面向种群的 Android 应用风险评估研究

余 琴[1],李 涛[2*],王 颉[3],万振华[3]
(1. 武汉科技大学 计算机科学与技术学院,湖北 武汉 430065;
2. 智能信息处理与实时工业系统湖北省重点实验室,湖北 武汉 430065;
3. 深圳开源互联网安全技术有限公司,广东 深圳 518000)

摘 要:应用程序的权限使用情况是进行安全风险评估和检测的重要因素,但权限调用合理性是一个不确定问题。不同功能的应用程序申请的权限是不同的,单个的应用程序很难判断所申请的权限是否满足最小特权原则。针对这一问题,提出了一种面向种群的 Android 应用风险评估模型。从种群的角度,判定申请的权限是否满足此类应用程序的基本特征行为。首先建立权限使用情况、评分值、下载量、好评率等多维度评价指标体系,对应用程序进行权限特征分析并使用聚类算法实现权限风险评估。为了提高风险评估聚类结果的可解释性,使用决策树进行调整,增强合理性判定。对比于仅研究权限使用情况,采用决策树将多方面信息纳入应用程序风险评估的研究范围,实验结果可以进一步准确有效地检测出应用程序的风险程度。

关键词:Android 应用;种群;风险评估;聚类;决策树
中图分类号:TP316;TP309 **文献标识码**:A **文章编号**:1673-629X(2022)12-0007-05
doi:10.3969/j.issn.1673-629X.2022.12.002

Research on Population-based Android Application Risk Assessment

YU Qin[1],LI Tao[2*],WANG Jie[3],WAN Zhen-hua[3]
(1. School of Computer Science and Technology,Wuhan University of Science and Technology,
Wuhan 430065,China;
2. Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System,
Wuhan 430065,China;
3. Shenzhen Kaiyuan Internet Security Technology Co.,Ltd.,Shenzhen 518000,China)

Abstract:The usage of application permission is an important factor for security risk assessment and detection,but the rationality of permission invocation is an uncertain problem. Applications with different functions have different permissions. It is difficult to judge whether the requested permissions meet the minimum privilege principle for a single application. To solve this problem,we propose a population-oriented Android application risk assessment model. From the perspective of population,we determine whether the applied permission meets the basic characteristic behavior of such applications. Firstly,a multi-dimensional evaluation index system such as the usage of permission,score,evaluation and downloads is established,the permission characteristics of the application is analyzed,and the clustering algorithm is used to realize the risk assessment of permission. In order to improve the interpretability of clustering results,decision tree is used to adjust and enhance rationality judgment. Compared with only studying the usage of permission,the decision tree is used to bring various information into the research scope of application risk assessment. The experimental results can further accurately and effectively detect the risk degree of application.

Key words:Android application;population;risk assessment;clustering;decision tree

0 引 言

近些年来,中国大力推动互联网普及工作,手机渐渐成为人们日常生活不可或缺的一部分。Felt 等人[1]对 940 个应用程序进行研究分析,发现大约有 1/3 的应用程序存在过度调用隐私权限的情况。《移动应用安全形势分析报告(2020 年)》[2]显示,2020 年度收录存在安全漏洞威胁的 APK 860 万余个,同一 App 普遍存在多个漏洞。其中流氓行为、恶意扣费、广告推送、

收稿日期:2021-12-15 **修回日期**:2022-04-19
基金项目:国家自然科学基金资助项目(61702383);湖北省教育厅重大项目(17ZD014)
作者简介:余 琴(1999-),女,研究方向为信息安全;通讯作者:李 涛(1979-),男,博士,教授,研究方向为推荐系统人工智能、信息安全。

隐私信息泄露等恶意行为高达 81%,对用户的个人信息以及财产安全带来巨大威胁。该文提出了一种面向种群的 Android 应用风险评估模型。从第三方应用市场收集多种功能的应用程序进行分析和研究,根据特征权限使用情况及权限等级进行风险预估,利用 k-means 算法划分风险等级,最后使用决策树对聚类结果进行调整和优化。

1 相关研究

Android 目前使用的权限机制在数据安全保护方面并没有取得很好的效果[3]。基于权限管理的安全机制是 Android 隐私保护的重要组成部分[4],故权限机制成为判断应用程序是否存在恶意行为的一个重要因素。

