基于 SVM-双加权 RkNN 模型的汉服月销分类问题

Hanfu monthly sales classification based on a SVM-double-weighted RkNN model

许小诺，张秋梅
（长春大学 数学与统计学院，长春 130022）

摘要：汉服作为彰显中华优秀传统文化的服饰之一，在近些年逐渐走进大众视野。本文旨在准确预测汉服销量，为店铺提供数据用于合理经营，共同助力汉服的传承和推广。以淘宝平台的汉服销量为研究对象，建立包含汉服的形制、款式和传统文化等影响汉服销量的指标体系。针对汉服月销的二分类问题，建立 SVM-双加权 RkNN 模型，有效解决支持向量机分离超平面附近点易错分的问题；针对高销汉服的三分类问题，在 SVM-双加权 RkNN 模型的基础上引入两点改进。分类结果表明，本文提出的模型在二分类和三分类问题上的准确率分别达到了 0.883 0 和 0.777 8，均具有较好的分类效果，可以帮助汉服原创品牌预测汉服月销，合理安排生产存储计划，及时调整营销策略。

关键词：汉服；销量分类；影响因素；指标体系；SVM；双加权 RkNN
中图分类号：TS941.7　　文献标志码：A　　文章编号：1001－7003(2025)10－0021－10
DOI：10.3969/j.issn.1001-7003.2025.10.003

汉服是中国传统服饰的重要组成部分。它始于黄帝，定型于周朝，又在唐、宋、明时期不断发展，经历了数千年的演变，逐渐形成了完整的服装体系，成为汉族人民的主要服饰。作为能够彰显中国传统文化的服饰，汉服却在清朝灭亡之后逐渐淡出人们的视野。中国改革开放之后，综合国力逐渐提升，一部分国人的大国意识开始觉醒，广大汉文化爱好者发起了一场“汉服运动”。根据艾媒咨询发布的《2022 年中国新汉服行业发展白皮书》显示，2021 年新汉服行业市场规模达到 101 亿元，同比增长 6.4%。根据京东零售大时尚事业部发布的《2024 汉服趋势白皮书》显示，汉服市场规模在 2023 年达到了 137 亿元，预计至 2028 年将进一步增长至 334 亿元。新汉服发展势头迅猛，而汉服的销售在汉服行业中占较大比重。因此，本文将关注点投向线上汉服的销量预测问题。

本文对现有文献进行了梳理。一方面，由于汉服本质上是一件商品，对商品销量的现有研究进行分析。Aksoy 等[1]选取商品属性特征和历史数据预测服装销量，Kulkarni 等[2]把搜索引擎日志数据引入模型，而吕杰妮[3]则把体感温度、风寒指数等天气信息加入到预测模型中，都取得了不错的预测效果。孙一文[4]运用 RBF 神经网络预测抖音平台的两家品牌店铺的服装销量。另一方面，对本文使用的相关模型进行了研究。Li 等[5]使用 k 近邻算法对 SVM 分离超平面附近的点进行二次分类，解决 SVM 核函数参数的选择问题。简强[6]在 SVM 的训练过程中引入基于 Levy 飞行策略的 AOA 算法，实现关键参数的优化。Abu 等[7]研究发现不同距离测度的选择对 kNN 模型分类性能有较大的影响。陈丽等[8]针对支持向量机决策超平面附近的点易错分的问题，引入反 k 近邻算法，提出 SVM-RkNN 分类模型，分类准确率相较于 SVM 模型有所提高。另外，考虑到汉服传播效果对销量的影响，分别对抖音[9]、新浪微博[10]和百度指数平台[11]的传播效果衡量进行了梳理。

目前学者对汉服的相关研究主要集中在汉服的评价[12]和购买意愿分析[13]上，很少对汉服的销量进行深入探讨。因此，本文以线上汉服的销量为切入点，建立较为准确的汉服销量预测模型。汉服作为包含文化属性的商品，虽然其销量相较于普通服装商品较低，但针对汉服销量的研究，可以为线上原创汉服品牌在生产和存储方面提供数据支撑。

本文以淘宝平台汉服店铺的月销为研究对象，建立包含汉服特有、传统文化因素在内的汉服销量影响因素指标体系。首先，为改善 SVM 模型分离超平面附近数据点易错分的情况，采用 SVM-双加权 RkNN 模型对汉服月销进行以 100 为界的低销高销的二分类划分。其次，提出 SVM-改进双加权 RkNN 模型，对高销汉服进行三分类划分。本文的创新点主

收稿日期：2024－07－13；修回日期：2025－09－15
基金项目：吉林省科技厅项目(20220101025JC)；吉林省教育厅项目(JJKH20240736KJ)
作者简介：许小诺(1999—)，女，硕士研究生，研究方向为大数据分析。通信作者：张秋梅，教授，zhangqm1110@163.com。

要有两点:第一,使用多种数据技术对不易量化的文化因素进行量化,构建影响汉服销量的指标体系;第二,使用 SVM-双加权 RkNN 模型提高 SVM 模型分离超平面附近数据点的分类准确率。本文得到的分类结果准确率较高,可以为汉服店铺调整营销策略、优化产品结构提供一定的理论和实践基础,助力汉服这一传统文化的推广和传承。

1 模型的建立

1.1 支持向量机模型(SVM)

支持向量机(support vector machine, SVM)是一种广泛使用的机器学习算法,主要用于分类和回归分析。该方法用于分类时,旨在找到一个超平面,该超平面能够将不同类别的数据尽可能地分开,并且保证与数据集最接近的点之间的间隔最大。

对于线性可分问题,支持向量机通过间隔最大化或等价地求解相应的凸二次规划问题,学习得到对应的分离超平面及决策函数。此时的支持向量机称为线性可分支持向量机。

$$\omega^* \cdot x + b^* = 0 \tag{1}$$

$$f(x) = sign(\omega^* \cdot x + b^*) \tag{2}$$

式中:ω^* 为超平面的法向量;b^* 为超平面的截距;x 为数据点的特征。$f(x)$ 为支持向量机的决策函数。

1.2 反 k 近邻模型(RkNN)

反 k 近邻作为 k 近邻模型的改进,其主要思想是对于测试点 p,找到将其作为 k 近邻点的点,并根据这些点的目标值确定预测值。反 k 近邻点的定义如下:

对任意正整数 k 和数据集 D,数据点 p 的反 k 近邻点是那些 k 近邻点中包含点 p 的数据点构成的集合,记作 $RkNN_k(p)$:

$$RkNN_k(p) = \{s \mid s \in D \cap p \in kNN_k(s)\} \tag{3}$$

式中:$RkNN_k(p)$ 是数据点 p 的反 k 近邻点;$kNN_k(p)$ 是数据点 p 的 k 近邻点。

反 k 近邻模型目前主要用于查询问题和离群值的检测,也有部分学者将其运用于分类问题。反 k 近邻模型用于分类问题时,具体步骤如下:

设测试点坐标为 $p = \{x_p, y_p\}$,训练集 $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$,其中 $y_i \in Y = \{w_1, w_2, \cdots, w_c\}$,$w_j$ 表示样本点的类别,$j = 1, 2, \cdots, c$。

输入:测试样本点 p 和训练样本集 D。

输出:测试样本点 p 的类别 w_j。

方法:

1)把测试点 p 和训练集 D 放在一起,依次找到每个训练点的 k 个近邻点;

2)如果第 i 个训练点的 k 近邻点中包含测试点 p,则认为 x_i 是测试点 p 的一个反 k 近邻点。假设共找到 m 个 p 的反 k 近邻点,$k_1, k_2, \cdots, k_c$ 分别是这 m 个反 k 近邻点中属于 $w_1, w_2, \cdots, w_c$ 类的样本数。

3)若 $k_j = \max\{k_1, k_2, \cdots, k_c\}$,则 $p \in w_j$。

1.3 SVM-双加权 RkNN 模型的建立

为提高 SVM 模型分离超平面附近数据点的分类准确率,本文构建了 SVM-双加权 RkNN 模型用于汉服月销分类问题,其主要思想为:首先,使用线性支持向量机对数据进行第一次分类;其次,计算出各数据点距分离超平面的距离,将距离较近的点筛选出来,运用双加权 RkNN 模型进行二次分类;最后,把两个阶段的分类结果结合起来,得到最终预测结果。该模型的流程图如图 1 所示:

目前 RkNN 用于分类问题的研究较少,由于 RkNN 模型与 kNN 模型在寻找近邻点时的思路相似,主要区别在于 kNN 以测试集的近邻点为分类依据,而 RkNN 模型则把测试点作为近邻点的点作为分类依据。鉴于二者的相同点,对 RkNN 的加权改进主要借鉴 kNN 模型的改进。本文对 RkNN 的加权主要集中于以下两点:

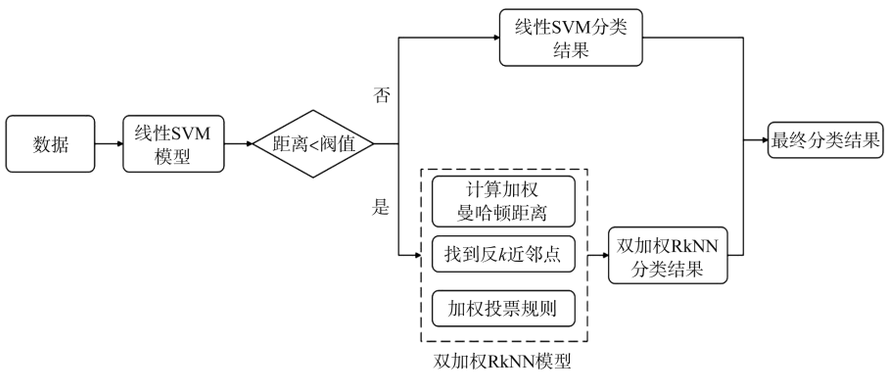


图 1 SVM-双加权 RkNN 模型流程示意

Fig. 1 Flowchart of a SVM-double-weighted RkNN model

1)距离公式中的特征加权。反 k 近邻算法可以选择的距离度量较多,其中,曼哈顿距离的计算公式如下:

$$d(x,y_i) = \sum_{k=1}^{n} |x_k - y_{ik}| \tag{4}$$

式中:$d(x,y_i)$ 为点 x 和点 y_i 之间的曼哈顿距离;x_k 为点 x 在第 k 个特征上的数值;y_{ik} 为点 y_i 在第 k 个特征上的数值。

由于反 k 近邻算法对不同的距离度量较敏感,且各指标的重要性程度往往不同。本文采用递归特征消除法确定的权重对曼哈顿距离公式进行改进,计算公式如下:

$$d_w(x,y_i) = \sum_{k=1}^{n} \omega_k \cdot |x_k - y_{ik}| \tag{5}$$

式中:$d_w(x,y_i)$ 为点 x 和点 y_i 之间的加权曼哈顿距离;ω_k 为递归特征消除法确定的第 k 个特征的权重值。

2)投票规则加权。反 k 近邻算法中的多数投票规则忽略了距离不同的点对测试点产生的影响不同,因此,对投票规则同样引入权重。本文选取倒数权重优化投票规则,权重计算公式如下:

$$w_i = \frac{1}{d_i + 0.1} \tag{6}$$

式中:w_i 表示第 i 个反 k 近邻点对应的距离权重;d_i 表示第 i 个反 k 近邻点距测试点的距离。

2　数据的收集与处理

2.1　数据的收集

汉服可以看作包含文化属性的商品。一方面,它包含服装类商品的功能;另一方面,它包含汉服所涵盖的文化特征。从服装类商品角度出发,主要考虑其作为商品所必须要具备的属性,比如说包含价格、销量和应季等与商品本身有关的变量及包含情感得分、评论数量和交互式评论这些评论信息中的变量。从汉服的文化特征角度出发,首先考虑汉服作为服装时的区分变量,就像现代服装会以衬衣、T 恤、牛仔裤来区分,汉服主要以形制和款式两个变量来区分;其次,考虑汉服所具有的传统文化指标,由于传统节日和大型线下汉服活动是促使消费者购买汉服的契机,考虑构建上述两个指标,另外,部分网友在媒体平台上可能会受到汉服的吸引进而购买汉服,从传播效果的角度,选取百度、微博、抖音三个平台衡量汉服的传播效果。由此,本文构建出影响汉服销量的指标体系,如图 2 所示。

