

面向短视频账号浅度安全等级识别方法

——短视频潜在优质账号的初步划分

吴志浩¹, 郭晓军¹, 闫宇辰¹, 李金涛²

¹西藏民族大学信息工程学院, 陕西 咸阳

²西藏民族大学马克思主义学院, 陕西 咸阳

收稿日期: 2023年3月13日; 录用日期: 2023年4月13日; 发布日期: 2023年4月21日

摘 要

针对短视频账号信息内容浅度安全等级分类识别的短视频官方未认证账号进行初级的潜在优质用户划分进行研究。由于账号信息内容浅度安全等级分类识别的特征值相对不多, 而且本文要的需求是保证其识别的准确率, 故改进了KNN算法, 提出选择使用GM-KNN模型来对其进行分类识别。通过网格搜索算法确定了K值的最优参数, 将不同距离进行对比, 选择了该模型的最优距离曼哈顿距离, 实现了潜在优质账号的初步划分。经过实验对比, GM-KNN模型准确率在短视频潜在优质账号的初步划分中均优于其他对比算法。

关键词

账号安全类型识别, KNN算法, 网格搜索算法, 曼哈顿距离

Short Video Account Shallow Security Level Identification Method

—Preliminary Division of Potential High-Quality Accounts for Short Videos

Zhihao Wu¹, Xiaojun Guo¹, Yuchen Yan¹, Jintao Li²

¹College of Information Engineering, Xizang Minzu University, Xianyang Shaanxi

²School of Marxism, Xizang Minzu University, Xianyang Shaanxi

Received: Mar. 13th, 2023; accepted: Apr. 13th, 2023; published: Apr. 21st, 2023

Abstract

The research is on the primary potential high-quality user division of short video official unau-

文章引用: 吴志浩, 郭晓军, 闫宇辰, 李金涛. 面向短视频账号浅度安全等级识别方法[J]. 数据挖掘, 2023, 13(2): 165-172. DOI: 10.12677/hjdm.2023.132016

thenticated accounts identified by shallow security level classification of short video account information content. Since there are relatively few eigenvalues for the classification and identification of shallow security levels of account information, and the requirement of this paper is to ensure the accuracy of its identification, the KNN algorithm is improved, and the GM-KNN model is proposed to be used for classification and identification. The optimal parameter of the K value was determined by the grid search algorithm, and the different distances were compared, and the optimal distance of the model was selected, the Manhattan distance, and the preliminary division of potential high-quality accounts was realized. After experimental comparison, the accuracy of GM-KNN model is better than other comparative algorithms in the preliminary division of potential high-quality accounts of short videos.

Keywords

Account Security Type Identification, KNN Algorithm, Grid Search Algorithm, Manhattan Distance

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

短视频属于一种互联网内容传播方式。短视频账号信息内容的浅度安全等级识别是从账号信息安全的基础信息风险角度入手, 对其账号的安全类型进行初步判断, 考虑划分出不同类型。而短视频平台内有获得平台的官方级认证的账号和未获得平台的官方级认证的账号两种账号类型, 其中获得平台官方级认证的账号相对风险较低, 即安全性较高。而未获得平台的官方级认证的账号相对风险较高, 但是未获得平台的官方级认证的账号风险也不完全一样, 有的作为潜在的优质账号安全性相较于其他未认证账号安全性更高, 通过这一方面的考虑入手, 我们可以通过短视频潜在优质账号分类识别来完成短视频账号信息内容的浅度安全等级识别。基于此, 在面向短视频潜在优质账号分类识别模型中, 在模型分析为认证账号而未获得平台的官方级认证的账号里, 将其认证为潜在优质账号。短视频账号信息内容浅度安全等级分类通过对短视频未认证账号预测该账号是否为潜在优质账号进行分类后, 分为实际未认证且潜在分类为优质未认证账号, 实际未认证且潜在分类为优质认证账号, 实际已认证优质账号三类账号类型。

预测分类的算法有 KNN 算法、逻辑回归算法、随机森林算法、朴素贝叶斯算法、SVM 算法、ELM 算法等。张炎亮[1]等使用 KNN 算法将具有不同特征的产品推荐给相应的用户, 从而完成个性化推荐。Guangxia Xu 等[2]提出了一种面向混合属性的动态 SOM 模糊聚类算法(D-SOMFCM-OMA)用于移动用户分类。Ming Yan 等[3]提出了一种基于移动用户分类的加权马尔可夫预测模型。通过机器学习算法对所有用户进行分类和根据各用户类别的特点, 优化加权马尔可夫预测模型的步长阈值和加权系数, 对各用户类别进行移动性预测。Dingding Hu [4]为了对电动汽车用户进行分类提出了一种结合扩展 RFM 模型、两阶段聚类方法和熵权法的集成方法, Prasad N. Achyutha [5]使用 SVM 算法和推特分析对用户进行了分类。

