基于特征点配准的真伪卷烟商标纸鉴别

冯伟华[1] 王 锐[1] 宗国浩[1] 赵志成[3,4] 罗 泽[3] 周明珠[2] 李晓辉[2] 邢 军[2]

[1](中国烟草总公司郑州烟草研究院 河南 郑州 450001)

[2](中国烟草总公司国家烟草质量监督检验中心 河南 郑州 450001)

[3](中国科学院计算机网络信息中心 北京 100190)

[4](中国科学院大学 北京 100049)

摘 要 为提高真伪卷烟商标纸鉴别的准确性和效率,降低鉴别的经验要求和主观性,提出一种基于特征点配准的真伪卷烟商标纸鉴别方法。使用一致的标准扫描采集卷烟样品图像,基于尺度不变特征转换算法提取图像特征点,通过特征匹配和基于单应性变换的图像配准获取判别预测变量。采用逻辑回归、梯度提升分类决策树算法构建二元分类模型对图像样本进行训练和评估。在 64 个卷烟规格、2 918 个样本数据集上进行实验,该方法准确率高于 95%。通过对比实验验证了该方法的稳定性和有效性。

关键词 卷烟商标纸 真伪鉴别 特征点 图像配准 模型算法 机器学习

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.03.030

IDENTIFICATION OF GENUINE AND FAKE CIGARETTE PACK LABEL BASED ON FEATURE POINT REGISTRATION

Feng Weihua[1] Wang Rui[1] Zong Guohao[1] Zhao Zhicheng[3,4] Luo Ze[3] Zhou Mingzhu[2] Li Xiaohui[2] Xing Jun[2]

[1](*Zhengzhou Tobacco Research Institute of CNTC*, *Zhengzhou 450001*, *Henan*, *China*)

[2](*China National Tobacco Quality Supervision & Test Center*, *Zhengzhou 450001*, *Henan*, *China*)

[3](*Computer Network Information Center*, *Chinese Academy of Sciences*, *Beijing 100190*, *China*)

[4](*University of Chinese Academy of Sciences*, *Beijing 100049*, *China*)

Abstract In order to improve the accuracy and efficiency of the identification of genuine and fake cigarette label paper, and reduce the experience requirements and subjectivity of identification, this paper proposes a method for identifying cigarette label paper based on feature point registration. Sample images were acquired based on a unified image acquisition standard, and discriminant predictors were acquired through feature point extraction and description based on SIFT algorithm, feature matching and image registration based on homograph transformation. Logistic regression algorithm and gradient boosting classification decision tree algorithm were selected to construct a binary classification model for training and evaluation. We performed experimental evaluation on 64 cigarette specifications and 2918 sample data sets. The results show that the accuracy of the proposed identification method is higher than 95%. The stability and effectiveness of the proposed identification method are verified through a comparative experiment.

Keywords Cigarette label paper Identification of genuine and fake Feature point Image registration Model algorithm Machine Learning

---

收稿日期:2020-10-09。国家烟草专卖局科技重大专项(110201901029(SJ-08))。**冯伟华**,工程师,主研领域:数据挖掘。**王锐**,工程师。**宗国浩**,工程师。**赵志成**,博士生。**罗泽**,研究员。**周明珠**,高工。**李晓辉**,高工。**邢军**,研究员。

0　引　言

感官鉴别检验、仪器鉴别检验和评吸鉴别检验是常见的卷烟真伪鉴别检验方法。感官鉴别检验方法主要通过人的感官,包括视觉和触觉等,对比检验样品和真品在印刷工艺、包装工艺、加工工艺等方面的差别,判定卷烟样品的真伪。条装、盒装检验是感官鉴别检验的主要内容之一,主要检验卷烟商标纸包装材料、包装工艺、防伪标志、印刷工艺等的特征及质量。该方法对检验人员及经验要求较高,在鉴别检验比对尤其是卷烟商标印刷比对时,由于视觉疲劳容易造成样品错检。对于印刷品颜色,不同检验人员对色彩的敏感度存在差异,导致判定结果主观性较强。仪器鉴别检验使用专业的检测仪器设备检测样品的物理、化学特性进行分析,对比真品的物理化学指标加以判定。具有判定准确率高、不受主观影响的优点,但与感官鉴别检验相比,其鉴别效率较低。评吸鉴别检验采用样品和真品对比评吸的方法,通过评吸者的主观感受,对样品和真品的香气、香型等差异进行主观判别。与感官鉴别类似,判定结果主观性较强。同时,评吸鉴别对样品造成了破坏,而且鉴别效率较低。

