糕的是利用本企业已感染的设备对第三方企业发起（Distributed Denial of Service，DDoS）攻击或其他攻击。UEBA 可以跟踪无数个连接设备，为每个设备或类似设备组建立行为基线，并立即检测设备是否在其常规边界之外运行。

数据泄露防护（Data leakage prevention，DLP）：UEBA 解决方案可以通过了解哪些事件是现有基线下的异常行为来获取 DLP 警报，确定报警日志的优先级并将日志合并。这为调查人员节省了时间，并帮助他们更快地发现真正的安全事件。

2.2.3 UEBA 算法

传统安全分析技术使用人工规则和威胁建模，效果好坏取决于安全管理人

员的技术水平。UEBA 技术使用机器学习技术，能够在缺少匹配的情况下通过行为异常发现攻击。UEBA 技术会用到的机器学习算法有监督机器学习、贝叶斯网络、无监督机器学习、增强/半监督机器学习、深度学习。

监督机器学习：将已知的良好和不良行为的数据集输入系统，用于学习分析新行为是与已知的良好行为相似，还是与不良行为相似。

贝叶斯网络：是一种概率图型模型，可以结合监督机器学习和相应的规则来创建行为画像。

无监督机器学习：不需要对训练样本数据进行人工标注，计算机代替人工解决模式识别中的各种问题。通过该算法计算机完成学习正常行为，检测并警告异常行为。但是这种算法没有办法判断异常行为是好还是坏，只能判断是否偏离了正常值。

增强/半监督机器学习：它是一种混合模型，其基础是无监督机器学习。在异常检测中，实际警报的误报百分比被反馈到系统中，以允许机器学习微调模型并降低信噪比，提升检测模型的准确率。这种方法的突出价值在于，能够学习到人无法描述的深层次特征。

深度学习：用于虚拟警报分类和调查，可以系统地训练代表安全警报及其分类结果的数据集，使其具备自我识别的功能，并能够预测新的安全警报集的分类结果。

2.3 K-Means 算法

2.3.1 聚类

介绍 K-Means 算法前，首先了解下聚类。聚类是指把已知的数据划分成多个组或者“簇”，同一个簇内的数据尽量相似，不同簇内的数据尽量不同，它是一种无监督的机器学习技术。聚类不仅可以将数据实现分割，还可以用于异常点的监控。

2.3.2 算法概述

K 均值聚类算法（K-Means clustering algorithm）是在 1967 年由 J.B.MacQueen 提出，并对该算法进行了详细描述。该算法是基于距离的聚类算法，是一种迭代求解的聚类分析算法[21]。目前该算法已被广泛应用在各领域。

2.3.3 算法实现

假设对一个 d 维向量的点集 D=$\{x_i|i=1,\dots,N\}$进行聚类，其中$x_i \in$d 表示第 i 条数据。集合 C=$\{C_j|j=1,\dots,K\}$表示聚簇。聚类算法一般是基于“紧密度”或者“相似度”等原则对数据集进行分组，默认的紧密度衡量标准是欧几里得距离。K-Means 算法要使每个点x_i和离它最近的聚簇中心C_j之间的欧几里得距离的平方和最小[22]，它的目标函数如下列公式(2.1)所示：

$$\text{Cost}=\sum_{i=1}^{N}(argmin_j||x_i-C_j||_2^2) \tag{2.1}$$

算法流程：

步骤一，从数据中随机挑选 K 个数据点作为原始的聚簇中心。

步骤二，计算其他数据与各聚簇中心的距离，并把数据集中非聚簇中心的数据分配给距离它最近的聚簇中心。

步骤三，重新计算各聚簇中样本点的均值，并以均值作为新的聚簇中心。

不断重复步骤二和步骤三这个过程，直到没有样本数据被重新分配给不同的聚类，聚簇中心也没有再发生改变，算法收敛。

2.3.4 算法优缺点

K-Means 聚类算法的优点：

1、在处理大数据集中，该算法效率高、伸缩性好；

2、原理简单、实现容易；

3、聚类效果中上；

4、适用于高维。

K-Means 聚类算法的缺点：

1、合理的 K 值难以确定，该算法的 K 需要事先确定，但在实际应用中并不清楚 K 选多少才合适；

2、对噪声和离群点敏感：该算法很容易受到离群点的影响，因为中心点是通过样本均值确定的；

3、不适于发现非球形的簇或大小差别很大的簇： 因为该算法是基于欧式距离的方式判断样本间的相似度[23]。

2.4 孤立森林

2.4.1 异常

异常(或异常对象、异常、罕见事件、特殊对象)是数据分析的一个重要概念，目前学术界对异常的定义有很多种，比如：数据对象明显偏离特定领域中常见数据行为的规则模式，一种不太可能的观察结果——一种低概率事件。在孤立森林中，异常被定义为“容易被孤立的离群点”。

2.4.2 算法概述

孤立森林(Isolation Forest，iForest)又叫隔离森林，是一种从异常点出发，通过指定规则进行划分，根据划分次数进行判断的异常检测方法，或者说离群点挖掘方法，总之是在一大堆数据中，找出与其它数据的规律不太符合的数据。它由南京大学周志华教授及其团队成员提出。孤立森林是由N棵树构成，是一种基于空间随机划分思想的集成算法，由多棵二叉树并行得到，再将输出结果进行加权平均[24]。在孤立森林中，递归地随机分割特征及特征值，直到所有的对象都是孤立的。在这种随机分割的方式下，异常点通常具有较短的路径长度。也就是说，那些高密度簇中的点是需要被分割很多次才能被孤立，而那些低密度簇中的点很少的分割次数就可以将其孤立。从图2-1和图2-2中可以看到，正常点X_i需要更多次的分割才能被孤立，而异常点X_0需要较少的分割次数就能被孤立。图2-3展示了异常点的平均路径长度小于正常点的路径长度。

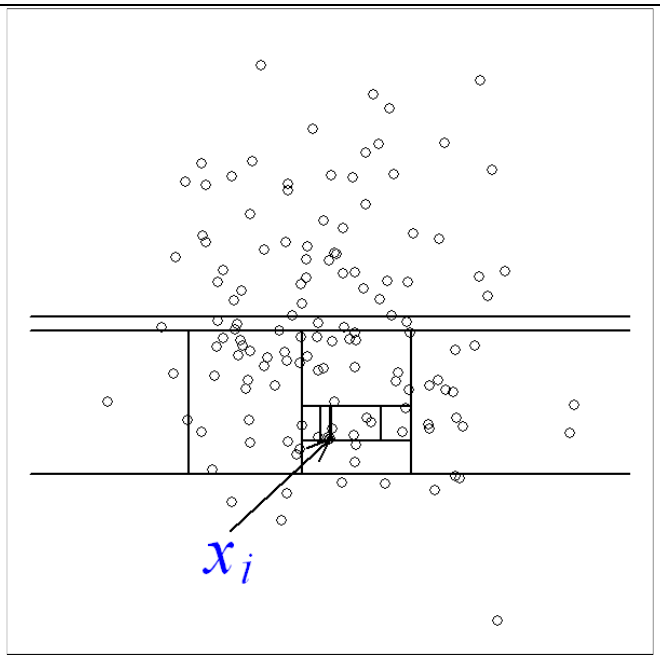


图 2-1 密度高的簇中点示例

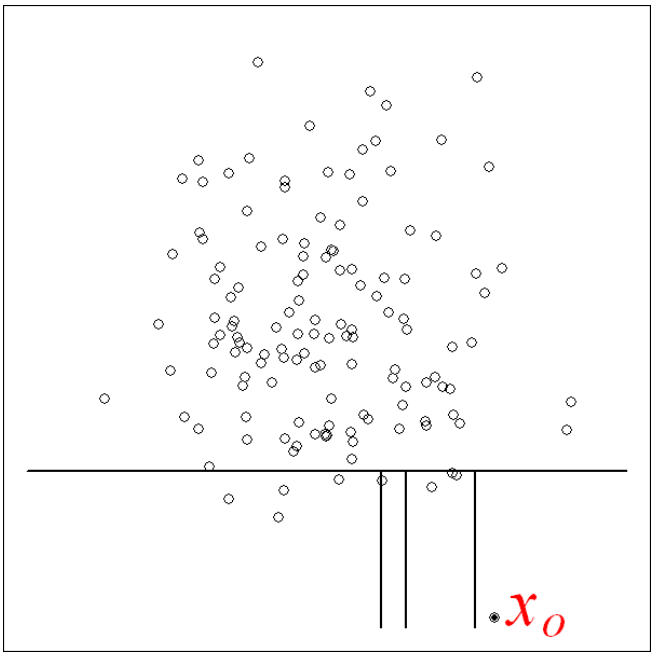


图 2-2 密度低的簇中点示例

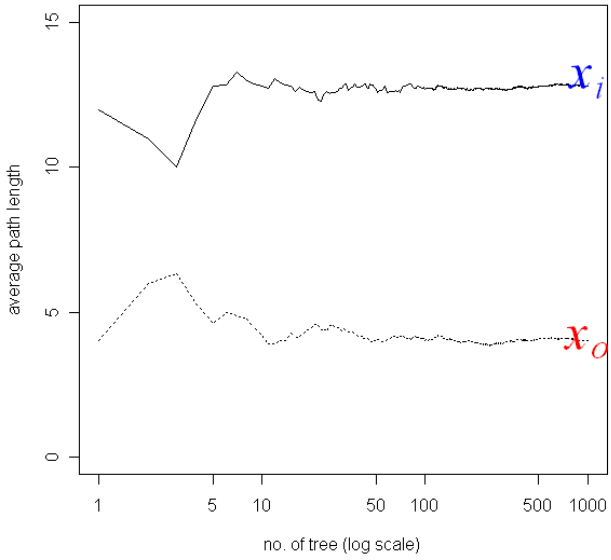


图 2-3 异常点和正常点的平均路径长度

2.4.3 孤立树

孤立树(Isolation Tree，iTree）是一种随机二叉树：假设T是孤立树的一个节点，它要么是没有子节点的叶子节点，要么是只有两个子节点(T_l,T_r)的内部节点。孤立树的构建过程：假设n个样本数据X={x_i,…,x_n}，特征的维度为d，随机选择一个特征q及其分割值p，递归地分割数据集X，直到满足以下任意一个条件：(1)树达到了限制的高度；(2)节点上只有一个样本；(3)节点上的样本所有特征都相同。