本研究根据短视频账号的粉丝数、作品数、历史获赞数、周期获赞数、周期评论数、周期收藏数、周期转发数、粉丝群聊数构建出账号潜在优质认证的评价特征。根据评价特征构建了 GM-KNN 模型(Grid Search CV and Manhattan Distance, K-Nearest Neighbor), 以 KNN 算法为基础, 使用网格搜索算法[6]进行参数寻优, 确定最佳 K 值。根据多距离公式[7]进行评价指标对比, 确定了本文中最优距离公式为曼哈顿

距离公式。构建模型后对未认证短视频账号进行分类预测, 可以将未认证短视频账号分为两类, 即可认证优质账号和不可认证优质账号。评价指标表明该模型在短视频潜在优质账号的初步划分方面具有较佳效果。

2. 相关技术基础

2.1. KNN 算法

KNN 分类算法常用于数据挖掘分类技术。K 最近邻即表示 K 个最近的邻居, 每个样本都可以用它最接近的 K 个邻近值来代表。工作原理是: 存在一个样本数据集合, 也称为训练集, 样本集中每个数据都存在标签。输入没有标签的新数据后, 将新的数据的每个特征与样本集中数据对应的特征进行比较, 然后提取样本最相似数据(最近邻)的分类标签。一般只选择样本数据集中前 k 个最相似的数据, 通常 k 是不大于 20 的整数。KNN 算法的基本步骤有:

- 1) 通过距离公式计算样本之间的距离;
- 2) 将未知样本和训练样本之间的距离按递增关系排序;
- 3) 选择距离最小的 k 个点;
- 4) 查看 k 个点的类别频率;
- 5) 将频率最多的类别作为未知样本的类别。

2.2. 网格搜索算法

网格搜索算法属于参数寻优算法, 其将变量区域网格化, 对所有网格点进行遍历, 求解满足约束函数的目标函数值, 通过比较选择出最优点, 其方法属于穷举搜索方法。

2.3. 多距离公式

2.3.1. 欧几里得距离

N 维空间点 $a(x_{11}, x_{12}, \dots, x_{1n})$ 和 $b(x_{21}, x_{22}, \dots, x_{2n})$ 欧式距离计算公式为公式(1)所示。

$$d_{12} = \sqrt{\sum_{k=1}^n (x_{1k} - x_{2k})^2} \quad (1)$$

2.3.2. 曼哈顿距离

N 维空间点 $a(x_{11}, x_{12}, \dots, x_{1n})$ 和 $b(x_{21}, x_{22}, \dots, x_{2n})$ 曼哈顿距离计算公式为公式所示。

$$d_{12} = \sum_{k=1}^n |x_{1k} - x_{2k}| \quad (2)$$

2.3.3. 切比雪夫距离

N 维空间点 $a(x_{11}, x_{12}, \dots, x_{1n})$ 和 $b(x_{21}, x_{22}, \dots, x_{2n})$ 切比雪夫距离计算公式为公式所示。

$$d_{12} = \max(|x_{1i} - x_{2i}|) \quad (3)$$

3. 数据特征及分析

3.1. 数据集

采用爬虫软件获取的抖音短视频平台数据, 通过指定关键字搜索进行爬取, 使用了在 2022 年 5 月 28 日至 2022 年 6 月 12 日获取的 2008 个账号信息。基本信息字段如表 1 所示。预处理后进行了多种算法以及方法的对比试验。经过参考和筛选, 数据集里采用了账号的粉丝数、作品数、历史获赞数、周期

获赞数、周期评论数、周期收藏数、周期转发数、粉丝群聊数作为主要特征，其中周期特征由表 2 可见并说明。

Table 1. Dataset user information data display

表 1. 数据集用户信息数据展示

属性	说明
昵称	用户昵称
抖音号	抖音 ID
介绍	个人介绍的信息
认证信息	即官方给的用户标签
粉丝群聊	群聊的数量
橱窗	电商卖货的橱窗
作品数	注册以来发布的作品数量
获赞数	注册以来作品获得点赞的总数量
关注数	该用户关注了其他多少人
粉丝数	该用户受到了多少人关注
主页链接	点击该链接到该用户个人信息界面

