文章编号：2096-1618(2020)04-0382-10

基于二维粒子谱仪的固态降水粒子自动分类研究—雪花和霰

林慧玲[1]， 肖 辉[2,3]， 姚振东[1]， 孙 跃[2,3]， 杨慧玲[2]， 冯启祯[4]， 饶 晨[1]

(1. 成都信息工程大学电子工程学院，四川 成都 610225；2. 中国科学院大气物理研究所中国科学院云降水物理与强风暴重点实验室，北京 100029；3. 中国科学院大学，北京 100049；4. 中国民用航空飞行学院绵阳分院，四川 绵阳 621000)

摘要：固态降水粒子进行准确而细致的分类对许多大气过程及天气雷达的应用是十分重要的。使用二维光学粒子谱仪(2DVD)对单个降水粒子进行测量，并基于测得的粒子微物理参数及特性提供降水过程中一分钟单位时间间隔内主要降水粒子类型的估测，对固态降水粒子进行自动分类。为实现自动分类任务，考虑将该工作与常用的机器学习分类算法相结合，应用朴素贝叶斯，支撑向量机(SVM)，决策树三种监督学习算法对单位时间间隔内的粒子分类。文中将降水粒子归类为雪花和霰两种主要类型，并结合人工检测进行结果验证，最终利用独立的数据集进一步验证，证明分类算法的准确性。

关 键 词：固态降水粒子；2DVD；粒子自动分类

中图分类号：TP731 **文献标志码**：A

doi：10.16836/j.cnki.jcuit.2020.04.004

0 引言

识别固态降水粒子形状对理解其微观特性十分重要[1]。识别降水粒子，尤其是固态降水粒子形状对于云降水微物理的认识研究具有重要意义。粒子形状会影响自身的散射特性、增长率和下落末速度[2]。准确分类降水粒子类型可以为许多大气相关研究提供有效的信息帮助，如：天气雷达降水估计的验证[3]，数值天气预报模式的开发、改进和验证[4]，云微物理结构变化如粒子的凝华、淞化过程等。

另外，随着天气雷达降水估测工作的开展及双偏振雷达的普及应用，有关降水粒子分类的工作也得以发展并基于不同探测仪器不断开展。例如，遥感中较为典型的目前开展广泛的基于双偏振雷达的模糊逻辑算法识别降水粒子的相关研究[5]，以及一些用于观测冰相云的机载雷达和激光雷达探测工作[6]。这些探测仪器能够在短时间内进行高分辨率的采样，但由于这些仪器间接反演参数的原理，分析中受到数值模拟的约束，并且难以进行广泛验证。相比之下，机载成像仪等仪器虽能进行直接的穿云观测，但受限于飞机的飞行姿态，观测过程不稳定，成本过高。可进行直接定点探测的地基探测仪器虽采样面积较小，但可以作为单个参考点补充降水资料。自地面滴谱仪被国外科研工作者提出以来，在降水过程相关研究方面的应用逐渐展开。一维光学雨滴谱仪(parsivel)可以提供单位测量时间间隔内降水粒子下落末速度、降水强度及降水分类等信息[7]，目前已广泛用于降水研究并相对成熟。二维视频粒子谱仪(以下称2DVD)能够直接测量固态降水粒子的自然下落速度，并实时记录单个粒子的尺寸、形状、体积等微观特征。因此，可用以描述液态和固态降水的微观物理和微观结构，显著提高地面降水观测的能力[8]。使用地基探测仪器-2DVD进行冬季近地面固态降水粒子的测量及自动分类工作。

1 数据处理

主要进行降水粒子的特征提取，根据人工监测给出的待分类项标签，形成训练样本集合。数据特征的充分表达以及训练样本集的质量是决定分类器实现效果及性能的关键，对整个过程将有重要影响。

1.1 仪器与试验

数据由二维视频粒子谱仪(2DVD)在两次冬季观测试验中收集提供。

2DVD粒子谱仪中共配置2个正交光源和(A、B)2个高速线性扫描相机，并根据其相耦合的结构构成

收稿日期：2020-03-17

基金项目：国家重点研发计划战略性国际科技创新合作重点专项资助项目(2016YFE0201900-02)；国家自然科学基金面上资助项目(41575037)；国家重点基础研究发展计划资助项目(2014CB441403)

约 10 cm×10 cm 的层叠测量区域[8]。当降水粒子下落经过测量区域时,2DVD 通过 A、B 2 个相机的同步扫描,在 2 个正交视图上记录降水粒子在测量区域的阴影,根据像素点信息重建粒子形状和大小。另外,两摄像机之间约6.2 ~7 mm的(确切的值由机械校准确定)垂直距离的设计能够用以测量降水粒子的下落速度,粒子纵横比等参数信息[9]。自从 2DVD 研制以来,因其对单个粒子信息记录准确而详细,已被广泛应用于许多其他相关研究。该仪器的性能也已经得到评估和不断改善。

2DVD 的测量原理及结构如图 1 所示。

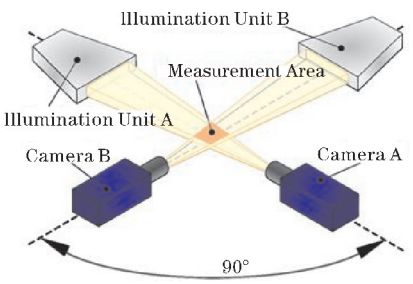


图 1　2DVD 测量原理图

使用 2DVD 对冬季固相降水粒子进行观测与分析。在降水过程中,记录测量区域内下落的固态降水粒子大小、形状和末速度等物理参数信息。文中使用的 2DVD 数据来源于 2015-2016 年及 2018-2019 年两次冬季观测。这两次观测分别于中国北京顺义国家气象观测站和张家口市崇礼区气象站展开。

