DOI:10.13878/j.cnki.jnuist.2018.06.015

史逸民[1]　史达伟[1]　郝玲[1]　张银意[1]　王鹏[1]

基于数据挖掘 CART 算法的区域夏季降水日数分类与预测模型研究

摘要

夏季降水日数的准确预测,对于保障农业、运输业、电力等行业的有序进行具有重要现实意义.利用连云港市气象局提供的 1951—2012 年夏季降水数据对连云港地区的降水日数特征进行分析,难以直观地发现夏季降水日数随时间分布的规律.为进一步探索降水日数的发生规律,结合国家气候中心网站提供的多种气候因子数据,基于 CART 决策树算法构建了连云港地区夏季降水日数是否偏多与是否偏少的分类与预测模型.该模型可以发现在多种气候因子不同条件下,夏季降水日数是否偏多(偏少)的规律,模型的分类与预测都具有良好的效果.利用 52 a 的数据样本训练模型,模型的训练准确率为 90.38%(86.54%),再用剩余 10 a 数据样本检验模型,测试准确率为 80%(80%),并且得到规则集,方便气象业务人员使用以及决策服务人员参考.同时,为降水日数的预测提供了数据挖掘的新思路.

关键词

数据挖掘;CART 算法;降水日数

中图分类号 TP242

文献标志码 A

收稿日期 2016-09-03

资助项目 江苏省科技厅社会发展项目(BE2011720);江苏省气象局预报员专项(JSYBY201612,JSYBY201811);江苏省气象局气象科研基金重点项目(KZ201406);淮河流域气象开放研究基金(HRM201602);连云港市科技支撑项目(SH1634)

作者简介

史逸民,男,工程师,研究方向为短期气候预测.lygsdw@163.com

1　江苏省连云港市气象局,连云港,222006

0　引言

夏季降水的过多或者过少对人类社会的发展而言都是一种气候灾害,而降水日数的多寡,往往对农业病虫害防治、电力部门输电线路的安全管理以及航空运输的安全起降有着重要的影响.汛期降水的短期气候预测不论是基于统计还是模式的准确率仍维持在 60%到 70%左右[1].对于降水日数的研究,往往停留在气候特征的分析上,并且,对降水日数的预测手段方法较少.以往的研究成果主要还是采用模式预测[2]以及统计预测[3]方法,而模式预测往往具有参数复杂、不易获取等特点,统计方法也存在准确率有限等不足.

我国华东地区的夏季降水,往往是由于中小尺度的对流系统造成的,这些尺度较小的系统常常受到大尺度的环流背景场的调制,诸如东亚夏季风、副热带高压以及其他的一些气候系统[4].高辉等[5]研究发现,当前期 ENSO 为暖(冷)位相状态时,则长江流域夏季降水偏多(偏少).西太平洋副热带高压的强度和位置变化是华东地区旱涝的最主要的影响因素.闵锦忠等[6]发现南海、孟加拉湾和阿拉伯海春季海温与夏季长江流域降水呈正相关;黄嘉佑等[7]运用奇异值分解方法发现北半球极涡指数和北半球副高指数对我国夏季降水有一定影响,并讨论了它们之间的具体关系;龚道溢[8]指出北极涛动(AO)指数与梅雨量呈负相关;李自强等[9]研究发现了 QBO 东西位相与华东地区夏季降水的显著关系.此外南方涛动[10]、印缅槽[11]、北太平洋海温[12]、极涡[13]等也影响着夏季降水的变化.

数据挖掘技术是一种基于机器学习的专家系统,其本质是从数据中发现对人们有用的知识和规律,其基本任务是对事物的预测和描述[14].决策树算法是数据挖掘中较为常用的分类与预测算法,相比于神经网络等算法的黑箱式操作及收敛速度慢等特点,决策树算法可以从数据中挖掘出决策规则集,并且计算的复杂度较低,具有较快的收敛速度.目前,决策树算法在气象上的应用越来越广泛.史达伟等[15]利用决策树算法对道路结冰灾害建立了较为准确的分类与预测模型;Zhang 等[16-17]利用决策树算法对台风路径是否转向与台风路径是否登陆建立了较为准确的分类与预测模型.本文将以连云港地区为例,利用数据挖掘技术中的经典的分类与预测算法——CART 算法,

对连云港地区的夏季降水日数进行分类和预测.