Table 2. Video information display of dataset users

表 2. 数据集用户的视频信息展示

属性	说明
页面网址	点击该链接跳转至该视频播放界面
标题	视频的名称
作者	发布或转发该视频的账号名称
抖音号	发布或转发该视频的账号抖音 ID
视频时长	该视频的时间播放长度
点赞数	其他用户给该视频的点赞数
评论数	其他用户给该视频的评论数
转发数	其他用户给该视频的转发数
收藏数	其他用户给该视频的收藏数
发布时间	该用户将该视频发布在抖音平台上的时间

3.2. 数据类别平衡性分析

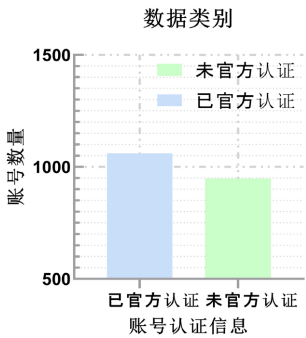


Figure 1. Data category

图 1. 数据类别

由图 1 可见，已认证账号人数和未认证账号人数数量基本均衡，所以不需要对数据进行平衡优化处理。

3.3. 数据特征相关性分析

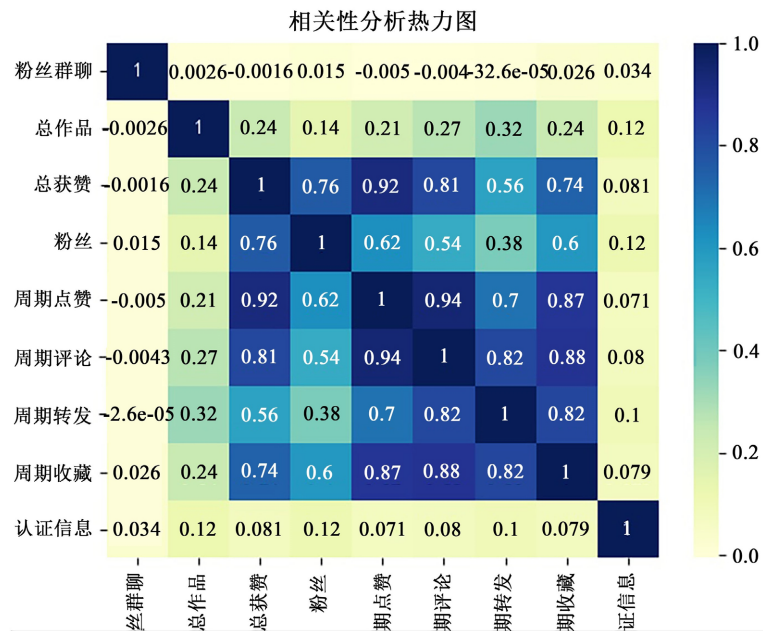


Figure 2. Data feature correlation analysis
图 2. 数据特征相关性分析

由相关性分析热力图图 2 可知，短视频账号的粉丝数、作品数、历史获赞数、周期获赞数、周期评论数、周期收藏数、周期转发数、粉丝群聊数特征都与认证信息正相关。

4. 基于网格搜索算法和曼哈顿距离的 KNN 算法面向短视频优质账号分类识别模型——GM-KNN 模型

4.1. 模型框架

针对官方认证短视频账号特征性不是特别明显，比如有的短视频账号粉丝量和其他特征较少，但是因为具有一定知名度或者是某领域专家为官方认证账号，另外有的自媒体账号虽然级别和各方面的要求以及达到了官方认证程度，但是该账号由于没有主动申请官方认证，而使得该账户暂时为非官方认证账号。再加上短视频本章节采用数据集随机性较大，难以保证潜在优质用户分类准确度的情况下，本章节结合基于网格搜索算法和曼哈顿距离的 KNN 算法提出了一种账号分类模型 GM-KNN 算法，进行潜在官方认证优质账号分类识别(图 3)。

4.2. 基于网格搜索算法对最优参数 K 值的确定

通过网格搜索算法对模型中的参数 n_neighbors 即 K 值寻找最优的参数值，一般 K 值为不大于 20 的数，设置了参数搜索范围为 1~20，算法结果本文中最优 K 值为 18。