已有不少研究和应用尝试利用计算机视觉技术和机器学习技术提高卷烟真伪鉴别的准确率和检验效率,同时降低鉴别检验的主观性。在 t 假设检验和支持向量机(Support Vector Machine,SVM)算法的基础上,结合卷烟物理指标特性,魏中华[1]提出了一种鉴定卷烟真伪的鉴别模型;聂磊等[2]提出了基于衰减全反射红外光谱法(ATR-FTIR)的鉴别方法,该方法利用真伪烟用材料表层化学成分存在差异,通过采集样品材料特定部位的衰减全反射红外光谱进行对比,实现真伪鉴别。钟宇等[3]利用图像处理技术对卷烟商标纸进行特征提取,为判定卷烟真伪,分别以支持向量机模型、相似性度量模型对特征向量进行分类。除卷烟鉴别检验外,近年来机器学习和人工智能在烟草行业研究和应用越来越多。基于遗传算法,叶安新[4]对烟草配送车路径优化问题开展研究。李敬[5]使用卷积神经网络等深度学习技术开展烟草病虫害识别研究。高震宇等[6]采用卷积神经网络技术,面向烟丝图像建立了分类识别模型,用于烟丝组层的识别和分类。结合烟草物流中不规则烟包码垛的组合匹配特性以及复杂性等特点,张毅等[7]研发了基于计算机视觉的不规则烟包校对码垛系统。王伟等[8]提出了一种能够进行阈值自动校正以及相似性分析的烟箱缺条智能检测技术。

为更好地支持烟草科研应用,建设了烟草科研数据分析模型服务平台,支持烟草科研领域开展基于机器学习算法和模型的行业应用。本文从系统设计、基于特征点的图像配准、观测数据生成等方面详述提出的真伪卷烟商标纸鉴别方法。在烟草科研数据分析模型服务平台上,基于逻辑回归和梯度提升分类决策模型开展实验分析。两种模型的鉴别准确率都达到了 95% 以上。说明提出的基于特征点配准的真伪卷烟商标纸鉴别方法,能够有效地提取出真伪分类的预测指标,进而较准确地完成鉴别任务。同时验证了烟草科研数据分析模型服务平台的模型应用支撑能力。

1　系统设计

基于特征点配准的真伪卷烟商标纸鉴别的系统设计如图 1 所示。包括特征检测、基于 brute-force 的特征匹配、基于单应性变换的图像配准、观测数据生成、分类模型等组成部分。应用流程如下:针对某个卷烟规格,分别对样本库中该规格所有真样本以及测试样本进行特征提取,分别提取每个样本图像的特征点,生成特征点描述向量,在测试样本和真样本图像上使用 brute-force 方法进行特征点匹配,使用单应性变换将测试样本图像配准到目标真样本图像上,采用随机抽样一致(Random Sample Consensus,RANSAC)算法[9],计算出两个图像之间最好的单应性变换。通过与样本库中该规格的所有真样本分别比对,找出其中最相似的真样本。配准过程中提取的参数作为预测变量输入烟草科研数据分析模型服务平台中训练的二元分类模型,做出真伪判别。

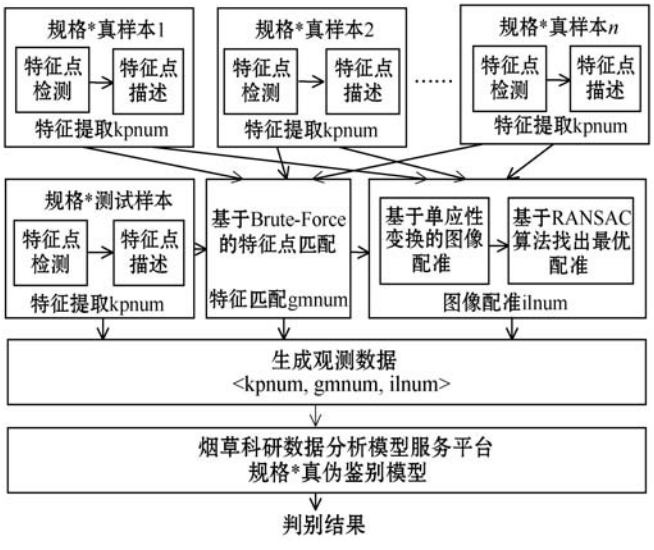


图 1　基于特征点配准的真伪卷烟商标纸鉴别系统设计

2 特征提取和图像配准

2.1 特征点检测和描述

当前常用的图像特征点检测算法有 SIFT[10]、SURF[11]、ORB[12]、AKAZA[13]等,在基于点特征的研究方法中,尺度不变特征转换(Scale-Invariant Feature Transform,SIFT)算法[10]检测的特征点在图像透视变换、尺度缩放、刚性(平移、旋转)变换,以及光照变化等情况下具有较好的稳定性,本文选择使用 SIFT 算法[10]进行图像特征点提取。图像上检测到的特征点的数目 k_{pnum}作为观测数据的第一个预测参数。

该算法包含以下部分:

(1) 尺度空间建立。图像的尺度空间 $L(x,y,\delta)$ 定义如式(1)所示。

$$L(x,y,\delta)=G(x,y,\delta)*I(x,y) \tag{1}$$

尺度空间表示为原图像 $I(x,y)$ 与高斯函数 $G(x,y,\delta)$ 的卷积,其中 * 表示卷积操作。尺度变化的高斯函数如式(2)所示。

$$G(x,y,\delta)=\frac{1}{2\pi\delta^2}\mathrm{e}^{\frac{-\left(x-\frac{m}{2}\right)^2+\left(y-\frac{n}{2}\right)^2}{2\delta^2}} \tag{2}$$

