doi: 10.15940/j.cnki.0001-5245.2025.06.011

基于机器学习的云污染图像识别研究*

刘 浏[1,2] 赵 赫[1,2] 孙荣煜[1†]

(1 中国科学院紫金山天文台 南京 210023)
(2 中国科学技术大学天文与空间科学学院 合肥 230026)

摘要 在空间目标与碎片光学巡天中, 云是影响观测效能的重要因素之一. 云层的遮挡会降低目标可见度, 给目标图像检测和后续准确的位置、亮度提取带来困难, 从而干扰空间目标与碎片的观测. 高效、准确地鉴别云污染图像, 可以给后续数据处理提供有价值的先验信息, 支撑业务流程常态、稳定地运行. 目前主流的基于图像分割判断图像中云的方法存在耗时、易受噪声影响等缺点. 因此结合图像特征指标评估与人工筛选, 基于空间碎片光学巡天实测数据建立云污染图像的数据集, 并使用支持向量机(Support Vector Machine, SVM)、Shufflenet V2和Resnet 18这3种机器学习方法, 对云污染图像和正常图像开展分类研究. 结果表明, Shufflenet V2在分类任务中的总准确率高于97%, 支持向量机对云污染图像的识别准确率高于98%, 深度学习方法可以有效完成云污染图像的识别, 并且计算速度满足观测数据处理时效的要求. 在未来观测中, 建立的方法可以与云量仪相配合, 协同应用于空间碎片观测计划的优化, 降低天气对观测设备效能的影响, 促进观测台站更加稳定的运行.

关键词 方法: 数据分析, 技术: 图像处理, 天文数据库: 巡天

中图分类号: P111; **文献标识码**: A

1 引言

随着空间目标和碎片数量的不断增加, 光学巡天是满足当前对其监测需求的重要手段[1]. 巡天模式下观测空间目标时, 望远镜会根据观测计划指向特定的天区, 以恒动模式或凝视模式获取连续帧图像, 并每隔一段时间切换所观测的天区. 对于空间目标与碎片的光学巡天, 当望远镜观测某一特定天区时, 如果该天区中有云, 那么云的遮挡会严重影响目标探测[2], 严重时甚至会使采集到的图像完全无法处理, 影响观测计划的实施. 后续在对这些图像中的空间目标进行位置与光度测量时, 精度也受云污染的影响随之下降. 图1为光学巡天中受云污染的典型图像, 图2为光学巡天获取的正常图像.

对于地基天文观测而言, 云量的测量是重要的观测辅助手段, 通过云量测定可以获得观测天区云层的分布, 从而优化观测天区的安排. 全天相机是一种广泛应用且有效的工具, 用于提取全天的云层信息, 通常在可见光或红外波段下工作[3]. 在可见光波段, 常用的地面成像设备是广角镜头或鱼眼相机, 通过获取大范围云层特征, 结合图像处理最终得到云量的信息[4]. 在红外波段, 全天相机通过测量红外辐射强度确定云的亮温, 从而准确推算云的温度和厚度, 这为云层分布的研究提

2025-01-02收到原稿, 2025-03-06收到修改稿

*国家自然科学基金项目(12473079)、国家重点研发计划项目(2023YFF0725300)、宇航动力学国家重点实验室基金项目(2022ADL-J002)资助

†rysun@pmo.ac.cn

供了宝贵的信息[5]. 红外云量仪相比于可见光云量仪精度更高, 然而红外云量仪通常更昂贵, 并且需要频繁的维护. 面向地基光学观测, 我们已经具备了一定的云层检测手段, 但实际应用中或多或少存在一些局限性. 为了确保观测计划的高效实施, 需要对云污染图像的识别方法开展进一步研究, 以便在观测过程中及时识别出云量较多的天区. 同时, 该方法应能够快速处理大量实测图像, 配合云量仪优化观测计划. 目前, 云层检测主要采用经典的图像分割、传统机器学习和深度学习等方法.

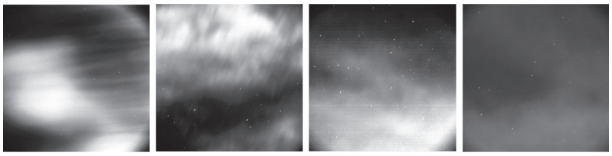

图 1 地基望远镜巡天观测获取的云污染图像. (a)–(d): 不同观测天区获取的云污染图像.

Fig. 1 Cloud-contaminated images acquired from ground-based telescope sky survey observations. (a)–(d): cloud-contaminated images taken at different fields.

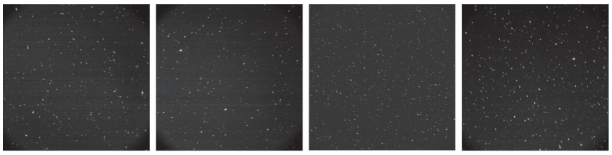

图 2 地基望远镜巡天观测获取的正常图像. (a)–(d): 不同观测天区获取的正常巡天图像.

Fig. 2 Normal images acquired from ground-based telescope sky survey observations. (a)–(d): normal survey images taken at different fields.

围绕图像分割方法的应用, 樊亮等[6]利用红外云量仪获得了全天高空间分辨率云图, 并且按照不同月份选取晴空样本, 通过统计建立晴空阈值, 然后对图像进行二值化, 最终通过连通域形状判断, 实现云层的检测. 秦永等[7]利用卫星图像中云和雪的不同波谱特征, 采用最大类间方差的方法对云层进行阈值化分割, 并据此提取云的形状特征, 成功地将云层和非云层分离. 在传统机器学习方法方面, 王利文等[8–9]将天文图像的纹理特征和灰度不一致度作为指标特征, 通过人工标记获取了包含数百帧图像的训练集, 利用支持向量机(Support Vector Machine, SVM)对这些图像进行二分类, 在测试集上分类准确率超过98%. Li等[10]首先人工筛选出被云层污染的图像, 随后利用SVM和轻量级梯度提升树(Light Gradient-Boosting Machine, LightGBM), 实现了较高精度的云污染和正常图像的分类. 在深度学习方法方面, 车蕾等[11]通过改进的通道加权-特征融合(channel weighting-feature fusion, CWFF)结构的U型网络(u-shaped network, U-Net)模型, 实现了全天相机成像日间云量计算方法, 通过和专家标注相比较, 与传统方法相比像素准确率和平均绝对差实现了提升. Mommert[12]基于人工标记的约2000张图像, 通过对全天相机图像进行分区, 利用深度学习方法实现了85%的准确率.

这些方法在对应场景中取得了较好的应用效果, 但在使用时也受到一定的限制. 例如对红外云量仪而言, 虽然可以准确地获取云层的分布, 但是它的成本较高, 而且需要频繁维护. 通过对图像分割与二值化处理获得云层分布, 虽然有着较高的准确性, 但是易受噪声影响. 通过对图像分区并训练深度学习模型分类云污染图像与正常图像, 在测试集上有着较好的效果, 但需要对每帧图像的区域分类并进行标记, 数据集的获取相当耗费人力. 此外, 通过传统机器学习或深度学习方法分类筛选有云和无云的图像, 虽然可以取得较高的准确率, 然而大多数方法中使用的训练集选取有一定的主观因素, 没有通用的参考标准. 越来越多的学者开始采用深度学习方法来解决云污染图像识别问题, 然而在应用过程中, 由于数据集规模的限制, 通常仅在少量实测图像中进行测试, 或者仅基于特定场景的数据进行测试. 因此, 利用深度学习方法对海量光学巡天数据中的云污染图像识别仍需要开展进一步的研究.