1.2　数据预处理

原始的 2DVD 降水资料在进行分析前必须要经过质量控制。2DVD 数据质量控制主要包括剔除与匹配两个方面[4]。降水粒子在下降过程中,由于碰撞,下落间隔与设备自身分辨率以及 2DVD 设备周围的空气动力学影响会引起获取资料中不可避免地包含一些不合理的测量,即伪粒子的数据记录。伪粒子的存在,导致从粒子图像获取的微物理参数发生一定偏差,这不仅不利于云降水物理研究,也将影响测量结果[6]。因此,首先要对 2DVD 降水粒子图像中的伪粒子进行识别和剔除。伪粒子一般包括破碎粒子、空白粒子、条形粒子等。为剔除伪粒子,对 A、B 两个相机的测量数据进行合理的条件约束。

2DVD 数据的匹配是指经过相机 A、B 对同一粒子一致性的简单匹配[9]。固态粒子由于其复杂的形状及结构特征使相机图像的匹配变得复杂。相机下粒子的匹配包括图像匹配与数量匹配。尽管实际的固态粒子从不同的视角将呈现出不同的形状,但是将两台相机截然不同的粒子图像滤除可以减少误差。文中关于 2DVD 图像的匹配,基于 Hanesch (1999) 和 Huang (2010)的工作并根据实际情况加以合理的修正。根据粒子在两相机下唯一相同的参数——粒子图像的高度 H 作为判断准则[6],测量下降速度用于设置匹配的时间窗口,在时间窗口内完成 A、B 相机图像对的匹配。

除了上述数据质量控制,对固态降水粒子进行形状识别还要考虑到其他因素的潜在影响。2DVD 对风具有敏感性,仪器周围的水平风场产生的局部风很可能导致测量区域内降水粒子和 2DVD 图像本身的空间分布水平失真,进而影响对降水粒子尺寸分布的测量,尤其是小密度、小尺寸的粒子。研究中,结合固态降水粒子的属性特征,为提高数据质量,粒径≤1 mm的降水粒子为可疑粒子并进行剔除,选用静风至中等风速条件下的 2DVD 降水资料,风速约束条件为 $v_{\mathrm{wind}} \leqslant 5\ \mathrm{m \cdot s^{-1}}$[9]。

1.3　数据特征提取及表达

降水粒子经测量区域下落时,2DVD 提供单个粒子的 A、B 两个视图的像素以及粒子粒径、体积、下落速度等信息。选取合理的图像和数据特征作为粒子特征参数,并通过这些参数进行粒子特征表达。

这些特征量包括通过直接测量得到的粒子参数和进一步处理得到的图像形状参数。直接测量得到的粒子参数包括粒子直径 De、下落速度 Ve、长轴 Obl_l、短轴 Obl_s,粒子高度 height,粒子宽度 longest line。这些属性由 A、B 两个相机联合计算得到。其中,粒子直径是指粒子的等体积直径,见表 1。

降水粒子的形状参数需要对两个视图的原始图像进行相应的图像处理。经过初步图像处理得到的一些参量定义如图 2 所示[10]。其中,Pa 为通过图像处理获得的粒子图像周长,Aa 为面积,W、H 分别代表粒子还原图像的最小外接矩形边长。

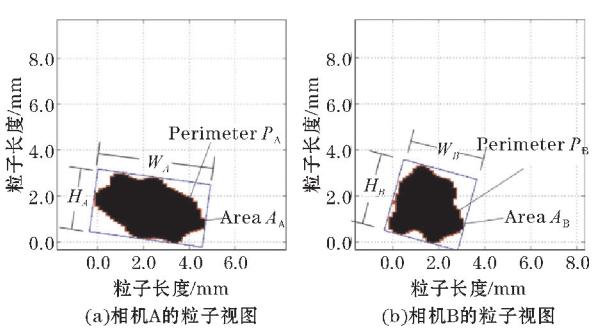


图 2　相机 A、B 对中对粒子周长、面积、高和宽的定义

利用这些特征量,可以进一步计算出更多详细的无量纲形状特征参量用以补充描述降水粒子的形状和尺寸参数。

表 1 粒子特征描述量

序号	特征量	含义
1	v	下落末速度/(m/s)
2	D_e	等体积直径/mm
3	Vol	粒子体积/mm^3
4	$A_{A,B}$	粒子测量面积/mm^2
5	$P_{A,B}$	粒子还原图像周长/mm
6	$Obl_{A,B}$	粒子横纵比
7	$Elong_{A,B}$	粒子伸长程度
8	$PF_{A,B}$	像素分位数(矩形)
9	$Roundness_{A,B}$	像素分位数(外接圆)
10	$Shape\_factor_{A,B}$	形状指数
11	$SI_{A,B}$	形状指数
12	$FD_{A,B}$	分形维数

其中,$Obl_{A,B}$为 A,B 相机测量的粒子的横纵比,$Elong_{A,B}$为还原的粒子图像的最小外接矩形的边长之比,$PF_{A,B}$和 $Roundness_{A,B}$特征量分别表示粒子还原图像在最小外接矩形和最小外接圆中所占的面积比例。$Shape\_factor_{A,B}$、$FD_{A,B}$和 $SI_{A,B}$则均作为形状指数从多个角度共同描述了单个降水粒子的形状特征及形状复杂度。A,B 分别表示由相机 A 或者 B 计算得到的该项特征参数。

根据定义的粒子特征量经过计算可以得到以下无量纲的形状特征参量,进一步详细描述降水粒子的形状特征。

$$Obl=\frac{Obl\_l_{A,B}}{Obl\_s_{A,B}} \quad [1/3,3] \tag{1}$$

$$E\mathrm{long}=\frac{W_{A,B}}{H_{A,B}} \quad [1,+\infty] \tag{2}$$

$$PF_{A,B}=\frac{Aa_{A,B}}{W_{A,B}\cdot H_{A,B}} \quad (0,1] \tag{3}$$

$$Roundness=\frac{4\cdot Aa_{A,B}}{\pi\cdot(\max(W_{A,B},H_{A,B}))^2} \quad (0,1] \tag{4}$$

$$Shape\_factor=\frac{4\cdot\pi\cdot Aa_{A,B}}{Pa_{A,B}^{\ 2}} \quad (0,1] \tag{5}$$

$$FD_{A,B}=2\frac{\ln(\frac{Pa_{A,B}}{4})}{\ln(Aa_{A,B})} \quad [1,2] \tag{6}$$

$$SI_{A,B}=\frac{Pa_{A,B}}{4\sqrt{Aa_{A,B}}} \quad [\sqrt{\pi}/2,+\infty] \tag{7}$$