本文首先对连云港地区的夏季降水特征进行分析,接着,将连云港地区夏季降水日数是否偏多、是否偏少抽象为两个二元分类问题,以国家气候中心及 NOAA 提供的多个气候因子作为模型的输入变量,利用 CART 决策树算法分析因子与降水日数之间的关系,并运用算法筛选后的因子建立了连云港地区夏季降水日数预测模型,最后对模型的预测效果进行检验.

1 资料与方法

1.1 资料来源

本文采用连云港市气象局提供的 1951—2012 年日降水数据,将缺测值进行了剔除,对有降水的日数标记为一个降水日数.同时,本文采用了国家气候中心及 NOAA 网站下载的多种气候因子数据诸如 ENSO 指数、副高指数,具体如表 1 所示,求得其 6、7、8 三个月的平均值作为夏季值.

1.2 CART 算法

CART 算法又称分类与回归树算法,是数据挖掘中常用的分类预测算法,它是一种二叉树非参数的统计方法,适用于离散型变量和连续型变量的分类.若目标变量是离散型,那么 CART 算法生成分类树;若目标变量是连续型,则 CART 算法生成回归树.本文运用的是 CART 的分类树算法.在分类树的构建中 CART 选择最小 Gini 系数的属性作为测试属性,Gini 系数越小,样本的异质性越小,分割效果越好.

CART 算法首先将数据按升序排序,从小到大以相邻数值的中间值将样本分为两组,然后通过 Gini 系数计算两组样本中输出变量取值异质性:

$$G(t)=1-\sum_{j=1}^{K}p^2(j\mid t), \tag{1}$$

其中,t 为节点,K 为输出变量的类别数,$p(j|t)$ 为节点 t 样本输出变量取 j 的概率.当节点样本为同一类别值时,输出变量取值的差异性最小,Gini 系数为 0,而当各类别概率相等时,输出变量取值差异性最大,Gini 系数也最大,为 $1-1/k$.

CART 算法利用 Gini 系数的减少量描述异质性的下降:

$$\Delta G(t)=G(t)-\frac{N_r}{N}G(t_r)-\frac{N_l}{N}G(t_l), \tag{2}$$

其中,$G(t)$ 和 N 分别为分组前输出变量的 Gini 系数和样本量,$G(t_r)$、N_r 和 $G(t_l)$、N_l 分别为分组后右子树的 Gini 系数、样本量及左子树的 Gini 系数、样本量.

按照这种方式,反复计算便可得到异质性下降最大的分割点,即使 $\Delta G(t)$ 达到最大的组限为当前最佳分割点.

2 连云港地区降水日数时间特征分析

连云港市位于江苏省东北部的黄海之滨,属于雨热同季的温带季风性气候,因此,对其夏季降水的研究具有较高的现实意义.在夏季,年均降水日数 35.22 d,最多的夏季降水日数出现在 1956 年,有 48 d,最少的夏季降水日数出现在 2002 年,夏季降水日数仅有 18 d.

为了研究连云港地区夏季降水日数与降水量的年际变化特征,本文绘制了降水量与降水日数随年份变化的折线图,如图 1 所示.可以看出,连云港地区夏季的降水日数年际分布较为复杂,难以直观地发现其变化规律,因此,从夏季降水日数年变化的角度建立预测模型具有重要意义.