深度学习已在天文学大数据处理中发挥了重要作用, 考虑在空间碎片广域光学巡天可以获取大量实测图像, 其中不乏受到云污染的图像, 因此, 可以基于这些数据开展利用深度学习方法的云污染图像分类. 本文通过图像特征指标计算与人工筛选相结合, 建立精确可靠的训练集, 在此基础上比较了多种机器学习方法的分类性能, 重点分析了处理的时效性与分类的准确性, 研究在实际观

测中的应用价值.

2　数据集获取

2.1　人工筛选

本文所用数据来自于一台专用于空间碎片观测的大视场光学望远镜, 该望远镜和CCD的参数如表1所示. 观测数据采集时相机使用了凝视模式, 这意味着望远镜在某一时间段指向相同方位角与高度角的天区, 并获取连续帧图像. 我们选取了云污染较严重的某个观测夜整晚的所有图像进行试验, 首先我们对这些图像进行人工筛选, 将受到云污染的图像和未受到云污染的图像进行分类, 随后我们计算出它们的图像特征指标以验证人工筛选的准确性.

表 1　望远镜和CCD详细参数

Table 1　Detail information of the telescope and CCD frame

Parameter	Value
Diameter	500 mm
Field of view	$2.2^\circ \times 2.2^\circ$
Frame size	2048×2048
Pixel scale	3.9°
Dynamic range	$0 \sim 65535$
CCD operating mode	Frame transfer

对于地基夜天文观测, 存在站址杂散光的影响, 例如地面光污染、月光等, 云层中的水滴或冰晶会散射这些光源, 尤其当云层较厚时, 散射效果更加显著. 这会使得含云图像中部分区域甚至全部区域的像素点灰度值较高, 据此我们可以对含云图像进行人工初筛. 我们从单个观测夜拍摄的5813帧图像中, 筛选出了2199帧云污染图像和2480帧正常图像供模型训练与验证使用, 剩余的620帧正常图像和514帧云污染图像作为测试集. 考虑到人工筛选依赖于经验, 存在一定的主观因素, 我们使用图像特征指标进一步验证筛选数据集的准确性与有效性.

2.2　图像特征指标

图像中云的像斑具有较强的延展性, 与没有云的图像相比其纹理特征存在一定差异. 为了更精确地筛选出云图, 可以通过计算图像的纹理特征参数进一步验证所分类云污染图像与正常图像的可靠性. 通常来说, 可以通过计算图像的灰度共生矩阵来获取图像的纹理特征[13]. 灰度共生矩阵中的元素是具有某种空间位置关系的两个像素灰度同时出现的概率. 具体来说: 灰度共生矩阵被定义为从灰度为i的像素点出发, 离开某个固定位置(相隔距离为d、夹角为θ)的点上灰度值为j的概率, 即所有估计的值可以表示成一个矩阵的形式, 因此被称为灰度共生矩阵. 灰度共生矩阵$\boldsymbol{g}$中的元素可以表示为:

$$g(i,j,d,\theta)=\frac{\#\{f(x_1,y_1)=i,f(x_2,y_2)=j\|(x_1,y_1),(x_2,y_2)\in M\times N\}}{MN}. \tag{1}$$

$f(x_1,y_1)$和$f(x_2,y_2)$是$\boldsymbol{f}$矩阵中的两个点, 两个点的距离为d, 它们与坐标横轴的夹角为θ. $\boldsymbol{f}$为图像的灰度值矩阵, $\boldsymbol{g}$为$\boldsymbol{f}$的灰度共生矩阵, $\#\{x\}$为集合中x元素的个数. M和N分别为图像的长和宽, 即$\boldsymbol{f}$矩阵的长和宽, $\boldsymbol{g}$矩阵的长和宽与图像灰度值的最大值相同, 本文使用的图像经过变换后最大灰度值为255. 此时可以得到不同距离、不同夹角的灰度共生矩阵. 在我们的研究中, 观测获取的图像满足旋转对称性, 这一特性说明在不同旋转角度上灰度共生矩阵纹理特征差异不大, 因此参数设置为$d=1$、$\theta=0$.

我们选取了4个典型的图像纹理特征指标:

(1)角二阶矩(Angular Second Moment, ASM), 又称能量, 是图像灰度分布均匀程度和纹理粗细的一个度量, 它代表矩阵$\boldsymbol{g}$中每个元素的平方和. 正常图像中, 灰度值分布均匀, 灰度共生矩阵中某些区域值较大, 而其他区域值较小, 此时ASM较大, 图像较为平整. 云污染图像中, 矩阵中大部分的值不会特别大, 此时ASM较小, 图像不平整. 因此ASM可以作为筛选训练集中云污染图像和正常图像的一个参考, ASM的表达式如下:

$$\mathrm{ASM}=\sum_{i=1}^{U}\sum_{j=1}^{V}[g(i,j)]^2\,. \tag{2}$$

其中U和V分别是矩阵$\boldsymbol{g}$的行、列数.

(2)熵(Entropy, ENT)是图像所具有的信息量的度量, 图像随机性越大, 即灰度共生矩阵$\boldsymbol{g}$元素的值分布均匀熵越大, 反之熵越小. 如果图像不含云, 则图像分布均匀, 灰度共生矩阵$\boldsymbol{g}$某些元素有较大的值, 分布不均匀. 它表示了图像纹理的非均匀性和复杂性. 对于相同观测模式采集的图像, 云污染图像的熵大于正常图像. 所以ENT可以作为我们筛选训练集图像是否有云的一个参考. ENT可以用如下公式表示:

$$\mathrm{ENT}=-\sum_{i=1}^{U}\sum_{j=1}^{V}g(i,j)\ln[g(i,j)]\,. \tag{3}$$

(3)对比度(Contrast, CON), 反映了图像某个位置像素值及其邻域像素值的大小对比情况. 纹理沟纹越深, 其对比度越大. 云污染图像的纹理沟纹和正常图像会存在一定的差别. CON具体表示为:

$$\mathrm{CON}=\sum_{i=1}^{U}\sum_{j=1}^{V}(i-j)^2g(i,j)\,. \tag{4}$$

(4)逆差距(Inverse Difference Moment, IDM), 它代表图像纹理局部的变化, 是图像纹理同质性的反映. 连续灰度的图像会有较大的IDM值, 所以云污染图像的纹理局部变化通常和正常图像也有一定的差别. IDM可以表示为:

$$\mathrm{IDM}=\sum_{i=1}^{U}\sum_{j=1}^{V}\frac{g(i,j)}{1+(i-j)^2}\,. \tag{5}$$

我们统计了云污染图像和正常图像4个纹理特征的均值和标准差, 表2是均值和标准差的统计信息. 对于4个纹理特征, 我们绘制了两两特征组合的散点图, 如图3所示. 从6个子图中可以发现, ASM和ENT是两个类别图像差异最大的特征. 对IDM和CON特征来说, 云污染图像和正常巡天图像的差异不显著.