式中:m、n 表示卷积模板矩阵的大小;(x,y) 表示像素在图像上的坐标位置;使用高斯金字塔构建图像尺度空间;δ 表示尺度空间因子。

(2) 特征点定位。基于高斯差分(Difference of Gaussian,DoG)算子与尺度归一化的高斯拉普拉斯(Laplacian of Gaussian,LoG)算子的近似特征,用 DoG 代替 LoG 进行极值检测。在高斯金字塔中,将每 octave 中相邻的两层图像对应相减,如式(3)所示,得到 DoG 尺度空间。

$$\begin{aligned}D(x,y,\delta)&=(G(x,y,k\delta)-G(x,y,\delta))*I(x,y)=\\&L(x,y,k\delta)-L(x,y,\delta)\end{aligned} \tag{3}$$

DoG 尺度空间中的局部极值点即检测出的特征点,在此基础上进一步确定特征点的准确位置和响应尺度。

(3) 特征点方向确定。对检测出的特征点,根据式(4)和式(5)分别计算特征点 3δ 邻域窗口内像素梯度的模值和方向。

$$m(x,y)=\sqrt{\begin{aligned}&(L(x+1,y)-L(x-1,y))^2+\\&(L(x,y+1)-L(x,y-1))^2\end{aligned}} \tag{4}$$

$$\theta(x,y)=\tan^{-1}\left(\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)}\right) \tag{5}$$

统计邻域内像素的梯度和方向,最多的统计方向即为特征点的主方向。

(4) 特征描述。在特征点所在尺度空间的 4×4 邻域窗口中,分别计算每个区域 8 个方向的梯度信息,形成 4×4×8=128 维向量,使用该向量描述特征点。

2.2 特征匹配和图像配准

(1) 特征匹配。在对两幅图像进行特征匹配时,使用 brute-force 方法计算两个特征向量的距离,根据距离确定两个特征点的匹配程度。Brute-force 算法的原理如下:基于每个特征点表示为 1 个 128 维的特征向量,对源图像中的每一个特征点,与目标图像中所有特征点分别计算欧氏空间距离,并将得到的结果按特征点对之间的欧氏空间距离由小到大进行排序。在此基础上,对最近邻和第二近邻匹配的距离比设定阈值进行筛选,选出可能性较高的匹配特征点,记为 G_{MP}。在实验中,选择 0.75 的阈值。满足要求的匹配对的数目记为 g_{mnum},作为模型观测数据的第二个预测参数。

(2) 单应性变换。单应性变换[14]是一种具有保线性质的线性变换,可以表示为一个 3×3 的非奇异矩阵。假设已经计算出两个图像之间的单应性矩阵 $\boldsymbol{H}$,则两幅图像之间可以建立如式(6)所示的关系。$\boldsymbol{H}$ 如式(7)所示。

$$\begin{pmatrix}x'_i\\y'_i\\1\end{pmatrix}=\begin{pmatrix}h_{11}&h_{12}&h_{13}\\h_{21}&h_{22}&h_{23}\\h_{31}&h_{32}&h_{33}\end{pmatrix}\begin{pmatrix}x_i\\y_i\\1\end{pmatrix} \tag{6}$$

$$\boldsymbol{H}=\begin{pmatrix}h_{11}&h_{12}&h_{13}\\h_{21}&h_{22}&h_{23}\\h_{31}&h_{32}&h_{33}\end{pmatrix} \tag{7}$$

式中:$(x'_i,y'_i,1)^{\mathrm{T}}$表示目标图像中的像素坐标;$(x_i,y_i,1)^{\mathrm{T}}$表示源图像中的对应像素坐标。通过单应性矩阵 $\boldsymbol{H}$ 将源图像变换到目标图像空间。实现基于单应性变换的图像配准。

根据式(6),可以得到:

$$\begin{cases}x'_i=h_{11}x_i+h_{12}y_i+h_{13}\\y'_i=h_{21}x_i+h_{22}y_i+h_{23}\\1=h_{31}x_i+h_{32}y_i+h_{33}\end{cases} \tag{8}$$

根据 $1=h_{31}x_i+h_{32}y_i+h_{33}$,可以得到:

$$\begin{cases}x'_i=\dfrac{x'_i}{1}=\dfrac{h_{11}x_i+h_{12}y_i+h_{13}}{h_{31}x_i+h_{32}y_i+h_{33}}\\[2ex]y'_i=\dfrac{y'_i}{1}=\dfrac{h_{21}x_i+h_{22}y_i+h_{23}}{h_{31}x_i+h_{32}y_i+h_{33}}\end{cases} \tag{9}$$

可以推导出:

$$\begin{cases}h_{31}x_ix'_i+h_{32}y_ix'_i+h_{33}x'_i-(h_{11}x_i+h_{12}y_i+h_{13})=0\\h_{31}x_iy'_i+h_{32}y_iy'_i+h_{33}y'_i-(h_{21}x_i+h_{22}y_i+h_{23})=0\end{cases} \tag{10}$$

