贝叶斯分类法在 MCS 移动路径预测中的应用

郭雅芬，过仲阳，苏君毅，戴晓燕
(华东师范大学 地理信息科学教育部重点实验室，上海 200062)

摘要：长江流域上出现的致洪大暴雨与青藏高原上中尺度对流系统(Mesoscale Convective System，简称 MCS)的东移密切相关。为了揭示高原上 MCS 的移动与其周围环境物理量场之间的关系，本文利用 1998 年 6~8 月的日本地球静止气象卫星(GMS)探测的逐时红外辐射亮温资料(Tbb)，并结合国家气象中心高分辨率有限区域分析预报系统(HLAFS)产品中的数值格点预报值，构建了 MCS 环境场特征值数据库。据在 400hPa 和 500hPa 两个层面上，分别运用朴素贝叶斯分类法对 MCS 周边环境物理量特征数据集进行训练并预测，结果表明，贝叶斯分类法能有效地运用于预测 MCS 的移动路径。

关键词：青藏高原；中尺度对流系统；贝叶斯分类

1 引 言

我国是自然灾害频繁发生的国家，其中由暴雨等灾害性天气造成的洪涝灾害是威胁我国国民经济和人民生命财产安全的主要灾害之一[1]。近年来研究表明[2-6]，长江流域历次出现的致洪大暴雨与青藏高原上中尺度对流系统(MCS)的东移密切相关，因此，探讨 MCS 移动和传播规律对于暴雨等灾害性天气的准确预报具有重要的现实意义。

为探讨暴雨发生的物理成因、提高灾害性天气的预报能力，近年来，一些学者及专家正试图运用空间数据挖掘技术从海量的气象云图数据中提取隐含的、有用的信息和模式。例如，何婧等人[7]利用 1961 年 1 月至 1997 年 12 月云南省气象数据提出了一种基于项目序列集的空间关联规则挖掘算法，并运用这一算法研究了影响云南省年降雨量的因素，得到了较好的结果。方兆宝[8]等人利用 1998 年夏季青藏高原逐时红外遥感云图及高分辨率有限区域数值预报资料，运用空间数据挖掘中的相关分析技术对青藏高原 MCS 的移动与其周围物理量之间的关系进行了研究，表明有 6 个物理量(位势高度、涡度、散度、水汽通量散度、垂直速度、K 指数)与 MCS 形状密切相关。过仲阳[9]等人利用 1998 年夏季日本地球静止气象卫星(GMS)记录的红外辐射亮温资料，运用空间数据挖掘中的决策树方法建立了移出高原的 MCS 与其环境物理量场之间的关系，结果表明，利用该法来预测 MCS 未来的移动路径是切实可行的。

由于影响 MCS 形成及移动的原因十分复杂，为揭示 MCS 的移动与其周围环境物理量之间的密切关系，以预测 MCS 的移动，本文利用 1998 年 6~8 月的日本地球静止气象卫星(GMS)探测的逐时红外辐射亮温资料(Tbb)，并结合国家气象中心高分辨率有限区域分析预报系统(HLAFS)产品中的数值格点预报值，运用数据挖掘技术中的贝叶斯分类方法，对 MCS 的移动方向进行了预测和研究。

2 数据采集、分析与预处理

2.1 数据采集

本文所用数据来源于 1998 年 6~8 月日本地球静止气象卫星(GMS)云顶黑体辐射温度(Tbb)资料以及国家气象中心高分辨率有限区域数值预报系统中的数值格点预报值(HLAFS)资料。其中，Tbb 数据的水平分辨率为 0.5°(经)×0.5°(纬)，时间分辨率为 1h；HLAFS 物理量场格点值的水平分辨率为 1°

收稿日期：2006-01-02；**修回日期**：2006-07-02.
资助项目：国家自然科学基金资助项目(40371080)，教育部重点基金资助项目(104083)，武汉大学测绘遥感信息工程国家重点实验室资助项目(WKL(03)0103)，教育部留学回国人员基金资助。
作者简介：郭雅芬(1981-)，女，山西省平遥人，硕士研究生，主要从事空间数据挖掘的研究。E-mail: girlyafen@sohu.com