对数据的预测是将测试数据在孤立树上游走一遍，计算其落入的叶子节点到根节点的路径长度h(x)，用其来判断数据点x是否是异常点。一个包含n个样本的数据集，树的平均路径长度如下列公式(2.2)所示[25]：

$$c(n) = 2H(n-1)-2(n-1)/n \tag{2.2}$$

公式(2.2)中，c(n)表示样本点 n 的平均路径长度，H(i)为调和数，可由 ln(i) + 0.5772156649(欧拉常数)估计。实例 x 的异常得分 S 定义如下列公式(2.3)所示。

$$S(x,n)=2^{-\frac{E(h(x))}{c(n)}} \tag{2.3}$$

公式(2.3)中，E(h(x))为样本x在孤立树中路径长度的期望值，S和E(h(x))的关系如图2-4所示。

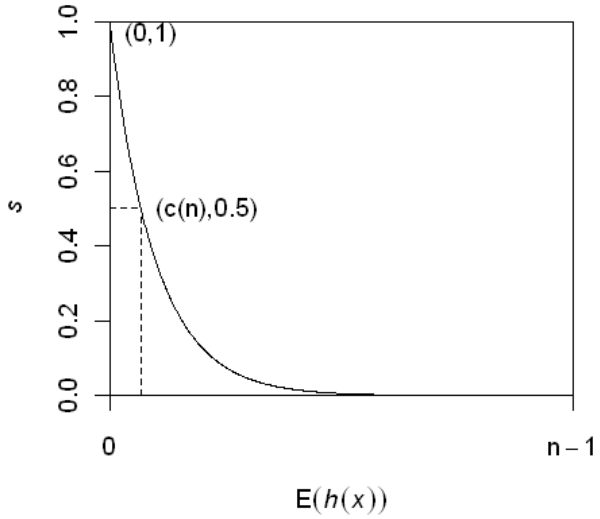


图 2-4 孤立森林异常得分

异常得分趋近于0.5分时，不能判断是否为异常，异常得分趋近于1分时，认为是异常点，异常得分趋近于0分时，认为是正常点。

2.4.4 异常检测

利用孤立森林进行异常检测包含两阶段：

第一个阶段使用训练集的子样本构建孤立树。

第二个阶段让测试集通过孤立树，以获得每个测试数据的异常得分。

在训练阶段，通过递归划分给定的训练集来构造孤立树，直到所有的样本点被孤立，或者树达到了指定的高度。树的高度限制l与子样本数量ψ的关系为l=ceiling($log_2(\psi)$)，它近似等于树的平均高度。将树生长到平均树高的基本原理：只对路径长度小于平均长度的数据点感兴趣，因为这些点更有可能是异常点。子样本大小ψ和树的数量t的经验值：ψ=256，t=100，在训练过程结束时，返回一组树，并为评估阶段做好准备。

在评估阶段，从每个测试实例的预期路径长度E(h(x))得到一个异常分数S，

异常分数S的计算公式见公式(2.3)。

2.4.5 算法优缺点

孤立森林算法的优点：

1、具有线性时间复杂度，具有低常数和低内存需求，此外，孤立森林算法的收敛速度快，集合规模小，能够高效检测异常；

2、可以部署在大规模分布式系统上加速运算；

3、不需要人工标注。

孤立森林算法的缺点：

1、不适用于特别高维的数据；

2、仅对全局稀疏点敏感。

2.5 本章小结

本章主要介绍研究课题的相关理论与技术。首先介绍了企业用户异常行为的概念，从两个维度对异常进行分类，同时介绍了6种不同的异常检测技术。然后介绍了UEBA的概念、及其适用的场景和常用算法。最后分别介绍K-Means聚类算法和孤立森林算法的概述、实现及优缺点。

第 3 章 特征分析与行为基线构造

用户行为特征提取是整个用户行为分析建模的基础，本文通过查阅 UEBA 相关的技术文献，结合业务实际需求，找出相关的数据实体，以数据实体为中心，规约数据维度类型和关联关系，形成符合业务实际情况的建模体系。一般的特征提取步骤包括用户数据与实体数据的分解和对应、实体间关联关系分解、用户特征维度分解以及用户行为特征的提取。

3.1 数据集介绍

本文的企业用户网络异常行为检测是基于企业内部用户的实体和行为的异常检测。因此，需要采集企业内部用户的实体和行为信息，即企业用户访问各种服务器产生的各种日志文件。考虑到数据安全，无法搜集企业这些日志文件作为研究。本文研究使用的数据集来自Datafountain的公开数据集，这些数据集已做了脱敏处理，共528690条，其中训练集的数据468889条，测试集的数据59801条。接下来，将通过表3.1描述的5条数据示例对研究使用的数据集进行介绍。

表 3.1 数据集示例

id	account	group	ip	url	port	vlan	Switchip	time
20120	anpeng@qq.com	市场战略发展中心	0	0	0	800	0	2021/1/21 18:59
22152	anpeng@qq.com	市场战略发展中心	1	1	1	800	0	2021/1/22 18:59
20165	anpeng@qq.com	市场战略发展中心	2	2	2	800	1	2021/1/23 18:59
20800	anpeng@qq.com	市场战略发展中心	3	3	3	800	2	2021/1/24 18:59
19690	anpeng@qq.com	市场战略发展中心	4	4	4	800	3	2021/1/25 18:59

表 3.2 企业用户日常上网日志字段描述

序号	字段	字段描述	字段类型
1	id	日志数据记录编号	数字类型
2	account	用户账号	字符类型
3	group	用户归属部门	字符类型
4	IP	终端 IP 地址	字符类型

序号	字段	字段描述	字段类型
5	url	终端上网网址	字符类型
6	port	终端上网应用端口	数字类型
7	vlan	终端所在虚拟网域编号	数字类型
8	switchIP	终端连接交换机 IP	字符类型
9	time	终端上网行为发生时间	日期类型
10	ret	异常分数	浮点类型

表 3.2 描述了企业用户日常上网日志所涉及到的主要 10 个字段，下面分别对每个字段在企业内部上网行为的意义进行描述。

编号id：主键字段，用于一条企业用户上网行为日志的唯一标识，方便对数据的统计与分析，提升数据库的处理速度。

用户账号：一般该账号必须经过企业内部审批通过，才有权限访问相关的网站或数据，所以通过这单一特征是无法判断用户行为是否异常，需要对其进行分析。

用户归属部门：每个企业用户都有对应的组织机构，每个公司的组织机构一般是固定不常变化的。所以通过这单一特征是无法判断用户行为是否异常，需要对其进行分析。

终端IP地址：对于企业用户，工作的设备通常是企业内部的设备，IP地址的范围是固定的。但是不能只通过这单一特征来判断用户行为是否异常，例如，企业用户可能去客户处开会或者借调到别的项目组出差等，通过VPN登录了企业内部网络进行工作，所以需要对其进行分析。

上网网址：一般来说，企业经常访问的网址范围是固定的，但是，我们不能只通过该用户访问的网址之前没有访问过就判定用户是否异常。例如，系统更换服务器，或者访问交互系统对应的网站。

应用端口：一般来说，企业经常访问的应用端口是固定的，但是，我们不能只通过该用户访问的端口来判断用户是否异常。例如，服务器重新部署。

虚拟网域编号：一般来说，终端所在虚拟网域编号是固定的，但是，我们不能只通过该用户所属的虚拟网域编号判断用户是否异常。例如，去客户处开会，用户就可以在范围外的虚拟网域编号登入系统。

终端交换机IP：一般来说，终端交换机IP是固定的，但是，我们不能只通过该用户访问的终端交换机来判断用户是否异常。例如，去客户处开会，用户上网对应的终端交换机IP就会发生变化。

上网行为发生时间：一般来说，企业用户正常的工作时间有一定的范围。

例如，早上8:30到晚上17:30。但是，我们不能只通过上网行为发生时间在这个时间范围外就判定用户是异常。例如，加班，用户就可以在范围外的时间登入系统。

3.2 面向企业用户网络异常行为检测需要的上网行为基线构造

所谓的上网行为基线是指人们在自然情形下，正常上网的行为表现。这涉及到确定分析的时间粒度，并确定一组特性来描述每个用户（使用终端或网络服务的人）在每个时间段内的访问模式。一旦知道这种基线行为，系统就可以将用户的上网行为日志与基线进行对比，确定偏离程度，从而判断该行为是否异常。

本文通过企业员工日常访问行为记录日志数据来挖掘三个维度的用户，分别是：企业级维度用户、部门级维度用户以及个人维度用户。通过每个维度的用户和实体（终端IP、终端连接交换机IP）的历史行为（访问的网址、访问的应用端口、访问的时间）数据，构建企业行为基线、部门行为基线、个人行为基线。

企业行为基线：包括的特征有企业所有用户在终端IP、访问时间基线，企业所有用户在终端IP、应用端口、访问时间基线，企业所有用户在终端连接交换机IP、访问时间基线，企业所有用户访问网址、访问时间基线，企业所有用户在终端IP、访问网址、访问时间基线。

部门行为基线：包括的特征有部门所有用户访问时间基线，部门所有用户在终端IP、访问时间基线，部门所有用户在终端IP、应用端口、访问时间基线，部门所有用户在终端连接交换机IP、访问时间基线，部门所有用户访问网址、访问时间基线，部门所有用户在终端IP、访问网址、访问时间基线。

个人行为基线：包括的特征有单个账号访问时间基线，单个账号在终端IP、访问时间基线，单个账号在终端IP、应用端口、访问时间基线，单个账号在终端连接交换机IP、访问时间基线，单个账号访问网址、访问时间基线，单个账号在终端IP、访问网址、访问时间基线。

3.3 特征构造

3.3.1 特征构造常见方法

数据对模型的预测起到至关重要的作用，好的数据除原始特征外，还有很多是通过特征构造所得。特征构造是通过属性的分割和组合构造成一个新特征的过程，是UEBA分析中非常重要的一个环节。常用的特征构造方法有时间序列处理法、统计值构造法、分解类别属性法、分箱和分区法、组合特征法等。

时间序列处理法：将时间戳属性值分割成多个维度的值，比如年、月、日、时、分、秒等；

统计值构造法：通过对单个或多个属性进行求最大值、最小值、平均值、总数、中位数、众数等，构造新的特征值。

分解类别属性法：对类别型的特征进行分解。如：由{红、黄、蓝}组成的颜色特征，最常用的方式是独热编码（One-Hot Encoding），把每个特征转换成二元属性值，即从{0,1}取一个值，1表示有，0表示无。上述示例中{红、黄、蓝}通过独热编码后的值为{100、010、001}。

分箱法：通过考察相邻数据特征，将一组较多连续值分割成多组较少的离散值，常用的分箱法有等深分箱法（每个箱子有相同的记录数）、等宽分箱法（每个箱子的区间相等）和用户自定义分箱法（自定义规则进行分箱）。数据分箱后，再对每个箱中的数据进行平滑处理，平滑处理的方法有按平均值平滑（同一个箱子中的平均值代替箱中的所有值）、按边界值平滑（距离较小的边界值代替箱中的所有值）、按中值平滑（每个箱中的中值代替箱中的所有值）[26]。

组合特征法：也叫特征交叉，是将两个或多个原始特征组合成包含更多信息的新特征。

3.3.2 企业用户上网行为的特征构造

从上节的数据介绍中可以看出，企业用户上网行为日志基本特征种类偏少，维度偏低，很难发现其中的规律，无法对正常行为和异常行为进行区分。所以需要对原始的数据特征进行数据挖掘，以提升特征维度，增加数据解释能力。本文通过特征构造法总共构造了90个特征，包括：时间序列处理法构造了5个特征，统计值构造法和组合特征构造法构造了85个特征。

通过时间序列处理法，可以将时间特征发散成：星期几、月、日、小时、分钟5个特征，如表3.3所示。

表 3.3 经过时间序列处理法得到的新特征

序号	原始特征	新特征	新特征描述	字段类型
1	time	dayofweek	星期几	数字类型
2	time	month	上网月份	数字类型
3	time	day	几号上网	数字类型
4	time	hour	几时上网	数字类型
5	time	minute	几分上网	数字类型

根据企业行为基线、部门行为基线、个人行为基线构建思路，通过统计值构造法和组合特征法将用户账号、用户归属部门、终端IP地址、终端上网网址、终端上网应用端口、终端连接交换机IP构造出85个特征，如表3.4所示。

