决策树方法在环境物理量场与暴雨之间关系研究中的应用

张海玲[1,2], 过仲阳[1], 吴健平[1], 林 珲[3]
(1 华东师范大学 教育部地球信息科学实验室,上海 200062;
2 同济大学海洋地质国家重点实验室, 上海 200092;
3 香港中文大学地理系 地球信息科学联合实验室, 香港)

摘要:根据香港天文台提供的2001年8月和2002年6~7月业务区域谱模式(ORSM)物理量场预报值,运用空间数据挖掘中的决策树方法研究了空间环境物理量场变化与暴雨中心降雨量之间的关系。结果表明,研究区暴雨中心降雨量的多少与周边相对湿度,海平面高度,温度和经向风速以及这些物理量场所处的纬度位置关系密切,而与这些物理量场所处的经度位置和纬向风速关系较小。

关键词:决策树; 环境物理量场; 暴雨
中图分类号:P208

1 引言

决策树是空间数据挖掘进行自动分类的方法之一。由于该法以图形化的方式表示数据挖掘结果,浅显易懂,易于做出判断,目前已在遥感影像处理、环境演变、灾害天气预测等方面得到了广泛应用。例如,赵萍等人[1]以南京江宁县为研究区域,根据SPOT卫星影像资料, 运用决策树方法对研究区居民地信息自动提取进行了研究,结果表明,采用该法可以将背景地物类型复杂的江南地区的城镇和村居民地自动提取出来,并且模型受时相影响较小,只是在域值大小上会存在一些差异;论文[2]则运用该法得到了1998年夏季青藏高原上影响中尺度对流系统东移的环境物理量场条件;此外,李飞雪等人[3]将Kohonen神经网络与决策树方法相结合来研究遥感图像的自动分类问题, 分类结果显示,与单一的Kohonen方法相比,两者的结合极大地提高了分类精度;李爽等人[4]将决策树方法与最大似然法在土地覆盖分类中的应用进行了比较,研究结果表明,决策树方法对于输入数据空间特征和分类标志有更好的弹性和鲁棒性;而Mills等人[5]利用澳大利亚区域业务数值天气预报模式的输出数据,运用决策树方法对澳大利亚的雷暴区、强雷暴、龙卷雷暴区进行预报并判断这些雷暴是否可能伴有暴洪、暴流、强风等;强天气目标个例研究表明,就所观测的强天气位置及类型而言,该法的准确度很高。

本文根据香港天文台提供的资料,运用空间数据挖掘中的决策树方法研究了空间环境物理量场变化与暴雨中心降雨量之间的关系,从而为暴雨形成机理研究提供了一种新的方法和思路。

2 决策树应用原理与研究思路方法

决策树是以规则形式对数据进行自动分类[6]。树的根节点是整个数据集合空间,每个分节点是对一个单一变量的测试,该测试将数据集合空间分割成两块或多块。每个叶节点是属于单一类别的记录。决策树的构造分为生长和剪除两个阶段。生长阶段时, 首先将整个训练集作为产生决策树的集合,且训练集每个记录必须是已经分好类的,在此

收稿日期:2004-08-16; **修回日期**: 2005-01-06.
资助项目: 本研究受国家自然科学基金 (40371080)、教育部重点基金(104083)、武汉大学测绘遥感信息工程国家重点实验室(WKL (03) 0103)基金资助。
作者简介:张海玲(1979-), 女,山东人, 华东师范大学地理系地图学与地理信息系统专业硕士生。E-mail: qfzhlsunny@126.com

基础上，寻找初始分裂；在决定哪个属性域作为目前最好的分类指标时，通常的做法是对每个属性域分裂的好坏做出量化，计算出最好的一个分裂。量化的标准是计算每个分裂的信息增益，选择信息增益最大的属性域进行分裂。其次，重复上述步骤，直至每个叶节点内的记录都属于同一类，增长到一棵完整的树。生长过程中，许多分支反映的可能是训练数据中的噪声或孤立点，因此，建树过程中需按照某种规则将相似或相近的分支进行合并，并将其剪除，从而得到一棵能反映数据集特性的决策树。

运用决策树方法探讨空间环境物理量场与暴雨中心暴雨量之间关系的研究思路如图 1：

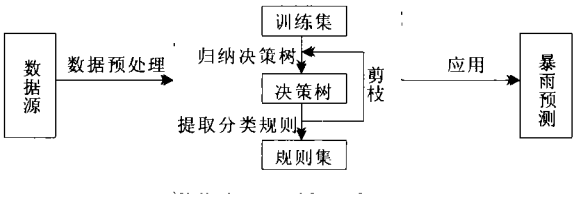

图 1 研究思路

Fig.1 The thinking process of research

(1)对数据源进行数据预处理。根据降雨等级确定暴雨中心，提取周边环境物理量场特征值，形成决策树的训练集。

(2)对训练集进行训练，计算每个环境物理量场参数的信息增益，选择信息增益最大的物理量场参数产生决策树结点，使决策树结点数最少，目的是为了提高识别例子的准确率。在此基础上，由该特征的不同取值建立分支，对该分支的实例子集递归用该方法建立决策树的结点和分支，直到某一子集中例子属于同一类，或没有物理量场参数可再供划分使用，生成初始的决策树。