将时间序列中每分钟内数据为单位进行样本划分,并使用这些属性在每分钟内的统计分布用作降水粒子分类的输入信息。这些统计量包括单位分钟内 N 个粒子各参数的中值、平均值、标准差、分位数、四分位距以及距平[11],以及 A、B 两个相机形状参数的相关系数。因为 A、B 两个相机信息高度相关[4],只选取其中一个相机(A)的记录数据进行处理作为待用样本。

1.4 训练数据集

监督式学习分类器主要的实现问题在于学习阶段的训练。研究在 2015-2016 年(顺义)、2018-2019 年(崇礼)共开展了 2 次冬季连续观测活动,期间分别收集了 4 次、5 次连续降雪过程,样本总数为 3379 分钟。

为给样本进行较为准确的类别标记,选择尽可能纯粹的可用样本并对该单位时间间隔内占比最多的粒子类型进行主要类型标记。

为准确地将降水粒子分类为雪花和霰,首先对采集的数据进行粒子图像还原及数据处理,应用 2DVD 仪器软件中.SNO(雪花文件,已进行粒子的初步固态筛选)文件进行粒子分类研究。研究中用于训练和测试分类器的粒子样本数有 2231 分钟数据。其中,包含 1996 个雪花样本和 235 个霰样本组成。

2 方法

2.1 冬季固态降水粒子分类

固态降水粒子不同于液态降水粒子仅由形状相似的雨滴组成,根据不同的参数特征,可分为多种类别。参数受各种因素影响,例如降雪过程的发展和宏观物理条件。

旨在结合常用的不同机器学习分类算法实现对冬季固态降水粒子关于雪花和霰两种基本类型的自动识别分类。

2.2 分类算法

分类是数据分析和机器学习领域的一个基本问题,也是近年来各领域进行研究的热点。选用了 3 种不同的监督式分类学习方法对冬季降水粒子进行自动分类,分别是朴素贝叶斯、支持向量机和决策树算法模型。样本数据集的分类标记基于观测者对于降水粒子的人工识别并结合粒径-下落速度关系[12-13]和粒子形状因子最终给出[14]。

2.2.1 朴素贝叶斯

朴素贝叶斯分类方法是基于贝叶斯定理与特征条件独立假设的分类方法[15]。朴素贝叶斯模型(naive Bayesian model,NBM)基于古典数学理论,有着坚实且成熟的数学基础以及稳定的分类性能。同时,该模型所需估计的参数较少,对缺失数据不太敏感,算法实现也比较简单。

但实际应用中,由于朴素贝叶斯分类模型很难满足数据样本属性之间相互独立的假设条件。因此,不

相互独立的样本数据往往会给朴素贝叶斯分类模型的分类性能效果带来影响。

朴素贝叶斯是最为简单且常用的一种贝叶斯分类器。其原理为计算某类的先验概率,利用贝叶斯公式计算出其后验概率,即该目标属于某一类的概率,选择具有最大后验概率的类作为该对象所属的类。即

$$y_k \mid_{y_k \in y} = \arg\max(P(y_k \mid x)) \tag{8}$$

根据贝叶斯理论,$P(yk|x)$ 计算方法为

$$P(y_k|x) = \frac{P(x|y_k)P(y_k)}{P(x)} \tag{9}$$

根据朴素贝叶斯算法的属性条件独立性假设,可以得到:

$$P(x|y_k)P(y_k) = P(y_k)\prod_{d=1}^{D} P(x_d|y_k) = P(y_k)P(x_1|y_k)P(x_2|y_k)P(x_3|y_k)\cdots P(x_D|y_k) \tag{10}$$

贝叶斯决策算法考虑基于概率分布和误判损失来选择最优的类别标记结果。应用的朴素贝叶斯分类算法中采用的概率密度函数为高斯分布。

2.2.2　SVM

支持向量机(support vector machine, SVM)分类器是一类以监督学习(supervised learning)方式对数据进行二元分类的广义线性分类器,其分类边界是对学习样本求解的最大边距超平面[16]。

SVM 分类模型的目标在于通过找到一个最佳超平面将数据分为两个独立的类。而且,这个超平面的选择应满足不同的类到该平面之间的距离都尽可能大[17]。当样本特征非线性时,SVM 可以通过引入核函数,即通过映射特征到另一个高维空间,得出非线性假设函数并在该空间中实现数据的线性分离。

SVM 是线性二元监督分类器,目标是寻求不同两类观测值之间的最佳分离。SVM 能够处理高维输入,比其他监督方法更不容易出现过度拟合问题。间隔最大化问题如下:

$$\min\left\{\frac{1}{2}\|w\|^2\right\} \quad s.t. \begin{cases} y_i[\langle x, w\rangle + b] \geqslant 1 \\ x_i \geqslant 0 \quad j = 1, \cdots, N \end{cases} \tag{11}$$

为解决粒子形状分类工作,对该目标函数进行修改及重写:

$$\min\left\{\frac{1}{2}\|w\|^2 + c\sum_{i=1}^{N_{train}}\zeta_i\right\} \quad s.t. \begin{cases} y_i[\langle x, w\rangle + b] \geqslant 1 - \zeta_i \\ \zeta_i \geqslant 0 \quad i = 1, \cdots, N_{\text{train}} \end{cases} \tag{12}$$