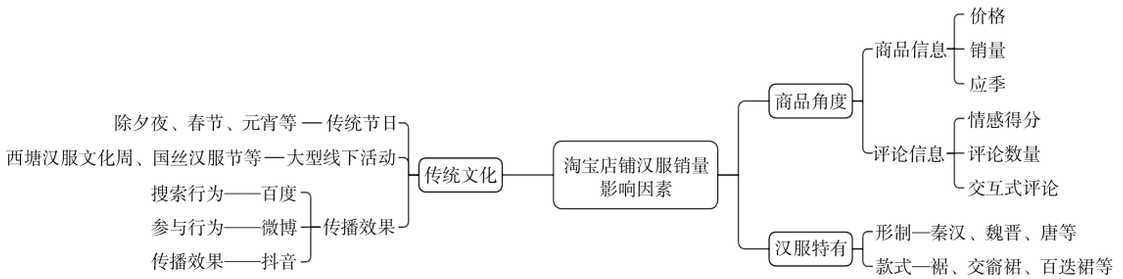


图 2　汉服销量影响因素指标体系

Fig. 2　Index system of influencing factors on Hanfu sales

本文所使用的数据主要分为商品、汉服特有和传统文化三个维度。商品维度的数据来源于淘宝平台,选取成立时间较长且较为被消费者熟知的“池夏 汉风原创铺子”“十三余小豆蔻儿国风工作室”和“花朝记汉服”三家汉服店铺,采集时间为 2023 年 12 月到 2024 年 5 月的每月月初,获取数据的时间跨度为 2023 年 11 月到 2024 年 4 月的 6 个月数据,共采集到 3 274 条商品相关数据,每月约采集到 4 000 条的评论数据;传统文化维度下传播效果方面的数据来源于百度指数、新浪微博和抖音平台,采集时间与商品维度的数据相对应。以上数据均通过 Python 爬虫得到。得到数据并清洗后,对数据进行了缺失值和异常值处理,保证数据的有效性。

2.2　指标的构建

本文所构建的指标体系中,有一些指标需要使用采集到的数据进行构建,主要有商品评论信息、汉服特有属性、汉服传播效果、传统节日和大型线下活动。

2.2.1　商品评论信息

结合现有文献的研究,在商品评论信息角度共选取三个指标,分别是情感得分[14]、综合评论深度[15]和评论数量[16]。

计算情感得分时,主要按照文本数据清洗、基于哈工大 LTP 模型的文本分词和词性标注、基于知网情感词典的情感得分计算三个步骤进行。

由于商品评论中既包含初次评论,也包含追加评论。关于商品评论对销量的影响,有学者认为追加评论比初次评论表达的情感倾向更浓,对商品销量的影响更大。因此,本文提出各商品的综合情感得分的计算方法,如下式所示:

$$\text{综合情感得分} = \begin{cases} \text{初评得分} & \text{只含初评} \\ 1.2 \times \text{追评得分} & \text{只含追评} \\ 0.4 \times \text{初评得分} + 0.6 \times \text{追评得分} & \text{包含两种} \end{cases} \tag{7}$$

评论深度指的是评论包含的字符数,通常字符数越多,所

包含的信息越多。本文分别计算初评和追评的评论深度,并将初评和追评的评论深度相加得到综合评论深度,其中初评或追评无效时,认为其深度为0。评论数量则为某商品截止到爬取时间之前累计的有效评论数,其中同一用户发布的初评和追评只计入一次。

2.2.2 汉服特有属性

汉服作为蕴含特殊文化的服装,和普通服装的类别划分有所不同,本文主要从形制(汉服所属朝代)和款式(汉服本身特点)两方面对汉服类别进行区分。汉服的形制和款式的分类情况如表1所示:

表1 汉服形制和款式分类

Tab. 1 Classification of Hanfu forms and styles

分类标准	分类情况
形制	秦汉、魏晋、唐、宋、明、汉元素
款式	裾、交裔裙、百迭裙、三涧裙、两片裙、马面裙、褶裙、改良、其他

注:本文提到的形制和款式与汉服研究中的形制、款式略有不同,只为对汉服进行区分,对汉服本身的演变历史研究不具有参考意义,仅做本文研究使用

2.2.3 汉服传播效果

汉服在线上平台的传播效果主要从三个平台出发进行衡量,分别是以百度指数平台为依据研究搜索行为、以新浪微博平台为依据研究参与行为和以抖音平台为依据研究各汉服店铺对自身的传播效果。

百度指数主要收录不同关键词在百度平台的搜索量。本文以“汉服”为关键词,利用百度指数平台提供的需求图谱和相关词热度进行关键词扩充;之后删除无关词,并根据相关系数的大小进行关键词选取,最终选定包括“汉服”“汉服图片”“马面裙”等在内的八个关键词;以关键词的相关系数大小所占比例作为权重,得到了百度指数传播效果。汉服在百度指数平台各月的传播效果如表2所示。

表2 百度指数传播效果

Tab. 2 Spread effects of Baidu index

月份	百度指数传播效果	月份	百度指数传播效果
2023年11月	29 545.462 7	2024年2月	45 033.572 8
2023年12月	28 958.659 7	2024年3月	53 097.678 2
2024年1月	32 717.330 3	2024年4月	52 309.985 2

微博平台爬取的数据为“#汉服#”话题词的阅读量、讨论量、互动量及原创量;抖音平台爬取的数据为各店铺每条视频的名称、发布时间、收藏数、评论数、点赞数和分享数。结合众多学者提出的微博、抖音平台传播效果的衡量方法以及通过熵权-CRITIC权重得到的各指标综合权重,本文按照公式(8)和公式(9)来计算微博和抖音平台汉服的传播效果。

$$S_{wi}=0.25log_{10}R_i+0.2log_{10}O_i+0.3log_{10}D_i+0.25log_{10}I_i \tag{8}$$

$$S_{dij}=\ln(0.3P_{ij}+0.15L_{ij}+0.2C_{ij}+0.2S_{ij}+0.15E_{ij}) \tag{9}$$

式中:S_{wi}为第i个月微博的传播效果;R_i为第i个月“#汉服#”话题词的阅读量;O_i为第i个月“#汉服#”话题词的原创量;D_i为第i个月“#汉服#”话题词的讨论量;I_i为第i个月“#汉服#”话题词的互动量。S_{dij}为第j个店铺第i个月抖音视频的传播效果;P_{ij}为第j个店铺第i个月抖音视频的发布量;L_{ij}为第j个店铺第i个月抖音视频的点赞数;C_{ij}为第j个店铺第i个月抖音视频的收藏数;S_{ij}为第j个店铺第i个月抖音视频的分享数;E_{ij}为第j个店铺第i个月抖音视频的评论数。