文献[5]提出了一种轻量级的快速检测方法,根据分析特征权限在恶意应用和正常应用中使用的频率来定量其恶意值,通过计算样本恶意值与规定的阈值进行比较来判断样本的恶意程度。文献[6]提取应用程序声明和自定义的各类权限以及动态检测获取执行过程中使用到的权限,利用层次分析法计算权重,评估应用风险值。但是,自定义权限在第三方应用程序中使用频繁,卜同同等[6]收集的软件样本有限,因此难免会出现自定义权限数据集不充分的问题。文献[7]设计了一种挖掘权限频繁项集的算法来处理应用程序权限列表,构建权限特征关系库。但是将不同功能的应用混合在一起构建数据集,忽略了应用程序的差异,即没有考虑到不同权限对不同功能的应用程序的敏感程度是不同的。

传统的应用检测方法主要是从个体的角度来判断应用的安全性,难以满足大规模用户检测的需要[8]。文献[9]借鉴生物学中种群的概念,提出了一种面向种群的适用于大规模 Android 应用评估和恶意应用检测的方法。通过群体特征分析和种群聚类,进行高效大规模的应用隐私泄露风险评估。但是,聚类结果缺乏可解释性。

2 基于聚类和决策树的应用风险评估方法

2.1 种群风险评估模型框架

该文提出了一种面向种群的 Android 应用风险评估模型,可以根据应用程序标签信息将爬取到的应用划分进不同的种群,并自动提取应用权限列表。然后,根据权限使用情况和 Android 系统权限等级对未知应用进行评估并给出对应的风险等级,为用户在选择 Android 应用时提供一个较好的参考依据。整个模型主要由数据采集模块和风险评估模块组成,系统的流程框架如图 1 所示。

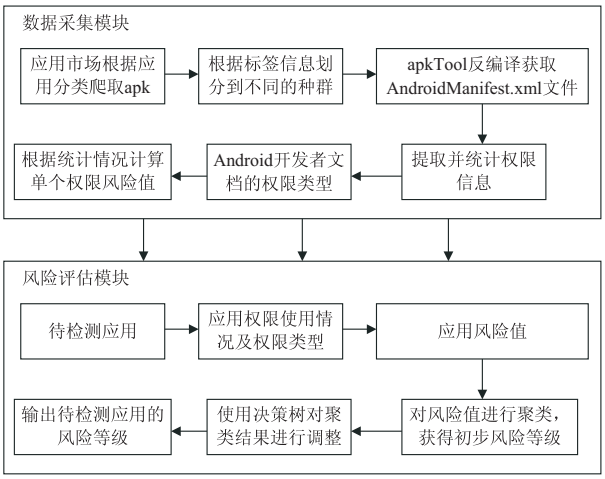

图 1 系统流程框架

具有相似功能的应用程序类似一个种群,它们所需的系统权限也相似。在数据采集模块,使用 python 编写的爬虫爬取网站提供的应用介绍、评分值、好评率、下载量等相关信息,同时,根据类别标签对应用程序进行分类爬取下载,并标上新的群体标签。把群体作为处理单元,使用逆向分析工具对 APK 进行反编译得到 AndroidManifest. xml 文件,获取应用程序的权限或权限组的信息。将获取到的所有信息存储在云端数据库中,并建立多维度评价指标体系。利用统计学相关知识对权限特征集合进行静态分析,根据统计情况计算单个权限风险指数,从而获得种群中应用程序的风险值。然后,利用数据挖掘中的 k-means 算法对应用程序风险值进行聚类,获得初步风险等级。为了提高风险评估聚类结果的可解释性,使用决策树将多维度评价指标体系纳入应用程序风险评估的研究范围,增强权限调用合理性判定。

2.2 风险值计算

通过研究 Android 应用权限的使用情况发现,不同功能的应用在权限申请方面存在较大的差异。将相同功能类型的应用作为一个种群,种群类别分为 x 个类,种群的类别会根据爬取 App 总数的增加而变化,于是类别集合为[8]:

定义 1 类别标签:Class = $\{C_1, C_2, \cdots, C_x\}$, C_x 为每个种群的类别标签,如新闻阅读、主题壁纸、社交聊天等类别。

定义 2 应用个体:Application = $\{C_x$, Introduction, Score, Download, Evaluate, PermissionMatrix$\}$ 。其中,C_x 为种群的类别标签,Introduction 为应用介绍,Score 为应用评分,Download 为下载量,Evaluate 为评价信息,PermissonMatrix 为每个应用经过权限预处理后的权限信息矩阵,具体定义如下:

