文章编号：1005-8451（2018）10-0007-04

基于指数权重算法的铁路互联网售票异常用户智能识别的研究与实现

李　雯[1]，朱建生[2]，单杏花[2]
（1.中国铁道科学研究院集团有限公司　研究生部，北京　100081；
2.中国铁道科学研究院集团有限公司　电子计算技术研究所，北京　100081）

摘　要：为了确保公平公正售票，保障百姓购票利益，利用大数据技术，结合现有用户购票行为数据，设计基于指数权重的铁路互联网异常用户智能识别算法，并用2017年的用户行为数据进行测试，异常用户预测准确度达80%。测试结果验证了该算法的可行性，可以有效提高异常用户识别准确度，为保障12306铁路互联网售票系统的安全稳定运行及维护公平公正的售票环境提供了技术支持。

关键词：指数权重；购票行为；互联网售票；异常用户
中图分类号：U293.22：TP39　**文献标识码：**A

Abnormal user intelligence recognition of railway Internet ticketing based on index weight algorithm

LI Wen[1], ZHU Jiansheng[2], SHAN Xinhua[2]
(1. Graduate Department, China Academy of Railway Sciences Corporation Limited, Beijing　100081, China;
2. Institute of Computing Technologies, China Academy of Railway Sciences Corporation Limited, Beijing　100081, China)

Abstract: In order to ensure fair and fair ticketing and protect the interests of people ticket buying, this paper designed an abnormal user intelligence recognition algorithm of railway Internet ticking based on index weight by using the big data technology and combining with existing user ticket buying behavior data. Using the user behavior data in 2017, the accuracy of abnormal user prediction reached 80%. The test results verify the feasibility of the algorithm, which can effectively improve the accuracy of abnormal user identification, and provide technical support for ensuring the safe and stable operation of the 12306 railway Internet Ticketing and Reservation System and maintaining fair and fair ticketing environment.

Keywords: index weight; ticket buying behavior; Internet ticketing; abnormal user

铁路客运自2011年推出12306互联网售票系统以来[1]，不断地进行技术攻关，至2018年春运，该系统的承载能力与以往相比已经有了质的飞跃，单日售票能力从1 000万张提高到了1 500万张，高峰时段1 s可以售出近700张票，已经可以满足春运购票期间大量用户购票时系统的正常稳定运行。但是第三方软件以“预付可提高排名”，“专享100 M提速光纤”等标题为“噱头”，吸引了大量的用户借助其进行购票。数据显示，2017年12月开始，各种抢票软件活跃用户环比增长近3成，抢票功能加速包费用从10元到50元不等，在利益的驱使下，第三方软件公司严重损害了用户的利益和公平公正的购票环境。目前，互联网交易相关的法律法规还不健全[2]，相关异常用户的行为对社会危害性极大，严重破坏了交易平台公平公正的环境。

通过技术手段对异常购票行为进行限制是确保公平公正售票，保障百姓购票利益的主要方式。目前，风险控制系统可以从用户登录IP更换频率、设备指纹更换频率、余票查询频率以及内容分发网络（CDN）地址更换频率等角度实时识别异常请求[3]。然而，对海量的用户行为历史数据还没有进行更深一层的分析，还不能有效地挖掘出历史数据的潜在价值，因此，急需构建一个基于海量历史数据对异常用户进行识别的模型。本文结合大数据技术及机器学习技术，研

收稿日期：2018-03-13
基金项目：中国铁道科学研究院青年课题项目（2017YJ104）。
作者简介：李　雯，在读硕士研究生；朱建生，研究员。

究识别异常用户的分析方法，设计了一套异常用户智能识别模型，通过对历史用户数据进行机器学习训练，实现对囤票、倒票等异常用户行为的高效识别。

1 异常用户识别基本方法

为了识别异常用户，需要对用户的异常性进行指数化，异常用户指数区间为 [0~1]，如果该指数越接近 1，则说明该用户是异常用户的概率越大。

铁路 12306 互联网售票系统在提供服务过程中，用户与系统交互产生了海量有关用户访问的行为日志数据，这些日志数据详细描述了用户对铁路 12306 互联网售票系统的使用情况，通过对海量的购票日志数据进行数据分析，挖掘并提取出用户异常购票行为特征，建立规则库，根据规则识别出异常购票用户。