得到的微博平台用户的参与行为和抖音平台各店铺的传播效果的衡量结果如表3、表4所示:

表3 微博“#汉服#”话题词传播效果

Tab. 3 Spread effects of Weibo's “# Hanfu #” topic keywords

时间	话题词传播效果	时间	话题词传播效果
2023年11月	118.987 3	2024年2月	110.739 7
2023年12月	118.244 0	2024年3月	121.526 7
2024年1月	116.050 9	2024年4月	113.488 8

表4 抖音平台的汉服传播效果

Tab. 4 Spread effect of Hanfu on Douyin platform

时间	池夏	十三余	花朝记
2023年11月	6.848 2	8.518 7	6.709 3
2023年12月	8.228 0	8.494 4	7.009 2
2024年1月	7.745 1	8.073 4	6.929 8
2024年2月	7.357 3	8.164 4	6.544 3
2024年3月	5.518 1	8.248 1	6.180 4
2024年4月	5.934 8	5.829 5	8.355 1

2.2.4 传统节日和汉服线下活动

传统节日和大型线下活动是宣传中华传统文化和汉服的合适时机,很多汉服爱好者会选择身着汉服参加相关的活动。目前大多汉服店铺采取新品预售的销售方法,预售期从30~60 d不等,考虑到快递运输和商品退换的情况,人们大概需要提前3个月考虑汉服的购买。因此,本文认为传统节日和大型线下活动所在日期的前三个月的销量会受到影响。本文考虑的传统节日和大型线下活动如表5所示。

表5 传统节日和大型线下活动

Tab. 5 Traditional festivals and large-scale offline events

活动类别	活动名称
传统节日	除夕夜、春节、元宵、清明节、端午节
大型线下活动	西塘汉服文化周、国丝汉服节、花朝节、华服日、海峡汉服文化节

根据传统节日和大型线下活动的日期,可以得到各月受到影响的传统节日和大型线下活动的数量,结果如表 6 所示。

表 6　各月份受到影响的节日数

Tab. 6　Affected festival number of each month

时间	传统节日数/天	大型线下活动数/天
2023 年 11 月	0	2
2023 年 12 月	3	0
2024 年 1 月	3	1
2024 年 2 月	4	3
2024 年 3 月	1	3
2024 年 4 月	2	3

2.3　数据预处理

2.3.1　分类数据的编码处理

分类数据作为一种非数值型变量,可以用于描述和区分数据集中的不同类别,但是却不能进行直接运算。因此,对于分类数据需要进行编码处理。由于本文的分类数据均涉及多种取值情况,为充分挖掘不同取值对汉服销量的影响,选择独热编码对分类数据进行编码处理。独热编码将分类变量改为二进制向量的表示,首先将分类值映射到整数值,然后将每个整数值转化为二维向量,向量中只有一个元素为 1,其余都是 0。需要处理的指标主要有四个,分别是应季、店铺、形制和款式,编码情况如表 7 所示。

表 7　分类数据编码情况

Tab. 7　Encoding of categorical data

变量	取值	编码情况
应季	是	(1, 0)
	否	(0, 1)
形制	秦汉	(1, 0, 0, 0, 0, 0, 0)
	魏晋	(0, 1, 0, 0, 0, 0, 0)
	⋮	⋮
	汉元素	(0, 0, 0, 0, 0, 0, 1)
店铺	池夏	(1, 0, 0)
	十三余	(0, 1, 0)
	花朝记	(0, 0, 1)
款式	裾	(1, 0, 0, 0, 0, 0, 0, 0, 0, 0)
	⋮	⋮
	百迭裙	(0, 0, 0, 0, 0, 0, 0, 0, 0, 1)

2.3.2　标准化处理

数据标准化可以将数据转化到同一尺度或标准上,以消除不同特征量纲和数值范围差异对分析结果的影响。本文选取归一化(最大值最小值标准化)方法,按照下式对数值型数据进行标准化处理。

$$X_i' = \frac{X_i - \min(X)}{\max(X) - \min(X)}, i = 1, 2, \cdots, n \tag{10}$$

式中:X_i 为数据 X 的第 i 个值;X_i' 为 X_i 标准化之后的值。

2.3.3　不平衡处理

由于大部分汉服的月销处于 100 件以下,月销大于 100 件的汉服数量较少,低销汉服和高销汉服的数量比达到 2 884∶390,超过 7∶1,存在不平衡问题。为了得到最佳采样效果,本文对不同采样方法、不同采样比例进行了对比(即使用不平衡前的训练集拟合 SVM 模型,得到分类 AUC 值),如表 8 所示。

表 8　不平衡处理方法效果对比

Tab. 8　Comparison effects of unbalanced sampling methods

采样方法	采样比例	模型效果	采样方法	采样比例	模型效果
未采样	—	0.679 4			
SMOTE	1∶1.5	0.886 0	SMOTE + Tomek Links	1∶1.5	0.886 0
	1∶2	0.872 9		1∶2	0.872 9
	1∶3	0.846 4		1∶3	0.844 8
ADASYN	1∶1.5	0.899 7	SMOTE + ENN	1∶1.5	0.886 8
	1∶2	0.893 6		1∶2	0.869 7
	1∶3	0.857 0		1∶3	0.842 1

综合对比采样结果后,选择 ADASYN 过采样,采样比例为 1∶1.5 的采样方法对数据进行过采样。

2.3.4　特征工程

在收集到的数据中,并不是所有指标都对因变量具有显著的影响。特征工程对指标进行筛选,选出对因变量有显著影响的指标,从而进行后续模型的建立和研究,是训练模型前的必要步骤。为更合理地选择特征,本文使用基于相关系数和非参显著性检验的过滤法和基于递归特征消除的包裹法两种特征选择方法,综合两种方法之后,筛选得到 27 个指标,如表 9 和图 3 所示。

表 9　特征工程筛选得到的指标

Tab. 9　Selected indicators by feature engineering

商品属性角度	评论信息角度	传统文化角度
价格 - 低	得分	传统节日数
价格 - 高	数量	抖音传播效果
上新时间	深度	百度指数传播效果
尺码数	总评价数	微博话题传播效果
颜色分类数	有图评论数	形制_唐
发货时间	追评数	形制_明
上月销量	大家印象总数 - 正	形制_汉元素
应季	大家印象总数 - 负	款式 - 破裙
店铺 - 十三余	问大家回复总数	
店铺 - 花朝记		