定义 3 PermissionMatrix = $\{P_{ij} \mid i = 1, 2, \cdots, n; j = 1, 2, \cdots, m\}$,在权限矩阵中,如果应用个体拥有权限

j,那么 P_{ij} 为 1,否则 P_{ij} 为 0。

定义 4 权限申请率 Permission_rate:

$$\text{Permission\_rate} = \frac{P_i\text{\_Num}}{C_x\text{\_Sum}}, 0 < P(C_i) \leqslant 1$$

其中,C_x_Sum 为种群 C_x 中应用程序总数,P_i_Num 为权限 P_i 在种群 C_x 中出现的次数。

定义 5 权限等级风险值:Android 将权限划分成不同的类型,包括安装时权限、运行时权限和特殊权限[10]。不同类型权限的风险程度是不同的,比如 READ_CONTACTS(读取手机联系人)属于危险权限,存在泄露用户隐私信息的可能性越大。现将权限等级风险值 Grisk_i 定义如下:

$$\text{Grisk}_i = \begin{cases} 10, & i \in \text{安装时权限} \\ 50, & i \in \text{运行时权限} \\ 100, & i \in \text{特殊权限} \end{cases}$$

该文从 Android 开发者文档[10]中关于权限介绍指南提取权限类型。例如,普通权限有 ACCESS_NETWORK_STATE、BLUETOOTH、CHANGE_NETWORK_STATE、INSTALL_SHORTCUT 等,危险权限有 ACCESS_BACKGROUND_LOCATION、READ_CALL_LOG 等,特殊权限有 LOADER_USAGE_STATS、INSTANT_APP_FOREGROUND_SERVICE 等。

如果某个群体使用某种权限的概率比较高,则表明该权限是该群体的必要权限,哪怕该权限属于危险权限,也应该具有较低的危险值。反之,如果权限的概率比较低,哪怕该权限属于普通权限,也应该具有较高的危险值。为了量化应用权限的风险指数,给出如下定义:

定义 6 样本中某个应用程序的权限特征向量为 $P=(P_1,P_2,\cdots,P_n)$,某种群应用程序 C_x 的权限等级风险指数向量 $\text{PGrisk}=(\text{Grisk}_1,\text{Grisk}_2,\cdots,\text{Grisk}_n)$。某个应用程序的风险值计算方法为:

$$\text{Prisk\_sum} = \sum_{i=1} P_i \, \text{PGrisk}_i(1 - \text{Permission\_rate}_i)$$
$$0 < i \leqslant n$$

选取 K-means 算法对风险值进行聚类,获得初步的风险等级,该聚类算法可以对大型数据集进行高效分类[11]。将所有应用程序的风险值收敛成 k 簇。结合欧氏距离,将 k 个簇设置成 k 个安全等级,表示为 $L=(\text{Level}_1,\text{Level}_2,\cdots,\text{Level}_k)$。$\text{Level}_1$ 到 Level_k 的等级逐渐升高,相应的风险值也依次增加。也就是说,Level_1 级的应用程序风险值最低,安全系数最高,存在窃取用户信息等危险行为的可能性就越低;level_k 级的应用程序风险值最高,安全系数最低,存在窃取用户信息等危险行为的可能性就越高。

应用程序的权限机制是进行风险评估的一个重要因素,但不是决定性因素,因为广告推送、恶意扣费等行为是很难通过研究权限机制发现的。此外,新互联网时代的应用程序风险危机来自方方面面[12-15]。应用程序的风险程度只通过研究权限或权限组合已经无法适用于当下互联网时代,故该文在使用 k-means 算法对风险值聚类后,再使用 CART 决策树算法[16]将多方面信息纳入应用程序风险评估的研究范围,对聚类结果进行优化,增强可解释性。

2.3　风险评估算法流程

面向种群的 Android 应用风险评估算法流程如图 2 所示。

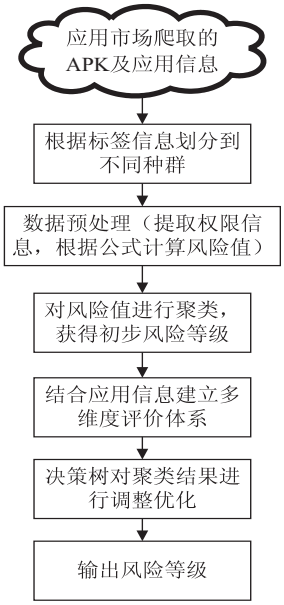

图 2　风险评估算法流程

风险评估算法具体步骤如下:

输入:某种群 C_x 数据集,风险值格式为 $\{\text{Prisk\_sum}_1,\text{Prisk\_sum}_2,\cdots,\text{Prisk\_sum}_n\}$

应用信息矩阵为 $\text{Application}=\{\text{Score}_i,\text{Download}_i,\text{Evaluate}_i,\text{Avg\_Risk}_i\}$;

输出:$C=(\text{Cluster}_1,\text{Cluster}_2,\cdots,\text{Cluster}_k)$ 和 CART 决策树。

(1)初始化 k 个聚类中心:$\{u_1,u_2,\cdots,u_k\}$,计算数据集中的每个样本 Prisk_sum_i 到各个聚类中心 u_j 的距离:$d_{ij}=\| \text{Prisk\_sum}_i - u_j \|_2^2$,样本距离哪个质点最近就归到哪一类,形成划分簇 $\text{Cluster}_1,\text{Cluster}_2,\cdots,\text{Cluster}_k$。重新计算新的质心:$u_j = \frac{1}{|\text{Cluster}_j|}\sum_{\text{Prisk\_sum}\in\text{Cluster}_j}\text{Prisk\_sum}$,直至质心不再发生改变。

(2)聚类后可以得到种群 C_x 的风险簇划分 $\text{Cluster}_1,\text{Cluster}_2,\cdots,\text{Cluster}_k$,进而可以得到种群中每个应用程序的风险等级 level_i,与应用信息矩阵 $\text{Application}=\{\text{Score}_i,\text{Download}_i,\text{Evaluate}_i,\text{Avg\_Risk}_i\}$

结合后得到多维度评价指标数据集，具体格式为 Multi_Data = $\{a_{1j}, a_{2j}, a_{3j}, a_{4j}, \text{level}_{ij}\}$，划分节点的属性集为 Attribute = {Score, Download, Evaluate, Avg_Risk}。其中，a_{1j} 为应用评分用户信息，a_{2j} 为应用下载量信息，a_{3j} 为应用评价信息，a_{4j} 为平均风险值(风险值/权限数)。

(3)对于多维度评价指标数据集，从根节点开始进行以下操作，构建二叉树：

①如果当前数据集中的样本都属于同一个属性，则设为一个叶节点并返回决策子树，当前节点停止递归。

②如果 Attribute = 空集(已经没有属性)或者 Multi_Data 中的数据在剩余属性中表现相同(属性无法划分)，则返回决策树子树，当前节点停止递归。

③将连续特征 A ($A \in$ Attribute)的属性值从小到大排列，取相邻两属性值的平均数作为划分点，其中第 j 个划分点 T_j 表示为：$T_j = (a_j + a_{j+1})/2$。根据划分点将节点数据集分为 D_1 和 D_2 子集，然后计算每个划分点下对应的基尼指数：

$$\text{Gini}(\text{Multi\_Data}) = 1 - \sum_{k=1}^{K} \left(\frac{|C_k|}{|\text{Multi\_Data}|}\right)^2$$

$$\text{Gini}(\text{Multi\_Data}, A) = \frac{|D_1|}{|\text{Multi\_Data}|}\text{Gini}(D_1) + \frac{|D_2|}{|\text{Multi\_Data}|}\text{Gini}(D_2)$$

④在计算出来的各个基尼系数中，选择基尼系数最小的属性 A 和对应的属性值 a，把多维度评价指标数据集划分成 D_1 子集和 D_2 子集。

⑤对 D_1 子集和 D_2 子集递归的调用① ~ ④步，生成决策树。

3 实验结果与分析

该文选取 360 应用市场和 Google 的官方应用市场 Google Play 作为数据源，使用 python 爬取大量 Android 应用以及相关信息，其中包括系统安全、通讯社交、生活休闲、地图旅游、教育学习等。将爬取的应用按照应用标签划分进不同的种群并存储在云端数据库中，建立多维度评价指标体系。新闻阅读、主题壁纸广泛用于人们的日常生活中，几乎每个用户都会根据个人需要安装这两类应用，故选取这两类应用程序作为研究对象。其中，新闻阅读有 1 744 个，主题壁纸有 1 207 个。图 3 所示为新闻阅读类应用程序和主题壁纸类应用程序的权限申请率。

