图文信息技术

面向不均衡训练集的印刷图像套准状态检测方法

简川霞，高健
（广东工业大学，广州 510006）

摘要：目的 针对不均衡的印刷图像套准状态检测中存在的印刷套不准图像识别准确率低的问题，研究不均衡印刷图像训练集的预处理方法。**方法** 提出不均衡印刷图像训练集数据的集成采样预处理方法。支持向量机先将不均衡的训练集数据分为支持向量和非支持向量，然后过采集少类样本（即印刷套不准图像）中的支持向量，欠采集多类样本（即印刷套准图像）中的非支持向量，实现训练集数据的均衡化。最后采用预处理后的均衡训练集对支持向量机模型进行训练，并优化模型参数。**结果** 采用文中提出的集成采样方法对不均衡训练集预处理后获得支持向量机模型，通过对印刷图像套准状态进行识别，获得的少类样本识别率 a^+为 0.9375，识别准确率几何平均数 G_{mean}为 0.9437，F 测度为 0.9574。**结论** 文中提出方法获得的印刷套不准图像识别准确率 a^+, G_{mean}和 F 测度均优于实验中的其他方法。

关键词：不均衡数据；印刷套准；集成采样；支持向量机
中图分类号：TP391.41 **文献标识码：**A **文章编号：**1001-3563(2018)11-0158-07
DOI：10.19554/j.cnki.1001-3563.2018.11.028

Registration Recognition Methods of the Printing Images Oriented to the Imbalanced Training Set

JIAN Chuan-xia, *GAO Jian*
(Guangdong University of Technology, Guangzhou 510006, China)

ABSTRACT: The work aims to study a preprocessing method for the imbalanced printing image training set in view of the problem of poor recognition accuracy of unaligned printing images in the detection of imbalanced printing image registration. An integrated sampling preprocessing method for the imbalanced printing image training set was proposed. Firstly, the imbalanced training set was divided into support vectors and non-support vectors through the support vector machine model. Secondly, in order to balance the training set, the support vectors in the minority class (i.e., the unaligned printing images) were oversampled and the non-support vectors in the majority class (i.e., the aligned printing images) were undersampled. Finally, the pre-processed balanced training set was used to train the support vector machine model, and the model parameters were optimized. The proposed integrated sampling method was used to pre-process the imbalanced training set to obtain the support vector machine model. The recognition rate of the minority class a^+ obtained through the recognition of printing image registration was 0.9375, the geometric mean G_{mean} of the recognition accuracy was 0.9437 and the F-score was 0.9574. The proposed method outperforms other methods in the experiment in terms of the recognition accuracy a^+ of the unaligned printing images, G_{mean} and F-score.

KEY WORDS: imbalanced data; printing registration; integrated sampling; support vector machine

印刷套准是实现印刷色彩复制和阶调复制的基础。当印刷的四色出现套准误差时，复制品就难以还原原稿的色彩和阶调。目前四色套准状态检测还依赖人工抽检，效率低，可靠性差，且不能实现全检。目前基于机器视觉的方法实现印刷套准自动检测已成为研究的热点[1—6]。简川霞等分别提取了印刷标志图

收稿日期：2018-01-28
基金项目：国家自然科学基金（51675106）；广东省自然科学基金（2015A030312008, 2016A030308016）；广东省科技计划（2015B010104008）；广东工业大学青年基金重点项目（17QNZD001）
作者简介：简川霞（1979—），男，博士，广东工业大学讲师，主要研究方向为机器视觉与图像处理。

像灰度共生矩阵特征和印刷标志图像的纹理特征，采用 Adaboost 分类器和支持向量机分类器进行印刷套准状态识别[7—8]。实验证明，2 种方法都取得了较好的印刷套准识别率。于丽杰等[9]提出一种接近人的视觉模型的颜色特征提取方法，提取了图像的灰度共生矩阵的纹理特征参数，并设计了不同距离测度进行分类。实验证明，基于纹理特征参数的分类效果优于颜色特征的分类。尽管这些印刷图像套准状态识别方法都取得了较好的识别效果，但这些方法有一个假设：用于分类器训练的套准图像和套不准图像的数量大体上是相当的，即训练集样本是均衡的，但在实际的印刷生产线上，套准图像的数量要远远多于套不准图像的数量，这样获得的用于分类器模型建立的印刷图像训练集往往是不均衡的。建立在不均衡印刷图像训练集上的分类器模型在对印刷图像测试集样本分类检测时（即分类为 2 种状态：印刷套准和印刷套不准），对少数类别的印刷套不准图像的识别准确率低[10]。在印刷套准状态检测时，往往更加关心套不准图像，即要从大量待检测图像中识别出套不准图像，因此提高少数类别的印刷套不准图像的识别准确率是非常重要的。在文中，少数（此处指样本数量少）类别的印刷套不准图像的类别标签设为+1，称为正类样本数据，多数（此处指样本数量多）类别的印刷套准图像的类别标签设为−1，称为负类样本数据。