(3)对初始决策树进行剪枝。将决策树生长过程中不能用来预测新数据的规则和过于细化的规则剪掉。然后由所得的决策树提取分类规则。对从根到树叶的每一条路径创建一个规则，形成规则集。当获取周边环境物理量场参数时，运用决策树所得规则进行分析，从而预测暴雨的发生。

依据香港天文台提供的 2001 年 8 月和 2002 年 6~7 月业务区域谱模式(ORSM)物理量场预报值，采用的参数包括降雨量(A:mm)、位势高度(H:gpm)、温度(T:K)、相对湿度(RH:%)、经向风速(V:m/s)和纬向风速(U:m/s)，观测数据的水平分辨率为 0.5°(经)×0.5°(纬)，时间分辨率为 6 小时；研究范围为 20°~40° N，90°~130°E，研究层次为 400hPa。根据日降雨量的大小，我们将降雨等级分为暴雨(50~100 mm)、大暴雨(100~200mm)和特大暴雨(200mm 以上)三大类。

运用决策树方法研究环境物理量场变化与降雨量之间的关系时，首先根据上述降雨等级确定暴雨中心 (x_0,y_0)，然后以该中心为原点，向四周各外延 3 格(图 2)，得到包括研究中心在内的 49 个格网的环境物理量场值，然后分别求取 5 个环境物理量场特征值 $H=(H_1+H_2+\cdots\cdots+H_m)/n$，$T=(T_1+T_2+\cdots\cdots+T_m)/n$，$RH=(RH_1+RH_2+\cdots\cdots+RH_m)/n$，$V=(V_1+V_2+\cdots\cdots+V_m)/n$，$V=(U_1+U_2+\cdots\cdots+U_m)/n$，$(m,n=49)$，因环境物理量场特征初始值为连续值，需将环境物理量场特征值进行离散化处理，在此基础上，将特征值划分为若个区段，从而构建出决策表，进而运用决策树方法来研究暴雨中心的降雨量与周边环境物理量场之间的关系。

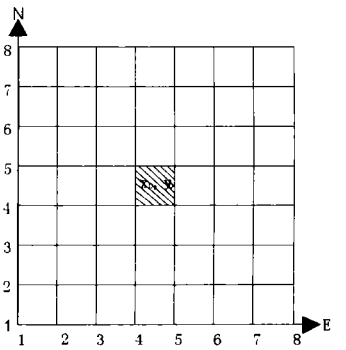

图 2 环境物理量场数据提取示意图

Fig.2 The sketch map of extracting environmental physical field values

3 应用规则分析

本文运用前述方法对 2001 年 8 月和 2002 年 6~7 月研究区的环境物理量场与暴雨之间的关系进行了详细研究，结果如表 1 所示，其中：

LONG 表示经度，LAT 表示纬度，其他物理量场的参数说明如上所述。这里我们以第一条规则为例来说明各规则的含义，该规则表明，在纬度大于等于 21.5° 的范围内，当相对湿度大于或等于 99.9%，温度小于 261.2K，经向风速小于 21.8 m/s 时，研究中心就会产生 74.8mm 的日平均暴雨量。其他规则的含义与此相似。

总的来说，由表可见，在 400hPa 这一层次，在北半球纬度小于 28°的范围内，当经向风速小于 21.8 m/s 时，研究中心形成暴雨，此时的降雨量多少主要受温度影响，且随温度升高而增多；在北半球纬度小于 25.5°的区域内，当位势高度小于 767.3 gpm，温度在 253.3K 与 263.8K 之间，经向风速小于 21.8m/s 时，这些物理量场纬度位置越低，研究中心暴雨量越多，同时随位势高度升高而增加，当位势高度大于 761.3 gpm 时，暴雨量会相应减少；在纬度在 21.5°至 40°的区域内，当温度小于 261.2K，经向风速小于 21.8 m/s 时，相对湿度越大，研究中心暴雨量随之增多。因此研究中心降雨量的多少与相对湿度，位势高度，温度和经向风速以及这些物理量场所处的纬度位置有密切关系，而与这些物理量场所处的经度位置和纬向风速关系较小。

表 1 400hPa 层次的环境物理量场与暴雨之间的关系

Tab.1 The relationship between environmental physical field values at the level of 400hPa and rainstorm

