

# 基于 C4.5 决策树算法的天气预警系统的手机终端设计

唐慧强, 杭丽娜\*, 范海娟

(南京信息工程大学 信息与控制学院, 南京 210044)

(\*通信作者电子邮箱 hlnhanglina@163.com)

**摘要:**为满足现代社会对气象预警预报服务的需求,研发了 Android 系统平台下实时天气预测和异常天气预警系统。根据决策树算法中的 C4.5 算法,解决天气预警分类问题。该方法通过提取训练样本中最大增益率属性作为属性特征建立决策树,经剪枝后得到天气预警评估的决策树模型,并对此模型进行分析和应用。实验结果表明这种方法在分类评估准确率上具有优势,分类正确率达到 85.8%。

**关键词:** Web Service; 天气预报; 决策树; C4.5 算法; 剪枝; 警报

**中图分类号:** TP301.6 **文献标志码:** A

## Design of mobile phone terminal of weather warning system based on C4.5 decision tree

TANG Huiqiang, HANG Lina\*, FAN Haijuan

(School of Information and Control, Nanjing University of Information Science and Technology, Nanjing Jiangsu 210044, China)

**Abstract:** In order to meet the needs of modern society for weather forecast and early warning service, a real-time weather forecast and abnormal weather early warning system was researched and implemented in the Android system. Based on the decision tree algorithm of C4.5 algorithm, the warning classification problem was resolved. By means of extracting the attributes with maximum gain rate as the features of training sample, a decision tree was built. A model of decision tree was got by the pruning weather warning evaluation and analysis and application were made on this model. The experimental results show that this method has advantages in the assessment of classification accuracy, with correct classification rate up to 85.8%.

**Key words:** Web Service; weather forecast; decision tree; C4.5 algorithm; pruning; alert

## 0 引言

目前在天气预测和气象灾害预警方面的研究取得了显著的成就,在气象服务中发挥了重要的作用,建立了预警共享平台,人们可以通过手机短信、12121 气象服务电话、电视、手机上网以及气象网站等媒介获知气象信息。

决策树算法采用自顶向下的递归方式对数据进行处理,把一个无序、无规则的实例集合归纳成一组树形结构表示的分类规则。决策树的典型算法有 ID3、C4.5、CART 等,而 C4.5 算法严格来说是 ID3 算法的一个改进算法,它继承了 ID3 算法的优点并进行了改进:用信息增益率代替信息增益来选择属性,在树建造过程中进行剪枝,能够完成对连续属性的离散化处理,能够对缺省数据进行处理等<sup>[1-2]</sup>。

在本预警系统中选择了强风、强降雨和高温警报三类异常天气,研究了基于经典决策树算法 C4.5 对异常天气预警的评估实现。预警系统定时从 Web 服务网上调取天气实况及预报资料,解析出的有效数据根据 C4.5 决策树算法生成的规则分类方法自动检索,将今后 4 天的天气预报信息和检索得出的预警信息显示在窗口界面并发出声音警报。

## 1 C4.5 算法的气象预警实现方法

基于 C4.5 决策树算法的预警类别评估步骤如下:

1) 对气象数据训练集  $T$  各项属性数据进行预处理,形成天气预警决策树的训练集。

2) 计算各个属性的信息增益和信息增益率。

C4.5 算法通过信息增益率来选择需要检验的属性,集合  $S$  的熵为:

$$Info(T) = - \sum_{i=1}^k ((freq(C_i, S) / |S|) \times \lg(freq(C_i, S) / |S|)) \quad (1)$$