即 1 对匹配点对应式(10)。

$$\frac{1}{h_{33}}\boldsymbol{H}=\begin{bmatrix}\frac{h_{11}}{h_{33}} & \frac{h_{12}}{h_{33}} & \frac{h_{13}}{h_{33}}\\ \frac{h_{21}}{h_{33}} & \frac{h_{22}}{h_{33}} & \frac{h_{23}}{h_{33}}\\ \frac{h_{31}}{h_{33}} & \frac{h_{32}}{h_{33}} & 1\end{bmatrix} \tag{11}$$

从式(11)可知,$\boldsymbol{H}$ 有 8 个自由度。因此至少需要 4 对匹配点,构建 4 个式(10),即可通过求解线性方程组得到单应性矩阵 $\boldsymbol{H}$。

(3) 最优单应性变换计算。基于 RANSAC 算法[9]计算最优单应性变换,算法描述如下:

步骤 1　从 G_{MP} 中随机选择 4 对配对的特征点,计算两幅图像间的单应性变换矩阵 $\boldsymbol{H}$。

步骤 2　计算在 $\boldsymbol{H}$ 表示的单应性变换下,G_{MP} 中所有配对的特征点是否外点或内点。判别依据如式(12)所示。

$$\left\| t_{\mathrm{argetPointS}_i} - \boldsymbol{H} * s_{\mathrm{ourcePointS}_i} \right\| > t_{\mathrm{hreshod}} \tag{12}$$

对 G_{MP} 中的每对配对点,将配对点中源图像上的特征点(以 $s_{\mathrm{ourcePointS}_i}$ 表示)根据式(6)进行单应性变换后,计算变换后的特征点与目标图像上对应特征点(以 $t_{\mathrm{argetPointS}_i}$ 表示)之间的距离是否大于阈值 t_{hreshold},如果大于阈值,则特征点 i 是外点,对应的配对是错误的配对。

步骤 3　计算单应性变换 $\boldsymbol{H}$ 的反向投影错误率。如式(13)所示。累加每对配对点投影后产生的误差。

$$b_{\mathrm{pe}} = \sum_i \left(x'_i - \frac{h_{11}x_i + h_{12}y_i + h_{13}}{h_{31}x_i + h_{32}y_i + h_{33}} \right)^2 + \left(y'_i - \frac{h_{21}x_i + h_{22}y_i + h_{23}}{h_{31}x_i + h_{32}y_i + h_{33}} \right)^2 \tag{13}$$

步骤 4　重复执行步骤 1 - 步骤 3,直到穷尽所有随机选择或达到预先设置的执行次数。反向投影错误率最小的单应性变换即为最优选择、最优变换下的内点数目,即正确匹配的特征点数目,记为 i_{lnum},是度量两幅图像相似度的重要指标,作为观测数据的第三个预测参数。

2.3　观测数据生成

真伪卷烟商标纸图像呈现出类内差异较大、类间差异较小的特点。导致类内差异较大的因素较多,包括同一规格的卷烟商标纸可能由不同的厂家印刷、由不同的包装机型封装打包、不同年份、批次设计略有不同等;较小的类间差异主要是因为假烟的商标纸在印刷和封装过程中,尽可能地模仿相应的真烟商标纸特征,目标是达到靠视觉感官难以区分的程度。因此在样本库的建设过程中,需要尽可能地收集能够体现类内差异的不同的真品样本。基于此,在训练和测试过程中,将卷烟规格中的所有样本(包括真样本和假样本)作为源图像,分别与本规格真样本中的每个图像作为目标图像,按前述的方法进行图像配准,计算出相应的 k_{pnum}、g_{mnum} 和 i_{lnum},并标记相应的样本类型(真为 1,假为 0)。对于假样本,选择 i_{lnum} 最高的记录作为该样本的观测数据。即从所有对比中,选择与其最相似的真样本的配对数据作为观测记录。对于真样本,因为 i_{lnum} 最高的记录是该真样本本身,所以选择 i_{lnum} 第二高的记录作为该样本的观测数据。获得观测数据后,将真伪卷烟商标纸鉴别建模为二元分类问题,在烟草科研数据分析模型服务平台上进行算法选择、模型构建、训练和评估。

3　烟草科研数据分析模型服务平台

3.1　平台介绍

烟草科研数据分析模型服务平台是一个支持分布式部署的协作环境,前端 Web 服务器支持用户发布、共享、访问模型。机器学习模型系统、RStudio、R Shiny 模型交互分析系统等应用系统运行在后端执行主机构成的集群服务器上。机器学习模型系统、RStudio、R Shiny 模型交互分析系统提供基于 Web 的访问界面和用户接口,系统功能通过运行在容器环境中的 Web 服务实例提供。烟草科研数据分析模型服务平台结构如图 2 所示。