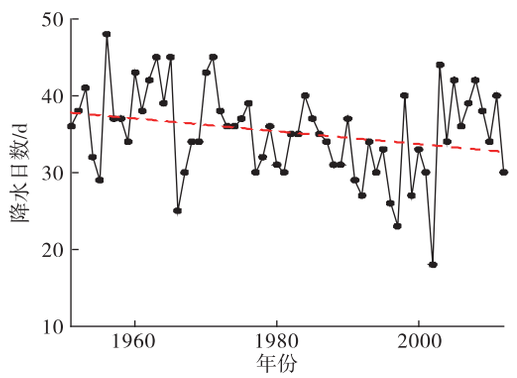


图 1 1951—2012 年连云港降水日数分布状况(虚线为趋势线)

Fig. 1 Distribution of number of precipitation days in Lianyungang during 1951-2012 (dashed line for trend)

从连云港地区夏季降水日数的月分布状况而言(图 2),最大值出现在 7 月,最小值出现在 6 月.降水日数的月分布规律较为简单,容易掌握.

3 基于 CART 决策树的降雨日数分类与预测模型

连云港地区的夏季降水多是由中小尺度天气系统的影响造成的.前文已经阐明,中小尺度引发的降水现象是受到大尺度环流系统调制的,连云港地区的夏季降水也不会例外.因此本文致力于挖掘连云

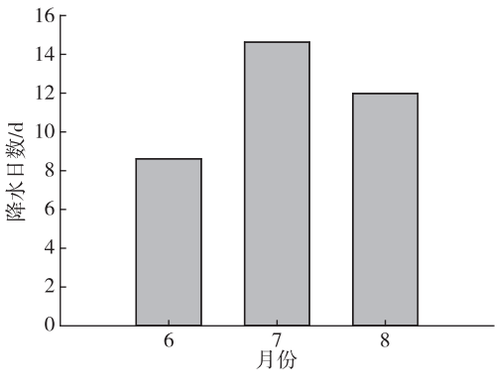


图 2　1951—2012 年连云港月平均降水日数

Fig. 2　Monthly average precipitation days in Lianyungang during 1951-2012

港地区夏季降水日数与大尺度环流系统气候因子间的关系.如表 1 所示,连云港地区夏季降水日数与 Niño4 及欧亚经向环流指数通过了显著性检验.从单一的气候因子角度也难以准确发现连云港地区夏季降水日数的变化规律,那么,能否建立连云港地区夏季降水日数与多种气候因子间的关系呢?

为了进一步探索连云港地区夏季降水日数与本文中所采用的气候因子之间的关系,本文利用 CART 算法,将连云港地区夏季降水日数作为目标变量与各个夏季的气候因子进行了联合建模.

3.1　模型的构建

首先,本文将连云港地区夏季降雨日数偏多(少)的标准定为夏季降水日数平均值加上正(负)0.5 倍的标准差.即当连云港地区某年夏季降水日数≥38.16(≤32.3)时,可以认为连云港地区夏季降水日数偏多(少).接着,利用 CART 算法将 1951—2012 年连云港地区夏季降水日数样本中随机产生 52 a 的样本作为模型的训练集,剩余 10 a 的样本作为模型的测试集,用来验证模型的有效性和鲁棒性.每年夏季的多种气候信号指数作为模型的学习属性,来确定目标变量夏季降水日数"是否偏多",当某年连云港地区夏季降水日数≥38.16(<38.16)时为"是"("否"),即连云港地区夏季降水日数偏多(偏少).在总共 62 a 的夏季降水日数数据中,降水日数偏多年样本为 17 个,降水日数偏少年样本为 19 个,剩余样本为正常年份.通过多次的随机数据建立模型,选取了测试集准确率最高的决策树作为最优决策树模型.

经过 CART 算法的筛选,参与连云港地区夏季降水日数是否偏多模型的属性为太平洋区涡强度指数、北半球副高北界指数、亚欧经向环流指数以及 QBO 指数,最终得到决策树,如图 3 所示.每条从根节点到叶节点的路径代表一条预测连云港地区夏季降水日数是否偏多的规则.以一个叶节点"0(23/1)"为例,括号前的 0 代表夏季降水日数偏少,数字 23 和 1 是这个叶节点中的样本总量为 23,其中有 23-1=22 个正确分类的连云港地区夏季降水日数的样本和 1 个没有正确分类的路径频数的样本.模型的训练(分类)准确率为 90.38%.运用同样的方式,本文又建立了连云港地区夏季降水是否偏少的决策树模型,如图 4 所示.模型的训练(分类)准确率为 86.54%.