在特征工程的过程中,主要删掉了快递费、大家印象种类、问大家问题总数、活动数、形制款式的部分编码变量。结合人们平常网购的习惯、较看重的方面及目前汉服的畅销款式,特征过程是合理的。

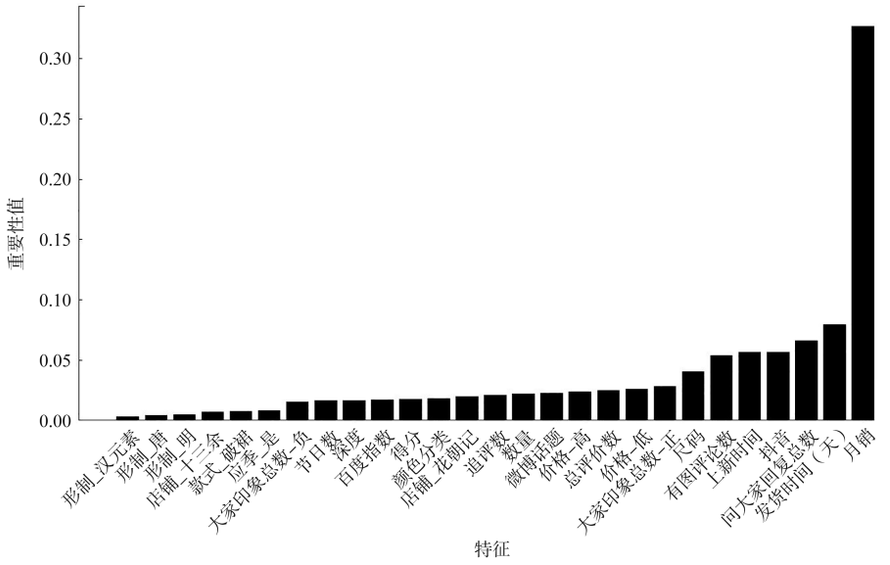


图 3　特征重要性

Fig. 3　Feature importance

3　基于 SVM-双加权 RkNN 的汉服月销二分类问题

由于淘宝平台的商品月销有两种形式，一种是当商品月销小于等于 100 时以具体数值展示；另一种是当商品月销大于 100 时，以形似“100 +”“200 +”这样的分类数据展示。因此，本文首先以 100 为界，解决汉服低销（月销≤100）高销（月销 > 100）的二分类问题。使用上文建立的 SVM-双加权 RkNN 模型，模型的分类效果如表 10 和图 4、图 5 所示。

表 10　SVM-双加权 RkNN 模型分类效果

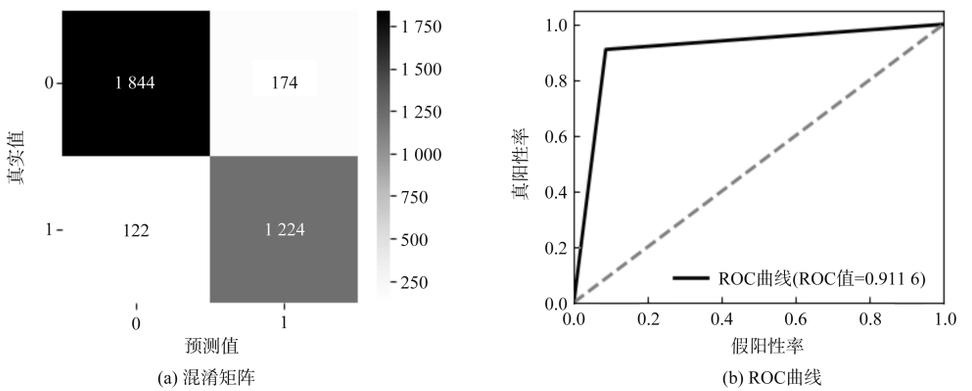
Tab. 10　Classification performance of a SVM-double-weighted RkNN model

评价指标	准确率/%	精确率/%	召回率/%	F1 分数	AUC 值
训练集	0.912 0	0.913 0	0.912 0	0.912 3	0.911 6
测试集	0.883 0	0.915 7	0.883 0	0.893 9	0.848 6

从训练集和预测集上的分类效果来看，本文提出的 SVM-双加权 RkNN 模型在训练集和测试集上都有较好的表现。虽然在测试集上的 AUC 值要小于训练集，但这种差距也是模型训练较为合理的结果。

为进一步证明本文提出的 SVM-双加权 RkNN 模型的分类效果，选取了多个模型进行对比实验，得到的模型效果如表 11 所示。

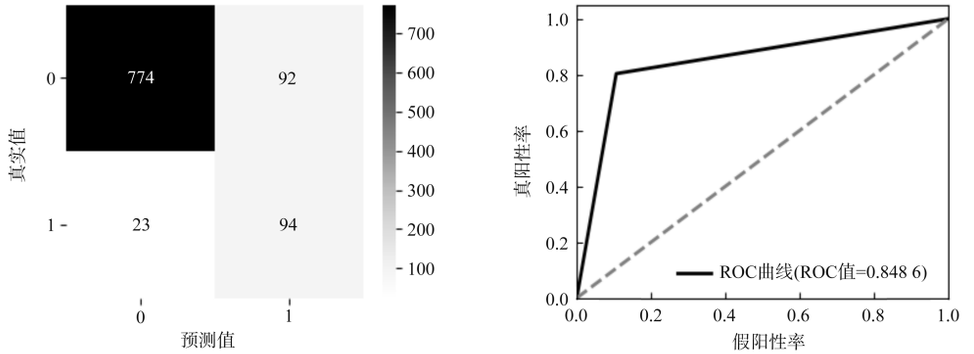
从表 11 可以看到，将本文提出的模型与优化算法调参的 SVM 模型对比，发现 SVM-双加权 RkNN 模型可以有效改善优化算法调参导致的过拟合和欠拟合问题，并且，在不同的评价指标上模型的性能均有所提升；将第二阶段的双加权 RkNN 模型与基础 kNN、RkNN、单一加权 RkNN 模型进行对比，本文建立的 SVM-双加权 RkNN 模型有更好的预测效果。从这两方面都证明本文提出的模型在二分类问题上的有效性。