印刷图像是采用四色套印实现的，任意两色之间都需套准。文中是采用图像识别的方法对任意两色之间的套准状态进行检测。针对由不均衡的印刷图像训练集导致的图像识别分类模型对印刷套不准图像的识别准确率低的问题，提出一种集成样本采样方法，获取均衡的印刷图像训练集样本，提高印刷套不准图像的识别准确率。

1　印刷图像套准状态检测流程

文中仅研究两色之间的印刷标志是否套准，不涉及各印版的颜色，为便于研究，用于套准的印刷标志图像都转化为灰度图像。从印刷图像上提取印刷标志，并转化为灰度图像，见图 1，在后文中均采用此印刷标志。印刷图像套准状态检测流程见图 2。从印刷标志图像中提取图像纹理特征数据，作为图像识别分类器模型的输入数据。在印刷工业生产中，印刷套准的图像数量多，属于多类数据，印刷套不准的图像数量少，属于少类数据。多类数据和少类数据组合在一起，构成不均衡的印刷图像训练集。然后采用集成采样的方法对训练集样本进行采样，即过采集少类数据样本，使得少类数据样本的数量增加，得到新的印刷套不准图像；同时欠采集多类数据样本，使得多类数据样本数量减少，得到新的印刷套准图像；最终使得新的印刷套准图像和新的印刷套不准图像的数量大致相等，即构成均衡的印刷图像训练集。在此均衡的训练集上进行学习以获得印刷图像套准识别分类器模型，并优化模型参数；最后用此模型对印刷图像套准状态进行识别。

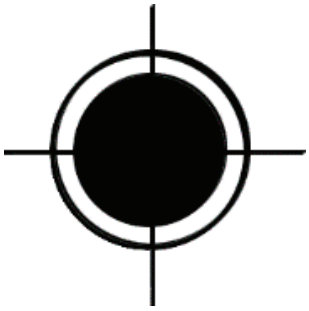

图 1　印刷标志
Fig.1 Printing mark

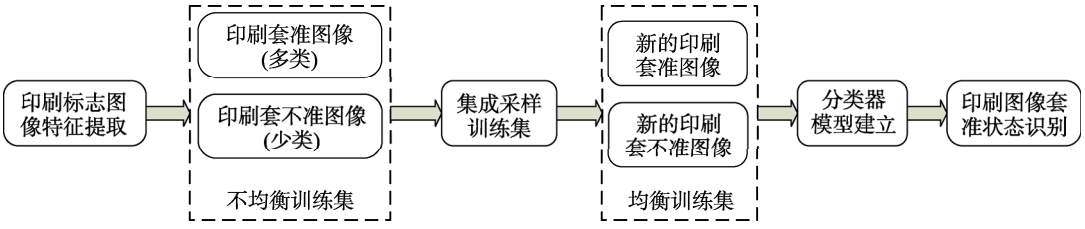

图 2　印刷图像套准状态检测流程
Fig.2 Printing image registration detection flowchart

2　支持向量机分类模型

在图像识别分类器模型中，支持向量机[11]（SVM）是基于统计学习理论提出一种监督式学习的方法，即通过寻求最小结构化风险来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。近年来 SVM 在模式识别领域越来越被大家普遍认可，已成为机器学习和数据挖掘领域的最有力工具[12—13]，因此在文中，采用 SVM 分类器来实现印刷图像套准状态的识别检测。下面重点介绍一下 SVM 的分类模型。

H 为 2 类之间的分类超平面，H_1 和 H_2 这 2 类中距离 H 最近的样本且平行于 H 的超平面，线性 SVM 的思想是用最大化 H_1 和 H_2 的“间隔”，即寻求最优分类超平面 H，将 2 类样本正确分开，见图 3。假定训

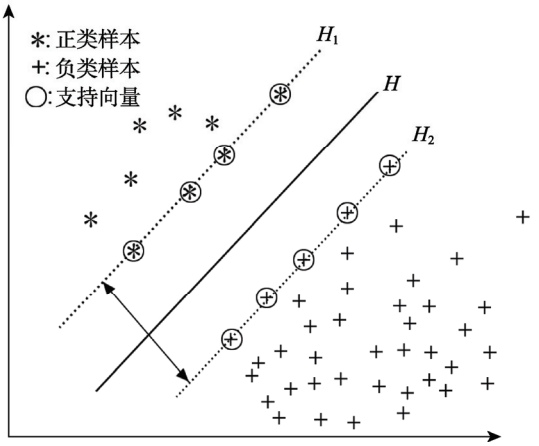

图 3 线性支持向量机的 2 类分类

Fig.3 Linear SVM classification of two-class data

练集：$T=\{(x_1, y_1), (x_2, y_2)\cdots(x_i, y_i)\}$，$x_i \in R^n$ 和 $y_i \in \{1, -1\}$，$i=1, 2\cdots N$。寻求最优分类超平面，可转化为求解下列对应变量 ω 和 b 的最优化问题，即二次规划问题：