(经) ×1 °(纬), 时间分辨率为 12h。此外, 由于 Tbb 资料在高原上东经 80 °以西误差较大, 因此将高原上 MCS 研究范围确定为 27 °~40 °N, 80 °~105 °E; 研究层次为 400hPa 和 500hPa。

本文选取了 HALFS 资料中位势高度(H)、温度(T)、涡度(VOR)、散度(DIV)、水汽通量散度(IFVQ)、垂直速度(W)、假相当位温(θse)、K 指数、相对湿度(RH)共 9 个物理量数据。此外, 研究过程中, 我们所考虑的 MCS 必须符合以下条件: 即 Tbb≤-32℃连续格点个数≥3 个, 生命史≥3 小时的云团。

2.2 数据分析及预处理

(1) 常用的 MCS 追踪方法主要有: 专家目视判别法、面积重叠跟踪法、基于形态特征的模板匹配追踪法和基于最大空间特征相关的追踪法等。前两种方法计算简单, 可操作性强, 故较为常用; 后两种方法计算复杂, 实用性较差。由于专家目视判别法追踪效率低, 且主观性较强, 不适合大范围、长时间的追踪。因此, 本文采用过仲阳等[9]运用的面积重叠比较法对 1998 年夏季青藏高原上 MCS 进行自动追踪, 并规定若干个 MCS 移出 105 °E, 则认为该 MCS 移出了高原。

(2) 根据追踪结果, 将待研究的 MCS 起始重心位置确定在 100 °E 附近, 在提取各物理量特征值时, 当 MCS 处于该位置的时次接近 00UTC 时, 就采用 00UTC 的 HLAFS 值; 反之, 当 MCS 在该位置的时次接近 12UTC 时, 就采用 12UTC 的 HLAFS 值, 其余类推。

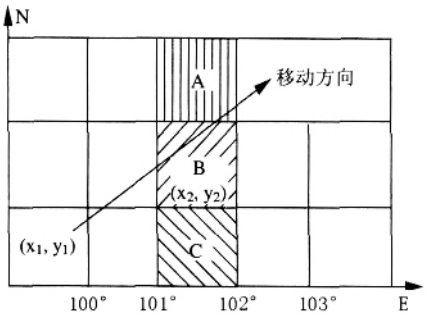

图 1 MCS 环境场平均特征值提取示意图

Fig.1 Diagram showing the abstracting features of MCS environment physical field values

如图 1 所示, 设某个 MCS 在起始中心位置为 (x_1,y_1), 经过 2 小时间隔后其位置为 (x_2,y_2), 现在我们以该点为中心, 取中心点及南北各 1 个纬度的 HLAFS 值, 将其平均得到中心区域 HLAFS 值的平均值 D_b, 然后再在南北方向各外延 3 个纬度, 分别得到该经度上三个区域 HLAFS 值的平均值 D_a, D_c, 最后根据这三个平均值, 求中心区域 HLAFS 值的平均值 D_b 与它们的差, 即 D_b- D_a, D_b- D_c。从而得到能反映 MCS 环境物理量场变化的特征值。

(3) 为便于数据的检索与管理, 本文使用 SQL 数据库软件来构建 MCS 环境物理量数据库, 将 400hPa 和 500hPa 层次上, 各时次 MCS 周边环境场物理量特征值导入数据库中, 以便利用 SQL 强大的数据库管理功能, 以及提供与其他编程软件接口的优势, 进一步运用贝叶斯分类模型对青藏高原上 MCS 的移动路径与周边环境场中各物理量特征值之间的关系进行深入的研究。