通过离线分析异常购票用户的识别模型，在用户使用铁路 12306 互联网售票系统购票的过程中，实时收集该用户的行为日志数据以及基本数据（包括个人信息、常用联系人信息等），并与离线分析出的识别模型进行匹配，从而达到实时管控和打击刷票等异常行为的目的，维护互联网售票交易的公平性。

2 数据采集及预处理

基于海量用户信息，传统的数据存储和处理方法无法满足算法的高效准确运行。本文主要通过 KETTLE 和 Flume 作为数据采集的主要工具 [4-5]，其中，KETTLE 主要采集关系型数据库的数据，Flume 主要采集用户操作日志即非结构化数据。关系型数据库的数据直接存储到 Hive 中等待数据预处理，Flume 采集的用户行为日志数据，进入 Kafka 消息队列，被 SparkStreaming 准实时消费解析成结构化数据并存储到 Hive 中 [6-7]。

对于采集到 Hive 中的用户行为数据以及购票信息，主要使用 Hive Sql 及其自定义函数（UDF），将数据处理操作转换成分布式 MapReduce 任务运行，使海量数据能够高效准确地进行处理。这种方法能够满足 Spark 机器学习数据模式的条件，从而对数据源指数权重算法模型进行高效训练。

3 指数权重算法

指数的大小源于数据集的特征，而处理这些特征需要一组权重。因此，在异常用户特征属性标签的实现处理问题上，采用一种自定义的指数权重算法，即通过指数权重算法动态计算指数的权重，实现修正人工确定指数的误判。

3.1 原理概述

根据研究目标的实际情况，异常用户指数的计算分为两部分：（1）用户特征权重值计算；（2）用户指数预测。因此，采用 Spark MLlib 中的 K-means 聚类算法和逻辑回归判别算法，构成一个组合指数算法模型，对异常用户指数进行预测 [8-9]，指数权重算法的具体结构如图 1 所示。

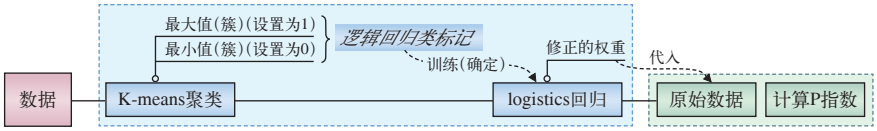

图1 指数权重算法结构

指数权重算法为两层结构：第 1 层主要用于修正特征值权重；第 2 层主要是用于计算指数 P 值，P 值的取值范围为 [0~1]。

3.2 主要流程

（1）利用 K-means 聚类算法将数据分成 K 类。其中，由于该场景中无法确定具体的 k 值，因此，通过肘部法则估计聚类数量。肘部法则模拟不同 k 值的成本函数值，k 值增大，平均畸变程度减小，从而每个类包含的样本数减少，样本离其重心更近。但是，随着 k 值继续增大，平均畸变程度的改善效果将不断减低，当畸变程度的改善效果下降幅度最大时，相应的 k 值称为肘部。

（2）利用 K-means 聚类算法 [10] 的结果筛选最大值簇和最小值簇。

（3）筛选得到的最大值簇数据类标记为 1，最小值簇数据类标记为 0，将二者数据作为逻辑回归的训练集数据。

（4）利用筛选得到的数据训练逻辑回归模型。

（5）利用确定的逻辑回归模型得到修正的特征权重。

（6）将修正后的权重带入原始数据，计算出最终的指数。

（7）对结果进行数据归一化（指数归一化，指数范围为 [0~1]）。

3.3 计算方式

该算法的最终目标是计算样本特征的某一指数（没有归一化的指数）。指数的大小来源于样本特征实现的特征值。通常有：

$$G=w_1x_1+w_2x_2+\cdots+w_nx_n \tag{1}$$

其中，w_1, w_2,…, w_n 表示权重系数，x_1, x_2,…, x_n 表示特征值。已知 w_1, w_2,…, w_n 初始值，为了得到准确度高的 G 值，需要修正权重值。具体修正权重值算法步骤如下。