$$\min \ \frac{1}{2}\|\omega\|^2$$

$$\text{s.t. } y_i(\omega x_i + b) \geqslant 1, \ i = 1, 2, \cdots N \tag{1}$$

式中：ω 为分类超平面 H 的法矢量；b 为截距或偏置。在实际应用中，数据常常是线性不可分的，通过引入松弛变量 ξ_i 和核函数 $K(x, x')$，非线性支持向量机的数学模型可写为：

$$\min \ \frac{1}{2}\|\omega\|^2 + C\sum_{i \in I}\xi_i^k$$

$$\text{s.t. } y_i(\omega^{\mathrm{T}}\phi(x_i) + b) - 1 + \xi_i \geqslant 0, \ \xi_i \geqslant 0 \ \forall i \tag{2}$$

式中：ξ_i 为松弛变量（或松弛因子）；C 为正则化参数，又叫惩罚参数，是一个大于 0 的常数，控制着对错分样本惩罚的程度；$\phi(x)$ 为映射函数，可以通过数据映射的方式，将空间 R^n（低维空间）中线性不可分的数据通过映射变换 $\phi: x = \phi(x)$ 映射到 Hilber 空间（高维空间）中，使得数据在 Hilber 空间中线性可分。变换 C 在算法中的作用是通过核函数 $K = (\phi(x_i) \cdot \phi(x))$ 来实现的。研究表明，当缺少过程的先验知识时，选择高斯核函数比选择其他核函数分类效果要好[14]。

使用 Lagrange 乘子 α_i（$\alpha_i > 0$），非线性支持向量机原始问题的对偶问题，即凸二次规划问题为：

$$\max_{\alpha} \ Q(\alpha) = J(\omega, b, \alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{j=1}^{N}\sum_{i=1}^{N}\alpha_i\alpha_j y_i y_j x_i^{\mathrm{T}} x_j$$

$$\text{s.t. } \sum_{i=1}^{N}\alpha_i y_i = 0, \ \alpha \geqslant 0 \tag{3}$$

由式（3）可求得解 $\alpha^* = (\alpha_1^*, \alpha_2^* \cdots \alpha_N^*)$，则

$$b^* = y_i - \sum_{i=1}^{N}\alpha_i^* y_i K(x_i, x_j) \tag{4}$$

最终的决策函数为：

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{N}\alpha_i^* y_i K(x, x_i) + b^*\right) \tag{5}$$

如果 α^* 的分量 α_i^* 非 0，则训练点 (x_i, y_i) 的输入 x_i 为支持向量（SV）；否则称 x_i 为非支持向量（NSV）。

3 新的集成采样方法

采样方法是对不均衡的训练集数据进行预处理，达到正类数据和负类数据在数量上大致相等，即使训练集数据均衡[15—16]。常用的采样方法有过采样、欠采样和集成采样（或混合采样）[17]。过采样是对少类数据进行采集，增加少类数据的数量，但该方法易出现数据过拟合现象或不重要信息增加的问题[18—19]。欠采样是对多类数据进行采样，减少多类数据的数量，但该方法易丢失对分类重要的样本信息[20—21]。集成采样方法既对少类数据进行过采样，也对多类数据进行欠采样，这可以在一定程度上减少欠采样和过采样带来的问题，但没有从根本上解决 2 种采样方法存在的问题[22]。欠采样和过采样存在问题的根本原因是没有考虑样本数据对确定样本分类超平面的贡献差异。在支持向量机分类中，样本数据中的支持向量（SV）对确定分类超平面的贡献大，非支持向量（NSV）贡献小，因此在采样中，要尽量保留或增加支持向量，保留或减少非支持向量。基于此考虑，文中提出一种新的集成采样方法处理不均衡的印刷图像训练集数据，以获得相对均衡的训练集，用于支持向量机模型训练。新的集成采样方法的流程见图 4。