其中:  $freq(C_i, S)$  代表  $S$  中属于类  $C_i$  ( $k$  个可能的类中的一个) 的样本数 (例如  $S$  中高温警报的样本数),  $|S|$  表示集合  $S$  中的样本数。那么对训练集  $T$  按离散属性  $x$  划分为  $T_1, T_2, \dots, T_n$  的  $n$  个子集, 检验  $x$  的输出可通过相应子集的熵的加权和求取:

$$Info_x(T) = \sum_{i=1}^n ((|T_i| / |T|) \times Info(T_i)) \quad (2)$$

其中,  $T$  为按照属性  $x$  进行分区的集合。属性  $x$  的信息增益率为:

$$GainRatio(x) = Gain(x) / Split\_Info(x) \quad (3)$$

其中:  $Split\_Info(x) = - \sum_{i=1}^n ((|T_i| / |T|) \times \lg(|T_i| / |T|))$ ,  $Gain(x) = Info(T) - Info_x(T)$  表示分区前后的集合的熵的差 (增益)。

3) 挑选具有最高信息增益率的属性 (比如天气状况) 作为决策树的根节点。

4) 在剩下的候选属性 (温度、风力、湿度) 中选择具有最高增益率的属性作为当前分叉节点, 递归直到形成决策树模型。

收稿日期: 2012-11-21; 修回日期: 2013-01-08。

**作者简介:** 唐慧强 (1965 -), 男, 浙江嘉兴人, 教授, 博士生导师, 主要研究方向: 智能仪器及气象仪器、无线传感器网络; 杭丽娜 (1988 -), 女, 江苏丹阳人, 硕士研究生, 主要研究方向: 智能仪器; 范海娟 (1988 -), 女, 江苏如皋人, 硕士研究生, 主要研究方向: 智能仪器。

5)从构造的天气预警决策树中提取分类规则,对新的数据集分类<sup>[3-4]</sup>。

在天气预警决策树数据模型中,高增益率确保高分枝属性不被选取,使得决策树的树形不会因为某个节点分枝过多而过于松散,分枝过多会使决策树依赖于某个属性。在创建决策树的过程中,数据中的噪声和孤立点会引起训练集的分枝异常,C4.5 决策树通过剪枝的方法处理这种数据过分拟合问题,即通过统计度量,剪去最不可靠的分枝。一般剪枝后的决策树在正确的对独立检验数据分类时比未剪枝的决策树更快更好<sup>[5]</sup>。

## 2 天气预警系统手机终端的设计

Android 操作系统是由谷歌公司和开放手机联盟共同提供的开源软件平台,包括一个基于 Linux 内核的操作系统、丰富的用户界面和一些最终用户应用程序,不仅具有非常好的开发调试环境,而且具有各种可扩展的设施。Web Service 是一种基于开放标准的新型分布式应用构件,是基于可扩展标记语言(XML)和以安全为目标的超文本传送协议(Hypertext Transfer Protocol,HTTP)通道的一种服务,主要基于简单对象访问协议(Simple Object Access Protocol,SOAP)和接口定义语言(Web Services Description Language,WSDL)来访问 Web Service,通过 Web Service 内部执行得到所需结果。

系统以 Android 操作系统为开发环境,Eclipse 为开发平台,主要通过第三方的 KSOAP2 库来调用 Web Service,解析处理返回的数据,将数据绑定到对应的应用流程,实现天气信息数据处理;经过 C4.5 算法处理得出预警信息,通过编程实现窗口报警和声音报警。系统结构如图 1 所示。

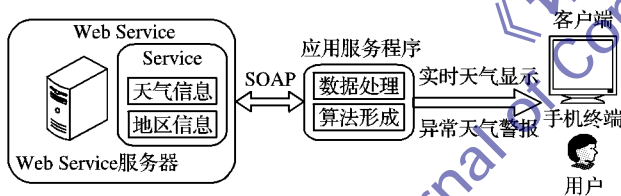


图 1 系统结构

### 2.1 天气预报的实现

系统基于 Java 语言设计实现了基于 C4.5 算法的天气预警系统,其中天气预报模块通过调用 getRegionProvince 方法和 getSupportCityString 方法获得中国省份和城市信息,建立二级联动 Spinner 下拉控件实现地区选择。用户选择的省级市作为调用 Web Service 的参数,根据 getWeather 方法来查询获得今后四天的天气情况等<sup>[6-7]</sup>。用户通过点击 UI 界面右上角的城市按钮,在跳出的二级联动下拉控件里选择需要查询的城市,确定后显示出该城市今后四天的天气情况信息,在该界面上得到查询目标城市的天气预报的天气现象、气温、湿度、风力、空气质量和紫外线强弱等详细信息。系统运行的效果如图 2 所示。