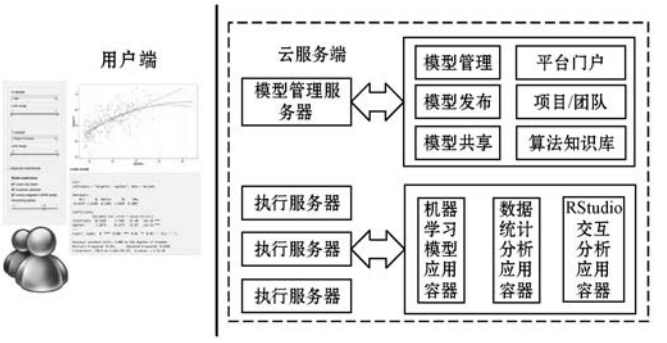


图 2　烟草科研数据分析模型服务平台

机器学习模型系统与 Hadoop[15]、Spark[16] 等大数据计算平台环境深度集成,支持将模型训练及预测计算提交到大数据分析处理平台进行高效计算。整合了主流的机器学习算法,例如广义线性模型,包括线性回归(Linear Regression)[17]、逻辑回归(Logistic Regression)[18] 等、朴素贝叶斯(Naïve Bayes Classifier)[19]、主成分分析(Principal Component Analysis)[20]、K-means 聚类(K-means Clustering)[21]、随机森林(Random For-

est)[22]、梯度提升(Gradient Boosting Machine)[23]和深度学习(Deep Learning)等[24-25]。提供了丰富的模型算法组件。用户训练的模型可以被编译为 Java 对象,容易地嵌入到 Java 环境中。提供了基于 Web 的用户接口,支持用户基于 Web 实现数据导入、算法选择、模型训练、模型预测和评估。同时,提供了面向 R、Python、Java 等语言的应用编程接口。

烟草科研数据分析模型服务平台将 RStudio IDE[26]的 Web 版本部署到应用容器中。同样的机制适用于基于 Shiny[27-28](Shiny 是基于 R 构建 Web 应用的软件包)构建的应用,以及基于 RMarkdown[29]的文档。基于 R 和 Shiny,构建了数据统计分析系统,实现了数据集载入、预览、统计分析、透视分析,支持数据集变换、合并、简单的可视化等数据操作。

3.2 真伪卷烟商标纸鉴别模型

在烟草科研数据分析模型服务平台上,分别采用逻辑回归算法和梯度提升分类决策树算法建模二元分类进行真伪卷烟商标纸鉴别。

(1) 逻辑回归算法。烟草科研数据分析模型服务平台上支持的逻辑回归算法原理描述如下。将响应向量 $\boldsymbol{y}$ 和输入向量 $\boldsymbol{x}$ 之间的关系建模,如式(14)所示。

$$\hat{\boldsymbol{y}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\beta}_0 \tag{14}$$

式中:$\boldsymbol{\beta}$ 表示参数向量;$\boldsymbol{\beta}_0$表示截距项。拟合的模型如式(15)所示。

$$\hat{\boldsymbol{y}} = \Pr(y = 1 \mid x) = \frac{\mathrm{e}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\beta}_0}}{1 + \mathrm{e}^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\beta}_0}} \tag{15}$$

可进一步转换为:

$$\log\left(\frac{\hat{y}}{1 - \hat{y}}\right) = \log\left(\frac{\Pr(y = 1 \mid x)}{\Pr(y = 0 \mid x)}\right) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\beta}_0 \tag{16}$$

通过最大化似然函数进行拟合,得到式(17)。

$$\max_{\boldsymbol{\beta},\boldsymbol{\beta}_0} \frac{1}{N}\sum_{i=1}^{N}\left(y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\beta}_0) - \log(1 + \mathrm{e}^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\beta}_0})\right) - \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + \frac{1}{2}(1 - \alpha)\|\boldsymbol{\beta}\|_2^2\right) \tag{17}$$

式中:参数 λ 控制正则化强度;参数 α 控制 L_1 和 L_2 范数之间的权重分布。

相应的偏差如式(18)所示。

$$D = -2\sum_{i=1}^{n}\left(\hat{y}_i\log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right) \tag{18}$$

(2) 梯度提升决策树。梯度提升是整合了基于梯度的优化和提升(boosting)两种工具的机器学习技术。基于梯度的优化使用梯度计算损失函数。提升指通过逐步增加弱模型,创建一个鲁棒的、用于预测任务的集成学习系统。烟草科研数据分析模型服务平台上的梯度提升分类决策树(Gradient Boosting Decision Tree,GBDT)算法描述如算法 1 所示。实现了一个 K 类分类模型。

算法 1 GBDT 实现 K 分类

1. 初始化 $f_{k0} = 0, k = 1,2,\cdots,K$
2. For $m = 1$ to M
3. 设置 $p_k(x) = \dfrac{\mathrm{e}^{f_k(x)}}{\sum_{l=1}^{K}\mathrm{e}^{f_l(x)}}, k = 1,2,\cdots,K$
4. For $k = 1$ to K
5. 计算 $r_{ikm} = y_{ik} - p_k(x_i), i = 1,2,\cdots,N$
6. 以 $r_{ikm}, i = 1,2,\cdots,N$ 为目标,构建 CART 进行拟合,叶子节点为 $R_{jkm}, j = 1,2,\cdots,J_m$
7. 计算 $\gamma_{jkm} = \dfrac{K-1}{K}\dfrac{\sum_{x_i \in R_{jkm}}(r_{ikm})}{\sum_{x_i \in R_{jkm}}|r_{ikm}|(1 - |r_{ikm}|)}, j = 1,2,\cdots,J_m$
8. 更新 $f_{km}(x) = f_{k,m-1}(x) + \sum_{j=1}^{J_m}\gamma_{jkm}I(x \in R_{jkm})$
9. 输出 $\hat{f}_k(x) = f_{kM}(x), k = 1,2,\cdots,K$