表 1　1951—2012 年夏季的多种气候因子与连云港地区夏季降水日数的相关系数

Table 1　Correlation coefficients between various climatic factors and summer precipitation days in Lianyungang during 1951-2012

气候因子	相关系数
Niño1+2	-0.022
Niño3	-0.110
Niño4	-0.271 *
Niño3.4	-0.167
MEI	-0.192
EMI	-0.200
PDO	-0.244
QBO	-0.205
NAO	-0.186
北半球副高北界	0.106
北半球副高面积指数	-0.099
南方涛动指数	0.169
南海副高北界	0.048
南海副高脊线	0.046
欧亚经向环流	0.259 *
欧亚纬向环流	-0.045
太平洋副高脊线	0.028
太平洋副高面积	-0.066
太平洋区涡强度指数	0.150
西太平洋副高北界(110~150°E)	0.198
西太平洋副高脊线(110~150°E)	0.108
西太平洋副高面积指数(110~180°E)	-0.032
亚洲经向环流指数	0.230
亚洲区极涡强度指数	-0.105
亚洲纬向环流指数	-0.168
印缅槽	-0.162
东亚夏季风指数	0.015

注 1:* 表示通过 0.05 显著性水平检验.

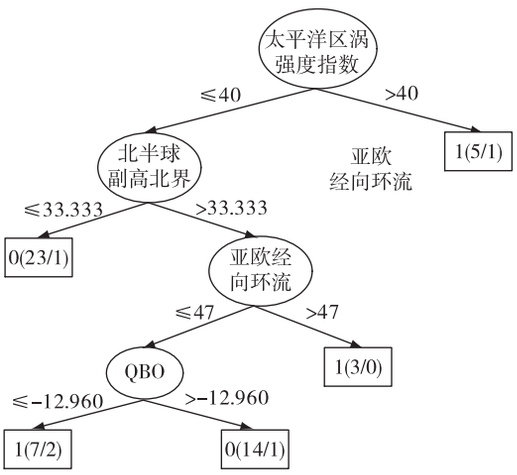


图 3 基于 CART 算法 1951—2012 年连云港地区夏季降水日数是否偏多的决策树模型

Fig. 3 Decision tree to judge positive anomaly of summer precipitation days in Lianyungang during 1951-2012 based on CART algorithm

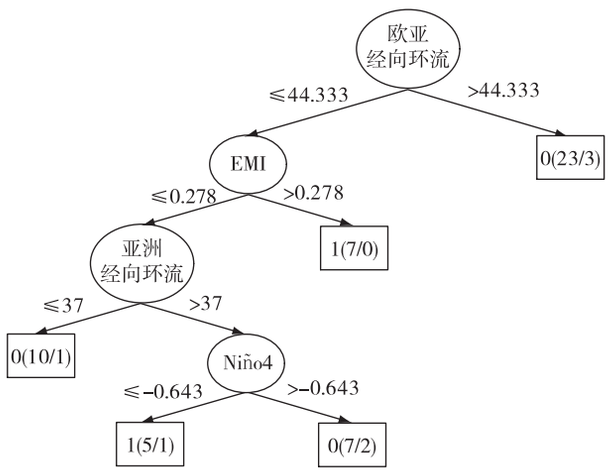


图 4 基于 CART 算法 1951—2012 年连云港地区夏季降水日数是否偏少的决策树模型

Fig. 4 Decision tree to judge negative anomaly of summer precipitation days in Lianyungang during 1951-2012 based on CART algorithm

3.2 模型的验证及决策树规则集

将测试集代入决策树模型进行验证,结果显示,对于夏季降水日数是否偏多验证(预测)准确率为80%;对于夏季降水日数是否偏少验证(预测)准确率同样为80%.通常为了防止模型的过度拟合会对决策树采取剪枝策略,由于参与本次实验计算的样本较少,采取剪枝与否对于实验结果的影响很小.通过对比发现本次实验在没有采取剪枝策略的情况下模型测试的准确率达到最高.