图 2 系统结果

### 2.2 C4.5 算法在预警评估中的应用

用户选择目标城市得到天气预报的详细信息,同时显示

满足预警评估规则的警报窗口和声音警报提醒。为了得出结合 C4.5 算法的警报类别评估规则,下面给出一个训练集,其中经过预处理的数据来自 2012 年夏季南京地区形成警报天气的部分影响数据。训练集  $T$  如表 1 所示。

表 1 训练集  $T$  数据

属性 1 (天气现象)	属性 2 (气温)	属性 3 (风力)	属性 4 (湿度)	警报类别
多云	炎热	微风	偏高	高温
雷阵雨	炎热	和风	偏高	高温
暴雨	热	和风	偏高	强降雨
晴	炎热	微风	正常	高温
晴	炎热	和风	正常	高温
多云	炎热	和风	正常	高温
多云	炎热	微风	正常	高温
晴	炎热	微风	偏高	高温
暴雨	暖	和风	偏高	强降雨
多云	炎热	和风	偏高	高温
雷阵雨	炎热	微风	偏高	高温
雷阵雨	热	和风	偏高	高温
晴	炎热	和风	偏高	高温
暴雨	暖	强风	偏高	强降雨, 强风
雷阵雨	热	强风	偏高	强风
雷阵雨	炎热	和风	正常	高温

训练集  $T$  中,属于“高温”警报的样本有 12 个,属于“强降雨”警报的样本有 3 个,属于“强风”天气的样本有 2 个,根据集合  $S$  的熵的计算式(1)得出分区前训练集的熵为:  $Info(T) = 1.139 \text{ bit}$ 。根据表 1 中的 4 个属性把训练集进行分区,根据  $Info_{x_n}$  检验  $x_n (1 \leq n \leq 4)$  来从相应的属性的离散值中选择其一(例如高温预警类中晴天的个数是 4),属性 1 和属性 3 可将训练集  $T$  分为 4 个子集,属性 2 可将训练集  $T$  分别分为 3 个子集,属性 4 可将训练集分别分为 2 个子集<sup>[8]</sup>,如表 2 所示。

表 2 训练集  $T$  的分区

类别	晴	多云	雷阵雨	暴雨	炎热	热	暖	微风	和风	强风	偏高	正常
高温	4	4	4	0	11	1	0	5	7	0	7	5
强降雨	0	0	0	3	0	1	2	0	2	1	3	0
强风	0	0	1	1	0	1	1	0	0	2	2	0

根据式(2)和(3)计算得出各个属性的信息增益率为:

天气现象:  $GainRatio(x_1) = 0.385$

气温:  $GainRatio(x_2) = 0.606$

风力:  $GainRatio(x_3) = 0.435$

湿度:  $GainRatio(x_4) = 0.209$

从上述公式计算可知属性 2 具有最高信息增益率,因此选取属性 2“气温”用于划分属性。将属性 2 作为决策树的根节点进行首次分区,每个属性值有个分支。首次分区后每个子节点重复首次的选择校验和优化的过程,生成预警决策树,如图 3 所示。

### 2.3 决策树的剪枝

决策树形成后所得到的决策树大而复杂,需要对决策树进行剪枝,直到建立最佳决策树。

对决策树的剪枝包括预剪枝和后剪枝两种方法,预剪枝算法相对简单效率也高,适合解决大规模问题,但缺点是树的生长可能过早停止,因此应用较少,对于过度拟合的树进行后

剪枝在实践应用中更为成功。后剪枝算法是先拟合后化简,首先生成与训练集完全拟合的决策树,然后根据剪枝算法删除树的某一个或几个子树,并用叶节点代替,该叶节点所标示的类别根据子树中占主导地位的实例所属的类来确定。

预警系统中采用后剪枝算法中的 Rule Post-pruning(规则后剪枝)算法,将决策树转化为规则集来简化决策树。规则后剪枝方法可描述如下:

1) 从图3 预警决策树的根节点(天气)到叶节点的每条路径对应一个规则,它的前件(“If”部分)是这条路径上所有属性的合取项,结论(“Then”部分)是叶节点包涵类预测,并

以 If-Then 形式的表示,将决策树转化为等价的规则集<sup>[9-10]</sup>(If“天气晴”and“风力微风”and“湿度偏高”and“气温炎热”,Then“高温预警”)。

2) 对规则集进行评价:对每一个生成规则的相关程度做出评价,如果省略它后该类实例的剩余规则在训练集的分类错误不会增加,则对该规则进行修剪。

3) 根据修剪后规则的估计精度将它们排序,并将这些规则应用于分类后来的实例<sup>[11-12]</sup>。

按照上述步骤对决策树进行剪枝后得到新的预警决策树,如图4所示。

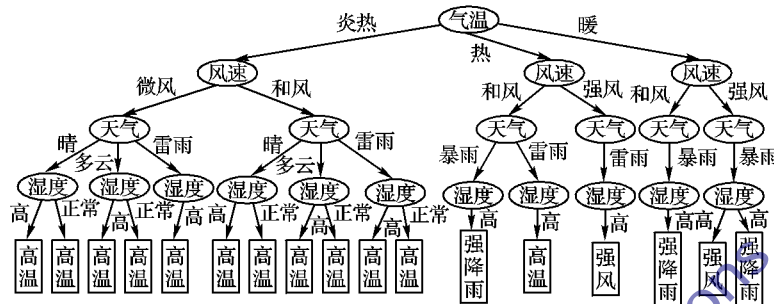


图3 预警决策树

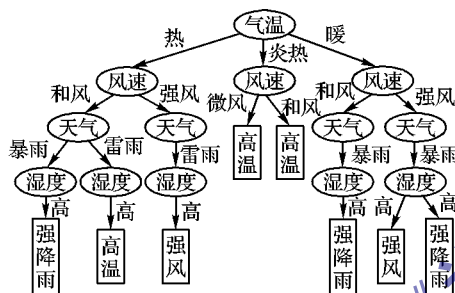


图4 剪枝后的预警决策树

C4.5 剪枝后生成的模型训练规则为:

If“气温炎热”and“风力微风”Then“预警类别为高温”;

If“气温炎热”and“风力和风”Then“预警类别为高温”等。

#### 2.4 预警结果与分析

天气预报预警分类评估子系统定时自动从网上调取天气信息,经过解析得到对应的天气数据信息,根据C4.5决策树算法生成的预警分类方法生成相应的分类规则,从而对天气信息进行评估分类,对达到预警类别条件的天气进行窗口显示和声音警示,警报系统界面如图5所示。

为了验证预警系统分类评估的有效性,评估系统以JBuilder2008R2为开发平台,SQL Server2005为后台数据库,从中随机抽取100条天气数据,利用C4.5算法从这些数据中找出达到天气预警级别的数据,如表3所示。将预处理好的天气信息分为2部分:其中80%的数据为训练集,根据上述方法使用这80%数据的训练集得出预警分类规则,

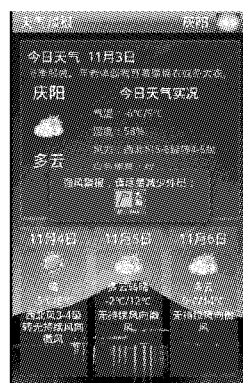


图5 警报显示界面

用剩余的20%数据位测试集估计其准确率。经过7次分组实验得出评估预警类别的分类正确率如图6所示,平均正确率达到85.8%,结果表明,将C4.5算法应用于预警分类具有较高的可信度。

表3 历史数据表

序号	天气现象	气温/℃	风力	湿度/%
1	晴	20~32	4~5级~3~4级	60
2	多云	21~33	3~4级	65
3	多云	22~32	3~4级	80
4	阵雨	22~25	3~4级	80
5	阵雨	21~27	3~4级	90
6	阵雨	21~26	3~4级	85
7	多云	20~30	3~4级~4~5级	75
8	晴	24~35	4~5级~3~4级	70
9	雷阵雨	24~36	3~4级~4~5级	76
10	中雨	24~28	3~4级	80
...	...	...	...	...