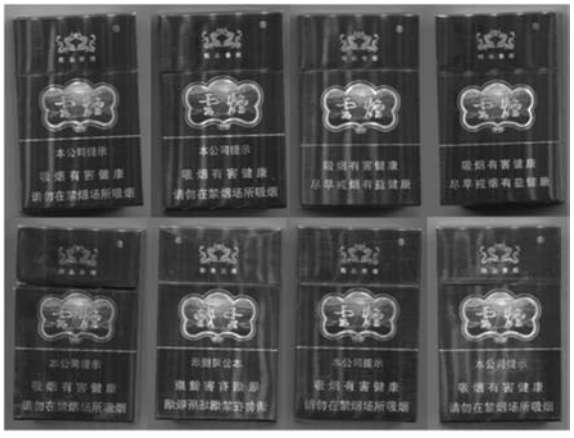
在算法 1 中构建了 k 个回归树,每个树表示一个目标类。m 表示加入到当前整体中的弱分类器的个数。在内循环中,第一步首先计算残差 r_{ikm}(算法中第 5 步),其实际上是分类回归决策树(Classification And Regression Trees,CART)的 N 个分箱上的梯度值。然后构建一个回归树来拟合这些梯度计算(算法中第 6 步)。对于生成的决策树,分别计算各个叶子节点最好的负梯度拟合的近似值(算法中第 7 步)。基于梯度下降的优化方法将构建的回归树加入整体集成学习模型中提高训练精度(算法中第 8 步)。经过 M 次迭代,完成训练,用于预测任务。

4 实验分析

4.1 实验数据集

采集经人工检验确认真伪的 64 个卷烟规格共计 2 918 个样本,其中:真样本 1 428 个,假样本 1 490 个。以规格 A 和规格 B 为例,其中:规格 A 真样品 55 份,假样品 49 份,合计 104 份;规格 B 真样品 22 份,假样品 27 份,合计 49 份。采用统一的图像采集标准,使用扫描仪扫描采集卷烟商标纸。扫描设备为 EPSON V370 扫描仪,扫描分辨率为 1 200 dpi,扫描模式为全彩模式。图 3 展示了采集图像示意图。其中:图 3(a)为规格 A 图像,图 3(b)为规格 B 图像,图的上半部分为真样品采集图像,下半部分为假样品采集图像。规格 A 需要进行 5 720 次图像配准,生成 104 条观测记录,规格 B 需要进行 1 078 次图像配准,生成 49 条观

测记录。生成的观测数据如表 1 所示。规格表示卷烟的规格号,k_{pnum}是从源样本图像中检测到的点特征数目,g_{mnum}是满足阈值要求的匹配特征点对数,i_{lnum}是配准后匹配的特征点的数量,类别表示真伪类别(1 为真,0 为伪)。



(a)

(b)

图 3　采集图像示意图

表 1　观测数据示例

规格	k_{pnum}	g_{mnum}	i_{lnum}	类别
6901028046886	42 957	2 705	1 922	1
6901028046886	73 985	2 038	1 025	1
6901028046886	…	…	…	…
6901028046886	48 521	641	213	0
6901028046886	24 194	634	230	0
6901028120692	16 638	1 672	1 089	1
6901028120692	24 640	1 751	1 191	1
6901028120692	…	…	…	…
6901028120692	21 507	919	505	0
6901028120692	15 857	853	454	0

4.2　逻辑回归模型

逻辑回归模型参数设置如表 2 所示。

表 2　逻辑回归模型参数设置

参数	值	描述
response_column	类别	响应变量列
family	binomial	逻辑回归二元分类
solver	IRLSM	设置求解器
alpha	0.5	L_1 和 L_2 惩罚之间正则化的分布
lambda	0.000 844	指定正则化强度
max_iterations	50	迭代最大次数
link	logit	Link 函数

在 64 个卷烟规格的实验数据上分别进行测试和评估,从实验结果看,最低准确率达到了 95.4%。以规格 A 和规格 B 为例,图 4 展示了在规格 A 和规格 B 的观测数据集上分别训练逻辑回归模型得到的两个规格真伪判别的混合矩阵。规格 A 的 104 个样本数据中,有 1 个发生了误判,总体准确率达到 99%;规格 B 的 49 个样本数据中,有 2 个样本发生了误判,总体准确率达到 96%。

实际值/预测值	0	1	错误	误判率	准确率
0	49	0	0	0/49	1
1	1	54	0.018	1/55	0.98
Total	50	54	0.010	1/104	
Recall	0.98	1			

(a) 规格 A 逻辑回归模型

实际值/预测值	0	1	错误	误判率	准确率
0	26	1	0.037	1/27	0.96
1	1	21	0.046	1/22	0.95
Total	27	22	0.041	2/49	
Recall	0.96	0.95			