使用拉格朗日乘数 α 求解该优化模型,允许将问题重写为:

$$\max\left\{\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\langle x_i, x_j\rangle\right\} \quad s.t. \begin{cases} 0 \leqslant \alpha_i \leqslant c \\ \sum_i y_i\alpha_i = 0 \end{cases} \tag{13}$$

对两类样本分类的 SVM 模型结构如图 3 所示。

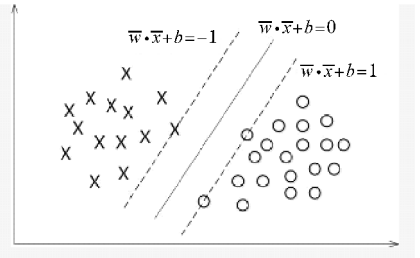


图 3　SVM 二元分类模型分类效果示意图

使用带有标准预测变量的二元 SVM ECOC 学习分类器。为解决输入特征线性不可分的问题,选择高斯径向基核函数(RBF)SVM 二元分类器实现粒子自动分类。

2.2.3　决策树

决策树是当下使用的最流行的非线性框架之一。分类决策树(decision tree)模型是一种简单易用的非参数分类器。它采用树形结构,每个分支代表一个测试输出,每个叶节点代表一种类别。简单地说,就是根据训练数据集构造一个类似树形的分类决策模型,然后用这个模型辅助分类决策。决策分类树不需要对数据有任何的先验假设,计算速度较快,而且稳健性较强[11]。但在应用决策树进行分类工作时,若不加以控制条件,树的生长会趋向于达到训练集的最佳拟合,此时容易使树的结构变得十分复杂。

学习时,利用训练数据,根据损失函数最小化的原则建立决策树模型。

对于分类树的构建,重要的是结点数和特征划分选择。特征选择表示从众多的特征中选择一个特征作为当前节点分裂的标准,并逐层生成树。建立生成的决策树分类模型结构如图 4 所示。

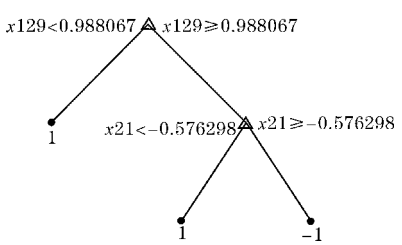


图 4　决策树分类模型示意图

应用以“基尼指数(Gini)”进行特征划分的 CART 分类决策树,并对构建的分类树进行剪枝防止过拟合,以提高其泛化能力。

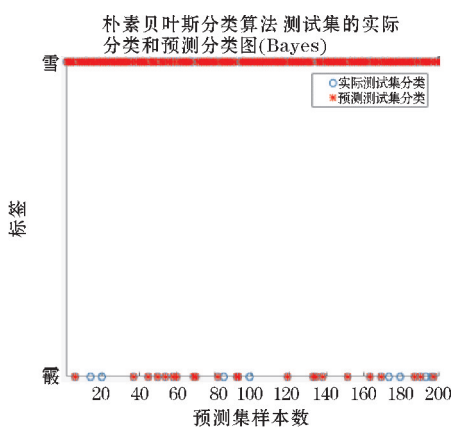
3　结果与分析

降水粒子分类算法设计中另一个具有挑战性的关键是分类器的性能评估。实际上,在仪器分辨率范围内收集独立且一致的降水粒子测量值十分困难。对分类结果进行分析并对文中所构建的分类器的分类性能进行合理性评估。

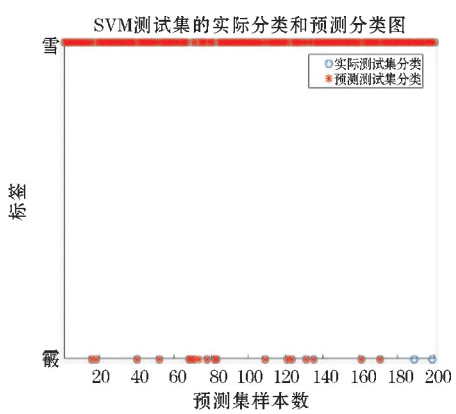
为评估预测样本分类的性能,通常进行交叉验证。交叉验证是为了避免将全部用于数据训练模型而造成没有数据集对该模型进行验证从而采用的一种合理的样本划分方法。采用 10 重交叉验证:将给定的数据集随机分为 10 份,轮流将其中 90% 的数据样本集用于训练分类器,剩余的 10% 则用以分类性能测试。将 10 次验证后每次试验得出相应正确率的平均值作为对算法精度的估计,一般还需要进行多次 10 折交叉验证(例如 10 次 10 折交叉验证),再求其均值,作为对算法准确性的估计。10 折交叉验证方法可以有效地避免过拟合与欠拟合的发生,最后得到的结果也比较具有可靠性。

分类器模型的超参数优化方式一般包括网格搜索、随机搜索和贝叶斯优化优化理论。文中进行二元粒子分类模型的构建与优化时,对 3 种不同的分类算法模型均采用不同的优化方式,并对优化结果进行比较。最后分别给出适用于不同分类算法的优化方式下分类效果最好的分类器模型结果。

图 5 显示样本集在不同分类模型下的预测集分类效果。图 5(a)给出了朴素贝叶斯分类器在预测集的分类结果,图 5(b)为随机优化方法的 SVM 算法分类器模型效果。



(a)朴素贝叶斯模型分类器预测集结果



(b)SVM 模型分类器预测集结果

图 5 样本集在不同分类模型下的预测集分类效果

为更加准确、定量的评价分类情况,引入常用的模型评估统计量-混淆矩阵进行说明。混淆矩阵(也称误差矩阵,confusion matrix)是分类型模型评判的指标中最常见的方法之一。

混淆矩阵多用于判断分类器性能的优劣,同时也是衡量分类模型准确度方法中最基本、最直观且计算最简单的一种[18]。表 2、表 3 分别给出了朴素贝叶斯分类器、SVM 分类器在样本不均衡下预测集结果的混淆矩阵 $\boldsymbol{C}$。