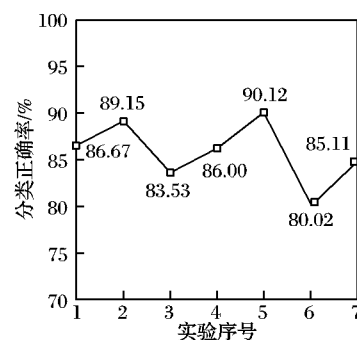


图6 预警分类实验正确率

#### 4 结语

本文在Android系统平台下调用Web Service的天气预警服务信息,并将C4.5决策树算法应用于异常天气预警分类评估。由于城市天气预报及异常天气警报与各个天气要素的错综复杂的关系,需要按照一定的标准和方法来对不同层次区域的天气要素进行定性定量的描述,采用了天气现象、气温、湿度和风力4个属性根据C4.5决策树算法来进行粗略的异常天气预警的分类评定。系统实现了从主动查找到被动接受天气预报信息和警报信息的转变,有利于人们及时获取实时天气信息和异常天气警报信息。但是由于本系统的预警数

(下转第1480页)



### 3.3 算法性能对比

仿真实验中,融合中心总共进行了 200 次关联。取所有目标同时正确关联作为一次标准关联,标准关联次数与总关联次数之比即为关联正确率。

为了测试数据更明显,数据总量定为 80,即 4 个传感器对 20 个目标的航迹关联。表 2 给出了随着观测误差的增加,分别运用 Kohonen 算法<sup>[10]</sup>和本文提出的改进的 Kohonen 神经网络算法的关联正确率的数据结果。

表 2 两种算法正确率比较(数据总量 80) %

观测误差/m	Kohonen 算法	本文算法
50	98.5	100.0
100	96.5	100.0
200	72.0	98.5
400	35.0	95.5
800	31.5	90.0

如果把观测误差定为 100 m,随着测量数据总量的范围分类统计,得到关联正确率见表 3。

表 3 两种算法的正确率比较(观测误差 100 m) %

测量数据总量	Kohonen 算法	本文算法
0 ~ 50	90.5	100.0
50 ~ 150	73.0	98.0
150 ~ 300	45.0	80.5
300 ~ 480	29.5	76.5

仿真结果表明,改进的 Kohonen 神经网络在系统观测误差不断增加的情况下仍然能保持较高的航迹关联正确率。随着测量数据总量的增加,改进的 Kohonen 神经网络算法关联正确率相对 Kohonen 神经网络算法下降不明显,能够有效解决目标密集环境下航迹关联问题。

## 4 结语

本文基于神经网络强大的并行处理和学习能力,提出了一种改进的 Kohonen 神经网络航迹关联算法,解决多传感器多目标航迹关联问题。所提出的改进自组织神经网络利用了神经网络良好的容错性和自适应性,使关联判断随着环境的变化而自适应地变化,避免了过多错、漏关联。同时,聚类算法将所有测量数据作为一个整体进行关联判决,从而使关联速度不会因为传感器和目标数量的增加而过度下降。仿真结

果也验证了方法的可行性和有效性。

本文提出的神经网络算法通过对 *bias* 的适时调整,避免了坏死神经元的出现。在一定范围内比 Kohonen 自组织竞争聚类神经网络在性能上有一定的提高,更好地保证了密集目标环境下航迹关联的正确率。然而,在聚类数据中不免出现个别孤立点,不同程度地影响网络聚类结果。因此,有关于初始聚类中心合理选择上还有待进一步的改善。另外,还需对大量更复杂的目标航迹进行仿真研究,以期将所提出方法应用于现代防空指挥自动化信息融合的实践中。