(a) 混淆矩阵　(b) ROC曲线

图 4　SVM-双加权 RkNN 模型训练集

Fig. 4　Training set of the SVM-double-weighted Rknn model



(a) 混淆矩阵　　(b) ROC曲线

图 5　SVM-双加权 RkNN 模型测试集

Fig. 5　Testing set of the SVM-double-weighted RkNN model

表 11　SVM-双加权 RkNN 模型对比实验结果

Tab. 11　Comparative experimental results of the SVM-double-weighted RkNN model

模型		准确率/%	精确率/%	召回率/%	F1 分数	AUC 值
SVM	训练集	0.869 7	0.870 9	0.869 8	0.870 2	0.867 5
	测试集	0.867 8	0.912 9	0.867 8	0.882 3	0.847 3
网格搜索-SVM	训练集	0.991 9	0.992 1	0.992 0	0.992 0	0.992 8
	测试集	0.929 8	0.927 7	0.929 8	0.928 6	0.816 0
GA-SVM	训练集	0.823 1	0.822 2	0.823 1	0.821 6	0.808 6
	测试集	0.873 8	0.907 3	0.873 9	0.885 5	0.824 9
PSO-SVM	训练集	0.984 8	0.985 1	0.984 8	0.984 9	0.986 3
	测试集	0.919 6	0.919 3	0.919 6	0.919 5	0.806 5
SVM-kNN	训练集	0.882 9	0.886 0	0.882 9	0.883 5	0.884 7
	测试集	0.855 5	0.910 6	0.855 5	0.873 0	0.844 1
SVM-RkNN	训练集	0.898 9	0.900 8	0.898 9	0.899 4	0.899 6
	测试集	0.873 9	0.912 6	0.873 9	0.886 6	0.843 4
SVM-特征加权 RkNN	训练集	0.898 9	0.900 8	0.898 9	0.899 4	0.899 6
	测试集	0.873 9	0.912 6	0.873 9	0.886 6	0.843 4
SVM-投票加权 RkNN	训练集	0.906 9	0.908 3	0.907 0	0.907 3	0.907 1
	测试集	0.876 9	0.914 7	0.876 9	0.889 3	0.848 8
SVM-双加权 RkNN	训练集	**0.912 0**	**0.913 0**	**0.912 0**	**0.912 3**	**0.911 6**
	测试集	**0.883 0**	**0.915 7**	**0.883 0**	**0.893 9**	**0.848 6**

4　基于 SVM-改进双加权 RkNN 的高销汉服三分类问题

在本文收集的数据中，商品月销大于 100 的数据仅有 294 条，占总数据量的 12% 左右，并且这些数据分属于不同的 11 个类，有的类别数据量很小，不便于进行多分类的讨论。因此，本文对高销汉服构建了三分类问题。

如图 6 所示，三分类问题的主要思想是构建两个二分类器，逐步进行两个二分类问题。首先，构建分类器 c1，用于区分类别 1 和月销大于等于 300 的数据；其次，构建分类器 c2，用于区分类别 2 和类别 3。

在构建分类器时发现，由于数据量较小，第一步的 SVM 模型中数据点距分离超平面的距离和数据点是否错分之间的联系减小，并且不同类别间的数据量差异对结果影响较大。因此，在二分类模型的基础上，提出了 SVM-改进双加权 RkNN 模型，该模型的流程图如图 7 所示。

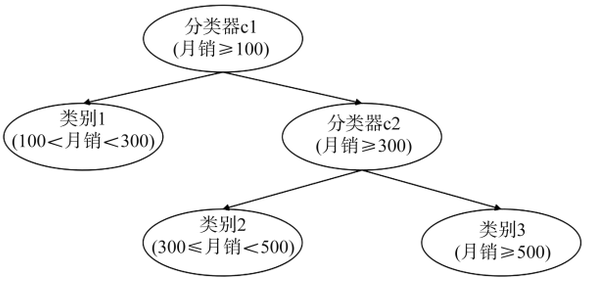


图 6　三分类模型思路

Fig. 6　Thinking of the trinary classification model

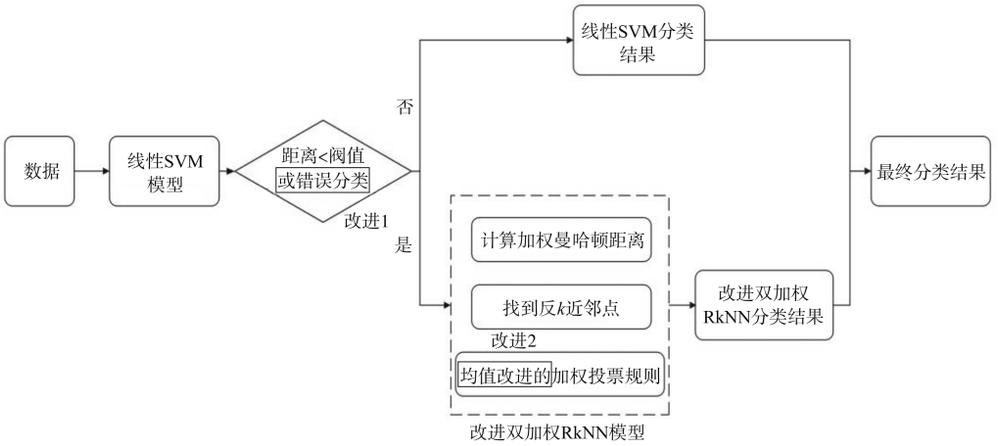


图 7　SVM-改进双加权 RkNN 模型流程示意

Fig. 7　Flowchart of the SVM-improved double-weighted RkNN model

在 SVM-改进双加权 RkNN 模型中,本文主要提出了两点改进:

1)将距分离超平面较近的点和错分点都进行二次分类,解决数据点距分离超平面的距离和数据点是否错分之间的联系减小的问题。

2)更改双加权 RkNN 的投票策略,提出使用平均权重的改进方法。即在投票阶段,把计算各类别样本点的权重之和,改为计算各类别样本点的权重均值,减小各类别样本点数量对分类结果的影响。

对于高销汉服的三分类问题,本文构造出两个 SVM-改进双加权 RkNN 模型。在这里仅展示三分类问题的总分类结果,各二分类分类器的分类结果不再赘述。得到的模型评价指标和混淆矩阵如表 12 和图 8 所示。

从表 12 和图 8 中可以看到,本文构建的三分类模型在训练集和预测集上都有较好的分类效果。其中,训练集上的误分数据主要在 1 类和 2 类之间,测试集上的误分数据主要在 0 类和 2 类之间。