3 贝叶斯分类处理与结果分析

3.1 贝叶斯分类的原理与模型应用

(1) 贝叶斯分类的基本原理

贝叶斯分类器是一种基于贝叶斯定理的统计分类器, 它能够预测类别所属的概率, 即一个对象属于某个类别的概率。朴素贝叶斯分类是一种监督分类方法。理论上, 朴素贝叶斯分类的应用前提是样本的属性值独立于样本的分类属性。其训练集由一组数据库记录或样本组成, 样本是由属性值组成的特征向量。此外, 训练样本还有一个类别标记。一个具体的样本形式为: $\{x_1, x_2, \wedge, x_n; c\}$, 其中 x_i, $(1\leq i\leq n)$表示属性值, c 表示类别。

理论上, 与其他分类器相比, 贝叶斯分类器具有较小的错误率, 且其学习效率很高。实际应用中, 由于其所依据的类别独立性假设和缺乏某些数据的准确概率分布, 预测准确率会受到影响, 但仍能取得较好的结果。有关研究结果表明[10], 当数据库中数据量较大, 贝叶斯分类器的分类效果相对较好。

(2) 贝叶斯分类模型的应用

由于环境物理量值为连续量, 则假设其满足高斯分布, 即用高斯密度函数来计算后验概率。运用朴素贝叶斯分类法对 MCS 周边环境物理量特征数据集进行分类的步骤如下:

①特征向量 X 为各物理量的两个特征值, 共 18 个变量。其 A_k 属性为 MCS 的移动方向即(E,

NE, SE)。因此,首先须输入格式为{H1, H2, T1, T2, VOR1, VOR2, DIV1, DIV2, IFVQ1, IFVQ2, W1, W2, θse1, θse2, K1, K2, RH1, RH2: A}的训练样本。

②计算训练样本类别的事前概率 $P(C_i)=s_i/s$, 其中为训练样本集合中类别 C_i 的个数, s 为整个训练样本集合的大小。即将各个移动方向的 MCS 记录的个数统计出来,则各个移动方向的先验分布就是:各个移动方向 MCS 的个数占 MCS 总数的比例。

③计算训练样本各类别的各个特征值的平均值和标准差。

平均值的求法: 先求出类别属性为 A_i 的某个特征值的总和,再求出类别属性为 A_i 的记录总数,总和除以总数即可得到结果。

标准差的求法:

$\sigma=\sqrt{\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2}$, 式中$\bar{x}$为上面所求的平均值, x_i 为训练样本的特征值, n 为同一类别属性的记录集的个数。

④输入验证记录。验证记录的格式和训练样本的格式相同。

⑤计算高斯函数的值。用验证记录中的特征值以及计算出的平均值和标准差代入到公式 $P(x_k|C_i)=g(x_k, \mu_{c_i}, \sigma_{c_i})=\frac{1}{\sqrt{2\pi}\sigma_{c_i}}e^{\frac{(x_k-\mu)^2}{2\sigma_{c_i}^2}}$ 中,计算该函数的值。

⑥计算后验概率。根据公式 $P(X|C_i)=\prod_{k=1}^{n}P(x_k|C_i)$, 把上一步中计算得到的高斯函数值连乘,再根据公式 $P(X|C_i)P(C_i)$, 把所得的连乘的积再乘以各类别的先验概率 $P(C_i)$, 就可得到后验概率。

⑦比较各个类别的后验概率, 选取后验概率最大的类别作为样本的类别。这样,就可以根据环境物理量值来预测 MCS 的移动方向。

以下采用朴素贝叶斯分类方法, 分别运用 400hPa 和 500hPa 上的环境物理量特征值对 MCS 移动方向进行预测。

3.2 分类结果分析

(1) 400hPa 分类结果

根据追踪结果,移出高原且往东(E)、东北(NE)、东南(SE)方向移动的 MCS 个数分别为 41、9、5。研究过程中,首先将这些 MCS 分成两大部分:一部分为训练样本,另一部分用来预测,且样本的提取都按随机的方式进行。为了达到最好的预测效果,我们将这两部分按不同的比例进行了测试。第一次将这两个部分(训练样本、预测样本)的比例设定为 75%、25%; 第二次将这两部分的比例设定为 80%、20%; 第三次将这两部分的比例设定为 85%、15%; 第四次则为 90%、10%。表 1 列出了两部分的比例及所需 MCS 的个数。我们用不同比例的样本进行训练,然后根据训练结果进行预测,进而将预测精度进行比较,以选出最好的比例分配。