### 参考文献:

- [1] 王晓伟,杨龙坡,胡军.多站雷达航迹关联的工程实现[J].空军雷达学院学报,2009,10(5):334-336.
- [2] KOSAKA M, MIYAMOTO S, IHARA H. A track correlation algorithm for multisensor integration[C]// Proceedings of the IEEE / AIAA 5th Digital Avionics Systems Conference. Seattle: IEEE Press, 1983:1-8.
- [3] 巴宏欣,赵宗贵,杨飞.多传感器多目标跟踪的 JPDA 算法[J].系统仿真学报,2004,16(7):1563-1566.
- [4] 黄树峰,秦超英.序贯处理的多传感器航迹融合算法研究[J].计算机工程与应用,2010,46(16):42-45.
- [5] ZHOU L, GAO Q, ZOU H L, et al. New optimal track correlation algorithm of three-local node and its applications[J]. Journal of Information and Computational Science, 2011,8(7):1189-1197.
- [6] KAPLAN L M, BLAIR W D, BAR-SHALOM Y. Simulations studies of multisensor track association and fusion methods[C]// Aerospace Conference. Big Sky, Montana: IEEE Press, 2006:1-16.
- [7] DUAN M, LIU J H. Track correlation algorithm based on neural network [C]// 2009 Second International Symposium on Computational Intelligence and Design. Washington, DC: IEEE Computer Society, 2009:181-185.
- [8] 任选宏,李希.模糊技术分布式多传感器系统量测-航迹相关算法[J].火力与指挥控制,2010,35(10):174-176.
- [9] 张池平,崔平远,张英俊.人工神经网络在航迹关联中的应用[J].黑龙江大学自然科学学报,2006,23(1):38-41.
- [10] 林岚,邱晓红.运用自组织神经网络进行多目标跟踪的算法[J].现代雷达,2005,27(1):24-28.
- [11] SHEN F R, OGURA T, HASEGAWA O. An enhanced self-organizing incremental neural network for online unsupervised learning[J]. Neural Networks, 2007,20(8):893-903.
- [12] 郑子扬,陈小惠.基于 Kohonen 神经网络的多传感器数据关联算法[J].华东船舶工业学院学报,2004,10(5):32-37.

(上接第 1469 页)

据全部基于实况,因而具有滞后问题,预警数据的及时性、完整性、正确性和数据采集过程的顺畅等,均是影响预警系统发挥水平的重要原因。

### 参考文献:

- [1] 王黎明.决策树学习及其剪枝算法研究[D].武汉:武汉理工大学,2007.
- [2] 樊建聪,张问银,梁永全.基于贝叶斯方法的决策树分类算法[J].计算机应用,2005,25(12):166-168.
- [3] 徐鹏,林森.基于 C4.5 决策树的流量分类方法[J].软件学报,2009,20(10):2692-2704.
- [4] 陈杰.基于遗传算法的决策树剪枝方法[D].保定:河北大学,2010.
- [5] 张福勇,齐德昱,胡镜林.基于 C4.5 决策树的嵌入型恶意代码检测方法[J].华南理工大学学报:自然科学版,2011,39(5):68-72.
- [6] 孙长征,朱小谦,张卫民.基于 Web 的数值天气预报系统的研究

与设计[J].计算机工程与科学,2009,31(A1):296-299.

- [7] SCHATTEL J L, BUNGE R. The national weather service shares digital forecasts using Web services [J]. Bulletin of the American Meteorological Society, 2008,89(4):449-450.
- [8] 巩固,张虹.决策树算法在天气评估中的应用[J].微计算机信息,2007,23(34):245-247.
- [9] 李会,胡笑梅.决策树中 ID3 算法与 C4.5 算法分析与比较[J].水电能源科学,2008,26(2):129-132,163.
- [10] 张晓龙,骆名剑.基于 IF-THEN 规则的决策树裁剪算法[J].计算机应用,2005,25(9):1986-1988.
- [11] 魏红宁.决策树剪枝方法的比较[J].西南交通大学学报,2005,40(1):44-48.
- [12] 陈杰.基于遗传算法的决策树剪枝方法[D].保定:河北大学,2010.