新的集成采样方法的步骤如下所述。

1）先使用 SVM 对不均衡的印刷图像训练集进行学习，识别不均衡训练集中支持向量样本和非支持向

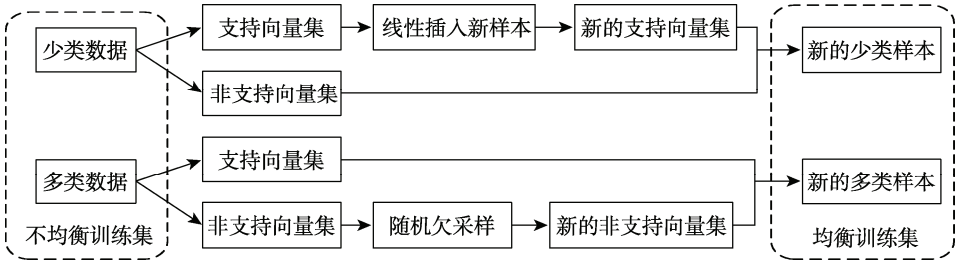

图 4 新的集成采样方法流程

Fig.4 Flowchart of a new integrated sampling method

量样本。具体如下：将训练集 $T=\{(x_1, y_1), (x_2, y_2)\ldots(x_i, y_i)\}$带入式（3）中，求解此凸二次规划问题，得解 $\alpha^* = (\alpha_1^*, \alpha_2^* \cdots \alpha_N^*)$。如果 $\alpha_i^* > 0$，则对应 α_i^* 的 x_i 是支持向量，训练样本（x_i, y_i）是支持向量样本。否则，对应 α_i^* 的 x_i 是非支持向量，训练样本（x_i, y_i）是非支持向量样本。

2）对少类数据中的支持向量进行过采样，以增加支持向量的数量。过采样的方法如下：在 2 个少类数据支持向量之间线性插入数据，以生成新的支持向量，见图 5。新的支持向量 x_i' 为：