3.3.1 计算聚类中心

在数据采集阶段，利用 K-means 聚类方法对数据集进行聚类，得到聚类的簇中心向量：β_1, β_1,…, β_n，根据公式（2）对簇中心进行指数化：

$$B_i=A\cdot\beta_i \tag{2}$$

其中，A 表示初始权重，B_i 为第 i 个簇中心的结果指数，最后筛选出簇中心结果指数最大和最小的两个簇。

假设总共分为 4 类，其中，簇中心向量分别为：β_1，β_2，β_3，β_4，那么，簇中心的结果指数取值有：B_1, B_2, B_3, B_4，若：$B_1>B_3>B_4>B_2$，则将结果为 B_1 的簇标记为 G_1 类用户，结果为 B_2 的簇标记为 G_2 类用户，这两类用户作为逻辑回归模型的数据集。

3.3.2 建立判别模型

建立一种基于逻辑回归模型的判别模型，对于指数值的大小可以用 G → {0, 1} 的映射进行表示，逻辑回归模型的结果也只能为 1 或 0，因此，可以假设 1 为异常行为指数值大的用户（即 G_1 类用户），0 为异常行为指数值小的用户（即 G_2 类用户）。

逻辑回归模型建立在 Sigmoid 函数基础之上，逻辑回归模型计算公式为：

$$P_\theta(x)=\frac{1}{1+e(\theta_0+x_1\theta_1+x_1\theta_2+\cdots+\theta_nx_n)} \tag{3}$$

其中，x_i 是提取的用户特征，θ_i 为 x_i 的对应参数（及回归模型特征 x_i 的回归系数）。当 $P_\theta(x)=0$ 时，被检测的用户指数值小，当 $P_\theta(x)=1$ 时，被检测的异常用户指数值大。

为计算最佳回归参数 θ，采用极大似然法：（1）输入人工设置初始参数值（初始权重）：θ'_0, θ'_1, θ'_2, …, θ'_i,…,θ'_n。（2）计算输出，得到训练后的参数值（修正后的权重）：θ_0, θ_1, θ_2, …, θ_i,…,θ_n。（3）利用上述的 G_1 类以及 G_2 类用户数据集作为逻辑回归模型的训练集数据进行训练，从而得到修正后的权重。

3.3.3 结果归一化

将得到的修正后的权重值代入原始数据，计算所有数据样本的异常用户指数。

以上是用户特征处理自定义的指数算法模型原理推导。主要是对用户的行为属性以及用户属性进行分析和选取，选择合适的属性作为输入参数，用以训练基于逻辑回归算法的分类模型，得到相应系数（特征权重），从而完成对指数算法模型的构建。

4 实验结果及分析

4.1 样本数据集特征提取

分析数据主要来源于 2017 年所有的购票数据和用户购票日志，主要包括用户的购票信息，常用联系人操作信息，用户购票行为日志信息等。通过 KETTLE 和 SparkStreaming 将数据进行处理并存储到 Hive 数据仓库中，产生分析数据的宽表，表中每个记录对应一个特征的数据项，总共包含 25 种特征向量，其主要特征向量包括：CDN 地址变换频率，退票比例，页面平均查询频率，改签比例，同一时间段购票次数，常用联系人更换频率等。特征向量是以结构化和数字化处理的用户基本购票信息，常用联系人操作信息，购票行为日志信息等 3 类价值特征，并对特征向量进行归一化处理，排除各价值变量因数据级差别造成的影响。

4.2 用户特征权重值计算

运用 Spark MLlib 中的 K-means 方法进行聚类分析。

4.2.1 聚类变量独立性检验

聚类算法需要输入变量彼此之间相互独立，故要对聚类变量进行相关性分析，运用 Spark MLlib 工具检验，结果表明，挑选的 25 种特征变量中，存在 3 对变量之间相对波动幅度的相关系数大于 0.3，故

删除这 3 个特征向量，剩余的 22 种特征变量通过独立性检验。

4.2.2 聚类分析及聚类群体分析

通过肘部法估计出聚类个数是 4。

K-means 聚类算法将用户群体分为 4 个群体，其中，各个用户群体的聚类中心点如表 1 所示。

表1 K-means聚类结果表

特征 \ 类别	第1类	第2类	第3类	第4类
CDN地址变换频率	0.9	0.1	0.3	0.4
退票比例	0.7	0.2	0.7	0.4
页面平均查询频率	0.7	0.1	0.3	0.5
……	……	……	……	……