表 2 云污染图像和正常图像纹理特征的统计均值与标准差

Table 2 The mean and standard deviation of feature values for cloud-contaminated images and normal images

Image type	Texture Features			
	ASM	CON	IDM	ENT
Cloud	0.168±0.122	8.136±25.278	0.928±0.055	2.624±1.144
Normal	0.854±0.166	4.302±59.544	0.971±0.025	0.445±0.385

统计结果显示, 云污染图像的ASM主要集中在[0, 0.25]区间, ENT主要集中在[1.8, 7]区间; 而正常图像的ASM主要分布在[0.45, 1]区间, ENT集中在[0, 1.5]区间. 二者的分布差异较大, 与我们的预期一致, 表明筛选结果具有较好的可靠性. 然而, 当ASM位于[0.25, 0.45]区间、ENT位于[1.5, 1.8]区间时, 正常图像和云污染图像呈现交错分布. 此现象归结于当图像中存在薄云时, 其灰度值分布较为平坦, 且图像的随机性较低, 接近正常图像的特征. 对于这部分数据, 依靠一定的观测经验能够通过人工甄别将其区分开来.

对于不同云污染程度的图像, 最显著的差别是图像中被遮挡源的数量不同, 被遮挡源数量越多, 云污染程度越严重. 然而, 不同观测天区中源的个数各不相同, 也会影响检测出源的数量. 熵作为衡量图像信息量的指标, 代表了图像的随机性. 通常来说, 图像中云越厚, 图像随机性越高, 即云污染程度越严重; 反之, 图像随机性越低, 云污染程度越轻微. 图4是不同云污染程度的典型图像, 其中图4 (a)、(d)为污染程度较为严重的图像, 其熵值分别为6.325、6.584, 图像中绝大部分的亮源因云遮挡而不可见. 图4 (e)中云污染集中在图像上方, 图4 (b)、(e)的熵值分别为4.233、3.845, 其云污染程度中等, 图像中部分亮源没有被遮挡.

图4(c)、(f)的熵值分别为2.135、1.986, 图像中只有极少的源不可见, 云污染程度相对较轻.

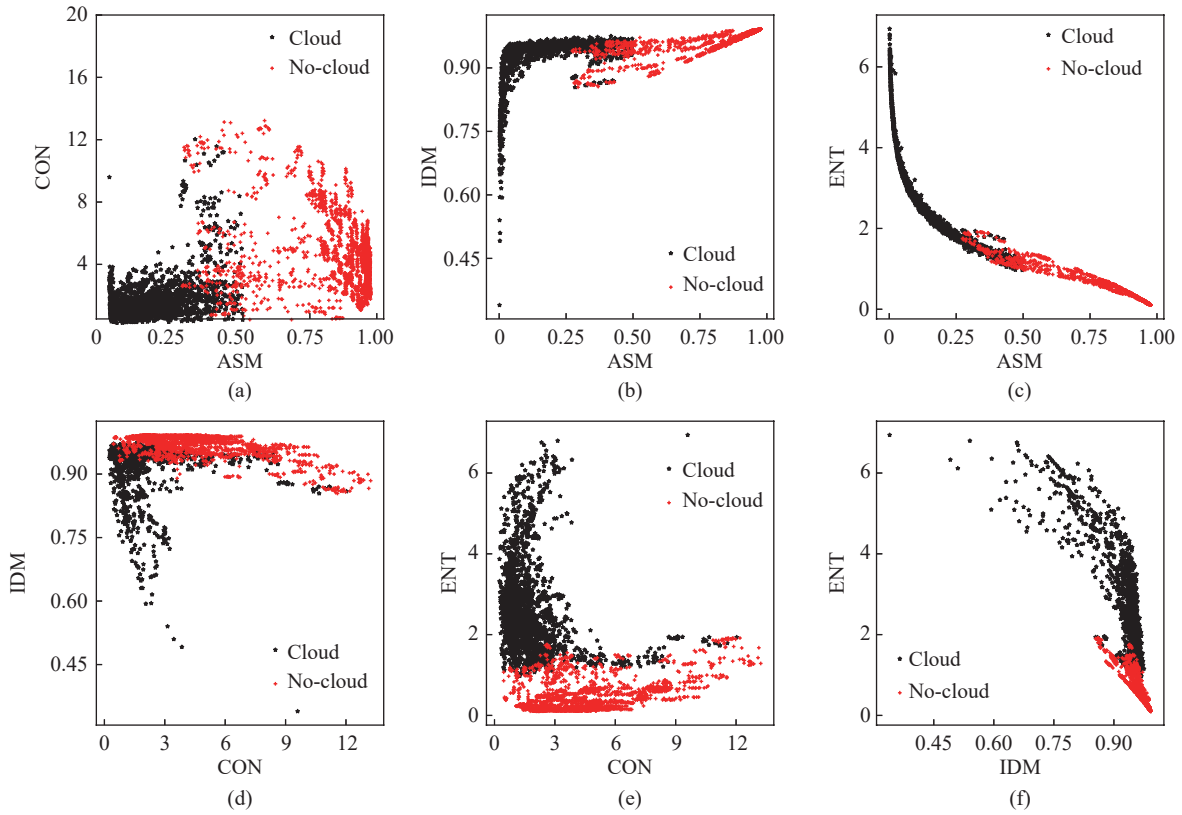

图 3　正常图像和云污染图像的纹理特征比较(其中(a)–(f)分别代表CON和ASM、ASM和IDM、ASM和ENT、CON和IDM、CON和ENT以及IDM和ENT的散点图).

Fig. 3　The comparison of texture features between normal images and cloud-contaminated images (where (a)–(f) represent scatter plots of the combinations of CON and ASM, ASM and IDM, ASM and ENT, CON and IDM, CON and ENT, and IDM and ENT, respectively).

3　分类算法

随着计算机硬件(特别是图形处理单元(Graphic Processing Unit, GPU)和张量处理单元(Tensor Processing Unit, TPU))和专用的深度学习框架(如Pytolch和Tensorflow)的发展, 深度学习模型的训练和推理速度得到了显著提高. 由于其强大的特征提取能力, 深度学习已被应用于大量的天文学项目中. 例如Zhao等[14]利用深度神经网络Yolo V5进行空间目标的检测. Duev等[15]利用Vgg 6、Resnet 50和Densenet 121深度学习模型检测图像中由小行星产生的条纹. He等[16]利用卷积神经网络搜寻引力透镜. 本文将尝试使用多种机器学习方法对云污染图像进行分类研究, 并比较它们的性能.

3.1　支持向量机

SVM是一种传统的监督机器学习模型, 其对噪声较为鲁棒, 特别在处理数据集中的异常点时, 由于只关注支持向量, 因此受异常数据的影响较小. SVM被广泛应用于分类和识别恒星[17]与星系[18]以及预测太阳高能粒子事件[19]等工作中. SVM由Cortes等[20]提出, 最早被用来解决二分类问题. 对于一个特征空间上的数据集:

$$\boldsymbol{T}=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_n,y_n)\},\qquad(6)$$

$\boldsymbol{x}_i \in \boldsymbol{R}^n$, $y_i = \{+1\text{or}-1\}$, n为维度, $\boldsymbol{x}_i$是特征空间上的某个点. y_i代表$\boldsymbol{x}_i$的类别, +1为正类, −1为负类. 我们希望找到一个超平面使这些特征空间上的点按照类别分成两部分, 一部分为正类, 一部分为负类. 图5为SVM算法超平面示意图. 超平面对应方程:

$$\boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{x} + b = 0\,. \tag{7}$$