表 3.4 经过统计值构造法和组合特征法得到的新特征

序号	原始特征	新特征	特征描述	基线类型	字段类型
1	ip、hour	ip_hour_max	终端 ip_上网时间最大值	企业行为基线	数字类型
2	ip、hour	ip_hour_min	终端 ip_上网时间最小值	企业行为基线	数字类型
3	ip、hour	ip_hour_mean	终端 ip_上网时间平均值	企业行为基线	数字类型
4	ip、hour	ip_hour_median	终端 ip_上网时间中位数	企业行为基线	数字类型
5	ip、hour	ip_hour_count	终端 ip_上网时间总数量	企业行为基线	数字类型
6	ip、port、hour	ip_port_hour_max	终端 ip_应用端口_访问时间最大值	企业行为基线	数字类型
7	ip、port、hour	ip_port_hour_min	终端 ip_应用端口_访问时间最小值	企业行为基线	数字类型
8	ip、port、hour	ip_port_hour_mean	终端 ip_应用端口_访问时间平均值	企业行为基线	数字类型
9	ip、port、hour	ip_port_hour_median	终端 ip_应用端口_访问时间中位数	企业行为基线	数字类型
10	ip、port、hour	ip_port_hour_count	终端 ip_应用端口_访问时间总数	企业行为基线	数字类型
11	switchIp、hour	switchIp_hour_max	交换机 ip_上网时间最大值	企业行为基线	数字类型
12	switchIp、hour	switchIp_hour_min	交换机 ip_上网时间最小值	企业行为基线	数字类型
13	switchIp、hour	switchIp_hour_mean	交换机 ip_上网时间平均值	企业行为基线	数字类型
14	switchIp、hour	switchIp_hour_median	交换机 ip_上网时间中位数	企业行为基线	数字类型
15	switchIp、hour	switchIp_hour_count	交换机 ip_上网时间中位数	企业行为基线	数字类型

序号	原始特征	新特征	特征描述	基线类型	字段类型
16	url、hour	url_hour_max	url_上网时间最大值	企业行为基线	数字类型
17	url、hour	url_hour_min	url_上网时间最小值	企业行为基线	数字类型
18	url、hour	url_hour_mean	url_上网时间平均值	企业行为基线	数字类型
19	url、hour	url_hour_median	url_上网时间中位数	企业行为基线	数字类型
20	url、hour	url_hour_count	url_上网时间总数量	企业行为基线	数字类型
21	ip、url、hour	ip_url_hour_max	ip_url_上网时间最大值	企业行为基线	数字类型
22	ip、url、hour	ip_url_hour_min	ip_url_上网时间最小值	企业行为基线	数字类型
23	ip、url、hour	ip_url_hour_mean	ip_url_上网时间平均值	企业行为基线	数字类型
24	ip、url、hour	ip_url_hour_median	ip_url_上网时间中位数	企业行为基线	数字类型
25	ip、url、hour	ip_url_hour_count	ip_url_上网时间总数量	企业行为基线	数字类型
26	group、hour	group_hour_max	部门_上网时间最大值	部门行为基线	数字类型
27	group、hour	group_hour_min	部门_上网时间最小值	部门行为基线	数字类型
28	group、hour	group_hour_mean	部门_上网时间最大值	部门行为基线	数字类型
29	group、hour	group_hour_median	部门_上网时间中位数	部门行为基线	数字类型
30	group、hour	group_hour_count	部门_上网时间总数	部门行为基线	数字类型
31	group、ip、hour	group_ip_hour_max	部门_终端 ip_上网时间最大值	部门行为基线	数字类型
32	group、ip、hour	group_ip_hour_min	部门_终端 ip_上网时间最小值	部门行为基线	数字类型
33	group、ip、hour	group_ip_hour_mean	部门_终端 ip_上网时间平均值	部门行为基线	数字类型
34	group、ip、hour	group_ip_hour_median	部门_终端 ip_上网时间中位数	部门行为基线	数字类型
35	group、ip、hour	group_ip_hour_count	部门_终端 ip_上网时间总数量	部门行为基线	数字类型
36	group、ip、port、hour	group_ip_port_hour_max	部门_终端 ip_应用端口_访问时间最大值	部门行为基线	数字类型

序号	原始特征	新特征	特征描述	基线类型	字段类型
37	group、ip、port、hour	group_ip_port_hour_min	部门_终端 ip_应用端口_访问时间最小值	部门行为基线	数字类型
38	group、ip、port、hour	group_ip_port_hour_mean	部门_终端 ip_应用端口_访问时间平均值	部门行为基线	数字类型
39	group、ip、hour	group_ip_port_hour_median	部门_终端 ip_应用端口_访问时间中位数	部门行为基线	数字类型
40	group、ip、hour	group_ip_port_hour_count	部门_终端 ip_应用端口_访问时间总数	部门行为基线	数字类型
41	group_switchIp、hour	group_switchIp_hour_max	部门_交换机 ip_上网时间最大值	部门行为基线	数字类型
42	group_switchIp、hour	group_switchIp_hour_min	部门_交换机 ip_上网时间最小值	部门行为基线	数字类型
43	group_switchIp、hour	group_switchIp_hour_mean	部门_交换机 ip_上网时间平均值	部门行为基线	数字类型
44	group_switchIp、hour	group_switchIp_hour_median	部门_交换机 ip_上网时间中位数	部门行为基线	数字类型
45	group_switchIp、hour	group_switchIp_hour_count	部门_交换机 ip_上网时间中位数	部门行为基线	数字类型
46	group、url、hour	group_url_hour_max	部门_url_上网时间最大值	部门行为基线	数字类型
47	group、url、hour	group_url_hour_min	部门_url_上网时间最小值	部门行为基线	数字类型
48	group、url、hour	group_url_hour_mean	部门_url_上网时间平均值	部门行为基线	数字类型
49	group、url、hour	group_url_hour_median	部门_url_上网时间中位数	部门行为基线	数字类型
50	group、url、hour	group_url_hour_count	部门_url_上网时间总数量	部门行为基线	数字类型
51	group、ip、url、hour	group_ip_url_hour_max	部门_ip_url_上网时间最大值	部门行为基线	数字类型
52	group、ip、url、hour	group_ip_url_hour_min	部门_ip_url_上网时间最小值	部门行为基线	数字类型
53	group、ip、url、hour	group_ip_url_hour_mean	部门_ip_url_上网时间平均值	部门行为基线	数字类型
54	group、ip、url、hour	group_ip_url_hour_median	部门_ip_url_上网时间中位数	部门行为基线	数字类型
55	group、ip、url、hour	group_ip_url_hour_count	部门_ip_url_上网时间总数量	部门行为基线	数字类型
56	account、hour	account_hour_max	个人账号_上网时间最大值	个人行为基线	数字类型
57	account、hour	account_hour_min	个人账号_上网时间最小值	个人行为基线	数字类型

序号	原始特征	新特征	特征描述	基线类型	字段类型
58	account、hour	account_hour_mean	个人账号_上网时间最大值	个人行为基线	数字类型
59	account、hour	account_hour_median	个人账号_上网时间中位数	个人行为基线	数字类型
60	account、hour	account_hour_count	个人账号_上网时间总数	个人行为基线	数字类型
61	account、ip、hour	account_ip_hour_max	个人账号_终端 ip_上网时间最大值	个人行为基线	数字类型
62	account、ip、hour	account_ip_hour_min	个人账号_终端 ip_上网时间最小值	个人行为基线	数字类型
63	account、ip、hour	account_ip_hour_mean	个人账号_终端 ip_上网时间平均值	个人行为基线	数字类型
64	account、ip、hour	account_ip_hour_median	个人账号_终端 ip_上网时间中位数	个人行为基线	数字类型
65	account、ip、hour	account_ip_hour_count	个人账号_终端 ip_上网时间总数量	个人行为基线	数字类型
66	account、ip、port、hour	account_ip_port_hour_max	个人账号_终端 ip_应用端口_访问时间最大值	个人行为基线	数字类型
67	account、ip、port、hour	account_ip_port_hour_min	个人账号_终端 ip_应用端口_访问时间最小值	个人行为基线	数字类型
68	account、ip、port、hour	account_ip_port_hour_mean	个人账号_终端 ip_应用端口_访问时间平均值	个人行为基线	数字类型
69	account、ip、port、hour	account_ip_port_hour_median	个人账号_终端 ip_应用端口_访问时间中位数	个人行为基线	数字类型
70	account、ip、port、hour	account_ip_port_hour_count	个人账号_终端 ip_应用端口_访问时间总数	个人行为基线	数字类型
71	account、switchIp、hour	account_switchIp_hour_max	个人账号_交换机 ip_上网时间最大值	个人行为基线	数字类型
72	account、switchIp、hour	account_switchIp_hour_min	个人账号_交换机 ip_上网时间最小值	个人行为基线	数字类型
73	account、switchIp、hour	account_switchIp_hour_mean	个人账号_交换机 ip_上网时间平均值	个人行为基线	数字类型
74	account、switchIp、hour	account_switchIp_hour_median	个人账号_交换机 ip_上网时间中位数	个人行为基线	数字类型

序号	原始特征	新特征	特征描述	基线类型	字段类型
75	account、switchIp、hour	account_switchIp_hour_count	个人账号_交换机 ip_上网时间中位数	个人行为基线	数字类型
76	account、url、hour	account_url_hour_max	个人账号_url_上网时间最大值	个人行为基线	数字类型
77	account、url、hour	account_url_hour_min	个人账号_url_上网时间最小值	个人行为基线	数字类型
78	account、url、hour	account_url_hour_mean	个人账号_url_上网时间平均值	个人行为基线	数字类型
79	account、url、hour	account_url_hour_median	个人账号_url_上网时间中位数	个人行为基线	数字类型
80	account、url、hour	account_url_hour_count	个人账号_url_上网时间总数量	个人行为基线	数字类型
81	account、ip、url、hour	account_ip_url_hour_max	个人账号_ip_url_上网时间最大值	个人行为基线	数字类型
82	account、ip、url、hour	account_ip_url_hour_min	个人账号_ip_url_上网时间最小值	个人行为基线	数字类型
83	account、ip、url、hour	account_ip_url_hour_mean	个人账号_ip_url_上网时间平均值	个人行为基线	数字类型
84	account、ip、url、hour	account_ip_url_hour_median	个人账号_ip_url_上网时间中位数	个人行为基线	数字类型
85	account、ip、url、hour	account_ip_url_hour_count	个人账号_ip_url_上网时间总数量	个人行为基线	数字类型

3.4 特征选择

3.4.1 特征选择常见方法

特征选择[27](Feature Selection)，是指从已有的m个原始特征中选择n个有效的特征，是对数据降维的一个过程，通过这个过程可以提升模型训练的性能，节省服务器的存储和开销，也可以简化模型，删掉不相关的特征会减少模型训练的难度，降低过拟合风险。在模型训练中，好的特征是训练模型的关键。一般，从两个方面来选择特征：一是特征是否发散？如果一个特征不发散，也就是所有实例数据在这个特征上基本相同，那么这个特征将无法把各实例区别开。二是特征与目标的相关性：与目标的相关性越高的特征，越应该被优先选择。

常用的特征选择方法有：

Filter(过滤法)：根据发散特性或者相关特性对所有的特征进行评分，设定边界值或者待选择边界值来选择特征。常用的方法有方差选择法、相关系数法、

卡方检验和互信息法。

Wrapper(包装法)：根据目标函数，每次选择一些特征，或者删除一些特征。常用的包装法有递归特征消除法。

Embedded(嵌入法)：把特征选择嵌入到模型的训练过程中，得到每个特征的权值系数值，根据系数值从大到小选择特征。常用的嵌入法有基于惩罚项的特征选择法和基于树模型的特征选择法。