从图 3 可以看出，同一种权限或者权限组合在不同群体的应用程序中的使用情况是不同的。例如，ACCESS_COARSE_LOCATION 权限，新闻阅读这一群体中申请率为 0.64，而主题壁纸群体中申请率只有 0.29。然后使用上文定义的风险值计算公式，对应用程序的危险程度进行量化。新闻阅读类风险值分布和主题壁纸类风险值分布如图 4 和图 5 所示。

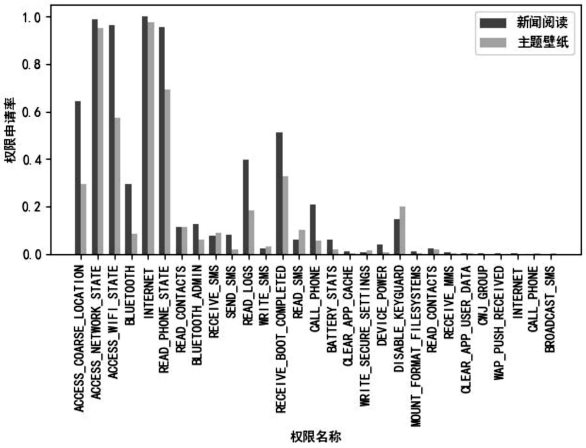

图 3 新闻阅读类应用程序和主题壁纸类应用程序的权限使用频率

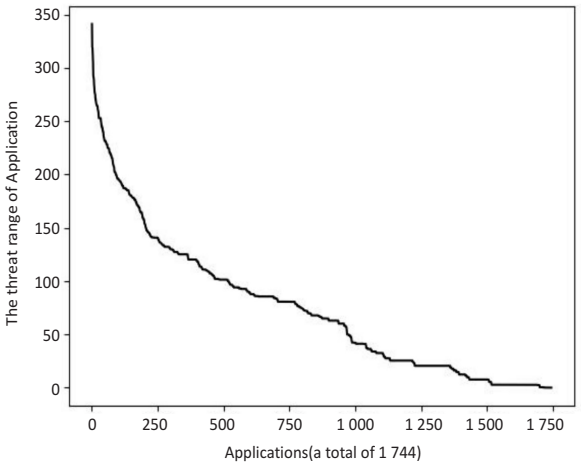

图 4 新闻阅读类风险值分布(总数 1 744)

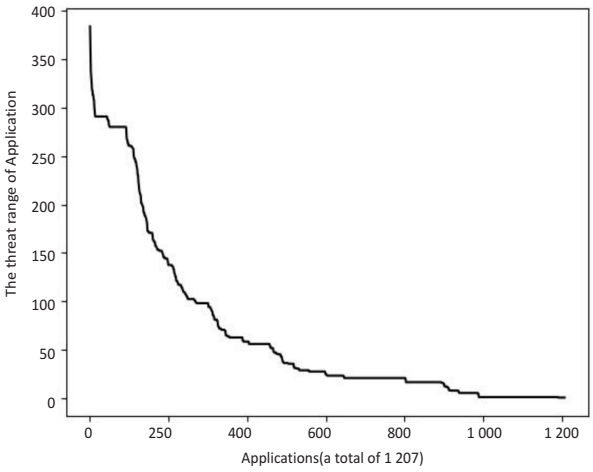

图 5 主题壁纸类风险值分布(总数 1 207)