其中$\boldsymbol{w}$为法向量, b为截距, $\boldsymbol{x}$代表数据向量. 图中, x1、x2代表数据点两个维度的元素值. 并且该超平面需要的是几何间隔距离最大的超平面, 几何间隔的定义为:

$$M = \min_{i=1,2,\cdots,n} M_i\,. \tag{8}$$

其中:

$$M_i = y_i \left(\frac{\boldsymbol{w}^{\mathrm{T}}}{\|\boldsymbol{w}\|} \cdot \boldsymbol{x}_i + \frac{b}{\|\boldsymbol{w}\|} \right). \tag{9}$$

M_i是每个样本点的几何间隔. 在每个样本点都大于M这一约束条件下求M的最大值问题是一个凸优化问题[21], 可以转化为:

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2. \tag{10}$$

$$\text{s.t}\ \ y_i\left(\boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{x}_i + b\right) - 1 \geqslant 0, i = 1, 2, \cdots, n\,. \tag{11}$$

然而该问题用于解决线性可分数据, 这是理想的情形. 现实问题中, 数据往往是线性不可分的, 即在样本中出现噪声或异常点, 此时约束条件需要引入一个松弛变量u_i, 此时对于每个松弛变量u_i, 目标函数相应增加一个代价u_i. 此时问题变为:

$$\min_{\boldsymbol{w},b,\boldsymbol{u}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n} u_i. \tag{12}$$

$$\text{s.t}\ \ y_i\left(\boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{x}_i + b\right) \geqslant 1 - u_i, i = 1, 2, \cdots, n\,, \tag{13}$$

这里C代表惩罚系数, 根据具体问题取值, $\boldsymbol{u}$代表$\{u_1, u_2, \cdots, u_n\}$. 对于非线性分类, 采用上述方法通常不能解决问题. 此时可以采用非线性支持向量机, 其核心为使用核技巧将非线性问题转换为线性问题. 常用的核函数包括高斯核函数(Gaussian Kernel Function)、多项式核函数(Polynomial Kernel Function)和字符串核函数(String Kernel Function)等.

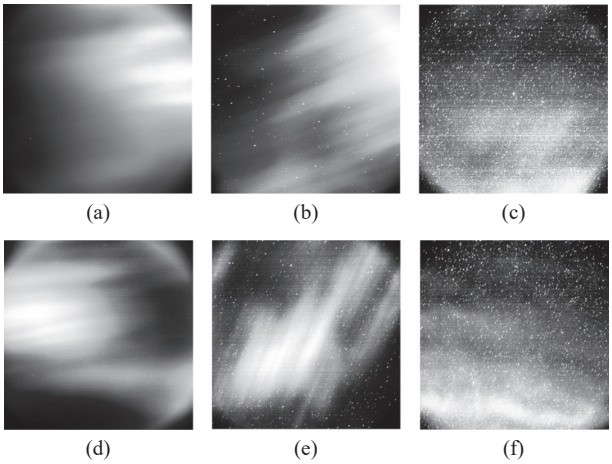

图 4 不同云污染程度的典型图像. (a)与(d)、(b)与(e)、(c)与(f)的云污染程度由重到轻.

Fig. 4 The typical images with different cloud contamination levels. (a) and (d), (b) and (e), (c) and (f) are arranged from heavy to light contamination.

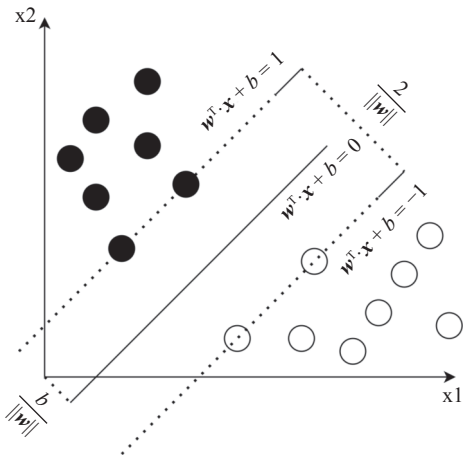

图 5 SVM算法中超平面的示意图

Fig. 5 Illustration of the hyperplane in the SVM algorithm

3.2 Shufflenet V2

随着光学巡天观测中数据量的显著提升, 对天文数据处理实时性的要求变得越来越高. 因此我们希望使用轻量级、高效的深度学习模型识别

云污染图像. 随着深度学习方法不断发展, 其性能显著提升. 然而, 伴随而来的则是模型计算复杂度的不断增加. 很多研究中考虑的是间接计算复杂度指标, 例如浮点运算量(Floating Point Operations, FLOPs). Ma等[22]在Shufflenet V1[23]的基础上, 主要考虑内存访问成本和并行化程度, 通过4项改进提出了一种新的架构Shufflenet V2.

图6展示了Shufflenet V2的模型结构. 作为一种轻量化模型, Shufflenet V2在训练和测试过程中均表现出较快的速度. 该模型在Shufflenet V1的channel shuffle基础上, 引入了channel split, 并将1×1的组卷积替换为普通的1×1卷积. 此外, 模型不仅考虑提高网络的并行度, 还将ReLU和Add等元素级操作考虑在内. channel split操作的核心就是在该模块的开始处, 将特征通道的输入分成两个相等的分支, 并且输入和输出通道数量相同. 这些改进不仅提升了模型的训练和预测速度, 还提高了模型的精度. 具体来说, 原始图像首先经过一个3×3大小、步长为2的卷积层, 使图像尺寸缩小4倍. 随后, 图像通过一个3×3大小、步长为2的最大池化层, 进一步缩小4倍. Convolution2、Convolution3和Convolution4均采用了相似的结构, 每一层都由一个改进后的卷积层进行两倍的下采样, 随后重复该卷积层3、7、3次, 其中被重复的卷积层不进行下采样. Convolution5由一个步长为1的1×1卷积、ReLU激活函数和批量归一化(Batch Normalization)组成. 激活函数引入了非线性, 使模型具有更强的表达能力. Convolution6通过全局池化将特征图尺寸压缩为$1\times1\times1024$, 最后通过全连接层输出一个2维向量. 这两个归一化值分别表示输入图像属于两个类别的概率.