在上述3.3小节中为了使用UEBA进行企业用户网络异常行为分析，首先需要对原始数据集通过三种方法，即时间序列处理法、统计值构造法和组合特征法构造了具有特征数为90个的数据集。其次，为了基于UEBA构建具有高质量的检测模型，应该分析所构造的数据集中哪些特征是有利于构建企业用户网络异常行为检测模型的。所以对构造好的数据集通过特征选择方法进行特征约简，以减少不发散的特征及特征间相关性很高的特征，对提升企业用户网络异常行为检测模型的性能具有重要意义。本文使用的特征选择方法是方差选择和皮尔逊相关系数法。

3.4.2 方差选择

在统计学中，方差是各样本值与其均值之差的平方值的平均数，是衡量随机变量离散程度最重要的指标，也是衡量随机变量和其数学期望之间偏离程度的方法。当数据比较集中时，数据的方差较小，当数据比较分散时，数据的方差较大。因此需要优先删除方差为 0 或较小的特征。特征选择中，方差选择是针对每一列特征，求其方差，然后通过设定一个阈值，来选取方差较大的特征，方差大说明信息量大。本文通过选取 0.3 的阈值，删除了特征：终端 ip_上网时间最大值(ip_hour_max)，终端 ip_上网时间最小值(ip_hour_min)，交换机 ip_上网时间最大值(switchip_hour_max)，交换机 ip_上网时间最小值(switchip_hour_min)，部门_上网时间最大值(group_hour_max)，部门_上网时间最小值(group_hour_min)。

3.4.3 皮尔逊相关系数

通过方差选择后，再使用相关系数法中的皮尔逊相关系数方法对特征进行进一步选择。

皮尔逊相关系数是用于度量两个变量X和Y之间的线性相关性，其值介于-1与1之间。在机器学习中可以用其来衡量特征间的相似度。

对于总体（由很多共同性质组成的集合），相关系数ρ等于变量X和变量Y之间的协方差cov(X,Y)除以X的标准差$\boldsymbol{\delta_X}$和Y的标准差$\boldsymbol{\delta_Y}$之间的乘积，表达式如下列公式(3.1)所示：

$$\rho(X,Y)=\frac{\boldsymbol{cov(X,Y)}}{\boldsymbol{\delta_X\delta_Y}}=\frac{\boldsymbol{E[(X-\mu_X)(Y-\mu_Y)]}}{\boldsymbol{\delta_X\delta_Y}} \tag{3.1}$$

对于同样本来说，样本的皮尔森相关系数r的表达式如下列公式(3.2)所示：

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{3.2}$$

公式(3.2)中 $\bar{X}$为X_i样本的均值， $\bar{Y}$为Y_i样本的均值，n为样本数量。r还可以由(X_i，Y_i)样本点的标准分数均值，得到与上式等价的表达式如下列公式(3.3)所示：

$$r=\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\delta_X}\right)\left(\frac{Y_i-\bar{Y}}{\delta_Y}\right) \tag{3.3}$$

公式(3.3)中δ_X为X_i样本的标准差，δ_Y为Y_i样本的标准差。

当r>0时，表明两个变量正相关；

当r<0时，表明两个变量负相关；

当r=0时，表明两个变量不相关；

当r=1和-1时，表明所有样本点都落在一条直线上。

在上节方差删除特征选择后的基础上，通过计算皮尔逊相关系数，按照相关系数由小到大排列后，根据相关系数阈值设置（0.5）得到17个特征。详细见表3.5。

表 3.5 经过方差选择和皮尔逊相关系数得到的最终特征

序号	特征	特征描述
1	ip_port_hour_max	终端 ip_应用端口_访问时间最大值
2	ip_PORT_hour_count	终端 ip_应用端口_访问时间总数
3	switchip_hour_count	交换机 ip_上网时间中位数
4	url_hour_max	url_上网时间最大值
5	ip_URL_hour_max	ip_url_上网时间最大值
6	group_hour_mean	部门_上网时间最大值
7	group_hour_count	部门_上网时间总数

序号	特征	特征描述
8	group_ip_hour_max	部门_终端 ip_上网时间最大值
9	group_ip_hour_median	部门_终端 ip_上网时间中位数
10	group_switchip_hour_max	部门_交换机 ip_上网时间最大值
11	group_url_hour_max	部门_url_上网时间最大值
12	group_ip_url_hour_max	部门_ip_url_上网时间最大值
13	account_hour_count	个人账号_上网时间总数
14	account_ip_hour_max	个人账号_终端 ip_上网时间最大值
15	account_ip_hour_count	个人账号_终端 ip_上网时间总数量
16	account_switchip_hour_max	个人账号_交换机 ip_上网时间最大值
17	account_url_hour_max	个人账号_url_上网时间最大值

每个特征间的相关系数见图3-1皮尔逊相关系数图。

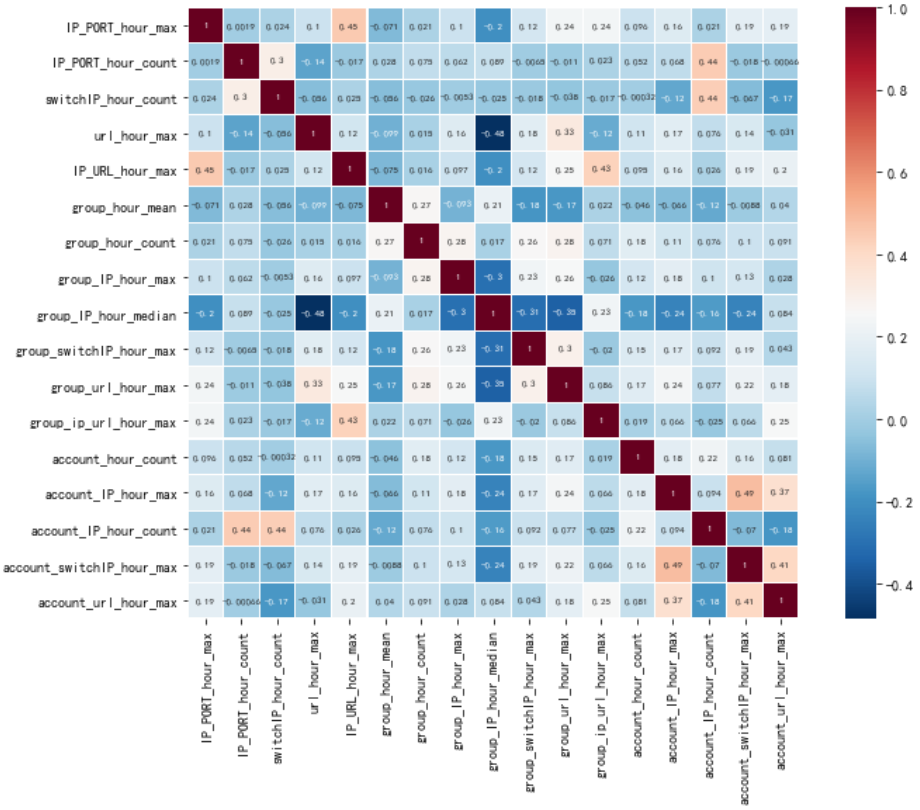


图 3-1 皮尔逊相关系数图

3.5 本章小结

本章第一小节对本文使用的数据集进行了介绍。第二小节介绍了面向企业用户网络异常行为检测需要的企业上网行为基线、部门上网行为基线、个人上网行为基线。第三小节介绍了特征构造的定义、常用的特征构造方法，本文的特征构造法及构造后的特征。 第四小节介绍了特征选择的常见方法及本文通过方差选择和皮尔逊相关系数方法最后确认的17个特征。

第 4 章 结合 K 均值聚类和孤立森林算法的异常行为检测模型

4.1 结合 K 均值聚类-孤立森林算法的异常行为检测模型图

用户、实体和行为特征是用户异常行为建模的 3 大要素。通过前面章节按照 UEBA 方法对实体和关系进行了分析，提取了 17 种有效的用户行为特征，并构造了三个维度的企业用户网络异常行为基线。按照 UEBA 方法，在特征获取后，就可以对获取的特征进行分析，即在数据集上开始构建企业用户网络异常行为检测模型。本文基于机器学习算法来训练模型，但由于数据集一般是无标签的，所以一般 UEBA 的分析方法是基于无监督学习的算法。结合无监督聚类的 K 均值聚类和孤立森林算法的优势，本文使用 K 均值聚类-孤立森林算法训练用于企业用户网络异常行为的检测模型，该模型图如图 4-1 所示。

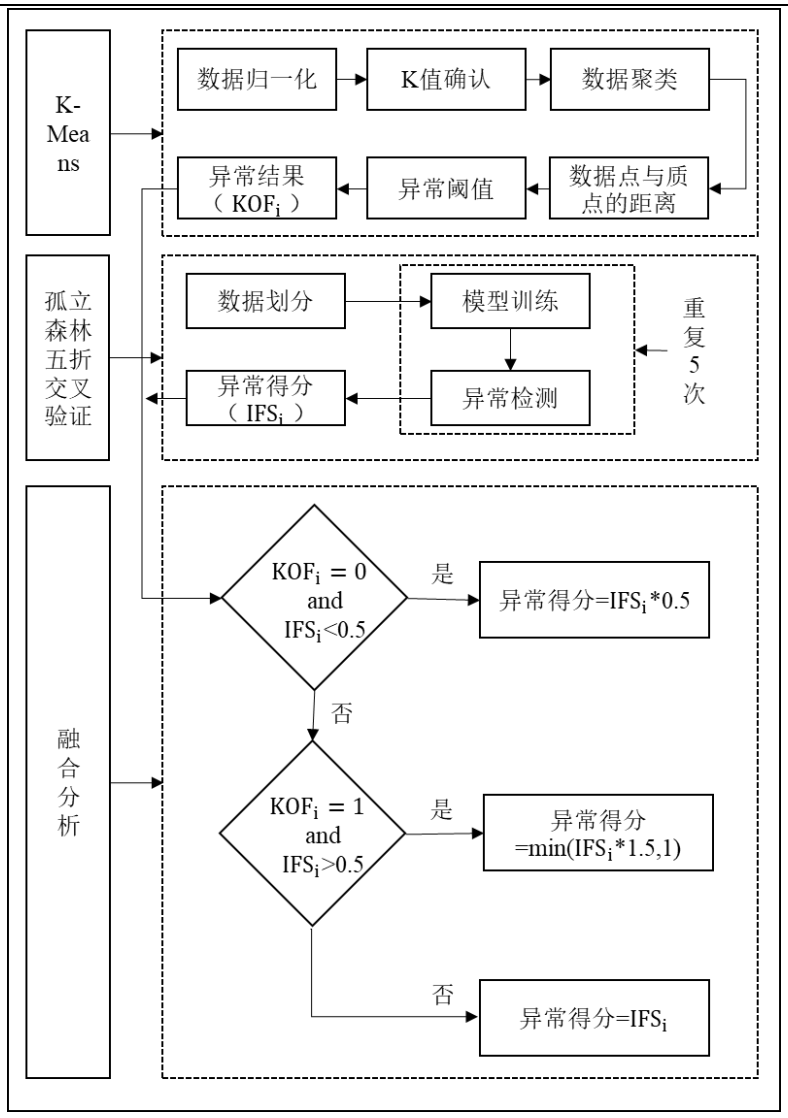


图 4-1 结合 K 均值聚类和孤立森林的异常行为检测模型图

第一步：通过 K-Means 算法检测异常行为，异常行为值为 1，正常行为值为 0：

1) 对数据进行归一化处理；

2) 根据肘部法则和轮廓系数找到适合本数据源的合理 K 值；

3) 通过 K-Means 聚类算法对数据进行聚类；

4) 计算每个数据点与其聚簇中心之间的距离，设置异常值的阈值，通过阈值来判断行为是否异常，异常行为值设为 1，正常行为值设为 0。

第二步：使用孤立森林 5 折交叉验证法对整体数据，不同的部门数据，不

同的账号数据分别建模并获取行为异常分数：

1) 首先不重复抽样将训练数据随机分成5份；

2) 拿出4份数据通过孤立森林算法训练异常检测模型，然后使用训练好的模型对1份验证数据进行异常预测，同时在整个一次的交叉验证完成之后，对测试集进行预测，总共进行5次训练；