(b) 规格 B 逻辑回归模型

图 4　训练逻辑回归模型得到的两个混合矩阵

4.3　梯度提升决策树模型

梯度提升决策树模型参数设置如表 3 所示。构建 50 个分类回归树,树的最大层级深度为 5。对二元分类问题,选择伯努利(bernoulli)分布函数作为响应变量分布函数。

表 3　梯度提升分类决策模型参数设置

参数	值	描述
response_column	类别	响应变量列
ntrees	50	构建的分类回归树的数量
max_depth	5	最大树深度
nbins	20	对数值列,构建具有该分箱数的直方图

续表 3

参数	值	描述
learn_rate	0.1	学习率
stopping_metric	logloss	提前停止的度量方式
stopping_tolerance	0.001	提前停止的相对容差
distribution	bernoulli	响应变量分布函数

在 64 个卷烟规格的实验数据上分别进行测试和评估,从实验结果看,最低准确率达到了 95.2%。以规格 A 和规格 B 为例,图 5 展示了在规格 A 和规格 B 的观测数据集上分别训练梯度提升分类决策模型得到的两个规格真伪判别的混合矩阵。规格 A 的 104 个样本数据中,有 2 个发生了误判,总体准确率达到 98%,总体召回率达到 98%;规格 B 的 49 个样本数据中,有 2 个样本发生了误判,总体准确率达到 96%,总体召回率达到 95%。

实际值/预测值	0	1	错误	误判率	准确率
0	48	1	0.020 4	1/49	0.98
1	1	54	0.018 2	1/55	0.98
Total	49	55	0.019 2	2/104	
Recall	0.98	0.98			

(a) 规格 A 梯度提升分类决策树模型

实际值/预测值	0	1	错误	误判率	准确率
0	26	1	0.037 0	1/27	0.96
1	1	21	0.045 5	1/22	0.95
Total	27	22	0.040 8	2/49	
Recall	0.96	0.95			

(b) 规格 B 梯度提升分类决策树模型

图 5 训练梯度提升分类决策模型得到的两个混合矩阵

4.4 基于不同特征点检测算法的对比实验

采用相同的实验方案和参数设置,将图像特征点检测算法由 SIFT 替换为 SURF、ORB、BRISK 和 AKAZE,分别在规格 A 和规格 B 上进行了误判率的对比实验。实验结果如图 6 所示。

模型	SIFT	SURF	ORB	BRISK	AKAZE
逻辑回归模型	1/104	2/104	2/104	3/104	4/104
梯度提升决策树模型	2/104	2/104	2/104	3/104	3/104

(a) 规格 A

模型	SIFT	SURF	ORB	BRISK	AKAZE
逻辑回归模型	2/49	3/49	2/49	2/49	3/49
梯度提升决策树模型	2/49	2/49	3/49	3/49	2/49

(b) 规格 B

图 6 误判率对比结果

可以看出:一方面,五类特征点检测算法都能够有效地提取出真伪卷烟商标纸图像特征,支持后续分析模型作出有效判别(规格 A 判别最低准确率为 96%,规格 B 判别最低准确率为 94%)。另一方面,基于 SIFT 算法的鉴别模型与其他模型相比,在两个规格上都具有最低的误判率。这验证了本文方法的稳定性和有效性。

4.5 讨 论

基于统一标准采集的 64 个卷烟规格的商标纸图像数据,通过特征点提取和描述、特征匹配和图像配准获得观测数据,在烟草科研数据分析模型服务平台上,基于逻辑回归模型和梯度提升决策树模型进行真伪卷烟商标纸鉴别,两种模型的鉴别准确率都达到了 95% 以上。说明提出的基于特征点配准的真伪卷烟商标纸鉴别方法,能够有效地提取出真伪分类的预测指标,进而较准确地完成鉴别任务。结合烟草科研数据分析模型服务平台提供的算法,通过模型构建、训练和评估,有效验证了本文方法的有效性。验证了烟草科研数据分析模型服务平台的模型应用支撑能力。

另一方面,如前所述,真伪卷烟商标纸图像呈现出类内差异较大、类间差异较小的特点。因此,选择配准得最好情况下的样本数据作为观测数据,体现了样本与真样品相似度的上限。该方法受限于以下两个方面:(1) 需要积累建设完备的真品样本库。如果真品样本缺失,可能导致误判。(2) 每次判别的过程,需要与样本库中本规格所有的真样品进行配准比对,导致计算成本较高。在后续工作中,将持续构建样本库,进一步探索提升鉴别效率的方法。

5 结 语

本文以经人工检验确认真伪的 64 个卷烟规格,共计 2 918 个样本为基础。从图像采集、基于特征点的图像配准、观测数据生成、分类算法等方面详述提出的真伪卷烟商标纸鉴别方法。采用统一的图像采集标准,使用扫描仪扫描采集卷烟商标纸图像。应用特征点检测算法提取卷烟商标纸图像特征点。检测到的特征点的数目 k_{pnum} 作为观测数据的第一个预测参数。找出两幅图像中每个特征点对应的最匹配的特征点,在此基础上,对最近邻和第二近邻匹配的距离比设定阈值进行筛选,选出可能性高的匹配对。在实验中,选择