表 2 朴素贝叶斯分类器样本不均衡下预测集结果混淆矩阵

	雪花	霰
雪花	190	9
霰	1	26

表 3 SVM 分类器样本不均衡下预测集结果混淆矩阵

	雪花	霰
雪花	189	7
霰	2	28

以上两种算法的 20 次交叉验证预测集平均分类准确率结果由表 4 给出。

表 4 朴素贝叶斯分类模型和 SVM 分类器模型的平均分类准确率

	朴素贝叶斯	SVM
雪花分类准确率/%	99.6324	98.4261
霰分类准确率/%	71.2857	78.9231
预测集分类准确率/%	93.9684	95.7164
10 重交叉检验误差	0.1794	0.0198

在正负样本不平衡的情况下,准确率这个评价指标有很大的缺陷,单纯靠准确率来评价一个算法模型远远不够全面、也不够可靠。因此,基于上文混淆矩阵的计算,进一步计算得到分类器的分类精确率和召回率。其中,分类精确率表示被分为正例的示例中实际为正例的比例,召回率是覆盖面的度量,度量有多个正例被分为正例。F1 score 是精确度和召回率的加权调和平均,平衡了两个指标的结果,当 F1 较高时则能说明试验方法比较有效,见表 5。

表 5 朴素贝叶斯分类模型和 SVM 分类器模型的评价指标

单位:%

	NBM	SVM
精确率	0.6650	0.9655
召回率	0.6775	0.9032
F1 Score	0.6687	0.9333

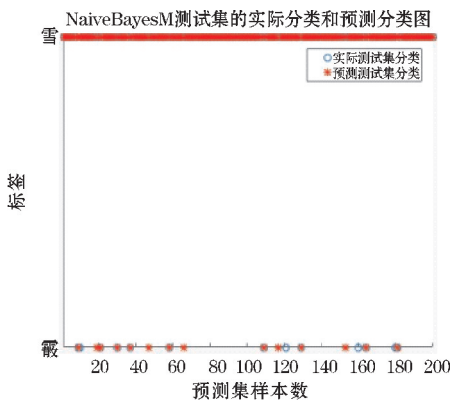
根据分类结果可知,朴素贝叶斯模型分类器和 SVM 模型分类器对于雪花类别的粒子具有很高的分类准确率,但对霰粒子的分类准确率却低,两类粒子整体分类准确率差异过大。由于雪花粒子在样本中占比较大,高准确率的雪花粒子分类带来了较高的预测集

分类准确率,但此时的预测集准确率忽略了小比例的霰类,过于乐观,可靠性低。

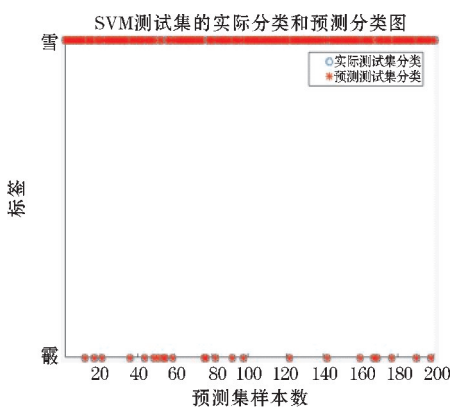
研究中,收集和应用的两类样本(雪花和霰)的数据量存在很大差异:雪花样本总数为 3051 min,而霰样本量仅有 306 min。根据对霰粒子的样本数据统计可知,收集的霰粒数量相对过少且数据过于集中,样本分布不完整。不同类别样本集的数据不均衡使分类结果的准确性受到明显影响。

文中应用的高斯朴素贝叶斯模型基于贝叶斯理论,通过对数据的概率分布拟合进行样本分类。当训练集中各个类别的样本分布不均匀且存在数据稀疏问题时,分类器容易倾向于大类别而忽略小类别,使朴素贝叶斯算法分类不够准确。

同样地,由于 SVM 算法对缺失数据敏感,构建的 SVM 分类器模型的分类性能对参与分类的不同类别样本集的数量有很大的依赖性,在数据偏斜时会造成一定程度的分类偏斜。类别数据不均衡是分类任务中一个典型的存在的问题。文中,雪粒子样本相对充足,分布区域覆盖广,而霰粒子由于发生受限,导致样本过少,数据分布不完整,霰粒子的分类准确率较雪花样本存在很大差异。为解决样本不均衡造成的数据偏斜,可以尝试通过修改正则化系数为偏斜数据赋权,即对模型加以惩罚进行改进[18]。



(a)朴素贝叶斯模型分类器预测集结果



(b)SVM 模型分类器预测集结果

图 6　改进后不同模型分类器预测集结果

文中首先按两类样本的比例设定正则化系数对模型进行修改:对雪花样本集使用小的正则化系数,对霰粒样本集则使用大的正则化系数。之后,在初步的训练结果上根据训练情况不断调整正负样本权重比例,得到最佳分类器模型。图 6(a)为引入比例权重后的朴素贝叶斯分类器模型预测集的分类效果,图 6(b)为改进后的 SVM 分类器的分类效果。混淆矩阵结果如表 6、表 7 所示。

表 6　朴素贝叶斯分类器数据偏斜改进后预测集结果混淆矩阵

	雪花	霰
雪花	187	6
霰	8	25

表 7　SVM 分类器数据偏斜改进后预测集结果混淆矩阵

	雪花	霰
雪花	188	5
霰	4	30

经过偏斜数据改进算法训练得到的 SVM 分类器模型对于冬季固态降水粒子的分类取得不错的结果:对于雪花分类的平均准确率达到98.7692%,而霰粒子的分类也达到91.3226%,预测集的最终平均分类准确率约为97.8089%。广义分类估计误差为0.0132,表明该分类器的泛化效果较好,见表 8。