3) 每次训练得到的异常得分进行加权平均后就是该行为的异常得分。

第三步：将两种算法的异常行为检测结果进行结合分析：

1) 如果 K-Means 算法得到的行为值为 0，孤立森林 5 折交叉验证法得到的行为异常得分小于 0.5，则该行为异常得分为 0.5*孤立森林行为异常得分，表达式如下列公式(4.1)所示：

$$\text{if } KOF_i = 0 \text{ and } IFS_i<0.5 \ then \ \text{OS}_i=IFS_i*0.5 \tag{4.1}$$

2) 如果 K-Means 算法得到的行为值为 1，孤立森林 5 折交叉验证法得到的行为异常得分大于 0.5，则该行为异常得分为取 1.5*孤立森林异常得分和 1 较小的值，表达式如下列公式(4.2)所示：

$$\text{if } KOF_i = 1 \text{ and } IFS_i>0.5 \ then \ \text{OS}_i=\min(IFS_i*1.5, 1) \tag{4.2}$$

3) 其他情况，包括：K-Means 算法得到的行为值为 0，孤立森林 5 折交叉验证法得到的行为异常得分大于 0.5，或 K-Means 算法得到的行为值为 1，孤立森林 5 折交叉验证法得到的行为异常得分小于 0.5，或者 K-Means 算法得到的行为值为 0 或 1，孤立森林 5 折交叉验证法得到的行为异常得分等于 0.5，则该用户和实体行为异常得分为孤立森林行为异常得分，表达式如下列公式(4.3)所示：

$$\text{OS}_i=IFS_i \tag{4.3}$$

其中KOF_i(K-Means Outlier Flag)表示第 i 条用户和实体行为日志通过 K-Means 算法计算得到的异常标识。IFS_i(Isolation Forest Score)表示第 i 条用户和实体行为日志通过孤立森林 5 折交叉验证法得到的行为异常得分。OS_i(Outlier Score)表示第 i 条用户和实体行为日志最终的行为异常得分。

4.2 基于 K-Means 聚类的异常行为检测

对数据进行 K-Means 聚类异常行为检测，具体检测步骤包括：

第一步：对数据进行归一化处理；

第二步：根据肘部法则和轮廓系数找到适合本数据源的合理 K 值；

第三步：通过 K-Means 聚类算法对数据进行聚类；

第四步：计算每个数据点与其聚簇中心之间的距离，设置异常值的阈值，通过阈值来判断行为是否异常，异常行为值为 1，正常行为值为 0。

4.2.1 数据归一化

归一化是对数据进行预处理，使其限定在一定的范围之内，通过归一化能够使各个维度之间的特征在数值上有一定的比较性，以使聚类效果更好。K-Means的归一化通常限定在区间[0, 1]或者区间[-1, 1]。

本研究选用的归一化方法是，StandardScaler归一化方法，本方法不是针对样本，而是针对每一个特征维度来。通过StandardScaler归一化处理后的数据符合标准正态分布，函数如下列公式(4.4)所示：

$$x = \frac{(x - \mu)}{\sigma} \tag{4.4}$$

其中μ表示所有样本的均值，σ表示所有样本的方差（或标准差）

根据特征构造、方差选择和皮尔逊相关系数法最终确定了17个特征，对这17个特征进行数据归一化，以下5条示例是17个特征经过数据归一化后的值。

表 4.1 归一化处理后的数据示例

特征	数据 1	数据 2	数据 3	数据 4	数据 5
IP_PORT_hour_max	0.652174	0.347826	0.391304	0.304348	0.347826
IP_PORT_hour_count	0.002328	0.000166	0.00266	0.003658	0.002827
switchIP_hour_count	1	1	0.92057	0.424601	0.417329
url_hour_max	0.904762	1	0.904762	1	0.809524
IP_URL_hour_max	0.391304	0.217391	0.304348	0.304348	0.391304
group_hour_mean	0.700481	0.700481	0.700481	0.700481	0.700481
group_hour_count	0	0	0	0	0
group_IP_hour_max	0.954545	1	0.954545	1	0.909091
group_IP_hour_median	0.324324	0.216216	0.27027	0.216216	0.324324
group_switchIP_hour_max	0.904762	0.904762	0.857143	0.952381	0.857143
group_url_hour_max	0.565217	0.913043	0.869565	0.478261	0.521739

特征	数据 1	数据 2	数据 3	数据 4	数据 5
group_ip_url_hour_max	0.347826	0.130435	0.304348	0.26087	0.217391
account_hour_count	0.611623	0.611623	0.611623	0.611623	0.611623
account_IP_hour_max	0.521739	0.347826	0.478261	0.347826	0.652174
account_IP_hour_count	0.036575	0.024312	0.047547	0.035929	0.02969
account_switchIP_hour_max	0.652174	0.652174	0.652174	0.652174	0.608696
account_url_hour_max	0.304348	0.304348	0.347826	0.347826	0.521739

4.2.2 K 值确认

对于K-Means算法来说，簇数K值的大小非常重要，影响聚类效果。在实际应用中，由于数据量过大且缺乏经验，K值选多少合适比较难确定。

怎样根据数据的特征确认合理的K值呢？本文通过肘部法则和轮廓系数综合评估，确认合理的K值。

肘部法则的原理是：计算不同K值对应的簇内离差平方和。簇内离差平方和是每个簇的中心点与簇内每个点的距离的平方和。对于一个簇，该值越低代表簇内成员越紧密，该值越高代表簇内结构越松散。随着K值的增加，簇内离差平方和慢慢变小，此时会有一个拐点也叫作“肘”点，当达到“肘”点时簇内离差平方和的大小会得到极大改善，随后该值的下降率慢慢变缓，这个“肘”点可以考虑为簇数K的值。为了更清晰的找到K值，将聚类个数和簇内离差平方和通过可视化方法呈现。

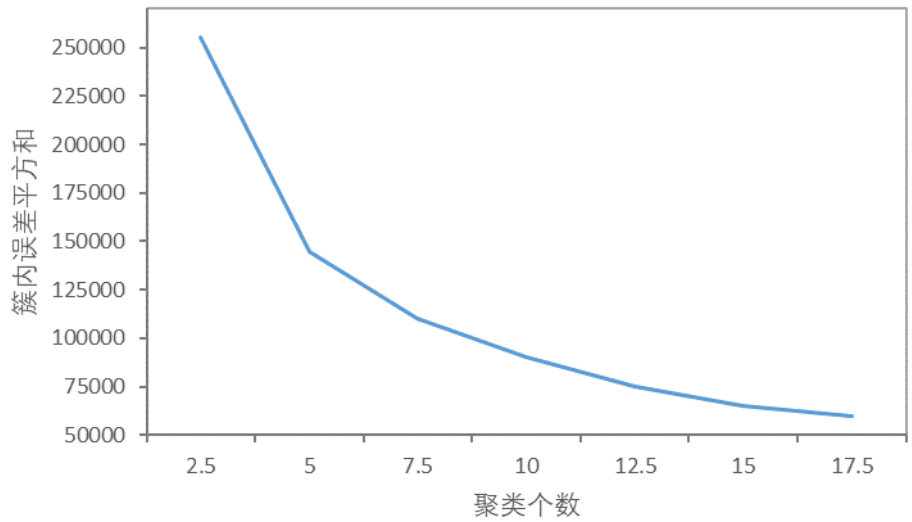


图 4-2 肘部法则估算 K-聚类中的 K 值

从图4-2中可以看出，在k在[5，12]区间时，簇内离差平方和得到大幅改善，可以考虑选取[5，12]中的一个值作为聚类数量。

轮廓系数综合考虑了簇的密集性和分散性，如果K值确定的合理，那么相同簇内的数据会很密集，不同簇间的数据会很分散。轮廓系数S的计算公式如下列公式(4.5)所示：

$$S=\frac{b-a}{\max(b,a)} \tag{4.5}$$

其中a代表样本点与同簇内其他样本点距离的平均值；b表示样本点与其他非同簇样本点的距离的平均值。通过公式(4.5)可知，S的取值范围为（-1, 1），当S接近于-1时，说明样本分配的不合理，需要将其分配到其他簇中；当S近似为0时，说明样本落在了模糊地区，即簇的边界处；当S接近1时，说明样本的分配是合理的。为了更清晰的呈现轮廓系数结果，将轮廓系数结果通过可视化方法呈现。

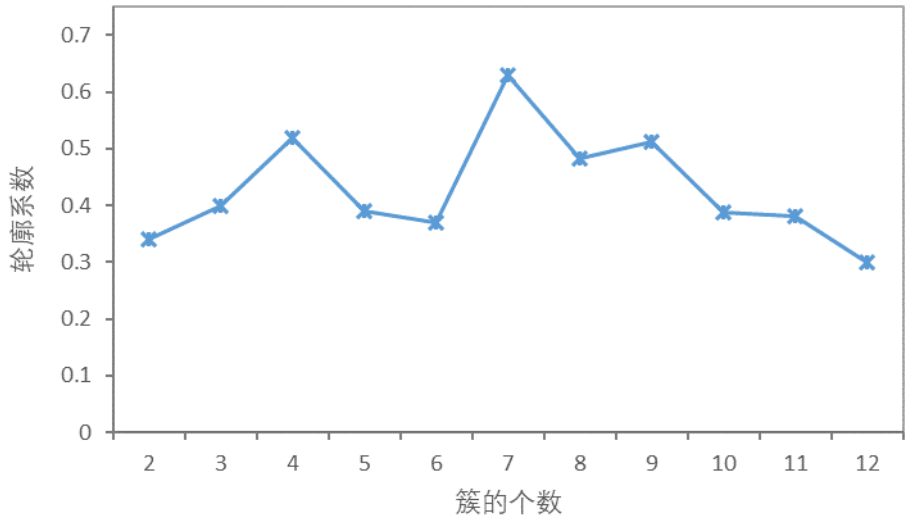


图 4-3 轮廓系数图

图4-3绘制了不同K值下对应的轮廓系数，当K等于4、7时，轮廓系数较大，通过与肘部法则选取的K值做交运算，最后本研究训练集数据K-Means聚类使用的数据K值选为7。

4.2.3 数据聚类

K-Means 聚类过程：

第一步：从数据集中随机抽取 K 个数据点作为初始的聚簇中心；

第二步：计算每个数据点和聚簇中心的欧几里得距离的平方和，将其分给当前与之最近的那个聚簇中心；

第三步：计算各聚簇内数据的均值，并将均值设为新的聚簇中心。

重复第二步、第三步，直到聚簇中心不再变化。

在训练集数据中，每个聚簇内的数量如下表 4.2 所示：

表 4.2 训练集数据聚类情况

训练集聚类标签	训练集聚类数量
0	121294
1	93001
2	46762
3	17111
4	127033
5	25236
6	38452
总计	468889

4.2.4 企业用户网络异常行为检测

基于 K-Means 聚类的企业用户网络异常行为检测步骤如下：

第一步：计算每个数据点与其最近的聚类中心之间的距离。表 4.3 是距离计算结果的 5 条示例数据。

表 4.3 数据点与其最近的聚类中心之间的距离示例数据

主键 id	聚类标签	与质点距离
20120	0	1.062875559
22152	2	1.160667393
20165	0	0.966324292
20800	0	0.925925087
22958	0	0.998813135