0.75 的阈值,满足要求的匹配对的数目记为 g_{mnum},作为模型观测数据的第二个预测参数。采用 RANSAC 算法,计算出两幅图像最好的配准结果,以及在此配准下,匹配的特征点对数目,记为 i_{Inum},作为观测数据的第三个预测参数。针对每个样本,选择与其最相似的真样本的配对数据作为观测记录。在烟草科研数据分析模型服务平台上,选择逻辑回归算法、梯度提升分类决策树算法构建二元分类模型,并对模型进行训练和评估。从实验结果看,本文方法准确率达到了 95% 以上,说明本文方法能够有效地提取出真伪分类的预测指标,结合烟草科研数据分析模型服务平台提供的算法,通过模型构建、训练和评估,有效验证了本文方法的有效性。同时验证了烟草科研数据分析模型服务平台的模型应用支撑能力。

参 考 文 献

[1] 魏中华. 基于 t 假设检验及 SVM 神经网络的卷烟真伪判定[J]. 烟草科技,2015,48(2):75 - 78.

[2] 聂磊,别振英,朱友,等. 基于烟用材料的衰减全反射红外光谱无损鉴别真假卷烟[J]. 烟草科技,2019,52(5):31 - 39.

[3] 钟宇,徐燕,刘德祥,等. 基于计算机视觉和机器学习的真伪卷烟包装鉴别[J]. 烟草科技,2020,53(5):83 - 92.

[4] 叶安新. 基于遗传算法的烟草配送车路径优化问题[J]. 计算机系统应用,2011,20(4):241 - 244.

[5] 李敬. 基于卷积神经网络的烟草病害自动识别研究[D]. 泰安:山东农业大学,2016.

[6] 高震宇,王安,董浩,等. 基于卷积神经网络的烟丝物质组成识别方法[J]. 烟草科技,2017,50(9):68 - 75.

[7] 张毅,王彦博,付华森,等. 基于机器视觉的不规则烟包校对码垛系统[J]. 烟草科技,2019,52(6):105 - 111.

[8] 王伟,朱立明,章强,等. 基于相似性分析和阈值自校正的烟箱缺条智能检测方法[J]. 烟草科技,2019,52(1):91 - 97.

[9] Fischler M, Bolles R. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography[J]. Communications of the ACM, 1981,24(6):381 - 395.

[10] Lowe D. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision,2004, 60(2):91 - 110.

[11] Bay H, Ess A, Gool L, et al. Speeded-up robust features (SURF)[J]. Computer Vision and Image Understanding, 2008,110(3):346 - 359.

[12] Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF[C]//International Conference on Computer Vision,2011:2564 - 2571.

[13] Alcantarilla P, Nuevo J, Bartoli A. Fast explicit diffusion for accelerated features in nonlinear scale spaces[C]//British Machine Vision Conference,2013:1 - 12.

[14] Chum O, Pajdla T, Sturm P. The geometric error for homographies[J]. Computer Vision and Image Understanding,2005, 97(1):86 - 102.

[15] Shvachko K, Kuang H, Radia S, et al. The Hadoop distributed file system[C]//2010 IEEE 26th Symposium on Mass Storage Systems and Technologies,2010:1 - 10.

[16] Zaharia M, Chowdhury M, Franklin M, et al. Spark: Cluster computing with working sets[C]//2nd USENIX Conference on Hot Topics in Cloud Computing,2010.

[17] Montgomery D, Peck E, Vining G. Introduction to linear regression analysis[M]. John Wiley and Sons,2001.

[18] Harrell F. Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis [M]. Springer,2015.

[19] John G, Langley P. Estimating continuous distributions in Bayesian classifiers[C]//11th Conference on Uncertainty in Artificial Intelligence,1995:338 - 345.

[20] Deutsch H. Principle component analysis[M]//Derivatives and Internal Models. Palgrave Macmillan,2004:615 - 623.

[21] Coates A, Ng Andrew Y. Learning feature representations with k-means[M]//Neural Networks: Tricks of the Trade. Springer,2012:561 - 580.

[22] Liaw A, Wiener M. Classification and regression by RandomForest[J]. R News,2002,2(3):18 - 22.

[23] Xi Y, Zhuang X, Wang X, et al. A research and application based on gradient boosting decision tree[C]//International Conference on Web Information Systems and Applications, 2018:15 - 26.

[24] LeCun Y, Hinton G. Deep learning[J]. Nature,2015,521: 436 - 444.

[25] 赵志成,罗泽,王鹏彦,等. 基于深度残差网络图像分类算法研究综述[J]. 计算机系统应用,2020,29(1):14 - 21.

[26] Racine J. RStudio: A platform-independent IDE for R and Sweave[J]. Journal of Applied Econometrics,2012,27:167 - 172.

[27] Beeley C, Sukhdeve S. Web application development with R using shiny[M]. Packt,2018.

[28] Vincent N. Radiant: Business analytics using R and shiny [EB/OL]. (2020 - 06 - 06). https://CRAN. R-project. org/package = radiant.

[29] Valdez A. Making reproducible research simple using RMarkdown and the OSF[C]//International Conference on Human-Computer Interaction,2020:27 - 44.