4.3. 基于多距离公式确定最优距离公式

通过在模型中更改不同的距离公式对各项评价指标进行对比，由于本文实际需求，主要对准确率进

行对比。采用了三种距离公式，分别为欧氏距离、曼哈顿距离、切比雪夫距离进行测试，最终确定本文最优距离为曼哈顿距离。

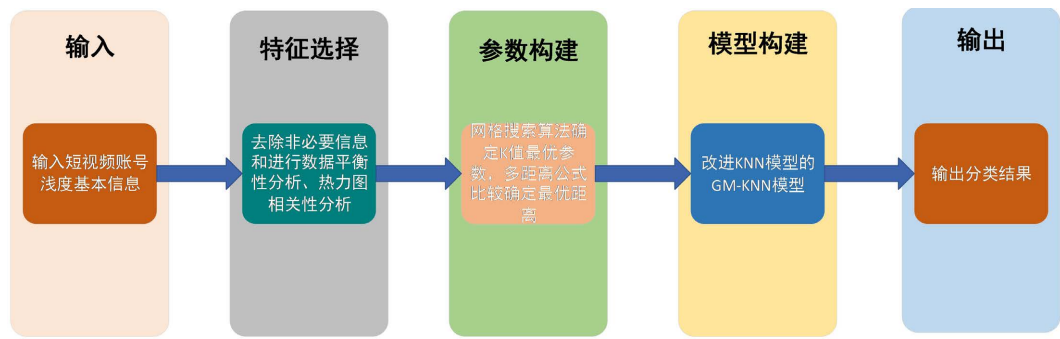


Figure 3. GM-KNN algorithm model diagram
图 3. GM-KNN 算法模型图

5. 短视频潜在优质账号实验过程对比和分析

5.1. 实验对象和具体设计

为了验证、评估本文方法的有效性，实验将数据集进行了 80%训练集、20%测试集的划分，选取了不同 KNN 算法的自身比较和目前较为经典的且比较适用短视频潜在优质账号分类识别算法对比，具体方法如下采用。

1) 多种 KNN 算法对比。通过对原始 KNN 算法即基于欧氏距离的 KNN 算法、以及基于曼哈顿距离的 KNN 算法、基于切比雪夫距离的 KNN 算法、基于网格搜索算法和欧氏距离的 KNN 算法、基于网格搜索算法和切比雪夫距离的 KNN 算法和本文提出的 GM-KNN 算法进行准确率、查准率、召回率、F1 值的整体比较，同时考虑本章节的最根本需求为短视频潜在优质账号分类识别的准确性，所以将准确率作为各算法的最根本评价指标。

2) 通过其他算法的比较。LogisticRegression 算法、SVM 算法、EIM 算法与 GM-KNN 算法进行准确率对比。

5.2. 实验结果分析

5.2.1. 多种 KNN 算法对比

Table 3. Algorithm and accuracy
表 3. 算法及准确率

算法	准确率	精准率	召回率	F1 分值
基于欧氏距离的 KNN 算法	0.590	0.604	0.612	0.608
基于切比雪夫距离的 KNN 算法	0.592	0.607	0.612	0.610
基于曼哈顿距离的 KNN 算法	0.614	0.630	0.627	0.628
基于网格搜索算法和欧氏距离的 KNN 算法	0.600	0.620	0.593	0.606
基于网格搜索算法和切比雪夫距离的 KNN 算法	0.610	0.630	0.603	0.616
GM-KNN 算法	0.622	0.649	0.593	0.620

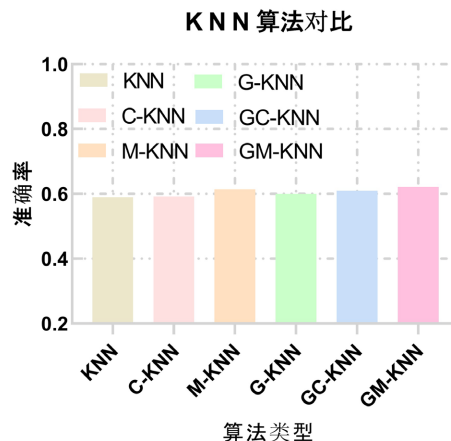


Figure 4. KNN algorithm comparison

图 4. KNN 算法对比

从表 3 可以看出, GM-KNN 算法在各项评价指标方面均优于其他 KNN 算法, 在本章节最主要评价指标标准准确率上, 从图 4 可看出, GM-KNN 算法高了基于欧氏距离的 KNN 算法 3.2%、以及基于曼哈顿距离的 KNN 算法 0.8%、基于切比雪夫距离的 KNN 算法 3%、基于网格搜索算法和欧氏距离的 KNN 算法 2.2%、基于网格搜索算法和切比雪夫距离的 KNN 算法百分点 1.2%。