通过对决策树根节点到叶节点的描述,可以总结出预测连云港地区夏季降水日数是否偏多与是否偏少的规则集,结果分别如表 2 与表 3 所示.

表 2 CART 算法发现的预测连云港地区夏季降水日数是否偏多规则集

Table 2 Rule set to predict positive anomaly of summer precipitation days in Lianyungang based on CART algorithm

规则	决策属性	学习准确率
If (太平洋涡强度指数>40) then 降水日数偏多	太平洋涡强度指数	4/5=80%
If (太平洋涡强度指数≤40 and 北半球副高北界指数≤33.333) then 降水日数偏少	太平洋涡强度指数、北半球副高北界指数	22/23=95.7%
If (太平洋涡强度指数≤40 and 北半球副高北界指数> 33.333 and 亚欧经向环流指数> 47) then 降水日数偏多	太平洋涡强度指数、北半球副高北界指数、亚欧经向环流指数	3/3=100%
If (太平洋涡强度指数≤40 and 北半球副高北界指数> 33.333 and 亚欧经向环流指数≤47 and QBO≤-12.960) then 降水日数偏多	太平洋涡强度指数、北半球副高北界指数、亚欧经向环流指数、QBO	5/7=71.4%
If (太平洋涡强度指数≤40 and 北半球副高北界指数> 33.333 and 亚欧经向环流指数≤47 and QBO>-12.960) then 降水日数偏少	太平洋涡强度指数、北半球副高北界指数、亚欧经向环流指数、QBO	13/14=92.9%

表 3 CART 算法发现的预测连云港地区夏季降水日数是否偏少规则集

Table 3 Rule set to predict negative anomaly of summer precipitation days in Lianyungang based on CART algorithm

规则	决策属性	学习准确率
If (欧亚经向环流指数> 44.333) then 降水日数偏少	欧亚经向环流指数	20/23=87.1%
If (欧亚经向环流指数≤44.333 and EMI> 0.278) then 降水日数偏多	欧亚经向环流指数、EMI	7/7=100%
If (欧亚经向环流指数≤44.333 and EMI≤0.278 and 亚洲经向环流指数≤37) then 降水日数偏少	欧亚经向环流指数、EMI、亚洲经向环流指数	9/10=90%
If (欧亚经向环流指数≤44.333 and EMI≤0.278 and 亚洲经向环流指数> 37 and Niño4≤-0.643) then 降水日数偏少	欧亚经向环流指数、EMI、亚洲经向环流指数、Niño4	4/5=80%
If (欧亚经向环流指数≤44.333 and EMI≤0.278 and 亚洲经向环流指数> 37 and Niño4>-0.643) then 降水日数偏多	欧亚经向环流指数、EMI、亚洲经向环流指数、Niño4	5/7=71.4%

从以上的实验结果可以看出,CART 算法对于连云港地区夏季降水日数的异常预测有较好的效果,并且可以得出简约的预测规则集,非常科学易用.CART 算法是数据挖掘中一种经典高效的决策树算法,利用 CART 对夏季降水日数进行研究,也为非线性的分类与预测夏季降水日数提供了一种新的研究思路.