表 8　朴素贝叶斯分类模型和 SVM 分类器模型的平均分类准确率

单位:%

	朴素贝叶斯	SVM
雪花分类准确率	96.9388	98.7692
霰分类准确率	80.6452	91.3226
预测集分类准确率	95.3846	97.8089
10 重交叉检验误差	0.0397	0.0132

改进后的朴素贝叶斯分类器分类性能也有一定的改进:雪花分类准确率为 96.9388%,霰类为 80.6452%,预测集平均分类准确率约为95.3846%,广义分类误差估测为0.0397。此时,训练得到的朴素贝叶斯分类器模型的准确率虽已达到 95% 左右,但由于霰类和雪花的先验概率相差过大([0.0957, 0.9043]),霰类的分类效果依然明显偏低。

由于降水过程的自然属性,固态降水粒子特征的分布是未知的。因此在假设概率密度分布前提下实现的朴素贝叶斯度算法分类器的分类性能会受到一定程度的影响。SVM 分类器模型通过参数的迭代优化得以实现,而并不需要降水粒子的分布信息。因此在对降水粒子的分类工作中,分类性能更加优秀。另外,霰类极小

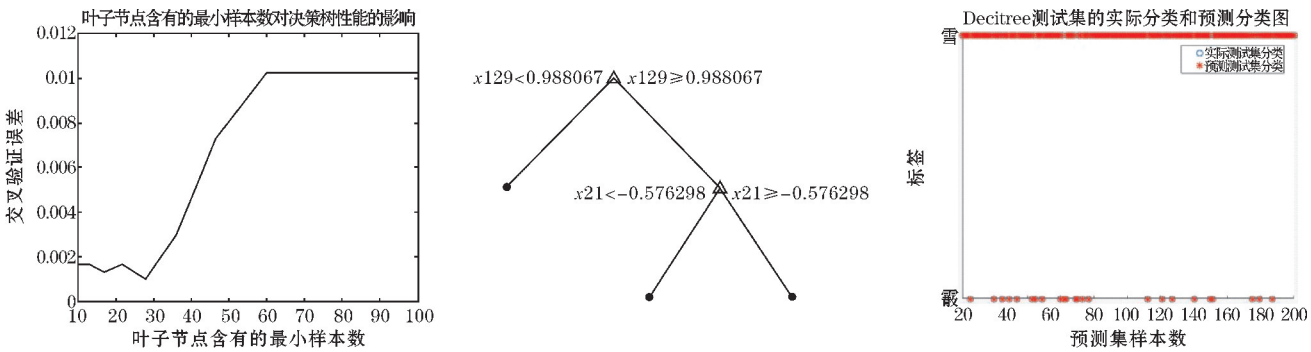
的先验概率以及样本数据的稀疏使得霰类的分类效果大大降低。在 SVM 分类器模型中,虽引入权重增加霰类错误分类的代价弥补样本数据的不足,但因为用于分类的霰类样本缺失值较多,分类依然偏向于雪花。SVM 分类器较朴素贝叶斯分类器有更高的泛化能力,见表9。

表 9 朴素贝叶斯分类模型和 SVM 分类器模型的评价指标

单位:%

	NBM	SVM
精确率	0.8750	0.9688
召回率	0.7797	0.9959
F1 Score	0.8279	0.9841

分类决策树是一种十分直观且分类性能很高的分类决策模型。但决策树在长成的过程中极易出现过拟合的情况,导致泛化能力低。在构建和训练分类决策树模型时,文中对构建的分类树进行改进从而防止和解决决策树分类的过拟合问题。图 7 给出了分类决策树对于降水粒子的分类情况。图 7(a)描述了叶子节点所含最小样本数与生成的决策分类树交叉验证误差的关系。训练中选择的叶子节点最小样本数为 28。图 7(b)为构建生成的决策分类树,训练时通过交叉验证、约束最大树深和剪枝的方法控制树的生长防止过拟合。图 7(c)为决策树分类模型对于预测集样本的分类结果。



(a)叶子节点最小样本数与决策分类树交验证误差 (b)生成的决策分类树(剪枝后) (c)决策分类树预测集分类结果

图 7 分类决策树模型的分类情况

决策树使用基于类变量的划分规则创建分类树,可以强制地将不同类别的样本分开。因此,往往在类别不均衡数据上表现不错。根据上述的分类情况可知,训练的决策树分类器对于雪花的平均分类准确率为99.2633%,霰类约为98.6667%,预测集整体平均分类准确率达到99.0013%,训练后的决策分类树的交叉验证误差为0.0286。通过最优化交叉验证误差训练得到的分类决策树获得很好的两类分类结果,且泛化能力良好,见表 10。另外,通过如表 11-12 的计算结果,可以明显看出决策树分类算法的分类精确性、可靠性较另外 2 种算法更高,误分比更低。

表 10 决策树分类器预测集分类平均准确率结果

雪花分类准确率/%	霰分类准确率/%	预测集分类准确率/%	交叉检验误差
99.2633	98.6667	99.0013	0.0286

表 11 决策树分类器预测集结果混淆矩阵

	雪花	霰
雪花	188	2
霰	2	34

表 12 决策树分类器预测集分类评价指标 %

精确率	召回率	F1 Score
1	0.9999	0.9999

4 应用

通过对 2015 年 11 月至 2016 年 2 月冬季(顺义)收集的一次独立的连续性降雪过程进行分类性能验证。这一次过程发生在 2015 年 11 月 22 日,等效液体的积雪量(持续时间累积)为11.4 mm(820 min)[19]。

4.1 2015 年 11 月 22 日案例

图 8 显示了 2015 年 11 月 22 日观测到的 2015-2016 年 1 例冬季降水过程。此次降雪过程在 0700 LST 时,环境温度略高于0 ℃。发展到 0700 到1000 LST 时,温度由0.3 ℃下降到-1.7 ℃,并稳定在-2.5 ℃,平均风速为1.4 ms^{-1}。