5.2.2. 各算法准确率对比

通过 Logistic 算法、SVM 算法、Elm 算法分别获得短视频潜在优质账号分类准确率, 然后对这三个方法进行对比, 实验结果如表所示, 结果表明 GM-KNN 算法在多次测试取平均值后的准确率上高于 Logistic 算法、SVM 算法、Elm 算法, 这是由于网格搜索算法寻优后参数 K 值最佳, 加上真实短视频账号数据集里面官方认证潜在优质用户的特征不够明显, 加上特征属性维数不算过低, 曼哈顿距离对于高维数据更好, 所以在此次实验中表现更佳。

从表 4 和图 5 可以看出, 测试算法准确率 N 次取平均值后(N 本章节取 10), 基于曼哈顿距离的 KNN 算法准确率均优于其他三种算法, 其中, 准确率相较于 Logistic 算法提升了 7.2%, 准确率相较于 SVM 算法提升了 9.9%, 准确率相较于 Elm 算法提升了 5.3%。

Table 4. Algorithm and Accuracy

表 4. 算法及准确率

算法	准确率
Logistic 算法	0.550
SVM 算法	0.523
Elm 算法	0.569
GM-KNN 算法	0.622

6. 本章小结

短视频潜在优质账号的初步划分实质上是未认证账号分类预测的问题, 本文先分析了数据集不同类别的均衡性, 不需要对数据进行均衡优化, 之后进行了相关性分析, 短视频账号的粉丝数、作品数、历史获赞数、周期获赞数、周期评论数、周期收藏数、周期转发数、粉丝群聊数等特征都与认证信息相关, 提出了本文的基于 KNN 算法的应用模型 GM-KNN 算法, 针对 KNN 算法如何确定参数的问题, 使用了

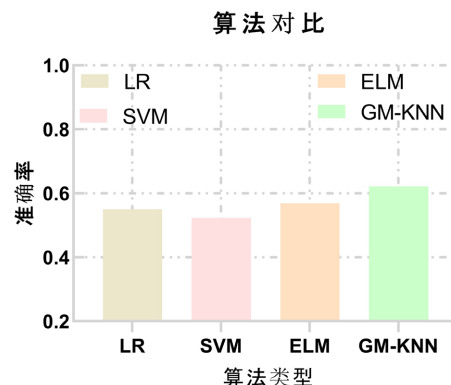


Figure 5. Algorithm comparison

图 5. 算法对比

网格搜索法对其进行了参数优化, 针对基础 KNN 算法欧式距离公式非最优的问题, 采用了多种距离公式进行对比, 使用曼哈顿距离公式进行距离改进, 在经过评价指标对比后, 结果表明该模型在该领域应用中具有更佳的效果, 给出了一种短视频账号信息内容的浅度安全等级识别新的思路方法。

参考文献

- [1] 张炎亮, 张超, 李静. 基于动态用户画像标签的 KNN 分类推荐算法研究[J]. 情报科学, 2020, 38(8): 11-15.
- [2] Xu, G., Zhang, L., Ma, C. and Liu, Y. (2020) A Mixed Attributes Oriented Dynamic SOM Fuzzy Cluster Algorithm for Mobile User Classification. *Information Sciences*, **515**, 280-293. <https://doi.org/10.1016/j.ins.2019.12.019>
- [3] Yan, M., Li, S., Chan, C.A., et al. (2021) Mobility Prediction Using a Weighted Markov Model Based on Mobile User Classification. *Sensors*, **21**, Article 1740. <https://doi.org/10.3390/s21051740>
- [4] Hu, D., Zhou, K., Li, F. and Ma, D. (2022) Electric Vehicle User Classification and Value Discovery Based on Charging Big Data. *Energy*, **249**, Article ID: 123698. <https://doi.org/10.1016/j.energy.2022.123698>
- [5] Reddy, S.R.G., Varma, G.P.S. and Davuluri, R.L. (2023) Resnet-Based Modified Red Deer Optimization with DLCNN Classifier for Plant Disease Identification and Classification. *Computers and Electrical Engineering*, **105**, Article ID: 108492. <https://doi.org/10.1016/j.compeleceng.2022.108492>
- [6] 张立秀, 张淑娟, 孙海霞, 薛建新, 任锐, 刘文俊. 高光谱技术结合网格搜索优化支持向量机的桃缺陷检测[J/OL]. 食品与发酵工业: 1-10. <https://doi.org/10.13995/j.cnki.11-1802/ts.033557>
- [7] 吴胜义, 王义贵, 王飞, 李伟坡. 基于多距离度量 kNN 模型的森林蓄积量反演[J]. 中南林业科技大学学报, 2023, 43(2): 10-18. <https://doi.org/10.14067/j.cnki.1673-923x.2023.02.002>