4 总结与讨论

本文首先分析了连云港地区夏季降水日数的特征,发现其年际变化规律较为复杂,并且发展趋势与降水量的发展趋势存在着不一致.为了进一步探索夏季降水日数的规律,本文利用 CART 算法揭示了连云港地区夏季降水日数与各个夏季气候因子间的关系,并得到了规则集,为从事短期气候预测的气象工作人员提供了可以参考的新思路.本文得到了以下结论:1)连云港地区夏季降水日数年际分布特征复杂,难以直观地发现准确的规律,降水日数有着下降的发展趋势;2)连云港地区的夏季降水日数仅与 Niño4 与欧亚经向环流指数取得了显著的相关;3)通过随机抽取将 62 a 的数据分割为建立模型的训练集样本(52 a)与验证模型可靠性的测试集样本(剩余的 10 a),通过 CART 算法对连云港地区的夏季降水日数和各个气候因子联合建立了降水日数是否偏多与是否偏少的分类与预测模型,降水日数是否偏多的模型训练准确率为 90.38%,是否偏少的模型训练准确率为 86.54%,两个模型的验证准确率均为 80%,达到了较好的分类与预测效果.

随着大数据时代的到来,气象数据也越来越多元和海量,数据挖掘技术作为这个时代的“破冰船”,在气象领域的应用也变得越来越广泛.相信随着气象学理论的不断发展,气象数据的不断丰富和积累,数据挖掘技术将会在气象领域发挥出更大的作用.

参考文献

References

[1] 范可,王会军,Choi Y J.一个长江中下游夏季降水的物理统计预测模型[J].科学通报,2007,52(24):2900-2905
FAN Ke,WANG Huijun,CHOI Y J.A physically-based statistical model to forecast summer precipitation in middle and lower reaches of Yangtze River[J].Chinese Science Bulletin,2007,52(24):2900-2905

[2] 刘绿柳,孙林海,廖要明,等.基于 DERF 的 SD 方法预测月降水和极端降水日数[J].应用气象学报,2011,22(1):77-85
LIU Lüliu,SUN Linhai,LIAO Yaoming,et al.Prediction of monthly precipitation and number of extreme precipitation days with statistical downscaling methods based on the monthly dynamical climate model[J].Journal of Applied Meteorological Science,2011,22(1):77-85

[3] 陆文秀,刘丙军,陈俊凡,等.近 50 年来珠江流域降水变化趋势分析[J].自然资源学报,2014,29(1):80-90
LU Wenxiu,LIU Bingjun,CHEN Junfan,et al.Variation trend of precipitation in the Pearl River Basin in recent 50 years[J].Journal of Natural Resources,2014,29(1):80-90

[4] 周秀骥.大气随机动力学与可预报性[J].气象学报,2005,63(5):806-811
ZHOU Xiuji.Atmospheric stochastic dynamics and predictability[J].Acta Meteorologica Sinica,2005,63(5):806-811

[5] 高辉,王永光.ENSO 对中国夏季降水可预测性变化的研究[J].气象学报,2007,65(1):131-137
GAO Hui,WANG Yongguang.On the weakening relationship between summer precipitation in China and ENSO[J].Acta Meteorologica Sinica,2007,65(1):131-137

[6] 闵锦忠,孙照渤,曾刚.南海和印度洋海温异常对东亚大气环流及降水的影响[J].南京气象学院学报,2000,23(4):542-548
MIN Jinzhong,SUN Zhaobo,ZENG Gang.Effect of South China Sea and Indian Ocean SSTA on East Asian circulation and precipitation[J].Journal of Nanjing Institute of Meteorology,2000,23(4):542-548

[7] 黄嘉佑,刘舸,赵昕奕.副高、极涡因子对我国夏季降水的影响[J].大气科学,2004,28(4):517-526.
HUANG Jiayou,LIU Ge,ZHAO Xinyi.The influence of subtropical high indexes and polar vortex indexes on the summertime precipitation in China[J].Chinese Journal of Atmospheric Sciences,2004,28(4):517-526

[8] 龚道溢.北极涛动对东亚夏季降水的预测意义[J].气象,2003,29(6):3-6
GONG Daoyi.Arctic oscillationcs significance for prediction of East Asian summer monsoon rainfall[J].Meteorological Monthly,2003,29(6):3-6