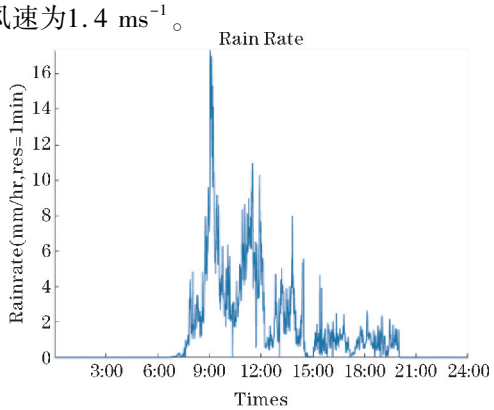


图 8 2015-11-22(0700-1959 LST)小时累积降水量的时间序列

4.2　分类验证结果

3 种分类器模型 20 次交叉验证下的验证集平均分类准确率结果如表 13 ~15 所示。

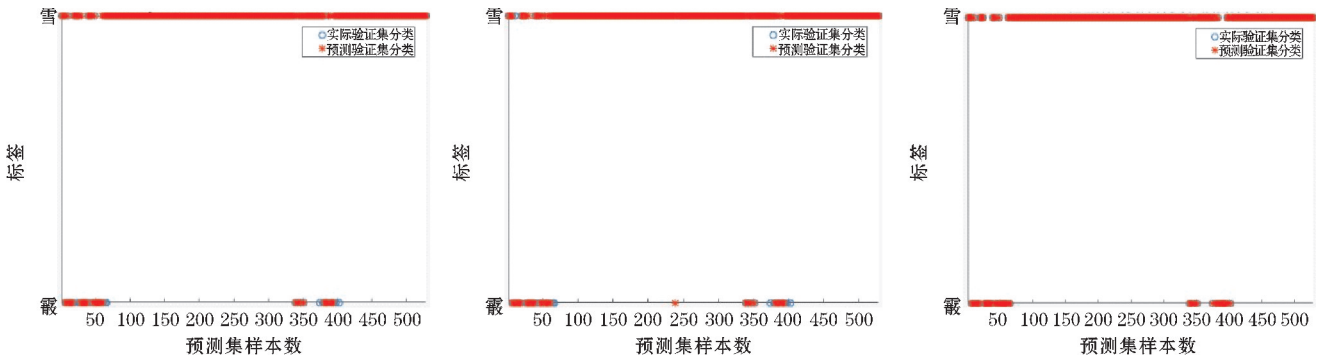


图 9　三种模型分类器验证集分类结果

根据 3 种方法对于独立验证集得到的分类效果可知,朴素贝叶斯方法分类器对于霰类的分类准确率过低,导致模型整体分类准确性较低。SVM 分类器模型较之朴素贝叶斯分类器模型获得更好的验证集分类效果,但对于霰类的分类准确率依然偏低。决策分类树模型对于雪花和霰类均获得不错的分类准确率,对于验证集的分类获得最好的整体分类效果。

表 13　朴素贝叶斯分类器模型验证集分类结果混淆矩阵

	雪花	霰
雪花	456	20
霰	2	50

表 14　SVM 分类器验证集分类结果混淆矩阵

	雪花	霰
雪花	455	13
霰	3	57

表 15　分类决策树分类器验证集分类结果混淆矩阵

	雪花	霰
雪花	453	2
霰	5	68

3 种分类器模型 20 次交叉验证下的验证集平均分类准确率结果如表 16 所示。分类器模型评价指标结果如表 17 所示。

表 16　3 种分类器模型验证集平均分类结果　　单位:%

	朴素贝叶斯	SVM	决策树
雪花分类准确率	99.7817	99.3450	98.4716
霰分类准确率	71.4286	81.4286	97.5714
验证集分类准确率	95.0227	97.9697	98.4427

表 17　3 种分类器模型验证集评价指标计算结果　　单位:%

	朴素贝叶斯	SVM	决策树
精确率	99.7817	0.9722	1
召回率	71.4286	0.9934	0.9857
F1 score	95.0227	0.9827	0.9989

5　结束语

为实现对冬季固态降水粒子的——雪-霰二元自动分类,分别应用朴素贝叶斯分类算法、SVM 和决策分类树 3 种常用的监督式分类算法构建分类器模型。

朴素贝叶斯模型是一种非常简单快速的分类算法,其原理和实现比较简单,学习和预测的效率高。但数据之间属性独立性的条件是朴素贝叶斯分类器的不足之处。在应用朴素贝叶斯算法时采用了普适性较强的正态函数。但由于朴素贝叶斯算法样本相互独立的假设前提在实际样本集中并未满足,对于降水粒子的分类性能较其他两种算法较差。

SVM 能够处理高维输入,比其他监督方法也更不容易过拟合。但 SVM 对缺失数据敏感,数据偏斜情况下的分类会导致分类偏斜,工作需要对模型进行改进。SVM 分类器模型获得不错的分类效果。

决策树分类模型高效且所需的数据准备工作简单,对于样本数不均衡的数据也能够进行比较高效的分类。但在构建时极容易发生过拟合。训练的决策树分类模型经过改进获得十分不错的分类效果。

文中应用不同的算法训练分类模型实现降水粒子的主要类型分类。在实际构建分类器模型时,通过引入正则化系数增加权重处理数据偏斜问题并利用交叉验证方法估算误差并使用独立的未分类的降雪过程数据作为验证集进行进一步验证。

总的来说,针对固态降水粒子分类训练得到的3种分类器模型在预测集和独立的待分类验证集上的总体平均准确率均在95%以上。基于之前相关工作的判断原则,实现的3种粒子分类器模型均获得良好的分类效果。文中3种分类器模型对于雪花的分类准确率均达到97%左右,获得相当好的分类效果;但朴素贝叶斯和SVM分类模型在验证集上对于霰类的分类准确率却明显偏低,分别为71.4286%和81.4286%。因此,由于雪花样本占比较大,朴素贝叶斯模型和SVM分类器模型的整体分类准确率由于大样本的倾斜过于乐观。