4.3 用户异常指数预测

根据 K-means 聚类的结果，将第 1 类和第 2 类作为异常用户指数预测模型的样本数据进行训练，其中，第 1 类指数设置为 1，第 2 指数设置为 0，将该样本数据源加入到训练模型中进行训练，通过 Spark MLlib 的逻辑回归模型分析，由式 3 可得各个影响因子逻辑回归的回归系数值，如表 2 所示。

表2 K-means聚类参数值结果表

常数项	CDN地址变换频率	退票比例	改签频率	……
-2.13	0.21	0.3	-0.2	……

由此获得的逻辑回归模型表达式为：

$$P_\theta(x)=\frac{e^{-2.13+0.21x_1+0.3x_2-0.21x_3+\cdots}}{1+e^{-2.13+0.21x_1+0.3x_2-0.21x_3+\cdots}} \quad (4)$$

根据异常用户指数预测的模型，对 K-means 算法的样本数据进行训练，计算出每个用户的异常用户指数，依据现有风控系统对用户的拦截频率，发现现有风控系统拦截频率高的用户在通过该指数模型计算出的异常用户指数几乎都在 [0.7，1] 之间，实验结果如表 3 所示。

表3 异常指数计算与分类结果对比表

预测指数类别划分 \ 聚类类别	第1类	第2类	第3类	第4类
[0,0.3]		0.75		
[0.3,0.4]			0.81	
[0.4,0.7]				0.78
[0.7,1]	0.8			

分析表明，异常用户预测准确度达到了 80%，该模型具有较好的检验效果。

5 结束语

目前，铁路互联网售票系统异常用户恶意抢票、囤票现象泛滥，本文研究并设计了一种指数算法模型，能够通过海量的用户购票信息以及交易行为日志等数据，对异常购票用户进行识别，并使用 2017 年互联网客票系统相关数据对算法模型进行验证，80% 异常用户预测准确度证明该算法模型能够比较有效地识别出异常购票用户，可以与当前实时风险控制系统相结合，更加高效地识别异常用户。

参考文献：

[1] 王明哲，张振利，徐 彦，等．铁路互联网售票系统的研究与实现 [J]. 铁路计算机应用，2012，21（4）：23-25.

[2] 袁红艳．网络交易中对个人消费者权利的保护及网络交易的法律适用和立法探究 [J]. 中国集体经济，2017（32）：69-70.

[3] 郝晓培，单杏花，杨立鹏，等．基于大数据技术的铁路互联网售票异常用户行为分析研究与实现 [J]. 铁路计算机应用，2017，26（5）：1-4.

[4] 孙国强，韩强飞，陈 俊．Kettle 在企业数据仓库建设中的应用与研究 [J]. 信息系统工程，2017（2）：28-28.

[5] 陈 飞，艾中良．基于 Flume 的分布式日志采集分析系统设计与实现 [J]. 软件，2016，37（12）：82-88.

[6] Ghesmoune M, Lebbah M, Azzag H. Micro-Batching Growing Neural Gas for Clustering Data Streams Using Spark Streaming [J]. Procedia Computer Science, 2015, 53(1):158-166.

[7] 林宗缪，郭先超，裴雨清，等．基于 Spark 的网络日志分析平台研究与设计 [J]. 自动化与仪器仪表，2017（11）：157-159.

[8] Zhang L, Pang X, Ozolins O, et al. Spectrally efficient digitized radio-over-fiber system with k-means clustering-based multidimensional quantization[J]. Optics Letters, 2018, 43(7):1546-1551.

[9] 邬长安，郑桂荣，孙艳歌，等．k-means 和逻辑回归混合策略的不平衡类学习方法 [J]. 小型微型计算机系统，2017，38（9）：2119-2124.

[10] 王宏杰，师彦文．结合初始中心优化和特征加权的 K-Means 聚类算法 [J]. 计算机科学，2017（b11）：457-459.

责任编辑 王 浩