[9] 李自强,马生春.平流层冬季 50 hPa QBO 与长江中下游地区夏季旱涝关系的阶段性[J].气象,1992,18(1):3-7
LI Ziqiang,MA Shengchun.The stage character of relationship between 50 hPa QBO in winter and summer drought/flood trender in the lower and middle reaches of Changjiang River[J].Meteorological Monthly,1992,18(1):3-7

[10] 赵振国,廖荃荪.南方涛动与我国夏季降水[J].气象,1991,17(6):33-37
ZHAO Zhenguo,LIAO Quansun.Southern oscillation and summer precipitation in China[J].Meteorological Monthly,1991,17(6):33-37

[11] 时珍玲.九十年代以来江淮流域夏季典型旱涝成因分析[J].气象,1996,22(9):35-38
SHI Zhenling.The cause analysis of the typical drought and flood years in the area between the Yangtze River

and Huaihe River in summer since 1990 [J]. Meteorological Monthly,1996,22(9):35-38

[12] 张庆云,吕俊梅,杨莲梅,等.夏季中国降水型的年代际变化与大气内部动力过程及外强迫因子关系[J].大气科学,2007,31(6):1290-1300
ZHANG Qingyun,LÜ Junmei,YANG Lianmei,et al.The interdecadal variation of precipitation pattern over China during summer and its relationship with the atmospheric internal dynamic processes and extra-forcing factors[J].Chinese Journal of Atmospheric Sciences,2007,31(6):1290-1300

[13] 王遵娅,丁一汇.夏季亚洲极涡的长期变化对东亚环流和水汽收支的影响[J].地球物理学报,2009,52(1):20-29
WANG Zunya, DING Yihui. Impacts of the long-term change of the summer Asian polar vortex on the circulation system and thewater vapor transport in East Asia[J].Chinese J Geophys,2009,52(1):20-29

[14] Han J,Kamber M.Data mining:Concepts and techniques [M].San Fransisco:Morgan Kaufmann,2006

[15] 史达伟,耿焕同,吉辰,等.基于CART决策树算法的道路结冰预报模型构建及应用[J].气象科学,2015,35(2):204-209
SHI Dawei,GENG Huantong,JI Chen,et al.Construction and application of road icing forecast model based on CART decision tree algorithm[J].Journal of the Meteorological Sciences,2015,35(2):204-209

[16] Zhang W,Leung Y,Chan J C L,et al.The analysis of tropical cyclone tracks in the Western North Pacific through data mining.part Ⅰ:tropical cyclone recurvature [J].Journal of Applied Meteorology and Climatology,2013,52:1394-1416

[17] Zhang W,Leung Y,Chan J C L,et al.The analysis of tropical cyclone tracks in the Western North Pacific through data mining.part Ⅱ:tropical cyclone landfall [J].Journal of Applied Meteorology and Climatology,2013,52:1417-1432

Model prediction of regional summer precipitation days based on CART algorithm

SHI Yimin[1] SHI Dawei[1] HAO Ling[1] ZHANG Yinyi[1] WANG Peng[1]
1 Lianyungang Meteorological Bureau of Jiangsu Province,Lianyungang 222006

Abstract The accurate prediction of the number of summer precipitation days has important practical significance for industries such as agriculture,transportation,and electric power supply.The data of summer precipitation during 1951-2012 provided by Lianyungang Meteorological Bureau were used to analyze the interannual characteristics of summer precipitation days,yet no obvious temporal variation trends were found.Thus a model to predict the regularity of precipitation days is established based on analysis of climate factors listed by National Climate Center website,and CART decision tree algorithm.Year with positive/negative anomalies of summer precipitation days in Lianyungang is defined by various climatic factors,which is trained by sample data of 52 years with training accuracy of 90.38%/86.54%.The remaining data of 10 years are used to test the model,resulting in accuracy of 80% for positive/negative anomalies of summer precipitation days prediction.The rule set is provided for meteorological business and decision-making.

Key words data mining;CART algorithm;number of precipitation days