第二步：设置异常值比例，并根据异常值比例设置异常值数量。

统计假设检验思想中提到小概率事件是指概率小于10%、5%或1%的事件。正态分布中提到，$(\mu-3\sigma,\mu+3\sigma)$以外的概率小于0.3%，几乎是不可能发生的事件。参考以上两种理论思想，本文将异常值比例设置为1%。根据异常值比例计算得

到训练集的异常数量4689条。

第三步：确认异常距离阈值，并根据阈值判断是否异常。

将训练集中每个数据点与其最近的聚类中心之间的距离进行降序排列。训练集中取距离最大的4689条数据中最小的距离值1.893261436，距离大于等于该值的数据认为是异常数据，将K-Means聚类异常标识置为1，其它数据置为0。将训练好的模型训练测试集数据，并标记K-Means聚类异常标识。

4.3 基于孤立森林 5 折交叉验证的异常行为分数获取

4.3.1 5 折交叉验证概述

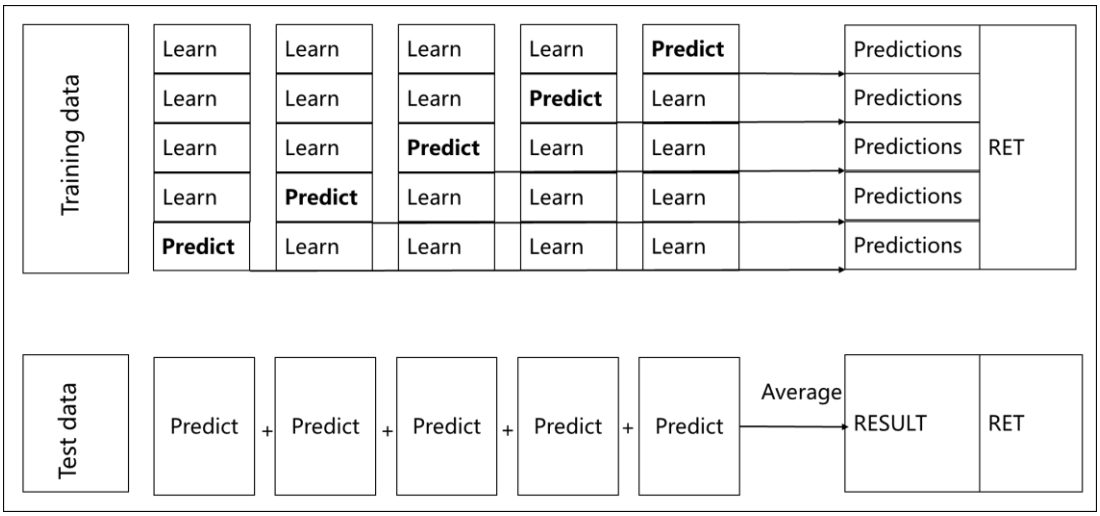


图 4-4 5 折交叉验证模型图

图4-4绘制了5折交叉验证的实施步骤。首先，将数据分成两部分，一部分是训练集，一部分是测试集。其次，不重复抽样将训练数据随机分成5份，先拿出4份作为训练数据集，另外1份作为验证数据集。再次，总共进行5次交叉验证，每次交叉验证包括两个部分，一是基于4份训练数据集对模型进行训练，二是使用训练好的模型对1份验证集数据进行异常预测，得到异常分数。最后，在整个一次的交叉验证完成之后，还要对测试集进行预测，每一次测试集上的异常分数进行加权平均后，就是最后测试集的异常分数。

4.3.2 基于整体数据的企业用户网络异常行为检测模型

通过特征构造以及方差选择方法和皮尔逊相关系数法对基础数据进行特征约简，整合得到能有效反映企业行为基线的5个特征：终端ip_应用端口_访问时

间最大值(ip_port_hour_max)，终端ip_应用端口_访问时间总数(ip_port_hour_count)，交换机ip_上网时间中位数(switchIP_hour_count)，url_上网时间最大值(url_hour_max')，终端ip_url_上网时间最大值(IP_URL_hour_max)。将企业行为基线特征和当前行为基本特征作为整体数据企业用户网络异常行为检测模型的特征。当前行为基本特征包括：终端ip地址(ip)，终端上网网址(url)，终端上网应用端口(port)，终端连接交换机IP(switchIP)，上网时间(hour)。基于企业行为基线特征和当前行为基本特征，通过孤立森林5折交叉验证的得到通过整体数据建模的企业用户网络行为异常分数。

4.3.3 基于部门数据的企业用户网络异常行为检测模型

以下以市场战略发展中心和人事行政中心两个部门在不同时间上网次数进行可视化示例分析和展示。

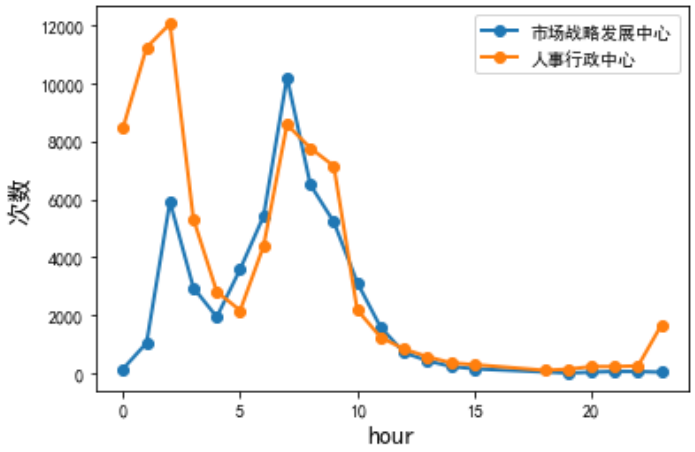


图 4-5 不同部门的上网时间分布对比图

图4-5描绘了市场战略发展中心和人事行政中心不同时间上网次数的分布情况。人事行政中心在3点左右访问次数高达12000次，而市场战略发展中心在3点左右访问次数仅6000次。在4:50之前，人事行政中心的访问次数要多于市场战略发展中心的访问次数。但4:50-8:00左右，市场战略发展中心的访问次数要多于人事行政中心的访问次数。23:00左右，市场战略发展中心几乎不上网，但是人事行政中心仍有2000次左右的访问。可见，即使是相同的特征，不同部门的行为特征值也是不一样的。因此，除了整体建模外，考虑对不同的部门不同的特征进行分析构建用户异常行为检测模型。

基于部门数据建模：首先确认模型特征，包括当前行为基本特征（“基于

整体数据的企业用户网络异常行为检测模型”中对此进行过介绍，此处不再赘述）和部门行为基线特征。部门行为基线特征包括：部门_上网时间平均值(group_hour_mean)，部门_上网时间总数(group_hour_count)，部门_终端ip_上网时间最大值(group_IP_hour_max)，部门_终端ip_上网时间平均值(group_IP_hour_median)，部门_交换机ip_上网时间最大值(group_switchIP_hour_max)，部门_url_上网时间最大值(group_url_hour_max)，部门_ip_url_上网时间最大值(group_ip_url_hour_max)。其次，根据部门维度，对数据进行划分。最后，将每个部门的数据输入孤立森林5折交叉验证模型中获取各部门中每条上网行为的异常分数。

4.3.4 基于账号数据的企业用户网络异常行为检测模型

以下以chengge@qq.com和yangxiaohong@qq.com两个不同账号在不同时间上网次数进行可视化示例分析和展示。

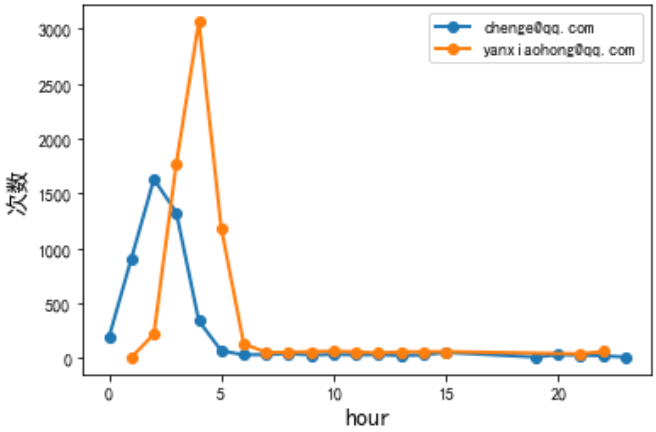


图 4-6 不同账号的上网时间分布对比图

图4-6描绘了账号chenge@qq.com和yangxiaohong@qq.com不同时间上网次数的分布情况。账号chenge@qq.com在2点的时候上网次数最多,最多每小时访问1600次，yangxiaohong@qq.com在5点的时候上网次数最多，最多每小时访问300次。有此可见，即使是相同的特征，不同账号的行为特征也是不一样的。因此，除了整体建模、不同的部门建模外，考虑对不同的账号不同的特征进行分析构建用户异常行为检测模型。

基于账号数据建模：首先确认模型特征，包括当前行为基本特征（“基于整体数据的企业用户网络异常行为检测模型”中对此进行过介绍，此处不再赘

述）和账号行为基线特征。账号行为基线特征包括：个人账号_上网时间总数(account_hour_count)，个人账号_终端ip_上网时间最大值(account_IP_hour_max)，个人账号_终端ip_上网时间总数量(account_IP_hour_count)，个人账号_交换机ip_上网时间最大值(account_switchIP_hour_max)，个人账号_url_上网时间最大值(account_url_hour_max)。其次，根据个人账号维度，对数据进行划分。最后，将每个账号的数据输入孤立森林5折交叉验证模型中获取各账号中每条上网行为的异常分数。

4.4 结合 K-Means 和孤立森林算法的异常行为结果融合分析

通过上述操作获取了每条上网行为日志对应的K-Means聚类异常标识和孤立森林计算得到的行为异常分数。接下来是将两种算法的异常检测结果进行融合分析。同为正常，说明该条数据为正常行为的概率较大，故将孤立森林计算得到的异常分数降低50%做为该行为最终的异常分数。同为异常，说明该条数据为异常行为的概率较大，考虑将孤立森林计算得到的异常分数进行放大，而异常分数的上限为1（也就是100%为异常）。故同为异常时，行为最终的异常分数取孤立森林计算得到的异常分数乘以1.5和1中较小值。其他情况以孤立森林模型计算的异常分数作为该行为最终的异常分数。详细过程见本章第一节“结合K均值聚类-孤立森林算法的异常行为检测模型图”中描述，本节不再赘述。

4.5 本章小结

本章主要介绍了结合K均值聚类和孤立森林算法的异常行为检测模型。首先对模型构建进行了概述。接着介绍了K-Means聚类的异常行为检测，包括数据归一化，K值确认及企业用户网络异常行为检测方法。然后介绍了基于孤立森林5折交叉验证的异常行为分数获取，包括5折交叉验证概述，基于整体数据的企业用户网络异常行为检测模型，基于部门数据的企业用户网络异常行为检测模型和基于账号数据的企业用户网络异常行为检测模型。最后介绍了结合K-Means和孤立森林算法的异常行为结果融合分析，计算对应行为的最终异常分数。

第 5 章 实验与结果分析

5.1 模型评价指标

本文使用的模型评估是均方根误差和平均绝对误差。

均方根误差的计算是用每一个实例的真实值减去预测值所得到偏差值的平方累加和除以观测次数的平方根。均方根误差是用来衡量观测值同真值之间的偏差，算式如下列公式(5.1)所示：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{X}_i-\mathrm{Y}_i)^2} \tag{5.1}$$