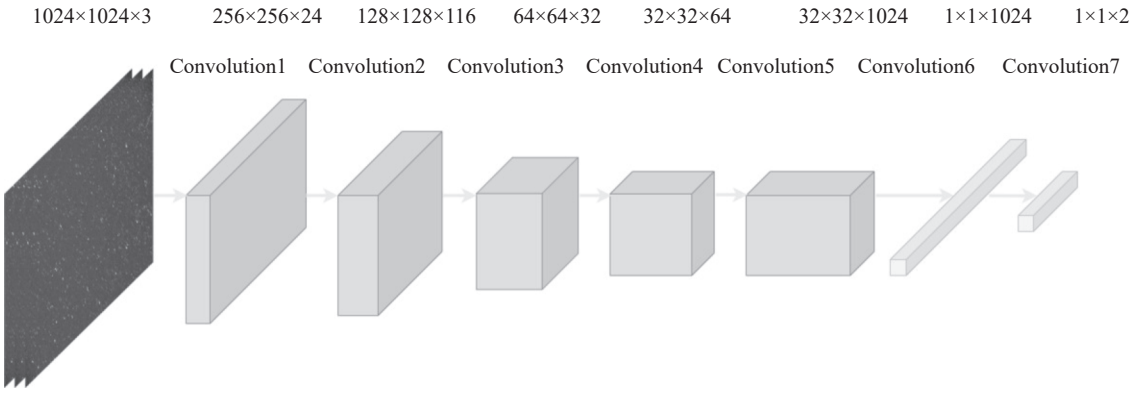

图 6　Shufflenet V2的模型结构

Fig. 6　Shufflenet V2 model architecture

3.3　Resnet 18

随着神经网络层数的增加, 传统卷积神经网络在训练过程中往往会面临退化问题. 具体而言, 随着网络层数的增多, 模型的训练精度不再持续提升, 甚至可能出现过拟合或梯度消失等问题, 从而导致更深的网络无法有效地学习. 这种退化现象在深度神经网络中尤为明显, 不仅影响训练效果, 还降低模型的泛化性能. 在此背景下, He等[24]为了克服深度神经网络训练中的困难, 提出了一种残差学习框架Resnet. Resnet已在天文研究中广泛应用, 例如小行星、恒星和空间目标的分类与检测[25−26]工作. Resnet常见的卷积层深度包括18层、34层和50层等, 根据问题的不同可以选择不同的深度, 对于目标特征较为复杂的问题, 可以选择层数较深的101层和152层; 对于目标特征较为明显的问题, 可以选择较浅的18层和34层. 本文由于云污染图像特征较为简单, 我们选用18层卷积层的Resnet 18.

图7展示了Resnet 18的模型结构, Resnet 18通过多次堆叠残差块来构建深度卷积神经网络模型. 在预处理后的图像输入模型后, 首先经过一个7×7大小、步长为2的卷积层(Convolution1), 接

着是一个3 × 3大小、步长为2的最大池化层(Convolution2), 完成4倍下采样并得到子特征图. Convolution2中还包含两个相同的残差块, 这两个残差块不进行下采样. 接下来的3个卷积层都包含两个部分, 结构相似: 每个部分由两个3 × 3大小的残差块组成, 其中第1个部分进行下采样, 第2个部分不进行下采样. 残差块能够有效地缓解深层卷积神经网络中常见的梯度消失问题. 经过32倍下采样后的特征图通过Convolution6层进行全局平均池化, 最终将特征图缩减至1 × 1 × 512的维度. 随后, 经过一个全连接层, 将512维的输出映射到2维输出, 这两个维度分别表示该图像属于两个类别的归一化概率. 此外, 除了最后两个卷积层外, 每个卷积层结束时均包含一个批量归一化过程, 能够加速模型训练并提升其泛化性能.

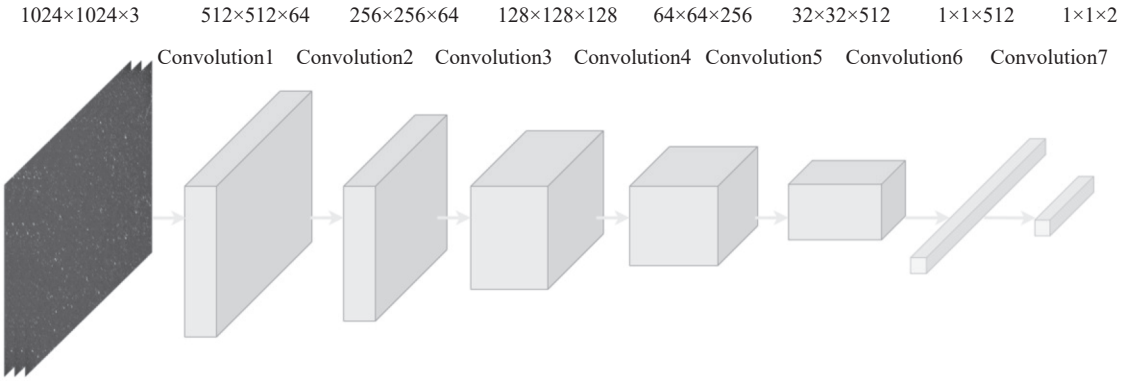

图 7 Resnet 18的模型结构

Fig. 7 Resnet 18 model architecture

4 分类实验结果

本文基于实测数据建立的数据集进行云污染图像分类, 使用Python语言和PyTorch框架实现模型, 并在NVIDIA RTX A2000显卡上进行训练. 训练集包含2199张含云图像和2480张正常图像, 训练集与验证集的比例为3:1. 我们采用了3种不同模型, 并在该观测夜的新数据集上进行了测试. 为了评估模型的性能, 我们计算了3种方法的准确率. 实验结果表明, 3种方法在训练集上的准确率均较高. 在测试阶段, 我们进一步计算了3种方法的混淆矩阵. 通过对比分析, 可以发现, Shufflenet V2在总体准确率上表现较好, 而SVM在云污染图像的识别准确率上相对较高. 此外, 我们使用新观测的数据对3种模型的泛化能力进行了测试, 发现深度学习模型在新数据上的泛化效果要明显好于SVM.

4.1 参数设置与数据处理

对于机器学习模型而言, 不同的参数设置以及输入数据对算法结果有着很大的影响, 下面给出经过参数调优之后的详细设置以及经过一定预处理后输入机器学习模型的数据. 分类实验的结果都基于以下参数设定和数据处理.

(1)对SVM而言, 本文只使用线性分类支持向量机求解, 根据经验惩罚系数C设置为1. 对于非线性分类问题, 可以使用核函数技巧求解. 在后续实际观测中, 如果使用线性核函数的结果不能满足观测需求, 可以根据实际需求将线性核修改为非线性核.

为了区分正常图像和云污染图像, 需要提取图像数据中具有显著类别特征的指标. 图像的纹理特征是区分云污染图像和正常图像的关键特征之一. 对于云污染图像而言, 云与背景之间的边界区域在局部的像素值上会发生显著变化, 这导致该区域的标准差与其他区域存在一定差异, 我们称为灰度不一致度. 该特征可以表示为:

$$G = \lg\left(\frac{P_{\mathrm{m}}}{P_{\mathrm{s}}}\right) . \tag{14}$$

P_{m}和P_{s}分别为局部区域标准差的最大值与最小

值. 由于观测图像中的星像呈点状分布, 根据典型星像的大小, 本文选取了9×9的边界框对整个图像进行遍历. 在实际观测中, 针对不同的设备和观测策略, 我们可以灵活调整边界框的大小. 因此, 纹理特征和图像的灰度不一致度可以作为区分云污染图像与正常图像的重要依据, 即数据点x_i的特征指标. 然而, 不同特征的量级可能存在较大差异. 例如, 图像的灰度不一致度通常大于10, 而某些图像的能量可能小于0.1. 为了消除不同指标间量级差异对分类结果的影响, 通常需要对这些指标进行归一化处理. 我们对训练集和验证集中的同一指标进行归一化, 最终得到5个归一化后的指标集合.

(2)对Shufflenet V2而言, 模型的损失函数选用交叉熵损失(CrossEntropy Loss), 其公式如下:

$$\mathrm{Loss}(x_i,\mathrm{class})=-\sum_{i=1}^{n}\left\{\lg\left[\exp(x_i[0])+\exp(x_i[1])\right]+x_i[\mathrm{class}]\right\}. \quad (15)$$

这里的class代表图像的真实类别(0表示正常图像, 1表示云污染图像), $x_i[j]$ $(j=0,1)$表示模型预测该图像为某类别的概率. 我们使用随机梯度下降(Stochastic Gradient Descent, SGD)算法作为优化算法, 初始学习率设置为0.01, 动量设置为0.9, 并将权值衰减设置为0.00004. 同时, 采用余弦退火(Cosine Annealing)策略动态调整每一轮的学习率. 该策略通过余弦函数平滑地调整学习率, 有助于避免训练过程中学习率的剧烈波动, 从而提高模型的稳定性和收敛性.