表 1 训练样本、预测样本的比例分配

Tab.1 The proportion of training samples and prediction samples

	训练样本个数及比例	预测样本个数及比例
1	41(75%)	14(25%)
2	44 (80%)	11(20%)
3	47(85%)	8(15%)
4	49(90%)	6(10%)

运用朴素贝叶斯分类法,先对训练样本进行训练,再用预测样本来预测,根据实际的移动方向,可以检验其预测的准确率(表 2)。从表中可见,前三次的准确率都大于 85%, 获得了较为满意的结果; 其中第一次的准确率最高,这时训练样本所占的比例为 75%,预测样本的比例为 25%。因此,在 400hPa 上, 运用贝叶斯分类法对 MCS 的移动方向进行预测时,使用 75%的特征数据作为训练样本,可以获得较为满意的预测结果, 同时, 该结果也证实了 MCS 周边的环境物理量场值与 MCS 的移动路径之间存在着密切的关系。

表 2 400hPa 上的预测结果

Tab.2 The prediction results at 400hPa

	预测结果的正确数	准确率
1	13	92.86%
2	10	90.91%
3	7	87.50%
4	5	83.33%

(2) 500hPa 分类结果

类似于 400hPa 高度层, 在 500hPa 层次上对 MCS 移动方向的预测也根据表 1 中训练样本、预测样本的比例, 用朴素贝叶斯分类法进行训练和预测, 表 3 为 500hPa 上预测的准确率。从表中可见, 第二次和第三次的准确率都大于 85%; 其中第二次的准确率最高, 这时训练样本占总数的 80%, 用来预测样本占总数的20%。因而, 在 500hPa 层次上运用朴素贝叶斯分类法对 MCS 的移动方向进行预测时, 使用 80%的数据作为训练样本较为合适。

表 3　500hPa 上的预测结果

Tab.3 The prediction results at 500hPa

	预测结果的正确数	准确率
1	11	78.57%
2	10	90.91%
3	7	87.5%
4	5	83.33%

上述结果表明, 贝叶斯分类能有效地运用于 MCS 移动路径的预测, 且从总体上说, 其预测精度较高。另外, 对同一高度层进行训练和预测时, 训练样本和预测样本的比例不同, 预测的精度也不同, 所以在预测中要按不同的比例进行预测, 选取准确率最高的为划分标准。对于不同高度层预测时, 各层所对应的预测准确率最高的训练样本和预测样本划分的比例是不同的, 具体操作中要针对不同层分别予以讨论。

4 结　论

空间数据挖掘技术已广泛应用于灾害天气的预测, 本文运用朴素贝叶斯分类方法, 对夏季青藏高原上 MCS 的移动方向进行了预测, 得到了令人满意的结果。在 400hPa 高度层上时, 当训练样本为 75%时, 预测的准确率最高, 达到 92.86%; 而在 500hPa 上时, 当训练样本为 80%时, 预测的准确率最高为 90.91%。

由于影响高原上 MCS 移动和传播的因素十分复杂, 进一步运用其他空间数据挖掘方法, 如分类规则、聚类规则、特征规则等来研究影响 MCS 移出高原的环境物理场条件是今后需深入研究的工作。

参考文献

[1] 孙绍聘. 遥感技术在洪涝灾害防治体系建设中的应用. 地理科学进展, 2002, 21(3): 282~288.

[2] 张顺利, 陶诗言, 张庆云, 卫　捷. 长江中下游致洪暴雨的多尺度条件. 科学通报, 2002, 47(6): 467~473.

[3] 王立琨, 陶祖钰. 1998 年长江洪水大暴雨的卫星云图分析. 北京大学学报(自然科学版), 2000, 36(1): 87~94.