公式(5.1)中，X_i表示真实值，Y_i表示预测值，RMSE的范围是[0,+∞)，RMSE的值越大，偏差越大，模型的准确度越低，RMSE的值越小，偏差越小，模型的准确度越高，当预测值与真实值无偏差时RMSE等于0，即完美模型（实际上几乎从未实现）。

平均绝对误差的计算是所有预测值与真实的偏差的绝对值的平均值，用来反映实际预测误差的大小，算式如下列公式(5.2)所示：

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|\mathrm{X}_i-\mathrm{Y}_i| \tag{5.2}$$

公式(5.2)中，X_i表示真实值，Y_i表示预测值， MAE的值越大，偏差越大，模型的准确度越低，MAE的值越小，偏差越小，模型的准确度越高。

正常采集到的企业内部用户和实体行为日志信息是无异常标签。但是为了验证模型的准确性，本研究对使用的数据进行了人工标注，即3.1“数据集介绍”介绍小节中的第10个字段ret字段，该值取值范围是[0,1]。本研究中，计算RMSE值和MAE值时，真实值取的该人工标注的值，预测值是通过K均值聚类-孤立森林5折交叉验证模型得到的异常得分值。

5.2 实验环境

本文实验过程中的相关实验环境，包括CPU、硬盘、内存、编程语言、编程

工具、主要使用的库和包，如下表5.1所示。

表 5.1 实验环境

硬件		软件	
CPU	intel i7	系统	win10
硬盘	512GB 固态硬盘	编程语言&工具	python3.7,Anaconda 4.10.1
内存	16GB	主要使用的库&包	sklearn, numpy,pandas,matplotlib

5.3 不同异常行为检测模型的实验与结果分析

使用K均值聚类-孤立森林5折交叉验证法和孤立森林5折交叉验证法分别对整体数据建模，不同的部门数据建模，不同的账号数据建模，并计算每种模型的RMSE值和MAE值。

5.3.1 基于整体数据的异常行为检测模型实验与结果分析

使用K均值聚类-孤立森林算法作为基础模型的5折交叉验证方法对数据的整体建模，和使用孤立森林算法作为基础模型的5折交叉验证方法对数据的整体建模，在训练集和测试集上得到的结果如下表5.2所示：

表 5.2 整体数据建模实验结果

数据集	K 均值聚类-孤立森林	孤立森林
训练集 RMSE	0.219391	0.255425
测试集 RMSE	0.225347	0.251737
训练集 MAE	0.167358	0.219336
测试集 MAE	0.171624	0.217007

K均值聚类-孤立森林5折交叉验证整体数据建模，在训练数据集上的均方根误差值为0.219391，平均绝对误差值为0.167358，测试数据集上的均方根误差值为0.225347，平均绝对误差值为0.171624。

孤立森林5折交叉验证整体数据建模，在训练数据集上的均方根误差值为0.255425，平均绝对误差值为0.219336，测试数据集上的均方根误差值为0.251737，平均绝对误差值为0.217007。

5.3.2 基于部门数据的异常行为检测模型实验与结果分析

使用K均值聚类-孤立森林算法作为基础模型的5折交叉验证方法对不同部门的数据建模，和使用孤立森林算法作为基础模型的5折交叉验证方法对不同部

门数据建模，在训练集和测试集上得到的RMSE和MAE结果分别如下表5.3和表5.4所示：

表 5.3 不同的部门数据建模实验结果-RMSE

部门	K 均值聚类-孤立森林	孤立森林
市场战略发展中心	0.203533	0.234634
研发中心	0.214900	0.271340
人事行政中心	0.213941	0.263272
渠道生态合作事业部	0.224998	0.231321
业务创新中心	0.208743	0.233086
政企事业部	0.223087	0.272567
通用市场部	0.198257	0.272138
各部门平均值	0.212494	0.254051
测试集	0.230979	0.264656

表 5.4 不同的部门数据建模实验结果-MAE

部门	K 均值聚类-孤立森林	孤立森林
市场战略发展中心	0.152688	0.203882
研发中心	0.166464	0.240391
人事行政中心	0.168558	0.220569
渠道生态合作事业部	0.172907	0.191404
业务创新中心	0.175936	0.198592
政企事业部	0.172091	0.237530
通用市场部	0.154024	0.241425
各部门平均值	0.166095	0.219113
测试集	0.177685	0.220545

使用K均值聚类-孤立森林5折交叉验证法对不同的部门建模在训练数据集上市场战略发展中心的均方根误差值为0.203533，平均绝对误差值为0.152688，研发中心的均方根误差值为0.214900，平均绝对误差值为0.166464，人事行政中心的均方根误差值为0.213941，平均绝对误差值为0.168558，渠道生态合作事业部的均方根误差值为0.224998，平均绝对误差值为0.172907， 业务创新中心的均方根误差值为0.208743，平均绝对误差值为0.175936，政企事业部的均方根误差值为0.223087，平均绝对误差值为0.172091，通用市场部的均方根误差值为0.198257，平均绝对误差值为0.154024。通过分部门建模方式在训练数据集上的加权平均后得到的均方根误差值为0.212494，平均绝对误差值为0.166095，在测试数据集上的均方根误差值为0.230979，平均绝对误差值为0.177685。

使用孤立森林5折交叉验证法对不同的部门建模在训练数据集上市场战略发展中心的均方根误差值为0.234634，平均绝对误差值为0.203882，研发中心的均方根误差值为0.271340，平均绝对误差值为0.240391，人事行政中心的均方根误差值为0.263272，平均绝对误差值为0.220569，渠道生态合作事业部的均方根误差值为0.231321，平均绝对误差值为0.191404， 业务创新中心的均方根误差值为0.233086，平均绝对误差值为0.198592，政企事业部的均方根误差值为0.272567，平均绝对误差值为0.237530，通用市场部的均方根误差值为0.27213，平均绝对误差值为0.241425。通过分部门建模方式在训练数据集上的平均均方根误差值为0.254051，平均绝对误差值为0.219113，在测试数据集上的均方根误差值为0.264656，平均绝对误差值为0.220545。

5.3.3 基于账号数据的异常行为检测模型实验及结果分析

使用K均值聚类-孤立森林的5折交叉验证方法对不同的账号建模，和使用孤立森林的5折交叉验证方法对不同的账号建模，最后在训练集（由于账号比较多，展示10个账号的结果）和测试集上得到的RMSE和MAE结果分别如下表5.5和表5.6所示：

表 5.5 不同的账号数据建模实验结果-RMSE

账号	K 均值聚类-孤立森林	孤立森林
gongxiaolong@qq.com	0.285791	0.320515
guowenjian@qq.com	0.171815	0.182271
hanxiang@qq.com	0.150326	0.166434
hanzhenguo@qq.com	0.217053	0.163543
haoxiaofang@qq.com	0.178699	0.180755
hecailing@qq.com	0.215029	0.384534
huangjunsheng@qq.com	0.190708	0.178019
jiakaiqiang@qq.com	0.161264	0.182060
jibin@qq.com	0.140521	0.222712
jingai@qq.com	0.161073	0.158528
.....	.....	.....
训练集	0.186232	0.229136
测试集	0.229619	0.290042

表 5.6 不同的账号数据建模实验结果-MAE

账号	K 均值聚类-孤立森林	孤立森林
gongxiaolong@qq.com	0.188669	0.296450
guowenjian@qq.com	0.126129	0.159083
hanxiang@qq.com	0.124077	0.135167
hanzhenguo@qq.com	0.172035	0.131963

账号	K 均值聚类-孤立森林	孤立森林
haoxiaofang@qq.com	0.143585	0.148125
hecailing@qq.com	0.199533	0.378537
huangjunsheng@qq.com	0.145199	0.149342
jiakaiqiang@qq.com	0.122676	0.159395
jibin@qq.com	0.113408	0.194035
jingai@qq.com	0.128829	0.133200
.....		.....
训练集	0.149231	0.205106
测试集	0.189932	0.253427

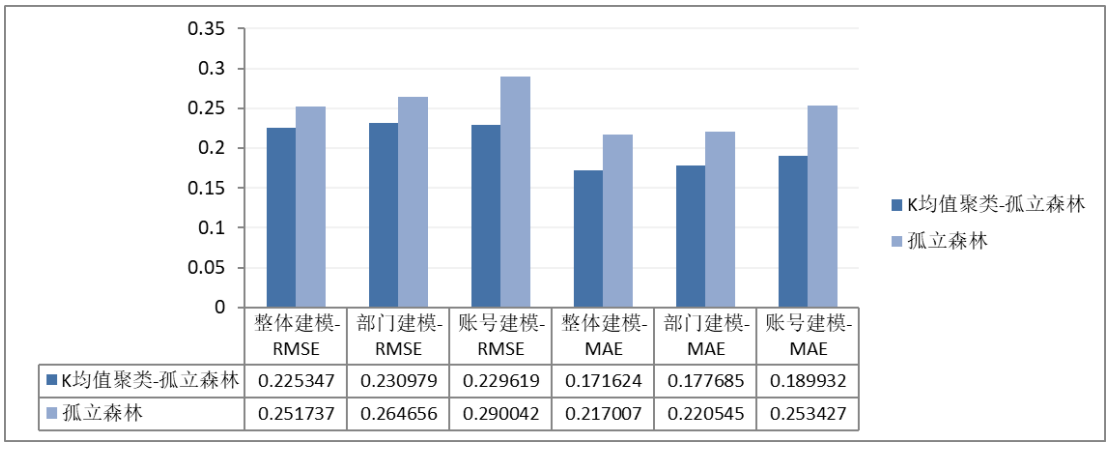
使用K均值聚类-孤立森林5折交叉验证法对不同的账号建模，通过各账号在训练数据集上的加权平均后得到的均方根误差值为0.186232，平均绝对误差值为0.149231，在测试数据集上的均方根误差值为0.229619，平均绝对误差值为0.189932。

使用孤立森林5折交叉验证法对不同的账号建模，通过各账号在训练数据集上的加权平均后得到的均方根误差值为0.229136，平均绝对误差值为0.205106，在测试数据集上的均方根误差值为0.290042，平均绝对误差值为0.253427。

5.4 两种算法构建的不同模型的对比实验与结果分析

分别对比K均值聚类-孤立森林算法和孤立森林算法对整体数据建模、不同部门数据建模、不同账号建模在测试集上的RMSE值和MAE值如下图5.1所示。

图 5.1 实验结果对比



通过K均值聚类-孤立森林5折交叉验证法对整体数据建模、不同部门建模、不同账号建模在测试集上的均方根误差值分别为：0.225347、0.230979、0.229619，

平均绝对误差值为：0.171624、0.177685、0.189932。通过孤立森林5折交叉验证法对整体数据建模、不同部门建模、不同账号建模在测试集上的均方根误差值分别为0.251737、0.264656、0.290042，平均绝对误差值为：0.217007、0.220545、0.253427。

实验结果表明基于整体数据的异常检测模型要优于分部门和分账号的异常检测模型。在数据集相同的情况下，基于K均值聚类-孤立森林算法的异常行为检测模型的均方根误差值和平均绝对误差值比基于孤立森林算法的行为检测模型的均方根误差值和平均绝对误差值更小，异常检测模型更精准。

5.5 本章小结

本章主要介绍基于K均值聚类-孤立森林5折交叉验证对整体数据、不同的部门数据、不同的账号数据建模的实验结果，并与同数据集下孤立森林5折交叉验证对整体数据、不同的部门数据、不同的账号数据建模的实验结果做对比分析。实验结果表明在数据集相同的情况下，K均值聚类-孤立森林算法对整体数据建模的异常检测结果更精准。