$$x_i' = x_i + rand(0,1) \times (x_i - x_0) \tag{6}$$

式中：x_0 和 x_i 为少类样本中的支持向量；$rand(0,1)$ 为[0, 1]之间的一个随机数。

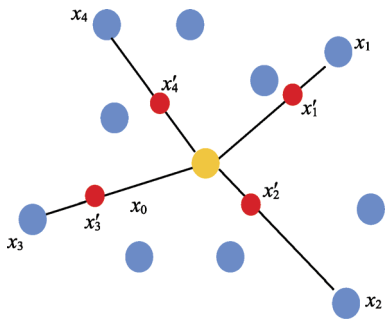

图 5　线性插入支持向量数据示意

Fig.5 Sketch map of inserting linearly support vector data

新的少类数据由新的支持向量（原有少类中的支持向量和新生成的少类支持向量组成）和原有少类非支持向量组成，见图 6。

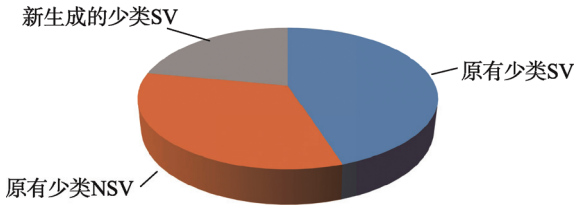

图 6　新的少类样本组成

Fig.6 Composition of new minority samples

3）对多类数据中的非支持向量进行欠采样，以减少非支持向量的数量，见图 7。新的多类样本由原有多类支持向量和新生成的多类非支持向量组成，见图 8。

4）新的少类样本和新的多类样本数量基本相等，构成均衡的训练集，用于支持向量机模型训练。

4　实验

4.1　不均衡印刷图像套准识别评价指标

常采用总体分类准确率 a 来评价均衡的数据分类性能，但不适合评价不均衡的数据分类性能。如：

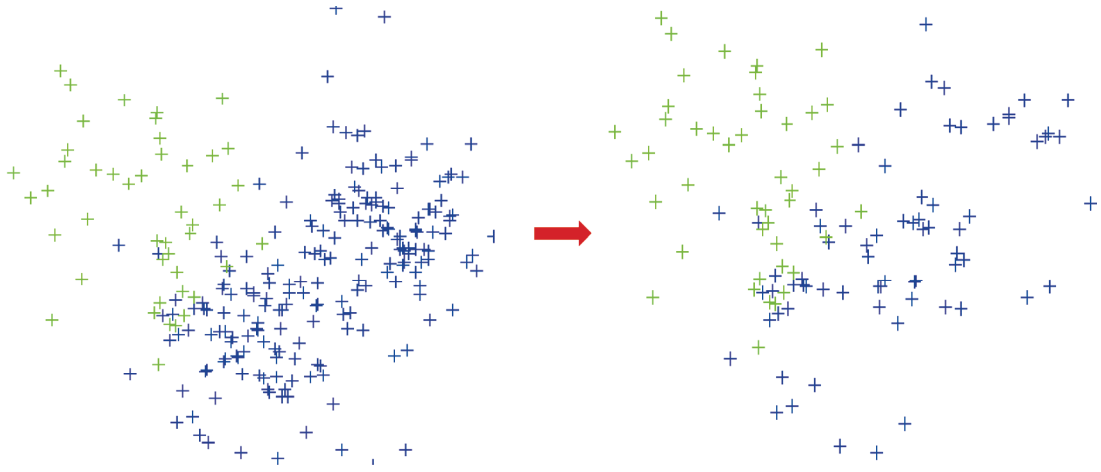

图 7　多类数据非支持向量欠采样

Fig.7 Undersampling of non-support vectors in the majority data

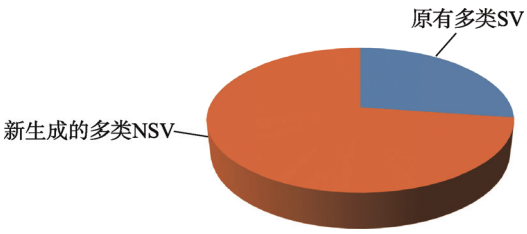

图 8　新的多类样本组成

Fig.8 Composition of new majority samples

假定不均衡的印刷图像数据集中包含 98 个印刷图像套准数据和 2 个印刷图像套不准数据，即使 2 个印刷图像套不准数据被错误分类，此不均衡数据分类仍可取得 98%的总的分类准确率，但印刷图像套不准数据的分类准确率却为 0。事实上，在不均衡印刷图像套准识别中，更关心少数类的印刷图像套不准的情况，因此，为了评价不均衡数据的分类性能，文中引入混合矩阵（见表 1）加以说明。

表 1 混合矩阵

Tab.1 Confusion matrix

实际样本	分类	
	正类样本	负类样本
正类样本	正类样本被正确识别的数量（T_p）	被错误识别为负类样本的正类样本数量（F_N）
负类样本	被错误识别为正类样本的负类样本数量（F_p）	负类样本被正确识别的数量（T_N）

文中引入分类准确率几何平均数 G_{mean} 和 F 测度（即 F-score）来评价不均衡数据的分类性能。

1）分类准确率几何平均数 G_{mean}。

G_{mean} 定义为：

$$G_{mean}=\sqrt{a^{+}\times a^{-}} \tag{7}$$

其中，正类准确率（即少类准确率）$a^{+}=T_p/(T_p+F_N)$，负类准确率（即多类准确率）$a^{-}=T_N/(T_N+F_p)$。G_{mean} 考虑了分类器对正类样本和负类样本的分类能力，在正类准确率和负类准确率之间做了一个几何权衡。G_{mean} 较大时，意味着分类器对不均衡样本的分类能力较强。

2）F 测度（F_{score}）

$$F_{score}=\frac{(1+\beta)^2\times p\times r}{\beta(p+r)} \tag{8}$$

式中：F_{score} 为 p 和 r 加权调和平均；$p=T_p/(T_p+F_p)$，代表真正的正类样本在所有预测为正类样本中的比例；r 为正类样本的召回率，$r=T_p/(T_p+F_N)$，代表正类样本被正确分类的比例，在数值上等于 a^{+}；p 为分类器对某一个类别准确率的评价。在不均衡样本数据分类时，希望不降低 p 的前提下，提高正类样本的召回率 r，但有时不降低 p 和提高 r 是冲突的。F_{score} 权衡了 p 和 r 这 2 个因素，反映了不均衡数据分类的性能。在式（8）中，通常 $\beta=1$。

4.2 数据

实验从印刷图像上提取 200 个标志图像（其中 100 个是印刷套不准图像，100 个是印刷套准图像）。计算这 200 个标志图像的灰度共生矩阵，采用文献[7]的方法提取每幅图像灰度共生矩阵的能量（E）、熵（H）、惯性矩（I）、相关（C）等 4 个纹理参数，并求 E, H, I 和 C 的均值和标准差，作为 8 维标志图像的纹理特征（$T(1)$, $T(2)$, $T(3)$, $T(4)$, $T(5)$, $T(6)$, $T(7)$ 和 $T(8)$），以每幅标志图像的 8 维纹理特征作为支持向量机分类器的输入参数，支持向量机根据每幅标志图像的输入参数来判断标志图像是处于套准状态或套不准状态（支持向量机输出标签为+1 为套不准状态，输出标签为−1 为套准状态）。部分印刷标志图像的纹理特征见表 2。为了验证文中提出的新的集成采样方法的有效性，实验采用了一个不均衡比率较大