[4] 江吉喜, 范梅珠. 夏季青藏高原上的对流云和中尺度对流系统. 大气科学, 2002, 26(2): 263~270.

[5] Tao S Y, Chen L. A review of recent research on the East Asian summer monsoon in China. In: Monsoon Meteorology. Oxford: Oxford University Press, 1987, 60~92.

[6] Zhu Qiangen, He Jinhai, Wang Panxing. A study of circulation difference between East Asian and Indian summer monsoons with their interaction. Advances in Atmospheric Sciences, 1986, 3: 446~477.

[7] 何　婧, 王丽珍, 邹力鹗. 基于云南气象数据的空间关联规则挖掘. 计算机工程与应用, 2003, 34: 187~190.

[8] 方兆宝, 吴立新, 林　珲, 江吉喜, 过仲阳. 面向空间数据挖掘的 MCSs 移动和传播影响因素分析. 热带气象学报, 2004, 20(5): 600~604.

[9] 过仲阳, 戴晓燕, 林　珲, 江吉喜, 黄　签. 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, 1: 67~72.

[10] 朱　明. 数据挖掘. 合肥: 中国科学技术出版社, 2002, 5.

下转 P42

Research of WebGIS Based upon RIA

ZHANG Hong, FENG Jiangfan, LV Guonian, TENG Xuewei
(Jiangsu Provincial Key Lab of GIS, Nanjing Normal University, Nanjing 210097,China)

Abstracts: Geographical Information Systems (GIS) are characterised by the ability to integrate geospatial data from a wide variety of sources. World Wide Web Geographical Information System (WebGIS) is the combination of Internet/Web technologies and Geographical Information System, it makes the Web publishing and sharing of geospatial information in the whole world possible and has exerted great influence on the national economic development and our everyday life. In the application of WebGIS, how to transfer efficiently the geospatial information, such as the user's spatial query submits and the result data from GIS database, through the network is a key factor to determine the capability of the system. This paper firstly analyses the problem existed in WebGIS, and then introduces the Rich Internet Applications, in short RIA, which is a brand- new Web application solution. Considering the trait of RIA, the thought for constructing WebGIS based on RIA is presented, and three key techniques adopted in RIA based WebGIS are summarized and analyzed. After that, this paper describes the framework of WebGIS based on RIA, consisting of five layers: client layer, presentation layer, business layer, integration layer and resource layer. In the design, the realization of client, server and protocol are discussed in detail. Finally, the thought given in this paper is proved feasible and effective with an experimental system, in addition, some future work in WebGIS based on RIA, such as security, interoperability are indicated and simply explored. The authors believe that with more research on WebGIS based on RIA, RIA will play greater role in publishing geospatial information, furthermore, the geospatial information web sharing will become easier.
Key words: WebGIS; geographic information; rich internet application; XML

上接 P23

Application of Bayes Classification in Forecasting the Trajectories of MCS

GUO Yafen, GUO Zhongyang, SU Junyi, DAI Xiaoyan
(Key Laboratory of Geographic Information Science, East China Normal University, Ministry of Education, Shanghai 200062, China)

Abstract: The relationship between the intensive precipitation caused severe flood in the Yangtze River Basin and the trajectories of Mesoscale Convective System (MCS) over the Tibetan Plateau is close. In order to find the relationship between the trajectories of MCS and the environmental physical field values around MCS, GMS (Geostationary Meteorological Satellite) brightness temperature (Tbb) data and High Resolution Limited Area Analysis and Forecasting System (HLAFS) data from June to August in 1998 are used to build the database of MCS environmental physical field feature values. Based on these, Bayes Classification is applied to train and predict the dataset of MCS environmental physical field feature values at the levels of 400hPa and 500hPa, respectively. Consequently, it is proven that Bayes Classification is effective to predict the trajectories of MCS.
Key words: Tibetan Plateau; mesoscale convective system; Bayes classification