第 6 章 总结与展望

6.1 总结

互联网的迅猛发展和数字经济的快速推进，企业数据也呈爆炸式的增长，企业关键数据的泄露会对个人甚至国家造成巨大的影响。所以，及时、准确、有效的识别或预测企业用户网络异常行为是非常重要且有意义的课题。本文的研究是在阅读大量关于网络异常行为检测及UEBA相关的文献下，结合国内外研究学者总结的机器学习理论和对机器学习的研究进展，以用户和实体为对象，利用机器学习技术对用户行为进行分析和异常检测，快速感知用户的异常行为。本文的主要研究工作和贡献如下：

（1）本文主要工作：首先对企业员工日常访问行为记录日志进行分析，通过特征构造和特征选择方法，构建能反映企业行为的基线特征、部门行为的基线特征、个人行为的基线特征。然后利用机器学习技术分别对当前行为与企业行为基线、当前行为与部门行为基线、当前行为与个人行为基线进行建模、异常检测。最后通过均方根误差和平均绝对误差对模型进行评估。

（2）本文的主要贡献：第一，利用用户和实体行为分析（UEBA）方法对企业员工日常访问行为记录进行异常评分，多维度多角度的构建企业正常活动基线，识别异常的访问行为。第二，提出了K均值聚类-孤立森林算法的异常行为检测模型，并与孤立森林算法的异常检测结果做对比实验，验证本文提出的检测方法对企业用户网络异常行为的检测更精准。

6.2 展望

未来，不管信息化技术如何发展，网络安全作为个人、企业、国家的基本屏障，只会越来越重要。UEBA作为识别异常行为的一种新方法，未来会被广泛集成到各种企业安全产品中。本文的研究虽然取得了初步的成功，但尚有许多有待进一步深入进行的研究工作，这里择其要者简要讨论如下：

1、目前的算法检测异常的结果准确性不能达到100%，也就是说存在一定的误报，后续可以进一步的优化特征和算法，提高检测准确性。

2、本文研究的数据，仅企业上网行为日志，未来还可以继续丰富数据源，

更多维度的检测异常。

致　谢

在论文完成之际，首先要感谢的是我的导师邱桃荣教授。工作十年后再进校园学习已是一件十分值得庆幸的事，更为庆幸的是，我遇到了一位好导师。导师渊博的专业知识、严谨的治学态度、精益求精的工作作风、诲人不倦的高尚师德，严于律己、宽以待人的崇高风范，朴实无华、平易近人的人格魅力对本人影响深远。不仅使本人树立了远大的学习目标、掌握了基本的研究方法，还使本人明白了许多为人处事的道理。在此，谨向导师表示崇高的敬意和衷心的感谢！

同时我还要感谢我的家人在此期间给予我的包容、关爱和鼓励，我的每一步都离不开他们在背后默默的付出，是他们给了我继续坚持的决心和勇气。

最后，由于我的学术水平有限，所写论文难免有不足之处，恳请各位老师和学友批评、斧正！

苏文英

2022 年 3 月

参考文献

[1] Swarnkar M, Hubballi N. OCPAD: One class Naive Bayes classifier for payload based anomaly detection [J]. Expert Systems with Applications, 2016, 64: 330-339.

[2] Wang H, Gu J, Wang S. An effective intrusion detection framework based on SVM with feature augmentation [J]. Knowledge-Based Systems, 2017, 136: 130-139.

[3] Subba B, Biswas S, Karmakar S. A neural network based system for intrusion detection and attack classification [C]. 2016 Twenty Second National Conference on Communication (NCC). IEEE. 2016.

[4] Saeed A, Ahmadinia A, Javed A, etal. Intelligent intrusion detection in low-power IoTs [J]. ACM Transactions on Internet Technology, 2016, 16(4): 1-25.

[5] David J, Thomas C. DDoS attack detection using fast entropy approach on flow- based network traffic [J]. Procedia Computer Science, 2015, 50: 30-36.

[6] 宁亚飞,赵英亮,吴美荣,等.时空卷积自编码网络异常行为检测[J].国外电子测量技术, 2020, 39(10): 104-108.

[7] 刘良鑫,林勉芬,钟良泉,等.基于 3D 双流卷积神经网络的异常行为检测[J].计算机系统应用, 2021, 30(05): 120-127.

[8] Singh K, Singh P, Kumar K. User behavior analytics-based classification of application layer HTTP-GET flood attacks [J]. Journal of Network and Computer Applications, 2018, 112: 97-114.

[9] 陆英.Gartner:2018 年十大安全项目详解(二)[J].计算机与网络, 2018, 44(23): 48-50.

[10] Martin Alejandro G, Beltrun Marta, Fernandez-Isabel Alberto, etal. An approach to detect user behaviour anomalies within identity federations [J]. Computers & Security, 2021, 108: 1-18

[11] Madhu Shashanka Niara, Inc, Min-Yi Shen, etal. User and Entity Behavior Analytics for Enterprise Security [C]. 2016 IEEE International Conference on Big Data (Big Data) . IEEE. 2016.

[12] 美创发布新一代数据安全管理平台[J].中国信息安全, 2021, (05): 90.

[13] 胡绍勇.基于 UEBA 的数据泄漏分析[J].信息安全与通信保密, 2018, (08): 26-28.

[14] 吴宏胜.基于可信计算和 UEBA 的智慧政务系统[J].信息网络安全, 2020, 20(01): 89-93.

[15] 徐飞,徐志斌,徐嘉宁,等.用户异常行为分析在智慧公路系统的应用[J].中国交通信息化, 2021, (05): 130-133.

[16] 莫凡,何帅,孙佳,等.基于机器学习的用户实体行为分析技术在账号异常检测中的应用[J].通信技术, 2020, 53(05): 1262-1267.

[17] 张兆信.计算机网络安全与应用技术[M].机械工业出版社, 2017: 902-904.

[18] 彭新光,王峥.信息安全技术与应用[M].人民邮电出版社, 2013.

[19] Jr G F, Rodrigues J J P C, Carvalho L F, etal. A comprehensive survey on network anomaly detection [J]. Telecommunication Systems, 2018, 70: 447-489.
[20] 石祖文.大型互联网企业安全架构[M]. 电子工业出版社, 2020.
[21] 周巍,崔艳林,蔡新雷,等.基于 k-Means 算法的电网调度辅助决策平台[J].自动化与仪器仪表, 2020, (11): 137-140.
[22] 吴信东,库玛尔.数据挖掘十大算法[M].清华大学出版社, 2014.
[23] 刘顺祥.从零开始学 python 数据分析与挖掘[M].清华大学出版社, 2018: 326-343.
[24] 梅子行,毛鑫宇.写给风控师的实操手册[M].机械工业出版社, 2020.
[25] 刘李梦玮.基于孤立森林的网络数据流在线异常检测方法研究[D].北京邮电大学, 2021.
[26] 黄源,涂旭东.数据清洗[M].机械工业出版社, 2020: 119-121.
[27] Li JD, Cheng KW, Wang SH, etal. Feature selection: A data perspective [J]. ACM Computing Surveys, 2018, 50(6): 94.
[28] Firdaus A, Anuar NB, Razak MFA, etal. Bio-inspired computational paradigm for feature investigation and malware detection: Interactive analytics [J]. Multimedia Tools and Applications, 2017, 77(14): 17519-17555.
[29] Ahmed M, Naser Mahmood A, Hu J. A survey of network anomaly detection techniques [J]. Journal of Network and Computer Applications, 2016, 60: 19-31.
[30] 于旭,王前龙,徐凌伟,等.基于有效特征子集提取的高效推荐算法[J].计算机系统应用, 2019, 28(07): 2-2.
[31] 杭州安恒信息技术股份有限公司.用户实体行为分析技术（UEBA）[R].中国信息通信研究院安全研究所, 2020.
[32] 涂伟阳. SDN 网络架构下基于网络异常行为的 DDoS 攻击检测方法研究[D].华中师范大学,2021.
[33] 陈俊杰.基于近邻传播的网络异常行为检测算法设计及应用[D].北京邮电大学, 2020.
[34] 王雪宁.基于深度森林的网络异常行为检测方法研究与实现[D].北京邮电大学, 2020.
[35] 尹隽,彭艳红,陆怡,等.基于深度神经网络的企业信息系统用户异常行为预测[J].管理科学, 2020, 33(01).
[36] 董文静.K-Means 算法综述[J].信息与电脑(理论版), 2021, 33(11): 76-78.
[37] 杨俊闯,赵超.K-Means 聚类算法研究综述[J].计算机工程与应用, 2019, 55(23): 7-14+63.
[38] 丛思安,王星星.K-Means 算法研究综述[J].电子技术与软件工程, 2018, (17): 155-156.
[39] 王千,王成,冯振元,等.K-Means 聚类算法研究综述[J].电子设计工程, 2012,20(07): 21-24.
[40] 侯泳旭,段磊,秦江龙,等.基于 Isolation Forest 的并行化异常探测设计[J].计算机工程与科学, 2017, 39(02): 236-244.
[41] 王诚,狄萱.孤立森林算法研究及并行化实现[J].计算机技术与发展, 2021, 31(06): 13-18.
[42] 吴元君.基于孤立森林挖掘算法的入侵检测系统研究[J].盐城工学院学报(自然科学版), 2020, 33(04):24-29.
[43] 肖峰.基于孤立森林算法的计算机网络潜在攻击检测方法[J].河北北方学院学报(自然科学版), 2021, 37(11): 13-18.

[44] 张珂嘉,黄树成.一种改进的 K-means 入侵检测算法[J].计算机与数字工程, 2021, 49(10): 1963-1966+2047.
[45] 严南.基于 K-means 算法的网络入侵信息分层检索系统设计[J].信息与电脑(理论版), 2021, 33(15): 38-40.
[46] 刘凤,戴家佳,胡阳.基于局部密度离群点检测 k-means 算法[J].重庆工商大学学报(自然科学版), 2021, 38(04): 30-35.
[47] 罗俊.基于企业应用的 K-means 算法的实现与改进[J].电脑知识与技术, 2021, 17(18): 29-31.
[48] 高德平.基于孤立森林的移动终端网络数据异常检测[J].信息技术, 2021(06): 125-129.
[49] 戴军.一种基于 K-Means 聚类的离群点检测方法[J].中国计量, 2021, (06): 102-103.
[50] 白璐,赵鑫,孔钰婷,等.谱聚类算法研究综述[J].计算机工程与应用, 2021, 57(14): 15-26.
[51] 孟庆杰,尧海昌.基于谱聚类算法的信息资产行为异常检测方法[J].南京理工大学学报, 2021, 45(02): 205-213.
[52] 乌旭东,容晓峰,仲姝锜.K-Means 和 KNN 日志异常检测方法[J].西安工业大学学报, 2021, 41(02):213-218.
[53] 李倩,韩斌,汪旭祥.基于模糊孤立森林算法的多维数据异常检测方法[J].计算机与数字工程, 2020, 48(04): 862-866.
[54] 司德睿,华程,杨红光,等.一种基于机器学习的安全威胁分析系统[J].信息技术与网络安全, 2019, 38(04): 37-41.
[55] 周青松.企业内部网络用户的异常行为分析[D].华中科技大学, 2016.

攻读学位期间的研究成果

已发表论文：

1. 苏文英.工程项目管理中大数据挖掘的应用研究[J].休闲, 2021, 249: 216-216.