序号	规　　则
1	21.5°≤LAT∧99.9≤R∧T<261.2∧V<21.8→74.8
2	LAT<28°∧T<251.2∧V<21.8→59.7
3	LAT<28°∧251.4≤T<253.3∧V<21.8→65.1
4	LAT<36.5°∧H<757.45∧253.3≤T<263.8∧V<21.8→72.1
5	36.5°≤LAT<25.5°∧H<757.5∧253.3≤T<263.8∧V<21.8→68
6	LAT<25.5°∧757.5≤H<7612.5∧253.3≤T<263.8∧V<21.8→71.4
7	LAT<25.5°∧761.3 ≤ H∧253.3≤T<263.8∧V<21.8→69.4
8	25.5°≤LAT<21.5°∧31≤Long∧T<261.2∧V<21.8→73.9
9	21.5°≤LAT<21°∧R<99.9∧T<261.2∧V<21.8→71.7
10	21°≤LAT∧R<99.9∧T<261.2∧V<21.8→68.3
11	25.5°≤LAT∧H<764∧R<23.2∧261.2≤T∧4.8≤V<21.8→70.3
12	21.5°≤LAT∧23.2≤R ∧261.2≤T∧V<16.4→72.5
13	21.8≤V→103.8

4 结论

本文采用决策树方法研究了环境物理量场变化与暴雨之间的关系，得到了一些有意义的结论，如：暴雨中心周边的位势高度、温度、相对湿度、经向风速、纬向风速等 5 个环境物理量场不同时，中心降雨量会发生明显变化，同时 5 个环境物理量场特征在暴雨中心降雨过程中发挥的作用也各有不同。但是由于受资料所限，挖掘结果具有一定的局限性，特别是在环境物理量场特征值的提取中，仅涉及到研究中心周边 3°以内的区域，并且 5 个环境物理量场各自对暴雨中心降雨量影响的重要程度，也没有定量的表示。因此，利用决策树方法来探讨环境物理量场与暴雨过程之间的关系还需做进一步的研究。在今后的工作中，我们将在收集更多资料的基础上，结合空间数据挖掘中的其他方法，如粗糙集、云理论等对它们的关系做进一步的分析研究。

参考文献

[1] 赵　萍，冯学智，林广发. SPOT 卫星影像居民地信息自动提取的决策树方法研究. 遥感学报，2003，7 (1)：310~314.

[2] 过仲阳，戴晓燕，林　珲. 影响 MCSs 移动的环境物理量场提取. 华东师范大学学报(自然科学版)，2004，3(1)：68~72.

[3] 李飞雪，李满春，赵书河. 基于人工神经网络与决策树结合模型的遥感图像自动分类研究. 遥感信息，2003，3(1)：23~26.

[4] 李　爽，丁圣彦，钱乐祥. 决策树分类法及其在土地覆盖分类中的应用. 遥感技术与应用. 2002，17(1)：7~11.

[5] G A Mills，J R Colquhoun. 强雷暴环境的客观预报：决策树与区域业务数值天气预报模式. Weather and Fororecasting，1998，4(1)：12~16.

[6] Jiawei Han，Micheline Kambr. Data Mining: Concepts and Techniques，北京：高等教育出版社，2001，286~316.

An Application of Decision Tree in Studying the Relationship Between Rainstorm and Environmental Physical Field Values

ZHANG Hailing[1,2], GUO Zhongyang[1], WU Jianping[1], LIN Hui[3]

(1 Laboratory of Geographic Information Science, East China Normal University, Ministry of Education, Shanghai 200062, China;
2 State key Laboratory of Marine Geology, Tongji University, Shanghai 200092, China;
3 Department of Geography & Joint Laboratory for GeoInformation Science, The Chinese University of Hong Kong, China)

Abstract: In this project, the operational regional spectral model (ORSM) environmental physical field forecast values, including geopotential height, temperature, relative humidity, meridional wind component and zonal wind component, provided by astronomical observatory of Hong Kong with rainfall data in August, 2001 and June-July, 2002, are used. The data cover an area of 20°~40°N and 90°~130°E at 0.5 degree resolution and six-hour time resolution at the vertical level of 400hPa. Based on the above-mentioned data, decision tree is used to analyze the relationship between the environmental physical field values and rainstorm. The achieved rules show that at the level of 400hPa, within the scope of latitude 28°,when meridional wind component is below 21.8 m/s, the center researched will form a rainstorm. Therefore, the rainfall changes with temperature and the higher the temperature is, the more the rainfall. In the region of latitude below 25.5°, the geopotential height is less than 767.3 gpm, the temperature ranges between 253.3K and 263.8K, and the meridional wind component is less than 21.8m/s, if the latitude of these physical field values becomes lower and the geopotential height larger, the rainfall will increase. When the geopotential height reaches 761.3 gpm, the rainfall will decrease. While in the region of latitude above or equal to 21.5°, when the temperature is below 261.2K, the meridional wind component is less than 21.8 m/s, the greater the relative humidity is, the more the rainfall of the center researched. In a word, the rainfall of the rainstorm center is related greatly to relative humidity, geopotential height, temperature, meridional wind component peripherally and latitude, but is less related to longitude and zonal wind component.

Key words: decision tree; environmental physical field values; rainstorm