在实验中，采用 k-means 算法对两个群体进行聚类。在分析数据规模和数据范围后，经过多次对比聚

类结果,最后将数据划分成 4 个等级,如表 1 和表 2 所示。等级越高,风险值就越高,存在恶意行为的可能性就越大。

表 1　新闻阅读种群聚类结果

风险值	风险等级	应用程序个数
[0,48]	良好	762
[48,105]	正常	523
[105,172]	偏高	289
[172,342]	危险	170

表 2　主题壁纸种群风险聚类结果

风险值	风险等级	应用程序个数
[0,38]	良好	717
[38,95]	正常	189
[95,203]	偏高	171
[203,385]	危险	130

接着,再使用决策树将评分、下载量、好评率、平均风险值(风险值/权限数)等信息纳入应用程序风险评估的研究范围,提高风险评估聚类结果的可解释性。

表 3　新闻阅读种群和主题壁纸种群风险评估结果

风险等级	新闻阅读种群 应用程序个数	主题壁纸种群 应用程序个数
良好	784	723
正常	508	198
偏高	295	168
危险	157	118

从表 3 可以看出,经过决策树修正优化聚类结果,新闻阅读种群和主题壁纸种群各风险等级的应用程序个数发生了改变。进一步研究发现,风险等级会因为程序功能、用户体验感等因素发生改变,而这些因素是很难通过研究权限或权限组获得的。例如,com. hnr. dxxw 属于新闻阅读种群,经过决策树修正风险等级从危险变为偏高,com. hnr. dxxw 比同种群其他应用增加了视频直播、线上学习功能,申请的权限更多,风险值也相对偏高。但是,并不存在窃取用户个人信息的情况,反而因为强大的功能深受大众喜爱。

实验结果表明,不同种群在权限申请方面存在一定的差异。从种群的角度来判定权限使用的合理性,可以考虑到不同功能应用程序之间的差异。使用 k-means 算法对种群风险值进行聚类,得到各风险等级的威胁值范围。最后,利用决策树将多维度指标体系纳入评估系统,提高风险评估结果的可解释性。在实际应用中,为用户选择 Android 应用提供一个较好的参考依据。

4　结束语

借用了生物种群的思想,提出了一种面向种群的 Android 应用权限分析和风险评估模型,将群体的权限使用情况和权限等级结合起来,并引入决策树对结果进行修正调优。实验结果表明,该方法可以量化计算 Android 应用风险值并且具有较好的自适应性,对用户选择安全低风险和体验感优越的 Android 应用程序提供较好的参考价值。接下来,将考虑恶意权限组合、动作组件、敏感 API、篡改检测等多方面,提高应用风险评估的准确性。

参考文献:

[1] FELT A,CHIN E,HANNA S,et al. Android permissions demystified[C]//Proc of the 18th ACM conference on computer and communications security. New York:ACM,2011:627-638.

[2] 中国互联网协会. 移动应用安全形势分析报告(2020 年)[R/OL]. 2021-07-05. https://www. isc. org. cn/article/40058. html.

[3] 朱佳伟,喻梁文,关　志,等. Android 权限机制安全研究综述[J]. 计算机应用研究,2015,32(10):2881-2885.

[4] LIU Z W,SUN Q B. Android application behavior detection based on rights management[J]. Netinfo Security,2014(6):72-77.

[5] 程运安,汪奕祥. 基于权限统计的 Android 恶意应用检测算法[J]. 计算机应用与软件,2017,34(1):306-310.

[6] 卜同同,曹天杰. 基于权限的 Android 应用风险评估方法[J]. 计算机应用,2019,39(1):131-135.

[7] 王家琰,徐开勇,戴乐育. 一种基于权限特征的 Android 恶意应用检测方法[J]. 计算机应用与软件,2018,35(3):316-320.

[8] XIAO Z,LI T,WANG Y. Feature analysis and risk assessment of android group based on clustering[C]//International conference on intelligent computing. [s. l.]:Springer,2018.

[9] 肖智婕. 面向种群的 Android 安全风险评估和恶意应用检测[D]. 武汉:武汉科技大学,2019.

[10] Android Developers. 面向应用开发者的文档[EB/OL]. 2016-12-08. https://developer. android. google. cn/docs.

[11] JAIN A K. Data clustering:50 years beyond K-means[J]. Pattern Recognition Letters,2010,31(8):651-666.

[12] 汤　玲. 智能手机风险分析与安全防护[J]. 数字技术与应用,2014(1):179-180.

[13] 吴星晨. 短视频 App 风险评估的方法和路径[J]. 青年记者,2019(34):53-54.

[14] 田　波,郑羽莎,刘鹏远,等. 移动 APP 用户隐私信息泄露风险评价指标及实证研究[J]. 图书情报工作,2018,62(19):101-110.

[15] 徐琬玥. 新媒体时代短视频 APP 的风靡——以抖音短视频为例[J]. 传播力研究,2018(8):82.

[16] BREIMAN L,FRIEDMAN J,STONE C J,et al. Classification and regression trees[M]. [s. l.]:CRC Press,1984.