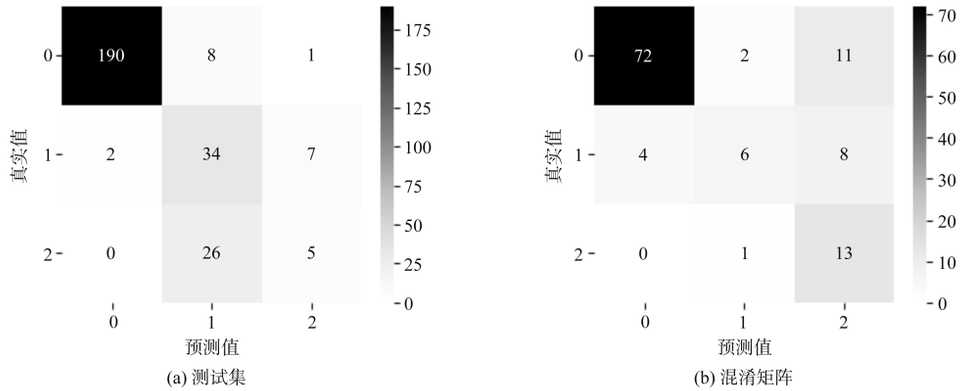
表 12　SVM-改进双加权 RkNN 模型分类效果

Tab. 12　Classification performance of the SVM-improved double-weighted RkNN model

评价指标	准确率/%	精确率/%	召回率/%	F1 分数
训练集	0.838 8	0.843 8	0.838 8	0.830 7
测试集	0.777 8	0.834 9	0.777 8	0.785 8

为进一步验证本文提出的两个改进方法的有效性,分别使用改进之间的模型、采用其中一种改进方法的模型和同时使用两种改进方法的模型进行对比,得到的对比实验结果如表 13 所示。

从表 13 中可以看到,引入改进方法 1 时,训练集和测试集的效果有所下降;引入改进方法 2 时,测试集上的部分评价指标得到了提升;但是,同时引入两个改进方法之后,模型的效果得到了较为明显的提升,训练集和测试集的效果均有所改进,这也证明了,本部分将两种改进方法引入模型是合理的。



(a) 测试集　(b) 混淆矩阵

图 8　三分类问题训练集

Fig. 8　Training set for the ternary classification problem

表 13　SVM-改进双加权 RkNN 模型对比实验结果

Tab. 13　Comparison experimental results of the SVM-improved double-weighted RKNN model

模型		准确率/%	精确率/%	召回率/%	F1 分数
SVM-双加权 RkNN	训练集	0.787 5	0.765 4	0.787 5	0.773 1
	测试集	0.760 7	0.714 6	0.760 7	0.728 4
改进 1 错分点	训练集	0.783 9	0.745 3	0.783 9	0.748 4
	测试集	0.743 6	0.702 5	0.743 6	0.700 9
改进 2 权重均值	训练集	0.787 5	0.765 4	0.787 5	0.773 1
	测试集	0.760 7	0.738 6	0.760 7	0.741 5
SVM-改进双加权 RkNN	训练集	**0.838 8**	**0.843 8**	**0.838 8**	**0.830 7**
	测试集	**0.777 8**	**0.834 9**	**0.777 8**	**0.785 8**

5　结　论

在汉服逐渐被大众接受的当今，本文关注淘宝店铺汉服的销量预测。首先，将汉服特有和传统文化因素纳入销量影响指标体系，考虑文化背景对销量的影响，充分利用多种数据技术，对一些不易量化的文化因素进行科学量化；其次，针对淘宝月销情况，建立二分类问题，提出 SVM-双加权 RkNN 模型，解决 SVM 分离超平面附近点易错分的问题，测试集分类精度和 AUC 值分别为 0.883 0 和 0.848 6，分类效果良好。对高销汉服，建立三分类问题，并进一步提出 SVM-改进双加权 RkNN 模型，测试集分类准确率为 0.777 8，分类精度较好。较精确的二分类和三分类问题的准确率，可以帮助商家较为准确地确定汉服下月的销量，从而根据预测结果和当前库存合理安排后续生产和存储工作，并且，可以针对销量较低的汉服开展促销活动，优化营销策略。

本文的研究只针对汉服的月销进行了分类，而对于低销量汉服是具有具体销量数值的，后续可以展开进一步的研究，对低销量汉服建立数值预测模型，更准确地预测销量信息。

本文研究可以较为精确地预测汉服月销分类情况，现实意义如下：首先，准确的销量预测可以帮助企业制订合理的生产计划，降低库存成本，提高资金周转率。其次，汉服作为传统文化代表，其销量预测有助于了解市场对传统文化的接受程度，推动其传承与发展。最后，汉服销量预测对文化产业的发展具有重要意义，可以帮助政府制定相关政策，推动文化产业繁荣发展。

参考文献：

[1] AKSOY A, OZTURK N, SUCKY E. A decision support system for demand forecasting in the clothing industry [J]. International Journal of Clothing Science and Technology, 2012, 24(4): 221-236.

[2] KULKARNI G, KANNAN P K, MOE W. Using online search data to forecast new product sales[J]. Decision Support Systems, 2012, 52(3): 604-611.

[3] 吕杰妮. 考虑天气信息的服装销售预测：基于机器学习的线下商店实证研究[D]. 杭州：浙江理工大学, 2023.
LÜ J N. Clothing Sales Forecasting Considering Weather Information: An Empirical Study in Brick-and-Mortarstores by Machine-Learning [D]. Hangzhou: Zhejiang Sci-Tech University, 2023.

[4] 孙一文. 基于 RBF 神经网络的服装销售预测研究：以抖音平台为例[D]. 杭州：浙江理工大学, 2023.
SUN Y W. Research on Clothing Sales Forecast Based on RBF Neural Network: Take the Tik Tok Platform as an Example[D]. Hangzhou: Zhejiang Sci-Tech University, 2023.

[5] LI R, YE S W, SHI Z Z. SVM-KNN classifier-a new method of improving the accuracy of SVM classifier[J]. Acta Electronica Sinica, 2002, 30(5): 745-748.

[6] 简强. 基于多特征提取和优化 SVM 的提花针织面料疵点检测算法研究[D]. 杭州：浙江理工大学, 2023.
JIAN Q. Research on Defect Detection Algorithm for Jacquard Knitted Fabric Based on Multi-feature Extraction and Optimized SVM [D]. Hangzhou: Zhejiang Sci-Tech University, 2023.