表 2 部分印刷标志图像的纹理特征

Tab.2 Texture features of some printing mark images

序号	$T(1)$	$T(2)$	$T(3)$	$T(4)$	$T(5)$	$T(6)$	$T(7)$	$T(8)$	标签
1	0.5225	0.0076	1.0983	0.0383	5.0970	1.7544	0.0209	4.5538×10^{-4}	+1
2	0.5119	0.0083	1.1592	0.0402	5.4313	1.8412	0.0207	4.7135×10^{-4}	+1
3	0.5137	0.0082	1.1311	0.0406	5.3055	1.8238	0.0207	4.6388×10^{-4}	+1
4	0.5019	0.0082	1.1654	0.0409	5.4766	1.8621	0.0203	4.5704×10^{-4}	+1
5	0.5093	0.0081	1.1640	0.0395	5.3524	1.7969	0.0206	4.5521×10^{-4}	+1
6	0.5255	0.0080	1.1242	0.0391	5.1308	1.7718	0.0212	4.7379×10^{-4}	+1
7	0.5211	0.0081	1.1370	0.0397	5.2426	1.8006	0.0211	4.7562×10^{-4}	+1
8	0.5177	0.0082	1.1457	0.0399	5.3185	1.8183	0.0210	4.7500×10^{-4}	+1
9	0.5146	0.0082	1.1531	0.0401	5.3807	1.8313	0.0208	4.7345×10^{-4}	+1
…	…	…	…	…	…	…	…	…	…
101	0.5552	0.0058	1.0270	0.0292	3.5124	1.1212	0.0224	3.3652×10^{-4}	−1
102	0.5477	0.0066	1.0477	0.0327	4.0662	1.3571	0.0220	3.8994×10^{-4}	−1
103	0.5552	0.0058	1.0259	0.0296	3.5429	1.1459	0.0224	3.4289×10^{-4}	−1
104	0.5553	0.0066	1.0282	0.0333	4.1766	1.4105	0.0223	4.1708×10^{-4}	−1
105	0.5370	0.0069	1.0911	0.0349	4.5069	1.5364	0.0216	4.2560×10^{-4}	−1
106	0.5479	0.0068	1.0446	0.0335	4.2393	1.4231	0.0220	4.0778×10^{-4}	−1
107	0.5407	0.0068	1.0714	0.0339	4.1523	1.3826	0.0218	3.8800×10^{-4}	−1
108	0.5408	0.0073	1.0718	0.0341	4.6119	1.5096	0.0217	4.2350×10^{-4}	−1
109	0.5469	0.0067	1.0537	0.0331	4.0675	1.3664	0.0220	3.9227×10^{-4}	−1
…	…	…	…	…	…	…	…	…	…
200	0.5653	0.0057	1.0023	0.0299	3.5570	1.1368	0.0229	3.5539×10^{-4}	−1

的训练集（不均衡比率为多类样本数量与少类样本数量的比例，文中采用的不均衡比率为 60∶4），即从 200 个标志图像中随机取出 60 个套准图像和 4 个套不准图像，组成不均衡的标志图像训练集，其余的套准图像（100−60=40 个）和套不准图像（100−4=96 个）组成测试集，以验证新的集成采样方法的有效性。

4.3　结果

为了验证新的集成采样方法的优越性，文中分别选取了欠采样方法和过采样方法中较为典型的 2 种方法作为比较。实验采用的方法如下所述。

1）没有对训练集进行任何采样，直接使用训练集训练 SVM 模型，然后用获得的 SVM 分类器[23]对测试集数据进行分类，标记为 SVM。

2）采用单边欠采样进行数据预处理[24]，使用预处理后的数据训练 SVM 模型，然后用获得的 SVM 分类器对测试集数据进行分类，标记为 US。

3）采用 SMOTE 过采样方法对数据进行预处理[25]，使用预处理后的数据训练 SVM 模型，然后用获得的 SVM 分类器对测试集数据进行分类，标记为 SMOTE。

4）采用文中提出的新的集成采样方法对不均衡训练集进行采样，使用预处理后的数据训练 SVM 模型，然后用获得的 SVM 分类器对测试集数据进行分类，即文中方法。

实验采用 C-SVM 作为支持向量机分类器，采用高斯径向基核函数（RBF）。分类模型中惩罚函数 C 和核宽度 σ 采用网格优化算法和五折交叉验证法获得[26]。由于训练集中的套准图像和套不准图像是随机从 200 个标志图像中选取的，为了增加测试的准确性，文中随机选取了 10 组训练集和测试集，分别采用上述的 4 种方法进行实验，取得的测试结果取平均值，见表 3。

表 3　不同方法在不均衡训练集上获得的测试结果
Tab.3 Testing results of addressing the imbalanced training sets with different methods