(3)对Resnet 18而言, 在本文中, 损失函数采用CrossEntropy Loss, 选择了自适应动量估计(Adaptive Moment Estimation, Adam)优化器, 因其具有较快的收敛速度. 学习率设置为0.0001, 学习率作为控制优化过程中每次参数更新步长的超参数, 较小的学习率能够有效避免更新过大而错过最优解的情况.

我们利用第2节中所筛选出的数据作为原始训练数据, 为了消除观测夜背景天光亮度随时间变化的影响, 需要对数据进行一定的预处理. 这些图像的格式是16位浮点型FITS (Flexible Image Transport System)图像, 我们首先利用Astropy库[27]中的Normalization和Stretching将FITS图像转换为深度学习模型训练常用的JPEG格式. Normalization代表将原始灰度值[0, 65535]映射到区间[0, 1], 公式如下:

$$K=\frac{j-j_{\min}}{j_{\max}-j_{\min}}. \quad (16)$$

这里K代表归一化之后的灰度值, j代表FITS图像中原始灰度值, $j_{\min}$代表图像原始灰度值中的最小值, $j_{\max}$代表图像原始灰度值中的最大值. Stretching代表把归一化后的数值通过线性或者非线性的函数映射到[0, 255]. 这里我们选择采用线性函数, 这意味着最终的灰度值是归一化后的255倍.

4.2　训练结果

图8为Shufflenet V2和Resnet 18在训练集和验证集上的损失变化情况. 从图中可以看出, 两个模型都在较短的训练轮次内迅速收敛至较低的损失值. 图9为两种模型在训练集和验证集上的准确率变化情况. 在这里, train mean loss代表训练集上损失大小, val mean loss代表验证集上损失大小. 其中, 准确率指模型正确分类的图像数量占所有待分类图像的比例, 定义如下:

$$\mathrm{accuracy}_{\mathrm{all}}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}. \quad (17)$$

其中, TP表示实际为云污染图像且被模型正确预测为云污染图像的数量; TN表示实际为正常图像且被模型正确预测为正常图像的数量; FP表示实际为正常图像但被模型误预测为云污染图像的数量; FN表示实际为云污染图像但被模型误预测为正常图像的数量. 在训练集上, 两个模型的准确率均超过99%. 然而, 在验证集上, Shufflenet V2的准确率可达97%, 而Resnet 18则为95%. 这表明Resnet 18的泛化能力相较于Shufflenet V2稍逊. 在SVM方法中, 我们选用了线性核函数. 实验结果显示, SVM在训练集和验证集上的准确率与两种深度卷积神经网络方法接近, 都能达到99%的准确率. 在验证集上的准确率为97%. 由此可以看出, 在训练

和验证过程中, 传统机器学习方法与深度学习方法的效果接近.

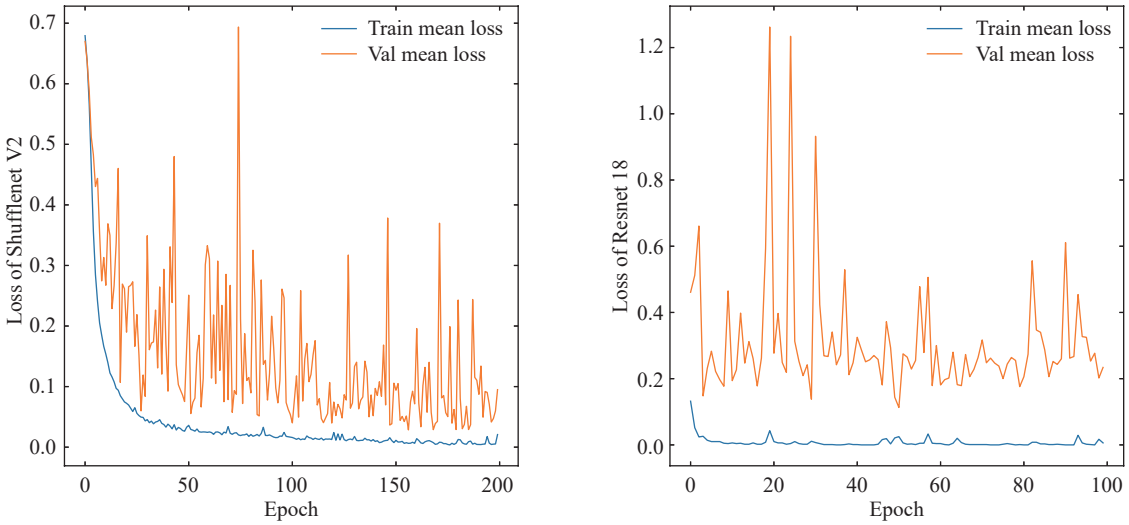

图 8 Shufflenet V2和Resnet 18模型在训练集与验证集(val)上的损失率变化. (a): Shufflenet V2, (b): Resnet 18.

Fig. 8 The loss of the Shufflenet V2 and Resnet 18 models on train and validation (val) data. (a): Shufflenet V2, (b): Resnet 18.

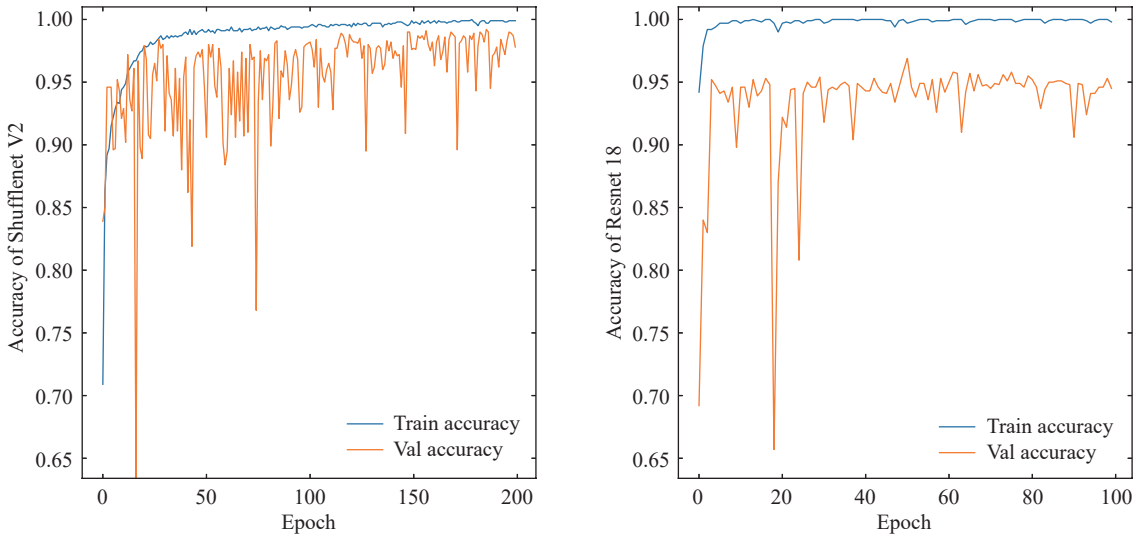

图 9 Shufflenet V2和Resnet 18模型在训练集与验证集上的准确率变化. (a): Shufflenet V2, (b): Resnet 18.