[7] ABU ALFEILAT H A, HASSANAT AHMAF B A, LASASSMEH O, et al. Effects of distance measure choice on K-nearest neighbor classifier performance: A review[J]. Big Data, 2019, 7(4): 221-248.

[8] 陈丽, 陈静, 高新涛, 等. 基于支持向量机与反 K 近邻的分类算法研究[J]. 计算机工程与应用, 2010, 46(24): 135-137.
CHEN L, CHEN J, GAO X T, et al. Classification algorithm research based on support vector machine and reverse K-nearest neighbor [J]. Computer Engineering and Applications, 2010, 46(24): 135-137.

[9] 张燕, 尹琰, 韦欣宜.《人民日报》抖音号短视频传播热度影响因素实证研究[J]. 中国传媒大学学报(自然科学版), 2020, 27(3): 6-17.
ZHANG Y, YIN Y, WEI X Y. An empirical study on influencing factors of the communication heat of *People's Daily*'s Tiktok account[J]. Information and Communication Research, 2020, 27(3): 6-17.

[10] 杨果. 高校官方微博社会主义核心价值观传播效果的影响因素及赋能路径：基于启发-系统模型的实证分析[J]. 湖南师范大学社会科学学报, 2023, 52(5): 149-156.
YANG G. Influencing factors and empowerment paths of the communication effect of core socialist values on university official microblogs: An empirical analysis based on the heuristic-systematic model[J]. Journal of Social Science of Hunan Normal University, 2023, 52(5): 149-156.

[11] 陆利军. 基于网络搜索指数和 EMD-ARIMA-BP 组合模型的游客量预测：以张家界为例[J]. 吉首大学学报(社会科学版), 2019, 40(1): 138-150.
LU L J. On the prediction of tourist volume based on network search index and EMD-ARIMA-BP combination model: A case study of Zhangjiajie[J]. Journal of Jishou University (Social Sciences), 2019, 40(1): 138-150.

[12] 俞翔, 吴巧英. 基于三角模糊数的汉服袖型感性评价[J]. 丝绸, 2022, 59(8): 62-68.
YU X, WU Q Y. Perceptual evaluation of Hanfu sleeve shapes based on triangular fuzzy number[J]. Journal of Silk, 2022, 59(8): 62-68.

[13] 王思维, 冯思媛, 朱殷, 等. "互联网 +"背景下汉服文化产业发展模式的研究：基于 1002 家汉服淘宝店截面数据的分析[J]. 社会科学前沿, 2020(5): 603-614.
WANG S W, FENG S Y, ZHU Y, et al. Research on the development mode of Hanfu cultural industry under the background of "Internet +": Analysis based on cross-sectional data of 1002 Hanfu Taobao stores[J]. Advances in Social Sciences, 2020(5): 603-614.

[14] CHEVALIER J A, MAYZLIN D. The effect of word of mouth on sales: Online book reviews [J]. Journal of marketing research, 2006, 43 (3): 345-354.

[15] SONNIER G P, MCALISTER L, RUTZ O J. A dynamic model of the effect of online communications on firm sales [J]. Marketing Science, 2011, 30 (4): 702-716.

[16] BLAL I, STURMAN M C. The differential effects of the quality and quantity of online reviews on hotel room sales [J]. Cornell Hospitality Quarterly, 2014, 55(4): 365-375.

Hanfu monthly sales classification based on a SVM-double-weighted RkNN model

XU Xiaonuo, ZHANG Qiumei

(School of Mathematics and Statistics, Changchun University, Changchun 130022, China)

Abstract: As a traditional costume with China's national characteristics, Hanfu has become increasingly prevalent in people's daily lives in recent years. In the development of the new Hanfu industry, the sale of Hanfu constitutes a significant portion. Therefore, this article focuses on predicting the online sales of Hanfu, aiming to provide original Hanfu brands with more future sales information through the forecast results.

First, this article gets the relevant data of Hanfu from Taobao, Douyin, Weibo and Baidu index platforms from November 2023 to July 2024 through Python crawlers. Secondly, In addition to the basic attributes of Hanfu as a commodity, indicators are also constructed from the perspectives of Hanfu's uniqueness and traditional culture, an index system that includes product information, review information, traditional festivals, large offline events, dissemination effects, styles, and designs for online Hanfu sales is established. Finally, based on the monthly sales data characteristics displayed on the Taobao platform, a binary classification is established between low sales (monthly sales less than 100) and high sales Hanfu (monthly sales greater than or equal to 100). The SVM-double-weighted RkNN model is used, which applies the double-weighted RkNN model to perform secondary classification on the points near the separating hyperplane of the SVM model, effectively addressing the issue of misclassification of points near the separating hyperplane in the support vector machine model. On this basis, a more detailed classification of high-sales Hanfu is performed. An improvement is proposed for the binary classification model to make it more suitable for multi-class classification problems based on the data characteristics of multi-classification issues. The innovations of this article are primarily reflected in the following two aspects. First, in today's context where Hanfu is gradually coming into people's view, this article focuses on the prediction of online Hanfu sales, starting from the cultural attributes of Hanfu, quantifying cultural attributes from the unique characteristics of Hanfu and traditional culture, and constructing a relatively complete indicator system that influences online Hanfu sales. Second, when studying the classification problem of online Hanfu sales, this article proposes the SVM-double-weighted RkNN model, which uses a double weighted inverse k-nearest neighbors model to perform secondary classification on points near the separating hyperplane of the support vector machine model, enhancing the classification accuracy. The classification results demonstrate that the proposed model achieves accuracies of 0.883 0 and 0.777 8 for the binary and ternary classification tasks, respectively, indicating robust performance. This model can assist original Hanfu brands in predicting monthly sales, optimizing production and inventory planning, and timely adjusting marketing strategies.

This study fully demonstrates that the monthly sales of Hanfu are closely related to its cultural attributes. A set of indicators that includes cultural attributes can better predict the monthly sales of Hanfu. In future research, further studies could focus on the specific numerical data of low-sales Hanfu by establishing a numerical prediction model. This would enable more accurate sales forecasts and provide more precise predictions of monthly Hanfu sales, so as to assist stores in achieving better and more sustainable development.

Key words: Hanfu; sales classification; influencing factors; evaluation index system; SVM; double-weighted RkNN