根据对3种分类模型的分类性能评估,由于粒子的随机性不适用于概率密度假设的前提,朴素贝叶斯分类模型分类效果相对较差,因此仅考虑作为粒子粗糙分类的测试分类模型。SVM模型受数据偏斜影响,对于小样本的分类准确率不够,需要不断地调整参数及权重,进一步改进。数据样本不均衡的影响对决策树分类器较小,因此训练的决策树模型获得了不错的分类结果。但在构建决策树分类模型时,尽管通过对生成的树按照一定规则进行剪枝以提前停止树的生长,在一定程度上解决了决策树分类模型过拟合问题,但也很难避免过拟合问题发生。尤其是在多类别分类的情形下,构建的决策树经常过拟合严重,稳定性低。因此,在后续即将进行的固态降水粒子多分类工作中,考虑使用集成方法来替代单棵决策树进行。

参考文献:

[1] 杨军,陈宝君,银燕. 云降水物理学[M]. 北京:气象出版社,2011.

[2] Christophe Praz1, Yves Alain Roulet, Alexis Berne. Solid hydrometeor classification and riming degree estimation from pictures collected with a Multi-Angle Snowflake Camera[J]. Atmos. Meas. Tech., 2017,10:1335-1357.

[3] K. Nurzyńska, Mamoru Kubo, Ken ichiro Muramoto. Texture operator for snow particle classification into snowflake and graupel[J]. Atmospheric Research,2012,118:121-132.

[4] Grazioli J, Tuia D, Monhart S, et al. Hydrometeor classification from two-dimensional video disdrometer data[J]. Atmos. Meas. Tech,2014,7:2869-2882.

[5] Dusan S Z, Alexander Ryzhkov, Jerry Straka. Testing a Procedure for Automatic Classification of Hydrometeor Types[J]. Journal of Atmospheric and Oceanic Technology,2011,18:892-913.

[6] 黄敏松. 云降水粒子图像识别方法研究及其在云微物理分析中的应用[D]. 北京:中国科学院大学,2015.

[7] Operating instructions Present Weather Sensor OTT Parsival2[Z]. OTTHydromet GmbH. (OTT Parsival2 用户手册(English)). 2006.

[8] MJ Bartholomew. Two-dimensional Video Disdrometer(VDIS) Instrument Handbook[Z]. Brookhaven National Laboratory,2017.

[9] Huang G, Bringi V N, Cifelli R, et al. A methodology to derive radar reflectivity liquid equivalent snow rate relations using C-band radar and a 2D video disdrometer[J]. J. Atmos. Oceanic Technol, 2010,27:637-651.

[10] Bernauer F, Hürkamp K, Rühm W, et al. On the consistency of 2-D video disdrometers in measuring micro physical parameters of solid precipitation[J]. Atmos. Meas. Tech,2015,8:3251-3261.

[11] 李舰,肖凯. 数据科学中的R语言[M]. 西安:西安交通大学出版社,2015.

[12] Lee J E, S H Jung, H M Park, S, et al. Classification of precipitation types using fall velocity diameter relationships from 2D-videodistrometer measurements[J]. Adv. Atmos. Sci., 2015, 32(9): 1277-1290.

[13] Locatelli J D, P V. Hobbs Fall speeds and masses of solid precipitation particles[J]. J. Geophys. Res., 1974, 79:2185-2197.

[14] Bernauer F, Hürkamp K, Rühm W, et al. Snow event classification with a 2D video disdrometer-A decision tree approach[J]. Atmos. Res., 2016, 172:186-195.

[15] 马刚. 朴素贝叶斯算法的改进与应用[D]. 合肥:安徽大学,2018.

[16] 周志华. 机器学习[M]. 北京:清华大学出版社,2016.

[17] Nicoletta Roberto, Luca Baldini. A Support Vector Machine Hydrometeor Classification Algorithm for Dual-Polarization Radar[J]. Atmosphere, 2017, 8:134.

[18] BenHur A. Weston J A user's guide to support vector machines In Data mining techniques for the life sciences[M]. New Jersey:Humana Press,2010,223-239.

[19] Wen G,H Xiao,H Yang,et al. Characteristics of summer and winter precipitation over northern China[J]. Atmos. Res. ,2017,197:390-406.

Auto-classification of Solid Precipitation Particles based on A 2DVD into Snowflake and Graupel

LIN Huiling[1], XIAO Hui[2,3], YAO Zhendong[1], SUN Yue[2,3], YANG Huiling[2], FENG Qizhen[4], RAO Chen[1]

(1. College of Electronic Engineering,Chengdu University of Information and Technology,Chengdu 610225,China;2. Institute of Atmospheric,Chinese Academy of Sciences,IAP,LACS,Beijing 100029,China;3. Institute of Atmospheric,Chinese Academy of Sciences, Beijing 100049,China;4. Mianyang Flight College, Civil Aviation Flight University, Mianyang 621000,China)

Abstract: Giving an accurate and detailed classification of solid precipitation particles is of a paramount importance to most of the atmospheric processes and the application of weather radar. This paper aims to use two-dimensional optical disdrometer (hereinafter referred to as 2DVD) to measure the precipitation of a single particle, and based on its microphysical parameters and characteristics of precipitation in the course of a minute of the main types of precipitation particles in the unit Interval estimation, this paper classifies the solid precipitation particles. In order to actualize the automatic classification, this paper also attempts to make this task and common classification of machine learning algorithms combined, and the three supervised learning algorithm, naive Bayesian algorithm, support vector machine (SVM),and Decision Tree, applied to classify particles in the unit interval. In this paper, precipitation particles is classified mainly as snowflake and graupel, and its result is tested with the help of Manual detection. Besides,the independent data has been searched to do further examination, proving the accuracy of classification algorithms.

Keywords: solid precipitation particles;2DVD;automatic classification algorithms