方法	SVM	US	SMOTE	文中方法
a^+	0.7813	0.8542	0.8750	0.9375
a^-	1.0000	1.0000	1.0000	0.9500
a	0.8456	0.8971	0.9118	0.9412
p	1.0000	1.0000	1.0000	0.9783
r	0.7813	0.8542	0.8750	0.9375
G_{mean}	0.8839	0.9242	0.9354	0.9437
F_{score}	0.8772	0.9213	0.9333	0.9574

4.4　分析与讨论

从表 3 中可以看出，文中方法对少类样本（即印刷套不准图像）的分类准确率 a^+为 0.9375，比 SVM 方法、US 方法和 SMOTE 方法分别提高了 15.6%（0.9375−0.7813=0.1560）、8.3%（0.9375−0.8542=0.0830）和 6.2%（0.9375−0.8750=0.0620）。文中方法对多类样本（即印刷套准图像）分类准确率为 0.9500，虽然比实验中的其它方法略微降低了一点（1−0.9500=0.05），但在分类准确率几何平均数 G_{mean} 上是最大的，即 G_{mean}=0.9437。与其他方法相比，文中方法虽然略微牺牲了一些多类样本的分类准确率，但大幅度提高了少类样本的分类准确率（最大提高了 15.6%）。从表 3 可以看到，文中方法取得的 p 和 r 分别为 0.9783 和 0.9375。文中方法比其他方法略微降低了 p（1−0.9783=0.0220），却较大幅度地提高了 r（最大提高了 15.6%）。权衡 p 和 r 这 2 个因素，文中方法取得了最大的 F 测度（为 0.9574），大于实验中其它方法的 F 测度。SVM 方法对不均衡训练集数据没有做任何处理，因此获取的分类模型对不均衡数据的分类性能最差。US 方法欠采样了多类样本，但没有考虑多类样本中不同样本对确定分类超平面的重要性，在欠采样的过程中，容易丢失一些重要的样本。SMOTE 方法过采样少类样本，尽管该方法避免了重复采样导致的数据过拟合问题，但在不重要的少类样本之间进行过采样，对提高不均衡数据分类性能的作用不大。文中方法考虑了不同样本在确定分类超平面时重要性不同，先将样本数据分为重要的支持向量样本和不重要的非支持向量样本，针对重要的支持向量样本和不重要的非支持向量样本采用不同的采样策略（即欠采样不重要的多类非支持向量样本，过采样重要的少类支持向量样本），这样避免了 US 方法造成的重要样本丢失问题和 SMOTE 方法造成的不重要样本增加的问题，因此文中方法获取的 G_{mean} 和 F 测度要优于实验中的 SVM, US 和 SMOTE 方法。

5　结语

针对不均衡的印刷图像套准状态识别中存在印刷套不准图像识别准确率低的问题，文中提出的集成采样方法对不均衡的印刷图像训练集进行预处理，实现了训练集的均衡化，由此建立的支持向量机模型在测试集上获得的少类样本（即印刷套不准图像）识别准确率为 0.9375，识别准确率几何平均数 G_{mean} 为 0.9437，F 测度为 0.9547，要优于实验中 SVM 方法、US 方法和 SMOTE 方法。

参考文献：

[1] 张锡福. 基于机器视觉的套印对准技术研究[D]. 济南: 山东师范大学, 2015.
ZHANG Xi-fu. Overprint Alignment Technology Research Based on Machine Vision[D]. Jinan: Shandong