Fig. 9 The accuracy of the Shufflenet V2 and Resnet 18 models on train and validation data. (a): Shufflenet V2, (b): Resnet 18.

4.3 测试结果

图10展示了3种模型在测试集上的混淆矩阵结果. 我们使用了620张正常图像和514张云污染图像作为测试集, 为了进一步评估分类性能, 我们引入了一个云污染图像准确率指标, 定义为正确分类的云污染图像占所有云污染图像的比例. 其计算公式为:

$$\text{accuracy}_{\text{cloud}} = \frac{\text{TP}}{\text{TP} + \text{FN}} . \tag{18}$$

从结果中可以看出, Shufflenet V2模型将26张

云污染图像误分类为正常图像, 占所有测试图像的2.29%, 占云污染图像的5.06%. 这意味着, 该模型对所有新样本的分类准确率超过97%, 且对云图的分类准确率超过94%. 相比之下, Resnet 18模型将52张云污染图像误分类为正常图像, 9张正常图像误分类为云污染图像, 总分类准确率为94.62%, 云污染图像分类准确率为89.88%. 这两个准确率均低于Shufflenet V2模型. 对于SVM模型, 其错误分类了7张云污染图像为正常图像, 68张正常图像为云污染图像, 整体分类准确率为93.39%, 而云污染图像分类准确率则高达98.63%.

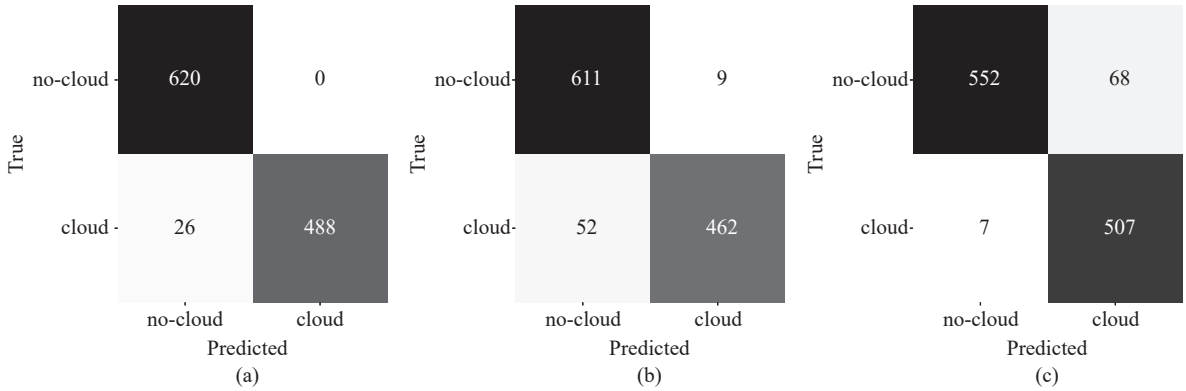

图 10　测试集数据得到的3个模型的混淆矩阵. (a): Shufflenet V2, (b): Resnet 18, (c): SVM.

Fig. 10　Confusion matrix of three models on test data. (a): Shufflenet V2, (b): Resnet 18, (c): SVM.

从混淆矩阵的结果来看, Shufflenet V2模型在3种方法中具有最高的总准确率, 表明其在所有图像的分类中出现错误的概率最小. 然而, SVM模型在云图分类方面的表现最佳, 显示其在识别云图时的错误分类率最低. 根据实际需求, 我们可以选择不同的模型来优化所有云污染图像的筛选效果. 例如, 在需要保证总分类准确率时, Shufflenet V2模型是较好的选择. 根据测试集结果, 若测试图像量为1000张, 则模型的误分类图像数量约为30张左右, 已能达到人工筛选的精度, 能够满足日常观测需求. 如果在观测时需要尽可能多地筛选出云污染图像, 并且要保证分类时的准确率较高, 则可以考虑使用SVM模型. 通常使用大视场望远镜进行空间碎片光学巡天时, 每台望远镜每晚能够获取数千甚至上万张图像, 此时较高的总体准确率能够有效减少误分类数量, 深度学习方法展现出了明显优势.

经过深入分析被模型错误分类的云污染图像, 发现其中大部分为污染程度轻微的图像, 仅有少量云污染程度严重的图像. 前者的分类错误可能由于轻微云污染的图像中云的特征不明显, 和正常巡天图像的特征相似; 而后者的分类错误则由于云污染严重图像的数量较少, 影响了模型训练, 随着后续观测获取更多数量的这类图像, 对其分类会更加准确.

我们也进一步利用不同观测夜的数据对模型的性能表现进行了测试. 对于新观测的数据, 我们筛选了3147帧云污染图像和2802帧正常图像, 图11为3种模型计算的新观测数据的混淆矩阵结果. 可以发现对于总准确率和云污染图像的识别准确率, Shufflenet V2都高于Resnet 18和SVM, 均优于95%, 而Resnet 18的总准确率不到92%, 云污染图像的识别准确率不到91%. SVM的总准确率和云污染图像的识别准确率都不超过91%. 传统机器学习方法在新观测夜的泛化能力远低于两种深度学习方法, 而Shufflenet V2算法的泛化能力高于Resnet 18. 3种方法在新观测数据上的表现都低于在原有数据集上的表现, 这是符合预期的. 主要原因是由于训练和验证样本数量较少, 而深度学习方法的优势在于从大批量数据中学习数据的特征,

并拥有较强的泛化能力. SVM方法的优势在于小批量数据的模型训练时泛化能力强, 且计算耗时较低. 随着观测数据的积累, 深度学习模型的表现会越来越好.

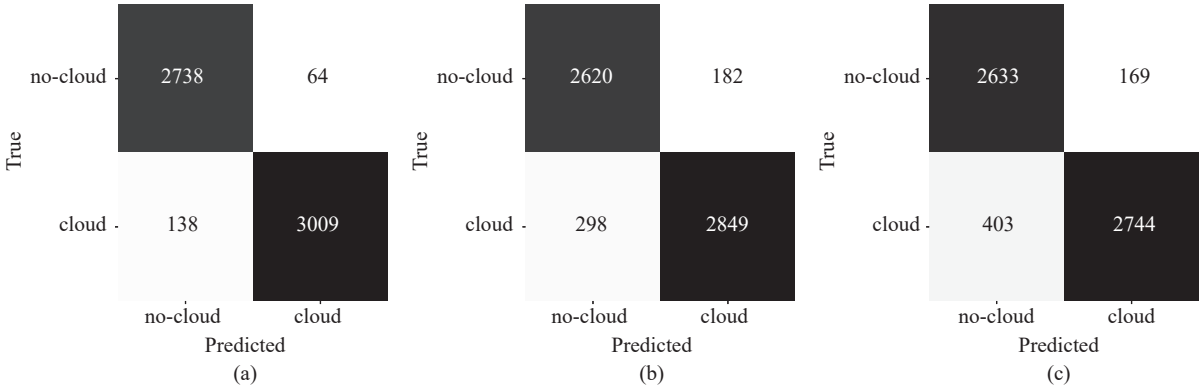

图 11 新观测数据计算得到的混淆矩阵. (a)–(c)分别为Shufflenet V2、Resnet 18和SVM的混淆矩阵.