Normal University, 2015.
[2] 许旭萍, 于跃飞, 双文杰. 彩色印刷套准误差的自动检测研究[J]. 包装工程, 2013, 34(9): 107—110.
XU Xu-ping, YU Yue-fei, SHUANG Wen-jie. Research of Automatic Registration Error Detection in Color Printing[J]. Packaging Engineering, 2013, 34(9): 107—110.
[3] 张秀珍. 彩色印刷品缺陷检测方法研究[D]. 洛阳: 河南科技大学, 2015.
ZHANG Xiu-zhen. The Method Research of Detection of Color Printing Defect[D]. Luoyang: Henan University of Science and Technology, 2015.
[4] 王文举, 赵萍, 陈伟, 等. 彩色印刷品缺陷快速精确检测方法研究[J]. 包装工程, 2015, 36(17): 112—118.
WANG Wen-ju, ZHAO Ping, CHEN Wei, et al. A Fast and Accurate Method of Defect Detection of Colour Printing Image[J]. Packaging Engineering, 2015, 36(17): 112—118.
[5] 刘建平, 朱方文, 袁振鹏. 基于机器视觉的 IC 卡印刷缺陷检测[J]. 计量与测试技术, 2016, 43(11): 33—36.
LIU Jian-ping, ZHU Fang-wen, YUAN Zhen-peng. Detection of IC Cards' Printing Defect Based on Machine Vision[J]. Metrology & Measurement Technique, 2016, 43(11): 33—36.
[6] 李靓. 基于机器视觉的印刷品缺陷快速在线检测方法研究[D]. 天津: 天津科技大学, 2015.
LI Liang. The Research of High Speed Inspection Method of Printed Matter Based on Machine Vision[D]. Tianjin: Tianjin University of Science &Technology, 2015.
[7] 简川霞, 高健, 李克天, 等. 印刷套准识别方法研究[J]. 包装工程, 2015, 36(7): 129—133.
JIAN Chuan-xia, GAO Jian, LI Ke-tian, et al. Printing Registration Recognition[J]. Packaging Engineering, 2015, 36(7): 129—133.
[8] 简川霞, 高健, 李克天, 等. 基于机器视觉的印刷套准识别方法研究[J]. 电视技术, 2015, 39(16): 69—72.
JIAN Chuan-xia, GAO Jian, LI Ke-tian, et al. A Study on Printing Registration Recognition Based on Machine Vision[J]. Video Engineering, 2015, 39(7): 129—133.
[9] 于丽杰, 李德胜. 彩色印刷套准识别方法研究[J]. 计算机工程与应用, 2011, 47(5): 163—165.
YU Li-jie, LI De-sheng. Study on Identifying Register State of Color Printing[J]. Computer Engineering and Applications, 2011, 47(5): 163—165.
[10] HWANG J P, SEONGKEUN P, EUNTAI K. A New Weighted Approach to Imbalanced Data Classification Problem via Support Vector Machine with Quadratic Cost Function[J]. Expert Systems with Applications, 2011, 38(7): 8580—8585.
[11] CORTES C, VAPNIK V. Support-vector Networks[J]. Machine Learning, 1995, 20(3): 273—297.
[12] HALFAWY M R, HENGMEECHAI J. Automated Defect Detection in Sewer Closed Circuit Television Images Using Histograms of Oriented Gradients and Support Vector Machine[J]. Automation in Construction, 2014, 38(5): 1—13.
[13] KUMAR A, KUMAR R. Time-Frequency Analysis and Support Vector Machine in Automatic Detection of Defect from Vibration Signal of Centrifugal Pump[J]. Measurement, 2017, 108: 119—133.
[14] SCHOLKOPF B, SMOLA A J. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond[M]. Cambridge: MIT Press, 2001.
[15] BATISTA G E A P A, PRATI R C, MONARD M C. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data [J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 20—29.
[16] GALAR M, FERNANDEZ A, BARRENECHEA E, et al. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches[J]. IEEE Transactions on Systems Man & Cybernetics Part C: Applications & Reviews, 2012, 42(4): 463—484.
[17] ALIBEIGI M, HASHEMI S, HAMZEH A. DBFS: An Effective Density Based Feature Selection Scheme for Small Sample Size and High Dimensional Imbalanced Data Sets[J]. Data & Knowledge Engineering, 2012, 81/82(4): 67—103.
[18] CHAWLA N V, JAPKOWICZ N, KOTCZ A. Editorial: Special Issue on Learning from Imbalanced Data Sets[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 1—6.
[19] ESTABROOKS A, JO T, JAPKOWICZ N. A Multiple Resampling Method for Learning from Imbalanced Data Sets[J]. Computational Intelligence, 2004, 20(1): 18—36.
[20] LIN Z, HAO Z, YANG X, et al. Several SVM Ensemble Methods Integrated with Under-sampling for Imbalanced Data Learning[C]// International Conference on Advanced Data Mining and Applications, Beijing China, 2009: 536—544.
[21] CHAWLA N V. Data Mining for Imbalanced Datasets: An Overview[M]. Data Mining and Knowledge Discovery Handbook, Germany: Springer, 2005.
[22] JIAN C X, GAO J, AO Y H. A New Sampling Method for Classifying Imbalanced Data Based on Support Vector Machine Ensemble[J]. Neurocomputing, 2016, 193(12): 115—122.
[23] VAPNIK V. The Nature of Statistical Learning Theory[M]. New York: Springer, 1995.
[24] KUBAT M, MATWIN S. Addressing the Curse of Imbalanced Training Sets: One-sided Selection[C]// The 14th International Conference on Machine Learning, Morgan Kaufmann, 1997: 179—186.
[25] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321—357.
[26] 王兴玲, 李占斌. 基于网格搜索的支持向量机核函数参数的确定[J]. 中国海洋大学学报(自然科学版)自然科学版, 2005, 35(5): 859—862.
WANG Xing-ling, LI Zhan-bin. Identifying the Parameters of the Kernel Function in Support Vector Machines Based on the Grid-search Method[J]. Periodical of Ocean University of China (Natural Science Edition), 2005, 35(5): 859—862.