Fig. 11 Confusion matrix of the three models for new data. (a)–(c) represent the confusion matrix of Shufflenet V2, Resnet 18, and SVM, respectively.

4.4 算法时间对比

在实际观测中, 除了考虑模型对云污染图像的识别效果外, 还需关注模型训练与测试所消耗的时间. 如果模型在训练和测试时所需时间较长, 可能无法满足实时观测中大量云污染图像的快速识别需求. 3个模型的训练和测试时间见表3. 从表中可以看出, 传统机器学习模型SVM在训练和测试过程中具有较快的速度, 单次收敛时间小于10 ms, 整体训练(200个epoch)可以在几秒钟内完成. 两种深度学习模型中, Shufflenet V2在训练和测试时均快于Resnet 18. 在测试阶段, 尽管SVM所需时间比Shufflenet V2稍短, 但由于每张图像的曝光时间通常远大于模型测试所需的时间(例如本文中每张图像的曝光时间为2 s), 因此SVM和Shufflenet V2都能够满足实时观测对云污染图像识别的时间要求. 值得注意的是, Shufflenet V2相比SVM具有更高的识别准确率. 综合来看, 深度学习模型相较于传统机器学习模型在云污染图像识别效果上表现更好, 且在速度上也能够满足实时观测的需求.

表 3 不同模型训练和测试的时间对比

Table 3 The comparison of training and testing times for different models

Model	SVM	Shufflenet V2	Resnet 18
Training (epoch)	< 0.01 s	∼90 s	∼160 s
Testing (frame)	< 0.01 s	∼0.030 s	∼0.034 s

5 结论与展望

本文旨在探究机器学习方法在地基大视场光学巡天云污染图像识别中的应用, 并基于实测数据比较了几种典型方法的性能与表现. 首先, 我们通过人工筛选出受云影响的观测图像, 并通过比较多项图像纹理特征指标, 确保所筛选出云污染图像的准确性. 然后, 基于获得的数据集, 分别使用两种深度学习模型和一种传统机器学习模型在测试集上进行验证. 结果表明: 测试集上的总体准确率达到97%, 云污染图像分类准确率达到98%. 进一步分析发现, 深度学习方法相较于传统机器学习方法在准确率上具有一定的优势. 此外, 3种方法在模型处理速度上均能满足实时识别的要求.

该方法可以在巡天观测图像采集过程中实时识别出云污染的图像, 通过与云量仪配合, 支撑观测策略优化, 维持观测业务流程的稳定运行.

本文试验的方法作为不增加额外成本与负担的云污染图像感知工具, 可以用于在观测台站没有云量仪时有效识别云污染图像, 给测站工作人员提供可靠的云量先验信息; 也可以和云量仪协同配合, 更加准确地优化观测计划, 维持观测业务流程的稳定运行. 对于识别出的云污染图像, 根据云污染程度的轻重, 我们可以采用不同的应对策略, 对于云污染程度严重的图像, 图像中能检测的源的数量相较正常图像大幅减少, 这些图像难以利用, 可以适当舍弃、节省存储空间; 对云污染程度轻微的图像, 图像中只有少量的源被遮挡, 在后续研究中, 可以考虑使用图像处理算法降低云的影响, 提升数据整体的科学产出.

参 考 文 献

[1] 王歆. 飞行器测控学报, 2015, 34: 201
[2] Zhong X, Du F, Hu Y, et al. Electronics, 2024, 13: 1503
[3] Klebe D I, Blatherwick R D, Morris V R. AMT, 2014, 7: 637
[4] Zhi H, Wang J F, Zhang X M, et al. PASP, 2024, 136: 035002
[5] Wang Y R, Liu D, Xie W Y, et al. RemS, 2021, 13: 1852
[6] 樊亮, 雷成明, 师冬冬, 等. 天文学报, 2017, 58: 49
[7] 秦永, 付仲良, 周凡, 等. 武汉大学学报, 2014, 39: 234
[8] 王利文, 贾鹏, 蔡冬梅, 等. 天文学报, 2018, 59: 25
[9] Wang L W, Jia P, Cai D M, et al. ChA&A, 2019, 43: 128
[10] Li H, Li R W, Shu P, et al. RAA, 2024, 24: 045025
[11] 车蕾, 李磊磊, 刘立勇. 天文学进展, 2024, 42: 349
[12] Mommert M. AJ, 2020, 159: 159
[13] Haralick R, Shanmugam K, Dinstein I. IEEE Transactions on Systems, Man, and Cybernetics. Piscataway: IEEE, 1973, 6: 610-621
[14] Zhao H, Sun R Y, Yu S X. RAA, 2024, 24: 115009
[15] Duev D A, Mahabal A, Ye Q Z, et al. MNRAS, 2019, 486: 4158
[16] He Z Z, Er X Z, Long Q. MNRAS, 2020, 497: 556
[17] Vázquez C V, Solano E, Ulla A, et al. A&A, 2024, 691: A223
[18] Krakowski T, Małek K, Bilicki M, et al. A&A, 2016, 596: A39
[19] Kasapis S, Kitiashvili I N, Paul K, et al. AJ, 2024, 974: 131
[20] Cortes C, Vapnik V. Machine Learning, 1995, 20: 273
[21] 周志华. 机器学习. 北京: 清华大学出版社, 2016: 60-229
[22] Ma N N, Zhang X Y, Zheng H T. European Conference on Computer Vision. Munich: Springer, 2018, 11218: 122-138
[23] Zhang X Y, Zhou X Y, Lin M X, et al. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 6848-6856
[24] He K M, Zhang X Y, Ren S Q. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 770-778
[25] Prithivraj G, Kumari A. arXiv, 2023: 2305.09596
[26] Jia P, Liu Q, Sun Y Y. AJ, 2020, 159: 212
[27] The Astropy Collaboration, Price-Whelan A M, Sipőcz B M. AJ, 2018, 156: 123

Cloud-contaminated Image Recognition Based on Machine Learning

LIU Liu[1,2] ZHAO He[1,2] SUN Rong-yu[1]

(1 Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing 210023)
(2 School of Astronomy and Space Science, University of Science and Technology of China, Hefei 230026)

Abstract In optical surveys of space targets and debris, clouds are one of the key factors affecting observation efficiency. Cloud cover can reduce target visibility, complicating target image detection and subsequent accurate position and brightness extraction, thus interfering with the observation of space targets and debris. Efficient and accurate identification of cloud-contaminated images can provide valuable prior information for subsequent data processing, supporting the normal and stable operation of the operational workflow. Currently, mainstream methods based on image segmentation to detect clouds in images have drawbacks such as being time-consuming and vulnerable to noise. This paper combines image feature evaluation and manual screening to establish a cloud-contaminated image dataset based on optical survey data of space debris. We experiment with three machine learning methods: support vector machine, Shufflenet V2, and Resnet 18, to classify cloud-contaminated and normal images. The results show that Shufflenet V2 achieves an overall classification accuracy greater than 97%, while the SVM (Support Vector Machine) model achieves a cloud-contaminated image recognition accuracy of over 98%. Deep learning methods can effectively identify cloud-contaminated images, and the computational speed meets the real-time processing requirements for observational data. In future observations, the proposed method can be integrated with cloud imager and jointly applied to optimize space debris observation plans, reducing the impact of weather on observational equipment performance and promoting more stable operation of observation stations.

Key words methods: data analysis, techniques: image processing, astronomical data bases: surveys